PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 115 OCTOBER 2020 163–186

Derivations and Connections: Word Formation in the LiLa Knowledge Base of Linguistic Resources for Latin

Eleonora Litta, Marco Passarotti, Francesco Mambrini

CIRCSE Research Centre.
Università Cattolica del Sacro Cuore, Milan, Italy

Abstract

The *LiLa* project aims to build a Knowledge Base of linguistic resources for Latin based on the Linked Data framework, with the goal of creating interoperability between them. To this end, LiLa integrates all types of annotation applied to a particular word/text into a common representation where all linguistic information conveyed by a specific linguistic resource becomes accessible. The recent inclusion in the Knowledge Base of information on word formation raised a number of theoretical and practical issues concerning its treatment and representation. This paper discusses such issues, detailing how they are addressed in the project, and introduces the web application to query the collection of lemmas of the Knowledge Base. A number of use-case scenarios that employ the information on word formation made available in the LiLa Knowledge Base are also presented, particularly focusing on the use of the Knowledge Base to compare the perspectives on word formation in different linguistic resources.

1. Introduction

The increasing quantity, complexity and diversity of the currently available linguistic resources for a wide range of languages has led, in recent times, to a growing interest in the sustainability and interoperability of (annotated) corpora, dictionaries,

This paper is an extended version of the work presented by Litta et al. (2019) at the Second Edition of the Workshop on Resources and Tools for Derivational Morphology (DeriMo 2019), 19-20 September 2019, Prague, Czech Republic.

 Corresponding author: marco.passarotti@unicatt.it
Cite as: Eleonora Litta, Marco Passarotti, Francesco Mambrini. Derivations and Connections: Word Formation in the LiLa Knowledge Base of Linguistic Resources for Latin. The Prague Bulletin of Mathematical Linguistics No. 115, 2020, pp. 163–186. doi: 10.14712/00326585.010.

thesauri, lexica and Natural Language Processing (NLP) tools (Ide and Pustejovsky, 2010). This effort, initially, resulted in the creation of databases and infrastructures hosting linguistic resources, such as CLARIN,[1] DARIAH,[2] META-SHARE[3] and EAGLE.[4] Such initiatives collect resources and tools, which can be used and queried from a single web portal, but they do not provide real interconnection between them. In fact, in order to make linguistic resources interoperable, all types of annotations applied to a particular word/text should be integrated into a common representation that enables access to the linguistic information conveyed in a linguistic resource or produced by an NLP tool (Chiarcos, 2012, p. 162).

To meet the need of interoperability, the *LiLa* project's objective (2018-2023)[5] is to create a Knowledge Base of linguistic resources for Latin based on the Linked Data framework,[6] 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 ultimate goal of the project is to exploit to the fullest the wealth of linguistic resources and NLP tools for Latin developed so far, and to bridge the gap between raw language data, NLP and knowledge description (Declerck et al., 2012, p. 111).

In its design, the structure of LiLa is highly lexically-based: the core component of the Knowledge Base is an extensive list of Latin lemmas extracted from the morphological analyser for Latin Lemlat (Passarotti et al., 2017). This list has been compiled into a database from three reference dictionaries for Classical Latin ((Georges, 1913); (Glare, 1982); (Gradenwitz, 1904)), the entire Onomasticon from Forcellini's (Forcellini, 1867) 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., 2018). The portion of the lexical basis of Lemlat concerning Classical and Late Latin (43,432 lemmas) was also enhanced with information taken from the Word Formation Latin (WFL) lexicon (Litta and Passarotti, 2019), a lexical resource that provides information about derivational morphology by connecting lemmas via word formation rules.

The consolidation of information taken from WFL into the LiLa Knowledge Base raises a number of theoretical and practical issues concerning the treatment and representation of word formation in LiLa. The present paper discusses such issues, presenting how they are addressed in the project. The paper is organised as follows. Section 2

[1] http://www.clarin.eu.

[2] http://www.dariah.eu.

[3] http://www.meta-share.org.

[4] http://www.eagle-network.eu.

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

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

introduces the LiLa Knowledge Base, sketching its fundamental architecture. Section 3 presents the WFL lexicon. Section 4 discusses how word formation is accounted for in LiLa, detailing the classes of the LiLa ontology concerned. Section 5 describes the main features of the web application built to query the collection of lemmas of the Knowledge Base. Section 6 presents a number of use-case scenarios that employ the information on word formation made available in LiLa, particularly focusing on the use of the Knowledge Base to compare the perspectives on word formation provided by different linguistic resources. Lastly, Section 7 concludes the paper.

2. The LiLa Knowledge Base

In order to achieve interoperability between distributed resources and tools, LiLa adopts a set of Semantic Web and Linked Data standards. These include ontologies that 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[7]). 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 value, e.g. a string or an integer). The SPARQL Protocol and RDF Query Language (SPARQL) is used to query the data recorded in the form of RDF triples in a triplestore (Prud'Hommeaux et al., 2008).[8]

The lexically-based nature of the LiLa Knowledge Base results from a simple, fundamental assumption: textual resources are made of (occurrences of) words, lexical resources describe properties of words, and NLP tools process words. In particular, the lemma is considered the ideal interconnection between lexical resources (such as dictionaries, thesauri and lexica), annotated corpora and NLP tools that lemmatise their input text. Lemmas are canonical forms of words that are used by dictionaries to cite lexical entries, and are produced by lemmatisers to analyse tokens in corpora. For this reason, as was said, the core of the LiLa Knowledge Base is represented by the collection of Latin lemmas taken from the morphological analyser Lemlat;[9] Lemlat has proven to cover more than 98% of the textual occurrences of the word forms recorded in the comprehensive *Thesaurus formarum totius latinitatis* (TFTL, Tombeur, 1998), which is based on a corpus of texts ranging from the beginnings of Latin literature up to present times, for a total of more than 60 million words (Cecchini et al., 2018). LiLa thus aims to achieve interoperability by linking all entries in lexical re-

[7] `https://www.w3.org/community/ontolex`.

[8] A prototype of the LiLa triplestore is accessible at `https://lila-erc.eu/sparql`.

[9] `https://github.com/CIRCSE/LEMLAT3`.

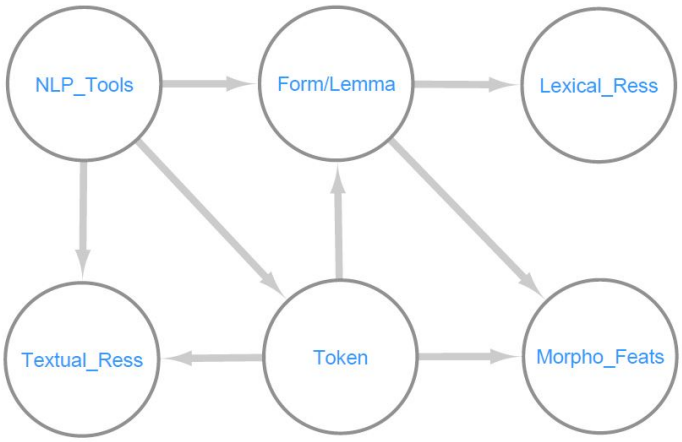

Figure 1. The fundamental architecture of LiLa.

sources and corpus tokens that refer to the same lemma, allowing a good balance between feasibility and granularity.

Figure 1 shows a simplified representation of the fundamental architecture of LiLa, highlighting the relations between the main components represented by the lemma and the other types of resources that interact with the Knowledge Base. There are two nodes representing as many kinds of linguistic resources linked to the core components: a) **Textual Resources**: they provide texts, which are made of **Tokens**; from a morphological standpoint, tokens can be analysed as occurrences of word forms;[10] b) **Lexical Resources**: they describe lexical items, which can include references to lemmas (e.g. in a bilingual dictionary), or to word forms (e.g. in a collection of forms like the aforementioned TFTL). A **Lemma** is one special type of (inflected) **Form** that is conventionally chosen as the citation form for a lexical item. Both tokens and forms (and thus lemmas, as a subclass of forms) are assigned **Morphological Features**, like part of speech (PoS), inflexional category and gender. Finally, **NLP tools** such as tokenisers, PoS taggers and morphological analysers can process respectively textual resources, tokens and forms.

Using the Lemma node as a pivot, it is thus possible to connect resources and make them interact, for instance by searching in different corpora all the occurrences of a lemma featuring some specific lexical properties (provided by one or more lexical resource).

[10]The degree of overlapping between tokens and forms depend on the criteria for tokenisation applied. Given the morphosyntactic properties of Latin, in LiLa this overlapping is complete.

3. The Word Formation Latin Lexicon

The WFL lexicon adds a layer of information on word formation to the lexical materials for Classical and Late Latin of the Lemlat database. The lexicon is based on a set of word formation rules (WFRs) represented as directed one-to-many input-output relations between lemmas. The lexicon was devised according to the Item-and-Arrangement (I&A) model of morphological description (Hockett, 1954): lemmas are either non-derived lexical morphemes, or a concatenation of a base in combination with affixes. This theoretical model was chosen because it emphasises the semantic significance of affixal elements, and because it had been previously adopted by other resources treating derivation, such as the morphological dictionaries Word Manager (Domenig and ten Hacken, 1992).

WFL is characterised by a step-by-step morphotactic approach: each word formation process is treated individually as the application of one single rule. For instance, the adjective *febricula* 'a slight fever' is recorded in WFL as derived from the noun *febris* 'fever' via a WFR that creates diminutive nouns with the suffix -(*us*/*un*)*cul*.

This approach results in a hierarchical structure, whereby one or more lemmas derive from one ancestor lemma. A set of lemmas derived from one common ancestor is defined as a "word formation family". In the web application for querying the WFL lexicon, this hierarchical structure is represented in a directed graph resembling a tree.[11] In the graph of a word formation family, nodes are occupied by lemmas, and edges are labelled with a description of the WFR used to derive the output lemma from the input one. For instance, Figure 2 shows the derivation graph for the word formation family whose ancestor (or "root") lemma is *febris*.

Each output lemma can only have one input lemma, unless the output lemma qualifies as a compound, as in the case of *febrifugia* 'a plant called centaury', a compound formed by the noun *febris* and the verb *fugo* 'to cause to flee, to drive away'. In WFL, simple conversion (i.e. change of PoS without further affixation) is treated as a separate WFR, like in the case of the the verb *fugo* derived from the noun *fuga* 'flight' in Figure 2. However, when considering formations involving both the attachment of an affix and a shift in PoS (as, for example, *febris*, noun >*febricito* 'to have a fever', verb), these are handled in one single step.

That being said, portraying word formation processes via directed graphs raises some significant theoretical issues, especially in cases where the derivational direction is uncertain or unsuitable to be represented by a single step-by-step process (Budassi and Litta, 2017). In such instances, WFL adheres to a strict methodology in order to work around fuzziness. An illustrative case in point is the difficulty in firmly establishing a direction in the derivation of conversion processes such as N-to-A or A-to-N. When considering, to give an example, the relation between the adjective *adversus* 'facing towards', the noun *adversarius* 'an opponent', and the adjective *adversarius* 'hostile',

[11] `http://wfl.marginalia.it`.

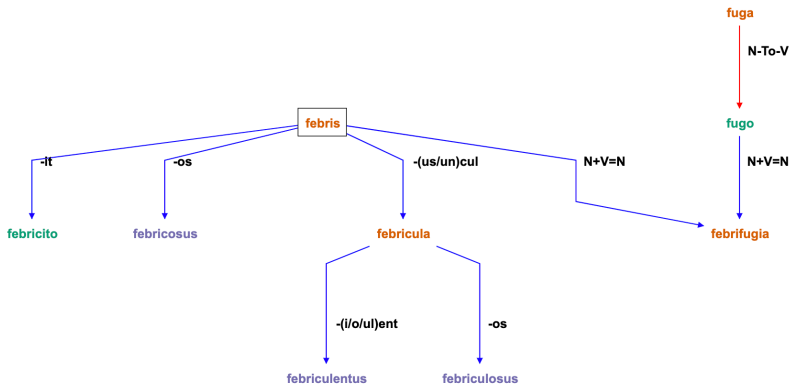

Figure 2. Derivation graph for the word formation family of febris *in WFL.*

did the word formation process work like *adversus* > *adversarius* A > *adversarius* N, or like *adversus* > *adversarius* N > *adversarius* A? When there is space for interpretation on which direction the change has happened, Oxford Latin Dictionary (OLD) (Glare, 1982) is used, as a rule, to "testify" the provenance of lemmas (in our case *adversus* > *adversarius* A > *adversarius* N). Even so, in a few occasions it has been necessary to take some independent choices: for instance, OLD states that diminutive noun *amiculus* 'a pet friend' derives from the adjective *amicus* 'friend'; we, however, chose to make it derive from noun *amicus* as it seems more probable that a diminutive noun was created to diminish a noun rather than an adjective.

The most controversial strategy adopted in WFL to work around non-linear derivations was the creation of "fictional" lemmas that act as placeholders between attested words in order to justify extra "mechanical" (morphotactic) steps. The verb *exaquesco* 'to become water', for example, is connected to the noun *aqua* 'water', through a made-up verb **aquesco*.[12] However, the existence of these fictional lemmas has proven to be less than ideal. User feedback has reported confusion and puzzlement at the presence of the fictional element in the derivational tree. Moreover, when browsing the data, the existence of fictional lemmas needs to be factored in. For instance, if looking for all lemmas created with the suffix *-bil* in WFL, 598 lemmas are given as a result.[13] In WFL, 103 of these are fictional lemmas, 17% of the total number of lemmas derived

[12]The asterisk used to indicate fictional lemmas in WFL does not have the same value as the asterisk employed in Indo-European studies to indicate a reconstructed word, but merely marks a fabricated "stepping stone" in a two-step derivational process.

[13]These are in Latin adjectives that have generally instrumental (e.g. *terribilis* 'by whom/which one is terrified') and/or passive and potential meaning (e.g. *amabilis* 'which/who can be loved') (Kircher-Durand, 1991 and Litta, 2019).

using the *-bil* suffix. The vast majority of these were fabricated in order to establish a derivational process between lemmas such as the adverb *imperabiliter* 'authoritatively' to their "next of kin", the verb *impero* 'to demand, to order'. In order to account for these two steps, i.e. the addition of the suffixes *-bil* and *-ter*, the fictional adjective **imperabilis* was created as a further step in the word formation process. The presence of fictional lemmas in the WFL dataset means that when making general considerations on the distribution of the *-bil* suffix in Classical and Late Latin, for instance, one should keep in mind that a portion of what is extracted from WFL might need to be disregarded.

4. Word Formation in *LiLa*

The inclusion of the WFL data into the LiLa Knowledge Base provided an opportunity to devise a different way to account for those processes that do not fit into a linear hierarchical structure. The recent emergence of interest in the application of Word and Paradigm (W&P) models to derivational morphology (Blevins, 2016) and, in particular, the theoretical framework of the word-(and sign)-based model known as Construction Morphology (CxM) (Booij, 2010), has been crucial for designing the inclusion of the WFL data into LiLa.[14] CxM revolves around the central notion of "constructions", conventionalised pairings of form and meaning (Booij, 2010, p. 6). For example, the English noun *driver* is analysed in its internal structure as $[[\texttt{drive}]_{V}\ \texttt{er}]_{N} \longleftrightarrow [\texttt{someone who drive(s)}_{V}]_{N}$. Constructions may be hierarchically organised and abstracted into "schemas". The following schema, for instance, describes a generalisation of the construction of all words displaying the same morphological structure as *driver*, like for instance *buyer*, *player* and *reader*: $[[\texttt{x}]_{Vi}\texttt{er}]_{Nj} \longleftrightarrow [\texttt{someone who SEM}_{Vi}]_{Nj}$.[15]

One of the most crucial fundamentals of CxM is that schemas are word-based and declarative, which means that they describe static generalisations, as opposed to explaining the procedure of change from one PoS to another like WFRs do (e.g. V-to-N *-er*). Also, schemas are purely output-oriented, so the focus is not on the derivational process anymore, but on the morphological structure of the word itself. This translates into a concept that is especially fit to be included in the LiLa Knowledge Base: if words can be described as a construction of formative elements, these can be organised into (connected) classes of objects in an ontology.

In particular, the LiLa ontology defines three classes of objects that are used for the treatment of derivational morphology: (1) Lemmas, (2) Affixes, divided into Prefixes and Suffixes, and (3) Bases. Each Affix is labelled with a citation form chosen to repre-

[14]For a full description of the theoretical justification of why W&P approaches such as CxM can be advantageous in describing word formation in Latin, see Litta and Budassi (Forthcoming).

[15]Subscript like *V*, *N*, *i* and *j* are traditionally used as placeholders for morphological (e.g. *V* and *N*) and semantic (e.g. *i* and *j*) features that are referred to separately.

sent it in the Knowledge Base. Bases are currently not assigned a further description, nor are they associated with any human-readable string (such as, for instance, a lexical stem); they are simply defined by their function of connectors between lemmas belonging to the same word formation family.[16] Like any object in LiLa, Affixes and Bases are assigned a unique identifier.

These three classes of objects are connected to each other via object properties that are also formalised in the ontology of LiLa. A Lemma node is linked (a) to the Affix nodes that are part of its construction through the relationship `hasPrefix` or `hasSuffix` and (b) to its Base (or Bases, in the case of compounds) through the relationship `hasBase`. No relation of derivational nature is posed between lemmas, so as not to take assumptions on the direction of the formative process.

Figure 3. The word formation family of febris *in LiLa.*

[16] In what follows, therefore, bases will be mentioned using the numeric ID that forms the last component of their URI (e.g. Base 217 is the base that has the URI: `http://lila-erc.eu/data/id/base/217`); occasionally, if no ambiguity arises, they are also cited by mentioning one of the lemmas that are attached to it (thus, the same base can be referred to as "the base of *gigno*").

Figure 3 shows the word formation family of *febris* as it is represented in LiLa. Nodes for Lemma objects are assigned a unique identifier, that can be read by hovering on the node. By expanding the toggles on each node it is possible to view all the information linked to it. For example, the lemma with ID 103049[17] has written representation 'febriculentus', this information can be found on the node and on a list that opens on the right hand side of the screen detailing the type of node, its label and written representation(s). Node 'febriculentus' (which belongs to the class Lemma) is connected through a `hasPOS` property to the node 'adjective', through `hasInflectionType` to 'first class adjective', through `hasDegree` to 'positive', to two suffixes with written representation '-(us/un)cul' and '-(i/o/ul)ent' respectively, and to Base 1633. This base node has 7 ingoing edges, one for each of the lemmas belonging to the word formation family *febris* belongs to. One of these lemmas, *febrifugia* is also related to another base (1719), which connects all members of the word formation family of *fugo*, because it is a compound.

5. Querying the Lemma Collection of LiLa

LiLa provides also a user-friendly interface to query its collection of lemmas.[18] The query results are shown as lists of lemmas, and all the information linked to a lemma in the Knowledge Base can be visualised via a simple LodLive application.[19]

In this section, we describe the query interface of the lemma collection, specifically focusing on the retrieval and visualisation of information about derivational morphology.

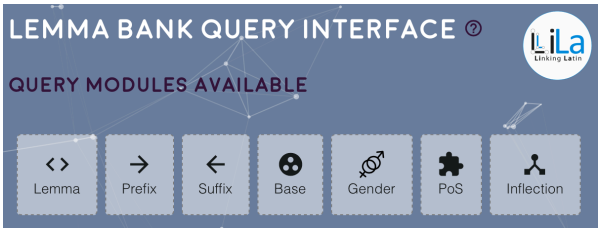

Figure 4. The query interface of the LiLa lemma collection.

Figure 4 shows a screenshot from the query interface home page. Users can select one or more query modules. On selection, the "Lemma" module allows free text (and

[17] `http://lila-erc.eu/data/id/lemma/103049`.

[18] `https://lila-erc.eu/query`.

[19] `http://en.lodlive.it`.

RegEx) input, while the others offer a range of values available from a drop-down menu. The list of lemmas resulting from the combination of values from the modules selected is updated dynamically. For instance, the module-value couples Prefix=*ante-* and PoS=*Verb* return a list of all verbs in the lemma collection that are formed with the prefix *ante-*.

For every query, it is possible to download the results as a CSV file (Comma-Separated Values) and also to copy the SPARQL code for the query, which can be reused (and obviously modified *ad libitum*) on the endpoint of the LiLa Triplestore.[20]

For example, the selection of the Lemma query module with free text value *febris* returns two records with written representation *febris* in the lemma collection, a common noun and a proper noun (both feminine of the third declension). From the resulting list it is possible to consult data on a chosen lemma from two different points of view: a data sheet and a graph view in LodLive.

The data sheet includes the URL for the relevant lemma, its label and written representation(s), its type, links to bases, affixes and suffixes (if any), gender, inflectional category and PoS. Figure 5 shows the data sheet for the common noun *febris*. All information on the data sheet is clickable and brings to other relevant data sheets. For example, the URL and number of the base leads to the dedicated web page of the data point.

The second, more dynamic way of visualising the data is the graph view. Here, lemmas, affixes, bases and other objects from the ontology are shown as circle shaped nodes surrounded by a number of smaller satellite circles. Clicking on each of them reveals all the information linked to the nodes in the Knowledge Base. Figure 6 shows the information linked to base node labelled Base 1633.[21] Beside the entity type of the node (Base), the lemmas that belong to the same family of *febris* are shown. All lemmas are linked to the base node via the `hasBase` property.

The LiLa interface also allows users to run complex queries on derivational information, by combining different modules. Figure 7 shows an example of a query that searches for verbs formed with prefix *de-*, suffix *-sc* and at least one written representation of their cita-

Figure 5. The data sheet of lemma febris.

[20] `https://lila-erc.eu/sparql`.

[21] `http://lila-erc.eu/data/id/base/1633`.

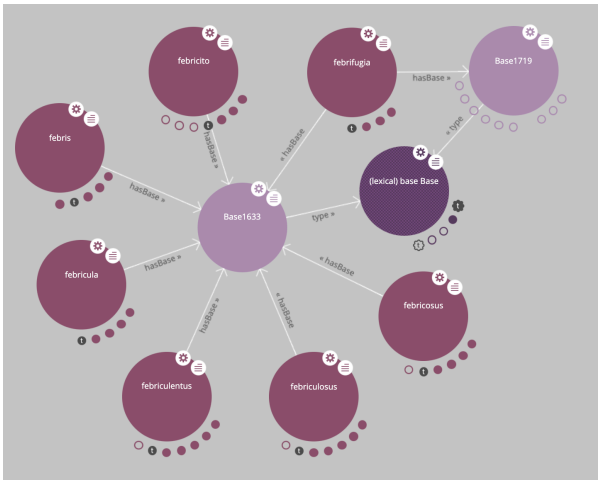

Figure 6. The graph view of the word formation family of febris *in the query interface of the LiLa lemma collection.*

tion form beginning with the characters *de* (expressed through the regular expression `^de`).
This last condition is added in order to exclude from the results (22 items) those verbs that contain prefix *de-* not at the start of the word, as it is the case with *condeliquesco* 'to dissolve', where prefix *con-* appears at the start of the word before *de*.

Figure 8 shows the graph view of the derivational information connected to the node for *condeliquesco*. This figure exemplifies how the idea of linking data is realised in the LiLa Knowledge Base. The node for *condeliquesco* is connected to three nodes concerning derivational information, namely those for prefixes *de-* and *con-* and that for the Base 1266. Each of these nodes connects all the words in the lemma collection of LiLa that respectively are formed with prefix *de-* (like, for instance, *debello* 'to fight a battle (or a war) out') or suffix *con-* (e.g., *accommodo* 'to fit'), and those that share the same lexical base of *condeliquesco*, like for instance the adjective *perliquidus* 'completely fluid'.

6. Use-case Scenarios

This section presents some examples of the use of derivational data in the LiLa Knowledge Base. Examples are organised in three subsections, respectively dedicated (1) to investigations that can be performed on derivational data alone, like for instance the distribution of affixes in the lemma collection of LiLa (Subsection 6.1), (2) to complex queries on different resources interlinked in LiLa, where derivational and textual

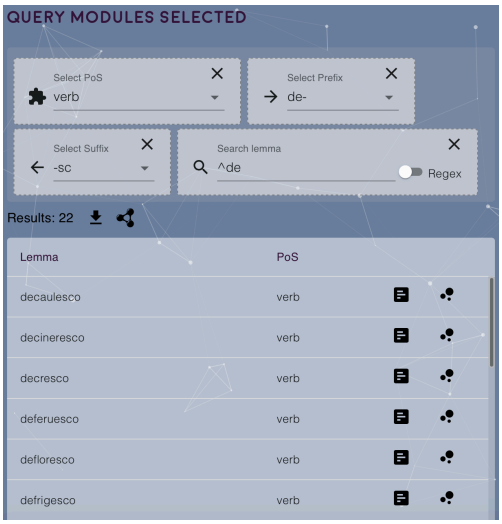

Figure 7. A complex query on derivational information in the query interface of the LiLa lemma collection.

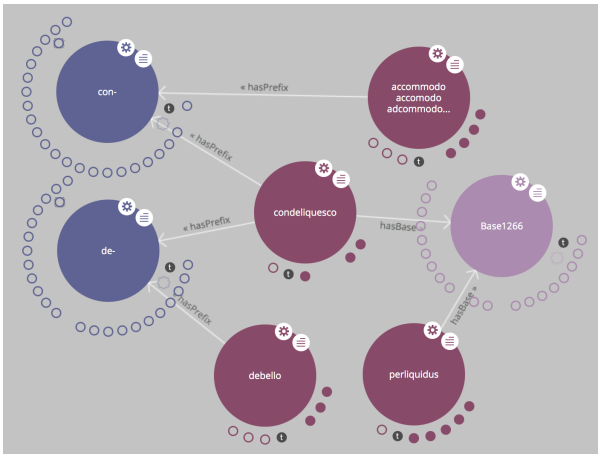

Figure 8. The graph view of lemma condeliquesco*.*

data from corpora are cross-referenced (Subsection 6.2) and (3) to the combination and comparison of two resources providing derivational information.

6.1. Inside Derivational Data

As it stands, the LiLa Knowledge Base can support a number of investigations on word formation that were not so comprehensively and instantly feasible before.

One of the most basic queries is the retrieval of all lemmas linked to the same lexical base (i.e. all the members of a word formation family) via the `hasBase` object property. The query starts by finding a given lemma, then identifies the lexical base linked to it, and finally lists all the other lemmas connected to the same base. For instance, starting from the adjective *formalis* 'of a form, formal', 67 lemmas are retrieved.[22] These can be grouped by PoS: 32 adjectives (including e.g. *serpentiformis* 'shaped like a snake' and *uniformis* 'uniform'), 25 nouns (e.g. *forma* 'shape', *formella* 'mould' and *informator* 'one who shapes'), 9 verbs (e.g. *informo* 'to shape, to inform' and *reformo* 'to transform'), and 1 adverb (*ambiformiter* 'ambiguously').

Similar queries can be performed using affixes as starting points. These can be useful, as an example, when considering that the same affixes have a tendency to be frequently associated in complex words. The LiLa Knowledge Base allows accurate empirical evidence on which among affixes are more often found together in a lemma. A query that performs this operation traverses all the lemmas in the LiLa Knowledge Base, counts all couplets of prefixes and/or suffixes, and finally reports statistics on those that are most frequently associated.

For example: with 121 instances, the most frequently associated prefixes in the LiLa lemma collection are *con-* and *in-* (with meaning of negation).[23] These two affixes are preponderantly found together in adjectives (96), such as *incommutabilis* 'unchangeable', while less frequently in nouns (23, e.g. *inconsequentia* 'lack of consistency') and adverbs (2, *incommote* 'immovably, firmly' and *incorribiliter* 'incorrigibly'). With 79 lemmas, the association of (negative) *in-* prefix and *ex-* is less frequent; examples are for instance adjective *inefficax* 'unproductive' and noun *inexperientia* 'inexperience'.

As for suffixes, the most frequent association is that of -(*i*)*t* and -(*t*)*io*(*n*), which are found in combination in 214 nouns such as *dissertatio* 'dissertation' and *excogitatio* 'a thinking out'. The second most attested combination (153 lemmas) involves again -(*i*)*t* and the suffix -(*t*)*or*, the latter mainly typical of agent or instrumental nouns. This association occurs in nouns like *dictator* 'dictator' and the adjective *gestatorius* 'that serves for carrying'.

The two most productive associations between a prefix and a suffix in LiLa are those between the negative *in-* prefix and the suffix *-bil* (296 lemmas, such as adjec-

[22] The starting word *formalis* is included in the count.

[23] In Latin there are two prefixes *in-*, one with negative and one with entering meaning.

tive *insuperabilis* 'that cannot be passed'), and between the prefix *con-* and the suffix *-(t)io(n)*, with 290 lemmas, which are mostly nouns like *contemplatio* 'viewing, contemplation' and *reconciliatio* 're-establishing'.

6.2. Outside Derivational Data

The data on word formation stored in the LiLa Knowledge Base can also be used to perform corpus-based queries. Links between lemmatised texts and the lemmas of the LiLa collection are then used to test how the different prefixes, suffixes or bases are distributed in texts.

As an example, we investigate the most frequently occurring derivational morphemes in a group of annotated textual resources, namely three Latin treebanks and one lemmatised corpus. The treebanks are the *Index Thomisticus* Treebank (IT-TB) (Passarotti, 2011), based on works written in the 13th century by Thomas Aquinas (approximately 400k nodes), the PROIEL corpus (Haug and Jøhndal, 2008), which includes the entire New Testament in Latin (the so called *Vulgata* by Jerome) along with other prose texts of the Classical and Late Antique period, and the Late Latin Charter Treebank (Korkiakangas and Passarotti, 2011) (LLCT; around 250k nodes), a syntactically annotated corpus of original VIIIth-IXth century charters from Central Italy. To those treebanks we add also the corpus of the Latin works of Dante Alighieri (13/14th century), distributed as part of the *Dante Search* project.[24]

All four resources include lemmatisation, which we use to connect the corpus tokens to the lemmas in LiLa following the procedure presented in Mambrini and Passarotti (2019). Once that the tokens in the annotated texts are linked to the LiLa lemmas, we use the SPARQL query language to extract information about the derivational morphemes attested in each corpus. While some lemmatised resources, like the IT-TB and the works of Dante, are already accessible via a dedicated endpoint provided by LiLa,[25] virtually any other lemmatised corpus can be linked and searched using local files with the methodology described in Mambrini and Passarotti (2019); the results reported here for PROIEL and LLCT were obtained by querying local files.

Table 1 reports some simple statistics on the incidence of verbs formed with the prefixes *de-* and *e(x)-* in the two corpora available in LiLa (IT-TB and Dante) and in the two treebanks queried locally (PROIEL and LLCT); in the table, we provide both the number of occurrences of any given verb formed with the two prefixes (Tokens), and of the different verbs attested (Lemmas).

The LiLa Knowledge Base can also help researchers with questions such as: what are the most frequent affixes in Latin texts? In order to observe the distribution of prefixes and suffixes in the lexicon of the PROIEL corpus, the most balanced Latin treebank in terms of textual genres, we can start from a SPARQL query that retrieves

[24] `https://dantesearch.dantenetwork.it`.

[25] `https://lila-erc.eu/sparql/corpora`.

	de-		e(x)-	
Corpus	Tokens	Lemmas	Tokens	Lemmas
IT-TB	1,013	52	1,098	76
Dante Search	299	81	379	126
PROIEL (UD)	1,011	128	1,328	152
LLCT	209	28	155	16

Table 1. Number of occurrences of verbs formed with the prefixes de- *and* e(x)- *in four corpora.*

Affix	Type	Lemmas	Tokens
-(t)io(n)	Suffix	393	2,157
con-	Prefix	344	3,297
ad-	Prefix	201	2,514
e(x)-	Prefix	197	2,713
-i	Suffix	194	2,052
de-	Prefix	182	1,294
in (entering)-	Prefix	178	1,559
-(i)t	Suffix	158	1,275
-tas/tat	Suffix	157	1,582
re-	Prefix	151	1,858

Table 2. The 10 affixes most frequently associated with a token in the PROIEL corpus.

all tokens linked with a LiLa lemma that is, in turn, connected to one or more derivational morphemes. The results are reported in Table 2. Here, while tokens of words derived with the suffix -$(t)io(n)$ rank only in the fourth place and are considerably outnumbered by tokens formed with the prefix *con-*, lemmas displaying the suffix -$(t)io(n)$ outnumber all the others; this means that, while there are more occurrences of tokens formed with *con-*, the PROIEL texts contain more words formed with -$(t)io(n)$. Such distribution reflects the greater productivity of this suffix as recorded in WFL: 2,686 lemmas formed with -$(t)io(n)$ vs. 748 with *con-*.

6.3. Comparing Derivation in Different Resources

One added value of including linguistic resources in a lexically-based Knowledge Base like LiLa, where data and metadata from distributed sources interact, is that

information about a lexical item provided by different resources can be combined and possibly compared.

Beside WFL, LiLa now includes another lexicon that provides derivational information. The Knowledge Base has been recently enriched with etymological data taken from the *Etymological Dictionary of Latin and the other Italic Languages* (de Vaan, 2008). By adopting the Ontolex-lemon model, now a *de facto* standard for the representation of lexical resources, and the lemonEty expansion designed to represent also etymological information (Khan, 2018), the etymologies were modelled to represent scientific hypotheses about the inheritance links between Latin words and the reconstructed forms of the Proto-Italic (PIt) and Proto-Indo-European (PIE) languages (Mambrini and Passarotti, 2020).

For each entry, the dictionary lists a number of derivatives of the head word in Latin, which are limited to those words whose first attestation is dated no later than the times of Cicero (106-43 BCE; de Vaan, 2008, 11-2). For instance, for the entry *donum* 'gift, present', the derivatives *donare* 'to present, give', *donabilis* 'worthy to be the recipient' and *donaticus* 'formally presented' are reported. Thus, with the inclusion in LiLa of the derivational data from WFL and of the etymological dictionary by de Vaan (2008), it becomes possible to compare (and possibly enhance) the two resources.

The comparison process starts from collecting the relevant data. We begin by selecting those lemmas of the LiLa collection that are assigned etymological information taken from de Vaan (2008). For each lemma selected, we then check whether it is connected to at least one base node in the Knowledge Base, which implies that the lemma is recorded in WFL as the member of a word formation family. We finally repeat the step for each of the Latin derivatives of the lexical entries in de Vaan (2008) included in LiLa, checking if they are present in the LiLa collection and if they are connected to at least one base node.

By using the data collected with the methodology described above, the derivatives from de Vaan (2008) are then compared to those from WFL as recorded in LiLa. This is performed in two steps:

- we calculate the number of different base nodes in LiLa connected to the derivatives listed in a lexical entry from de Vaan (2008). This informs us on whether the derivatives match the same word formation family as WFL or whether they are part of different families;
- for each word formation family in WFL, we collect all members not included among the set of derivatives of de Vaan (2008). As the derivatives in de Vaan (2008) are selected to represent only the earliest phases of the history of Latin until Cicero, this step allows us to extend the number of derivatives with words of later attestation.[26]

[26]In turn, also word formation families in WFL can be enhanced with derivatives from de Vaan (2008). There are, indeed, cases where a derivative reported by de Vaan (2008) is not recorded in WFL. In such

To give an example of how the comparison process works, in the case of the lexical entry *donum*, all the 3 derivatives reported by de Vaan (2008) are present in the LiLa Knowledge Base and all of them are connected to the same base,[27] and therefore belong to the same word formation family in WFL. Therefore, de Vaan (2008) and WFL agree that *donum* and its 3 derivatives share the same 'derivational history'. By collecting all the members of the WFL word formation family of *donum*, on the other hand, it is possible to enhance the set of derivatives reported by de Vaan (2008) with 14 further words.[28] Figure 9 shows the graphical representation of the connection of the lemmas *condonatio*, *donatiuncula*, *donarius* and *dono* which share the same base node of *donum* in LiLa.[29] Note that the lemma *donum* (with graphical variants *donom*, *dunom* and *dunum*) is connected to a Lexical Entry (*dōnum*) in de Vaan (2008) via the property `canonicalForm` and, from there, to its PIE and PIt reconstructed forms.[30].

Entries/derivatives of de Vaan (2008) with 1 base in LiLa	675
Entries/derivatives of de Vaan (2008) with 2+ bases in Lila	429
Entries of de Vaan (2008) not in WFL	14
Total de Vaan (2008) in LiLa	1,118

Table 3. Comparison between de Vaan (2008) and WFL.

Table 3 reports the number of lexical entries from de Vaan (2008) whose Latin derivatives included in WFL are connected respectively to one (675) and to two or more base nodes (429) in LiLa. Furthermore, Table 3 reports that 14 entries of the etymological dictionary are not in WFL (although 9 of them are indeed contained in the LiLa collection of lemmas).

The 675 entries of de Vaan (2008), whose derivatives included in WFL are all connected to only one base in LiLa, match those cases where the two lexical resources agree on the derivational history of the words concerned. Despite such agreement, the two resources diverge largely in the number of derivatives reported for each lex-

cases, the word in question is either absent from the LiLa collection or, if present, it is not connected to any base node (as this information should come from WFL).

[27] `https://lila-erc.eu/data/id/base/1012`.

[28] *condonatio* 'grant, donation, remission', *condonatrix* 'one who remits' (feminine), *condono* 'to deliver up, to remit', *donarium* 'temple treasury, endowment', *donarius* 'donee', *donatio* 'donation', *donatiuncula* 'small donation', *donativum* 'largess', *donator* 'donor' (masculine), *donatrix* 'donor' (feminine), *donifico* 'to make presents', *indonatus* 'unrewarded, unendowed', *redonator* 'restorer' and *redono*'to restore'.

[29] The verbal lemma *dono* corresponds to the derivative *donare* in de Vaan (2008), as in the etymological dictionary the citation form for verbs is the present infinitive instead of the first singular person of the present indicative (used in LiLa and WFL).

[30] See Mambrini and Passarotti (2020) for details.

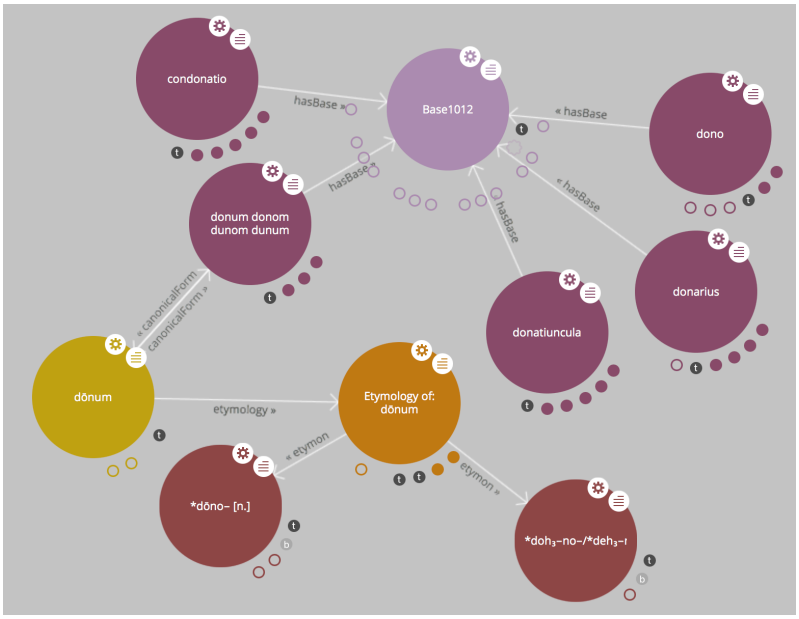

Figure 9. Lemmas connected to the same Base node of donum *in LiLa.*

ical entry or word formation family. Only 12 entries of de Vaan (2008) show a full overlap with the corresponding family in WFL, such as in the case of the adjective *sons* 'guilty', whose derivatives in both de Vaan (2008) and WFL are *insons* 'innocent' and *sonticus* 'genuine, valid'. In LiLa, these three lemmas are the only ones sharing a connection to the same base.[31]

The remaining 663 entries of de Vaan (2008) whose derivatives included in WFL are all connected to the same base node in LiLa display a different number of derivatives than those included in the same family in WFL. This happens either because a derivative reported by de Vaan (2008) is not included in WFL (and possibly in LiLa, too), or because a member of the word formation family of WFL is not reported among the derivatives for its corresponding entry in de Vaan (2008). In the former scenario, the WFL families could be enhanced with the additional data provided by de Vaan (2008); in the latter, WFL could provide candidate words to expand the range of derivational and etymological explanation provided by de Vaan (2008) with fur-

[31] https://lila-erc.eu/data/id/base/3278.

ther (and later attested) derivatives.[32] This would mean that de Vaan (2008) could contribute to the addition of 321 derivatives missing in the corresponding WFL families,[33] and that WFL could enhance the pool of the early attested derivatives in de Vaan (2008) with a total 10,438 extra lemmas.

As mentioned, it is not surprising that the number of candidate words for the inter-resource enhancement provided by WFL is much bigger than that by de Vaan (2008). While de Vaan (2008) focuses on Indo-European etymology and on the earliest stages of linguistic history, WFL aims to be as exhaustive as possible in lexical coverage, and thus includes the entire Classical Latin vocabulary well after the Republican period. Such different approach pursued by the two resources becomes an added value when these are compared and joined through a Knowledge Base like LiLa, because they provide different, yet compatible, information about the same items. Hence the added value is not only to contribute 10,438 additional lemmas to the total of derivatives reported by de Vaan (2008) in his entries, but also to obtain, for these lemmas, etymological information inherited through their connection to the same base in LiLa.

One example showing mutual enhancement is the verb *fluo* 'to flow, run (of waters)'. The lexical entry for *fluo* in de Vaan (2008) reports 26 derivatives, 23 out of which are connected to the same base in LiLa.[34] Indeed, although all the 26 derivatives of *fluo* are present in the LiLa collection, 3 of them are not recorded in WFL, which means that they are not connected to any base in LiLa. This is a case of enhancement of WFL (and, as a consequence, of LiLa, too) from de Vaan (2008), as the 3 derivatives concerned are all good candidates to be connected to the same base node of the other 23 of the same entry of de Vaan (2008). The opposite enhancement, from WFL/LiLa to de Vaan (2008), is much bigger, as there are 121 lemmas connected to the same base of *fluo* in LiLa that are not reported among its derivatives in the etymological dictionary. We manually checked that all these 121 lemmas are good candidates to be included in the list of derivatives of *fluo* in de Vaan (2008). They can all inherit the etymological information offered by the dictionary's entry.

On the other hand, there are 429 entries in the etymological dictionary whose Latin derivatives are connected to 2, or more, base nodes in LiLa (i.e. they belong to different word formation families in WFL). The reason for this falls in two main categories.

First, there can be errors in WFL, namely cases of words that must belong to the same family but are instead spread in two, or more families. Most of the cases result-

[32] We speak of 'candidate words', because each of them must be checked manually, as it cannot be taken for granted that all derivatives provided by de Vaan (2008), as well as all members of the WFL families, can be transferred from one resource to the other. However, the fact that all the derivatives of an entry of the etymological dictionary present in WFL are connected to the same base node in LiLa is a good argument in support of the portability of the information between the two resources.

[33] These 321 derivatives can be either absent from the LiLa collection, or they can be present but not connected to the base node of the WFL word formation family in question.

[34] `https://lila-erc.eu/data/id/base/183`.

ing from the 429 entries in question fall in this category. The identification of such errors must be considered a positive outcome of joining the two resources through LiLa, in that it helps to improve the quality of the connected resources. One example is the verb *eviro* 'to unman', which is listed among the derivatives of *vir* 'man' in de Vaan (2008), but it is not connected to the same base node of *vir* in LiLa,[35] and thus it does not belong to the same word formation family of *vir* in WFL. This kind of error, once discovered, can be rectified.

Second, there are cases of discrepancy due to the different perspective of the two lexical resources, reflecting the approach to word formation they pursue, their background motivation, or a different stance on the history of words. For example, in de Vaan (2008) the entry *mens* 'mind' records 8 derivatives, 7 of which match lemmas connected in LiLa to the same base of *mens*;[36] however, the noun *mentio* 'mention' in LiLa is connected to another base,[37] namely the one that connects words belonging to the WFL word formation family whose ancestor is the verb *miniscor* 'to remember'. As mentioned above, in WFL decisions on derivation are mostly based on OLD. Here the lexical entry for *mentio* is recorded as originating from the reconstructed root **men* plus suffix *-tio*, and the entries for *mens* and *miniscor* are referred to for comparison. The entry for *miniscor* states that it is cognate with *memini* 'to remember' and refers to *mentum*, i.e. the perfect participle of *miniscor*. In WFL, *mentio* is recorded as derived from *miniscor* and does not belong to the same family of *mens*, because the suffix *-tio* tends to form nouns from verbs, in particular from the base of the perfect participle of the input verb. This is exactly what happens in *mentio*, which is derived from the base of *mentum* (perfect participle of *miniscor*). In de Vaan (2008), on the other hand, *mentio* is listed as derivative in the entry of *mens*, while the verb *miniscor* is recorded as cognate with *memini*. Although the PIE words that *mens* and *memini* are derived from are etymologically related (a fact that is reflected in the cross references between the entries in the dictionary), the two are discussed under different lemmas and thus they are linked to multiple WFL families.

Table 4 sums up the recording of the words concerned in OLD, LiLa, WFL and de Vaan (2008).

7. Conclusions

In this paper, we have described the treatment of word formation in the LiLa Knowledge Base, which links together distributed linguistic resources for Latin. By reporting a number of use-case scenarios of the Knowledge Base on different issues related to derivational morphology, we have shown how helpful linguistic resources

[35] Base node of *eviro* in LiLa: `https://lila-erc.eu/data/id/base/1554`. Base node of *vir* in LiLa: `https://lila-erc.eu/data/id/base/790`.

[36] `https://lila-erc.eu/data/id/base/259`.

[37] `https://lila-erc.eu/data/id/base/961`.

Word	OLD	LiLa	WFL	de Vaan (2008)
memini	underived in Latin	Base: 2353	ancestor	head word
mens	underived in Latin	Base: 259	ancestor	head word
mentio	<**men+-tio*	Base: 961	<*miniscor*	derivative of *mens*
miniscor	cognate with *memini*	Base: 961	ancestor	derivative of *memini*

Table 4. Comparison between OLD, LiLa, WFL and de Vaan (2008) on single words.

are when they are made interoperable. Indeed, the steady work done across the last decades on new digital corpora and lexica for Latin, together with the century-long tradition of lexicography for Classical languages, has led to the current availability of a large set of linguistic resources for Latin. In different ways, all these resources concern words. For this reason, LiLa's starting point is based on the idea of linking through lemmas; each connected resource then provides its contribution to the overall picture resulting from the joining of the appropriate (meta)data from all sources.

As for derivational morphology, the information recorded in the list of Latin lemmas in LiLa is based on the WFL lexicon, which was built on the portion for Classical and Late Latin of the Lemlat lexical basis. However, since LiLa is not meant to be limited to a specific era of Latin only, extending the coverage of WFL to the Medieval Latin lemmas included in Lemlat (around 86,000) represents a major next step for the coming years. Although probabilistic models can be used in the first phase of this task (like, for instance, the one described by Sumalvico, 2017), manual disambiguation of the results, as well as the retrieval of both false positives and negatives, is to be expected.

Another potential development of the description of word formation in the LiLa Knowledge Base would be to assign some kind of linguistic information to the base nodes, which are currently just empty connectors of lemmas belonging to the same word formation family. One possible solution could be to assign to each base a written representation consisting of a string describing the lexical "element" that lies behind each lemma in the word formation family (e.g. DIC- for *dico* 'to say', or *dictio* 'a saying'). This procedure is however complicated by the fact that different bases can be used in the same word formation family: for example *fer-*, *tul-* and *lat-* can all be found as bases in the word formation family the verb *fero* 'to bring' belongs to.

Acknowledgements

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.

Bibliography

Bird, Steven and Mark Liberman. A formal framework for linguistic annotation. *Speech communication*, 33(1-2):23–60, 2001. doi: 10.1016/S0167-6393(00)00068-6.

Blevins, James P. *Word and paradigm morphology*. Oxford University Press, Oxford, UK, 2016. doi: 10.1093/acprof:oso/9780199593545.001.0001.

Booij, Geert. Construction morphology. *Language and linguistics compass*, 4(7):543–555, 2010. doi: 10.1093/acrefore/9780199384655.013.254.

Budassi, Marco and Eleonora Litta. In Trouble with the Rules. Theoretical Issues Raised by the Insertion of -sc- Verbs into Word Formation Latin. In *Proceedings of the Workshop on Resources and Tools for Derivational Morphology (DeriMo)*, pages 15–26. Educatt, 2017.

Budassi, Marco and Marco Passarotti. 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, 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-2110.

Buitelaar, Paul, Philipp Cimiano, John McCrae, Elena Montiel-Ponsoda, and Thierry Declerck. Ontology lexicalisation: The lemon perspective. In *WS 2 Workshop Extended Abstracts, 9th International Conference on Terminology and Artificial Intelligence*, pages 33–36, 2011.

Cecchini, Flavio Massimiliano, Marco Passarotti, Paolo Ruffolo, Marinella Testori, Lia Draetta, Martina Fieromonte, Annarita Liano, Costanza Marini, and Giovanni Piantanida. Enhancing the Latin Morphological Analyser LEMLAT with a Medieval Latin Glossary. In Cabrio, Elena, Alessandro Mazzei, and Fabio Tamburini, editors, *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018). 10-12 December 2018, Torino*, pages 87–92. aAccademia university press, 2018. doi: 10.4000/books.aaccademia.3121.

Chiarcos, Christian. Interoperability of corpora and annotations. In *Linked Data in Linguistics*, pages 161–179. Springer, 2012. doi: 10.1007/978-3-642-28249-2_16.

Chiarcos, Christian and Christian Fäth. CoNLL-RDF: Linked Corpora Done in an NLP-Friendly Way. In Gracia, Jorge, Francis Bond, John P. McCrae, Paul Buitelaar, Christian Chiarcos, and Sebastian Hellmann, editors, *Language, Data, and Knowledge*, pages 74–88, Cham, 2017. Springer International Publishing. ISBN 978-3-319-59888-8. doi: 10.1007/978-3-319-59888-8_6. URL https://link.springer.com/content/pdf/10.1007%2F978-3-319-59888-8_6.pdf.

Chiarcos, Christian and Maria Sukhareva. OLiA - Ontologies of Linguistic Annotation. *Semantic Web Journal*, 6(4):379–386, 2015. doi: 10.3233/SW-140167. URL http://www.semantic-web-journal.net/content/olia-%E2%80%93-ontologies-linguistic-annotation.

de Vaan, Michiel. *Etymological Dictionary of Latin: and the other Italic Languages*. Brill, Amsterdam, 2008. ISBN 978-90-04-16797-1. URL https://brill.com/view/title/12612.

Declerck, Thierry, Piroska Lendvai, Karlheinz Mörth, Gerhard Budin, and Tamás Váradi. Towards linked language data for digital humanities. In *Linked Data in Linguistics*, pages 109–116. Springer, 2012. doi: 10.1007/978-3-642-28249-2_11.

Domenig, Mark and Pius ten Hacken. *Word Manager: A system for morphological dictionaries*, volume 1. Georg Olms Verlag AG, Hildesheim, 1992.

Forcellini, Egidio. *Totius latinitatis lexicon: Onomasticon ; 1 (A - B)*. Typis Aldinianis, 1867.

Georges, Karl Ernst. *Ausführliches Lateinisch-Deutsches Handwörterbuch*. Hannover: Hahn, 1913.

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

Gradenwitz, Otto. *Laterculi Vocum Latinarum*. Verlag Von S. Hirzel, Leipzig, 1904.

Haug, Dag TT and Marius Jøhndal. 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, Marrakesh, Morocco, 2008. European Language Resources Association (ELRA).

Hellmann, Sebastian, Jens Lehmann, Sören Auer, and Martin Brümmer. Integrating NLP using Linked Data. In *12th International Semantic Web Conference, Sydney, Australia, October 21-25, 2013*, 2013. doi: 10.1007/978-3-642-41338-4_7. URL `https://svn.aksw.org/papers/2013/ISWC_NIF/public.pdf`.

Hockett, Charles F. Two Models of Grammatical Description. *Words*, 10:210–231, 1954. doi: 10.1080/00437956.1954.11659524.

Ide, Nancy and James Pustejovsky. What does interoperability mean, anyway. *Toward an Operational*, 2010.

Khan, Fahad. Towards the Representation of Etymological and Diachronic Lexical Data on the Semantic Web. In McCrae, John P., Christian Chiarcos, Thierry Declerck, Jorge Gracia, and Bettina Klimek, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, 2018. European Language Resources Association (ELRA). ISBN 979-10-95546-19-1.

Kircher-Durand, Chantal. Syntax, morphology and semantics in the structuring of the Latin lexicon, as illustrated in the -lis derivatives. In Coleman, Robert, editor, *New Studies in Latin Linguistics, Proceedings of the 4th International Colloquium on Latin Linguistics, Cambridge, April 1987*, Cambridge, 1991. John Benjamins.

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

Lassila, Ora, Ralph R. Swick, World Wide, and Web Consortium. Resource Description Framework (RDF) Model and Syntax Specification, 1998.

Litta, Eleonora. On the Use of Latin -bilis Adjectives across Time. *Quaderni Borromaici. Saggi studi proposte*, 6:149–62, 2019.

Litta, Eleonora and Marco Budassi. What we talk about when we talk about paradigms. In Fernández-Domínguez, Jesús, Alexandra Bagasheva, and Cristina Lara-Clares, editors, *Paradigmatic relations in derivational morphology*. Forthcoming.

Litta, Eleonora and Marco Passarotti. (When) inflection needs derivation: a word formation lexicon for Latin. In Holmes, Nigel, Marijke Ottink, Josine Schrickx, and Maria Selig, editors, *Words and Sounds*, pages 224–239. De Gruyter, Berlin, Boston, December 2019. ISBN 978-3-11-064758-7. doi: 10.1515/9783110647587-015. URL `http://www.degruyter.com/view/books/9783110647587/9783110647587-015/9783110647587-015.xml`.

Litta, Eleonora, Marco Passarotti, and Francesco Mambrini. The Treatment of Word Formation in the LiLa Knowledge Base of Linguistic Resources for Latin. In *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology*, pages 35–43, Prague, Czechia, September 2019. Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics. URL https://www.aclweb.org/anthology/W19-8505.

Mambrini, Francesco and Marco Passarotti. Harmonizing Different Lemmatization Strategies for Building a Knowledge Base of Linguistic Resources for Latin. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 71–80, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4009. URL https://www.aclweb.org/anthology/W19-4009.

Mambrini, Francesco and Marco Passarotti. Representing Etymology in the LiLa Knowledge Base of Linguistic Resources for Latin. In *Proceedings of the 2020 Globalex Workshop on Linked Lexicography*, pages 20–28, Marseille, France, May 2020. European Language Resources Association. URL https://www.aclweb.org/anthology/2020.globalex-1.3.

McCrae, John P., Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. The OntoLex-Lemon Model: development and applications. In *Proceedings of eLex 2017*, pages 587–597, 2017. URL https://elex.link/elex2017/wp-content/uploads/2017/09/paper36.pdf.

Passarotti, Marco. Language resources. The state of the art of Latin and the *Index Thomisticus* treebank project. In Ortola, Marie-Sol, editor, *Corpus anciens et Bases de données*, number 2 in ALIENTO. Échanges sapientiels en Méditerranée, pages 301–320, Nancy, France, 2011. Presses universitaires de Nancy. ISBN 978-2-8143-0104-7.

Passarotti, Marco, Marco Budassi, Eleonora Litta, and Paolo Ruffolo. The Lemlat 3.0 Package for Morphological Analysis of Latin. In *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*, pages 24–31. Linköping University Electronic Press, 2017.

Prud'Hommeaux, Eric, Andy Seaborne, et al. SPARQL query language for RDF. W3C. *Internet: https://www.w3.org/TR/rdf-sparql-query/*[*Accessed on February 27th, 2019*], 2008.

Sumalvico, Maciej. Unsupervised Learning of Morphology with Graph Sampling. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing* (*RANLP 2017*), Varna, Bulgaria, 2017. doi: 10.26615/978-954-452-049-6_093.

Tombeur, Paul. *Thesaurus formarum totius latinitatis a Plauto usque ad saeculum XXum*. Brepols, Turnhout, Belgium, 1998.

Address for correspondence:
Marco Passarotti
marco.passarotti@unicatt.it
Università Cattolica del Sacro Cuore. Largo Gemelli, 1 - 20123 Milan, Italy