What is Word Formation Latin (WFL)?

Word Formation Latin (WFL) is a language resource for Classical Latin that connects lexical items on the basis of word-formation rules (WFRs). The scope of WFL is to assign a WFR to each morphologically-complex lexeme (i.e. one word morphologically derived from another word) and to link each complex lexeme to its ancestor. All those lexemes that share a common (not derived) ancestor belong to the same "word formation family". For instance, the noun _bellatrix_ 'she who wages war', the verb _rebello_ 'to revolt, rebel', and the adjective _bellicosus_ 'fond of war' all belong to the word formation family whose ancestor is noun _bellum_ 'war'. The semi-automatic insertion of lemmas into the WFL database establishes input-output relations for a set of lemmas matching the features that characterise each WFR.

WFL has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 658332-WFL. The project is based at the Centro Interdisciplinare di Ricerche per la Computerizzazione dei Segni dell'Espressione (CIRCSE), at the Università Cattolica del Sacro Cuore, Milan, Italy. The project ran from November 2015 to the end of October 2017, and resulted in the publication of the word formation based lexicon, which is accessible digitally through this website and in connection to the morphological analyser and lemmatiser for Latin Lemlat.

This documentation collects all information regarding how the WFL lexicon was built, together with instructions on how to navigate its site and how and where to find the database for your own personal research.

Lexical Basis

The lexical basis for WFL is the same as that of the morphological analyser and lemmatiser for Latin Lemlat which has been collated from three Classical Latin Dictionaries: *Oxford Latin Dictionary* (Glare 1982); *Ausführliches lateinisch-deutsches Handwörterbuch* (Georges and Georges 1913-18); *Laterculi vocum latinarum* (Gradenwitz 1904). It contains 40,014 lexical entries and 43,432 lemmas (as more than one lemma can be part of the same lexical entry). Additionally, the lexical basis of Lemlat has recently been enriched with the integration of most of the Onomasticon (26,250 lexemes out of 28,178) contained in the Forcellini lexicon (Budassi and Passarotti 2016).

Lemlat contains every string of characters required in the inflectional paradigm of each lexeme, like the uninflected parts of irregular supines (duc-, duct- for duco ‘to lead’), or the stem of the genitive of imparisyllaba nouns and adjectives (crimen, crimin- ‘accusation’), fundamental for the automatic processing of WFRs, as well as including graphical variants, like obf-/off- in offero ‘to put oneself forward, cause to be encountered’ (Passarotti and Mambrini, 2012).

These strings of characters are used by Lemlat while morphologically analysing and lemmatising input word forms, which are automatically segmented into formative elements. Among these, the lexical element is called les (for “LExical Segment”). This is the invariable part of the inflected forms, i.e. the sequence – or one of the sequences – of characters that remains the same in the inflectional paradigm of a lexeme. The les does not necessarily match the word stem. For example, poetis the les for the lexeme poeta ‘poet’, as it is the sequence of characters that does not change in the different forms of the lexeme poeta: poet-a, poet-ae, poet-am, poet-ae, poet-arum, poet-as, poet-is. Lemlat includes a les archive, in which each les is assigned a number of inflectional features. Among these, there is a tag for the gender of the lexeme (for nouns) and a code (codles) for its inflectional category. For instance, the codles for the les poet is n1e (first declension irregular nouns) and its gender is m (masculine). In the case of irregular nouns, as for poeta, there is also a field (lem) containing information on how to recognise an irregular ending (poeta can sometimes appear with nom. sg. poetes) during the lemmatisation.

WFL makes use of the les archive together with a list of 43,432 lemmas automatically extracted from the Lemlat dataset. Both lists were added as tables to the relational database used while building WFL.

For more details about Lemlat and its lexical basis please refer to Passarotti et al. 2017 (in bibliography).

In the WFL lexical basis, there are three codes for three different kinds of lemmas:

- **B** for "basic": lemmas taken from the original Lemlat lexical basis.
- **O** for "onomastic": lemmas added from the Forcellini onomastic lexicon. These have not been used to build word formation relations, unless they are the input for a non onomastic lemma, e.g. antonius 'Anthony' > antonesco 'to behave like Anthony'.
- **F** for "fictional": fictional lemmas that have been added during the building of WFL to account for relationships that are otherwise difficult to fit in the WFL morphotactic database (see Budassi and Litta 2017). Fictional entries are marked with a preceding asterisc ('') that indicates a "reconstructed" lemma, although the lemma does not necessarily need to have ever existed, but it only acts as a trait d'union between two attested lemmas.

Methodology

For what the construction of the word formation relationships is concerned, WFL uses a step-by-step morphotactic approach: derivational and compounding rules are modelled as directed one-to-many input-output relations between lexemes, and each word formation process is treated individually. The lexeme resulting from a WFR is usually richer (containing more morphemes) than the input. with the exception of conversion, which only involves a change of PoS. Each output lexeme can only have one source, except in the case of compounds, where it is possible to have two (or three) input lexemes for one output lexeme.

Building the Lexicon

The word formation lexicon is built in two steps. First, word formation rules are detected. Then, they are applied to lexical data.

Detecting Word Formation Rules

Word formation rules (WFRs) are conceived according to the so-calledItem-and-Arrangement model, outlined by Hockett (1954), which considers word forms either as simple morphemes (not derived word forms) or as a concatenation of morphemes (derived word forms). The following conditions on bases and affixes do hold: (1) Baudoin's assumption that both bases and affixes are lexical elements (i.e. they are both morphemes); (2) as a consequence, they exist in the lexicon (Bloomfield's "lexical morpheme" theory); (3) they are dualistic, i.e. they have both form and meaning (Bloomfield's "sign-base" morpheme theory). The first two conditions motivate the fact that in our word formation lexicon affixes are recorded with the same status of lexical bases; the third condition concerns the semantic properties of WFRs.

In Latin, WFRs fall into two main types: (1) derivation and (2) compounding. Derivation rules are further organised into two subcategories: (a) affixal, in its turn split into prefixal and suffixal, and (b) conversion, a derivation process that changes the PoS of the input word without affixation.

Compounding and conversion WFRs are automatically detected, by considering all the possible combinations of main PoS (verbs, nouns, adjectives), regardless of their actual instantiations in the lexical basis. For instance, there are four possible types of conversion WFRs involving verbs: V-To-N (claudo > clausa; 'to close' > 'cell'), V-To-A (eligo > elegans; 'to pick out' > 'accustomed to select, tasteful'), N-To-V (magister > magistro; 'master' > 'to rule'), A-To-V (celer > celero; 'quick' > 'to quicken'). Each compounding and conversion WFR type is further specified

by the inflectional category of both input and output. For instance, A1-To-V1 is the conversion WFR from first class adjectives to first conjugation verbs.

Affixal WFRs are found both according to previous literature on Latin derivational morphology (Jenks, 1911; Fruyt, 2011; Oniga, 1988) and in semi-automatic fashion. The latter is performed by extracting from the list of lemmas of Lemlat the most frequent sequences of characters occurring on the left (prefixes) and on the right (suffixes) side of lemmas. The PoS for WFR input and output lemmas as well as their inflectional category are manually assigned. Further affixal WFRs are found by confrontation with data.

We recorded the rules in a table of a MySQL relational database where each WFR is classified by type and it is assigned the required PoS, inflectional category and gender for its input and output.

Applying Word Formation Rules

Each morphologically derived lemma is assigned a WFR. All those lemmas that share a common (not derived) ancestor belong to the same “word formation family”. For instance, lemmas formatio 'formation', formo 'to form' and formosus (“beautiful”, lit. “finely formed”) all belong to the word formation family whose ancestor is the lemma forma(“form”).

WFL uses a morphotactic approach. Each word formation process is treated individually, and the lexeme resulting from a WFR is usually richer (containing more morphemes) than the input. with the exception of conversion, which only involves a change of PoS. Each output lexeme can only have one source, except in the case of compounds, where it is possible to have two (or three) input lexemes for one output lexeme.

Lemmas and WFRs are paired by using a MySQL relational database whose main tables are the les archive of Lemlat, the list of its lemmas (each assigned its PoS, inflectional category and, for nouns only, gender) and the list of WFRs.

A number of MySQL queries provide the candidate lemmas for each WFR. Some of these queries run on the list of lemmas, while others on thelesarchive. In particular, most candidate lemmas of prefixal WFRs are found by running queries on the list of lemmas, as such rules tend to just add the characters of the prefix to the input lemma, like in the case ofaccuso→sub+accuso(“to blame” → “to blame somewhat”). Instead, suffixal WFRs are mostly assigned to their candidate input and output lemmas by running queries on thelesarchive, because suffixes attach tolesinstead of modifying full lemmas, like inamo→amabilis(“to love” → “lovable”) where suffix–bil–attaches tolesam(plus the thematic vowel –a–, used for first conjugation verbs) instead of full lemmaamo. Also, there are suffixal WFRs whose input is the basis of the irregular perfect participle of the input verb, like induco→ductilis(“to lead” → “that

may be led”) where suffix–il–attaches to the basis of the irregular perfect participle of the verbduco(duct). Such irregular bases are recorded explicitly in the les archive with a specific codles.

Making Choices

Homography can occur at different levels when we search for pairings. Let us consider, for example, the application of prefixal rules to the list of lemmas. An SQL query searches through all verbs contained in the list of lemmas and is instructed to return them in two main groups: an input list of lemmas (i_lemma) and an output list of lemmas (o_lemma), where the output needs to look the same as the input with the addition of a string of characters, the prefix, preceding the same string of characters as it is displayed in the input. Such a query will return something like this:[1]

input_lemma	**output_lemma**
sero 1	subsero 1
sero 1	subsero 2
sero 2	subsero 1
sero 2	subsero 2
sero 3	subsero 1
sero 3	subsero 2

There are three lemmas sero in Latin, with different inflections and meanings (1 ‘to plant’, 2 ‘to entwine’ and 3 ‘to bolt’), and two lemmas subsero (1‘to plant as a replacement’ and 2 ‘to insert below’). The simple pairing SQL query currently used does not discern which subsero comes from which sero. However, there are two ways in which this can be solved. First, we can exclude the lemma sero 3 as, unlike the other third conjugation lemmas, it is a first conjugation

verb. The codles for both les ser is v3r (third conjugation regular verb). All les related to a common lemma are grouped within the same clem (“costellazione lemmatica”) and referred to through the same identifier n_id (hereafter 1/2/3). In the sero1 clem, for example, we have the forms seu with codles v7s (perfect) and sat with n6p1(irregular perfect participles); for sero 2 we have seru for v7r and sert for n6p1; for subsero 1 we have the forms subseu with codles v7s and subsat n6p1; for subsero 2 we have subseru for v7r and subsert for n6p1.

One can either write a query that looks deeper into the clem of a lemma to match lemmas on the basis of les for v7r and/or n6p1or make the distinction manually, in this case on the basis of inflection: sero(1), sevi, satus => subsero(1), subsevi, subsatus/ sero(2), serui, sertus => subsero (2), subserui, subsertus).

At times, the disambiguation only works on meaning, as there might not be real formal differences between totally homographic lemmas (i.e. lemmas that look completely identical even in their inflectional paradigms). The only way to rectify these dubious pairs is to consult both the LEMLAT archive (for additional morphological information) and the two main reference dictionaries , Oxford Latin Dictionary and Georges and Georges.

Homography happens even more frequently when the query performs searches directly onles, as in the case of suffixal rules. The probability that onelesmight look like another, even if the resulting lemma does not, becomes higher.

For example:

input_lemma	**output_lemma**
'uerna' 1	'uernalis' a
'uerna' 1	'uernalis' b
'uernum' 2	'uernalis' b
'uernum' 2	'uernalis' a

The pairings above are among the results of the query that pairs denominal adjectives of the second class with suffix -al with their supposed ‘root’ nouns (N => A2). As above, we have two identical entries for the adjective uernalis. This time the two adjectives are not distinguishable by any inflectional feature because their appearance is exactly the same as that in the LEMLAT database.

Only a dictionary consultation will reveal that the first entry for uernalis (a) means ‘of spring’, while the second entry uernalis (b) means ‘of or belonging to the house slave’. This allows one to establish that ‘a’ is connected to uernum ‘spring’ (2) and ‘b’ is connected to uerna ‘house slave’ (1).

Sometimes, deciding on which WFR to treat first can help reduce the manual work. Therefore, another factor that can affect the amount of manual checking is workflow efficiency.

Carefully choosing which WFRs to apply first is of crucial importance. As previously mentioned, there can be only one derivation relation for each output lemma. This allows one to exclude output lemmas that have already been paired, while building new relations. This can be very useful when working through the uncorrected outputs of very productive rules (with thousands of parallels to check manually), but it is only advantageous if the most productive rule has been treated first.

If all adjectives derived from the past participle of a verb through conversion, have conclusively been assigned, they will not turn up when processing another rule involving perhaps denominal adjectives that could be paired to other homographic roots.

Some morphotactically obscure word formation processes, such as certain kinds of compounding rules, are very difficult to formalise so that they can be found by automatic procedures. In these cases, derivations are added to the database almost entirely manually.

Theoretical Issues

As mentioned above, WFRs are conceived according to the Item-and-Arrangement model. This means that the construction of the lexicon favours a morpheme-based approach to morphology. Moreover, this model is put in practice through the use of input-output relationships between lexemes, in a morphotactic approach that has been prioritised over philological considerations. The use of directed edges in our derivational trees implies that one word formation process has happened before the other, and that one given lexeme specifically derives from another. The main issue in the development of a derivational morphology resource for a language like Latin - even though WFL focusses on Classical Latin - is that of the diachronic distribution of the lexicon, together with the fact that there are no native speakers to help with considerations about the transparency or opacity of a dubious derivation.

As a result, we are forced to commit to a clear-cut work policy in order to ensure consistency throughout the lexicon. Three major factors are considered when in doubt:

a) theoretical statements (Unitary Base Hypothesis,[1] Item and Arrangement), and previous research on word formation;

b) dictionaries: Oxford Latin Dictionary and Georges and Georges, we consider whether or not they support our analysis;

c) semantic motivations.

Overall coherence in the design of the lexicon is what generally drives decisions, but case-by-case analysis is necessary in order to establish a motivated derivation graph.

Compounding or suffixation?

In the treatment of certain lexemes, we had to take a decision on whether they are the result of compounding or whether certain second constituents need to be considered as suffixoids. For example, -fex -fico and -ficus (from base form fec- from facio ‘to make’) are considered suffixes denoting ‘making’ in Oxford Latin Dictionary. Yet, against Oxford Latin Dictionary’s position, we have chosen to consider -fex and -fico as second constituents for compounds, following authoritative bibliography on the subject such as Brucale 2012, Jenks 1911, Oniga 1988 and 1992.

For what -ficus (again from facio, forming adjectives) is concerned, on the other hand, we have taken a non-uniform approach: it was decided to treat certain lexemes ending in -ficus which had a corresponding verb ending in -fico, as V-to-N conversions instead. The reasoning behind this decision originated from the fact that, generally speaking, linguistic research on Latin word formation considers compounding - comparatively to what happens in other Indo-European languages - as a poorly productive phenomenon (Fruyt 2002). For this reason, one can imagine a different formation process to have happened instead. In the case of -fico and -ficus, for example, we hypothesise the verbal compound to have formed before the adjective, as the main meaning of such compounds is almost always the result of a performed action (damnificus ‘that causes loss’ <damnifico ‘to cause loss’ <damnum + facio). This also means that rather than having another separated compounding rule for damnificus, which would have connected the two lexemes at a higher, more distant, level (from damnum and facio, but not with each other), the two lexemes are directly connected in the same word formation family.

However, on 92 adjectives ending in -ficus contained in the lexical basis, 65 do not have a -fico verb counterpart to be connected to through conversion, hence they are considered the result of compounding. This has ultimately created an inconsistency in the resource, that needs to be either rectified or justified theoretically.

Prefixation or conversion?

Occasionally, we had to decide whether to prioritise one derivation process over another. In certain cases, it is difficult to ascertain whether prefixation happened before or after conversion - or suffixation - diachronically. Let us consider, for example, the choice between the following two options: 1. sono ‘to make a noise’ >resono ‘to produce a prolonged sound’ >resonus ‘reverberating/echoing’, or 2. sono>sonus ‘sounding’ >resonus. In this case the derivation trail 1, is preferred over 2, as it absolves all three of the deciding factors outlined above. Generally speaking, conversion and prefixation could have happened at any point of the derivation process, but when trying to establish whether the prefixation process happened before or after certain conversive processes, we tend to give precedence to verbal prefixation, as it appears more productive (Fruyt 2002, Oniga 1992).[2]Georges and Georges avails this hypothesis by stating that resonus comes from resono, and semantically the meaning of ‘reverberating/echoing’ seems to be more likely an inheritance from the verb meaning ‘to produce a prolonged sound’, rather than merely from the adjective ‘sounding’. In a similar way, supported by all or a combination of the three factors above, the derivation properus ‘quick’ >propero ‘to hurry’>depropero ‘to hasten’ >deproperus ‘hastening’ is chosen over properus>deproperus. On the other hand, notice how, for semantic reasons, sonus>absonus ‘of unpleasant sound’ >absono ‘to have an unpleasant sound’, as the semantic emphasis for the verbal counterpart here seems to derive from the adjective.

However, sometimes, when looking into prioritising prefixation over suffixation, relying on what the dictionary tells us makes things uneven within the same word formation family. See for example the case of horreo ‘to become stiff/ tremble (with fear) >abhorreo ‘to recoil from’ >abhorresco ‘to become disgusted’, but horreo>horresco ‘to stand up stiffly / become agitated’ >inhorresco ‘to become stiff (with cold)’ described in Budassi and Litta 2017.

Fictional entries

Another phenomenon that required an unconventional solution was the presence of parasynthetic phenomena formed through the simultaneous addition of a suffix and a prefix, or a prefix and conversion. See for example indolesco ‘to feel painful’, which, according to the Oxford Latin Dictionary, is formed by in + dolor + e-sco, or bicolor ‘of two colours’, an adjective resulting from prefixation of noun color ‘colour’ with prefix bi- indicating something ‘consisting of two things’. Bicolor has been inserted into WFL simply as the result of a prefixal WFR involving a change of PoS (N-To-A), while the first has created more complicated issues.

The project's morphotactic approach forced us to create a number of fictional lexemes in order to fill a gap in the derivation tree. These are not reconstructed lexemes, and are not expected to have ever existed, but only serve as a “mechanical” connection between two lexemes. For

instance, indoleo only acts as a trait d'union between dolor ‘pain’, and indolesco ‘to feel painful’, hence creating two formative processes instead of one.

Backformation

Backformation is a derivation process that involves the removal of an affix (or a supposed one), so that a new word is created by analogy with similar looking existing ones, as for example galeo ‘to equip with a helmet’ <galeatus (adj.) ‘wearing a helmet’, grego ‘to collect into a flock’ <aggrego/congrego ‘to bring people together’, irascor ‘to be angry’ < iratus ‘angy’, (g)nascor ‘to be born’ < natus ‘born’.

At the time of writing, these and other backformation processes are not marked in WFL, but are portrayed as follows: galeo<galea ‘helmet’, galeatus<galea, aggrego/congrego<grego, but iratus>irascor (A-To-V -sc-), (g)nascor = root verb (i.e. not derived). In the future, we are planning on adding backformation as a de facto WFR, and mark backformations to the three graphs so that the edge can point in the opposite direction, or be recognisable by a different colour.

[1]According to the Unitary Base Hypothesis Word Formation Rules may only operate over a single type of syntactically or semantically defined base (Scalise 1983, Aronoff 1976).

[2]In WFL there are 4,281 prefixed V-To-V vs. 786 V-To-N and 101 V-To-A conversions.

Parts of Speech (PoS)

Keys for PoS codes used in the resource

A: adjectives. A is followed by a number indicating the class of the adjective, i.e. 1 for adjectives of the first class and 2 for adjectives of the second class. For example, A2 means adjective of the second class.

N: nouns. The number following the initial letter indicates the declension, e.g. N1, N2, N3 etc. The small letter following the declension number indicates the gender of the noun, i.e. m for masculine, f for feminine, n for neuter. For example, N2m means masculine noun of the second declension. N on its own means uninflected nouns, such as numerals, letters (beta) and unassimilated borrowings.

V: verbs. V is followed by a number indicating the conjugation, or by A to indicate an auxiliary verb, e.g. V1, V3, VA, etc. V5 means e/i conjugation verbs e.g. capio 'to take'.

I: invariable lemmas, i.e. adverbs, interjections, conjunctions.

PR: pronouns

Which Word Formation Rules?

Word Formation Rules (WFRs) were conceived according to the Item-and-Arrangement (IA) model, which considers word forms either as simple (non-derived) morphemes or as a sequence of morphemes (base and affixes) having both form and meaning (Hockett 1954). IA was chosen as the theoretical model supporting WFL for two main reasons: first, it emphasises the semantic significance of affixal elements as they are found in the lexicon; secondly, IA was the model adopted by other derivational lexica like Word Manager (Domenig and ten Hacken, 1992), after which WFL was designed.

There are two types of word formation processes in Latin: derivation and compounding.

Derivation can be further split into:

1) Affixation, where one or more morphemes, called affixes, can be attached to the base of a word. Affixation can be of two types, and can involve (or not) a change of part of speech:

Prefixation: where the affix is attached before the base.
Suffixation: where the affix is attached after the base.
2) Conversion, where the derived word incurs only in a change of part of speech without the addition of any affix.

Compounding is the formation of a new lexeme from two or more lexemes.

Prefixation

In general, we consider prefixes those morphemes that are often equivalent to a preposition, which are attached to the beginning of a word, e.g. ad, ab, ex, prae etc. as indicated by Oxford Latin Dictionary.

Moreover, we have included among prefixes all those numeral affixes that are considered prefixes by the Oxford Latin Dictionary. These are bi-, tri-, quadri-.

The following is a list of prefixes that can be found in WFL:

a(b)-
ad-
am(b)(i)-
ante-
archi-
bi-
circum-
con-
contra-
de-
dis-
e(x)-
ec-
extra-
in (entering)-
inter-
intro-
multi-
ne-
ob-
per-
post-
prae-
praeter-
pro-
quadri-
re-
retro-
se-/sed-/so-
semi-
sub-
subter-
super-
tra(ns)-
tri-

Suffixation

A suffix is an affix that is placed after the stem of a word, to form another word.

A list of suffixes that can be found in WFL:

(at)im
(i)cul
(i/a)n
(i/e)ll
(t)io(n)
(t)iu
(t)or
(t)ric
(t)ur
ac
ace
al
an
ar
at
atil
bil
bund
cell/cill
cr
cul
e
edo/edin
el
et
far
fex
go/gin
i
ic
ici
id
il
iss
ist
it
iti
itud/itudin
men/min

ment
n
ol
or
os
ri
sc
str
tas/tat
tori
tr
udo/udin
ul
uncul
ur

Conversion

We have included change of inflection class, or gender, such as masculine to feminine, or 3rd declension to 1st declension, among N-to-N conversions.

examples: aera N1 < aes N3:n

Conversions in WFL are of the following kind:

V-To-V
V-To-N
V-To-A
V-To-I
N-To-V
N-To-N
N-To-A
N-To-I
A-To-V
A-To-N
A-To-I
I-To-A
PR-To-N

Compounding

Compound words collected in WFL are created through 59 WFRs.

A+A=A
A+A=N
A+I=I
A+N=A
A+N=I
A+N=N
A+PR=PR
A+V=A
A+V=N
A+V=V
I+A=A
I+A=I
I+I=I
I+N=A
I+N=I
I+N=N
I+N=N
I+PR=A
I+PR=I
I+PR=PR
I+V=A
I+V=I
I+V=N
I+V=V
N+A=A
N+A=N
N+I=I
N+N=A
N+N=N
N+V=A
N+V=I
N+V=N
N+V=V
PR+A=A
PR+A=N
PR+A=PR
PR+I=I

PR+I=PR
PR+N=A
PR+N=I
PR+PR=I
PR+PR=PR
PR+V=N
PR+V=PR
PR+V=PR
PR+V=V
V+A=A
V+N=A
V+N=I
V+N=N
V+PR=PR
V+V=A
V+V=N
V+V=V

Accessing the Data

Important caveats:

1) WFL only contain derived and compounded lemmas and their input lemmas. You will not find a lemma that is not derived, or that does not have a derived output.

2) Type 'u' instead of 'v'.

The word formation lexicon can be accessed on-line through a visualisation query system (http://wfl.marginalia.it). The lexicon can be browsed either by WFR, affix, or input and output PoS or lemma. Drop down menus provide the available options for each selection, like for instance the list of affixes and lemmas.

Results are visualised as tree graphs, whose nodes are lemmas and edges are WFRs. Trees are interactive. Clicking on a node shows the full derivation tree (“word formation cluster”, which is calculated dynamically) for the lemma reported in that node. For example, Figure 1 shows part of the word formation cluster for the lemma amo 'to love'. One can see that amabilis 'lovable' derives from amo and it is in turn the input for two other derived lemmas: amabilitas 'loveliness' and inamabilis 'unlovely'. Clicking on an edge shows the lemmas built by the WFR concerned in that edge. Lemmas are provided both as a derivation graph and as an alphabetical list. For instance, clicking on the edge going from amo to amabilis in Figure 1 shows the lemmas

built by the derivation WFR that builds second class adjectives (A2) from first conjugation verbs (V1) with suffix –bil–.

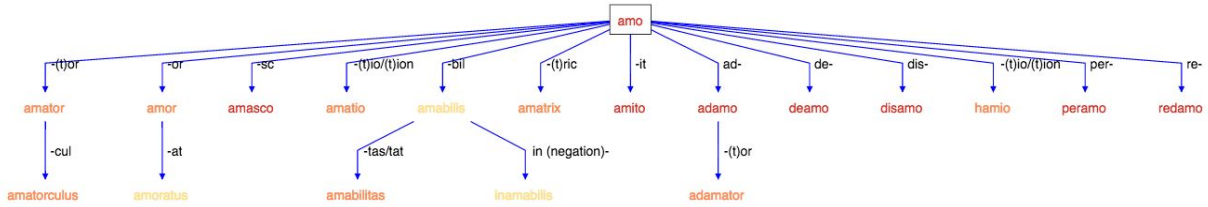

Figure 1: partial word formation family for amo.

Figure 2 presents a portion of the derivation graph for this rule.

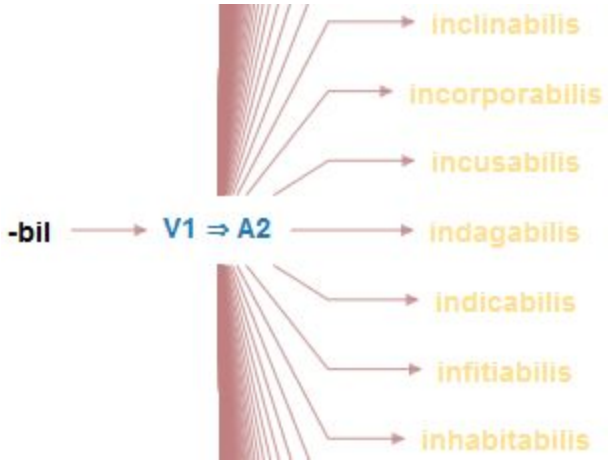

Figure 2. Derivation graph for a WFR.

In order to enable users to access WFL, we have developed a specific web application, where the relationships between lexemes of the same word formation family are represented as a tree-graph. In this graph, a node is a lexeme, and an edge is the WFR used to derive the output lexeme from the input one (or two/three, in the case of compounds), along with any affix used.The entire database is thus like a big graph represented as a collection of edges, and the set of word formation families is simply the set of connected subgraphs.

The website has been designed keeping in mind the kinds of queries and results that a linguist would be interested in. There are four distinct perspectives to query WFL:

1.By WFR – the primary interest is the behaviour of a specific WFR. For example, it is possible to view and download a list of all verbs that derive from a noun with a conversive derivation process (e.g. radix ‘root’ >radicor ‘to grow roots’);

2.By affix – it acts similarly as above, but works more specifically on affixal behaviour. For example, this perspective enables to retrieve all masculine nouns featuring the suffix -tor and to verify how many of them correspond to a female equivalent ending in -trix;

3.By PoS – the primary interest is the part of speech (PoS) of input and output lexemes. This view is useful for studies on macro-categories of morphological transformation, like nominalisation and verbalisation;

4.By lexeme – it focuses on both derived and non-derived lexemes. It supports studies on the productivity of one specific morphological family or a set of morphological families.

The results of these browsing options are of three types:

a.lists of lexemes resulting from a query, that can be downloaded in a .txt file;

b.derivational graphs: this type of graph represents the derivational chain (or cluster) for a specific lexeme, which includes all the lexemes derived from the lexeme selected, as well as all those it is derived from;

c.a summary of the application of a given WFR to different PoS and the resulting lexemes.

An important design aspect of the WFL web application is the fact that it limits queries that produce no results. Queries could produce no results if they search either for unattested WFRs, or for WFRs not yet included in WFL. Providing users with all the possible combinations of PoS, WFRs and affixes would result in quite long lists, thus requiring users to run single queries to find manually which of them have no occurrences in WFL. Instead, this can be easily inferred from the interface, as it is expected that one possible combination that is unavailable in the interface does not correspond to any word in WFL. For instance, the suffix -ace- is available only for denominal adjectives (N-To-A, argilla ‘clay’ >argillaceus ‘made of clay’), which means that it is not at work for all the other possible combinations of input/output PoS.

In the web application, the four perspectives on queries mentioned above are implemented as four different screens, accessed via a top-level menu.

For WFRs and affixes, the basic type (e.g. “Prefixation” for WFRs, or “Prefixes” for Affixes) is chosen via tab buttons, and for all perspectives the finer grained choices are specified via drop-down menus. The difference between querying the database by WFRs and by affixes is reflected in the priority of drop-down menus. For WFRs, first a WFR type (or types) is chosen (e.g. V-to-V for deverbal verbs), and then any desired affixes. The choice of the WFR type updates the second drop-down menu to restrict the affixes to just those that occur with the selected WFR type. A similar interaction holds for affixes.

The PoS-based query option does not have an intermediate level of selection, but the choice is made via a series of drop-down menus. For each possible item involved in a WFR (one or two base input lexemes - the latter for compounds - and the output), there is the choice of PoS, and then refinements of that PoS: these are inflectional categories for all PoS (declension for nouns, classes for adjectives and conjugation for verbs), as well as gender for nouns. The options for the inflectional categories are limited to those appropriate for the PoS chosen.

Querying WFL by lexeme is performed by radio buttons, which allow for the selection of “all lexemes”, “only roots” of derivational clusters (not derived lexemes) or “only derived lexemes”. The list of lexemes with their PoS (and gender, for nouns) is shown in a list, which can be filtered with the employment of common regular expression queries.

The three types of query results are visualised in distinct ways in separate windows, interacting across the result types. Clicking on a lexeme in the list opens its derivational graph in a separate window.

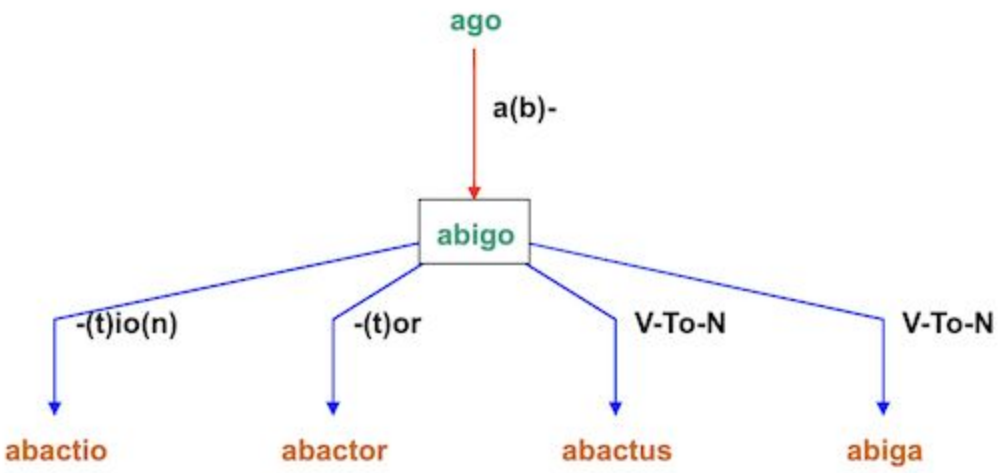

Figure 3. Derivational graph of abigo ‘to drive away’.

In the graph of Figure 3, nodes are filled with lexemes and edges are labelled with affixes or input-output PoS (in the case of compounding and conversions). The selected lexeme is shown inside a box. Clicking on any lexeme in the graph replaces the current derivational graph with the one for the clicked lexeme, moving the focus of the derivational trail. Clicking on an edge label in the derivational graph opens a new window (Figure 2) which provides a visualisation summarising the application of the corresponding WFR by PoS, a left-rooted tree, with the name of the affix as the root (first level of the tree), and all the combinations of the input and output

PoS with their refinements (e.g. conjugation for verbs) as second level branches, giving the number of lexemes for each input-output combination.

Figure 4. Derivation Graph for WFR N-to-A -ace

The graph is collapsable so the user can focus on certain subsets only. As the subsets change, the list of lexemes is updated to reflect just the subsets that are selected.

An additional feature of querying the lexicon by WFRs and by affixes is to search across the full derivational path of lexemes, thus providing results that go beyond the “outermost” WFR. By selecting the “include as intermediate” option, one can search not only for all the lexemes derived by a specific WFR but also for those that include at least one lexeme produced by that WFR along their derivational path. For instance, with this option selected, among the results of a query that searches for deverbal adjectives formed with suffix -bil is not only the adjective affabilis ‘that can be easily spoken’, but also the noun affabilitas ‘courtesy’ which has a deverbal adjective formed with suffix -bil along its derivational path as it is derived from affabilis.

All results can be downloaded: the list of lexemes as a tab-delimited text file, while the derivation graphs and WFR trees as images.

Moreover, in the case of compounds, the user can choose whether to visualise or not both the roots of compounds, thus resulting in a multi-tree graph rather than a simple tree.

It has to be remembered that, while searching by WFRs or PoS, a few peculiarities of the Lemlat lexical basis can result in a rather unconventional classification of the rules, which impacts especially (but not solely) on searches performed on compounds. For instance, participial adjectives are not included in the Lemlat lexical basis, because they are considered part of the verbal paradigm. This means that certain compounds that would be expected to have an adjective (A) as one of their constituents have a verb (V) instead, e.g. adjective altisonus (altus + sono) ‘that sounds high up, sublime’ can be found among V+V=A compounds rather than among A+V=A. Also, adjectival adverbs are considered in Lemlat adverbial cases of the adjectival declension, hence a word such as dulciloquus (dulce + loquor) ‘sweet talking’ is to be found among A+V=A, rather than I+V=A.

Across the free text search options it is possible to use regular expressions.

Project Publications

The present documentation on WFL is covered partly by the following publications that result from the work done during the MSCA Fellowship:

Budassi, Marco, Eleonora Litta, and Marco Passarotti. 2017. '-io Nouns through the Ages. Analysing Latin Morphological Productivity with Lemlat'. In Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017), 65-70. aAccademia University Press, Roma.
http://www.aaccademia.it/component/search/?searchword=clic-it&searchphrase=all&Itemid=118

Budassi, Marco, and Eleonora Litta. 2017. 'In Trouble with the Rules. Theoretical Issues Raised by the Instertion of -sc- verbs into Word Formation Latin'. In Proceedings of the Workshop on Resources and Tools for Derivational Morphology (DeriMo), 15–26. Milan: Educatt. http://itreebank.marginalia.it/doc/2017_Litta-Passarotti_Proceedings-DeriMo.pdf

Culy, Chris, Eleonora Litta, and Marco Passarotti. n.d. 'Visual Exploration of Latin Derivational Morphology'. In Proceedings of the Thirtieth International Florida Artificial Intelligence Research Society Conference. Marco Island, Florida. May 22–24, 2017, 601–6. Palo Alto, California - USA: The AAAI Press. https://www.aaai.org/Library/FLAIRS/flairs17contents.php

Litta Eleonora, and Marco Passarotti. 2017. 'Preface'. In Proceedings of the Workshop on Resources and Tools for Derivational Morphology (DeriMo). Milan: Educatt. http://itreebank.marginalia.it/doc/2017_Litta-Passarotti_Proceedings-DeriMo.pdf

Litta, Eleonora, Marco Passarotti, and Paolo Ruffolo. 2017. 'Node Formation: Using Networks to Inspect Productivity in Affixal Derivation in Classical Latin'. In Proceedings of the 2Nd International Conference on Digital Access to Textual Cultural Heritage, 103–8.DATeCH2017. New York, NY, USA: ACM. doi:10.1145/3078081.3078092.

Litta, Eleonora, Marco Passarotti, and Chris Culy. n.d. 'Formatio Formosa Est. Building a Word Formation Lexicon for Latin'. In Third Italian Conference on Computational Linguistics (CLiC–it 2016), 185–89. Naples: aAccademia University Press. http://www.aaccademia.it/component/search/?searchword=CliC-it 2016&searchphrase=all&Itemid=118.

Micheli, Silvia, and Eleonora Litta. 2017. 'E pluribus unum. E pluribus unum. Representing compounding in a derivational lexicon of Latin.' In Proceedings of the Fourth Italian Conference

on Computational Linguistics (CLiC-it 2017), 65-70. aAccademia University Press, Roma. http://www.aaccademia.it/component/search/?searchword=clic-it&searchphrase=all&Itemid=118

Passarotti, Marco, 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, 24–31. Linköping University Electronic Press. http://www.ep.liu.se/ecp/article.asp?issue=133&article=006&volume=.

Bibliography

This is a list of works that were used and consulted during the compilation of Word Formation Latin. Some of these titles are mentioned in this documentation.

Aronoff, Mark. 1976. ‘Word Formation in Generative Grammar’. Linguistic Inquiry Monographs Cambridge, Mass., no. 1:1–134.

Bader, Françoise. 1962. La Formation Des Composés Nominaux Du Latin.Vol. 46. Presses Univ. Franche-Comté.

Benedetti, Marina. 1988. I composti radicali latini: esame storico e comparativo.Giardini.

Clackson, James. 2002a. ‘Composition in Indo-European Languages’. Transactions of the Philological Society 100 (2):163–67.

———. , ed. 2011. A Companion to the Latin Language. Blackwell Companions to the Ancient World. Chichester ; Malden, Mass: Wiley-Blackwell.

Domenig, Marc. 1988. ‘Word Manager: A System for the Definition, Access and Maintenance of Lexical Databases’. In Proceedings of the 12th Conference on Computational Linguistics-Volume 1, 154–59. Association for Computational Linguistics.

Domenig, Marc, and Pius Ten Hacken. 1992. Word Manager: A System for Morphological Dictionaries. Vol. 1. Georg Olms Verlag AG.

Fruyt, M. 2002. ‘Les Dérivés En-Cus,-Ca,-Cum’. C. Kircher-Durand (éd.), Création Lexicale: La Formation Des Noms Par Dérivation Suffixale, Louvain-Paris-Dudley, MA, 67–108.

Fruyt, Michèle. 1990. ‘Complex Lexical Units in Latin’. In New Studies in Latin Linguistics: Proceedings of the 4th International Colloquium on Latin Linguistics, Cambridge, April 1987, 21:75. John Benjamins Publishing.

———. 2000. ‘La Création Lexicale: Généralités Appliquées Au Domaine Latin’. In Michèle Fruyt et Christian Nicolas (éd.), La Création Lexicale En Latin.Actes de La Table Ronde Du Neuvième Colloque International de Linguistique Latine (Madrid, 16 Avril 1997), Paris, Presses de L’université de Paris-Sorbonne, 11–48.

———. 2002. ‘Constraints and Productivity in Latin Nominal Compounding’. Transactions of the Philological Society 100 (3):259–87.

———. 2011. ‘Word-Formation in Classical Latin’. In A Companion to the Latin Language, edited by James Clackson, 157–75. Wiley-Blackwell.

Gaide, F. 1988. Les Substantifs Masculins Latins En...(i) O,...(i) Onis. éditions Peeters Louvain-Paris.

Hacken, Pius ten. 1998. ‘Word Formation in Electronic Dictionaries’. Dictionaries: Journal of the Dictionary Society of North America 19 (1):158–87.

Hacken, Pius ten, Stephan Bopp, Marc Domenig, Dieter Holz, Alain Hsiung, and Sandro Pedrazzini. 1994. ‘A Knowledge Acquisition and Management System for Morphological Dictionaries’. In , 2:1284. Association for Computational Linguistics.

Hacken, Pius Ten, and Dorota Smyk. 2003. ‘Word Formation versus Etymology in Electronic Dictionaries’, 221–30.

Hathout, Nabil, and Fiammetta Namer. 2014. ‘Démonette, a French Derivational Morpho-Semantic Network’. LiLT (Linguistic Issues in Language Technology) 11.

Hockett, Charles F. 1954. ‘Two Models of Grammatical Description’. Morphology: Critical Concepts in Linguistics 1:110–38.

ICCU. n.d. ‘Risultati sintetici’.OPAC SBN. Accessed 16 December 2015. http://www.sbn.it/opacsbn/opaclib.

Jenks, P.R. 1911. A Manual of Latin Word Formation for Secondary Schools. DC Heath & Company.

Kircher-Durand, Chantal. 2002. Grammaire Fondamentale Du Latin Tome IX Création Lexicale: La Formation Des Noms Par Dérivation Suffixale. Vol. 9. Peeters Publishers.

Kircher-Durand, Chantal. 2002a. ‘Les Adjectifs En-Eus,-A,-Um’. Création Lexicale: La Formation Des Noms Par Dérivation Suffixale, Louvain-Paris-Dudley Ma, 85–108.

Kircher-Durand, Chantal. 2002b. ‘Les Dérivés En-ENSIS’.Grammaire Fondamentale Du Latin9:185–94.

———. 2002c.‘Les Dérivés En-Nus,-Na,-Num’. Grammaire Fondamentale Du Latin 9:125–84.

Leumann, Manu. 1944a. Der Anteil Der Schweiz an Der Sprachforschung. Huber.

———. 1944b. Gruppierung Und Funktionen Der Wortbildungssuffixe Des Lateins.

Lieber, Rochelle, and Pavol Stekauer. 2009. The Oxford Handbook of Compounding. OUP Oxford.

———. 2014. The Oxford Handbook of Derivational Morphology. OUP Oxford.

Linguistic Studies on Latin : Selected Papers from the 6th International Colloquium on Latin Linguistics (Budapest, 23–27 March 1991). 1994. Amsterda, NLD: John Benjamins Publishing Company.

Maire, Brigitte. 2014. ‘Greek’ and ‘Roman’ in Latin Medical Texts: Studies in Cultural Change and Exchange in Ancient Medicine. BRILL.

Nielsen, B. 2000. ‘On Latin Instrument-Nouns In /-Lo’. In Indo-European Word-Formation. Proceedings of the Conference Held at the University of Copenhagen October 20th-22nd 2000.

Oniga, Renato. 1988. I composti nominali latini: una morfologia generativa. Pàtron.

———. 2007. Il latino: breve introduzione linguistica. FrancoAngeli.

———. n.d. ‘Compounding in Latin, Rivista Di Linguistica 4, 1992, 97-116’.

Pala, Karel, and Radek Sedláček. 2005. ‘Enriching Wordnet with Derivational Subnets’. In Computational Linguistics and Intelligent Text Processing, 305–11. Springer.

Panichi, Emidio. 1972. La Formazione Nominale in Latino.Liguori.

Pecman, Mojca. 2002. ‘Les Adjectifs En-Ax’.Grammaire Fondamentale Du Latin. Tome IX.Création Lexicale: La Formation Des Noms Par Dérivation Suffixale, pp – 25.

Pedrazzini, Sandro, and Pius ten Hacken. 1993. Phrase Manager. Universität Basel. Institut für Informatik.

Rasmussen, Jens Elmeg\a ard. 2002. ‘The Compound as a Phonological Domain in Indo–European’. Transactions of the Philological Society 100 (3):331–50.

Rosén, Hannah. 1983. The Mechanisms of Latin Nominalization and Conceptualization in Historical View. de Gruyter.

SBLENDORIO CUGUSI, M.T. 1991. I Sostantivi Latini in -Tudo. Pàtron.

Ševčíková, Magda, and Zdeněk Žabokrtskỳ. 2014. 'Word-Formation Network for Czech'. In Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014), 1087–93.

Sgroi, Salvatore Claudio. 2006. 'Dizionari a Confronto: A Proposito Della "Wortbildung" Nella Lessicografia Italiana', 1181–92.

Ten Hacken, Pius. 1994. Defining Morphology: A Principled Approach to Determining the Boundaries of Compounding, Derivation, and Inflection. Vol. 4. Georg Olms Verlag.

———. 2000. 'Derivation and Compounding'. Morphology/Morphologie, 349–59.

Ten Hacken, Pius, and Marc Domenig. 1996. 'Reusable Dictionaries for NLP: The Word Manager Approach'. LEXICOLOGY-BERLIN- 2:232–55.

Zeller, Britta D., Jan Snajder, and Sebastian Padó. 2013. 'DErivBase: Inducing and Evaluating a Derivational Morphology Resource for German.' In ACL (1), 1201–11. http://anthology.aclweb.org/P/P13/P13-1118.pdf.

Zeller, Britta, Sebastian Padó, and Jan Šnajder. 2014. 'Towards Semantic Validation of a Derivational Lexicon'. In Proceedings of COLING, 1728–39. http://anthology.aclweb.org/C/C14/C14-1163.pdf.n.d.