Of Families and Occurrences. Derivation and Word Usage in Latin

Marco Passarotti, Eleonora Litta
Università Cattolica del Sacro Cuore / Milano
marco.passarotti@unicatt.it
eleonoramaria.litta@unicatt.it

Abstract

In this paper we present the results of an investigation on the relation between derivational morphology, represented in terms of derivational families from a word formation lexicon for Latin, and the number of textual occurrences of their members in a large set of Latin corpora made interoperable in a Linked Data Knowledge Base.

1 Introduction

The current availability of several linguistic resources for the Latin language has raised the issue of their dispersion, which affects the full exploitation of the (meta)data they provide. This means that, even when they are published in common repositories or infrastructures (like, for instance, CLARIN),[1] resources still stay confined in separate silos that do not communicate with each other.

This situation impacts negatively on the use of data, because it prevents scholars from running federated queries across different resources, although this is a typical need when linguistic (meta)data are concerned. Particularly, this is the case when Classical and ancient languages are concerned, as scholars for centuries have been joining information from texts in different collections, as well as from lexical resources like dictionaries and glossaries.

To address the issue of dispersion and lack of interaction among the available linguistic data for Latin, the *LiLa: Linking Latin* ERC project (2018-2023)[2] has built a Knowledge Base of interoperable lexical and textual resources for Latin based on the principles of the Linked Data paradigm (Berners-Lee et al., 2001), by representing and publishing the (meta)data from these resources using common vocabularies (provided by ontologies) for knowledge description.

The resources currently made interoperable by the LiLa Knowledge Base include several corpora, which cover a wide chronological and typological span of Latin texts, and a number of lexical resources, like a bilingual dictionary, an etymological lexicon and a polarity lexicon.[3] Among the lexical resources published in LiLa is *Word Formation Latin*, a derivational lexicon for Latin where derived words are assigned a word formation rule and a link to the lexical item (or items, in the case of compounds) from which they are derived. The interoperability between the derivational information provided by *Word Formation Latin* and the (meta)data of all the other resources published in LiLa makes it possible to collect lexical information and textual evidence to empirically test hypotheses (or assumptions) about the relation between word formation processes in the lexicon and the use of derived words in texts.

In this paper, we want to address and evaluate empirically the hypothesis that, given a derivational family (i.e., a set of words sharing the same ancestor, henceforth the 'root'), the member with the highest number of occurrences in texts is derivationally simple (i.e., not featuring any affix), the root of the family being the most typical case. After providing the quantitative results taken from the corpora linked to the LiLa Knowledge Base, we focus on some cases that exceed the prototype, i.e., those where the root of a family is not also the most frequently attested member in texts. Finally, we compare the distribution of

[1] https://www.clarin.eu
[2] https://lila-erc.eu
[3] For the full list of the resources currently linked to LiLa see https://lila-erc.eu/data-page/.

Proceedings of the 4th Int. Workshop on Resources and Tools for Derivational Morphology (DeriMo 2023), pages 81–87, Dubrovnik, Croatia, 5-6 October 2023.

the root and the most frequent members of the derivational families common to two different corpora, showing that this helps to highlight some lexical properties of the texts included in the corpora concerned.

2 LiLa and Word Formation Latin

To make distributed linguistic resources interact following the Linked Data principles, the LiLa Knowledge Base adopts the data model of the Resource Description Framework (Lassila et al., 1998) (RDF). According to RDF, information is coded in terms of triples, that connect a subject – a labelled node – to an object – another labelled node or a literal – by means of a property – a labelled edge. The vocabulary used for properties and the criteria for their application are provided by a set of ontologies developed and widely adopted by the Linguistic Linked Open Data community and adopted in LiLa 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).

LiLa connects resources building on the intuition that words play a central role in both lexical and textual resources, and that through words these can be interlinked and interact. Following this intuition, the core of the Knowledge Base is its Lemma Bank, an ever growing collection of around 215,000 canonical citation forms of Latin words. Through the Lemma Bank, the entries of the various lexical resources published in LiLa and the word occurrences in the corpora included therein are linked to their appropriate citation form in the Lemma Bank, thus achieving interoperability (Passarotti et al., 2020).

Word Formation Latin (WFL) is a derivational lexicon for Classical Latin that includes 41,977 entries connected by input-output relations, grouping all members of the same derivational family in a hierarchical structure 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 word formation rules (Litta et al., 2020).

In building the LiLa Lemma Bank, derivational data were extracted from WFL to describe the word formation construction of the WFL entries in a flat way, i.e. without inferences on their derivational history, but only with details about the presence of affixes and affiliation to a derivational family. In the LiLa ontology, this information was encoded in two classes (sub-classes of the class `Morpheme`), namely `Affix` – divided into `Prefix` and `Suffix` – and `Base`. Bases are abstract connectors between lemmas that belong to the same family. These connectors are labelled with the lemma of the root word of the family concerned. In the Lemma Bank, a lemma is linked to the base to which it is related by means of the property `lila:hasBase`, and to the affixes it contains by means of the property `lila:hasPrefix` or `lila:hasSuffix` (Litta et al., 2019). Hence for example lemma *aduersaria* is connected through the property hasPrefix to the prefix *ad-*, through the property hasSuffix to suffix *-ari*, and through the property hasBase to the connector node labelled *uerto*. WFL was subsequently linked as a lexical resource to the LiLa Knowledge Base, in order to preserve precious data about more detailed, hierarchical information on the order of application of different word formation processes (Pellegrini et al., 2021).

3 Data and Discussion

Among the 4,769 derivational families provided by WFL, we select those that feature at least 10 members (1,086 families), which means that in the LiLa Knowledge Base the individual representing the base that connects all the members of a family has an in-degree via the property `lila:hasBase` ≥ 10. Also, among such families, we select those where the total number of occurrences of the members in all the textual resources currently linked to the LiLa Knowledge Base[4] is ≥ 100, leading to 878 families.[5]

3.1 Derivation, Frequency and Lexicalisation

In 582 out of the 878 families under investigation, the root member is also the most frequent one in the corpora linked to the LiLa Knowledge Base (e.g., *pono* 'to put'), while this is not the case for the

[4]The textual resources in LiLa contain more than 3 million occurrences in Latin texts of different period (from Classical era to Medieval times) and genre (including literary, documentary, historical and philosophical texts).

[5]The script providing the SPARQL queries that we used to collect the data described in this Section is available at https://github.com/CIRCSE/DevAttFreq.

remaining 296 families (e.g., *accipio* 'to accept' is the most frequently encountered word from the family rooted by *capio* 'to take'). In 89 out of these 296 families, the most frequent member is derivationally simple, i.e., it does not include any affix (either prefix, or suffix), as it formed by a conversion process, like in the case of *cursus* 'course', converted from the base of the supine of the root verb *curro* 'to run'). Given these figures, we can confirm that in most families (671 out of 878) the member with the highest number of textual occurrences is derivationally simple and, very often (582 out of 671), it is also the root word. Conversely, in 207 (296 - 89) out of the 878 families concerned, the most frequently attested member is a derivationally complex word, formed with one, or more affixes.

Table 1 shows the 10 most attested affixes in the most frequent derived words of a family. For instance, the prefix *con-* appears in the most frequent word of 25 families, like in the case of the verb *cognosco* 'to know', which is the most frequent word (2,543 occurrences in the LiLa corpora) of the family whose root is the derivationally simple verb *nosco* 'to know' (1,497 occurrences).

Moreover, Table 1 shows the ranking of each of the 10 affixes in the LiLa Lemma Bank, resulting from the number of lemmas therein formed with that suffix. For instance, the prefix *con-* is present in 2,204 entries of the Lemma Bank, which makes it the third most attested affix in the Bank. The difference between the lexical ranking of an affix (i.e., the number of lemmas in the LiLa Bank formed with that affix) and its textual ranking (i.e., the number of occurrences of the lemmas formed with that affix in the LiLa corpora) is remarkably positive as for the suffixes *-i* (from 11 to 2), *-id* (from 36 to 3) and *-in* (from 19 to 5), and negative as for *-(t)io* (from 1 to 6). As for the latter, this means that, although in the Latin lexicon the number of available derived words featuring the suffix *-(t)io* (3,418) is higher than for any other affixed word, the number of families whose most frequent member features the suffix *-(t)io* is quite low (8). The opposite holds, for instance, for suffix *-i*: although the number of words in the Lemma Bank formed with this suffix is much lower (1,323) than those featuring *-(t)io*, 22 out of them are the most frequently occurring member in as many families, against only 8 formed with *-(t)io*.

Table 2 shows the 5 words with the highest frequency in the *-i* and *-(t)io* sets. The words of the *-i* set have more occurrences than those of the *-(t)io* set, which is headed by one much frequent word (*ratio*), while the others show a lower number of occurrences. We notice that some of the non-root members of a derivational family that are the most frequently attested in corpora are cases of lexicalisation.[6] According to Lehmann (2002, pp. 1-2), "grammar is concerned with those signs which are formed regularly and which are handled analytically, while the lexicon is concerned with those signs which are formed irregularly and which are handled holistically. [....] The analytic approach consists in considering each part of the object and the contribution that it makes to the assemblage by its nature and function, and thus to arrive at a mental representation of the whole by applying rules of composition to its parts. The holistic approach is to directly grasp the whole without consideration of the parts". For instance, the first sense of the noun *substantia* provided by the Oxford Latin Dictionary (Glare, 2012) is "the quality of being real". Other senses are "underlying, or essential nature", "the material of which a thing is made", "possessions" and "the basic unit of measurement (in any calculation)". Clearly, this is a case of lexicalisation, as the meaning of the word does not result from the simple composition of the semantic contribution from each of its parts, but underwent a process of shift from the original spatial semantic field to a metaphorical meaning. As for the derivation process of *substantia*, the Oxford Latin Dictionary reports "substo+ia". The verb *substo* means "to hold one's ground", still pertaining to the spatial semantic field that the lexicalisation process has made *substantia* loose.

Some interesting insights come from comparing the distribution of the parts of speech (PoS) of the root words with those of the most frequent words of the families.[7] If we focus on adjectives, common nouns and verbs only (as the PoS with most words here concerned), Table 3 shows that adjectives and verbs are the root of a family more often than playing the role of the most frequent word. The opposite holds when common nouns are concerned: while the root word is a common noun in 364 families, the most frequently attested word of a family is a common noun in 415 cases. The great majority of these are shifts from adjectives or verbs as the root word to common nouns as the most frequent word in texts,

[6] According to Lehmann (2002, pp. 1-2), lexicalisation is a lexical semantic process "concerned with those signs which [...] are handled holistically", which means "to directly grasp the whole without consideration of the parts".

[7] The LiLa Lemma Bank adopts the Universal PoS tagset (Petrov et al., 2012).

Affix	Number of families	Lemma Bank ranking	Example
con-	25	3	cognosco
-i	22	11	substantia
-id	11	36	frigidus
-or	11	4	calor
de-	11	9	detrimentum
ad-	10	10	accipio
-in	9	19	dominus
ex-	9	5	exsulto
in(entering)-	9	8	instruo
-(t)io	8	1	oratio

Table 1: The 10 most attested affixes in the most frequent derived words of a family.

Ranking	*-i* set	*-(t)io* set
1	consilium (2,147)	ratio (3,513)
2	gratia (2,051)	oratio (1,250)
3	substantia (1,697)	opinio (504)
4	sententia (1,606)	fornicatio (179)
5	memoria (1,039)	satisfactio (175)

Table 2: The 5 most frequent words in the *-i* and *-(t)io* sets.

often due to conversion, like in the case of the family whose root word is the verb *lugeo* 'to mourn' (frequency: 174) and whose most frequent one is the common noun *luctus* 'sorrow' (274), which is derived by conversion from the perfect participle of *lugeo*. Moreover, if we compare the distribution of adjectives, common nouns and verbs playing either the role of root or most frequent word in a family with their number in a large collection of Latin words like the LiLa Lemma Bank, we notice the importance of verbs in derivational families. Indeed, if we consider that the total number of verbs in the Lemma Bank (16,618) is much lower than nouns (80,892) and adjectives (65,006), the figures in Table 3 show that verbs are either the root or the most frequent word in a family much more often (351+291 = 642) than nouns (364+415 = 779) and adjectives (133+114 = 247).

PoS	Root	Most frequent
adjective	133	114
common noun	364	415
verb	351	291

Table 3: PoS distribution of root words and most frequent words in derivational families.

3.2 Comparing Corpora

As mentioned, the Latin corpora currently interlinked through LiLa include very diverse texts, which belong to different periods and genres. If such a diversity makes the corpora of LiLa quite a representative set of data to draw conclusions about the Latin language, merging the data from all the corpora prevents from identifying the characteristics of the lexicon of one specific corpus.

To this aim, we compare two of the largest corpora in LiLa, namely the LASLA collection of Classical Latin texts (around 1.7 million words) (Fantoli et al., 2022) and the *Index Thomisticus* Treebank (ITTB), which includes the full text of *Summa contra Gentiles*, a Medieval Latin philosophical treatise by Thomas

Aquinas, for a total of approximately 350,000 words (Passarotti, 2019).

Table 4 shows the quantitative results of the LASLA-ITTB comparison. Out of the 878 families selected, 214 are common to the LASLA and the ITTB data sets, i.e. the total number of the occurrences of their members in the two corpora is ≥ 100. 116 out of these 214 families have the same most frequent word in the two corpora, while for 98 families it is different. Focusing on the latter, we notice (1) that there are 34 families where the most frequent word is different in the data from the two corpora and, for both of them, it is not the root, and (b) that the number of cases where the most frequent word of a family is also the root in the LASLA corpus, while it is not in the ITTB, is much higher than the opposite (55 vs 9). This result is worth noticing. Indeed, in several such cases from the ITTB, the most frequent word of a family is a derivationally complex word, featuring one (e.g., *transeo* 'to cross over') or two affixes (e.g., *differentia* 'difference'). The fact that the texts of Thomas Aquinas tend to make use of derived words more than those from the LASLA corpus might be due to the specific characteristics of the lexicon found in Aquinas' works. As a matter of fact, often in the ITTB the most frequent word of a derivational family is a technical word of the philosophical terminology of Thomas Aquinas (and, overall, of Medieval Scholasticism), like *substantia*, which belongs to the family rooted by *sto* 'to stay' (the most frequent family member in the LASLA data). Other cases are *passio* 'passion' (from the family rooted by *patior* 'to bear'), *sensibilis* 'sensible' (*sentio* 'to feel') and *virtus* 'strength' (*vir* 'man'). Instead, looking at the 9 cases, where the LASLA most frequent word in a family does not correspond to its root while the opposite holds for the ITTB, we find terms related to the political area, like *rex* 'king' (from the family of *rego* 'to guide'), *gubernator* 'steersman' (*guberno* 'to steer') and *libertas* 'liberty' (*liber* 'free').

	LI-r	**LI-no-r**	**L-r I-no-r**	**L-no-r I-r**	**Total**
Same most frequent word	89	27	NA	NA	116
Different most frequent word	NA	34	55	9	98
Common families					214

Table 4: Comparing the LASLA and ITTB corpora. L=LASLA. I=ITTB. r=root.

Also in order to verify this hypothesis, we focus on the 15 families (among the 878 selected) with the highest number of members. Table 5 shows for each of them its root word and the name of the most frequently attested one in the LASLA and ITTB corpora, respectively. Finally, it is worth noticing that a quite frequent family in LASLA like that of *gero* does not reach the minimum number of occurrences (100) in the ITTB.[8]

4 Conclusion and Future Work

In this paper we presented the results of an investigation about the relation holding between derivational morphology, represented in terms of derivational families, and the number of textual occurrences of their members in a large set of Latin corpora.

In the near future, we plan to exploit the evidence that we collected in order to explore some trends. For instance, for those families where the most frequent word is not the family root but another derivationally simple word (related to the root proper by conversion), we shall investigate whether the evidence suggests that the root-derivative relation should be reversed. Indeed, the criteria followed by dictionaries to identify the order in derivations are not always consistent and, in most cases, do not take into account the frequency of use of the words in texts. As an example, for the verb *nuntio* 'to announce', the Oxford Latin Dictionary reports that it derives from the noun *nuntium* 'announcement' ("nuntium+o"), while in our textual data *nuntio* is far more frequent than *nuntium* (411 vs. 28 occurrences). However, it is clear that frequency cannot be the only criterion to identify derivational roots when conversion takes place. This is especially true for a language like Latin, that shows a wide diachronic span of more than two millennia: we must investigate the role possibly played by a chronological shift of prominence between two derivationally

[8]The case of the family of *fluo* is less surprising, because the high number of occurrences of the noun *flumen* in LASLA makes alone the family suitable for selection of the experiment described here.

Root	Most frequent in LASLA	Most frequent in ITTB
facio	facio	facio
fero	fero	differentia
capio	accipio	principium
ago	ago	actus
verto	versus	universalis
gero	gero	NA
pes	pes	impedio
lego	legio	intellectus
eo	eo	transeo
fluo	flumen	NA
pario	pario	comparo
sto	sto	substantia
mitto	mitto	praemitto
loquor	loquor	loquor
duco	duco	produco

Table 5: Most frequent word of the 15 largest families in the LASLA and ITTB corpora.

simple items candidate to be the root of a family, considering, for instance, the possibility that the original root had become obsolete in Late, or even already in Classical Latin. Such investigation would shed light also on the thin line holding between etymology and derivation.

We collected the results discussed in the paper thanks to the interoperability among the lexical and textual resources for Latin made possible by the LiLa Knowledge Base, showing how making the (meta)data provided by different linguistic resources interact is helpful and much needed. This is particularly relevant when an ancient language is concerned, because the absence of native speakers requires any investigation on the lexicon to be grounded on a steady confrontation with the evidence provided by texts, as the only still surviving witnesses of the real use of the words of a dead language. However, having the possibility to make the wealth of lexical and textual data from several available resources interact would prove helpful also for living languages. In such respect, it is necessary that, in the near future, the research community makes an effort towards making real (and effective) the interoperability between distributed digital resources for as many languages as possible, with the outlook of making all their data finally interact in multi-lingual fashion. To this aim, Latin might play the important role of connection among, at least, romance languages.

5 Credits

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.

References

Tim Berners-Lee, James Hendler, and Ora Lassila. 2001. The semantic web. *Scientific american* 284(5):34–43.

Paul Buitelaar, Philipp Cimiano, John McCrae, Elena Montiel-Ponsoda, and Thierry Declerck. 2011. Ontology Lexicalisation: The *lemon* Perspective. In *9th International Conference on Terminology and Artificial Intelligence (TIA11) – Proceedings of the Workshops*. Paris, France, pages 33–36. https://pub.uni-bielefeld.de/record/2486962.

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. https://www.semantic-web-journal.net/system/files/swj518$_0$.*pdf*.

Margherita Fantoli, Marco Passarotti, Francesco Mambrini, Giovanni Moretti, and Paolo Ruffolo. 2022. Linking the lasla corpus in the lila knowledge base of interoperable linguistic resources for latin. In *Proceedings of the Linked Data in Linguistics Workshop@ LREC2022*. pages 26–34.

Peter Geoffrey William Glare. 2012. *Oxford Latin Dictionary*. Oxford Languages. Oxford University Press, Oxford, UK, 2 edition. https://global.oup.com/academic/product/oxford-latin-dictionary-9780199580316?cc=uslang=en.

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. Also see http://persistence.uni-leipzig.org/nlp2rdf/.

Ora Lassila, Ralph R. Swick, World Wide, and Web Consortium. 1998. Resource Description Framework (RDF) Model and Syntax Specification. https://www.w3.org/TR/1999/REC-rdf-syntax-19990222/.

Christian Lehmann. 2002. New reflections on grammaticalization and lexicalization. *New reflections on grammaticalization* pages 1–18.

Eleonora Litta, Marco Passarotti, Marco Budassi, and Marco Pappalepore. 2020. Of nodes and cells. two perspectives on (and from) word formation latin. *Lingue antiche e moderne* 9:131–155.

Eleonora Litta, Marco Passarotti, and Francesco Mambrini. 2019. The treatment of word formation in the LiLa knowledge base of linguistic resources for Latin. In *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology*. Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, Prague, Czechia, pages 35–43. https://www.aclweb.org/anthology/W19-8505.

John P McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. The OntoLex-Lemon Model: Development and Applications. In *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference*. Lexical Computing CZ s.r.o., Brno, Czech Republic, pages 587–597. https://elex.link/elex2017/wp-content/uploads/2017/09/paper36.pdf.

Marco Passarotti. 2019. *The Project of the Index Thomisticus Treebank*, De Gruyter Saur, Berlin, Germany; Boston, MA, USA, pages 299–320. Number 10 in Age of Access? Grundfragen der Informationsgesellschaft. https://doi.org/10.1515/9783110599572-017.

Marco Passarotti, Francesco Mambrini, Greta Franzini, Flavio Massimiliano Cecchini, Eleonora Litta, Giovanni Moretti, Paolo Ruffolo, and Rachele Sprugnoli. 2020. Interlinking through Lemmas. The Lexical Collection of the LiLa Knowledge Base of Linguistic Resources for Latin. *Studi e Saggi Linguistici* LVIII(1):177–212. https://doi.org/10.4454/ssl.v58i1.277.

Matteo Pellegrini, Eleonora Litta, Marco Passarotti, Francesco Mambrini, and Giovanni Moretti. 2021. The Two Approaches to Word Formation in the LiLa Knowledge Base of Latin Resources. In *Proceedings of the Third International Workshop on Resources and Tools for Derivational Morphology (DeriMo 2021)*. ATILF, Nancy, France, pages 101–109. https://doi.org/10.5281/zenodo.5532501.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A Universal Part-of-Speech Tagset. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA), Istanbul, Turkey, pages 2089–2096. http://www.lrec-conf.org/proceedings/lrec2012/pdf/274$_{P}$*aper.pdf*.