E pluribus unum.
Representing compounding in a derivational lexicon of Latin

Silvia Micheli
Università degli Studi di Pavia
Corso Strada Nuova, 75
27100 Pavia
silvia.micheli@unibg.it

Eleonora Litta
Università Cattolica del Sacro Cuore
Largo Gemelli, 1
20123 Milano
e.littamodignani@gmail.com

Abstract

English. This paper describes how compounding is treated in the *Word Formation Latin* derivational lexicon. Through the analysis of some types of Latin compounds, perspectives and limitations of the resource are highlighted; its contribution to theoretical and computational linguistic issues is also outlined.

Italiano. *Questo contributo descrive come viene trattata la composizione nel lessico derivazionale* Word Formation Latin. *Attraverso l'analisi di alcuni aspetti della composizione latina, vengono messi in luce potenzialità e limiti della risorsa e delineato il suo contributo in campo teorico e computazionale.*

1 Introduction: the *Word Formation Latin* lexicon

Word Formation Latin (WFL, (Litta et al., 2016)) is a derivational morphology resource for Latin where words are analysed in their formative components and related to each other on the basis of word formation rules (WFRs).[1] It represents a wide lexical resource not only for the study of Latin derivational morphology (i.e. affixal and conversive processes), but also for compounding, which has often been neglected in other most recent resources for other languages.[2] The lexical basis behind WFL is the same as the morphological analyser and lemmatiser for Latin Lemlat (Passarotti et al., 2017). All lemmas have been collected from three main Classical Latin dictionaries ((Georges and Georges, 1913-1918); (Glare, 1982); (Gradenwitz, 1904)) plus the Onomasticon of Forcellini's (Forcellini, 1940) 5th edition of *Lexicon Totius Latinitatis* (Budassi and Passarotti, 2016). All those lemmas that share a common (not derived) ancestor belong to the same "morphological family", (Litta et al., 2016) represented in the web application (http://wfl.marginalia.it/) as a tree-graph.

The aim of this paper is twofold: on the one hand, it describes how compounding is represented into the WFL derivational lexicon; on the other hand, it aims at highlighting the theoretical and computational contribution of this resource through the analysis of some aspects (i.e. WFRs, input and output lexical categories) of the Latin compounds collected in it.

2 Latin compounding

Compared to other Indo-European languages (e.g. Sanskrit or Greek), compounding in Latin is generally considered to be not very productive. According to (Grenier, 1912) and (Puccioni, 1944), most of Latin compounds are *hapax legomena* and mainly occur in poetic, religious and legal texts. Furthermore, they seem to be strongly influenced by Greek models.

In the last decades, Latin compounding (henceforth LC) has received more attention ((Oniga, 1992); (Oniga, 1988); (Benedetti, 1988); (Fruyt, 2002); (Brucale, 2012)). However, most of the available studies are qualitative descriptions of compounding mechanism, which are based on a small amount of data, usually extracted from dictionaries, and cited as examples of the main types of compounds. These studies have mainly focussed on formal features of LC, which is

[1] The WFL project, still ongoing, received funding from EU Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie Individual Fellowship grant agreement No 658332-WFL

[2] Among them, notable ones are the lexical network for Czech DeriNet (Ševčíková and Žabokrtský, 2014) and (Žabokrtský and al., 2016), the derivational lexicon for German DErivBASE (Zeller et al., 2013) and that for Italian derIvaTario (Talamo et al., 2016).

essentially stem-based: Latin compounds are almost always made up of bound units (i.e. roots, stems) connected by a linking element (LE) *-i-*, as in (1).

(1) *purifico*V
pur-i-fic-o
purus+LE+*facio*+INFL
A+V+INFL=V

The nature of the linking element *-i-*,[3] the relationship between compounding and derivation in Latin, and the classification of Latin compounds, are the main theoretical topics on which attention is focused. However, there are still many questions that so far could not be answered exhaustively due to the scarcity of data collected so far: which were the most productive types of compound in Latin? Through which rules were Latin compounds formed? What PoS did Latin compounds consist of most frequently? What kinds of meaning are expressed by compounding in Latin? WFL allows to fill to answer these questions by providing a large account of quantitative data which can help to better understand the mechanisms of LC.

3 Compounding in WFL

The methodology behind WFL is consistent with the Item-and-Arrangement model outlined in (Hockett, 1954), which considers morphemes, not words, the basic units for the study of utterances, containing both form and meaning. The resource relies on a fairly strict morphotactic approach, where, to the basic component of the uninflected word, the so-called les ("LExical Segment"), one derivational morpheme (prefix/suffix) or phenomenon (conversive PoS change) is attached at a time. This means that the output of a WFR is always a lemma richer (containing more morphemes, or different inflection) than the input one.

During the compilation of WFL, an initial list of possible compounds has been drawn by taking into account all possible combinations of V (verb), N (nouns), A (adjectives), PR (pronouns), and I (invariables - e.g. adverbs). Some categories have been filled semi-automatically with the help of SQL queries. These usually matched a string that combines a certain lexical element + -i- + another lexical element or lemma (this one sometimes in the form of a customised string). This method was applicable to morphotactically transparent compounds like those verbs including *-fico* (from verb *facio* 'to make', e.g. *clarifico* 'to make illustrious'), or those adjectives featuring noun *pes* 'foot' as a second constituent (e.g. celer-i-pes, lit. 'fast foot'). However, morphotactically obscure compounds like *fidicina* 'lyre player' (fides 'lyre' + cano 'to sing'), needed to be inserted manually.

The WFL web application allows compounds to be browsed in three ways:

1. By WFR - opens research questions on a specific word formation behaviour; for example, it is possible to view and download a list of all adjectives formed by a A+V=A rule.
2. By PoS - useful for studies on macro-categories, it allows for deeper refinement of constituent PoS.
3. By Lemma - allows for quick search of a specific lemma.

For each compound, a derivational tree-graph is provided (as in Figure 1). In each graph, nodes are lemmas, and edges are relations showing the kind of WFR involved. Special provisions are made in order to collapse and hide compounding relations according to the user's choice. This is useful when very productive constituents are displayed in massive multi-tree graphs.

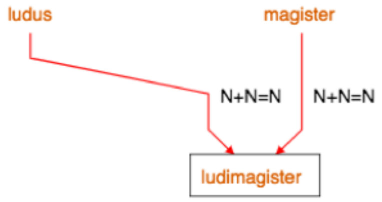

Figure 1: Derivation graph of *ludimagister*

The sample collected from the WFL lexical basis consists of 1744 compounds. The fact that all compounds collected from the three dictionaries mentioned above are for the first time categorised and labelled into a language resource allows for a more in-depth overview and for a quantitative analysis on many aspects of LC (e.g. productivity, WFRs, lexical categories involved in compounding). In the following sections, some preliminary

[3] A survey of the literature on the nature of this linking element is in (Brucale, 2012).

considerations on the data currently included in WFL are provided.

3.1 Word Formation Rules

Compound words collected in WFL are created through 59 WFRs. In table 1, the first twenty most productive WFRs are shown.[4]

	WFRs	Compounds
1	N+V=A	429
2	N+V=N	239
3	N+N=N	135
4	A+V=A	134
5	A+N=A	131
6	N+N=A	120
7	V+V=V	64
8	A+V=V	59
9	N+V=V	56
10	A+N=N	35
11	V+V=A	33
12	A+V=N	32
13	V+N=A	28
14	I+I=I	27
15	A+A=A	22
16	PR+PR=PR	15
17	I+N=N	15
18	N+A=A	14
19	N+A=N	13
20	PR+V=PR	13

Table 1: Compounding WFRs in WFL

The most productive pattern in LC is Noun+Verb: the rule creates both adjectives and nouns, e.g. *soporifer* 'soporific' (*sopor+fero*) or *artifex* 'artisan'(*ars+facio*). This word formation process is no longer productive in Romance Languages, in which the reverse order (i.e. the Verb+Noun pattern, e.g. Italian *portafoglio* 'wallet' or French *porte-parole* 'spokesman') is the most frequent.

In almost all cases, Latin compounds are made up of two constituents. There are only very few (and not productive) cases in which there are three elements, e.g. *turpilucricupidus* (turpis 'vile' + lucrum 'gain' + cupidus 'desirous'; WFR: A+N+N=N) or *suovetaurilia* (sus 'pig' + ovis 'sheep' + taurus 'bull'; WFR: N+N+N=N).

The V+V pattern, that in Italian creates nouns (e.g. *dormiveglia* 'half-sleep', lit. 'to sleep-to stay awake'), in Latin forms mainly new verbs, such as *patefacio* 'to reveal' (*pateo* 'to be evident' + *facio* 'to do').

In addiction to other patterns already identified as productive in previous literature (i.e. A+N=A, N+N=N, N+N=A), it is interesting to notice the presence of a significant number of compounds consisting of two invariable forms (e.g. *etiamtum*, *etiam+tum* 'even then, yet') or two pronouns (e.g. *aliquis*, *alis+quis* 'anyone, someone') which are generally neglected in studies on Latin word-formation.

3.2 Input and output lexical categories

As already pointed out by (Brucale, 2012), verbs and nouns are the most frequent input elements in Latin compounds. While nouns can be found both in first and in second constituent, verbs show a clearer tendency to appear in second position. Data collected in WFL confirms these observations.[5]

Lexical cat.	1° const.	2° const.	Output
A	428	69	942
I	96	55	63
N	1008	491	491
PR	63	32	53
V	141	1089	187

Table 2: Input and output lexical categories in WFL compounds

Table 2 shows the quantitative distribution of the lexical categories (i.e. how many times adjectives are present as the input or as the output PoS) in WFL compounds. More than half of the sample (i.e. 1089 forms, 62.7%) has a verbal second element (e.g. compounds with *-facio* or a related stem, such as *aedifico* 'to build' or *candefacio* 'to whitewash').

As far as the output of whole compounds are concerned, it is worth noticing that LC creates mostly adjectives (e.g. compounds with *-fer* as second constituent, such as *alifer* 'winged'), followed by nouns and verbs. Conversely, in Romance languages, compounding is exploited to create primarily nouns and less frequently adjectives. In Italian, there are very few cases of verbs obtained through compounding, which are made up of a noun and a verb (e.g. *manomettere* 'to tamper

[4] N: noun; V: verb; A: adjective; I: invariable form (i.e. adverb, conjunction); PR: pronoun.

[5] However, as reported below in section 3.3, in order to interpret correctly the data in Table 2, a distinction should be made between adjectives and adjectival participles, which are categorised here as V.

with'); the formation of pronouns and invariable forms through compounding does not seem to be productive anymore.

3.3 Some *caveats*

The main bedrock of WFL methodology lies in its strict relation to the morphological analyser Lemlat and on the PoS categorisation dictated by its lexical basis. As a consequence, the way compound constituents are pigeonholed can sometimes be unconventional. This impacts the representation of compounds in WFL in the following ways:

1. Adjectives that derive or function like participles are not included in the Lemlat lexical basis, because they are seen as part of the verbal paradigm, this means that certain compounds that would be expected to have a A as one of their constituents have a V instead. e.g *altivolans* (altus + volo) 'high flying' can be found among V+V=A compounds rather than among A+V=A.
2. certain type of adverbs ending in *-e* are considered in Lemlat ablative cases of the adjectival declension, so *dulciloquus* (dulce + loquor) 'sweet talking' is to be found among A+V=A, rather than I+V=A.

Another principle lying behind WFL's methodology is that *Oxford Latin Dictionary* acts like a sort of manual for solving a number of theoretical issues. For instance, unlike some traditional studies on Latin word-formation (i.e. (Benedetti, 1988), (Fruyt, 2002) and (Fruyt, 2011)), prepositions (e.g. *cum* 'with' or *in* 'in') are not included among compounding input elements in WFL, due to the overlap with prefixes. However, this can lead to inconsistencies. For instance, in OLD there is a clear distinction between affixes and isolated words where the lemmas' formative elements are specified. This means that words including what OLD considers a prefix, such as *quadriennium* 'period of four years' (*quadri-* 'consisting of four of the things following' + *annus* 'year', and not *quatuor* 'four' + *annus*) are included among prefixes, while other similar lemmas formed by numerals, like *sexennium* 'period of six years', on the other hand, are labelled as N+N=N compounds, because OLD categorises *sex* 'six' as a noun.

Moreover, in certain cases, it was decided to treat certain lemmas, which are generally seen as compounds, as conversions instead. For example, A+V=V and N+V=V compounds ending in *-fico*, i.e. involving the verb*facio* 'to do', which have often a corresponding adjective ending in *-ficus*. The assumption here is that the verbal compound must have been born before the adjective, as the main meaning of such compounds is almost always the result of a performed action (*amplifico* = amplus facio, 'to make (something) bigger'). In WFL, the corresponding adjective *amplificus* 'magnificent', has been connected to 'amplifico' through a conversion relationship V-to-A. This allows the two lemmas to appear in the same derivational tree.

4 Conclusions and future work

This paper has provided an overview of how compounding is represented in WFL, a derivational lexicon for Latin. This preliminary study, with its quantitative analysis in the field of LC, shows the potential for raising new questions and issues offered by a resource that for the first time collects all compounds used in Classical Latin. For instance, representing all compounding rules into a network, as it has been already successfully done for the affixal rules listed in WFL, (Litta et al., 2017), could lead to further research questions. These could be the investigation on constituent typologies or on the productivity of the different types of compounds. Future developments in WFL should be a way of searching through constituents by original lemma (currently still missing), and implementing a way of marking those PoS that appear differently in the resource's lexical basis. This would also allow for a more precise quantitative investigation on constituent typologies.

References

Marina Benedetti. 1988. *I composti radicali latini. Esame storico e comparativo*. Giardini, Pisa.

Luisa Brucale. 2012. Latin compounds. *Probus*, 24: 93-117.

Marco Budassi and Marco Passarotti. 2016. Nomen omen. Enhancing the Latin Morphological Analyser Lemlat with an Onomasticon. *Proceedings of the 10^{th} SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2016)*, 90-94. Berlin: The Association for Computational Linguistics.

Egidio Forcellini. 1940. *Lexicon Totius Latinitatis / ad Aeg. Forcellini lucubratum, dein a Jos. Furlanetto emendatum et auctum; nunc demum Fr. Cor-*

radini et Jos. Perin curantibus emendatius et auctius meloremque in formam redactum adjecto altera quasi parte Onomastico totius latinitatis opera et studio ejusdem Jos, Perin. Typis Seminarii, Padova.

Michèle Fruyt. 2011. Word-Formation in Classical Latin. *A companion to the Latin language*, 157-175.

Michèle Fruyt. 2002. Constraints and productivity in Latin nominal compounding. *Transactions of the Philological Society*, 100(3): 259-287.

Karl Ernst Georges and Heinrich Georges. 1972. *Ausführliches Lateinisch-Deutsches Handworterbuch*, Hahn, Hannover.

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

Otto Gradenwitz. 1904. *Laterali Vocum Latinarum*. Hirzel, Leipzig.

Albert Grenier. 1912. *Ètude sur la formation et l'emploi des composès nominaux dans le latin archaique*. Berger-Levrault, Paris.

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

Eleonora Litta, Marco Passarotti, and Chris Culy. 2016. Formatio formosa est. Building a Word Formation Lexicon for Latin. *Proceedings of the Third Italian Conference on Computational Linguistics (CLiC–it 2016). Napoli, aAccademia University Press*, 185-189.

Eleonora Litta, Marco Passarotti and Paolo Ruffolo. 2017. Node Formation. Using Networks to Inspect Productivity in Affixal Derivation in Classical Latin. In *Proceedings of DATeCH2017, Göttingen, Germany, June 01-02, 2017*, 6 pages. DOI: http://dx.doi.org/10.1145/3078081.3078092

Renato Oniga. 1992. Compounding in Latin. *Rivista di linguistica*, 4(1): 97-116.

Renato Oniga. 1988. *I composti nominali latini: una morfologia generativa*. Patron, Bologna.

Marco Passarotti, Marco Budassi, Eleonora Litta, and Paolo Ruffolo. 2017. The Lemlat 3.0 Package for Morphological Analysis of Latin. *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*, 133, 24-31.

Giulio Puccioni. 1944. L'uso stilistico dei composti nominali latini. *Atti della Accademia d'Italia. Memorie della classe di scienze morali e storiche*, Series 7, 4(10): 372-481.

Magda Ševčíková and Zdeněk Žabokrtský. 2014. Word-Formation Network for Czech. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014), Reykjavik, Iceland*: 1087-1093, ELRA.

Luigi Talamo, Chiara Celata and Pier Marco Bertinetto. 2016. Derivatario: an annotated lexicon of Italian derivatives. *Word Structures*, 9(1): 72-102.

Zdeněk Žabokrtský, Magda Ševčíková, Milan Straka, Jonáš Vidra and Adéla Limburská. 2016. Merging Data Resources for Inflectional and Derivational Morphology in Czech, In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*: 1307-1314, ELRA.

Britta D. Zeller, Jan Snajder, and Sebastian Padò. 2013. DErivBase: Inducing and Evaluating a Derivational Morphology Resource for German. *ACL*, 1: 1201-1211.