Nomen Omen. Enhancing the Latin Morphological Analyser Lemlat with an Onomasticon

Marco Budassi
Università di Pavia
Corso Strada Nuova, 65
27100 Pavia - Italy
marcobudassi@hotmail.it

Marco Passarotti
Università Cattolica del Sacro Cuore
Largo Gemelli, 1
20123 Milan - Italy
marco.passarotti@unicatt.it

Abstract

Lemlat is a morphological analyser for Latin, which shows a remarkably wide coverage of the Latin lexicon. However, the performance of the tool is limited by the absence of proper names in its lexical basis. In this paper we present the extension of Lemlat with a large Onomasticon for Latin. First, we describe and motivate the automatic and manual procedures for including the proper names in Lemlat. Then, we compare the new version of Lemlat with the previous one, by evaluating their lexical coverage of four Latin texts of different era and genre.

1 Introduction

Since the time of the *Index Thomisticus* by father Roberto Busa (Busa, 1974-1980), which is usually mentioned among the first electronic (nowadays called "digital") annotated corpora available, NLP tools for automatic morphological analysis and lemmatisation of a richly inflected language like Latin were needed. Over the last decades, this need was fulfilled by a number of morphological analysers for Latin. Among the most widespread ones are *Morpheus* (Crane, 1991), Whitaker's *Words* (http://archives.nd.edu/words.html) and Lemlat (Passarotti, 2004). Over the past ten years, such tools have become essential, in light of a number of projects aimed at developing advanced language resources for Latin, like treebanks.[1]

The most recent advances in linguistic annotation of Latin treebanks are moving beyond the level of syntax, by performing semantic-based tasks like semantic role labelling and anaphora and ellipsis resolution (Passarotti, 2014). In particular, in the area of Digital Humanities there is growing interest in Named Entity Recognition (NER), especially for purposes of geographical-based analysis of texts.

NER is a sub-branch of Information Extraction, whose inception goes back to the Sixth Message Understanding Conference (MUC-6) (Grishman and Sundheim, 1996). NER aims at recognising and labelling (multi)words, as names of people, things, places, etc. Since MUC-6, NER has largely expanded, with several applications also on ancient languages (see, for example, Depauw and Van Beek, 2009).

Although Lemlat provides quite a large coverage of the Latin lexicon, its performance is limited by the absence of an Onomasticon in its lexical basis, which would be helpful for tasks like NER. Given that in Latin proper names undergo morphological inflection, in this paper we describe our work of enhancing Lemlat with an Onomasticon. The paper is organised as follows. Section 2 presents the basic features of Lemlat. Section 3 describes our method to enhance Lemlat with an Onomasticon, by detailing the rules for the automatic enhancement and discussing the most problematic kinds of words. Section 4 evaluates the rules and presents one experiment run on four Latin texts. Section 5 is a short conclusion and sketches the future work.

2 Lemlat

The lexical basis of Lemlat results 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.

[1] Three dependency treebanks are currently available for Latin: the Latin Dependency Treebank (Bamman and Crane, 2006), the *Index Thomisticus* Treebank (Passarotti, 2009) and the Latin portion of the PROIEL corpus (Haug and Jøndal, 2008).

Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH), pages 90–94,
Berlin, Germany, August 11, 2016. ©2016 Association for Computational Linguistics

Given an input wordform that is recognised by Lemlat, the tool produces in output the corresponding lemma(s) 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 wordform (e.g. singular nominative), as well as the identification number (ID) of the lemma(s) in the lexical basis of Lemlat. No contextual disambiguation is performed.

For instance, receiving in input the wordform *abamitae* ("great-aunt"), Lemlat outputs the corresponding lemma (*abamita*, ID: A0019), the tags for its inflectional paradigm (N1: first declension noun) and those for the morphological features of the input wordform (feminine singular genitive and dative; feminine plural nominative and vocative).

The basic component of the lexical look-up table used by Lemlat to analyse input wordforms is the so-called LES ("LExical Segment"). The LES is defined as the invariable part of the inflected form (e.g. *abamit* for *abamit-ae*). 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).

Lemlat includes a LES archive, in which LES are assigned an ID and a number of inflectional features among which are 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 of its inflectional paradigm. For instance, the CODLES for the LES *abamit* is N1 (first declension nouns) and its gender is F (feminine). The wordform *abamitae* is thus analysed as belonging to the LES *abamit* because the segment -*ae* is recognised as an ending compatible with a LES with CODLES N1.

3 Enhancing Lemlat. Method

The bedrock of our work is Busa's (1988) *Totius Latinitatis Lemmata*, which contains the list of the lemmas (92,052) from the 5th edition of *Lexicon Totius Latinitatis* (Forcellini, 1940). In Busa (1988), three kinds of metadata are assigned to each lemma: (a) a code for the section of the dictionary in which the lemma occurs (e.g. ON: the lemma occurs in the *Onomasticon*), (b) a code for the inflectional paradigm the lemma belongs to and its gender (e.g. BM: second declension masculine nouns) and (c) the number of lines of the lexical entry for the lemma in Forcellini.

In order to enhance Lemlat with Forcellini's Onomasticon, we first extracted from Busa (1988) the list of those lemmas that occur in the ON section. This list counts 28,178 lemmas. Then, we built a number of rules to automatically include the lemmas of the Onomasticon into the lexical basis of Lemlat.

3.1 Types of Rules

Including the Onomasticon of Forcellini into Lemlat means converting the list of proper names provided by Busa (1988) into the same format of the LES archive. In order to perform this task as automatically as possible, we built a number of rules to extract the relevant information for each lemma in the list, namely its LES, CODLES and gender. By exploiting the morphological tagging of Busa (1988), which groups sets of lemmas showing common inflectional features, our rules treat automatically such inflectionally regular groups. In total, we wrote 122 rules, which fall into four types.

The first type (60 rules) builds the LES by removing one or more characters from the right side of the lemma. Such a removal is constrained by the code for the inflectional paradigm of the lemma, which is then used to create both the CODLES and the tag for the gender. For instance, the lemma *marcus* ("Mark") is assigned the inflectional paradigm BM in Busa (1988). One rule states that the LES for BM lemmas ending in -*us* is built by removing the last two characters from the lemma (*marcus* > *marc*) The inflectional code BM stands for second declension (B) masculine (M) nouns: this is converted into the CODLES of Lemlat for second declension nouns (B > N2) and into the tag for masculine gender (M > m).

The second type of rules (19) adds one or more characters on the right side of the lemma to build the LES. Again, this is done according both to the inflectional paradigm and to the ending of the lemma in Busa (1988). For instance, the LES for lemmas with inflectional code CM (third declension masculine nouns) and ending in -*o* is built by adding an -*n* after the last character. One example is the lemma *bappo* ("Bappo"), whose LES is *bappon*, as third declension imparisyllable nouns are analysed by Lemlat by using the basis for their singular genitive (*bappon-is*).

The third type of rules (19) replaces one or more characters on the right side of the lemma with others. For instance, the LES of *clemens* ("Clement", third declension masculine noun ending in *-s*, with singular genitive *clement-is*) is built by replacing the final *-s* with a *-t* (*clement*).

The last type of rules (24) deals with those lemmas that are equal to their LES (no change is needed). These are uninflected nouns, (like *hamilcar* - "Hamilcar"), which can be easily retrieved because they are assigned a specific inflectional code in Busa (1988).

3.2 Problematic Cases

Not all inflectional paradigms are as much regular as to allow for a fully automatic rule-based treatment.

For instance, third declension feminine nouns represent an entangled class. The lemma *charybdis, -is* ("Charybdis") is a third declension parisyllable feminine noun ending in *-is*. Instead, *phegis, -gidis* ("daughter of Phegeus") is a third declension imparisyllable feminine noun ending in *-is*. One common rule cannot be used for these two kinds of words. We overcome such problem by building two more specific rules: one accounting for third declension feminine nouns ending in *-dis* and one for third declension feminine nouns ending in *-gis*. However, there are sub-groups of nouns for which such a solution does not work, like third declension feminine nouns ending in *-mis*, which can be both imparisyllable nouns (e.g. *salamis*, *-minis*, "Salamis") and parisyllable nouns (e.g. *tomis*, *-is*, "Tomis"). For these lemmas we checked manually their inflection in Forcellini and assigned LES and CODLES accordingly.

Another group of tricky words includes those lemmas that show two (or even more) different inflectional paradigms. For instance, *apollonides* ("Apollonides") shows both a singular genitive of the second declension (in *-i*) and one of the first declension (in *-ae*). We treated these cases manually by checking their lexical entries in Forcellini.

A further problem is represented by graphical variants, which are managed by Lemlat through so-called "exceptional forms". These are wordforms that are hard-coded in the LES archive and are assigned the same ID of the LES used to build their base lemma. For instance, the nominative singular of the lemma *jesus* ("Jesus") is attested also as *hiesus*, *ihesus* and *zesus*. Beside the LES *jes* (used for the base lemma *jesus*), in the LES archive also the wordforms *hiesus*, *ihesus* and *zesus* are recorded and assigned the same ID of the LES *jes*.

4 Evaluating the Enhancement

We evaluated the enhancement of Lemlat with the Onomasticon of Forcellini in two steps. First, we focused on the accuracy of the rules for automatic enhancement. Then, we compared the new version of Lemlat with the previous one by the lexical coverage they provide for four Latin texts.

4.1 Rules

We evaluated the quality of the rules for automatic enhancement by precision and recall (Van Rijsbergen, 1979).

Measuring the precision of our rules is straightforward. As said, while writing the rules, we focused on inflectionally regular groups of lemmas. As a consequence, we never had to modify the output of rules neither in terms of removal of results (i.e. wrong results due to overproduction) nor in terms of completion of results (i.e. wrong results due to underproduction). Thus, the precision of our rules is always 100%.

To calculate recall, we grouped all those rules that treat lemmas of the same inflectional class (e.g. all rules for nouns of the first declension). We measured the recall of such groups of rules by comparing the number of lemmas automatically inserted into Lemlat by one group of rules with the total number of lemmas in the Onomasticon of Forcellini belonging to the inflectional class addressed by that group of rules. Table 1 shows the results.

Inflectional Class	Lemmas per Class	Lemmas per Rules	Recall
1st decl.	6,597	6,597	100%
2nd decl.	12,968	12,961	99.946%
3rd decl.	5,397	3,923	72.688%
4th decl.	50	11	22%
5th decl.	6	6	100%
Uninflected	1,166	11,66	100%
	26,184	**24,664**	**94.194%**

Table 1: The recall of rules.

The most problematic inflectional class is that of third declension nouns.[2] As mentioned above, this is motivated by the fact that it is not always

[2] The rules for fourth declension nouns show an even lower recall than those for third declension, but the results for such class must be evaluated carefully as the lemmas of the fourth declension in the Onomasticon are just a few (50).

possible to match regularly an inflectional paradigm (e.g. third declension imparisyllable nouns) with one specific ending. Hence, given such a low recall, the amount of manual work required for enhancing Lemlat with third declension proper names was quite considerable. To provide an example, the number of third declension feminine nouns in the Onomasticon is 1,200. Our rules covered only 542 out of them. Thus, 658 nouns had to be inserted into Lemlat manually (54.833% of the total for that class).

There are also entire inflectional classes for which writing a rule was not possible, like for instance Busa's class of irregularly inflected nouns (146 wordforms). All these lemmas were inserted into the LES archive manually.

In total, the number of lemmas transferred manually into Lemlat is 1,752 (6.632% of all the lemmas of the Onomasticon).

4.2 Coverage

We evaluated the enhancement of Lemlat with the Onomasticon of Forcellini by comparing the lexical coverage provided by the two versions of the tool for four Latin texts of similar size and different genre (prose and poetry) and era (Classical and Late Latin).[3] Table 2 presents the number of distinct words (types) analysed by the original version of Lemlat and by the one enhanced with the Onomasticon (LemlatON).

Text	Types	Lemlat	LemlatON	Improv.
(1)	3,092	2,888 (93.4%)	3,039 (98.1%)	+4.7%
(2)	5,057	4,717 (93.27%)	5,005 (98.97%)	+5.7%
(3)	3,542	3,357 (94.78%)	3,487 (98.45%)	+3.67%
(4)	4,589	4,292 (93.53%)	4,537 (98.87%)	+5.34%
Avg	**4,070**	**93.74%**	**98.6%**	**+4.86%**

Table 2: Type-based evaluation.

The coverage of Lemlat on the four test texts improved of 4.86% on average after the enhancement with Forcellini's Onomasticon. The highest improvement is on Virgil (+5.7%).

Most of the words not analysed by LemlatON are graphical variants (e.g. *creüsa* for *creusa* - "Creusa") or part of the inflectional paradigm of lemmas not available in its lexical basis. Beside these words, there are Roman numbers (e.g. *XV*, "fifteen"), abbreviations (e.g. *kal* for *kalendae*, "calends") and foreign words (e.g. *μητέρα*, "mother").[4] Table 3 shows the results by category of unknown words (types).

Text	Unk	RN	FW	Abb	Misc.
(1)	**53**	19	0	2	32
(2)	**51**	0	1	0	52
(3)	**55**	0	5	0	50
(4)	**52**	0	1	3	48

Table 3: Categories of unknown words.[5]

Roman numbers are frequent in Caesar's text (1). The fact that Lemlat does not analyse Roman numbers is not a major concern, as their form is regular, easily predictable and interpretable. Only a few of them can raise ambiguity when written lowercase. For instance, *vi* ("six") is homograph with the singular ablative of the third declension noun *vis* ("power").

Homography can hold also between items of of the Onomasticon and the original lexical basis of Lemlat. For instance, the lemma *augustus* occurs both in the original Lemlat (a first class adjective, "solemn") and in the Onomasticon (a proper name, "Augustus").

If we look at tokens instead of types, coverage rates remain quite similar, as it is shown by Table 4.

Text	Tokens	Lemlat	LemlatON	Improv.
(1)	8,171	7,558 (92.49%)	8,100 (99.13%)	+6.64%
(2)	10,045	9,478 (94.36%)	9,971 (99.26%)	+4.9%
(3)	7,317	7,059 (96.47%)	7,260 (99.22%)	+2.75%
(4)	6,991	6,604 (94.46%)	6,931 (99.14%)	+4.68%
Avg	**8,131**	**94.39%**	**99.19%**	**+4.8%**

Table 4: Token-based evaluation.

[3] (1) Caesar, *De Bello Gallico,* lib. 1 (Classical Lat., prose); (2) Virgil, *Aeneid*, lib. 1 & 2 (Classical Lat., poetry); (3) Tertullian, *Apologeticum* (Late Lat., prose); (4) Claudian, *De Raptu Proserpinae* (Late Lat., poetry). All the texts were downloaded from the Perseus Digital Library (www.perseus.tufts.edu).

[4] We do not consider as foreign words Greek proper names transliterated into Latin characters (e.g. *cytherea*).

[5] "Unk": total number of words per text not analysed by LemlatON. "RN": Roman numbers. "FW": foreign words. "Abb": abbreviations. "Misc": graphical variants and missing lemmas.

It is worth noting that, while the text of Virgil shows the highest improvement in type-based evaluation (+5.7%), Caesar's *De Bello Gallico* is the one that mostly benefits from the extension of Lemlat with the Onomasticon in token-based evaluation (+6.64%). This is due to the higher number of occurrences of proper names in Caesar than in Virgil. Indeed, although the number of new word types analysed by LemlatON in comparison to Lemlat is lower for Caesar than for Virgil, the opposite holds when tokens are concerned.[6] In more detail, the average number of occurrences (tokens) of the new word types analysed by LemlatON for Caesar is 3.59 (542/151), while it is 1.71 for Virgil (493/288).

5 Conclusion and Future Work

In this paper we described the enhancing of the morphological analyser for Latin Lemlat with a large Onomasticon provided by a reference dictionary for Latin (Forcellini).

Although we have included most of the words of the Onomasticon into Lemlat, the work is far from being complete. Indeed, we have just started to enhance the analyser with graphical variants. Furthermore, around 2,000 words of the Onomasticon belonging to minor and irregular inflectional classes still have to be included into Lemlat. Although this promises to be a largely manual and time-consuming work, it is worth doing for achieving the lexicographically motivated completeness of the tool's lexical basis.

Once completed, the lexical look-up table of the Onomasticon will become part of the overall Lemlat suite, which will be shortly made available for free download and on-line use.

References

David Bamman and Gregory Crane. 2006. The Design and Use of a Latin Dependency Treebank. *TLT 2006: Proceedings of the Fifth International Treebanks and Linguistic Theories Conference*, 67–78.

Roberto Busa. 1974-1980. *Index Thomisticus: sancti Thomae Aquinatis operum omnium indices et concordantiae, in quibus verborum omnium et singulorum formae et lemmata cum suis frequentiis et contextibus variis modis referuntur quaeque / consociata plurium opera atque electronico IBM automato usus digessit Robertus Busa SJ.* Frommann-Holzboog, Stuttgart-Bad Cannstatt.

Roberto Busa. 1988. *Totius Latinitatis lemmata quae ex Aeg. Forcellini Patavina editione 1940 a fronte, a tergo atque morphologice opera IBM automati ordinaverat Robertus Busa SJ.* Istituto Lombardo, Accademia di scienze e lettere, Milano.

Gregory Crane. 1991. Generating and Parsing Classical Greek. *Literary and Linguistic Computing*, 6(4): 243–245.

Mark Depauw and Bart Van Beek. 2009. People in Greek Documentary Papyri: First Results of a Research Project. *JJurP*, 39: 31–47.

Egidio Forcellini. 1940. *Lexicon Totius Latinitatis / ad Aeg. Forcellini lucubratum, dein a Jos. Furlanetto emendatum et auctum; nunc demum Fr. Corradini et Jos. Perin curantibus emendatius et auctius meloremque in formam redactum adjecto altera quasi parte Onomastico totius latinitatis opera et studio ejusdem Jos. Perin*. Typis Seminarii, Padova.

Karl E. Georges and Heinrich Georges. 1913-1918. *Ausführliches Lateinisch-Deutsches Handwörterbuch*. Hahn, Hannover.

Peter G. W. Glare. 1982. *Oxford Latin Dictionary*. Oxford University Press, Oxford.

Otto Gradenwitz. 1904. *Laterculi Vocum Latinarum*. Hirzel, Leipzig.

Ralph Grishman and Beth Sundheim. 1996. Message Understanding Conference-6. A Brief History. *COLING*, 96: 466–471.

Dag T. T. Haug and Marius L. Jøhndal. 2008. Creating a Parallel Treebank of the Old Indo-European Bible Translations. *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, 27–34.

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

Marco Passarotti. 2014. From Syntax to Semantics. First Steps Towards Tectogrammatical Annotation of Latin. *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences and Humanities (LaTeCH 2014)*, 100–109.

Marco Passarotti. 2009. Theory and Practice of Corpus Annotation in the *Index Thomisticus* Treebank. *Lexis*, 27: 5–23.

Cornelis J. Van Rijsbergen. 1979. *Information Retrieval*. Butterworths, London.

[6] Caesar: 151 types (3,039–2,888) and 542 tokens (8,100–7,558). Virgil: 288 types (5,005–4,717) and 493 tokens (9,971–9,478).