-io Nouns through the Ages. Analysing Latin Morphological Productivity with Lemlat

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

Eleonora Litta, Marco Passarotti
Università Cattolica del Sacro Cuore
Largo Gemelli 1
Milan, Italy 20123
e.littamodignani@gmail.com
marco.passarotti@unicatt.it

Abstract

English. This paper aims at examining the diachronic distribution of one of the richest classes of nouns in Latin, namely those ending in *-io*. The work is performed through the combined use of a morphological analyser for Latin (Lemlat), and a database collecting all word forms occurring through different periods of Latin language (TF-CILF).

Italiano. *Questo articolo presenta un'analisi della distribuzione diacronica di una delle più ricche classi di nomi in latino, ossia quelli che terminano in -io. Metodologicamente, il lavoro viene condotto attraverso l'uso incrociato di un analizzatore morfologico per il latino (Lemlat) e di una risorsa lessicale contenente tutte le forme di parole latine che occorrono in testi che vanno dall'antichità al neo-latino (TF-CILF).*

1 Introduction

The investigation of lexical data of Classical languages through the use of linguistic resources and Natural Language Processing (NLP) tools has witnessed a surge of interest in the past decade. As far as Latin is concerned, today several textual and lexical resources, as well as NLP tools, are being used in lexicographic research.[1] One of the bedrocks of this type of research is the use of morphological analysers, that is, tools that, given an input word form, output its corresponding lemma(s) and morphological features.

First released at the beginning of the 1990s and recently made freely available in its version 3.0 (Passarotti et al., 2017), Lemlat is one of the best performing morphological analysers and lemmatisers for Latin.[2] Lemlat is currently in the process of being enriched with all lemmas contained in the glossary of Medieval Latin *Glossarium mediae et infimae latinitatis* compiled by Charles Du Cange et alii in 1883-1887 (Glorieux, 2010).

One of the first groups of lemmas from Du Cange which was included into the lexical basis of Lemlat was that collecting all 3rd declension nouns ending in *-io*, one of the most productive affixes in all periods of Latin, up to Romance languages (Fruyt, 2011). The aim of this study is to perform a diachronic quantitative evaluation of 3rd declension nouns ending in *-io*. To do so, first we use Lemlat to lemmatise all word forms of such nouns contained in *Thesaurus formarum totius latinitatis a Plauto usque ad saeculum XXum* (TF-CILF) (Tombeur, 1998). Then we evaluate the results of the lemmatisation in both quantitative and qualitative terms.

2 Lemlat and Du Cange

Lemlat relies on a lexical basis resulting from the collation of three Classical Latin dictionaries,[3] for a total of 40,014 lexical entries and 43,432 lemmas (as more than one lemma can be included in one lexical entry). In the context of the development of Lemlat version 3.0, its lexical basis was further enlarged by adding semi-automatically most of the Onomasticon (26,415 lemmas out of 28,178) provided by the 5th edition of the Forcellini dictionary for Latin (Budassi and Passarotti, 2016). Furthermore, the inflectional information provided by Lemlat has been enhanced with information on derivational morphology taken from the *Word For-*

[1] See (Bamman and Crane, 2008), (McGillivray and Passarotti, 2009), (McGillivray, 2013) and (Passarotti et al., 2016).

[2] www.lemlat3.eu. See (Springmann et al., 2016) for a comparative evaluation of the morphological analysers currently available for Latin.

[3] (Georges and Georges, 1913-1918), (Glare, 1982) and (Gradenwitz, 1904).

mation Latin (WFL) lexicon (Litta et al., 2016).[4]

However, being based on dictionaries for Classical Latin, one of the current limitations of Lemlat is the fact that its lexical basis is not large enough yet to provide a wide coverage of the word forms occurring in Late and Medieval Latin texts. For this reason an upgrade of Lemlat 3.0 with the Medieval Latin lemmas contained in the Du Cange glossary (Glorieux, 2010), made available online by the École National des Chartes,[5] is underway.

3 Nouns Ending in *-io*

In the Lemlat lexical basis, nouns of the 3rd declension ending in *-io* (with genitive in *-ionis*) are mostly feminine. Only 294 out of 3,065 *-io* nouns in Lemlat are masculine, more than half of which are proper names.[6] Most frequently, nouns in *-io* derive from verbs. WFL contains 2,510 deverbal nouns in *-io*, 87 denominal, and 36 deadjectival. There are also not derived *-io* nouns, like for instance *bacrio* 'trowel'.

Resulting from one of the main mechanisms for Latin nominalisation (Rosén, 1983), deverbal nouns in *-io* are generally called processes or verbal nouns. Semantically, they can be either "nomina actionis", referring to the process of the action expressed by the input verb (e.g. *aberro* 'to wander from the way' > *aberratio* 'diversion', as the process of wandering from the way), or "nomina rei actae", referring to the result of such process (e.g. *aberratio* as the result of wandering from the way).[7]

An investigation on productivity in affixal derivation performed on the data extracted from WFL has proved that deverbal nouns in *-io* are the most numerous formations in Classical Latin (Litta et al., 2017). Such a high presence of nouns in *-io* in Latin lexicon motivates the choice of them as the object of this work.

For this study, we have grouped the nouns in *-io* as follows:

1. Group D: nouns that are only contained in Du Cange (tot. no. 1,416);
2. Group L: nouns that are only contained in Lemlat (tot. no. 2,246);
3. Group L&D: nouns that are contained in both Du Cange and Lemlat (tot. no. 1,494).

Du Cange contains a total of 2,910 nouns ending in *-io*. One of the characteristics of the Du Cange glossary is indeed that no Classical Latin lemma is included in its lexical basis, and if the same lemma is contained in both lexical bases, it means that it has undergone a major semantic or morphological change. 1,416 *-io* nouns out of 2,910 are listed only in Du Cange (Group D), which means that they were absent in the Classical Latin dictionaries used for compiling Lemlat.

Group L contains all those *-io* nouns whose meaning (or morphology) did not change from Classical Latin throughout time, or that were not used anymore in Medieval Latin. Such words are then exclusive only of the Lemlat lexical basis. Even if they were used in Medieval times, they did not undergo a semantic or morphological change, hence they were not included in Du Cange.

Group L&D contains all those *-io* nouns that are recorded both in Lemlat and Du Cange. These are mostly words that have undergone a semantic change, but there are also cases of words that are spelled differently in Medieval sources (e.g. Med. *adsumtio* or *assumtio* for Cl. *assumptio* 'acquisition'), or that in Medieval times acquired a different inflection (e.g. Cl. *beneficium* 'kindness', 2nd declension > Med. *beneficio*, 3rd declension). Because Du Cange treats different meanings in different entries, there is also a number of words appearing more than once (e.g. *defensio* 'defense' x4, *invocatio* 'invocation' x2).

4 Methodology

In order to perform a diachronic evaluation of the frequency of distribution of these three groups, we have used data extracted from the TF-CILF database (Tombeur, 1998). TF-CILF is a database collecting the vocabulary of the entire Latin world drawn from (a) the ancient Latin literature, (b) the literature of the patristic period, (c) a vast body

[4]Funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 658332-WFL, *Word Formation Latin* is a derivational morphology resource for Latin that links lemmas on the basis of word formation processes (`http://wfl.marginalia.it`).

[5]`http://ducange.enc.sorbonne.fr/doc/sources`.

[6]Because at the moment of writing there is no implemented distinction between onomastic and non-onomastic lemmas for what lemmas in Du Cange are concerned, we have taken into consideration onomastic data also in the Lemlat lexical basis.

[7]An ample bibliography on *-io* nouns in Latin is available. See for example (Fruyt, 1995) and (Fruyt, 2011).

of Medieval material and (d) collections of Neo-Latin works. Word forms are assigned their number of occurrences in each of these four periods.

Lemlat has been already proven to perform very efficiently on the TF-CILF dataset, as it is able to analyse 98.345% of the approximately 63 millions textual occurrences of the word forms it contains (Budassi and Passarotti, 2017).

We extracted from TF-CILF a list including those word forms that feature one of the possible inflectional endings of *-io* nouns (*-io*, *-ionis*, *-ionem*, *-ioni* etc.), together with data on their frequency of occurrence in the four periods of Latin mentioned above. In total we extracted 25,510 candidate word forms.

Then we processed these word forms with both Lemlat 3.0 and an enhanced version of it containing nouns ending in *-io* taken from Du Cange. This version of Lemlat was able to analyse 17,775 word forms out of the 25,510 extracted from TF-CILF. Such a low word coverage (69.79%) is consistent with the overall coverage of TF-CILF word forms provided by Lemlat 3.0 (72.25%) (Budassi and Passarotti, 2017). However, if we look at the number of textual occurrences of these unknown forms, they are extremely rare, which makes the textual coverage of Lemlat 3.0 largely reliable. The automatic processing allows not only to match each word form with a lemma, but also to exclude homographs like *capio* 'to seize' (verb). The resulting output (lemmas + frequency) can be graphically mapped on a temporal axis in order to have a complete view on the distribution of *-io* nouns through the ages.

5 Distribution of *-io* Nouns in Latin

Table 1 offers an overview of the total number of occurrences by period.[8] The vast majority of *-io* nouns are attested in the Middle Ages.

However, any evaluation of these results is going to be biased by the fact that the datasets for each period are not balanced. The size of the subsets covering respectively the Patristic and the Medieval period is bigger than that for Classical Latin. The subset for Neo-Latin is considerably smaller than those for the other periods. To give an idea of the difference in size between the four chronological subsets, Table 2 reports the total number of word forms and lemmas in TF-CILF by period.

	L	**L&D**	**D**
Antiquity	30,282	36,570	1,638
Patres	133,042	255,235	5,740
Medieval	216,220	541,049	14,299
Neo-Latin	19,551	45,145	1,812

Table 1: Absolute frequencies by period.

	Word Forms	**Lemmas**
Antiquity	5,726,051	229,587
Patres	21,982,097	310,348
Medieval	33,285,740	359,262
Neo-Latin	2,184,025	105,857
Total	**63,177,913**	**554,828**

Table 2: Number of word forms and lemmas in TF-CILF by period.

In order to flatten the difference in size between the subsets, relative values need to be used instead of absolute. Table 3 displays the distribution of *-io* nouns in Latin texts in terms of relative frequencies of occurrence by period.

	L	**L&D**	**D**
Antiquity	0.528%	0.638%	0.028%
Patres	0.605%	1.161%	0.026%
Medieval	0.649%	1.625%	0.042%
Neo-Latin	0.895%	2.067%	0.082%

Table 3: Relative frequencies by period.

For instance, looking at Table 3, it turns out that *-io* nouns that are only contained in Lemlat are 0.649% of the total number of occurrences in Medieval texts. Those contained in both Lemlat and Du Cange are 1.625%, and those contained in Du Cange (hence exclusively Medieval) are only 0.042%. An overview of the diachronic distribution of relative frequencies of occurrence of *-io* nouns is provided in Figure 1.

Figure 1 clarifies the variation of the presence of *-io* nouns in different chronological phases of Latin. The distribution of the occurrences of those *-io* nouns that were in the lexicon of Classical Latin (Lemlat line) remains fairly constant across all the diachronic phases of the language. In Neo-

[8] L stands for Lemlat only, L&D stands for Lemlat and Du Cange, D stands for Du Cange only. 'Antiquity' (i.e. up to the end of 2nd century AD), 'Patres' (i.e. 3rd century - 735 AD), 'Medieval' (i.e. 736 - 1499 AD) and 'Neo-Latin' (i.e. 1499 AD henceforth) are chronological parameters adopted by TF-CILF.

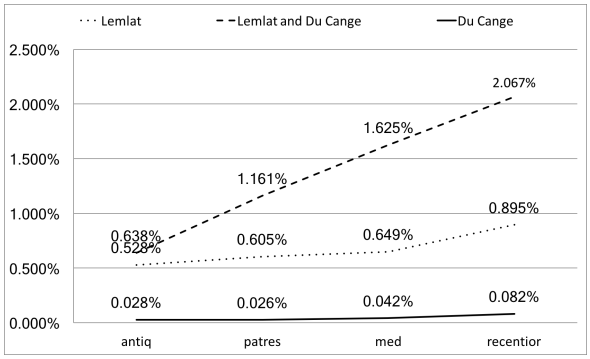

Figure 1: Distribution of relative frequencies of occurrence of *-io* nouns.

Latin times, however, a sharp increase is registered (from 0.649% to 0.895% in terms of relative frequencies). This peak is observable also as far as Medieval Latin *-io* nouns are concerned (Du Cange line). From a value of 0.042% in the Medieval period, the relative frequency raises until 0.082% in Neo-Latin. Nevertheless, the majority of *-io* nouns stored in both Lemlat's and Du Cange's lexical bases (which mostly underwent some semantic change across centuries) are the ones that live the best fate (Lemlat and Du Cange line): they constantly keep growing from the relative frequency value of 0.638% in the Antiquity to the relative frequency value of 2.067% in Neo-Latin.

The odd presence of words from Du Cange in Classical times is due to non-disambiguated homography. For instance, this is the case of the word *dubio*, which is analysed by Lemlat both as a form of the first class adjective *dubius* 'uncertain' (recorded in the original lexical basis of Lemlat, hence here left out) and as the nominative/vocative singular of the *-io* noun *dubio* (a type of hooked tool) from the Du Cange lexical basis.

6 General Discussion

The distribution of *-io* nouns reflects Zipf's law (Zipf, 1949), stating that the frequency of any word in a corpus is inversely proportional to its rank in the frequency table. To put it another way, there are a few *-io* nouns that are massively used, and a lot of *-io* nouns that are used only a few times.

The top most used nouns in *-io* throughout all periods are *ratio* 'reckoning', *passio* 'passion (of Christ)',[9] *oratio* 'speech' and *actio* 'action'. The most used words in Antiquity are *ratio*, *oratio*, *legio* 'legion' and *regio* 'region'. The top most frequent *-io* nouns in Patristic and Medieval times can all be found both in Lemlat and Du Cange. In Patristic literature, the most frequent words (from now on, after *ratio*) are *oratio*, *actio*, *passio* and *resurrectio* 'resurrection'. In Medieval times, they are *passio*, *oratio*, *operatio* 'activity' and *perfectio* 'perfection/completion'.

On another note, the high peak in the relative frequency of *-io* nouns in Neo-Latin texts suggests that these were used more often than others in more recent times. This can be explained by looking at the kind of texts included in the corpus. The texts contained in the Neo-Latin subset are mainly scientific and philosophical treatises, judicial texts, and the text of the Second Vatican Council. When these texts were written, Latin was not the spoken language anymore, as its place was mainly taken by Italian and French, two languages that inherited the suffix *-io* straight from Latin, especially for what learned vocabulary was concerned.[10] The assumption is that learned texts contained a large number of words resembling those used in Italian and French learned language, at least for what *-io* nouns are concerned. A look at the most used *-io* nouns in Neo-Latin texts confirms that once again *ratio* was the most used, followed by *propositio* 'statement of facts', *actio*, *notio* 'judicial enquiry', *definitio* 'definition' and *cognitio* 'examination'. These are also all contained in the Lemlat + Du Cange group.

7 Conclusions and Future Work

In this paper, we presented a study of the diachronic distribution of Latin nouns ending in *-io* by processing word forms from the TF-CILF corpus with the morphological analyser Lemlat. We demonstrated that the *-io* suffix is very productive across all periods of Latin language, showing a particularly high frequency in both Medieval and Neo-Latin texts. *Ratio* remains always the most used *-io* noun across the entire diachronic span covered by the corpus used in our work.

One step further in the study of *-io* nouns would be to establish derivational relationships for each lemma and to verify which of the two lexical groups (Lemlat or Du Cange) the input lemma belongs to. Also, an evaluation of the unknown word

[9] *Passio* is absent in Antiquity texts.

[10] See (Thornton, 1990), (Thornton, 1991) and (Štichauer, 2015).

forms after the lemmatisation process should be performed.

Given the wide lexical coverage provided by Lemlat, our work represents a positive example of how much NLP tools can help to investigate diachronic aspects of language. The wide diachronic as well as diatopic span over which Latin texts are spread opens an appealing challenge for research in NLP, which has to address the problem of portability of NLP tools across time, place and genre. In this sense, Latin texts represent a perfect dataset both for developing and for evaluating techniques of domain-adaptation of NLP tools.

References

David Bamman and Gregory Crane. 2008. Building a Dynamic Lexicon from a Digital Library, In *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2008, Pittsburgh)* ACM: New York.

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)*, 90-94, Association for Computational Linguistics: Berlin.

Marco Budassi and Marco Passarotti. 2017. The Impact of Unassimilated Loanwords on the Latin Lexicon. A Qualitative and Quantitative Analysis. In *Proceedings of DATeCH2017, Göttingen, Germany, June 01-02, 2017*, 85-90. DOI: http://dx.doi.org/10.1145/3078081.3078083.

Charles du Fresne Du Cange 1678-1887. *Glossarium Mediae et Infimae Latinitatis*, éd. augm., Niort, L. Favre http://ducange.enc.sorbonne.fr/.

Michèle Fruyt. 2011 Word-Formation in Classical Latin. In *A Companion to the Latin Language*, ed. James Clackson, 157-175, Wiley-Blackwell: Malden, Mass.

Michèle Fruyt. L'accusatif et les noms en-tio chez Plaute. *De usu, Études de syntaxe latine offertes en hommage à Marius Lavency*, 131-141.

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.

Thuillier Glorieux. 2010. Pourquoi informatiser un vieux glossaire? Présentation du Du Cange en ligne. *ÉLA* n°156, octobre-décembre 2009, Klincksieck.

Otto Gradenwitz. 1904. *Laterali Vocum Latinarum*. Hirzel: Leipzig.

Eleonora Litta, Marco Passarotti, and Chris Culy. 2016. Formatio formosa est. Building a Word Formation Lexicon for Latin. *Proceedings of the Third Italian Conference on Computational Linguistics (CLiC–it 2016). Napoli, aAccademia University Press*, 185-189.

Eleonora Litta, Marco Passarotti and Paolo Ruffolo. 2017. Node Formation. Using Networks to Inspect Productivity in Affixal Derivation in Classical Latin. In *Proceedings of DATeCH2017, Göttingen, Germany, June 01-02, 2017*, 103-108. DOI: http://dx.doi.org/10.1145/3078081.3078092.

Barbara McGillivray. 2013. *Methods in Latin Computational Linguistics* Brill: Leiden.

Barbara McGillivray and Marco Passarotti. 2009. The Development of the Index Thomisticus Treebank Valency Lexicon. In *Proceedings of LaTeCH-SHELT&R Workshop 2009, Athens, Greece*, 43-50, ACL.

Marco Passarotti, Berta González Saveedra and Christophe Onambele. 2016. Latin vallex. A treebank-based semantic valency lexicon for latin. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016) Portorož, Slovenia*, 2599–2606.

Marco Passarotti, Marco Budassi, Eleonora Litta and Paolo Ruffolo 2017. The Lemlat 3.0 Package for Morphological Analysis of Latin. *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*, 24–31.

Hannah Rosén. 1993. The mechanisms of Latin nominalization and conceptualization in historical view. In *ANRW 11I29. 1*, 178-211. De Gruyter: Berlin.

Uwe Springmann, Helmut Schmid and Dietmar Najock. 2016. LatMor: A Latin Finite-State Morphology Encoding Vowel Quantity. In Giuseppe Celano and Gregory Crane (eds.), *Treebanking and Ancient Languages: Current and Prospective Re-search (Topical Issue), Open Linguistics vol. 2*, 386–392.

Pavel Štichauer. 2015. From emergent availability to full profitability: The diachronic development of the Italian suffix -zione from the 16th to the 20th century. In Augendre S., Couasnon-Torlois G., Lebon D., Michard C., et al. *Proceedings of the Décembrettes 8th International conference on morphology*, 319-326, Université de Toulouse: Toulouse.

Anna Maria Thornton. 1990. Sui deverbali italiani in -mento e -zione (I). *Archivio glottologico italiano, LXXV/II*, 169-207. Le Monnier: Torino.

Anna Maria Thornton. 1991. Sui deverbali italiani in -mento e -zione (II). *Archivio glottologico italiano, LXXVI/I*, 79-102. Le Monnier: Torino.

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

George K. Zipf. 1949. *Human Behavior and the Principle of Least Effort*. Addison-Wesley Press: Cambridge, Mass.