First Steps towards the Semi-automatic Development of a Wordformation-based Lexicon of Latin

Marco Passarotti, Francesco Mambrini
CIRCSE Research Centre
Università Cattolica del Sacro Cuore
Largo A. Gemelli 1, 20123 Milan, Italy
marco.passarotti@unicatt.it, f.mambrini@gmail.com

Abstract

Although lexicography of Latin has a long tradition dating back to ancient grammarians, and almost all Latin grammars devote to wordformation at least one part of the section(s) concerning morphology, none of the today available lexical resources and NLP tools of Latin feature a wordformation-based organization of the Latin lexicon. In this paper, we describe the first steps towards the semi-automatic development of a wordformation-based lexicon of Latin, by detailing several problems occurring while building the lexicon and presenting our solutions. Developing a wordformation-based lexicon of Latin is nowadays of outmost importance, as the last years have seen a large growth of annotated corpora of Latin texts of different eras. While these corpora include lemmatization, morphological tagging and syntactic analysis, none of them features segmentation of the word forms and wordformation relations between the lexemes. This restricts the browsing and the exploitation of the annotated data for linguistic research and NLP tasks, such as information retrieval and heuristics in PoS tagging of unknown words.

Keywords: wordformation, Latin, lexicography

1. State of the Art

Lexicography of Classical languages has a long tradition, dating back to ancient Latin and Greek grammarians. Over many centuries, this has resulted in several resources, such as dictionaries, thesauri and lexica.

As far as Latin is concerned, some of the most relevant are the following, a number of which is today made available on-line:

- *Glossarium Ad Scriptores Mediae et Infimae Latinitatis* (du Cange, 1678, extended in 1766 by P. Carpentier)[1];
- *Lexicon totius latinitatis* (Forcellini, 1771; extended in 1896 by R. Klotz, G. Freund & L. Doderlein);
- *Ausführliches Lateinisch-Deutsches Handwörterbuch* (Georges, 1913-1918);
- *A Latin Dictionary* by Lewis & Short (1969)[2];
- *Oxford Latin Dictionary* (Glare, 1982);
- *Thesaurus Formarum totius latinitatis a Plauto usque ad saeculum XXum*, developed by the Centre 'Traditio Litterarum Occidentalium' (CTLO) in Turnhout (Tombeur, 1998)[3];
- *Thesaurus Linguae Latinae* (ongoing; presently, arrived at letter P) from the Bayerische Akademie der Wissenschaften in Munich[4];
- *Neulateinische Wortliste* by Johann Ramminger (2003 ff.)[5];
- Latin WordNet (Minozzi, 2008), integrated into Multi-WordNet[6];
- IT-VaLex (McGillivray & Passarotti, 2009), a valency lexicon built by induction from the *Index Thomisticus* Treebank data[7];
- the Dynamic Lexicon automatically built from the textual collection of the Perseus Digital Library (Bamman & Crane, 2009).

Latin morphology can be processed automatically with three available morphological analyzers. They are LEMLAT (Passarotti, 2004), Whitaker's *Words* and *Morpheus* (Crane, 1991), this latter being first developed for Ancient Greek in 1985 and extended to support Latin in 1996.

None of the aforementioned lexical resources features a wordformation-based organization of the Latin lexicon. This means that wordformation relations between lexical entries are not described in any available lexical resource of Latin.

The same limitation holds for morphological analyzers too, as all of them process the word forms without both providing segmentation of the wordformation affixes and establishing relations between the input and output lexemes of the wordformation rules (WFRs)[8]. Thus, while morphology is traditionally divided into inflection (formation of word forms of a lexeme) and wordformation (formation of new lexemes), the latter is not presently taken into account by any of the available Latin morphological analyzers.

In this paper, we describe the first steps towards the semi-automatic development of a wordformation-based lexicon of Latin. The paper is organized as follows:

[1] http://www.uni-mannheim.de/mateo/camenaref/ducange.html.
[2] http://www.perseus.tufts.edu.
[3] http://www.corpuschristianorum.org/centres/turnhout.html.
[4] http://www.thesaurus.badw.de/.
[5] www.neulatein.de.
[6] http://multiwordnet.itc.it/english/home.php.
[7] http://itreebank.marginalia.it/itvalex.
[8] According to Matthews (1974) and Aronoff (1994), our notion of lexeme is a word considered as an inflectional paradigm. The 'lemma', instead, is the citation form as usually reported in dictionaries.

section 2 provides the background motivation of building the lexicon; section 3 describes the overall organization of the work, the results and their evaluation; section 4 details several problems occurring while developing the lexicon and presents the solution of some specific issues by sampling two morphological families; section 5 discusses a number of issues that remain still open and draws general conclusions.

2. Motivation

Although almost all Latin grammars (especially the historical ones) devote to wordformation at least one part of the section(s) concerning morphology, no Latin dictionary is organized according to wordformation, and neither a complete nor a partial description of the Latin lexicon according to wordformation is today available.

Wordformation-based lexica are important language resources which allow to query the word forms of texts not only as independent lexical units (in a way similar to that provided by traditional dictionaries), but also according to worformation relations and WFRs.

Such lexica are relevant for NLP purposes, too. As a WFR is not only a mechanism to build new words, but it also creates new words that share a common (often predictable) semantic core, tasks like information retrieval and heuristics in PoS tagging of unknown words require wordformation-based lexica which, together with suffix-stripping algorithms for stemming, allow morphological anlyzers to perform segmentation of the formative elements of the words.

Developing a wordformation-based lexicon of Latin is nowadays of outmost importance, as the last years have seen a large growth of annotated corpora of Latin texts of different eras, like the Latin Dependency Treebank (Bamman & Crane, 2007), the *Index Thomisticus* Treebank (Passarotti, 2010) and the PROIEL corpus (Haug & Jøndal, 2008). While these corpora feature lemmatization, morphological tagging and syntactic analysis, they neither include segmentation of the word forms nor describe the wordformation relations between the lexemes, thus restrincting the way the annotated data can be browsed and exploited for linguistic research and NLP purposes.

Moreover, wordformation is a research field showing a huge amount of scientific literature, spread over a long time span (especially when dealing with ancient languages). This work has resulted in a number of different theoretical frameworks and approaches to wordformation. A wordformation-based large-scale exploration of a lexicon allows to evaluate and refine these theories by confrontation with the empirical evidence provided by data.

3. Contribution

The data of our lexicon are taken from the list of lemmas provided by the *Lexicon totius latinitatis* (Forcellini, 1771) made available by Busa (1988)[9]. The total number of lemmas in Forcellini is 92,052 (24,879 of which are registered in the onomasticon, i.e. the list of proper names). Our work has two main aims:

- to assign a WFR to each morphologically complex lexeme (i.e. a lexeme morphologically derived from another lexeme[10]);
- to link each morphologically complex lexeme to its parent lexeme(s).

Data are organized and represented according to the style of *Word Manager*, a system for morphological dictionaries available for German, English and Italian (Domenig & ten Hacken, 1992).

We conceive WFRs according to the so-called Item-and-Arrangement model (IA), which follows a morpheme-based approach to morphology. In IA, word forms are analyzed as arrangement of morphemes according to the following three axioms:

a) roots and affixes have the same status of morphemes (Baudoin's single morpheme hypothesis);
b) they are dualistic, as they have both a form and a meaning (Bloomfield's sign base morpheme hypothesis);
c) they are stored in the lexicon (Bloomfield's lexical morpheme hypothesis)[11].

3.1 Organization of the Work

The wordformation-based lexicon of Latin is built in a two-step fashion.

A)

Manual and data-driven finding of the WFRs:

1. lemmas are grouped into classes according to their PoS and inflectional category (declension and conjugation);
2. an 'incipitarium'[12] and an 'explicitarium'[13] of each class of lemmas is built;
3. candidate prefixal derivation rules are automatically extracted from the incipitarium; candidate suffixal rules are in turn derived from the explicitarium. The PoS and the inflectional category of the input lexeme(s) are manually assigned to each candidate rule;
4. WFRs are grouped into two main classes: (a) compounding and (b) derivational.
   Derivational rules are divided into two categories: (i) affixal and (ii) conversive[14]. Affixal rules are divided into prefixal and suffixal.

[9] In Busa (1988), each lexical entry is PoS tagged and assigned an inflexional category.

[10] In case of compounding, the input lexemes are more than one.

[11] The model alternative to IA is called Item-and-Process (IP). IP states that WFRs are processes that apply to a base which is modified by the rule to produce a new word. The two models are outlined by Hockett (1954).

[12] A standard alphabetical list in which lemmas beginning with the same characters are close to each other.

[13] An alphabetical list in which lemmas are ordered according to right-to-left reading. Thus, lemmas ending with the same characters are close to each another in the list.

[14] Conversion is a derivation process that does not include any affix. Conversive WFRs are manually definied.

B)
Application (and evaluation) of the WFRs resulting from (A), and creation of the "morphological families"[15]. New rules can be added in this phase by confrontation with data.
(B) is divided into two subtasks:

1. each complex lexeme is assigned a WFR.
   This task is performed in semi-automatic fashion by using a program that assigns to each (possibly) complex lexeme its most likely WFR according to the PoS of the lexeme and the string of its initial (prefixal rules) and final (suffixal rules) characters;
2. morphological families are built.
   Latin words may contain three kinds of morphemes: the root, one or more affixes, which are attached to a root to form a stem, and the inflectional ending (Palmer, 1954).
   Each simple lexeme is regarded as the possible ancestor of a morphological family [16]. The inflectional ending of the lemma of a simple lexeme is removed, in order to detect the string of characters that remains the same in the inflectional paradigm (which roughly corresponds to the root). For instance, the invariable part of the lemma *amo* is *am* (*am-o*).
   The same holds for the morphologically complex lexemes, with the difference that in these cases stems and not roots are concerned. This means that not only the endings, but also the affixes (listed in A) are removed from lemma. For instance, the invariable part of the lemma *amabilis* is *am* (*am-a-bil-is*).
   All those (simple, or complex[17]) lexemes that share the same invariable part are automatically assigned to the same morphological family.
   Finally, the members of each family are automatically linked with each other according to their PoS, inflectional category and affixes by means of the WFR assignment (B.1). The simple lexeme member is assigned the role of ancestor of the family. Given the high number of homographs in Latin, this automatic procedure is regarded as non-ultimate for building the morphological families. However, it is helpful as it provides filtered data that must be checked manually[18].

3.2 Results and Evaluation

3.2.1 Finding and Writing the Rules
The automatic procedure described in (A) found 118 different WFRs. Each rule includes the following information:

- PoS of the input lexeme (or lexemes, in case of compounding WFRs);
- PoS of the output lexeme;
- class of the rule (see (A.4) above);
- input root/stem;
- an alphanumeric key that identifies the rule;
- the prefix (optional);
- the inflexional category of the input lexeme(s);
- the thematic vowel (optional);
- the suffix (optional);
- the inflexional category of the output lexeme;
- one example.

For instance, the WFR that derives nouns in *-mentum* from verbs is registered as follows:

- PoS of the input lexeme: verb;
- PoS of the output lexeme: noun:
- class of the rule: derivational-suffixal;
- input root/stem: present infinitive;
- an alphanumeric key that identifies the rule: rule 50;
- no prefix;
- the inflexional category of the input lexeme: verb of any (regular or irregular) Latin conjugation;
- the thematic vowel:
  - if the input verb is of first conjugation: *-a-*;
  - if the input verb is of second conjugation: *-e-*;
  - if the input verb is of third/fourth/irregular conjugation: *-i-*;
- the suffix: *-ment-*;
- the inflexional category of the output lexeme: second declension neuter ending in *-um*;
- one example: *imit-o* > *imit-a-ment-um*.

3.2.2 Applying the Rules
So far, we have applied to the list of lemmas of Forcellini 23 of the 118 WFRs found. This led to the automatic tagging of 6,720 morphologically complex lexemes.
These 23 rules were chosen among the "simplest" ones, i.e. those showing the highest morphological transparence and, thus presenting less problems in the automatic finding of the input-output relations. For instance, the deverbal WFRs chosen are only those that take the stem of the regular present infinitive in input, which is easy to detect by just removing the ending of the lemma (for instance: *am-o* > *am-*).
All the 23 WFRs are derivational. 19 out of them are of the Verb-to-Verb type (all prefixal); 2 are Noun-to-Adjective, 1 is Noun-to-Noun and 1 is Verb-to-Noun (all suffixal).
The 19 Verb-To-Verb WFRs are those involving the following prefixes (the number of complex lexemes formed by each WFR is reported in parenthesis): *ab-* (69), *ad-* (150), *circum-* (158), *con-* (614), *de-* (412), *dis-* (89), *inter-* (117), *intro-* (15), *ob-* (156), *per-* (307), *prae-* (253), *praeter-* (20), *pro-* (137), *re-* (379), *retro-* (9), *sub-* (173), *subter-* (20), *super-* (179), *trans-* (62). All these WFRs form a new verb belonging to the same conjugation of the

[15] By "morphological family" we mean the set of lexemes morphologically derived from one common ancestor-lexeme.
[16] In principle, we consider simple lexemes all those not assigned a WFR in (B.1).
[17] WFRs do not take as input morphologically simple lexemes only, but also morphologically complex ones. For example, the noun *excubatio* derives by suffixation from the verb *excubo*, which is morphologically complex, as it is derived (by prefixation) from the verb *cubo*.
[18] Relations between lemmas are defined not on a etymological basis, but according to derivational-morphological criteria only (ten Hacken & Smyk, 2002).

base-lexeme. For instance, *abduco* (*ab-duc-o*) is a third conjugation verb derived from a base-verb of the third conjugation (*duco*: *duc-o*).
Table 1 reports the first 11 lines of the input and output lexemes automatically assigned the V-To-V WFR with the prefix *ab-*. For each input and output lexeme, the table reports the stem and a label formed by two tags: the first tag informs about the conjugation (J: first, K: second; L: third; M: fourth), the second about the PoS (in this case, only the tag A is concerned: verb). The lemma is automatically produced by adding an ending according to the conjugation: *-o* in case of first and third conjugation, *-eo* for the second one, and *-io* for the fourth. For instance, in table 1 there are two input stems *dic-*: one is the stem of the verb *dico, -are* (JA), while the other is the stem of the verb *dico, -ere* (LA). The same distinction is retained in the corresponding output lexemes: *abdico, -are*, and *abdico, -ere*.

Input_stem	Input_PoS	Output_stem	Output_PoS
aestim	JA	abaestim	JA
alien	JA	abalien	JA
brevi	JA	abbrevi	JA
dic	JA	abdic	JA
dic	LA	abdic	LA
duc	LA	abduc	LA
em	LA	abem	LA
equit	JA	abequit	JA
err	JA	aberr	JA
horr	KA	abhorr	KA
horresc	LA	abhorresc	LA

Table 1. First 11 lines of the V-To-V (prefix *ab-*) WFR

The 2 Noun-To-Adjective WFRs concern the following suffixes: *-ic-us* (*naut-a* > *naut-ic-us*) (288), and *-os-us* (*tenebr-a* > *tenebr-os-us*) (473).
The only Verb-To-Noun WFR deals with the *-io/-ion-is* suffix (*possid-eo* > *possess-io/-ion-is*) (2,626).
Finally, the Noun-To-Noun WFR collects those nouns formed with the suffix *-uncul-a/-us* (*ran-a* > *ran-uncul-a*; *lemb-us* > *lemb-uncul-us*) (14).
The precision rate of the WFRs application to data is generally high (ranging from 100% to 95.7), while the recall is lower and shows wide variability (from 97.8% to 63.2%). This means that usually WFRs are automatically assigned to the correct lexemes and that input-output relations are well detected. However, the automatic assignment does cover a quite low percentage of the total of the lexemes and relations involved by a WFR. As reported in the next section, this is due to several reasons, among which are graphical variations in the inflexional paradigm of the lexemes.

4. Discussion

The development of the lexicon is at its very beginning, as we just started to face the simplest WFRs. Many issues remain still open.
One traditional aspect affecting wordformation concerns several restrictions on WFRs, whose application on data can result in overgeneration of outputs. However, while such constraints (dealing with syntax, semantics, morphology and phonology) can affect the productivity of WFRs, this is not a problem for our wordformation-based lexicon, as WFRs are applied only if a candidate input/output stem is found in Forcellini, thus preventing overgeneration.
One important open issue concerning the overall organization of the work is that we are not aware of the exact ratio of the morphologically simple and morphologically complex lexemes present in Forcellini. Our purpose is to refine the data by tagging automatically the highest number of complex lexemes as possible, by using data-driven WFRs. These rules must be of increasing complexity and able to manage wordformation issues that are well documented in literature, such as the following:

- stem change involving internal vowel alternation (apophony): *fac-io* > *per-fic-io*;
- assimilation of the prefix (the sound of the ending of the prefix becomes similar to the sound of the beginning of the following word): *fer-o* > **ob-fer-o* > *of-fer-o*;
- derivation from the stem of the genitive of the imparisyllaba nouns and adjectives of the third declension: *crimen* (gen. *crimin-is*) > *crimin-al-is*;
- unclear segmentation: a word like *creator* can be segmented either *cre-a-tor*, or *cre-at-or*, according to which form of the suffix is chosen (*-tor* vs. *-or*) (Scalise, 1996; see 4.1.2 below);
- cases where the boundary between compounding and derivation is not fully clear. For instance, in the complex lexeme *primiformis*, *primi-* can be regarded both as a lexeme (compounding), or as a prefix (derivation) (see 4.1.1 below);
- complex lexemes including both a prefix and a suffix raise the problem of determining the base-lexeme. One example is the complex lexeme *reclamatio*. The root of the morphological family of *reclamatio* is the verb *clamo*. From *clamo* the noun *clamatio* is derived by suffixation, and the verb *reclamo* by prefixation. The lexeme *reclamatio* can, thus, derive from either *reclamo* (by suffixation) or *clamatio* (by prefixation) (see 4.1.1 below);
- one lexeme that is required in the wordformation chain is missing in the dictionary. For instance, the noun *insuasibilitas* derives either from *suasibilitas* (by prefixation), or from *insuasibilis* (by suffixation). However, neither *suasibilitas*, nor *insuasibilis* are lemmas reported by Forcellini. Thus, either we add a fictional entry in the wordformation chain, or we register *insuasibilitas* as formed by two WFRs at the same time.

In order to discuss into more details some of these problems, we report below a number of relevant cases extracted from two morphological families.

4.1 Samples from Two Morphological Families

4.1.1 Forma

According to Forcellini, the morphological family of the ancestor-noun *forma* includes 91 complex lexemes, among which 52 are adjectives, 25 nouns, 9 verbs and 5 adverbs[19].

In order to describe how we deal with some of the problematic issues related to the building of the wordformation-based lexicon, we report those 7 lexemes of the *forma* family that are formed with the prefix *de-*: *deforma* (N), *deformatio* (N), *deformis* (A), *deformitas* (N), *deformo* (V), *deformosus* (A), *deformus* (A).

First, building the wordformation relations among these lexemes arises several problems about choosing the correct base-lexemes. For instance, the noun *deformatio* can be derived either from the noun *formatio* (by prefixation) or from the verb *deformo* (by suffixation). Equally, the base-lexeme of *deformitas* can be either the noun *formitas*, or the adjective *deformis*.

In such cases, we follow a semantic-based approach, by looking at the meaning of the derived lexeme. As the meaning of *deformatio* is "the act of deforming something", *deformatio* is considered as derived from the verb *deformo* (to deform) instead than from the noun *formatio* (formation).

The same holds for *deformitas*, which means "the property of being deformed": *deformitas* is, thus, derived from *deformis* (deformed) instead than from *formitas* (shaping, fashioning, forming).

As a general guideline, in case of lexemes including both a prefix and a suffix, we consider prefixation as acting before suffixation along the wordformation chain (*formo* > *deformo* > *deformatio*). This is not only due to semantic aspects, but also to the need of collecting all the words formed with a common prefix under one common base-lexeme.

Another problematic issue concerns the form of the suffix in a lexeme like *deformatio*, which can be segmented in two different ways:

(a) *de-form-a-tio*: the base is the stem of the present infinitive (*deform-*) and the suffix is *-tio*, preceded by the thematic vowel of the first conjugation verbs (*-a-*);
(b) *de-form-at-io*: the base is the stem of the supine (*deform-at-*) and the suffix is *-io*[20].

Looking at the overall building of the lexicon, choosing which segmentation is working in *deformatio* means to decide if we want to follow an historical-based or a graphical-based approach while describing the wordformation relations.

In the lexicon there are several cases of third declension nouns ending in *-io* where the ending is graphically attached to the root/stem of the irregular supine. For instance, the noun *inclusio* looks like derived from the stem of the irregular supine of the verb *includo* (*inclus-*). Indeed, the historical formation of the noun *inclusio* results from the attachment of the suffix *-tio* to the stem of the present infinitive of *includo* (*includ-*): **includ-tio*. Like in the stem of the supine (**includ-t-um* > *inclus-um*), the final form *inclusio* thus results from a number of phonological modifications.

This means that solution (a) is more correct than (b) if historical morphology is concerned[21], because the correct form of the suffix is *-tio*, and not *-io*. However, solution (a) is less economic than (b) for what concerns the overall organization of the lexicon. Indeed, solution (a) requires the following:

- WFR: suffixation V-To-N; input root/stem: present infinitive; optional thematic vowel (missing in case of athematic roots: see 4.1.2 below, about *traductor*);
- suffix: *-tio* with a graphical variant *-io*.

Further, solution (a) requires to add a graphical variant of the root/stem of the present infinitive to those verbs showing irregular supine, like *includo*: *includ-* (stem of the regular present infinitive) and *inclus-* (graphical variant of *includ-* used in some wordformation processes).

Solution (b) is more economic than (a) because it does not require to add any variant of both the suffix and the present infinitive root/stem of certain verbs to the lexicon:

- WFR: derivation V-To-N; input root/stem: supine;
- suffix: *-io*.

In order to decide which solution (and, thus, which approach) to choose, our starting point is the correct form of both the WFR and the affix. In this case, the correct WFR that produces words like *deformatio* and *inclusio* is that described by solution (a), because the suffix indeed attaches to the root/stem of the present infinitive of the input verbs. Further, the original form of the suffix is *-tio*, while *-io* is a graphical variant of *-tio* that can be motivated according to phonological modifications.

This problem is raised by many other WFRs, one of which is discussed below about the word *traductor* (see 4.1.2).

The relations between the 7 lexemes reported above result as follows:

- *deforma*: *de-form-a* < *de-form-o*. WFR: conversion V-To-N;
- *deformatio*: *de-form-a-tio* < *de-form-o*. WFR: suffixation V-To-N;
- *deformis*: *de-form-is* < *form-is*. WFR: prefixation A-To-A;
- *deformitas*: *de-form-itas* < *de-form-is*. WFR: suffixation A-To-N;
- *deformo*: *de-form-o* < *form-o*. WFR: prefixation V-To-V;

[19] Forcellini assigns a separate entry to some adverbs derived from adjectives. In the morphological family of *forma*, the adverbs holding a separate entry are the following: *ambiformiter* (no entry for the adjective *ambiformis* is provided), *deformiter* (< *deformis*), *informiter* (< *informis*), *multiformiter* (< *multiformis*) and *uniformiter* (< *uniformis*). In our wordformation-based lexicon, adverbs derived from adjectives do not receive a separate entry on their own, but they are included into the inflectional paradigm of the base-adjective.

[20] The affix *-at-* can be further segmented in two parts: the thematic vowel (*-a-*) and the supine affix (*-t-*).

[21] Historical morphology and etimology are different aspects of wordformation and their boundaries must be carefully considered while building the lexicon.

- *deformosus*: *de-form-os-us* < *de-form-is*. WFR: suffixation A-To-A;
- *deformus*: *de-form-us* < *form-us*. WFR: prefixation A-To-A.

The word *deformitas* is another example of unclear segmentation of the formative elements, as *deformitas* can be segmented either *de-form-itas*, or *de-form-i-tas*.

In the former solution, *-i-* (*-itas*) is considered part of the suffix, while in the latter *-i-* (*-i-tas*) is a linking element between the lexical root/stem and the suffix itself (*-tas*).

The second solution allows to manage cases of variation of the linking vowel, like *empietas* (*em-pi-e-tas*), where an *-e-* instead of an *-i-* appears. Moreover, choosing the form *-tas* for the suffix allows to collect under one common WFR (A-To-N+*-tas*) both cases like *deformitas* and *empietas*, without separating them in two different WFRs, namely A-To-N-*itas* and A-To-N-*etas*.

Nonetheless, we chose the *-itas* solution, because the vowel *-i-* in *-itas* is part of the suffix itself and it is not a thematic vowel. The alternation *-i-/-e-* (*-itas/-etas*) is due to phonetic reasons, as *-e-* substitutes *-i-* in those cases where the root/stem of the input adjective ends with an *-i-* (*empi-/empi-etas*). Thus, the suffix is recorded in the lexicon with two possible graphical variants: *-itas* (regular) and *-etas*, the latter occurring with roots/stems ending in *-i-*.

Another problematic issue raised by several lexemes concerns the boundary between compounding and derivation[22]. Beside complex lexemes that are clearly compounds (*serpenti+formis*, *tauri+formis*) or clear derivations by affixation (*re-formo*, *trans-formo*), there is a number of cases where choosing between compounding and derivation is not trivial (*biformis, primiformis*).

In this respect, we follow a lexicalist approach, by distinguishing between lexemes and affixes, and considering affixation as a kind of derivation and not as a compounding mechanism (Aronoff, 1976; Scalise, 1984). In order to distinguish between lexemes and affixes, we apply the following tests:

- when a word can appear both as the first and the second element in a complex lexeme, it is a lexeme. For instance, *primus* appears as first element in *primiformis*, and as second element in *undecimprimus*;
- when a word can be used as input of a suffixation WFR, it is a lexeme. For instance, *primus* is the base lexeme of *primitivus*, which is formed by attaching the suffix *-i-tiv-* to the root of *primus* (*prim-*);

According to these tests, we consider as a prefix the formative element *bi-* in the word *biformis*. Thus, this word is registered in the lexicon as formed by a prefixation WFR.

In turn, *primiformis* is considered as a compound of an adjective (*primus*) plus a noun (*forma*).

[22] Determining the boundaries between compounding and derivation is a topic widely discussed in literature. Among the several papers about this issue, we mention here ten Hacken (2000) and Booij (2005).

4.1.2 Duco

In Forcellini, the morphological family of the ancestor-verb *duco* includes 188 complex lexemes, among which 83 are nouns, 63 adjectives, 36 verbs and 6 adverbs.

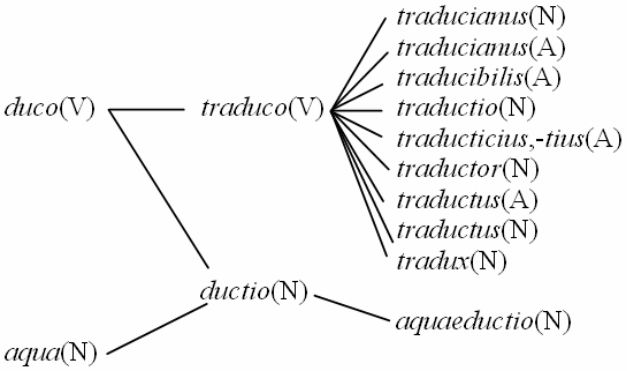

Figure 1. Part of the family of *duco*

Some of the lexemes reported in figure 1 raise discussion about their segmentation and the form of their formative elements:

- *traductor*: like for *traductio*, which is formed in the same way as *inclusio* (see 4.1.1 above), the suffix occurring in *traductor* can be analysed in two different ways:
  (a) *-tor* attaching to the root/stem of the present infinitive of the input verb: *tra-duc-tor*;
  (b) *-or* attaching to the root/stem of the supine of the input verb: *tra-duct-or*.
  Like for the suffix *-tio*, solution (a) is more correct according to historical grammar, as the form of the suffix is indeed *-tor*. The suffix can be attached to the root/stem of the present infinitive of the input through a thematic vowel (*-a-/-i-*: *elimin-a-tio*, *cred-i-tor*) or not, like in cases of athematic roots (*tra-duc-tio*). This can yeld to graphical variations of the suffix (*-tor/-or*), as for instance in *possessor*, resulting from **possid-tor*.
  Solution (b) is more economic and reflects a graphical-based approach to the development of the lexicon.
  Following solution (a) implies that the derivation of *possessor* is made through adding in the lexicon a graphical variant of both the suffix (*-tor/-or*) and the root of the present infinitive of *possideo* (*possess-*). Instead, according to solution (b), the suffix is attached to the root/stem of the supine (like in *tra-duct-or*) and no graphical variant must be added in the lexicon.
  For the same reasons reported above about the suffix *-tio*, we follow the solution (a), which is more correct, although less economic;
- *tradux*: like *dux* (from *duco*), *tradux* is a noun formed by conversion from the verb *traduco*. By considering *dux* as produced by a conversion process and not by a derivational one, we follow the analysis of this kind of nouns proposed by Palmer (1954), who considers *dux* a "root noun" showing a zero suffix: *dux* < *dŭc-s*;

- *traducibilis*: as a general guideline, we keep trace of the thematic vowel in the segmentation process. The adjective *traducibilis* is thus segmented *tra-duc-i-bil-is* (instead of *tra-duc-ibil-is*). The thematic vowel *-i-* is used for those adjectives derived from verbs of the second, third and fourth conjugation, while *-a-* appears in those derived from verbs of the first conjugation (*am-a-bil-is*).
  In this case, we keep the thematic vowel separated from the root not only because they are indeed two different parts of the word, but also in order to collect under one common WFR (V-To-N+*-bil*) all those deverbal adjectives featuring the suffix *-bil-*. Otherwise, if the thematic vowel were considered part of the suffix itself, this would result in two different suffixes (*-abil* and *-ibil*), which cannot be regarded as graphical variants of the same suffix (as for *itas/-etas*).
  This would lead to the consequence of registering words that are indeed produced by one common rule under two different WFRs;
- *aquaeductio*: this is a compound noun, which thus belongs to two different morphological families, i.e. that of *duco* (verb) and that of *aqua* (noun).
  The compounding WFR (N+N-To-N) states that the output noun is formed by attaching the genitive form of the first noun (*aquae* is the genitive of *aqua*) to the second noun (*ductio*), whose inflection class and paradigm is kept in the output word.
  We developed this WRF according to the evidence provided by several words formed with the same structure of *aquaedutio*. See for instance: *auricaesor* (*aurum+caesor*) and *linitextor* (*linum+textor*).

4.2 Georges and LEMLAT

Although Forcellini is the Latin dictionary that comprises the highest number of lemmas, and the only one providing an onomasticon, Lomanto (1980) demonstrates that Georges (1913-1918) shows both a higher lexical richness and a better quality of the lexical entries. Thus, following Lomanto, we want to collate the Forcellini lexicon with that provided by LEMLAT, which includes all the lexical entries of Gradenwitz (1904), Georges (1913-1918) and Glare (1982), for a total of 40,014 lemmas.
Another reason in favour of the use of LEMLAT for our aims, is that LEMLAT includes every different string of characters that is required in the inflectional paradigm of each lexeme, but that is not automatically produced by a rule. For instance, LEMLAT provides the uninflected parts of irregular supines (*duc-*, *duct-*) and the stem of the genitive of imparisyllaba nouns and adjectives (*crimen*, *crimin-*). Moreover, LEMLAT manages automatically many graphical variants, like *obf-/off-* in *offero*.

5. Open Issues and Conclusions

The main issue concerning the development of a wordformation-based lexicon of Latin is the way the boundary between diachronic and synchronic morphology is managed. In several cases, it is not easy to decide about which is the correct segmentation of the complex words and, thus, about the form itself of the affixes. Moreover, while NLP-oriented researchers build lexica for such tasks as word-sense disambiguation or topic classification, traditional humanists (like classicists) are interested in the way lexemes themselves are registered in the lexicon, according to a wide literature spread over many centuries.
Thus, roughly speaking, our general guideline is to register the lexical entries as most correctly as possible and in a way such that they can be retrieved without ambiguities. This leads to generally favour solutions that reflect the real wordformation process instead of just accounting for the synchronic graphical appearance of the complex lexemes. Our aim is to avoid to add wrong affixes or incorrect WFRs to the lexicon, just to easily face the graphical form of the lexical entries. The case of *-tor/-tio* vs. *-or/-io* discussed above is representative of the different criteria that must be taken into account when dealing with issues concerning the development of a wordformation-based lexicon of an ancient language.
A strong desideratum of the lexicon is, thus, the writing of clear guidelines where how each WFR was designed in terms of segmentation of the lexemes and form of the affixes is explained and reasoned out in full.
Making the lexicon accessible through an on-line interface is a another requirement of the project. The interface must feature at least the following ways to access the lexicon:

- by single lexical entry, providing the possibility of looking at both all its derived words and its ancestor(s);
- by morphological family;
- by WFR, according to WFR class, affix, input and output PoS and inflectional category.

Once all the WFRs that can be induced from data are applied and evaluated, we must look at those lexemes to which no WFR is assigned, in order to distinguish those that are indeed morphologically simple lexemes from those complex lexemes that are not detected by automatic tagging. We hope that this will allow us to find new WFRs, that are able to include also these lexemes. However, a certain amount of manual hard-coding of lexemes produced by complex, or poorly productive WFRs will be finally required.
In the near future, we would like to assign to each WFR a prototypical semantic description of its input and output lexemes and evaluate them by comparison with the information provided by Latin WordNet. Indeed, in our purpose the wordformation-based lexicon must become a new lexical resource of Latin that interacts with others already available, such as Latin WordNet, IT-VaLex and the Perseus Dynamic Lexicon. These lexica must be linked each other, to result in one common lexical resource of Latin providing information about inflection, wordformation, valency and semantics of the Latin words.

6. References

Aronoff, M.H. (1976). *Word formation in generative grammar*. Cambridge (Mass.): MIT Press.

Aronoff, M.H. (1994). *Morphology by Itself: Stems and Inflectional Classes*. Cambridge (Mass.): MIT Press.

Bamman, D. & Crane, G. (2007). The Latin Dependency Treebank in a cultural heritage digital library. In *Proceedings of LaTeCH 2007. Prague, Czech Republic*, pp. 33--40.

Bamman, D. & Crane, G. (2009). Computational Linguistics and Classical Lexicography. *Digital Humanities Quarterly*, 3(1). http://digitalhumanities.org/dhq/vol/3/1/000033/000033.html.

Booij, G. (2005). Compounding and derivation: evidence for Constructional Morphology. In W.U. Dressler, P. Kastovsky & F. Rainer (Eds.), *Morphology and its demarcations*. Amsterdam: Benjamins, pp. 109--132.

Busa, R. (1988). *Totius latinitatis lemmata*. Milano: Istituto Lombardo - Accademia di Scienze e Lettere.

Crane, G. (1991). Generating and Parsing Classical Greek. *Literary and Linguistic Computing*, vol. 6, n. 4, pp. 243--245.

Domenig, M. & ten Hacken, P. (1992). *Word Manager: A system for morphological dictionaries*. Hildesheim: Olms.

Du Cange, C. (1678). *Glossarium ad scriptores mediae et infimae latinitatis*. Paris.

Forcellini, A. (1771). *Totius Latinitatis lexicon, consilio et cura Jacobi Facciolati opera et studio Aegidii Forcellini, lucubratum*. Patavii: typis Seminarii, 4 voll., 1771.

Georges, K.E. (1913-1918). *Ausführliches Lateinisch-Deutsches Handwörterbuch*, Hannover: Hahn.

Glare, P.G.W. (1982). *Oxford Latin Dictionary*. Oxford: At the Clarendon press.

Gradenwitz, O. (1904). *Laterculi Vocum Latinarum*, Leipzig: Hirzel.

Haug, D.T.T. & Jøndal, M.L. (2008). Creating a Parallel Treebank of the Old Indo-European Bible Translations. In *Proceedings of LaTeCH Workshop - LREC 2008. Marrakech, Morocco*, pp. 27--34.

Hockett, C.F. (1954). Two Models of Grammatical Description. *Words*, 10, pp. 210--231.

Lewis, C.T. & Short, C. (1969). *A Latin Dictionary*. Oxford: At the Clarendon press.

Lomanto, V. (1980). Lessici latini e lessicografia automatica. In *Memorie dell'Accademia delle Scienze di Torino*, 5.4.2, pp. 111--269.

Matthews, P.H. (1974). *Morphology: An Introduction to the Theory of Word Structure*. Cambridge: Cambridge University Press.

McGillivray, B. & Passarotti, M. (2009). The Development of the *Index Thomisticus* Treebank Valency Lexicon. In *Proceedings of LaTeCH-SHELT&R Workshop 2009, Athens, March 30, 2009*.

Minozzi, S. (2008). La costruzione di una base di conoscenza lessicale per la lingua latina: Latinwordnet. In *Studi in onore di Gilberto Lonardi*. Verona: Fiorini, pp. 243--258.

Palmer, L.R. (1954). *The Latin Language*. London: Faber and Faber.

Passarotti, M. (2004). Development and perspectives of the Latin morphological analyser LEMLAT. In A. Bozzi, L. Cignoni & J.L. Lebrave (Eds.), *Digital Technology and Philological Disciplines*. *Linguistica Computazionale*, XX-XXI, pp. 397--414.

Passarotti, M. (2010). Leaving Behind the Less-Resourced Status. The Case of Latin through the Experience of the *Index Thomisticus* Treebank. In *7th SaLTMiL Workshop on Creation and Use of Basic Lexical Resources for Less-Resourced Languages, LREC 2010, La Valletta, Malta, 23 May 2010,* pp. 27--32.

Ramminger, J. (2003 ff.). *Neulateinische Wortliste. Ein Wörterbuch des Lateinischen von Petrarca bis 1700*. www.neulatein.de.

Scalise, S. (1984). *Generative morphology*. Dordrecht: Foris.

Scalise, S. (1996). Preliminari per lo studio di un affisso: -tore o -ore. In P. Benincà, G. Cinque, T. De Mauro & N. Vincent (Eds.), *Italiano e dialetti nel tempo. Saggi di grammatica per Giulio C. Lepschy*. Roma: Bulzoni, pp. 291--307.

ten Hacken, P. (2000). Derivation and Compounding. In G. Booij, C. Lehmann & J. Mugdan (Eds.), *Morphologie - Morphology: Ein Handbuch zur Flexion und Wortbildung - A Handbook on Inflection and Word Formation*. Berlin: Walter de Gruyter, pp. 349--360.

ten Hacken, P. & Smyk, D. (2002). Word Formation versus Etymology in Electronic Dictionaries. In A. Braasch, & C. Povlsen (Eds.), *Proceedings of the Tenth Euralex International Congress, Copenhagen - Denmark, August 13-17, 2002*, pp. 221--230.

Tombeur, P. (Ed.) (1998). *Thesaurus formarum totius latinitatis a Plauto usque ad saeculum XXum*. Turnhout: Brepols.