May the Goddess of Hope Help Us.

Homonymy in Latin Lexicon and Onomasticon

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

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

Abstract

English. We present a study on the degree of homonymy between the lexicon of a morphological analyser for Latin and an Onomasticon. To understand the impact of homonymy, we discuss an experiment on four Latin texts of different era and genre.

Italiano. *L'articolo presenta uno studio sul grado di omonimia tra il lessico di un analizzatore morfologico per il latino e un Onomasticon. Al fine di comprendere l'impatto dell'omonimia, viene descritto un esperimento condotto su quattro testi latini di diversa epoca e genere.*

1 Introduction

Ambiguity affects linguistic analysis at various levels. In particular, homonymy plays a substantial role in the analysis of single words. Indeed, when considered out of context, one same word can be assigned different Parts of Speech (PoS), morphological features, lemmas and meanings. Contextual disambiguation is the task of Natural Language Processing (NLP) tools like PoS-taggers, morphological analysers, lemmatisers and word-sense disambiguators.

The problem of ambiguity is particularly remarkable for NLP when Named Entity Recognition (NER) is concerned. In order to automatically classify the textual occurrences of (multi)words into categories such as names of persons, locations and organisations, NER faces that specific kind of ambiguity consisting in the homonymy between proper names and other words in the lexicon (Nadeau and Sekine, 2007). For instance, the word *mark* in English can be a proper name, a noun or a verb. Although such homonymy is often tackled by using the upper/lowercase distinction for the initial letter of words, this solution is neither decisive (as uppercase letters can also be motivated by punctuation) nor always available. The latter is especially true for historical languages, as a large amount of texts in such languages comes with no upper/lowercase distinction and it may follow different editorial criteria.

The recent extension of the lexical basis of the morphological analyser and lemmatiser for Latin Lemlat with an Onomasticon (i.e. a list of proper names) makes it possible to evaluate the degree of homonymy of proper names in Latin and, thus, to understand the extent of the disambiguation task (Passarotti and Ruffolo, 2004). To this aim, in this paper we explore the lexical basis of Lemlat as providing the empirical evidence supporting our analysis on homonymy between names in the Onomasticon and words in the Latin lexicon.

2 Lemlat

Together with *Morpheus* (Crane, 1991) and Whitaker's *Words*, Lemlat (Passarotti, 2004) is one of the most widespread tools for automatic analysis of Latin morphology available. The original lexical basis of Lemlat (L) 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). Such lexical basis was recently merged with most of the Onomasticon (O) (26,250 lemmas out of 28,178) provided by the 5th edition of *Lexicon Totius Latinitatis* (Forcellini, 1940) (Budassi and Passarotti, 2016).

Since the large majority of lemmas in O are nouns (19,599 out of 26,250), we will focus on them here, first by comparing their distribution in L and O. Table 1 shows the number of nouns and their percentage (on the total of nouns) in L and O by inflectional category.

Infl. Cat.	Lemlat	Onomasticon
I decl.	5,009 (22.26%)	6,651 (33.94%)
II decl.	7,466 (33.17%)	7,235 (36.92%)
III decl.	8,677 (38.54%)	4,464 (22.77%)
IV decl.	980 (4.35%)	58 (0.29%)
V decl.	101 (0.45%)	6 (0.03%)
Uninfl.	278 (1.23%)	1,185 (6.05%)
TOTAL	22,511	19,599

Table 1. Nouns in L and O.

While third declension nouns are more frequent in L than in O, the opposite holds for first declension and (to a lesser extent) second declension nouns. The main difference between L and O concerns uninflected nouns, which are much more in O than in L because of the large number of loans recorded in O.

Also gender-based distribution of nouns by inflectional category shows substantial differences between L and O. Among the most relevant is that O includes more first declension masculine nouns than L (1,626 vs. 562). Instead, the number of second declension neuter nouns is larger in L than in O (4,005 vs. 1,523), because O tends to include more proper names of persons than of places, the latter being often assigned the neuter gender. As for third declension, feminine nouns are more than masculine in L (5,112 vs. 2,590), while the opposite holds in O (2,847 masculine vs. 1,185 feminine).

3 Mining Nominal Homonymy

To categorise nominal homonymy in L and O, we defined three kinds of homonymy: (a) Full Homonymy (FH): words with the same lemma, PoS, inflectional category and gender in L and O; (b) Partial Homonymy (PH): words with the same lemma in L and O, but with different PoS, inflectional category or gender (the last for nouns only); (c) Mixed Homonymy (MH): words with the same lemma in L and O and with more than one PoS, inflectional category or gender, thus resulting partly into FH and partly into PH.

An example of FH in our data is the word *spes*, which means “hope” in L and “the Goddess of Hope” in O. PH is represented, for instance, by the word *augustus*, which is an adjective in L (“majestic”) and a noun in O (a cognomen given to Octavius Caesar as emperor). The word *spina* is a case of MH, being a first declension feminine noun in L (“thorn”) and both a first declension feminine noun (an old town in Aemilia) and a third declension masculine noun with genitive in *–anis* (a river God) in O, the former thus showing FH and the latter PH.

Table 2 presents the rates of homonymy in L and O by each kind per inflectional category. The total number of homonyms is provided as well (column “H”). This corresponds to the number of nouns of an inflectional category that are graphically identical in L and O. For instance, the first row of table 2 shows that there are 556 lemmas recorded as first declension nouns (in L or O) that are identical to a lemma occurring in the other section of the lexical basis of Lemlat. 383 lemmas out of these show FH, i.e. they share not only the same graphical form but also the same PoS, inflectional category and gender in L and O (column “FH”). Instead, 163 lemmas occur as graphically identical in L and O but do not have in common at least one among PoS, inflectional category or gender (column PH”). Finally, 10 lemmas show MH.

Infl. Cat.	H	FH	PH	MH
I decl.	556	383	163	10
II decl.	752	307	389	56
III decl.	584	334	226	24
IV decl.	85	9	73	3
V decl.	6	5	1	0
Uninfl.	60	0	60	0
TOTAL	2,043	1,038	912	93

Table 2. Kinds of homonymy.

Most of the PH instances for first declension lemmas are due to different gender. An example is the first declension noun *caligula*, which is feminine in L (“a small military boot”) while it is masculine in O (a cognomen). Second declension shows several cases of PoS change, like in the case of *severus*, which is an adjective in L (“serious”) and a noun in O (a proper name). Instead, a large number of verb-noun changes holds for third declension. This mostly occurs for imparisyllable nouns ending in *–o*, like *cato*, which is a first conjugation verb in L (“to see”) and a noun in O (a proper name).

PH does not raise any tricky issue for NLP, the task of PoS/morphological taggers being just that of disambiguating contextually PoS and morphological features. Conversely, FH (including the FH-like part of MH) represents a challenging question for NLP. Indeed, if upper/lowercase distinction is not available in input data, only context-based semantic properties can disambiguate between candidate lemmas affected by FH. For instance, in the clause "spes est expectatio boni" ("hope is expectation of good", Cicero, *Tusculanae*, 4, 37, 80) there is nothing but semantics to help us to understand that the word *spes* is an occurrence of the noun from L instead of the proper name from O. In order to evaluate the extent of homonymy in real texts and to understand how much big the impact of FH is, we performed the experiment discussed in the next section.

4 Homonymy in Texts. A Case-study

We run Lemlat on four Latin texts of similar size and different genre and era.[1] Table 3 shows the number of distinct words out of the total (column "Types") analysed by the original version of Lemlat (column "Lemlat") and by the one enhanced with the Onomasticon (column "LemlatON").

Text	Types	Lemlat	LemlatON	Improv.
(1)	3,092	2,888	3,039	+151
(2)	5,057	4,717	5,005	+288
(3)	3,542	3,357	3,487	+130
(4)	4,589	4,292	4,537	+245

Table 3. Results of Lemlat(ON) on four texts.

Beside the words analysed by LemlatOn only (column "Improv."), there is a certain degree of overlapping between Lemlat and LemlatOn. The words falling in this 'grey zone' are those that are analysed both by Lemlat and by LemlatOn, as they are lemmatised both under a lemma from L and under one from O. Among these words, those affected by homonymy are to be found.

Text	L/O	H	FH	PH	MH
(1)	618	405	303	88	14
(2)	1,207	799	546	186	67
(3)	686	486	330	120	36
(4)	1,062	706	469	177	60

Table 4. Overlapping and homonymy rates.

Column "L/O" in table 4 reports the number of words for each text that are analysed both by Lemlat and by LemlatON. The other columns show the homonymy rates by the kinds described in Section 3. For instance, in the text of Caesar (1) there are 618 words analysed by both the versions of Lemlat (L/O). 405 out of them share the same lemma in at least one analysis (H). This is further detailed: 303 out of 405 show FH, 88 PH and 14 MH. An example of a word analysed by both the versions of the tool that does not share the same lemma in all analyses is *acie*, which is lemmatised under *acies* ("dagger") by Lemlat (fifth declension feminine noun) and also under the proper name *acius* by LemlatON (second declension masculine noun). The word *constantia* is an example of H: it is lemmatised as a form of both lemmas *consto* ("to agree"; first conjugation verb) and *constantia* ("steadiness"; second declension feminine noun) by Lemlat, and also as a form of both proper names *constantius* (second declension masculine noun) and *constantia* (second declension feminine noun) by LemlatON. The word *constantia* is also an example of FH, as the analyses provided by the two versions of Lemlat that share the same lemma have in common even the same inflectional category and gender. PH is shown by the word *crassi*, which is assigned the same lemma (*crassus*) both by Lemlat and by LemlatON, but while it is a first class adjective in the former ("solid"), it is a second declension masculine noun in the latter (a proper name). An example of MH is the word *amico*, which is lemmatised under the lemma *amicus* ("friend") both by Lemlat and LemlatON. The lemma *amicus* is both an adjective and a second declension masculine noun in Lemlat, but only the latter analysis is shared with LemlatON, because the lemma *amicus* in the Onomasticon is recorded only as a noun and not also as an adjective. Thus, when the word *amico* is assigned PoS noun it shows FH, while when it is assigned PoS adjective it shows PH.

The proportions between the kinds of homonymy remain quite similar for all the texts.

[1] (1) Caesar, De Bello Gallico, 1 (Classical Lat., prose); (2) Virgil, Aeneid, 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).

Words affected by H tend to be more than half of L/O; among them the large majority is affected by FH. By comparing columns "FH" and "MH" in table 4 with column "Types" in table 3, one can see that slightly more than 10% of the words of all the texts is affected by FH. This is the percentage rate of words whose lemmatisation cannot be disambiguated by a PoS tagger, because semantic features only are here at work to choose between candidate lemmas. For instance, if a PoS tagger assigns to one occurrence of the word *constantia* PoS noun and gender feminine, it cannot disambiguate between the two (fully morphologically identical) lemmas *constantia* provided by LemlatON.

If we focus on textual occurrences (tokens) instead of distinct words (types), the rates of FH (+MH) range between 8.44% (Caesar) and 13.19% (Tertullian), as shown by table 5. This result represents the extent of the impact of FH on the texts that we used in the case-study.

Text	**Tokens**	**FH+MH**
(1)	8,171	690 (8.44%)
(2)	10,045	1,325 (13.19%)
(3)	7,317	668 (9.13%)
(4)	6,991	797 (11.4%)

Table 5. Token-based homonymy rates.

Most of the words showing FH can be easily disambiguated (at least, manually) according to peculiarities of single texts. For instance, the word *amicitiam* (from Caesar's text) belongs to lemma *amicitia* both in L ("friendship") and in O ("the Goddess of Friendship"), thus showing FH. However, it is more likely that the former is the one occurring in Caesar than the latter. Conversely, in the same text the word *galli* (lemma *gallus*) is more likely a proper name from O ("Gauls") than a noun from L ("cock").

5 Conclusion

We presented a study about the degree of homonymy between the lexical basis of a morphological analyser for Latin and an Onomasticon recently added in the tool. We have shown the impact of nominal homonymy on a number of Latin texts of different era and genre.

Since the analysis of many homonymous words can be disambiguated according to the features of single texts (and authors), in the near future we foresee to enhance such words in Lemlat with information about their distribution in a number of manually tagged reference texts.

References

Marco Budassi and Marco Passarotti. 2016. Nomen Omen. Enhancing the Latin Morphological Analyser Lemlat with an Onomasticon. *Proceedings of the 10th Workshop on Language Technology for Cultural Heritage, Social Sciences and Humanities (LaTeCH 2016).* The Association for Computational Linguistics, Berlin, Germany, 90–94.

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

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 Ernst 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.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

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

Marco Passarotti and Paolo Ruffolo. 2004. L'utilizzo del lemmatizzatore LEMLAT per una sistematizzazione dell'omografia in latino. *Euphrosyne*, XXXII:99–110.