

Interlinking Valency Frames and WordNet Synsets in the LiLa Knowledge Base of Linguistic Resources for Latin

Francesco MAMBRINI^{a,1}, Marco PASSAROTTI^a Eleonora LITTA^a
Giovanni MORETTI^a

^a*Università Cattolica del Sacro Cuore, Milan (Italy)*

Abstract. This paper describes the steps taken to model a valency lexicon for Latin (Latin Vallex) according to the principles of the Linked Data paradigm, and to interlink its valency frames with the lexical senses recorded in a manually checked subset of the Latin WordNet. The valency lexicon and the WordNet share lexical entries and are part of the LiLa Knowledge Base, which interlinks multiple linguistic resources for Latin. After describing the overall architecture of LiLa, as well as the structure of the lexical entries of Latin Vallex and Latin WordNet, the paper focuses on how valency frames have been modeled in LiLa, in line with a submodule of the Predicate Model for Ontologies (PreMON) specifically created for the representation of grammatical valency. A mapping of the valency frames and the WordNet synsets assigned to the lexical entries shared by the two resources is detailed, as well as a number of queries that can be run across the interoperable resources for Latin currently included in LiLa.

Keywords. Linguistic valency, Linguistic Linked Open Data, Latin, WordNet.

1. Introduction

In lexicography, a widespread approach to the representation of lexical meaning is based on the fundamental assumption of frame semantics by C. Fillmore [1], namely that the meaning of some words can be fully understood only by knowing the frame elements that those lexical items evoke. Such an approach is related to the concept of linguistic valency [2] [3]. The latter concept is used to denote the number of obligatory complements (called 'arguments', but also 'actants' [3] or 'inner participants' [4]) controlled by a word, usually a content verb. The different types of argument are generally represented in valency frames through labels for semantic roles, such as Agent, Patient and Beneficiary.

Several lexical resources are built around the concept of valency and argument frames that describe the argument structure of words with the help of various sets of labels for semantic roles. For instance, the criteria for distinguishing obligatory and non-obligatory complements and the degree of granularity of the set of semantic roles are

¹Corresponding Author: Francesco Mambrini, CIRCSE, Università Cattolica del Sacro Cuore, Largo Gemelli 1, 20123 Milano, Italia; E-mail: francesco.mambrini@unicatt.it.

what mostly distinguishes valency-based resources like PropBank [5] (and NomBank [6]), VerbNet [7] and FrameNet [8] one from the other.

Another kind of lexical resource largely used in both theoretical and computational linguistics is WordNet (WN) [9], which is built around the idea of synonymy in the broad sense. In WordNet, words are included in synsets, which are sets of lexical items that share the same sense, so “that [they] are interchangeable in some context without changing the truth value of the proposition in which they are embedded”.²

Despite their differences, not only are the valency-based and the WN approaches to lexical meaning not incompatible, but they are strictly related, because valency frames tend to correspond to lexical senses. This is precisely the case in the PDT-VALLEX valency lexicon for Czech [10], where one valency lexicon entry (i.e., valency frame) is created for each sense of a word. Given that in WN the different senses of a polysemic word are represented through assignments to different synsets, mapping the frame entries of a valency lexicon with the synsets of a WN promises to be an effective way to provide a comprehensive representation of lexical meaning that joins the valency-based and the synset-based approaches.

In order to perform such mapping, two conditions must hold. Firstly, both a WN and a valency lexicon must be available for a specific language, with a substantial number of lexical entries in common. Secondly, the mapping process entails the design of a specific technique for interlinking these two kinds of lexical resources in the most standard fashion as possible, so as to make the process applicable to data in any language.

For Latin, the former condition is satisfied by the existence of both a Latin WN and a Latin valency lexicon. In this paper, we will address the challenges in the second condition with the help of a set of models and ontologies developed by the community working in the area of the Linguistic Linked Open Data (LLOD), whose objective is to make (meta)data of distributed linguistic resources interoperable on the web. We describe the process to represent a valency lexicon for Latin according to the principles of LLOD and to interlink its valency frames with the lexical senses recorded in a manually checked subset of the Latin WN. The valency lexicon and the WN are included together (with shared lexical entries) in the LiLa Knowledge Base, which makes distributed linguistic resources for Latin interoperable by applying LLOD principles. The outcome of the work is a new sub-module of the *PreMOn* ontology (see Section 5.1) dedicated to valency frames and the two linked datasets of the Latin WordNet,³ and Latin Vallex.⁴

The paper is organized as follows. Section 2 provides an overview of the related work about (inter)linking valency-based lexical resources between them and with WNs. Section 3 presents the fundamental architecture of the LiLa Knowledge Base. Section 4 describes the Latin WN and the Latin Vallex resources. Section 5 details how we modelled valency in LiLa and the mapping between the valency frames of Latin Vallex and the synsets of WN. Finally, Section 6 concludes the paper.

²Quoted from the glossary of WordNet: <http://wordnet.princeton.edu>.

³<http://lila-erc.eu/lodview/data/lexicalResources/LatinWordNet/Lexicon>.

⁴<http://lila-erc.eu/lodview/data/lexicalResources/LatinVallex/Lexicon>.

2. Related Work

Over the last decade, several attempts at (inter)linking different lexical resources together have been performed. One of the best known projects is Semlink, which makes use of a set of mappings to link PropBank, VerbNet, FrameNet and WN [11]. Paziienza et alii [12] study the semantics of verb relations by mixing WN, VerbNet and PropBank. Shi and Mihalcea [13] integrate FrameNet, VerbNet and WN into one knowledge-base for semantic parsing purposes.

In the LLOD context, the *PreMON* (*predicate model for ontologies*) resource [14] exposes predicate models for PropBank, NomBank, VerbNet, and FrameNet and mappings between them. More information on *PreMON* will be provided in Section 5.1.

Regarding the relations between valency lexica and WNs, Hlaváčková [15] describes the merging of the Czech WordNet (CWN) with the database of verb valency frames for Czech VerbaLex, whose lexical entries are related to each other according to the CWN synsets. Hajič et alii [16] use CWN while performing the lexico-semantic annotation of the Prague Dependency Treebank for Czech (PDT), which is in turn exploited to improve the quality and the coverage of CWN. To pick out the semantic constraints of the verbal arguments in the Polish WordNet (PolNet), the valency structure of verbs is used as a property of verbal synsets, because it is “one of the formal indices of the meaning (it is so that all members of a given synset share the valency structure)” (page 402 of [17]). Finally, Passarotti et alii [18] compare the different views on lexical meaning conveyed by the Latin WN and by a valency lexicon for Latin, evaluating the degree of overlapping between a number of homogeneous lexical subsets extracted from the two resources.

3. The LiLa Knowledge Base

The *LiLa: Linking Latin* project (2018-2023)⁵ was awarded funding from the European Research Council (ERC) to build a Knowledge Base (KB) of linguistic resources for Latin based on the Linked Data paradigm. Our aim is to build a collection of multifarious, interlinked data sets of Latin resources represented with the same vocabulary of knowledge description (by using common data categories and ontologies) [19].

According to the Linked Data paradigm, data in the Semantic Web [20] are interlinked through connections that can be semantically queried, so as to make the structure of web data better serve the needs of users. In order to achieve interoperability between distributed resources for Latin, LiLa makes use of a set of Semantic Web and Linked Data standards and practices. These include ontologies to describe linguistic annotation (OLiA: [21]), corpus annotation (CoNLL-RDF: [22]) and lexical resources (Ontolex-Lemon: [23]).

Following Bird and Liberman [24], the Resource Description Framework (RDF) [25] is used to encode graph-based data structures to represent linguistic annotations in terms of triples: (i) a predicate-property (a relation; in graph terms: a labeled edge) that connects (ii) a subject (a resource; in graph terms: a labeled node) with (iii) its object (another resource/node, or a literal, e.g. a string).

Given the presence and role played by lemmatization in various linguistic resources, and the good accuracy rates achieved by the best performing lemmatizers for Latin (up

⁵<https://lila-erc.eu>.

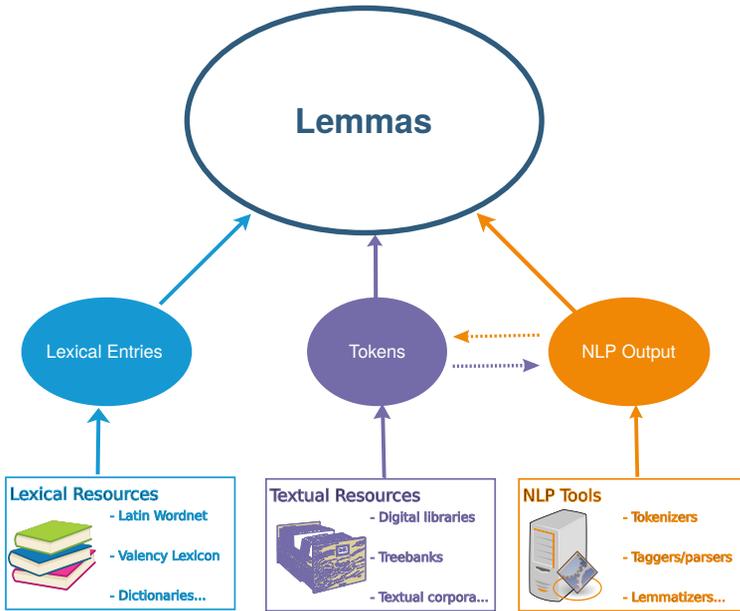


Figure 1. The main components of the LiLa ontology model.

to 0.96, as per the results of the EvaLatin 2020 evaluation campaign [26]), LiLa uses the lemma as the most productive interface between lexical resources, annotated corpora and Natural Language Processing (NLP) tools. Consequently, the ontology model of the LiLa KB is highly lexically based, grounded on a simple, but effective assumption that strikes a good balance between feasibility and granularity: textual resources are made of (occurrences of) words, lexical resources describe properties of words, and NLP tools process words.

Fig. 1 presents the main components of the ontology model of the LiLa KB, showing the key role of interlinking played by lemmas.

Fig. 1 shows that there are three kinds of (meta)data sources in LiLa, namely: (1) lexical resources, which include lexical entries, (2) textual resources, which are made of occurrences of words (tokens), and (3) NLP tools, which produce different outputs according to their task(s) of automatic linguistic analysis.⁶ In LiLa, lexical entries, tokens and NLP outputs are interlinked via lemmas.

To make this conceptual architecture work, the core of the LiLa KB consists of a large collection of Latin lemmas: interoperability is achieved by linking all those entries in lexical resources and tokens in corpora that point to the same lemma. The lexical basis of the Latin morphological analyzer Lemlat [27] was used to populate the LiLa collection. Lemlat's database reconciles three reference dictionaries for Classical Latin [28] [29] [30], the entire Onomasticon from Forcellini's *Lexicon Totius Latinitatis* [31]

⁶The arrows that double link the nodes for 'Tokens' and 'NLP Output' in Fig. 1 represent that tokens can be both the output of an NLP tool (a tokenizer) and the input of another (e.g., a PoS tagger).

and the Medieval Latin *Glossarium Mediae et Infimae Latinitatis* by du Cange et alii [32], for a total of over 150,000 lemmas.

Beside the collection of Latin lemmas, the linguistic resources linked so far via the LiLa KB are the following:⁷

- the *Index Thomisticus* Treebank, both in its original and in the *Universal Dependencies* version [33];⁸
- the Latin works of Dante Alighieri, taken from the *DanteSearch* corpus [34];⁹
- the text of the Late Antiquity comedy *Querolus sive Aulularia*;¹⁰
- the *Etymological Dictionary of Latin and the other Italic Languages* by M. De Vaan [35] [36];¹¹
- the *LatinAffectus* sentiment lexicon of Latin [37];¹²
- the collection of Ancient Greek loanwords in the Latin language *Index Graecorum vocabulorum in linguam Latinam translatorum quaestiunculis auctus* by G. Saalfeld [38] [39];¹³
- a manually checked set of 1,424 entries of the Latin WordNet [40];¹⁴
- the valency lexicon *Latin Vallex* [41];¹⁵
- the derivational lexicon *Word Formation Latin*[42].¹⁶

Both the lemma collection and the source (meta)data of the resources linked to LiLa (together with their Turtle files, which provide the RDF triples) are freely available from the GitHub page of the CIRCSE research center.¹⁷

4. Latin WordNet and Latin Vallex

The LatinWordNet (LWN) [43] was initiated in the context of the MultiWordNet project [44], whose aim was to build a number of semantic networks for specific languages aligned with the synsets of the Princeton WordNet (PWN) [45]. Unfortunately, the automatic process employed to set up such an alignment resulted in a dataset that is at the same time largely incomplete (i.e. lacking a number of crucial lemmas e.g. *amo* “to love”) and inaccurate, especially due to the presence of various modern senses inherited from MultiWordNet.

⁷To connect the entries of lexical resources as well as the lemmatized tokens of annotated corpora to the LiLa collection of lemmas, we perform a simple string match between the lemmas in the resource to connect and those in the LiLa collection. Ambiguous links, which happen when one lemma in the resource is connected to more than one lemma in the LiLa collection, are disambiguated manually. Missing links are solved by including the missing lemmas in LiLa. The procedure is detailed with reference to the *Index Thomisticus* Treebank in [19].

⁸<https://lila-erc.eu/data/corpora/ITTB/id/corpus>.

⁹<https://lila-erc.eu/data/corpora/DanteSearch/id/corpus>; <https://dantesearch.dantenetwork.it/>.

¹⁰<https://lila-erc.eu/data/corpora/Querolus/id/citationUnit/QuerolussiveAulularia>.

¹¹<https://lila-erc.eu/data/lexicalResources/BrillEDL/Lexicon>.

¹²<https://lila-erc.eu/data/lexicalResources/LatinAffectus/Lexicon>.

¹³<https://lila-erc.eu/data/lexicalResources/IGVLL/Lexicon>.

¹⁴<http://lila-erc.eu/data/lexicalResources/LatinWordNet/Lexicon>.

¹⁵<https://lila-erc.eu/data/lexicalResources/LatinVallex/Lexicon>.

¹⁶<https://lila-erc.eu/data/lexicalResources/WFL/Lexicon>.

¹⁷<https://github.com/CIRCSE>.

Before its inclusion in LiLa, LWN counted 8,973 synsets and 9,124 lemmas. These are currently undergoing substantial revision to refine and extend its contents.¹⁸ There are currently 1,424 fully reviewed entries in LiLa, for a total of 2,809 unique synsets. Synset are found or checked against two main Classical and Late Latin dictionaries [29] [46]. For each sense found in the dictionaries, one or more corresponding synsets are extracted from the 3.0 version of Princeton WordNet [9]. For example, the Latin verb *concordo* has the following senses:

- “to be in good terms, be friendly, live in harmony (with someone)”;
- “to be in agreement, harmonise, agree”;
- “to bring about harmony/an harmonious relationship (between things), bring into union”.

Thereby, the following synsets have been chosen for *concordo*:

- 02700104-v “go together”;
- 02657219-v “be compatible, similar or consistent; coincide in their characteristics”;
- 00805376-v “be in accord; be in agreement”;
- 01035530-v “achieve harmony of opinion, feeling, or purpose”;
- 00482473-v “bring into consonance or accord”.

Each of these are then linked to a valency frame in the valency lexicon Latin Vallex (LV).

LV [41] was originally developed while performing the semantic annotation of two Latin treebanks, namely the *Index Thomisticus* Treebank, which includes works of Thomas Aquinas [47], and the Latin Dependency Treebank, featuring works of different authors of the Classical era [48]. All valency-capable lemmas occurring in the semantically annotated portion of the two treebanks are assigned one lexical entry and one valency frame in LV.

The structure of LV resembles that of the valency lexicon for Czech PDT-VALLEX [10]. On the topmost level, the lexicon is divided into lexical entries. Each entry consists of a sequence of frame entries relevant for the lemma in question. A frame entry contains a sequence of frame slots, each corresponding to one argument of the given lemma. Each frame slot is assigned a semantic role. The set of semantic roles is the same used for the semantic annotation of the PDT [49]. Since the development of the lexicon is directly related to the annotation of the texts in the PDT, the surface form of the semantic roles run across during the annotation process is recorded as well.

In view of its extension and linking to LiLa, a different choice has been made in regards to the inner structure and theoretical principles of LV. In order to increase the lexical coverage of LV, the writing of the valency frames was disassociated from the treebank annotation process and valency is now defined on the basis of senses. For each sense, a valency frame is established intuitively, listing only its obligatory complements. This process involves finding a valency frame for each sense of the dictionary headword, as there might be differences in the number and/or type of arguments for different senses.

Since senses of words in LiLa are represented as PWN synsets evoked by LWN lexical entries, each (valency-bearing) synset is linked to a valency frame of LV, thus

¹⁸For an evaluation of the original content in LWN, see [40].

interlinking the two resources (see Section 5.1). The job is currently being performed manually, the valency frames included in the first version of LV have been updated, cleaned or rectified, and applied to each valency-bearing synset. Currently, 1,064 lexical entries have been annotated, for a total of 9,806 valency frames, while 1,424 entries of the LWN have been checked and revised manually.

5. Modeling Valency in LiLa

5.1. *Ontolex and PreMON*

In order to provide a common infrastructure to link the different lexical and textual resources using lemmatization as the connecting point, the LiLa KB adopts the *Ontolex-Lemon* ontology to model information of its comprehensive collection of Latin lemmas. *Ontolex-Lemon* [23] is a de-facto standard for the description of lexical entries. Particularly relevant for the aims of the projects is the property *canonical form*, as defined in the ontology, that allows us to express the relation between an entry in any given lexical resource and a lemma in the LiLa collection.

The core properties and classes of *Ontolex* are sufficiently expressive to account for the relations between words, senses and synsets as defined in WN. The meaning of a lexical entry is captured by expressing the relation to either a denoted entity or an evoked lexical concept in an ontology. An instance of the class `Lexical Sense` is used to reify this relation. Following the model of a complete publication of the PWN as LLOD, LiLa now includes all the 1,424 entries of the manually revised LWN.¹⁹

While a number of extensions of *Ontolex* allow for the expression of many properties of the lexical items, including the syntactic frames and their elements, the ontology is not well suited to conveyexpress the notion of a predicate structure of semantic roles. The *Predicate Model for Ontologies* (PreMON) builds also on *Ontolex*, but is explicitly designed to provide a description of the predicate structure and of the semantic roles connected to each lexical entry. At the same time, it also allows to map different semantic descriptions of any given word (such as a predicate structure and a link to a WN synset) to each other. This feature makes it ideally suited to represent both the structure of the valency frames and their connection to the senses described in the LWN.

PreMON is based on a core module, whose main elements are the `Semantic Class` and `Semantic Role` classes. The former represents the the semantic classes from the various predicate models; thus, for examplinstance, rolesets from PropBank and frames from FrameNet are all instances of semantic classes. The different sub-classes of this general class, however, can be further specified in dedicated submodels, where the terminology and the relations that are peculiar to each specific project can also be defined.

5.2. *The Vallex Submodule*

To properly capture the structure of the valency frames in LV, we created a submodule of *PreMON* that introduces a series of new subclasses and subproperties (which are prefixed with the namespace `pmo1v` in what follows). The structure of the LV submodule is

¹⁹Metadata and a list of the entries can be browsed at: <http://lila-erc.eu/data/lexicalResources/LatinWordNet/Lexicon>.

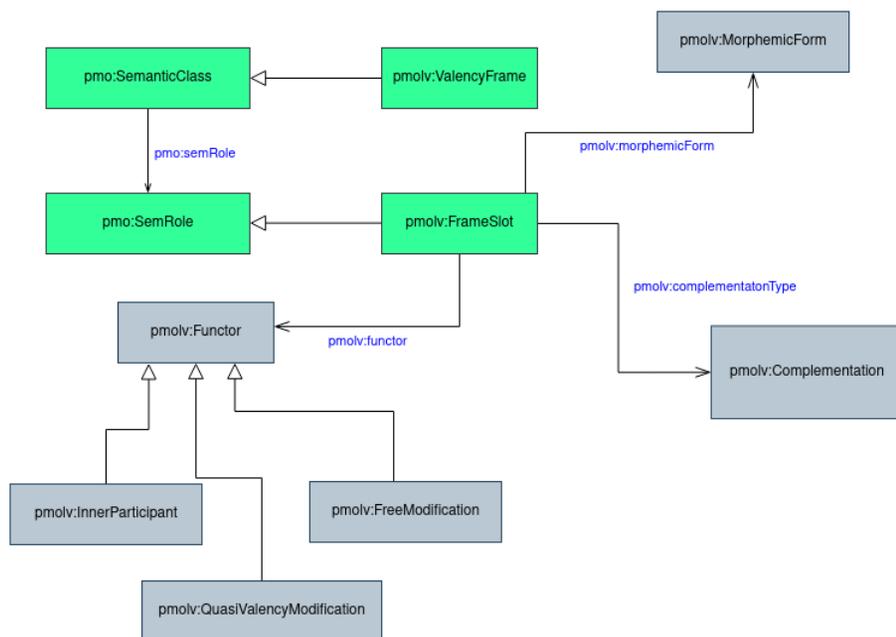


Figure 2. The Latin Vallex submodule for *PreMON*.

illustrated in Fig. 2. We define the valency frame as a subclass of `pmo:SemanticClass`; each different frame of any given entry in our lexicon is an instance of this class. The arguments involved in the valency frames of LV, called *frame slots*, are treated as a subclass of `pmo:SemRole`. These slots, which, as it is mandated for *PreMON*'s semantic roles, are defined locally to each semantic class, correspond to the so-called “functors” (i.e. semantic values of syntactic dependency relations) of the Functional Generative Description [49], which are classified into the three main categories of inner participants, free modifications, and quasi-valency modifications [49] [50]. We also use the `pmolv:complementationType` property to distinguish between obligatory and optional modification. Although the morphemic form, i.e. the morpho-syntactic realization of the role used in language, is not expressed in LV, the submodule identifies one property and one class to specify this information.

5.3. Mapping between Valency Frames and Synsets

PreMON allows users to map pairs of words and predicate structure from different predicate models (e.g. from PropBank and FrameNet) to each other. In order to express this link, the core module defines a special reification of the relation between a given semantic class and a lexical entry, called “Conceptualization”. Mapping itself is performed with instances of the class `pmo:Mapping`, which is defined as a set of conceptualizations, semantic classes, or semantic roles. Thus, in the *PreMON* data, words-synsets pairs are matched with the predicate analyses from resources like PropBank and VerbNet by means of mapping instances linking the corresponding conceptualizations.²⁰

²⁰The mapping at the following address, for instance, links the conceptualization between the verb “to leave out” and synset 200616690-v to the one between the same verb and the class `neglect-75.1` for VerbNet:

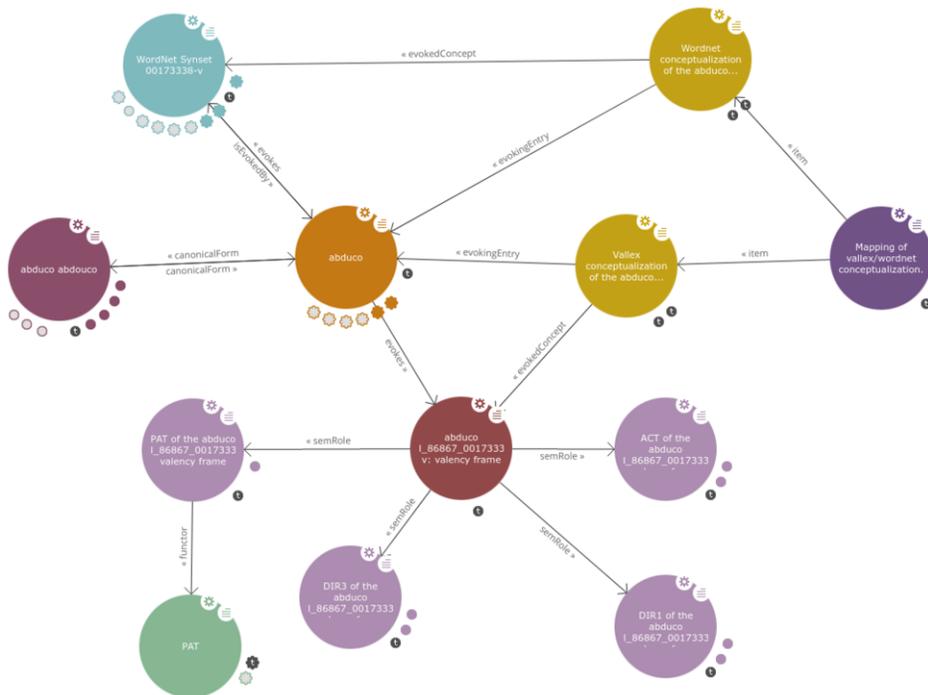


Figure 3. The verb *abduco* from the Latin Vallex and WordNet in LiLa.

The version of LV included in LiLa adopts the same approach to connect valency frames and senses from the LWV. Fig. 3 represents the predicate structure and one of the WN links for the Latin verb *abduco*, in the sense “to take away, to remove”.²¹ A mapping instance (furthest node on the right, in the figure) holds together the two conceptualizations, that connect the verb to, respectively, synset 00173338-v of WN²² and a valency frame for that specific sense.²³

The valency frame requires four (obligatory) frame slot, whose roles are filled by the functors: ACT (generally, the Actor), PAT (generally, the Patient),²⁴ DIR1 (Direction From) and DIR3 (Direction To). Note that in the figure only the link between the semantic role for PAT and the functor is shown.

5.4. Querying Interlinked Lexical and Textual Resources

As an example of the benefits of interoperability between different resources enabled by the Linked-Data approach to the publication of the joint LWN and LV data, we consider

<http://premon.fbk.eu/resource/sense-Ep7UGYgbEXbB3B2uGhZamc>.

²¹Following the habit of Latin lexicography, the 1st person singular of active present indicative is conventionally chosen as the citation form for verbs.

²²See <http://wordnet-rdf.princeton.edu/pwn30/00173338-v>.

²³The frame is: <http://lila-erc.eu/data/lexicalResources/LatinVallex/id/Conceptualization/co-val-1\86867\00173338-v>.

²⁴It must be kept in mind that, in the PDT-VALLEX formalism, the first arguments are subject to the “argument shift” [49, sec. 6.2.1.4], and are always labeled ACT, PAT and EFF, regardless of their semantic value.

Table 1. Synsets associated with tetravalent verbs in LV (5 with highest nr. of lexical entries associated in LWN).

Synset	Gloss	Nr. Entries
00173338-v	remove something concrete, as by lifting, pushing, or taking off, or remove something abstract	13
01354673-v	connect, fasten, or put together two or more pieces	10
01494310-v	put into a certain place or abstract location	10
02304982-v	get or gather together	9
00179311-v	take out or remove	9

Table 2. Most frequent tetravalent verbs in the *Index Thomisticus* Treebank (ITTB) and in *DanteSearch* (DS).

Corpus	Verb	Freq.
ITTB	compono	423
	confero	300
	committo	136
DS	confero	60
	expello	32
	committo	28

a couple of complex queries that span over several resources. At present, the majority of valency frames in LV has two argument slots (5,505, i.e. 66% of the frames). The highest value of obligatory arguments is 4, a number that is required by 55 frames (0.6%).

The mapping between LV and the LWN allows for an easy investigation on the semantics of these verbs. By querying the SPARQL endpoint of LiLa,²⁵ it is possible to list the synsets that are mapped with the 55 valency frames that require 4 arguments. Table 1 lists the top five synsets associated with tetravalent verbs, ranked by the number of lexical entries associated to them.

LiLa's connections via lemmatization enables also to scrutinize, at least in a preliminary survey, the distribution of these tetravalent verbs in the linked corpora. The three most frequent verbs that have at least one 4-argument frame in the *Index Thomisticus* Treebank and *DanteSearch* are reported in Table 2. It must be kept in mind, however, that the raw frequencies reported in the table do not distinguish between the multiple valency frame associated with each verb; thus, the 300 and 60 occurrences of the verb *confero* "to bring, to carry together" might be instances of any of the 32 valency frames associated with it.²⁶ Only with a corpus enhanced with word-sense disambiguation it will be possible to distinguish between them.

6. Conclusions

In this paper we described how we modeled the contents of a valency lexicon for Latin strictly connected to a WordNet, to interlink it with other linguistic resources for Latin in a Knowledge Base built upon LLOD principles.

²⁵<https://lila-erc.eu/sparql>.

²⁶Translations of the Latin verbs in Table 2 beside *confero*: *committo* "to bring together, to join", *compono* "to bring together, to collect", *expello* "to drive out".

The modeling was performed by building a specific submodule of the *PreMON* Linked Data resource for representing predicate-based lexical resources. In particular, the OWL ontology provided by *PreMON* was used for modeling the valency frames and the semantic roles adopted in the Latin valency lexicon. We reused (and extended) an already existing vocabulary to meet one of the main tenets of the LLOD world, where distributed linguistic data, metadata and resources are made interoperable just thanks to the fact that they are represented through common data categories and ontologies maintained by the large and active LLOD community.

Although the LiLa Knowledge Base aims first of all to interlink the available linguistic resources for Latin, by grounding on standard LLOD modules, it also wants to represent a reference architecture for the publication of the several (kinds of) resources that were developed for many languages over the last decades.

Thanks to LiLa, a number of linguistic resources for Latin are already made interoperable. The inclusion of a valency lexicon, with entries shared with a manually checked subset of the Latin WordNet, is a major achievement for the Knowledge Base and, hopefully, for the entire community interested in accessing and using linguistic resources for Latin, as it enhances the queries that can be performed on the textual resources connected to LiLa with lexical semantic information. In such respect, one of the near future objectives of the LiLa project is to include a reference Latin-English dictionary in the Knowledge Base [46], while continuing to enlarge the number of corpora that are made interoperable thanks to their connection to LiLa.

References

- [1] Fillmore CJ. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences: Conference on the origin and development of language and speech*. 1976;280(1):20–32.
- [2] Ágel V, Fischer K. Dependency grammar and valency theory. In: Heine B, Narrog H, editors. *The Oxford Handbook of Linguistic Analysis*. Oxford University Press; 2009. p. 225–258.
- [3] Tesnière L. *Éléments de syntaxe structurale*. Paris: Klincksieck; 1959.
- [4] Panevová J. In favour of the argument-adjunct distinction (from the perspective of FGD). *Prague Bulletin of Mathematical Linguistics*. 2016;106(1):21–30.
- [5] Kingsbury P, Palmer M. From TreeBank to PropBank. In: *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*. Las Palmas: European Language Resources Association (ELRA); 2002. p. 1989–1993.
- [6] Meyers A, Reeves R, Macleod C, Szekely R, Zielinska V, Young B, et al. The NomBank Project: An Interim Report. In: *Proceedings of the Workshop Frontiers in Corpus Annotation at HLT-NAACL 2004*. Boston: Association for Computational Linguistics; 2004. p. 24–31.
- [7] Schuler KK. *VerbNet: A broad-coverage, comprehensive verb lexicon* [Dissertation]. University of Pennsylvania. Philadelphia; 2005. Available from: <https://repository.upenn.edu/dissertations/AAI3179808>.
- [8] Baker CF, Fillmore CJ, Lowe JB. The Berkeley FrameNet Project. In: *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*. Montreal, Quebec, Canada: Association for Computational Linguistics; 1998. p. 86–90.
- [9] Miller GA. WordNet: a lexical database for English. *Communications of the ACM*. 1995;38(11):39–41.
- [10] Hajic J, Panevová J, Uřešová Z, Bémová A, Kolářová V, Pajas P. PDT-VALLEX: Creating a large-coverage valency lexicon for treebank annotation. In: *Proceedings of the second workshop on treebanks and linguistic theories (TLT)*. Växjö: Växjö University Press; 2003. p. 57–68.
- [11] Palmer M. Semlink: Linking PropBank, VerbNet, FrameNet. In: *Proceedings of the generative lexicon conference*. Pisa: Association for Computational Linguistics; 2009. p. 9–15.
- [12] Paziienza MT, Pennacchiotti M, Zanzotto FM. Mixing WordNet, VerbNet and PropBank for studying verb relations. In: *Proceedings of the Fifth International Conference on Language Resources and*

- Evaluation (LREC 2006). Genoa, Italy: European Language Resources Association (ELRA); 2006. p. 1372–1377.
- [13] Shi L, Mihalcea R. Putting pieces together: Combining FrameNet, VerbNet and WordNet for robust semantic parsing. In: Gelbukh A, editor. International conference on intelligent text processing and computational linguistics. Berlin and Heidelberg: Springer; 2005. p. 100–111.
- [14] Corcoglioniti F, Rospocher M, Aprosio AP, Tonelli S. 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); 2016. p. 877–884.
- [15] Hlaváčková D. The relations between semantic roles and semantic classes in VerbaLex. In: Sojka P, Horák A, editors. Proceedings of Recent Advances in Slavonic Natural Language Processing (RASLAN 2007). Brno: Masaryk University; 2007. p. 97–101.
- [16] Hajič J, Holub M, Hučínová M, Pavlík M, Pecina P, Straňák P, et al. Validating and Improving the Czech WordNet via Lexico-Semantic Annotation of the Prague Dependency Treebank. In: Proceedings of the 4th International Conference On Language Resources And Evaluation (LREC). Lisbon: European Language Resources Association; 2004. p. 25–30.
- [17] Vetulani Z, Kochanowski B. “PolNet - Polish WordNet” project: PolNet 2.0 - a short description of the release. In: Proceedings of the Seventh Global Wordnet Conference. Tartu, Estonia: University of Tartu Press; 2014. p. 400–404.
- [18] Passarotti M, González Saavedra B, Onambélé Manga C. Somewhere between valency frames and synsets. Comparing Latin Vallex and Latin WordNet. In: Bosco C, Zanzotto FM, editors. Proceedings of the Second Italian Conference on Computational Linguistics CLiC-it 2015. Torino: Accademia University Press; 2015. p. 221.
- [19] Passarotti M, Mambrini F, Franzini G, Cecchini FM, Litta E, Moretti G, et al. Interlinking through Lemmas. The Lexical Collection of the LiLa Knowledge Base of Linguistic Resources for Latin. *Studi e Saggi Linguistici*. 2020;58(1):177–212.
- [20] Berners-Lee T, Hendler J, Lassila O. The semantic web. *Scientific american*. 2001;284(5):34–43.
- [21] Chiarcos C, Sukhareva M. Olia – ontologies of linguistic annotation. *Semantic Web*. 2015;6(4):379–386.
- [22] Chiarcos C, Fäth C. CoNLL-RDF: Linked Corpora Done in an NLP-Friendly Way. In: Gracia J, Bond F, McCrae JP, Buitelaar P, Chiarcos C, Hellmann S, editors. Language, Data, and Knowledge. Cham, Switzerland: Springer; 2017. p. 74–88.
- [23] McCrae JP, Bosque-Gil J, Gracia J, Buitelaar P, Cimiano P. The OntoLex-Lemon Model: Development and Applications. In: Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference. Brno, Czech Republic: Lexical Computing CZ s.r.o.; 2017. p. 587–597.
- [24] Bird S, Liberman M. A formal framework for linguistic annotation. *Speech communication*. 2001;33(1-2):23–60.
- [25] Lassila O, Swick RR. Resource Description Framework (RDF) Model and Syntax Specification; 1998. Available from: <https://www.w3.org/TR/1999/REC-rdf-syntax-19990222/>.
- [26] Sprugnoli R, Passarotti M, Cecchini FM, Pellegrini M. Overview of the EvaLatin 2020 Evaluation Campaign. In: Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages. Marseille, France: European Language Resources Association (ELRA); 2020. p. 105–110. Available from: <https://www.aclweb.org/anthology/2020.lt4hala-1.16>.
- [27] Passarotti M, Budassi M, Litta E, Ruffolo P. The Lemlat 3.0 Package for Morphological Analysis of Latin. In: Bouma G, Adesam Y, editors. Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language. vol. 133. Gothenburg: Linköping University Electronic Press; 2017. p. 24–31.
- [28] Georges KE, Georges H. Ausführliches lateinisch-deutsches Handwörterbuch. Hannover: Hahn; 1913.
- [29] Glare PG. Oxford Latin Dictionary. Oxford: Oxford University Press; 1982.
- [30] Gradenwitz O. Laterculi vocum Latinarum: voces Latinas et a fronte et a tergo ordinandas. Leipzig: Hirzel; 1904.
- [31] Budassi M, Passarotti M. Nomen Omen. Enhancing the Latin Morphological Analyser Lemlat with an Onomasticon. In: Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities. Berlin, Germany: Association for Computational Linguistics; 2016. p. 90–94.
- [32] Cecchini FM, Passarotti M, Ruffolo P, Testori M, Draetta L, Fieromonte M, et al. Enhancing the Latin Morphological Analyser LEMLAT with a Medieval Latin Glossary. In: Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018). Torino: aAccademia University Press; 2018.

- p. 87–92.
- [33] Cecchini FM, Passarotti M, Marongiu P, Zeman D. Challenges in Converting the Index Thomisticus Treebank into Universal Dependencies. In: Proceedings of the Second Workshop on Universal Dependencies (UDW 2018). Brussels, Belgium: Association for Computational Linguistics; 2018. p. 27–36.
- [34] Cecchini FM, Sprugnoli R, Moretti G, Passarotti M. UDante: First Steps Towards the Universal Dependencies Treebank of Dante’s Latin Works. In: Seventh Italian Conference on Computational Linguistics. Bologna: CEUR-WS.org; 2020. p. 1–7.
- [35] Vaan Md. Etymological Dictionary of Latin: and the other Italic Languages. Leiden and Boston: Brill; 2008. Available from: <https://brill.com/view/title/12612>.
- [36] Mambrini F, Passarotti M. Representing Etymology in the LiLa Knowledge Base of Linguistic Resources for Latin. In: Proceedings of the 2020 Globalex Workshop on Linked Lexicography. Marseille, France: European Language Resources Association; 2020. p. 20–28.
- [37] Sprugnoli R, Moretti G, Passarotti M. Towards the Modeling of Polarity in a Latin Knowledge Base. In: Adamou A, Daga E, Meroño-Peñuela A, editors. WHiSe 2020 Workshop on Humanities in the Semantic Web 2020. Heraklion, Greece: CEUR; 2020. p. 59–70.
- [38] Saalfeld GAEL. Tensaurus Italograecus: ausführliches historisch-kritisches Wörterbuch der griechischen Lehn- und Fremdwörter im Lateinischen. C. Gerold’s Sohn; 1884.
- [39] Franzini G, Zampedri F, Passarotti M, Mambrini F, G. Græcissâre: Ancient Greek Loanwords in the LiLa Knowledge Base of Linguistic Resources for Latin. In: Monti J, Dell’Orletta F, Tamburini F, editors. Proceedings of the Seventh Italian Conference on Computational Linguistics. Bologna, Italy, March 1-3, 2021. Bologna: CEUR-WS.org; 2020. p. 1–6.
- [40] Franzini G, Peverelli A, Ruffolo P, Passarotti M, Sanna H, Signoroni E, et al. Nunc Est Aestimandum. Towards an evaluation of the Latin WordNet. In: Bernardi R, Navigli R, Semeraro G, editors. Sixth Italian Conference on Computational Linguistics (CLiC-it 2019). Bari, Italy: CEUR-WS.org; 2019. p. 1–8.
- [41] Passarotti M, Saavedra BG, Onambele C. Latin Vallex. A Treebank-based Semantic Valency Lexicon for Latin. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). Portorož, Slovenia: European Language Resources Association (ELRA); 2016. p. 2599–2606.
- [42] Litta E, Passarotti M, Mambrini F. 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. Prague, Czechia: Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics; 2019. p. 35–43. Available from: <https://www.aclweb.org/anthology/W19-8505>.
- [43] Minozzi S. The Latin WordNet project. In: Anreiter P, Kienpointner M, editors. Latin Linguistics Today. Akten des 15. Internationalen Kolloquiums zur Lateinischen Linguistik. Innsbruck: Institut für Sprachen und Literaturen der Universität Innsbruck; 2010. p. 707–716.
- [44] Pianta E, Bentivogli L, Girardi C. MultiWordNet: developing an aligned multilingual database. In: Proceedings of the First International Conference on Global WordNet. Mysore, India: Global WordNet Association; 2002. p. 293–302. Available from: <http://multiwordnet.fbk.eu/paper/MWN-India-published.pdf>.
- [45] Fellbaum C. WordNet. In: Chapelle C, editor. The Encyclopedia of Applied Linguistics. Wiley Online Library; 2012. .
- [46] Lewis CT, Short C. A Latin Dictionary. Oxford: Clarendon Press; 1879.
- [47] Passarotti M. The Project of the Index Thomisticus Treebank. Digital Classical Philology. 2019;10:299–320.
- [48] Bamman D, Crane G. The Design and Use of a Latin Dependency Treebank. In: TLT 2006: Proceedings of the Fifth International Treebanks and Linguistic Theories Conference; 2006. p. 68–78.
- [49] Mikulová M, Bémová A, Hajic J, Hajičová E, Kolářová V, Kučová L, et al. Annotation on the tectogrammatical layer in the Prague Dependency Treebank. Annotation manual. Prague: UFAL; 2006. 30. Available from: <https://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/t-layer/pdf/t-man-en.pdf>.
- [50] Lopatková M, Žabokrtský Z, Benešová V. Valency Lexicon of Czech verbs VALLEX 2.0. Prague: UFAL; 2006. TR-2006-34. Available from: <https://ufal.mff.cuni.cz/~lopatkova/literatura/06-TR-vallex-2.0.pdf>.