Morphology beyond inflection. Building a word-formation-based lexicon for Latin

Eleonora Litta, CIRCSE - Università Cattolica del Sacro Cuore, eleonoramaria.litta@unicatt.it
Marco Passarotti, CIRCSE - Università Cattolica del Sacro Cuore, marco.passarotti@unicatt.it
Chris Culy, independent scholar, chrisculy@mac.com

In the construction of computational language resources, derivational morphology has constantly been overlooked compared to inflectional morphology, which plays a central role in fundamental annotation tasks such as PoS tagging. Yet enhancing textual data with derivational morphology tagging has the potential to provide solid results.

The study of derivation allows for a fine-grained organisation of the lexicon by linking words that share the same lexical ancestor or the same word formation process. Likewise, because core semantic properties are shared at different extent by derivate words built by a common word formation process, derivational morphology can act like a kind of interface between morphology and semantics. This especially holds true in the context of linguistic investigations of textual material.

In recent times, the computational linguistics world has gradually been focussing its interests on researching and building new derivational morphology resources and tools. Among them are the lexical network for Czech DeriNet (Ševciková e Zabokrtský 2014), the derivational lexicon for German DErivBASE (Zeller, Snajder e Padó 2013) and that for Italian derIvaTario (Talamo, Celata e Bertinetto 2016).

On the Classical languages front, although the number of lexical resources and NLP tools, especially for Latin, is now large and varied (corpora, treebanks, computational lexica, and digital libraries)\(^1\), there has not been any attempt to create a derivational morphology tool, where lemmas are segmented and analysed into their derivational components, to establish relationships between them on the basis of word formation.

The first steps towards building such a word formation lexicon for Latin were made by Passarotti e Mambrini (2012), who described a model for the semi-automatic extraction of word formation rules from the list of lemmas of Lexicon Totius Latinitatis by Forcellini (1940, fifth edition) and the subsequent pairing of lexical entries and their derivational ancestor(s).

In this context, the Word Formation Latin project has received funding from the European Union’s Horizon 2020 research and innovation programme (Marie Skłodowska-Curie grant agreement No 658332-WFL) to create the definitive derivational lexicon for Classical Latin. This will ultimately be included in the automatic morphological lemmatiser for Latin LEMLAT (http://www.ilc.cnr.it/lemlat/lemlat/index.html, accessed 20/05/2016), creating a 360° resource for the study of Latin morphology.

This proposal contains a brief description of the methodology used to build the resource, followed by a preview on how the data will be ultimately accessed and presented to the user in the finished resource.

The lexical basis used for the word formation based resource is the one featured in LEMLAT (http://www.ilc.cnr.it/lemlat/lemlat/index.html), which results from the collation of three Latin dictionaries (Georges and Georges, 1913-1918; Glare 1982; Gra- denwitz 1904). It counts 40,014 lexical entries and 43,432 lemmas (as more than one lemma can be included into the same lexical entry). Recently, the lexical basis of Lemlat was further enlarged by adding 26,250 lemmas from the Onomasticon provided by Forcellini (1940).

The word formation lexicon is built in two steps:

- Word formation rules are detected.
- They are applied to lexical data.

Word formation rules (WFRs) are grouped in two classes: 1. compounding; 2. derivational. Derivational rules are divided in two further categories: a. affixal (in turn split into prefixal and suffixal), and b. conversion, a derivation process that affects PoS without including any affix. Compounding and conversion WFRs are automatically detected, by considering all the possible combinations of main PoS (verbs, nouns, adjectives). Affixal WFRs are found both according to previous literature on Latin derivational morphology (Jenks 1911; Fruyt 2011; Oniga 1988) and in semi-automatic fashion. Each morphologically derived lemma is assigned a WFR. All those lemmas that share a common (not derived) ancestor belong to the same “morphological family”. For instance, lemmas *formatio* (“formation”), *formo* (“to form”) and *formosus* (“beautiful”, lit. “finely formed”) all belong to the morphological family whose ancestor is the lemma *forma* (“form”).

Lemmas and WFRs are paired by using a MySQL relational database, and a number of MySQL queries provide the candidate lemmas for each WFR.

Given the high number of homographs in Latin, the procedure described above is not necessarily sufficient for building the morphological families. Thorough manual checking allows the identification of false results, duplication and lacunas resulting from the automatic process. Such duplicate results need to be analysed and rectified.

Some morphotactically obscure word formation processes, like some kinds of compounding, need to be completely hardcoded. The word formation based lexicon is accessible on-line through a visualization query system (temporarily at http://wfl.marginalia.it). The lexicon can be browsed either by WFR, affix, 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 visualized as tree graphs, whose nodes are lemmas and edges are WFRs. Trees are interactive. Clicking on a node shows the full derivational tree (“word formation cluster”) for the lemma reported in that node. For example, Fig. 1 shows the word formation cluster for the lemma *bellum*. Clicking on an edge shows the lemmas...
built by the WFR concerned in the edge. Lemmas are provided both as a derivational graph and as an alphabetical list. For instance, double clicking on the edge going from *bello* to *re-bello* in Fig. 1 shows the lemmas built by the derivational WFR that builds new verbs from first conjugation verbs (V1) with prefix re–. Fig. 2 presents a close-up of the derivational graph for this rule.

![Figura 1: Word formation cluster for *bello*](image1)

![Figura 2: Derivational graph for a WFR](image2)
Phase 2 in the project workflow will be to integrate the information extracted from the resulting derivational morphological lexicon into the morphological layer of annotation of the *Index Thomisticus Treebank* (IT-TB). The *Index Thomisticus* (IT) is considered a pioneer in digital humanities: started by father Roberto Busa SJ in 1949, it is a database retaining the *opera omnia* by Thomas Aquinas (118 texts), plus 61 works by other authors related to Thomas.

The integration of the word formation-based information into the IT-TB will be executed through the embedding of the lexicon data within the morphological layer of annotation of the treebank, using TEI P5 conformant XML encoding to favour data exchange and linking to other lexical resources.

The final resource will be both a standalone lexicon and tool accessible through its own website, and interconnected with the IT-TB. Finally the whole resource will be made available through the CLARIN infrastructure (www.clarin.eu).

**Bibliografia**


