Of nodes and cells. Two perspectives on (and from) Word Formation Latin

Eleonora Litta - Marco Passarotti - Marco Budassi - Marco Pappalepore

ABSTRACT

The "LiLa: Linking Latin" project involves the creation of a Knowledge Base of linguistic resources for Latin based on the Linked Data framework. The ultimate goal is to reach full interoperability on the web between distributed lexical and textual resources. LiLa integrates all types of annotation applied to a particular word/text into a common representation where all linguistic information contained in a linguistic resource becomes accessible. The LiLa Knowledge Base is thus a collection of resources represented with a shared vocabulary of (meta)linguistic knowledge description. The inclusion in the Knowledge Base of information on word formation, extracted from the Word Formation Latin lexical resource, raised a number of theoretical and practical issues concerning its treatment and representation. This paper discusses such issues, presents how they were addressed in the project with the help and implementation of a Word&Paradigm theoretical model, and describes how the word formation data were included in the LiLa ontology.

1. LiLa

The increasing number of complex and diverse linguistic resources for several languages has raised expectations, in recent years, for the potentials of interoperability of (annotated) corpora, dictionaries, thesauri, lexica and Natural Language Processing (NLP) tools (Ide – Pustejovsky 2010). However, besides infrastructural initiatives collecting resources and tools, that act as web portals to query their data

Lingue antiche e moderne 9 (2020) ISSN 2281-4841
DOI: 10.4424/lam92020-6

(like, for instance, CLARIN and META-SHARE)[1], there is nothing yet that can provide real interconnection between them.

The "LiLa: Linking Latin" project[2] aims at creating a Knowledge Base of linguistic resources for Latin based on the Linked Data framework, i.e. a collection of several data sets described using the same vocabulary for knowledge description, so that they can be linked together and interact. This makes possible a better exploitation of the linguistic resources and NLP tools for Latin developed so far.

The LiLa Knowledge Base is highly lexically-based: the main assumption behind the design of the interactions that are going to make LiLa work is that textual resources are made of (occurrences of) words, lexical resources describe properties of words, and NLP tools process words. Hence, words are the pivotal elements that link resources together. Particularly, the lemma is considered the ideal interconnection between lexical resources, annotated corpora and NLP tools that lemmatise their input text.

One of the LiLa's core components is a collection of Latin lemmas extracted from the morphological analyser Lemlat (Passarotti *et al.* 2017). The Lemlat lexical basis was originally collated from three Classical Latin dictionaries (Georges 1972; Glare 1982; Gradenwitz 1904). Subsequently, the lexical basis was further enlarged by adding most of the Onomasticon (person names, place names, names linked to ethnicity, and adjectives derived from these names, e.g. *aaroneus* "of Aaron") provided by Forcellini's dictionary (Budassi – Passarotti 2016), and the full list of lemmas of the Medieval Latin glossary by du Cange (Cecchini *et al.* 2018).

In the LiLa Knowledge Base, interoperability is achieved by linking all entries in lexical resources and all *corpus* tokens that refer to the same lemma. The repository of lemmas serving as a hub in LiLa was built in such a way that it is able to harmonise different lemmatisation strategies. Indeed, although selecting canonical forms to be used as

[1] https://www.clarin.eu; http://www.meta-share.org.
[2] https://lila-erc.eu.

lemmas is a process that tends to follow a standardised series of language-dependent conventions (e.g. for Latin, the form in the nominative singular for nouns, or the first singular person of the indicative present tense for verbs), different corpora, lexica and tools may adopt different strategies to solve conceptual and linguistic challenges posed by lemmatisation. Such challenges follow under two main categories (Mambrini – Passarotti 2019):

a) the form of the lemma. Different citation forms can be chosen to represent the lemma for the same lexical item. These include alternations (a) in spelling (e.g. *sulphur* vs. *sulfur* "brimstone"), (b) in ending and possibly inflectional type (e.g. *diameter* vs. *diametros* vs. *diametrus* "diameter"), or (c) in the paradigmatic slot representing the lemma (e.g. *sequor* "to follow", first person singular of the passive/deponent present indicative vs. *sequo*, first person singular of the active present indicative). In LiLa, these cases are managed through different 'Written Representations' of a lemma for (a) and 'Lemma Variants' for (b) and (c);
b) the lemmatisation criteria. Various lemmas can be assigned to the same word form in different resources. For instance, participles can be considered either part of the inflectional paradigm of verbs, or independent lemmas provided with an autonomous entry in lexical resources (e.g. *doctus* "learned", morphologically the perfect participle of *doceo* "to teach"). The same holds true for deadjectival adverbs (e.g. *aequaliter* "evenly" from *aequalis* "equal"), which are either lemmatised as forms of their base adjective, or treated as independent lemmas. In LiLa, this is solved by using different 'Hypolemmas' for the same lexical entry (see Section 4, Figure 2).

The first lexical resource to be linked to the LiLa Knowledge Base was Word Formation Latin (Litta – Passarotti 2019)[3], a derivational

[3] Word Formation Latin has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 658332-WFL. The project (2015-2017) was based at the Centro

morphology-based lexical resource for Classical and Late Latin that connects lemmas on the basis of word formation rules. The inclusion of WFL into LiLa has highlighted a number of theoretical and methodological issues concerning the treatment and representation of word formation in a lexical resource, that emerged at the end of the project (Budassi – Litta 2017). This paper discusses such issues and presents how they were addressed during the inclusion of WFL into LiLa.

2. Word Formation Latin

Word Formation Latin (WFL; http://wfl.marginalia.it/) is a lexical resource describing word formation in Classical and Late Latin. In WFL, derivational and compounding word formation rules (WFRs) are modelled as directed one-to-many input-output relations between lemmas. The structure of the lexicon was designed on the basis of the Item-and-Arrangement (I&A) model of morphological description by Hockett (1954), according to which lemmas are either non-derived lexical morphemes, or a concatenation of a base in combination with affixes. I&A was chosen as a theoretical model for the resource, both because it emphasises the semantic significance of affixal elements and because it had been previously adopted by other resources treating derivation, such as the morphological dictionaries Word Manager (Domenig – Ten Hacken 1992).

WFL uses a step-by-step morphotactic approach to account for word formation processes. In the specific case of affixation, each time an affix is added to a simpler word to form a more complex one, this process is considered a single WFR. Prefixation and suffixation rules are hence described as individual steps in the word formation process, but the same happens also with other rule types. For instance, a lemma

Interdisciplinare di Ricerche per la Computerizzazione dei Segni dell'Espressione (CIRCSE), at the Università Cattolica del Sacro Cuore, Milan, Italy.

derived from another without affixes and showing a part-of-speech (PoS) change, is said to be the result of conversion: *albus* "white" > *albeo* "to be white" (A-To-V conversion) > *albesco* "to become white" (V-To-V *-sc*, suffixation) > *exalbesco* "to turn pale" (V-To-V *ex-*, prefixation). Hence, the output of a WFR usually is the result of the application of one morphotactic step in the derivation chain, whether that be the addition of a morpheme (prefixes or suffixes), or a change of PoS (like in the case of conversions). However, sometimes the rule involves both the addition of a suffix and a change of PoS, like for example in cases such as *albus* > *albedo* "whiteness", involving both a shift from adjective to noun and the addition of suffix *-edo/-edin*. Each output lemma can only have one input lemma, unless the output lemma qualifies as a compound.

Such organisation of data results in hierarchical structures represented as directed rooted graphs resembling a tree, whereby one or more lemmas derive from one ancestor lemma. In the graph of Figure 1 (the 'tree'), nodes are occupied by lemmas, and edges are labelled with a description of the WFR used to derive the output lemma from the input one. The group of lemmas contained in the same full derivational tree is described as a 'word formation family'.

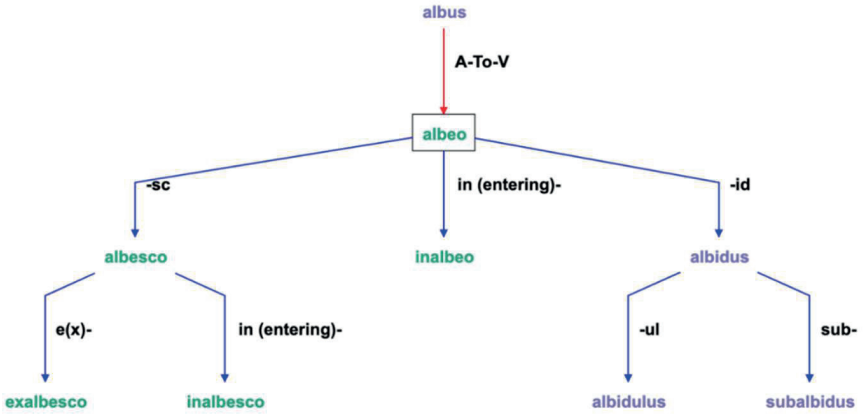

Figure 1. Partial directed rooted graph for the word formation family of *albus* "white".

Towards the end of the WFL project, it was evident that portraying word formation processes via directed rooted graphs had created some significant theoretical concerns, the main reason being that sometimes the derivational relation is ambiguous or unsuitable to be represented straightforwardly by a single step-by-step process, as shown in Budassi – Litta (2017). The types of issues encountered can be listed as follows:

1. Doubtful directionality (e.g. *nuntio* "to announce" > *renuntio* "to report" > *abrenuntio* "to renounce" > *abrenuntiatio* "renounciation" vs. *nuntio* > *renuntio* > *renuntiatio* "proclamation" > *abrenuntiatio*);
2. Double/triple affixation (e.g. *perturbo* "to disturb" > *imperturbabilis* "that cannot be disturbed");
3. Backformation (e.g. *lido* "to break against" < *allido* "to crush against");
4. Diachronic discrepancies (e.g. *exhorresco* "to be terrified" is attested before *exhorreo* "to dread" even if it contains an extra *-sc* suffix) (Haverling 2000);
5. Conversion in case of borrowings (e.g. *astrologus* N / *astrologus* A "astrologer" / *astrologia* "astrology" N: N-To-N -i, vs. A-To-N -i).

In cases such as (1), WFL resorts to a series of tactics to work around the problem. To give an example, when considering the relation between *abrenuntiatio*, *renuntiatio* and *abrenuntio*, in which the origin of a word (*abrenuntiatio*) can be either one or another word, there is a lot of space for interpretation on which direction the change has happened from/to, and which between *abrenuntio* or *renuntiatio* generated the derived lemma *abrenuntiatio*. The *Oxford Latin Dictionary* (*OLD*) by Glare (1982) is employed first in the compilation of WFL to verify the derivational history of lemmas, followed by Georges (1972) where *OLD* does not contain the given lemma. In the example case, it is the Georges dictionary that reports how *abrenuntio* > *abrenuntiatio* is the correct process, which is thus recorded as such in WFL. This is confirmed also by *TLL* (2009). Even so, the description of this relationship is still to be considered rather arbitrary, considering

the absence of native speakers to inform us on the nuances of meaning, much needed to describe such a word formation process.

Another method used in WFL to work around non-linear derivations caused by simultaneous double/triple affixation, as described at point (2.), is the creation of 'fictional' lemmas that act as stepping stones between attested words in order to justify extra morphotactic steps. In WFL there are 379 fictional lemmas acting as intermediate steps from one lemma to another. In addition to these, 7 fictional roots have been inserted. This has been necessary in order to group in a family a number of lemmas that otherwise would have not been linked. For example, the root **cello* was created to keep together *antecello* "to surpass", *decello* "to tend from", *excello*/*excelleo* "to rise", *percello* "to strike down", *praecello* "to be superior", *procello* "to throw violently forward", *recello* "to recoil", and their derivates[4]. The existence of fictional lemmas in WFL has however proven to be less than ideal, as exposed in Budassi – Litta (2017) and Litta – Budassi (2020). For example, a big portion of fictional lemmas (103) consists in second class adjectives with the *-bil* suffix (17% of the total number of lemmas derived using the *-bil* suffix). Most of these were created to keep together lemmas such as the adverb *imperabiliter* "peremptorily" to their 'next of kin', the verb *impero* "to command". Since, as mentioned above, in WFL it is not possible to connect two lemmas using two suffixes at the same time (*-bil* and *-ter* in this case), the fictional adjective **imperabilis* was created to function as a further step in the step-by-step word formation process. The presence of fictional lemmas in the WFL dataset means

[4] The other 6 fictional roots in WFL are: **cumbo* (*accumbo* "to lie down", *percumbo* – unused but contained in Georges as mentioned by Varro, *De Lingua Latina*, 9, 49, *succumbo* "to fall down", *discumbo* "to recline", *incumbo* "to lean", *occumbo* "to fall in death", *procumbo* "to fall forwards", *recumbo* "to lie down again", *concumbo* "to lie together"); **gruo* (*congruo* "to coincide" and its family, with *ingruo* "to assail"); **nuo* (*abnueo*/*abnuo* "to deny", *adnuo* "to nod to" and *nuto* "to nod"); **insuasibilis* (*insuasibilitas* "incomprehensibility" and *insuasibiliter* "inaccessibly"); **perior* (*experior* "to prove", *supperior* "to undergo", *opperior* "to attend", *periculum* "attempt"); **temerus* (*temere* "by chance", *temeritas* "chance", *temeriter* "by chance/accident", *temero* "to treat rashly").

that when making general considerations on the distribution of the *-bil* suffix in Classical and Late Latin, for instance, one should keep in mind that a good portion of what is extracted from WFL needs to be discarded.

These solutions, relying on dictionaries and creating fictional lemmas, appear temporary and tentative, and do not offer full support when dealing with e.g. gaps in attestation, backformation, diachronic discrepancies, analogy, or doubtful borrowings from other languages, as exemplified in points (3)-(5) above.

3. Derivational Paradigms

The recent interest for the application of Word and Paradigm (W&P) models to derivational morphology (see e.g. Štekauer 2014) led to the exploration of their potential in considering a less problematic approach to word formation, as described in Litta – Budassi (2020). According to Hathout – Namer's (2019: 160-161) definition, a «derivational paradigm is […] an arrangement of derivational families. These families are arranged with respect to a set of arrangement relations», that is relations that «connect lexemes formed by a same derivational process».

Bonami – Strnadová (2019) describe the 'paradigmatic system' as a combination of morphological families that are related (or aligned) in exactly the same way. More precisely, if two pairs of morphologically related words hold the same content and/or form relation, this relation is an 'aligning relation'. Each pair of morphologically related words is a (partial)[5] family, a set of two or more (partial) families forms a

[5] A family can be 'partial' when other members (words) not taken in consideration exist; a set of families can also be 'partial' when it is not necessarily exaustive (hence when other families exist that could align into the system). A set of (partial) families constitutes a (partial) paradigmatic system. The majority of examples given in literature when describing paradigmatic systems are constituted by partial families.

'paradigmatic system'. This concept can be applied to both inflection and derivation. In the case of derivation, the focus rests on alignment based on content rather than strictly on form. Hence, in the case of Latin derivation, we can have (partial) families such as *mon-eo* "to admonish", *ad-mon-eo* "to bring to mind", *ad-mon-i-tio* "suggestion", that aligns, in form and content, with *iuv-o* "to help", *ad-iuv-o* "to help", *ad-iuv-a-tio* "help", but we could also have *cresco* "to grow", *ac(ad)-cresco* "to become larger by growing" that aligns with *acuo* "to sharpen", *ex-acuo* "to make very sharp", because of the intensification meaning that both prefixes *ad-* and *ex-* hold in these instances.

The first advantage that the paradigmatic system offers is that it does not need to describe word formation relationships in a necessarily linear way directed from an input to an output form (i.e. not in a tree-like structure with a root and branches). Instead, derivation can be intended in purely networking terms.

The other advantage of pursuing a W&P perspective on derivational morphology is that, by stacking morphological families, a model can be built in which the cells in the paradigm have descriptive and predictive power, where the availability of slots is more important than the form filling them (Bauer 1997), thus highlighting general behaviours. Such an approach might not seem useful when applied to a dead language, as there might seem to be no need for regularity and predictability for the formation of new, potential words. However, the readiness of a slot – consider for example the slot usually containing negative adjectives prefixed by *in-* – makes it possible for *imperturbabilis* to have formed without the need for the existence of a corresponding positive adjective such as **perturbabilis*.

As a consequence, the fundamental characteristic of derivational paradigms – besides the total absence of directionality in the description of word formation processes – is the focus on the concept of 'cell', a slot that allows for the rise of a word without the need for a direct relationship with a simpler word. In W&P, each cell should be thus enhanced with information on semantic features, due to the underlying role of semantics in accounting for derivational processes. However,

throughout the wide diachronic span of Latin, some affixes tend to undergo a semantic shift, thus making their semantic labelling difficult; for instance, the suffix *-sc* loses its inchoative semantic value throughout time (Haverling 2000), while the negation prefix *in-* always keeps its privative meaning. Additionally, cells are required to preserve morphological integrity, that is to say that while assigning semantic features we cannot neglect morphological features: the same semantic feature of an affix (e.g., *-(t)or*, masculine agent/instrument) does not blur morphological differences. For instance, *dictor* "the one who says", which shares the same base with the verb *dico* "to say", is rather semantically different from *dictator* "dictator", even though in both cases the same suffix *-(t)or* is added to the same base, since the latter features also the iterative affix *-it* (Litta – Budassi 2020).

4. WFL in LiLa

The components of the LiLa Knowledge Base and their relations are formalised in an ontology made of:

a) Individuals: instances of objects (e.g. one specific token, or lemma);
b) Classes: types of objects/concepts (e.g. Token, Lemma, Form);
c) Data properties: attributes that objects can have (e.g. morphological features for lemmas/tokens, like PoS, inflectional category, gender etc.);
d) Object properties: ways in which classes and individuals can be related to one another.

Object properties are expressed in terms of RDF triples (Resource Description Framework; Lassila *et al.* 1999). These are sets of statements that describe semantic data in the form of subject-predicate-object expressions: (1) a predicate-property (a relation; in graph terms: a labelled edge) connects (2) a subject (a resource; in graph terms: a

labelled node) with (3) its object (another resource/node, or a literal, e.g. a string).

Relations are assigned labels taken from a restricted vocabulary of knowledge description, such as 'hasLemma', 'hasPOS', 'hasGender' and 'hasInflectionType'. Each component of the ontology, as well as its instantiations in the Knowledge Base, is uniquely identified through a Uniform Resource Identifier (URI).

Including WFL into LiLa represented an opportunity to implement an approach to Latin word formation that could fit in a structure that needs to be declarative rather than procedural, and that omits postulations on directionality and step-by-step derivational processes, in order to avoid the theoretical problems described above.

The theoretical framework of the word-(and sign)-based model known as Construction Morphology (CxM) (Booij 2010) was crucial as a starting point for theorising a model for including WFL data into the LiLa Knowledge Base. According to Booij (2009: 201), «word formation patterns can be seen as abstract schemas that generalise over sets of existing complex words with a systematic correlation between form and meaning». Such abstract schemas describe words in their internal structure in pair with details on the meaning of this structure. For instance, the adjective *imperturbabilis* can be analysed as follows[6].

(1) $[in_i\ [per[turb]_y]_j\ (a)\ bil_k]_{A2} \leftrightarrow [that\ can_k\ not_i\ be\ [DISTURB]_j ed]_A$

This 'construction' can be further abstracted into what in CxM terms is called a 'schema':

(2) $[in_i\ [per[x]_y]_j\ (a)\ bil_k]_{A2} \leftrightarrow [that\ can_k\ not_i\ be\ [SEM]_j ed]_A$

[6] Constructions and schemas use subscripted letters as placeholders for morphological and semantic features that are usually described elsewhere. For example, subscripted *i* indicates which of the two *in-* prefixes (negative or entering) is being used in this specific construction, whereas subscripted *y*, *j* and *k* may signify matches between the various elements in the construction, e.g. *bil* = *can*, and *A* indicates that the PoS of this lexeme is an adjective.

This schema can be applied to the description of all adjectives of the second class that include the suffix *-bil* and the negative prefix *in-*, such as *inaccessibilis* "unapproachable", *incogitabilis* "thoughtless", *inconsolabilis* "inconsolable" and *inconvertibilis* "unchangeable".

CxM schemas are word-based and declarative, which means that they describe static generalisations, as opposed to explaining the procedure of change from one PoS to another like WFRs do, and are merely output-oriented. This is particularly appropriate for the needs of LiLa, because if words can be described into their formative elements, these can consequently be organised into classes of objects in an ontology and thus connected in the Knowledge Base.

Thus, to include WFL into LiLa, data was extracted from WFL through a process that turned it into triples fitting the requirements of the ontology behind the LiLa Knowledge Base. In such ontology, three classes are being used for the treatment of word formation: (1) Lemmas, (2) Affixes, including two subclasses: Prefixes and Suffixes, and (3) Bases.

The process of 'triplication' of the WFL data flattens the input-output WFR-based relations, by assigning, for each lemma, one triple for each suffix/prefix found along the derivational path for that lemma in WFL. For instance, considering the WFL tree reported in Figure 1, the lemma *albidulus* "whitish" is connected via a triple with suffix *-ul* and via another triple with suffix *-id*. All the lemmas belonging to the same word formation family are then connected via a triple with the same Base.

In LiLa, bases are for the moment not further described and are just identified through a sequential number, only acting as connectors of Lemmas belonging to the same word formation family. The problem with providing a Base with some kind of linguistic information is both theoretical and practical. From the theoretical point of view, it is hard to decide whether a Base can be represented simply by a root, or whether it should contain all the graphical variants it appears in throughout the whole of the word formation family. Consider as a chief example, the family where verb *fero* "to bring/carry" belongs: it

includes both lemmas featuring the base *fer-* (e.g. *affero* "to bring/carry to a place", *aurifer* "goldbearing", *circumferentia* "circumference", *fertilis* "fertile", *praefero* "to carry in front") and lemmas with the base *lat-* for the perfect participle/supinum from the inflectional paradigm of *fero* (e.g. *collator* "he who bears", *illatio* "a carrying in", *latifundium* "large estate", *latus* "carried", *relatio* "a retort"). The issue here would be to decide whether a label such as *fer-*/*tul-* would be an acceptable way of describing a Base in LiLa. The practical problem is mainly due to finding a way of automating the extraction of a Base from the lemmas belonging to the same family throughout the whole resource, and to perform an obligatory manual checking on this automatisation for the 3,852 Bases currently contained in LiLa.

Like any component in LiLa, also Affixes and Bases are assigned a URI. Affixes and Lemmas are assigned one, or more, Written Representation(s), i.e. the actual string(s) they are realized as[7].

Affixes and Lemmas are connected to each other via labelled edges (thus, forming triples). A Lemma node in the LiLa Knowledge Base is linked (a) to the Affix nodes that are part of its derivational process through edges labelled hasPrefix or hasSuffix[8] and (b) to its Base (or Bases, in the case of compounds) through an edge labelled hasBase. Lemmas are never related to each other, so as not to take assumptions on the direction of the formative process.

For instance, the adjective *imperturbabilis* has the following characteristics:

[7] The terminology is taken from the Lemon RDF model for representing lexical information relative to ontologies (https://lemon-model.net).

[8] In the LiLa ontology, prefixes and suffixes belong respectively to classes Prefix and Suffix, which are subclasses of the class Affix, in turn a subclass of the class Morpheme (which includes Affixes and Bases). Homonymous affixes, like for instance the two prefixes *in-* (respectively, in the negative and entering meanings), are kept separate.

- type ‘Lemma’;
- Written Representations: ‘imperturbabilis’ and ‘inperturbabilis’ (spelling variant);
- hasBase ‘1102’;
- hasDegree ‘positive’;
- hasInflectionType ‘second class adjectives (nom. sing. in: *-is*, *-e*)’;
- hasPOS ‘adjective’;
- hasPrefix ‘prefix14’ (*per-*) and ‘prefix20’ (*in-* negation);
- hasSuffix ‘suffix25’ (*-bil*)[9].

Figure 2 offers a view of the elements describing lemma *imperturbabilis* in the graphical representation of the contents of the LiLa Knowledge Base available at https://lila-erc.eu/lodlive.

[9] Figure 2 includes also an isHypolemma relation holding between the node for the adverb *imperturbabiliter* and that for *imperturbabilis/inperturbabilis*. As mentioned in Section 1, in LiLa participles and deadjectival adverbs are treated as Hypolemmas (a subclass of Lemma), which are connected to their reference lemmas through the isHypolemma relation, to harmonise different lemmatisation criteria in the various resources connected in LiLa.

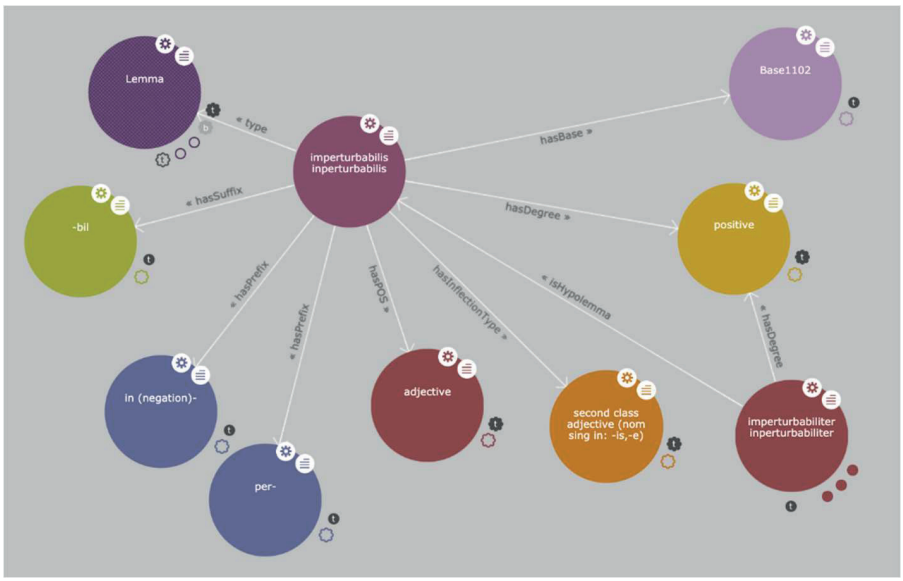

Figure 2. Graphical representation of the triples around the lemma *imperturbabilis* in LiLa.

The Base node numbered 1102 has 43 ingoing hasBase edges, one for each of the lemmas belonging to the word formation family *imperturbabilis* belongs to. Figure 3 shows a selection out of these that are also connected to other affixes, like the prefixes *per-* (*perturbatio* "confusion", *perturbator* "disturber", *perturbidus* "very unquiet", *perturbo* "to confuse/disturb") and the suffix -(*t*)*or* (*perturbator*). As it can be seen from the Figure, the representation of relationships between lemmas is flattened, and there is no implication on whether *imperturbabilis* derives from *perturbo*, as they only share the same base and both feature a relationship with the prefix *per-*.

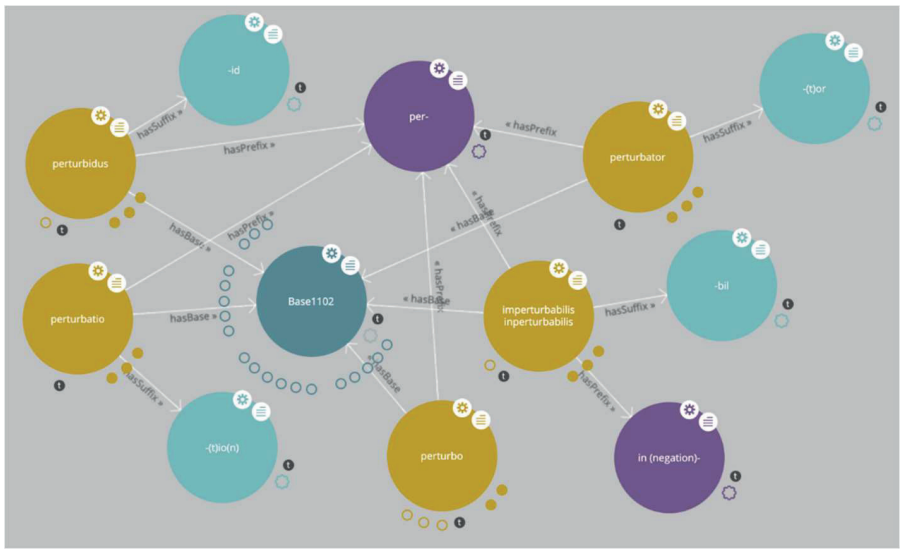

Figure 3. Portion of the word formation family of imperturbabilis in LiLa.

Such an organisation of the lexical data makes it possible to query the Knowledge Base in order to find out, for example, that there are 296 second class adjectives featuring both the negative prefix *in-* and the suffix *-bil* over a total of 2,906 second class adjectives in the Classical Latin portion of the LiLa collection of lemmas.

Backformation does not represent a theoretical issue in LiLa, as the flat organisation of data considers all derivational relations being on the same level, without assumptions on directionality. Consider, for instance, the example given in point (3) Section 2 above: the verb *lido*, found in only one occurrence in Lucr. 5, 1001 *nec turbida ponti aequora lidebant navis ad saxa virosque* "Nor did the dark billows of the sea smash the ship and the men on the rocks", is said by *OLD* to be perhaps a backformation from *allido*, a prefixed verb related to *laedo* "to injure" and to a series of prefixed verbs characterised by apophony (*con+laedo* > *collīdo* "to clash", *dilīdo* "to batter to pieces", *elīdo* "to tear out", *illīdo* "to strike against", *interlīdo* "to strike out", *oblīdo* "to squeeze to pieces", *relīdo* "to strike back", *sublīdo* "to press out"), or not (*illaedo* "to not hurt", *relaedo* "to pound/crush"). It seems

reasonable that any of these verbs, carrying similar meaning of clashing/beating/crushing, could have been the one from which *lido* might have been backformed. The fact that *lido* is only attested in Lucretius could suggest that the author adapted the "clashing/striking" meaning lying beneath the base (*laed-*/*līd-*), shared by all these verbs, for economic reasons due to metric syllabification. However, this does not mean that *lido* was never used by speakers. In the LiLa Knowledge Base, all these verbs are just connected to the affix(es) they feature and, most importantly, to the same base, thus representing that they belong to the same word formation family, but without providing any assumption about the direction of their word formation process(es).

Another situation that has been vastly improved by the rendition of word formation relationships in LiLa is that of a few lemmas whose derivational history was forced in WFL without entirely sticking to the resource's declared methodology. It is the case of those verbs that can be described as belonging to the so-called Caland System (Rau 2009). Consider certain semantic alternations such as *caleo* "to be hot" ~ *calesco* "to become hot" ~ *calefacio* "to make something/someone hot" ~ *calefio* "to be made hot", or *liqueo* "to be fluid" ~ *liquesco* "to become fluid" ~ *liquefacio* "to make fluid" ~ *liquefio* "to be made fluid"[10]. In this case, two oppositions are identifiable: (i.) that between a 'basic' (i.e., non-suffixed) verb and a suffixed verb (namely, *caleo* ~ *calesco* and *liqueo* ~ *liquesco*), and (ii.) that between those two verbs and a compound verb in *-facio* (active) or *-fio* (passive), namely, *calefacio*/*calefio* and *liquefacio*/*liquefio*. When such alternations occur in the paradigm, there is usually also an adjective in *-idus* (e.g. *calidus* "warm", *liquidus* "liquid"). The peculiarity of these derivations is that it is indistinguishable which are primary and which secondary derivations (to put it another way, which words are derived from which within the given suffix family). Thus, what we have is a fully paradigmatic set of correspondences characterised by a specific number

[10] For a more complete description of this system of suffixes and a list of lemmas that are part of this 'paradigm', see Litta – Budassi (2020).

of suffixes/formatives through which nouns/adjectives/verbs are derived from one root (rather than one base)[1], although it is undistinguishable which are primary and which secondary derivations. Such system, although it explains plainly the relationship between e.g. *assuesco ~ assuefacio ~ assuefio*, *consuesco ~ consuefacio ~ consuefio* (all having the general meaning of "to get/be made accustomed"), *desuesco ~ desuefacio ~ desuefio* ("to disaccustom/to be made disaccustomed"), *insuesco ~ insuefacio ~ insuefio* ("to become/be made accustomed") does not fit easily into the morphotactic model employed in WFL. According to this model, *assuefacio*, *assuefio*, *consuefacio*, *consuefio*, *desuefacio*, *desuefio*, *insuefacio*, *insuefio* have been simply connected to respectively *assuesco*, *consuesco*, *desuesco*, *insuesco.*

This linear procedure, however, does not account for the removal of suffix *-sc* from the verb, before deriving e.g. *assuefacio* from *assuesco*, just because input-output relations are not represented in LiLa[2]. This issue, which remains not ideally described in WFL, is resolved in LiLa, because for instance *assuefacio* is not connected to *assuesco* anymore, but to a base shared by *assuesco*, *assuefio*, *desuesco*, etc. as well as to a base connected to the verb *facio*, thus without any need to justify the absence of the *-sc* suffix, which is natural in a paradigmatic view on derivation.

5. Conclusions

In this paper, we have described the representation of word formation in the LiLa Knowledge Base of linguistic resources for Latin. Our objective is to address the variety of requirements needed for describing a derivational paradigm (absence of directionality, labels for both morphology and semantics), pursuing a theoretically-sound balance

[11] In Latin, from the synchronic standpoint, such a system of suffixes is made of adjectival *-id*, substantival *-(t)or* and verbal *-sc* suffixes.
[12] On derivations on the basis of *-sc* verbs see Budassi *et al.* (2019).

between ease of consultation and connection between the members of the same word formation family.

The theoretical framework used to represent word formation in LiLa adheres to state-of-the-art research regarding theoretical models on derivational morphology more than the original approach pursued in WFL. Network-like representations of word formation families, such as the one shown here, highlight the key features of the W&P framework, that is potentiality and non-directionality. Furthermore, the representation of word formation in LiLa is a step forward in both applicative and theoretical terms. Linguistic resources tend to focus on one type only of (meta)data, like PoS tagging, syntactic analysis and word sense disambiguation. Interlinking the (meta)data provided by a resource focussed on derivational morphology, such as WFL, with those from other linguistic resources, such as annotated corpora, lexica and thesauri, is an efficient way to exploit to the best the specific information available in (still) scattered resources. This is extremely valuable, since – as noted above in the case of morphology and semantics – one single linguistic level cannot be investigated without a deep comprehension of the others.

In regards to this, as far as future developments are concerned, we are planning on connecting further lexical resources to LiLa. Among these are the Latin WordNet (Franzini *et al.* 2019), a semantic dictionary where concepts are lexicalised by sets of synonymous lemmas, Latin Vallex (Passarotti *et al.* 2016), a lexicon where verbs, nouns and adjectives are described in their valency, and the *Etymological Dictionary of Latin and the Other Italic Languages*, containing information on Proto-Italic and Proto-Indoeuropean reconstructed forms (De Vaan 2018). The inclusion of these lexical resources in LiLa, together with the connection of its lemmas to their empirical usage in the texts of the Latin corpora connected to the Knowledge Base, will provide scholars with an easy and harmonised access to a wide and diverse amount of linguistic (meta)data stored in the several available resources for Latin. We hope that our effort to collect and connect such (meta)data will help to shed light also about

the chronology of word formation on a morphosemantic level, an aspect not yet explicitly represented in LiLa.

ACKNOWLEDGMENTS

The "LiLa: Linking Latin" project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme – Grant Agreement No 769994.

Università Cattolica del Sacro Cuore
Centro Interdisciplinare di Ricerche per la Computerizzazione dei Segni dell'Espressione (CIRCSE)
eleonoramaria.litta@unicatt.it

Università Cattolica del Sacro Cuore
Centro Interdisciplinare di Ricerche per la Computerizzazione dei Segni dell'Espressione (CIRCSE)
marco.passarotti@unicatt.it

Università degli Studi di Pavia
Dipartimento di Studi Umanistici
marco.budassi01@universitadipavia.it

Independent Scholar
crillion@tiscali.it

References

Bauer, L.
1997 *Derivational paradigms*, in Booij, G. – van Marle, J. (eds.), *Yearbook of Morphology 1996*, Amsterdam, Springer Netherlands, pp. 243-256.

Bonami, O. – Strnadová, J.
2019 *Paradigm structure and predictability in derivational morphology*, in «Morphology», 29:2, pp. 167-197.

Booij, G.
2005 *Compounding and Derivation: Evidence for Construction Morphology*, in Dressler, W.U. – Kastovsky, D. – Pfeiffer, O.E. – Rainer, F. (eds.), *Morphology and its demarcations: selected papers from the 11th morphology meeting, Vienna, February 2004 (current issues in linguistic theory 264)*, Amsterdam-Philadelphia, John Benjamins, pp. 109-132.

2009 *Compounding and Construction Morphology*, in Lieber, R. – Štekauer, P. (eds.), *The Oxford Handbook of Compounding*, Oxford, Oxford University Press, pp. 201-216.

2010 *Construction morphology*, Oxford, Oxford University Press.

2018 *The Construction of Words: Advances in Construction Morphology*, Cham, Springer.

Budassi, M. – Passarotti, M.
2016 Nomen omen. *Enhancing the Latin Morphological Analyser Lemlat with an Onomasticon*, in *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural*

Heritage, Social Sciences, and Humanities (LaTeCH 2016), Berlin, The Association for Computational Linguistics, pp. 90-94.

Budassi, M. – Litta, E.
2017 *In Trouble with the Rules. Theoretical Issues Raised by the Insertion of* -sc- *Verbs into Word Formation Latin*, in Litta, E. – Passarotti, M. (eds.), *Proceedings of the Workshop on Resources and Tools for Derivational Morphology (DeriMo)*, Milano, EDUCatt, pp. 15-26.

Budassi, M. – Litta, E. – Passarotti, M.
2019 *What's beyond "inchoatives"? Derivation types on the basis of* -sc- *verbs*, in Holmes, N. – Ottink, M. – Schrickx, J. – Selig, M. (eds.), *Lemmata linguistica Latina; volume 1 Words and Sounds*, Berlin-Boston, De Gruyter, pp. 240-257.

Cecchini, F.M. – Passarotti, M. – Testori, M. – Ruffolo, P. – Draetta, L. – Fieromonte, M. – Liano, A. – Marini, C. – Piantanida, G.
2018 *Enhancing the Latin Morphological Analyser LEMLAT with a Medieval Latin Glossary*, in Cabrio, P.E. – Mazzei, A. – Tamburini, F. (eds.), *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*, Torino, aAccademia University Press, pp. 87-92.

Cuzzolin, P.
2014 *L'espressione della totalità in latino*, in Manco, A. (ed.), *L'espressione linguistica della totalità*, Napoli, Quaderni di AIΩN 2, pp. 53-70.

Domenig, M. – Ten Hacken, P.
1992 *Word Manager: A System for Morphological Dictionaries. Vol. 1*, Hildesheim, Georg Olms Verlag AG.

Franzini, G. – Peverelli, A. – Ruffolo, P. – Passarotti, M. – Sanna, H. – Signoroni, E. – Ventura, V. – Zampedri, F.
2019 *Nunc Est Aestimandum Towards an Evaluation of the Latin WordNet*, in Bernardi, R. – Navigli, R. – Semeraro, G. (eds.), *Proceedings of the Sixth Italian Conference on Computational Linguistics, Bari, Italy, November 13-15*, Bari, CEUR Workshop Proceedings, AI*IA Series Vol. 2481, pp. 1-8.

Georges, K.E. – Georges, H.
1972 *Ausführliches Lateinisch-Deutsches Handworterbuch*, Hannover, Hahn.

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

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

Hathout, N. – Namer, F.
2019 *Paradigms in word formation: what are we up to?*, in «Morphology», 29:2, pp. 153-165.

Haverling, G.
2000 *On* sco-*verbs, prefixes and semantic functions: a study in the development of prefixed and unprefixed verbs from early to late Latin*, Gothenburg, Acta Universitatis Gothoburgensis.

Hockett, C.F.
1954 *Two Models of Grammatical Description*, in «Words», 10, pp. 210-231.

Ide, N. – Pustejovsky, J.
2010 *What does interoperability mean, anyway? Toward an operational definition of interoperability for language technology*,

in *Proceedings of the Second International Conference on Global Interoperability for Language Resources*, Hong Kong, China, City University of Hong Kong.

Jackendoff, R.
1975 *Morphological and semantic regularities in the lexicon*, in «Language», 51:3, pp. 639-671.

Lassila, O. – Swick, R.R.
1999 *Resource description framework (rdf) model and syntax specification*, in *W3C Recommendation*, Cambridge (MA), World Wide Web Consortium.

Litta, E. – Passarotti, M.
2019 *(When) Inflection Needs Derivation. A Word Formation Lexicon for Latin*, in Holmes, N. – Ottink, M. – Schrickx, J. – Selig, M. (eds.), *Lemmata linguistica Latina; volume 1 Words and Sounds*, Berlin-Boston, De Gruyter, pp. 224-239.

Litta, E. – Budassi, M.
2020 *What we talk about when we talk about paradigms*, in Fernández-Domínguez, J. – Bagasheva, A. – Lara-Clares, C. (eds.), *Paradigmatic relations in word formation*, Leiden-Boston, Brill, pp. 128-163.

Mambrini, F. – Passarotti, M.
2019 *Harmonizing Different Lemmatization Strategies for Building a Knowledge Base of Linguistic Resources for Latin*, in Friedrich, A. – Zeyrek, D. (eds.), *Proceedings of the 13th Linguistic Annotation Workshop (LAW XIII). August 1, 2019. Florence, Italy*, Florence, Association for Computational Linguistics, pp. 71-80.

Passarotti, M. – González Saavedra, B. – Onambele, C.
2016 *Latin Vallex. A Treebank-based Semantic Valency Lexicon for Latin*, in Calzolari, N. – Choukri, K. – Declerck, T. – Grobelnik, M. – Maegaard, B. – Mariani, J. – Moreno, A. – Odijk, J. – Piperidis, S. (eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, European Language Resources Association (ELRA), pp. 2599-2606.

Passarotti, M. – Budassi, M. – Litta, E. – Ruffolo, P.
2017 *The Lemlat 3.0 Package for Morphological Analysis of Latin*, in Bouma, G. – Adesam, Y. (eds.), *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*, Götheborg, Northern European Association for Language Technology (NEALT), Proceedings Series, Vol. 32, pp. 24-31.

Rau, J.
2009 *Indo-european nominal morphology: The decads and the Caland system*, Innsbrucker Beiträge zur Sprachwissenschaft, Bd. 132, Innsbruck, Institut für Sprachen und Literaturen der Universität Innsbruck, pp. 106–143.

TLL
2009 *Thesaurus Linguae Latinae Online*, Berlin-Boston, De Gruyter: https://www.degruyter.com/view/db/tll

Štekauer, P.
2014 *Derivational Paradigms*, in Lieber, R. – Štekauer, P. (eds.), *The Oxford Handbook of Derivational Morphology*, Oxford, Oxford University Press, pp. 354-369.

De Vaan, M.
2018 *Etymological Dictionary of Latin and the Other Italic Languages*, Leiden-Boston, Brill.