Formatio formosa est.

Building a Word Formation Lexicon for Latin

Eleonora Litta, Marco Passarotti, Chris Culy
CIRCSE Research Centre
Università Cattolica del Sacro Cuore
Largo Gemelli, 1 – 20123 Milan, Italy
{eleonoramaria.litta, marco.passarotti}@unicatt.it,
chrisculy@mac.com

Abstract

English. This paper presents the steps undertaken for building a word formation lexicon for Latin. The types of word formation rules are discussed and the semi-automatic procedure to pair their input and output lexical items is evaluated. An on-line graphical query system to access the lexicon is described as well.

Italiano. *Questo articolo presenta le procedure di realizzazione di un lessico morfologico derivazionale per il latino. Sono descritti i tipi di regole di formazione di parola e viene valutata la qualità del sistema semi-automatico di individuazione delle parole in input e in output ad esse. Il sistema grafico d'interrogazione on-line dei dati è altresì presentato.*

1 Introduction

In the area of Natural Language Processing (NLP), derivational morphology has always been neglected if compared to inflectional morphology, which plays a central role in fundamental annotation tasks like PoS tagging. Yet enhancing textual data with derivational morphology tagging promises to provide strong outcomes. First, it organises the lexicon at higher level than words, by building word formation based sets of lexical items sharing a common derivational ancestor. Secondly, derivational morphology acts like a kind of interface between morphology and semantics, since core semantic properties are shared at different extent by words built by a common word formation process.

Lately, some lexical resources for derivational morphology have been made available. Among them are the lexical network for Czech DeriNet (Ševčíková and Žabokrtský, 2014), the derivational lexicon for German DErivBASE (Zeller et al., 2013) and that for Italian derIvaTario (Talamo et al., 2016). Furthermore, stemming is a technique largely used for detecting word formation processes (Goldsmith, 2001), and language independent NLP tools were trained to extract derivation information from inflectional lexica (Baranes and Sagot, 2014).

On the Classical languages front, although the number of resources and NLP tools for Ancient Greek and Latin is now manifold and varied (ranging from digital libraries, treebanks and computational lexica to PoS taggers and parsers), no lexical resource for derivational morphology is available yet, where words are connected by word formation processes. The first steps towards building such a word formation lexicon for Latin were made by Passarotti and 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 (fifth edition; 1940) and the subsequent pairing of lexical entries and their derivational ancestor(s).

The *Word Formation Latin* project has received funding from the EU Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie Individual Fellowship to expand on these efforts and create a word formation lexicon (working as an NLP tool as well) for Latin. In this paper, we describe the steps undertaken to build such a lexicon.

The paper is organised as follows. Section 2 presents the lexical basis supporting the lexicon; section 3 details the way the lexicon is built; section 4 describes how to access the data; section 5 concludes the paper and sketches the future work.

2 Lemlat

The lexical basis used for building the word formation lexicon is the one provided by the morphological analyser for Latin Lemlat (Passarotti, 2004). Resulting from the collation of three Latin dictionaries (Georges and Georges, 1913-1918; Glare, 1982; Gradenwitz, 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 most of the Onomasticon (26,250 lemmas out of 28,178) provided by Forcellini (1940).

The basic component of the lexical look-up table used by Lemlat to morphologically analyse (and lemmatise) the input wordforms is the so-called les ("LExical Segment"), which roughly corresponds to the invariable part of the inflected forms. In other words, the les is the sequence (or one of the sequences) of characters that remains the same in the inflectional paradigm of a lemma (hence, the les does not necessarily correspond to the word stem). For instance, puell is the les for the lemma puell–a ("girl").

Lemlat includes a LES archive, in which each LES is assigned a number of inflectional features among which are a tag for the gender of the lemma (for nouns only) and a code (CODLES) for its inflectional category. For instance, the CODLES for the LES *puell* is N1 (first declension regular nouns) and its gender is F (feminine).

3 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.

3.1 Detecting Word Formation Rules

Word formation rules (WFRs) are conceived according to the so-called *Item-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 mentioned in Section 1.

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. So far, we have detected 167 affixal WFRs: 71 prefixal and 96 suffixal.

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.

3.2 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 "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 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 the LES archive. 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 of *accuso* → *sub+accuso* ("to blame" → "to blame somewhat"). Instead, suffixal WFRs are mostly assigned to their candidate input and output lemmas by running queries on the LES archive, because suffixes attach to LES instead of modifying full lemmas, like in *amo* → *amabilis* ("to love" → "lovable") where suffix *–bil–* attaches to LES *am* (plus the thematic vowel *–a–*, used for first conjugation verbs) instead of full lemma *amo*. Also, there are suffixal WFRs whose input is the basis of the irregular perfect participle of the input verb, like in *duco* → *ductilis* ("to lead" → "that may be led") where suffix *–il–* attaches to the basis of the irregular perfect participle of the verb *duco* (*duct*). Such irregular bases are recorded explicitly in the LES archive with a specific CODLES.

3.3 State of Affair and Evaluation

The procedure described above is not sufficient neither for detecting nor for applying the WFRs and, ultimately, for building the morphological families. Manual checking is largely needed for identifying false results and disambiguating duplication, as well as for filling lacunas resulting from the automatic process.

For example, while looking for the candidates of the WFR that forms adjectives from nouns with the addition of the suffix *–ax/–acis*, two candidate input nouns are found for the adjective *fugax* ("swift, transitory"): *fuga* ("flight") and *fugium* (rare, scarcely used in place of *fuga*). Such duplicate results need to be checked and disambiguated manually, as there must be only one input lemma for each output lemma resulting from a WFR of the derivation type, just like there must be only one WFR associated with each derived lemma.

Morphotactically obscure word formation processes, like most compounding WFRs, are examples of lacunas of the automatic process of assigning WFRs, which are thus fully manually hard-coded. For instance, the compound lemma *matricida* ("matricide") is derived by compounding the input lemmas *mater* ("mother") and *caedo* ("to cut"), thus showing quite an obscure morphotactic configuration.

So far, we have applied to data 134 WFRs (45 prefixal, 80 suffixal, 6 conversion and 3 compounding), which corresponds to having assigned a WFR to 18,774 lemmas. Evaluation is performed by calculating the precision rate (Van Rijsbergen, 1979) of MySQL queries, i.e. the percentage of the correct candidate input-output pairs that are automatically assigned to a WFR by a query.

As expected, precision is higher when morphotactic mutations are lower. Indeed, while precision rates for prefixal rules range between 0.95 and 0.8, as they imply quite a few graphical mutations, precision for suffixal rules can vary heavily, ranging from 0.75 to as little as 0.3. Instead, the recall of queries has to be calculated later in the project, as currently we are unable to verify how many derived lemmas are not automatically picked up by queries.

4 Accessing the Data

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 the currently available word formation cluster for the lemma *amo*. One can see that *amabilis* 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 2 presents a portion of the derivation graph for this rule.

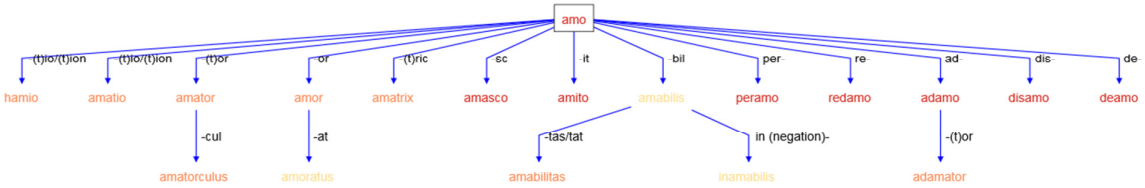

Figure 1. Word formation cluster for *amo*.

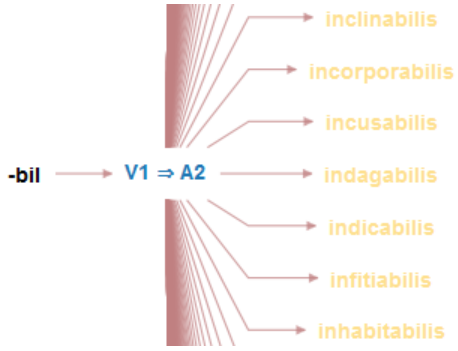

Figure 2. Derivation graph for a WFR.

5 Conclusion and Future Work

The building process of the word formation lexicon for Latin is ongoing. We still have to fully exploit the potential of querying the lexical basis of Lemlat to automatically detect candidates for WFRs. Furthermore, a substantial amount of manual work is needed to pick up morphotactically obscure formations, like those resulting from compounding.

The word formation lexicon is meant to enhance Lemlat by providing its processing with word formation analysis of input data, thus building a wide lexical resource and NLP tool for Latin morphology, which will be made available through CLARIN infrastructure (www.clarin.eu).

References

Marion Baranes and Benoît Sagot. 2014. A Language-Independent Approach to Extracting Derivational Relations from an Inflectional Lexicon. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. ELRA, Reykjavik, Iceland, 2793–2799.

Egidio Forcellini. 1940. *Lexicon totius latinitatis*. Typis Seminarii, Padova.

Michele Fruyt. 2011. Word Formation in Classical Latin. J. Clarckson (ed.), *A Companion to the Latin Language*, Wiley-Blackwell, Chichester/Malden, Mass, 157–175.

Karl Ernst Georges and Heinrich Georges. 1913-1918. *Ausfuhrliches Lateinisch-Deutsches Handwôrterbuch*. Hahn, Hannover.

Peter GW. Glare. 1982. *Oxford Latin Dictionary*. At the Clarendon Press, Oxford.

John Goldsmith. 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, 27(2): 153–198.

Otto Gradenwitz. 1904. *Laterculi vocum latinarum*. Hirzel, Leipzig.

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

Paul Rockwell Jenks. 1911. *A manual of Latin word formation for secondary schools*. DC Heath & Company, Harvard.

Renato Oniga. 1988. *I composti nominali latini: una morfologia generativa* (Vol. 29). Pàtron, Bologna.

Marco Passarotti. 2004. Development and perspectives of the Latin morphological analyzer LEMLAT. *Linguistica Computazionale*, XX-XXI: 397-414.

Marco Passarotti and Francesco Mambrini. 2012. First Steps towards the Semi-automatic Development of a Wordformation-based Lexicon of Latin. *Proceedings of the Eighth International Conference on Language Resources and*

Evaluation (LREC'12). ELRA, Istanbul, Turkey, 852–859.

Magda Ševčíková and Zdeněk Žabokrtský. 2014. Word-Formation Network for Czech. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. ELRA, Reykjavik, Iceland, 1087–1093.

Luigi Talamo, Chiara Celata and Pier Marco Bertinetto. 2016. DerIvaTario: An annotated lexicon of Italian derivatives. *Word Structure*, 9(1): 72–102.

Cornelis Joost Van Rijsbergen. 1979. *Information retrieval*. Butterworths, London, 2nd edition.

Britta D. Zeller, Jan Snajder and Sebastian Padó. 2013. DErivBase: Inducing and Evaluating a Derivational Morphology Resource for German. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. ACL, Sofia, Bulgaria, 1201-1211.