Proceedings of 1st Workshop on Language Technologies for Historical and Ancient Languages, pages 37–46
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Using LatInfLexi for an Entropy-Based Assessment of Predictability in Latin Inflection

Matteo Pellegrini
Università di Bergamo
piazza Rosate 2 – 24129 Bergamo (BG) – Italia
matteo.pellegrini@unibg.it

Abstract

This paper presents LatInfLexi, a large inflected lexicon of Latin providing information on all the inflected wordforms of 3,348 verbs and 1,038 nouns. After a description of the structure of the resource and some data on its size, the procedure followed to obtain the lexicon from the database of the Lemlat 3.0 morphological analyzer is detailed, as well as the choices made regarding overabundant and defective cells. The way in which the data of LatInfLexi can be exploited in order to perform a quantitative assessment of predictability in Latin verb inflection is then illustrated: results obtained by computing the conditional entropy of guessing the content of a paradigm cell assuming knowledge of one wordform or multiple wordforms are presented in turn, highlighting the descriptive and theoretical relevance of the analysis. Lastly, the paper envisages the advantages of an inclusion of LatInfLexi into the LiLa knowledge base, both for the presented resource and for the knowledge base itself.

Keywords: Lexicon, Morphology, Paradigm, Predictability, Entropy

1. Introduction

This paper presents LatInfLexi, an inflected lexicon of Latin verbs and nouns, and shows its place in the larger field of resources for the Latin language in general, and its usefulness in allowing for an entropy-based analysis of predictability in verb inflection in particular.

In studies on morphological theory, inflected wordforms are often considered to be composed of smaller, meaningful units, morphemes. Such an approach to word structure has been called 'constructive' by Blevins (2006; 2016). In this perspective, the goal is analyzing how exactly the relevant units are assembled in order to realize different Morphosyntactic Property Sets (MPS) for a given lexical item, in a 'syntagmatic' (Boyé and Schalchli, 2016), 'exponence-based' (Stump, 2015) fashion. Conversely, a different line of research, finding its roots in work on the implicative structure of paradigms within the framework of Natural Morphology (Wurzel, 1984), takes full inflected wordforms as the starting point, with smaller units possibly inferred only *a posteriori*, in an 'abstractive' (Blevins, 2006; Blevins, 2016) perspective. Similar approaches can be defined as implicative, in Stump (2015)'s terms, and 'paradigmatic', in Boyé and Schalchli (2016)'s terms: the focus is on implicative relations between wordforms, allowing to infer the content of a given paradigm cell assuming knowledge of the content of other cells.

This task has been stated in the question that Ackerman et al. (2009) call the 'Paradigm Cell Filling Problem' (PCFP): «What licenses reliable inferences about the inflected (and derived) surface forms of a lexical item?». In the last decade, this question has received remarkable attention in the morphological literature, especially within two related but different frameworks. A set-theoretic approach is represented by Stump and Finkel (2013)'s Principal Part Analysis, that aims at finding sets of inflected wordforms ('Principal Part Sets') from which the content of the whole paradigm of a lexeme can be inferred. Another way of tackling the PCFP is quantifying the contribution of each inflected wordform to predictability, estimating the uncertainty in guessing the content of individual cells, rather than trying to fill the whole paradigm as in Principal Part Analysis. This second possibility has been modelled in information-theoretic terms, using conditional entropy (Ackerman et al., 2009). In this way, it is also possible to weigh the impact of different inflectional patterns according to their type frequency (Bonami and Boyé, 2014; Beniamine, 2018).

However, this presupposes the availability of large, representative inflected lexicons for the languages under investigation. Indeed, similar resources are being increasingly developed for modern Indo-European languages: see, among else, the CELEX database (Baayen et al., 1996) for Dutch, English, and German, Flexique (Bonami et al., 2014) and GLÀFF (Hathout et al., 2014) for French, Morph-it! (Zanchetta and Baroni, 2005) and GLÀFF-IT (Calderone et al., 2017) for Italian. The availability of inflected lexicons is much more limited for historical languages like Latin, despite the growing amount of resources and NLP tools developed for such languages in the last years (Piotrowski, 2012; Bouma and Adesam, 2017), among which also lexical resources, like the derivational lexicon Word Formation Latin (Litta et al., 2016). As for inflected lexicons, the only easily available resource is the one provided within the Unimorph[1] project (Sylak-Glassman et al., 2015). However, the data of this resource display issues of lack of homogeneity and systematicity, due to the collaborative design of the source from which they are taken, namely Wiktionary. On the other hand, it would be possible to obtain an inflected lexicon without such shortcomings semi-automatically, using the information contained in morphological analyzers such as *Words*,[2] *Morpheus*,[3] *LatMor*,[4] and the PROIEL Latin morphology

[1] `http://unimorph.org/`.
[2] `http://archives.nd.edu/words.html`.
[3] `https://github.com/tmallon/morpheus`.
[4] `http://cistern.cis.lmu.de`.

system.[5]

This paper details, in Section 2., the procedure that was followed to exploit one of these morphological analyzers – namely, the recently renewed Lemlat 3.0 (Passarotti et al., 2017) – in order to obtain LatInfLexi, a paradigm-based inflected lexicon of Latin. Section 3. shows how the data in LatInfLexi allow for a quantitative, entropy-based analysis of predictability in Latin verb inflection that on the one hand recovers traditional notions such as Principal Parts on a more solid ground, on the other hand sheds new light on Latin paradigm structure, revealing patterns of inter-predictability between wordforms that are less trivial than the ones that are usually identified. Section 4. discusses the possible use of LatInfLexi to enhance the LiLa knowledge base (Passarotti et al., 2019), providing information not only on wordforms that are attested in the texts included therein, but also on unattested, but nevertheless possible wordforms, also highlighting the advantages for LatInfLexi itself of a connection with the textual resources in LiLa. In conclusion, Section 5. summarizes the main points of the paper.

2. The Resource: LatInfLexi

This section is devoted to a careful description of LatInflexi, starting in 2.1. from a few words on its design and overall structure. Some quantitative data on the size of the resource and its coverage of the Latin lexicon are then given in 2.2.. In 2.3., the procedure followed to generate inflected wordforms from the information provided in Lemlat 3.0 is detailed, regarding both verbs and nouns. Lastly, 2.4. explains and motivates the choices made in the resource for cases of non-canonical filling of paradigm cells, namely defectiveness and overabundance.

2.1. Design

The overall structure of LatInfLexi is based on lexemes and paradigm cells, rather than on attested wordforms. This means that for each nominal and verbal[6] lexeme, we list all the paradigm cells, providing the following information for each of them:

- the lexeme to which the cell refers, notated through the citation form used in Lemlat;
- its PoS-tag and the MPS realized by the cell, notated through Petrov et al. (2011)'s 'Universal Part-Of-Speech Tagset' and the features used in the Universal Dependencies[7] project (Nivre et al., 2016);
- the inflected wordform filling the cell, in both orthographical and phonological, IPA, transcription;
- its frequency according to Tombeur (1998)'s *Thesaurus Formarum Totius Latinitatis*, across different epochs: *Antiquitas*, from the origins to the end of the 2nd century A.D.; *Aetas Patrum*, from the 2nd century to 735; *Medium Aeuum*, from 736 to 1499; *Recentior Latinitas*, from 1500 to 1965.

2.2. Size

The selection of lexemes is frequency-based. LatInfLexi contains all the 3,348 verbs reported in the *Dictionnaire fréquentiel et Index inverse de la langue latine* (Delatte et al., 1981). Regarding nouns, only those with a frequency of 30 or more are kept, for a total of 1,038.

For each noun, a 12-cells paradigm is given, as generated by various combinations of different values of the inflectional categories of number – singular vs. plural – and case – nominative, genitive, dative, accusative, vocative, ablative. In the currently distributed version of LatInfLexi, the locative case is not considered because of its marginality, being attested almost only in names of towns and small islands. This exclusion is due to practical reasons: since the resource was originally conceived to allow for a quantitative analysis of predictability, for a cell attested in so few lexemes it would not have been possible to obtain significant results. However, the plan is to add the locative too, to make the resource more complete.

As for verbs, the provided paradigms are made up of 254 cells, generated by the combinations of values of tense-aspect (present, perfect, future, imperfect, pluperfect, future perfect), mood (indicative, subjunctive, imperative, infinitive), voice (active vs. passive), person and number. They include also nominal and adjectival forms inflected for case and (only the adjectival ones) for gender, for instance gerunds and participles. On the other hand, paradigm cells that are always filled analytically by means of a periphrasis, rather than with a dedicated, synthetic inflected wordform, are excluded: for instance, there is no cell PRF.PASS.IND.1SG, since passive perfective cells are always realized by means of a periphrasis composed by the perfect participle of the relevant verb and the appropriately inflected form of the verb 'to be', e.g. *amātus sum* 'I was loved'.

Table 1 summarizes some data on the overall size of the lexicon.

	verbs	nouns
lexemes	3,348	1,038
paradigm cells	850,392	12,456
wordforms	752,537	12,355
distinct wordforms	434,040	7,307

Table 1: The size of LatInfLexi

The number of wordforms does not match the number of cells because there are cells that are marked as defective (#DEF#) in LatInfLexi: they do not contain any inflected wordform. Further details on such cases can be found in 2.4. On the other hand, the difference between the sheer number of wordforms and the number of distinct wordforms is due to cases of more or less systematic syncretism, where the same surface wordform appears in different cells: for instance, in nominal inflection the dative and ablative plural are always realized in the same way. It is interesting to compare the number of distinct wordforms in our resource to the ones reported in the very extensive database of Tombeur (1998), that lists all the

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

[6] Adjectives have not been included in the current version because LatInfLexi was originally conceived to allow for an entropy-based analysis of verb and noun inflection, but the plan for the near future is to add adjectives too.

[7] http://universaldependencies.org/u/feat/index.html.

forms attested in a very large corpus of Latin, also providing information on their frequency in different epochs (see above, 2.1.). Out of the 554,828 wordforms attested in Tombeur (1998), 183,579 are present also in LatInfLexi, that thus cover for about one third of the forms of Tombeur (1998). This proportion is remarkable, especially considering that LatInfLexi only contains verbs and nouns, systematically excluding other lexical categories, even open ones like adjectives and adverbs. Furthermore, it should be noticed that LatInfLexi, thanks to its previously mentioned paradigm-based design, also contains many inflected wordforms (257,768 distinct wordforms) that are not attested in the texts on which Tombeur (1998) is based.

2.3. Generation of Wordforms

The database of Lemlat 3.0, a large and recently renewed morphological analyzer for Latin, was exploited to generate full paradigms for all the lexemes of our sample. For each lemma, Lemlat reports one or more 'LExical Segment(s) (LES), roughly corresponding to the stem(s) appearing in the various inflected wordforms. Every LES is equipped with a CODLES, from which plenty of information can be inferred, for instance on the subset of paradigm cells where the CODLES can be used and on the inflectional endings that are compatible with it. As an example, for the verb STO 'stay', Lemlat lists the LESs and CODLESs given in Table 2 below.

LES	CODLES
st	v1i
ist	v1i
stet	v7s
stat	n41
stat	n6p1
statūr	n6p2

Table 2: LESs and CODLESs of STO 'stay'

The CODLES 'v1i' is used for LESs that correspond to the stem traditionally labelled as 'present stem', appearing in the so-called 'present system' – i.e., in imperfective cells – in intransitive ('i') 1st conjugation ('1') verbs ('v'). The CODLES 'v7s' instead marks LESs that correspond to the 'perfect stem', appearing in the 'perfect system' – i.e., in perfective cells. The remaining CODLESs identify stems used in nominal forms ('n'), namely the supine ('n41') and the perfect ('n6p1') and future ('n6p2') participle, corresponding to what Aronoff (1994) calls the 'third stem', and other stems derived from it, like the one of the future participle.

The first step of the procedure consists in extracting all LESs and CODLESs for each of the selected lexemes and matching them to the stems used in the principal parts provided by Latin dictionaries – in particular, Lewis and Short (1879), that is used as the primary source of information, due to its easy availability in machine-readable format. On the one hand, this allows to decide what LES should be selected in cases – like the one of Table 2 – where more than one LES with the same CODLES is present in Lemlat. For instance, the principal parts of STO in Lewis and Short (1879) are *stō*, *stetī* and *statum*, filling the cells PRS.ACT.IND.1SG, PRF.ACT.IND.1SG and SUP.ACC, respectively. Therefore, between the two LESs with CODLES 'v1i' given in Table 2, only the first one is kept, since it corresponds to the stem appearing in the wordform used as principal part, while the second one is in Lemlat only because it is reported in dictionaries as a marginal variant sometimes attested in texts. On the other hand, the information that can be inferred from the principal parts of Lewis and Short (1879) and other dictionaries is more detailed than the one in Lemlat regarding the phonological shape of the stems, since there is also a coding of vowel length and of the distinction between the vowels /i/, /u/ (<i>, <u>) and the semivowels /j/, /w/ (<j>, <v>). Since our lexicon aims to be as surface-true as possible, the LESs of Lemlat are enhanced with this additional information. This also allows to automatically obtain phonological transcriptions in IPA notation.

After the extraction of LESs, by attaching the endings of the 1st conjugation to the ones with CODLES 'v1i', the imperfective forms of the present system are generated – but not the passive ones, that are defective because the verb is intransitive, except for the ones referring to the third-person singular, attested in an impersonal usage (e.g. *stātur* 'one stays'). The LESs with CODLES v7s can be used to generate perfective forms of the perfect system, again by attaching the appropriate endings, that are the same for all conjugations. The other LESs are used to generate supine and participial wordforms, adding the relevant nominal/adjectival endings. The procedure is illustrated in Table 3 below.

LES	CODLES	cell	wordform
st	v1i	PRS.ACT.IND.1SG	*st-ō*
		PRS.ACT.IND.3SG	*st-at*
		PRS.PASS.IND.1SG	#DEF#
		PRS.PASS.IND.3SG	*st-ātur*
		...	...
stet	v7s	PRF.ACT.IND.1SG	*stet-ī*
		PRF.ACT.IND.3SG	*stet-it*
		...	...
stat	n41	SUP.ACC	*stat-um*
		SUP.ABL	*stat-ū*
stat	n6p1	PRF.PTCP.NOM.M.SG	*stat-us*
		...	...
statūr	n6p2	FUT.PTCP.NOM.M.SG	*statūr-us*
		...	...

Table 3: Generation of some inflected wordforms of STO 'to stay'

The procedure followed for nouns was very similar, the only difference being that for a given lexeme there are not multiple LESs with different CODLESs to be used in different sections of the paradigm, but only one (or more) LES with a CODLES corresponding to the inflectional (sub)class. In most cases, all the inflected wordforms can be generated from the LES and CODLES alone. For instance, Table 4 and Table 5 illustrate the generation of some wordforms of the 1st declension noun ROSA 'rose' and of the 5th declension noun RES 'thing', respectively.

On the other hand, in 3rd declension nouns and in some 2nd declension nouns, a different stem allomorph appears in some cells, namely NOM.SG and VOC.SG in masculine and feminine nouns and ACC.SG too in neuter nouns, where this cell is systematically syncretic with NOM.SG and VOC.SG.

LES	CODLES	cell	wordform
ros	n1	NOM.SG	*ros-a*
		GEN.SG	*ros-ae*
		ACC.SG	*ros-am*
		...	...

Table 4: Generation of some inflected wordforms of ROSA 'rose'

LES	CODLES	cell	wordform
r	n5	NOM.SG	*r-ēs*
		GEN.SG	*r-eī*
		ACC.SG	*r-em*
		...	...

Table 5: Generation of some inflected wordforms of RES 'thing'

Differently than what happens for verbs, the shape of this allomorph is not explicitly coded with a dedicated LES and a specific CODLES. However, in Lemlat, for all lemmas, under the heading LEM, information on how to produce the citation form is provided. Since the citation form used for nouns is exactly NOM.SG, and the other cells are syncretic with NOM.SG whenever they display a different allomorph, this information was exploited to fill the cells displaying stem allomorphy in our resource, as illustrated below in Table 6 by the allomorphic 2nd declension noun APER 'boar' and in Table 7 by the 3rd declension noun AGMEN 'multitude (of men/animals)'.

LES	CODLES	LEM	cell	wordform
apr	n2	aper	NOM.SG	*aper*
			GEN.SG	*apr-ī*
			ACC.SG	*apr-um*
			...	...

Table 6: Generation of some inflected wordforms of APER 'boar'

LES	CODLES	LEM	cell	wordform
agmin	n3	agmen	NOM.SG	*agmen*
			GEN.SG	*agmin-is*
			ACC.SG	*agmen*
			...	...

Table 7: Generation of some inflected wordforms of AGMEN 'multitude (of men/animals)'

2.4. Defectiveness and Overabundance

As was hinted above, LatInfLexi aims at providing full paradigms for all its lexemes. Therefore, every paradigm cell is filled with a wordform, whenever this is possible. This choice is reasonable, since in the usual, 'canonical' (Corbett, 2005) situation each paradigm cell is expected to be realized by exactly one inflected wordform.

However, it is a well-known fact that there are non-canonical cases of defectiveness (Sims, 2015), i.e. empty cells, for which the corresponding inflected wordform is not only unattested, but indeed non-existent. For instance, in Latin intransitive verbs are defective of passive wordforms, except for the third-person singular that can be used with an impersonal meaning (cf. above, 2.3., Table 3). Conversely, deponent verbs (Grestenberger, 2019) are always defective of morphologically active wordforms. Impersonal verbs only display third-person singular wordforms, as well as infinitives, gerunds and participles, but are systematically defective in all other cells. Regarding nouns, *pluralia tantum* do not have singular wordforms. In all such cases, the defective paradigm cells are not filled with a wordform, but simply marked as such (#DEF#) in LatInfLexi. In verb paradigms, also cells for which the stem that should be used to generate the corresponding wordform is not reported in Lemlat are marked as defective: for instance, for the verb ALBEO 'to be white', only the LES corresponding to the present stem is reported in Lemlat, thus perfective forms and the nominal forms based on the third stem are marked as #DEF#.

Another non-canonical phenomenon concerning paradigms is overabundance – multiple filling of the same cell by different wordforms (Thornton, 2019). In the current version of LatInfLexi, each non-defective cell contains exactly one wordform. In cases where more than one wordform could potentially be generated for the same paradigm cell – either because more than one LES with the same CODLES is available, or because different endings would be compatible with a given LES – a choice was made on which wordform to keep and which one(s) to discard, based on the principal parts reported in dictionaries in the former case (as showed above in 2.3.), while in the latter case the wordforms outputted in the inflectional tables of the Collatinus toolkit[8] are used.

3. An Entropy-Based Assessment of Predictability in Latin Verb Paradigms

This section illustrates how the data of LatInfLexi can be used for a quantitative, entropy-based analysis of predictability in Latin verb inflection. After an explanation, in 3.1., of the procedure that was followed, the results obtained on Latin verb paradigms are presented in 3.2., first focusing on predictions from one form (3.2.1.) and then extending the investigation to predictions from more than one form (3.2.2.).

3.1. The Method

In general, entropy (H) is a measure of uncertainty about the outcome of a random variable: the more the uncertainty, the higher the entropy value. Entropy increases with the number of possible outcomes: for instance, the entropy of a coin flip, with two possible outcomes, is higher than the entropy of rolling a dice, where the possible outcomes are six. Conversely, entropy decreases if the different outcomes are not equiprobable: the entropy of a coin flip is lower if the coin is rigged to always or often come up heads. Bonami and Boyé (2014) propose a method to estimate the uncertainty in predicting one cell from another one by means of conditional entropy – $H(A|B)$, a measure of the uncertainty about the outcome of a random variable A, given the value of another random variable B. To illustrate

[8] https://outils.biblissima.fr/fr/collatinus-web/.

their procedure, let us consider in Table 8 the phonological shape of the inflected wordforms filling the paradigm cells PRS.ACT.IND.1SG and PRS.ACT.IND.2SG for Latin verbs belonging to different conjugations, explaining how the conditional entropy of guessing the latter given the former can be computed.

lexeme	conj.	PRS.ACT. IND.1SG	PRS.ACT. IND.2SG
AMO 'love'	1st	amo:	ama:s
MONEO 'warn'	2nd	moneo:	mone:s
SCRIBO 'write'	3rd	skri:bo:	skri:bis
CAPIO 'take'	mix.[9]	kapio:	kapis
VENIO 'come'	4th	wenio:	weni:s

Table 8: PRS.ACT.IND.1SG and PRS.ACT.IND.2SG of Latin verbs of different conjugations

The first step of Bonami and Boyé (2014)'s methodology consists in extracting alternation patterns between the wordforms, and contexts where such alternation patterns can be applied, as the second column of Table 9 illustrates. The second step is a classification of lexemes according to the patterns that can potentially be applied, based on the phonological makeup of the patterns themselves and of the extracted contexts. The outcome of this classification is given in the third column of Table 9. Verbs of the 1st and 3rd conjugation are in the same class, because patterns 1 and 3 can both be applied to a PRS.ACT.IND.1SG ending in /o:/ preceded by a consonant; similarly, verbs of the 4th and mixed conjugation are in the same class, because faced with a PRS.ACT.IND.1SG ending in /io:/ preceded by a consonant, both pattern 4 and pattern 5 can be applied.

lexeme	pattern/context (1SG ↔ 2SG)	applicable patterns	n. verbs
AMO	1. _o: ↔ _a:s / C_#	A. (1,3)	1,332
MONEO	2. _eo: ↔ _e:s / C_#	B. (2)	298
SCRIBO	3. _o: ↔ _is / C_#	A. (1,3)	1,152
CAPIO	4. _o: ↔ _s / i_#	C. (4,5)	132
VENIO	5. _io: ↔ _i:s / C_#	C. (4,5)	169

Table 9: Information used to compute $H(PRS.ACT.IND.2SG|PRS.ACT.IND.1SG)$

Given these two cross-cutting classifications and information on the number of verbs in which the various alternation patterns occur (given in the last column of Table 9 with data taken from LatInfLexi), it is possible to compute the conditional entropy of guessing PRS.ACT.IND.2SG from PRS.ACT.IND.1SG in each of the classes based on applicable patterns, using the type frequency of alternation patterns as an estimate of their probability of application. In class B (see (1), b.) there is no uncertainty: given a PRS.ACT.IND.1SG in /eo:/, the PRS.ACT.IND.2SG cannot but be in /e:s/.[10] In classes A and C (cf. (1), a. and c.) there are competing patterns (1 vs. 3 and 4 vs. 5), and therefore there is some uncertainty, whose impact can be quantified by means of the number of verbs in which each pattern occurs. The results regarding the different classes can then be put together – again weighing them on the basis of type frequency, as is shown in (1)d. – to obtain a single entropy value, estimating the uncertainty in guessing the content of PRS.ACT.IND.2SG knowing the wordform filling PRS.ACT.IND.1SG. This value is called 'implicative entropy' by Bonami (2014).

(1) $H(PRS.ACT.IND.2SG|PRS.ACT.IND.1SG)$

a. Class A:
$H = -\left(\left(\frac{1,332}{2,484} \times log_2 \frac{1,332}{2,484} \right) + \left(\frac{1,152}{2,484} \times log_2 \frac{1,152}{2,484} \right) \right) = 0.996$

b. Class B:
$H = -(1 \times log_2 1)$

c. Class C:
$H = -\left(\left(\frac{132}{301} \times log_2 \frac{132}{301} \right) + \left(\frac{169}{301} \times log_2 \frac{161}{309} \right) \right) = 0.989$

d. Overall:
$H = \left(\frac{2,484}{3,083} \times 0.996 \right) + \left(\frac{298}{3,083} \times 0 \right) + \left(\frac{301}{3,083} \times 0.989 \right) = 0.899$

This procedure has two crucial advantages with respect to other entropy-based quantitative measurements of inflectional predictability proposed in the literature (cf. e.g. Ackerman et al. (2009) and subsequent work). Firstly, this methodology takes the type frequency of different patterns into account, rather than relying on the simplifying assumption that all inflection classes are equiprobable. Secondly, it does not require a pre-existing classification of inflection classes, since alternation patterns and contexts can simply be inferred from the surface phonological shape of the inflected wordforms.

3.2. Applying the Method to Latin Verb Paradigms

Thanks to the freely available Qumin[11] toolkit (Beniamine, 2018), it is possible to automatically perform implicative entropy computations according to Bonami and Boyé (2014)'s procedure on all the inflected wordforms of LatInfLexi, obtaining the results that will be presented in the following sub-sections.

3.2.1. Predicting from One Form: Zones of Interpredictability in Latin Verb Inflection

To have a first overall picture of predictability in Latin verb paradigms, implicative entropy values are computed for each pair of cells. A first relevant fact that should be noticed is that for a lot pairs of cells (A,B) the entropy values of both $H(A|B)$ and $H(B|A)$ are null, meaning that knowing one of the two inflected wordforms involved, the other one can be predicted with no uncertainty, since they are in systematic covariation: for instance, given the present active infinitive of a verb, the cells of the imperfect active subjunctive can always be obtained by adding personal endings to it, no matter how irregular the infinitive, and vice-versa, as is shown in (2).

[9] The conjugation of CAPIO is called 'mixed', as in Dressler (2002), because it displays the endings of the 3rd conjugation in some cells and the endings of the 4th conjugation in other cells.

[10] For the sake of simplicity, in this example we disregard highly irregular verbs, as well as verbs whose PRS.ACT.IND.1SG ends in /eo:/ that belong to the 1st conjugation and thus have PRS.ACT.IND.2SG in /ea:s/ (e.g. CREO 'create', PRS.ACT.IND.1SG *creō*, PRS.ACT.IND.2SG *creās*).

[11] https://github.com/XachaB/Qumin

(2) PRS.ACT.INF X ↔ PRS.ACT.SBJV.1SG X*m*

a. AMO 'love':
PRS.ACT.INF *amāre* ↔ PRS.ACT.SBJV.1SG *amārem*

b. FERO 'bring':
PRS.ACT.INF *ferre* ↔ PRS.ACT.SBJV.1SG *ferrem*

Similar categorical implicative relations can be exploited to obtain a mapping of the Latin verbal paradigm in zones of full interpredictability: within such zones, all cells can be predicted from one another with no uncertainty. This mapping is sketched in Table 10 (for active[12] verbal forms) and Table 11 (for nominal and adjectival forms) below, with cells that belong to the same zone sharing the same color and index (Z1-15), and different shades of the same color used to visualize zones that are closer to one another in terms of mutual predictability.

ACT	1SG	2SG	3SG	1PL	2PL	3PL
IPRF.IND	Z1	Z1	Z1	Z1	Z1	Z1
IPRF.SBJV	Z2	Z2	Z2	Z2	Z2	Z2
PRS.IMP		Z3			Z2	
PRS.IND	Z4	Z5	Z6	Z2	Z2	Z7
FUT.IMP		Z2	Z2		Z2	Z7
FUT.IND	Z8	Z8	Z8	Z8	Z8	Z8
PRS.SBJV	Z9	Z9	Z9	Z9	Z9	Z9
PRF.IND	Z10	Z10	Z10	Z10	Z10	Z10
PLUPRF.IND	Z10	Z10	Z10	Z10	Z10	Z10
FUTPRF.IND	Z10	Z10	Z10	Z10	Z10	Z10
PRF.SBJV	Z10	Z10	Z10	Z10	Z10	Z10
PLUPRF.SBJV	Z10	Z10	Z10	Z10	Z10	Z10

Table 10: Zones of interpredictability in Latin verb paradigms: verbal forms (active only)

			GDV	PRS. PTCP	PRF. PTCP	FUT. PTCP
PRS.INF.ACT	Z2	NOM.SG	Z12	Z13	Z14	Z15
PRS.INF.PASS	Z11	GEN	Z12	Z12	Z14	Z15
PRF.INF.ACT	Z10	DAT	Z12	Z12	Z14	Z15
GER.GEN	Z12	ACC	Z12	Z12	Z14	Z15
GER.DAT	Z12	VOC.N.SG	Z12	Z13	Z14	Z15
GER.ACC	Z12	VOC.M/F.SG	Z12	Z12	Z14	Z15
GER.ABL	Z12	ABL	Z12	Z12	Z14	Z15
SUP.ACC	Z14	NOM.PL	Z12	Z12	Z14	Z15
SUP.ABL	Z14	VOC.PL	Z12	Z12	Z14	Z15

Table 11: Zones of interpredictability in Latin verb paradigms: nominal and adjectival forms

Therefore, although the sheer number of cells in Latin verb paradigms is very high, in many cases the presence of different wordforms does not contribute to uncertainty in the PCFP, since such wordforms can be predicted from other wordforms in the same zone. In this way, the 254-cells paradigm of LatInfLexi can be reduced to only 15 zones between which there is not full interpredictability. To go into some more detail, Z10 corresponds to what traditional descriptions call the 'perfect system', containing cells based on the perfect stem. The cells that Aronoff (1994) identifies as based on the 'third stem' correspond to two different zones (Z14 and Z15) in our mapping because there actually are a few cases where the future participle is based on a different stem than the perfect participle and supine (e.g. PRF.PASS.PTCP.NOM.SG *mortu-us* vs. FUT.ACT.PTCP.NOM.M.SG *morit-ūrus*). As for what traditional descriptions label the 'present system', containing imperfective wordforms based on the present stem, it proves to be split between several (13) zones. This happens because with the adopted methodology not only the uncertainty generated by stem allomorphy is taken into account, but also the impact of the opacity of some endings with respect to inflection class assignment – witness the example provided above in Table 8, where the endings of PRS.ACT.IND.1SG are partly uninformative on the inflectional behavior of PRS.ACT.IND.2SG, because the ending *-ō* is ambiguous between the 1st and 3rd conjugation, and the ending *-iō* between the 4th and mixed conjugation.

It is interesting to observe that, if compared with the picture that would emerge by only considering the role of stem allomorphy, the mapping of the paradigm summarized in Table 10 and Table 11 is much more similar to the situation found in Romance verb inflection, with several zones of interpredictability, as shown e.g. by Bonami and Boyé (2003) for French, Pirrelli and Battista (2000) and Montermini and Bonami (2013) for Italian, Boyé and Cabredo Hofherr (2006) for Spanish. For instance, Table 10 shows that the cells PRS.ACT.IND.1SG and PRS.ACT.IND.3PL are very distant from the other present active indicative cells in terms of interpredictability. Thus, the overall picture is similar to the one produced by what Maiden (2018, pp. 84 ff.) calls 'U-pattern' in Romance languages. This suggests that there might be more continuity from Romance to Latin regarding paradigm structure than is usually assumed in diachronic accounts of this topic, like e.g. Maiden (2009).

Having identified these 15 zones of interpredictability, it is possible to take advantage of them to obtain a more compact version of the Latin paradigm, where only one cell per zone is kept. This allows to focus on the cases where there is some uncertainty and compare the different levels of predictability of different zones. To this aim, for each selected cell X, the values of average cell predictability – i.e., the average implicative entropy of predicting cell X knowing each of the other chosen cells – and average cell predictiveness – i.e., the average implicative entropy of predicting each of the other cells knowing cell X – are computed and given in Table 12a-b, sorted by decreasing entropy values. It can be observed that while the values of predictability are in a narrower range, the various zones display remarkable differences in their predictiveness: in particular, Z4 (the zone of the first-person singular of the present indicative) has a very low predictiveness, because of the above-mentioned opacity of the endings of that cell, that is poorly informative on the overall inflectional behavior of the lexemes (see again Table 8 above).

3.2.2. Predicting from More than One Form: (Near) Principal Parts

In the previous sub-section, the implicative entropy of guessing the content of one cell given knowledge of in-

[12] Passive wordforms can be inferred from their active counterpart with no uncertainty, and they are therefore not reported in Table 10 for reasons of space.

a

zone	average cell predictability
Z13	0.208271
Z1	0.229066
Z4	0.231378
Z12	0.240871
Z11	0.244131
Z14	0.255304
Z15	0.263901
Z6	0.269721
Z9	0.302484
Z8	0.309003
Z7	0.311636
Z3	0.315026
Z5	0.315126
Z2	0.342413
Z10	0.343957

b

zone	average cell predictiveness
Z8	0.079394
Z7	0.089352
Z9	0.127819
Z2	0.130643
Z3	0.13107
Z5	0.166161
Z11	0.189036
Z15	0.257111
Z14	0.266108
Z1	0.3122
Z12	0.348084
Z6	0.355214
Z13	0.370468
Z10	0.442993
Z4	0.916636

Table 12: Average cell predictability and predictiveness

dividual wordforms – what Bonami and Beniamine (2016) call 'unary implicative entropy' – was used in order to obtain an overall assessment of predictability in Latin verb paradigms. However, Bonami and Beniamine (2016) argue that, in languages with large paradigms, in many cases speakers are exposed to more than one inflected wordform of a lexeme without being exposed to all of them: therefore, it is reasonable to extend the investigation to predictions from more than one wordform, using what Bonami and Beniamine (2016) call '*n*-ary (binary, ternary etc.) implicative entropy'. Table 13 compares average unary implicative entropy – i.e., the entropy of guessing paradigm cells from one another, averaged across all pairs of cells – with average *n*-ary implicative entropy at different cardinalities – i.e., using combinations of *n* forms as predictors. These results show that knowledge of multiple wordforms reduces uncertainty in the PCFP drastically: already with two predictors, the average implicative entropy value drops below 0.1, and with five predictors uncertainty is virtually eliminated.

cardinality	average implicative entropy
1	0.28
2	0.06
3	0.03
4	0.02
5	0.01

Table 13: Average *n*-ary implicative entropy

The idea of predictions from more than one form is what stands behind the traditional notion of principal parts and their contemporary and more principled recovery by Stump and Finkel (2013): in an entropy-based perspective, principal parts are sets of inflected wordforms knowing which the entropy of guessing the content of all the remaining cells of the paradigm – what Bonami and Beniamine (2016) call 'residual uncertainty' – is exactly 0. As can be seen from Table 14 below, in Latin verb inflection there are no principal part sets composed of two or three paradigm cells. The smallest combinations of cells that work as principal parts are composed of four cells: there are 56 combinations of four cells that allow to eliminate residual uncertainty. If five predictors are used, there are more principal part sets, both in absolute terms and in percentage on the number of possible combinations of cells.

cardinality	principal parts	
	n.	%
2	0	0
3	0	0
4	56	4.1%
5	336	11.2%

Table 14: Principal part sets at different cardinalities

This confirms on a more empirically-based ground the descriptions of Latin grammars and dictionaries, where the four principal parts are PRS.ACT.IND.1SG, PRS.ACT.IND.2SG, PRF.ACT.IND.1SG and, lastly, PRF.PASS.PTCP.NOM.M.SG or SUP.ACC, depending on the choices made by different authors.[13] Our results are also in line with the findings obtained by Finkel and Stump (2009) with a different, set-theoretic rather than information-theoretic, methodology: also in their study, four principal parts prove to be sufficient in order to be able to guess the rest of the paradigm with no uncertainty. An advantage of the information-theoretic methodology is that it makes it possible to take into consideration not only categorical principal parts, but also what Bonami and Beniamine (2016) call 'near principal parts', i.e., sets of cells that allow to infer the rest of the paradigm with very low – but not null – residual uncertainty. In Table 15, the threshold of residual uncertainty is set at 0.001 and 0.01, and the number and percentage of near principal parts at different cardinalities is reported.

cardinality	near principal parts			
	$H < 0.001$		$H < 0.01$	
	n.	%	n.	%
2	0	0	15	14.3%
3	15	3.3%	196	43.1%
4	122	8.9%	834	61.1%
5	471	15.7%	2,190	72.9%

Table 15: Near principal part sets at different cardinalities

It can be observed that already with the very low threshold of 0.001, there are sets of near principal parts composed of three cells. If the threshold is set at 0.01, there are even combinations of two cells that work as near principal parts; furthermore, almost half of the available combinations of three cells, more than half of the combinations of four cells, and the relevant majority of combinations of five cells allow to infer the rest of the paradigm with a residual uncertainty of less than 0.01. This means that knowledge of a limited number of cells yields a very relevant reduction of uncertainty in the PCFP, giving further confirmation to Ackerman and Malouf (2013)'s 'low entropy conjecture', according to which the surface complexity of the inflectional patterns of languages with a rich morphology – like Latin – does not make unpredictability in such systems so great as to make them hard to learn and master for speakers.

[13] Lewis and Short (1879) use only three principal parts, but only because the conjugation is stated explicitly.

4. Inclusion of LatInfLexi into the LiLa Knowledge Base

The topic of this section is a discussion of the perspectives opened by the planned inclusion of the data of LatInfLexi into the LiLa knowledge base (Passarotti et al., 2019). The goal of the LiLa (Linking Latin) project[14] is to connect and make interoperable the wealth of digital resources – like corpora and lexicons – and NLP tools – like lemmatizers, morphological analyzers and dependency parsers – that are already available for Latin. To this aim, LiLa makes use of a set of Semantic Web and Linguistic Linked Open Data standards, among which here at least the ontology used for lexical resources (Lemon, Buitelaar et al. (2011), Ontolex[15]) should be mentioned, that is based on the 'Lexical Entry' to which all the relevant forms can be associated. The architecture of LiLa thus has the 'lemma' as its core. A lemma is defined as an inflected 'form' that is conventionally chosen as the citation form of a lexical entry. Lemmas are then directly linked to 'tokens' – i.e., actual occurrences in textual resources. Both forms and tokens can be analyzed by NLP tools.

Within this architecture, it would be useful to make the coverage of LatInfLexi more systematic – adding also the nouns with less than 30 occurrences in Delatte et al. (1981) and including adjectives – and incorporate the wordforms reported in LatInfLexi in the knowledge base. Both LatInfLexi and the Lila knowledge base would benefit greatly from such interaction, due to their different design. The LiLa knowledge base takes a concrete perspective, including only wordforms that are either attested in corpora, or reported in lexical resources that are in turn based on actual usage in texts, like for instance Tombeur (1998). Conversely, we have seen in 2.1. that in LatInfLexi a much more abstract perspective drives the selection of different inflected wordforms: for each lexeme, the content of all non-defective paradigm cells is given, regardless of the actual attestation of the generated wordforms in actual texts.

Therefore, the inclusion of the data of LatInfLexi into the LiLa knowledge base would greatly enrich the latter: lemmas would be linked to all their possible inflected wordforms, rather than only to attested ones. The relevance of such enrichment would be more relevant than one could think, since recent quantitative work on the attestation of inflected wordforms in large paradigms (Chan, 2008; Bonami and Beniamine, 2016; Blevins et al., 2017) shows that, even using very large corpora, 'saturation' – i.e., the situation in which all the inflected wordforms of a lexeme occur in a given corpus (Chan, 2008) – is reached only for a handful of very frequent lexemes, while in all other cases only some cells are actually filled by a wordform, and for many lexemes only a couple of wordforms are attested, or even only one. On the other hand, LatInfLexi too would benefit from being included into LiLa, because the linking of the possible wordforms of the former to the real occurrences in the lemmatized (and sometimes, e.g. in treebanks, even equipped with fine-grained morphosyntactic analyses) texts of the latter would allow for a more accurate assessment of the frequency of wordforms,[16] and thus for a more careful discrimination between forms that are possible but are not attested and those that actually occur in texts. This could also be useful in order to have a more satisfactory, corpus-based treatment of overabundance, where the marginality of a 'cell-mate' (Thornton, 2019) with respect to the other one(s) is not decided according to lexicographical sources, but rather on the basis of the actual usage of the competing wordforms in texts.

[14] https://lila-erc.eu/.

[15] https://www.w3.org/community/ontolex/.

5. Conclusions

This paper has presented LatInfLexi, a large, freely available, paradigm-based inflected lexicon of Latin verbs and nouns, detailing how the wordforms have been generated starting from the information provided in the morphological analyzer Lemlat 3.0.

It has then illustrated the usefulness of such a lexicon, firstly to perform a quantitative analysis of predictability in inflectional morphology by means of the information-theoretic notion of implicative entropy. From this analysis, by means of unary implicative entropy a mapping of the verbal paradigm in 15 zones of complete interpredictability has been proposed: this picture is less straightforward than the traditional one, based on the three different stems appearing in the paradigm, and therefore more similar to the situation found in Romance verb paradigms, suggesting that there is more continuity from Latin to Romance than is traditionally assumed, at least if patterns of interpredictability are considered. Secondly, *n*-ary implicative entropy has been used to recover the traditional notion of principal parts on more solid grounds, confirming the analysis of grammars and dictionaries in this respect, as well as results recently obtained for Latin verb inflection with Finkel and Stump (2009)'s Principal Part Analysis, but also highlighting the usefulness of extending the investigation to non-categorical 'near principal parts', that allow for a relevant – although not complete – reduction of residual uncertainty regarding other paradigm cells.

Lastly, another possible use of the resource that has been discussed in this paper is its inclusion in the LiLa knowledge base, that in this way would be enhanced with possible inflected wordforms that can be linked to lemmas, besides the ones attested in textual resources, while LatInfLexi would benefit from this interaction in that it would have access to more detailed frequency data.

6. Availability of Data and Tools

The data and tools used in this study are freely available online, allowing for an easy replication of the presented results. LatInfLexi can be found at https://github.com/matteo-pellegrini/LatInfLexi. The Qumin toolkit that was used to automatically perform entropy computations can be freely downloaded at https://github.com/XachaB/Qumin.

[16] As we have seen in 2.1., LatInfLexi provides information on frequency, but with the same shortcomings of the source from which it takes it, Tombeur (1998), where there is no disambiguation of wordforms with multiple possible analyses. For a more detailed discussion of the issues related to frequency data in LatInfLexi, the reader is referred to Pellegrini and Passarotti (2018).

7. Bibliographical References

Ackerman, F. and Malouf, R. (2013). Morphological organization: The low conditional entropy conjecture. *Language*, 89(3):429–464.

Ackerman, F., Blevins, J. P., and Malouf, R. (2009). Parts and wholes: Patterns of relatedness in complex morphological systems and why they matter. In James P Blevins et al., editors, *Analogy in grammar: Form and acquisition*, pages 54–82. Oxford University Press, Oxford.

Aronoff, M. (1994). *Morphology by itself: Stems and inflectional classes*. MIT press, Cambridge.

Baayen, R. H., Piepenbrock, R., and Gulikers, L. (1996). The CELEX lexical database (cd-rom). University of Pennsylvania.

Beniamine, S. (2018). *Classifications flexionnelles. Étude quantitative des structures de paradigmes*. Ph.D. thesis, Université Sorbonne Paris Cité-Université Paris Diderot.

Blevins, J. P., Milin, P., and Ramscar, M. (2017). The zipfian paradigm cell filling problem. In Ferenc Kiefer, et al., editors, *Perspectives on Morphological Organization*, pages 139–158. Brill, Leiden-Boston.

Blevins, J. P. (2006). Word-based morphology. *Journal of Linguistics*, 42(3):531–573.

Blevins, J. P. (2016). *Word and paradigm morphology*. Oxford University Press, Oxford.

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

Bonami, O. and Boyé, G. (2003). Supplétion et classes flexionnelles. *Langages*, 37(152):102–126.

Bonami, O. and Boyé, G. (2014). De formes en thèmes. In Florence Villoing, et al., editors, *Foisonnements morphologiques. Études en hommage à Françoise Kerleroux*, pages 17–45. Presses Universitaires de Paris-Ouest, Paris.

Bonami, O., Caron, G., and Plancq, C. (2014). Construction d'un lexique flexionnel phonétisé libre du français. In *Congrès Mondial de Linguistique Française -- CMLF 2014*, volume 8, pages 2583–2596. EDP Sciences.

Bonami, O. (2014). La structure fine des paradigmes de flexion. études de morphologie descriptive, théorique et formelle. Mémoire d'habilitation à diriger des recherches. Université Paris Diderot (Paris 7).

Bouma, G. and Adesam, Y. (2017). *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*. Linköping University Electronic Press, Gothenburg.

Boyé, G. and Cabredo Hofherr, P. (2006). The structure of allomorphy in spanish verbal inflection. *Cuadernos de Lingüística del Instituto Universitario Ortega y Gasset*, 13:9–24.

Boyé, G. and Schalchli, G. (2016). The status of paradigms. In Andrew Hippisley et al., editors, *The Cambridge handbook of morphology*, pages 206–234. Cambridge University Press, Cambridge.

Buitelaar, P., Cimiano, P., McCrae, J., Montiel-Ponsoda, E., and Declerck, T. (2011). Ontology lexicalization: The lemon perspective. In *Proceedings of the Workshops-9th International Conference on Terminology and Artificial Intelligence (TIA 2011)*.

Calderone, B., Pascoli, M., Sajous, F., and Hathout, N. (2017). Hybrid method for stress prediction applied to GLÀFF-IT, a large-scale Italian lexicon. In *International Conference on Language, Data and Knowledge*, pages 26–41, Cham. Springer.

Chan, E. (2008). *Structures and distributions in morphology learning*. Ph.D. thesis, University of Pennsylvania.

Corbett, G. G. (2005). The canonical approach in typology. In Zygmunt Frajzyngier, et al., editors, *Linguistic diversity and language theories*, pages 25–49. John Benjamins, Amsterdam.

Delatte, L., Evrard, É., Govaerts, S., and Denooz, J. (1981). *Dictionnaire fréquentiel et index inverse de la langue latine*. LASLA, Liège.

Dressler, W. U. (2002). Latin inflection classes. In A Machtelt Bolkestein, et al., editors, *Theory and description in Latin linguistics*, pages 91–110. Brill, Leiden-Boston.

Finkel, R. and Stump, G. (2009). What your teacher told you is true: Latin verbs have four principal parts. *Digital Humanities Quarterly*, 3(1).

Grestenberger, L. (2019). Deponency in morphology. In *Oxford Research Encyclopedia of Linguistics*. Oxford University Press.

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

Lewis, C. and Short, C. (1879). *A Latin Dictionary*. Clarendon, Oxford.

Litta, E., Passarotti, M., and Culy, C. (2016). Formatio formosa est. Building a word formation lexicon for Latin. In *Proceedings of the Third Italian Conference on Computational Linguistics (CLiC-it 2016)*, pages 185–189.

Maiden, M. (2009). From pure phonology to pure morphology: the reshaping of the romance verb. *Recherches linguistiques de Vincennes*, 38:45–82.

Maiden, M. (2018). *The Romance verb: Morphomic structure and diachrony*. Oxford University Press, Oxford.

Montermini, F. and Bonami, O. (2013). Stem spaces and predictability in verbal inflection. *Lingue e linguaggio*, 12(2):171–190.

Nivre, J., De Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., et al. (2016). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC '16)*, pages 1659–1666.

Passarotti, M., Budassi, M., Litta, E., and Ruffolo, P. (2017). The Lemlat 3.0 package for morphological analysis of Latin. In Gerlof Bouma et al., editors, *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*, pages 24–31. Linköping University Electronic Press, Gothenburg.

Passarotti, M. C., Cecchini, F. M., Franzini, G., Litta, E., Mambrini, F., and Ruffolo, P. (2019). The LiLa knowledge base of linguistic resources and NLP tools for latin.

In *2nd Conference on Language, Data and Knowledge (LDK 2019)*, pages 6–11. CEUR-WS. org.

Pellegrini, M. and Passarotti, M. (2018). LatInfLexi: an inflected lexicon of Latin verbs. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*.

Petrov, S., Das, D., and McDonald, R. (2011). A universal part-of-speech tagset. *ArXiv*, pages 2089–2096.

Piotrowski, M. (2012). Natural language processing for historical texts. *Synthesis lectures on human language technologies*, 5(2):1–157.

Pirrelli, V. and Battista, M. (2000). The paradigmatic dimension of stem allomorphy in italian verb inflection: 2628. *Italian Journal of Linguistics*, 12(2):307–380.

Sims, A. D. (2015). *Inflectional defectiveness*. Cambridge University Press, Cambridge.

Stump, G. and Finkel, R. A. (2013). *Morphological typology: From word to paradigm*. Cambridge University Press, Cambridge.

Stump, G. (2015). *Inflectional paradigms: Content and form at the syntax-morphology interface*. Cambridge University Press, Cambridge.

Sylak-Glassman, J., Kirov, C., Post, M., Que, R., and Yarowsky, D. (2015). A universal feature schema for rich morphological annotation and fine-grained cross-lingual part-of-speech tagging. In *International Workshop on Systems and Frameworks for Computational Morphology*, pages 72–93, Cham. Springer.

Thornton, A. M. (2019). Overabundance: a canonical typology. In Francesco Gardani, et al., editors, *Competition in inflection and word-formation*, pages 223–258. Springer, Berlin.

Tombeur, P. (1998). *Thesaurus formarum totius Latinitatis: a Plauto usque ad saeculum XXum; TF.[2]. CETEDOC Index of Latin forms: database for the study of the vocabulary of the entire Latin world; base de données pour l'étude du vocabulaire de toute la latinité*. Brepols, Turnhout.

Wurzel, W. U. (1984). *Flexionsmorphologie und Natürlichkeit: ein Beitrag zur morphologischen Theoriebildung*. Akademie-Verlag, Berlin.

Zanchetta, E. and Baroni, M. (2005). Morph-it!: A free corpus-based morphological resource for the Italian language. In *Proceedings of corpus linguistics,* `http://dev.sslmit.unibo.it/linguistics/morph-it.php`. Citeseer.