SSL

Interlinking through lemmas. The lexical collection of the *LiLa* Knowledge Base of linguistic resources for Latin

Marco Passarotti, Francesco Mambrini, Greta Franzini, Flavio Massimiliano Cecchini, Eleonora Litta, Giovanni Moretti, Paolo Ruffolo, Rachele Sprugnoli

Abstract

This paper presents the structure of the *LiLa* Knowledge Base, i.e. a collection of multifarious linguistic resources for Latin described with the same vocabulary of knowledge description and interlinked according to the principles of the so-called Linked Data paradigm. Following its highly lexically based nature, the core of the *LiLa* Knowledge Base consists of a large collection of Latin lemmas, serving as the backbone to achieve interoperability between the resources, by linking all those entries in lexical resources and tokens in corpora that point to the same lemma. After detailing the architecture supporting *LiLa*, the paper particularly focusses on how we approach the challenges raised by harmonizing different strategies of lemmatization that can be found in linguistic resources for Latin. As an example of the process to connect a linguistic resource to *LiLa*, the inclusion in the Knowledge Base of a dependency treebank is described and evaluated.

Keywords: linguistic resources, linguistic linked open data, lemmatization, interoperability, Latin.

1. *Introduction*

Linguistic resources are machine-readable collections of language data and descriptions typically divided into two categories depending on the kind of content they include: (i) textual resources, such as written and spoken corpora, featuring either partial or full texts of various typologies, which may differ in genre, author or time period, and (ii) lexical resources, for instance lexica, dictionaries and terminological databases, providing information on lexical items for one or more languages, including definitions, translations and morphological properties. In most cases, linguistic resources do not only feature data, namely texts and lists of lexical items, but also metadata, which enhance the resource with a medley of annotations ranging from descriptive information (e.g. structural division into books, chapters, etc.) to linguistic

Received: January 2020
Accepted: April 2020

traits, such as lemmatization, Part-of-Speech (PoS) tags and syntactic function.

Over the past two decades the research area dedicated to building, improving and evaluating linguistic resources has seen substantial growth and, today, covers a wide span of languages and language varieties. This progress speaks to the need of larger (meta)data sets to support empirically-based studies and to the fact that most (stochastic) systems, tools or algorithms for Natural Language Processing (NLP) currently rely on the linguistic and meta-linguistic evidence stored in corpora or lexica. The strict relation holding between NLP tools and linguistic resources is twofold. On the one hand, NLP tools exploit the empirical data provided by resources to build trained models, whose accuracy rates heavily depend on the size (and quality) of the training data. On the other, the development of new resources, as well as the extension of existing ones, is supported by NLP tools, which automatically enrich (textual or lexical) data with linguistic metadata.

Despite the increase in the quantity and coverage of linguistic resources, most of these are locked in data silos, which prevent users from honing both their individual and joint potential in interoperable ways. While resources tend to focus on providing annotation at one or more levels of linguistic analysis – be those lexical, morphological, syntactic, semantic or pragmatic – linking them to one another helps to draw the overall picture and to maximize their individual contribution. Indeed, linguistic data and metadata today are scattered in distributed resources, thus failing to provide a comprehensive overview of the annotations available in these separate collections. One of the main challenges at the present time is interlinking the motley amount of linguistic data and metadata stored in the resources developed over the past five decades of Computational Linguistics and empirical language studies (Chiarcos *et al.*, 2012: 1). Overcoming this challenge is no simple task because: (a) linguistic resources are often designed for particular tasks (e.g. PoS tagging and syntactic analysis); (b) linguistic resources and NLP tools may use different conceptual models (e.g. different PoS tagsets); (c) linguistic data might be represented using different formalisms (e.g. annotation schemas), which are often incompatible between systems (van Erp, 2012: 58).

We owe this predicament to the fact that, throughout the years, more attention has been given to making linguistic resources grow in size, complexity and diversity, rather than making them interact. Tentative solutions

to the problem of resource isolation, such as the CLARIN[1], DARIAH[2] and META-SHARE[3] linguistic infrastructures and databases, are but upshots of the last decade. What these initiatives provide, however, is a single query access point to multiple meta-collections of resources and tools, rather than connections between them. Instead, making linguistic resources interoperable requires that all types of annotation applied to a particular word/text be integrated into a common representation for indiscriminate access to any linguistic information provided by a resource or tool (Chiarcos, 2012: 162).

A current approach to interlinking linguistic resources takes up Linked Data principles, so that «it is possible to follow links between existing resources to find other, related data and exploit network effects» (Chiarcos *et al.*, 2013: iii). According to the Linked Data paradigm, data in the Semantic Web (Berners-Lee *et al.*, 2001) 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 the area of linguistic resources, the Linguistic Linked Open Data cloud (*LLOD*)[4] is a collaborative effort pursued by several members of the Open Linguistics Working Group[5], with the general goal of developing a Linked Open Data (sub-)cloud of linguistic resources (McCrae *et al.*, 2016). Indeed, the application of Linked Data to linguistic data ultimately connects Linguistics to other domains that have adopted the paradigm, including Geography (Goodwin *et al.*, 2008), Biomedicine (Ashburner *et al.*, 2000) and Government[6].

What this fervent area of research still lacks, however, is a fine-grained level of interaction between linguistic resources capable of stretching beyond descriptive metadata over to individual word occurrences in a text or entries in a lexicon.

One subfield that has enjoyed particular prosperity over the past decade is that devoted to ancient languages. Owing to their key role in accessing and understanding the so-called Classical tradition, Latin and Ancient Greek are among the main beneficiaries.

Although Latin was among the first languages to be automatically processed with computers thanks to the pioneering work on the texts of

[1] Cf. *https://www.clarin.eu/*.
[2] Cf. *https://www.dariah.eu/*.
[3] Cf. *http://www.meta-share.org/*.
[4] Cf. *http://linguistic-lod.org/llod-cloud*.
[5] Cf. *https://linguistics.okfn.org/index.html*.
[6] Cf. *https://data.gov.uk/*.

Thomas Aquinas by the Italian Jesuit Roberto Busa in the 1940s, throughout its sixty-year history Computational Linguistics has been mainly focusing on living languages, whose commercial and social impact is larger than that of their dead counterparts. In 2006, however, the launch of two independent (but related) projects aimed at building the first syntactically annotated corpora (called 'treebanks') for Latin brought about a research renaissance of linguistic resources and NLP tools for ancient languages (Bamman *et al.*, 2008). This came as no surprise given the vast amount of texts written in Latin spread all over Europe and covering a time span of almost two millennia. These texts bear testament to a common yet diverse past and have contributed to shaping European cultural heritage. Making full use of the most advanced techniques for preserving, investigating and sharing this legacy is at the same time a challenge and an obligation for the research community.

Thanks to international efforts, several textual and lexical resources, as well as NLP tools, are currently available for Latin. Despite the launch of a number of projects for automatic extraction of structured knowledge from ancient sources in the last decade (see Section 2), much like other languages, linguistic resources and tools for Latin often live in isolation, a condition which prevents them from benefiting a large research community of historians, philologists, archaeologists and literary scholars.

To this end, the *LiLa*: Linking Latin project (2018-2023)[7] was awarded funding from the European Research Council (*ERC*) to build a Knowledge Base of linguistic resources for Latin based on the Linked Data paradigm, i.e. a collection of multifarious, interlinked data sets described with the same vocabulary of knowledge description (by using common data categories and ontologies). The project's ultimate goal is to make full use of the linguistic resources and NLP tools for Latin developed thus far, in order to bridge the gap between raw language data, NLP and knowledge description (Declerck *et al.*, 2012: 111).

This paper presents the structure of the lexical basis of *LiLa*, which serves as the backbone of the Knowledge Base to achieve interoperability between textual and lexical resources for Latin. Following a summary of the linguistic resources currently available for Latin (Section 2), we detail the architecture supporting *LiLa*, with special focus on how we approach the challenges raised by harmonizing different strategies of lemmatization

[7] Cf. *https://lila-erc.eu*.

(Section 3). The inclusion in *LiLa* of a dependency treebank is described and evaluated in Section 4. Finally, Section 5 discusses a number of open questions to be addressed by the project in the near future.

2. *Linguistic resources for Latin*

A wealth of linguistic resources is digitally available for Latin today as a result of decades' worth of work spent turning paper-based textual and lexical data into machine-readable formats. This section seeks to provide a brief overview of these efforts to delineate the quantity and diversity of the linguistic data currently at our disposal.

With regard to textual resources, among the most prominent collections of digital texts are the Perseus Digital Library[8], the corpus of Latin texts developed by the Laboratoire d'Analyse Statistique des Langues Anciennes (*L.A.S.L.A.*)[9], the Bibliotheca Teubneriana Latina by De Gruyter[10], the collection of Classical Latin texts prepared by the Packard Humanities Institute (*PHI*)[11], the Loeb Classical Library[12], and a set of collections published by Brepols, such as the Library of Latin Texts[13], the Archive of Celtic Latin Literature[14] and the Aristoteles Latinus Database[15]. More recently, the Digital Latin Library project[16] set out to publish and curate critical editions of Latin texts of all types, genres and eras. A similar objective is pursued by the Open Greek and Latin project[17], whose ultimate goal is to represent every source text produced in Classical Greek or Latin in Antiquity (through c. 600 AD) with a view to covering also the Post-classical era until modern times. The project places the total number of Ancient Greek and Latin words surviving from Antiquity at 150 million, and the number of Post-classical Latin words available in some 10,000 books in the Internet Archive at 200 million.

[8] Cf. *http://www.perseus.tufts.edu/hopper/*.
[9] Cf. *http://web.philo.ulg.ac.be/lasla/*.
[10] Cf. *https://www.degruyter.com/view/db/btl*.
[11] Cf. *http://latin.packhum.org/*.
[12] Cf. *https://www.loebclassics.com/*.
[13] Cf. *https://about.brepolis.net/library-of-latin-texts/*.
[14] Cf. *http://www.brepols.net/Pages/BrowseBySeries.aspx?TreeSeries=ACLL-O*.
[15] Cf. *http://www.brepols.net/Pages/BrowseBySeries.aspx?TreeSeries=ALD*.
[16] Cf. *https://digitallatin.org/*.
[17] Cf. *http://www.opengreekandlatin.org/*.

By virtue of infrastructural efforts conducted over the past decade, this large amount of textual resources is now accessible via aggregating initiatives, including Corpus Corporum[18], a meta-collection containing more than 150 million words in texts written in Ancient Greek or Latin provided by more than twenty different corpora and collections; Trismegistos[19], a portal of papyrological and epigraphical resources formerly covering Egypt and the Nile valley (800 BC-800 AD) and now expanding to the Ancient World in general; and the eAqua project[20], conceived to support the search of co-occurrences and citations in a number of collections of Ancient Greek and Latin texts, including Perseus and *PHI*.

Beside these catchall (meta)collections comprising large number of texts, genres and authors diachronically spread from Antiquity to Neo-Latin, some corpora provide more specific data. The Patrologia Latina database[21], for instance, features more than 100 million words from the writings of the Church Fathers; the Musisque Deoque digital archive[22] contains poetic works by some 200 authors; late-antique Latin texts are made available by the *digilibLT* Digital Library, which currently boasts 265 works written before the 6th century AD[23]; the Corpus Grammaticorum Latinorum[24] gathers the *Grammatici Latini*, that is, Latin grammar manuals written between the 2nd and 7th centuries AD and edited by Heinrich Keil in Leipzig from 1855 to 1880. As for Medieval Latin, the Index Thomisticus by father Roberto Busa SJ (Busa, 1974-1980)[25] collects the opera omnia of Thomas Aquinas, for a total of over 11 million words, the *ALIM* corpus[26] provides texts of the Italian Latinity of the Middle Ages, and the Computational Historical Semantics project[27] is a large database of Medieval Latin texts from various sources.

Among other distinctive digital corpora for Latin, noteworthy examples are the School of Salamanca[28], a digital text corpus of 116 works of Sal-

[18] Cf. *http://www.mlat.uzh.ch/MLS/*.
[19] Cf. *https://www.trismegistos.org/index.php*.
[20] Cf. *http://www.eaqua.net/*.
[21] Cf. *http://pld.chadwyck.co.uk/*.
[22] Cf. *http://mizar.unive.it/mqdq/public/*.
[23] Cf. *http://digiliblt.lett.unipmn.it/index.php*.
[24] Cf. *https://bibliotheque.univ-paris-diderot.fr/bases-de-donnees/cgl-corpus-grammaticorum-latinorum*.
[25] Cf. *http://www.corpusthomisticum.org/*.
[26] Cf. *http://www.alim.dfll.univr.it/*.
[27] Cf. *https://www.comphistsem.org/home.html*.
[28] Cf. *https://www.salamanca.school/en/works.html*.

mantine jurists and theologians found in selected printed books published between the 16th-17th centuries; the *CroALa* corpus brings together some 450 writings by 181 Croatian Latin authors, for a total of over 5 million words produced between the 10th and 20th centuries[29], the Domus sermonum compilatorium archive[30] provides the texts of the sermons of the Franciscan preacher Osvladus de Lasko; the Roman Inscriptions of Britain[31] hosts multiple corpora, including the Vindolanda tablets; Epistolae[32] is a collection of medieval Latin letters written between the 4th and 13th centuries to and from women; DanteSearch[33] provides both the vernacular and the Latin writings of Dante Alighieri, the Latin portion of the corpus counting approximately 46,000 words; finally, *CLaSSES*[34] is a collection of more than 1,200 non-literary Latin texts, such as epigraphs and letters, from different eras (between the 4th century BC and the 6th century AD) and sources (Rome, Central Italy, Britain, Egypt and the Eastern Mediterranean Sea).

A subset of the Latin texts carries linguistic annotation. The most common layer of linguistic annotation available in Latin corpora is lemmatization, which in some cases is also enriched with PoS and morphological tagging. For instance, while the data provided by *CLaSSES* and Roman Inscriptions from Britain are lemmatized, the large collection of texts assembled by *L.A.S.L.A.*, the Index Thomisticus, DanteSearch, as well as roughly one million tokens of the Computational Historical Semantics corpus are all fully lemmatized and morphologically tagged.

Syntactic annotation, on the other hand, is still limited to a small set of texts. Four treebanks are currently available for Latin. These are: (i) the Index Thomisticus Treebank (*IT-TB*) (Passarotti, 2019), based on the works of Thomas Aquinas; (ii) the Latin Dependency Treebank (*LDT*) (Bamman and Crane, 2006) of texts belonging to the Classical era, now part of the Ancient Greek and Latin Dependency Treebank 2.0 under development at the University of Leipzig (Celano, 2019); (iii) 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 from

[29] Cf. *http://www.ffzg.unizg.hr/klafil/croala/*.
[30] Cf. *http://sermones.elte.hu/szovegkiadasok/latinul/laskaiosvat/index.php?file=os_index*.
[31] Cf. *https://romaninscriptionsofbritain.org/*.
[32] Cf. *https://epistolae.ctl.columbia.edu/*.
[33] Cf. *http://www.perunaenciclopediadantescadigitale.eu:8080/dantesearch/*.
[34] Cf. *http://classes-latin-linguistics.fileli.unipi.it/*.

both the Classical and Late eras (Haug and Jøhndal, 2008); and (iv) the Late Latin Charter Treebank (*LLCT*), a syntactically annotated corpus of original 8th-9th century charters from Central Italy (Korkiakangas and Passarotti, 2011). While the *LDT*, the *IT-TB* and the *LLCT* have shared the same syntactic annotation schema since their inception (Bamman *et al.*, 2007), resembling that of the so-called analytical layer of annotation of the Prague Dependency Treebank for Czech (Hajič *et al.*, 1999), the *PROIEL* treebank follows a slightly different style (Haug, 2010). At present, with the exception of the *LLCT*, all Latin treebanks are also available in the Universal Dependencies collection (*UD*) (Nivre *et al.*, 2016)[35]. In terms of size, the *IT-TB* currently counts some 350,000 annotated words, *LDT* counts 55,000, the Latin section of the *PROIEL* corpus 200,000 and *LLCT* counts 250,000 annotated words.

With regard to lexical resources, among the many dictionaries and lexica available in digital format today are the Lewis and Short dictionary accessible through Perseus, the Thesaurus Linguae Latinae of the Bayerische Akademie der Wissenschaften in Munich[36], and Johann Ramminger's Neulateinische Wortliste[37]. Brepols provides an extensive list of Latin word forms, known as Thesaurus Formarum Totius Latinitatis[38], with number of occurrences for each in the Library of Latin Texts, and the comprehensive Database of Latin Dictionaries[39], which itself consists of a large number of different types of lexical resources. Another noteworthy initiative is Logeion[40], a cross-dictionary search tool, providing simultaneous lookup of entries in the many lemmatized works from the Perseus Classical collection by way of the PhiloLogic system[41]. Within the Computational Historical Semantics project there is the Frankfurt Latin Lexicon, a lexical resource built upon assorted source lexicons and taggers and used for NLP tasks, such as morphological tagging, lemmatization, and PoS tagging[42].

The availability of Latin treebanks has made it possible to induce subcategorization lexica from the *IT-TB* (*IT-VaLex*) (McGillivray and Passarotti, 2009) and the *LDT* (*VaLex*) (McGillivray, 2013). Latin Vallex is a valency

[35] Cf. *https://universaldependencies.org/.*
[36] Cf. *https://www.degruyter.com/view/db/tll.*
[37] Cf. *http://www.neulatein.de/.*
[38] Cf. *http://www.brepols.net/Pages/BrowseBySeries.aspx?TreeSeries=TF.*
[39] Cf. *https://about.brepolis.net/database-of-latin-dictionaries/.*
[40] Cf. *https://logeion.uchicago.edu/.*
[41] Cf. *http://philologic.uchicago.edu/.*
[42] Cf. *https://www.comphistsem.org/lexicon0.html.*

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 LatinWordNet (*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)[43]. The language-specific synsets were created by translating *PWN* data with the help of bilingual dictionaries. The *LWN* counts 8,973 synsets and 9,124 lemmas, and is currently undergoing substantial revision with a view to refining and extending its contents (Franzini *et al.*, 2019). The Word Formation Latin (*WFL*) lexicon (Litta and Passarotti, 2019) provides information about derivational morphology by connecting lemmas via word formation rules[44].

LiLa seeks to maximize the use of these (and many other) resources for Latin by making them interoperable, thus allowing users to run complex queries across linked and distributed resources, like, for instance, searching the four Latin treebanks for occurrences of verbs featuring a specific (a) dependency relation, e.g. subject (source: treebanks), (b) prefix (source: *WFL*), (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 describe the architecture of the *LiLa* Knowledge Base, built to structure the information of the Latin linguistic resources in a centralized hub of interaction.

In order to achieve interoperability between distributed resources, *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: McCrae *et al.*, 2017).

[43] Synsets are unordered sets of cognitive synonyms, i.e. words that denote the same concept and are interchangeable in many contexts. In WordNets, synsets are interlinked by means of conceptual-semantic and lexical relations.

[44] Cf. *http://wfl.marginalia.it/*.

Following Bird and Liberman (2001), the Resource Description Framework (*RDF*) (Lassila and Swick, 1998) 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). The *SPARQL* language is used to query the data recorded in the form of *RDF* triples (Prud'Hommeaux and Seaborne, 2008).

3.1. *Linking through lemmatization*

Lemmatization is a layer of annotation and organization of linguistic data common to different kinds of resources. Dictionaries tend to index lexical entries using lemmas. Thesauri organize the lexicon by collecting all related entries, and use lemmas to index them; so, for instance, the nominal synset n#07202206 of the *PWN*, glossed as "a female human offspring", is lexicalized in *LWN* by the lemmas: *filia* "daughter", *nata* "daughter" and *puella* "girl". Lemmas are also used to facilitate lexical search in corpora. This is particularly helpful for languages, like Latin, with rich inflectional morphology; a regular Latin verb, for instance, can have up to 130 forms (if we exclude the nominal inflection of the participles or gerundives), with varying endings and, at times, different stems.

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 to 95.30%, as per Eger *et al.*, 2015)[45], *LiLa* uses the lemma as the most productive interface between lexical resources, annotated corpora and NLP tools. Consequently, the *LiLa* Knowledge Base is highly lexically based, grounding on a simple, but effective assumption that strikes a good balance between feasibility and granularity: textual resources

[45] Such high rates of automatic lemmatization of Latin should be taken with a grain of salt. Indeed, performances of stochastic NLP tools heavily depend on the training set on which their models are built, and so decrease when they are applied to out-of-domain texts. This problem is particularly challenging for Latin owing to its wide diachrony (spanning two millennia), genre diversity (ranging from literary to philosophical, historical and documentary texts) and diatopy (Europe and beyond). For the state of the art in automatic lemmatization and PoS tagging for Latin, see the results of the first edition of *EvaLatin*, a campaign devoted to the evaluation of NLP tools for Latin (Sprugnoli *et al.*, 2020). The first edition of *EvaLatin* focused on two shared tasks (i.e. lemmatization and PoS tagging), each featuring three sub-tasks (i.e. Classical, Cross-Genre, Cross-Time). These sub-tasks were specifically designed to measure the impact of genre variation and diachrony on NLP tool performances.

are made of (occurrences of) words, lexical resources describe properties of words, and NLP tools process words.

Figure 1 presents the main components of the *LiLa* Knowledge Base, showing the key interlinking role played by the Lemma node. A 'Lemma' is an (inflected) 'Form' chosen as the citation/canonical form of a lexical item. Lemmas occur in 'Lexical Resources' as citation/canonical forms of lexical entries. Forms, too, can occur in lexical resources, like in a lexicon containing all of the forms of a language (for instance, Tombeur, 1998). Both Lemmas and Forms can have 'Morphological Features', such as PoS, gender, mood and tense. The occurrences of Forms in real texts are 'Tokens', which are provided by 'Textual Resources'. Finally, on NLP tools performances can process either Textual Resources (e.g. a tokenizer), Forms, regardless of their contextual use (e.g. a morphological analyzer), or Tokens (e.g. a PoS tagger).

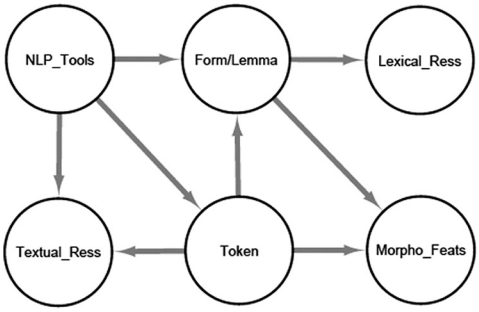

Figure 1. *The main components of* LiLa.

The core of the *LiLa* Knowledge Base 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. While the process of selecting the canonical forms to be used as lemmas tends to follow a standardized series of language-dependent conventions (e.g. for Latin, the nominative singular form for nouns, or the first person singular of the active indicative present tense for verbs), building and structuring a repository of canonical forms that may serve as a hub in *LiLa* is complicated by the fact that different corpora, lexica and tools adopt different strategies to solve the conceptual and linguistic challenges posed by lemmatization, namely (a) the form of the lemma and (b) lemmatization criteria.

Citation forms for the same lexical item chosen to represent the lemma differ in (a) graphical representation (*voluptas* vs *uoluptas* "satisfaction"), (b)

spelling (*sulfur* vs *sulphur* "brimstone"), (c) ending and possibly inflectional type (*diameter* vs *diametros* vs *diametrus* "diameter"), or (d) in 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). Furthermore, homographic lemmas, like *occīdo* (*ob+caedo* "to strike down") and *occĭdo* (*ob+cado* "to fall down"), can either be left ambiguous by using the same character string *occido* for the forms of both lemmas, or told apart. For instance, in the Index Thomisticus corpus *occīdo* and *occĭdo* are recorded as *occido^caedo* and *occido^cado*, respectively, while in the *LDT* (and in the Perseus Digital Library in general) as *occido1* and *occido2*.

As for lemmatization criteria, differences are such that, on occasion, a word form can be reduced to multiple lemmas. This is the case of participles, which can be considered either as part of the verbal inflectional paradigm or as independent lemmas deserving of a separate entry in lexical resources. Accordingly, participles can either be lemmatized under the main verb or under a dedicated participial lemma, which in turn may be used either systematically or only when the participle has grown into an autonomous lexical item (e.g. *doctus* "learned", morphologically the perfect participle of *doceo* "to teach"). The same holds true for deadjectival adverbs (e.g. *aequaliter* "evenly" from *aequalis* "equal"), which are either lemmatized as forms of their base adjective, as happens in the *IT-TB*, or treated as independent lemmas, like in the *PROIEL* treebank. Another issue is raised by polythematic words for which missing forms are taken from other stems, as is the case of *melior* used as the comparative of *bonus* (see English "good" and "better"). These are sometimes subsumed under the (positive degree of the) adjective or given a self-standing lemma.

3.2. *The* LiLa *ontology of Latin canonical forms*

Cases like the disambiguation of the ambiguous forms *occīdo* and *occĭdo* attest to the variety of lemmatization solutions different resources may adopt. In this respect, it is important to note that the approach of *LiLa* is not to harmonize resources by choosing one lemmatization standard over another or by imposing prescriptive guidelines to which all lemmatized resources must be converted. Rather, *LiLa* aims to provide a descriptive set of concepts and properties capable of integrating *all* solutions adopted by different Latin resources.

To this end, *LiLa* implements a formal ontology, expressed in the Web Ontology Language (*OWL*; McGuinness and Van Harmelen, 2004), that defines the classes, properties and instances involved in the task of lemmatization, as well as the possible interactions between lemmas, lemmatized corpora and lexica. Since the ultimate goal of the project is to establish a network of linguistic resources fully interoperable within the *LLOD* cloud, this ontology reuses as many existing standards as possible. In this way, we ensure that the data amassed by *LiLa* are immediately compatible with other Linked (Open) Data resources.

The *LiLa* ontology starts by defining the class of the Lemma, the pivotal concept in our domain. In our definition, lemmatization is the task of indexing all inflected forms under one that is conventionally identified as canonical. As such, the Lemma is safely subsumed under the general class of Form as defined in the Ontolex ontology, a *de facto* standard in the Linked Data publication of lexical resources. Relying on the concepts of Ontolex, we define the Lemma as a Form that is linked to a Lexical Entry via the property 'canonical form'. This structural choice allows us to potentially connect all other lexical resources compiled using the Ontolex (or Lemon) formalism to our collection.

Forms are grammatical realizations of words or of any other class of Lexical Entries that have at least one written representation. The Ontolex 'written representation' property can be used to accommodate the different spellings or peculiar inflections of canonical forms: in the case of the examples discussed above, *sulfur* and *sulphur* become two written representations of the same lemma, and so do the loan words that display either the Greek or the Latin endings (like *diametros* and *diametrus*)[46]. We, therefore, use this property whenever the variation in the realization of a lemma affects only the orthography of a form (including the word ending), provided that its morphological analysis and the inflectional paradigm are not altered.

What Ontolex also permits is the inclusion of a phonetic representation of a form. As vocalic quantity is often used to disambiguate between homographic words (again, *occīdo* and *occĭdo*), we add a special sub-property for prosodic representation, which carries all the relevant transcriptions of a form with long and short vowel diacritics. The variation, however, may involve changes in PoS, inflectional paradigm or other morphological features.

[46] But note that if the variation also entails a different type of inflection (such as *diameter* on the one hand and *diametrus/diametros* on the other), we represent the lemmas as two different forms linked to one another via the property 'lemma variant' (see below).

Some Latin words belong to more than one PoS, as is, for example, the case of prepositions that can be used as adverbs. Since Lexical Entries in Ontolex cannot have more than one PoS[47], the same restriction applies also to canonical forms. Accordingly, *LiLa* will provide two lemmas with written representation *ante* "before", one for the preposition and one for the adverb.

Participles and inflectional variation are harder to model and require an extension of the Ontolex ontology. Some words present two or more alternative inflectional paradigms, which entail different lemmas. Verbs with both a deponent and an active inflection, for example, are often found in Latin lexica. Although one of the paradigms might be more frequent and more 'regular' than another from a traditional lexicography or grammar standpoint[48], we cannot exclude that corpora in which the 'irregular' instances are met lemmatize these under the less typical canonical form. As a consequence, *LiLa* records all possible canonical forms as lemmas; so, in our collection, the verbs *sequor*[49] and *sequo*[50], for example, exist as independent lemmas. Since these forms can both be used to lemmatize instances of the same words, we link them to one another with the symmetric property 'lemma variant', thus making it possible to retrieve from the textual resources connected to *LiLa* all the tokens that belong to the same lexical item, regardless of the lemmatization criteria followed in individual corpora.

Participles, again, behave differently. As previously mentioned, participles like *docti* "learned" can be reduced to a form of either *doceo* "to learn" or *doctus* "learned". In these cases, that is, whenever a form can be interpreted as part of the (regular) inflectional paradigm or as a Lemma in itself, we associate that form to a special sub-class of Lemma called Hypolemma. Hyper- and hypolemmas are linked to one another via the symmetric property 'has hypolemma'/'is hypolemma'[51].

A Lemma is also defined by a series of morphological features. All lemmas are assigned a PoS (which, as we have already seen, must be exclusive for each form), and can be analyzed by those traits that are typical of nominal (gender, number, case), adjectival (gender, number, case, degree) and verbal

[47] See the definition of Lexical Entry at *https://www.w3.org/2016/05/ontolex/#lexical-entries*.

[48] In the case of verbs *sequor*/*sequo*, the active form *sequo* is mentioned by grammarians only: see Gell. 18.9.8 and Prisc. *Ars Gramm.* 9.28.

[49] Cf. *https://lila-erc.eu/lodview/data/id/lemma/124461*.

[50] Cf. *https://lila-erc.eu/data/id/lemma/124462*.

[51] Note that, with respect to its hyperlemma, a hypolemma entails a change in the PoS: *faciliter* "easily" is an adverb, while *facilis* "easy" is an adjective; *doctus* (as an autonomous lemma) is an adjective, while *doceo* is a verb.

(tense, mood, person, number, voice) inflection; additionally, lemmas have an inflectional type (i.e. the conjugations and declensions of traditional grammars). *LiLa*'s ontology formalizes these linguistic properties together with the relevant restrictions, so that, for instance, tense cannot be predicated of nouns. The PoS tags adopted in *LiLa* are based on the universal tagset of Universal Dependencies (Petrov *et al.*, 2011). However, in order to ensure compatibility with other tagsets used for Latin, *LiLa*'s categories for linguistic annotation are aligned with the *OLiA* ontology. So, for instance, *LiLa*'s class 'Adjective' is a sub-class of *OLiA*'s 'Adjective', which also subsumes all other tags used to annotate the same grammatical category.

Lemmas can also be analyzed in terms of their derivational morphology. This level integrates the information recorded in the *WFL* lexicon into the *LiLa* collection. Since an Ontolex extension for derivational morphology is currently under development, this module is still not available for immediate deploying. Ontolex allows lexical resources to describe derivational morphemes as regular lexical entries, provided with written representations. However, for our ontology, we opted for a minimal extension only. In *LiLa*, morphemes belong to their own class, and are grouped into Affixes (distinguishing between prefixes and suffixes) and Bases. We define the Base as the lexical morpheme of a word that is neither a prefix nor a suffix. Words that are derived, even in several steps, from the same root (for instance, *adduco* "to lead to", *adductio* "bringing in", *duco* "to lead", *produco* "to lead forth" and *productivus* "productive") are therefore linked to the same base.

This conceptual architecture was first put to the test with a comprehensive list of Latin canonical forms based on the one provided by the Latin morphological analyzer *Lemlat* (Passarotti *et al.*, 2017), which was used to populate the *LiLa* collection[52]. *Lemlat*'s database reconciles three reference dictionaries for Classical Latin (*GGG*: Georges and Georges, 1913-1918; Glare, 1982; Gradenwitz, 1904)[53], the entire *Onomasticon* from Forcellini's (1940) *Lexicon Totius Latinitatis* (Budassi and Passarotti, 2016) and the Medieval Latin *Glossarium Mediae et Infimae Latinitatis* by du Cange *et al.* (1883-1887), for a total of over 150,000 lemmas (Cecchini *et al.*, 2018b).

[52] Cf. *https://github.com/CIRCSE/LEMLAT3*.

[53] The choice of lexicographic sources for Classical Latin in *Lemlat* is based on the remarks by Lomanto (1980).

The linguistic properties of these lemmas are expressed as *RDF* triples using the *LiLa* ontology formalism and are stored in a triplestore publicly accessible via a *SPARQL* endpoint[54]. *Lemlat*'s lemmas have undergone a twofold process of revision: firstly, we removed overlapping or duplicate lemmas between the Classical and Medieval forms; secondly, we generated hypolemmas for all the canonical forms of present, future and perfect participles, as well as for deadjectival adverbs, and connected them to their main hyperlemmas via the symmetric property 'has hypolemma'/'is hypolemma'.

The *LiLa* collection currently includes 130,925 lemmas, 92,947 hypolemmas, 292,657 written representations of (hypo)lemmas, 59,945 'has/is hypolemma' properties, and 6,120 links between lemma variants[55].

3.3. *Examples from the lexical collection of* LiLa

In this section, we report on examples taken from the Knowledge Base to show the way in which a lemma and its connected information are stored in the *LiLa* lexical collection. More specifically, we detail how lemma variants, morphological features, hypolemmas, information on derivational morphology and prosodic representations are treated.

We first consider the lemma *claudeo*/*claudo* "to limp". In the *Oxford Latin Dictionary* (Glare, 1982), the entry for this lemma includes both the second conjugation (*claudeo*) and third conjugation verbs (*claudo*), the latter also featuring the graphical variant *cludo* (Lucil. 250). The lemma is recorded as deriving from the first class adjective *clausus* "closed, inaccessible".

In the *Ausführliches Lateinisch-Deutsches Handwörterbuch* (Georges and Georges, 1913-1918), alongside the citation forms *claudeo* and *claudo* we also find their respective and semantically identical deponent counterparts *claudeor* and *claudor*.

In the du Cange Medieval Latin *Glossarium*, lexical entries are provided neither for *claudeo*/*-eor* nor *claudo*/*-or*.

As previously mentioned, the *Lemlat* lexical basis integrates the *GGG* dictionaries. In *Lemlat*, the information about *claudeo*/*claudo* provided by these three reference dictionaries is merged into one single entry; here, a

[54] Cf. *https://lila-erc.eu/sparql/*. A network-based access point to the collection is available at *https://lila-erc.eu/lodlive/* and a user-friendly query interface is accessible at *https://lila-erc.eu/query/*.

[55] Numbers subject to change as the process of elimination of duplicate lemmas is still ongoing.

common ID is assigned to all lexical bases used to build the citation forms of the lexical entry. In the example case, *Lemlat* contains five different citation forms for the same lexical entry, all bearing the same ID: *claudeo*, *claudeor*, *claudo*, *claudor*, and *cludo*. In *LiLa*, these citation forms are represented by four lemmas distinguished by inflectional category. *Claudeo* and *claudo*, as well as their corresponding deponent forms *claudeor* and *claudor*, are citation forms for different lemmas, as they follow two different inflectional categories (active and deponent second conjugation, respectively)[56]. *Cludo*, on the other hand, is merged with *claudo*, as these share the same inflection. Just like *sequor* and *sequo*, *LiLa* connects these four lemmas via the 'lemma variant' property, while *cludo* and *claudo* are represented as different written representations, i.e. graphical variants of the same lemma (Figure 2)[57].

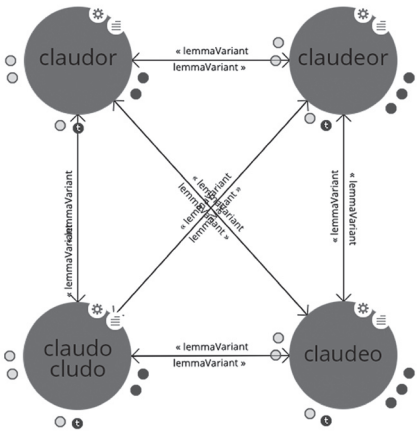

Figure 2. *Four citation forms of the same lexical entry in* LiLa.

In doing so, and as previously mentioned, *LiLa* harmonizes different lemmatization strategies and annotation styles, thus granting interoperability. In the example of *claudeo/claudo*, all the tokens of this lexical item occur-

[56] The homographic lemma of the third conjugation *claudo* "to close" is an independent node in *LiLa*, separate from *claudeo/claudo* and, thus, given a different unique identifier in the Knowledge Base.

[57] In all *LiLa* Figures henceforth (taken from the Lodlive interface), the small 'satellite' nodes circling the larger ones represent links to other nodes in the Knowledge Base, e.g. the PoS of the lemma.

ring in the lemmatized corpora and lexica available in *LiLa* can be joined together by using a set of five connected citations, regardless of whether the citation form used in a specific textual resource is *claudeo*, *claudeor*, *claudor*, or *claudo*/*cludo*.

The criterion used to distinguish between the different citation forms and different written representations of the same lexical item is purely morphological and, specifically, inflectional. If two citation forms for the same item belong to different inflectional categories, they are considered (and thus represented in *LiLa*) as two separate lemmas connected via the 'lemma variant' property. If not, they are stored in the lexical collection of *LiLa* as two written representations of the same lemma. Indeed, each Lemma node in *LiLa* is connected to a number of morphological features, among which is the inflectional category, as indicated by the 'has inflection type' property. Figure 3 shows the different categories to which the possible citation forms for *claudeo/claudo* are connected.

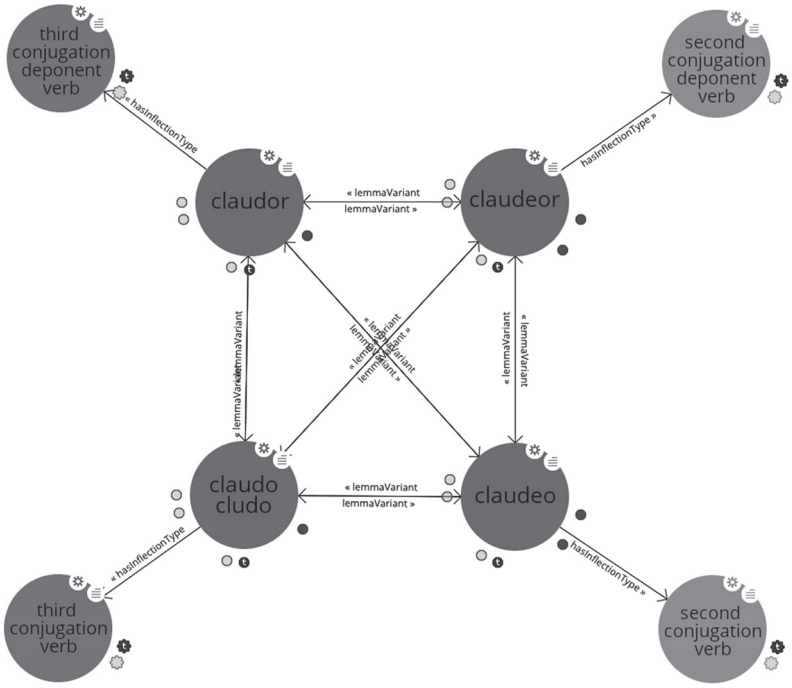

Figure 3. *Inflectional categories in* LiLa.

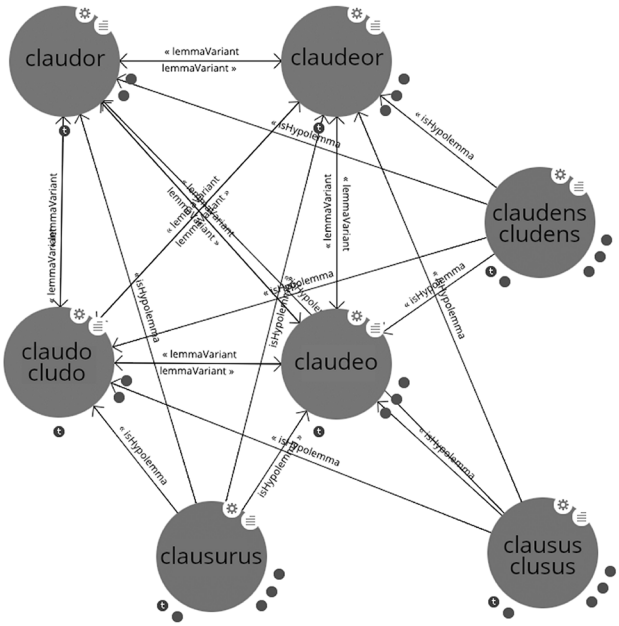

Figure 4. *Hypolemmas of verbs in* LiLa.

As we have already seen, Lemma nodes in *LiLa* can be connected to those for hypolemmas. In the case of lemmas for verbs, these are all connected to their hypolemmas for present, future and perfect participles. As Figure 4 shows, the node for *claudeo* (lemma) is connected to those for its participial citation forms *clausurus/clusurus*, *claudens/cludens* and *clausus/clusus* (hypolemmas) via the relation 'is hypolemma', making it possible to join different lemmatization strategies for participles. The same holds true for the other three lemmas connected via 'lemma variant'. In this way, whether in a lemmatized corpus a form like *claudentem* is assigned lemma *claudeo* (or *claudo*, *cludo*, *claudor*, *claudeor*) or *claudens*, in *LiLa* the form is always connected to the same lemma, as *claudens* is the written representation of the hypolemmas of all four lemmas for *claudeo/claudo*. Once again, *LiLa* does not perform any analysis but merely reflects the disambiguation provided by the connected resources. This means that, be it assigned to *claudo* or *claudens*, the form *claudentem* in *LiLa* is connected to both *claudeo/claudo* and *claudo* "to close". If the source corpus (or lexicon) includes morphological annotation,

the connection of the form to the correct lemma can be partly disambiguated on the basis of inflection, seeing as *claudeo* and *claudo* belong to two different categories[58]. Instead, if the resource to be included in *LiLa* does not provide morphological annotation but lemmatization and PoS tagging only, any form associated with the lemma *claudo* or *claudens* would be connected to both *claudeo/claudo* and *claudo*.

Beside inflectional morphological features, *LiLa* lemmas also carry information on derivational morphology. Two types of information about word formation are provided. Firstly, all lemmas belonging to a derivational family, i.e. a set of (derived) lemmas sharing the same lexical base, are connected to a node common to all family members (Base)[59]. Secondly, lemmas formed with one or more derivational affixes are connected to the nodes for such affixes (prefixes or suffixes). The information on derivational morphology is taken from the *WFL* resource by flattening the hierarchical relations of derivation recorded therein. Indeed, while *WFL* represents derivational families in terms of rooted trees, where one lemma is hierarchically derived from another (or from others, in the case of compounds), *LiLa* does not include such hierarchical relations between lemmas, but represents derivational morphology via flat connections between lemmas and their base(s) and affix(es) (Litta *et al*., 2019). Figure 5 shows the derivational family tree of *claudeo* in *WFL*.

In the derivational tree of Figure 5, each node represents a lemma belonging to the same derivational family. Nodes are connected by hierarchical relations labelled with the respective word formation rule. For instance, the lemma *claudeo/-eor* is the result of an adjective-to-verb conversion rule (A-To-V) applied to the adjective *claudus* "limping". The verb *claudico* "to limp", in turn, is derived from *claudeo/-eor* as a deverbal verb with the suffix *-ic*.

Like *LiLa*, *WFL* too makes use of the *Lemlat* lexical basis and so inherits the tool's lemma merges (e.g. *claudeo/-eor*). In *LiLa*, however, *claudeo* and *claudeor* are separate lemmas connected via the property 'lemma variant'. Furthermore, *LiLa* uses 'lemma variant' also to connect the third conjugation lemmas *claudo/cludo* and *claudor*; these are missing from *WFL* despite being recorded in *Lemlat* as variant forms of *claudeo/-eor*. Figure 6 shows how the derivational family of *claudeo* is represented in *LiLa*.

[58] This disambiguation is only partial. In order to disambiguate between *claudo* "to limp" and *claudo* "to close" (both third conjugation verbs) the resource must provide additional information other than morphology, e.g. a reference to the semantics of the lexical item.

[59] Compounds are connected to more than one Base node.

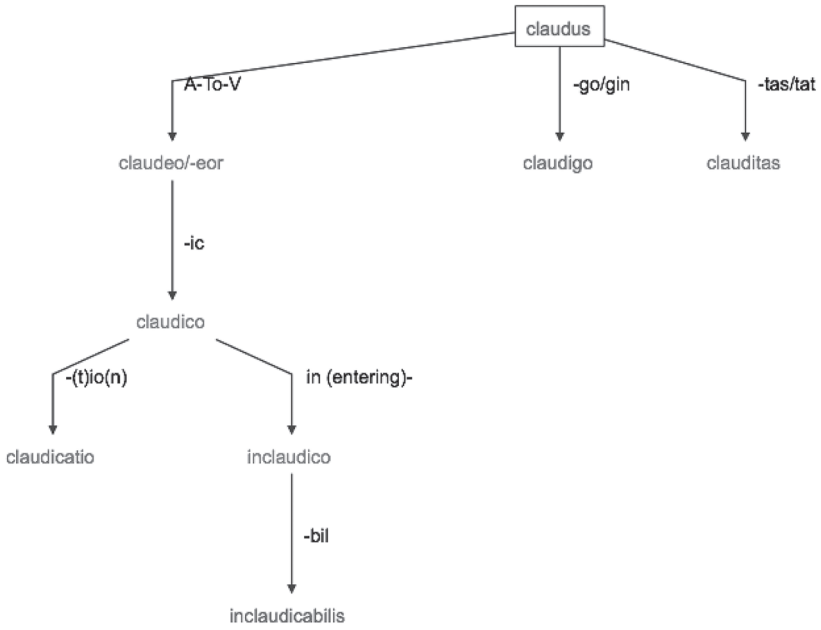

Figure 5. *The derivational family tree of* claudeo *in* WFL.

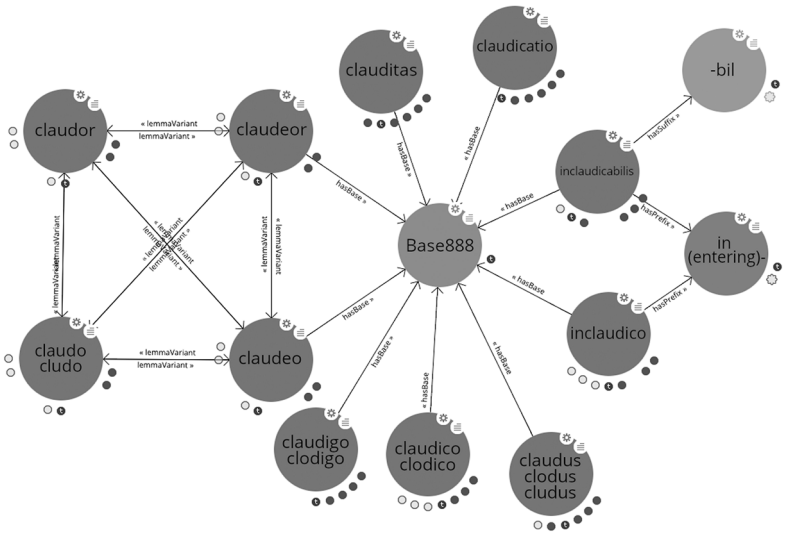

Figure 6. *The derivational family of* claudeo *in* LiLa.

In Figure 6, each lemma of the derivational family of *claudeo* is connected to a common Base node via the relation 'has Base'. As a connector between lemmas of a family, the Base node is unspecific and is instead given a numeric label (in this case, 888)[60]. Those lemmas that include one or more affixes are connected to the nodes for such affixes via the 'has prefix' and 'has suffix' properties, respectively. In Figure 6, this is the case of *inclaudico* "to limp / to be lame" and *inclaudicabilis* "not limping": while both lemmas are connected to the prefix node *in (entering)-* via the relation 'has prefix', *inclaudicabilis* alone is connected to the suffix node *-bil* via the 'has suffix' relation. Since the lemma variants *claudo*/*cludo* and *claudor* do not occur in *WFL* but in *LiLa* only, they are not explicitly connected to the Base node 888. These relations, however, are automatically induced in the ontology of *LiLa* in that all lemmas connected via 'lemma variant' share, possibly via inheritance, the same base and affixes (where present).

As mentioned in Section 3.2, cases like *occīdo* vs *occĭdo* are handled by attaching a 'prosodic representation' with vowel length to the lemma. Figure 7 shows the representation in *LiLa* of the verb *occīdo*.

The lemma node for the verb *occīdo* (with Type 'Lemma'), is connected to (a) its participial hypolemmas (*occisurus*, *occisus* and *occidens*), (b) its PoS ('Verb'), (c) the prefix *ob-*, (d) the inflection type 'third conjugation verb' and (e) Base 37, which is shared with, for instance, the verb *peroccido*, "to kill thoroughly". Moreover, the node *occīdo* is connected to the written representation *occido* and to the prosodic representation *occīdo*.

[60] Base nodes lack any kind of explicitly recorded linguistic information, as doing so would require a clear definition of the linguistic status of Base nodes stretching beyond that of connectors between lemmas belonging to the same derivational family. Indeed, such definition would open up a number of issues. One possible solution could be to assign each Base node a written representation consisting of a string describing the lexical 'element' (a root? a stem?) underlying each lemma in the derivational family (e.g. *dic-* for *dico* "to say", or *dictio* "a saying"). This procedure is complicated by the fact that different bases can be used in the same family, as is the case of, for example, *fer-*, *tul-* and *lat-*, which can all be found as bases in the family to which the verb *fero* "to bring" belongs. However, the current treatment of Base nodes does not prevent from integrating etymological information in the *LiLa* Knowledge Base (Mambrini and Passarotti, 2020).

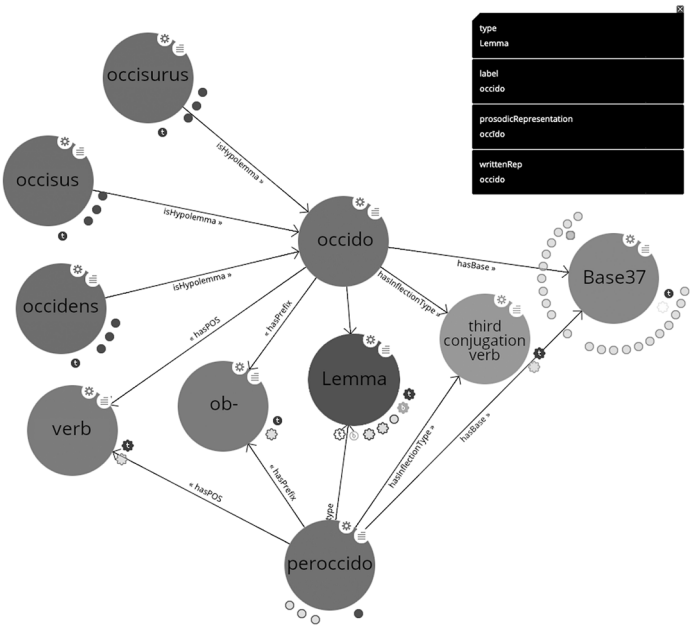

Figure 7. *Prosodic Representation in* LiLa.

4. *Including linguistic resources into* LiLa

Compiling the collection of lemmas described in previous sections is not the ultimate objective of *LiLa*, but a necessary step towards achieving interoperability between the linguistic resources included in the Knowledge Base.

In metaphysical terms, the collection of lemmas in *LiLa* represents a set of *noumena* (and it is, in itself, a *noumenon*), and a resource is a provider of *phenomena* (and it is, in itself, a *phenomenon*). The definition of these terms in Webster's Online Dictionary reads[61]:

The *noumenon* (plural: *noumena*) classically refers to an object of human inquiry, understanding or cognition. The term is generally used in contrast with, or in relation to, *phenomenon* (plural: *phenomena*), which refers to appearances, or objects

[61] Cf. *http://www.websters-dictionary-online.org*.

of the senses. A *phenomenon* is that which is perceived; A *noumenon* is the actual object that emits the *phenomenon* in question.

In *LiLa*, lemmas exist regardless of their actual realizations in textual and/or lexical resources. The first step of the *LiLa* project was to build this 'lexical *noumenon*'. The second step is to connect the *noumenon* to the *phenomenon*, i.e. to its actual realizations.

So far, the only textual resource to have been connected to *LiLa* is the *IT-TB* in its original annotation schema. This section describes the process of connecting the *IT-TB* to *LiLa* and details how the (meta) data provided by this treebank are linked to the lemma collection of the Knowledge Base.

The *IT-TB* exists in *LiLa* in its version downloadable from the *IT-TB* website (December, 2019)[62]. This version includes a selection of the concordances of the lemma *forma* "form" extracted from three works of Thomas Aquinas and the full text of the first three books of the *Summa contra gentiles*, for a total of 277,547 tokens (239,496 lexical tokens and 38,051 punctuation marks), corresponding to 3,901 different lemmas[63].

To connect the lemmatized lexical tokens of the *IT-TB* to the *LiLa* collection of lemmas, we perform a simple string match between the lemmas in the treebank and the written representations of lemmas in the Knowledge Base. As a result of this strategy, 3,627 out of 3,901 lemmas in the *IT-TB* (corresponding to 233,291 lexical tokens) were linked to at least one lemma in *LiLa*, while 274 (corresponding to 6,205 lexical tokens) found no match. Out of 3,697 lemmas, 778 were linked ambiguously[64] or, in other words, connected to more than one lemma in *LiLa*; in *LiLa*, for example, there exist two lemmas with written representation *venio*, both of which are verbs, one first conjugation ("to genuflect", a rare Medieval word from the du Cange glossary) and the other fourth conjugation ("to come")[65].

[62] Cf. *https://itreebank.marginalia.it*.

[63] Details on the composition of the *IT-TB* can be found in Passarotti (2019).

[64] Unambiguous linking obtained through simple string match may be risky in the case of homographic lemmas missing from the *LiLa* lexical collection, i.e. when a lemma in the incoming resource is a homograph of only one written representation of a lemma in *LiLa*, but belongs to another homographic lemma not present in the collection.

[65] The integration in *LiLa* of lexical resources providing information like, for instance, the date of first attestation of a lemma, its frequency, or its prevalence in a specific genre, will help to reduce ambiguity in the linking process.

To disambiguate cases like *venio*, we use the morphological tagging provided by the *IT-TB*, which assigns to each word form its PoS and inflectional category (declension, conjugation)[66]. For instance, in the sentence (1):

(1) *Nam primo habet formam seminis, postea sanguinis, et sic inde quousque veniat ad ultimum complementum.* (Thom. *Summa contra gentiles* II 89,9)
"At first it possesses the form of semen, afterwards of blood, and so on, until at last it arrives at that wherein it finds its fulfilment." [67]

the word form *veniat* in the *IT-TB* is assigned the PoS 'Verb' and the fourth conjugation, thus making it possible to unambiguously link it to the correct lemma in *LiLa*. This strategy disambiguated 650 lemmas out of the ambiguous 778 previously linked.

This leaves us with 128 ambiguously linked lemmas, because the lemmatization and morphological tagging of the *IT-TB* preclude an automated choice between the candidate lemmas. This is the case of the lemma *campus* (a second declension masculine noun), which links to *campus* "field" and *campus* (*marinus*) for *hippocampus* "sea-horse".

Finally, a number of lemmas were still left unlinked. These were found to fall under one of the following categories:

– the lemma does not exist in the *LiLa* collection, as is the case of the third declension feminine noun *actualitas* "actuality" (as opposed to potentiality). The *IT-TB* counts 223 of these cases, besides which 4 are new hypolemmas (e.g. the adverb *quantum* "as much as" recorded as hypolemma of *quantus* "how much") and 24 are lemmas of the type *occido^caedo*/*occido^cado*, for which disambiguation was performed manually: *IT-TB* tokens connected to *occido^caedo* were linked to the lemma with prosodic representation *occīdo*, while those connected to *occido^cado* were linked to *occĭdo*;

– the lemma of the *IT-TB* is a new written representation of a lemma already present in *LiLa*; this is the case of the written representation *annuncio* for the first conjugation verb *adnuntio* "to announce". Eight cases;

– the lemma of the *IT-TB* is a new lemma variant of a lemma already present in *LiLa*. For example, the singular first declension masculine noun

[66] PoS tagging in the *IT-TB* does not make use of the usual PoS labels, but follows three inflectional classes: nominal inflection (for nouns, adjectives and pronouns, with a separate tag for the nominal forms of the verbal paradigms: gerunds, gerundives, participles and supines), verbal inflection (for verbs) and no inflection (for adpositions, adverbs, conjunctions and interjections). Further details on the tripartite tagging of the *IT-TB* can be found in Cecchini *et al.* (2018a).

[67] English translation from *https://dhspriory.org/thomas/english/ContraGentiles2.htm#89*.

anthropomorphita is a lemma variant of the corresponding pluralia tantum *anthropomorphitae* (a group of heretics who attributed human form to God). Three cases;

– so-called 'pseudo-lemmas', which are used in the *IT-TB* for non Latin words (*non latina vox*), numbers (*num. arab.* and *num. rom.* for Arabic and Roman numbers, respectively) and abbreviations (e.g. *breviata loci notatio*). Eleven cases;

– lemmatization errors in the *IT-TB*. Six cases, e.g. *pbiectum* instead of *obiectum* "object".

After classifying lemmas into these categories, we expanded the *LiLa* collection with the new lemmas, written representations and lemma variants needed to fully connect the *IT-TB* to *LiLa*[68]. This strategy exemplifies *LiLa*'s empirical approach, whereby the lexical basis of the Knowledge Base grows with the number of linguistic resources connected.

The syntax of the *IT-TB* is annotated in dependency trees. Figure 8 shows the *IT-TB* dependency tree of sentence (1).

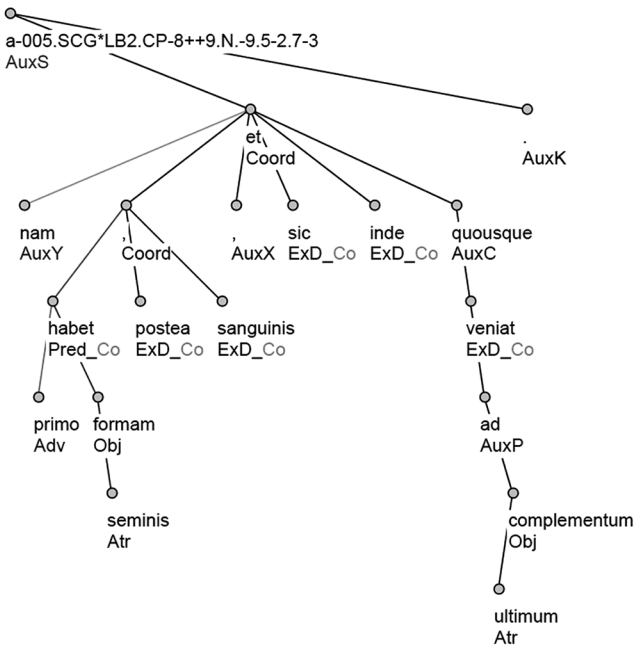

Figure 8. *A dependency tree from the Index Thomisticus Treebank.*

[68] Pseudo-lemmas and lemmatization errors remain unlinked.

The tree in Figure 8 features as many nodes as there are tokens in the sentence, including punctuation. Each token is assigned a syntactic function, known in dependency treebank jargon as 'dependency relation' (DepRel)[69]. Figure 9 shows the graphical representation of the connections holding between the tokens of the clause *quousque veniat ad ultimum complementum* (part of sentence 1) and the lemmas in the *LiLa* collection.

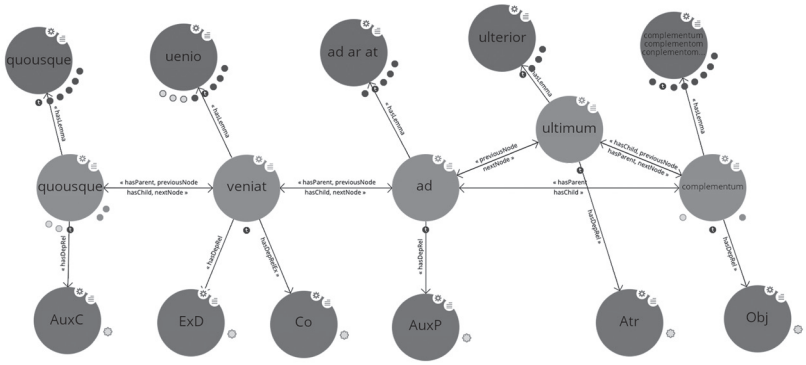

Figure 9. *A clause of the Index Thomisticus Treebank in* LiLa.

In Figure 9, each token of the example clause in the *IT-TB* is connected to exactly one lemma in *LiLa* via the relation 'has lemma', and to its previous/next node in the sentence via the symmetric relation 'previous node'/'next node'[70].

In the *LiLa* Knowledge Base, two pieces of information can be extracted from the trees of a dependency treebank:

(i) tokens are connected to their syntactic function via the property 'has DepRel'. The dependency relations shown in Figure 9 are AuxC (for subordinating conjunctions, here *quousque*), ExD (for nodes missing their head node in the dependency tree, i.e. ellipsis, here *veniat*), AuxP

[69] For a detailed description of the annotation rules and the set of dependency relations used in the *IT-TB*, see Bamman *et al.* (2007).

[70] Each token is also connected to a number of descriptive metadata taken from the original linguistic resource. In the case of the *IT-TB*, each token is linked to descriptive metadata recording its position in the texts of Thomas Aquinas (e.g. work, book, chapter, etc.) and to the sequence of morphological tags originally attached to it in the *IT-TB* (e.g. 3-MB1--6--1 for the third person singular of the present subjunctive of fourth conjugation verbs, e.g. *veniat*). The full morphological tagset of the *IT-TB* is available at *https://itreebank.marginalia.it/doc/Tagset_IT.pdf*.

(for prepositions, here *ad*), Atr (for Attributes, here *ultimum*) and Obj (for direct/indirect objects, i.e. arguments, here *complementum*). In the *IT-TB*, the syntactic functions of nodes in coordinated constructions are indicated by the extension _Co, as evidenced by *veniat* in Figure 8. In *LiLa*, this is represented via the relation 'has DepRelEx', which in Figure 9 connects the token *veniat* to the node *Co*;

(ii) dependencies between head and dependent nodes are represented through the symmetric property 'has parent'/'has child'. In Figure 9, for instance, the relation 'has child' holding between *veniat* and *ad* indicates that *veniat* is the head of *ad* in the dependency tree of this *IT-TB* clause[71].

5. *Conclusion*

In this paper, we have presented the overall architecture of the *LiLa* Knowledge Base of linguistic resources for Latin. Interweaving the large amount of linguistic (meta)data developed thus far in an interoperable whole is key to promoting the use of resources and tools. Today, this is made possible thanks to Linked Data technologies.

The first objective of the *LiLa* project was to compile a large collection of Latin lemmas in Linked Data form. This collection, described here in Section 3, represents the backbone of *LiLa*, given the central role played by the lemma in making resources interact. The collection was derived from a number of reference dictionaries and glossaries covering different chronological eras. However, as demonstrated by the inclusion of the first linguistic resource in the Knowledge Base (the Index Thomisticus Treebank; Section 4), a complete lexical coverage is far from being achieved (if not impossible), seeing as future resources are expected to introduce new lexical items and/or new citation forms of lemmas already recorded in *LiLa*. The greater the number of resources connected in *LiLa*, the larger its lemma collection will become.

The important role of the lemma in *LiLa* implies that only lemmatized resources can fully exploit the (lexical) connections in the Knowledge Base. Nowadays, this is a restrictive condition as, despite growing numbers, many Latin corpora do not carry this layer of linguistic annotation. One core chal-

[71] When the 'has parent'/'has child' property overlaps with the 'previous node'/'next node' one, these are merged into one edge in the visualization, as exemplified by *veniat* and *ad* in Figure 9: *veniat* both precedes *ad* in the word order of the clause and it is its parent node in the dependency tree.

lenge for *LiLa* will be to collect and evaluate the tools and trained models available for automatic lemmatization and, next, to build a new set to allow data providers to process their resource(s) for ready inclusion in the Knowledge Base. Indeed, even if lemmatized, texts might nevertheless cause trouble in cases such as ambiguous homographic lemmas (e.g. *occīdo* vs *occĭdo*). *LiLa*, after all, reflects the degree of annotation granularity provided by the resources attached to the Knowledge Base.

Another important issue that *LiLa* must address is how to deal with resources in closed and/or proprietary formats. While most Computational Linguistics resources and tools are freely available, popular collections of scholarly editions of Latin and Ancient Greek texts, such as the Bibliotheca Teubneriana Latina by De Gruyter and all Brepols corpora, are locked behind paywalls. In line with the 'as open as possible, as closed as necessary' approach, proprietary resources will be connected in the Knowledge Base but access to them will be subject to charges. In doing so, we hope to influence policy change and to establish *LiLa* as a leading publication venue of Latin's linguistic legacy.

Acknowledgments

We are thankful to Daniela Corbetta and Andrea Peverelli for their invaluable support in building and extending the *LiLa* collection of lemmas. The *LiLa*: Linking Latin 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

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. and Sherlock, G. (2000), *Gene ontology: tool for the unification of biology*, in «Nature genetics», 25, 1, p. 25.

Bamman, D. and Crane, G. (2006), *The design and use of a Latin dependency treebank*, in Hajič, J. and Nivre, J. (2006, eds.), *Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories*, Institute of Formal and Applied Linguistics, Prague, pp. 67-78.

Bamman, D., Passarotti, M., Busa, R. and Crane, G. (2008), *The annotation guidelines of the Latin Dependency Treebank and Index Thomisticus Treebank. The treatment of some specific syntactic constructions in Latin*, in Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S. and Tapias, D. (2008, eds.), *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, European Language Resources Association (ELRA), Paris, pp. 71-76.

Bamman, D., Passarotti, M., Crane, G. and Raynaud, S. (2007), *Guidelines for the syntactic annotation of Latin treebanks*, Tufts University Digital Library, Medford / Somerville.

Berners-Lee, T., Hendler, J. and Lassila, O. (2001), *The Semantic Web*, in «Scientific American», 284, 5, pp. 28-37.

Bird, S. and Liberman, M. (2001), *A formal framework for linguistic annotation*, in «Speech communication», 33, 1-2, pp. 23-60.

Budassi, M. and Passarotti, M. (2016), Nomen Omen. *Enhancing the Latin morphological analyser* Lemlat *with an onomasticon*, in Reiter, N., Alex, B. and Zervanou, K.A. (2016, eds.), *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2016)*, Association for Computational Linguistics, Berlin, pp. 90-94.

Buitelaar, P., Cimiano, P., McCrae, J., Montiel-Ponsoda, E. and Declerck, T. (2011), *Ontology lexicalisation: The lemon perspective*, in Slodzian, M., Valette, M., Aussenac-Gilles, N., Condamines, A., Hernandez, N. and Rothenburger, B. (2011, eds.), *Proceedings of the Workshops. 9th International Conference on Terminology and Artificial Intelligence*, INALCO, Paris, pp. 33-36.

Busa, R. (1974-1980), *Index Thomisticus: sancti Thomae Aquinatis operum omnium indices et concordan-tiae, in quibus verborum omnium et singulorum formae et lemmata cum suis frequentiis et contextibus variis modis referuntur quaeque consociata plurium opera atque electronicoIBM automato usus digessit Robertus Busa SJ.*, Frommann / Holzboog, Stuttgart / Bad Cannstatt.

DU CANGE, C., BÉNÉDICTINS DE SAINT-MAUR, CARPENTIER, P., HENSCHEL, L. and FAVRE, L. (1883-1887), *Glossarium Mediae et Infimae Latinitatis*, L. Favre, Niort.

CECCHINI, F.M., PASSAROTTI, M., MARONGIU, P. and ZEMAN, D. (2018a), *Challenges in converting the Index Thomisticus Treebank into Universal Dependencies*, in DE MARNEFFE, M.C., LYNN, T. and SCHUSTER, S. (2018, eds.), *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, The Association for Computational Linguistics, Bruxelles, pp. 27-36.

CECCHINI, F.M., PASSAROTTI, M., TESTORI, M., RUFFOLO, P., DRAETTA, L., FIEROMONTE, M., LIANO, A., MARINI, C. and PIANTANIDA, G. (2018b), *Enhancing the Latin morphological analyser* LEMLAT *with a Medieval Latin glossary*, in CABRIO, E., MAZZEI, A. and TAMBURINI, F. (2018, eds.), *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, Accademia university press, Torino, pp. 87-92.

CELANO, G.G.A. (2019), *The Dependency Treebanks for Ancient Greek and Latin*, in BERTI, M. (2019, ed.), *Digital Classical Philology. Ancient Greek and Latin in the Digital Revolution*, De Gruyter, Berlin / Boston, pp. 279-298.

CHIARCOS, C. (2012), *Interoperability of corpora and annotations*, in CHIARCOS, C., HELLMANN, S. and NORDHOFF, S. (2012, eds.), *Linked Data in Linguistics*, Springer, Berlin, pp. 161-179.

CHIARCOS, C., HELLMANN, S. and NORDHOFF, S. (2012), *Introduction and overview*, in CHIARCOS, C., HELLMANN, S. and NORDHOFF, S. (2012, eds.), *Linked Data in Linguistics*, Springer, Berlin, pp. 1-12.

CHIARCOS, C., CIMIANO, P., DECLERCK, T. and MCCRAE, J.P. (2013), *Linguistic linked open data (*llod*). Introduction and overview*, in CHIARCOS, C., CIMIANO, P., DECLERCK, T. and MCCRAE, J.P. (2013, eds.), *Proceedings of the 2nd Workshop on Linked Data in Linguistics (LDL-2013): Representing and Linking Lexicons, Terminologies and Other Language Data*, Association for Computational Linguistics, Pisa, pp. i-xi.

CHIARCOS, C. and SUKHAREVA, M. (2015), OLiA - *Ontologies of Linguistic Annotation*, in «Semantic Web Journal», 6, 4, pp. 379-386.

CHIARCOS, C. and FÄTH, C. (2017), CoNLL-RDF*: Linked corpora done in an* NLP*-friendly way*, in GRACIA, J., BOND, F., MCCRAE, J., BUITELAAR, P., CHIARCOS, C. and HELLMANN, S. (2017, eds.), *Language, Data, and Knowledge*, Springer, Berlin, pp. 74-88.

Declerck, T., Lendvai, P., Mörth, K., Budin, G. and Váradi, T. (2012), *Towards linked language data for digital humanities*, in Chiarcos, C., Hellmann, S. and Nordhoff, S. (2012, eds.), *Linked Data in Linguistics*, Springer, Berlin, pp. 109-116.

Eger, S., vor der Brück, T. and Mehler, A. (2015), *Lexicon-assisted tagging and lemmatization in Latin: A comparison of six taggers and two lemmatization methods*, in Zervanou, K., van Erp, M. and Alex, B. (2015, eds.), *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, Association for Computational Linguistics, Beijing, pp. 105-113.

Fellbaum, C. (2012), *Wordnet*, in Chapelle, C. (2012, ed.), *The Encyclopedia of Applied Linguistics*, Wiley Online Library [doi:*10.1002/9781405198431.wbeal1285*, accessed on 28.11.2019].

Forcellini, E. (1940), *Lexicon Totius Latinitatis / ad Aeg. Forcellini lucubratum, dein a Jos. Furlanetto emendatum et auctum; nunc demum Fr. Corradini et Jos. Perin curantibus emendatius et auctius meloremque in formam redactum adjecto altera quasi parte Onomastico totius latinitatis opera et studio ejusdem Jos. Perin*, Typis Seminarii, Padova.

Franzini, G., Peverelli, A., Ruffolo, P., Passarotti, M., Sanna, H., Signoroni, E., Ventura, V. and Zampedri, F. (2019), *Nunc Est Aestimandum: Towards an Evaluation of the Latin WordNet*, in Bernardi, R., Navigli, R. and Semeraro, G. (2019, eds.), *Proceedings of the Sixth Italian Conference on Computational Linguistics* (AI*IA Series, 2481), CEUR Workshop Proceedings, Bari, pp. 1-8.

Georges, K.E. and Georges, H. (1913-1918), *Ausführliches Lateinisch-Deutsches Handwörterbuch*, Hahn, Hannover.

Glare, P.G.W. (1982), *Oxford Latin Dictionary*, Oxford University Press, Oxford.

Goodwin, J., Dolbear, C. and Hart, G. (2008), *Geographical linked data: The administrative geography of great britain on the semantic web*, in «Transactions in GIS», 12, pp. 19-30.

Gradenwitz, O. (1904), *Laterculi Vocum Latinarum*, Hirzel, Leipzig.

Hajič, J., Panevová, J., Buráňová, E., Urešová, Z. and Bémová, A. (1999), *Annotations at analytical level. Instructions for annotators*, UK MFF ÚFAL, Prague [available online at *https://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/a-layer/pdf/a-man-en.pdf*, accessed on 28.11.2019].

Haug, D.T.T. and Jøhndal, M. (2008), *Creating a parallel treebank of the old Indo-European Bible translations*, in Sporleder, C. and Ribarov, K. (2008, eds.), *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, European Language Resources Association (ELRA), Paris, pp. 27-34.

Haug, D. (2010), Proiel *guidelines for annotation* [available online at *http://folk.uio.no/daghaug/syntactic_guidelines.pdf*, accessed on 28.11.2019].

Hellmann, S., Lehmann, J., Auer, S. and Brümmer, M. (2013), *Integrating NLP using Linked Data*, in Alani, H., Lalana, K., Fokoue, A., Groth, P., Biemann, C., Xavier Parreira, J., Aroyo, L., Noy, N., Welty, C. and Janowicz, K. (2013, eds.), *The Semantic Web – ISWC 2013. 12th International Semantic Web Conference, Sydney, Australia, October 21-25, 2013*, Springer, Berlin / Heidelberg, pp. 98-113.

Korkiakangas, T. and Passarotti, M. (2011), *Challenges in annotating medieval Latin charters*, in «Journal for Language Technology and Computational Linguistics», 26, 2, pp. 103-114.

Lassila, O. and Swick, R.R. (1998), *Resource description framework (rdf) model and syntax specification* [available online at *https://www.w3.org/TR/1999/REC-rdf-syntax-19990222/*, accessed on 28.11.2019].

Litta, E. and Passarotti, M. (2019), *(When) inflection needs derivation: a word formation lexicon for Latin*, in Holmes, N., Ottink, M., Schrickx, J. and Selig, M. (2019, eds.), *Lemmata Linguistica Latina*. Vol. 1: *Words and Sounds*, De Gruyter, Berlin / Boston, pp. 224-239.

Litta, E., Passarotti, M. and Mambrini, F. (2019), *The treatment of word formation in the* LiLa *Knowledge Base of linguistic resources for Latin*, in Žabokrtský, Z., Ševčíková, M., Litta, E. and Passarotti, M. (2019, eds.), *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology (DeriMo 2019). 19-20 September 2019, Prague*, Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, Prague, pp. 35-43.

Lomanto, V. (1980), *Lessici latini e lessicografia automatica*, in «Memorie dell'Accademia delle Scienze di Torino. Classe di Scienze Morali, Storiche e Filologiche», 5, 4, pp. 113-270.

Mambrini, F. and Passarotti, M. (2020), *Representing etymology in the* LiLa *knowledge base of linguistic resources for Latin*, in Kernerman, I. and Krek, S. (2020, eds.), *Proceedings of the Globalex Workshop on Linked Lexicography (@LREC 2020)*, European Language Resources Association (ELRA), Paris.

McCrae, J.P., Chiarcos, C., Bond, F., Cimiano, P., Declerck, T., De Melo, G., Gracia, J., Hellmann, S., Klimek, B., Moran, S., Osenova, P., Pareja-Lora, A. and Pool, J. (2016), *The open linguistics working group: Developing the Linguistic Linked Open Data cloud*, in Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J. and Piperidis, S. (2016, eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, European Language Resources Association (ELRA), Paris, pp. 2435-2441.

McCrae, J.P., Bosque-Gil, J., Gracia, J., Buitelaar, P. and Cimiano, P. (2017), *The Ontolex-Lemon model: development and applications*, in Kosem, I., Tiberius, C., Jakubíček, M., Kallas, J., Krek, S. and Baisa, V. (2017, eds.), *Proceedings of eLex 2017 conference*, Lexical Computing, Brno, pp. 19-21.

McGillivray, B. and Passarotti, M. (2009), *The Development of the Index Thomisticus Treebank Valency Lexicon*, in Borin, L. and Lendvai, P. (2009, eds.), *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education (LaTeCH – SHELT &R 2009)*, Association for Computational Linguistics, Athens, pp. 43-50.

McGillivray, B. (2013), *Methods in Latin Computational Linguistics*, Brill, Leiden.

McGuinness, D.L. and Van Harmelen, F. (2004), OWL *web ontology language overview*, in Web Ontology Working Group (2004, ed.), *W3C recommendation*, 10.10 [available online at *https://www.w3.org/TR/2004/REC-owl-features-20040210/*, accessed on 28.11.2019].

Minozzi, S. (2010), *The Latin WordNet project*, in Anreiter, P. and Kienpointner, M. (2010, eds.), *Latin Linguistics Today. Akten des 15. Internationalen Kolloquiums zur Lateinischen Linguistik, Innsbrucker Beiträge zur Sprachwissenschaft*, Institut für Sprachen und Literaturen der Universität Innsbruck Bereich Sprachwissenschaft, Innsbruck, pp. 707-716.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R. and Zeman, D. (2016), *Universal Dependencies v1: A multilingual treebank collection*, in Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J. and Piperidis, S. (2016, eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, European Language Resources Association (ELRA), Paris, pp. 1659-1666.

Passarotti, M. (2019), *The project of the Index Thomisticus Treebank*, in Berti, M. (2019, ed.), *Digital Classical Philology. Ancient Greek and Latin in the Digital Revolution*, De Gruyter, Berlin / Boston, pp. 299-319.

Passarotti, M., González Saavedra, B. and Onambele, C. (2016), *Latin Vallex. A treebank-based semantic valency lexicon for Latin*, in Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J. and Piperidis, S. (2016, eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, European Language Resources Association (ELRA), Paris, pp. 2599-2606.

Passarotti, M., Budassi, M., Litta, E. and Ruffolo, P. (2017), *The* Lemlat *3.0 package for morphological analysis of Latin*, in Bouma, G. and Adesam, Y. (2017, eds.), *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language. 22nd May 2017 Gothenburg*, Linköping University Electronic Press, Linköping, pp. 24-31.

Petrov, S., Das, D. and McDonald, R. (2011), *A Universal Part-of-Speech Tagset*, in «ArXiv Preprint» [available online at *https://arxiv.org/abs/1104.2086*, accessed on 28.11.2019].

Pianta, E., Bentivogli, L. and Girardi, C. (2002), *MultiWordNet: Developing an aligned multilingual database*, in Hamdan, H. and Boubiche, D.E. (2002, eds.), *Proceedings of the First International Conference on Global WordNet*, The Association for Computational Linguistics, Liverpool, pp. 55-63.

Prud'Hommeaux, E. and Seaborne, A. (2008), *Sparql query language for rdf. W3c* [available online at *https://www.w3.org/TR/rdf-sparql-query/*, accessed on 28.11.2019].

Sprugnoli, R., Passarotti, M., Cecchini, F.M. and Pellegrini, M. (2020), *Overview of the EvaLatin 2020 evaluation campaign*, in Sprugnoli, R. and Passarotti, M. (2020, eds.), *Proceedings of the LT4HALA 2020 Workshop - 1st Workshop on Language Technologies for Historical and Ancient Languages, satellite event to the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*, European Language Resources Association (ELRA), Paris, pp. 105-110.

Tombeur, P. (1998, ed.), *Thesaurus formarum totius latinitatis a Plauto usque ad saeculum Xxum*, Turnhout, Brepols.

van Erp, M. (2012), *Reusing linguistic resources: Tasks and goals for a linked data approach*, in Chiarcos, C., Hellmann, S. and Nordhoff, S. (2012, eds.), *Linked Data in Linguistics*, Springer, Berlin, pp. 57-64.

Marco Passarotti
Facoltà di Scienze Linguistiche
e Letterature Straniere
Università Cattolica del Sacro Cuore
Largo Gemelli 1
20123 Milano (Italy)
marco.passarotti@unicatt.it

Francesco Mambrini
Facoltà di Scienze Linguistiche
e Letterature Straniere
Università Cattolica del Sacro Cuore
Largo Gemelli 1
20123 Milano (Italy)
francesco.mambrini@unicatt.it

Greta Franzini
Facoltà di Scienze Linguistiche
e Letterature Straniere
Università Cattolica del Sacro Cuore
Largo Gemelli 1
20123 Milano (Italy)
greta.franzini@unicatt.it

Flavio Massimiliano Cecchini
Facoltà di Scienze Linguistiche
e Letterature Straniere
Università Cattolica del Sacro Cuore
Largo Gemelli 1
20123 Milano (Italy)
flavio.cecchini@unicatt.it

Eleonora Litta
Facoltà di Scienze Linguistiche
e Letterature Straniere
Università Cattolica del Sacro Cuore
Largo Gemelli 1
20123 Milano (Italy)
eleonoramaria.litta@unicatt.it

Giovanni Moretti
Facoltà di Scienze Linguistiche
e Letterature Straniere
Università Cattolica del Sacro Cuore
Largo Gemelli 1
20123 Milano (Italy)
giovanni.moretti@unicatt.it

Paolo Ruffolo
Facoltà di Scienze Linguistiche
e Letterature Straniere
Università Cattolica del Sacro Cuore
Largo Gemelli 1
20123 Milano (Italy)
paolo.ruffolo@posteo.eu

Rachele Sprugnoli
Facoltà di Scienze Linguistiche
e Letterature Straniere
Università Cattolica del Sacro Cuore
Largo Gemelli 1
20123 Milano (Italy)
rachele.sprugnoli@unicatt.it