PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 119 OCTOBER 2022 67–92

Enhancing Derivational Information on Latin Lemmas in the LiLa Knowledge Base. A Structural and Diachronic Extension

Matteo Pellegrini,[a] Marco Passarotti,[a] Eleonora Litta,[a]
Francesco Mambrini,[a] Giovanni Moretti,[a] Claudia Corbetta,[a]
Martina Verdelli[b]

[a] CIRCSE Research Centre, Università Cattolica del Sacro Cuore, Milano
[b] Università di Pavia

Abstract

In this paper[1] we document both the structural and the diachronic extension of the derivational information provided in the LiLa Knowledge Base of interoperable linguistic resources for Latin. Structurally, to the flat information on families (i.e., groups of lemmas that share the same base) and affixes that is already available for the collection of lemmas of the LiLa Lemma Bank, we add hierarchical information on derivation processes provided by the Word Formation Latin (WFL) lexical resource, which in turn is characterised by a step-to-step morphotactic approach, where lexemes that are directly derived from one another are connected through word formation rules of different kinds. This is done by modelling WFL data into an ontology that adheres to the principles of the Linked Data paradigm, and connecting these data to the LiLa Lemma Bank. From a diachronic point of view, while the previous version of WFL only took Classical Latin lemmas into account, in this paper we describe the work conducted to produce a new version of WFL that is enhanced with derivational information on Medieval Latin lemmas. We then show how the data of this new version of WFL were used to extract derivational information in the format required by the LiLa Lemma Bank.

[1] This paper is an extended version of the work presented by Matteo Pellegrini, Eleonora Litta, Marco Passarotti, Francesco Mambrini and Giovanni Moretti at the Third International Workshop on Resources and Tools for Derivational Morphology (DeriMo), 9-10 September 2021, Nancy, France.

 Corresponding author: matteo.pellegrini@unicatt.it
Cite as: Matteo Pellegrini, Marco Passarotti, Eleonora Litta, Francesco Mambrini, Giovanni Moretti, Claudia Corbetta, Martina Verdelli. Enhancing Derivational Information on Latin Lemmas in the LiLa Knowledge Base. A Structural and Diachronic Extension. The Prague Bulletin of Mathematical Linguistics No. 119, 2022, pp. 67–92. doi: 10.14712/00326585.023.

1. Introduction

In recent years, the principles of the so-called Linked Data paradigm[2] have increasingly been applied to language data and metadata, with the aim of improving interoperability between resources that were originally developed for different purposes, and are therefore characterised by different formalisms and conceptual models. As a consequence, several resources are continuously being added to the so-called Linguistic Linked Data Cloud (Cimiano et al., 2020). Within this framework, the aim of the *LiLa* project[3] is to add Latin to this cloud, by creating a Knowledge Base (KB) of interlinked resources using a common vocabulary for knowledge description for the existing textual (i.e. corpora) and lexical (e.g. dictionaries and lexica) resources, as well as for Natural Language Processing (NLP) tools such as morphological analysers and part-of-speech taggers.

To do so, LiLa adopts the data model of the Resource Description Framework (Lassila and Swick, 1998), making use of a series of Semantic Web and Linked Data standards, including ontologies, to describe linguistic annotation (OLiA, cf. Chiarcos and Sukhareva 2015), corpus annotation (NIF, cf. Hellmann et al. 2013; CoNLL-RDF, cf. Chiarcos and Fäth 2017) and lexical resources (Lemon, cf. Buitelaar et al. 2011; OntoLex, cf. McCrae et al. 2017). According to the RDF data model, information is coded in terms of triples, that connect a subject – a labelled node – to an object – another labelled node or a literal – by means of a property – a labelled edge.

The backbone of the architecture of the LiLa KB is the Lemma Bank, a large collection of lemmas – i.e. citation forms – to which tokens from textual resources and entries from lexical resources, as well as the output of NLP tools, are connected. The Lemma Bank initially included a limited amount of derivational information on Classical Latin lemmas, taken from the Word Formation Latin (WFL) lexical resource (Litta and Passarotti, 2019). Initially, a choice was made not to include the entire information provided by WFL. That, however, might prove useful in certain circumstances.

In this paper, we start by detailing the organisation of derivational information in the Lemma Bank – in contrast with the one of the source from which it is extracted, namely WFL – and its coverage with respect to all the lemmas in the Lemma Bank (Section 2). We then show how we have extended the derivational information available in the LiLa Knowledge Base in two directions: structurally, and in terms of diachronic coverage. As for the former (Section 3), we describe a new ontology to model WFL data, so as to include it into the Lila KB in its entirety and link it to the Lemma Bank, thus having its information available in both formats within the same framework. We also discuss how our model interacts with other models developed by the Linked Data community – namely, the OntoLex-Lemon vocabulary for describing

[2] https://www.w3.org/DesignIssues/LinkedData.html.

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

lexical resources (McCrae et al., 2017; Buitelaar et al., 2011) and, more specifically, its Morphology Module (Klimek et al., 2019; Chiarcos et al., 2022). As for the latter (Section 4), we document the work done to produce a new version of WFL, enhanced to incorporate new derivational relations regarding Medieval Latin lemmas, in addition to the Classical Latin ones of the first version. We also show how these new relations have been exploited to provide additional "flat" derivational information in the Lemma Bank for the same Medieval Latin lemmas. Lastly, we draw a number of conclusions and highlight a few directions for future work (Section 5).

2. Derivational information in the LiLa Knowledge Base

The intuition behind the way in which LiLa connects different resources and tools is based on the central role of words: the idea is that textual resources are made of occurrences of words, lexical resources describe some properties of words, and NLP tools process words. As a consequence, in LiLa's architecture, a pivotal role is played by the class `Lemma` in LiLa's ontology,[4] a subclass of the class `Form` from OntoLex-Lemon. A lemma is defined as the canonical form of a lexical item, i.e. the one that is used for citation purposes by dictionaries and lemmatisers. The core of the LiLa KB is its Lemma Bank, a collection of around 200,000 Latin lemmas taken from the database of the morphological analyser Lemlat (Passarotti et al., 2017). Lemlat's database includes Classical Latin lemmas taken from Glare (2012), Georges (1998) and Gradenwitz (1904), Medieval Latin lemmas taken from du Cange et al.'s (1883-1887) glossary and proper names taken from Forcellini's (1965) *Onomasticon*. Through the Lemma Bank, the entries of the various lexical resources represented in LiLa and the tokens of the corpora included therein can be linked to the appropriate lemma, thus achieving the desired interoperability.

WFL is a word formation lexicon of Latin, characterised by a step-to-step morphotactic approach. This means that lexemes that are considered as deriving from one another are connected via word formation rules (WFR) of different kinds, by the application of one affix or one part-of-speech change at a time (note that circumfixation is not productive in Latin). There are compounding rules – with two, or more input lexemes and one output lexeme – and derivation rules – with only one lexeme as input and one as output. Among derivation rules, depending on the presence or not of affixes and their nature, there are affixal rules (more specifically, prefixal and suffixal) and conversion, when only a change of part of speech is involved. Furthermore, rules are classified according to the part of speech of the lexemes they take as input and output. All these features are illustrated in the examples of Table 1.

In WFL all the members of the same word formation family are grouped in a hierarchical structure, resembling that of a directed tree-graph, taking root from the ancestor – the lexeme from which all the members of the family ultimately derive –

[4] `https://lila-erc.eu/lodview/ontologies/lila/`.

input lexeme(s)	output lexeme	prefix	suffix	WFR
FELIX$_A$ 'happy'	FELICITAS$_N$ 'happiness'	-	-tas	A-To-N -tas
FELIX$_A$ 'happy'	INFELIX$_N$ 'unhappy'	in-	-	A-To-A in-
MALUS$_A$ 'bad'	MALUM$_N$ 'bad thing'	-	-	A-To-N
AGER$_N$ 'field' + COLO$_V$ 'cultivate'	AGRICOLA$_N$ 'farmer'	-	-	N+V=N

Table 1. Examples of Word Formation Rules in WFL

and branching out to all derivatives by means of the successive application of individual WFRs. For example, Figure 1 shows a portion of the family taking root from the ancestor lexeme FELIX$_A$ 'happy' in WFL: the four lexemes are linked by edges labelled by the affix involved in the WFR at work.

The Lemma Bank of the LiLa KB includes only a selection of the word formation information contained in WFL. Whenever a lemma is considered "derived", it is accompanied by information related to its morphological segmentation. So each derived lemma has a relation to one or more affixes and one or more (in case of compounding) bases, merely defined as abstract connectors between lemmas that belong to the same family. Hence besides `Lemmas`, two other classes are involved, namely `Affixes` – in their turn divided into `Prefixes` and `Suffixes` – and `Bases`. Each lemma is linked to the base to which it is related by means of the property `hasBase`, and to the affixes it contains by means of the property `hasPrefix` or `hasSuffix`.[5] As a result, the organization of derivational information in the Lemma Bank is flat, rather than hierarchical. Figure 2 shows how the four lexemes in the portion of the word formation family of FELIX$_A$ of Figure 1 are linked to the same base and to their affixes in the Lemma Bank, without any representation of both the WFR and the derivational hierarchical order.

Two different perspectives on derivational morphology are thus taken by WFL and by the Lemma Bank. In the 4-way classification of resources specialized in word formation operated by Kyjánek (2020), WFL can be considered as lexeme-oriented, since it describes the relationship among individual derivationally related lexemes. The approach of the Lemma Bank, on the other hand, is family-oriented, since it identifies groups of derivationally related lexemes sharing the same base.[6]

As is argued by Litta et al. (2020), the choice of a flat organization of derivational information in the Lemma Bank is due to its compatibility with more recent, Word-and-Paradigm theoretical approaches, such as Construction Morphology (Booij, 2010). Furthermore, this approach allows for a more natural treatment of cases that were

[5]These properties are all defined in LiLa's ontology.

[6]Kyjánek's (2020) classification also identifies morpheme-oriented resources – that decompose morphologically complex words into sub-word units – and paradigm-oriented resources – that aim at a modelling consisting of aligned morphological relations.

Figure 1. Word Formation in WFL

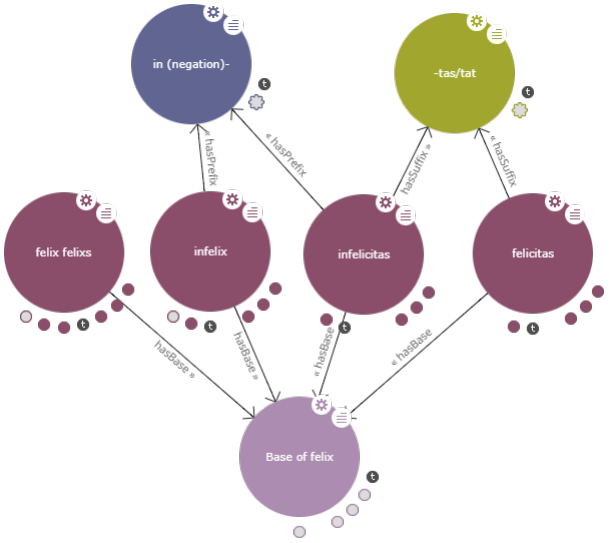


Figure 2. Word Formation in the Lemma Bank

problematic for the rigidly hierarchic structure in WFL (Litta and Budassi, 2020). For instance, WFL is forced to take a stance on the directionality of conversion processes, even when cases are not clear-cut, for instance ADVERSARIUS$_A$ 'opposed' vs. ADVERSARIUS$_N$ 'opponent'. An even more significant phenomenon is exemplified by a word like EXAQUESCO 'to become water': in this case, the step-by-step procedure of WFL requires the application of one affixation process at a time, but since neither *EXAQUO nor *AQUESCO are actually attested as intermediate steps, it has been necessary to add one of them (namely, *AQUESCO) as a fictional entry, so to comply with the requirements of WFL's general structure.

On the other hand, LiLa's flat representation of Latin word formation overlooks many details on the order of derivation. Since such information can still be potentially useful, we have decided to model the data from WFL so that it could be added to the LiLa KB. In this way, we achieve a structural enrichment of the knowledge of word formation represented in LiLa: both the flat organization of the Lemma Bank and the hierarchical organization of WFL are made available within a unified framework, leaving up to data users the choice about the kind of information that proves to be more appropriate for their specific needs. The details of the architecture of the WFL ontology designed for this purpose are the topic of Section 3.

One further direction to increase the degree of informativity of the LiLa KB on word formation concerns its diachronic coverage. At the time when WFL was com-

piled, entries from the Du Cange's medieval latin glossary had not yet been added into the Lemlat database, so WFL revolved around Classical and Late Latin only. Because the lexical basis of the LiLa KB is richer, we have felt the need to enrich its coverage even for what word formation information is concerned, and decided to keep this information in both theoretical formats, in order to offer the same level of flexibility as for the Classical Latin data. Since the bases, prefixes and affixes listed in the Lemma Bank are ultimately derived from WFL, that needed to remain the starting point for this enrichment phase. In Section 4, we describe the procedure that we followed to enhance WFL with new relations whose output is a Medieval Latin lemma, and to exploit this information to infer the base and the prefix and/or suffix of the corresponding lemmas in the Lemma Bank.

3. Modelling WFL with LiLa and Morph

The full inclusion of a lexical resource into the LiLa KB involves the modellisation of its data into an ontology that respects the Linguistic Linked Open Data (LLOD) standards. Figure 3 illustrates the details of our proposed ontology for WFL. Properties are represented as labelled directed arrows, and Classes as boxes. Boxes are colour-coded, according to the ontology where they are defined. This information is also expressed in the portion of the name that precedes the colon (e.g. `morph:Rule` means that "Rule" is a Class described in the "Morph" module of OntoLex). The arrows that are not labelled and have a white head are shortcuts for subclass relations.

Consistently with the spirit of Linked Data, our model makes use of classes and properties already defined in other ontologies. The most relevant for our purpose is OntoLex (cf. above in Section 1), both in its core model – where the class `LexicalEntry` is defined – and in more specific modules. In particular, we use the properties `source` and `target` from the Variation & Translation module (`vartrans`),[7] devised to handle relations of different kinds between lexical entries and senses, and several classes (the ones in blue in Figure 3) defined in the above-mentioned (cf. Section 1) Morphology module (`morph`). Furthermore, we refer to the classes already used in LiLa to treat derivational information (the ones in light green in Figure 3). Besides the ones taken from existing ontologies, we had to define some new classes and properties – identifiable by the `wfl` prefix and their white colour in Figure 3 – in order to properly model the information contained in WFL, as we will detail below.

There is one instance of the class `ontolex:LexicalEntry` for each lexeme contained in WFL. The entries of WFL that are directly derived from one another are linked by a specific instance of the class `morph:WordFormationRelation`, through properties taken from the `vartrans` module of OntoLex, having the entry of the base as `source` and the one of the derivative as `target`. Each relation is then connected to the WFR it instantiates (`wfl:WFLRule`) by means of the property `wfl:hasWordFormationRule`.

[7] `https://www.w3.org/2016/05/ontolex/#variation-translation-vartrans`.

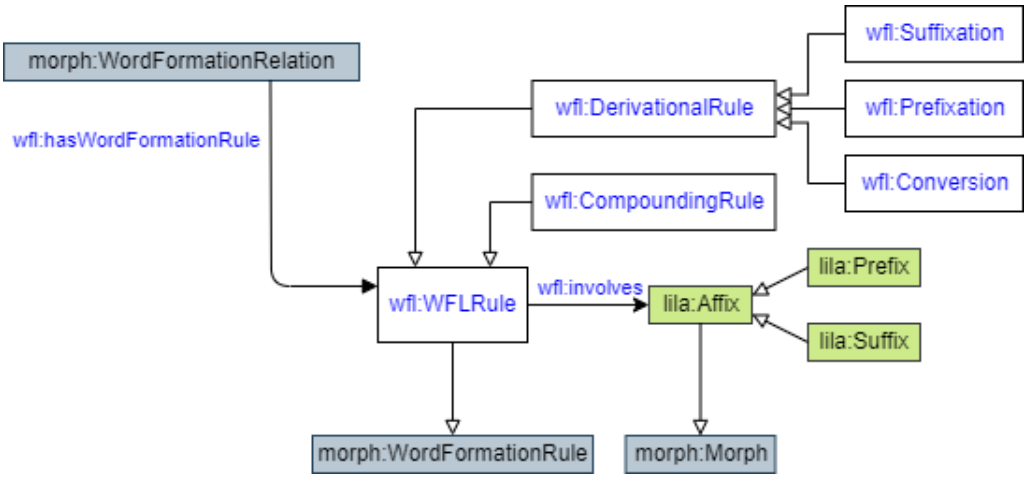


Figure 3. Architecture of the WFL ontology

The class `wfl:WFLRule` has two subclasses `wfl:DerivationalRule` and `wfl:CompoundingRule`, with the former having in its turn three subclasses `wfl:Suffixation`, `wfl:Prefixation` and `wfl:Conversion`, to reflect the organization of WFL.[8] Lastly, an object property `wfl:involves` links affixal rules to the prefix or suffix they display, as they are coded in LiLa – i.e. to an instance of either `lila:Prefix` or `lila:Suffix`, both subclasses of `lila:Affix`. Besides the use of `morph:WordFormationRelation`, the integration with the Morphology Module (`morph`)[9] of OntoLex is achieved by establishing a subclass relation between the rules of WFL and the ones of `morph` (`morph:WordFormationRule`) on the one hand, and between the affixes of the Lila ontology and the ones of `morph` (`morph:Morph`) on the other hand.

Figure 4 shows the model at work with specific pairs of related words with the Linked Data treatment of the derivation of INFELIX$_A$ 'unhappy' from FELIX$_A$ 'happy' on the one hand (left side of the image), of INFELICITAS$_N$ 'unhappiness' from INFELIX$_A$ 'unhappy' on the other hand (right side of the image).

[8]For the sake of completeness, we should mention that there is also a class `wfl:Backformation`, to account for a few cases of words that have been (probably) created by analogy, having been interpreted as the base of an already existing complex word that, however, has actually been formed by a different process. A clear example is the word CONSUEO 'to be used to', back-formed from CONSUESCO 'to become used to', that has actually been created by prefixing *con-* to SUESCO 'to become used to'. Since this phenomenon is very marginal in our data (there are only 5 cases in WFL), we do not go into more detail here.

[9]Note that this module is still the object of discussion in the Linked Data community: our proposal reflects its current state, but some details might change in the future.

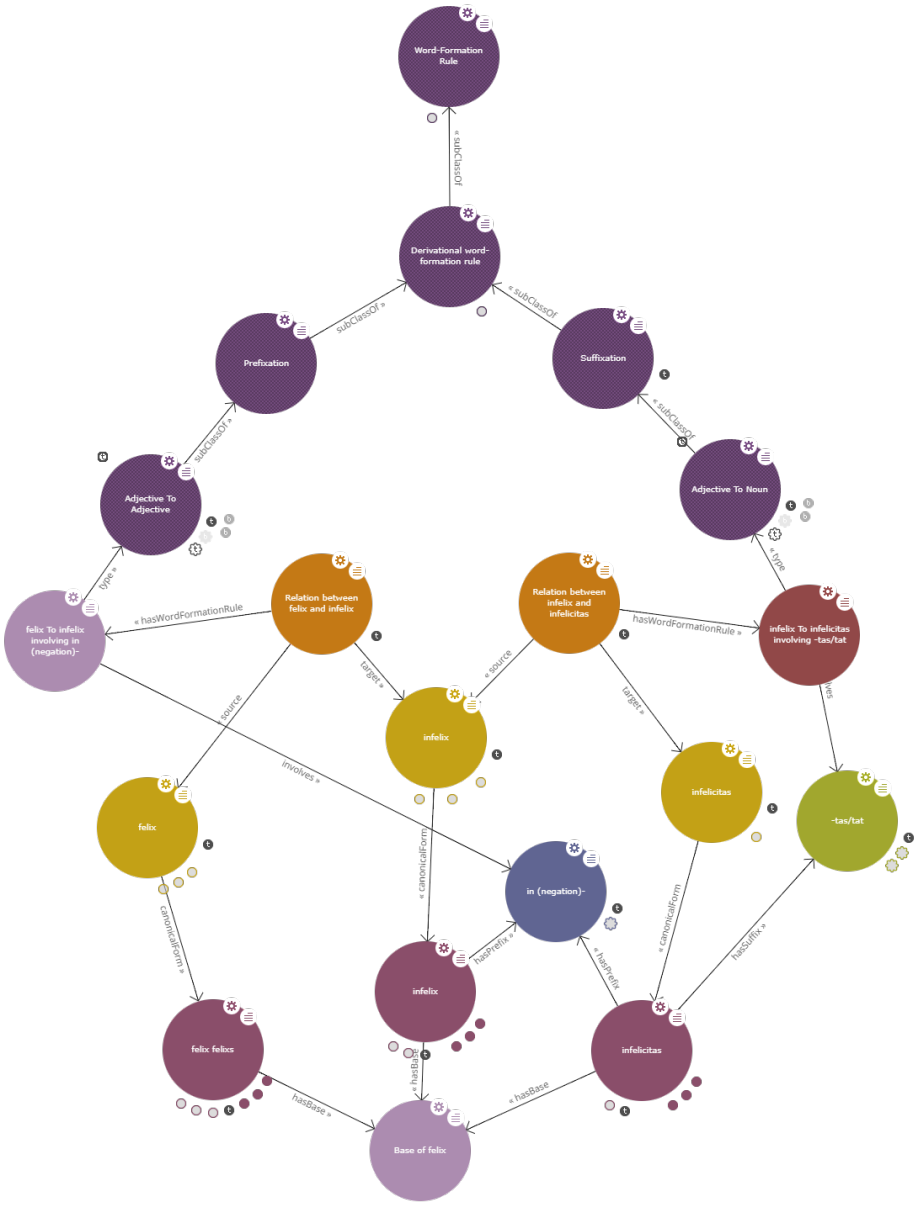


Figure 4. Modelling of prefixation and suffixation in the WFL ontology. Colours represent the classes in the LiLa ontology. E.g., dark purple: Classes; yellow: Lexical Entries; orange: WFL Relations; brownish purple: Lemmas, etc.

There is a specific word formation relation – in orange in the picture – between each of the entries of WFL that are considered as derived from one another, i.e. one between FELIX$_A$ and INFELIX$_A$ and one between INFELIX$_A$ and INFELICITAS$_N$. Each relation is instantiated by a specific WFR: see the nodes labelled as "felix To infelix involving in (negation)-"[10] and "infelix To infelicitas involving -tas/tat",[11] respectively. Starting from the one that forms INFELIX$_A$ from FELIX$_A$, it belongs to the class of prefixation rules creating adjectives from other adjectives: see the node with label "Adjective to Adjective" connected to the node with label "Prefixation" by means of the property `subClassOf` in Figure 4. Furthermore, this rule is also said to involve the prefix "in (negation)-". As for the WFR that forms INFELICITAS$_N$ from INFELIX$_A$, it belongs to the class of suffixation rules creating deadjectival nouns, and it involves the suffix "-tas/tat". Both prefixation and suffixation are sub-classes of the class of (affixal) derivational word formation rules, that on its turn is a sub-class of the class including all the rules of WFL. The bottom part of Figure 4 shows the connection with the Lemma Bank and the derivational information included therein. The lexical entries of WFL (above, in yellow) are connected to the lemmas of the Lemma Bank (below, in purple) by means of the OntoLex-Lemon property `canonicalForm`, and lemmas are connected to their shared base and to all the prefixes and suffixes they display, through the properties `hasBase`, `hasPrefix` and `hasSuffix` respectively.

There is one fact that is worth stressing in the description of this model: word formation relations always link a single source to a single target in our model. This restriction is inherited from the class of which `morph:WordFormationRelation` is stated to be a subclass, i.e. `LexicalRelation` from the `vartrans` module, that has been defined as connecting exactly two lexical entries. This has consequences on the treatment of compounding, as illustrated by Figure 5, showing the case of AGRICOLA$_N$ 'farmer' (from AGER$_N$ 'field' + COLO$_V$ 'to cultivate'). In this case, two relations are needed (one between the compound and its first member, one between the same compound and its second member), both of them pointing to the same WFR. A last remark should be made on the order of constituents, that is explicitly coded on each relation by means of the property `wfl:positionInWFR`: for instance, in the case of AGRICOLA$_N$ the value of this property is 1 for the relation between AGER$_N$ and AGRICOLA$_N$, 2 for the relation between COLO$_V$ and AGRICOLA$_A$.

For the sake of completeness, we also exemplify the treatment of noun-to-adjective conversion in Figure 6 below. It can be observed that the picture is similar to the one of affixal derivation (see Figure 4 above), the only difference being that the rule is not stated to involve any affix, consistently with the definition of conversion.

[10]The negative meaning of the prefix *in-* is specified to distinguish it from its omograph meaning 'entering', appearing for instance in INEO 'to go into, enter' from EO 'to go'.

[11]The notation of the shape of the suffix reflects the presence of different stem allomorphs in different forms, e.g. NOM.SG *infelici-tas* vs. GEN.SG *infelici-tat-is*.

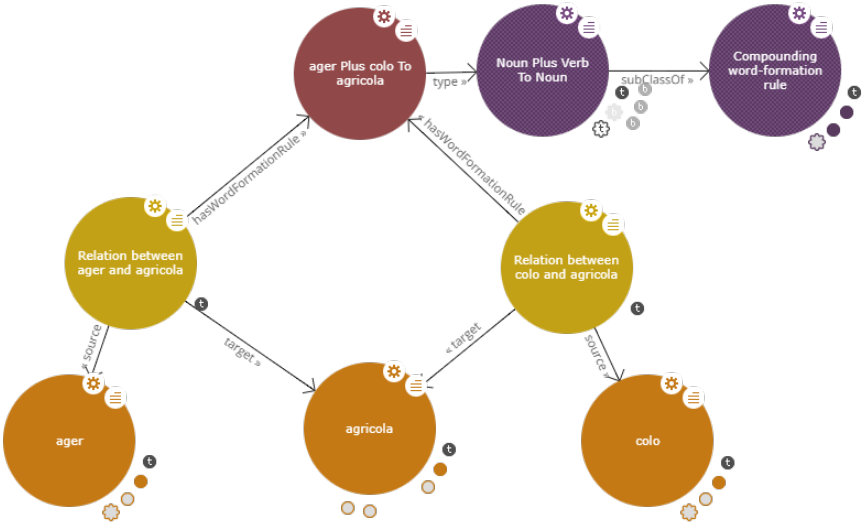


Figure 5. Modelling of compounding in the WFL ontology

4. Extracting derivational information for Medieval Latin lemmas

This section describes the procedure that we followed to enhance WFL with relations regarding Medieval Latin lemmas, and consequently to provide derivational information on those lemmas also in the Lemma Bank.

Our procedure is articulated (a) in an automatic extraction of words that might be derivationally related with one another based on their form – described in Section 4.1 – and (b) in a manual validation of the pairs that have been identified – addressing a number of issues raised during the process, some of which are discussed in Section 4.2. We then add the newly created relations to WFL and to Lemlat's database and we use them to establish new triples having an existing lemma of the Lemma Bank as subject and `hasBase`, `hasPrefix` and `hasSuffix` as properties, as detailed in Section 4.3. This section concludes by presenting some quantitative data on the outcome of this procedure in Section 4.4.

4.1. The methodology

The first step in our procedure is to identify the derivational processes that are most relevant in Classical Latin in terms of frequency: these are our starting point for the addition of new derivational information. To this end, we exploit the data of WFL, looking at the number of relations instantiating each WFR. For instance, the

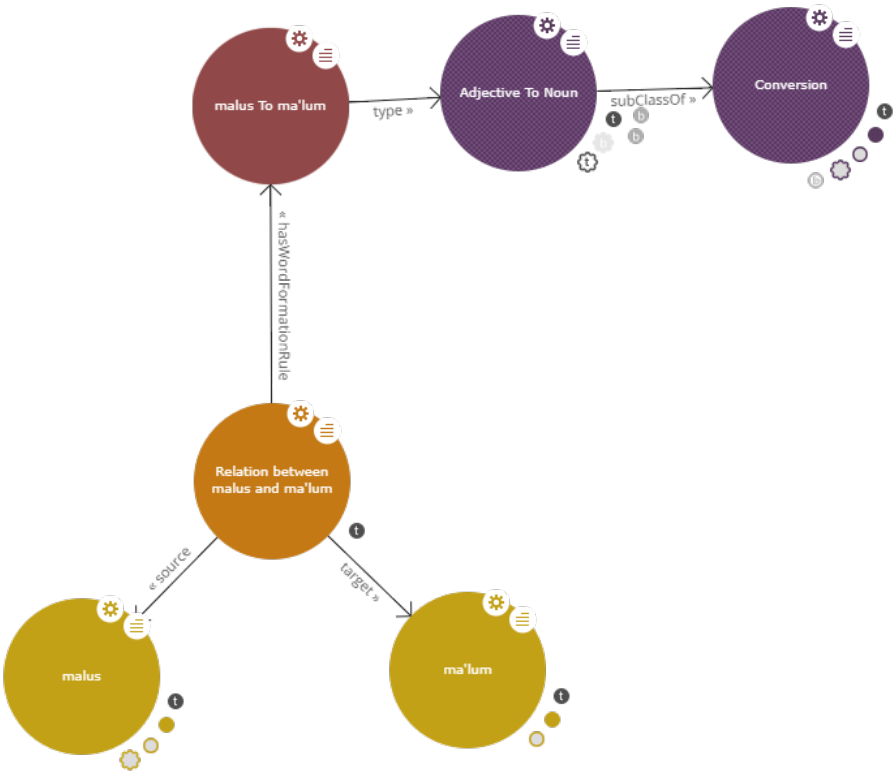


Figure 6. Modelling of conversion in the WFL ontology

single most frequent rule proves to be the one that derives (mostly action) nouns from verbs using the suffix labelled as -(*t*)*io*(*n*)- in WFL: there are 2,555 relations that instantiate this rule – among them, DUCTIO$_{N}$ 'leading' from DUCO$_{V}$ 'lead'.

The following step is to look for Medieval Latin lemmas that are potentially the output of one of these rules, based on their form. In cases of suffixation, this is done by first identifying Medieval Latin lemmas that end with a sequence of segments matching the suffix at hand, and then going through all lemmas – both Classical and Medieval Latin – and checking if they can be the input of the rule introducing that suffix, on the basis of the form of the potential input and output.

For instance, as for the above-mentioned action nouns ending in -(*t*)*io*(*n*)-, we can see that these are normally derived by taking the Third Stem (cf. Aronoff 1994) of the verb – in the example above, *duct-* – and adding the suffix *-ion-*, followed by the appropriate case endings – with the final nasal segment being regularly dropped in the nominative and vocative singular, thus yielding citation forms in *-io*. Our procedure consists in selecting all Medieval Latin lemmas whose citation form ends in *-io*, and keeping only the ones for which, among all verbs, there is at least one that can potentially be the base according to the regular formal procedure described above.

To give an example, among the Medieval Latin lemmas provided by Lemlat we find RETRUSIO$_{N}$ 'inclusion'; and among all lemmas from Lemlat (both Classical and Medieval), we find RETRUDO$_{V}$ 'thrust back'. The latter has *retrus-* as its Third Stem, as is shown by its perfect participle *retrusus*, that is listed in the Lemma Bank as a member of the class `Hypolemma`.[12] Hence, we identify a potential derivational relation between these two lemmas, as it is plausible that RETRUSIO$_{N}$ is an action noun formed from RETRUDO$_{V}$.

It should be noted that in some cases we also extract bases belonging to a part of speech different than the one that is required as input by the relevant rule. For instance, as potential inputs for adjectives in *-alis* we extract not only nouns, as in the most frequent case (e.g. ABYSSUS$_{N}$ 'bottomless pit' for ABYSSALIS$_{A}$ 'bottomless'), but also adjectives (e.g. ASSIDUUS$_{A}$ 'constantly present' for ASSIDUALIS$_{V}$ 'assiduous'). This is justified by the fact that adjectives are also attested – although much less frequently – as bases of adjectives in *-alis* already in Classical Latin (e.g. NOVALIS 'that is ploughed anew' from NOVUS 'new').

As for prefixation, we rely on a similar intuition, but follow a procedure that goes in the opposite direction: we start from all the lemmas, both Classical and Medieval, and look for the Medieval ones that, based on their form, can be analysed as being derived by adding one of the several Latin prefixes to it. For instance, for MELIORO$_{V}$ 'improve' we find three potential prefixed verbs that come Du Cange's glossary, namely

[12]This makes it possible to accomodate the different ways in which a participle like *retrusus* can be lemmatised, i.e. as an adjective (hypolemma RETRUSUS, `http://lila-erc.eu/data/id/hypolemma/38805`) or as the verb from which the participial form is created (lemma RETRUDO$_{V}$ `http://lila-erc.eu/data/id/lemma/122792`). For further details on the architecture of the LiLa KB and its use of hypolemmas, see Passarotti et al. (2020).

EMELIORO$_V$ 'improve/correct' (with prefix *e(x)-*), IMMELIORO$_V$ 'improve' (with prefix *in-*) and REMELIORO$_V$ 'correct' (with prefix *re-*). The application of this different procedure is partly due to practical reasons. The fact that inflection in Latin is mostly suffixal makes it difficult to predict the shape of the outcome of the application of a suffixation rule, as it is preliminarily necessary to strip the inflectional endings in order to identify the base to which the suffix should be added, and allomorphy plays a relevant role due to inflection class distinctions and other less systematic facts – for instance, the different stem variants of 3rd declension nouns, usually due to phonological adjustments triggered by the addition of particular endings (e.g. NOM.SG *dent-s* > *dens* 'tooth' vs. GEN.SG *dent-is*), but also simply suppletive stems (e.g. NOM.SG *iecur* 'liver' vs. GEN.SG *iecinoris*). On the left side of words, instead, there is much less allomorphy. As a consequence, it is much easier to predict the shape of the outcome of a prefixation rule, at least as long as we are dealing with category-preserving rules, for which we do not even need to guess the final part of the derived word, and can simply assume it to be the same as in the base. From a theoretical standpoint, this choice is also motivated by the fact that different preverbs in Latin, and in Indo-European languages in general (cf. Lehmann 1983, Booij and van Kemenade 2003) appear to have common characteristics that make it reasonable to treat them as a unique process – for instance, most of them have a basic spatial or temporal meaning, besides other metaphoric or idiosyncratic uses. On the other hand, suffixation rules differ markedly from each other in their morphological and semantic behaviour.

For the time being, we do not deal with compounding, as it would require a completely different treatment, which we leave for future work. Conversion is also currently out of the picture. By definition, in cases of conversion there is no formal marker of the derivation process except (possibly) for the different inflectional endings, hence it would not be possible to apply the procedure outlined above as it is. The only exception is the process of conversion that takes the Third Stem of a verb as input and produces a noun as output. The inclusion of such cases in our procedure is made possible by the fact that we can exploit LiLa's `Hypolemma` information (cf. Footnote 12 above) to extract lemmas that can plausibly be considered as potential bases of the conversion process, in a way similar to the one we have described above for suffixes that apply to the Third Stem like *-(t)io(n)-*. For instance, the Medieval Latin lemma DISPLICITUS$_N$ 'offence' can be presumed to be a conversion from the Third Stem *displicit-* of the verb DISPLICEO 'displease'. The reason for the inclusion of this conversion process is that among the rules that we consider there are also denominal nouns in *-atus*; but many of the nominal lemmas of LiLa that end in *-atus*, given their meaning, seem to be better analysed as conversions from the Third Stem of a verb rather than as obtained by adding the suffix *-at-* to a nominal or adjectival base. Hence, by extracting all the possible bases we are in a position to manually select the most appropriate one on a case-by-case basis (cf. Section 4.2 below).

Among the pairs of lemmas extracted following the procedure described above there are also false positives, either because the formal similarity between the two

lemmas is coincidental, or due to the noisy nature of the data we are dealing with (cfr. Section 4.4 below). Since we aim at having high-quality data, we also performed a manual validation of the base-derivative pairs that were extracted in this first phase, as detailed in the next section.

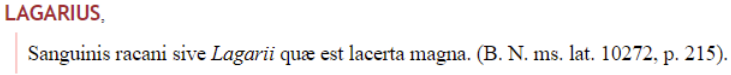
4.2. Issues in the annotation process

This section concerns issues related to the manual validation of the base-derivative pairs automatically extracted from the procedure described above. We discuss two issues in particular: in Section 4.2.1 we focus on the choice between two or more candidates that are automatically extracted as input for the same output; in Section 4.2.2, we deal with problematic entries of Du Cange's glossary.

4.2.1. Manual selection among two or more potential inputs for the same output lemma

First of all, alongside cases in which for a given output only one input was extracted, which were checked by looking at their semantics, in several cases the automatic selection detected two or more candidates as possible input lemmas. Therefore, we manually identified the most appropriate input lemma among all the ones that had been extracted. For instance, for the Medieval derivative noun COALITIO$_N$ 'assemblage, meeting', we automatically extracted two candidate inputs, COALO$_V$ 'feed, sustain, nourish' and COALESCO$_V$ 'grow together with something, unite', both of them coming from the Classical Latin section of Lemlat's database. This is motivated by the fact that both these verbs have *coalit-* as their Third Stem. In this case, COALESCO$_V$ is semantically closer to the derivative COALITIO$_N$ than COALO$_V$. Consequently, we selected the former as the input.

We also had to deal with several cases in which we automatically extracted two or more candidates with a similar meaning. These cases usually involved a Classical and a Medieval Latin lemma, as we can see in the noun COLLEGARIUS$_N$ ('one of the colleagues'). For this lemma, the automatic selection detected both COLLEGA$_N$ 'colleague' and COLLEGUS$_N$ 'companion' as possible inputs. The first one is a Classical Latin lemma, while the second one is attested only in Du Cange's glossary. Because of their semantic similarity, both candidates are potentially the input from which COLLEGARIUS$_N$ is formed. In such cases, we choose the Classical lemma, for practical reasons: as we will see in more detail in Section 4.3, when the input of the relation is a Classical Latin lemma that is present in WFL – as happens in this case – we are often in a position to attach the new relation to the corresponding tree – in this case, the one having LEGO$_V$ 'read' as its root. As a consequence, we can also infer information on the base and prefixes/affixes to which the corresponding lemma in LiLa should be linked. If we had selected the Medieval Latin lemma as input, on the other hand, we would have missed the relation of COLLEGARIUS$_N$ with the verb LEGO$_V$ and with the prefix *con-*, as no information on the derivational history of COLLEGUS$_N$ is provided in WFL and in the LiLa Lemma Bank.


LAGARIUS,

Sanguinis racani sive *Lagarii* quæ est lacerta magna. (B. N. ms. lat. 10272, p. 215).

Figure 7. Entry for lagarius *in Du Cange's glossary*

CONSTABILITOR, in Charta Roberti II. Principis Capuani apud Michaelem Monachum in Sanctuario Capuano pag. 643.

Figure 8. Entry for constabilitor *in Du Cange's glossary*

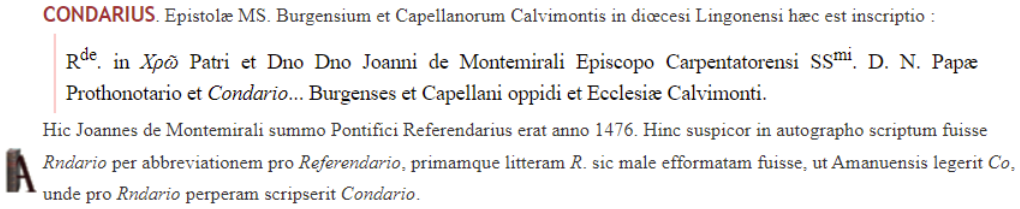
4.2.2. Issues related to the nature of Du Cange's entries

In addition to the previous issue, we also had to deal with some peculiarities of our source of Medieval Latin lemmas, which is a glossary, rather than a dictionary, and also has a complex publishing history (Géraud, 1839): for these reasons, there are many issues related to the descriptions or definitions of lemmas provided by glossators. More specifically, the glossary presents several cases of lemmas with quotes or bibliographic references, but without a definition, as we can see in the entries reported in Figures 7 and 8.

In both cases the glossator does not provide a definition, but just quotes the text where the lemma derives from (as we can see in 7) or provides the bibliographic reference where the lemma was found (as we can see in 8). Since we did not have an explanation of the meaning of the derivatives LAGARIUS$_{A}$ and CONSTABILITOR$_{N}$, we cannot be sure on their derivation history, therefore we decided not to include them in our validation process.

Moreover, there are philological issues taken into account by the glossator in the definition of the lemmas. An example is given in Figure 9 below.

In the entry, the glossator provides the source of the lemma and quotes the text where the lemma is taken from, without giving an explanation of the meaning. He actually proposes to amend the form *condario* to *Rndario* (hypothesizing an abbreviation for the word Referendario 'referendum'), identifying the former as an amanuensis' mistake. Therefore, it is not even sure that such a word ever existed in Latin: hence, also in such cases we decided not to consider relations involving those lemmas as valid.



CONDARIUS. Epistolæ MS. Burgensium et Capellanorum Calvimontis in diœcesi Lingonensi hæc est inscriptio :

> R^de^. in *Xpõ* Patri et Dno Dno Joanni de Montemirali Episcopo Carpentatorensi SS^mi^. D. N. Papæ Prothonotario et *Condario*... Burgenses et Capellani oppidi et Ecclesiæ Calvimonti.

Hic Joannes de Montemirali summo Pontifici Referendarius erat anno 1476. Hinc suspicor in autographo scriptum fuisse *Rndario* per abbreviationem pro *Referendario*, primamque litteram *R*. sic male efformatam fuisse, ut Amanuensis legerit *Co*, unde pro *Rndario* perperam scripserit *Condario*.

Figure 9. Entry for condarius *in Du Cange's glossary*

4.3. Organizing the information

As an outcome of the procedure described above in Sections 4.1 and 4.2, we obtain data in the form of a series of derivational relations between an input and output lemma. To release this information, first of all we need to include it into the relational database that contains the information displayed by WFL. One of its tables reports a list of relations, with fields for the identifier of the WFR at play and input and output lemmas. After adding our new relations to this table, we can release WFL 2.0, a new version of WFL that also covers Medieval Latin.[13] This is the version of WFL that is modelled as an ontology as described in Section 3 above and consequently included into the LiLa KB.[14] Its data can also be accessed with the same tree-based graphical interface used for the previous version of WFL.[15]

Furthermore, the morphological analysis of a wordform performed by Lemlat exploits this knowledge to provide the user with derivational information on the lemma that is assigned to the analysed wordform, if such information is available. Specifically, the analysis outputs i) the lemma (or lemmas, in cases of compounding) from which the wordform's lemma is derived, ii) the derivational process at play, iii) its type and iv) the affix (prefix or suffix) involved in this process, as illustrated in Figure 10 below. Hence, thanks to the work described in the previous sections, such derivational information will be provided in Lemlat's analyses also for some Medieval Latin lemmas, and not only for Classical Latin lemmas.

A caveat needs to be made on the structure of the data of WFL 2.0. While the output of the newly added relations is always a Medieval Latin lemma, the input can be either a Classical Latin or a Medieval Latin lemma. In the latter case, it is not possible to attach it to any of the trees of the previous version of WFL, as only Classical Latin is

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

[14] http://lila-erc.eu/data/lexicalResources/WFL/Lexicon.

[15] https://wfl.marginalia.it/

```
SEGMENTATION:  felicitat -em
---------- morphological feats ----
--afs--

Case:   Accusative
Gender: Feminine
Number: Singular
        =================LEMMA ================
        felicitas         N3B f0322 f
        --------------morphological feats-------
        NcC

        PoS:   Noun
        Type:  Common
        Inflexional Category:  III decl
        --------------derivational info--------
        IS DERIVED: YES
        --------------rule id: 48--------------
        Lexical Basis:
                felix           N3A f0327 * Af-
        Derivational Type: Derivation_Suffix
        Derivational Category: A-To-N
        Affix: tas/tat
```

Figure 10. Morphological analysis of a wordform by Lemlat

represented there: hence, we end up with unconnected branches.[16] In some cases, this reflects the fact that both input and output are Medieval Latin innovations, that cannot be related to any Classical Latin lemma: for instance, the Medieval Latin lemma ACUNYDO$_V$ 'be suspicious', that we identify as the input of the derivational process forming ACUNYDAMENTUM$_N$ 'suspiciousness', cannot be linked to any Classical Latin lemma – it is tagged as *vox catalonica* in Du Cange's glossary. In other cases, it would be possible to attach these unconnected branches to an existing WFL tree, by identifying the intermediate step(s) in the derivational history. For instance, we find a derivational relation between the two medieval Latin lemmas VICINO$_V$ 'be close' and ABVICINO$_V$ 'separate'. The base VICINO$_V$ could be connected to a WFL tree by establishing a relation between it and the Classical Latin adjective VICINUS$_A$ 'close' as input – relation that is not identified by our automatic procedure because it is a conversion process. Hence, these missing links will eventually be found when progressing on the coverage of Medieval Latin lemmas.

In addition to releasing this new version of WFL and enhancing the informativity of Lemlat's analyses, we have also decided to release our data in the format required by the flat structure of the Lemma Bank, using our new relations to infer new triples

[16] Unless we have identified also a relation where the same lemma appears as output in our procedure, and whose input is a Classical Latin lemma for which we do have a tree in WFL.

Lemma	hasBase	hasPrefix	hasSuffix
DEFECTIVUS$_A$	Base of FACIO$_V$	*de-*	*-(t)iu-*
INDEFECTIVUS$_A$	**Base of FACIO$_V$**	***de-, in-***	***-(t)iu-***

Table 2. New derivational information in the Lemma Bank: transfer of base, prefix and suffix from input to output lemma

lemma	hasBase	hasPrefix	hasSuffix
ACUNYDO$_V$	**Base of ACUNYDO$_V$**		
ACUNYDAMENTUM$_N$	**Base of ACUNYDO$_V$**		***-ment-***

Table 3. New derivational information in the Lemma Bank: generation of new bases

that connect lemmas to their base, prefix and/or suffix through the dedicated properties `hasBase`, `hasPrefix`, `hasSuffix`. When the input of the relation is a Classical Latin lemma, we often already have information on its base and (possibly) on affixes displayed by it. In such cases, we can simply transfer this information also to the output, that inherits the base and (possibly) affixes of its input, and additionally displays the prefix or suffix involved in the rule that relates the two (except for cases of conversion). For instance, we find a relation between the Classical Latin lemma DEFECTIVUS$_A$ 'missing' and the Medieval Latin lemma INDEFECTIVUS$_A$ 'not missing'. In the Lemma Bank, the former is related to the base of FACIO$_V$ 'make', to the prefix *de-*, and to the suffix *-(t)iu-*. Hence, also the latter will inherit all these connections, and it will be additionally related to the prefix *in-* that is involved in the WFR that relates the two lemmas, as illustrated in Table 2.[17]

On the other hand, when the input is a Medieval Latin lemma, we do not have derivational information on it. Therefore, all we can do is expressing the fact that the input and output lemmas are part of a same family, by generating a new base and linking both of them to it. For instance, the input of the relation between ACUNYDO$_V$ 'be suspicious' and ACUNYDAMENTUM$_N$ 'suspiciousness' is a Medieval Latin lemma itself, hence it is not connected to a base, prefix nor suffix. Therefore, we generate a new base – that we label with the label of the input – and connect both ACUNYDO$_V$ and ACUNYDAMENTUM$_N$ to it. Furthermore, the output is connected to suffix *-mentum* involved in the rule relating it to the input (see Table 3).

[17]In Tables 2 and 3, we distinguish the pieces of information that are already in the Lemma Bank from the ones that we add in this phase by highlighting the latter in bold.

Note that the same treatment is applied when there is no derivational information on the input even if it is a Classical Latin lemma. This happens when the lemma is the only member of its family, and hence it is not derivationally related to any other lemma.

4.4. A quantitative evaluation

In this section, we offer a quantitative evaluation of the different types of derivational information that have been obtained. Table 4 summarizes the rules of WFL that we select as the ones for which we look for new potential derivatives in Medieval Latin lemmas, sorted by descending frequency (i.e., number of relations in WFL). For each rule, we also report how many new potential relations were extracted automatically.[18] For suffixal rules, we report in how many cases more than one input candidate was extracted following the procedure described in Section 4.1, and hence a choice has to be made on which of them is the most appropriate.[19] Lastly, we report how many relations are kept after the manual validation described in Section 4.2. This allows us to evaluate the precision of the automatic procudure to identify candidate pairs. It can be observed that there is a lot of variation across rules regarding the proportion of relations that are considered to be valid after the manual checking among all the relations that are extracted automatically: the percentage ranges from 95.24% for adjectives in *(-t)iu-* to as low as 34.42% for diminutives in *-ul-*, with an average of about 63%. In some cases, the decision to exclude a relation that was extracted automatically is ultimately due to the remarkable amount of noise in the source of Medieval Latin lemmas: we have already seen in Section 4.2 that there are words whose meaning is doubtful, or even entries that refer to forms that are plausibly a copyist mistake, or for which there are philological issues of other kinds. However, the main reason behind this quite low precision is a principled choice in the design of the rules used to extract input and output candidates. These are intended to capture as many potential pairs as possible, including cases where the formal relation between input and output lemmas is less than fully regular, at the cost of a higher number of false positives. Indeed, the phase of manual validation is intended to identify exactly these false positives, that consequently do not affect the final quality of the data. On the other hand, if the rules had been designed to be more restrictive, precision would have been increased at

[18] We do not provide this information for diminutive nouns in *-(us/un)cul-* because in that case we did not extracted candidates automatically. They are included because diminutives that end in *-(us/un)culus/a/um* also end in *-ulus/a/um*, hence they are sometimes identified as the output of the latter rule, and manually corrected when they are actually the output of the former, like in the case of BELLICULUM$_{N}$ 'simulated war', that is a diminutive of *bellum*$_{n}$ 'war' with suffix *-cul-* rather than a diminutive of *bellicum*$_{n}$ 'war signal' with suffix *-ul-*.

[19] For prefixal rules, this never happens because we start from the potential input and look for potential outputs, rather than the reverse: hence, the corresponding cell is left empty.

process	n. relations in WFL 1.0	n. potential new pairs	n. outputs with >1 possible input (%)	n. valid pairs (%)
V-To-V prefixation	4,850	2,194	–	1,129 (51.46%)
V-To-N -(t)io(n)- suffixation	2,555	458	41 (8.95%)	423 (92.36%)
V-To-N -(t)or- suffixation	1,419	382	44 (11.52%)	321 (84.03%)
V-To-N conversion	1,074	210	33 (15.71%)	140 (66.67%)
A-To-N -tas/tat- suffixation	623	225	27 (12.00%)	192 (85.33%)
N-To-A -os- suffixation	563	203	115 (56.37%)	152 (75.00%)
N-To-A -al- suffixation	547	176	100 (56.82%)	126 (71.59%)
A-To-A in- prefixation	508	101	–	64 (63.37%)
N-To-A -ari- suffixation	467	586	315 (53.75%)	396 (67.58%)
N-To-N -ari- suffixation	452	596	355 (59.56%)	387 (64.93%)
N-To-N -ul- suffixation	427	491	299 (60.90%)	169 (34.42%)
V-To-N -(t)ric- suffixation	415	33	6 (18.18%)	25 (75.76%)
N-To-A -at- suffixation	404	342	162 (47.37%)	198 (57.89%)
V-To-A -bil- suffixation	390	145	27 (18.62%)	117 (80.69%)
N-To-N -(us/un)cul- suffixation	370	–	–	21
V-To-V -(i)t- suffixation	343	89	40 (44.94%)	46 (51.69%)
N-To-A -ic- suffixation	339	126	75 (59.52%)	67 (53.17%)
N-To-A -in- suffixation	307	86	51 (59.30%)	43 (50.00%)
V-To-A -(t)iu- suffixation	289	63	18 (28.57%)	60 (95.24%)
V-To-N -ment- suffixation	277	380	95 (25.00%)	304 (80.00%)
N-To-A -e- suffixation	242	56	32 (57.14%)	31 (55.36%)
V-To-I -(at)im- suffixation	203	95	39 (41.05%)	58 (61.05%)
V-To-V -sc- suffixation	199	34	8 (23.53%)	18 (52.94%)
N-To-N -at- suffixation	85	192	135 (70.31%)	71 (36.98%)
TOTAL	–	7,263	–	4,558 (62.76%)

Table 4. Automatically extracted and manually validated relations for Medieval Latin lemmas

number of relations	Classical Latin	34,960
	Medieval Latin	4,558
n. lemmas in Lemlat's database	Classical Latin	43,407
	Medieval Latin	86,745

Table 5. Coverage of Classical and Medieval Latin lemmas in WFL 2.0

the expense of recall, but then there would have been no way to recover the marginal cases that were left out.

Another consequence of this choice is the remarkable proportion of cases for which more than one potential input is found by the automatic procedure: on average, this happens for about 40% of the output candidates, and for some rules this is the case for the majority of them, as shown in the third column of Table 4.

In Table 5, we provide data on the coverage of Classical Latin and Medieval Latin lemmas in WFL 2.0, by giving the number of relations that have a Classical Latin lemma as output (i.e. the ones of WFL 1.0) vs. the ones that have a Medieval Latin lemmas as output (i.e. the ones that are added to WFL 2.0). We also give the number of Classical and Medieval Latin lemmas in Lemlat's database for reference.

Unsurprisingly, the coverage of Medieval Latin is much lower: while all Classical Latin lemmas have been taken into account, for Medieval Latin we have focused only on the most frequent processes reported in Table 4 above. Therefore, the coverage of Medieval Latin can still increase when other relations will be added (see Section 5 below).

For what an evaluation of the derivational information added for Medieval Latin lemmas in the Lemma Bank is concerned, Table 6 gives the number of new triples that connect lemmas to their base, prefix, and/or suffix through the respective dedicated properties. As for the property `hasBase`, we also report how many of the new triples have a Classical Latin lemma as subject. This happens when a relation is established whose input is a Classical Latin lemma that is not related to any other Classical Latin. Since a LiLa's `Base` is nothing but an abstract connector between lemmas of the same family, such lemmas have no base as long as only Classical Latin is considered; if a relation is found with a Medieval Latin lemma, a new `Base` has to be established, as we have seen above in Section 4.3. Table 6 also reports the number of new `Bases` introduced into the Lemma Bank.

5. Conclusions

In this paper, we have described the work that was conducted to extend the derivational information available in the LiLa KB (Section 2) in two directions: structurally, by making also the hierarchical organization of WFL data available into the KB, along-

new triples with `hasBase` property	Classical Latin	96
	Medieval Latin	5,696
new triples with `hasPrefix` property	Classical Latin	–
	Medieval Latin	2,043
new triples with `hasSuffix` property	Classical Latin	–
	Medieval Latin	4,156
new `Bases`		1,143

Table 6. New derivational information in the Lemma Bank

side the flat organization of those same data in the Lemma Bank (Section 3); and diachronically, by extending WFL to cover also Medieval Latin lemmas, and consequently providing information on those lemmas also in the Lemma Bank (Section 4).

In Section 2, we have hinted at the reasons behind the choice of adopting a paradigmatic approach to word formation in the LiLa Lemma Bank – thus yielding a flat structure of related lexemes belonging to the same family. However, there are cases where the more detailed, hierarchical information provided by WFL on the order of application of different word formation processes can prove helpful. For instance, an advantage of the hierarchical structure of WFL is that it allows to focus on smaller, more tightly connected sub-sections of word formation families. This can be helpful especially when dealing with very large and quite heterogeneous families, e.g. the one of the verb FACIO 'to make', which includes 689 lemmas in the Lemma Bank. Since the semantic connection between some members of this family is quite loose, it might be useful to be able to zoom on smaller sub-families with a higher degree of internal semantic cohesion, isolating e.g. only those lexemes that are directly related to the adjective DIFFICILIS 'difficult' (e.g. PERDIFFICILIS and SUBDIFFICILIS 'very/somewhat difficult'), or only the verbs formed by adding a prefix to FACIO itself (e.g. INFICIO 'to put into' and PERFICIO 'to achieve'[20]). Such a focus on sub-families cannot be performed with the representation of word formation in the Lemma Bank, where all lemmas belonging to the same word formation family are simply connected to their common base without any further information about the hierarchy of derivations, whereas in WFL each derived lexeme is directly linked to its source lexeme.

In other cases, however, the flat organization of derivational information in the Lemma Bank can prove helpful. As an example, when considering prefixed and suffixed words, for some purposes it can be useful to focus only on those words that are actually formed by means of a WFR that involves a specific affix, while for other pur-

[20]The different shape of the stem in the base vs. derivative is due to a phonological process of weakening of short vowels in non-initial syllables.

poses it might be better to collect all those words that display that affix somewhere along their word formation history. Consider for instance the structural difference between the adjectives INFRUCTUOSUS 'unfruitful' and INIURIOSUS 'injurious': the former is created by prefixing *in-* (negation) to FRUCTUOSUS 'fruitful' (*INFRUCTUS is not attested as a Latin word), while the latter is formed by suffixing *-os* to INIURIA 'injury' (*IURIOSUS). Therefore, when investigating e.g. *in-* prefixation, it is a matter of choice whether to include also cases like *iniuriosus*. If we want to exclude them, this has to be done using the hierarchical information of WFL. Conversely, however, if we decide to include such cases, then the relevant information can be obtained by exploiting the flat structure of the Lemma Bank, where all lemmas are linked to all the prefixes and suffixes they display, regardless of their order of application in the word formation history. Although, in this specific case, it would be possible to construct a query that goes down one step in the hierarchy of WFL, things would be even more difficult in cases featuring more than two affixes – consider for instance a word like the adverb INADDUCIBILITER 'unobstructively' (lit. 'not in a way that can be pulled back and forth'), with prefixes *in-* (negation) and *ad-* and suffixes *-bil-* and *-ter*.

One of the main advantages of adopting Linked Data principles and models to represent and publish linguistic information provided by distributed resources is that this makes it possible to represent different approaches within a unified framework, as it is clearly shown in Figure 4. Scholars can choose the approach that is more compatible with their theoretical view, or simply the one that provides the kind of information more appropriate for the case at hand, also allowing to make different approaches interact easily, in case several pieces of information from different sources are needed.

This motivates our structural decision of modelling WFL data into an ontology and linking them to the LiLa KB. As for the diachronic extension documented in this paper, increasing the coverage of our derivational resources to Medieval Latin represents a first step toward filling a gap that is widespread in digital lexical resources for Latin, which tend to focus on the Classical and Late periods. However, it emerges from the data of Table 5 that a full coverage of Medieval Latin lemmas is still far. Hence, it is necessary to explore some possibilities for future work in this direction.

The most obvious thing to do would be to carry on with the same procedure, extracting pairs of lemmas that are potentially related by means of other derivational processes. However, it can be observed that the number of potential pairs is already quite low for some of the more frequent processes already considered and listed in Table 4 above. Therefore, moving on to processes that are even more marginal, the number of additional pairs is likely to progress quite slowly.

Another possibility that might be explored is a machine learning approach to the task, following the hint of recent work such as Lango et al. (2021), Svoboda and Ševčíková (2022) and Kyjánek et al. (2022). In our case, it could be interesting to try using WFL 1.0 as a gold standard to train a machine learning algorithm whose task is to identify derivational relations within a lexicon, and then applying this algorithm to Medieval Latin lemmas for which derivational information is still lacking. How-

ever, it should be highlighted that the data provided in WFL and in the LiLa KB are meant to be used as a reference for philological and linguistic work, rather than for massive NLP on big data. The accuracy of the fully automatic methods proposed in other studies, although variable,[21] do not in any case reach the close-to-perfect value that would be needed for these purposes. Hence, a phase of *ex post* manual check – possibly enhanced with semi-automatic means – would probably be necessary in any case.

Acknowledgments

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

Aronoff, Mark. *Morphology by Itself: Stems and Inflectional Classes*. MIT Press, Cambridge, MA, 1994.

Booij, Geert. Construction morphology. *Language and linguistics compass*, 4(7):543–555, 2010. doi: 10.1017/9781139814720.016.

Booij, Geert and Ans van Kemenade. Preverbs: an introduction. In Booij, Geert and Jaap van Marle, editors, *Yearbook of morphology 2003*, pages 1–11. Kluwer, Dordrecht, 2003. doi: 10.1007/978-1-4020-1513-7_1.

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

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, Switzerland, 2017. Springer. ISBN 978-3-319-59888-8. doi: 10.1007/978-3-319-59888-8_6.

Chiarcos, Christian and Maria Sukhareva. OLiA – Ontologies of Linguistic Annotation. *Semantic Web*, 6(4):379–386, 2015. doi: 10.3233/SW-140167.

Chiarcos, Christian, Katerina Gkirtzou, Fahad Khan, Penny Labropoulou, Marco Passarotti, and Matteo Pellegrini. Computational Morphology with OntoLex-Morph. In *8th Workshop on Linked Data in Linguistics*, pages 78–86. European Language Resources Association (ELRA), 2022.

Cimiano, Philipp, Christian Chiarcos, John McCrae, and Jorge Gracia. *Linguistic Linked Data*. Springer, 2020. doi: 10.1007/978-3-030-30225-2.

du Cange, Charles du Fresne sieur, bénédictins de la congrégation de Saint-Maur, d. Pierre Carpentier, Johann Christoph Adelung, G. A. Louis Henschel, Lorenz Diefenbach, and Léopold Favre. *Glossarium mediae et infimae latinitatis*. Favre, Niort, France, 1883–1887.

[21] The reader is referred to the publications cited above for more details.

Forcellini, Egidio. *Lexicon totius latinitatis*. Arnaldo Forni, Bologna, Italy, 1965.

Georges, Karl Ernst. *Ausführliches lateinisch-deutsches Handwörterbuch*. Wissenschaftliche Buchgesellschaft, Darmstadt, Germany, 1998. URL `http://www.zeno.org/Georges-1913`. Reprint of first edition of 1913–1918, Hannover, Germany: Hahnsche Buchhandlung.

Géraud, Hercule. Historique du Glossaire de la basse latinité de Du Cange. *Bibliothèque de l'École des chartes*, pages 498–510, 1839. doi: 10.3406/bec.1840.461649.

Glare, Peter G. W. *Oxford Latin Dictionary*. Oxford Languages. Oxford University Press, Oxford, UK, 2012. ISBN 978-0-19-958031-6.

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

Hellmann, Sebastian, Jens Lehmann, Sören Auer, and Martin Brümmer. Integrating NLP using Linked Data. In *Proc. 12th International Semantic Web Conference, 21-25 October 2013*, Sydney, Australia, 2013. doi: 10.1007/978-3-642-41338-4_7.

Klimek, Bettina, John McCrae, Julia Bosque-Gil, Maxim Ionov, James K. Tauber, and Christian Chiarcos. Challenges for the Representation of Morphology in Ontology Lexicons. In *Proceedings of eLex*, pages 570–591, 2019.

Kyjánek, Lukáš. Harmonisation of Language Resources for Word-Formation of Multiple Languages. Master's thesis, Univerzita Karlova, Matematicko-fyzikální fakulta, 2020.

Kyjánek, Lukáš, Olga Lyashevskaya, Anna Nedoluzhko, Daniil Vodolazsky, and Zdeněk Žabokrtskỳ. Constructing a Lexical Resource of Russian Derivational Morphology. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 2788–2797, 2022.

Lango, Mateusz, Zdeněk Žabokrtskỳ, and Magda Ševčíková. Semi-automatic construction of word-formation networks. *Language Resources and Evaluation*, 55(1):3–32, 2021. doi: 10.1007/s10579-019-09484-2.

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

Lehmann, Christian. Latin preverbs and cases. In Pinkster, Harm, editor, *Latin linguistics and linguistic theory: Proceedings of the 1st International Colloquium on Latin Linguistics*, pages 145–161. John Benjamins, Amsterdam, 1983. doi: 10.1075/slcs.12.15leh.

Litta, Eleonora and Marco Budassi. What we talk about when we talk about paradigms: representing Latin word formation. In *Paradigmatic relations in word formation*, pages 128–163. Brill, 2020.

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, *Lemmata Linguistica Latina. Volume 1. Words and Sounds*, pages 224–239. De Gruyter, Berlin, Boston, December 2019. ISBN 978-3-11-064758-7. doi: 10.1515/9783110647587-015.

Litta, Eleonora, Marco Passarotti, and Francesco Mambrini. Derivations and Connections: Word Formation in the LiLa Knowledge Base of Linguistic Resources for Latin. *The Prague Bulletin of Mathematical Linguistics*, (115):163–186, 2020. doi: 10.14712/00326585.010.

McCrae, John, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. The OntoLex-Lemon Model: Development and Applications. In *Proceedings of eLex*, pages 587–597, 2017.

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, 2017.

Passarotti, Marco, Francesco Mambrini, Greta Franzini, Flavio Massimiliano Cecchini, Eleonora Litta, Giovanni Moretti, Paolo Ruffolo, and Rachele Sprugnoli. Interlinking through Lemmas. The Lexical Collection of the LiLa Knowledge Base of Linguistic Resources for Latin. *Studi e Saggi Linguistici*, LVIII(1):177–212, 2020.

Svoboda, Emil and Magda Ševčíková. Word Formation Analyzer for Czech: Automatic Parent Retrieval and Classification of Word Formation Processes. *The Prague Bulletin of Mathematical Linguistics*, (118):55–73, 2022. doi: 10.14712/00326585.019.

Address for correspondence:
Matteo Pellegrini
matteo.pellegrini@unicatt.it
CIRCSE Research Centre, Università Cattolica del Sacro Cuore
Largo Agostino Gemelli 1, 20123 Milano, Italy