Including the *Word Formation Latin* Resource in the LiLa Knowledge Base

Matteo Pellegrini, Eleonora Litta, Marco Passarotti, Francesco Mambrini, Giovanni Moretti
Università Cattolica del Sacro Cuore, Milano

1 Background and Motivation

Nowadays, a continuously increasing quantity of resources (like corpora, dictionaries and lexica) and Natural Language Processing (NLP) tools is available for several languages. However, such resources and tools are often not able to interact with each other, making it difficult to search for pieces of information coming from different sources. To tackle this problem, in recent years there has been a trend towards applying techniques of the so-called Linked Data paradigm to linguistic data, creating a Linguistic Linked Data Cloud of interoperable resources (Cimiano et al., 2020).

The aim of the *LiLa* project[1] is to include Latin into this framework, by creating a Knowledge Base (KB) of interlinked resources for Latin using a common vocabulary for knowledge description. Here, we focus on the treatment of word formation in the LiLa KB, that already provides some derivational information taken from the Word Formation Latin (WFL) database (Litta & Passarotti, 2019). WFL on its part adopts a step-by-step, morphotactic approach where each lexeme is linked to the one from which it is directly derived by means of a specific word formation rule (WFR), thus providing a detailed, hierarchical information that is not currently encoded in the KB.

In this contribution, we describe a model designed to represent all the information contained in WFL in the LiLa KB, highlighting the theoretical principles underlying the differences in the current treatment of word formation in LiLa *vis-à-vis* the one of WFL. In Section 2, we describe the architecture of the LiLa KB on the one hand and of the WFL database on the other hand. We outline the model that we propose in order to include WFL in LiLa, showing how it interacts with other more general-purpose models developed by the Linked Data community, particularly the Morphology Module (Klimek et al., 2019) of the OntoLex-Lemon vocabulary for describing lexical resources (McCrae et al., 2017). In Section 3, we discuss some cases of research questions that are not easily answered with the information currently provided in the KB alone and require the use of WFL, and *vice versa*, exemplifying the benefit of having different pieces of information in a unified fashion, as it is allowed by the inclusion of WFL into LiLa.

2 LiLa and WFL

By adopting the data model of the Resource Description Framework (RDF), LiLa expresses information in terms of triples, that connect a subject – a labeled node in the graphical representations that follow – to an object – another labeled node, or a literal – by means of a property – a labeled edge. More specifically, the intuition behind the way in which LiLa achieves the desired interoperability between distributed resources is based on the central role of words: a pivotal role is played by the `Lemma`, defined as the canonical form of a lexical item, i.e. its citation form. The backbone of LiLa's architecture is a Lemma Bank that contains all the lemmas of the database of the morphological analyzer Lemlat (Passarotti et al., 2017). Among else, the Lemma Bank currently also provides some derivational information. Besides lemmas, two other classes

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

of entities are involved in the treatment of word formation in LiLa, namely `Affixes` (in their turn divided into `Prefixes` and `Suffixes`) and `Bases`, defined simply 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 displays by means of the properties `hasPrefix` and `hasSuffix`. This results in a flat structure, as shown in Figure 1.

As for WFL, it is a derivational lexicon of Latin whose structure is devised according to the Item-and-Arrangement (I&A) morphology model (Hockett, 1954). Lexemes that are considered as deriving from one another are connected via WFRs of different kinds. More specifically, there are compounding rules and derivation rules; in turn, within derivation rules, affixation (divided into prefixation and suffixation) and conversion are distinguished. Furthermore, WFRs are classified according to the Part-of-Speech of the lexemes they take as input and output. This results in a hierarchical structure represented by a directed tree-graph, that takes root from the ancestor – the lexeme from which all the lexemes belonging to the same word formation family ultimately derive – and links it to all derivatives by means of the successive application of individual rules, as shown in Figure 2.

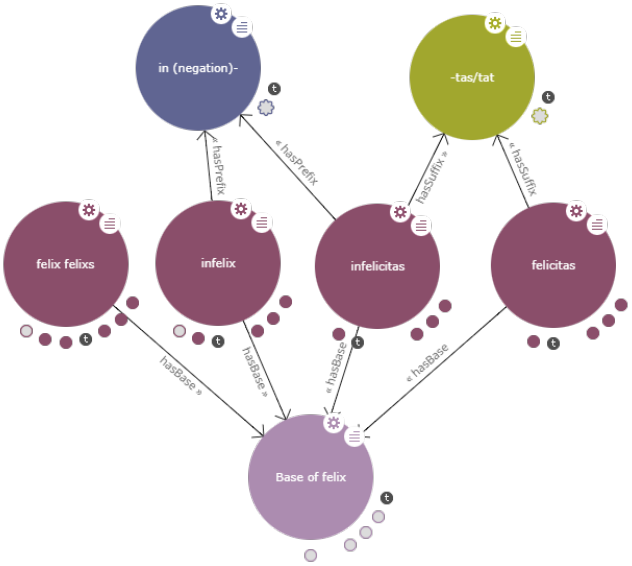

Figure 1: Word Formation in the Lemma Bank

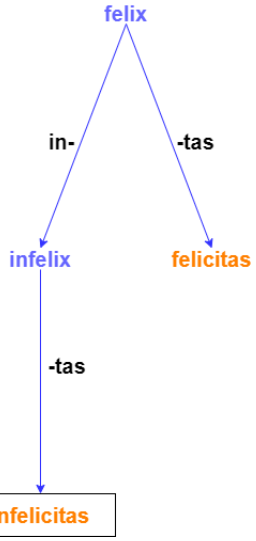

Figure 2: Word Formation in WFL

The flat organization of derivational information in the Lemma Bank was specifically envisaged to overcome some of the limits that have been observed regarding the treatment of word formation in WFL (Budassi & Litta, 2017). Indeed, the rigidly hierarchical structure of WFL forces it to make a choice about the directionality of conversion processes, even when there are doubts (e.g., does the noun ADVERSARIUS 'opponent' derive from the adjective ADVERSARIUS, or *vice versa*?), and to create fictional entries to account for cases where more than one affix appears to be simultaneously added to a base (e.g. EXAQUESCO 'to become water' from AQUA 'water', for which neither *AQUESCO nor *EXAQUO are attested as intermediate steps). As is argued by Litta et al. (2020), the flat approach adopted in LiLa allows for a more natural treatment of such cases, by providing a different modelling strategy compatible with Word-and-Paradigm (W&P) frameworks, and especially Construction Morphology (Booij, 2010). Nevertheless, this means that a lot of potentially useful information provided by WFL is not currently represented in LiLa. In what follows, we will describe the work that has been done in order to include such information within the architecture of LiLa, as summarized in the example in Figure 3.

In our model, the words of WFL that are derived from one another are treated as instances of the class `LexicalEntry` of OntoLex-Lemon, and they are seen as connected by an individual word

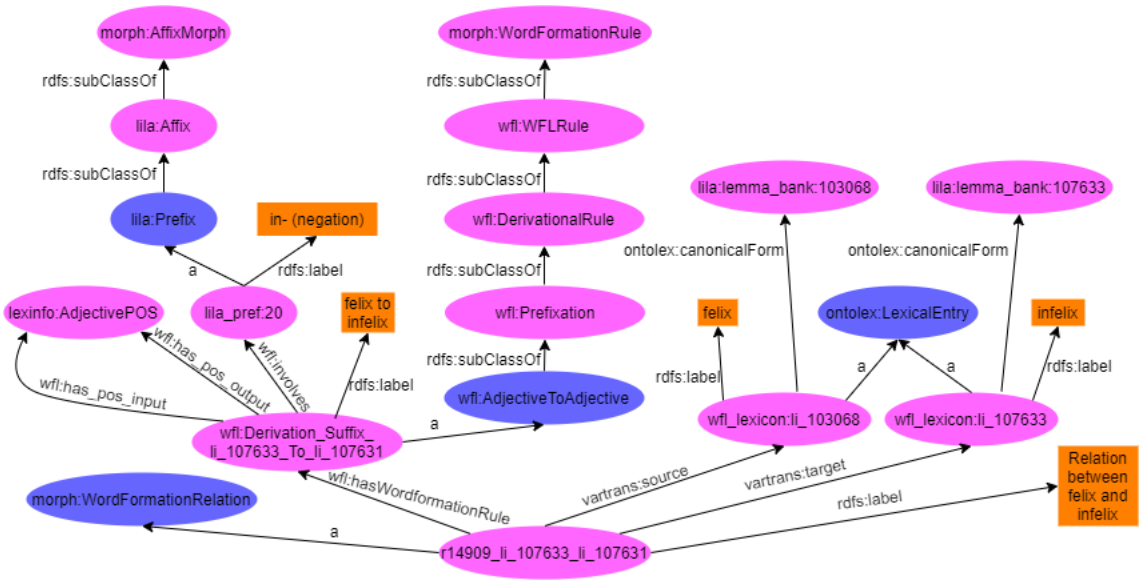

Figure 3: Example of prefixation in the WFL ontology

formation relation – i.e., an instance of the class `WordFormationRelation` of the Morphology Module.[2] More precisely, each relation is linked to the lexical entries of the input and output of the word formation process using properties taken from the Variation and Translation (vartrans) Module, namely `source` and `target`, respectively. Importantly, this allows to express the directionality of the word formation process as stated in WFL, thus ensuring that its hierarchical structure is preserved. Each relation is then linked to the WFR it instantiates according to WFL – in this case, an instance of the class of rules forming deadjectival adjectives – by means of the property `hasWordFormationRule`. The connection with the Morphology Module is achieved by establishing a sub-class relation between its rules (`WordFormationRule`) and the ones of WFL (`WFLRule`). Rules are classified in a way that accurately reflects the structure of WFL, as described above: firstly, there are two sub-classes `CompoundingRule` and `DerivationalRule`, with the latter in its turn displaying three sub-classes, `Suffixation`, `Prefixation` and `Conversion`; secondly, they are distinguished on the basis of the lexical category of the base and derivative, by providing a connection to the Parts-of-Speech of LexInfo (Cimiano et al., 2011) using the properties `has_pos_input` and `has_pos_output`, respectively. Lastly, each affixal rule is also linked to the prefix or suffix it displays, as expressed in the LiLa ontology, by means of the property `involves`, and again a sub-class relation is established between the affixes of LiLa and the ones of the Morphology Module to ensure interoperability.

3 Discussion and Conclusions

We have seen in Section 2 that the choice of a flat approach to word formation in the Lemma Bank of LiLa was motivated by the difficulties raised by the rigidly morphotactic approach of WFL in treating specific phenomena. However, it should be stressed that there are also relatively uncontroversial cases where the more detailed, hierarchical information provided by WFL on the order of application of different processes can prove helpful.

To give an example, it might be useful to be able to distinguish lexemes that are actually formed by means of an affix from the ones that simply display that affix because it has been

[2]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.

introduced in a previous step of their derivational history. This kind of information is not available in the Lemma Bank, that simply records the affixes present in each lemma. For instance, the adjectives INFRUCTUOSUS 'unfruitful' and INIURIOSUS 'injurious' have a similar structure on the surface, both displaying the prefix *in-* (negation) and the suffix *-os*. However, their derivational history is quite different, the former being undisputably formed by prefixation of *in-* to FRUCTUOSUS 'fruiful' (as *INFRUCTUS is not only not attested, but arguably not even possible as a Latin word), and the latter being clearly a result of the suffixation of *-os* to INIURIA 'injury' (*IURIOSUS). Therefore, the additional information provided by WFL is crucial.

Conversely, if the objective is to obtain all the lexemes that display a given affix, regardless of its position in the linear order of morphs and/or of its derivational history, this can be done with the information provided in the Lemma Bank. For instance, all the verbs containing the prefix *ob-* can trivially be extracted from the Lemma Bank, as they are all linked to that suffix by means of the property `hasPrefix`, while an analogous search in WFL would have trouble finding a verb like OBDURESCO 'to become hard', since it is not considered as formed directly by prefixation of *ob-*, but rather by suffixing *-sc* to the prefixed verb OBDURO 'to harden'.

One of the main advantages of the adoption of the Linked Data standards mentioned in Section 1 is exactly the possibility of not having to force a decision between the two approaches: both of them are made available within a unified framework, leaving up to scholars the choice of the one that is more compatible with their theoretical view, or that merely provides the kind of information more appropriate for the case at hand. This also allows to make the two approaches easily interact in case pieces of information from different sources are needed.

References

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

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

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

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

Hockett, Charles F. 1954. Two models of grammatical description. *Word* 10. 210–234.

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

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

Litta, Eleonora, Marco Passarotti & 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.

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

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