Harmonizing Different Lemmatization Strategies for Building a Knowledge Base of Linguistic Resources for Latin

Francesco Mambrini, Marco Passarotti
CIRCSE Research Centre
Università Cattolica del Sacro Cuore
Largo Gemelli, 1 - 20123 Milan, Italy
{francesco.mambrini}{marco.passarotti}@unicatt.it

Abstract

The interoperability between lemmatized corpora of Latin and other resources that use the lemma as indexing key is hampered by the multiple lemmatization strategies that different projects adopt. In this paper we discuss how we tackle the challenges raised by harmonizing different lemmatization criteria in a project that aims to connect linguistic resources for Latin using the Linked Data paradigm. The paper introduces the architecture supporting an open-ended, lemma-based Knowledge Base, built to make textual and lexical resources for Latin interoperable. Particularly, the paper describes the inclusion into the Knowledge Base of its lexical basis, of a word formation lexicon and of a lemmatized and syntactically annotated corpus.

1 Introduction

In spite of the growth in the quantity and coverage of linguistic resources for several languages, the greatest part of these resources are still not interoperable. Lack of interoperability is an issue that severely limits their potential for exploitation and use. Indeed, linking linguistic resources to one another would maximize their contribution to, and use in linguistic analysis at multiple levels, be those lexical, morphological, syntactic, semantic or pragmatic.

Interlinking the tremendous wealth of linguistic (meta)data accumulated in more than half a century of Computational Linguistics and empirical study of language is one of the main challenges of the present time (Chiarcos et al., 2012, p. 1). However, the task is not straightforward, in particular on account of the existence of several different formalisms (e.g. various annotation schemas) or different conceptual models (e.g. different PoS tagsets) that each project may use to represent linguistic data and which are often incompatible between systems (van Erp, 2012, p. 58).

Part-of-Speech (PoS) tagging and lemmatization are key annotation tasks that are often performed to produce empirical data for research on linguist problems, to train stochastic Natural Language Processing (NLP) tools or to support the automatic processing of higher annotation levels (like, for instance, syntactic parsing). Especially for highly inflected languages (like Latin), harmonization of lemmatization and PoS tagging strategies could already promote joint exploitation, querying and interlinking of several available resources (see Section 2). Instead, annotated corpora as well as lexical resources and NLP tools show frequent problems of mismatch (see Section 3.1).

In this paper we discuss how we tackle the challenges raised by harmonizing different lemmatization criteria in the *LiLa: Linking Latin* project, which aims to make resources for Latin interoperable.[1] To this aim, the LiLa project builds a Knowledge Base of linguistic resources based on the Linked Data paradigm, i.e. a collection of several data sets described using the same vocabulary and linked together.[2] In the LiLa Knowledge Base (henceforth LiLa), lemmas are used as a pivotal node in a dense network of linguistic information, making lexical resources, NLP tools and annotated (at least, lemmatized and PoS-tagged) corpora interact. To this end, it is crucial to harmonize the different lemmatization strategies adopted so far in the currently available linguistic resources for Latin.

The LiLa project responds to the growing need in the fields of Classics, Humanities Computing and Computational Linguistics to create an inter-

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

[2] See Tim Berners-Lee's note at https://www.w3.org/DesignIssues/LinkedData.html.

Proceedings of the 13th Linguistic Annotation Workshop, pages 71–80
Florence, Italy, August 1, 2019. ©2019 Association for Computational Linguistics

operable ecosystem of resources and NLP tools for Latin. In particular, the work of harmonization of lemmatizations for Latin is motivated by two main reasons that make Latin an optimal use case: (a) the diachrony and diversity of the language present complex challenges for NLP, especially with regard to the portability of the tools across different eras, genres and domains; (b) an interconnected network of the numerous linguistic resources currently available for Latin would greatly support a large and diverse community made of historians, philologists, archaeologists and literary scholars, whose research work is strictly bound to the empirical evidence provided (also) by textual data.

This paper discusses the results of a first attempt to: (a) create and organize a collection of lemmas that would serve as a "hub" point for different resources; (b) link one annotated corpus of Latin texts to it, by solving the different harmonization problems. After providing a brief summary of the linguistic resources currently available for Latin (Section 2), we describe the LiLa Knowledge Base, particularly discussing the harmonization process of the different annotation strategies concerning lemmatization for Latin (Section 3). The inclusion into the Knowledge Base of its fundamental lexical basis, of a word formation lexicon and of a syntactically annotated corpus (a dependency treebank) is described and evaluated in Section 4. Finally, Section 5 discusses a number of open challenges to be addressed by the LiLa project in the near future.

2 Linguistic Resources for Latin

A huge amount of Latin texts is currently available in digital format. Among the most prominent providers and collections of digital texts in Latin are the *Perseus Digital Library* at Tufts University in Boston, MA,[3] the *Open Greek and Latin* project in Leipzig, Germany,[4] the *Laboratoire d'Analyse Statistique des Langues Anciennes* (LASLA) in Liège, Belgium,[5] the *Patrologia Latina* database,[6] the digital archive of Latin poetry *Musisque Deoque*,[7] the collection of Medieval Italian Latinity *ALIM*,[8] and the *Monumenta Germaniae Historica*.[9]

Despite such a large availability, only a few Latin texts are currently enhanced with linguistic annotation, while most of them still lack any linguistic tagging at all. In particular, three treebanks are currently available for Latin, all featuring also a version included in the *Universal Dependencies* (UD) collection.[10] These are the *Index Thomisticus* Treebank (IT-TB) (Passarotti, 2009), based on the works of Thomas Aquinas, the *Latin Dependency Treebank* (LDT) (Bamman and Crane, 2011), including texts of the Classical era, and the PROIEL corpus (*Pragmatic Resources in Old Indo-European Languages*), which features the syntactic annotation of the oldest extant versions of the New Testament in Indo-European languages and Latin texts of both the Classical and Late eras (Haug and Jøhndal, 2008). The size of these treebanks is presently around 350,000 annotated words for the IT-TB, 55,000 for the LDT and 200,000 for the Latin section of the PROIEL corpus.

In regards to Latin digital lexical resources, many Latin dictionaries and lexica are today available in digital format. Some of the most important are the Lewis-Short dictionary available at Perseus, the *Thesaurus Linguae Latinae* by the Bayerische Akademie der Wissenschaften in Munich,[11] and the *Neulateinische Wortliste* by Johann Ramminger.[12]

The availability of Latin treebanks made it possible to induce subcategorization lexica from the IT-TB (*IT-VaLex*) (McGillivray and Passarotti, 2009) and from the LDT (*VaLex*) (McGillivray, 2013). *Latin Vallex* is a recently created lexical resource for Latin consisting in a semantic-based valency lexicon built in conjunction with the semantic and pragmatic annotation of the IT-TB and the LDT (Passarotti et al., 2016). Presently, *Latin Vallex* includes around 1,350 lexical entries.

The *Latin WordNet* (LWN) (Minozzi, 2010) was built in the context of the *MultiWordNet* project (Pianta et al., 2002), whose aim was to build a number of semantic networks for specific languages aligned with the synsets of the Princeton WordNet (PWN) (Fellbaum, 2012). The language-specific synsets were built by importing

[3] http://www.perseus.tufts.edu/hopper/

[4] https://www.dh.uni-leipzig.de/wo/projects/open-greek-and-latin-project/

[5] http://web.philo.ulg.ac.be/lasla/

[6] http://pld.chadwyck.com/

[7] http://www.mqdq.it/

[8] http://www.alim.dfll.univr.it/

[9] http://www.dmgh.de/

[10] https://universaldependencies.org/

[11] http://www.thesaurus.badw.de/

[12] http://www.lrz-muenchen.de/~ramminger/

the semantic relations among the synsets for English provided by the PWN. At the moment, the LWN includes 8,973 synsets and 9,124 lemmas.

The recently built *Word Formation Latin* (WFL) lexicon (Litta et al., 2016) describes the Latin lexicon in terms of derivational morphology, by connecting lemmas via word formation rules.[13] For instance, the noun *amator*, "lover" is connected to the verb *amo*, "to love" via a rule that derives nouns from verbs by adding the agentive/instrumental suffix *-tor*.

The LiLa project wants to maximize the use of these (and other) resources for Latin by making them interoperable, thus allowing to run queries across linked and distributed resources, for instance making it possible to search in the three Latin treebanks all the occurrences of verbs featuring a specific (a) dependency relation (source: treebanks), (b) prefix (source: WFL) and (c) valency frame (source: Latin Vallex), and (d) belonging to a particular WordNet synset (source: LWN).

3 The LiLa Knowledge Base

In this section we present the first steps undertaken in order to structure the information of the Latin linguistic resources (and, then, NLP tools) in a centralized architecture representing the backbone of the LiLa Knowledge Base.

In order to achieve interoperability between distributed resources and tools, LiLa makes use of a set of Semantic Web and Linked Data standards and practices. These include ontologies to describe linguistic annotation (*OLiA*, Chiarcos and Sukhareva (2015)), corpus annotation (*NLP Interchange Format* (NIF), Hellmann et al. (2013); *CoNLL-RDF*, Chiarcos and Fäth (2017)) and lexical resources (*Lemon*, Buitelaar et al. (2011); *Ontolex*[14]). Furthermore, following Bird and Liberman (2001), the *Resource Description Framework* (RDF) (Lassila et al., 1998) is used to encode graph-based data structures to represent linguistic annotations in terms of triples: (1) a predicate-property (a relation; in graph terms: a labeled edge) that connects (2) a subject (a resource; in graph terms: a labeled node) with (3) its object (another resource, or a literal, e.g. a string). The SPARQL language is used to query the data recorded in the form of RDF triples (Prud'Hommeaux et al., 2008).

By applying the principles of Linked Data to linguistic resources, "it is possible to follow links between existing resources to find other, related data and exploit network effects" (Chiarcos et al., 2013, p. iii). The *Linguistic Linked Open Data cloud* (LLOD) is a good example of a set of linked linguistic resources.[15]

Publishing linguistic resources using Linked Data allows existing resources to be connected, thereby creating a web of linguistic data, which supports complex querying across different and distributed resources. Consequently, Linked Data is at the core of recent research efforts in linguistics, like the *Open Linguistic Working Group* (OLWG).[16] Moreover, applying the Linked Data paradigm to linguistic data enables to connect linguistics to other disciplines and, ultimately, to the world. As a matter of fact, Linked Data has achieved success in a wide variety of domains, like geography (Goodwin et al., 2008), biomedicine (Ashburner et al., 2000) and government data.[17]

3.1 Linking Through Lemmatization

Like for many languages, modern and early-modern Latin dictionaries index each lexical entry using a canonical form known as the lemma. Selecting the canonical forms is a fundamental annotation step, which tends to follow a standardized series of conventions (e.g. the form in nominative singular for nouns, or the first person of present tense for verbs). Thesauri, including the most modern ones like the LWN, organize the lexicon by collecting all related entries, and use the canonical form to index them; so, for instance, the synset `n#07202206` of the LWN, glossed as "a female human offspring", includes the nouns with lemmas: *filia*, "daughter", *nata*, "daughter" and *puella*, "girl". Similarly, other resources, like word formation based or valency lexica, use lemmas to group together entries that share certain features, like derivative morphemes or valency arguments.

Lemmas are also used to enable lexical search in corpora, given the very rich inflectional morphology of Latin; a regular Latin verb, for instance, can have up to 130 forms (not including the nominal inflection of the participles or gerundives), with

[13] https://github.com/CIRCSE/WFL

[14] https://www.w3.org/community/ontolex/

[15] http://linguistic-lod.org/llod-cloud

[16] http://linguistics.okfn.org

[17] https://data.gov.uk/

varying endings and, at times, different stems. Although the task of lemmatization is far from trivial just because of such rich morphology, the most accurate lemmatizers of Latin achieve an accuracy up to 95.30 (Eger et al., 2015). However, such quite high rate for automatic lemmatization of Latin must be considered carefully. Indeed, performances of stochastic NLP tools depend heavily on the training set which their models are built on, thus decreasing when they are applied to out-of-domain texts. This problem is particularly hard when Latin is concerned, because Latin texts show an enormous diversity resulting from (a) a wide time span (covering more than two millennia), (b) a large variety of genres (ranging from literary to philosophical, historical and documentary texts) and (c) a diatopic spread all over Europe (and beyond).

LiLa is highly lexically-based, grounding on a simple, but effective assumption that allows 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. Particularly, the level of lemma is considered the ideal interface between the lexical resources (dictionaries, thesauri and lexica), annotated corpora and NLP tools that lemmatize their input text. For this reason, we have identified the collection of canonical forms of Latin as the core of LiLa. Interoperability can be achieved by linking all entries in lexical resources and corpus tokens that refer to the same lemma.

The task of building and organizing a repository of canonical forms that may serve as a hub in this architecture is, however, complicated by the fact that different corpora, lexica or tools for Latin may adopt different strategies to solve conceptual and linguistic challenges posed by lemmatization. These include:

- different citation forms for the same word, resulting from alternation in (a) the graphical representation (*voluptas* vs. *uoluptas*, “satisfaction”), (b) the spelling (*sulphur* vs. *sulfur*, “brimstone”), (c) the ending (*diameter* vs. *diametros* vs. *diametrus*, “diameter”) or (d) the paradigmatic slot representing the lemma (*sequor*, “to follow”, first person singular of the passive/deponent present indicative vs. *sequo*, first person singular of the active present indicative);

- the existence of homographic lemmas, like *occido* (*occīdo* < *ob* + *caedo*, “to strike down”) vs. *occido* (*occĭdo* < *ob* + *cado*, “to fall down”);

- ambiguity in choosing the lemma: certain forms, such as participles or deadjectival adverbs, can be considered either part of the inflectional paradigm of verbs or adjectives, or independent lemmas provided with an autonomous entry in lexical resources;

- polythematic words, for which missing forms are taken from other stems, like *melior* used as comparative of *bonus* (see En. “good” and “better”).

When dealing with homographs, corpora may choose to index the different entries, but most of the times the string of the lemma is not disambiguated. Participles can either be lemmatized always under the main verb, or have a dedicated participial lemma, which in turn may be used systematically or only when the participle has grown into an autonomous lexical item (e.g. *doctus*, “learned”, morphologically the past participle of *doceo*, “to teach”). Deadjectival adverbs (e.g. *aequaliter*, “evenly” from *aequalis*, “equal”) or peculiar forms such as comparatives (both regular and irregular) are sometimes subsumed under the (positive degree of the) adjective, or given a self-standing lemma.

3.2 Lemmas and Forms. Towards an Ontology of Latin Canonical Forms

Given the challenges and the degree of variation raised by different lemmatization strategies for Latin, our approach in LiLa is to be as descriptive and inclusive as possible: our aim is rather to collect as many word forms as may be used for lemmatization and attempt to model their relations. In order to do that, LiLa builds upon a series of ontologies for lexical resources to describe the word forms used in lemmatization, and use the *Web Ontology Language* (OWL) (McGuinness et al., 2004) in order to model the relations between them.

Building upon the Ontolex ontology, we define a Lemma as a Form of a word. In this way, lexical resources compiled using the Ontolex or Lemon formalism can already be connected to our collection. Forms have one or more written representations and are linked to one or more PoS. PoS are linked to the appropriate OLiA concepts, and we plan to represent the most widespread tagsets used in Latin PoS-tagging via dedicated OLiA ontologies.

Relations between the lemma and the other forms of the same word are defined horizontally, i.e. via direct relations between forms. Although the architecture is ready to accommodate all the inflected forms of a lexical item that are either attested in a text or morphologically possible, we are currently populating it only with those forms that are potentially used as lemmas, to create the collection of canonical forms representing the core of LiLa. The fundamental list of Latin lemmas used in the Knowledge Base is taken from the one provided by the Latin morphological analyzer Lemlat (Passarotti et al., 2017).[18] In particular, following the practice of Lemlat, we define a special subclass of lemmas, called "hypolemmas", to harmonize different strategies for the lemmatization of participles. Hypolemmas are defined as forms of the inflectional paradigm of a word that may be used in annotated corpora or by NLP tools to lemmatize certain forms instead of the main lemma. Namely, these are the nominal inflected forms of verbal paradigms (participles, gerunds, gerundives, supines). Currently, we generated hypolemmas for all the canonical forms of present, future and perfect participles of all verbs in Lemlat, and connected them with their main (verbal) lemma via a subclass of the property "Form variant" defined by the Lemon ontology.[19] Thus, for instance, the present participle *subsistens*, "taking a stand" is hypolemma of the main lemma *subsisto*, "to take a stand". The same subclass is used also for alternative paradigmatic slots representing the lemma.

Alternations in spelling and ending are managed as different written representations of the same lemma, while systematic graphical variations (e.g. *u*/*v*) are preprocessed automatically.

4 Populating the LiLa Knowledge Base

In this section we present the current status of the LiLa Knowledge Base obtained by (a) including the lemma collection taken from Lemlat and (b) linking one lexical resource and one treebank, using the principles discussed in the previous Section.

The data and resources currently linked in LiLa are stored in a triple store using the Jena framework;[20] the Fuseki component exposes the data as SPARQL end-point accessible over HTTP.[21]

4.1 The Lemma Collection

As mentioned, our database of canonical forms is built on top on the lemma collection used by Lemlat. Lemlat relies on a lexical basis resulting from the collation of three Latin dictionaries (Georges and Georges, 1913–1918; Glare, 1982; Gradenwitz, 1904) for a total of 40,014 lexical entries and 43,432 lemmas, as more than one lemma can be included in one lexical entry. This lexical basis was recently enlarged by adding most of the *Onomasticon* (26,415 lemmas out of 28,178) provided by the 5th edition of the Forcellini dictionary (Budassi and Passarotti, 2016) and the entries from a large reference glossary for Medieval Latin, namely the *Glossarium Mediae et Infimae Latinitatis* (du Cange et al., 1883–1887; Cecchini et al., 2018).

In Lemlat, lemmas are annotated with up to two PoS tags expressed using the Universal PoS tagset adopted in UD (Petrov et al., 2011), as well as with other information such as the grammatical gender for nouns and the inflectional class for verbs, adjectives and nouns. While the linking between the Universal PoS tags and OLiA is already in place, the process of aligning the other morphological features is in progress.

4.2 Lexical Resources. The Word Formation Latin Lexicon

The WFL lexicon is strictly bound to Lemlat, as it enhances its lexical basis with information on derivational morphology.

The information provided by WFL can be readily linked to the lemma collection of LiLa. In the Knowledge Base, each Lemma is connected to a series of Morphemes, including at least a Lexical Base, and possibly Prefixes and Suffixes. This conceptualization yields a network representation of the morphological derivation of Latin words, where lemmas belonging to the same word formation family are linked to the same Lexical Base, which in LiLa is not assigned any written representation and functions just as a connector of the

[18] https://github.com/CIRCSE/LEMLAT3

[19] https://www.lemon-model.net/lemon-cookbook/node17.html

[20] A prototype of the LiLa triple store is available at https://lila-erc.eu/data/.

[21] https://jena.apache.org/

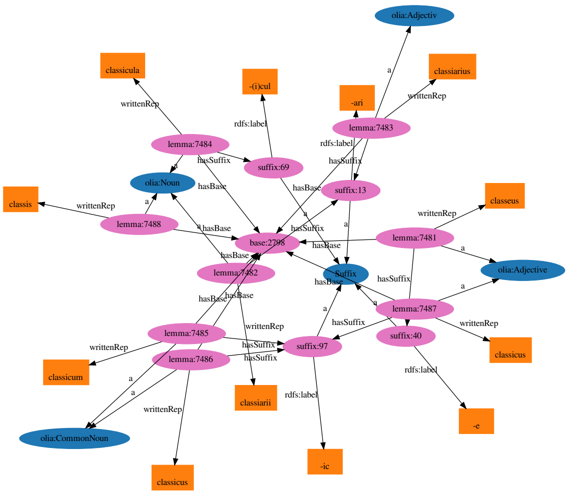

Figure 1: The word formation family including *classis* and *classicus*, with lemmas and suffixes.

lemmas of the same family; words derived with the same affixe(s) can also be readily retrieved.

Figure 1 shows a word formation family, i.e. a set of lemmas connected to a common lexical base. The family includes, among others, lemmas like noun *classis*, "class/division", and adjective *classicus*, "of the fleet/classic", the latter being derived with suffix *-ic*. By following the links to suffix 97, labelled "-ic", it would be possible to retrieve all the other lemmas that are formed with that morpheme, like for instance *ethicus*, "ethic".

4.3 Textual Resources. The PROIEL Treebank

Lemmatized Latin corpora, regardless of genre, date or provenance, are already fit to be linked to LiLa. As a preliminary experiment, we integrated one of the largest and most diverse annotated corpora of Latin, the PROIEL treebank, focusing on the version distributed in UD release 2.3. With the help of the CoNLL-RDF application, we generated an RDF graph out of the treebank, where the main nodes are the corpus tokens and sentences, as defined in NIF. Most annotations recorded in the corpus file are expressed as data attributes (strings) of the nodes, while some information (PoS, syntactic annotation) is recorded as edges between nodes.

Figure 2 gives a (simplified) representation of the nodes and relations attached to a single token in our architecture. The word *inferni* ("hell", genitive singular) from Jerome's *Vulgate* (*Revelation* 1.18) is part of sentence `proiel:s17835_0`[22] and is governed in the UD tree by the word `proiel:s17835_4` (see the attribute HEAD) via the UD relation "conj" ("conjunct"; EDGE).[23]

Although LiLa is a lexically-based resource, it integrates information about sentences if they are available in the original corpus, i.e. if the corpus, as treebanks are, is split into sentences. In this way, users have the opportunity to use also the sentence boundaries as context information for their research.

The word shares the same string in the LEMMA attribute with the written representation of one Lemma object in LiLa (`lemma:20369`; written representation: *infernus*); the two nodes point also to the same PoS concept from the OLiA ontology (CommonNoun). In this case, the token can be straightforwardly and unambiguously matched to the lemma, so that all the lexical information (currently, the links to derivational morphemes) attached to it becomes retrievable.

The figure reproduces three other lemmas that are attached to the same lexical base with id `639` (*infernalis*, "nether" and *inferiae*, "sacrifices to the dead") and the same suffix (*arcanus*, "hidden, secret") of our target word *infernus*. *Arcanus* is in fact built from the stem of *arca* ("coffin", not reported in the Figure, but retrievable following the edges) via the same suffix *-n* that produces *infernus* from *inferus* ("lower"). All this information is taken from the WFL lexicon; the image illustrates how the network of connections in LiLa can be leveraged to move from the level of lexicon to corpora and vice versa, in order to extract complex linguistic information from distributed resources. Note, therefore, that Figure 2 incorporates the type of information represented in Figure 1, although only a part of the dense network of connections can be displayed here.

4.4 Evaluation of Lemma Matching

Table 1 reports the results of our matching between the strings used in PROIEL to lemmatize the tokens and the Lemma objects in LiLa.

The PROIEL UD 2.3 corpus includes 18,400 sentences and 199,958 tokens; in total, the corpus uses 8,536 different strings (i.e. lemmas) to lemmatize them. 5,806 out of these, corresponding

[22] Full sentence: "et habeo claves mortis et inferni.". Translation: "And I hold the keys of death and of Hell.".

[23] `https://universaldependencies.org/u/dep/conj.html`

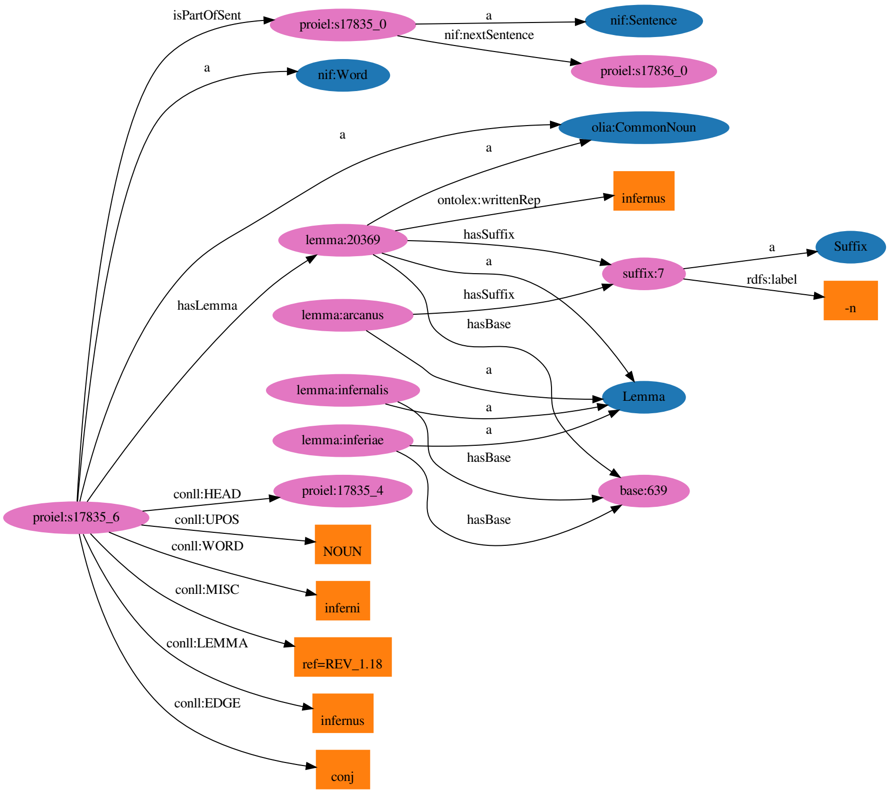

Figure 2: A token from the PROIEL UD 2.3 treebank linked to the LiLa Knowledge Base.

Type of Match	Nr. Tokens	Nr. Lemmas
String match	162,998	5,806
PoS disambig.	6,262	209
Hypolemmas	1,026	152
Onomasticon	7,252	974
Multiple matches	11,865	242
No match	10,555	1,164

Table 1: Matching scores between corpus tokens and Lemmas in the LiLa Knowledge Base.

to 162,998 tokens (81.52%), were matched unambiguously to a lemma in LiLa through a simple string comparison between the written representation of the lemma in the Knowledge Base and the PROIEL string.[24]

For 6,262 additional tokens, a single match was obtained by using the PoS tag to disambiguate between possible candidates. This is the case of the match illustrated in Figure 2, as the string *infernus* can point either to an adjective ("lower") or, as it is for the token in the Figure, to a noun ("hell"). Simple match and PoS-driven disambiguation cover together 84.64% of the PROIEL tokens.

This workflow of using PoS tags to disambiguate the greedy match of simple-string comparison is more productive than comparing tuples of string and PoS tag from the onset. For the use of tagsets of different granularity or strategy might result in loss of connection even for tokens that could be unambiguously matched. For instance, the lemma *ille* (demonstrative "that"), which occurs with a frequency of 109.62 per 10k words in PROIEL, can be matched to a single Lemma using string comparison. However, while the Lemma is tagged only as an Adjective in our collection, it is annotated as Determiner in 445 cases (vs. Adjective 1,747) in PROIEL. Those 445 tokens would not be matched if we used the tuple comparison.

After PoS-driven disambiguation, 11,865 tokens (about 6% of the total) remain associated with more than one Lemma in LiLa (with a maximum of 4 links for 5 tokens). In several cases, this is due to actual ambiguity: some high-frequency

[24] It might be the case that simple string comparison leads to wrong connections. This can happen when a lemma provided by a corpus is not present in the lexical basis of LiLa and it is homographic to one of the lemmas there included. However, we have not found such a case so far in our data.

lemmas admit multiple interpretations, even after PoS-driven disambiguation. For instance, the string *dico* (120.86 per 10k tokens in PROIEL) can be matched to two different entries with PoS Verb (one corresponding to a verb with infinitive *dicere*, "to say", the other with infinitive *dicare*, "to dedicate"), as does *tempus* (14.60 per 10k), which can be reduced to two different nouns with the same inflection (one meaning "time", the other "temple").

In the case of *omnis*, "all/every" (87.42 per 10k tokens), the multiple links point to an error in the Knowledge Base inherited from the Lemlat database: the lemma was wrongly duplicated and one of the entries must be deleted. Mismatches or multiple matches can thus provide a useful testbed to diagnose problems in the architecture.

Although the lemma collection of LiLa does not currently include the named entites of the Forcellini's Onomasticon provided by Lemlat, 7,252 proper names in PROIEL can be matched unambiguously to one entry in the Forcellini dictionary (see Section 5).

Unresolved mismatches (10,555 tokens) are due to different factors. Tokenization of the enclitic -*que*, "and" in PROIEL produces a lemma which is very frequent (86.97 per 10k words), but not yet present in LiLa (as Lemlat presupposes a different tokenization). Deadjectival adverbs (e.g. *vehementer*, "violently") are treated as lemmas in PROIEL, but reduced to their base adjective by Lemlat (e.g. *vehemens*, "violent"), so that no lemma like *vehementer* exists yet in LiLa. Deadjectival adverbs used as lemmas cover about 1,300 tokens in PROIEL. Named entities missing in the Onomasticon (e.g. *Iudaei*, "the Jewish people"), strings written with diacritics in PROIEL (e.g. *appr(eh)endo*, "to seize", corresponding to two written representations: *apprehendo* and *apprendo*), numerals and non-Latin expressions (e.g. the string: "Greek expression", used to lemmatize Greek words) also affect the matching.

5 Conclusions and Future Work

In this paper, we have introduced the fundamental components (and their relations) of a Knowledge Base, called LiLa, built to make linguistic resources for Latin interoperable according to the Linked Data paradigm.

As LiLa is highly lexically-based, we have discussed some issues concerning the repository of Latin lemmas that we have included so far therin, particularly focusing on some challenges raised by lemmatization. Indeed, one of the main tasks of LiLa is harmonizing between different strategies of linguistic annotation, namely lemmatization and PoS tagging, which currently still affects the interoperability between different annotated corpora (not only for Latin). Furthermore, we have described the inclusion in the Knowledge Base of a lexical resource (WFL) and of a treebank (PROIEL).

The LiLa project started in June 2018 and has a duration of five years. Thus, there are several open issues to address. Some of the most urgent and related to this paper are mentioned in what follows.

Given the central role played by the object Lemma in the Knowledge Base, one challenge of the project is building an efficient strategy for automatic PoS tagging and lemmatization of the (many) corpora of Latin texts still missing this level of linguistic annotation. Indeed, if connecting raw textual data to LiLa still remains possible (limiting interoperability at the level of tokens), the real added value results from exploiting the connecting power of the Lemma object in the Knowledge Base. We are now testing the already available tools and trained models on Latin texts of different eras and genres, to evaluate how much the application of these tools and models to out-of-domain texts affects their accuracy.

As mentioned, since Lemlat lemmatizes deadjectival adverbs under the adjective they are derived from, these are missing from the list of lemmas we populated LiLa with so far. However, generating all morphologically possible deadjectival adverbs from Lemlat is straightforward. Once generated, these will be included in the Knowledge Base as lemmas.

In the near future, we also plan to extend the lemma collection of LiLa with the lemmas provided by the Onomasticon of the Forcellini dictionary, as well with those by the du Cange glossary. Another short time goal is including the LWN as a key resource to support semantic-based search.

Our hope is that LiLa will help to foster the exploitation and accessibility of linguistic resources for Latin, enlarging the number of their users and impacting the diverse scholarly community concerned. Thanks to its open-ended nature, LiLa aims to become the main venue where publishing linguistic resources and, more generally, digital objects concerning the Latin cultural heritage.

Acknowledgments

We gratefully acknowledge the support of the project *LiLa* (*Linking Latin. Building a Knowledge Base of Linguistic Resources for Latin*). This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme - Grant Agreement No 769994.

References

Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. 2000. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25.

David Bamman and Gregory Crane. 2011. The ancient greek and latin dependency treebanks. In *Language technology for cultural heritage*, pages 79–98. Springer.

Steven Bird and Mark Liberman. 2001. A formal framework for linguistic annotation. *Speech communication*, 33(1-2):23–60.

Marco Budassi and Marco Passarotti. 2016. 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 (LaTeCH)*, pages 90–94, Berlin, Germany. Association for Computational Linguistics.

Paul Buitelaar, Philipp Cimiano, John McCrae, Elena Montiel-Ponsoda, and Thierry Declerck. 2011. Ontology lexicalisation: The lemon perspective. In *Proceedings of the Workshops. 9th International Conference on Terminology and Artificial Intelligence*, pages 33–36.

Flavio Cecchini, Marco Passarotti, Paolo Ruffolo, Marinella Testori, Lia Draetta, Martina Fieromonte, Annarita Liano, Costanza Marini, and Giovanni Piantanida. 2018. Enhancing the latin morphological analyser lemlat with a medieval latin glossary. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018). 10-12 December 2018, Torino*, pages 87–92.

Christian Chiarcos, Philipp Cimiano, Thierry Declerck, and John P McCrae. 2013. Linguistic linked open data (llod). introduction and overview. In *Proceedings of the 2nd Workshop on Linked Data in Linguistics (LDL-2013): Representing and linking lexicons, terminologies and other language data*, pages i–xi.

Christian Chiarcos and Christian Fäth. 2017. CoNLL-RDF: Linked Corpora Done in an NLP-Friendly Way. In *Language, Data, and Knowledge*, pages 74–88, Cham. Springer International Publishing.

Christian Chiarcos, Sebastian Hellmann, and Sebastian Nordhoff. 2012. Introduction and overview. In *Linked Data in Linguistics*, pages 1–12. Springer.

Christian Chiarcos and Maria Sukhareva. 2015. OLiA - Ontologies of Linguistic Annotation. *Semantic Web Journal*, 6(4):379–386.

Charles du Fresne du Cange, Bénédictins de Saint-Maur, Pierre Carpentier, Louis Henschel, and Léopold Favre. 1883–1887. *Glossarium mediae et infimae latinitatis*. Niort, France.

Steffen Eger, Tim vor der Brück, and Alexander Mehler. 2015. Lexicon-assisted tagging and lemmatization in latin: A comparison of six taggers and two lemmatization methods. In *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 105–113.

Marieke van Erp. 2012. Reusing linguistic resources: Tasks and goals for a linked data approach. In *Linked Data in Linguistics*, pages 57–64. Springer.

Christiane Fellbaum. 2012. Wordnet. *The Encyclopedia of Applied Linguistics*.

Karl Ernst Georges and Heinrich Georges. 1913–1918. *Ausführliches lateinisch-deutsches Handwörterbuch*. Hahn, Hannover, Germany.

Peter GW Glare. 1982. *Oxford Latin dictionary*. Clarendon Press. Oxford University Press, Oxford, UK.

John Goodwin, Catherine Dolbear, and Glen Hart. 2008. Geographical linked data: The administrative geography of great britain on the semantic web. *Transactions in GIS*, 12:19–30.

Otto Gradenwitz. 1904. *Laterculi Vocum Latinarum: voces Latinas et a fronte et a tergo ordinandas*. Hirzel, Leipzig, Germany.

Dag TT Haug and Marius Jøhndal. 2008. Creating a parallel treebank of the old indo-european bible translations. In *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, pages 27–34.

Sebastian Hellmann, Jens Lehmann, Sören Auer, and Martin Brümmer. 2013. Integrating NLP using Linked Data. In *12th International Semantic Web Conference, Sydney, Australia, October 21-25, 2013*.

Ora Lassila, Ralph R. Swick, World Wide, and Web Consortium. 1998. Resource description framework (rdf) model and syntax specification.

Eleonora Litta, Marco Passarotti, and Chris Culy. 2016. Formatio formosa est. building a word formation lexicon for latin. In *Proceedings of the third italian conference on computational linguistics (clic–it 2016)*, pages 185–189.

Barbara McGillivray. 2013. *Methods in Latin computational linguistics*. Brill.

Barbara McGillivray and Marco Passarotti. 2009. The development of the “index thomisticus” treebank valency lexicon. In *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education (LaTeCH–SHELT&R 2009)*, pages 43–50.

Deborah L McGuinness, Frank Van Harmelen, et al. 2004. Owl web ontology language overview. *W3C recommendation*, 10(10):2004.

Stefano Minozzi. 2010. The latin wordnet project. In *Latin Linguistics Today. Latin Linguistics Today. Akten des 15. Internationalen Kolloquiums zur Lateinischen Linguistik*, pages 707–716.

Marco Passarotti. 2009. Theory and practice of corpus annotation in the index thomisticus treebank. *Lexis*, 27(A):5–23.

Marco Passarotti, Marco Budassi, Eleonora Litta, and Paolo Ruffolo. 2017. The lemlat 3.0 package for morphological analysis of latin. In *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*, 133, pages 24–31. Linköping University Electronic Press.

Marco Passarotti, Berta González Saavedra, and Christophe Onambele. 2016. Latin vallex. a treebank-based semantic valency lexicon for latin. In *LREC*.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2011. A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086*.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. Multiwordnet: developing an aligned multilingual database. In *First international conference on global WordNet*, pages 293–302.

Eric Prud’Hommeaux, Andy Seaborne, et al. 2008. Sparql query language for rdf. w3c. *Internet: https://www.w3.org/TR/rdf-sparql-query/[Accessed on February 27th, 2019].*