Linking Latin

Interoperable Lexical Resources in the LiLa Project

Marco C. Passarotti*, Francesco Mambrini*

This paper introduces the overall architecture of the LiLa Knowledge Base, which makes distributed language resources for Latin interoperable on the Web through the application of principles, ontologies and models developed by the Linguistic Linked Open Data community. In particular, the paper focuses on some linguistic aspects of the Latin lexicon that the lexical resources already linked to LiLa allow to investigate, showing how the network of connections that the LiLa Knowledge Base builds between lexical and textual resources for Latin is bigger than the parts considered singularly.

Keywords: Latin, lexicon, lemmatization, Language Resources, linguistic Linked Open Data, interoperability

1. *Introduction: The quest for interoperability of (research) data*

A recent trend that has gained traction in the area of scientific infrastructures is the emphasis on reusability and accessibility of scholarly data. A growing consensus has emerged on a set of principles that are now popularized in the often-quoted acronym FAIR – Findability, Accessibility, Interoperability and Reusability (Wilkinson et al. 2016). One of the purposes behind these guidelines is to overcome obstacles in the discovery and reuse of data, a problem that is particularly urgent, as the current COVID-19 pandemic has proved, in fields like the bio-medical sciences, where an effective and quick access to information is of the essence. Nevertheless, the emphasis to adopt models that lead to more integrated and discoverable digital datasets is gaining momentum in the community of language resources too. In particular, the growing interest in standards for representing linguistic collections as Linked Open Data (LOD) is also a response to the need for more carefully documented and more interconnected data in the field.

Latin and the ecosystem of digital projects of linguistic tools, lexica and corpora dedicated to that language represents a small but compelling example of the importance of such initiatives, as well as of the limitations that they intend to overcome. Over the last decade, the amount and diversity of the (often freely) available resources for Latin has grown exponentially.[1] However, most tools and collections

* Università Cattolica del Sacro Cuore.

1. See Passarotti et al. (2020) for an overview of the currently available language resources for Latin.

of textual or lexical material still live in insulated online environments, such as institutional websites, and are often unknown beyond the circle of the already knowledgeable experts.

Even though discoverability is a serious issue, more damaging still is the lack of interoperability. In the last years the community of Latin language learners and researchers has witnessed the publication, to name just a few interesting resources, of a Latin WordNet in at least two different projects, a series of Latin treebanks (i.e. corpora with word-by-word morphosyntactic annotation), and many other text collections with some forms of linguistic annotation, like lemmatization. However, how would a user leverage the combined power of these datasets to, for instance, discover all the subjects of verbs belonging to a certain WordNet synset? The problem can be readily summarized in the following terms: although digital corpora and lexical resources intuitively deal with the same *entities*, all connections between them exist (if at all) only in the mind of the human user.

The LiLa project was built to answer this very issue, by creating an infrastructure to link potentially all the resources that provide information about the same entities; by taking such steps, the project aims to respond to the challenge of interoperability highlighted by the FAIR best practices. In order to connect all the resources that attach some information to Latin words, LiLa builds a Knowledge Base, meant as a network of structured information about lemmas, the canonical forms that are used (or may potentially be used) by digital language resources to lemmatize word forms or to index dictionary entries.

In this paper we first introduce the model of the LiLa Knowledge Base and its architecture; in the following sections then we focus on some linguistic aspects of the Latin lexicon that the lexical resources already linked to LiLa allow investigating. Finally, we briefly address the question of why and how the whole, i.e. the network of connections that the LiLa architecture builds between those lexical resources and the corpora, is potentially more powerful than a simple sum of its parts.

2. *The LiLa Knowledge Base*

2.1. The role of lemmatization

As was said, an impressive array of digital resources for the study of Latin is currently available over the internet. The most obvious types of datasets in this respect are the digital libraries of Latin texts from all genres, media and periods, including such diverse typologies of documents as Late-Latin legal charters, inscriptions, ecclesiastical, historical and technical treatises, as well as the works of literature from the Classical era. A second group of resources that can be identified includes lexicons, both in the form of retro-digitized editions of printed dictionaries, and of digi-

tal-born databases. A third class includes tools for either automatic linguistic analysis and Natural Language Processing (NLP), or language learning, such as applications for generating exercises on vocabulary or syntactic constructions.

This situation is in fact an ideal use case for applying the paradigm of Linked Open Data. The expression "Linked Open Data" (LOD) points to a set of guidelines for the publication of "smarter" data on the web, which are interlinked through connections that can be semantically queried. Among others, two tenets that are particularly relevant for our discussion are: (1) the prescription to use Uniform Resource Identifiers (URIs), i.e. unambiguous and stable identifiers compliant to a formalized syntax, as the name of the data points; possibly, those URIs should be in the form of HTTP Uniform Resources Locators (URLs) that can be looked up in a web browser; (2) to link data across different data collections, so that information about the same entity from multiple sources may be attainable.

In our particular case of Latin corpora, dictionaries, lexica and NLP tools, all the resources are not only conceptually linked to the same "entities," but they also use comparable steps to identify them. Such "entities" are the words of the Latin lexicon, and the way words are identified in corpora, recognized by NLP tools in their input texts, and indexed in dictionaries is via lemmatization. Lemmas, then, are the ideal candidates to provide links across all the types of language resources, according to the principles of the LOD paradigm.

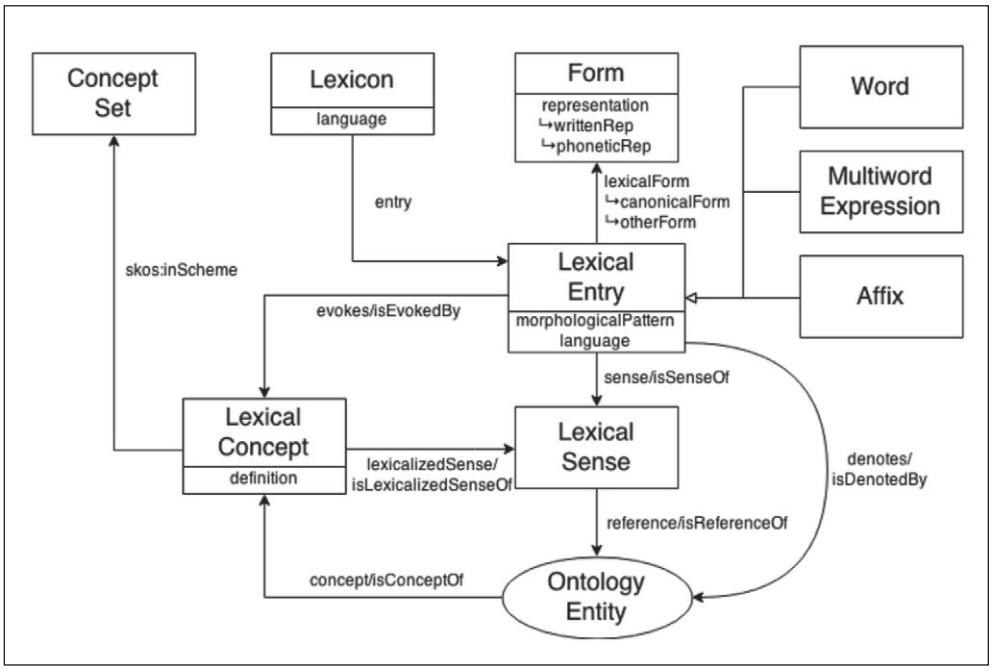
In standard Latin lexicography and corpus annotation, lemmatization is defined as the task of reducing the multiple inflected forms of a word to a form conventionally recognized as canonical. Accordingly, to lemmatize a noun form (e.g., the genitive singular *lupi*) means to reduce it to the nominative singular (*lupus* 'wolf'). Thus, the approach that LiLa adopted in order to connect the different resources is precisely to rely on this process: a corpus with a series of lemmatized tokens, as well as the output of NLP software that includes lemmatization, together with entries in lexicons that are indexed under a lemma, are all making statements about the same objects.

2.2. Form and meaning: LiLa and the OntoLex-Lemon model

While the emphasis on the practical task of lemmatization is peculiar to it, the lexically-based approach of LiLa and its emphasis on the special relation between canonical forms and words is entirely compatible with one of the best established model adopted by the Linguistic LOD community.

The OntoLex-Lemon module (Cimiano et al. 2020: 45-60), developed by the W3C Ontolex Group, has now become the *de-facto* standard for the representation of lexical resources. Figure 1 illustrates how the ontology provides a simple, but sophisticated vocabulary to describe lexical items, such as words, multi-word

Figure 1 The OntoLex-Lemon core model



expressions and affixes. The "Lexical Entry," the central concept in the core model, can be defined in both its formal and semantic properties. In the upper part of the diagram, the entry is in relation with a series of its (inflected) forms, which, in turn, have at least one (or more) written representations and, possibly, a phonetic representation. The semantic aspect of a word can be captured either in terms of the relation of denotation towards an entity defined in a formal ontology or knowledge base (for example, an entry in DBPedia representing a Wikipedia page), or by a reference to an evoked mental concept ("Lexical Concept"). In both cases, as shown in the diagram, the relation between the lexical item and the concept or the entity can be either expressed directly and/or be mediated via a "Lexical Sense."[2]

The OntoLex core model provides a suitable framework for LiLa. In particular, the working hypothesis about lemmatization can be converted into a formal definition that aligns itself with the rest of the classes and properties of the on-

2. See the definition of Lexical Sense in the official documentation at https://www.w3.org/2016/05/ontolex/#lexical-sense-reference.

tology. According to the schema of Figure 1, a lemma is defined as an instance of an OntoLex Form that can be linked to a Lexical Entry via the property "canonical form".

This design choice carries important consequences. To begin with, in OntoLex, a lexical entry cannot be assigned more than one part of speech (POS). Accordingly, if a word is licensed to being used in more than one syntactic function (as, for instance, an adverb or an adjective) and being annotated with different POS, then it must be differentiated into two different lexical entries. Moreover, a lexical entry cannot have more than one canonical form, but canonical forms can have more than one written representation. For Latin, this feature is particularly useful, as it can readily accommodate multiple variant and non-standard spellings of a wordform, which, in the case of a language with more than two millennia of written attestations, are particularly abundant. Thus, for instance, we can attribute to the lemma of the adjective *exspes* 'without hope' both the quoted spelling and the variant *expes*.[3] In the OntoLex ontology, however, written representations are modeled as data properties, i.e. properties that link resources to data values like strings or numbers; data properties do not point to other resources, and therefore cannot become in turn subjects of other statements. As a consequence, written representations cannot be assigned any other property, and it is impossible, within the current version of OntoLex, to make statements about them, such as in which testimonia a given variant spelling is attested, from what date or place, or how many occurrences of each of the variants are documented.

2.3. The Lemma Bank

The backbone of the network of resources in LiLa is made of a set of lemmas (called Lemma Bank) that is sufficiently large as to allow for all resources that deal with any kind of Latin texts or lexical collections to identify the forms used for lemmatization. According to the principles of LOD, the lemmas in the LiLa Lemma Bank are all identified by a unique identifier, which complies to the format of URIs. Moreover, each of them is described by a series of features and a series of relations that are formalized in the dedicated LiLa ontology.[4]

Among the linguistic features attached to lemmas, a special importance is given to the POS. As said, whenever a form is susceptible of multiple interpretations in terms of POS assignment, the solution within the OntoLex-Lemon model is to distinguish as many lexical entries as the POS concerned and, therefore, as many ca-

3. See http://lila-erc.eu/data/id/lemma/102584.
4. See https://lila-erc.eu/ontologies/lila/.

nonical forms. Accordingly, for instance, LiLa has three lemmas with written representation *cum* 'with, along, as', corresponding to the preposition, the adverb and the conjunction.[5] Other features includes the relevant morphological tags (e.g. gender and number for nouns) and the verbal or nominal inflection type, according to the definitions of traditional grammars.[6]

In some cases, deciding whether the orthographic and morphological variation related to a single lemma or multiple instances, each with its own URI, proved more challenging. Purely orthographic variations of the canonical form, that do not modify even a single trait of the morphological analysis, as in the case of *expes/exspes* quoted above, clearly entail a single lemma with multiple written representations. Whenever the variation brings about also a different morphological interpretation or a change in the inflectional category, on the other hand, we decided to create distinct instances. This is often the case with verbs attested with either a deponent or an active inflection, such as *somnio* and *somnior* 'to dream'.[7]

By applying these criteria, we generated the Lemma Bank of LiLa out of the lexical base provided by the database of the morphological analyzer LEMLAT 3.0 (Passarotti et al. 2017). As the software includes independent word lists targeted to the analysis of Classical Latin, Medieval Latin and proper names respectively, a considerable amount of repeated lemmas had to be identified and collapsed under a singular item.

Currently, the LiLa lemma bank includes 196,365 canonical forms, with a total of 232,340 written representations, ready to be linked to lexical resources or lemmatized texts.

3. *Lexical resources in LiLa*

At the moment of writing, six lexical resources are connected to the Lemma Bank of the LiLa Knowledge Base. Table 1 provides an overview of them. Although their coverage in terms of Latin lexical entries is variable, and in some cases quite low, these resources account for a rather wide spectrum of lexical and semantic phenomena.

The following subsections discuss how the linguistic aspects that each of the

5. See respectively http://lila-erc.eu/data/id/lemma/97201, http://lila-erc.eu/data/id/lemma/97207, and http://lila-erc.eu/data/id/lemma/97202.

6. See for instance the definition for the first verbal conjugation in the LiLa ontology at: http://lila-erc.eu/ontologies/lila/v1r.

7. See http://lila-erc.eu/data/id/lemma/125124 (*somnio*), and http://lila-erc.eu/data/id/lemma/125123 (*somnior*). For a more detailed discussion of the different classes of lemmas and of the properties linking them in the LiLa ontology see Passarotti et al. (2020).

Table 1 Lexical resources currently in LiLa

Title	Content	Status	Tot Entries
WFL	Word formation and derivation	Completed	36,138
Brill EDLIL	Etymology (I.-E. and Proto-Italic)	Completed	1,452
IGVLL	Etymology (Greek loan words)	Completed	1,759
Latin Affectus	Polarity	Ongoing	1,998
Latin WordNet	Word senses and synsets	Ongoing	1,424
Vallex 2.0	Valency lexicon	Ongoing	1,064

lexical resources currently in LiLa attempts to describe are represented by applying the LOD principles and the Semantic Web ontologies that were chosen to model the data.

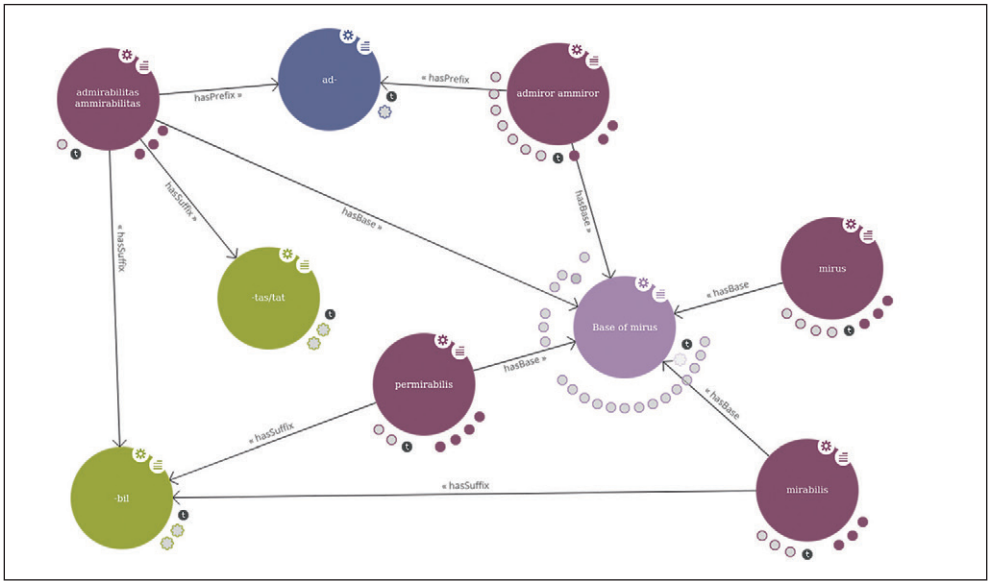
3.1. Word formation

Information on how Latin words are formed and are analyzable in terms of derivational processes is linked to the LiLa Knowledge Base in two different forms (Litta et al. 2019, Litta et al. 2020). The data used in both representations come from the Word Formation Latin (WFL) lexicon, a database where Latin words are described (and related to each other) in connection with word-formation rules. Following a step-by-step morphotactic approach, each process of word formation is regarded as the application of one rule (Litta 2018).

On the one hand, information on derivation is already attached to the canonical forms stored in the LiLa Lemma Bank. A total of 36,250 lemmas from the collection are linked to two special classes of morphemes that are recognizable in their derivational process. Affixes, further sub-specified as either prefixes or suffixes, are connected to forms where each of them is identifiable at any step in the derivational history of the word, so that, for instance, the prefix *per-* links forms such as *pernobilis* 'very famous', *perueho* 'to convey (through)', but also *imperfectus* 'imperfect'.[8] Lexical bases, on the other hand, are those morphemes that are left once all the affixes have been removed, and correspond to the lexical element that is shared by all the derivational family: so, for instance, the base of *ueho* 'to

8. For the prefix *per-* see http://lila-erc.eu/data/id/prefix/14, where all the 843 connected lemmas in the Lemma Bank are also listed.

Figure 2 Affixes, bases, and lemmas in the LiLa Lemma Bank



transport' links lemmas like *perueho*, *conuector* 'one who carries', or *inuecticius* 'imported, exotic'.[9]

The result of this representation is a network of derivational information like the one shown in Figure 2, which represents a lexical base surrounded by a series of connected canonical forms, together with two suffixes (*-bil* and *-tas/tat*) and one prefix (*ad-*) that are involved in the formation of the connected forms.

The output-oriented and descriptive model adopted in the LiLa Lemma Bank does not include any information on derivation processes (in terms of both word formation rules and order of their application), in accordance with the paradigm of Construction Morphology (Booij 2010, Litta, et al. 2020). At the same time, the LiLa Knowledge Base leverages the OntoLex ontology, with the help of some classes taken from its Morph extension that is currently under development (Klimek et al. 2019), in order to link also the entries and the word formation rules as represent-

9. See http://lila-erc.eu/data/id/base/134, with the 104 lemmas connected. Note that, although the OntoLex-Lemon ontology allows representing the morphemes as regular lexical entries with their own canonical form, we did not adopt this representation. Indeed, canonical forms of lexical entries *must* have at least one written representation, but, at the current stage of the work, we are not sure whether lexical bases comply to this constraint, as it is disputed which canonical form is to assign to lexical bases (a root? a stem?). Affixes and bases are therefore independent concepts of the LiLa ontology, not linked to OntoLex. In particular, lexical bases are just used as connectors between the lemmas that belong to the same derivational family in the Lemma Bank.

ed and applied WFL. In such representation, the LiLa lemmas are linked (via the OntoLex property "canonical form") to the lexical entries of the WFL resource. In their turn, each of these entries can be the source (input) and/or the target (output) of a word-formation relation, which is linked to a word-formation rule. In WFL 239 rule types are defined, distinguishing compounding from derivational rules, which are in turn sub-specified as suffixation, prefixation and conversion. In the LOD representation of WFL, classes of rules are described also in terms of POS of their input and output, such as for instance a suffixation rule that outputs an adjective from a verb.[10] To go back to the example mentioned above, the canonical form *imperfectus* from the LiLa Lemma Bank is linked to its lexical entry in WFL, which is, in turn, put in relation with both the verb *perficio* 'accomplish' and with *imperfectio* 'imperfection'. With the former, *imperfectus* is the output of a verb(participle)-to-adjective rule involving the negative prefix *in-*. With the latter, the relation is produced by a rule of the type adjective to noun that involves the suffix *-(t)io(n)*.[11]

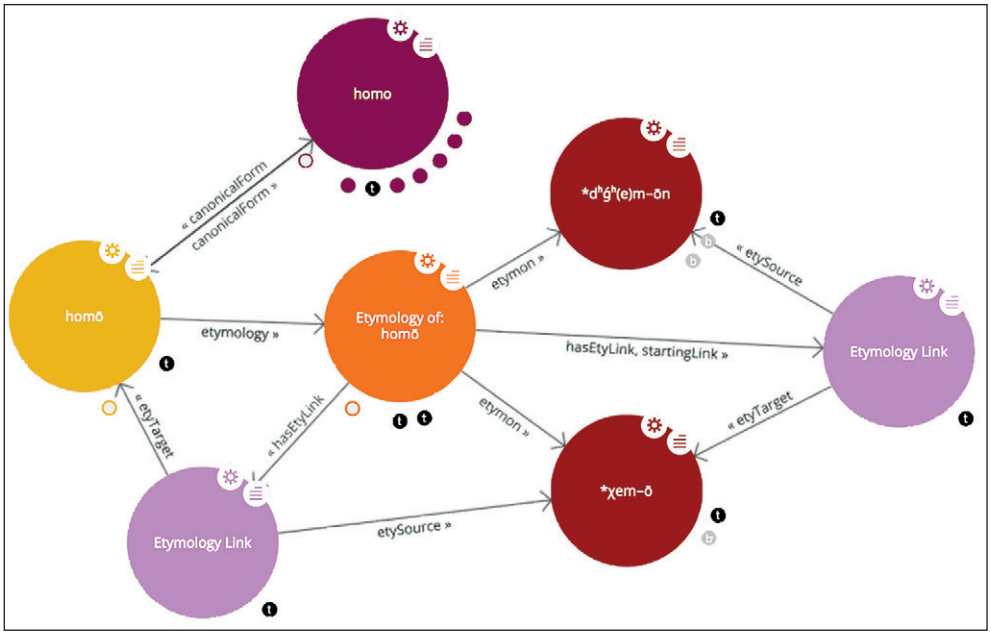
3.2. Etymology

The lemonEty ontology (Khan 2018) extends the OntoLex-Lemon model with classes and properties to express the etymological relations between words and forms. The module introduces a special sub-class of the OntoLex Lexical Entry called "Etymon", which includes all those lexical items that are used to discuss an etymological hypothesis, and that generally belong to a different language or a different diachronic phase as the entry whose etymology is being discussed. Reconstructed Indo-European words or borrowed terms from neighbor languages in an etymological dictionary of Latin are all possible examples of etymons. Etymologies are also defined as resources (in the technical sense that they are entities provided with a URI and which can become subjects or objects of statements). Instances of the class Etymology reify a scientific hypothesis about the origin of an entry and consist of a set of "etymology links" that connect a source to a target. One special advantage of this modeling strategy is the fact that both reified etymological hypotheses and links can be assigned any type of descriptive properties, from a bibliographical reference, to possible truth values. The full sequence of the argumentative steps on which the etymology relies can also be expressed, using a formalism such as the CRM_{inf} (Stead et al. 2019; Mambrini and Passarotti 2020).

Etymology links can be further specified in terms of the relation type that they postulate between a source word and a target. The prototypical instances are inheritance relation from an ancestor language or borrowing. As a matter of fact, LiLa

10. See http://lila-erc.eu/ontologies/lila/wfl/Suffixation/VerbToAdjective.
11. The WFL lexicon in LiLa can be accessed at https://lila-erc.eu/data/lexicalResources/WFL/Lexicon.

Figure 3 Etymology of *homo* in LiLa (according to de Vaan 2008)



makes use of both types of links to express the etymological hypotheses advanced in two lexical resources that are connected to the Knowledge Base.

The entries of the *Etymological Dictionary of Latin and the other Italic Languages* (de Vaan 2008) are all connected to etymologies that encompass a series of links to Proto-Indo-European and Proto-Italic source etymons. Figure 3 represents the etymology of *homo* 'man, human being' in LiLa, as reconstructed by de Vaan (2008).[12] The reified etymological hypothesis is represented by the node at the center of the picture ("Etymology of: homo"); the etymology connects the lexical entry ("homō") to a chain of etymological links (the red nodes) that go from the Proto-Indo-European reconstructed ancestor **dʰǵʰ(e)m-ōn* back to the Latin word via the properties etySource and etyTarget.

The retro-digitized *Index Graecorum Vocabulorum in Linguam Latinam Translatorum* (IGVLL, Saalfeld 1874) integrates these data with a list of loan words from Ancient Greek. In this case too, we chose to model the information with the lemon-

12. See http://lila-erc.eu/data/lexicalResources/BrillEDL/id/etymology/116. The *Etymological Dictionary* by de Vaaan can be accessed in LiLa at https://lila-erc.eu/data/lexicalResources/BrillEDL/Lexicon. Note that LiLa does *not* include a full version of the printed dictionary, but only the etymological links between the Latin words and the I.-E. and Proto-Italic etymons. The lexical entries are linked to their pages on the website of the publisher, so that subscribing readers can access the full text of the dictionary.

Ety extension of the OntoLex core ontology. The lexical entries of IGVLL are also linked to reified etymologies, which consist of one single etymology link from the Greek to the Latin word.[13]

3.3. Polarity

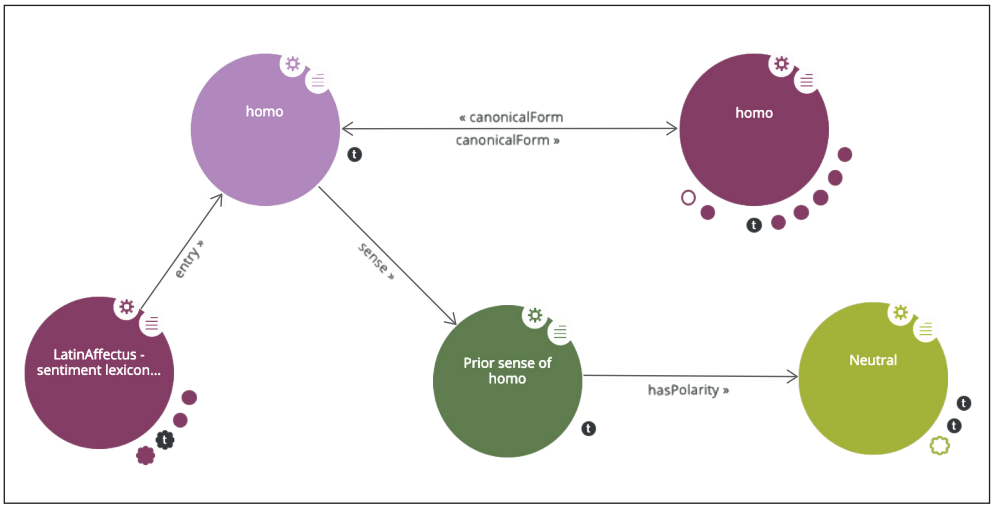
As shown in Figure 1, the OntoLex-Lemon model provides a flexible set of properties and classes to describe the plurality of senses and meanings of a word. Whether the lexical entry is set in relation to an evoked mental concept or a denoted entity, these relations can be either direct and/or mediated through a lexical sense.

The *LatinAffectus - sentiment lexicon for Latin* is a lexical resource that records the prior polarity of a selection of Latin adjectives and nouns (Sprugnoli et al. 2020a). By "prior polarity" we intend the positive or negative value associated to an item in the lexicon of a language, independently from the actual usages in context. Therefore, the polarity value is attached to a single, general sense of a word, and it is measured on a scale of five scores: -1, -0.5 (negative pole), 0 (neuter), +0.5, +1 (positive).

The scores were originally assigned manually by experts working independently, whose annotation underwent an extensive reconciliation phase, then extended with information from derivational morphology (Sprugnoli et al. 2020b). Further iterations of manual annotation and reconciliation are in progress.

Figure 4 shows how the polarity values provided by LatinAffectus are repre-

Figure 4 Polarity of *homo* from LatinAffectus in LiLa



13. The IGVLL lexicon in LiLa can be accessed at https://lila-erc.eu/data/lexicalResources/IGVLL/Lexicon.

sented in LiLa. In particular, Figure 4 shows the polarity of the noun *homo*. The word does not carry any *a-priori* positive or negative connotations, and is therefore recognized as neutral (score of 0). The node for the lexical entry in LatinAffectus is linked to the lemma *homo* in the Lemma Bank through the property "canonical form" and to its prior sense via the property sense. In turn, the prior sense of *homo* is linked to its polarity value (Neutral) via the property "has polarity."[14]

3.4. Senses, synonyms and valency

WordNet is a lexical database of English that groups certain categories of words (nouns, adjectives, verbs and adverbs) into sets of cognitive synonyms known as "synsets" (Fellbaum 1998). Although originally developed for English, several projects have extended the application of the synsets to the lexicons of many more languages (Pianta et al. 2002; Bond and Foster 2013). In 2004, Minozzi (2017) created a Latin WordNet with a total 9,378 lemmas, spread across 8,973 synsets, that where automatically classified using the Italian and English WordNet and bilingual dictionaries to match the Latin words. This dataset represents a foundational resource, but its usefulness is limited by a series of shortcomings, such as the arbitrary selection of the included lemmas and the existence of wrong connections to synsets inherited from English (Franzini et al. 2019). More recently, a larger Latin WordNet including more than 70,000 entries, has been developed, by following the same automatic procedure as the one built by Minozzi.[15] The precision and recall of the synset assignment of this Latin WordNet still has to be assessed.

The English WordNet is also one of the largest datasets that were converted and distributed as LOD.[16] Particularly, an official RDF version of the Princeton WordNet is available, which uses OntoLex-Lemon to model the relations between words, senses and synsets (McCrae et al. 2014; Cimiano et al. 2020: 215–28). The synset is there interpreted as an OntoLex Lexical Concept, i.e. as an "abstraction, concept, or unit of thought that can be lexicalized by a given collection of senses."[17]

Starting from Minozzi's Latin WordNet and the RDF Princeton distribution, the LiLa team has worked on two different tasks. Firstly, we decided to revise manually as many lemma-synset associations from the available Latin WordNet as possible, in order to correct the instances of misalignment (precision) and to integrate the senses established in Latin lexicography that were not represented in the origi-

14. The LatinAffectus lexicon in LiLa can be accessed at https://lila-erc.eu/data/lexicalResources/LatinAffectus/Lexicon.
15. See Short in this volume.
16. See Cimiano et al. (2020: 217) for a history and an overview of the different projects dealing with the publication of the WordNet(s) as LOD.
17. https://www.w3.org/2016/05/ontolex/#lexical-concept.

nal version (recall) (Franzini et al. 2019). Secondly, our goal was to publish this refined resource as LOD, following the model of the RDF WordNet closely. Since this double effort goes on in parallel, the published LOD version of the Latin WordNet[18] now includes 1,424 lexical entries, distributed among 5,220 synsets.[19]

Following the OntoLex-Lemon model (see Figure 1), the relation between a word and a synset is mediated through a lexical sense. A second resource for Latin that is being actively developed for inclusion into the LiLa network also draws on the list of word senses associated with the entries of the Latin WordNet. Passarotti et al. (2016) built the first version of a valency lexicon, named Latin Vallex, on the evidence of the syntactic annotation from two Latin treebanks, namely the *Index Thomisticus* Treebank (Passarotti 2019), and the Latin Dependency Treebank (Bamman and Crane 2006). All valency-capable lemmas occurring in the semantically annotated portion of the two treebanks are assigned one lexical entry and one valency frame in Latin Vallex.

The structure of Latin Vallex is closely modeled on that of the Czech PDT-VALLEX (Hajič et al. 2003). Each entry of the lexicon consists of a sequence of frame entries that contain each a sequence of frame slots corresponding to the arguments of the given lemma. Each frame slot is assigned a semantic role labeled with the same tags used for the semantic annotation of the Prague Dependency Treebank (Mikulová et al. 2006). In the current stage of the work, in order to enhance the coverage of the Latin Vallex, the process of creation of the valency frames is running independently from the treebank annotation and is fully intuition-based. The task is currently being performed manually: the valency frames included in the first version of Latin Vallex have been updated, cleaned or rectified. Currently, 1,064 lexical entries have been annotated, for a total of 8,327 valency frames.

Valency frames are strictly linked to senses: for each recognized sense of a valency-capable word, a frame is established intuitively, and assigned the set of its obligatory complements. The senses to be annotated are taken directly from the repertoire of word senses in the Latin WordNet; thus, each entry-synset pair for the valency-capable words in the Latin WordNet is annotated (or will be annotated, once the work is completed) with a valency frame.

As the core module of OntoLex is not sufficiently expressive to capture the predicate structure of a lexical entry, we have adopted the PreMOn extension to model the information in the Latin Vallex and to map the entities to other schemas such as the Latin WordNet (Corcoglioniti et al. 2016). The property and classes that are

18. See http://lila-erc.eu/data/lexicalResources/LatinWordNet/Lexicon.

19. Note that the LiLa dataset also includes all relations between synsets that are stipulated in the Princeton WordNet (like antonymy, hypernymy and hyponymy). In total, the LiLa LatinWordNet provides information on 22,742 synsets.

needed to describe the valency frames are formalized in a dedicated extension of the PreMOn core ontology.

Following the model of PreMOn, each different frame of any given entry in Latin Vallex is an instance of the Valency Frame class. The arguments involved in the valency frames of Latin Vallex are called "frame slots," and are defined as a subclass of PreMOn's Semantic Roles. The slots, which are defined locally for each semantic class, correspond to the so-called "functors" (i.e. semantic values of syntactic dependency relations) of the Functional Generative Description (Mikulová et al. 2006). One of the main use cases of PreMOn was the mapping of different predicate models, namely those for PropBank (Palmer et al. 2005), NomBank (Meyers et al. 2004), VerbNet (Schuler 2005) and FrameNet (Baker et al. 1998). Therefore, the ontology is ideally suited to express the link between the word-synset pairs and the predicate analyses in the Latin Vallex. The PreMOn core module defines a special reification of the relation between a given semantic class and a lexical entry, called "Conceptualization." The linking itself is performed with instances of the class Mapping, which is defined as a set of conceptualizations, semantic classes, or semantic roles. Following this schema, which is also applied in the PreMOn data itself,[20] we match the words-synsets pairs of Latin WordNet and the predicate analyses in Latin Vallex by means of mapping instances linking the corresponding conceptualizations.

Figure 5 shows the complex of the WordNet and Valency annotation for one of the 12 senses recorded for the Latin verb *dono* 'to give, donate', namely the one connected to the synset 00887463-v of the Princeton WordNet (version 3.0).[21] The lexical entry (yellow node at the center of the image) is connected to both a valency frame (left-hand side) and a synset (on the right). A mapping node (in purple, directly below the entry)[22] connects the two conceptualizations.[23]

4. C*onclusion. Parts of a whole: interoperability in LiLa*

The diagram in Figure 6 provides a plastic representation of the interconnection between different layers of information linked to a canonical form in the LiLa Lemma Bank. The lemma of the adjective *malus* 'bad, evil' is described with a series of fea-

20. See for instance the mapping between a synset and a predicate analysis for the English verb "to leave out" at http://premon.fbk.eu/resource/sense-Ep7UGYgbEXbB3B2uGhZamc.

21. The synset encompasses the English lemmas: "devote, commit, give, dedicate, consecrate", with the following definition: "give entirely to a specific person, activity, or cause". See http://wordnet-rdf.princeton.edu/pwn30/00887463-v. Note that the figure also shows a second synset that is recorded as hyperonym of 00887463-v.

22. http://lila-erc.eu/data/lexicalResources/LatinVallex/id/Mapping/wn-val-l_100087_00887463-v.

23. The Latin Vallex connected to Latin WordNet in LiLa can be accessed at https://lila-erc.eu/data/lexicalResources/LatinVallex/Lexicon.

Figure 5 WordNet and valency annotation for a sense of *dono*

tures, some of which (namely, the POS and the inflection paradigm) are represented in the image. Moreover, the lemma is linked to a lexical base that is common to all the derived words belonging to the same derivational family of *malus*, like *malitia* 'malice', *maleficus* 'evil-doing, nefarious', or the rare verb *maleficio* 'to practice black magic' (shown in Figure 5).

In addition to the properties of the lemma, the canonical form is directly linked to three lexical entries from as many different resources (yellow nodes). The entry for *malus* from the etymological dictionary by de Vaan (2008) lists the inheritance relations from the Indo-European and Proto-Italic reconstructed forms. The entry from the WFL lexicon is connected to several formations in which the adjective is involved, the one with the verb *maleficio* being the only one represented in the diagram. Finally, on the left-hand side of the lemma, the entry for *malus* in the LatinAffectus lexicon registers the a-priory negative sense of the adjective.

The series of connections illustrated in Figure 6 (which, by the way, omits reference to the Latin WordNet or Vallex, as no information of the sort is available for the lemma in question) is already sufficient to provide a plastic visualization of the strong "network effect" that the model adopted by LiLa achieves. One of the most immediate applications to leverage the power of interoperability is to cross the information from one resource to the another in order to study the Latin lexicon. Traditionally, for instance, etymological dictionaries like de Vaan's (2008) do not discuss *all and every* word whose roots can be traced back to an Indo-European ancestor. Rather, the authors proceed by identifying a key lemma for a whole entry, where all the lexical items that are derived from it by regular word-formation processes are also listed. Even such list of "derivatives" is far for complete, both for the chronological limits that the dictionary authors would set to their work, and for the obvious limitations of space (in printed books) and time (available to the compiler). In a LOD scenario, these two tasks can be decoupled and assigned to two different resources, one dedicated to etymology, the other to derivational morphology. Students and scholars interested in a full list of items in the lexicon that trace back their etymology to a certain Indo-European root can interrogate the two datasets simultaneously.[24] Other possibilities offered by the interconnections between lexical resources include, for instance, a study on the semantic aspects of derivational processes. Indeed, the coverage of the LatinAffectus lexicon was extended by targeting words associated with morphemes capable of altering or conveying a polarity value, such as the prefix *in-* with negative meaning (Sprugnoli et al. 2020b).

24. See Litta et al. (2020: 177–82) for a comparison between the data on derivative words in the dictionary of de Vaan (2008) and in LiLa.

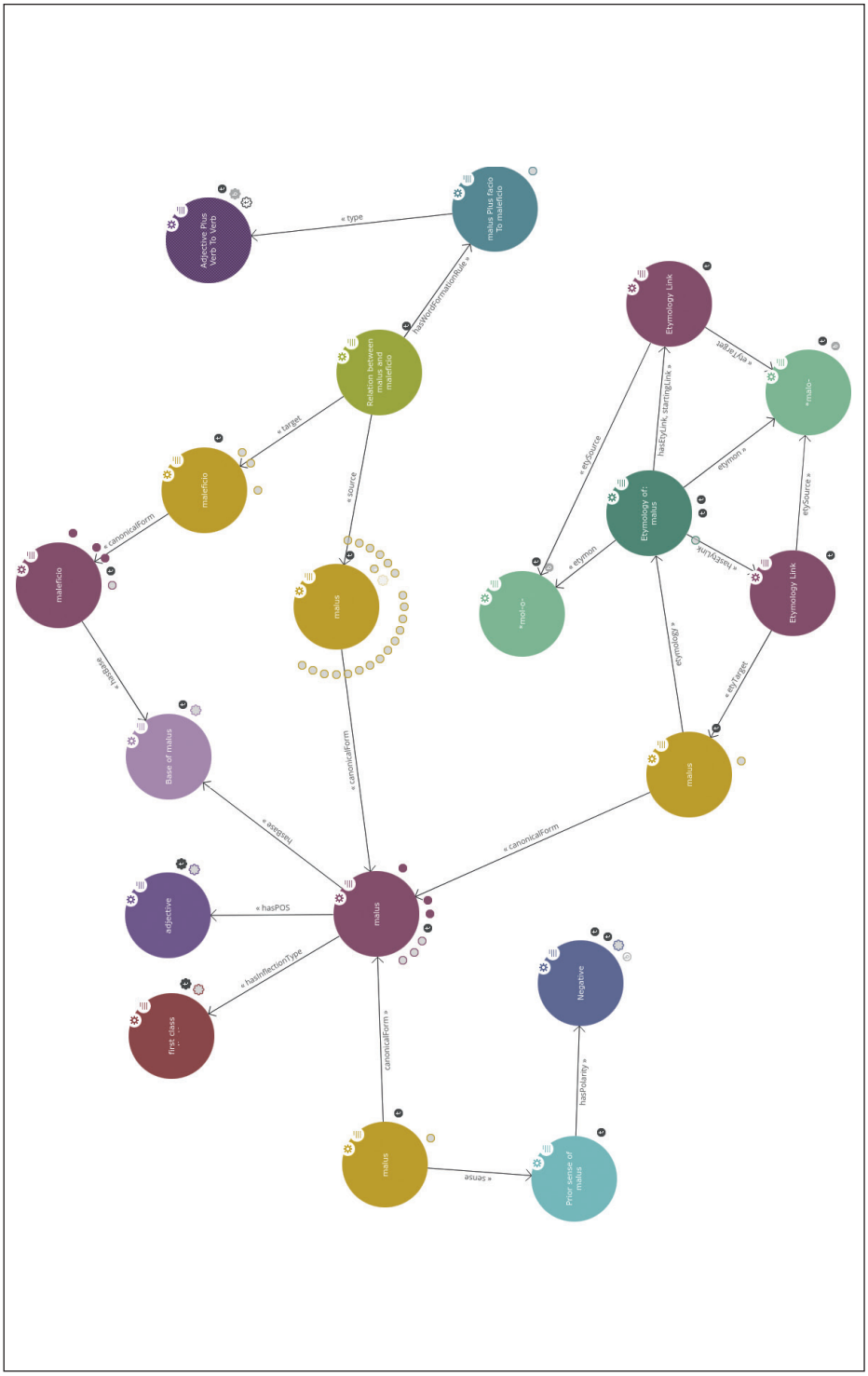


Figure 6 *Malus* (adjective) in LiLa: properties and lexical resources

Further possibilities are opened when tokens from textual corpora are integrated into the network. At present, all lemmatized corpora register the lemma of each token as a string associated to the form in the text; the same type of output is produced by automatic lemmatizers. A connection to the LiLa network is obtained when this lemma string is associated unambiguously to one of the lemmas in the Knowledge Base, for instance by matching it to one of the written representations of the canonical forms. Mambrini and Passarotti (2019) report on the results of a preliminary experiment of matching: up to 81.52% of the tokens in the Latin PROIEL UD corpus (v. 2.3) could be unambiguously associated with a LiLa lemma with a simple string match. Considering the central role played by textual resources in LiLa, the project developed a tool to automatically link a Latin raw text (i.e. without any linguistic annotation) to the LiLa Knwoledge Base. The tool, called Text Linker, makes use of an automatic lemmatizer, built upon a large training corpus that collects more than 6 million words taken from Latin texts of different eras.[25]

One of the added values of the LiLa Knowledge Base is interoperability between the different kinds of information about words provided by lexical resources (ranging from mono-/bilingual definitions to etymologies, polarity, morphology etc.) and their actual usage in texts stored in corpora, which makes of LiLa the natural venue where publishing any available or newly created language resource for Latin. By applying the principles of the LOD paradigm, it is today possible to interlink the (meta)data from any Latin resource, thus exploiting to the best its specific contribution in relation to the overall picture. This feature is essential when dealing with ancient languages that can be studied only through the attestations that survived throughout the centuries. Furthermore, interoperability between resources in LiLa is achieved by using (and sometimes extending) data models, categories and ontologies widely adopted in the larger community of Linguistic LOD. This design strategy is what makes Latin resources speak "the same language" as the resources of many other languages, both ancient and modern.

25. The training corpus was compiled by joining texts from various resources, including the LASLA corpus, the Latin treebanks available in Universal Dependencies, a subset of the Computational Historical Semantics corpus and the full text of *Confessiones* by Augustinus. Lemmatization criteria were harmonized among the corpora and the Universal POS tags were assigned (https://universaldependencies.org/u/pos/index.html).

Websites

Computational Historical Semantics corpus: https://comphistsem.org/home.html
English WordNet: https://wordnet.princeton.edu/
LASLA corpus: http://web.philo.ulg.ac.be/lasla/
Latin PROIEL UD corpus (v.2.3): http://hdl.handle.net/11234/1-2895
Latin WordNet: https://latinwordnet.exeter.ac.uk
Text Linker (beta version): http://lila-erc.eu:8080/LiLaTextLinker
UD: https://universaldependencies.org

References

Baker, Collin F., Fillmore, Charles J. & Lowe, John B. 1998. The Berkeley FrameNet Project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, 86–90. Montreal, Quebec: Association for Computational Linguistics. https://doi.org/10.3115/980845.980860

Bamman, David & Crane, Gregory. 2006. The Design and Use of a Latin Dependency Treebank. In *TLT 2006: Proceedings of the Fifth International Treebanks and Linguistic Theories Conference*. Prague: Institute of Formal and Applied Linguistics.

Bond, Francis & Foster, Ryan. 2013. Linking and Extending an Open Multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Volume 1*, Hinrich Schuetze, Pascale Fung & Massimo Poesio (eds), 1352–1362. Sofia, Bulgaria: Association for Computational Linguistics.

Booij, Geert. 2010. Construction Morphology. *Language and Linguistics Compass* 4: 543–55.

Cimiano, Philipp, Chiarcos, Christian, McCrae, John P. & Gracia, Jorge. 2020. *Linguistic Linked Data: Representation, Generation and Applications*. Cham: Springer. https://doi.org/10.1007/978-3-030-30225-2

Corcoglioniti, Francesco, Rospocher, Marco, Aprosio, Alessio P. & Tonelli, Sara. 2016. PreMOn: A Lemon Extension for Exposing Predicate Models as Linked Data. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds), 877–884. Portorož, Slovenia: European Language Resources Association.

Fellbaum, Christiane (ed). 1998. *WordNet: An Electronic Lexical Database*. Cambridge: MIT Press.

Franzini, Greta, Peverelli, Andrea, Ruffolo, Paolo, Passarotti, Marco C., Sanna, Helena, Signoroni, Edoardo, Ventura, Viviana & Zampedri, Federica. 2019. Nunc Est Aestimandum. Towards an Evaluation of the Latin WordNet. In *Sixth Italian Conference*

on Computational Linguistics (CLiC-It 2019), Raffaella Bernardi, Roberto Navigli & Giovanni Semeraro (eds), 1–8. Bari, Italy: CEUR-WS.org.

Hajič, Jan, Panevová, Jarmila, Urešová, Zdeňka, Bémová, Alevtina, Kolárová, Veronika & Pajas, Petr. 2003. PDT-VALLEX: Creating a Large-Coverage Valency Lexicon for Treebank Annotation. In *Proceedings of the Second Workshop on Treebanks and Linguistic Theories (TLT 2)*, Joakim Nivre & Erhard W. Hinrichs (eds), 57–68. Växjö: Växjö University Press.

Khan, Anas F. 2018. Towards the Representation of Etymological Data on the Semantic Web. *Information* 9: 304. https://doi.org/10.3390/info9120304

Klimek, Bettina, McCrae, John P., Ionov, Maxim, Tauber, James K., Chiarcos, Christian & Bosque-Gil, Julia. 2019. Challenges for the Representations for Morphology in Ontology Lexicons. In *Proceedings of Sixth Biennial Conference on Electronic Lexicography, ELex 2019*, Iztok Kosem, Tanara Zingano Kuhn, Margarita Correia, José Pedro Ferreira, Maarten Jansen, Isabel Pereira, Jelena Kallas, Miloš Jakubíček, Simon Krek & Carole Tiberius (eds), 570–591. Brno: Lexical Computing. https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_2019_33.pdf

Litta, Eleonora. 2018. Morphology Beyond Inflection. Building a Word Formation-Based Lexicon for Latin. In *Formal Representation and the Digital Humanities*, Paola Cotticelli-Kurras & Federico Giusfredi (eds), 97–114. Newcastle upon Tyne: Cambridge Scholars Publishing.

Litta, Eleonora, Passarotti, Marco C. & Mambrini, Francesco. 2019. The Treatment of Word Formation in the LiLa Knowledge Base of Linguistic Resources for Latin. In *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology (DeriMo 2019)*, Eleonora Litta, Marco C. Passarotti, Žabokrtský Zdeněk & Ševčíková Magda, 35–43. Prague: Institute of Formal and Applied Linguistics, Charles University.

Litta, Eleonora, Passarotti, Marco C. & Mambrini, Francesco. 2020. Derivations and Connections: Word Formation in the LiLa Knowledge Base of Linguistic Resources for Latin. *The Prague Bulletin Of Mathematical Linguistics* 115: 163–86.

Mambrini, Francesco & Passarotti, Marco C. 2019. Harmonizing Different Lemmatization Strategies for Building a Knowledge Base of Linguistic Resources for Latin. In *Proceedings of the 13th Linguistic Annotation Workshop*, Annemarie Friedrich, Daniz Zeyrek & Jet Hoek (eds), 71–80. Florence, Italy: Association for Computational Linguistics.

Mambrini, Francesco & Passarotti, Marco C. 2020. Representing Etymology in the LiLa Knowledge Base of Linguistic Resources for Latin. In *Proceedings of the Globalex Workshop on Linked Lexicography. LREC 2020 Workshop*, Ilan Kernerman, Simon Krek, John P. McCrae, Jorge Gracia, Sina Ahmadi & Besim Kabashi (eds), 20–28. Paris: European Language Resources Association.

McCrae, John P., Fellbaum, Christiane & Cimiano, Philipp. 2014. Publishing and Linking WordNet Using Lemon and RDF. In *Proceedings of the 3rd Workshop on Linked Data in Linguistics*, Christian Chiarcos, Petya Osenova, John P. McCrae & Cristina Vertan. Reykjavik, Iceland: European Language Resources Association.

Meyers, Adam, Reeves, Ruth, Macleod, Catherine, Szekely, Rachel, Zielinska, Veronika, Young, Brian & Grishman, Ralph. 2004. The NomBank Project: An Interim Report. In *Proceedings of the Workshop Frontiers in Corpus Annotation at HLT-NAACL 2004*, 24–31. Boston: Association for Computational Linguistics.

Mikulová, Marie, Bémová, Allevtina, Hajič, Jan, Hajičová, Eva, Kolářová, Veronika, Kučová, Lucie, Lopatková, Markéta et al. 2006. Annotation on the Tectogrammatical Layer in the Prague Dependency Treebank. Annotation Manual. 30. Prague: UFAL. https://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/t-layer/pdf/t-man-en.pdf

Minozzi, Stefano. 2017. Latin WordNet, una rete di conoscenza semantica per il latino e alcune ipotesi di utilizzo nel campo dell'information retrieval. In *Strumenti digitali e collaborativi per le Scienze dell'Antichità*, Paolo Mastrandrea (ed), 123–133. Venezia: Ca' Foscari. https://doi.org/10.14277/6969-182-9/ANT-14-10

Palmer, Martha, Gildea, Daniel & Kingsbury, Paul. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics* 31: 71–106. https://doi.org/10.1162/0891201053630264

Passarotti, Marco C. 2019. The Project of the Index Thomisticus Treebank. In *Digital Classical Philology. Ancient Greek and Latin in the Digital Revolution*, Monica Berti (ed), 299–319. Berlin: De Gruyter.

Passarotti, Marco C., Budassi, Marco, Litta, Eleonora & Ruffolo, Paolo. 2017. The Lemlat 3.0 Package for Morphological Analysis of Latin. In *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*, Gerlof Bouma & Yvonne Adesam (eds), 24–31. Gothenburg: Linköping University Electronic Press.

Passarotti, Marco C., Mambrini, Francesco, Franzini, Greta, Cecchini, Francesco M., Litta, Eleonora, Moretti, Giovanni, Ruffolo, Paolo & Sprugnoli, Rachele. 2020. Interlinking through Lemmas. The Lexical Collection of the LiLa Knowledge Base of Linguistic Resources for Latin. *Studi e Saggi Linguistici* 58: 177–212. https://doi.org/10.4454/ssl.v58i1.277

Passarotti, Marco C., Saavedra, Berta G. & Onambele, Christophe. 2016. Latin Vallex. A Treebank-Based Semantic Valency Lexicon for Latin. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds), 2599–2606. Portorož, Slovenia: European Language Resources Association.

Pianta, Emanuele, Bentivogli, Luisa & Girardi, Christian. 2002. MultiWordNet: Developing an Aligned Multilingual Database. In *Proceedings of the First International Conference on Global WordNet*. Vol. 152. Mysore, India: Global WordNet Association.

Saalfeld, Günther A. 1874. *Index graecorum vocabulorvm in linguam latinam translatorum quaestiunculis auctus*. Berolini: Berggold.

Short, William M. This Volume. WordNets, Sembanks, and the Challenge of Semantic Polyvalency.

Schuler, Karin K. 2005. *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. PhD

dissertation, University of Pennsylvania. https://repository.upenn.edu/dissertations/AAI3179808

Sprugnoli, Rachele, Mambrini, Francesco, Moretti, Giovanni & Passarotti, Marco C. 2020a. Towards the Modeling of Polarity in a Latin Knowledge Base. In *WHiSe 2020 Workshop on Humanities in the Semantic Web 2020*, Alessandro Adamou, Enrico Daga, & Albert Meroño-Peñuela (eds), 59–70. Heraklion, Greece: CEUR-WS.org.

Sprugnoli, Rachele, Passarotti, Marco C., Corbetta, Daniela & Peverelli, Andrea. 2020b. Odi et Amo. Creating, Evaluating and Extending Sentiment Lexicons for Latin. In *Proceedings of the 12th Language Resources and Evaluation Conference*, Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds), 3078–3086. Marseille, France: European Language Resources Association.

Stead, Stephen, Doerr, Martin, Ore, Christian-Emila, Kritsotaki, Athina et al. 2019. *CRMinf: The Argumentation Model, Version 0.10.1 (Draft)*. http://new.cidoc-crm.org/crminf/sites/default/files/CRMinf%20ver%2010.1.pdf

de Vaan, Michiel. 2008. *Etymological Dictionary of Latin: And the Other Italic Languages*. Leiden and Boston: Brill.

Wilkinson, Mark D., Dumontier, Michel, Aalbersberg, Ijsbrand J., Appleton, Gabrielle, Axton, Myles, Baak, Arie, Blomberg, Niklas et al. 2016. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Scientific Data* 3: 160018. https://doi.org/10.1038/sdata.2016.1