The Two Approaches to Word Formation in the LiLa Knowledge Base of Latin Resources

Matteo Pellegrini Eleonora Litta Marco Passarotti Francesco Mambrini Giovanni Moretti
CIRCSE Research Centre
Università Cattolica del Sacro Cuore
Largo Gemelli, 1 - 20123 Milan, Italy
{matteo.pellegrini}{eleonoramaria.litta}{marco.passarotti}
{francesco.mambrini}{giovanni.moretti}@unicatt.it

Abstract

In this paper, we propose a model to include a derivational lexicon for Latin (Word Formation Latin) within the LiLa Knowledge Base of interlinked linguistic resources for Latin. After a brief introduction on the architecture of LiLa, we discuss the differences between the flat organization of derivational information in LiLa's Lemma Bank and the hierarchical structure of Word Formation Latin, showing that the latter contains potentially useful information that is not already available in the former. We describe the modelling of such information in LiLa, exemplifying how different word formation processes are treated. We conclude the paper by showing the complementarity of the two approaches, and outlining the advantages offered by their interconnection.

1 Background and Motivation

In recent years, the principles of the so-called Linked Data paradigm[1] are increasingly being applied to language data and metadata, aiming to improve interoperability between resources originally developed for different purposes, hence characterised by different formalisms and conceptual models. As a consequence, a Linguistic Linked Data Cloud is being developed, to which several resources are continuously being added (Cimiano et al., 2020). Within this framework, the aim of the *LiLa* project[2] 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 like 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). As a consequence, 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. More specifically, the backbone of the architecture of the LiLa KB is the Lemma Bank, a large collection of lemmas – i.e. citation forms – to which both the tokens of textual resources and the entries of lexical resources can be connected, as well as the output of NLP tools. The Lemma Bank initially included a limited amount of derivational information on lemmas from the Word Formation Latin (WFL) lexical resource (Litta and Passarotti, 2019). A choice was made not to include the entire information provided by WFL, that, however, might prove useful in certain circumstances.

In this contribution, we describe a model designed to include all the information contained in WFL in the LiLa KB. In Section 2, we detail the architecture of the KB on the one hand and of WFL on the other hand. In Section 3, we describe the model that we propose in order to include WFL within the architecture of LiLa, showing how different word-formation processes are treated. Also, this section describes how our work interacts with other models developed by the Linked Data community – namely, the LexInfo ontology of data categories (Cimiano et al., 2011), the OntoLex-Lemon vocabulary for describing lexical

---

[1] https://www.w3.org/DesignIssues/LinkedData.html.
[2] https://lila-erc.eu.

resources (McCrae et al., 2017; Buitelaar et al., 2011) and, more specifically, its Morphology Module (Klimek et al., 2019). We conclude in Section 4 by reviewing the dissimilarities between the modelling of the original derivational information in the LiLa Lemma Bank and the one of the WFL resource linked to the KB, showing how the application of Linked Data principles and techniques can benefit the communication between diverse linguistic resources.

2   LiLa and Word Formation Latin

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[3], 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 130,000 Latin lemmas taken from the database of the morphological analyser Lemlat (Passarotti et al., 2017). 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, on its part, is a derivational lexicon of Latin, characterised by a step-to-step morphotactic approach: 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. More specifically, 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. In turn, within derivation rules, affixation (more specifically, prefixation and suffixation) and conversion are distinguished, depending on the presence of an affix and its nature. 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.

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

Table 1: Examples of Word Formation Rules in WFL.

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 – and branching out to all derivatives by means of the successive application of individual WFR. For example, Figure 1 shows a portion of the family taking root from the ancestor lexeme FELIX '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 currently includes only a selection of the derivational information contained in WFL. Besides `Lemmas`, two other classes are involved, namely `Affixes` – in their turn divided into `Prefixes` and `Suffixes` – and `Bases`, merely defined as abstract connectors between lemmas that belong to the same family. 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`.[4] As a consequence, 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 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.

---

[3]https://lila-erc.eu/lodview/ontologies/lila/.
[4]These properties are all defined in LiLa's ontology.
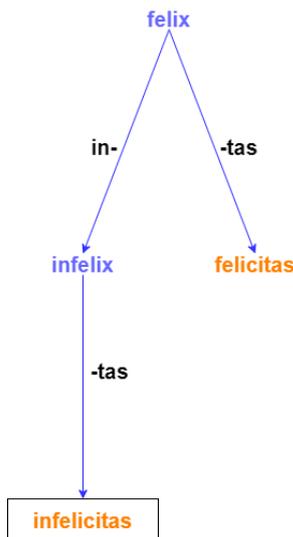
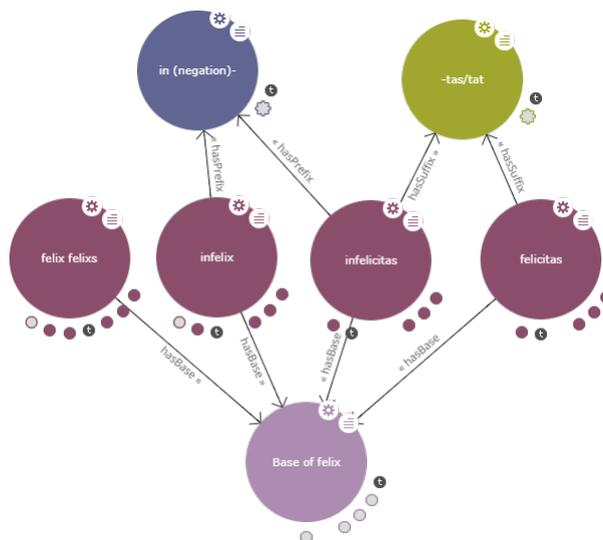Figure 1: Word Formation in WFL.

Figure 2: Word Formation in the Lemma Bank.

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.[5]

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, like Construction Morphology (Booij, 2010). Furthermore, such an approach allows for a more natural treatment of cases that were 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 included into the LiLa KB.

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

---

[5]Kyjánek (2020)'s 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.

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`),[6] 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 take the class `PartOfSpeech` from LexInfo (see again Section 1 for references), an ontology created to provide data categories for the OntoLex model, and we also 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.
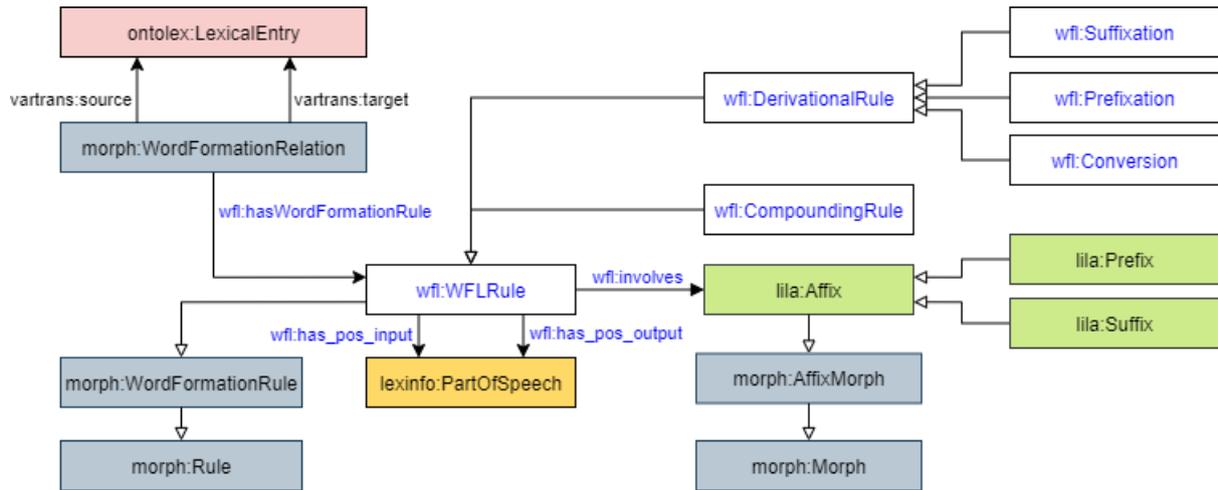

Figure 3: Architecture of the WFL ontology.

Let us now delve into some detail on the architecture of our model. We have 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`. The class `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 organi-zation of WFL.[7] For the same reason, rules are distinguished according to the lexical categories of the source and derivative, by providing a link to the `PartOfSpeech` of LexInfo through the properties `wfl:has_pos_input` and `wfl:has_pos_output`. Lastly, a 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`)[8] 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 Lila and the ones of `morph` (`morph:AffixMorph`) on the other hand.

To show the model at work with specific pairs of related words, Figure 4 shows the Linked Data treatment of the derivation of INFELIX 'unhappy' from FELIX 'happy' on the one hand (left side of the

---

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

[7]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.

[8]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.

image), of ɪɴꜰᴇʟɪᴄɪᴛᴀꜱ 'unhappiness' from ɪɴꜰᴇʟɪx 'unhappy' on the other hand (right side of the image).
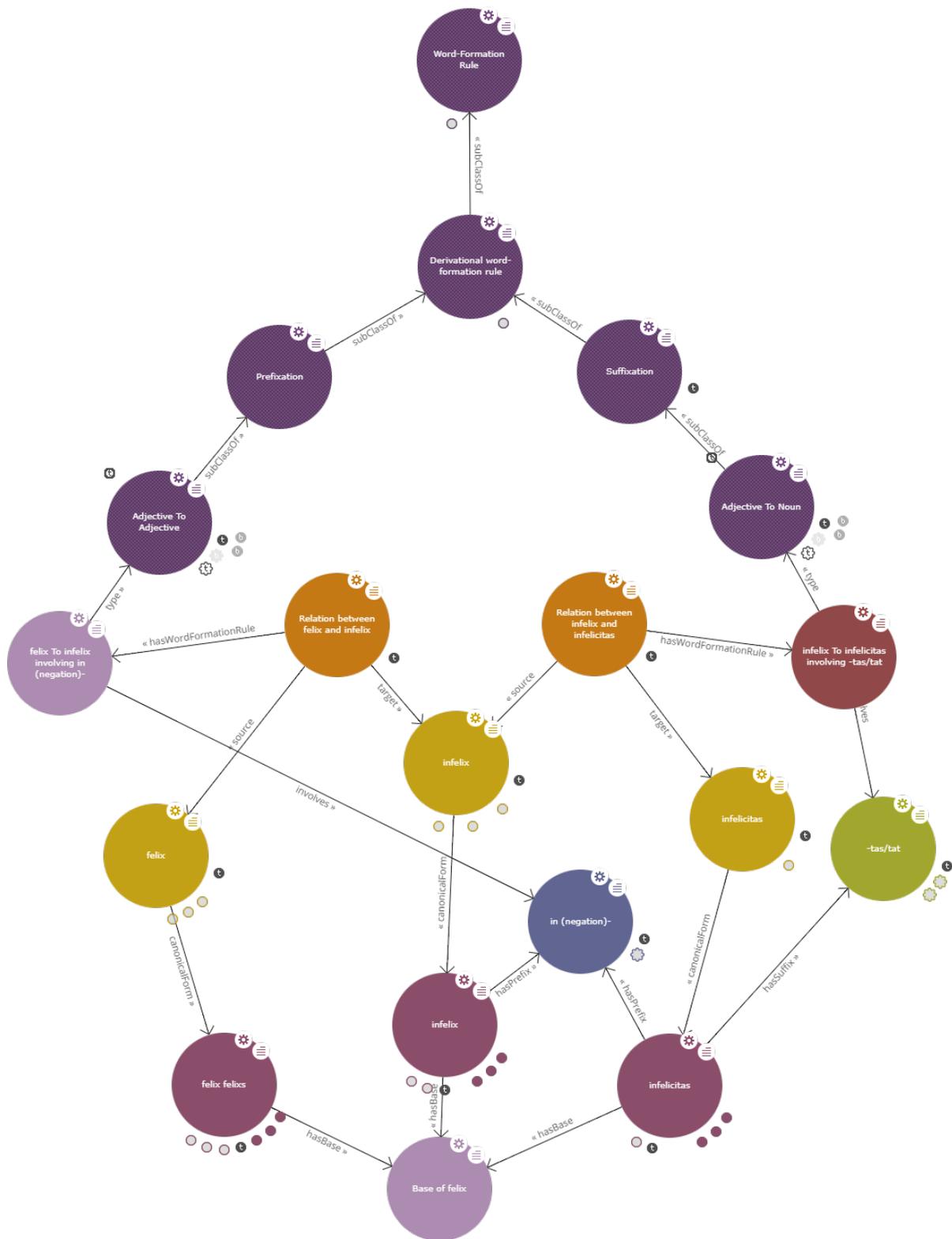

Figure 4: Modelling of prefixation and suffixation in the WFL ontology.

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 ꜰᴇʟɪx and ɪɴꜰᴇʟɪx and one

between INFELIX and INFELICITAS. Each relation is instantiated by a specific WFR: see the nodes labelled as "felix To infelix involving in (negation)-"[9] and "infelix To infelicitas involving -tas/tat",[10] respectively. Starting from the one that forms INFELIX from FELIX, 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 from INFELIX, 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 'farmer' (from AGER 'field' + COLO '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 the value of this property is 1 for the relation between AGER and AGRICOLA, 2 for the relation between COLO and AGRICOLA.
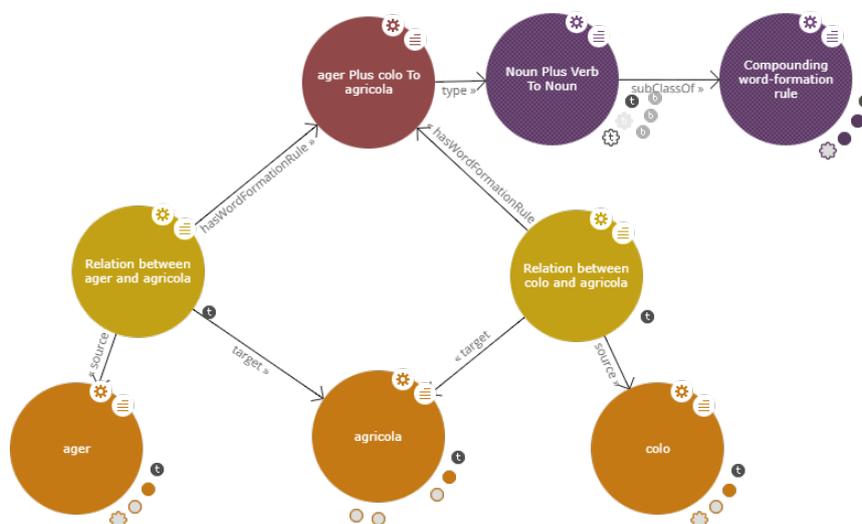

Figure 5: Modelling of compounding in the WFL ontology.

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.

---

[9]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'.

[10]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*.
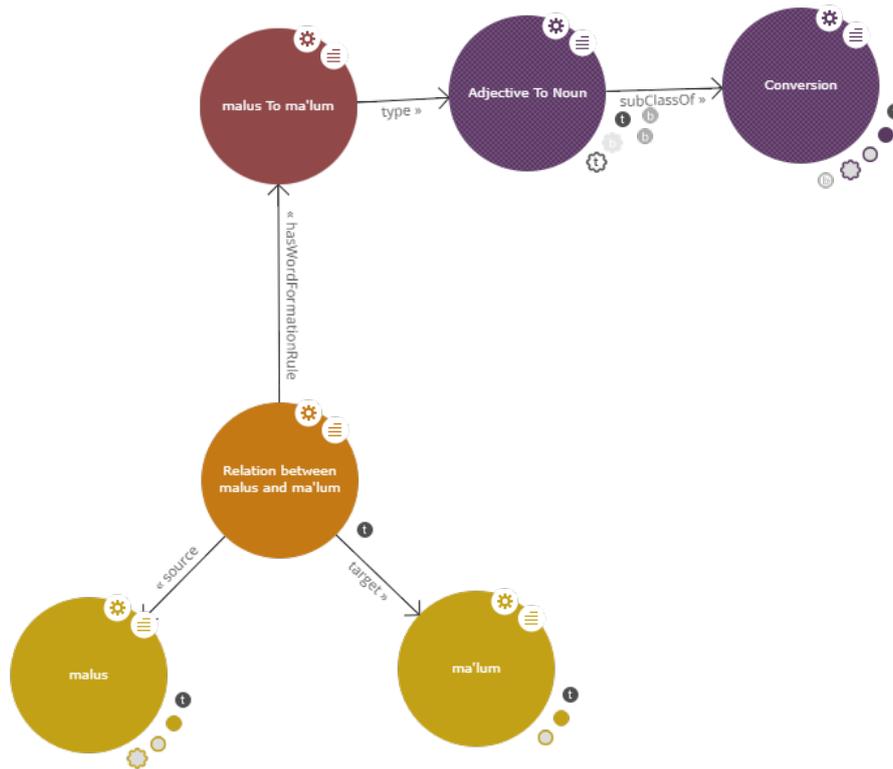
Figure 6: Modelling of conversion in the WFL ontology.

4 Discussion and Conclusion

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'[11]). 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 purposes 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

---

[11]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.

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.

Acknowledgments

References

Geert Booij. 2010. Construction morphology. *Language and linguistics compass* 4(7):543–555.

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

Christian Chiarcos and Christian Fäth. 2017. CoNLL-RDF: Linked Corpora Done in an NLP-Friendly Way. In Jorge Gracia, Francis Bond, John P. McCrae, Paul Buitelaar, Christian Chiarcos, and Sebastian Hellmann, editors, *Language, Data, and Knowledge*. Springer, Cham, Switzerland, pages 74–88.

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

Philipp Cimiano, Paul Buitelaar, John McCrae, and Michael Sintek. 2011. LexInfo: A Declarative Model for the Lexicon-Ontology Interface. *Journal of Web Semantics* 9(1):29–51.

Philipp Cimiano, Christian Chiarcos, John McCrae, and Jorge Gracia. 2020. *Linguistic Linked Data*. Springer.

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

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

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

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

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

Eleonora Litta and Marco Passarotti. 2019. (When) inflection needs derivation: a word formation lexicon for Latin. In Nigel Holmes, Marijke Ottink, Josine Schrickx, and Maria Selig, editors, *Lemmata Linguistica Latina. Volume 1. Words and Sounds*, De Gruyter, Berlin, Boston, pages 224–239.

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

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

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