The Treatment of Word Formation in the LiLa Knowledge Base of Linguistic Resources for Latin

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

Abstract

The *LiLa* project consists in the creation of a Knowledge Base of linguistic resources for Latin based on the Linked Data framework and aimed at reaching interoperability between them. To this goal, LiLa integrates all types of annotation applied to a particular word/text into a common representation where all linguistic information conveyed by a specific linguistic resource becomes accessible. The recent inclusion in the Knowledge Base of information on word formation raised a number of theoretical and practical issues concerning its treatment and representation. This paper discusses such issues, presents how they were addressed in the project and describes a number of use-case scenarios that employ the information on word formation made available in the LiLa Knowledge Base.

1 Introduction

The increasing quantity, complexity and diversity of available linguistic resources has led, in recent times, to a growing interest in the sustainability and interoperability of (annotated) corpora, dictionaries, thesauri, lexica and Natural Language Processing (NLP) tools (Ide and Pustejovsky, 2010). This, initially, led to the creation of databases and infrastructures hosting linguistic resources, like CLARIN,[1] DARIAH,[2] META-SHARE[3] and EAGLE.[4] These initiatives collect resources and tools, which can be used and queried from a single web portal, but they do not provide real interconnection between them. In fact, in order to make linguistic resources interoperable, all types of annotations applied to a particular word/text should be integrated into a common representation that enables access to the linguistic information conveyed in a linguistic resource or produced by an NLP tool (Chiarcos, 2012, p. 162).

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

The LiLa Knowledge Base is highly lexically-based: one of its core components is an extensive list of Latin lemmas extracted from the morphological analyser for Latin Lemlat. The portion of the lexical basis of Lemlat concerning Classical and Late Latin (43,432 lemmas) was recently enhanced with information on word formation taken from the Word Formation Latin lexicon (WFL) (Litta, 2018), which was also included in the Knowledge Base. This has raised a number of theoretical and practical issues concerning the treatment and representation of word formation in LiLa. This paper discusses such issues, presents how they were addressed in the project and describes a number of use-case scenarios that make use of the information on word formation made available in the LiLa Knowledge Base.

[1] http://www.clarin.eu
[2] http://www.dariah.eu
[3] http://www.meta-share.org/
[4] http://www.eagle-network.eu
[5] https://lila-erc.eu/
[6] See Tim Berners-Lee's note at https://www.w3.org/DesignIssues/LinkedData.html.

Proceedings of the 2nd Int. Workshop on Resources and Tools for Derivational Morphology (DeriMo 2019), pages 35–43,
Prague, Czechia, 19-20 September 2019.

2 The LiLa Knowledge Base

In order to achieve interoperability between distributed resources and tools, LiLa adopts a set of Semantic Web and Linked Data standards and practices. These include ontologies that describe linguistic annotation (OLiA, Chiarcos and Sukhareva, 2015), corpus annotation (NLP Interchange Format (NIF), Hellmann et al., 2013; CoNLL-RDF, Chiarcos and Fäth, 2017) and lexical resources (Lemon, Buitelaar et al., 2011; Ontolex[7]). Furthermore, following Bird and Liberman (2001), the Resource Description Framework (RDF) (Lassila et al., 1998) is used to encode graph-based data structures to represent linguistic annotations in terms of triples: (1) a predicate-property (a relation; in graph terms: a labeled edge) that connects (2) a subject (a resource; in graph terms: a labeled node) with (3) its object (another resource, or a literal, e.g. a string). The SPARQL Protocol and RDF Query Language (SPARQL) is used to query the data recorded in the form of RDF triples (Prud'Hommeaux et al., 2008).[8]

The highly lexically-based nature of the LiLa Knowledge Base results from a simple, fundamental assumption: textual resources are made of (occurrences of) words, lexical resources describe properties of words, and NLP tools process words. Particularly, the lemma is considered the ideal interconnection between lexical resources (such as dictionaries, thesauri and lexica), annotated corpora and NLP tools that lemmatise their input text. Lemmas are canonical forms of words that are used by dictionaries to cite lexical entries, and are produced by lemmatisers to analyse tokens in corpora. For this reason, the core of the LiLa Knowledge Base is represented by the collection of Latin lemmas taken from the morphological analyser Lemlat[9] (Passarotti et al., 2017), which has proven to cover more than 98% of the textual occurrences of the word forms recorded in the comprehensive *Thesaurus formarum totius latinitatis* (TFTL, Tombeur, 1998), which is based on a corpus of texts ranging from the beginnings of Latin literature to the present, for a total of more than 60 million words (Cecchini et al., 2018). Interoperability can be achieved by linking all entries in lexical resources and corpus tokens that refer to the same lemma, thus allowing a good balance between feasibility and granularity.

Figure 1 shows a simplified representation of the fundamental architecture of LiLa, highlighting the relations between the main components and the (meta)data providers of the Knowledge Base. The components of the Knowledge Base and their relations are formalised as classes of objects in an ontology. There are two nodes representing as many kinds of linguistic resources providing data and metadata: a) **Textual Resources**: they provide texts, which are made of **Tokens** (class: *Word*, as defined by the NIF vocabulary), i.e. occurrences of word forms (class: *Form*, as defined by Ontolex)[10]; b) **Lexical Resources**: they describe lexical items, which can include references to lemmas, e.g. in a bilingual dictionary, or to word forms, e.g. in a collection of forms like TFTL. A **Lemma** (class: *Lemma*, subclass of *Form*)) is an (inflected) **Form** conventionally chosen as the citation form for a lexical item. Both tokens and forms/lemmas are assigned **Morphological Features**, like part-of-speech (PoS), inflexional category and gender. Finally, **NLP tools** such as tokenisers, PoS taggers and morphological analysers can process respectively textual resources, tokens and forms.

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

3 The Word Formation Latin Lexicon

The WFL lexicon is a resource that deals with word formation in Classical and Late Latin. The lexicon is based on a set of word formation rules (WFRs) represented as directed one-to-many input-output relations between lemmas. The lexicon was devised according to the Item-and-Arrangement (I&A) model of morphological description (Hockett, 1954): lemmas are either non-derived lexical morphemes, or a concatenation of a base in combination with affixes. This theoretical model was chosen because it emphasises the semantic significance of affixal elements, and because it had been previously adopted by

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

[8] A prototype of the LiLa triplestore is available at https://lila-erc.eu/data/.

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

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

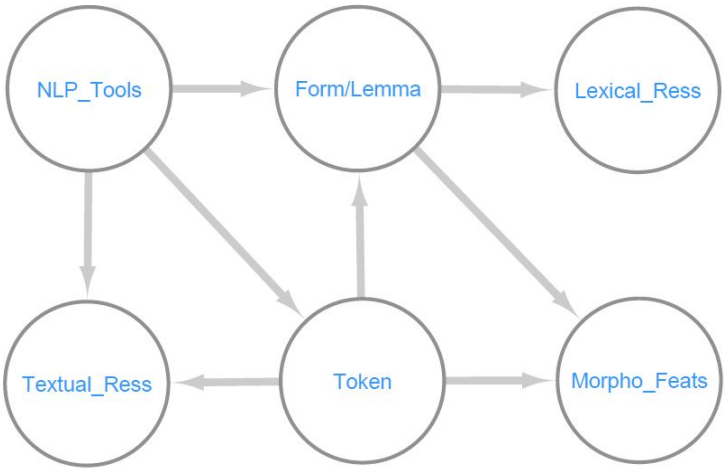

Figure 1: The fundamental architecture of LiLa.

other resources treating derivation, such as the morphological dictionaries Word Manager (Domenig and ten Hacken, 1992).

WFL is characterised by a step-by-step morphotactic approach: each word formation process is treated individually as the application of one single rule. For instance, the adjective *classiarius* 'of the fleet' is recorded in WFL as derived from the noun *classis* 'class, great division' via a WFR that creates denominal adjectives with the suffix *-ari*. In WFL, simple conversion (i.e. change of PoS without further affixation) is treated as a separate WFR, like in the case of the noun *classicum* 'trumpet-call' derived from the adjective *classicus* 'belonging to the highest class of citizens/connected with the fleet/with the trumpet call'. However, when considering formations involving both the attachment of an affix and a shift in PoS (as, for example, *classis*>*classiarius*), these are handled in one step. Each output lemma can only have one input lemma, unless the output lemma qualifies as a compound. This results in a hierarchical structure, whereby one or more lemmas derive from one ancestor lemma. A set of lemmas derived from one common ancestor is defined as a "word formation family". In the web application for querying the WFL lexicon, this hierarchical structure is represented in a directed graph resembling a tree.[11] In the graph of a word formation family, nodes are occupied by lemmas, and edges are labelled with a description of the WFR used to derive the output lemma from the input one. For instance, Figure 2 shows the derivation graph for the word formation family whose ancestor (or "root") lemma is *classis*.

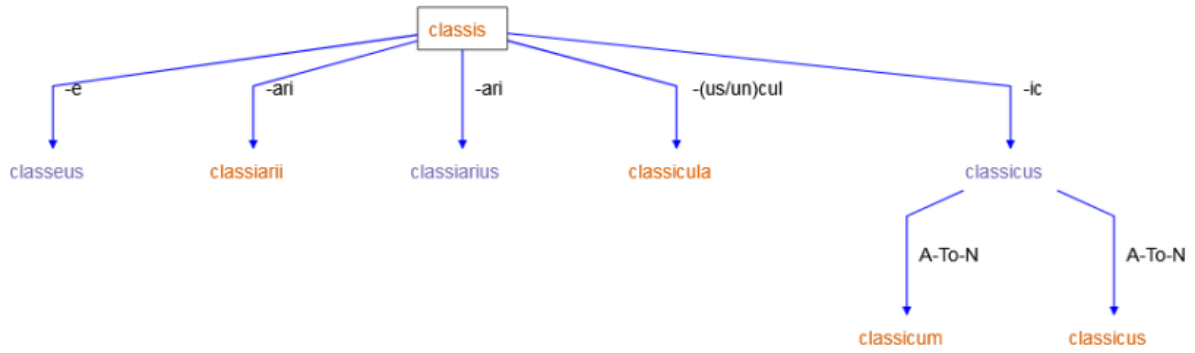

Figure 2: Derivation graph for the word formation family of *classis* in WFL.

However, portraying word formation processes via directed graphs has raised some significant theoretical issues, especially in cases where the derivational relation is ambiguous or unsuitable to be represented by a single step-by-step process, as shown in Budassi and Litta (2017). In such cases, WFL resorts

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

to a series of tactics to work around the problem. When considering, to give an example, the relation between the verb *amo* 'to love', the noun *amicus* 'friend', and the adjective amicus 'friend', did the word formation process work like *amo* > *amicus* A > *amicus* N, or like *amo* > *amicus* N > *amicus* A? In cases like this, in which there has been a conversion from noun to adjective or the reverse, there is a lot of space for interpretation on which direction the change has happened from-to, and which between noun or adjective generated the children lemmas: *Oxford Latin Dictionary* (OLD) (Glare, 1982) is usually employed in the compilation of WFL to verify the provenance of lemmas, and reports how *amo* > *amicus* A > *amicus* N is the correct process. Even so, in other occasions it has been necessary to take some independent choices: for instance, OLD states that diminutive noun *amiculus* 'a pet friend' derives from the adjective *amicus*; we, however, chose to make it derive from noun *amicus* as it seems more probable that a diminutive noun was created to diminish a noun rather than an adjective. Another method used in WFL to work around non-linear derivations is the creation of "fictional" lemmas that act as placeholders between attested words in order to justify extra "mechanical" steps. The existence of these fictional lemmas has however proven to be less than ideal. User feedback has reported confusion and puzzlement at the existence of the fictional element in the derivational tree. Moreover, when browsing the data, the existence of fictional lemmas needs to be factored in. For instance, if looking for all lemmas created with the suffix *-bil* in WFL, 598 lemmas are given as a result.[12] In WFL, 103 of these are fictional lemmas (17% of the total number of lemmas derived using the *-bil* suffix), most of which were created to connect lemmas such as adverb *imperabiliter* 'authoritatively' to their "next of kin", verb *impero* 'to demand / to order'. Because in WFL it is not possible to connect two lemmas using two suffixes at the same time (*-bil* and *-ter*), adjective **imperabilis* was created as a further step in the word formation process. The presence of fictional lemmas in the WFL dataset means that when making general considerations on the distribution of the *-bil* suffix in Classical and Late Latin, for instance, one should keep in mind that a good portion of what is extracted from WFL needs to be discarded.

4 Word Formation in *LiLa*

The recent emergence of interest in the application of Word and Paradigm (W&P) models to derivational morphology led to the exploration of their potential in describing those processes that do not fit into a linear hierarchical structure. In particular, the theoretical framework of the word-(and sign)-based model known as Construction Morphology (CxM) (Booij, 2010), has been crucial for including the WFL data into the LiLa Knowledge Base.[13] CxM revolves around the central notion of "constructions", conventionalised pairings of form and meaning (Booij, 2010, p. 6). For example, the English noun *walker* is analysed in its internal structure as $[[\texttt{walk}]_{V}\ \texttt{er}]_{N} \longleftrightarrow [\texttt{someone who walk}_{V}]_{N}$. Constructions may be hierarchically organised and abstracted into "schemas". The following schema, for instance, describes a generalisation of the construction of all words displaying the same morphological structure as *walker*, like for instance *buyer*, *player* and *reader*: $[[\texttt{x}]_{Vi}\texttt{er}]_{Nj} \longleftrightarrow [\texttt{someone who SEM}_{Vi}]_{Nj}$.[14]

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 (e.g. V-to-N *-er*), and are purely output-oriented. This is particularly fit for the needs of LiLa, as words are described into their formative elements, which can be organised into (connected) classes of objects in an ontology.

In particular, in the ontology the LiLa Knowledge Base is based on, three classes of objects are used for the treatment of derivational morphology: (1) Lemmas, (2) Affixes, divided into Prefixes and Suffixes, and (3) Bases. Bases are currently not assigned a further description, and play the role of connectors of the lemmas belonging to the same word formation family. Like any object in LiLa, Affixes and Bases are assigned a unique identifier. Each Affix is labelled with a citation form chosen to represent it in the Knowledge Base, while lemmas are connected to their Written Representation(s).

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

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

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

These three classes of objects are connected to each other via labelled edges. A Lemma node is linked (a) to the Affix nodes that are part of its construction through the relationship `hasPrefix` or `hasSuffix` and (b) to its Base (or Bases, in the case of compounds) through the relationship `hasBase`. Lemmas are never related to each other, so as not to take assumptions on the direction of the formative process. Figure

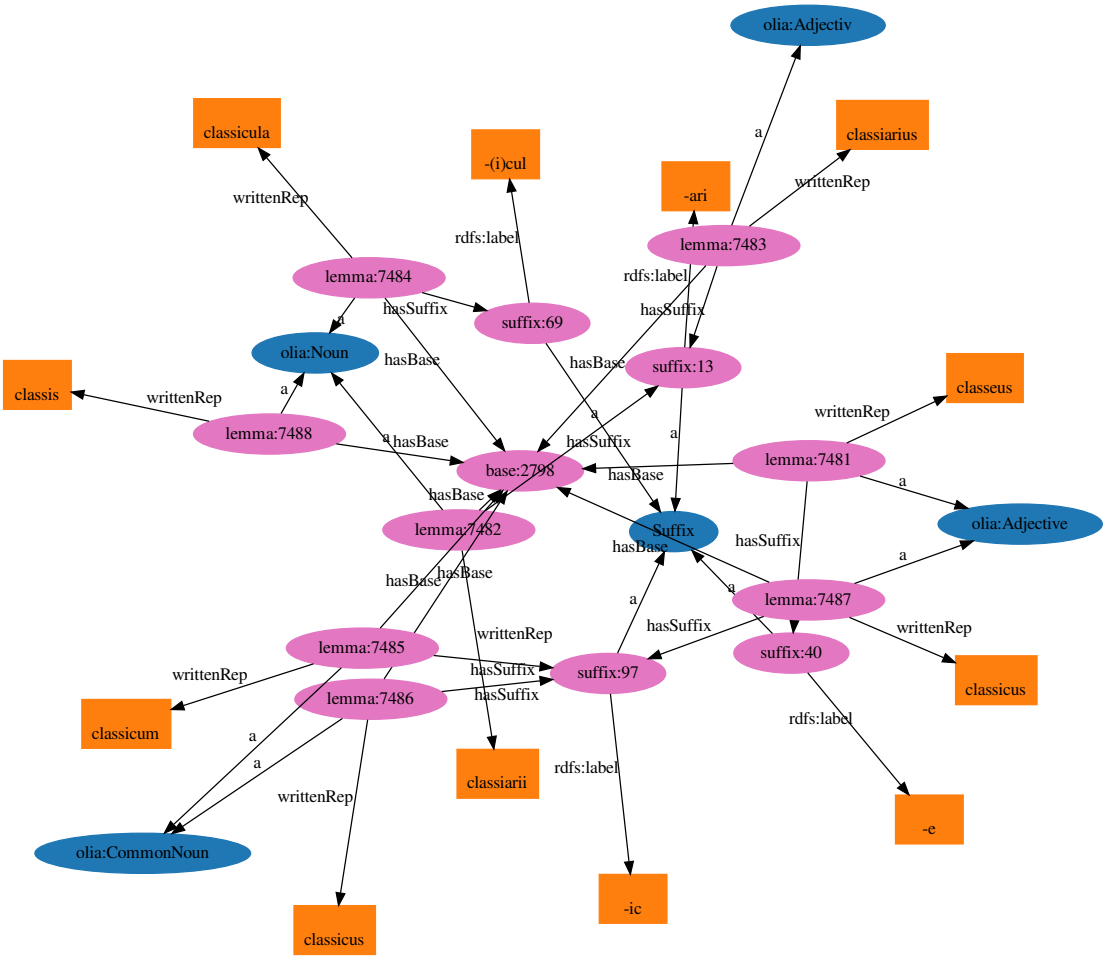

Figure 3: The word formation familiy of *classis* in LiLa.

3 shows the word formation family of *classis* as it is represented in LiLa. Nodes for Lemma objects are assigned a unique identifier and are connected to (a) their Written Representation, (b) their PoS, (c) a Base and (d) [optional] an Affix. For instance, lemma:7483 has Written Representation 'classiarius', PoS adjective (see the '(is-)a' edge connecting to the OLiA class olia:Adjective), suffix:13 (with Written representation 'ari') and base:2798. This Base node has 8 ingoing edges, one for each of the lemmas belonging to the word formation family *classis* belongs to. Conversion is not marked: lemmas such as *classicus* adjective and *classicum* noun are simply related to their Base and to the Suffix node *-ic*.

5 Use-case Scenarios

5.1 Inside Derivational Data

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

One of the most basic queries is the retrieval of all lemmas linked to the same lexical base (i.e. all the members of a word formation family) via the `hasBase` object property. The query starts by finding a

given lemma, then identifies the lexical base linked to it, and finally lists all the other lemmas connected to the same base. Starting from the adjective *formalis* 'of a form, formal', 67 lemmas are retrieved,[15] These can be grouped by PoS: 32 adjectives (including e.g. *serpentiformis* 'shaped like a snake' and *uniformis* 'uniform'), 25 nouns (e.g. *forma* 'shape', *formella* 'mould' and *informator* 'one who shapes'), 9 verbs (e.g. *informo* 'to shape', 'to inform' and *reformo* 'to transform'), and 1 adverb (*ambiformiter* 'with double meaning').

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

For example: with 121 instances, the most frequently associated prefixes in the LiLa lemma collection are *con-* and *in-* (with meaning of negation).[16] These two affixes are preponderantly found together in adjectives (96), such as *incommutabilis* 'unchangeable', less frequently nouns (23, e.g. *inconsequentia* 'lack of consistency') and adverbs (2, *incommote* 'immovably/firmly' and *incorribiliter* 'incorrigibly'). The association of (negative) *in-* prefix and *ex-* is however less frequent (79 lemmas); examples are for instance adjective *inefficax* 'unproductive' and noun *inexperientia* 'inexperience'.

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

The two most productive associations between a prefix and a suffix in LiLa are those between the negative *in-* prefix and the suffix *-bil* (296 lemmas, such as adjective *insuperabilis* 'that cannot be passed'), and between the prefix *con-* and the suffix *-(t)io(n)*, with 290 lemmas, which are mostly nouns like *contemplatio* 'viewing/contemplation' and *reconciliatio* 're-establishing'.

5.2 Outside Derivational Data

The data on word formation stored in the LiLa Knowledge Base can also be used to perform corpus-based queries. Users can use the link between lemmatised texts and the lemmas of the LiLa collection to maximum advantage to explore which are the most frequently occurring derivational morphemes in the textual resources connected so far in LiLa. These are three Latin treebanks, namely (1) the *Index Thomisticus* Treebank (IT-TB) (Passarotti, 2011), based on works written in the XIIIth century by Thomas Aquinas (approximately 400k nodes), (2) the PROIEL corpus (Haug and Jøhndal, 2008), which includes the entire New Testament in Latin (the so called *Vulgata* by Jerome) along with other prose texts of the Classical and Late Antique period and (3) the Late Latin Charter Treebank (Korkiakangas and Passarotti, 2011) (LLCT; around 250k nodes), a syntactically annotated corpus of original VIIIth-IXth century charters from Central Italy. Both the IT-TB and the PROIEL treebanks were queried in their Universal Dependencies (UD) version (Nivre et al., 2016).[17]

For instance, if we are looking for statistics on the incidence of verbs formed with prefixes *de-* and *ex-* in Latin texts, we can design a query to observe the distribution of the forms of such verbs in the corpora linked to the LiLa Knowledge Base. The results are shown in Table 1, where we report both the number of occurrences of any given verb formed with the two prefixes (Tokens), and of the different verbs attested (Lemmas).

The LiLa Knowledge Base can also be used to answer such questions as: what are the most frequent affixes in Latin texts? For instance, the use of prefixes and suffixes in the lexicon of the PROIEL corpus, the most balanced Latin treebank in terms of textual genres, can be observed with a SPARQL query that retrieves all tokens and all affixes linked with a LiLa lemma. The results are reported in Table 2. It can

[15]The starting word *formalis* is included in the count.
[16]In Latin there are two prefixes *in-*, respectively with negative and entering meaning.
[17]http://universaldependencies.org/

	de-		ex-	
Corpus	Tokens	Lemmas	Tokens	Lemmas
IT-TB (UD)	1,274	59	1,326	76
PROIEL (UD)	1,011	128	1,328	152
LLCT	209	28	155	16

Table 1: Occurrences of verbs formed with the prefixes *de-* and *ex-* in the corpora linked to LiLa.

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

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

be noted that, while tokens of words derived with the suffix *-(t)io(n)* rank only in the fourth place and are considerably outnumbered by tokens formed with the prefix *con-*, the lemmas displaying the suffix *-(t)io(n)* outnumber all the others. Such distribution reflects the greater productivity of this suffix as recorded in WFL: 2,686 lemmas formed with *-(t)io(n)* vs. 748 with *con-*.

6 Conclusions

In this paper, we have described the treatment of word formation in the LiLa Knowledge Base, which links together distributed linguistic resources for Latin.

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

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

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme - Grant Agreement No 769994.

References

Steven Bird and Mark Liberman. 2001. A formal framework for linguistic annotation. *Speech communication* 33(1-2):23–60.

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

Marco Budassi and 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)*. Educatt, pages 15–26.

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

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

Christian Chiarcos. 2012. Interoperability of corpora and annotations. In *Linked Data in Linguistics*, Springer, pages 161–179.

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 International Publishing, Cham, pages 74–88. https://link.springer.com/content/pdf/10.1007%2F978-3-319-59888-8_6.pdf.

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

Thierry Declerck, Piroska Lendvai, Karlheinz Mörth, Gerhard Budin, and Tamás Váradi. 2012. Towards linked language data for digital humanities. In *Linked Data in Linguistics*, Springer, pages 109–116.

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

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

Dag TT Haug and Marius Jøhndal. 2008. Creating a parallel treebank of the old Indo-European Bible translations. In *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*. European Language Resources Association (ELRA), Marrakesh, Morocco, pages 27–34.

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

Charles F. Hockett. 1954. Two Models of Grammatical Description. *Words* 10:210–231.

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

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

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

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

Eleonora Litta. 2018. Morphology Beyond Inflection. Building a Word Formation-Based Lexicon for Latin. In Paola Cotticelli-Kurras and Federico Giusfredi, editors, *Formal Representation and the Digital Humanities*. Cambridge Scholars Publishing, Newcastle upon Tyne, pages 97–114.

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

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

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA), Portorož, Slovenia, pages 1659–1666.

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

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*. Linköping University Electronic Press, 133, pages 24–31.

Eric Prud'Hommeaux, Andy Seaborne, et al. 2008. Sparql query language for rdf. w3c. *Internet: https://www.w3.org/TR/rdf-sparql-query/[Accessed on February 27th, 2019]* .

Maciej Sumalvico. 2017. Unsupervised Learning of Morphology with Graph Sampling. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2017)*. Varna, Bulgaria.

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