Enhancing the Latin Morphological Analyser LEMLAT with a Medieval Latin Glossary

**Flavio Cecchini*, Marco Passarotti*, Paolo Ruffolo*, Marinella Testori*,
Lia Draetta°, Martina Fieromonte°, Annarita Liano°, Costanza Marini°, Giovanni Piantanida°**
*Università Cattolica del Sacro Cuore - °Università di Pavia
*Largo Gemelli 1, 20123 Milan, Italy - °Corso Strada Nuova 65, 27100 Pavia, Italy
{flavio.cecchini, marco.passarotti}@unicatt.it

Abstract

English. We present the process of expanding the lexical basis of the Latin morphological analyser LEMLAT with the entries from the Medieval Latin glossary Du Cange. This process is performed semi-automatically by exploiting the morphological properties of lemmas, a previously available word list enhanced with inflectional information, and the contents of the lexical entries of Du Cange.

Italiano. *L'articolo descrive il processo di ampliamento della base lessicale dell'analizzatore morfologico per il latino* LEMLAT *con il glossario di latino medievale Du Cange. Il processo è realizzato semiautomaticamente ricorrendo ad alcune proprietà morfologiche dei lemmi, a un lemmario completo d'informazione flessionale e ai contenuti delle entrate lessicali del Du Cange.*

1 Introduction

Latin raises particular challenges for Natural Language Processing (NLP). Given that accuracy rates of stochastic NLP tools heavily depend on the training set on which their models are built, this becomes a particularly problematic issue when Latin is concerned, because Latin texts show an enormous linguistic variety resulting from (a) a wide time span (covering more than two millennia), (b) a large set of genres (ranging from literary to philosophical, historical and documentary texts) and (c) a big diatopic diversity (spread all over Europe and beyond).

Such complexity impacts NLP to the point that building NLP tools claiming to be suitable for all Latin varieties is an unrealistic task. One practical example comes from an experiment described by Ponti and Passarotti (2016), who show that the performance of a dependency parser trained on Medieval Latin data drops dramatically when the same trained model is applied to texts from the Classical era.

This issue affects all layers of linguistic annotation, including fundamental ones, like lemmatisation and morphological analysis. Today, a handful of morphological analysers are available for Latin, chiefly Words,[1] LEMLAT 3.0,[2] Morpheus[3] –reimplemented in 2013 as Parsley[4]–, the PROIEL Latin morphology system[5] and LatMor.[6]

Although LEMLAT, together with LatMor,[7] has proved to be the best performing morphological analyser for Latin and the one boasting the largest lexical basis, its lexical coverage is still limited to Classical and Late Latin only. First released as a morphological lemmatiser at the end of the 1980s at ILC-CNR in Pisa (Bozzi and Cappelli, 1990; Marinone, 1990, v 1.0), where it was enhanced with morphological features between 2002 and 2005 (Passarotti, 2004, v 2.0), LEMLAT relies on a lexical basis resulting from the collation of three Latin dictionaries (Georges and Georges, 1913 1918; Glare, 1982; Gradenwitz, 1904) for a total of 40 014 lexical entries and 43 432 lemmas, as more than one lemma can be included in one lexical entry. This lexical basis was further enlarged in version 3.0 of LEMLAT by semi-automatically adding most of the Onomasticon (26 415 lemmas out of 28 178) provided by the 5th edition of the Forcellini dictionary (Budassi and

[1] http://archives.nd.edu/words.html

[2] www.lemlat3.eu. Binaries and database available at https://github.com/CIRCSE/LEMLAT3.

[3] https://github.com/tmallon/morpheus

[4] https://github.com/goldibex/parsley-core

[5] https://github.com/mlj/proiel-webapp/tree/master/lib/morphology

[6] http://cistern.cis.lmu.de

[7] For an evaluation of morphological analysers for Latin see (Springmann et al., 2016).

Passarotti, 2016).

In order to equip LEMLAT to process Latin texts beyond the Classical period, we recently enhanced its lexical basis with the lexical entries from a large reference glossary for Medieval Latin, namely the *Glossarium Mediae et Infimae Latinitatis* by Du Cange et alii (1883 1887, hereafter DC). This paper details the process performed to include DC in LEMLAT's lexical basis.

2 Word Form Analysis in LEMLAT

LEMLAT is a lemmatiser and morphological analyser of types (i. e. no contextual disambiguation is performed). Given a word form in input (e. g. *coniugae*), LEMLAT's output produces the corresponding lemma(s) (e. g. *coniuga* 'wife') and a number of tags conveying (a) the inflectional paradigm of the lemma(s) (e. g. first declension noun) and (b) the morphological features of the input word form (e. g. feminine singular genitive and dative; feminine plural nominative and vocative).

LEMLAT makes use of a database that includes multiple tables recording the different formative elements (segments) of word forms. The core table is the lexical look-up table, whose basic component is the so-called LES (LExical Segment). The LES is defined as the invariable part of the inflected form (e. g. *coniug* for *coniug-ae*). In other words, the LES is the string (or one of the strings) of characters that remains the same in the inflectional paradigm of a lemma; hence, the LES does not necessarily correspond to either the word stem or the root.

LEMLAT includes a LES archive, in which LES are assigned an ID and a number of inflectional features, among which a tag for the gender of the lemma (for nouns only) and a code (called CODLES) for its inflectional category. According to the CODLES, the LES is compatible with the endings (called SF, "Final Segment") of its inflectional paradigm, which are collected in a separate table in the LEMLAT database. For example, the CODLES for the LES *coniug* is N1 (first declension nouns) and its gender is F (feminine). The word form *coniugae* is thus analysed as belonging to the LES *coniug*, the segment *ae* being recognised as an ending compatible with a LES with CODLES N1.

3 Adding the Du Cange Glossary

Adding DC to LEMLAT is a challenging task mostly because DC is not a dictionary in the modern sense of the word, but a glossary, i. e. a mere collection of words where information about parts of speech (PoS) and inflectional categories is almost absent, and therefore has to be deduced or reconstructed before an entry can be included in LEMLAT.[8] In addition, lemmatisation criteria are often inconsistent, even for words belonging to the same class (e. g. verbs are cited either by their present active infinitive or by their first person singular present indicative).

This is partly due to the fact that five different authors contributed to the glossary over a period of two centuries (Géraud, 1839), not always coherently with respect to their predecessors. Nonetheless, it is possible to distinguish some recurring patterns, which can be exploited to automatically include in LEMLAT as many of the 85 999 lemmas in DC as possible, or at least to expedite the manual recording of lexical entries.

3.1 Suffixes and Bon's Word List

The preliminary step to extend LEMLAT with DC consists in selecting a set of derivational suffixes that are morphologically-unambiguous in terms of PoS and inflectional category, and hence the set of all lemmas displaying these suffixes. These lemmas require no further analysis for entry in LEMLAT. Examples are *-itas* for feminine imparysillabic third declension nouns, or *-icum* for neuter second declension nouns. On the contrary, suffixes like, e. g. *-anus* or *-atus* are considered morphologically-ambiguous, as they can belong to different PoS (adjective or noun) and/or different inflectional categories (first or fourth declension). In these cases the corresponding lemmas require manual annotation (see Section 3.2). Approximately 30 000 DC lemmas are retrieved and added to LEMLAT in this way.

To extend the automatic acquisition of DC's lemmas, we also take advantage of a list of 71 908 Latin lemmas collected by Bruno Bon from various lexicographic sources and corpora.[9] This list supplies information about inflectional morphology.[10] Of these lemmas, 22 628 are found among

[8] For this work, we use the digital version of DC provided by the École nationale des chartes (Paris). Source data are available in XML format at `http://svn.code.sf.net/p/ducange/code/xml/`. The glossary can be accessed online at `http://ducange.enc.sorbonne.fr/`.

[9] Available at `http://glossaria.eu/outils/lemmatisation/` and presented in (Bon, 2011).

[10] Specifically: PoS; genitive endings of nouns; nominative

those in DC that are not analysed in the preliminary step; and out of these, 21 805 showing a one-to-one correspondence with lemmas in Bon's list are added to LEMLAT with no further check.[11]

3.2 Definitions and Quotations

Each lexical entry in DC comprises (a) the name of the lemma, (b) usually, a short **definition** and (c) possibly one or more **quotations** (taken from explicitly-cited textual sources), where most of the times a form of the lexical entry is capitalised. By making use of all these elements, we automatically assign a PoS and an inflectional category (i. e. a CODLES, in LEMLAT's terms) to the lemma.

In particular, to assess the PoS of a lemma we follow a principle of "lexical osmosis", that is, we assume that a lemma's definition core (see below) will most probably use terms belonging to the same PoS of that lemma. By cross-checking this information with the citation form of the lemma and possibly with its inflected forms in a quotation, we are able to assign it also its inflectional category.

With regard to the **definition**, we take into consideration only its initial part, maximally up to the first quotation; what comes after are mostly more in-depth discussions of the term, secondary interpretations or later interpolations. More precisely, we focus on the **definition's core**, i. e. a short capitalised phrase, enclosed in commas and/or ending with a full-stop, providing a short explanation or paraphrase of the lemma immediately after the lemma itself. Its terms are lemmas in typical quotation form, e. g. the nominative case for nouns. Moreover, the definition's core makes use of a standardised and Classical variety of Latin lexicon so as to be as clear as possible to the reader. This means that most of the terms in a definition's core can also be found in the list of lemmas of LEMLAT 3.0. Of the recognised forms, we retain only those that are univocally assigned only one PoS. We ignore a small set of both function and content words often recurring in definitions (e. g. *pro* 'for' and *omnis* 'all, every'), and discard as noise a set of very common lexicographical annotations and abbreviations (e. g. *Italus* or *Ital.*, *f.* = fortasse, *lib.*, *cap.*).

With regard to **quotations**, we only consider the first one as the most significant. Given the lemma's citation form in DC, we exploit the list of all Latin endings and their agreements with inflectional categories available in LEMLAT's database to construct all of its *a priori* possible inflectional paradigms; of these (partly artificial) forms, we retain only those that allow us to unambiguously discriminate a PoS and/or an inflectional category from the others. For example, the entry for *mansaticus* 'mansion, house' illustrates this method:

> MANSATICUS, **Mansio**, domus. Annal. Bertin. ad ann. 874. tom. 7. Collect. Histor. Franc. pag. 118 : *Inde per Attiniacum et consuetos* ***Mansaticos*** *Compendium adiit* [. . .]

Since the definition's core *mansio* can only be a noun for LEMLAT, we can conclude that *mansaticus* is almost surely a noun too, even if the *-icus* ending tends to be associated with denominal adjectives in Latin. The *-us* ending tells us that *mansaticus* can be either a masculine second or fourth declension noun;[12] a first class adjective might theoretically be possible, but is ruled out by the definition's core *mansio*. The second declension is confirmed by the ending *-os* found in the quotation, thus excluding the fourth declension (which should yield *-us*).

Thanks to this process, more than 10 000 additional lemmas are automatically included in LEMLAT. This process is applied very carefully, covering only decidedly unambiguous cases, i. e. when content words in the definition's core are found to belong to only one PoS or to a phrase of a fixed type (e. g. a phrase ending with an infinitive assigns PoS verb to the lemma) and when the inflectional category of the word form possibly found in the quotation can be univocally discriminated. This leads to high precision (1.0), but affects recall (0.18). For the remaining cases we have to resort to manual annotation; this happens most frequently when we correctly identify the PoS and the inflectional category of a lemma, but cannot infer its gender *a priori*. For instance, approxi-

endings of adjectives; infinitive endings of regular verbs and full paradigms of irregular verbs.

[11] The remaining lemmas are manually-checked because they correspond to multiple entries in one and/or the other source. For example, the lemma *fedus* appears once in DC (as a masculine second declension noun, 'fief') but three times in Bon's list: as a masculine second declension noun (but with the different meaning 'goat'), as a neuter third declension noun (with the genitive *federis*, 'alliance') and as a first class adjective ('hideous').

[12] Feminines are so rare in these declensions that we exclude them from the automated analysis.

mately 10% of first declension nouns are found to be masculine, and not feminine as expected.

4 Discussion

Not all of the 85 999 lemmas of DC are included in LEMLAT. We exclude the entries of some 3 000 fixed or idiomatic multi-word expressions and of around 300 adverbs derived either from an adjective (e. g. *affectuose* 'tenderly' from *affectuosus* 'tender') or from a verb (e. g. *attendenter* 'watchfully' from *attendere* 'to keep, to watch') in the lexical basis of the DC-enhanced LEMLAT. This is because LEMLAT considers derived adverbs as part of the inflectional paradigm of the source adjective or verb.

At the end of the process, 82 556 DC lemmas are added to LEMLAT. Since DC shows a tendency to treat different nuances of the same lemma as distinct entries, the total number of DC distinct lemmas inserted in LEMLAT is 73 131. The lemmas with the highest number of separate entries are *forma* 'form' (17), *scala* 'stairs, staircase, ladder' (15) and *status* 'mode, state, position, size' (15). These are all already attested in Classical Latin, but are also recorded in DC because of their semantic change over time.[13] This happens frequently; there are, in fact, 10 168 shared lemmas (corresponding to 14 469 entries in DC) in LEMLAT 3.0 and DC, with respect to the name of the lemma, its PoS and inflectional category (and gender, when applicable). Additionally, 1 820 lemmas share the same quotation form in both sources (often incidentally), despite being morphologically different. An example is *amo*: in DC, it is the third declension noun *amo*, *amonis*, a variant of *ammo*, *ammonis* (a unit of measure for wine), while in LEMLAT it is the verb *amare* 'to love'.

The remaining 66 267 lemmas are to be considered lexical innovations of "*media et infima Latinitas*". Looking at these Medieval lemmas, we notice some tendencies in the distribution of PoS and inflectional categories. Whereas nouns are the prevalent PoS both in LEMLAT and DC (albeit at very different rates, respectively 52% and 75%), in the former the most attested declension is the third (37% of nouns), while in the latter it is the first and second declensions that dominate (34% and 39% of nouns, against 20% of the third declension), showing a trend towards more transparent lexical items. While similar figures can be observed for verbs, in DC we notice a reduced presence of adjectives (12% against LEMLAT's 25%), revealing that they represent a less diachronically-productive PoS than nouns and verbs.

5 Evaluation

As conducted for the previous major update of LEMLAT (Passarotti et al., 2017), we evaluate LEMLAT's coverage of the Latin lexicon against the *Thesaurus formarum totius latinitatis* (TFTL) by Tombeur (1998), in order to assess the impact of LEMLAT's acquisition of DC. A primary reference for the study of the Latin lexicon, TFTL is a comprehensive diachronic collection of all Latin word forms as they occur in texts from the archaic period up to the Second Vatican Council (20th century), listing their respective frequencies in the sources from different eras.[14]

Passarotti et alii (2017) report a coverage of 72.254% of TFTL's forms, corresponding to 98.345% of the 62 922 781 total occurrences in the source texts.[15] This is partly explained by the fact that many forms in TFTL are either extremely rare, include punctuation in their spelling, or are merely sequences of numbers, letters and punctuation marks. When we add DC to LEMLAT, our coverage of TFTL raises by 3.264% to 75.518%, corresponding to 17 224 newly-recognised forms, whereas the covered occurrences increase to 98.665%.

We also perform a coverage evaluation over three Medieval Latin texts of comparable size, available from ALIM, the Archive of Italian Medieval Latinity (Ferrarini, 2017).[16] The texts belong to three different periods and genres; these are: the *Codex diplomaticus Cavensis* I (documents 33-210), a collection of documentary sources from Southern Italy dating to the 9th century; the *Historia Mongalorum*, a 13th century report of a journey and diplomatic mission; and the *De falso credita et ementita Constantini donatione*, a philological treatise dating back to the end of the 15th century.

[13] Indeed, DC does not at all record lemmas already available in Classical Latin, unless they show a different meaning and/or morphology.

[14] Archaic Latin (up to IInd c. AD), Patristic Latin (IInd c. AD – AD 735), Medieval Latin (AD 736 – AD 1499) and Modern Latin (AD 1500 – AD 1965), respectively.

[15] The statistics in this paper are based on updated, marginally corrected statistics with respect to those presented in Passarotti et alii (2017).

[16] `http://it.alim.unisi.it/`

Work (century)	Tokens	Types	LEMLAT	LEMLAT + DC	Only DC
Codex dipl. Cavensis (IX)	19428	3262	54.1%	59.2%	166 (5.1%)
Historia Mongalorum (XIII)	20360	4649	90.3%	92.2%	87 (1.9%)
De Constantini donatione (XV)	19805	6514	93.9%	94.8%	56 (0.9%)

Table 1: Comparison of the lexical coverage of DC-enhanced LEMLAT of three Medieval texts. The "Only DC" column lists the number of terms to be found exclusively in the added DC vocabulary.

Table 1 shows the improvements in lexical coverage obtained thanks to the enhancement of LEMLAT through DC. The results are in line with those for TFTL. Remarkably, the highest increase in performance is recorded for the least-standardised of the three texts, the *Codex diplomaticus*, which remains the most demanding for LEMLAT to analyse. This can be explained by the large presence of local names of people and places (e. g. *Sichelpertus*, *Eboli*), and especially by the very frequent deviations from the orthographic standard (e. g. *abentes* for *habentes* 'having (pl.)', *ecclesie* for *ecclesiae* 'of/to the church; churches'); the latter are also the source of many false positives, which LEMLAT does not discriminate from true positives. Names are challenging, too, as can be observed, for example, from the fact that among the 363 unrecognised forms in the *Historia Mongalorum*, the majority are ethnonyms, toponyms and anthroponyms (e. g. *Caracoron* 'Karakorum', *circassos* 'Circassians', *Mengu* 'Möngkh').

At the same time, LEMLAT is now able to analyse words which, while absent from the vocabulary of Classical Latin, are tied to key, widespread concepts in the Middle Ages. For example, in the *Historia Mongalorum* the enhanced LEMLAT can now detect terms like *orda* 'horde' (11 occurrences) or *protonotarius* 'prothonotary' (4 occurrences), both important in the 13th century onward in the context of conflicts and diplomatic missions between Western Europe and the Mongol Empire. Interestingly, the source for these lemmas in DC is not the *Historia Mongalorum* itself, which is an indication of the effective circulation of such words.

6 Conclusion

In this paper we present the rule-based process performed to semi-automatically enhance the Latin morphological analyser LEMLAT with the Du Cange glossary. While dated, such an approach is still necessary if the intent is to minimise the error rate resulting from the automatic PoS-tagging of the glossary's definitions and quotations. Indeed, unless tuned on an in-domain training set, existing stochastic PoS-taggers for Latin are not yet reliable enough when it comes to processing the complex, raw and "freestyle" definitions of DC.

The ever-growing availability of digitised Latin texts from various eras urges us to build NLP tools capable of automatically analysing such varied sets of linguistic data. In this respect, enhancing the lexical basis of LEMLAT with a Medieval Latin dictionary is a first step towards the development of well-performing tools on diachronic data. Conversely, even if building a tool suitable for different diachronic varieties of Latin were feasible for low-level annotation tasks (like e. g. lemmatisation and morphological analysis), this does not seem to be the case for tasks such as syntactic parsing or word sense disambiguation, for which either highly flexible or highly specialised tools will be needed.

This is an open issue not only for Latin. Indeed, the portability of NLP tools across domains and genres is currently one of the main challenges in NLP. Thanks to its highly diverse corpus, Latin is a perfect case-study language to tackle these problems.

For the future, we plan to expand LEMLAT's lexical database with all of the graphical variants reported in DC and possibly also with other Medieval Latin thesauri, such as the *Dictionary of Medieval Latin from British Sources* (Ashdown et al., 2018), so as to improve both its diatopic and diachronic coverage. In general, we aspire to make LEMLAT's algorithm better able to cope with the most widespread and predictable orthographic variations recorded in Medieval manuscripts and texts.[17]

[17] An introduction and an approach to this issue can be found in Kestemont and De Gussem (2017).

References

Richard K Ashdown, David R Howlett, and Ronald E Latham, editors. 2018. *Dictionary of Medieval Latin from British Sources*. Oxford University Press for the British Academy, Oxford, UK.

Bruno Bon. 2011. OMNIA : outils et méthodes numériques pour l'interrogation et l'analyse des textes médiolatins (3). *Bulletin du centre d'études médiévales d'Auxerre* BUCEMA, (15). Online at http://journals.openedition.org/cem/12015.

Andrea Bozzi and Giuseppe Cappelli. 1990. A project for Latin lexicography: 2. A Latin morphological analyzer. *Computers and the Humanities*, 24(5-6):421–426.

Marco Budassi and Marco Passarotti. 2016. Nomen omen. Enhancing the Latin morphological analyser Lemlat with an onomasticon. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 90–94, Berlin, Germany. Association for Computational Linguistics.

Charles du Fresne du Cange, Bénédictins de Saint-Maur, Pierre Carpentier, Louis Henschel, and Léopold Favre. 1883–1887. *Glossarium mediae et infimae latinitatis*. Niortm France.

Edoardo Ferrarini. 2017. ALIM ieri e oggi. *Umanistica Digitale*, 1(1). Online at https://umanisticadigitale.unibo.it/article/view/7193.

Karl Ernst Georges and Heinrich Georges. 1913–1918. *Ausführliches lateinisch-deutsches Handwörterbuch*. Hahn, Hannover, Germany.

Hercule Géraud. 1839. Historique du glossaire de la basse latinité de Du Cange. *Bibliothèque de l'École Nationale des Chartes*, 1:498–510.

Peter GW Glare. 1982. *Oxford Latin dictionary*. Clarendon Press. Oxford University Press, Oxford, UK.

Otto Gradenwitz. 1904. *Laterculi Vocum Latinarum: voces Latinas et a fronte et a tergo ordinandas*. Hirzel, Leipzig, Germany.

Mike Kestemont and Jeroen De Gussem. 2017. Integrated Sequence Tagging for Medieval Latin Using Deep Representation Learning. *Journal of Data Mining & Digital Humanities*, Special Issue on Computer-Aided Processing of Intertextuality in Ancient Languages, August. Online at https://jdmdh.episciences.org/3835.

Nino Marinone. 1990. A project for Latin lexicography: 1. Automatic lemmatization and word-list. *Computers and the Humanities*, 24(5-6):417–420.

Marco Passarotti, Marco Budassi, Eleonora Litta, and Paolo Ruffolo. 2017. The 3.0 Package for Morphological Analysis of Latin. In *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*, pages 24–31, Gothenburg, Sweden. Northern European association for language technology (NEALT), Linköping University Electronic Press.

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

Edoardo Maria Ponti and Marco Passarotti. 2016. Differentia compositionem facit. a slower-paced and reliable parser for Latin. In *Proceedings of the tenth international Conference on Language Resources and Evaluation (*LREC '16*)*, pages 683–688, Portorož, Slovenia. European Language Resources Association (ELRA).

Uwe Springmann, Helmut Schmid, and Dietmar Najock. 2016. LatMor: A Latin finite-state morphology encoding vowel quantity. *Open Linguistics - Topical Issue on Treebanking and Ancient Languages: Current and Prospective Research*, 2(1):386–392.

Paul Tombeur. 1998. *Thesaurus formarum totius Latinitatis: a Plauto usque ad saeculum XXum*. Brepols, Turnhout, Belgium.