LatInfLexi: an Inflected Lexicon of Latin Verbs

Matteo Pellegrini
Università di Bergamo/Pavia
Piazza Rosate, 2 –
24129 Bergamo, Italy
matteo.pellegrini@unibg.it

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

CEUR-WS.org/Vol-2253/paper23.pdf

Abstract

English. We present a paradigm-based inflected lexicon of Latin verbs built to provide empirical evidence supporting an entropy-based estimation of the degree of uncertainty in inflectional paradigms. The lexicon contains information on the inflected forms that occupy the 254 morphologically possible paradigm cells of 3,348 verbal lexemes extracted from a frequency lexicon of Latin. The resource also includes annotation of vowel length and the frequency of each form in different epochs.

Italiano. *Presentiamo un lessico di forme flesse basato sui paradigmi per i verbi latini, costruito per fornire evidenza empirica che permetta di quantificare il grado di incertezza nei paradigmi flessivi tramite l'entropia. Il lessico contiene informazioni sulle forme flesse che occupano le 254 celle possibili dal punto di vista morfologico di 3.348 lessemi verbali estratti da un dizionario frequenziale del latino. La risorsa include anche l'annotazione della lunghezza vocalica e la frequenza di ogni forma in diverse epoche*.

1 Introduction

In this paper, we describe the construction of LatInfLexi, an inflected lexicon of Latin verbs organized in lexemes[1] and paradigm cells.

In morphological theory, there is a recent trend towards a more realistic modelling of complex inflectional systems: for instance, Ackerman et al. (2009) and Bonami and Boyé (2014) propose that the analysis should take a full inflected form as a starting point, without assuming any segmentation *a priori*. In such approaches, what is investigated is not the construction of forms from smaller units like stems and inflectional endings, but rather their predictability given knowledge of other forms. This can be done by using the information theoretic notion of conditional entropy to estimate the uncertainty in guessing the content of the paradigm cell of a lexeme knowing another inflected form of the same lexeme, by weighting the probability of application of each inflectional pattern based on their type frequency in real data.

To do so, large-scale inflected lexicons listing all forms of a representative selection of lexemes are needed. Such resources are increasingly being developed for modern languages – see among else Zanchetta and Baroni (2005) and Calderone et al. (2017) for Italian, Neme (2013) for Arabic, Bonami et al. (2014) and Hathout et al. (2014) for French. However, to the best of our knowledge, there are no resources of this kind for Latin, although their (semi-)automatic building is made possible by the current availability of several morphological analyzers for Latin, including *Words* (http://archives.nd.edu/words.html), Lemlat (www.lemlat3.eu), *Morpheus* (https://github.com/tmallon/morpheus), the PROIEL Latin morphology system (https://github.com/mlj/proiel-

[1] The term "lexeme" is used for the abstract theoretical concept normally adopted in morphology and lexicology, while "lemma" refers to the concrete citation form representing an entry in dictionaries. Since we aim at a resource suitable for theoretical inquiries, we use the first term as a label in our resource.

webapp/tree/master/lib/morphology) and *LatMor* (http://cistern.cis.lmu.de). Our resource was created to fill this gap and to enable a quantitative, entropy-based analysis of Latin verb inflection.

2 Design

A distinctive feature of our inflected lexicon is that it is based on lexemes and paradigm cells, rather than on forms. This means that for each lexeme, all the morphologically possible paradigm cells are filled with a form, and not only those forms that are indeed attested in Latin texts are stored in paradigm cells. In this respect, our resource is similar to other recently developed inflected lexicons, like for instance Flexique for French (Bonami et al., 2014).

For each paradigm cell, the following information is provided:

(i) the inflected form that occupies the paradigm cell;
(ii) a univocal identifier of the lexeme to which it belongs;
(iii) the set of its morphological features;
(iv) information on the frequency of the form in different epochs.

As for (i), it should be noted that there is never more than one form per paradigm cell. In cases of overabundance (i.e. cells that are filled by more than one form, cf. Thornton, 2012), a choice was made to decide which "cell-mate" (Thornton, 2012: 183) should be kept, and which one discarded.

On the other hand, in some cases a paradigm cell could be empty, either because it is defective – like for instance the passive cells of intransitive verbs – or because it is not filled by a synthetic form, but rather it is analytically expressed, by means of a phrase – like for instance, in Latin, the perfective cells of deponent verbs, for which the periphrasis PRF.PTCP[2] + AUX *esse* 'to be' is used (e.g. PRF.IND.1SG *hortātus sum* 'I incited'). In both cases, the cell is marked as #DEF# in the resource. This convention is adopted also in Flexique (Bonami et al., 2014: 2585), and it fits the requirements of the Qumin package for entropy calculations on the predictability of implicative relations between inflected forms (Bonami and Beniamine, 2016; Beniamine, 2017).

As for (ii), the identifier corresponds to the citation form of the lexeme, almost always the first-person singular of the present indicative, following the Latin lexicographical and didactical tradition. A diacritic is added in those rare cases where different verbs have the same citation form (see *infra*, §3.2).

Regarding (iii), we use the PoS-tags of the Universal Part-of-Speech Tagset by Petrov et al. (2012) and the morphological features used in Universal Dependencies (http://universaldependencies.org/u/feat/index.html).

Lastly, the frequency data in (iv) are taken from Tombeur's (1998) *Thesaurus Formarum Totius Latinitatis* (see *infra*, §3.3).

3 Building the Lexicon

This section details the procedure followed to build the lexicon.

3.1 Selecting the Lexemes

Our first objective is to build an inflected lexicon of Latin featuring all the possible inflected forms of verbs only. To this aim, we include all the verbal entries contained in Delatte et al.'s (1981) *Dictionnaire fréquentiel et Index inverse de la langue latine* (henceforth DFILL). This yields a total of 3,348 verbs. In rare cases, more than one entry of DFILL corresponds to one and the same lexeme in our resource. This happens because some verbs are lemmatized twice in DFILL. For instance, for the verb *verso* two different entries appear in DFILL, using as citation form both the first-person singular of the present active indicative *verso* and the corresponding morphologically passive form *versor*. This choice is likely to be motivated by the different semantics of the two verbs, with the first one meaning 'to turn' and the second one meaning 'to remain'. However, in such cases our resource gives priority to collecting into one common inflectional paradigm all the forms that can be assigned to the same lexeme based on their morphological relatedness, rather than separating them in paradigms of different lexemes according to semantic criteria. Therefore, our lexicon includes only one lexeme *verso*, for which both active and passive forms are listed.

[2] Throughout the paper, we will refer to grammatical features by using the standard abbreviations of the Leipzig Glossing Rules.

3.2 Generating the Forms

In order to fill all of the paradigm cells of the selected lexemes, we exploit the database of Lemlat (Passarotti et al., 2017). For each lexeme, the database of Lemlat contains a list of segments called LES – roughly corresponding to the stems that are used in different subparadigms – each with a corresponding CODLES that provides (among else) information on the inflectional endings that can be attached to a LES. We make use of this information to generate the relevant forms.

To illustrate the details of the procedure, let's consider the verb *rumpo* 'to break'. For this verb, the database of Lemlat features the LESs and CODLESs shown in Table 1.

LES	CODLES
rump	v3r
rumpisse	fe
rup	v7s
rupsit	fe
rupt	n41
rupt	n6p1
ruptur	n6p2

Table 1: the verb *rumpo* in Lemlat 3.0

The two LESs with CODLES "fe" ("forma eccezionale", 'exceptional form') were discarded, since they are full irregular forms that are stored as such. As for the other LESs, the one with CODLES "v3r" is used to fill all the cells of the present system, by adding the inflectional endings of the conjugation represented by the CODLES (i.e. the 3rd conjugation). Similarly, the LES with CODLES "v7s" is used to fill the cells of the perfect system. From the remaining LESs, some nominal forms built upon the so-called "third stem" (Aronoff, 1994) can be derived, namely the supine *rupt-um* and *rupt-ū* from the LES with CODLES "n41", the perfect participle *rupt-us, -a, -um* from the LES with CODLES "n6p1" and the future participle *ruptūr-us*, *-a, -um* from the LES with CODLES "n6p2".

This given, our first step is to extract information on the LESs and CODLESs of each lexeme. Since Lemlat is a tool built to analyze rather than produce forms, it contains also several LESs occurring only in irregular and/or rare forms. To avoid the risk of overgeneration, we choose and keep only one LES for each CODLES. The choice is based on lexicographical sources, namely Lewis and Short (1879) and Glare (1982). In these dictionaries, at the very beginning of each verbal entry there is a set of four "principal parts" (Bennett, 1908: 55), i.e. exemplary inflected forms from which the whole paradigm of the lexeme can be inferred. We keep only those LESs that correspond to such principal parts, excluding the ones that correspond to more marginal forms that do appear in dictionaries but are given less prominence in the entry. For instance, Lemlat includes two LESs with CODLES "v3r" for the verb *dico* 'to say': "dic" and "deic". However, in both the lexicographical sources we use, the relevant principal parts are *dico* and *dicere*, corresponding to the first LES, while the second one is only mentioned later in the entries as an alternative form. Therefore, the LES selected for our resource is "dic".

We use the same dictionaries also to manually annotate the vowel length for each LES. This is a necessary enhancement, because in Latin verb inflection there are homographic forms that can be distinguished only based on that, like for instance PRS.ACT.IND.3SG *fugit* '(s)he flees' vs. PRF.ACT.IND.3SG *fūgit* '(s)he fleed'.

Following this process, we fill all the 254 paradigm cells of each of the 3,348 lexemes. However, because of Lemlat's design, for some quite frequent verbs with a highly irregular inflectional paradigm, it was not possible to apply the same procedure, at least for the cells of the present system, which is where most irregularity of the inflectional endings of Latin verbs happens. For the verbs shown in Table 2 and for those derived from them by prefixation (e.g. *abeo* 'to go away' from verb *eo* 'to go'), although it was technically possible to adopt a similar approach by using more than one LES for a CODLES, it proved to be faster and practical to manually record the correct forms as such.

Lemma	Meaning
aio	to say
eo	to go
fero	to bring
fio	to become
inquam	to say
malo	to prefer
nolo	not to want
possum	can
sum	to be
volo	to want

Table 2: irregular verbs

To each of the 850,392 generated paradigms cells, a univocal lexeme identifier is assigned,

which corresponds to the lemma used in Lemlat. In those rare cases where two or more verbs have the same lemma in Lemlat (although they inflect differently), a numeric diacritic is added to make the relevant distinction: for instance, we have *volo1* 'to fly' and *volo2* 'to want'.

3.3 Frequency Data

Many forms included in the paradigm cells of our lexicon are never attested in Latin texts. In order to make it possible to distinguish between plausible but unattested forms and those indeed occurring in texts, we enhance forms with information on their frequency. This information is taken from Tombeur's (1998) *Thesaurus Formarum Totius Latinitatis* (henceforth TFTL), where each form is assigned the number of its occurrences in four different epochs, respectively called *Antiquitas* (from the origins to the end of the 2nd century A.D.), *Aetas Patrum* (2nd century-735 A.D.), *Medium Aeuum* (736-1499) and *Recentior Latinitas* (1500-1965).

By including the frequency of each form in the lexicon, we know how many of the 752,537[3] forms recorded in the lexicon are never actually attested. Table 3 reports the relevant data[4].

TFTL epoch	unattested forms (%)
Antiquitas	544,395 (72.34%)
Aetas Patrum	482,324 (64.1%)
Medium Aeuum	484,421 (64.37%)
Recentior Latinitas	640,552 (85.12%)
all epochs	401,690 (53.38%)

Table 3: not attested forms

It can be observed that a significant amount of forms recorded in our lexicon are not attested, even in such a large corpus as the one the TFTL is based on. However, this is not surprising: recent large-scale corpus-based investigations (e.g. Bonami and Beniamine, 2016: 158 ff.) show that in languages with large inflectional paradigms – like the ones of Latin verbs – it is perfectly normal that many plausible forms do not appear, even in very large datasets, and the lexemes for which the full paradigm is attested are very few.

4 Discussion and Future Work

We described the design and building of a lexeme-based inflected lexicon consisting of 850,392 paradigm cells of 3,348 Latin verbs. Our first objective in the near future is to make the resource complete in terms of lexical coverage, including the lexemes of the other PoS. The lexicon is available for download as a .csv file at `https://github.com/matteo-pellegrini/LatInfLexi`.

We also plan to include phonetic annotation, by giving the IPA transcription of each form, which can be obtained semi-automatically by applying a script provided by the Classical Language Toolkit (Johnson et al., 2014-17) to stems and endings.

Another welcome addition would be to account for cases of overabundance, by allowing more than one form to appear in the same paradigm cell. However, to decide which cell-mates to keep and which ones to discard, their frequency in Latin texts should be preliminarily evaluated. In this respect, it has to be noted that the frequencies in the TFTL refer to bare surface forms, with no contextual disambiguation. For instance, the frequency of *veniam* comprises not only occurrences of both the PRS.ACT.SBJV.1SG and FUT.ACT.IND.1SG of the verb *venio* 'to come', but also of the ACC.SG of the noun *venia* 'indulgence'.

To get an idea of the impact of morphological ambiguity on our lexicon, we analyzed all the generated forms with Lemlat (version 3.0). We found that only for about 23% (170,735) of the 752,537 forms Lemlat outputs only one analysis (i.e. one lemma and one set of morphological features), the remaining 581,802 (about 77%) being ambiguous. This result weakens the reliability of the frequency data provided in the lexicon. Therefore, disambiguation is needed, although this would require a very time-consuming work.

However, to tackle the problem of ambiguity, a first useful step is distinguishing between cases like *veniam* above, which can be analyzed as an inflected form of two different lemmas, and cases where the different analyses only refer to different forms of the same lemma, e.g. *laudatis*, that appears both in the PRS.ACT.IND.2PL and in

[3] The 97,855 paradigm cells marked as #DEF# are excluded from this count.

[4] In total, the TFTL includes 554,828 different forms, corresponding to 62,922,781 occurrences in the reference corpus used by the Thesaurus. Our lexicon contains 165,898 of these unique forms (forms appearing in more than one paradigm cell are counted only once), for a total of 18,261,179 occurrences. This means that our resource covers around 30% of the forms of the TFTL, in terms of both type and token frequency. In addition, it also contains several other forms that are not attested in the TFTL (245,623 unique forms).

the PRF.PTCP.DAT/ABL.PL of *laudo* 'to praise', but cannot be a form of other lemmas. We call these different types 'exolemmatic' and 'endolemmatic' ambiguity, respectively (cf. Passarotti and Ruffolo, 2004). Cases of exolemmatic ambiguity are clearly more problematic, but they are also much rarer: only 79,490 (about 10%) of the forms in our resource belong to this type. The great majority of ambiguous forms only give rise to endolemmatic ambiguity, as can be observed in Table 4 below, where the relevant data are summarized.

	n.	%
unambiguous forms	170,735	22.69%
ambiguous forms	581,802	77.31%
only endolemmatic amb.	502,312	66.75%
exolemmatic amb.	79,490	10.56%

Table 4: the impact of ambiguity on frequency data

As far as endolemmatic ambiguity is concerned, although its quantitative impact is far greater, it could be considerably reduced in a principled manner. Indeed, it should be noted that in many cases this kind of ambiguity is due to systematic syncretism. For instance, the cells FUT.ACT.IMP.2SG and FUT.ACT.IMP.3SG are never unambiguously analyzed, because they are always identical for a same verb. Given the full systematicity of this syncretism, which holds for all lexemes, these cells could be considered as only one from a purely morphological point of view. Therefore, the problem of endolemmatic ambiguity could be at least reduced by adopting an approach based on "morphomic paradigms" (Boyé and Schalchli, 2016), where always syncretic cells are conflated, rather than on morphosyntactic paradigms. This would be helpful especially in nominal forms like participles and gerundives, where such cases of systematic syncretism are widespread.

When such ambiguity issues will have been resolved, it will also be possible to exploit the frequency data in a more systematic fashion, e.g. to perform diachronic investigations on how the frequency of specific (groups of) forms or paradigm cells change across the four considered epochs, or to model Latin inflectional morphology in an even more realistic way, by considering also the token frequency of inflected forms, as has been recently proposed by Boyé (2016).

References

Farrell Ackerman, James P. Blevins, and Robert Malouf. 2009. Parts and wholes: Implicative patterns in inflectional paradigms. In James P. Blevins and Juliette Blevins, editors, *Analogy in Grammar: Form and Acquisition*. Oxford University Press, Oxford: 54–82.

Mark Aronoff. 1994. *Morphology by itself: Stems and inflectional classes*. MIT Press, Cambridge/London.

Sacha Beniamine. 2017. Un algorithme universel pour l'abstraction automatique d'alternances morpho-phonologiques. In *24e Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*.

Charles Edwin Bennett. 1908. *New Latin Grammar*. Bolchazy-Carducci Publishers.

Olivier Bonami and Sarah Beniamine. 2016. Joint predictiveness in inflectional paradigms. *Word Structure* 9(2): 156–182.

Olivier Bonami and Gilles Boyé. 2014. De formes en thèmes. In Florence Villoing, Sophie David and Sarah Leroy, editors, *Foisonnements morphologiques: Études en hommage à Françoise Kerleroux*. Presses universitaires de Paris Ouest, Paris: 17–45.

Olivier Bonami, Gauthier Caron and Clément Plancq. 2014. Construction d'un lexique flexionnel phonétisé libre du français. In Franck Neveu, Peter Blumenthal, Linda Hriba, Annette Gerstenberg, Judith Meinschaefer and Sophie Prévost, editors, *Actes du quatrième congrès mondial de linguistique française*: 2583–2596.

Gilles Boyé. 2016. Pour une modélisation surfaciste de la flexion. Le cas de la conjugaison du français. In *SHS Web of Conferences*. Vol. 27. EDP Sciences.

Gilles Boyé and Gauvain Shalchli. 2016. The status of paradigms. In Andrew Hippisley and Gregory Stump, editors, *The Cambridge Handbook of Morphology*. Cambridge University Press, Cambridge: 206–234.

Basilio Calderone, Matteo Pascoli, Nabil Hathout and Franck Sajous. 2017. Hybrid method for stress prediction applied to GLAFF-IT, a large-scale Italian lexicon. In *International Conference on Language, Data and Knowledge*. Springer, Cham: 26–41.

Louis Delatte, Étienne Evrard, Suzanne Govaerts and Joseph Denooz. 1981. *Dictionnaire fréquentiel et index inverse de la langue latine*. L.A.S.L.A, Liege.

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

Nabil Hathout, Franck Sajous and Basilio Calderone. 2014. GLÀFF, a large versatile French lexicon. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*: 1007–1012.

Kyle P. Johnson et al. 2014-2017. *CLTK: The Classical Language Toolkit*. DOI 10.5281/zenodo593336.

Charlton Lewis and Charles Short. 1879. *A Latin Dictionary*. Clarendon, Oxford.

Alexis Amid Neme. 2013. A fully inflected Arabic verb resource constructed from a lexicon of lemmas by using finite-state transducers. *Revue RIST: revue de l'information scientifique et technique* 20(2): 7–19.

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

Marco Passarotti and Paolo Ruffolo. 2004. L'utilizzo del lemmatizzatore LEMLAT per una sistematizzazione dell'omografia in latino. *EUPHROSYNE* 32(A): 99–110.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2011. A universal part-of-speech tagset. *ArXiv*:1104–2086

Anna M. Thornton. 2012. Reduction and maintenance of overabundance. A case study on Italian verb paradigms. *Word Structure* 5(2): 183–207.

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

Eros Zanchetta and Marco Baroni. 2005. Morph-it!: a free corpus-based morphological resource for the italian language.