UNIVERSITÀ DEGLI STUDI DI BERGAMO

Scuola di Alta formazione Dottorale

Corso di Dottorato in Scienze Linguistiche

Ciclo XXXII

Settore scientifico disciplinare L-LIN/01

Predictability in Latin inflection

An entropy-based approach

Supervisore:

Chiar.mo Prof. Pierluigi Cuzzolin

Tesi di Dottorato

Matteo PELLEGRINI

Matricola n. 1043632

Anno Accademico 2018/19

Ringraziamenti

Tre anni di dottorato mi hanno mostrato con chiarezza quanto è importante confrontarsi con altre persone nel corso del lavoro di ricerca. Desidero quindi spendere qualche riga per ringraziare tutti coloro che hanno offerto il loro aiuto, a diverso titolo. Ovviamente, tutti i difetti di questa tesi restano responsabilità esclusivamente mia, ma se ci sono dei pregi è anche grazie al contributo di queste persone.

Innanzitutto, ringrazio il mio tutor Pierluigi Cuzzolin per avermi permesso di portare avanti questo progetto, per la sua costante supervisione, e per la sua lettura attenta dell'intero lavoro.

Fondamentali per l'avanzamento del progetto in ogni sua fase sono state le numerose discussioni con Marco Passarotti, che ringrazio in particolare per l'aiuto enorme che mi ha dato per ottenere i dati necessari, oltre che per la sua capacità di suggerire soluzioni pratiche ai problemi che mi sono trovato ad affrontare. Ringrazio anche gli altri membri del CIRCSE che mi hanno fornito consigli utili, e soprattutto mi hanno fatto sentire a casa durante le mie frequenti visite da "intruso" all'Università Cattolica di Milano: grazie a Flavio Cecchini, Greta Franzini, Francesco Mambrini, Paolo Ruffolo, Rachele Sprugnoli, e in particolar modo a Eleonora Litta.

Il dottorato mi ha poi dato l'opportunità di trascorrere due periodi di studio e ricerca all'estero. Il primo di questi è stato a Tolosa, sotto la guida di Fabio Montermini: oltre che per il suo aiuto nel definire in maniera più precisa obiettivi e cornice teorica della tesi, desidero ringraziarlo in special modo per la disponibilità con cui ha fatto sì che potessi sfruttare appieno il mio soggiorno malgrado gli ostacoli legati alla non facile situazione politica in cui si trovava l'Università di Tolosa "Jean Jaurès" nei tre mesi che ho passato lì.

Il secondo soggiorno è stato a Parigi, sotto la guida di Olivier Bonami, il cui contributo alla riuscita del lavoro è andato ben oltre le mie già alte aspettative. In particolare, gli sono debitore per il contenuto dell'ultimo capitolo, frutto del lavoro svolto insieme a lui durante il periodo trascorso al Laboratoire de Linguistique Formelle, presso l'Università Paris 7 - Diderot.

Ringrazio Davide Ricca, innanzitutto per avermi iniziato allo studio della morfologia fin dalla laurea triennale, poi per gli spunti di riflessione che anche in seguito non ha mai mancato di offrirmi nelle nostre discussioni su svariati ambiti della morfologia, infine per i consigli più puntuali che mi ha dato sul contenuto della tesi.

Un aiuto molto concreto mi è arrivato da Sacha Beniamine, che ringrazio per il supporto nell'utilizzo degli script da lui elaborati, di cui mi sono ampiamente servito in questo lavoro, ma soprattutto per aver apportato a tali script le modifiche necessarie per ottenere i risultati presentati nel quinto e nel sesto capitolo.

In diverse occasioni, ho avuto modo di esporre il mio progetto anche a Farrell Ackerman, Gilles Boyé, Bernard Fradin, Nicola Grandi e Vito Pirrelli: li ringrazio per gli utili consigli che mi hanno dato, nonostante la natura poco sistematica delle nostre discussioni.

Ringrazio poi gli altri dottorandi con cui ho condiviso tanti momenti in questi tre anni, in particolare quelli del XXXII ciclo: Marco Budassi, Cristina Lo Baido, Vittorio Napoli, Laura Restivo e Alessio Salomoni. Un ringraziamento speciale va poi alle due Silvie – Silvia Ballarè e Silvia Micheli: fin dal primo giorno di lezione, in virtù della loro maggiore esperienza mi hanno guidato passo dopo passo in tutti gli aspetti più concreti del mio percorso da dottorando.

Chiudo con una nota più personale. L'ultimo anno di dottorato ha coinciso con un periodo molto particolare della mia vita. Il ringraziamento più grande va quindi alle persone che sanno perché è così, e che in quel periodo mi sono state accanto, ognuna a modo suo: Ilaria, mia madre, mio padre, mio fratello, Luca, Matteo, Marco, Orsetta, Marianna, Edoardo, Davide, Lucia, Serena, Alessandra, Anna. A loro, grazie di cuore.

Contents

Introduction **1**

Chapter 1. The theoretical framework **4**

1.1 Some definitions 4

1.2 Classifications of theories of inflectional morphology 7

1.3 Implicative relations, words and sub-word units 14

1.3.1 Implicative relations and exponents 15

1.3.2 Implicative relations and stem allomorphy 16

1.3.3 Putting it all together: implicative relations between words 19

1.4 The quantitative dimension 24

1.5 Conclusion: a quantitative abstractive approach to the implicative structure of paradigms 26

Chapter 2. The method **27**

2.1 Basic information-theoretic notions 27

2.2 Predicting exponents 31

2.3 Predicting alternation patterns 35

2.4 Predicting from more than one wordform 44

2.5 Predicting forms knowing more than just forms 48

2.6 Conclusion: an entropy-based approach to the PCFP 49

Chapter 3. The data **51**

3.1 The structure of LatInfLexi 52

3.2 The selection of lexemes 53

3.3 The generation of wordforms 56

3.3.1 Verb paradigms 56

3.3.2 Noun paradigms 60

3.4 The treatment of overabundance 64
3.4.1 Overabundance due to the presence of more than one LES 65
3.4.2 Overabundance due to compatibility of a LES with more than one SF.. 69
3.5 Phonetic transcriptions 71
3.6 Conclusion 76
Chapter 4. Predictability and paradigm organization in Latin verb inflection 78
4.1 Latin verb inflection: the traditional account and previous theoretical research 79
4.2 The cell paradigm of Latin verbs 92
4.3 Predictability in Latin verb inflection: wordforms that are based on different stems 97
4.4 Predictability in Latin verb inflection: wordforms that are based on the same stem 103
4.5 Zones of interpredictability in Latin verb inflection 109
4.6 *n*-ary implicative entropy and principal parts 116
4.7 Conclusion 126
Chapter 5. Predictability in Latin noun inflection and the role of gender ... 128
5.1 Latin noun inflection: the traditional account 129
5.2 Predictability in Latin noun inflection 140
5.2.1 The cell paradigm of Latin nouns 140
5.2.2 Results: unary implicative entropy 141
5.2.3 *n*-ary implicative entropy and principal parts 146
5.3 Predictability and gender 151
5.3.1 Some examples 151
5.3.2 Gender and inflection classes in Latin 156
5.3.3 Results 158

5.3.4 Discussion 163
5.4 Conclusion 167
Chapter 6. The impact of derivational relatedness on inflectional predictions 169
6.1 The question 169
6.2 Verbs that derive from the same ancestor: derivational-inflectional families 176
6.2.1 Coding derivational-inflectional families 177
6.2.2 The inflectional behaviour of verbs in the same family 182
6.2.3 Results 186
6.3 Nouns that are formed by means of the same derivational suffix: derivational-inflectional series 190
6.3.1 Derivational-inflectional series: coding and inflectional behaviour 190
6.3.2 Results 193
6.4 Discussion 196
6.5 Conclusion 198
Conclusion 200
References 204

Introduction

In the last few years, the field of morphology has started to question some fundamental assumptions on the structure of wordforms. In particular, the idea is gaining ground that wordforms should not be viewed as obtained by concatenating smaller meaningful pieces one to another, as in classical morphemic analysis. Instead, the opposite is considered to be true: from the comparison of full inflected wordforms, recurrent partials are extracted which can be thought as having a discriminative function within the paradigm – i.e., what matters is that they are useful in order to distinguish wordforms from one another, rather than their association with a particular meaning (cf. Blevins 2016: 197).

A problem that has been widely investigated in this context is the possibility of predicting full inflected wordforms from one another within the inflectional paradigm of a lexeme, exploiting the presence of more or less reliable implicative relations (Wurzel 1984), in what has been labelled the "Paradigm Cell Filling Problem" (Ackerman et al. 2009). As a way of quantifying the difficulty of this task, the information-theoretic notion of conditional entropy has been used in much recent work (Ackerman et al. 2009, Bonami & Boyé 2014, Sims & Parker 2016, Beniamine 2018).

In this work, the above-mentioned theoretical and methodological innovations are applied to the Latin verbal and nominal paradigm, to obtain a quantitative analysis of the reliability of implicative relations, and thus of the patterns of interpredictability between inflected wordforms – i.e., of the difficulty of the Paradigm Cell Filling Problem.

Chapter 1 and Chapter 2 provide a more detailed picture of the theoretical framework within which this work is located and of the adopted, entropy-based, methodology, respectively. As we will see in more detail in Chapter 1, our theoretical framework can be considered as abstractive – i.e., considering morphemes as possibly extracted *a posteriori* from full inflected wordforms, rather than starting from morphemes and assembling them to obtain wordforms – and implicative – i.e., focusing on implicative relations, rather than on exponence of morphosyntactic properties. Our approach is also quantitative, as the entropy-based

assessment of predictability in inflectional paradigms is obtained by taking the type frequency of different inflectional patterns into account – as is shown in Chapter 2, where the details of the adopted methodology are outlined, namely the one proposed in Bonami & Boyé (2014) and Beniamine (2018).

To obtain information on the type frequency of inflectional patterns, an inflected lexicon listing the wordforms of a representative selection of lexemes is necessary. In Chapter 3, the lexical resource that was created for the purposes of this work – LatInfLexi – is presented, showing how it was obtained from the large database of a recently renewed morphological analyser of Latin, Lemlat 3.0 (Passarotti et al. 2017).

We can then move to the presentation of our results on verb paradigms – in Chapter 4 – and on noun paradigms – in Chapter 5. On the one hand, such results are exploited to obtain a mapping of the paradigm in zones of interpredictability – i.e., groups of cells that can be predicted from one another with no uncertainty. On the other hand, if not only predictions from one cell but also predictions from more than one cell are taken into account, principal parts – i.e., sets of cells from which the whole paradigm of a lexeme can be inferred without uncertainty – or at least near principal parts – which reduce uncertainty greatly, but not completely – can be found in a more principled way than in traditional descriptions.

In the last section of Chapter 5, a methodological innovation with respect to the standard procedure outlined in Bonami & Boyé (2014) and Beniamine (2018) is introduced. In §5.3, uncertainty in predicting one cell from another is quantified assuming that not only the phonotactic shape of the wordforms is known, but information of a different kind too – namely, the gender of a noun, that is partly predictive of its inflection behaviour, as is already acknowledged in traditional descriptions. The entropy-based methodology allows us to quantify the degree of the reduction in uncertainty obtained by including gender information.

In Chapter 6, another piece of information is assumed to be known beside phonotactics, namely the derivational relatedness of lexemes in our sample, in terms of both families – that for practical reasons we investigate in verb paradigms – and of series – studied in noun paradigms. The interpretation of the results of this last chapter raises interesting methodological and theoretical questions on how to count

on the one hand different lexemes that share the same lexical base (cf. the classification in families), on the other hand different lexemes that are built by means of the same derivational process (cf. the classification in series). Do these derivationally related lexemes constitute different types when quantifying the type frequency of different patterns, as usual in entropy-based analyses, or should they rather be grouped under the same type?

In conclusion, we summarize the contribution provided by this work to the set of language resources available for Latin, to the description of Latin inflectional morphology and to the theoretical and methodological framework of abstractive, implicative approaches. Finally, we briefly sketch some ideas for future work on the comparison of predictability and paradigm organization in Latin and in the Romance languages.

Chapter 1. The theoretical framework

In this chapter, we will locate the approach that is adopted in this work in the larger field of theoretical morphology. To do so, we will discuss several points: each of them will be illustrated using examples taken from Latin inflectional morphology.
The point of departure will be an overview of the terminology that is used throughout this work (§1.1). We will then discuss several classifications of morphological theories (§1.2), from the by now traditional distinction between Item-and-Arrangement, Item-and-Process and Word-and-Paradigm models operated by Hockett (1954) up to the recent characterization of constructive and abstractive approaches proposed in Blevins (2006, 2016), highlighting the specific aspects on which each of the discussed classifications is based. In §1.3, we will focus on implicative relations, contrasting different ways in which they can be formulated, in terms of generalizations on exponents (§1.3.1), on stems (§1.3.2) or on inflected wordforms (§1.3.3). We will then add a quantitative dimension to the picture (§1.4), showing the importance of considering also non-categorical implicative relations. Lastly, in §1.5 we will explain the choices that have been made in this work regarding each of the topics discussed in the previous sections.

1.1 Some definitions

It is useful to start by giving a precise definition of some basic terms: although most of them are well known at least since Matthews (1974), they sometimes appear with slightly different meanings in different studies, so this will be useful to clarify the way in which they are used in this work.
Let us consider the data in Table 1.[1]

[1] The examples used in this chapter constitute only small fragments of the much more complex inflectional system of Latin: therefore, in principle the generalizations that are drawn are to be considered as valid only regarding the data of the examples, and not in Latin in general (although in many cases their reliability is indeed wider).

Table 1: The paradigm of two Latin nouns

	NOM. SG	GEN. SG	DAT. SG	ACC. SG	VOC. SG	ABL. SG	NOM. PL	GEN. PL	DAT. PL	ACC. PL	VOC. PL	ABL. PL
LUPUS 'wolf'	*lupus*	*lupī*	*lupō*	*lupum*	*lupe*	*lupō*	*lupī*	*lupōrum*	*lupīs*	*lupōs*	*lupī*	*lupīs*
MURUS 'wall'	*mūrus*	*mūrī*	*mūrō*	*mūrum*	*mūre*	*mūrō*	*mūrī*	*mūrōrum*	*mūrīs*	*mūrōs*	*mūrī*	*mūrīs*

In the terminology proposed by Matthews (1991: 24 ff.), each case of the table is said to contain a different **word** (in italics in our table). For instance, the first case contains the word *lupus*. A word is defined by Matthews (1991: 31) as a concrete linguistic sign, whose *signifiant* is called a **wordform**: for instance, the word *lupus* consists of the sequence of phonemes /lupus/.[2] On the other hand, the *signifié* of a word comprises both lexical and morphosyntactic information, i.e. information on the **lexeme** and **morphosyntactic property set** which the word expresses.

A **morphosyntactic property** is the pairing of a feature with the specific value that it takes. In Latin, there are two morphosyntactic features for which a noun is inflected: there is the feature CASE, whose different possible values are NOMINATIVE, GENITIVE, ACCUSATIVE, DATIVE, VOCATIVE and ABLATIVE,[3] and the feature NUMBER, whose different values are SINGULAR and PLURAL. If we combine the different values of these two morphosyntactic features, we obtain 12 different **morphosyntactic property sets** (henceforth sometimes referred to also by means of the abbreviation MPS), corresponding to the headers of the columns of the table. Morphosyntactic property sets thus refer to the portion of meaning of a wordform that is relevant for syntax, rather than to its referential semantics. To express morphosyntactic property sets, throughout this work we will use the abbreviations of the Leipzig Glossing Rules, with values of different features separated by a dot: therefore, NOM.SG is to be taken as a shorthand for the feature:value pairings CASE:NOMINATIVE and NUMBER:SINGULAR.

In our table, the headers of the lines are **lexemes**, i.e. abstract lexical units comprising all the inflected wordforms that share a core lexical meaning, differing

[2] The IPA transcriptions provided in this work are the ones that are given in LatInfLexi, on which see Chapter 3 below.

[3] For some lexemes, and only in the singular, there is also another value of this feature, namely the LOCATIVE, that we omit here for the sake of simplicity (see also the discussion in §3.3.2 and §5.1 below).

only in the morphosyntactic property sets that are expressed. Lexemes, like LUPUS and MURUS, will be notated in small capitals (as is usual in morphology), using the inflected wordform realizing the morphosyntactic property set NOM.SG (as is usual in Latin linguistics) as a **citation form**.[4]

Throughout this work, we will stick to Matthews' terminology, the only difference being that we will avoid the term "word", that could generate some unclarity since it is normally used in the literature for what we call "lexemes". We will rather follow the standard usage of the specifically morphological literature, extending the coverage of the term "(inflected) wordform" also to cases where we need to refer to the elements that appear in paradigms as two-sided linguistic signs, considered not only in their formal aspects, but also in their semantic ones.

With these definitions in mind, we can now follow Stump & Finkel (2013: 9) in defining the **paradigm** of a lexeme as a complete set of **cells**, where each cell is the pairing of that lexeme with the morphosyntactic property set for which it is inflected. Following Stump (2006: 284) and Boyé & Schalchli (2016: 207), a distinction can be made between the abstract **content paradigm**, given by the different morphosyntactic property sets for which lexemes of a given category can be inflected (the headers of the columns in our table can be taken as a notation of the content paradigm of Latin nouns), and the concrete **form paradigm**, consisting of the wordforms that are used to realize the content paradigm for a specific lexeme (the different lines in our table, that give us the form paradigm of the Latin nouns LUPUS and MURUS).

Lastly, it is useful to have a working definition of "stem" and "exponent", although the notions to which these terms refer will be problematized in §1.3. For a given wordform, the **stem** is the portion of form that contributes lexical information, while the **exponent** is the portion of form that contributes information on the morphosyntactic property set that is expressed (cf. Beniamine 2018: 37). For instance, in the wordform *lupe*, the sequence *lup-* can be identified as the stem, since it appears in all the wordforms of the lexeme LUPUS and it does not appear in any wordform of the lexeme MURUS. On the other hand, the ending *-e* will be the

[4] It should be noticed that throughout this work information on vowel length will be systematically omitted in the notation of lexemes – but not of wordforms.

exponent of the morphosyntactic property set VOC.SG, since it only appears in that cell across lexemes – in this example, both in LUPUS and in MURUS. Since LUPUS and MURUS use the same exponents for each morphosyntactic property set, they can be said to belong to the same **inflection class**, whereas other lexemes would use different exponents for the same morphosyntactic property sets, thus being assigned to different inflection classes.

1.2 Classifications of theories of inflectional morphology

Many classifications of theories of inflectional morphology have been proposed, each of them based on specific aspects that are taken to be more relevant. Perhaps the most classical distinction was first stated by Hockett (1954), who distinguishes three different models, that he calls "Item-and-Arrangement" (henceforth IA; cf. e.g. Harris 1942), "Item-and-Process" (henceforth IP, defended by Hockett 1954 himself) and "Word-and-Paradigm" (henceforth WP, further elaborated by Robins 1959 and Matthews 1972).

This classification is not completely symmetric in the characteristics that are used to distinguish the different models. In both IA and IP, the basic unit is the morpheme: the two models only differ in the type of operations – arrangements or processes – that need to be applied to such units in order to obtain larger units – namely, fully inflected wordforms. On the other hand, in WP models inflected wordforms are considered as the basic unit of morphosyntactic content, and their associations with morphosyntactic properties is specified by their location in paradigms (cf. Robins 1959: 60).

Another well-known classification of morphological theories has been proposed by Stump (2001) and is based on two cross-cutting distinctions, both related to the different ways in which the associations between morphosyntactic properties and exponents are treated in different models. The first distinction, between what Stump calls "lexical" and "inferential" theories, is similar to the difference between IA an IP, in that it concerns the nature of the associations. In lexical theories, they are listed in the lexicon, in the same way as the association between "roots" and "their grammatical and semantic properties" (Stump 2001: 1). If we consider again the

vocative singular of LUPUS, according to lexical theories there would be a lexical entry for the affix *-e*, specifying its phonological shape /e/ and the morphosyntactic property set that it expresses (namely, VOC.SG), exactly like there is a lexical entry for the lexeme, specifying the phonology of the stem /lup/ and the meaning 'wolf', as well as other grammatical information – for instance the masculine gender. In inferential theories, on the other hand, the associations are expressed via morphological rules that relate a stem to a specific inflected wordform: in this example, there would be a process of suffixation of *-e* that is applied to the stem *lup-* to yield the inflected wordform *lupe*.

The second distinction, between "incremental" and "realizational" theories, concerns the direction of the association. In incremental theories, exponents enable wordforms to acquire morphosyntactic information: when *-e* is added, the wordform acquires the morphosyntactic property set VOC.SG. Conversely, in realizational theories, the fact that a wordform expresses a morphosyntactic property set licenses the introduction of the corresponding exponent: when the lexeme LUPUS appears in a context that require the properties VOC.SG, the exponent *-e* is used.

Table 2 summarizes the possible combinations of the different distinctions, providing examples for each type of theory (taken from Stump 2001: 2-3).

Table 2: Stump (2001)'s classification of morphological theories

	incremental theories (exponents → MPS)	realizational theories (MPS → exponents)
lexical theories (exponents as lexical items)	lexical-incremental (Lieber 1992)	lexical-realizational (Halle & Marantz 1993)
inferential theories (exponents as processes)	inferential-incremental (Steele 1995)	inferential-realizational (Stump 2001)

In such a framework, WP models are considered by Stump (2001) as realizational, in that they take full inflected wordforms as the basic units of content. However, this does not imply that wordforms are also considered to be the basic units of form: this point is explicitly rejected by Robins (1959: 52), according to which "WP must recognize the morpheme as the minimal grammatical (not semantic!) unit of a language".

This discrepancy was put to the fore by Blevins (2006; see also Blevins 2016 for a more detailed picture), who distinguished between **constructive** and **abstractive** approaches. Constructive approaches are morph-based (Blevins 2006: 533), since sub-word units like stems and exponents are taken to be the basic units of form, the pieces with which larger units – i.e., inflected wordforms – are, so to speak, assembled. On the contrary, according to abstractive approaches full inflected wordforms are the basic unit of form; smaller units like stems and exponents are taken to be secondary abstractions that are (possibly) extracted from wordforms.

In this sense, not only IA and IP models, but also most contemporary WP models can be considered as constructive and morph-based, since the wordform is the basic unit of meaning, but not of form (cf. Blevins 2016: 6 ff.): in realizational approaches like Paradigm Function Morphology (Stump 2001, Bonami & Stump 2016) and Network Morphology (Corbett & Fraser 1993, Brown & Hippisley 2012), inflected wordforms are generated starting from a set of stems to which a series of inflectional rules are applied.

Let us now consider a different aspect, by looking at the set of paradigm cells given in Table 3 for Latin nouns belonging to different declensions.

Table 3: Some paradigm cells of five Latin nouns

lexeme (declension)	GEN.SG	DAT.SG	GEN.PL	DAT.PL
ROSA 'rose' (1st declension)	*rosae*	*rosae*	*rosārum*	*rosīs*
LUPUS 'wolf' (2nd declension)	*lupī*	*lupō*	*lupōrum*	*lupīs*
DUX 'leader' (3rd declension)	*ducis*	*ducī*	*ducum*	*ducibus*
FRUCTUS 'fruit' (4th declension)	*frūctūs*	*frūctuī*	*frūctuum*	*frūctibus*
RES 'thing' (5th declension)	*reī*	*reī*	*rērum*	*rēbus*

We can follow Bonami (2014: 23) in distinguishing **exponence** and **implicative relations**. In the graphical structure of Table 3, exponence is the "vertical" relationship between a wordform and the morphosyntactic property set that it expresses (cf. also Stump & Finkel 2013: 263). For instance, a generalization about exponence in this dataset can be formulated as in (1), stating that if a noun ends in *-rum* it will be a genitive plural.

(1) An example of exponence relation

N*rum* → GEN.PL

Implicative relations, on the other hand, refer to the "horizontal" relationship between a wordform in a given paradigm cell and another wordform in a different paradigm cell. For instance, knowing that a noun ends *-ārum* in the genitive plural unambiguously identify it as belonging to the 1st declension, thus allowing to infer its nominative singular by replacing *-ārum* with *-a*. This generalization can be expressed as in (2).

(2) An example of implicative relation

N*ārum*, GEN.PL → N*a*, NOM.SG

There have been proposals to classify morphological theories according to the kind of relations that are given more prominence. For instance, according to Stump (2016: 257) the realization of paradigm cells can be given both an **exponence-based** and an **implicative** definition. Boyé & Schalchli (2016) accordingly identify **paradigmatic** frameworks as the ones that aim at an implicative definition, starting from known wordforms to predict the content of other paradigm cells, while in **syntagmatic** frameworks wordforms are assembled starting from stems and exponents that realize morphosyntactic properties, in an exponence-based fashion. Inferential-realizational theories (in the sense of Stump 2001) are therefore syntagmatic in this sense, while an example of a model that can be defined as paradigmatic is Natural Morphology: it was in this framework that Wurzel (1984: Chapter 5) first introduced the notion of "implicative structure of inflectional paradigms".

If we now look at the relationship between the abstractive (word-based) vs. constructive (morph-based) distinction and the paradigmatic (implicative) vs. syntagmatic (exponence-based) distinction, it can be observed that a constructive perspective is usually adopted to investigate exponence relations. Thus, syntagmatic theories in Boyé & Schalchli (2016)'s terms are usually constructive

in Blevins (2006)'s terms. The basic question that is addressed in such models can be formalized as follows.

(3) The basic question in a constructive approach to exponence

> Given a lexeme and the relevant morphosyntactic property sets, what are the wordforms that realize the paradigm cells expressing each morphosyntactic property set for the given lexeme?

In many languages, the answer to this question is based on the inflection class to which the lexeme belongs and on realization rules that are sensible to such inflection class distinction: for instance, since the lexeme LUPUS belongs to the 2nd declension, its genitive singular can be obtained by suffixing *-ī* to the stem *lup-*, while for a 3rd declension noun like DUX the suffix to be attached to the stem *duc-* will be *-is*.

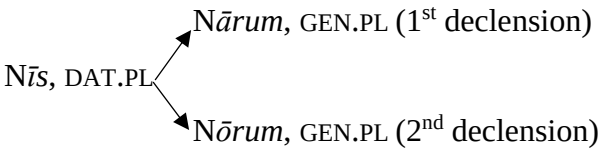
On the other hand, abstractive models were originally conceived to account for implicative relations, and are therefore identified as paradigmatic by Boyé & Schalchli (2016). The question that has been widely investigated in recent years using an abstractive approach is what Ackerman et al. (2009) call the Paradigm Cell Filling Problem, stating it in the form of the question: "What licenses reliable inferences about the inflected (and derived) surface forms of a lexical item?" (Ackerman et al. 2009: 54). Here, however, it seems useful to give a slightly different formulation, to make the difference with (3) more explicit. We can thus define the Paradigm Cell Filling Problem more precisely as in (4).

(4) The Paradigm Cell Filling Problem

> Given the content of one (or more) paradigm cell for a given lexeme (i.e., the pairing of a wordform with a morphosyntactic property set), what are the wordforms that realize other paradigm cells for the given lexeme?

If we go back to the data in Table 3, it can be observed that in some cases there are reliable generalizations concerning the implicative relations between wordforms, like the one in (2) between the genitive and nominative singular of 1st declension nouns. In other cases, however, wordforms are less informative on the content of other paradigm cells: for instance, knowing that the dative plural of a noun ends in *-īs*, it is not possible to predict with certainty what will be the content of other cells – for instance, the genitive plural, cf. the example in (5): the ending *-īs* is common to 1st and 2nd declension nouns, generating uncertainty on the endings that appear in other cells. This shows that there is variation in the informativity of wordforms in different paradigm cells with respect to the rest of the paradigm, and this makes the implicative structure of paradigms an interesting empirical domain to investigate in morphological theories.

(5) Uncertainty in the Paradigm Cell Filling Problem



The link between constructive theories and exponence relations, on the one hand, and abstractive theories and implicative relations, on the other hand, is so strong that the two are sometimes explicitly equated: see e.g. Stump (2016: 257 f.), according to whom constructive approaches would be the ones that aim at an exponence-based definition of the realization of paradigm cells, while abstractive approaches would be the ones that aim at an implicative definition.

However, as Bonami (2014: 24 f.) observes, this link appears to be only the product of an historical accident. On the one hand, an abstractive approach to exponence is conceivable:[5] instead of the aforementioned Paradigm Cell Filling Problem, the focus would be on what can be called the Paradigm Cell Recognition Problem (cf. Beniamine 2018: 308), which can be formulated as in (6).

[5] See Beniamine & Bonami (2018) for a first concrete proposal in this sense.

(6) The Paradigm Cell Recognition Problem

Given a wordform of a lexeme, what is the morphosyntactic property set that is expressed?

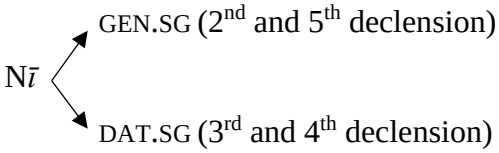
From this formulation, it is clear that the question concerns exponence relations, but it is formulated in an abstractive way by taking the wordform – and not the morphosyntactic property set, as in (4) – as the starting point. The difference between a constructive and an abstractive approach to exponence is summarized in Table 4.

Table 4: Constructive and abstractive approaches to exponence

approach	nature of units	
	given:	unknown:
constructive	lexeme, morphosyntactic property set	wordform
abstractive	lexeme, wordform	morphosyntactic property set

Considering again the data in Table 3, also generalizations about exponence are reliable in some cases: see the example in (1) stating the reliable relation between the ending *-rum* and the genitive plural. However, similarly to what we saw for implicative relations in (5), also for exponence relations there are cases where the shape of a wordform is less informative about the morphosyntactic property set that is expressed. For instance, by relying only on morphology, if we are faced with a wordform that ends in *-ī*, we cannot be sure whether it is a genitive singular or a dative singular, since that ending is used for the former morphosyntactic property set in 2nd and 5th declension nouns, but for the latter in 3rd and 4th declension nouns.

(7) Uncertainty in the Paradigm Cell Recognition Problem



As for the relationship between constructive approaches and implicative relations, first of all it should be noticed that implicative rules are part of the theoretical toolkit of such approaches in the form of rules of referral, which state that some paradigm cells systematically have the same exponents as other cells (cf. Zwicky 1985: 372, Stump 2001: 36 f.). More generally, Stump (2016: 260) argues that implicative relations can be derived from a complete exponence-based definition of the realization of paradigms, simply based on principles of logical inference.

There is also another, more concrete sense in which implicative relations can be treated in a constructive fashion. In this section, terms referring to the internal morphemic structure of wordforms were systematically avoided when expressing generalizations on implicative structure, but in most of the literature on this topic such generalizations are often stated in an implicitly constructive fashion, by referring to stems and affixes. For instance, principles that aim at limiting the possible formal complexity of inflectional paradigms, like the Paradigm Economy Principle (Carstairs 1987) and the No-Blur Principle (Carstairs-McCarthy 1994), are concerned with implicative structure, but are nevertheless explicitly stated to be valid only in terms of affixes, disregarding stem allomorphy (cf. Carstairs-McCarthy 1994: 739 f.).

In the next section, I will elaborate on this point, by contrasting generalizations formulated constructively in terms of sub-word units like stems and affixes with implicative relations based on full wordforms, in a purely abstractive approach.

1.3 Implicative relations, words and sub-word units

To exemplify the different ways in which implicative relations can be formulated, let us consider a different fragment of Latin paradigms, taken from verb inflection, with one verb for each of the traditional four conjugations.

Table 5: Some paradigm cells of four Latin verbs

	AMO 'to love' (1st conj.)	MONEO 'to warn' (2nd conj.)	RUMPO 'to break' (3rd conj.)	AUDIO 'to hear' (4th conj.)
PRS.ACT.IND.1SG	*amō*	*moneō*	*rumpō*	*audiō*
PRS.ACT.IND.2SG	*amās*	*monēs*	*rumpis*	*audīs*
PRS.ACT.IND.3SG	*amat*	*monet*	*rumpit*	*audit*
PRS.ACT.IND.1PL	*amāmus*	*monēmus*	*rumpimus*	*audīmus*
PRS.ACT.IND.2PL	*amātis*	*monētis*	*rumpitis*	*audītis*
PRS.ACT.IND.3PL	*amant*	*monent*	*rumpunt*	*audiunt*
PRF.ACT.IND.1SG	*amāvī*	*monuī*	*rūpī*	*audīvī*
PRF.ACT.IND.2SG	*amāvistī*	*monuistī*	*rūpistī*	*audīvistī*
PRF.ACT.IND.3SG	*amāvit*	*monuit*	*rūpit*	*audīvit*
PRF.ACT.IND.1PL	*amāvimus*	*monuimus*	*rūpimus*	*audīvimus*
PRF.ACT.IND.2PL	*amāvistis*	*monuistis*	*rūpistis*	*audīvistis*
PRF.ACT.IND.3PL	*amāvērunt*	*monuērunt*	*rūpērunt*	*audīvērunt*
SUP.ACC[6]	*amātum*	*monitum*	*ruptum*	*audītum*
PRF.PTCP.M.NOM.SG	*amātus*	*monitus*	*ruptus*	*audītus*
FUT.PTCP.M.NOM.SG	*amātūrus*	*monitūrus*	*ruptūrus*	*audītūrus*

1.3.1 Implicative relations and exponents

A first possible way to state generalizations on the implicative structure of paradigms is in terms of exponents – in this case, suffixes. For instances, in this example, knowing the exponent of the morphosyntactic property set PRS.ACT.IND.2SG of a lexeme, it is possible to infer the exponent of PRS.ACT.IND.3SG, thanks to the presence of the following set of implicative relations, related to the set of exponents used in different conjugations.

[6] We follow Aronoff (1994: 35) in making reference to the category of case, distinguishing the accusative supine, in this case *amātum* 'in order to love', from the ablative supine *amātū* 'to love' (in a context like 'easy to love'), rather than to the category of voice, as in the traditional distinction between active and passive supine. The adopted terminology is clearly more reasonable from a purely morphological point of view, since the endings *-um* and *-u* are indeed used as exponents of ACC.SG and ABL.SG in nominal inflection, and it is also capable to account for the semantic difference between the two forms.

(8) Implicative relations between exponents: PRS.ACT.IND.2SG and PRS.ACT.IND.3SG

-ās, PRS.ACT.IND.2SG → *-at*, PRS.ACT.IND.3SG (1st conjugation)

-ēs, PRS.ACT.IND.2SG → *-et*, PRS.ACT.IND.3SG (2nd conjugation)

-is, PRS.ACT.IND.2SG → *-it*, PRS.ACT.IND.3SG (3rd conjugation)

-īs, PRS.ACT.IND.2SG → *-it*, PRS.ACT.IND.3SG (4th conjugation)

Similar generalizations are obviously related to the traditional notion of inflection classes. In the framework of Canonical Typology (cf. e.g. Corbett 2005, Brown et al. 2012), the inflection classes that differ more consistently across paradigm cells are identified as more canonical (Corbett 2009: 4). Therefore, in the most canonical situation knowing the exponent used in one cell makes it possible to infer the exponents used in the rest of the paradigm, although of course more complex situations arise due to cases of exponents that are shared by different classes.

The one showed in (8) is perhaps the most usual way to express implicative relations, starting from the seminal study of Wurzel (1984), up to very recent works on this topic, even the ones that aim at an abstractive approach, e.g. Ackerman et al. (2009), where reference to exponents can be considered as a residue of a constructive approach.

1.3.2 Implicative relations and stem allomorphy

Implicative relations between exponents certainly account for an important part of the morphological complexity of a language, but they are not sufficient to cover all the facts related to the implicative structure of morphological paradigms. For instance, looking at the data in Table 5, if we only consider inflectional suffixes as identified in the traditional segmentation, there would be also a trivial but exceptionless implicative relation between every cell of the perfect active indicative and any other cell in the paradigm (in (9) the example of the implicative relation between PRS.ACT.IND.1SG and PRF.ACT.IND.1SG is given). Actually, there would not even be the need for an implicative relation in order to know the exponent of

PRF.ACT.IND.1SG, since the endings of perfective cells are invariable for lexemes of all conjugations.

(9) Implicative relations between exponents: PRS.ACT.IND.1SG and PRF.ACT.IND.1SG

-ō, PRS.ACT.IND.1SG → *-ī*, PRF.ACT.IND.1SG (1st and 3rd conjugation)

-eō, PRS.ACT.IND.1SG → *-ī*, PRF.ACT.IND.1SG (2nd conjugation)

-iō, PRS.ACT.IND.1SG → *-ī*, PRF.ACT.IND.1SG (4th conjugation)

However, such a generalization tells us very little about the morphology of perfective cells in Latin verb paradigms, since all the formal variation in such cells is located in the stem. If we now look not at the exponents, but rather at the stems used in the same cells considered in (9), it can be observed that there is a good deal of variation – summarized in (10) – in the relationship between the relevant stems, in some cases even involving non-concatenative alternations as in (10c).

(10) Patterns of formal alternation between different stems

(a) *am-*, PRS.ACT.IND.1SG → *amāv-*, PRF.ACT.IND.1SG

(b) *mone-*, PRS.ACT.IND.1SG → *monu-*, PRF.ACT.IND.1SG

(c) *rump-*, PRS.ACT.IND.1SG → *rūp-*, PRF.ACT.IND.1SG

(d) *aud-*, PRS.ACT.IND.1SG → *audīv-*, PRF.ACT.IND.1SG

A very similar situation is found in the nominal forms listed in the last three lines of Table 5, with the same endings in all conjugations, but unpredictable stem alternants.

Stem allomorphy has been extensively investigated in the last decade of the 20th century and in the first decade of the 21st century in a completely different family of works, among which we should cite at least the seminal studies of Maiden (1992) and Aronoff (1994), followed by a series of more detailed investigations especially on the inflectional morphology of verbs in the Romance languages, e.g. Pirrelli (2000), Pirrelli & Battista (2000), Montermini & Boyé (2012) and Montermini & Bonami (2013) on Italian, Bonami & Boyé (2003) on French, Boyé & Cabredo

Hofherr (2006) on Spanish; see also Maiden (2018) for a very recent and detailed survey on the inflectional morphology of the Romance verb. All these works share the core observation that stem alternants are sometimes manifold, but they are not randomly distributed throughout the paradigm: a given stem alternant will appear in the same set of cells for all lexemes.

For instance, in Latin we find two stem alternants whose distribution can be defined in morphosyntactic terms: there is one stem alternant, called the "present stem" in traditional descriptions, e.g. Leumann et al. (1977: 521), or more precisely the "*infectum* stem" (cf. e.g. Ernout & Thomas 1951: 183 f.; see also below, §4.1), since it always appears in cells with an imperfective meaning, and another one – the "perfect" or "*perfectum*" stem, cf. Leumann et al. (1977: 585), Ernout & Thomas (1951: 183) – that always appears in cells with a perfective meaning. Furthermore, there is another stem alternant, that Aronoff (1994) calls the "third stem", that also recurs in the same cells for all lexemes, although it arguably cannot be considered as expressing a morphosyntactically coherent meaning: Aronoff calls such stem alternants "morphomes" – as opposed to "morphemes" – since they have a form, but they lack the meaning that would be required in a full-fledged linguistic sign.

The fact that stem alternants recur always in the same set of cells – no matter if it is morphosyntactically or morphomically defined – has obvious consequences on the implicative structure of inflectional paradigms: knowing that a stem appears in a given cell, it is possible to infer the stem that will appear in another cell. For instance, in Latin the presence of a stem alternant in PRS.ACT.IND.1SG – no matter how irregular – implies that the same stem alternant will also appear in PRS.ACT.IND.2SG (as well as in all other cells with an imperfective meaning), and similar implicative relations can be formulated also for cells containing the perfect stem or the third stem, as is summarized in (11).

(11) Implicative relations between stems

rump-, PRS.ACT.IND.1SG → *rump-*, PRS.ACT.IND.2SG

rūp-, PRF.ACT.IND.1SG → *rūp-*, PRF.ACT.IND.2SG

rupt-, SUP.ACC → *rupt-*, PRF.PTCP.M.NOM.SG

1.3.3 Putting it all together: implicative relations between words

We have shown that implicative relations can be expressed as generalizations on the stems and exponents that appear in paradigm cells. However, deciding what is the stem and what is the exponent in a wordform is not a trivial task at all. In §1.1 we defined the stem as the portion of form that gives us information on the lexeme and the exponent as the portion of form that gives us information on the morphosyntactic property set. From this definition, a portion of form that appears in all cells of a lexeme can be certainly identified as a stem of that lexeme, and a portion of form that appears in a given cell for all lexemes can be certainly identified as an exponent of that morphosyntactic property set. However, there are also portions of form that are partially informative both on the lexeme that is used and on the morphosyntactic property set that is expressed. Intuitively, this is because of two facts that are typical of inflectional paradigms. On the one hand, it has been shown that stem alternants appear only in a specific set of cells, thus knowing the form of the stem will reduce the number of possible morphosyntactic property sets that we could be facing; in the data of Table 5, for instance, for the lexeme RUMPO, given the stem *rump-*, we already know that we are in an imperfective cell (although we still do not know which one precisely), since in other cells we would find different stems. On the other hand, exponents are sometimes different in lexemes belonging to different inflection classes, thus knowing the form of an exponent will reduce the number of possible lexemes that we could be facing: considering again the data of Table 5, if we know that the exponent of PRS.ACT.IND.2SG of a lexeme is *-ās*, we already know that we are facing a 1st conjugation verb (although we still do not know which one exactly), since in other conjugations we would find different endings. Theme vowels are a textbook example of phonetic material that simultaneously provides lexical and morphosyntactic information. Therefore, in such cases different choices can be made regarding segmentation. To exemplify some of the possibilities, for the sake of simplicity we will consider only the imperfective cells of Table 5, here given in Table 6 and Table 7.

In the analysis in §1.3.1-2, the segmentation that we assumed aimed at preserving the identity of the (present) stem, locating all the formal variation (of imperfective

cells) in the exponents, following a procedure that we can label as "Maximize Ending" (cf. Loporcaro 2012).

Table 6: A Maximize Ending approach to segmentation

	AMO 'to love' (1st conj.)	MONEO 'to warn' (2nd conj.)	RUMPO 'to break' (3rd conj.)	AUDIO 'to hear' (4th conj.)
PRS.ACT.IND.1SG	*am-ō*	*mon-eō*	*rump-ō*	*aud-iō*
PRS.ACT.IND.2SG	*am-ās*	*mon-ēs*	*rump-is*	*aud-īs*
PRS.ACT.IND.3SG	*am-at*	*mon-et*	*rump-it*	*aud-it*
PRS.ACT.IND.1PL	*am-āmus*	*mon-ēmus*	*rump-imus*	*aud-īmus*
PRS.ACT.IND.2PL	*am-ātis*	*mon-ētis*	*rump-itis*	*aud-ītis*
PRS.ACT.IND.3PL	*am-ant*	*mon-ent*	*rump-unt*	*aud-iunt*

An alternative possibility for this dataset would be to segment the forms in such a way that the exponents are the same for all verbs, at the cost of losing the identity of the stem across imperfective cells, in a "Maximize Stem" strategy (cf. Spencer 2012).

Table 7: A Maximize Stem approach to segmentation

	AMO 'to love' (1st conj.)	MONEO 'to warn' (2nd conj.)	RUMPO 'to break' (3rd conj.)	AUDIO 'to hear' (4th conj.)
PRS.ACT.IND.1SG	*am-ō*	*mone-ō*	*rump-ō*	*audi-ō*
PRS.ACT.IND.2SG	*amā-s*	*monē-s*	*rumpi-s*	*audī-s*
PRS.ACT.IND.3SG	*ama-t*	*mone-t*	*rumpi-t*	*audi-t*
PRS.ACT.IND.1PL	*amā-mus*	*monē-mus*	*rumpi-mus*	*audī-mus*
PRS.ACT.IND.2PL	*amā-tis*	*monē-tis*	*rumpi-tis*	*audī-tis*
PRS.ACT.IND.3PL	*ama-nt*	*mone-nt*	*rumpu-nt*	*audiu-nt*

Of course, if we state implicative relations in terms of stems and exponents the picture will change with different segmentations. Taking the cells PRS.ACT.IND.1SG and PRS.ACT.IND.2SG as an example, with a Maximize Ending segmentation there would a trivial bidirectional implicative relation among the stems that appear in those cells, since they are the same (cf. 12a). As far as exponents are concerned (cf. 12b), there would be a set of reliable implicative relations allowing to infer PRS.ACT.IND.1SG from PRS.ACT.IND.2SG, but not *vice versa*, since given the ending

-ō two possibilities arise, namely *-ās* in 1st conjugation verbs and *-is* in 3rd conjugation verbs.

(12) Implicative relations with a Maximize Ending segmentation

(a) implicative relations between stems

X-, PRS.ACT.IND.1SG ↔ X-, PRS.ACT.IND.2SG

(b) implicative relations between exponents

-ās, PRS.ACT.IND.2SG (1st conjugation)

-ō, PRS.ACT.IND.1SG

-is, PRS.ACT.IND.2SG (3rd conjugation)

-eō, PRS.ACT.IND.1SG → *-ēs*, PRS.ACT.IND.2SG (2nd conjugation)

-iō, PRS.ACT.IND.1SG → *-īs*, PRS.ACT.IND.2SG (4th conjugation)

-ās, PRS.ACT.IND.2SG → *-ō*, PRS.ACT.IND.1SG (1st conjugation)

-ēs, PRS.ACT.IND.2SG → *-eō*, PRS.ACT.IND.1SG (2nd conjugation)

-is, PRS.ACT.IND.2SG → *-ō*, PRS.ACT.IND.1SG (3rd conjugation)

-īs, PRS.ACT.IND.2SG → *-iō*, PRS.ACT.IND.1SG (4th conjugation)

Conversely, if we assume a Maximize Stem segmentation the reliable and bidirectional implicative relation would be between the exponents, since they are now the same for all conjugations (cf. 13b). However, the unpredictability of the second-person singular from the first-person singular would not disappear, but would just be moved to the stem, as is shown in (13a).

(13) Implicative relations with a Maximize Stem segmentation

(a) implicative relations between stems

X*ā*-, PRS.ACT.IND.2SG (1st conjugation)

X-, PRS.ACT.IND.1SG

X*i*-, PRS.ACT.IND.2SG (3rd conjugation)

X*e*-, PRS.ACT.IND.1SG → X*ē*-, PRS.ACT.IND.2SG (2nd conjugation)

X*i*-, PRS.ACT.IND.1SG → X*ī*-, PRS.ACT.IND.2SG (4th conjugation)

X*ā*-, PRS.ACT.IND.2SG → X-, PRS.ACT.IND.1SG (1st conjugation)

X*ē*-, PRS.ACT.IND.2SG → X*e*-, PRS.ACT.IND.1SG (2nd conjugation)

X*i*-, PRS.ACT.IND.2SG → X-, PRS.ACT.IND.1SG (3rd conjugation)

X*ī*-, PRS.ACT.IND.2SG → X*i*-, PRS.ACT.IND.1SG (4th conjugation)

(b) implicative relations between exponents

-*ō*, PRS.ACT.IND.1SG ↔ -*s*, PRS.ACT.IND.2SG

Although it is possible to find ways to evaluate the quality of different segmentations, e.g. by means of the information theoretic notion of description length, as is proposed in Walther & Sagot (2011), it should be stressed that for a given dataset there is no way of finding the "right" segmentation strategy algorithmically, as is argued in Beniamine (2018: 74 ff.). While for certain languages – most notably French, cf. Bonami & Boyé (2003) – almost all the formal variation can be considered to be located in the stem, all endings being invariable across lexemes, this does not hold for other languages, even Romance ones, as is shown for instance in Loporcaro (2012) on Logudorese and Ricca (2017) on Piedmontese verb inflection.

Therefore, the strategy that is envisaged in such works as Bonami & Boyé (2014) and Beniamine (2018) is to investigate implicative relations without assuming any fixed, global segmentation strategy valid for all the paradigm, but rather starting from the full inflected wordforms in the two different cells that are considered, in a purely abstractive approach, and looking at the actual pattern of alternation between the wordforms in such cells only: the result is a local segmentation which is only valid for the given pair of cells. If we look at the data of Table 6 and Table 7 from this perspective, indeed we end up with different segmentations for different pairs of cells. If we take PRS.ACT.IND.1PL and PRS.ACT.IND.2PL, the only difference between them is that we systematically find the final segment *-mus* in the former and the final segment *-tis* in the latter, the remaining portion of form being invariable. We can therefore formulate a very general implicative relation between such wordforms, that can be expressed as in (14a). If we instead consider PRS.ACT.IND.1PL and PRS.ACT.IND.3PL, we should formulate more specific but

equally reliable implicative relations that concern only verbs that belong the same conjugation, as is summarized in (14b).

(14) implicative relations between words

(a) X*mus*, PRS.ACT.IND.1PL ↔ X*tis*, PRS.ACT.IND.2PL

(b) X*āmus*, PRS.ACT.IND.1PL → X*ant*, PRS.ACT.IND.3PL (1st conjugation)
X*ētis*, PRS.ACT.IND.1PL → X*ent*, PRS.ACT.IND.3PL (2nd conjugation)
X*itis*, PRS.ACT.IND.1PL → X*unt*, PRS.ACT.IND.3PL (3rd conjugation)
X*ītis*, PRS.ACT.IND.1PL → X*iunt*, PRS.ACT.IND.3PL (4th conjugation)

The alternation pattern that is assumed locally in (14a) coincides with the result a Maximize Stem segmentation, while the one of (14b) would be the same in a Maximize Ending segmentation: the same wordform *rumpimus* should be segmented in a way when matched with *rumpitis* (*rumpi-mus* vs. *rumpi-tis*) and in another one when matched with *rumpunt* (*rump-imus* vs. *rump-unt*).

Using full inflected wordforms directly appears to be a better strategy than segmenting them in stems and exponents when trying to investigate the implicative structure of paradigms. Firstly, in unsegmented wordforms both information on stem allomorphy and on inflection class membership is encapsulated: wordforms are the most compact way of encoding information on both stems and exponents. Less trivially, it has been shown that segmentation sometimes obscures patterns that can be expressed more clearly and regularly in terms of full wordforms (cf. Blevins 2016: 51 ff.). Furthermore, the wordform is a more ecological unit of analysis than sub-word components: it is to wordforms that speakers are exposed, and not to stems and exponents in isolation. Indeed, there are works where morphemes are given no theoretical relevance, being explicitly treated as nothing more than epiphenomena resulting from the comparison of full inflected wordforms: cf. the already mentioned studies by Blevins (2006, 2016), but also the monograph by Bochner (1993) as a precursor. Lastly, it has been shown that it is possible to use fruitfully full inflected wordforms in a computational implementation of the solution to the PCFP: Malouf (2017) has presented a recurrent neural network that, given an abstract lexeme identifier and a morphosyntactic property set, is able to

generate the corresponding inflected wordform with good accuracy, with the parameters of the model being learned directly from a lexicon of unsegmented wordforms, the training consisting simply in mapping such wordforms to the morphosyntactic property set they express.

These advantages have already been exploited in the traditional descriptions of Ancient Greek and Latin,[7] with their use of principal parts, i.e. a set of inflected wordforms of a lexeme from which all the other wordforms of the same lexeme can be inferred. This tradition has been recently rediscovered in morphological theory, and a formal implementation can be found in Stump & Finkel (2013)'s Principal Part Analysis. However, even in this implementation the generalizations are still formulated in term of exponents, in a residually constructive manner. Only in works like Bonami & Boyé (2014) and Beniamine (2018) a purely abstractive perspective is adopted, since the analysis starts from unsegmented wordforms, as will be detailed in Chapter 2.

1.4 The quantitative dimension

In the examples provided in the previous sections, some of the implicative relations that have been formulated were intuitively said to be more reliable than others. In this section, it will be shown that implicative relations can indeed vary in two respects, namely their **coverage** and their **accuracy** (cf. Bonami & Beniamine 2016). Let us start from the example in (15).

(15) X, PRS.ACT.INF → X*m*, IPRF.ACT.SBJV.1SG

The one in (15) is a maximally general implicative relation. In the antecedent, the wordform is not required to have any particular phonological shape, thus the implication covers the whole Latin verbal lexicon, and also the change described in the consequent happens in all verbs: in Latin, imperfect subjunctive forms can always be obtained by adding agreement suffixes to the form of the present

[7] See Blevins (2013) for a detailed account of the relationship between grammatical descriptions in different traditions and recent morphological theories.

infinitive, even in highly irregular verbs (e.g. for the verb meaning 'to be' *esse*, PRS.ACT.INF → *essem*, IPRF.ACT.SBJV.1SG).

However, it does not seem advisable to limit the investigation to such categorical implicative relations. Consider, for instance, the data in Table 8 concerning the cells PRS.ACT.INF and PRF.ACT.INF.

Table 8: The present and perfect active infinitive of four Latin verbs

lexeme	PRS.ACT.INF	PRF.ACT.INF
AMO 'to love'	*amāre*	*amāvisse*
CUBO 'to lie down'	*cubāre*	*cubuisse*
DELEO 'to destroy'	*delēre*	*delēvisse*
MONEO 'to warn'	*monēre*	*monuisse*

No categorical implicative relation can be formulated on this data: given a present infinitive in *-āre*, the perfect infinitive can be in *āvisse* or in *-uisse*; given a present infinitive in *-ēre*, the perfect infinitive can be in *-ēvisse* or in *-uisse*. Similarly, in the opposite direction, the present infinitive can be in *-āre* or in *-ēre*, both from a perfect infinitive in *-visse* and from a perfect infinitive in *-uisse*.

However, if information on the number of verbs instantiating each type is added to the picture, interesting generalizations emerge, as can be seen in Table 9: for 1st conjugation verbs with a present infinitive in *-āre*, the perfect infinitive is overwhelmingly in *-āvisse*, while perfect infinitives in *-uisse* are more frequent for 2nd conjugation verbs with a present infinitive in - *ēre*.

Table 9: The formation of the perfect active infinitive in different conjugations: quantitative data[8]

PRS.ACT.INF in X*āre* (1st conj.)		PRS.ACT.INF in X*ēre* (2nd conj.)	
PRF.ACT.INF in X*āvisse*	1,214	PRF.ACT.INF in X*uisse*	144
PRF.ACT.INF in X*uisse*	34	PRF.ACT.INF in X*ēvisse*	11
other patterns[9]	18	other patterns	120

[8] These data are taken from LatInfLexi, on which see Chapter 3 below. Here, we exclude verbs that are defective in one or both of the involved cells.

[9] I.e., minor inflectional patterns, found in irregular verbs.

Therefore, we might still formulate implicative relations like in Table 10: the accuracy – which can be defined as the proportion of cases that satisfy the consequent among those that satisfy the antecedent, cf. Bonami & Beniamine (2016: 157) – of the relation X*āre* → X*āvisse* will not be perfect (i.e., 1), but it is still very high; the accuracy of the relation X*ēre* → X*uisse* is lower, but it still holds in more than half of the relevant cases.

Also the coverage of the two implicative relations – i.e., the proportion of cases that satisfy the antecedent among all the lexemes of the lexicon, see again Bonami & Beniamine (2016) – will be very different: 1st conjugation verbs – with present infinitive in *-āre* – are far more frequent than 2nd conjugation verbs – with present infinitive in *-ēre*.

Table 10: Accuracy and coverage of the implicative relation between PRS.ACT.INF and PRF.ACT.INF in Latin

implicative relation	accuracy	coverage
X*āre* → X*āvisse*	0.91 (1,214/1,266)	1,329/2,948 = 0.45
X*ēre* → X*uisse*	0.52 (144/275)	275/2,948 = 0.09

1.5 Conclusion: a quantitative abstractive approach to the implicative structure of paradigms

The present work is devoted to the study of the implicative structure of Latin paradigms, considering both verb and noun inflection. The theoretical framework of the work (cf. §1.2) can thus be defined as paradigmatic in the sense of Boyé & Schalchli (2016). It is also abstractive in the sense of Blevins (2016), since implicative relations will be expressed directly on full inflected wordforms, rather than on stems and affixes (cf. §1.3). Lastly, a quantitative, gradient perspective will be taken, by looking at the number of verbs for which the different patterns of formal alternation are attested, rather than limiting the investigation to categorical implicative relations (cf. §1.4). These are the theoretical choices that have been made: the next chapter will be devoted to a detailed description of the technical implementation of such ideas, based on the information-theoretic notion of conditional entropy.

Chapter 2. The method

This chapter is devoted to a detailed description of the methodology that will be applied in this work in order to perform an analysis of Latin inflectional paradigms that satisfy the theoretical desiderata outlined in Chapter 1. The applied method makes use of notions and procedures taken from information-theory (Shannon 1948), notably surprisal and entropy. In §2.1, a basic introduction to such information-theoretic notions is offered, pointing out the properties that make them useful to investigate topics related to implicative relations and the Paradigm Cell Filling Problem. The first proposal to use entropy for this purpose was outlined in Ackerman et al. (2009), whose procedure – aiming at an estimate of the degree of uncertainty associated with morphological realizations – is described in §2.2. However, the tools and algorithm that are used throughout this work are based on a similar but refined procedure (cf. Bonami 2014, Bonami & Boyé 2014, Beniamine 2018), that only requires a lexicon of inflected wordforms with no *a priori* morphological analysis. This method can be used not only to estimate the uncertainty in guessing the content of the paradigm cell of a lexeme knowing one inflected wordform, as explained in §2.3, but also given knowledge of multiple wordforms, as detailed in §2.4. The possible impact of additional information in making such tasks easier will be discussed in §2.5, where two possible variables are mentioned, namely i) the gender of a noun, and ii) the fact that a given lexeme is derivationally related to another one. To conclude, in §2.6 I will summarize the most important characteristics of the adopted methodology and clarify how they relate to the theoretical principles discussed in the first chapter.

2.1 Basic information-theoretic notions

The point of departure of this chapter cannot but be an explanation of some fundamental information-theoretic notions, namely surprisal and entropy.

Given an event *E* with a probability *P*, the **surprisal** or **information content** *I(E)* of the event measures the degree of surprise that is experienced when that event occurs, and therefore, in a sense, the amount of information that it expresses.

(1) $I(E) = -\log_2 P(E)$

In principle, different bases of the logarithm can be chosen. If the chosen base is 2, as happens here, then the unit of information used for surprisal is the bit. Surprisal is 0 bits if the event always occurs, approaching ∞ if the outcome almost never occurs.

On this ground, given a random variable *X* that can take one of a set of possible values $x_1, x_2, \ldots, x_n$, each with a given probability $P(x_1), P(x_2), \ldots, P(x_n)$, the entropy *H(X)* of the random variable is defined as the average surprisal of each possible outcome x_i, weighted according to their probability of occurring $P(x_i)$.

(2) $H(X) = -\sum_{x \in X} P(x) \log_2 P(x)$

Therefore, entropy provides an estimate of the degree of uncertainty on the actual value that is taken by a random variable, again measured in bits.

To understand the principles underlying these measures, and to underline their properties, let us consider some very easy examples. For instance, let us take a coin flip as our set of possible events. In this case, we have two equiprobable outcomes, namely heads and tails.

Table 1: Outcomes and probabilities of a fair coin flip

outcome	*P*
heads	0.5
tails	0.5

The entropy of a coin flip can thus be computed as in (3), putting together the surprisal of getting heads and of getting tails ($\log_2 0.5$, in both cases), weighted according to their probability of occurrence (0.5 in both cases):

(3) $H(X) = -\sum_{x \in X} P(x) \log_2 P(X)$

$= -((P(heads) \times \log_2 P(heads)) + (P(tails) \times \log_2 P(tails)))$

$= -(0.5 \times \log_2 0.5 + 0.5 \times \log_2 0.5)$

$= 1$ bit

The entropy of a random variable with two equiprobable outcomes is thus 1 bit. Two important properties of entropy that have to be kept in mind to understand their morphological applications are the following ones:

(i) all else being equal, the less the probability distribution is balanced, the lower the entropy value, since there is less uncertainty on the outcome that will occur;
(ii) all else being equal, the more the possible outcomes are numerous, the higher the entropy value, since there is more uncertainty on the outcome that will occur.

Let us start from property (i). In Table 1 we have considered a fair coin flip where the probability of getting heads and the probability of getting tails were equal, but if the coin were rigged to always come up heads, then there would be no uncertainty on the possible outcomes – trivially, because there would be only one possible outcome. The surprisal of getting heads would be 0 bits, and therefore also entropy would be 0. With different probabilities of getting heads or tails, we would have entropy values higher than 0 and lower than 1. The details of the computation are given in Tables 2a-b and in examples (4a-b).

Table 2: Outcomes and probabilities of some rigged coin flips

a. always heads (*X1*)

outcome	*P*
heads	1

b. 80% heads, 20% tails (*X2*)

outcome	*P*
heads	0.8
tails	0.2

(4a) $H(X1) = -(1 \times \log_2 1) = 0$ (always heads)

(4b) $H(X2) = -(0.8 \times \log_2 0.8 + 0.2 \times \log_2 0.2) \approx 0.72$ (80% heads, 20% tails)

Regarding property (ii), if the set of possible events is constituted by the different outcomes of throwing a dice, entropy is higher than in a coin flip, since there are six possibilities instead of two, and therefore the surprisal associated with each possible outcome is higher – cf. Table 3 and example (5).

Table 3: Outcomes and probabilities of throwing a dice

outcome	P
1	$\frac{1}{6}$
2	$\frac{1}{6}$
3	$\frac{1}{6}$
4	$\frac{1}{6}$
5	$\frac{1}{6}$
6	$\frac{1}{6}$

(5) $H(X) = -\left(\frac{1}{6} \times \log_2 \frac{1}{6} + \frac{1}{6} \times \log_2 \frac{1}{6} + \frac{1}{6} \times \log_2 \frac{1}{6} + \frac{1}{6} \times \log_2 \frac{1}{6} + \frac{1}{6} \times \log_2 \frac{1}{6} + \frac{1}{6} \times \log_2 \frac{1}{6}\right) \approx 2.58$

Information theory was presented by Shannon (1948) as a mathematical theory of communication. Since natural language is probably the most familiar means of communication, it does not come as a surprise that information-theoretic notions have been applied to language and linguistics in many different ways and for many different purposes, beginning from Shannon (1948) himself, and even more specifically Shannon (1951), where a way of measuring the entropy of written English is proposed. A similar, very general problem is tackled in subsequent work such as Cover & King (1978) and Brown et al. (1998). It is also worth mentioning that approaches based on the principle of Maximum Entropy have been widely used in several NLP applications (cf. e.g. Berger et al. 1996, Skut & Brants 1998, Charniak 2000). Specific aspects related to various levels of analysis, including

morphology, have been dealt with in information theoretic terms: cf. among else Moscoso del Prado Martín et al. (2004), who propose probabilistic measures of the informational complexity and informational residual of a word, showing that such measures are a good predictor of response latencies in visual lexical decision; Milin et al. (2009), where the processing cost of an inflected wordform is considered to be a function of the amount of information, in information-theoretic terms; Milizia (2013), where information-theoretic measures of the notion of "morphological equilibrium" are proposed and used to account for the location in paradigms of phenomena such as syncretism and semi-separate exponence.

Here, however, we are interested in the application of entropy as a way of modelling the implicative structure of morphological paradigms, and specifically the Paradigm Cell Filling Problem (henceforth abbreviated as PCFP), that was introduced in §1.2 and is repeated here in (6).

(6) The Paradigm Cell Filling Problem (PCFP)

> Given the content of one (or more) paradigm cell for a given lexeme (i.e. the pairing of a wordform with a morphosyntactic property set), what are the wordforms that realize other paradigm cells for the given lexeme?

In a series of recent studies, several ways of using information-theoretic notions to quantify the degree of uncertainty associated with this task have been proposed. Some of these proposals will be reviewed in the following sections.

2.2 Predicting exponents

An information-theoretic, entropy-based approach to the PCFP has been used as early as in the first study where such problem was formulated, i.e. Ackerman at al. (2009). The details of the procedure will be sketched in this section, again using Latin examples. The point of departure is a measure of the entropy of the random variable constituted by the different exponents that compete for the realization of a

given paradigm cell – two cells of the Latin verb paradigm are used as an example in Table 5.

Table 5: Different exponents realizing the cells PRS.ACT.IND.1SG and PRS.ACT.IND.2SG in Latin

lexeme (conj.)	PRS.ACT.IND.1SG	PRS.ACT.IND.2SG
AMO 'to love' (1st conj.)	*am-**ō***	*am-**ās***
MONEO 'to warn' (2nd conj.)	*mon-**eō***	*mon-**ēs***
SCRIBO 'to write' (3rd conj.)	*scrib-**ō***	*scrib-**is***
CAPIO 'to take' (mixed conj.)[1]	*cap-**iō***	*cap-**is***
VENIO 'to come' (4th conj.)	*ven-**iō***	*ven-**īs***

The entropy of the realization of these two cells can be calculated as in (7) and (8), where the different endings are the possible outcomes, and their probability is given by the proportion of inflection classes where they are used (see Table 6), based on the simplifying assumption that the conjugations of Table 5 are equiprobable (see the discussion in §2.3 below). Therefore, in the first-person singular the endings *-ō* and *-iō* are more likely to occur, since they appear in two inflection classes each (*-ō* in the 1st and 3rd conjugation, *-iō* in the 4th and mixed conjugation), as opposed to *-eō* that is only found in 2nd conjugation verbs; in the second-person singular, only the ending *-is* appears in two inflection classes (3rd and mixed), while all the other endings are only found in one inflection class each.

Table 6: Outcomes and probabilities of the realizations of PRS.ACT.IND.1SG and PRS.ACT.IND.2SG

6a – PRS.ACT.IND.1SG

outcome	*P*
-ō	$\frac{2}{5}$
-eō	$\frac{1}{5}$
-iō	$\frac{2}{5}$

6b – PRS.ACT.IND.2SG

outcome	*P*
-ās	$\frac{1}{5}$
-ēs	$\frac{1}{5}$
-is	$\frac{2}{5}$
-īs	$\frac{1}{5}$

[1] On this terminological choice, see §4.1 below.

(7) $H(\text{PRS. ACT. IND. 1SG}) = -\left(\frac{2}{5} \times \log_2 \frac{2}{5} + \frac{1}{5} \times \log_2 \frac{1}{5} + \frac{2}{5} \times \log_2 \frac{2}{5}\right) \approx 1.52$

(8) $H(\text{PRS. ACT. IND. 2SG}) = -\left(\frac{1}{5} \times \log_2 \frac{1}{5} + \frac{1}{5} \times \log_2 \frac{1}{5} + \frac{2}{5} \times \log_2 \frac{2}{5} + \frac{1}{5} \times \log_2 \frac{1}{5}\right) \approx 1.92$

This value is an estimate of the uncertainty in guessing the exponent that realizes a paradigm cell of a lexeme without any additional information. However, as pointed out by Ackerman & Malouf (2013: 440), this task is artificially difficult: a speaker is never faced with the problem of guessing the content of the paradigm cell of a lexeme, unless (s)he has already encountered an inflected wordform of that same lexeme – it is of course impossible to guess a wordform of a lexeme that is not known. Therefore, another information-theoretic measure is proposed, based on the notion of **conditional entropy**, measuring the uncertainty that remains about the outcome of a random variable *Y* when the value of another random variable *X* is known. Conditional entropy is calculated as in (9), where *P(y|x)* is the conditional probability of the value of the random variable *Y* being *y* knowing that the value of the random variable *X* is *x*.

(9) $H(Y|X) = -\sum_{x \in X} P(x) \log_2 P(X) \sum_{y \in Y} P(y|x) \log_2 P(y|x)$

If we apply this notion to the data in Table 5, trying to estimate the uncertainty in predicting PRS.ACT.IND.2SG knowing PRS.ACT.IND.1SG, the two variables at play can be shown in Table 7.

Table 7: Outcomes, probabilities and conditional probabilities of guessing PRS.ACT.IND.2SG from PRS.ACT.IND.1SG

known value *x* of the random variable *X*	*P(x)*	possible values *y* of the random variable *Y* given the value of *X*	*P(y|x)*
-ō	$\frac{2}{5}$	*-ās*	$\frac{1}{2}$
		-is	$\frac{1}{2}$
-eō	$\frac{1}{5}$	*-ēs*	1
-iō	$\frac{2}{5}$	*-is*	$\frac{1}{2}$
		-īs	$\frac{1}{2}$

It can be observed that if PRS.ACT.IND.1SG ends in *-eō*, then PRS.ACT.IND.2SG cannot but end in *-ēs*. However, there is uncertainty on the ending of PRS.ACT.IND.2SG when the PRS.ACT.IND.1SG ends in *-ō*, since that exponent is common to verbs of the 1st and 3rd conjugation, that differ in the exponent they use for PRS.ACT.IND.2SG (*-ās* and *-ēs*, respectively). Similarly, the 4th and mixed conjugations have the same ending *-iō* in the first-person singular, but different exponents of the second-person singular (*-is* and *-īs* respectively).

These facts concerning the implicative structure of Latin verb paradigms can nicely be captured by means of conditional entropy, as shown in (10): once the content of PRS.ACT.IND.1SG is known, the uncertainty in guessing the exponent realizing PRS.ACT.IND.2SG is considerably reduced, and so is the entropy value – 0.8 bits, as opposed to the value approaching 2 given in (8).

(10) $$H(2\text{SG}|1\text{SG}) = -\left(\frac{2}{5}\times\left(\frac{1}{2}\times\log_2\frac{1}{2}+\frac{1}{2}\times\log_2\frac{1}{2}\right)+\frac{1}{5}\times(1\times\log_2 1)+\frac{2}{5}\times\left(\frac{1}{2}\times\log_2\frac{1}{2}+\frac{1}{2}\times\log_2\frac{1}{2}\right)\right)=0.8$$

This procedure is applied by Ackerman et al. (2009) to a few case studies. Ackerman & Malouf (2013) extend such measures to a small sample of 10 typologically diverse languages, formulating the **low conditional entropy conjecture**, showing that the average conditional entropy across cells tends to remain low (around 1 bit or below) even in languages with a complex inflectional

morphology in terms of number of distinct exponents: therefore, the PCFP can be quantitatively shown to remain a feasible tasks for speakers even in such languages.

2.3 Predicting alternation patterns

The procedure outlined by Ackerman et al. (2009) can capture interesting facts on the implicative structure of paradigms, as shown in §2.2. However, there are some issues that have been pointed out in work by Bonami (2014), Bonami & Boyé (2014) and Beniamine (2018). Such issues will be shortly reviewed in this section. In the light of the facts discussed in §1.3.3, a first problem of Ackerman et al. (2009)'s strategy is that it relies on a segmentation in an invariant stem and a variable exponent that one needs to guess. This could constitute a limitation, for instance in case the language under investigation does not have a reliable morphological description, and therefore information on segmentation is lacking. But even for very well-documented and well-described languages – like Latin – there are often different options that can be chosen regarding segmentation, as was shown in §1.3.3. For instance, the segmentation problem arises very clearly in the data shown in Table 8 below.

Table 8: The cells PRS.ACT.INF and PRF.ACT.IND.1SG in Latin

lexeme (conj.)	PRS.ACT.INF	PRF.ACT.IND.1SG
AMO 'to love' (1st conj.)	*amāre*	*amāvī*
MONEO 'to warn' (2nd conj.)	*monēre*	*monuī*
SCRIBO 'to write' (3rd conj.)	*scrībere*	*scrīpsī*
CAPIO 'to take' (mixed conj.)	*capere*	*cēpī*
VENIO 'to come' (4th conj.)	*venīre*	*vēnī*

In the forms of PRF.ACT.IND.1SG it is very difficult to draw the line between stems and exponents. Only *-ī* is certainly a part of the exponent, since it appears in that cell in all lexemes. However, for this very reason it is not interesting to evaluate the predictability of that segment: since it appears in all lexemes, both its entropy and its conditional entropy will always be 0. Therefore, we might try to segment wordforms in such a way that also informative preceding segments are included in the ending. It would not be very problematic to consider the segment *-āv-* and *-u-*

as part of the ending in AMARE and MONERE, since they appear in many verbs in perfective cells. Similarly, for SCRIBERE, *-s-* could be considered as a part of the exponent. The status of the preceding *-p-*, however, is more problematic: it can be observed that this segment appears in PRF.ACT.IND.1SG but not in PRS.ACT.INF, and it is therefore partially informative on the cells that one could be facing (cf. the discussion in §1.3.3 above), although its appearance is due to a regular phonological process of assimilation of the feature [-VOICED][2].

An even more serious challenge is of course posed by cases of non-concatenative processes like the ones that happen in CAPERE and VENIRE, whose perfective stems are formed by replacing the vowel that appear in the stem of PRS.ACT.IND.1SG with another vowel – only different in length for VENIRE, also qualitatively different in CAPERE: in such cases, a segmentation in an invariant stem and a variable suffix is simply impossible.

To tackle this problem – and also some other issues that will be discussed later in this section – Bonami & Boyé (2014) and Bonami (2014) propose a revised procedure, inspired by Ackerman et al. (2009)'s proposal but designed so as to make it possible to avoid a fixed segmentation of forms *a priori*: what they do is simply looking at the alternation patterns between the considered forms. Beniamine (2018) builds on this work to make the algorithm computing what he calls **implicative entropy** applicable to typologically and structurally diverse languages – in previous work, the applicability was limited to the language under investigation, e.g. French in Bonami & Boyé (2014), Portuguese in Bonami & Luís (2014), Mauritian in Bonami et al. (2011).

The results presented in this work have been obtained by applying the algorithms of Beniamine 2018 – distributed and freely available in the form of the Qumin (Quantitative Modelling of Inflection) toolkit[3] – to Latin data. Therefore, in what follows, I will describe the design of the algorithms in some detail. To do so, I will

[2] This example raises an additional interesting issue, highlighting how in some cases considering segments – phonemes or phones – as the minimal unit of analysis is not unproblematic: in this case, for instance, only the feature [- VOICED] could be considered as part of the exponent, signalling that one is facing a perfective cell, with other features composing the sound /p/ remaining as part of the stem. Therefore, it might be interesting to compute implicative entropy starting from phonological features, rather than segments, although this would require a major revision of the methodological approach what we cannot tackle in the present work, and that we thus leave to future research.

[3] Available at https://github.com/XachaB/Qumin.

consider the same cells as in Table 5, with some additional lexemes (cf. Table 9). The task we focus on is again guessing PRS.ACT.IND.2SG knowing PRS.ACT.IND.1SG.

Table 9: The cells PRS.ACT.IND.1SG and PRS.ACT.IND.2SG in Latin – A sample of 15 verbs

conj.	lexeme	PRS.ACT.IND.1SG	PRS.ACT.IND.2SG
1st	AMO 'to love'	amo:	ama:s
	CUBO 'to lie down'	kubo:	kuba:s
	PLICO 'to fold'	pliko:	plika:s
2nd	DELEO 'to destroy'	de:leo:	de:le:s
	HABEO 'to have'	habeo:	habe:s
	MONEO 'to warn'	moneo:	mone:s
3rd	DICO 'to say'	di:ko:	di:kis
	RUMPO 'to break'	rumpo:	rumpis
	SCRIBO 'to write'	skri:bo:	skri:bis
mixed	CAPIO 'to take'	kapio:	kapis
	FACIO 'to make'	fakio:	fakis
	RAPIO 'to snatch'	rapio:	rapis
4th	AMBIO 'to surround'	ambio:	ambi:s
	AUDIO 'to hear'	awdio:	awdi:s
	VENIO 'to come'	wenio:	weni:s

Table 10: Alternation patterns and contexts for the data of Table 9

alternation patterns and contexts	lexemes instantiating the pattern
1. _o: ↔ _a:s / C_#	AMO, CUBO, PLICO
2. _eo: ↔ _e:s / C_#	DELEO, HABEO, MONEO
3. _o: ↔ _is / C_#	DICO, RUMPO, SCRIBO
4. _o: ↔ _s / i_#	CAPIO, FACIO, RAPIO
5. _io: ↔ _i:s / C_#	AMBIO, AUDIO, VENIO

As can be seen, in the table there are only forms in phonetic transcription, with no segmentation. This given, the first step of the procedure consists in aligning the forms and finding the alternation patterns that occur between them. For the 15 verbs of Table 9, we find five different alternation patterns (as displayed in the first part

of the first column of Table 10)[4], corresponding to the four traditional conjugations, plus the so-called mixed conjugation.

The second step is finding a generalization on the phonological contexts where such alternation patterns occur in the data, as in the second part of the first column of Table 10. For this purpose, Beniamine (2018)'s algorithm is based on a strategy that is inspired by the Minimal Generalization Learner (cf. Albright 2002, Albright & Hayes 2003), that starts from the individual contexts of each pair of wordforms sharing the same alternation pattern, and then merge them, finding the tightest rule able to cover all the relevant cases.

For instance, in our data it can observed that pattern 4 only appears after the vowel /i/, while for all the other patterns we could just state that they occur after a consonant. Of course, such contexts are a very rough simplification, but the algorithm can express complex restrictions on the occurrence of the different patterns.[5]

Based on the alternation patterns and their contexts of application, as summarized in Table 10, it is possible to move on to the third step of the procedure, i.e. classifying the verbs in our sample according to the patterns that can be applied, based on the pattern itself and its context of application, as in Table 11. While the alternation patterns are bidirectional, in this case the classification is based on the patterns that could be applied to the inflected wordform in PRS.ACT.IND.1SG in order to obtain the one in PRS.ACT.IND.2SG. It can be observed in Table 11 that patterns 1 and 3 could be applied to verbs of the 1st and 3rd conjugations, that are therefore grouped together in this classification. Verbs of the 4th and mixed conjugations also belong to a same class, since patterns 4 and 5 can be applied to both. Verbs of the 2nd conjugation have their separate class, since only pattern 2 can be applied to them.

[4] In this notation, the symbol "_" stands for the segment of the wordform that does not vary; the double arrow "↔" is intended to stress the bidirectionality of the alternation. "C", "V" and "#" mean 'consonant', 'vowel' and 'end of word', with "/" separating the pattern from the context, as usual in the SPE rule format (Chomsky & Halle 1968).

[5] For further details on the exact way in which alternation patterns and context are applied, the reader is referred to Beniamine (2018: §2).

Table 11: The cells PRS.ACT.IND.1SG and PRS.ACT.IND.2SG – patterns, contexts and applicable patterns

conj.	lexeme	PRS.ACT. IND.1SG	PRS.ACT. IND.2SG	pattern used / context	applicable patterns[6]
1st	AMO	amo:	ama:s	1. _o: ↔ _a:s / C_#	1,3
	CUBO	kubo:	kuba:s	1. _o: ↔ _a:s / C_#	1,3
	PLICO	pliko:	plika:s	1. _o: ↔ _a:s / C_#	1,3
2nd	DELEO	de:leo:	de:le:s	2. _eo: ↔ _e:s / C_#	2
	HABEO	habeo:	habe:s	2. _eo: ↔ _e:s / C_#	2
	MONEO	moneo:	mone:s	2. _eo: ↔ _e:s / C_#	2
3rd	DICO	di:ko:	di:kis	3. _o: ↔ _is / C_#	1,3
	RUMPO	rumpo:	rumpis	3. _o: ↔ _is / C_#	1,3
	SCRIBO	skri:bo:	skri:bis	3. _o: ↔ _is / C_#	1,3
mixed	CAPIO	kapio:	kapis	4. _o: ↔ _s / i_#	4,5
	FACIO	fakio:	fakis	4. _o: ↔ _s / i_#	4,5
	RAPIO	rapio:	rapis	4. _o: ↔ _s / i_#	4,5
4th	AMBIO	ambio:	ambi:s	5. _io: ↔ _i:s / C_#	4,5
	AUDIO	awdio:	awdi:s	5. _io: ↔ _i:s / C_#	4,5
	VENIO	wenio:	weni:s	5. _io: ↔ _i:s / C_#	4,5

This procedure allows us to obtain the two random variables that we need to perform the conditional entropy calculations. The first random variable, of which we assume that the value is known, is the classification based on applicable patterns – since, in order to know that, it suffices to know the form of PRS.ACT.IND.1SG. The other random variable – the one whose value we want to guess – is the alternation pattern: when the wordform in PRS.ACT.IND.1SG and the alternation pattern to obtain PRS.ACT.IND.2SG from it are known, the wordform in PRS.ACT.IND.1SG follows. This situation is summarized in Table 12.

[6] For reasons of space, here and in similar tables we refer to patterns by means of the numbers assigned to them in the preceding column.

Table 12: Outcomes, probabilities and conditional probabilities of guessing PRS.ACT.IND.2SG from PRS.ACT.IND.1SG

known value *x* of the random variable *X*	*P(x)*	possible values *y* of the random variable *Y* given the value of *X*	*P(y|x)*
1,3	$\frac{2}{5}$	_o: ↔ _a:s / C_#	$\frac{1}{2}$
		_o: ↔ _is / C_#	$\frac{1}{2}$
2	$\frac{1}{5}$	_eo: ↔ _e:s / C_#	1
4,5	$\frac{2}{5}$	_o: ↔ _s / i_#	$\frac{1}{2}$
		_io: ↔ _i:s / C_#	$\frac{1}{2}$

Thus, we can now compute entropy as in (11).

(11) $H(2\text{SG}|1\text{SG}) = -\left(\frac{2}{5} \times \left(\frac{1}{2} \times \log_2 \frac{1}{2} + \frac{1}{2} \times \log_2 \frac{1}{2}\right) + \frac{1}{5} \times (1 \times \log_2 1) + \frac{2}{5} \times \left(\frac{1}{2} \times \log_2 \frac{1}{2} + \frac{1}{2} \times \log_2 \frac{1}{2}\right)\right) = 0.8$

It can be observed that, in this case, the result is the same that was obtained with Ackerman et al. (2009)'s procedure, which indeed is an efficient way of capturing generalizations on implicative relations in relatively unproblematic cases like this one. However, this refined procedure has the advantage of being much more easily extendable to more complex cases, for instance the two cells of Table 8. Although in such cases it is difficult to come up with a satisfying segmentation in stems and exponents, there is no problem in obtaining alternation patterns between the involved wordforms, and the fact that such patterns are sometimes non-concatenative – as shown in Table 13 – does not change the way in which the procedure can be applied.

Table 13: Alternation patterns and contexts for the data of Table 8

lexeme (conj.)	PRS.ACT.INF	PRF.ACT.IND.1SG	pattern used / context[7]
AMO 'to love' (1st conj.)	ama:re	ama:wi:	_re ↔ _wi: / ama:_#
MONEO 'to warn' (2nd conj.)	mone:re	monui:	_e:re ↔ _ui: / mon_#
SCRIBO 'to write' (3rd conj.)	skri:bere	skri:psi:	_bere ↔ _psi: / skri:_#
CAPIO 'to take' (mixed conj.)	kapere	ke:pi:	_a_ere ↔ _e:_i: / k_p_#
VENIO 'to come' (4th conj.)	weni:re	we:ni:	_e_re ↔ _e:_ / w_ni:_#

Another important improvement suggested by Bonami & Boyé (2014) and subsequent work concerns the way of estimating the probability of the possible outcomes – i.e. of the different realizations of paradigm cells. In the examples above, the probability of a given realization was estimated by looking at the number of distinct inflection classes exhibiting such realizations. Going back to the data in Table 5, given the five major inflection classes relevant for imperfective forms in Latin verb paradigms, since for PRS.ACT.IND.1SG the ending *-eō* is only used in the 2nd conjugation, a probability of ⅕ was assigned to it, while the endings *-ō* and *-iō* were considered as having a probability of ⅖ (since they both appear in two conjugations, 1st-3rd and 4th-mixed, respectively). This means that an assumption is made that all inflection classes are equiprobable, as would be the case if the verbs of Latin were only the ones given in Table 9, where there are three verbs for each conjugation. However, of course there are many more verbs in Latin, and their distribution across inflection classes is not at all balanced: while the 1st and 3rd conjugations both have many members, the 2nd, 4th and mixed conjugation are much more marginal in terms of number of verbs.

[7] Since in this example there is only one verb per pattern, the context is simply given by the unchanged segments in the two wordforms. Of course, adding more verbs the context would become accordingly less specific.

Therefore, it would be useful to be able to use information on the type frequency[8] of the different inflection classes as a better estimate of their probability of occurrence, and the aforementioned Qumin toolkit allows for a principled and automatic way of doing so.[9] Suppose that, instead of the 15 verbs of Table 9, we have a large, representative sample of Latin verbs: as an estimate of the probability of application of the different patterns (and similarly for the classes based on applicable patterns), we can then use the actual number of verbs where those patterns and classes are attested. In Table 14, data from LatInfLexi (cf. below, Chapter 3) are provided. It can be observed that patterns 1 and 3 appear in a large number of verbs: thus, they can be considered as more likely outcomes than the other, rarer, patterns. Therefore, the entropy calculation can be performed as in (12), where the probabilities given in Table 12 and used in (11) have been replaced by probabilities based on the type frequency of different patterns and classes of applicable patterns, as shown in Table 15.

Table 14: Data on the type frequency of Latin conjugations

conj.	lexeme	PRS.ACT. IND.1SG	PRS.ACT. IND.2SG	pattern/ context	applicable patterns	n. verbs
1st	AMO	amo:	ama:s	1. _o: ↔ _a:s / C_#	1,3	1,332
2nd	MONEO	moneo:	mone:s	2. _eo: ↔ _e:s / C_#	2	298
3rd	SCRIBO	skri:bo:	skri:bis	3. _o: ↔ _is / C_#	1,3	1,152
mix.	CAPIO	kapio:	kapis	4. _o: ↔ _s / i_#	4,5	132
4th	VENIO	wenio:	weni:s	5. _io: ↔ _i:s / C_#	4,5	169

[8] The use of type frequency, rather than token frequency, is in line with Bybee's (1995: 433 ff.) observation that it is the former that correlates with the productivity of morphological patterns. On the contrary, high token frequency actually detracts from the strength of a given alternation pattern, since it makes the wordforms involved in it more likely to be stored as such.

[9] For a different approach to the role of type frequency in reducing the complexity of the PCFP, see Sims & Parker (2016).

Table 15: outcomes, probabilities and conditional probabilities of guessing PRS.ACT.IND.2SG from PRS.ACT.IND.1SG (based on type frequency data)

known value *x* of the random variable *X*	*P(x)*	possible values *y* of the random variable *Y* given the value of *X*	*P(y|x)*
1,3	$\frac{2,484}{3,083}$	_o: ↔ _a:s / C_#	$\frac{1,332}{2,484}$
		_o: ↔ _is / C_#	$\frac{1,152}{2,484}$
2	$\frac{298}{3,083}$	_eo: ↔ _e:s / C_#	1
4,5	$\frac{301}{3,083}$	_o: ↔ _s / i_#	$\frac{132}{301}$
		_io: ↔ _i:s / C_#	$\frac{169}{301}$

(12) $$H(2\text{SG}|1\text{SG}) = -\left(\frac{2,484}{3,083} \times \left(\frac{1,332}{2,484} \times \log_2 \frac{1,332}{2,484} + \frac{1,152}{2,484} \times \log_2 \frac{1,152}{2,484}\right) + \frac{298}{3,083} \times (1 \times \log_2 1) + \frac{301}{3,083} \times \left(\frac{132}{301} \times \log_2 \frac{132}{301} + \frac{169}{301} \times \log_2 \frac{169}{301}\right)\right) \approx 0.9$$

Ackerman et al. (2009: 65, Footnote 8) acknowledge the fact that type frequency can play a role, claiming that the entropy values they provide can be considered as upper bounds to the uncertainty in the PCFP. However, the example that was provided shows that actually entropy can turn out to be higher than it would be if we assumed an equiprobable distribution of verbs among inflection classes – see the entropy value of 0.8 bits in (11) as opposed to the value of about 0.9 in (12): this is due to the fact that also the weight of 0-entropy cases can be overestimated by that assumption, as is the case of 2nd conjugation verbs in the example, whose impact is quantitatively less relevant if type frequency data are taken into account. Notice also that this way of expressing the different inflectional behaviours of the lexemes does not rely on the availability of pre-existing descriptions of inflection classes, but can be directly derived from the data – i.e. simply given the pair of wordforms that is considered, from which the actual alternation pattern that is instantiated can easily be inferred.

2.4 Predicting from more than one wordform

We saw in §2.2 that the reason for using conditional entropy is the fact that guessing the content of a paradigm cell of a lexeme without knowing any other inflected wordform of that lexeme is an unrealistically difficult task. However, recent corpus-based investigations on the statistical distribution of the various wordforms in large morphological paradigms (e.g. Chan 2008: 79 ff., Bonami & Beniamine 2016: 159 ff., Blevins et al. 2017) show that it is often the case that for a given lexeme more than one inflected wordform is attested, but in many cases – in very large paradigms, virtually always – the paradigm is not "saturated" (Chan 2008: 79), i.e. not all the wordforms are attested. Therefore, in many cases speakers know more than one wordform of a lexeme, without knowing all of them. This suggests that speakers, when faced with the PCFP, can actually rely on information on more than one wordform when they need to guess the content of a given paradigm cell.

Indeed, in some cases information on two wordforms is more useful in reducing uncertainty in the PCFP than knowing each one of the wordforms. Consider for instance the Latin data in Table 16, supposing that the content of the cells PRS.ACT.IND.1SG and PRS.ACT.IND.2SG is known, while the cell PRS.ACT.IND.3PL is the one whose content we need to guess. If we only knew the wordforms realizing PRS.ACT.IND.1SG we would not be able to guess the PRS.ACT.IND.3PL of verbs ending in *-ō*, since in some cases (namely, 1st conjugation verbs) the ending would be *-ant* and in other cases (3rd conjugation verbs) it would be *-unt*. By contrast, if we only knew PRS.ACT.IND.2SG we would not be able to guess the PRS.ACT.IND.3PL of verbs ending in *-is*, since 3rd conjugation verbs would end in *-unt* and verbs of the mixed conjugation would end in *-iunt*. However, if we assume that the content of both PRS.ACT.IND.1SG and PRS.ACT.IND.2SG is known, than all the ambiguities are resolved, since the ambiguity between 1st and 3rd conjugation in the first-person singular is disambiguated by the different realizations (*-ās* and *-is*) in the second-person singular, and conversely the ambiguity between the 3rd and mixed conjugation in the second-person singular is disambiguated by the different realizations (*-ō* and *-iō*) in the first-person singular. Therefore, in this example no

single cell is a reliable predictor of PRS.ACT.IND.3PL, but taken together the two cells allow for a categorical inference.

Table 16: The cells PRS.ACT.IND.1SG, PRS.ACT.IND.2SG and PRS.ACT.IND.3PL in Latin

lexeme	PRS.ACT. IND.1SG	PRS.ACT. IND.2SG	PRS.ACT. IND.3PL
AMO 'to love' (1st conj.)	*amō*	*amās*	*amant*
MONEO 'to warn' (2nd conj.)	*moneō*	*monēs*	*monent*
SCRIBO 'to write' (3rd conj.)	*ducō*	*ducis*	*ducunt*
CAPIO 'to take' (mixed conj.)	*capiō*	*capis*	*capiunt*
VENIO 'to come' (4th conj.)	*veniō*	*venīs*	*veniunt*

Because of such facts, it could be useful to have a way of applying the procedure described in §2.3 to estimate the uncertainty in the PCFP given knowledge of more than one paradigm cell. Bonami & Beniamine (2016) describe a way to do so, by computing what they call ***n*-ary implicative entropy**, which will be detailed in this section.

For practical reason, rather than using alternation patterns between more than one form, Bonami & Beniamine (2016) outline a strategy that only relies on binary alternation patterns. Let us begin from the simpler case of predicting the content of a paradigm cell given the content of two different paradigm cells (**binary implicative entropy**). In this procedure, the random variable whose value has to be predicted is a joint variable consisting of: i) the alternation pattern between the first known form and the unknown form, notated as A↔C and ii) the alternation pattern between the second known form and the same unknown form, notated as B↔C. Knowing such alternation patterns is equivalent to knowing the form of C, which is what we are trying to guess.

On the other hand, the random variable whose value is known is again a joint variable, consisting of: i) the patterns that could be applied to A in order to obtain C, notated as A(A↔C); ii) the patterns that could be applied to B in order to obtain C, notated B(B↔C); iii) the alternation patterns that occur between A and B, notated A↔B. All this information can be reasonably considered to be known once the forms A and B are given.

Therefore, the entropy of guessing the first joint random variable that was described knowing the second one can be considered as equivalent to guessing the wordform in an unknown cell C given the wordforms in cells A and B:

(13) $H(C|A,B) = H(A \leftrightarrow C, B \leftrightarrow C | A(A \leftrightarrow C), B(B \leftrightarrow C), A \leftrightarrow B)$

Let us apply this procedure to the example in Table 16. The relevant forms and variables are given in Table 17. In Table 18, it is shown that in each class, when the value of the joint random variable *Y* is known, then there is only one possible outcome of the joint random variable *X*: therefore, there is no uncertainty, and entropy is 0, as shown in (15).

A trivial property of the binary entropy of predicting cell C knowing cells A and B is that it can never be greater than the unary entropy of predicting cell C knowing only cell A or only cell B: knowledge of multiple forms can reduce the uncertainty in the PCFP (although it does not necessarily do so), but cannot make this task more difficult.

Such procedure can easily be generalized to predictions from more than two forms (*n*-ary implicative entropy), as shown in Bonami & Beniamine (2016: 172):

(14) $H(B|A1, \dots, An) = H(A1 \leftrightarrow B, \dots, An \leftrightarrow B | A(A1 \leftrightarrow B), \dots, A(An \leftrightarrow B), [A1 \leftrightarrow An])$

Taking into account more than one wordform when predicting the content of other paradigm cells clearly shows the relationship with a related but different way of investigating the implicative structure of morphological paradigms, namely the use of Principal Parts, as implemented in recent work on what has been called "Principal Part Analysis" (cf. notably Stump & Finkel 2013, but also Finkel & Stump 2007, 2009a). We saw in §1.3.3 that principal parts can be defined as a set of cells of a lexeme's paradigm from which all the other cells of the same lexeme can be filled with no uncertainty. In terms of *n*-ary entropy, this means that any set of cells that allows to guess the content of all the other cells with H=0 can be considered as a reliable principal part set. Therefore, *n*-ary entropy can be used as a principled way

Table 17: The cells PRS.ACT.IND.1SG, PRS.ACT.IND.2SG and PRS.ACT.IND.3PL – wordforms, patterns and applicable patterns

lexeme	PRS.ACT.IND.1SG	PRS.ACT.IND.2SG	PRS.ACT.IND.3PL	A↔C	B↔C	A(A↔C)	B(B↔C)	A↔B
AMO	amo:	ama:s	amant	1. _o: ↔ _ant / C_#	1. _a:s ↔ _ant / C_#	1,3	1	1. _o: ↔ _a:s / C_#
MONEO	moneo:	mone:s	monent	2. _o: ↔ _nt / e_#	2. _e:s ↔ _ent / C_#	2	2	2. _eo: ↔ _e:s / C_#
SCRIBO	skri:bo:	skri:bis	skri:bunt	3. _o: ↔ _unt / _#	3. _is ↔ _unt / C_#	1,3	3,4	3. _o: ↔ _is / C_#
CAPIO	kapio:	kapis	kapiunt	3. _o: ↔ _unt / _#	4. _s ↔ _unt / i_#	3	3,4	4. _o: ↔ _s / i_#
VENIO	wenio:	weni:s	weniunt	3. _o: ↔ _unt / _#	5. _i:s ↔ _iunt / C_#	3	5	5. _io: ↔ _i:s / C_#

Table 18: outcomes, probabilities and conditional probabilities of guessing PRS.ACT.IND.3SG knowing PRS.ACT.IND.1SG and PRS.ACT.IND.2SG

known value *x* of the joint random variable *X*: A(A↔C);B(B↔C);A↔B	*P(x)*	possible values *y* of the joint random variable *Y*: A↔C; B↔C	*P(y|x)*
1,3; 1; _o: ↔ _a:s / C_#	$\frac{1}{5}$	_o: ↔ _ant / C_#; _a:s ↔ _ant / C_#	1
2; 2; _eo: ↔ _e:s / C_#	$\frac{1}{5}$	_o: ↔ _nt / e_#; _e:s ↔ _ent / e_#	1
1,3; 3,4; _o: ↔ _is / C_#	$\frac{1}{5}$	_o: ↔ _unt / _#; is ↔ _unt / C_#	1
3; 3,4; _o: ↔ _s / i_#	$\frac{1}{5}$	_o: ↔ _unt / _#; s ↔ _unt / i_#	1
3; 5; _io: ↔ _i:s / C_#	$\frac{1}{5}$	_o: ↔ _unt / _#; _i:s ↔ _iunt / C_#	1

(15) $$H(3\text{PL}|1\text{SG}, 2\text{SG}) = -\left(\frac{1}{5} \times (1 \times \log_2 1) + \left(\frac{1}{5} \times (1 \times \log_2 1)\right) + \left(\frac{1}{5} \times (1 \times \log_2 1)\right) + \left(\frac{1}{5} \times (1 \times \log_2 1)\right) + \left(\frac{1}{5} \times (1 \times \log_2 1)\right)\right) = 0$$

to find principal parts (and near-principal parts, i.e. sets of cells that allow for quasi categorical inference, with close-to-0 entropy values, as done by Bonami & Beniamine 2016) for a given inflectional system. With the aforementioned Qumin toolkit, also *n*-ary entropy calculations can be performed automatically, and results on Latin paradigms and their principal parts will be shown in the following chapters.[10]

2.5 Predicting forms knowing more than just forms

The procedure described above provides a method to assess the uncertainty in the PCFP having as the only information the phonotactic shape of the known wordform(s), and the paradigm cell that is occupied by it/them. However, this can be considered as an upper bound to the complexity of the implicative structure of inflectional paradigms: speakers do often have much more information on the wordforms that they learn and use, and this additional information can prove useful in reducing the uncertainty in the PCFP. In this work, we will also exploit an additional feature of the aforementioned Qumin toolkit – namely, the possibility of assigning lexemes to different classes[11] – in order to investigate the reduction in uncertainty that can be obtained when other aspects of the inflected wordform(s) are assumed to be known, beside phonotactics.

A first additional piece of information on lexemes that can help speakers when facing the PCFP concerning nouns is gender. In Latin, nouns traditionally considered as belonging to the same declension actually display different endings in some cells – namely, the nominative, accusative and vocative – exactly according to their gender: for instance, the 2nd-declension masculine noun LUPUS 'wolf' has NOM.SG *lupus*, ACC.SG *lupum* and VOC.SG *lupe*, while a neuter noun of the same declension, like BELLUM 'war', has *bellum* in all of those cells. Therefore, the gender of a noun is a crucial information when facing the PCFP, and its quantitative impact on our results on nominal inflection will be discussed in §5.3.

[10] In the provided example, information on type frequency has been omitted: since we obtain a null entropy value, in this case such information would not have made a difference. However, in the results presented in this work information on type frequency is always taken into account, also when considering binary and *n*-ary implicative entropy.

[11] We wish to thank Sacha Beniamine for working on the addition of such additional feature to the toolkit for the purposes of this work.

In Chapter 6 we will focus on information on derivational relatedness. Intuitively, the fact that the PRF.ACT.IND.1SG of CONFERO is the suppletive *contulī* is made much less unpredictable by the fact the it comes from FERO, whose PRF.ACT.IND.1SG is *tulī*. If we assume that we know that CONFERO comes from FERO, then its irregularity in PRF.ACT.IND.1SG can be derived straightforwardly from the irregularity of the base verb. Furthermore, when using quantitative data, if we only consider the base lexeme as a type, instead of counting each lexeme derived from the same base as a separate type – as is usually done in previous work, and also in Chapter 4 and Chapter 5 of this work – the quantitative estimation of the probability of different inflectional patterns can be quite different. This is especially relevant given the high number of verbs that are formed by adding a preverb to a given base in Latin, as we will see in §6.2. In a similar fashion, the inflectional behaviour of the noun MUTATIO 'change' is made predictable by the fact that it is inherited by the suffix with which it is formed – namely, *-tio*, forming action nouns from verbs. This aspect of derivational relatedness will be investigated in noun inflection (cf. §6.3), where it is suffixation – rather than prefixation as in verbs – that has the lion's share.

2.6 Conclusion: an entropy-based approach to the PCFP

In this chapter, we have outlined a method to measure the uncertainty in the PCFP – and, more generally, to investigate the implicative structure of morphological paradigms – by using the information-theoretic notion of conditional entropy. The described procedure allows to operate just on the basis of inflected wordforms, with no segmentation, by looking at the alternation patterns between them and their context of application, thus adopting a fully abstractive approach, as was deemed preferable in §1.3.3. Another desired characteristic that was discussed in §1.4 and is met by this method is the possibility to take into account information on the type frequency of different inflectional patterns, and thus to not disregard non-categorical but equally interesting implicative relations. Furthermore, the tools that will be used allow to weight the impact of knowledge of more than one wordform and of other information such as the derivational history of a lexeme or the gender of a noun in making the PCFP easier.

However, to make all this possible, a large, representative inflected lexicon of Latin verbs and nouns is needed. In the next chapter, the procedure to obtain such a lexicon will be described in detail.

Chapter 3. The data

To perform a quantitative, entropy-based analysis like the one that has been sketched out in the previous chapter, a large, representative lexicon of inflected wordforms in phonetic transcription is needed. Similar resources are being increasingly developed for modern languages: cf. among else Zanchetta & Baroni (2005) and Calderone et al. (2017) for Italian, Bonami et al. (2014) and Hathout et al. (2014) for French, the lexicon described in Bonami & Luís (2014) for Portuguese.

However, there is no freely and easily accessible inflected lexicon of this kind for Latin, although the current availability of several morphological analysers – e.g. *Words* (`http://archives.nd.edu/words.html`), *Morpheus* (`https://github.com/tmallon/morpheus`), the PROIEL Latin morphology system (`https://github.com/mlj/proiel-webapp/tree/master/lib/morphology`) and *LatMor* (`http://cistern.cis.lmu.de`) – allows for the semi-automatic creation of such a resource.

This is exactly what we did to obtain LatInfLexi (cf. Pellegrini & Passarotti 2018), the inflected lexicon on which the results of this work are based: the database of a recently renewed Latin morphological analyser – namely, Lemlat 3.0 (Passarotti et al. 2017) – was used as a source of the pieces of information that are necessary to get full paradigms for a large enough number of verbs and nouns. This chapter provides a description of the general structure and characteristics of our lexicon (§3.1), detailing the procedure that has been followed regarding the selection of lexemes (§3.2) and the generation of wordforms (§3.3). This is followed in §3.4 by a discussion of the choices that were made to reduce the impact of the phenomenon of overabundance, by selecting only one wordform for each paradigm cell when more than one option was potentially available. In §3.5, theoretical and practical issues related to the use of phonetic transcription are treated. We conclude by providing some information on the size of the resource in terms of number of lexemes, paradigm cells and wordforms in §3.6.

3.1 The structure of LatInfLexi

A remarkable feature of our lexicon is that it is based on lexemes and paradigm cells, rather than on wordforms: this means that for each lexeme, all the morphologically possible wordforms are included regardless of their actual attestation and usage in texts. This characteristic is due to the fact that the Qumin toolkit that we used to compute entropy excludes empty cells, but given the aforementioned intrinsically sparse nature of morphological paradigms, empty cells would have been numerous – especially in the very large verbal paradigm of Latin – if only attested wordforms were provided.

In this respect, our resource is therefore similar to other lexicons on which entropy-based analyses of implicative structure were based, notably Flexique (Bonami et al. 2014), where there is one line per lexeme and one column per paradigm cell, and the relevant inflected wordforms are displayed in the corresponding intersection.

However, our lexicon is distributed in a different format: there is one line for each paradigm cell, on which the following information is provided:

(i) a univocal identifier of the lexeme to which the paradigm cell belongs, corresponding to the lemma used in Lemlat's database, plus a numerical diacritic in the rare cases of different lexemes with the same citation form, e.g. VOLO$_1$ 'to fly' and VOLO$_2$ 'to want';

(ii) the morphosyntactic property set that is expressed, using the PoS-tags of the Universal Part-of-Speech Tagset by Petrov et al. (2011) and the morphological features of Universal Dependencies (`http://universaldependencies.org/u/feat/index.html`);

(iii) the inflected wordform filling the cell, both in orthographic and in phonetic transcription, with #DEF# marking cells that are empty, not because they are simply not attested in texts, but rather because they are systematically defective, like for instance passive forms of intransitive verbs and active forms

of deponent verbs (cf. below, §3.3.1), but also singular forms of *pluralia tantum* nouns;

(iv) information on the frequency of the wordform, taken from Tombeur (1998)'s *Thesaurus Formarum Totius Latinitatis*, a resource providing data on the number of occurrences of Latin forms in four different eras: *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).[1]

Table 1 exemplifies the structure of LatInfLexi, by providing the content of two paradigm cells of the *plurale tantum* NUPTIAE 'marriage'.

Table 1: LatInfLexi: some examples of paradigm cells

lexeme	MPS	wordform		frequency (TFTL)				
		graph.	phon.	*Antiquitas*	*Aetas Patrum*	*Medium Aeuum*	*Recentior Latinitas*	tot.
NUPTIAE	NOUN: Nom+Sing	#DEF#	#DEF#	0	0	0	0	0
NUPTIAE	NOUN: Nom+Plur	*nuptiae*	[nuptiaj]	61	440	326	4	831

3.2 The selection of lexemes

The database of Lemlat, the analyser that was used as a source of information in order to generate the inflected wordforms of our lexicon, is very extensive and includes many marginal lexical items: in its "lemmario", listing all the lemmas in the database, currently[2] there are 17,979 verbal entries and 103,916 nominal entries, including proper names taken from Forcellini (1940)'s Onomasticon (cf. Budassi & Passarotti 2016) and a large number of items from a Medieval Latin glossary, Du Cange et al. (1883-1887) – cf. Cecchini et al. (2018).

While such a wide coverage is undoubtedly a desired feature for a tool designed in order to analyse forms, much more caution is necessary when one wants to obtain

[1] In this work, we do not use frequency data. Therefore, we omit the discussion of the ambiguity issues related with such data in LatInfLexi and in the source from which they are taken, that the reader can find in Pellegrini & Passarotti (2018).

[2] The count was performed on January 22, 2019.

full paradigms, to avoid the risk of overgeneration. Therefore, a selection of lexemes was made, in order to i) reduce the considered time span excluding at least Medieval Latin and focusing on Classical Latin, and ii) keep only the most frequent lexemes, avoiding very marginal items.

To satisfy such requirements, the first move was, of course, to exclude the items of Lemlat that are taken from Du Cange et al. (1883-1887), thus avoiding lexemes that are attested only in Medieval Latin.[3] However, even if such items are not included, there is still a very large number of lexemes, especially nouns (42,144 nominal entries). Even if we remove proper names taken from Forcellini's Onomasticon, we are still left with 22,544 nouns, including very rare items. In order to exclude such marginal entries and further restrict the considered time span, a frequency lexicon of Classical Latin was used, namely Delatte et al. (1981)'s *Dictionnaire fréquentiel et Index inverse de la langue latine* – henceforth DFILL. This lexicon is based on a 794.662-token corpus developed by the *Laboratoire d'Analyse statistique des Langues anciennes* (L.A.S.L.A.) in Liège. It only includes authors ranging from the beginning of the 1st century B.C. to the early beginning of the 2nd century A.D.,[4] thus providing a tighter constraint on the diachronic variation in our data. Only items that are attested in DFILL are kept as lexemes in our resource. Regarding verbs, all of them are included in our lexicon, while for nouns the data used in this work only comprises the nominal entries with at least 30 occurrences.[5] The figures are summarized in Table 2.

Table 2: The number of lexemes in LatInfLexi

Part-of-Speech	n. lexemes
N	1,038
V	3,348

[3] Du Cange et al. (1883-1887) also contains words that were already attested in Classical Latin, but underwent some kind of formal or semantic change in Medieval Latin: of course, at least some of these words will be present in our resource, since they also appear in other sections of Lemlat's database, being also listed in lexicons of Classical Latin.

[4] Cf. Delatte et al. (1981: 1) for the full list of texts and authors.

[5] The lexicon of nouns is not very large in size if compared with the one of verbs, especially if one considers that nouns are actually much more numerous than verbs in Lemlat's database. This limitation is simply due to reasons of time: however, we plan to include all the nouns of DFILL in our nominal lexicon in the near future.

It should be noticed that there are some minor discrepancies between the number of entries of DFILL and the number of lexemes in our resource.

Firstly, there are some entries of DFILL that are not reported as such in Lemlat's database. In some cases, this is simply due to different choices regarding the citation form: for instance, the lexeme meaning 'to stamp (the foot)' is lemmatized as SUPPLAUDO in DFILL, as SUPPLODO in Lemlat. Similar cases were resolved by manually checking the correspondence between the two graphical variants. In other cases, however, the entry of DFILL does not have any correspondence in Lemlat's database: for instance, in DFILL the form *inexercitatum* 'unexercised' is considered to be the past participle of the verb INEXERCITO, which is accordingly present in the list of lemmas, while in Lemlat it is considered as an adjective in its own right, since the verb INEXERCITO is not reported in the lexicons from which its database has been built. In this case, the entry of DFILL cannot but be left out from our resource – trivially, because it would be impossible to automatically generate its inflected wordforms with the same procedure used for other lexemes.

Secondly, we have some verbal lexemes in LatInflexi that correspond to more than one entry of DFILL. For instance, in LatInfLexi there is only one lexeme VERSO, while DFILL lists two distinct entries, namely VERSO – using as citation form PRS.ACT.IND.1SG – and VERSOR – using as citation form the corresponding, morphologically passive form, PRS.PASS.IND.1SG. This implies an analysis where VERSOR, meaning 'to turn', and VERSO, meaning 'to remain', constitute two separate lexemes because of their different meaning. However, since the entropy-based analysis that we want to perform is essentially linked to phonotactic aspects, and is not influenced by semantics, it seems more reasonable to take a formal criterion as decisive in such cases: therefore, all the forms that can be assigned to the same lexeme based on their formal relatedness – like the morphologically active and passive forms *versō* (PRS.ACT.IND.1SG) and *versor* (PRS.PASS.IND.1SG) – were collected in the paradigm of a single lexeme, regardless of the fact that they are sometimes used with quite different meanings.

Another discrepancy, this time specifically concerning nouns, is due to the fact that DFILL contains nominal entries that refer to nominal uses of adjectives and participles: for instance INIMICUS is explicitly listed as a noun (meaning 'enemy')

in DFILL, although many dictionaries of Latin only mention this meaning as a nominal use of forms of the adjective INIMICUS ‘unfriendly’; similarly, DFILL contains an entry for the noun VERUM ‘the truth’, although this can be considered simply as a nominal use of neuter forms of the adjective VERUS ‘true’. In such cases, the choice made in LatInfLexi is simply the one that was made in Lemlat: while there is an entry for the noun VERUM in Lemlat’s “lemmario”, there is no nominal entry for INIMICUS, which only appears in Lemlat as an adjective. Therefore, the first one was kept in our resource, but the second one was discarded, simply because it would have been impossible to obtain the information necessary to generate the wordforms (see below, §3.3) without it being listed in Lemlat’s database.

3.3 The generation of wordforms

Having selected the lexemes, the next step is generating the wordforms that occupy the various cells of the nominal and verbal paradigm of Latin. To do so, we exploited the “lessario” table of Lemlat’s database, where for each lexeme a list of LESs ‘LExical Segments’ (roughly corresponding to different stems of the involved lexeme) is provided, each of them associated with a CODLES providing information on the endings that can be added to them (roughly corresponding to the traditional inflection classes), as well as on other properties of the lexeme, as we will see below in §3.3.1-2. The details of the procedure are slightly different for verbs and for nouns and are described in this section in §3.3.1 and §3.3.2, respectively.

3.3.1 Verb paradigms

The easiest and clearest way to explain the procedure that was followed is by means of an example. Let us consider the verb STO ‘to stay’: Lemlat’s database lists the LESs and associated CODLESs displayed in Table 3.

Table 3: LESs and CODLESs listed in Lemlat's database for the verb STO 'to stay'

LES	CODLES
st	v1i
ist	v1i
stet	v7s
stat	n41
stat	n6p1
statūr[6]	n6p2

The LESs correspond to the various stems that are used in different sections of the paradigm: to obtain full wordforms, we only need to add the appropriate inflectional endings. In the design of Lemlat's database, the CODLES associated with each LES is informative on the "segmenti finali" ('final segments'; henceforth referred to as SF), roughly corresponding to different inflectional endings that can be added to the LES.

If the CODLES consists of "v" (short for 'verb') in the first slot followed by a digit from 1 to 5 in the second slot, this means that the involved LES corresponds to the so called "Present Stem" and can thus be used to generate imperfective wordforms of the "Present System",[7] by adding the SFs corresponding to the various endings used in the different conjugations relevant in the Present System, identified by the digit in the CODLES, with 1-4 standing for the four conjugations in the traditional order, and 5 for the heteroclitic mixed conjugation. Furthermore, the third slot of the CODLES provides additional information that is useful to identify verbs for which some paradigm cells should not be filled by a wordform, but should rather be marked as defective (#DEF#): deponent verbs (signalled by a "d" in the third slot of the CODLES), are defective of morphologically active wordforms; conversely, intransitive verbs ("i") lack passive wordforms, except for third person singular wordforms, that admit an impersonal usage (e.g. PRS.ACT.IND.3SG *stātur* 'one stays'; cf. also Table 4 below); lastly, for impersonal verbs ("e") only third-person singular wordforms are generated, alongside with some nominal forms like the infinitive and gerunds.

[6] It should be noticed that in the forms of Lemlat's database vowel length is never marked: this information has been added by using lexicographical sources (cf. §3.5 below).
[7] For a more detailed account of the Latin verbal system, cf. Chapter 4 below.

On the other hand, if the "v" in the first slot of the CODLES is followed by a "7" in the second one, the LES corresponds to the Perfect Stem on which the perfective wordforms of the Perfect System are built. Since the Latin conjugation system is only strictly relevant in the Present System, the same endings can be used for all verbs in the Perfect System.

The remaining LESs of Table 3 all have a CODLES beginning with "n", short for 'nominal': this is because they roughly correspond to the morphomic stem that Aronoff (1994) labels as the "Third Stem", used in a series of nominal forms that are not unitary from a semantic point of view (see above, §1.3.2). In particular, the LESs with CODLES "n41" can be used to generate supine forms like *statum*; the ones with CODLES "n6p1" for the perfect participle *status, -a, -um*; the ones with CODLES n6p2 for the future participle *statūrus, -a, -um*.

It should be noted that in Table 3 there are two LESs ("st" and "ist") with the same CODLES ("v1i"), corresponding to two possible variants of the same stem – the Present Stem. In such cases, only one variant is used to generate the wordforms. Our choice is based on lexicographical sources: for instance, in this case the verb STO has its own entry in the main Latin dictionaries, while *istō* is only given as an alternative form. More generally, in cases like this one, only the LESs corresponding to the stems appearing in the principal parts of the relevant entry are kept.[8]

Table 4 summarizes the way in which the content of some of the paradigm cells contained in LatInfLexi for the verb STO can be inferred from the LESs, CODLESs and SFs listed in Lemlat. Table 5 and Table 6 show how the situation concerning imperfective cells built on the Present Stem changes for different verbs, whose LESs are marked by different CODLESs.

[8] For further details on the selection of LESs in similar cases, cf. below, §3.4.

Table 4: The content of some paradigm cells of STO 'to stay' in LatInfLexi

LES	CODLES	compatible SFs	content of the cell
st	v1i (intransitive 1st conj. verb)	PRS.ACT.IND.1SG: -ō	*stō*
		PRS.ACT.IND.3SG: -at	*stat*
		PRS.PASS.IND.1SG: (none)	#DEF#
		PRS.PASS.IND.3SG: -ātur	*stātur*
		…	…
stet	v7s	PRF.ACT.IND.1SG: -ī	*stetī*
		PRF.ACT.IND.3SG: -it	*stetit*
		…	…
stat	n41	SUP.ACC: -um	*statum*
		SUP.ABL: -ū	*statū*
stat	n6p1	-us	*status*
		-a	*stata*
		-um	*statum*
		…	…
statūr	n6p2	-us	*statūrus*
		-a	*statūra*
		-um	*statūrum*
		…	…

Table 5: The content of some imperfective paradigm cells of MONEO 'to warn' in LatInfLexi

LES	CODLES	compatible SFs	content of the cell
mon	v2r (transitive 2nd conj. verb)	PRS.ACT.IND.1SG: -eō	*moneō*
		PRS.ACT.IND.3SG: -et	*monet*
		PRS.PASS.IND.1SG: -eor	*moneor*
		PRS.PASS.IND.3SG: -ētur	*monētur*
		…	…

Table 6: The content of some imperfective paradigm cells of NASCOR 'to be born' in LatInfLexi

LES	CODLES	compatible SFs	content of the cell
nasc	v3d (deponent 3rd conj. verb)	PRS.ACT.IND.1SG: (none)	#DEF#
		PRS.ACT.IND.3SG: (none)	#DEF#
		PRS.PASS.IND.1SG: -or	*nascor*
		PRS.PASS.IND.3SG: -itur	*nascitur*
		…	…

There are also other, rarer CODLESs that appear in Lemlat in addition to the ones mentioned above, concerning verbs displaying some kind of irregularity. For instance, to account for the inflected wordforms of the verb EO 'to go', in Lemlat's database we find a LES 'i' marked with the CODLES 'v6ic', meaning that it can be

used to generate the wordforms of the future indicative (e.g. *ibō* FUT.ACT.IND.1SG, *ibis* FUT.ACT.IND 2SG, etc.). Furthermore, there are full inflected wordforms that are simply reported as such in Lemlat's database, due to their irregularity: they are marked by the codles FE, short for it. "forma eccezionale" 'exceptional form' (e.g. *eō* PRS.ACT.IND.1SG, *eunt* PRS.ACT.IND.3PL). LESs marked by such CODLESs have been discarded. As a consequence, for a few, highly irregular verbs, the wordforms had to be manually generated as such, without following the procedure described above. These verbs are the following ones: 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', and verbs that derive from them (e.g. ABEO 'to go away' from EO).

As a result, we obtain a 254-cell paradigm for each verbal lexeme in LatInfLexi. This figure includes all participles, in their different (nominally) inflected wordforms. On the other hand, passive perfective cells are not included in our resource, since they are always filled periphrastically, by means of the perfect participle of the involved verb, followed by the appropriately inflected form of the verb 'to be': e.g. *amātus sum* 'I was loved' (AMO, PRF.PASS.IND.1SG). The purpose of our analysis is evaluating the uncertainty in guessing one form knowing another one: in such cases, this task concerns the two elements of the periphrasis individually, rather than the whole construction, which is completely predictable. Since the two elements are already contained in LatInfLexi (in the perfect participle cells of the corresponding lexeme, and in the corresponding cell of the lexeme SUM, respectively), it seems reasonable to simply exclude such systematically periphrastic cells from our data.

3.3.2 Noun paradigms

The procedure that was followed to obtain an inflected lexicon of nouns is similar to the one used for verbs in that it exploits information inferable from the LESs and CODLESs of each noun, but it is slightly different because unpredictable stem allomorphy in some cells had to be inferred from another source, namely the

information on how to obtain the lemma starting from a LES. The details of the procedure are provided in this section.

Latin nouns are traditionally described as belonging to one of five major inflection classes, therefore displaying different endings. Given the organization of Lemlat's database, a single LES is sufficient to generate the full paradigm of a noun, differently than what happened for verbs, where different LESs are used in different sections of the paradigm, as shown above in §3.2.1. Therefore, we only need to select the main LES that we want to keep in the (many) cases where more than one LES is listed in Lemlat's database for a given noun. As happened for verbs, the selection is based on lexicographical sources (see above, §3.3.1, and below, §3.4, for further details).

Regarding (most) nouns belonging to the 1st, 2nd, 4th and 5th declension, the procedure to obtain full paradigms from the selected LESs is rather straightforward: it suffices to add to each LES the compatible SFs, i.e. the endings of the declension indicated by the corresponding CODLES. One example for each declension is given in Tables 7-10.

Table 7: The content of the paradigm cells of ROSA 'rose' in LatInfLexi

LES	CODLES	compatible SFs	content of the cell
ros	n1 (1st decl. noun)	NOM.SG: -a	*rosa*
		GEN.SG: -ae	*rosae*
		DAT.SG: -ae	*rosae*
		ACC.SG: -am	*rosam*
		VOC.SG: -a	*rosa*
		ABL.SG: -ā	*rosā*
		NOM.PL: -ae	*rosae*
		GEN.PL: -ārum	*rosārum*
		DAT.PL: -īs	*rosīs*
		ACC.PL: -ās	*rosās*
		VOC.PL: -ae	*rosae*
		ABL.PL: -īs	*rosīs*

Table 8: The content of the paradigm cells of FOCUS 'hearth' in LatInfLexi

LES	CODLES	compatible SFS	content of the cell
foc	n2	NOM.SG: -us	*focus*
	(2nd decl. noun)	GEN.SG: -ī	*focī*
		DAT.SG: -ō	*focō*
		ACC.SG: -um	*focum*
		VOC.SG: -e	*foce*
		ABL.SG: -ō	*focō*
		NOM.PL: - ī	*focī*
		GEN.PL: -ōrum	*focōrum*
		DAT.PL: -īs	*focīs*
		ACC.PL: -ōs	*focōs*
		VOC.PL: -ī	*focī*
		ABL.PL: -īs	*focīs*

Table 9: The content of the paradigm cells of CANTUS 'singing' in LatInfLexi

LES	CODLES	compatible SFS	content of the cell
cant	n4	NOM.SG: -us	*cantus*
	(4th decl. noun)	GEN.SG: -ūs	*cantūs*
		DAT.SG: -uī	*cantui*
		ACC.SG: -um	*cantum*
		VOC.SG: -us	*cantus*
		ABL.SG: -ū	*cantū*
		NOM.PL: -ūs	*cantūs*
		GEN.PL: -uum	*cantuum*
		DAT.PL: -ibus	*cantibus*
		ACC.PL: -ūs	*cantūs*
		VOC.PL: -ūs	*cantūs*
		ABL.PL: -ibus	*cantibus*

Table 10: The content of the paradigm cells of RES 'thing' in LatInfLexi

LES	CODLES	compatible SFS	content of the cell
r	n5	NOM.SG: -ēs	*rēs*
	(5th decl. noun)	GEN.SG: -eī	*reī*
		DAT.SG: -eī	*reī*
		ACC.SG: -em	*rem*
		VOC.SG: -ēs	*rēs*
		ABL.SG: -ē	*rē*
		NOM.PL: -ēs	*rēs*
		GEN.PL: -ērum	*rērum*
		DAT.PL: -ēbus	*rēbus*
		ACC.PL: -ēs	*rēs*
		VOC.PL: -ēs	*rēs*
		ABL.PL: -ēbus	*rēbus*

There are other subclasses of nouns that display different endings in some cells – notably, neuter nouns of the 2nd and 4th declension, with endings *-um* and *-ū*, respectively, in NOM.SG, ACC.SG and VOC.SG, and *-a* and *-ua*, respectively, in NOM.PL, ACC.PL and VOC.PL – but their identification, and consequently the generation of their wordforms, is always deducible from the CODLES. However, in many 3rd declension nouns and in a specific subclass of 2nd declension nouns the situation is not that simple, since the content of the paradigm cells NOM.SG, VOC.SG and sometimes ACC.SG is not completely predictable from the LES and CODLES alone, due to stem allomorphy in those cells. However, in Lemlat's "lessario" information on how to obtain the lemma corresponding to a given LES is provided under the column LEM, either in the form of a specific ending to be added to that LES or as a full inflected wordform. Since the citation form used as lemma in Lemlat (and more generally in the literature on Latin) is exactly the nominative singular, and the other cells mentioned above are always syncretic with the nominative singular when they are unpredictable, this information can be exploited to obtain the missing wordforms in such cases. Two relevant examples – one of them referring to the 2nd declension lemma APER 'boar', with les "apr", the other one to the 3rd declension lemma AGMEN 'train (of people)', with les "agmin" – are given in Table 11 and Table 12.

Table 11: The content of some paradigm cells of APER 'boar' in LatInfLexi

LES	CODLES	LEM	compatible SFs	content of the cell
apr	n2	aper	NOM.SG: (none – see LEM)	*aper*
			GEN.SG: -ī	*aprī*
			…	…

Table 12: The content of some paradigm cells of AGMEN 'train (of people)' in LatInfLexi

LES	CODLES	LEM	compatible SFs	content of the cell
agmin	n3n1	agmen	NOM.SG: (none – see LEM)	*agmen*
			GEN.SG: -is	*agminis*
			…	…

Another issue concerning the 3rd declension is the fact that this class consists of several subclasses, displaying different endings in some cells. In some cases, the

choice of the appropriate endings can be made on the basis of the CODLES given in Lemlat's database.[9] However, there are also cases in which the CODLES identifies nouns that can take two different endings. In such cases, as usual, only one wordform is generated and listed in LatInfLexi. This brings us to the issue that will be tackled in the following section, namely the treatment of cases of overabundance in our resource.

3.4 The treatment of overabundance

In LatInfLexi every paradigm cell is filled by a single form. Within the framework of Canonical Typology (cf. above, §1.3.1), this is considered to be the canonical situation inside morphological paradigms, on the basis of the so-called principle of "uniqueness of realization" or "univocality" (cf. Thornton 2011, 2019). Nevertheless, cases of overabundance – defined as the availability of more than one wordform in a given paradigm cell – are well documented, and they are, of course, present also in Latin.

The exclusion of competing forms from our lexicon is due to the primary purpose of the resource, that is to quantify uncertainty in predicting one wordform from another one, by applying the scripts of the Qumin toolkit. A limitation of this toolkit is that it cannot take as input more than one wordform for the same paradigm cell when computing entropy: overabundant cells would simply be dropped. Thus, it seemed more reasonable to select only one "cell-mate" (Thornton 2011: 360) in such cases. In this section, we will justify our choices in this respect. Rather than a complete account of overabundance in Latin, which goes beyond the purpose of this work, we will provide some examples to illustrate the principles underlying our decisions in cases where the database of Lemlat would have allowed to generate more than one wordform in a given paradigm cell. Two main typologies will be individuated and discussed in the following sub-sections: cases where overabundance would arise because more than one LES can be used to obtain the content of at least some paradigm cells of a lexeme (§3.4.1), and cases where this

[9] For a complete list of CODLESs and their correspondence with different inflection classes and subclasses, the reader is referred to the documentation of Lemlat, available at `http://www.lemlat3.eu/download/documentation/`.

would be due to the presence of more than one SF compatible with the same LES (§3.4.2).

3.4.1 Overabundance due to the presence of more than one LES

We saw above that the procedure to fill a paradigm cell of a lexeme requires that an ending is added to the LES corresponding to the stem used in that paradigm cell for that lexeme. Although in the simplest situation only one LES can be used to infer the content of a given paradigm cell, there are (groups of) cells for which more than one LES is available in Lemlat's database. The fact that Lemlat is designed as a tool to analyse forms, rather than to produce them, makes this problem severe, since in many cases even very marginal variants are reported as a LES, exactly like regular forms.

Given the design of Lemlat, for nouns the presence of more than one LES would generate a systematic overabundance over the whole set of paradigm cells of the involved lexeme, since from a given LES all the inflected wordforms of the lexeme can be generated (cf. §3.3.2 above).

For instance, for the noun AQUA 'water', two LESs with the same CODLES "n1" are listed in Lemlat, namely "aqu" and "acu": if both of them were used to generate wordforms, we would get cells containing two wordforms in the whole paradigm of the lexeme,[10] as shown in Table 13.

[10] In this particular case, we are probably dealing with graphical variation, rather than true overabundance. However, in Latin there are also cases where the different LESs are also different in their phonetic shape, as is witnessed by the examples in Table 14 and Table 15.

Table 13: Potential overabundance in the lexeme AQUA 'water'

LES	CODLES	compatible SFs	content of the cell	
aqu	n1	NOM.SG: -a	*aqua*	*acua*
acu	(1st decl.	GEN.SG: -ae	*aquae*	*acuae*
	noun)	DAT.SG: -ae	*aquae*	*acuae*
		ACC.SG: -am	*aquam*	*acuam*
		VOC.SG: -a	*aqua*	*acua*
		ABL.SG: -ā	*aquā*	*acuā*
		NOM.PL: -ae	*aquae*	*acuae*
		GEN.PL: -ārum	*aquārum*	*acuārum*
		DAT.PL: -īs	*aquīs*	*acuīs*
		ACC.PL: -ās	*aquās*	*acuās*
		VOC.PL: -ae	*aquae*	*acuae*
		ABL.PL: -īs	*aquīs*	*acuīs*

Regarding verbs, since – as we saw above in §3.3.1 – different LESs with different CODLESs have to be used in different sections of the paradigm, the availability of more than one LES would produce overabundance only in the involved sub-paradigm. For instance, in the verb ERIGO 'to erect', two LESs with CODLES "v7s" are listed, namely "ērēg" and "ērēx": if both were used to generate wordforms, there would be overabundance in the perfective forms built on this LES, but not in all the other cells, where only one LES is available.

Table 14: Potential overabundance in the lexeme ERIGO 'to erect'

les	codles	compatible SFs	content of the cell	
ērig	v3r	PRS.ACT.IND.1SG: -ō	*ērigō*	
		PRS.ACT.IND.3SG: -it	*ērigit*	
		…	…	
ērēg	v7s	PRF.ACT.IND.1SG: -ī	*ērēgī*	*erēxī*
ērēx		PRF.ACT.IND.3SG: -it	*ērēgit*	*erēxit*
		…	…	…
ērēct	n41	SUP.ACC: -um	*ērēctum*	
		SUP.ABL: -ū	*ērēctū*	

On the other hand, in TORQUEO 'to twist' there would be a different pattern of overabundance, only concerning the supine forms built on the LES with CODLES "n41"

Table 15: Potential overabundance in the lexeme TORQUEO 'to twist'

les	codles	compatible SFs	content of the cell	
torqu	v2r	PRS.ACT.IND.1SG: -eō	*torqueō*	
		PRS.ACT.IND.3SG: -et	*torquet*	
		…	…	
tors	v7s	PRF.ACT.IND.1SG: -ī	*torsī*	
		PRF.ACT.IND.3SG: -it	*torsit*	
		…	…	
tors	n41	SUP.ACC: -um	*torsum*	*tortum*
tort		SUP.ABL: -ū	*torsū*	*tortū*

Lastly, in ABNUO 'to refuse' more than one LES can be used to generate imperfective wordforms of the present system. This case is also different from the previous ones in that rather than segmentally different LESs we have the same sequence of characters listed twice, with different CODLESs implying different inflection class assignments (3rd vs. 2nd conjugation verb), and therefore there are different endings available for the same set of cells, rather than stem allomorphy.

Table 16: Potential overabundance in the lexeme ABNUO 'to refuse'

les	codles	compatible SFs	content of the cell	
abnu	v3r	PRS.ACT.IND.1SG: -ō, -eō	*abnuō*	*abnueō*
abnu	v2r	PRS.ACT.IND.3SG: -it, -et	*abnuit*	*abnuet*
		…	...	
abnu	v7s	PRF.ACT.IND.1SG: -ī	*abnuī*	
		PRF.ACT.IND.3SG: -it	*abnuit*	
		…	…	

It should be stressed that there is no way to select the LES that is more reasonable to keep in a principled way that can be applied systematically to the whole lexicon by using only information reported in Lemlat's database itself: while in some cases there is one LES that is given a somewhat prominent status in that it is the one on which the citation form of the lemma is built, this would not hold for the example in Table 14, where there is no way to know which of the two LESs with CODLES "v7s" can be considered as more marginal than the other.

Therefore, external criteria have to be invoked to choose what wordform should be generated in such cases. Ideally, the criterion would be the relative frequency of the different variants: of course, one would like to keep the most frequently used

variant. However, there are serious practical issues with this solution: to implement it, we would need very detailed frequency data, equipped not only with lemma information, but also with disambiguation of the morphosyntactic property set expressed by the wordforms, in order to be able to know what paradigm cells are involved; and we would need a very big amount of such data, at least for verbs, given the intrinsically sparse nature of large morphological paradigms. These data are not easily available, since only small corpora reach the desired level of granularity in annotation.

Our choice was thus to use information taken from lexicographical sources. In Latin dictionaries, any lexical entry begins with a set of principal parts. For nouns, there are two of them, namely the wordforms filling the cells NOM.SG and GEN.SG. Regarding verbs, there is some variation concerning the exact number (between three and five) of cells used as principal parts, but at least one cell for each of the sections of the paradigm described above (Present System, Perfect System, and nominal forms built on the third stem) is always reported. This allows to exploit such information in order to decide what variant should be used to generate paradigms. For instance, for the noun AQUA, in dictionaries like Glare (2012) and Lewis and Short (1978) the principal parts are the inflected wordforms *aqua*, NOM.SG and *aquae*, GEN.SG. The graphical variant of the LES with <c> is listed in Lemlat because it is reported in Glare (2012) in another section of the entry, as a variant sometimes attested in inscriptions. Therefore, it is reasonable to keep only the form written with <q>, and to discard the other one as marginal. Similarly, the principal parts used by the aforementioned dictionaries of Latin for the verbs ERIGO and TORQUEO are based on the stem alternants *erex-* and *tors-*, respectively: thus, only the corresponding LESs are used to generate the inflected wordforms of these verbs. Lastly, the citation form *abnuō* (PRS.ACT.IND.1SG) is unanimously attested in such dictionaries, and it implies that 3rd conjugation forms should be kept, rather than 2nd conjugation forms (PRS.ACT.IND.1SG *abnueō*), that again appear only in other sections of the entry as marginal variants.

Our main lexicographical source is Lewis & Short (1879), because its easy availability in machine readable format allows for a semi-automatic extraction of

the relevant information.[11] However, in some cases we had to rely on other dictionaries too: for instance, regarding the third stem of the verb INNITOR 'to lean upon', Lewis and Short (1879) mention both *innīx-us* and *innīs-us* as principal parts: our choice to keep only the first one is therefore based on the principal parts reported in other dictionaries, namely Georges & Georges (1913-18) – where only *innīxus* is given as principal part – and Glare (2012) – where *innīsus* is present, but only in brackets, as a more marginal form.

3.4.2 Overabundance due to compatibility of a LES with more than one SF

Another fact that could potentially produce overabundance in LatInfLexi given the organization of Lemlat's database is the fact that in some cells there is more than one SF that is compatible with the LES(s) that should be used in that cell. In this section different examples of this kind will be discussed to exemplify the varying levels of systematicity across lexemes of these cases of overabundance.
In some cases, overabundance is completely systematic across lexemes, potentially occurring in all of them. A very clear example is given by the different endings available for the second-person singular of passive verbal forms, that can always end in *-ris* or in *-re*. Examples are given in Table 17.

Table 17: Potential overabundance in second-person singular passive forms: the lexeme AMO 'to love'

les	codles	compatible SFs	content of the cell	
am	v1r	PRS.PASS.IND.2SG: -āris, -āre	*amāris*	*amāre*
		PRS.PASS.SBJV.2SG: -ēris, -ēre	*amēris*	*amēre*
		IPFV.PASS.IND.2SG: -ābāris, -ābare	*amābāris*	*amābāre*
		IPFV.PASS.SBJV.2SG: -ārēris, -ārēre	*amārēris*	*amārēre*
		…	…	

In this case, we have simply decided to keep the variant ending in *-ris*: given the full systematicity of this kind of overabundance, the alternative forms in *-re* can be

[11] Since we are only considering frequent lexemes that are attested in all the main Latin dictionaries, the choice of a lexicon different than the ones on which Lemlat is based does not create problems of compatibility – namely, lexemes that are kept in LatInfLexi, but are not attested in Lewis & Short (1879).

considered as trivially predictable with no uncertainty, and therefore adding them would not substantially influence the results that will be presented in the next chapters.

At the other end of the scale of systematicity across lexemes, there are cases of overabundance that only concern a few lexemes. One extreme example is given by the ending *-ās* for the cell GEN.SG in the 1st declension, alongside the regular *-ae*: this ending is only attested for the lexeme FAMILIA 'family' in classical Latin. The occurrence of this ending is restricted not only in terms of number of lexemes, but also in terms of syntactic contexts – it only appears in frozen expressions like *pater familiās* and *mater familiās* ('family man/woman'). In this case, and in similar cases where the alternative ending is comparably marginal, the more reasonable choice is to exclude it altogether, and only generate the regular form – in this case, the GEN.SG *familiae*.

Again, the problem is amplified by the organization of Lemlat's database, where in some cases the actual impact of the competition between different endings turns out to be overestimated. For instance, the dative and ablative plural SFs "-ibus" and "-ubus" are both marked as compatible with LESs with CODLES "n4", meaning that both endings could potentially be used to generate the wordforms in the cells DAT.PL and ABL.PL of any 4th declension noun. However, the ending *-ubus* is actually attested as a variant only for a few lexemes according to Latin grammars (cf. e.g. Bennett 1908: §48): in all the other lexemes of the 4th declension, there is no overabundance at all in these cells, and only the ending *-ibus* is used. As was the case for competing stem alternants, the ideal way to choose what wordform should be generated would be a corpus-based comparison of the frequency of the competing endings for each lexeme, but the same problems described in the previous section would arise. Since in this case dictionaries are not helpful, we used another source of information on what ending is more marginal for a given lexeme, and thus should not be used to generate inflected wordforms. For this purpose, we exploited the wordform generator provided by the Collatinus toolkit (`https://outils.biblissima.fr/fr/collatinus-web/`): the form that is reported in there is the one that we generate in our resource. Therefore, for TRIBUS 'tribe' and for a few other nouns LatInfLexi only contains the irregular DAT./ABL.PL *tribubus*

and the like, while for the rest of 4th declension nouns only the regular form in *-ibus* is reported.

A similar problem arises in a more systematic fashion for 3rd declension nouns. What is traditionally called the 3rd declension can actually be divided in many different sub-classes, all displaying the same set of endings in most cells, but different endings in ABL.SG, ACC.SG, GEN..PL and ACC.PL.[12] On historical grounds, one can distinguish nouns with *-i-* stems, inflected as in Table 18, and consonant stems, inflected as in Table 19; additionally, a series of "mixed classes" can be individuated, displaying the endings of *-i-* stems in some cells and the ones of consonant stems in other cells, as described by Wurzel (1984).

Table 18: some paradigm cells of the *-i-* stem noun PUPPIS 'stern (of a ship)'

cell	wordform
ACC.SG	*puppim*
ABL.SG	*puppī*
GEN.PL	*puppium*
ACC.PL	*puppīs*

Table 19: some paradigm cells of the consonant stem noun REX 'king'

cell	wordform
ACC.SG	*regem*
ABL.SG	*rege*
GEN.PL	*regum*
ACC.PL	*regēs*

Sometimes the endings that should be selected for a given cell of a 3rd declension noun can be inferred from the CODLES reported in Lemlat's database: for instance, the CODLES "n31" identifies 3rd declension lexemes that take *-um* in GEN.PL, and conversely the CODLES "n32" is used for lexemes that take *-ium* in that cell. But if the CODLES is just "n3", no information is provided on the GEN.PL ending that should be used. Therefore, also in such cases our choice is based on the wordform reported in Collatinus.

3.5 Phonetic transcriptions

The ultimate purpose for which LatInfLexi was built is the application of the Qumin toolkit in order to assess the uncertainty in the PCFP in Latin paradigms by means of implicative entropy. As we saw above in §2.3, the scripts of this toolkit are

[12] See Chapter 5 for further details on Latin nominal inflection.

designed to capture phonological restrictions on the context of application of the various morphological patterns. This is a desirable feature, but, as a consequence, the input data need to be coded in phonetic transcription, rather than in Latin orthography. Additionally, the phonological features relevant for each segment are required to be provided to the toolkit. The segments and features used in this work are the ones assumed by Cser (2016), listed in Table 20 below, with some minor differences. In this section, I will discuss both theoretical and practical issues related to the use of phonetic transcriptions of Latin wordforms.

Table 20: Segments and features used for Latin data[13]

segment	consonantal	sonorant	approximant	voice	spread glottis	continuant	nasal	lateral	labial	labiodental	coronal	dorsal	high	low	front	back	long	round
b <b>	+	-	-	+	-	-	-	-	+	-	-	-	-	-		-		
d <d>	+	-	-	+	-	-	-	-	-	-	+	-	-	-		-		
g <g>	+	-	-	+	-	-	-	-	-	-	-	+	+	-		+		
m <m>	+	+	-	+	-	-	+	-	+	-	-	-	-	-		-		
n <n>	+	+	-	+	-	-	+	-	-	-	+	-	-	-		-		
l <l>	+	+	+	+	-	+	-	+	-	-	+	-	-	-		-		
r <r>	+	+	+	+	-	+	-	-	-	-	+	-	-	-		-		
p <p>	+	-	-	-	-	-	-	-	+	-	-	-	-	-		-		
p^h <ph>	+	-	-	-	+	-	-	-	+	-	-	-	-	-		-		
f <f>	+	-	-	-	-	+	-	-	+	+	-	-	-	-		-		
t <t>	+	-	-	-	-	-	-	-	-	-	+	-	-	-		-		
t^h <th>	+	-	-	-	+	-	-	-	-	-	+	-	-	-		-		
s <s>	+	-	-	-	-	+	-	-	-	-	+	-	-	-		-		
k <c>	+	-	-	-	-	-	-	-	-	-	-	+	+	-		+		
k^h <ch>	+	-	-	-	+	-	-	-	-	-	-	+	+	-		+		
h <h>	+	-	-	-	+	+	-	-	-	-	-	-		-		-		
j <i>/<j>	-	+	+	+	-	+	-	-	-	-	-	+	+	-		-		
w <u>/<v>	-	+	+	+	-	+	-	-	+	-	-	+	+	-		+		
a <a>	-	+	-	+	-	+	-	-	-	-		+	-	+	-	+	-	-
a: <ā>	-	+	-	+	-	+	-	-	-	-		+	-	+	-	+	+	-
e <e>	-	+	-	+	-	+	-	-	-	-		+	-	-	+	-	-	-
e: <ē>	-	+	-	+	-	+	-	-	-	-		+	-	-	+	-	+	-
i <i>	-	+	-	+	-	+	-	-	-	-		+	+	-	+	-	-	-
i: <ī>	-	+	-	+	-	+	-	-	-	-		+	+	-	+	-	+	-
o <o>	-	+	-	+	-	+	-	-	+	-		+	-	-	-	+	-	+
o: <ō>	-	+	-	+	-	+	-	-	+	-		+	-	-	-	+	+	+
u <u>	-	+	-	+	-	+	-	-	+	-		+	+	-	-	+	-	+
u: <ū>	-	+	-	+	-	+	-	-	+	-		+	+	-	-	+	+	+
y <y>	-	+	-	+	-	+	-	-	+	-		+	+	-	+	-	-	+

[13] For reference, we also provide the grapheme(s) that normally – although not always – correspond(s) to each of the sounds.

From a theoretical point of view, it is obvious that the epistemological status of the reconstructed phonetic transcriptions provided for a classical language like Latin is by far different if compared to modern languages like the ones to which the Qumin toolkit has been applied before: while modern languages are currently spoken and pronounced, of course there is no direct evidence on the pronunciation on Latin. However, many sources of indirect evidence have been mentioned in the literature on Latin phonetics and phonology. First of all, there are aspects that can be inferred from the phonology of other Indo-European languages, on the one hand, and of Romance languages, on the other hand. In some cases, ancient grammarians tell us explicitly how a particular letter should be pronounced. In other cases, indirect evidence on the pronunciation of a word can be provided by its use in puns or wordplays. The spelling of Latin words borrowed into other languages is another source of information on how the loanword was pronounced in Latin at the time when the borrowing occurred. All this evidence has indeed been used to obtain a reconstruction of the phonetics of Classical Latin on which there is a reasonable consensus, from Allen (1965) up to McCullagh (2011), although of course there remain some items on whose exact phonetic nature there are doubts.

Another interesting theoretical issue concerns the level of phonetic detail of the transcriptions. In principle, even completely regular and exceptionless phonological processes can produce inflectional opacity and, consequently, unpredictability in the PCFP. A clear example is provided by Bonami et al. (2011): we repeat it here in Table 21.

Table 21: The cells IPFV.1SG and IPFV.1PL of some French verbs

lexeme	IPFV.1SG	IPFV.1PL
LAVER 'to wash'	[lavɛ]	[lavjõ]
CONTRER 'to counter'	[kõtʁɛ]	[kõtʁijõ]
QUADRILLER 'to divide into squares'	[kadʁijɛ]	[kadʁijõ]

From a phonological perspective, it is reasonable to consider /jõ/ as the underlying form of the suffix expressing IPFV.1PL. On the one hand, there are verbs whose stem ends in [j]: in such cases, the [j] of the suffix is always cancelled, according to a phonological rule that can be expressed as in (1), yielding for instance a form like

[kadʁijõ]. On the other hand, there are verbs whose stem does not end with [j]. In such cases, the IPFV.1PL suffix is always /jõ/ underlyingly (see e.g. the form [lavjõ]), but a sequence [ijõ] emerges after a branching onset, according to rule (2), yielding a form like [kõtʁijõ].

(1) j → ø / j_

(2) j → ij / BranchingOnset

Even if the different allomorphs are due to completely regular phonological processes, inflectional opacity nevertheless arises: faced with an IPFV.1PL in -[ijõ] preceded by a branching onset, a speaker has no way to know if the IPFV.1SG should be in -[ijɛ], with the [j] being part of the stem of the verb, as in [kadʁijɛ] from [kadʁijõ], or if it should be in -[ɛ], with the [j] belonging to the suffix, as in [kõtʁɛ] from [kõtʁijõ].

Therefore, Bonami et al. (2011) argue that the transcription format of the data should be as surface-true as possible. However, dealing with an historical language this requirement sometimes conflicts with the level of phonetic detail that it is possible to achieve without excessive speculation. Concretely, we opted for a lax phonetic transcription where some (but not all) of the phonological processes are abstracted away, depending on their impact on inflectional predictability.[14]

Let us take the form *absum* (the PRS.ACT.IND.1SG of the verb ABSUM 'to be away from') as an example. The sequence <bs> is transcribed as [ps], as stated in the reconstructions of Latin pronunciation, although the presence of [p] is the output of a regular phonological process of assimilation of an underlying /b/ before voiceless obstruents. This is because this process generates inflectional opacity in the same way as the French example of Table 21: faced with a form containing [ps], a speaker has no way of knowing if the [p] is produced by the aforementioned rule of assimilation, or if it simply corresponds to an underlying /p/. Conversely, <um> is transcribed as [um] even if it is more likely that the sequence of phonemes /um/ was

[14] See also Maiden (1995) for other arguments supporting the use of phonetic rather than phonological transcriptions in Romance linguistics, in a view where phonetics and phonology are not strictly separated.

phonetically realized as a long, nasalized vowel [ũ], since this difference would not have any impact on the interpredictability of wordforms.

Moving on to more practical issues, phonetic transcriptions of Latin inflected wordforms can be obtained automatically from the orthographic notation, as long as some distinctions that are optional in Latin orthography are taken care of. These details are i) the distinction between long and short vowels, optionally marked by the macron (<ā>, <ē>, <ī>, <ō>, <ū>), and ii) the distinction between [j] and [i] on the one hand and [u] and [w] on the other one, where the vowels are always written as <i> and <u>, respectively, and the semi-vowels can (but do not have to) be written as <j> and <v>. Since in Lemlat's database none of these distinctions is made, we had to add it to our data from a source where they are consistently present. For this purpose, again we have used the dictionary by Lewis & Short (1879), because of its easy accessibility, by projecting such distinctions from the principal parts of that lexicon, where they are always marked, to the LESs used to generate the inflected wordforms of LatInfLexi. Sometimes, however, we have also taken into account other lexicographical sources, and namely Georges and Georges (1913-1918), whose choices sometimes appear to be closer to the phonetic reality of Latin, especially concerning vowel length, where Lewis and Short (1879) sometimes disregard the effect of some regular phonological processes. For instance, Lewis & Short (1879) do not take into account the outcome of the phonological rule sometimes referred to as *ns* lengthening (cf. Weiss 2009: 129), according to which all vowels followed by the sequence of /n/ and a fricative are lengthened. Therefore, in the verb PENSO 'to weigh (out carefully)' the vowel is not marked as long: in this and similar cases, we prefer to follow Georges & Georges (1913-1918), whose choice of marking the vowel as long is clearly closer to the phonetic reality of the wordforms, as is required by our surface-based approach.

3.6 Conclusion

In this section, we have presented LatInfLexi, an inflected lexicon of Latin verbs and nouns organized in lexemes and paradigm cells, detailing the procedure that was followed in order to create the resource. In Table 22, we provide some details

on the size of the resource in terms of number of lexemes, paradigm cells and wordforms.

Table 22: The size of LatInfLexi

	verbs	nouns
n. lexemes	3,348	1,038
n. paradigm cells	850,392	12,456
n. wordforms	752,537	12,355

Having introduced the theoretical framework of this work in Chapter 1, the method that has been used in Chapter 2 and the data on which this method was applied in Chapter 3, we can now move on to the results that have been obtained accordingly, regarding verb inflection in Chapter 4 and noun inflection in Chapter 5.

Chapter 4. Predictability and paradigm organization in Latin verb inflection

This chapter will be devoted to Latin verb inflection. As a starting point, it is necessary to provide some preliminary information on the verbal system, as it is outlined in traditional descriptions: we will do so in §4.1, where we will also review previous theoretically grounded studies on Latin inflectional morphology regarding verbs. We will then move to our analysis of implicative relations, which is performed not on the full paradigm of Latin verbs, but on a reduced version that abstracts away from all cases of systematic syncretism, called the "cell paradigm" following Boyé & Schalchli (2016): see §4.2 for a more detailed elaboration. Results on various fragments of the Latin paradigm will be presented in §4.3 and §4.4: in the former section, we will focus on the alternation patterns that hold between wordforms that are based on different stems, and consequently on the uncertainty in predicting the cells involved from one another; in the latter, we will look at the situation in wordforms that are based on the same stem. In §4.5, we will try to give an idea of the overall structure of the Latin verb paradigm, as it emerges from our entropy-based analysis. Firstly, we will draw a map of the paradigm in different zones that contain cells between which there is full mutual predictability. Secondly, we will compute entropy values on a so-called "distillation" (cf. Stump & Finkel 2013) of the paradigm, where we keep only one cell for each zone. Lastly, in §4.6 we will extend our investigation to predictions from more than one cell, whose uncertainty will be measured by means of *n*-ary implicative entropy: these results will also be exploited to extract principal part sets and near-principal part sets, and compare them to the ones of the traditional analysis and to the ones that have been found with different methodologies – notably, Stump & Finkel (2013)'s Principal Part Analysis.

4.1 Latin verb inflection: the traditional account and previous theoretical research

A considerable amount of work has been devoted to Latin verb inflection. The facts are relatively well known from traditional descriptions like the grammars by Bennett (1908) and Leumann et al. (1977), and from Ernout (1914)'s historical morphology. These traditional accounts constitute the main source of the brief synopsis provided in this section. Furthermore, there are many studies that deal with theoretical issues related to specific aspects of verb inflection, from the complete monograph devoted to it by Matthews (1972) up to the recent morphophonological account of allomorphy in Latin inflectional morphology provided by Cser (2015, 2016). In this section, we will review in some detail Aronoff (1994), that focuses on the stems on which the various inflected wordforms of Latin verbs are based, and on their distribution throughout the paradigm. We will then summarize Dressler (2002), that proposes a detailed hierarchy of the inflection classes of Latin verbs and nouns.[1] Lastly, a few words will be devoted to Finkel & Stump (2009b), that provide a more principled, automatic implementation of the traditional notion of principal parts, applied to Latin verbs. Other studies concerned with more specific aspects will be cited in different places in this chapter.

Let us start from the complete paradigm of the lexeme AMO 'to love', which is given for reference in Table 1.

[1] Of course, in this chapter we will focus on the inflection classes of verbs; the ones of nouns will be reviewed in the next chapter (see §5.1 below).

Table 1: The complete paradigm of AMO 'to love'

1a. – *infectum*: the present system

	indicative					
	present		imperfect		future	
	active	passive	active	passive	active	passive
1SG	*amō*	*amor*	*amābam*	*amābar*	*amābō*	*amābor*
2SG	*amās*	*amāris*	*amābās*	*amābāris*	*amābis*	*amāberis*
3SG	*amat*	*amātur*	*amābat*	*amābātur*	*amābit*	*amābitur*
1PL	*amāmus*	*amāmur*	*amābāmus*	*amābāmur*	*amābimus*	*amābimur*
2PL	*amātis*	*amāmini*	*amābātis*	*amābāmini*	*amābitis*	*amābimini*
3PL	*amant*	*amāntur*	*amābant*	*amābantur*	*amābunt*	*amābuntur*

	subjunctive			
	present		imperfect	
	active	passive	active	passive
1SG	*amem*	*amer*	*amārem*	*amārer*
2SG	*amēs*	*amēris*	*amārēs*	*amārēris*
3SG	*amet*	*amētur*	*amāret*	*amārētur*
1PL	*amēmus*	*amēmur*	*amārēmus*	*amārēmur*
2PL	*amētis*	*amēminī*	*amārētis*	*amārēminī*
3PL	*ament*	*amentur*	*amārent*	*amārentur*

	imperative			
	present		future	
	active	passive	active	passive
2SG	*amā*	*amāre*	*amātō*	*amātor*
3SG			*amātō*	*amātor*
2PL	*amāte*	*amāmini*	*amātōte*	
3PL			*amantō*	*amantōr*

	infinitive			
	present		future	
	active	passive	active	passive
	amāre	*amārī*	*amātūrus, -a, -um esse*	*amātum īrī*

	gerund	supine
GEN.SG	*amandī*	
DAT.SG	*amandō*	
ACC.SG	*amandum*	*amātum*
ABL.SG	*amandō*	*amātū*

	gerundive		
	masculine	feminine	neuter
NOM.SG	*amandus*	*amanda*	*amandum*
GEN.SG	*amandī*	*amandae*	*amandī*
DAT.SG	*amandō*	*amandae*	*amandō*
ACC.SG	*amandum*	*amandam*	*amandum*
VOC.SG	*amande*	*amanda*	*amandum*
ABL.SG	*amandō*	*amandā*	*amandō*
NOM.PL	*amandī*	*amandae*	*amanda*
GEN.PL	*amandōrum*	*amandārum*	*amandōrum*
DAT.PL	*amandīs*	*amandīs*	*amandīs*
ACC.PL	*amandōs*	*amandās*	*amanda*
VOC.PL	*amandī*	*amandae*	*amanda*
ABL.PL	*amandīs*	*amandīs*	*amandīs*

	participle					
	present			future		
	masculine	feminine	neuter	masculine	feminine	neuter
NOM.SG	*amāns*	*amāns*	*amāns*	*amātūrus*	*amātūra*	*amātūrum*
GEN.SG	*amantis*	*amantis*	*amantis*	*amātūrī*	*amātūrae*	*amātūrī*
DAT.SG	*amantī*	*amantī*	*amantī*	*amātūrō*	*amātūrō*	*amātūrō*
ACC.SG	*amantem*	*amantem*	*amāns*	*amātūrum*	*amātūram*	*amātūrum*
VOC.SG	*amāns*	*amāns*	*amāns*	*amātūre*	*amātūra*	*amātūrum*
ABL.SG	*amante*	*amante*	*amante*	*amātūrō*	*amātūrā*	*amātūrō*
NOM.PL	*amantēs*	*amantēs*	*amantia*	*amātūrī*	*amātūrae*	*amātūra*
GEN.PL	*amantium*	*amantium*	*amantium*	*amātūrōrum*	*amātūrārum*	*amātūrōrum*
DAT.PL	*amantibus*	*amantibus*	*amantibus*	*amātūrīs*	*amātūrīs*	*amātūrīs*
ACC.PL	*amantēs*	*amantēs*	*amantia*	*amātūrōs*	*amātūrās*	*amātūra*
VOC.PL	*amantēs*	*amantēs*	*amantia*	*amātūrī*	*amātūrae*	*amātūra*
ABL.PL	*amantibus*	*amantibus*	*amantibus*	*amātūrīs*	*amātūrīs*	*amātūrīs*

1b. – *perfectum:* the perfect system

	indicative					
	perfect		pluperfect		future perfect	
	active	passive	active	passive	active	passive
1SG	*amāvī*	*amātus sum*	*amāveram*	*amātus eram*	*amāverō*	*amātus ero*
2SG	*amāvistī*	*amātus es*	*amāverās*	*amātus erās*	*amāveris*	*amātus eris*
3SG	*amāvit*	*amātus est*	*amāverat*	*amātus erat*	*amāverit*	*amātus erit*
1PL	*amāvimus*	*amātī sumus*	*amāverāmus*	*amātī erāmus*	*amāverimus*	*amātī erimus*
2PL	*amāvistis*	*amātī estis*	*amāverātis*	*amātī erātis*	*amāveritis*	*amātī eritis*
3PL	*amāvērunt*	*amātī sunt*	*amāverant*	*amātī erant*	*amāverint*	*amātī erunt*

	subjunctive			
	perfect		pluperfect	
	active	passive	active	passive
1SG	*amāverim*	*amātus sim*	*amāvissem*	*amātus essem*
2SG	*amāverīs*	*amātus sīs*	*amāvissēs*	*amātus essēs*
3SG	*amāverit*	*amātus sit*	*amāvisset*	*amātus esset*
1PL	*amāverīmus*	*amātī sīmus*	*amāvissēmus*	*amātī essēmus*
2PL	*amāverītis*	*amātī sītis*	*amāvissētis,*	*amātī essētis*
3PL	*amāverint*	*amātī sint*	*amāvissent*	*amātī essent*

infinitive	
perfect	
active	passive
amāvisse	*amātus esse*

	participle		
	perfect		
	masculine	feminine	neuter
NOM.SG	*amātus*	*amāta*	*amātum*
GEN.SG	*amātī*	*amātae*	*amātī*
DAT.SG	*amātō*	*amātae*	*amātō*
ACC.SG	*amātum*	*amātam*	*amātum*
VOC.SG	*amāte*	*amāta*	*amātum*
ABL.SG	*amātō*	*amātā*	*amātō*
NOM.PL	*amātī*	*amātae*	*amāta*
GEN.PL	*amātōrum*	*amātārum*	*amātōrum*
DAT.PL	*amātīs*	*amātīs*	*amātīs*
ACC.PL	*amātōs*	*amātās*	*amāta*
VOC.PL	*amātī*	*amātae*	*amāta*
ABL.PL	*amātīs*	*amātīs*	*amātīs*

The organization of the material in Table 1 is based on an aspectual opposition. The cells of the so-called "**present system**" (cf. Ernout & Thomas 1951: 236) have an imperfective meaning: the action is viewed as *infectum* 'not accomplished'. Conversely, the cells of the "**perfect system**" are perfective: the action is viewed as *perfectum*, 'accomplished'.

Both in the *infectum* and in the *perfectum* aspect we find a three-way temporal opposition between present, future, and past: in the *infectum*, we have the present, imperfect (i.e., past imperfective) and future; in the *perfectum*, the perfect (i.e., present perfective), pluperfect (i.e., past perfective) and future perfect (i.e., future perfective). Finite forms are also inflected for the categories of mood (whose possible values are indicative, subjunctive and imperative), voice (active and passive), person (1st, 2nd and 3rd) and number (singular and plural). The inflectional categories relevant for finite forms and their respective possible values are summarized in Table 2, where also the abbreviations of the Leipzig Glossing Rules – that will be used throughout this work – are given.

Table 2: Inflectional categories and values of Latin finite verbal forms

category	value
tense/aspect	present (PRS), imperfect (IPRF), future (FUT), perfect (PRF), pluperfect (PLUPRF), future perfect (FUTPRF)
mood	indicative (IND), subjunctive (SBJV), imperative (IMP)
voice	active (ACT), passive (PASS)
person	1, 2, 3
number	singular (SG), plural (PL)

Additionally, there are verbal adjectives – namely, the present, perfect and future participle (PTCP) and the gerundive (GDV) – that are also inflected for case – nominative (NOM), genitive (GEN), dative (DAT), accusative (ACC), vocative (VOC), and ablative (ABL) – and gender – masculine (M), feminine (F) and neuter (N). Lastly, there are verbal nouns like the present, perfect and future infinitive (INF), the gerund (GER) and the supine (SUP).[2]

[2] On our choice of distinguishing between accusative and ablative rather than active and passive supine, cf. §1.3.1, Footnote 6.

It should be observed that in many cells there is not a synthetic wordform: all the passive forms of the perfect system are filled by a periphrase composed of the perfect participle of the involved verb, displaying agreement in gender and number with the subject, followed by an appropriately inflected wordform of the verb SUM 'to be', e.g. *(puer) amātus est* '(the kid) has been loved'. The future active infinitive displays a similar construction, but with the future participle instead of the perfect participle (*amātūrus esse* 'to be going to love'), while the future passive infinitive is composed by the supine accusative of the involved verb followed by the impersonal passive infinitive of the verb EO 'to go' (*amātum īrī* 'to be going to be loved').

To the already mentioned semantic opposition between *infectum* and *perfectum* corresponds also a formal opposition based on the stems that appear in different inflected wordforms: forms of the present system contain the **present stem** *am-* (cf. e.g. PRS.ACT.IND.1SG *amō*, 2SG *amās*), forms of the perfect system contain the **perfect stem** *amāv-* (cf. e.g. PRF.ACT.IND.1SG *amāvī*, 2SG *amāvistī*). However, it was noted by Aronoff (1994: 54 ff.), and it can also be observed from the data in Table 1, that this correspondence is far from perfect: both in the present system and in the perfect system there are wordforms that are based on a different stem, *amāt-* in our example. Traditional descriptions of Latin call this stem in different ways: for instance, Bennett (1908: §95) refers to it as the "Participial Stem", while Ernout (1914: §363) defines it as the stem of the verbal adjective in *-to-*. Indeed, this stem normally displays a *-t-*, and it appears in the perfect and future participle, but also in supine forms. In this work, I will follow the terminological proposal of Aronoff (1994: Chapter 2), who calls it the "**third stem**", without any reference to its semantic content, since he convincingly argues that it is not possible to find some meaning that is shared by all the inflected wordforms that are based on this stem, whose identity can only be found in the fact that its distribution in the paradigm is the same for all verbs: it is a morphomic stem, in Aronoff's terminology.

Aronoff (1994: 56) goes perhaps too far in considering also periphrastic forms like, for instance, FUT.ACT.INF *amātūrus esse* or PRF.PASS.IND.1SG *amātus sum* as being based on the third stem, despite having an imperfective and perfective meaning respectively. The exceptionality of these cells is the fact that they are filled

periphrastically, rather than the stem they are based on: in such cases, it appears that what is based on the third stem is the participle that is contained in the periphrase, rather than the cell itself, which is simply filled by a construction that happens to contain an inflected wordform based on the third stem. However, this does not affect the substance of Aronoff's point, since there are nevertheless both imperfective and perfective synthetic inflected wordforms that are based on the third stem themselves, namely the future participle *amātūrus*, *-a*, *-um* etc. and the perfect participle *amātus*, *-a*, *-um* etc., beside the two supine forms *amātum* and *amātū*.

In recent work on Romance languages, the distribution of stems in verbal paradigms has been represented as a "stem space", where cells that share the same stem in (virtually) all verbs are marked with the same index (cf. the works already cited in §1.3.2). The stem space of the Latin verbal paradigm, as it emerges from the traditional descriptions that have been mentioned and as is confirmed in the more recent and theoretically grounded account by Aronoff (1994), can be schematically displayed as in Table 3, distinguishing three different stems (with PrS = present stem, PeS = perfect stem, S3 = third stem), each of them appearing in a given zone – i.e., a set of cells – of the paradigm. In Table 3, only synthetic wordforms are considered: periphrastic cells – for instance, all the passive perfective ones – are dashed, since they can be considered as defective from a purely morphological perspective, as there are no dedicated synthetic inflected wordforms. Three separate tables are given, one for verbal forms (3a.), one for nominal forms (3b.) and one for adjectival forms (3c.).

Table 3: The stem space of Latin verb paradigms

a. verbal forms

	1SG		2SG		3SG		1PL		2PL		3PL	
	ACT	PASS	ACT	PASS	ACT	PASS	ACT	PASS	ACT	PASS	ACT	PASS
IPRF.IND	PrS	PrS	PrS	PrS	PrS	PrS	PrS	PrS	PrS	PrS	PrS	PrS
IPRF.SBJV	PrS	PrS	PrS	PrS	PrS	PrS	PrS	PrS	PrS	PrS	PrS	PrS
PRS.IMP			PrS	PrS					PrS	PrS		
PRS.IND	PrS	PrS	PrS	PrS	PrS	PrS	PrS	PrS	PrS	PrS	PrS	PrS
FUT.IMP			PrS	PrS	PrS	PrS			PrS	PrS	PrS	PrS
FUT.IND	PrS	PrS	PrS	PrS	PrS	PrS	PrS	PrS	PrS	PrS	PrS	PrS
PRS.SBJV	PrS	PrS	PrS	PrS	PrS	PrS	PrS	PrS	PrS	PrS	PrS	PrS
PRF.IND	PeS		PeS		PeS		PeS		PeS		PeS	
PLUPRF.IND	PeS		PeS		PeS		PeS		PeS		PeS	
FUTPRF.IND	PeS		PeS		PeS		PeS		PeS		PeS	
PRF.SBJV	PeS		PeS		PeS		PeS		PeS		PeS	
PLUPRF.SBJV	PeS		PeS		PeS		PeS		PeS		PeS	

b. nominal forms

PRS.INF.ACT	PrS
PRS.INF.PASS	PrS
PRF.INF.ACT	PeS
GER.GEN	PrS
GER.DAT	PrS
GER.ACC	PrS
GER.ABL	PrS
SUP.ACC	S3
SUP.ABL	S3

c. adjectival forms

	GDV	PRS. PTCP	PRF. PTCP	FUT. PTCP
NOM.M.SG	PrS	PrS	S3	S3
NOM.F.SG	PrS	PrS	S3	S3
NOM.N.SG	PrS	PrS	S3	S3
GEN.M.SG	PrS	PrS	S3	S3
GEN.F.SG	PrS	PrS	S3	S3
GEN.N.SG	PrS	PrS	S3	S3
DAT.M.SG	PrS	PrS	S3	S3
DAT.F.SG	PrS	PrS	S3	S3
DAT.N.SG	PrS	PrS	S3	S3
ACC.M.SG	PrS	PrS	S3	S3
VOC.F.SG	PrS	PrS	S3	S3
VOC.N.SG	PrS	PrS	S3	S3
ABL.M.SG	PrS	PrS	S3	S3
ABL.F.SG	PrS	PrS	S3	S3
ABL.N.SG	PrS	PrS	S3	S3
NOM.M.PL	PrS	PrS	S3	S3
NOM.F.PL	PrS	PrS	S3	S3
NOM.N.PL	PrS	PrS	S3	S3
GEN.M.PL	PrS	PrS	S3	S3
GEN.F.PL	PrS	PrS	S3	S3
GEN.N.PL	PrS	PrS	S3	S3
DAT.M.PL	PrS	PrS	S3	S3
DAT.F.PL	PrS	PrS	S3	S3
DAT.N.PL	PrS	PrS	S3	S3
ACC.M.PL	PrS	PrS	S3	S3
ACC.F.PL	PrS	PrS	S3	S3
ACC.N.PL	PrS	PrS	S3	S3
VOC.M.PL	PrS	PrS	S3	S3
VOC.F.PL	PrS	PrS	S3	S3
VOC.N.PL	PrS	PrS	S3	S3
ABL.M.PL	PrS	PrS	S3	S3
ABL.F.PL	PrS	PrS	S3	S3
ABL.N.PL	PrS	PrS	S3	S3

The identification of these three stems as the ones on which the Latin verbal system is based is justified by the fact that, although for the verb AMO given in Table 1 the different stems are predictable from one another (for instance, if the present stem is known one can obtain the perfect stem *amāv-* by means of a regular process of suffixation of the stem formative *-āv-*), there are verbs whose different stems display unpredictable allomorphic variation. For instance, in the verb RUMPO we find a nasal infix in the present stem *rump-* (e.g. PRS.ACT.IND.1SG *rumpō*) and vowel lengthening in the perfect stem *rūp-* (e.g. PRF.ACT.IND.1SG *rūpī*), and it is not possible to predict these stems from one another with certainty.

Given this mapping, if the shape of the perfect stem or of the third stem is known, then it is possible to infer the inflected words that fill all the paradigm cells of the respective zones without any uncertainty, since the same endings are used in all verbs. On the other hand, in the present system knowing the shape of the stem is not enough, because different verbs may attach different endings to the present stem, according to the inflection class they belong to.

Four conjugations are identified by all traditional descriptions of Latin, based on the theme vowel that precedes the ending *-re* in the present infinitive: *ā* in the 1st conjugation, *ē* in the 2nd, *e* in the 3rd, *ī* in the 4th. Furthermore, there is a fairly large group of heteroclitic lexemes that display the endings of the 3rd conjugation in some cells and the ones of the 4th conjugation in other cells.[3] This group is sometimes considered as a subclass of the 3rd conjugation (cf. e.g. Bennett 1908: §109, Aronoff 1994: 45), since the infinitive, which is usually taken as the basis of the classification, ends in *-ere* like in the 3rd conjugation. Other descriptions treat this group as another inflection class in its own right, calling it the "mixed" class (cf. Dressler 2002: 107), as it seems more reasonable since the paradigm is split quite evenly between cells displaying 3rd conjugation endings and cells displaying 4th conjugation endings, and actually the cells that pattern with the 4th conjugation are even more numerous. The wordforms realizing some paradigm cells of verbs belonging to different conjugations are given below in Table 4.

[3] For a thorough investigation of the phenomenon of heteroclisis in Latin verb morphology, the reader is referred to Kaye (2015).

Table 4: Latin conjugations

cell	1st: AMO ('to love')	2nd: MONEO ('to warn')	3rd: LEGO ('to read')	mixed: CAPIO (to take)	4th: VENIO ('to come')
PRS.ACT.INF	*amāre*	*monēre*	*legere*	*capere*	*venire*
PRS.ACT.IND.1SG	*amō*	*moneō*	*legō*	*capiō*	*veniō*
PRS.ACT.IND.2SG	*amās*	*monēs*	*legis*	*capis*	*venīs*
PRS.ACT.IND.3SG	*amat*	*monet*	*legit*	*capit*	*venit*
PRS.ACT.IND.1PL	*amāmus*	*monēmus*	*legimus*	*capimus*	*venīmus*
PRS.ACT.IND.2PL	*amātis*	*monētis*	*legitis*	*capitis*	*venītis*
PRS.ACT.IND.3PL	*amant*	*monent*	*legunt*	*capiunt*	*veniunt*
FUT.ACT.IND.1SG	*amābo*	*monēbo*	*legam*	*capiam*	*veniam*
FUT.ACT.IND.2SG	*amābīs*	*monēbīs*	*legēs*	*capiēs*	*veniēs*
FUT.ACT.IND.3SG	*amābit*	*monēbit*	*leget*	*capiet*	*veniet*
FUT.ACT.IND.1PL	*amābimus*	*monēbimus*	*legēmus*	*capiēmus*	*veniēmus*
FUT.ACT.IND.2PL	*amābitis*	*monēbitis*	*legētis*	*capiētis*	*veniētis*
FUT.ACT.IND.3PL	*amābunt*	*monēbunt*	*legent*	*capient*	*venient*

In some cells, the inflectional realizations of the different conjugations only differ in the theme vowel, with the proper endings being the same for all classes: we have already seen above that this is the case of PRS.ACT.INF, with ending *-re* in all classes, preceded by *ā* in the 1st conjugation, *ē* in the 2nd conjugation, *e* in the 3rd and mixed conjugation, and *ī* in the 4th conjugation. A similar situation arises in PRS.ACT.IND.2SG, although in that cell the vowel that precedes the personal ending *-s* is *i* – rather than *e* – in the 3rd and mixed conjugation.

In other cases, however, the difference between the realizations displayed by the various classes cannot be simply reconducted to theme vowels: this is very clear in the future indicative, where verbs of the 1st and 2nd conjugations use the suffix *-b-*, preceded by the appropriate theme vowel and followed by personal endings, but the inflected wordforms of verbs of the 3rd, 4th and mixed classes do not display such formative, but rather the vowel *-a-* in the first-person singular and *-e/ē-* in all the other persons, always followed by the appropriate personal endings.

The Latin conjugation system is only strictly relevant in the present system, since, as we have seen above, in the inflected wordforms built on the perfect stem or on the third stem, the same endings are used in all verbs, and different patterns of formal alternation between stems are used by verbs that belong to the same conjugation, as is shown in Table 5.

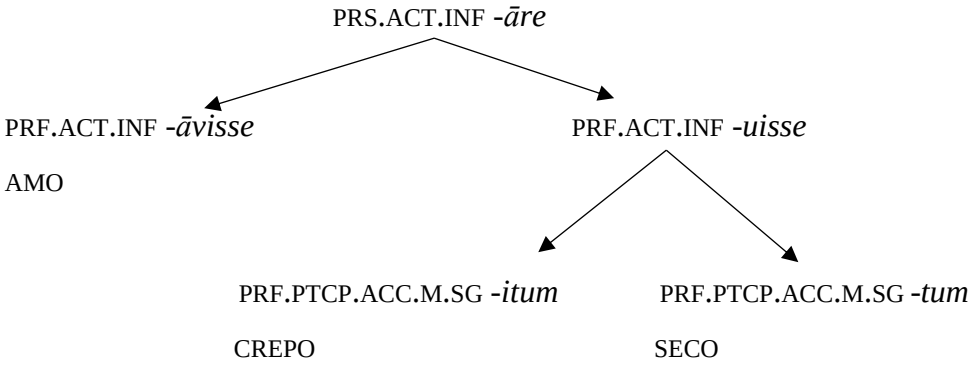
Table 5: The cells PRS.ACT.INF, PRF.ACT.INF and PRF.PTCP.ACC.M.SG of some 1st conjugation verbs

lexeme	PRS.ACT.INF	PRF.ACT.INF	PRF.PTCP.ACC.M.SG
AMO 'to love'	*amāre*	*amāvisse*	*amātum*
CREPO 'to rattle'	*crepāre*	*crepuisse*	*crepitum*
SECO 'to crack'	*secāre*	*secuisse*	*sectum*

Dressler (2002) provides a sketchy, but more comprehensive picture of Latin verb inflection classes within the framework of Natural Morphology, focusing on the so-called "micro-classes", defined as groups of verbs "that share exactly the same morphological and morphonological generalizations" (Dressler 2002: 95), rather than only the same inflectional endings, as in the traditional classification. For instance, AMO and CREPO belong to the same conjugation in the traditional account, since they display the same endings, as can be seen in Table 5: e.g. *-āre* in PRS.ACT.INF, *-isse* in PRF.ACT.INF, *-um* in PRF.PTCP.ACC.M.SG. However, they belong to different micro-classes in Dressler (2002)'s terms, since while in the lexeme AMO the cell PRF.ACT.INF displays the stem formative *-āv-*, in CREPO there is a different formative *-u-*. Similarly, CREPO and SECO belong to different micro-classes because of the different stem formatives they display in the perfect participle – *-it-* and *-t-*, respectively.

In Dressler's account, the traditional 1st and 2nd conjugations are therefore considered as constituting two "macro-classes", that capture the similarities – in this case, the fact that the same realizations are used in the cells based on the present stem – between the different micro-classes that can be found on the basis of different formal patterns of stem allomorphy. For instance, the facts summarized in Table 4 concerning the 1st conjugation are captured by the following hierarchy of micro-classes in Dressler (2002: 106).

Figure 1: Latin verb inflection: macro-class I



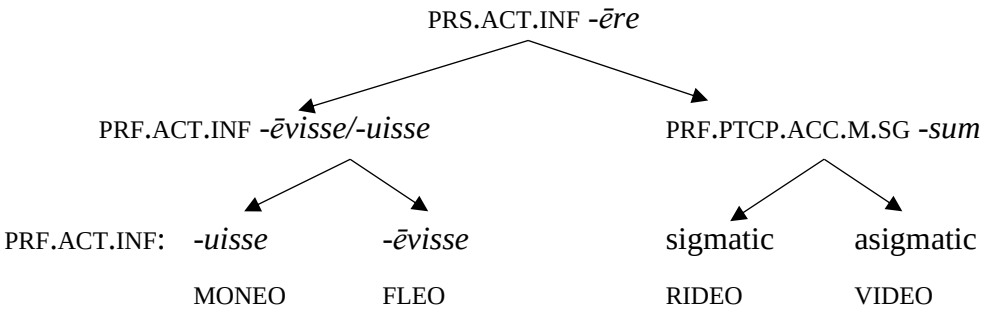
(from Dressler 2002, with simplifications)

In the 2nd conjugation, the situation is even more complex, due to the presence of other alternation patterns between stems: the most relevant ones (given in Table 6) can be schematized as in Figure 2.

Table 6: The cells PRS.ACT.INF, PRF.ACT.INF and PRF.PTCP.ACC.M.SG of some 2nd conjugation verbs

lexeme	PRS.ACT.INF	PRF.ACT.INF	PRF.PTCP.ACC.M.SG
MONEO 'to warn'	*monēre*	*monuisse*	*monitum*
FLEO 'to cry'	*flēre*	*flēvisse*	*flētum*
RIDEO 'to laugh'	*rīdēre*	*rīsisse*	*rīsum*
MORDEO 'to bite'	*vidēre*	*vīdisse*	*vīsum*

Figure 2: Latin verb inflection: macro-class II



(from Dressler 2002, with simplifications)

Regarding verbs of the 3rd, 4th and mixed conjugation, Dressler (2002) considers them as constituting a same macro-class, but he does not provide a detailed hierarchy of its different micro-classes. However, the picture would probably be similar to the one of Figure 1 and 2, with many different micro-classes due to the many different ways in which the various stems can be formed. However, it should be noticed that a tree-shaped hierarchy like the ones proposed by Dressler would not be able to capture the fact that the mixed, heteroclitic conjugation shares some properties with the 3rd conjugation and other properties with the 4th conjugation, as is argued in Beniamine (2018: 261 ff.)

After having sketched the main facts of Latin verb inflection concerning both stems and inflectional endings, it is interesting to move to another notion that is often used in traditional descriptions, but that has also been exploited in recent theoretical frameworks. We refer to the notion of principal parts, that can be defined as a set of paradigm cells from which the whole paradigm of a lexeme can be inferred (see above, §1.3.3). In Bennett (1908: §99), the cells PRS.ACT.IND.1SG, PRS.ACT.INF, PRF.ACT.IND.1SG and PRF.PTCP.NOM.M.SG are used as principal parts. If those cells are known, all the other inflected wordforms of the Latin verb paradigm can be inferred without any uncertainty. Of course, at least one cell for each of the three zones of the stem space is needed in order to know the shape of the three stems. Additionally, to be able to fill all the cells that are based on the present stem, one needs to know the conjugation to which the lexeme belongs. Because of the heteroclitic mixed conjugation, two cells – PRS.ACT.IND.1SG and PRS.ACT.INF in Bennett's usage – are needed to have a reliable information for all lexemes: for instance, given the principal parts *capiō capere cēpī captus* of the lexeme CAPIO, we know that it belongs to the mixed conjugation, since its first principal part displays the ending of the 4th conjugation, while the second one displays the ending of the 3rd conjugation.

This is due to the fact that the principal parts that are used in Bennett (1908), but also in many Latin dictionaries, are "static" in Stump & Finkel (2013)'s terminology, meaning that the same set of cells is used for all verbs in the lexicon. In an "adaptive" approach, where we have a different number of principal parts for

verbs belonging to different inflection classes, the infinitive would be sufficient to infer all the remaining cells of the present system of 1st, 2nd and 4th conjugation verbs, since the endings *-āre*, *-ēre* and *-īre* unambiguously reveal the inflection class of those verbs: the additional principal part PRS.ACT.IND.1SG would only be necessary for verbs of the 3rd and mixed conjugation, that share the infinitive in *-ere* and thus need another inflected wordform where they differ in their behaviour in order to be able to distinguish their conjugation. Therefore, in an adaptive approach, for a 1st conjugation verb like AMO the three principal parts *amāre amāvī amātus* would be sufficient, while for a verb belonging to the mixed conjugation, e.g. CAPIO, we would need a set of four principal parts that also includes *capio* alongside *capere cēpī captus*.

Finkel & Stump (2009b) provide a principal part analysis of the Latin verb paradigm obtained automatically by giving an exemplary paradigm for each inflection (sub-)class of verbs as input to their Principal Part Analyzer. Such an automatic analysis confirms that four static principal parts are needed for Latin verbs, but it also highlights the fact that they do not need to be the ones that are used in Latin traditional descriptions. For instance, also the alternative principal part set PRS.ACT.IND.3PL, IPRF.ACT.SBJV.1SG, PRF.ACT.IND.2SG, SUP.ACC (i.e. for instance *amant amārem amāvistī amātum*) would be reliable in the same way as the set proposed by Bennett (1908) mentioned above, since the same amount of information is provided, although by means of different cells. A lot of different and equally reliable principal part sets can therefore be provided for Latin verbs, given the fact that many cells are in systematic covariation one with another, and are therefore completely interchangeable as principal parts.

The notion of principal parts is particularly interesting in the context of the present work, since it exploits the presence of reliable implicative relations between full inflected wordforms, exactly as we will do with a different, entropy-based method. The similarities and differences between Finkel & Stump (2009b)'s principal part analysis and the one that can be obtained with our methodology will be treated more extensively below in §4.6.

4.2 The cell paradigm of Latin verbs

Having sketched out the main facts of Latin verb inflection, we can now move to the analysis of the system in terms of predictability and implicative relations. However, our starting point will not be the paradigm shown in Table 1, but a slightly reduced version of it. Firstly, cells that are never filled synthetically, but always by means of a periphrase, will not be taken into account, since the PCFP concerning those cells can be taken as being tackled for each of the forms constituting the periphrase separately, and such forms are already present in our dataset. Therefore, all the passive perfective cells and the future (active and passive) infinitive have been excluded: this choice reduces the paradigm of Table 1 to a 254-cell paradigm – the one that is given in LatInfLexi.

Another principled reduction of the paradigm of Table 1 can be achieved by following a suggestion of Boyé and Schalchli (2016), who distinguish between tabular paradigms, cell paradigms, and morphomic paradigms. In Table 1, the inflected wordforms of the paradigm are represented in the way that is common in language descriptions: Boyé and Schalchli (2016) call it the **tabular paradigm**. This representation format is content-driven: the starting point is constituted by the combinations of morphosyntactic properties that are morphologically possible – thus excluding combinations that are not allowed, for instance PRS.PASS.IMP.1SG in Latin. The wordform realizing each morphosyntactic property set is then given in the corresponding cell. In some cases, however, different morphosyntactic property sets are realized by the same surface wordform in a given paradigm: for instance, in Table 1 it can be observed that the wordform *amātīs* can realize 6 different morphosyntactic property sets, namely the dative and ablative plural of the perfect participle, in all three genders. This makes this representation format redundant from a purely formal point of view: the same wordform is repeated six times because of the different morphosyntactic contexts where it could appear. This redundancy emerges clearly in Table 7, where the relevant section of the paradigm is given in the tabular representation.

Table 7: The tabular representation format for a section of the paradigm of AMO 'to love'

MPS	wordform
PRF.PASS.PTCP.DAT.M.PL	*amātīs*
PRF.PASS.PTCP.DAT.F.PL	*amātīs*
PRF.PASS.PTCP.DAT.N.PL	*amātīs*
PRF.PASS.PTCP.ABL.M.PL	*amātīs*
PRF.PASS.PTCP.ABL.F.PL	*amātīs*
PRF.PASS.PTCP.ABL.N.PL	*amātīs*

Alternatively, we could represent the same facts in a different, form-driven rather than content-driven, way, in what Boyé and Schalchli (2016) call the **morphomic paradigm**, where the starting point is given by the phonologically contrasting wordforms of a given lexeme, and all the morphosyntactic property sets that can be associated with each of these wordforms are consequently listed. A shortcoming of this much more compact representation format is that the size of the paradigm can vary in different lexemes: for instance, in the lexeme RUMPO the FUT.ACT.IND.1SG and the PRS.ACT.SBJV.1SG are realized by the same wordform *rumpam*, but those same morphosyntactic property sets are realized by different wordforms – *amābō* and *amem* – in the lexeme AMO. Therefore, regarding the fragment of Latin verb inflection shown in Table 8 and Table 9, the morphomic paradigm of AMO has three cells, whereas the morphomic paradigm of RUMPO only has two.

Table 8: The morphomic representation format for a section of the paradigm of AMO 'to love'

wordform	MPS
amātīs	PRF.PASS.PTCP.DAT.M.PL, PRF.PASS.PTCP.DAT.F.PL, PRF.PASS.PTCP.DAT.N.PL, PRF.PASS.PTCP.ABL.M.PL, PRF.PASS.PTCP.ABL.F.PL, PRF.PASS.PTCP.ABL.N.PL
amābō	FUT.ACT.IND.1SG
amem	PRS.ACT.SBJV.1SG

Table 9: The morphomic representation format for a section of the paradigm of RUMPO 'to break'

wordform	MPS
ruptīs	PRF.PASS.PTCP.DAT.M.PL, PRF.PASS.PTCP.DAT.F.PL, PRF.PASS.PTCP.DAT.N.PL, PRF.PASS.PTCP.ABL.M.PL, PRF.PASS.PTCP.ABL.F.PL, PRF.PASS.PTCP.ABL.N.PL
rumpam	FUT.ACT.IND.1SG, PRS.ACT.SBJV.1SG

As a more balanced representation format for morphological paradigms, Boyé & Schalchli (2016) propose what they call the **cell paradigm**. It is important to observe that this label is based on a definition of "cell" that is different from the one that is usually adopted in the literature, given above in §1.1. Boyé & Schalchli (2016) define a "cell" as in (1).

(1) a. A cell is a group of contents whose phonological forms never contrast for any lexeme.

b. Two contents do not belong to the same cell only if their inflected forms contrast for at least one lexeme.

(Boyé & Shcalchli 2016: 209)

Based on the definitions in (1), the dative and ablative plural of the perfect participle belong to the same cell (cell 1 in Table 10), since there is no verb in the Latin lexicon where those morphosyntactic property sets are realized by different wordforms. On the other hand, FUT.ACT.IND.1SG and PRS.ACT.SBJV.1SG require two separate cells (cells 2 and 3), since the syncretic pattern displayed by RUMPO is only valid for verbs of the 3rd, 4th and mixed conjugation, and not for the whole lexicon.

Table 10: The cell paradigm representation format for a section of the paradigm of RUMPO 'to break'

cell	wordform	MPS
1	*ruptīs*	PRF.PASS.PTCP.DAT.M.PL, PRF.PASS.PTCP.DAT.F.PL, PRF.PASS.PTCP.DAT.N.PL, PRF.PASS.PTCP.ABL.M.PL, PRF.PASS.PTCP.ABL.F.PL, PRF.PASS.PTCP.ABL.N.PL
2	*rumpam*	PRS.ACT.SBJV.1SG
3	*rumpam*	FUT.ACT.IND.1SG

This representation format is similar to the morphomic one in cell 1, where the systematically syncretic morphosyntactic property sets of the dative and ablative plural perfect participle are conflated in a single cell, but it is similar to the tabular format in cells 2 and 3, that contain the same wordform for this lexeme, since such syncretism is not valid for all verbs. Therefore, a more compact encoding is achieved without the shortcoming of having lexemes with different number of cells: since we only abstract away from completely systematic syncretism, the size of the paradigm is the same for all verbs.

If we apply this procedure to the whole Latin verbal system, we obtain a paradigm composed of 152 cells, considerably smaller in size if compared to the 254 cells of the tabular paradigm – the one given in Table 1, with the exclusion of periphrastic cells. As far as finite forms are concerned, the difference in size between the tabular paradigm and the cell paradigm is very limited: the only systematic patterns of syncretism that are captured are the one between the second and third person singular of the future imperative (cf. the wordform *amātō* in Table 1) and the one between the future perfect indicative and the perfect subjunctive in the third person singular and plural (cf. the wordforms *amāverit* and *amāverint* in Table 1).

The reduction in size is mainly due to participles and other nominal forms, where cases of systematic syncretism like the one of the plural dative and ablative of perfect participles are widespread. As an example, the tabular and cell paradigm of present participles is given in Table 11, highlighting the relevant reduction in paradigm size that can be achieved.

Table 11: The tabular and cell paradigm of the present participle of AMO 'to love'

a. Tabular paradigm

MPS	wordform
NOM.M.SG	*amāns*
NOM.F.SG	*amāns*
NOM.N.SG	*amāns*
ACC.N.SG	*amāns*
VOC.M.SG	*amāns*
VOC.F.SG	*amāns*
VOC.N.SG	*amāns*
GEN.M.SG	*amantis*
GEN.F.SG	*amantis*
GEN.N.SG	*amantis*
DAT.M.SG	*amantī*
DAT.F.SG	*amantī*
DAT.N.SG	*amantī*
ACC.M.SG	*amantem*
ACC.F.SG	*amantem*
ABL.M.SG	*amante*
ABL.F.SG	*amante*
ABL.N.SG	*amante*
NOM.M.PL	*amantēs*
NOM.F.PL	*amantēs*
ACC.M.PL	*amantēs*
ACC.F.PL	*amantēs*
VOC.M.PL	*amantēs*
VOC.F.PL	*amantēs*
NOM.N.PL	*amantia*
ACC.N.PL	*amantia*
VOC.N.PL	*amantia*
GEN.M.PL	*amantium*
GEN.F.PL	*amantium*
GEN.N.PL	*amantium*
DAT.M.PL	*amantibus*
DAT.F.PL	*amantibus*
DAT.N.PL	*amantibus*
ABL.M.PL	*amantibus*
ABL.F.PL	*amantibus*
ABL.N.PL	*amantibus*

b. Cell paradigm

cell	wordform
1	*amāns*
2	*amantis*
3	*amantī*
4	*amantem*
5	*amante*
6	*amantēs*
7	*amantia*
8	*amantium*
9	*amantibus*

Now, cell paradigms appear to be the right starting point for an analysis of predictability and implicative relations between inflected wordforms:[4] the

[4] However, it should be stressed that we do not claim that this is the best representation format for paradigms in general: for instance, tabular paradigms remain useful to capture generalizations on

conflation of systematically syncretic cells is unproblematic, since the very fact that those cells are systematically syncretic means that they can predicted from one another with no uncertainty. Therefore, in the next sections we will present results obtained on the cell paradigm of Latin verbs.

4.3 Predictability in Latin verb inflection: wordforms that are based on different stems

We saw in §4.1 that the Latin verbal system is traditionally described as based on three stems – the present stem, the perfect stem and the third stem – on which the various inflected wordforms are built. While in regular verbs there is more predictability between wordforms built on different stems, there are also many cases of fully unpredictable stem allomorphy in the lexicon. Traditional descriptions of Latin verb inflection list many different patterns of stem allomorphy for each of the three stems. Some of the patterns given by Bennett (1908: §188) for the perfect stem are reported in Table 12.

Table 12: Patterns of stem allomorphy in the perfect stem

pattern	lexeme	perfect stem
-v- suffixation	AMO 'to love'	*amāv-*
-s- suffixation	CARPO 'to pick'	*carps-*
reduplication	CURRO 'to run'	*cucurr-*
lengthening	LEGO 'to read'	*lēg-*

As a starting point of our analysis, we can evaluate the impact of similar phenomena of stem allomorphy on predictability in Latin verb paradigms, as measured by means of implicative entropy. To do so, we will now select from the data of LatInfLexi three different paradigm cells, each one containing one of the three stems – PRS.ACT.INF containing the present stem, PRF.ACT.IND.1SG containing the perfect stem and PRF.PTCP.NOM.M.SG containing the third stem – and give them as input to the Qumin toolkit.

the content of inflected wordforms in an economic way, since the various morphosyntactic property sets can be obtained by simply crossing the possible values of a closed set of morphosyntactic features.

As we saw above in Chapter 2, the first step of the automatic analysis consists in finding the patterns of formal alternation between these inflected wordforms. For instance, let us have a look at the alternation patterns between PRS.ACT.INF and PRF.ACT.IND.1SG. In Table 13, the patterns that are attested for more than 20 verbs are given, sorted according to their type frequency in LatInfLexi.

Table 13: Alternation patterns between PRS.ACT.INF and PRF.ACT.IND.1SG

n.	pattern[5]	example		type
		PRS.ACT.IND	PRF.ACT.IND.1SG	freq.
1	re ↔wi:	ama:re	ama:wi:	1336
2	ere ↔ i:	solwere	solwi:	214
3	e:re ↔ ui:	mone:re	monui:	134
4	e:skere ↔ ui:	kande:skere	kandui:	88
5	ere ↔ ui:	kolere	kolui:	70
6	i_ere ↔ e:_i:	adigere	ade:gi:	57
7	V_gere ↔ V:_ksi:	kingere	kinksi:	50
8	dere ↔ si:	klawdere	klawsi:	48
9	ere ↔ si:	du:kere	du:ksi:	47
10	skere ↔ wi:	kre:skere	kre:wi:	43
11	a:re ↔ ui:	krepa:re	krepui:	34
12	V_ere ↔ V:_i:	fugere	fu:gi:	33
13	[bdzg]ere ↔ [kpst]si:	fi:gere	fi:ksi:	33
14	a_ere ↔ e:_i:	agere	e:gi:	31
15	V_e:re ↔ V:_i:	mowe:re	mo:wi:	30
16	Vhere ↔ V:ksi:	trahere	tra:ksi:	26
17	Vn_ere ↔ V:_i:	fundere	fu:di:	26
18	uere ↔ u:ksi:	fluere	flu:ksi:	22
19	ere ↔ idi:	wendere	wendidi:	20
20	ittere ↔ i:si:	mittere	mi:si:	20
21	o:nere ↔ osui:	po:nere	posui:	20

One alternation pattern (pattern 1) – clearly emerges as the most frequent, covering about 45% of the 2,939 lexemes of the lexicon that are not defective in one or both of the involved cells. Although in Table 13, for the sake of simplicity, only the alternation pattern is displayed, the Qumin toolkit also provides its context of application, that is given in (2).

(2) re ↔ wi: / X*[i:-e:-a:]

[5] While in the rest of this work for simplicity and readability patterns and wordforms are given in orthographic transcription, in this table and in the other ones provided in this section, where the focus is on the alternation patterns themselves, we use the IPA notation in which they are outputted by the Qumin toolkit.

This alternation pattern only appears when the *-re* ending of PRS.ACT.IND is preceded by the long theme vowels [aː], [eː] or [iː], that is to say in 1st, 2nd and 4th conjugation verbs. However, it should be noticed that this pattern is not actually the most frequent one for 2nd conjugation verbs, where the dominant pattern is actually pattern 3: the high frequency of pattern 1 is mainly due to the fact that it corresponds to the productive pattern of stem formation for verbs of the 4th conjugation and especially for the very frequent 1st conjugation. Pattern 2 is already far less frequent than pattern 1, and it is restricted to verbs of the 3rd and mixed conjugation – the only ones whose present infinitive ends in *-ere*. However, in those classes there is not a pattern that clearly outweighs the others, as is shown by patterns 4-10, that are all restricted to infinitives in *-ere* and do not display a great difference in their type frequency.

The presence of such a large number of alternation patterns has an impact on the mutual predictability of the two involved cells, since some of the patterns of Table 13 are in competition, meaning that they can be applied in contexts that are at least partly overlapping. For instance, both pattern 1 and pattern 11 are compatible with verbs with a PRS.ACT.INF in *-āre*, at least in cases where this ending is preceded by segments that are compatible with the context computed by the toolkit.[6] Therefore, it is not possible to predict with certainty if the PRS.ACT.IND.1SG of those verbs will be in *-āvī* or in *-uī*, causing some uncertainty in the PCFP. In order to assess the quantitative impact of this uncertainty by means of implicative entropy, the toolkit counts the number of lexemes where the patterns are actually applied, and then uses this information to evaluate the probability of application of the different patterns (cf. Table 14). In this case, pattern 1 proves to be much more likely to be applied than pattern 11. Therefore, this competition will generate only a small increase in uncertainty, since, as we saw above in §2.3, a more skewed probability distribution corresponds to a lower entropy value: it is likely that the more frequent pattern will be used.

[6] Indeed, there are many other verbs in *-āre* in LatInfLexi beside the ones considered in Table 14 below, but in all such cases the preceding segments are not compatible with the context of application of the less common pattern 11, as computed by the Qumin toolkit. Therefore, since only pattern 1 is considered to be applicable, there is no competition between different patterns and, as a consequence, no uncertainty regarding these verbs.

Table 14: A case of competition between patterns that can be applied to obtain PRF.ACT.IND.1SG from PRS.ACT.INF

n.	pattern	example lexeme	PRS.ACT. INF	PRF.ACT. IND.1SG	type freq.	probability of application
1	re ↔ wi:	VOCO 'to call'	woka:re	woka:wi:	331	0.907
11	a:re ↔ ui:	VETO 'to forbid'	weta:re	wetui:	34	0.093

If we look at the prediction in the opposite direction – i.e., predicting PRS.ACT.INF knowing PRF.ACT.IND.1SG – we find a case of competition between patterns 2, 3, 4, 5 and 11 in some of the verbs whose PRF.ACT.INF ends in *-uī* (when the contexts of such patterns overlap).[7] In this case, the probability distribution is quite balanced (cf. Table 15), although the impact on the overall entropy value will not be very high because of the relatively small number of verbs for which this competition is relevant.

Table 15: A case of competition between patterns that can be applied to obtain PRS.ACT.INF from PRF.ACT.IND.1SG

n.	pattern	example lexeme	PRF.ACT. IND.1SG	PRS.ACT. INF	type freq.	prob. of appl.
2	ere ↔ i:	TRIBUO 'to assign'	tribui:	tribuere	8	0.066
3	e:re ↔ ui:	VIGEO 'to be lively'	wigui:	wige:re	38	0.311
4	e:skere ↔ ui:	VIGESCO 'become lively'	wigui:	wige:skere	25	0.205
5	ere ↔ ui:	VOMO 'to puke'	womui:	womere	32	0.262
11	a:re ↔ ui:	VETO 'to forbid'	wetui:	weta:re	19	0.156

In Table 16 and Table 17, the alternation patterns between each of the two wordforms considered above and another one containing the so-called third stem – namely, PRF.PTCP.NOM.M.SG – are reported, again showing only patterns that are attested in more than 20 lexemes. Without going in too much detail, it should at

[7] Again, it is the fact that the preceding segments are required to match the context of application of each pattern that explains the difference between the type frequencies of Table 15 and the ones of Table 13 (cf. Footnote 6 above).

least be observed that the situation is similar to the one of Table 13 in the presence of one clearly prevailing pattern, covering for about half of the verbs of the lexicon, and a series of much less frequent patterns that are also far closer to each other in their frequency. In both cases, the majority pattern is the one that is predominant in the productive 1st conjugation, but also in the 4th conjugation, also appearing in a few 2nd conjugation verbs – again, exactly as it was the case for the alternation patterns between PRS.ACT.INF and PRF.ACT.IND.1SG, shown in Table 13.

Table 16: Alternation patterns between PRS.ACT.INF and PRF.PTCP.NOM.M.SG

n.	pattern	example		type freq.
		PRS.ACT.IND	PRF.ACT.IND.1SG	
1	re ↔ tus	ama:re	ama:tus	1294
2	ere ↔ tus	fakere	faktus	74
3	dere ↔ sus	ra:dere	ra:sus	65
4	i_ere ↔ e_tus	abripere	abreptus	61
5	_ere ↔ u:t_s	akuere	aku:tus	50
6	ere ↔ itus	fugere	fugitus	45
7	e:re ↔ itus	mone:re	monitus	43
8	tere ↔ sus	wertere	wersus	37
9	V_gere ↔ V:_ktus	kingere	ki:nktus	34
10	V:_ere ↔ V_tus	du:kere	duktus	30
11	i:re ↔ tus	inweni:re	inwentus	28
12	here ↔ ktus	trahere	traktus	26
13	V_dere ↔ V:_sus	pendere	pe:nsus	25
14	_kere ↔ tu_	kre:skere	kre:tus	25
15	[bdzg]ere ↔ [kpst]tus	skri:bere	skri:ptus	23
16	V[bdzg]ere ↔ V:[kpst]tus	agere	a:ktus	22
17	_ndere ↔ u:s_s	fundere	fu:sus	22
18	o:nere ↔ ositus	po:nere	positus	20
19	ttere ↔ ssus	mittere	missus	20

Table 17: Alternation patterns between PRF.ACT.IND.1SG and PRF.PTCP.NOM.M.SG

n.	pattern	example		type freq.
		PRF.ACT.IND.1SG	PRF.PTCP.NOM.M.SG	
1	wi: ↔ tus	ama: wi:	ama:tus	1366
2	_i: ↔ tu_	re:psi:	re:ptus	199
3	i: ↔ us	diwi:si:	diwi:sus	161
4	_i: ↔ it_s	monui:	monitus	93
5	V:_i: ↔ V_tus	ka:wi:	kawtus	71
6	V:__i: ↔ V_tu_	du:ksi:	duktus	54
7	_i: ↔ u:t_s	spui:	spu:tus	50
8	di: ↔ sus	fu:di:	fu:sus	47
9	_i: ↔ t_s	dokui:	doktus	43
10	di: ↔ tus	wendidi:	wenditus	31
11	V_di: ↔ V:_sus	prandi:	pra:nsus	30
12	e:_i: ↔ a_tus	ke:pi:	kaptus	28
13	V:di: ↔ Vssus	se:di:	sessus	27
14	e:gi: ↔ a:ktus	e:gi:	a:ktus	22
15	i:_i: ↔ i_sus	mi:si:	missus	20

In Table 18, we show the entropy values that are computed by the Qumin toolkit to quantify the interpredictability between the three involved paradigm cells. To familiarize the reader with the format in which the results are displayed here and in the rest of this work, it should be observed that the entropy value of each case of the table quantifies the uncertainty in guessing the content of the paradigm cell in the corresponding column assuming knowledge of the inflected wordform that occupy the paradigm cell in the corresponding line. Therefore, the lines of the table quantify the predictiveness of the cell, while the columns provide information on its predictability.

Table 18: Results: interpredictability between wordforms that are based on different stems

	PRS.ACT.INF	PRF.ACT.IND.1.SG	PRF.PTCP.M.NOM.SG
PRS.ACT.INF		0.252	0.2158
PRF.ACT.IND.1.SG	0.4736		0.3
PRF.PTCP.M.NOM.SG	0.2362	0.2246	

These results show that there is a non-negligible impact of phenomena of stem allomorphy on the predictability and predictiveness of paradigm cells: obviously, unpredictable patterns of stem allomorphy generate uncertainty. However, the entropy values are never too high, being always by far less than 1, in line with the

Low Entropy Conjecture proposed by Ackerman & Malouf (2013) and already introduced here in § 2.2, indicating that despite the presence of plenty of different formal patterns, as listed in Table 13, Table 16 and Table 17, the uncertainty in the PCFP still remains inside limits that make it manageable to tackle for speakers.

4.4 Predictability in Latin verb inflection: wordforms that are based on the same stem

In the previous section, we focused on the uncertainty that arises in the PCFP between inflected wordforms that are based on different stems, thus weighting the impact of phenomena of stem allomorphy on predictability in inflectional paradigms. We can now move to inflected wordforms that are based on the same stem, evaluating the role that is played by the presence of inflection classes. In this section, we will comment the results obtained by running the Qumin toolkit on each of the paradigm zones shown in Table 3 in turn.

Let us start from the case where the situation is simpler, i.e. from the cells that are based on the perfect stem, where there is no allomorphy whatsoever in inflectional endings. This means that all cells based on the perfect stem are in completely systematic covariation, and can therefore be predicted from one another with no uncertainty, i.e. with a null entropy value in both directions, as is indeed confirmed by our empirical results: entropy is always 0 between cells based on the perfect stem.

We saw in § 4.1 that also in paradigm cells that are based on the third stem the endings are the same for all verbs. Therefore, we would expect to find only null entropy values also there. However, a very small residual uncertainty emerges from our results concerning future participles, that cannot be predicted with absolute certainty from other cells in this paradigm zone, as can be seen in Table 19, where the results concerning some cells that contain the third stem are reported, followed by the average implicative entropy computed on all the relevant cells.

Table 19: Results: interpredictability between wordforms that are based on the third stem

	SUP.ACC	SUP.ABL	PRF.PTCP. M.NOM.SG	PRF.PTCP. F.NOM.SG	FUT.PTCP. M.NOM.SG	FUT.PTCP. F.NOM.SG
SUP.ACC		0	0	0	0.00831	0.00831
SUP.ABL	0		0	0	0.01537	0.01537
PRF.PTCP. M.NOM.SG	0	0		0	0.00818	0.00818
PRF.PTCP. F.NOM.SG	0	0	0		0.00818	0.00818
FUT.PTCP. M.NOM.SG	0.00396	0.006687	0.003252	0.003252		0
FUT.PTCP. F.NOM.SG	0.00396	0.006687	0.003252	0.003252	0	

Average implicative entropy: 0.003154

This uncertainty is due to the fact that there are a few verbs that actually use another stem allomorph in future participles, differently than the overwhelming majority of Latin verbs. Some examples of such irregular verbs are given below in Table 20.

Table 20: Verbs that display a different stem in SUP.ACC and PRF.PTCP.NOM.M.SG

lexeme	SUP.ACC	FUT.PTCP.NOM.M.SG
RUO 'to fall down'	*rutum*	*ruitūrus*
ORIOR 'to rise'	*ortum*	*oritūrus*
NASCOR 'to be born'	*natum*	*nascitūrus*
MORIOR 'to die'	*mortuum*	*moritūrus*
PARIO 'to beget'	*partum*	*paritūrus*

Both traditional descriptions (cf. e.g. Bennett 1908: §119) and the more recent account by Aronoff (1994) acknowledge this fact, but because of the rarity of such verbs they nevertheless consider the Latin verbal system as being based on three stems, rather than introducing a fourth one that would only be relevant for a handful of lexemes. On the other hand, in a procedure such as the one used by e.g. Bonami & Boyé (2003) to identify stems, such a state of affairs would lead to posit a 4th stem for cells of the future participle.

An advantage of the quantitative, entropy-based approach that is used in this work is that the impact of such facts can be given its right weight within the overall system. In the results shown in Table 19, entropy is not 0 between e.g. SUP.ACC and FUT.PTCP.NOM.M.SG, since for a verb like NASCOR knowing the SUP.ACC *ortum* it is not possible to infer the FUT.PTCP.NOM.M.SG *oritūrus*, which is based on a different stem. However, the entropy value is very low, since this unpredictable stem allomorphy is only displayed by a handful of lexemes, while for all the other verbs the two inflected wordforms are based on the same stem, and consequently there is no uncertainty.

Let us now move to cells that are based on the present stem. Here, the situation in terms of interpredictability is much more complex, as is shown in the cells given in Table 21, due to the fact that there is a relevant amount of uncertainty caused by allomorphy in inflectional endings – cf. the five conjugations described above in § 4.1. To improve the readability of the table, different shades of grey are used to indicate different levels of predictability, as measured by conditional entropy, with darker shades corresponding to higher entropy values and therefore to cases where there is more uncertainty, and white corresponding to pairs of cells that can be predicted from one another with no uncertainty, and therefore with H = 0.

Table 21: Results: interpredictability between forms that are based on the present stem

	PRS.ACT. IND.1.SG	PRS.ACT. IND.2.SG	PRS.ACT. IND.3.SG	PRS.ACT. IND.1.PL	PRS.ACT. IND.2.PL	PRS.ACT. IND.3.PL	IPRF.ACT. IND.1.SG	PRS.ACT. INF
PRS.ACT. IND.1.SG		1.049	1.051	1.024	0.9736	0.9556	0.9443	0.9365
PRS.ACT. IND.2.SG	0.2324		0	0.004562	0.004562	0.2142	0.2379	0.004562
PRS.ACT. IND.3.SG	0.3823	0.4778		0.9653	0.9653	0.383	0.3953	0.2651
PRS.ACT. IND.1.PL	0.2131	0	0		0	0.2152	0.2236	0
PRS.ACT. IND.2.PL	0.2201	0	0	0		0.2144	0.2334	0
PRS.ACT. IND.3.PL	0.00859	0.10126	0.0902	0.4106	0.1051		0.00844	0.05695
IPRF.ACT. IND.1.SG	0.3647	0.4375	0.4417	0.4133	0.4363	0.3647		0.3862
PRS.ACT .INF	0.1528	0	0	0	0	0.1528	0.1561	

Average implicative entropy: 0.0638

It should be observed that the presence of inflection classes does not necessarily entail uncertainty in the PCFP. Indeed, there are cells that can be reliably predicted from one another despite the fact that they display different endings in different conjugations, since those endings are in systematic covariation. An example is given below in Table 22. Even if both in PRS.ACT.INF and in PRS.ACT.IND.2PL there are different endings in each conjugation, there is no effect on interpredictability, since there is a set of completely reliable, bidirectional implicative relations linking one cell to the other one: if PRS.ACT.INF ends in *-re* preceded by a long vowel (or by *-r-*, as in the irregular verb FERO), then PRS.ACT.IND.2PL will be obtained by replacing *-re* with *-tis*, and vice versa; if PRS.ACT.INF ends in *-ere*, then PRS.ACT.IND.2PL will end in *-itis*, and vice versa.

Table 22: The cells PRS.ACT.INF and PRS.ACT.IND.2PL in lexemes of different conjugations[8]

lexeme (conj.)	PRS.ACT.INF	PRS.ACT.IND.2PL
AMO 'to love' (1st)	*amāre*	*amātis*
MONEO 'to warn' (2nd)	*monēre*	*monētis*
LEGO 'to read' (3rd)	*legere*	*legitis*
CAPIO 'to take' (mixed)	*capere*	*capitis*
VENIO 'to come' (4th)	*venīre*	*venītis*
FERO 'to bring' (irr.)	*ferre*	*fertis*

In cases of pairs of cells where the overall situation is similar, but there are a few, highly irregular verbs that have an unpredictably different inflectional behaviour, entropy values are close to 0, indicating that the interpredictability of such cells is not complete, but it is still very high.

For instance, PRS.ACT.IND.1PL and PRS.ACT.IND.2PL are normally linked by an implicative relation according to which, if the former ends in *-mus*, then the latter can be obtained by replacing that ending with *-tis*, as exemplified in Table 23.

[8] In this table, we do not report the PRS.ACT.INF and PRS.ACT.IND.2PL of highly irregular verbs like SUM 'to be' (*esse* and *estis*, respectively) and VOLO 'to want' (*velle* and *vultis*, respectively). The allomorphic alternation patterns displayed by such verbs, however, do not produce any additional uncertainty in our results: in this cases too, there is a completely reliable implicative relation between the cells involved, the only difference being that such implicative relations have a much smaller coverage, since they are only applied to one verb, and possibly to other verbs derived from it (e.g. SUPERSUM 'to be left', with PRS.ACT.INF *superesse* and PRS.ACT.IND.2PL *superestis*).

Although this implicative relation has a very wide coverage, it does not hold for the whole lexicon: in a verb like FERO, the PRS.ACT.IND.2PL is not **feritis* (as one would expect on the basis of the PRS.ACT.IND.1PL *ferimus* and of the aforementioned implicative relation), but *fertis*. Therefore, given a lexeme with first-person plural in *-imus*, there is some uncertainty on the realization of the second-person plural, whether it will be *-itis* or *-tis*.

Table 23: The cells PRS.ACT.IND.1PL and PRS.ACT.IND.2PL in lexemes of the 3rd conjugations and in FERO

lexeme	PRS.ACT.IND.1PL	PRS.ACT.IND.2PL
LEGO 'to read'	*legimus*	*legitis*
FERO 'to bring'	*ferimus*	*fertis*

Allomorphy in inflectional endings only generates high entropy values when the endings that appear in a given cell are not fully informative on the endings that are used in different cells. The most systematic reason for this to happen has to do with the heteroclitic mixed class: as was hinted above in §4.1, the very fact that the mixed conjugation shares the endings with the 3rd conjugation in some cells and with the 4th conjugation in other cells makes those endings not reliably informative about the inflection class membership of the verb displaying them. For instance, Table 24 shows that given a lexeme with PRS.ACT.IND.3PL ending in *-iunt*, since this ending is common to verbs of the 4th and mixed conjugation, the involved lexeme cannot be assigned to an inflection class with certainty, thus it is not possible to know if the PRS.ACT.INF will be in *-ere* or *-īre*. Conversely, the ending *-ere* of the PRS.ACT.INF is common to verbs of the 3rd and mixed conjugation, thus given an infinitive with that ending one cannot know if the PRS.ACT.IND.3PL will be in *-unt* or in *-iunt*.

Table 24: The cells PRS.ACT.IND.3PL and PRS.ACT.INF in lexemes of the 3rd, 4th and mixed conjugation

lexeme (conj.)	PRS.ACT.IND.3PL	PRS.ACT.INF
LEGO 'to read' (3rd)	*legunt*	*legere*
CAPIO 'to take' (mixed)	*capiunt*	*capere*
VENIO 'to come' (4th)	*veniunt*	*venire*

In other cases, unpredictability is due to less systematic cases of neutralization of inflection class distinctions: in the cell PRS.ACT.IND.1SG, for instance, the distinction between the 1st and the 3rd conjugation is neutralized, with both classes displaying the ending *-ō* (cf. Table 25).[9] On the other hand, in the imperfect indicative it is the distinction between 2nd and 3rd conjugation that is neutralized (cf. Table 26). These opacities generate high entropy values when predicting the content of other paradigm cells – for instance PRS.ACT.IND.3PL, as in Table 25 and Table 26.

Table 25: The cells PRS.ACT.IND.1SG and PRS.ACT.IND.3PL in lexemes of different conjugations

lexeme (conj.)	PRS.ACT.IND.1SG	PRS.ACT.IND.3PL
AMO 'to love' (1st)	*amō*	*amant*
MONEO 'to warn' (2nd)	*moneō*	*monent*
LEGO 'to read' (3rd)	*legō*	*legunt*
CAPIO 'to take' (mixed)	*capiō*	*capiunt*
VENIO 'to come' (4th)	*veniō*	*veniunt*

Table 26: The cells IPRF.ACT.IND.1SG and PRS.ACT.IND.3PL in lexemes of different conjugations

lexeme (conj.)	IPRF.ACT.IND.1SG	PRS.ACT.IND.3PL
AMO 'to love' (1st)	*amābam*	*amant*
MONEO 'to warn' (2nd)	*monēbam*	*monent*
LEGO 'to read' (3rd)	*legēbam*	*legunt*
CAPIO 'to take' (mixed)	*capiēbam*	*capiunt*
VENIO 'to come' (4th)	*veniēbam*	*veniunt*

When the impact of such opacities is combined with the effect of heteroclisis, as happens in the pair of cells given in Table 27, higher entropy values arise (cf. the corresponding case of Table 21).

[9] Ironically, it is exactly this cell that is used as a citation form for verbal lexemes in Latin, despite its opacity in revealing the inflectional behaviour of lexemes, that makes it actually the least predictive cell (see below the results in Table 33b).

Table 27: The cells PRS.ACT.IND.1SG and PRS.ACT.INF in lexemes of different conjugations

lexeme (conj.)	PRS.ACT.IND.1SG	PRS.ACT.INF
AMO 'to love' (1st)	*amō*	*amāre*
MONEO 'to warn' (2nd)	*moneō*	*monēre*
LEGO 'to read' (3rd)	*legō*	*legere*
CAPIO 'to take' (mixed)	*capiō*	*capere*
VENIO 'to come' (4th)	*veniō*	*venīre*

The entropy value concerning the latter pair is much higher than all the preceding ones because the effect neutralization between the 1st and the 3rd conjugation is very relevant from a quantitative standpoint: their relative frequency is similar, giving rise to a balanced probability distribution and therefore to high entropy values; furthermore, both classes are very frequent, thus the impact on the whole lexicon is huge. This clearly emerges from the quantitative data given below in Table 28 on the type frequency of different patterns and consequently on their probability of application in a case where the patterns are in competition for the same input wordform. The opacity between 2nd and 3rd conjugation in imperfective cells is less relevant from a quantitative point of view, since 2nd conjugation verbs are much rarer: therefore, the probability distribution is less balanced, and the impact of this source of unpredictability on the whole lexicon is minor, as can be seen from the results of Table 21.

Table 28: A case of competition between patterns that can be applied to obtain PRS.ACT.INF from PRS.ACT.IND.1SG

pattern	example lexeme	PRS.ACT.INF	PRF.ACT.IND.1SG	type freq.	probability of application
o: ↔ a:re	AMO 'to love'	amo:	ama:re	1215	0.539
o: ↔ ere	LEGO 'to read'	lego:	legere	1041	0.461

4.5 Zones of interpredictability in Latin verb inflection

In the previous section, we have analysed different sections of the Latin verb paradigm separately, by looking at entropy values regarding the present system, the perfect system and the nominal forms based on the third stem in turn. From now on, we will try to focus on generalizations and results that can be considered to be

valid on the whole paradigm. To do so, the first step consists in drawing a map of zones of full mutual interpredictability between cells. This is what we do in Table 29, where cells that can be predicted from one another with no uncertainty – i.e., with H = 0 in both directions – are given a same index – from Z1 to Z15, where Z is short for "zone" – and filled with the same colour. The colours are also meant to indicate how closely related different zones are in terms of predictability: different shades of the same colour are used when the interpredictability between the involved zones is high but not complete, i.e. with entropy values approaching 0, the average of the implicative entropy of predicting cell A from cell B and *vice versa* being below 0.1. Such areas of high, but not complete interpredictability are the one comprising Z2, Z3, Z5 and Z11 (different shades of red), the one comprising Z7, Z8 and Z9 (different shades of blue), the one comprising Z1, Z12 and Z13 (different shades of yellow) and the one comprising Z14 and Z15 (different shades of grey). In all these cases, the average of the implicative entropy values estimating the uncertainty in predicting zones in the same area from one another – i.e., predicting Z14 from Z15 and predicting Z15 from Z14 – is below 0.1. On the other hand, Z4 and Z6 are more isolated, and are not in a relation of high interpredictability with any of the other zones, because of opacities like the ones discussed in the previous section, as exemplified in Table 25 for Z4.

Table 29: Zones of interpredictability in Latin verb paradigms

a. verbal forms

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

b. nominal forms

PRS.INF.ACT	Z2
PRS.INF.PASS	Z11
PRF.INF.ACT	Z10
GER.GEN	Z12
GER.DAT	Z12
GER.ACC	Z12
GER.ABL	Z12
SUP.ACC	Z14
SUP.ABL	Z14

c. adjectival forms

	GDV	PRS. PTCP	PRF. PTCP	FUT. PTCP
NOM.M.SG	Z12	Z13	Z14	Z15
NOM.F.SG	Z12	Z13	Z14	Z15
NOM.N.SG	Z12	Z13	Z14	Z15
GEN.M.SG	Z12	Z12	Z14	Z15
GEN.F.SG	Z12	Z12	Z14	Z15
GEN.N.SG	Z12	Z12	Z14	Z15
DAT.M.SG	Z12	Z12	Z14	Z15
DAT.F.SG	Z12	Z12	Z14	Z15
DAT.N.SG	Z12	Z12	Z14	Z15
ACC.M.SG	Z12	Z12	Z14	Z15
VOC.F.SG	Z12	Z13	Z14	Z15
VOC.N.SG	Z12	Z13	Z14	Z15
ABL.M.SG	Z12	Z12	Z14	Z15
ABL.F.SG	Z12	Z12	Z14	Z15
ABL.N.SG	Z12	Z12	Z14	Z15
NOM.M.PL	Z12	Z12	Z14	Z15
NOM.F.PL	Z12	Z12	Z14	Z15
NOM.N.PL	Z12	Z12	Z14	Z15
GEN.M.PL	Z12	Z12	Z14	Z15
GEN.F.PL	Z12	Z12	Z14	Z15
GEN.N.PL	Z12	Z12	Z14	Z15
DAT.M.PL	Z12	Z12	Z14	Z15
DAT.F.PL	Z12	Z12	Z14	Z15
DAT.N.PL	Z12	Z12	Z14	Z15
ACC.M.PL	Z12	Z12	Z14	Z15
ACC.F.PL	Z12	Z12	Z14	Z15
ACC.N.PL	Z12	Z12	Z14	Z15
VOC.M.PL	Z12	Z12	Z14	Z15
VOC.F.PL	Z12	Z12	Z14	Z15
VOC.N.PL	Z12	Z12	Z14	Z15
ABL.M.PL	Z12	Z12	Z14	Z15
ABL.F.PL	Z12	Z12	Z14	Z15
ABL.N.PL	Z12	Z12	Z14	Z15

If we compare this map to the one shown above in Table 3, drawn according to the traditional subdivision of the Latin verb paradigm in three zones on the basis of the stems that appear in different wordforms, it can be observed that the picture is much more complex in the present system – i.e., concerning wordforms based on the present stem. This happens because the map given in Table 29 also takes into account the impact of allomorphy in inflectional endings on predictability: therefore, the various opacities related to the conjugation system – as described in §4.4 – generate uncertainty in the PCFP, and consequently different zones in our table. Furthermore, two zones are found regarding wordforms that are traditionally described as being based on the third stem, because of the aforementioned

observation that there actually are a few verbs that display a different stem allomorph in the cells of the future participle (cf. §4.4 above).

The patterns of interpredictability that emerge from Table 29 sometimes correspond to cells that share a relevant portion of morphosyntactic content: for instance, Z1 comprises cells of the imperfect indicative, Z8 cells of the future indicative, Z9 cells of the present subjunctive, Z10 cells with a perfective meaning.

However, in other cases the set of cells that constitutes a same zone is morphomic, in that it cannot be considered as being defined by the sharing of morphosyntactic properties: this is clearly the case of Z2, that includes some – but not all – of the cells of the present indicative, imperative and infinitive, of the imperfect subjunctive and of the future imperative; on the other hand, Z8 includes cells with the same value of the morphosyntactic feature of number (plural) and person (3rd), but none of them can be taken as being the defining property of the zone, since all the other third-person plural wordforms belong to different zones.

It can be noticed that the present indicative is the sub-paradigm where mutual interpredictability is lowest: its 12 cells belong to 4 different zones, as can be seen from the fourth line of Table 29a. This is hardly surprising, given the high frequency of such paradigm cells, that can thus be more plausibly stored as such. This is also comparable to the situation that is found in verb inflection in most of the Romance languages, where there is much unpredictability in the present indicative, often due to the presence of different stem allomorphs, as it emerges from descriptions like e.g. Pirrelli (2000) and Montermini & Bonami (2013) for Italian, Bonami & Boyé (2003, 2014) for French.

Another notable fact is that passive forms are almost always predictable from their active counterparts (and *vice versa*), even when they are not predictable from other cells – cf. e.g. Z7, that only includes the cells PRS.ACT.IND.3SG and PRS.PASS.IND.3SG. The only exceptions are the cells PRS.ACT.IMP.2SG and PRS.PASS.INF: each of them constitutes a zone on its own right, and there is no mutual interpredictability with any other cells. In PRS.ACT.IMP.2SG, this is due to the fact that a few verbs exceptionally display a consonant-ending form that cannot be predicted from any other cell: cf. e.g. the imperative *duc* for the 3rd conjugation verb DUCO ‘to lead’ and *fac* for the mixed conjugation verb FACIO ‘to make’ – vs.

the regular imperatives in *-e* in those conjugations (e.g. *lege* for the 3rd conjugation verb LEGO 'to read' and *cape* for the mixed conjugation verb CAPIO 'to take').
As for PRS.PASS.INF, the uncertainty is caused by the opacities exemplified in Table 30 below.

Table 30: The cells PRS.PASS.INF and PRS.ACT.INF for some lexemes

lexeme (conj.)	PRS.PASS.INF	PRS.ACT.INF
PARIO 'to bring forth' (mixed)	*parī*	*parere*
DO 'to give' (irr-1st)	*darī*	*dare*
VERRO 'to scrape' (3rd)	*verrī*	*verrere*
FERO 'to bring' (irr.)	*ferrī*	*ferre*

If a PRS.PASS.INF ends in *-arī* , we are probably facing a verb belonging to the 3rd or mixed conjugation where *-ar-* is part of the stem, and *-ī* is the regular passive infinitive marker of verbs of those classes (e.g. PARIO in our table), thus PRS.ACT.INF can be obtained by substituting that ending with the one of the active infinitive (i.e., *-ere*), as is shown in the first line of the table. However, if we look at the PRS.PASS.INF *darī* in the second line, it can be observed that in this case the active indicative is simply obtained by replacing the final *-ī-* with *-e-*, since DO 'to give' is actually a 1st conjugation verb that exceptionally displays a short *-a-*, rather than a long one as all the other verbs of that conjugation (e.g. *amārī* in AMO 'to love'). A similar opacity arises for passive infinitives ending in *-errī*, that are usually inflected wordforms of 3rd conjugation verbs with *-ī-* marking the passive infinitive and *-err-* belonging to the stem, thus yielding a PRS.ACT.INF in *-errere* (e.g. VERRO in our table), but not in a case like the one of the irregular FERO 'to bring', where to the PRS.PASS.INF *ferrī* corresponds a PRS.ACT.INF *ferre*.
The next step of our analysis consists in exploiting the mapping given above to obtain a more compact version of the Latin verb paradigm, where we abstract away from all cases of full mutual interpredictability, and therefore only keep one cell for each paradigm zone. We call such an abridged version a **distillation** of the paradigm, following Stump & Finkel (2013) – where the term was first introduced – and Bonami (2014) and Beniamine (2018) – that exploit this notion in the entropy-based framework that is also used in this work.

From now on, we will present the results of our entropy-based analysis on a distillation of the verbal paradigm: this makes them much more manageable than if they were given for the full paradigm, and the loss of information is minimal, since we are simply conflating cells that are already known to be mutually interpredictable.

In Table 31, we give the information on the cells that we (more or less arbitrarily)[10] decide to keep for each of the paradigm zones. The entropy values that estimate the mutual interpredictability between the various zones are then given in Table 32, with different shades of grey to indicate higher and lower entropy values, as above in Table 21.

Table 31: Cells used as representative for each zone of interpredictability

zone	cell
Z1	IPRF.ACT.IND.3.SG
Z2	PRS.ACT.INF
Z3	PRS.ACT.IMP.2.SG
Z4	PRS.ACT.IND.1.SG
Z5	PRS.ACT.IND.2.SG
Z6	PRS.ACT.IND.3.SG
Z7	PRS.ACT.IND.3.PL
Z8	FUT.ACT.IND.3.SG
Z9	PRS.ACT.SBJV.3.SG
Z10	PRF.ACT.INF
Z11	PRS.PASS.INF
Z12	GER.GEN.SG
Z13	PRS.PTCP.M.NOM.SG
Z14	SUP.ACC
Z15	FUT.PTCP.M.NOM.SG

[10] One criterion that was sometimes adopted to select one cell rather than another one was the number of lexemes for which the involved paradigm cell is attested in LatInfLexi: where there is a difference, we keep the one for which this number is higher.

Table 32: Overall results on a distillation of Latin verb paradigms

	Z1	Z2	Z3	Z4	Z5	Z6	Z7	Z8	Z9	Z10	Z11	Z12	Z13	Z14	Z15
Z1		0.3901	0.4587	0.3647	0.4375	0.4458	0.3647	0.3718	0.3694	0.3572	0.3347	0	0	0.2318	0.2444
Z2	0.1556		0.0371	0.1528	0	0	0.1528	0.1556	0.1523	0.2527	0	0.1533	0.1254	0.2394	0.252
Z3	0.1644	0.004566		0.1489	0	0	0.1514	0.1646	0.1565	0.2627	0.005318	0.1482	0.1289	0.2432	0.2563
Z4	0.944	0.9365	1.089		1.049	1.051	0.9556	0.944	0.979	0.7114	1.027	0.95	0.953	0.613	0.6304
Z5	0.2388	0.004562	0.02267	0.2324		0	0.2142	0.2389	0.2334	0.255	0.005318	0.1891	0.1866	0.2468	0.2585
Z6	0.3948	0.2651	0.2769	0.3823	0.4778		0.383	0.3865	0.3818	0.3738	0.2598	0.3838	0.3562	0.3245	0.3267
Z7	0.00843	0.05695	0.12445	0.00859	0.10126	0.0902		0.00844	0.01848	0.3025	0.0553	0.05273	0	0.2058	0.2178
Z8	0	0.04837	0.11975	0	0.0967	0.0853	0		0	0.299	0.05	0	0	0.2026	0.2098
Z9	0.0379	0.08624	0.2411	0	0.2125	0.1272	0.0367	0.0379		0.315	0.1691	0.03333	0.0336	0.2245	0.2344
Z10	0.4653	0.4753	0.471	0.4731	0.4841	0.4817	0.4753	0.4746	0.475		0.3416	0.508	0.4456	0.331	0.3003
Z11	0.2703	0.1296	0.1556	0.1967	0.1295	0.052	0.2046	0.2703	0.1967	0.2264		0.2109	0.1583	0.2175	0.2281
Z12	0.004578	0.921	0.43	0.3652	0.4238	0.441	0.3674	0.3718	0.365	0.357	0.3418		0	0.2339	0.2507
Z13	0.03842	0.932	0.4612	0.3865	0.449	0.4526	0.389	0.398	0.3906	0.376	0.3464	0.03363		0.2563	0.2769
Z14	0.2433	0.264	0.2583	0.2617	0.2769	0.2866	0.3625	0.2534	0.2568	0.372	0.2343	0.3762	0.2712		0.00831
Z15	0.2411	0.2795	0.2646	0.2664	0.2737	0.2627	0.3057	0.2502	0.2598	0.3547	0.2472	0.333	0.257	0.00396	

Average implicative entropy: 0.278819295

From Table 32, it can be observed that there is considerable variation in the distance between the various zones in terms of interpredictability. On the one hand, there are zones that can be predicted from one another with very little uncertainty, sometimes even with H = 0 in one direction: this is the case for instance of Z12 and Z13, as can be seen in the table. On the other hand, in some cases quite high entropy values appear, sometimes even with H > 1.

As was hinted above, for each zone the values in the lines of the table can be taken as an indicator of its predictiveness, the ones in the columns of its predictability. Therefore, to obtain an indicator of the overall predictability and predictiveness of each zone, we compute the average of the entropy values in the column and in the line corresponding to the involved zone, respectively. We call these indicators "average predictability" and "average predictiveness": their values are reported in Table 33a-b, ranked from the most predictive/predictable cell to the least predictive/predictable one.

Table 33: Average predictability and predictiveness in verb paradigms

a.

zone	average 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 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

It can be observed that Z4 – the one that includes the active and passive first-person singular of the present active indicative – proves to be by far the least predictive on the content of other paradigm cells, because of the already mentioned opacity of the inflected wordforms contained in Z4 on the conjugation to which the involved verb belongs (cf. §4.4 above). As for predictability, the difference in entropy between the different zones appears to be much less relevant.

4.6 *n*-ary implicative entropy and principal parts

In this section, we will see how the situation changes when two or more cells are used as predictors, by presenting results on *n*-ary implicative entropy (cf. §2.4 above). For practical reasons, we will always work on the distillation of Latin verb paradigms already used above in §4.5. In Table 34, binary implicative entropy values are shown – i.e., the entropy values estimating the uncertainty in predicting each cell used as representative for its paradigm zone assuming knowledge of two cells belonging to different zones. In the last column, the average predictiveness of each combination of zones when guessing the cells used as representative of any of

the other zones is reported: our results are sorted according to this indicator, in ascending order – i.e., from the most informative pair, with the highest average predictiveness value, to the least informative pair, with the lowest value.

The first fact that clearly emerges from such results is that there is a dramatic decrease in average implicative entropy: with one predictor, it was around 0.28, with two predictors it drops to about 0.06. This is consistent with the findings of Bonami & Beniamine (2016) on French and Portuguese and of Beniamine (2018: 170 ff.) on a wider sample of languages, and it supports the claim that knowledge of multiple cells is useful in making the PCFP easier.

Let us now have a closer look at the pairs of cells that work better as predictors, focusing on the ones whose average predictiveness is less than 0.01. It can be observed that such highly predictive pairs always include one cell that belongs to the present system; the other cell of the pair belongs either to Z10 – corresponding to the wordforms of the perfect system – or to Z14/Z15 – corresponding to the nominal forms based on the third stem. On the other hand, the worst predictors – with $H > 0.1$ – are the pairs composed of two cells that are based on the same stem – either the present stem, ore the third stem in the case of the pair including Z14 and Z15.

The fact that knowledge of cells based on different stems proves to be more useful is not surprising at all, given the relevance of unpredictable stem allomorphy in Latin verb inflection, as summarized by the many different alternation patterns listed above in Table 13, Table 16 and Table 17. What is remarkable is that joint knowledge of a wordform based on the perfect stem (i.e., belonging to Z10) and another one based on the third stem (i.e., belonging either to Z14 or to Z15) seems to be much less helpful: the pairs including Z10 and Z14 on the one hand and Z10 and Z15 on the other one are ranked much lower in our table (in 70th and 71st position, respectively), displaying considerably higher entropy values, with $H > 0.01$ in both cases. This suggests that wordforms based on the perfect stem and on the third stem are more informative about each other than they are on wordforms based on the present stem. Therefore, at least one wordform belonging to the present system should be known in order to allow for more reliable predictions.

Table 34: Binary implicative entropy in verb paradigms

	Z1	Z2	Z3	Z4	Z5	Z6	Z7	Z8
Z11, Z15	0.000825	0	0.002226	0.00652	0	0	0.00481	0.000825
Z9, Z10	0	0.0041	0.004116	0	0.005306	0	0	0
Z8, Z10	0	0.0041	0.004116	0	0.005306	0	0	
Z2, Z15	0.00474		0.002115	0.006054	0	0	0.004482	0.00448
Z7, Z10	0	0.00411	0.004116	0	0.005306	0		0
Z11, Z14	0.001968	0	0.002234	0.006886	0	0	0.004887	0.001968
Z3, Z15	0.004707	0		0.006004	0	0	0.004433	0.004433
Z2, Z14	0.004826		0.002136	0.006428	0	0	0.004574	0.004566
Z3, Z14	0.00479	0		0.00638	0	0	0.004524	0.004524
Z5, Z15	0.00828	0	0.002085	0.00663		0	0.004482	0.005714
Z5, Z14	0.006496	0	0.002104	0.00687		0	0.00534	0.00534
Z7, Z15	0	0.006775	0.006775	0	0.00783	0		0
Z8, Z15	0	0.006767	0.006775	0	0.00783	0	0	
Z7, Z14	0	0.00754	0.00754	0	0.008606	0		0
Z8, Z14	0	0.007534	0.00754	0	0.008606	0	0	
Z9, Z15	0.01503	0.02176	0.01846	0	0.02266	0.015	0.01483	0.01578
Z9, Z14	0.01804	0.02132	0.01941	0	0.0254	0.01697	0.0138	0.01398
Z3, Z10	0.02863	0		0.02652	0	0	0.02501	0.02501
Z5, Z10	0.02835	0	0	0.02785		0	0.02635	0.02635
Z4, Z10	0.01956	0.02094	0.0182		0.02254	0.02026	0.0195	0.0195
Z2, Z10	0.03098		0	0.02815	0	0	0.02666	0.02658
Z10, Z11	0.03146	0	0	0.03018	0	0	0.02885	0.02705
Z6, Z14	0.03043	0.0252	0.02734	0.02328	0.02965		0.0227	0.02267
Z4, Z15	0.02556	0.03198	0.03017		0.0344	0.02657	0.02657	0.02632
Z6, Z15	0.03098	0.03247	0.03406	0.02745	0.0325		0.02795	0.02843
Z13, Z14	0.00288	0.03766	0.03522	0.03598	0.03687	0.0348	0.02623	0.02803
Z12, Z14	0	0.03928	0.03885	0.0375	0.03702	0.0344	0.02826	0.02841
Z1, Z14		0.03967	0.0378	0.0375	0.03702	0.03497	0.02826	0.02881
Z4, Z14	0.02917	0.0328	0.0321		0.03815	0.02953	0.02652	0.02545
Z1, Z15		0.03772	0.03824	0.03976	0.04236	0.03787	0.0266	0.03436
Z12, Z15	0	0.03696	0.04007	0.03992	0.04153	0.03757	0.02669	0.03378
Z13, Z15	0.00286	0.04242	0.04102	0.04175	0.0484	0.03925	0.03003	0.03738
Z6, Z10	0.0487	0.01686	0.02025	0.0414	0.017		0.0411	0.041
Z4, Z11	0	0	0.002392		0	0	0	0
Z11, Z12	0	0	0	0	0	0	0	0
Z9, Z11	0	0	0.002392	0	0	0	0	0
Z1, Z11		0	0.002392	0	0	0	0	0
Z7, Z11	0	0	0.002392	0	0	0		0
Z8, Z11	0	0	0	0	0	0	0	
Z1, Z3		0		0	0	0	0	0
Z1, Z5		0	0.00201	0		0	0	0
Z3, Z4	0.003399	0.0034			0	0	0	0.0034

Z9	Z10	Z11	Z12	Z13	Z14	Z15	predictiveness
0.00196	0.01758		0.007	0.007164	0.001875		0.003906538
		0.001287	0	0	0.01834	0.01921	0.004027615
0		0.001287	0	0	0.01942	0.01892	0.004088385
0.001803	0.01845	0	0.004726	0.00747	0		0.004178462
0		0.001287	0	0	0.02023	0.01976	0.004216077
0.003109	0.01764		0.004738	0.006863		0.00539	0.004283308
0.001806	0.02412	0	0.004684	0.007435	0		0.004432462
0.002874	0.01836	0	0.00475	0.007378		0.002443	0.004487308
0.002878	0.0241	0	0.004707	0.007343		0.002445	0.004745462
0.002935	0.01859	0	0.004726	0.01394	0		0.005183231
0.00396	0.01851	0	0.00475	0.01312		0.002445	0.005302692
0	0.0543	0.00849	0	0	0		0.006474615
0	0.05435	0.00849	0	0	0		0.006477846
0	0.055	0.007694	0	0		0.002634	0.006847231
0	0.05505	0.007694	0	0		0.002632	0.006850462
	0.0616	0.02496	0.01392	0.01473	0		0.018363846
	0.06177	0.02164	0.01752	0.01846		0.002632	0.019303231
0.02501		0	0.02856	0.009346	0.04538	0.0464	0.019989692
0.02635		0	0.0256	0.009346	0.04538	0.0464	0.020152
0.0195		0.01607	0.01964	0.01964	0.02475	0.02446	0.020350769
0.02658		0	0.02856	0.009315	0.0456	0.04657	0.020691923
0.02885			0.0309	0.01039	0.04446	0.0443	0.021264615
0.02338	0.0234	0.02574	0.03055	0.03638		0.002882	0.024892462
0.01243	0.0585	0.03525	0.0257	0.0263	0		0.027673077
0.0295	0.024	0.03406	0.0285	0.03412	0.0012455		0.028097346
0.03458	0.06046	0.03232	0.0027			0.004936	0.028666615
0.03458	0.06052	0.03323		0		0.00494	0.028999231
0.03537	0.06058	0.03452	0	0		0.002632	0.029010154
0.01332	0.0575	0.03214	0.03024	0.0315		0.002634	0.029311846
0.0374	0.05945	0.03662	0	0	0		0.030029231
0.03647	0.05884	0.03973		0	0		0.03012
0.04044	0.02309	0.04318	0.00559		0.002157		0.030582077
0.041		0.01343	0.04333	0.02672	0.04028	0.04053	0.0332
0	0.1589		0	0	0.1415	0.1548	0.035199385
0	0.1803			0	0.1433	0.1573	0.036992308
	0.176		0	0	0.1519	0.1621	0.037876308
0	0.1809		0	0	0.1509	0.1613	0.038114769
0	0.1777		0	0	0.1567	0.1638	0.038507077
0	0.1814		0	0	0.1613	0.1708	0.0395
0	0.2084	0	0	0	0.1646	0.1775	0.042346154
0	0.2065	0	0	0	0.1666	0.1796	0.04267
0	0.1984	0.00397	0	0	0.1633	0.1794	0.042713

	Z1	Z2	Z3	Z4	Z5	Z6	Z7	Z8
Z4, Z5	0.003397	0.003399	0.00201			0	0	0.003399
Z2, Z4	0		0.02664		0	0	0	0
Z2, Z12	0		0.01881	0	0	0	0	0
Z1, Z2			0.02664	0	0	0	0	0
Z3, Z12	0.00458	0.00458		0	0	0	0	0.00458
Z5, Z12	0.00458	0.00458	0	0		0	0	0.00458
Z5, Z8	0	0	0	0		0	0	
Z3, Z8	0	0		0	0	0	0	
Z3, Z9	0.004562	0.004566		0	0	0	0	0.004566
Z5, Z9	0.004562	0.004562	0.00201	0		0	0	0.004562
Z2, Z9	0		0.02664	0	0	0	0	0
Z2, Z7	0		0.02664	0	0	0		0
Z3, Z7	0.004562	0.004566		0	0	0		0.004566
Z5, Z7	0.004562	0.004562	0.00201	0		0		0.004562
Z2, Z8	0		0.0254	0	0	0	0	
Z11, Z13	0.03062	0	0.002392	0.03062	0	0	0.03062	0.03062
Z10, Z12	0	0.08203	0.075	0.0812	0.0778	0.08154	0.0783	0.0783
Z1, Z10		0.08167	0.07825	0.0811	0.0804	0.0824	0.0783	0.07886
Z3, Z13	0.02824	0.004597		0.02365	0	0	0.02365	0.02826
Z1, Z6		0.04572	0.04364	0	0.07306		0	0
Z2, Z13	0.02809		0.02681	0.02817	0	0	0.02817	0.02809
Z5, Z13	0.03217	0.004597	0.002024	0.02756		0	0.02756	0.03217
Z6, Z12	0.00458	0.05243	0.03918	0	0.06122		0	0.00458
Z4, Z6	0.003397	0.04926	0.04105		0.08777		0	0.003399
Z6, Z8	0	0.04572	0.04166	0	0.07306		0	
Z6, Z7	0.004562	0.0504	0.04364	0	0.07306			0.004562
Z6, Z9	0.004547	0.05026	0.04364	0	0.08813		0	0.004547
Z10, Z14	0.0679	0.08844	0.0864	0.06976	0.08746	0.0775	0.0767	0.074
Z10, Z15	0.0736	0.08246	0.0844	0.07294	0.0881	0.07855	0.07465	0.07623
Z10, Z13	0.0198	0.08276	0.0789	0.0998	0.0815	0.0827	0.0969	0.0974
Z4, Z12	0.003412	0.05212	0.08606		0.0809	0.08655	0	0.003412
Z8, Z12	0	0.04868	0.08606	0	0.0809	0.08575	0	
Z1, Z4		0.0485	0.1184		0.0971	0.08624	0	0
Z9, Z12	0.004578	0.05325	0.08606	0	0.0809	0.0865	0	0.004578
Z4, Z13	0.00342	0.05228	0.1063		0.0978	0.08685	0	0.003422
Z8, Z13	0	0.0487	0.1073	0	0.09735	0.08594	0	
Z7, Z12	0.00458	0.05328	0.08606	0	0.0809	0.08655		0.00458
Z1, Z9		0.04834	0.1216	0	0.09735	0.0861	0	0
Z4, Z8	0	0.04852	0.1166		0.0964	0.0854	0	
Z1, Z7		0.0485	0.1212	0	0.09735	0.0863		0
Z9, Z13	0.004578	0.05328	0.10956	0	0.098	0.0868	0	0.00458
Z7, Z13	0.004593	0.05344	0.1092	0	0.098	0.0869		0.004593
Z1, Z8		0.04837	0.11975	0	0.0967	0.0853	0	

Z9	Z10	Z11	Z12	Z13	Z14	Z15	predictiveness
0	0.1967	0.00397	0	0	0.165	0.1796	0.042882692
0	0.1964	0	0	0	0.1616	0.1748	0.043033846
0	0.2086	0		0	0.1638	0.1748	0.043539231
0	0.2087	0	0	0	0.1646	0.1774	0.044410769
0	0.2122	0.0052		0	0.1676	0.1787	0.044418462
0	0.2104	0.0052		0	0.1697	0.1808	0.044603077
0	0.2197	0	0	0	0.1779	0.1849	0.044807692
0	0.2216	0	0	0	0.1763	0.1848	0.044823077
	0.2175	0.00397	0	0	0.1721	0.1835	0.045443385
	0.2157	0.00397	0	0	0.1738	0.1836	0.045597385
	0.2146	0	0	0	0.1704	0.1816	0.045633846
0	0.2148	0	0	0	0.1765	0.1843	0.046326154
0	0.2185	0.005318	0	0	0.1804	0.1882	0.046624
0	0.2167	0.005318	0	0	0.1821	0.1884	0.046785692
0	0.2222	0	0	0	0.1766	0.1852	0.046876923
0.03062	0.1819		0.03003		0.1617	0.1823	0.054724769
0.0776		0.06616		0	0.03317	0.03317	0.05879
0.07776		0.06616	0	0	0.03207	0.03302	0.05923
0.02365	0.2139	0.005318	0.0282		0.1831	0.2097	0.059405
0	0.2163	0.05	0	0	0.1696	0.1849	0.060247692
0.02809	0.2106	0	0.02818		0.1791	0.2056	0.060838462
0.02756	0.2122	0.005318	0.02814		0.1832	0.2098	0.060946077
0	0.22	0.05518		0	0.1753	0.1871	0.061505385
0	0.2073	0.05396	0	0	0.1716	0.1859	0.061818154
0	0.2302	0.05	0	0	0.1804	0.1873	0.06218
0	0.2284	0.0553	0	0	0.1842	0.1909	0.064232615
	0.2263	0.05396	0	0	0.1793	0.1871	0.064444923
0.0767		0.06244	0.0876	0.0678		0	0.070976923
0.0765		0.06274	0.08887	0.06946	0		0.071423077
0.0967		0.0668	0.01987		0.05545	0.05423	0.071754615
0	0.2585	0.05386		0	0.1743	0.1907	0.076139538
0	0.278	0.05		0	0.1785	0.1896	0.07673
0	0.2563	0.05	0	0	0.1705	0.187	0.078003077
	0.2756	0.05386		0	0.1793	0.192	0.078202
0	0.258	0.05396	0		0.1666	0.1913	0.078456308
0	0.2776	0.05	0		0.1683	0.1866	0.078599231
0	0.2769	0.05518		0	0.1873	0.199	0.079563846
	0.2715	0.05	0	0	0.1776	0.191	0.080268462
0	0.2742	0.05	0	0	0.1831	0.1929	0.080547692
0	0.2727	0.05	0	0	0.1819	0.194	0.080919231
	0.2751	0.05396	0		0.1737	0.1941	0.081050615
0	0.2764	0.0553	0		0.1774	0.1962	0.081694308
0	0.2776	0.05	0	0	0.1864	0.1981	0.081709231

	Z1	Z2	Z3	Z4	Z5	Z6	Z7	Z8
Z4, Z7	0.003397	0.0519	0.1181		0.0971	0.08624		0.003397
Z6, Z13	0.0372	0.05893	0.05045	0.03265	0.08044		0.0327	0.0372
Z8, Z9	0	0.04837	0.11975	0	0.0967	0.0853	0	
Z7, Z8	0	0.04852	0.11945	0	0.0967	0.0855		
Z7, Z9	0.00456	0.05307	0.1212	0	0.09735	0.0863		0.004562
Z3, Z11	0.1506	0		0.1506	0	0	0.1506	0.1506
Z3, Z6	0.157	0.004566		0.1373	0		0.146	0.1506
Z2, Z6	0.1555		0.02267	0.1465	0		0.1465	0.1461
Z6, Z11	0.1632	0	0.02608	0.1586	0		0.1586	0.156
Z4, Z9	0.0366	0.0851	0.2379		0.2122	0.1271	0.0367	0.0366
Z2, Z3	0.1461			0.1423	0	0	0.1423	0.1461
Z5, Z11	0.1622	0	0.02608	0.163		0	0.163	0.1622
Z2, Z11	0.1632		0.02608	0.1632	0	0	0.1632	0.1632
Z2, Z5	0.1547		0.02267	0.1522		0	0.1522	0.1547
Z3, Z5	0.1637	0.004566		0.1487		0	0.1511	0.1637
Z5, Z6	0.2196	0.004562	0.02267	0.2068			0.2079	0.2125
Z14, Z15	0.2147	0.2378	0.2382	0.2374	0.2427	0.2467	0.2825	0.2222
Z1, Z13		0.3923	0.4368	0.3535	0.44	0.449	0.3535	0.3606
Z1, Z12		0.3877	0.4238	0.3652	0.4204	0.4373	0.3652	0.3674
Z12, Z13	0.004597	0.9243	0.426	0.3538	0.4233	0.439	0.3562	0.3608

Average implicative entropy: 0.064055767

Z9	Z10	Z11	Z12	Z13	Z14	Z15	predictiveness
0	0.2695	0.05396	0	0	0.1896	0.1971	0.082330308
0.0326	0.2218	0.063	0.03268		0.1887	0.2084	0.082826923
	0.2905	0.05	0	0	0.193	0.2002	0.083370769
0	0.2937	0.05	0	0	0.1969	0.2013	0.084005385
	0.2905	0.05396	0	0	0.1971	0.2026	0.085477077
0.1506	0.1875		0.1565	0.1276	0.1942	0.2133	0.125546154
0.146	0.2426	0.005318	0.1431	0.129	0.2054	0.22	0.129760308
0.1461	0.2379	0	0.1471	0.1254	0.2002	0.2139	0.129836154
0.1586	0.2021		0.1586	0.1342	0.1827	0.1915	0.130013846
	0.2942	0.1691	0.03333	0.03342	0.2047	0.2205	0.132880769
0.1423	0.2491	0	0.1482	0.1205	0.239	0.2517	0.132892308
0.163	0.1885		0.1632	0.1342	0.1954	0.2135	0.133406154
0.1632	0.1906		0.1632	0.1342	0.1945	0.2133	0.133683077
0.1522	0.2505	0	0.1534	0.1257	0.2385	0.252	0.139136154
0.1562	0.251	0.005318	0.1482	0.129	0.2424	0.2563	0.140014154
0.2079	0.2407	0.005318	0.1813	0.1866	0.2058	0.2201	0.163211538
0.2377	0.329	0.2256	0.3103	0.2261			0.250069231
0.358	0.3567	0.3298	0		0.2114	0.2313	0.328684615
0.365	0.354	0.3347		0	0.2219	0.234	0.328969231
0.3535	0.3562	0.3374			0.2213	0.246	0.369415154

These observations on the different informativity of various pairs of cells bring us to Bonami & Beniamine (2016: 175)'s remark that *n*-ary implicative entropy values can be used to recognize pairs of cells that work as principal parts: they are the ones that display an average predictiveness of exactly 0, meaning that if they are assumed to be known, then the rest of the paradigm can be inferred with no uncertainty whatsoever. In Table 34, there is no pair of cells with average predictiveness of exactly 0: therefore, none of them can be taken as a completely reliable principal part set.

However, Bonami & Beniamine (2016: 176) also notice that it can be useful to focus not only on categorical inferences, but also on what they call "near principal parts" – i.e., pairs of cells whose average predictiveness at least approaches 0. Indeed, in languages with a rich and complex inflectional morphology comparable to the one of Latin verbs, not even adult native speakers reach a complete mastery of the system, and they are sometimes observed to make mistakes when producing unknown wordforms. Therefore, it is reasonable to take into account also cases where uncertainty is very low, although it is not absent.

For instance, in Table 35 there are 15 pairs of cells whose average predictiveness is less than 0.01. Such values indicate a very low level of uncertainty: to get an idea, such an entropy value roughly corresponds to the uncertainty in guessing the value of a binary random variable where one of the outcomes has a probability of 98,7%. Therefore, although from such pairs it is not possible to infer the paradigm cells of all the other zones with absolute certainty, they nevertheless allow for a very relevant reduction in uncertainty, and they can therefore be reasonably considered as near principal parts.

In Table 35, we summarize the results obtained by using up to 5 cells as predictors, by reporting the average *n*-ary implicative entropy value and the number of principal part sets and near principal part sets at different "cardinalities" – i.e., with different numbers of predictors, with Bonami & Beniamine (2016)'s terminology – and with different thresholds of predictiveness that need to be satisfied for a given combination of cells to be considered as a near principal part set – $H < 0.001$, $H < 0.01$ and $H < 0.1$.

Table 35: Verb paradigms: principal part sets and near principal part sets of different cardinalities

cardinality	average implic. entropy	principal parts (H = 0)		near principal parts (H < 0.001)		(H < 0.01)		(H < 0.1)	
		n.	%	n.	%	n.	%	n.	%
2	0.06	0	0	0	0	15	14.3%	90	85.7%
3	0.03	0	0	15	3.3%	196	43.1%	444	97.6
4	0.02	56	4.1%	122	8.9%	834	61.1%	1,360	99.6%
5	0.01	336	11.2%	471	15.7%	2,190	72.9%	3,001	99.9%

A first general observation is that the more paradigm cells we use as predictors, the better the overall predictability in the inflectional system: the average implicative entropy decreases, and the number of combinations of cells that work as principal parts increases, both in absolute terms and – more interestingly – in percentage on the number of available combinations.

Furthermore, it can be observed that no categorical principal part set can be found with less than 4 predictors. The principal part sets of cardinality 4 include two cells based on the present stem, one based on the perfect stem ad one based on the third stem, exactly like the sets of cells that are traditionally used in Latin grammars and lexicons.[11] Therefore, our results support Finkel & Stump (2009b)'s crucial claim that gives the title to their paper, i.e. that "What your teacher told you is true: Latin verbs have four principal parts". The traditional analysis is thus confirmed by different works conducted in a more principled, systematic and data-driven way, even if they are performed with different methodologies and datasets.

However, the approach used in the current work is different than both the traditional description and Finkel & Stump (2009b)'s Principal Part Analysis in that it allows to go beyond categorical inference of the rest of the paradigm, focusing also on near principal part sets, from which the remaining cells can be predicted with very little uncertainty. In this way, we can see that with three predictors, we do find near principal part sets even with a very low threshold of average predictiveness (H <

[11] It should be noticed that there are categorical principal part sets of cardinality 4, despite the presence of verbs for which the future participle is based on a stem different than the one of the supine and perfect participle, as shown in Table 20. This is due to subtle restrictions on the context of application of the relevant patterns computed by the Qumin toolkit for the various pairs of wordforms involved. As was hinted above (cf Footnote 6 and 7 of this chapter), such restrictions can have the effect of reducing the number of lexemes where there is competition between different patterns. When the verbs with a different inflectional behaviour are few, competition can be completely removed, as happens in this case.

0.001, roughly corresponding to the uncertainty in guessing the value of a binary random variable where one of the outcomes has a probability of 99,99%). Given the very high predictiveness of such cells, it is reasonable to assume that speakers can take advantage of such possible inferences, although they are not categorical.

4.7 Conclusion

In this chapter, we have presented our results on Latin verb inflection regarding implicative relations between paradigm cells and their mutual predictability, as well as their consequences on paradigm structure. After having summarized the facts in §4.1, in §4.2 we have introduced Boyé & Schalchli (2016)'s notion of cell paradigm, where all systematically syncretic cells are conflated, thus producing a reduced version of the verb paradigm that can be used as the point of departure of our analysis. In §4.3 and §4.4, we have discussed the impact on predictability of stem allomorphy and inflection classes, respectively, presenting results regarding wordforms based on different stems in §4.3 and regarding wordforms based on the same stem in §4.4. We have seen that despite the considerable formal variation in the alternation patterns between wordforms based on the same stem, entropy is not too high between such cells; on the other hand, we do sometimes find relevant entropy values between cells based on the present stem, because of the opacities generated by heteroclisis and by other less systematic aspect of the Latin conjugation system. In §4.5, we have exploited our results to obtain a mapping of the Latin verb paradigm in zones of full mutual interpredictability, where cells can be predicted from one another with no uncertainty (i.e., with H = 0 in both directions). We have then presented results obtained on a distillation of the paradigm – i.e., a reduced version that only contains one cell for each zone – concerning both unary implicative entropy and *n*-ary implicative entropy (cf. §4.6), with up to five cells used as predictors. *n*-ary implicative entropy results have been used to find combinations of cells from which the rest of the paradigm can be predicted with no uncertainty whatsoever – principal parts – or with very little uncertainty – near principal parts. Our findings are consistent with both the traditional analysis and Finkel & Stump (2009b)'s account in requiring 4

categorical principal parts. However, near principal part sets can be found even with lower number of predictors: with 3 predictors, there are combinations of cells from which the rest of the paradigm can be inferred with the threshold set at 0.001, and already with 2 predictors there are near principal part sets if the threshold is set at 0.01.

Chapter 5. Predictability in Latin noun inflection and the role of gender

We will now switch our focus from verbs to nouns. The structure of the first part of this chapter is analogous to the previous one. In §5.1, we provide some preliminary information on the inflectional behaviour of Latin nouns, as it is described in traditional grammars. We also review some (more or less) recent theoretical account of the well-known facts. In §5.2, we present the results of our systematic analysis of implicative relations and predictability in noun inflection. The shape of the cell paradigm of Latin nouns – on which our computations are done – is shown in §5.2.1. The results concerning unary and *n*-ary implicative entropy are then given in turn in §5.2.2 and §5.2.3. In the latter section, the principal parts and near principal parts that can be inferred from the entropy-based analysis are discussed.
The remaining part of the chapter is devoted to a topic that, unlike verbs, proves to be relevant for nouns, namely the role of information on a lexeme's gender in reducing uncertainty in inflectional predictions: how do entropy values change if we assume that we do not only know the phonotactic shape of an inflected wordform, but also the gender of the lexeme it belongs to? We begin by providing some examples that show that information on gender is at least potentially available to speakers and capable of yielding a reduction in implicative entropy (§5.3.1). We then briefly show some quantitative data on the distribution of Latin nouns among the three genders, both in absolute terms and in their relationship with the classification in the traditional five declensions, highlighting some clear preferential associations between a noun's declension and its gender and their impact on uncertainty in the PCFP (§5.3.2). We finally present the results that are obtained by taking gender information into account in §5.3.3, discussing the relevance of such results in the debate on the function(s) of gender, but also some caveats that should be made on their reliability, in §5.3.4. The main findings of the chapter are then summarized in §5.4.

5.1 Latin noun inflection: the traditional account

Let us begin with a brief synopsis of the main facts regarding Latin noun inflection, based on the same traditional descriptions that have already been used as a source in Chapter 4: grammars like Bennett (1908), Leumann et al. (1977), and Ernout (1914)'s historical morphology. In addition to these descriptive works, also the study specifically devoted by Risch (1977) to the Latin declension system has been used as a source in this chapter.

Latin nouns are inflected for number – singular and plural – and for case – the six main values of the category of case being nominative, genitive, dative, accusative, vocative and ablative. Additionally, nouns denoting towns and small islands and a few other lexemes also display a locative form in the singular; however, because of its marginality in terms of the number of lexemes in which it is attested, the locative will be excluded from this synopsis, and also from the analysis performed in this chapter.

Traditional descriptions agree on the presence of five major declensions in Latin noun inflection. The basis of the classification is the cell GEN.SG, where all these five declensions differ in their endings: *-ae* in the 1st declension, *-ī* in the 2nd, *-is* in the 3rd, *-ūs* in the 4th and *-eī* in the 5th. The realizations of the other cells are summarized in Table 1.

Table 1: The five Latin declensions

	ROSA 'rose' (1st decl.)	LUPUS 'wolf' (2nd decl.)	CONSUL 'consul' (3rd decl.)	FRUCTUS 'fruit' (4th decl.)	RES 'thing' (5th decl.)
NOM.SG	*rosa*	*lupus*	*consul*	*fructus*	*rēs*
GEN.SG	*rosae*	*lupī*	*consulis*	*fructūs*	*reī*
DAT.SG	*rosae*	*lupō*	*consulī*	*fructuī*	*reī*
ACC.SG	*rosam*	*lupum*	*consulem*	*fructum*	*rem*
VOC.SG	*rosa*	*lupe*	*consul*	*fructus*	*rēs*
ABL.SG	*rosā*	*lupō*	*consule*	*fructū*	*rē*
NOM.PL	*rosae*	*lupī*	*consulēs*	*fructūs*	*rēs*
GEN.PL	*rosārum*	*lupōrum*	*consulum*	*fructuum*	*rērum*
DAT.PL	*rosīs*	*lupīs*	*consulibus*	*fructibus*	*rēbus*
ACC.PL	*rosās*	*lupōs*	*consulēs*	*fructūs*	*rēs*
VOC.PL	*rosae*	*lupī*	*consulēs*	*fructūs*	*rēs*
ABL.PL	*rosīs*	*lupīs*	*consulibus*	*fructibus*	*rēbus*

However, this classification does not cover all the formal variation in exponents that can be found in Latin noun inflection. An important aspect to consider is the presence of gender-based subclasses, at least inside some declensions. As we will see in more detail below (cf. §5.3 below), Latin nouns can be assigned to one of three genders: masculine, feminine and neuter.[1] Masculine and feminine nouns that belong to the same declension do not display any difference in their inflectional behaviour. On the other hand, a notable characteristic of Latin noun inflection is the systematic syncretism of the nominative, accusative and vocative in neuter nouns, both in the singular and in the plural. In the plural, neuter nouns also share a specific, dedicated ending *-a*. This segment is found in those cells in all the declensions that contain neuter nouns – namely, the 2nd, 3rd and 4th – and it does not appear in masculine and feminine nouns.[2]

A consequence of these facts is that neuter nouns of the 2nd, 3rd and 4th declension will display realizations that are different from the ones shown in Table 1, at least in the aforementioned cells. On the other hand, the endings of the genitive, dative and ablative do not change in nouns of the same declension with different gender, with the exception of the DAT.SG of nouns of the 4th declension, that end in *-uī* in masculine and feminine nouns and in *-ū* in neuter nouns.

These differences in the inflectional behaviour of masculine/feminine and neuter nouns that belong to the same declension are summarized in Table 2, where cells whose realizations vary on the basis of gender are highlighted in grey. In the 1st and 5th declension, there are no such sub-classes, simply because there are no neuter nouns.

[1] Some lemmas (401 in Lemlat's lemma list, if lemmas from the Onomasticon and from Du Cange (1883-1887) are excluded) can be assigned to more than one gender: for instance, they can be both masculine and feminine, as. e.g. CUSTOS (M/F) 'protector/protectress'.

[2] On the relationship between gender and inflection classes in Latin, and on the generalizations that can be drawn on the inflectional behaviour of Latin nouns of different gender, cf. Aronoff 1994: 79-85.

Table 2: Gender-based sub-classes of the 2nd, 3rd and 4th declension

	LUPUS ‘wolf’ (2nd, masc.)	BELLUM ‘war’ (2nd, neut.)	CONSUL ‘consul’ (3rd, masc.)	ANIMAL ‘animal’ (3rd, neut.)	FRUCTUS ‘fruit’ (4th, masc.)	CORNUS ‘horn’ (4th, neut.)
NOM.SG	*lupus*	*bellum*	*consul*	*animal*	*fructus*	*cornū*
GEN.SG	*lupī*	*bellī*	*consulis*	*animālis*	*fructūs*	*cornūs*
DAT.SG	*lupō*	*bellō*	*consulī*	*animālī*	*fructuī*	*cornū*
ACC.SG	*lupum*	*bellum*	*consulem*	*animal*	*fructum*	*cornū*
VOC.SG	*lupe*	*bellum*	*consul*	*animal*	*fructus*	*cornū*
ABL.SG	*lupō*	*bellō*	*consule*	*animāle*	*fructū*	*cornū*
NOM.PL	*lupī*	*bella*	*consulēs*	*animālia*	*fructūs*	*cornua*
GEN.PL	*lupōrum*	*bellōrum*	*consulum*	*animālum*	*fructuum*	*cornuum*
DAT.PL	*lupīs*	*bellīs*	*consulibus*	*animālibus*	*fructibus*	*cornibus*
ACC.PL	*lupōs*	*bella*	*consulēs*	*animālia*	*fructūs*	*cornua*
VOC.PL	*lupī*	*bella*	*consulēs*	*animālia*	*fructus*	*cornua*
ABL.PL	*lupīs*	*bellīs*	*consulibus*	*animālibus*	*fructibus*	*cornibus*

Alongside this gender-based distinction, there are also other relevant sub-classes within the 2nd, 3rd and 5th declension.

As for the 5th declension, two sub-classes should be individuated because the ending of the genitive and dative singular is *-eī* if it is preceded by a consonant (e.g. in the noun FIDES ‘trust’, GEN.SG and DAT.SG *fideī*), *-ēī* if it is preceded by a vowel (e.g. in the noun FACIES ‘figure’, GEN.SG and DAT.SG *faciēī*).

In the 2nd declension, there are two sub-classes of nouns that in the NOM.SG and VOC.SG end in *-er* – rather than in *-us* and *-e* as the noun given in Table 1. In nouns like PUER, the stem used in these two cells is the same that appears in the rest of the paradigm, while in nouns like LIBER there is also stem allomorphy in such cells: we find the stem *liber-* in NOM.SG and VOC.SG (highlighted in grey), different from the stem *libr-* of the other cells.

Table 3: Nouns in *-er* of the 2nd declension

	PUER ‘child’ (2nd decl., *-er*)	LIBER ‘book’ (2nd decl., *-er*, with stem allomorphy)
NOM.SG	*puer*	*liber*
GEN.SG	*puerī*	*librī*
DAT.SG	*puerō*	*librō*
ACC.SG	*puerum*	*librum*
VOC.SG	*puer*	*liber*
ABL.SG	*puerō*	*librō*
NOM.PL	*puerī*	*librī*
GEN.PL	*puerōrum*	*librōrum*
DAT.PL	*puerīs*	*librīs*
ACC.PL	*puerōs*	*librōs*
VOC.PL	*puerī*	*librī*
ABL.PL	*puerīs*	*librīs*

Furthermore, there is another sub-class to account for the slightly different inflectional behaviour of nouns in *-ius* like FILIUS, displaying VOC.SG *filī* (vs. *-e* in regular masculine 2nd declension nouns, e.g. *lupe*) and GEN.SG *filī* (alongside the regular *filiī*, cf. regular *lupī*).

Several further subdivisions need to be established within the traditional 3rd declension, where many differences in the actual inflectional realizations that are displayed by different lexemes can be found.

Firstly, in many cases there is (more or less) unpredictable stem allomorphy or even suppletion in the nominative and vocative singular – and, consequently, in neuter nouns also in the accusative singular, which is systematically syncretic with those cells. This is shown in Table 4, where only the singular cells of some representative lexemes are given, with cells that display stem allomorphy (or weak suppletion, as in FEMUR) again highlighted in grey.

Table 4: Stem allomorphy in 3rd declension nouns

	MILES ‘soldier’	PATER ‘father’	TEMPUS ‘time’	FEMUR ‘thigh’
NOM.SG	*miles*	*pater*	*tempus*	*femur*
GEN.SG	*militis*	*patris*	*temporis*	*feminis*
DAT.SG	*militī*	*patrī*	*temporī*	*feminī*
ACC.SG	*militem*	*patrem*	*tempus*	*femur*
VOC.SG	*miles*	*pater*	*tempus*	*femur*
ABL.SG	*milite*	*patre*	*tempore*	*femine*

Furthermore, in some cells there is also variation in the endings that are used in different lexemes that are traditionally assigned to the 3rd declension. Let us consider the data in Table 5.

Table 5: 3rd declension: consonantal declension and *-i-* declension

	REX 'king' (3rd, consonantal decl.)	PUPPIS 'poop' (3rd, *-i-* decl.)
NOM.SG	*rex*	*puppis*
GEN.SG	*regis*	*puppis*
DAT.SG	*regī*	*puppī*
ACC.SG	*regem*	*puppim*
VOC.SG	*rex*	*puppis*
ABL.SG	*rege*	*puppī*
NOM.PL	*regēs*	*puppēs*
GEN.PL	*regum*	*puppium*
DAT.PL	*regibus*	*puppibus*
ACC.PL	*regēs*	*puppīs* (also *puppēs)*
VOC.PL	*regēs*	*puppēs*
ABL.PL	*regibus*	*puppibus*

The nouns REX and PUPPIS are very different in their inflectional behaviour, not only because of the aforementioned unpredictable stem allomorphy in the nominative and vocative singular, but also because of the different endings that are used in the cells highlighted in grey, namely ACC.SG (*-em* vs. *-im*), ABL.SG (*-e* vs. *-ī*), GEN.PL (*-um* vs. *-ium*) and ACC.PL (*-ēs* vs. *-īs*).

Following Wurzel (1984: 119), we will call the sub-class of PUPPIS "*-i-* declension" and the one of REX "consonantal declension". Such labels refer to the fact that, from an historical perspective, the former sub-class contains *-i-* stems (hence, e.g., ACC.SG *puppi-m*), the latter consonantal stems (hence ACC.SG *reg-em*). However, what matters in the context of this work are the different realizations as they surface in synchrony, whatever their diachronic origins.

While PUPPIS and REX exemplify the "pure" *-i-* declension and consonantal declension respectively, since the former only displays forms with *-i-* and the latter only forms without *-i-* in the involved cells, there are also other nouns that are not so straightforward in their inflectional behaviour. This is exemplified by AURIS in Table 6, ending in *-em* and *-e* in the accusative and ablative singular (like in the

consonantal declension) and in *-ium* and *-īs* in the genitive and accusative plural (like in the *-i-* declension).

Table 6: A mixed *-i-*/consonantal declension noun

	AURIS 'ear' (3rd, mixed *-i-*/consonantal decl.)
NOM.SG	*auris*
GEN.SG	*auris*
DAT.SG	*aurī*
ACC.SG	*aurem*
VOC.SG	*auris*
ABL.SG	*aure*
NOM.SG	*aurēs*
GEN.SG	*aurium*
DAT.SG	*auribus*
ACC.SG	*aurīs*
VOC.SG	*aurēs*
ABL.SG	*auribus*

However, different combinations of endings of the pure *-i-* and consonantal sub-classes are also possible; the picture is quite complex and has consequently been devoted considerable attention in the literature. A detailed description of the facts – i.e., of the endings that are actually used in different 3rd declension nouns – is provided by Janson (1971), while Carstairs (1984) is an attempt to explain the diachronic changes that took place in 3rd declension nouns on the basis of the need to restore the "Paradigm Economy Principle".[3]

However, it is Wurzel (1984)'s treatment of such facts that proves to be particularly interesting in the context of the present work, since it is claimed that implicative relations among inflected words of the 3rd declension can account for the endings that appear in different lexemes. Beside the pure *-i-* declension and the pure consonantal declension, three mixed classes are individuated, where endings of the two sub-classes appear in different cells, as exemplified in Table 7.

[3] But see Nyman (1987) for a critique of that account.

Table 7: *-i-* declension, consonantal declension and mixed classes in Wurzel (1984)

	-i-: PUPPIS 'poop'	mixed (a): IGNIS 'fire'	mixed (b): AURIS 'ear'	mixed (c): CIVIS 'citizen'	consonant: REX 'king'
ACC.SG	*puppim*	*ignem*	*aurem*	*civem*	*regem*
	↓		↑	↑	↑
ABL.SG	*puppī*	*ignī*	*aure*	*cive*	*rege*
	↓	↓		↑	↑
ACC.PL	*puppīs*	*ignīs*	*aurīs*	*civēs*	*regēs*
	↓	↓	↓		↑
GEN.PL	*puppium*	*ignium*	*aurium*	*civium*	*regum*

The crucial claim of Wurzel (1984) is that the actual occurrence of the different endings can be inferred based on the interaction between two very general implicative chains, that can be expressed as in (1).

(1) a. X*im*, ACC.SG → X*ī*, ABL.SG → X*īs*, ACC.PL → X*ium*, GEN.PL

b. X*um*, GEN.PL → X*ēs*, ACC.PL → X*e*, ABL.SG → X*em*, ACC.SG

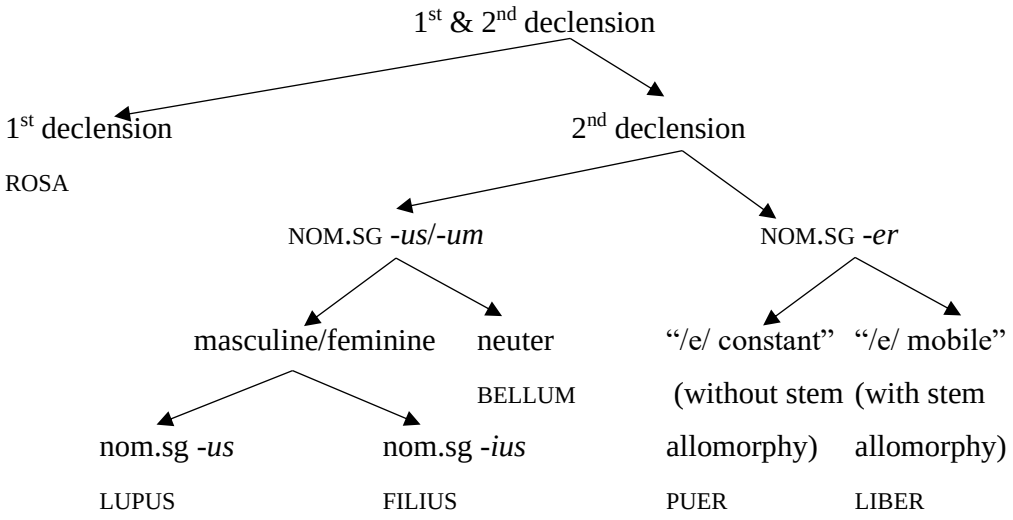
(from Wurzel 1984: 120, with adaptations)

In Wurzel's account, such chains of implicative relations make it possible to minimize the number of cells that need to be known in order to fill the whole paradigm of a nominal lexeme. In nouns of the pure *-i-* declension, starting from an ACC.SG in *-im*, all the other endings can be inferred based on the chain of implicative relations in (1a); conversely, in nouns of the pure consonantal declension, starting from a GEN.PL in *-um*, all the other endings can be inferred based on the chain of implicative relations in (1b).

On the other hand, in the mixed declension of AURIS, the ACC.SG *aurem* can be inferred from the ABL.SG *aure* based on the implicative chain (1b), while the GEN.PL *aurium* can be inferred from the ACC.PL *aurīs* based on the implication (1a). In a similar way, in the mixed classes (a) and (c) some of the implicative relations stated in (1a-b) also make it possible to reduce the number of forms that need to be known in order to fill the remaining paradigm cells, as is shown by the arrows of Table 7.

Having presented the main facts concerning the inflectional behaviour of Latin nouns, we are now in a position to review how such facts are accounted for in the detailed analysis of inflectional micro- and macro-classes provided by Dressler (2002). The 1st and 2nd declension are considered by Dressler as two sub-classes of a same macro-class, as shown in Figure 1.

Figure 1: Latin noun inflection: macro-class I



(from Dressler 2002, with modifications)

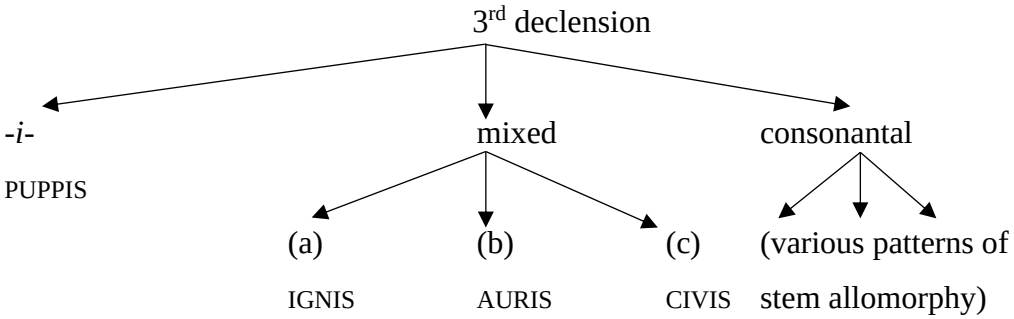
This choice is due to the fact that these two declensions have many points in common: as can be seen from the examples of Table 1, the exponent of the dative and ablative plural is the same (*-īs*), and other cells display the same ending, the only difference being the theme vowel that precedes it (e.g. ACC.PL *-am/-um*, GEN.PL *-ārum/-ōrum*, etc.).

Going down the hierarchy, while in the micro-class of 1st declension nouns we do not need further subdivisions, the one of 2nd declension nouns includes several sub-classes. The first branching is between nouns with NOM.SG in *-er* and other nouns. Within nouns in *-er*, we find two micro-classes, one for nouns like PUER in Table 3 – "/e/ constant" in Dressler's terminology – and another one for nouns like LIBER – "/e/ mobile".

Within other 2nd declension nouns, two sub-classes are individuated on the basis of gender, with masculine nouns in *-us* (like LUPUS in Table 2) and neuter nouns in *-um* (like BELLUM in Table 2). Lastly, within the class of masculine nouns in *-us* there is a further branching to account for the slightly different inflectional behaviour of nouns in *-ius* (cf. above, p. 132).

In Dressler (2002)'s account, the other declensions – 3rd, 4th and 5th – constitute macro-classes in their own right. Let us start from the macro-class of 3rd declension nouns, whose hierarchy of micro-classes is given here in Figure 2.

Figure 2: Latin noun inflection: macro-class II



(from Dressler 2002, with simplifications)

Here, we find three main branches: the first node on the left contains nouns of the pure *-i-* declension; the last one on the right contains nouns of the pure consonantal declension, including all the micro-classes generated by different patterns of stem allomorphy in NOM.SG, ACC.SG and VOC.SG. Lastly, all the mixed types showed in Table 7 are covered by a single macro-class, which in Dressler's account is simply characterised by the possibility of displaying *-i-* endings in ACC.PL, GEN.PL and ABL.SG, with the actual occurrence of such forms being regulated by other factors, including "Paradigm Structure Conditions" – i.e., implicative relations like the ones proposed by Wurzel (1984) and showed above in (1a-b).

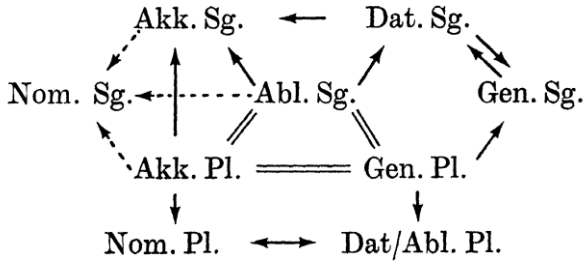
The situation in the remaining macro-classes is far less complex: the macro-class of the 4th declension only contains two micro-classes, one for masculine and feminine nouns in *-us* and one for neuter nouns in *-ū* (cf. Table 2), while for the one

of the 5th declension Dressler (2002) does not mention any further subdivision – despite the abovementioned distinction between nouns like FIDES and nouns like FACIES in GEN.SG and DAT.SG, cf. above, p. 131.

Having dealt with inflection classes and sub-classes of Latin nouns, we can now turn to the matter of how they can be revealed by means of principal parts. Traditionally, the cells NOM.SG and GEN.SG are used as principal parts in grammatical and lexicographical sources. As we saw at the beginning of this section, the different endings used in GEN.SG allow us to know to which of the five major declensions the involved noun belongs. Regarding NOM.SG, it is one of the cells where neuter nouns differ in their inflectional behaviour from masculine and feminine nouns (cf. Table 2), thus revealing the gender-based inflectional sub-class of the involved noun. Furthermore, regarding nouns in *-er* of the 2nd declension (cf. Table 3) and 3rd declension nouns displaying unpredictable stem allomorphy (cf. Table 4), NOM.SG is one of the cells that are based on a stem different from the one of GEN.SG: therefore, inflected words based on different stem allomorphs can be predicted once the content of these two cells is known. On the other hand, some of the variation in the endings of nouns of the *-i-* vs. consonantal declension (and the various mixed classes), as described in Table 7, cannot be predicted from these principal parts: for instance, the different endings used by AURIS and PUPPIS in ACC.SG (*aurem* vs. *puppim*) and ABL.SG (*aure* vs. *puppī*) cannot be guessed from their NOM.SG and GEN.SG, that are alike in their inflectional behaviour (*auris* like *puppis*). In the next sections – in particular, in §5.2.3 – data on *n*-ary implicative entropy will be exploited to obtain principal part sets for Latin in a more principled and data-driven way, and the relationship of the obtained results with the traditional description will be discussed in detail.

Another work on Latin nominal inflection that deserves to be mentioned in this context is Risch (1977): a section of that paper is devoted to a comprehensive picture of the implicative relations that can be exploited to fill the paradigm of a Latin noun. The overall account is summarized in a graph, that we reproduce here in Figure 3.

Figure 3: Risch (1977): a graph of implicative relations in Latin noun paradigms



(from Risch 1977: 236)

In Risch's graph, the three cells ABL.SG, ACC.PL and GEN.PL are considered to constitute the core of the Latin nominal paradigm – as is suggested by the double lines that connect them – from which all other cells can be inferred – as is illustrated by the arrows, with their direction representing the direction of the prediction. For instance, there is an arrow that starts from ACC.PL and goes to NOM.PL, since from the former cell we can predict the latter (see 2b), but not vice versa: NOM.PL fails to distinguish between the *-i-* declension on the one hand and the consonantal and 5th declension on the other one: in both cases, the NOM.PL ending is *-ēs*, but different realizations are displayed in ACC.PL (*-īs* in the *-i-* declension, *-ēs* in the 5th; see 2b).

(2) a. predicting NOM.PL from ACC.PL

-ās, ACC.PL → *-ae*, NOM.PL (1st decl.)

-ōs, ACC.PL → *-ī*, NOM.PL (2nd decl.)

-ūs, ACC.PL → *-ūs*, NOM.PL (4th decl.)

-īs, ACC.PL → *-ēs*, NOM.PL (*-i-* decl.)

-ēs, ACC.PL → *-ēs*, NOM.PL (consonantal and 5th decl.)

b. predicting ACC.PL from NOM.PL

-ae, NOM.PL → *-ās*, ACC.PL (1st decl.)

-ī, NOM.PL → *-ōs*, ACC.PL (2nd decl.)

-ūs, NOM.PL → *-ūs*, ACC.PL (4th decl.)

-ēs, NOM.PL → *-īs*, ACC.PL (*-i-* decl.)

-ēs, NOM.PL → *-ēs*, ACC.PL (consonantal and 5th decl.)

Risch's (1977) usage of implicative relations is much closer to ours than Wurzel's (1984). The latter proposes implicative relations that only work inside a given declension, while the former aims at implicative relations that are valid across declensions. This is similar to what happens in our entropy-based analysis, which is performed on a representative lexicon that of course includes nouns of different declensions. Nevertheless, the account of the implicative structure of Latin noun paradigms that emerge from our results – cf. below, §5.2.2 – is considerably different than the one proposed by Risch: in many cases, we find non-negligible levels of uncertainty in some cells that are linked – and therefore expected to display low entropy values, at least in one direction – in Risch's (1977) picture. However, a detailed comparison between these two accounts is made difficult by two systematic differences in the way in which the results are obtained. The first is the different methodology: Risch's approach can be considered constructive, in that its generalizations are stated in terms of exponents, rather than on full inflected words (cf. Chapter 1). This causes an underestimation of the uncertainty in guessing the content of some paradigm cells, particularly those where there is also unpredictable stem allomorphy. Our fully abstractive approach, on the other hand, is capable of capturing and quantifying this kind of uncertainty. The second difference is that the validity of Risch's account is more limited, since neuter nouns are left out of the picture (cf. Risch 1977: 236). Therefore, the uncertainty that is caused by the presence of gender-based sub-classes cannot be detected. Conversely, such uncertainty emerges clearly in our analysis, which is performed on all the nouns of LatInfLexi, including, of course, neuter ones.

5.2 Predictability in Latin noun inflection

5.2.1 The cell paradigm of Latin nouns

After having summarized in the previous section the main facts concerning Latin noun inflection and their treatment in some theoretically grounded accounts, in this section we can move to the analysis of the structure of the nominal paradigm as it

emerges from relations of interpredictability between inflected words. We saw in §4.2 that a first level of simplification in this sense can be achieved by replacing the tabular paradigm with the cell paradigm, where cells that are systematically syncretic across all lexemes are considered to constitute only one cell, since they are trivially predictable one from another with no uncertainty.

Regarding verbs, this allowed for a relevant reduction of the size of a very large paradigm. Noun paradigms are already much smaller in size to begin with, and the reduction is much less considerable – from 12 to 10 cells.[4] Nevertheless, there are two cases of fully systematic syncretism, between NOM.PL and VOC.PL on the one hand, and between DAT.PL and ABL.PL on the other hand, as shown in Table 8 below.

Table 8: The tabular and cell paradigm of CONSUL 'consul'

a. Tabular paradigm

cell	MPS	wordform
1	NOM.SG	*consul*
2	GEN.SG	*consulis*
3	DAT.SG	*consulī*
4	ACC.SG	*consulem*
5	VOC.SG	*consul*
6	ABL.SG	*consule*
7	NOM.PL	*consulēs*
8	VOC.PL	*consulēs*
9	GEN.PL	*consulum*
10	DAT.PL	*consulibus*
11	ABL.PL	*consulibus*
12	VOC.PL	*consulēs*

b. Cell paradigm

cell	MPS	wordform
1	NOM.SG	*consul*
2	GEN.SG	*consulis*
3	DAT.SG	*consulī*
4	ACC.SG	*consulem*
5	VOC.SG	*consul*
6	ABL.SG	*consule*
7	NOM.PL, VOC.PL	*consulēs*
8	GEN.PL	*consulum*
9	DAT.PL, ABL.PL	*consulibus*
10	VOC.PL	*consulēs*

Other syncretic patterns are less systematic, and cannot therefore be abstracted away in the cell paradigm representation format: for instance, the NOM.SG is overwhelmingly realized in the same way as the VOC.SG – as happens also in the noun CONSUL given in Table 8 above; however, this is not the case for masculine 2nd declension nouns ending in *-us* in NOM.SG, that display a VOC.SG in *-e*, or more rarely in *-ī*: e.g. NOM.SG *lupus*, but VOC.SG *lupe*, NOM.SG *filius* but VOC.SG *filī*. Therefore, these two contents cannot be conflated in the cell paradigm of Latin

[4] As we have seen above in §5.1, the locative cell has already been excluded because of its marginality.

nouns, in order to preserve the requisite of uniformity in paradigm size across all lexemes (cf. §4.2 above).

5.2.2 Results: unary implicative entropy

Since noun paradigms are much smaller in size than verb paradigms, we can simply show the overall results that have been obtained through the Qumin toolkit. The values of unary implicative entropy, and the resulting average implicative entropy regarding the 10-cell paradigm of Table 8b are given in Table 9. As usual, different shades of grey are used to help the visualization of more and less relevant levels of uncertainty, with darker shades used for higher entropy values.

Table 9: Predictability in Latin noun paradigms: unary implicative entropy

	NOM.SG	GEN.SG	DAT.SG	ACC.SG	VOC.SG	ABL.SG	NOM.PL	GEN.PL	DAT.PL	ACC.PL
NOM.SG		0.3525	0.378	0.2798	0.2017	0.3164	0.259	0.399	0.3271	0.292
GEN.SG	0.4133		0.02422	0.06586	0.4053	0.07837	0.4905	0.2986	0.010796	0.524
DAT.SG	0.4612	0		0.0664	0.476	0.0807	0.48	0.3203	0.01292	0.5386
ACC.SG	0.517	0.3047	0.2751		0.669	0.392	0.593	0.6294	0.3213	0.6157
VOC.SG	0.01883	0.2451	0.2413	0.2861		0.2039	0.1423	0.2462	0.2251	0.1876
ABL.SG	0.4043	0	0.2498	0.03644	0.4192		0.4856	0.264	0.01248	0.5205
NOM.PL	0.147	0.1605	0.1761	0.02682	0.147	0.254		0.5244	0.2245	0.0569
GEN.PL	0.541	0.2032	0.3755	0.5107	0.523	0.2957	0.7173		0.2148	0.789
DAT.PL	1.018	0.787	0.7246	0.817	1.018	0.8574	1.251	1.076		1.293
ACC.PL	0.1271	0.1613	0.3467	0.01715	0.1511	0.2356	0	0.5103	0.2292	

Average implicative entropy: 0.361849

It can be observed that there are no cases of full interpredictability among these paradigm cells, i.e. no pairs of cells that can be predicted from one another with no uncertainty (H = 0) in both directions. This means that the dramatic simplification of paradigm structure that was obtained for verb inflection (cf. §4.5 above) by means of what was called a "distillation" plays no role in nominal inflection. The only cells that are fully interpredictable – namely, NOM.PL and VOC.PL on the one hand, DAT.PL and ABL.PL on the other one – have already been merged by using the cell paradigm, since they are trivially interpredictable simply because they are always syncretic: they can be considered to constitute only one cell from a strictly morphological perspective.

On the other hand, some degree of simplification of paradigm structure can also be achieved for noun paradigms if the distillation is obtained by conflating cells that are not fully interpredictable – i.e., with null entropy values – but are nevertheless predictable with very little uncertainty – i.e, with entropy values that approach to 0 in both directions. Setting the threshold at 0.1, also GEN.SG, DAT.SG and ABL.SG can be conflated in a single distillation, since both DAT.SG and ABL.SG are in quasi-systematic covariation with GEN.SG.

For instance, GEN.SG can be predicted with no uncertainty from DAT.SG, as shown by the null entropy value in the corresponding case of Table 9. In the other direction, however, predictability is not complete. Given a GEN.SG ending in *-eī*, we could be facing a 5th declension noun like RES, but also a 2nd declension noun where the *e* is part of the stem – for instance, REUS. Furthermore, given a GEN.SG ending in *-ūs*, we know that we are facing a 4th declension noun but we do not know if it is masculine – thus with DAT.SG in *-uī* – or neuter – thus with DAT.SG in *-ū*. The situation is summarized below in Table 10.

Table 10: Sources of uncertainty in predicting DAT.SG from GEN.SG

lexeme (decl.)	GEN.SG	DAT.SG
RES 'thing'(5th)	*reī*	*reī*
REUS 'guilty' (2nd)	*reī*	*reō*
$SINUS_M$ 'curve' (4th)	*sinūs*	*sinuī*
$CORNU_N$ 'horn' (4th)	*cornūs*	*cornū*

However, the quantitative impact of these ambiguities is very limited. This is because there are very few lexemes whose GEN.SG ends in *-eī.*: in LatInfLexi, there are only 11.[5] Therefore, the impact of the uncertainty regarding those lexemes on the overall lexicon (the 1,038 lexemes of LatInfLexi) is not particularly relevant. On the other hand, although lexemes with a GEN.SG in *-ūs* are more numerous (90 lexemes in LatInfLexi), neuter 4th declension nouns are very rare (in LatInfLexi there are only two[6]). Thus, given a GEN.SG in *-ūs* it is much more likely that the noun will be masculine, and thus its DAT.SG will be in *-uī*; the impact on the entropy

[5] Namely, the 2nd declension nouns ALVEUS 'hollow', BALNEUM 'bath', CLIPEUS 'shield', CUNEUS 'wedge', DEUS 'god', LAQUEUS 'noose', REUS 'guilty', and the 5th declension nouns FIDES 'trust', PLEBES 'common people', RES 'thing' and SPES 'hope'.

[6] Namely, CORNU 'horn' and GENU 'knee'.

value will accordingly be very low. These quantitative observations explain the close-to-0 entropy value in the corresponding case of Table 9.

Also NOM.PL and ACC.PL can be lumped in a single quasi-distillation: the only uncertainty in this cell is due to the few 3rd declension nouns with ACC.PL in *-īs* , as opposed to the prevailing ACC.PL in *-ēs*, as shown in Table 11.

Table 11: Sources of uncertainty in predicting ACC.PL from NOM.PL

lexeme (decl.)	NOM.PL	ACC.PL
URBS 'town' (3rd, cons.)	*urbēs*	*urbēs*
TURRIS 'tower' (3rd, *-i-*)	*turrēs*	*turrīs*

If we abstract away from these marginal sources of uncertainty and lump those cells together in a single distillation, we can obtain a mapping of the Latin noun paradigm in zones that include cells that can be predicted one from another with very little uncertainty – the average of the entropy of predicting cell A from cell B and *vice versa* being lower than 0.1. Such mapping is given in Table 12 below, where a strong mutual interpredictability between oblique cases in the singular and direct cases in the plural can be observed.

Table 12: Zones of high interpredictability in Latin noun inflection

	SG	PL
NOM	Z1	Z5
ACC	Z2	Z5
VOC	Z3	Z5
GEN	Z4	Z6
DAT	Z4	Z7
ABL	Z4	Z7

Let us now consider the predictability and predictiveness of each of the 10 cells. Based on the results presented in Table 9, for each paradigm cell, on the one hand we can compute the average of the entropy values measuring the uncertainty in predicting the involved cell from each of the other cells, as an indicator of the predictability of that cell; on the other hand, we can use the average of the entropy values measuring the uncertainty in predicting each of the other cells from that cell as an indicator of the predictiveness of that cell. These indicators are given below

in Table 13a-b, in both cases in ascending order of entropy values – i.e., from the most predictable/predictive cell to the least predictable/predictive one.

Table 13a: Average predictability and predictiveness in noun paradigms

a.

cell	average predictability
DAT.PL	0.175355
ACC.SG	0.23403
GEN.SG	0.246033
ABL.SG	0.301563
DAT.SG	0.310147
NOM.SG	0.405303
VOC.SG	0.445589
GEN.PL	0.474244
NOM.PL	0.490967
ACC.PL	0.535256

b.

cell	average predictiveness
NOM.PL	0.190802
ACC.PL	0.197606
VOC.SG	0.199603
GEN.SG	0.256772
ABL.SG	0.265813
DAT.SG	0.27068
NOM.SG	0.311722
GEN.PL	0.463356
ACC.SG	0.479689
DAT.PL	0.982444

It can be observed from a comparison between these two tables that cells that are high in the ranking of predictability tend to be low in the ranking of predictiveness, and *vice versa*. This is not unexpected, since cells that display many different realizations across lexemes are difficult to predict, but highly informative on the content of other cells, and therefore highly predictive; conversely, cells where there is less variation are less predictive, but easier to predict.

A particularly clear example is provided by DAT.PL, which proves to be the most predictable cell, but also (by large) the least predictive one. This happens because in some cases lexemes that are inflected differently in other paradigm cells display the same realization in that cell. The distinction between the 1st and 2nd declension is neutralized in DAT.PL, and so is the one between 3rd and 4th: both 1st and 2nd declension nouns end in *-īs*, while both 3rd and 4th declension nouns end in *-ibus*, as can be seen in Table 14. Furthermore, the endings of DAT.PL are the same in nouns that belong to the same declension, but differ in gender, making that cell poorly informative on the content of the nominative, accusative and vocative cells that have a different realization according to the gender of the lexeme. All of these factors make it very difficult to infer the realization of other cells – e.g. NOM.PL, cf. Table 14 – knowing only DAT.PL, hence the very low predictiveness of that cell.

However, those very same factors also make DAT.PL a highly predictable cell, since there is no need to have a completely reliable information on the inflectional class (and sub-class) to which the involved lexeme belongs in order to be able to correctly infer the content of the cell.

Conversely, there are cells like NOM.PL whose exponents differ more systematically across lexemes – although not completely so: see for instance in Table 14 the endings *-a*, common to neuter nouns of the 2nd and 3rd declension, and *-ēs*, common to masculine and feminine nouns of the 3rd and 5th declension. Such cells will be very informative on the inflectional behaviour displayed in other cells, and therefore highly predictive, but also harder to predict from other cells if those cells fail to signal the relevant distinctions, as happens e.g. in DAT.PL in Table 14.

Table 14: The realizations of the cells NOM.PL and DAT.PL for some Latin nouns

lexeme (decl.)	NOM.PL	DAT.PL
ROSA 'rose' (1st)	*rosae*	*rosīs*
$LUPUS_M$ 'wolf' (2nd)	*lupī*	*lupīs*
$BELLUM_N$ 'war' (2nd)	*bella*	*bellīs*
$CONSUL_M$ 'consul'(3rd)	*consulēs*	*consulibus*
$TEMPUS_N$ 'time' (3rd)	*tempora*	*temporibus*
$FRUCTUS_M$ 'fruit' (4th)	*fructūs*	*fructibus*
$CORNU_N$ 'horn' (4th)	*cornua*	*cornibus*
RES 'thing' (5th)	*rēs*	*rēbus*

5.2.3 *n*-ary implicative entropy and principal parts

Let us now move to predictions from more than one inflected wordform, as measured by *n*-ary implicative entropy, starting from the situation that arises when two predictors are considered. In Table 15, we show the entropy values estimating the uncertainty in predicting the various cells (listed in the columns' headers) from every possible pair of cells (listed in the lines' headers), as well as the average predictiveness of each pair of cells (given in the last column), and lastly the overall average binary implicative entropy (at the bottom of the table). The pairs of cells are sorted according to their average predictiveness, in ascending order – thus, from

the most predictive pairs with values approaching 0, to the least predictive ones with higher values.

Table 15: Binary implicative entropy in noun paradigms

	NOM.SG	GEN.SG	DAT.SG	ACC.SG	VOC.SG	ABL.SG	NOM.PL	GEN.PL	DAT.PL	ACC.PL	average predictiveness
VOC.SG,ABL.SG	0	0	0	0.001955			0.002699	0	0.005917	0.00588	0.002056375
NOM.SG,ABL.SG		0	0	0.001953	0.001955		0.002699	0	0.005917	0.00587	0.00229925
VOC.SG,ACC.PL	0	0	0	0.001959		0.006714	0	0.00652	0.00861		0.002975375
GEN.SG,VOC.SG	0		0	0.001955		0.00977	0.002699	0	0	0.01128	0.003213
NOM.SG,GEN.SG			0	0.001953	0.001955	0.009766	0.002699	0	0	0.01126	0.003454125
NOM.SG,ACC.PL			0	0	0.001957	0.004864	0.006706	0	0.00651	0.00861	0.003580875
NOM.SG,DAT.PL			0	0	0.007835	0	0.007835	0.002699	0		0.003706125
VOC.SG,DAT.PL	0	0	0	0	0.007835		0.007835	0.002699	0		0.003706125
DAT.SG,VOC.SG	0	0		0.005867		0.0144	0.002699	0	0.005917	0.01724	0.005765375
NOM.SG,NOM.PL		0.001959	0.001959	0	0	0.01028		0.00652	0.006714	0.01944	0.005859
VOC.SG,NOM.PL	0	0.001959	0.001959	0		0.01028		0.00652	0.006714	0.01944	0.005859
NOM.SG,DAT.SG		0		0.00586	0.001955	0.01438	0.002699	0	0.005917	0.01723	0.006005125
NOM.SG,GEN.PL		0	0	0.02107	0	0.01441	0.002699		0.006355	0.02393	0.008558
VOC.SG,GEN.PL	0	0	0	0.02109		0.01443	0.002699		0.006355	0.02394	0.00856425
GEN.PL,ACC.PL	0.02069	0	0.03888	0.003817	0.0207	0.003532	0		0.012535		0.01251925
ACC.SG,VOC.SG	0	0.011734	0.011734			0.01878	0.009834	0.01707	0.02142	0.01842	0.013624
NOM.PL,GEN.PL	0.02309	0	0	0.01134	0.02309	0.01775			0.005325	0.03836	0.014869375
DAT.PL,ACC.PL	0.06445	0	0	0.001959	0.06445	0.002699	0	0.2224			0.04449475
GEN.SG,ACC.PL	0.07605		0	0	0.07806	0.001957	0	0.2231	0.005398		0.048070625
DAT.SG,ACC.PL	0.07605	0		0.001957	0.0955	0.003532	0	0.2231	0.01262		0.051594875
ABL.SG,NOM.PL	0.0824	0	0	0	0.0824			0.2299	0.005398	0.0174	0.05218725
NOM.PL,DAT.PL	0.0715	0	0	0.01287	0.0715	0.0216		0.2382		0.05496	0.05882875
GEN.SG,NOM.PL	0.0861		0	0.004753	0.0861	0.02182		0.2426	0.005398	0.04056	0.060916375
DAT.SG,NOM.PL	0.0861	0		0.01287	0.0861	0.02182		0.2426	0.005398	0.04727	0.06276975
ABL.SG,ACC.PL	0.07605	0	0.1641	0	0.0955		0	0.2208	0.01219		0.07108
ACC.SG,ACC.PL	0.1138	0.00809	0.02382		0.1378	0.07745	0	0.3003	0.0761		0.09217
ACC.SG,NOM.PL	0.1229	0.00863	0.0239		0.1229	0.0794		0.302	0.0727	0.02202	0.09430625
NOM.SG,ACC.SG		0.11395	0.1295		0.1515	0.12494	0.10645	0.1694	0.0961	0.1154	0.125905
ACC.SG,GEN.PL	0.2998	0	0		0.2913	0.002695	0.2808		0.01066	0.2993	0.148069375
ACC.SG,DAT.PL	0.3396	0	0		0.3284	0.001959	0.281	0.2322		0.3027	0.185732375
GEN.SG,ACC.SG	0.3586		0		0.3335	0.002691	0.281	0.2362	0.005398	0.304	0.190173625
ACC.SG,ABL.SG	0.3555	0	0		0.3718		0.276	0.2318	0.01066	0.297	0.192845
DAT.SG,ACC.SG	0.356	0			0.3713	0.002691	0.2786	0.2362	0.010574	0.3013	0.194583125
NOM.PL,ACC.PL	0.1272	0.1595	0.176	0.01531	0.1272	0.2333		0.4993	0.222		0.19497625
GEN.SG,GEN.PL	0.3145		0.00539	0.02614	0.288	0.04044	0.437		0.005398	0.4575	0.196796
GEN.PL,DAT.PL	0.3103	0	0	0.0348	0.2988	0.04263	0.442			0.4607	0.19865375
ABL.SG,GEN.PL	0.3088	0	0.0456	0.0184	0.304		0.443		0.01248	0.4573	0.1986975
NOM.SG,VOC.SG		0.2185	0.2045	0.2493		0.1827	0.1355	0.2252	0.2178	0.1655	0.199875
DAT.SG,GEN.PL	0.3135	0		0.03003	0.3086	0.05157	0.4348		0.01292	0.4714	0.2028525
ABL.SG,DAT.PL	0.3618	0	0	0.03296	0.3513		0.459	0.2551		0.469	0.241145
GEN.SG,ABL.SG	0.38		0.005383	0.03452	0.3547		0.4592	0.264	0.005398	0.4739	0.247137625
DAT.SG,ABL.SG	0.387	0		0.03174	0.4019		0.4592	0.264	0.012085	0.479	0.254365625
GEN.SG,DAT.PL	0.3848		0	0.0523	0.3538	0.06946	0.4756	0.2812		0.504	0.265145
DAT.SG,DAT.PL	0.3752	0		0.06152	0.3647	0.0702	0.4727	0.2812		0.5103	0.2669775
GEN.SG,DAT.SG	0.3938			0.0526	0.3745	0.073	0.4788	0.2932	0.005398	0.5093	0.27257475

Average implicative entropy: 0.100434

The first observation that should be made is that there is no categorical principal part set of cardinality 2 – i.e., composed of 2 inflected wordforms, cf. § 4.6 above – that emerges from the entropy-based analysis: the average predictiveness is never 0, meaning that there is no pair of cells from which the rest of the paradigm can be inferred with no uncertainty whatsoever. However, it is at least possible to

individuate several near principal part sets – pairs of cells whose average predictiveness approaches 0, corresponding to a very limited amount of uncertainty in the PCFP. If the threshold is set at 0.01, 14 of the 45 possible pairs of cells constitute near principal parts; if it is set at 0.1, we obtain 27 near principal part sets. As is argued by Bonami & Beniamine (2016: 176), and as we have already observed here in §4.6 above, it is perfectly reasonable to focus not only on cases where there is no uncertainty whatsoever, since a small level of uncertainty is present even in the competence of fluent native speakers, that do sometimes make mistakes in guessing the inflected wordforms of lexemes.

We saw above in §5.1 that the principal parts that are traditionally used in Latin dictionaries and grammars are NOM.SG and GEN.SG. Indeed, this pair of cell proves to be at the top of the ranking of Table 15, although there still is a very little amount of uncertainty in the prediction of other paradigm cells from this pair. As was hinted above in §5.1, the reason for the high informativity of this pair stands in the complementarity of the information that is provided by the two cells on the overall inflectional behaviour of a lexeme.

In NOM.SG, the main source of uncertainty is given by the fact that the distinction between masculine nouns of the 2nd and 4th declension is neutralized: given a noun in *-us*, a speaker cannot be sure whether it belongs to the 2nd declension, and therefore has GEN.SG *-ī*, DAT.SG *-ō* etc., or to the 4th declension, thus with GEN.SG *-ūs*, DAT.SG *-uī* etc. Furthermore, a noun in *-us* could also be a neuter noun of the 3rd declension, like TEMPUS, with GEN.SG *temporis*, DAT.SG *tempori* etc. Similarly, a noun ending in *-ēs* in NOM.SG could be a 3rd declension noun – e.g. CAEDES, with GEN.SG *caedis*, DAT.SG *caedī* etc. – or a 5th declension noun – e.g. FIDES, with genitive and dative singular *fideī* etc. All these neutralizations generate uncertainty in predicting other cells from NOM.SG, as can be observed in Table 9 above.

However, if GEN.SG is added to the picture, then almost all of this uncertainty disappears: given a noun with NOM.SG in *-us*, if its GEN.SG ends in *-ī*, then we are facing a 2nd declension masculine nouns, if it ends in *-ūs* we are facing a masculine 4th declension noun, if it ends in in *-is*, we are facing a neuter 3rd declension noun; similarly, given a noun with NOM.SG in *-es*, if it ends in *-is* in GEN.SG it belongs to the 3rd declension, if it ends in *-eī* or *-ēī* to the 5th declension.

On the other hand, from GEN.SG alone, a speaker would not be able to predict the remaining cells of the paradigm with certainty, since knowing only that cell it is not possible to know with certainty the gender of the involved nouns, and therefore its different inflectional behaviour in other cells, as summarized in Table 2: given two GEN.SG forms like *lupī* and *bellī*, there is no way of knowing that the first noun is masculine, and will therefore have NOM.SG *lupus*, while the second one is neuter, and will therefore have NOM.SG *bellum*. However, the gender of a noun is more systematically revealed by its NOM.SG: *-us* in masculine nouns of the 2nd and 4th declension, *-um* in neuter 2nd declension nouns and *-u* in neuter 4th declension nouns.

Therefore, if both the NOM.SG and the GEN.SG of a lexeme are given, enough information is provided to infer the rest of its paradigm with almost no uncertainty. The very small amount of uncertainty that is left is due to the fact that in the 3rd declension NOM.SG is not fully informative on gender. For instance, both ARBOR 'tree' and MARMOR 'marble' end in *-or* in NOM.SG and belong to the 3rd declension, but ARBOR is feminine, while MARMOR is neuter. They are therefore inflected differently in cells where there is systematic syncretism in neuter nouns – e.g., ACC.SG *arborem* vs. *marmor* – but this difference cannot be predicted on the basis of NOM.SG and GEN.SG alone, since in these two cells these nouns are inflected in the same way – *arbor arboris* like *marmor marmoris*. Furthermore, from NOM.SG and GEN.SG it is not possible to predict the inflectional behaviour of nouns belonging to the different sub-classes of the 3rd declension in the relevant cells (cf. Table 7 in §5.1): for instance CIVIS and PUPPIS behave in the same way in NOM.SG and GEN.SG (*civis civis* like *puppis puppis*), thus it is not possible to predict the different realizations they display, for instance, in ACC.SG (*civem* vs. *puppim*)

The results shown in Table 15 also indicate that, alongside the pair constituted by NOM.SG and GEN.SG that is traditionally used in dictionaries and grammars, there are several other pairs of cells that constitute a near principal part set for Latin noun paradigms, since they allow us to infer roughly the same amount of information on the overall inflectional behaviour of a lexeme.

For instance, ABL.SG is similar to GEN.SG in displaying different endings for each of the five traditional declensions, but failing to be informative on a noun's gender,

which is revealed also from its VOC.SG, as it was shown above to be revealed from its NOM.SG. Therefore, the pair of cells comprising VOC.SG and ABL.SG is also a near principal part set, with average predictiveness < 0.01. On the other hand, pairs of cells that provide similar information on the inflectional behaviour of a lexeme, like e.g. GEN.SG and DAT.SG or NOM.SG and VOC.SG, will obviously be much worse as predictors of the remaining paradigm cells, and indeed they rank low in Table 15.

Despite the absence of categorical principal part sets, the presence of many near principal part sets is obviously helpful in reducing the overall difficulty of the PCFP: once a speaker has been exposed to any of the pairs of cells that function as near principal parts, then (s)he will be able to infer the remaining paradigm cells with very little uncertainty, and the bigger the number of principal part sets, the higher the probability that a speaker will be exposed to at least one of them.

Let us now move to predictions from more than two forms. The number of principal part sets and near principal part sets of different cardinalities – together with the average implicative entropy value with the corresponding number of predictors – is summarized in Table 16. With three predictors, we finally find 3 categorical principal part sets, from which the remainder of the paradigm can be inferred with no uncertainty whatsoever: their average predictiveness is 0.[7] We also find many more near principal part sets of cardinality 3 than we did of cardinality 2: there are 10 sets if the threshold is set at 0.001, 69 with the threshold at 0.01, 100 at 0.1.

Quite unsurprisingly, the more forms that are used as predictors, the more numerous the sets of cells that function as principal parts (or at least as near principal parts) are, both in absolute terms and in percentage on the total of possible combinations of cells. With four predictors, 12.9% of the combinations constitute a completely reliable principal part set, and 92.9% at least function as near principal parts with the threshold fixed at 0.1. In Table 16, we stop at five predictors: at this cardinality,

[7] The combinations that prove to be categorically predictive are NOM.SG-DAT.PL-ACC.PL, GEN.SG-VOC.SG-ACC.PL and VOC.SG-DAT.PL-ACC.PL. The presence of categorical principal part sets of cardinality 3, despite the intricate patterns of implicative relations regarding 3rd declension nouns, as summarized in Table 7, §5.1, is due to subtle restrictions on the context of application of the relevant patterns, as they are computed by the Qumin toolkit. Some of these restrictions are perhaps not representative of the situation in the whole lexicon of Latin nouns: some of the sub-classes are attested only for a handful of nouns in LatInfLexi, thus the generalization might be based on too few lexemes. Therefore, it is likely that with a larger and more representative sample more than 3 wordforms would be required to have a categorical principal part set.

almost one third of the combinations (29%) can be taken as a fully reliable principal part set, and virtually all of them (97.6%) are near principal parts if the threshold is set at 0.1.

Table 16: Noun paradigms: principal part sets and near principal part sets of different cardinalities

cardinality	average implicative entropy	principal parts (H = 0)		near principal parts (H < 0.001)		(H < 0.01)		(H < 0.1)	
		n.	%	n.	%	n.	%	n.	%
2	0.100434	0	0	0	0	14	31.1%	27	60%
3	0.050904	3	2.5%	10	8.33%	69	57.5%	100	83.3%
4	0.02731	27	12.9%	65	31%	160	76.2%	195	92.9%
5	0.013505	73	29%	112	44.4%	216	85.7%	246	97.6%

5.3 Predictability and gender

5.3.1 Some examples

Let us now move to an evaluation of how the results change if information on gender is taken into account, starting from a quick overview of the Latin gender system.[8] There are three genders in Latin: masculine, feminine and neuter. Nouns belonging to different genders display different agreement patterns – as shown in (3) – with determiners, adjectives and participles.

(3) a. *ille* *puer*
that:NOM.M.SG boy(M):NOM.SG

b. *illa* *puella*
that:NOM.F.SG girl(F):NOM.SG

c. *illud* *mālum*
that:NOM.N.SG apple(N):NOM.SG

[8] For a recent, detailed account, the reader is referred to Loporcaro (2018: Chapter 2).

From a semantic standpoint, nouns denoting male humans and superior animates are almost[9] always assigned to the masculine gender (cf. PUER in 3a) and nouns denoting female humans and superior animates to the feminine gender (cf. PUELLA in 3b). Nouns denoting inanimates can be assigned to any of the three genders: for instance, PANIS 'bread' is masculine, TURRIS 'tower' is feminine and RETE 'net' is neuter.

We saw above in §5.1 that nouns of different gender are also inflected differently, even when they belong to the same declension according to the traditional description: several inflectional sub-classes can be individuated on the basis of gender, as exemplified in Table 2. Knowing the gender of a noun can therefore have consequences on predictability in the PCFP. Let us consider for instance the different inflectional behaviour of masculine/feminine vs. neuter nouns of the 2nd declension. A GEN.SG in *-ī* identifies a noun as belonging to the 2nd declension, but does not allow to predict the NOM.SG with certainty, since the ending would be *-us* if the noun is masculine or feminine, *-um* if it is neuter: if the gender is assumed to be known, this uncertainty disappears.

Therefore, it is interesting to evaluate the impact of gender on uncertainty in inflectional predictions, as measured by implicative entropy: this is what we will do in this section. Let us start from some examples to understand the way in which information on the gender of nouns can be taken into account in entropy calculations, and to show that such information is potentially available to speakers – at least in some cases.

In Table 17, we focus on the task of predicting GEN.SG from NOM.SG for 8 Latin nouns whose NOM.SG ends in *-us*.

[9] With a handful of exceptions, e.g. MANCIPIUM 'slave', which is neuter.

Table 17: Predicting GEN.SG from NOM.SG (reduced dataset)

lexeme (decl.)	gender	NOM.SG	GEN.SG	alternation pattern
PONTUS 'sea' (2nd)	M	*pontus*	*pontī*	1. *_us* ↔ *_ī*
ANNUS 'year' (2nd)	M	*annus*	*annī*	1. *_us* ↔ *_ī*
PORTUS 'harbour' (4th)	M	*portus*	*portūs*	2. *_us* ↔ *_ūs*
CANTUS 'song' (4th)	M	*cantus*	*cantūs*	2. *_us* ↔ *_ūs*
PECTUS 'breast' (3rd)	N	*pectus*	*pectoris*	3. *_us* ↔ *_oris*
CORPUS 'body' (3rd)	N	*corpus*	*corporis*	3. *_us* ↔ *_oris*
LATUS 'side' (3rd)	N	*latus*	*lateris*	4. *_us* ↔ *_eris*
SCELUS 'crime' (3rd)	N	*scelus*	*sceleris*	4. *_us* ↔ *_eris*

If we only look at the inflected wordforms filling the two paradigm cells of the lexemes involved, we find four different alternation patterns that can be applied to a noun that ends in *-us* in NOM.SG, as shown in the last column of Table 17. If the Latin lexicon comprised only these nouns, the entropy of predicting GEN.SG from NOM.SG would be 2, since there would be four equiprobable outcomes. The details of the computation are summarized in (4).

(4) $H(\text{GEN. SG}|\text{NOM. SG}) = -\left(\frac{2}{8} \times \log_2 \frac{2}{8} + \frac{2}{8} \times \log_2 \frac{2}{8} + \frac{2}{8} \times \log_2 \frac{2}{8} + \frac{2}{8} \times \log_2 \frac{2}{8}\right) = 2$

However, if a speaker is exposed to the inflected words of NOM.SG of those nouns in contexts where they are modified by a determiner, an adjective, or a partciple, the gender of such nouns would be revealed by the different agreement markers displayed on the agreement target, as exemplified in (5) with the determiner ILLE 'that'.

(5) a. ***ille*** *pontus / annus / portus / cantus*
'that (NOM.M.SG) sea / year / harbour / song'
b. ***illud*** *pectus / corpus / latus / scelus*
'that (NOM.N.SG) breast / body / side / crime'

Given the potential availability of information on the gender of the involved nouns, it is interesting to see how the situation changes if such information is assumed to

be known. It can be observed in Table 17 that patterns 1 and 2 are only attested in masculine nouns, while patterns 3 and 4 only in neuter nouns. Thus, we obtain an entropy of 1 for both masculine and neuter nouns, and since the number of masculine and neuter nouns is the same in this sample the overall entropy value is also 1, as shown in (6) – the half of the value obtained without taking gender into account, with a remarkable reduction in uncertainty.

(6) $$H(\text{GEN. SG}|\text{NOM. SG, GENDER}) = -\left(\frac{4}{8} \times \left(\frac{2}{4} \times \log_2 \frac{2}{4} + \frac{2}{4} \times \log_2 \frac{2}{4}\right) + \frac{4}{8} \times \left(\frac{2}{4} \times \log_2 \frac{2}{4} + \frac{2}{4} \times \log_2 \frac{2}{4}\right)\right) = 1$$

Let us now look at a different example, estimating the uncertainty in predicting NOM.SG from DAT.PL on a sample of 10 Latin nouns.

Table 18: Predicting NOM.SG from DAT.PL (other reduced dataset)

lexeme (decl.)	gender	DAT.PL	NOM.SG	alternation pattern
AMICA '(female) friend' (1st)	F	*amicīs*	*amica*	1. *_īs ↔ _a*
DOMINA 'mistress' (1st)	F	*dominīs*	*domina*	1. *_īs ↔ _a*
LUPA '(female) wolf' (1st)	F	*lupīs*	*lupa*	1. *_īs ↔ _a*
PUELLA 'girl' (1st)	F	*puellīs*	*puella*	1. *_īs ↔ _a*
NAUTA 'sailor' (1st)	M	*nautīs*	*nauta*	1. *_īs ↔ _a*
AMICUS '(male) friend' (2nd)	M	*amicīs*	*amicus*	2. *_īs ↔ _us*
DOMINUS 'master' (2nd)	M	*dominīs*	*dominus*	2. *_īs ↔ _us*
LUPUS '(male) wolf' (2nd)	M	*lupīs*	*lupus*	2. *_īs ↔ _us*
TYRANNUS 'monarch' (2nd)	M	*tyrannīs*	*tyrannus*	2. *_īs ↔ _us*
PINUS 'pine' (2nd)	F	*pinīs*	*pinus*	2. *_īs ↔ _us*

As we saw, the ending *-īs* of DAT.PL is not fully informative on the inflectional behaviour of the lexemes in NOM.SG, since it appears in both 1st and 2nd declension nouns. Therefore, given the inflected wordform filling the cell DAT.PL, there is a remarkable amount of uncertainty in guessing the content of NOM.SG: patterns 1 and 2 are both applied in half of the cases, thus yielding an entropy value of 1.

(7) $$H(\text{NOM. SG}|\text{DAT. PL}) = -\left(\frac{5}{10} \times \log_2 \frac{5}{10} + \frac{5}{10} \times \log_2 \frac{5}{10}\right) = 1$$

Now, in this case the agreement patterns are not helpful in inferring gender, since in DAT.PL the agreement targets are inflected in the same way.

(8) a. *illīs amicīs* (M/F) / *dominīs* (M/F) / *lupīs* (M/F) / *puellīs* / *nautīs* / *tyrannīs* / *pinīs*
'to those (DAT.M/F.PL) friends / master/mistress / wolf / girl / sailor / monarch / pine'

However, information on gender can be assumed to be known for such nouns based on their meaning. In Latin, as we saw, nouns denoting male humans and superior animates – like AMICUS, DOMINUS, LUPUS, TYRANNUS, NAUTA in our sample – are masculine, while the ones denoting female humans and superior animates – like AMICA, DOMINA, LUPA, PUELLA – are feminine. Another, more language-specific, rule of semantic assignment of Latin states that nouns denoting trees – like PINUS – are feminine.
In this example, adding gender to the picture allows for a relevant reduction in uncertainty: masculine nouns ending in *-īs* in DAT.PL are overwhelmingly 2nd declension nouns, and thus display a NOM.SG in *-us*, while feminine nouns with DAT.PL in *-īs* are usually 1st declension nouns with NOM.SG in *-a*. Therefore, the probability distribution in both masculine and feminine nouns is skewed, generating a lower entropy in both cases, and consequently in the overall value, as can be seen in (9).

(9) $$H(\text{NOM. SG}|\text{DAT. PL, GENDER}) = -\left(\frac{5}{10} \times \left(\frac{4}{5} \times \log_2 \frac{4}{5} + \frac{1}{5} \times \log_2 \frac{1}{5}\right) + \frac{5}{10} \times \left(\frac{4}{5} \times \log_2 \frac{4}{5} + \frac{1}{5} \times \log_2 \frac{1}{5}\right)\right) = 0.72$$

With these two examples, we have shown that information on the gender of nouns is often available to speakers – either because of the different markers displayed by agreement targets, as in (5), or on the basis of semantic clues, as in the nouns of Table 18 – and that this information is potentially helpful in reducing uncertainty

in the PCFP.[10] In what follows, we will evaluate the impact of such facts on real data. The gender of each of the 1,038 nouns contained in LatInfLexi was obtained from the database of Lemlat 3.0. This information was then exploited to automatically compute implicative entropy[11] for nouns after having assigned them to different classes according to their gender. This was done by means of a specific functionality of the Qumin toolkit. The details of the analysis are given in §5.3.3, but before that it is useful to have a look at the relationship between the gender of a noun and its inflectional behaviour, as we will do in the next section.

5.3.2 Gender and inflection classes in Latin

As a starting point, let us have a look in Table 19 at the classification of the nouns of LatInfLexi among the different genders and to how it is related to their inflectional behaviour in terms of belonging to one of the major declensions. To confirm the trends that we show on a larger scale, in Table 20 we also provide data from Lemlat's lemma list (excluding lemmas taken from the Onomasticon and from Du Cange 1883-1887).

Table 19: The gender of Latin nouns of different declensions in LatInfLexi[12]

	M	F	N
1st decl.	6	198	0
2nd decl.	115	3	206
3rd decl.	105	197	69
4th decl.	83	5	2
5th decl.	0	13	0
TOTAL	309	416	277

[10] Additionally, speakers can know a lexeme's gender because it is formed by means of a given derivational suffix: for instance, action nouns formed by means of the suffix *-tio* (e.g. MUTATIO$_N$ 'change' from MUTO$_V$ 'to change') are all feminine. More generally, almost all Latin nominal derivational suffixes can be considered as also assigning a specific gender to the nouns they create. For a more detailed account of noun-forming derivational suffixes in Latin, the reader is referred to §6.3 below.

[11] It is important to observe that gender information is not taken into account in the phase of extraction of alternation patterns, where only phonotactics is considered in order to obtain a maximally general context of application.

[12] Here and in Table 20, nouns that are marked as having more than one gender in Lemlat – for instance CUSTOS (M/F) 'protector/protectress' – have been excluded from the count.

Table 20: The gender of Latin nouns of different declensions in Lemlat

	M	F	N
1st decl.	562	4,409	0
2nd decl.	2,994	233	4,004
3rd decl.	2,590	5,112	840
4th decl.	940	26	4
5th decl.	0	98	0
TOTAL	7,086	9,878	4,848

Globally, Latin nouns are divided quite evenly in the three genders. However, if we look at their distribution in the major declensions, some clear tendencies emerge. Neuter nouns are not found in the 1st and 5th declension, and also in the 4th declension they are very rare: most neuters belong to the 3rd and to the 2nd declension. On the other hand, there are masculine nouns in all declension except the 5th – where, however, there is at least DIES 'day' that can take both masculine and feminine agreement. Also in the 1st declension masculines are rare, while in all the other declensions they are quite common. Lastly, we find feminine nouns in all declensions, although in the 2nd and 4th declension they constitute a minority.

Therefore, the only declension that comprises a fair number of nouns of each of the three genders is the 3rd. All the other declensions display some kind of preferences as to the gender of the nouns that they include. The 1st and 5th declensions are mostly devoted to feminines, with a few masculines and no neuter whatsoever. The core of the 4th declension is constituted by masculine nouns, with a couple of feminine and neuter nouns. Lastly, the 2nd declension is basically composed of masculine and neuter nouns, with few feminines.

These preferences make information on the gender of the involved noun useful in reducing uncertainty in the PCFP, not only – as is obvious – when predicting cells whose realization is systematically different in neuter nouns than it is in masculine and feminine ones (see again Table 2 above), but also in cases where the inflectional realization of a cell is the same in two declensions where the distribution of nouns among the three genders is strongly skewed. One clear example is given by the ending *-īs* of DAT.PL, which is shared by the 1st and 2nd declension, generating uncertainty when trying to predict other cells, as was shown for NOM.SG in Table 18. However, as was already suggested in that example, given the distribution of

nouns with different gender in the major declensions, as illustrated in Table 19 and Table 20, it is clear that if the involved lexeme is feminine, then it is much more likely to belong to the 1st declension, while if it is masculine, then it will probably be a 2nd declension noun: we thus expect a relevant reduction in uncertainty in predictions from DAT.PL if the gender of the noun is assumed to be known. Let us now see if similar predictions are confirmed, by moving to the results of the entropy-based analysis, presented in the following section.

5.3.3 Results

The values of unary implicative entropy obtained by taking into account the classification of nouns according to their gender are given in Table 21 below, with the usual scale of lighter and darker shades of grey to help the visualization of cells where there is more uncertainty.

Table 21: Unary implicative entropy assuming gender information as known

	NOM.SG	GEN.SG	DAT.SG	ACC.SG	VOC.SG	ABL.SG	NOM.PL	GEN.PL	DAT.PL	ACC.PL
NOM.SG		0.1898	0.2358	0.10846	0.1886	0.1854	0.149	0.2761	0.1796	0.1708
GEN.SG	0.1027		0	0.006695	0.1095	0.0656	0.0256	0.288	0.005398	0.05704
DAT.SG	0.1576	0		0.01274	0.1694	0.07245	0.02316	0.3071	0.01292	0.0754
ACC.SG	0.1142	0.1979	0.1929		0.2773	0.277	0.22	0.5127	0.2264	0.2408
VOC.SG	0.005867	0.10175	0.10425	0.1156		0.09375	0.05634	0.1304	0.1024	0.09344
ABL.SG	0.1163	0	0.1663	0	0.128		0.0129	0.2502	0.012245	0.02803
NOM.PL	0.1165	0.1582	0.174	0.02415	0.1165	0.2437		0.505	0.2153	0.0527
GEN.PL	0.1812	0.138	0.2303	0.288	0.1743	0.1953	0.1511		0.1517	0.1959
DAT.PL	0.1836	0.2585	0.1761	0.2275	0.1823	0.3105	0.2551	0.5376		0.2908
ACC.PL	0.09344	0.1578	0.3408	0.01688	0.1165	0.228	0	0.4944	0.2198	

Average implicative entropy: 0.159481

These results should be compared with the ones obtained with no information on gender, shown above in Table 9. It should be observed that the entropy values of Table 21 are never higher than the ones of Table 9, and that there is often a relevant reduction in uncertainty. As a consequence, the average implicative entropy is remarkably lower.

To ease a more detailed comparison, below we provide a table that, for each paradigm cell, reports the difference between the entropy value computed without information on gender and the one computed with such information. This time,

therefore, progressively darker shades of grey are used for cells where the difference is larger, and therefore the reduction in uncertainty is more relevant.

Table 22: Difference between entropy values computed with and without gender information

	NOM.SG	GEN.SG	DAT.SG	ACC.SG	VOC.SG	ABL.SG	NOM.PL	GEN.PL	DAT.PL	ACC.PL
NOM.SG	0	0.1627	0.1422	0.17134	0.0131	0.131	0.11	0.1229	0.1475	0.1212
GEN.SG	0.3106	0	0.02422	0.059165	0.2958	0.01277	0.4649	0.0106	0.005398	0.46696
DAT.SG	0.3036	0	0	0.05366	0.3066	0.00825	0.45684	0.0132	0	0.4632
ACC.SG	0.4028	0.1068	0.0822	0	0.3917	0.115	0.373	0.1167	0.0949	0.3749
VOC.SG	0.012963	0.14335	0.13705	0.1705	0	0.11015	0.08596	0.1158	0.1227	0.09416
ABL.SG	0.288	0	0.0835	0.03644	0.2912	0	0.4727	0.0138	0.000235	0.49247
NOM.PL	0.0305	0.0023	0.0021	0.00267	0.0305	0.0103	0	0.0194	0.0092	0.0042
GEN.PL	0.3598	0.0652	0.1452	0.2227	0.3487	0.1004	0.5662	0	0.0631	0.5931
DAT.PL	0.8344	0.5285	0.5485	0.5895	0.8357	0.5469	0.9959	0.5384	0	1.0022
ACC.PL	0.03366	0.0035	0.0059	0.00027	0.0346	0.0076	0	0.0159	0.0094	0

Difference in average implicative entropy: 0.202367

At first sight, it can already be observed that there is a huge difference in the cases of the table that refer to predictions from DAT.PL, which was pointed out to be by far the least informative cell in terms of predictiveness. This happens because two of the most quantitatively relevant causes of uncertainty in predictions from such cell have a far less relevant impact on entropy when gender information is taken into account. The first of such causes is the scarce informativity of DAT.PL on the gender-based differences in the inflectional behaviour of lexemes of the same declension (cf. again Table 2 above). Of course, if the gender of the noun is assumed to be known, such uncertainty disappears.

Another source of uncertainty in predictions from DAT.PL is given by the fact that, as we saw above, the ending *-īs* is shared by the 1st and 2nd declension. In this case too, knowing the gender of the lexeme strongly reduces the uncertainty associated with this ambiguity, since if the noun is feminine, then it is more likely to belong to the 1st declension (therefore displaying e.g. ACC.SG in *-am*) while if it is masculine or neuter it will probably a 2nd declension noun (with e.g. ACC.SG in *-um*).

However, even taking into consideration the gender of the lexeme, we are still left with a non-negligible uncertainty when predicting other forms from DAT.PL. This happens because there are other ambiguities on which gender is not informative, namely the fact that the ending *-ibus* is shared by nouns of the 3rd and 4th declension:

given a noun whose DAT.PL ends in *-ibus*, even if we know its gender we cannot be sure if its ACC.SG will be in *-em* – with the ending of the 3rd declension – or in *-um* – with the ending of the 4th declension.

If we look at predictability, rather than predictiveness (i.e., at the columns rather than at the lines of the tables), it can be observed that a relevant decrease in entropy values emerges in NOM.PL and ACC.PL. This happens because such cells sometimes display different realizations in nouns of the same declension, because of the usual gender-based inflectional sub-classes. Again, if the gender of the lexeme is assumed to be known, the realizations of such cells are of course much easier to predict even from cells that do not make the relevant distinctions.

Let us now compute the average predictability and the average predictiveness of each cell when information on gender is considered and compare these values to the ones obtained without such information. We can rank cells according to the relevance of the reduction in uncertainty, as measured by the difference between the two values. This is what we do in Table 23 for predictability and in Table 24 for predictiveness.

Table 23: Average predictability with and without gender information

	without gender	with gender	difference
ACC.PL	0,535256	0,133879	0,401377
NOM.PL	0,490967	0,099244	0,391722
NOM.SG	0,405303	0,119045	0,286258
VOC.SG	0,445589	0,162489	0,2831
ACC.SG	0,23403	0,088892	0,145138
DAT.SG	0,310147	0,18005	0,130097
ABL.SG	0,301563	0,185744	0,115819
GEN.SG	0,246033	0,13355	0,112483
GEN.PL	0,474244	0,366833	0,107411
DAT.PL	0,175355	0,125085	0,05027

Table 24: Average predictiveness with and without gender information

	without gender	with gender	difference
DAT.PL	0,982444	0,269111	0,713333
GEN.PL	0,463356	0,189533	0,273822
ACC.SG	0,479689	0,251022	0,228667
ABL.SG	0,265813	0,079331	0,186483
GEN.SG	0,256772	0,073393	0,183379
DAT.SG	0,27068	0,092308	0,178372
NOM.SG	0,311722	0,187062	0,12466
VOC.SG	0,199603	0,089311	0,110293
NOM.PL	0,190802	0,17845	0,012352
ACC.PL	0,197606	0,185291	0,012314

These tables allow us to observe some tendencies on the role of gender in helping inflectional predictions.

Let us start from cells where there are differences in the inflectional behaviour of lexemes that belong to the same declension, but have a different gender – namely, the nominative, accusative and vocative. In such cells, predictability is highly increased (they all rank high in Table 23), but not predictiveness (they tend to rank low in Table 24).[13] This is due to the fact that gender in these cells is basically overt (cf. Corbett 1991: 62), i.e. it can be inferred from the phonological shape of the inflected wordform. Therefore, coding it explicitly is not very useful, as the role of phonotactics is already taken care of by the Qumin toolkit.

Conversely, in cells where the inflectional behaviour of lexemes of different gender-based sub-classes is the same, predictiveness is highly increased (they are high in the ranking of Table 24), but not predictability (they are low in the ranking of Table 23), because in such cells gender is not overt in Corbett's terms: adding this information makes them much more predictive, but has a lesser impact on their predictability, since the gender-based sub-classes play no role in these cells.

The reduction in uncertainty that we have observed in our results is consistent with the findings of another work that investigated the role of information on a lexeme's gender in easing the PCFP, namely Stump & Finkel (2013: Chapter 5). In that chapter, a Principal Part Analysis is performed on two different datasets – with and

[13] The exception of ACC.SG is due to the fact that in 2nd declension nouns, the same ending *-um* is used for masculine and neuter nouns. Therefore, knowing the gender of the lexeme will make this cell more predictive of other cells where the gender distinction is relevant.

without information on gender – concerning Sanskrit nominal inflection. Despite the different language, the fact that both nouns and adjectives are considered (while our dataset only includes nouns), and the already mentioned methodological differences between Principal Part Analysis and our entropy-based approach (cf. §5.2.3 above), also Stump & Finkel (2013)'s results show that gender is useful in reducing the uncertainty in the PCFP, changing in a relevant way many of the quantitative measurements that are used throughout their work. Here, we will focus on the ones that can be more easily compared with our findings.

Firstly, in Stump & Finkel (2013: 151) a relevant increase of both the predictability and the predictiveness of paradigm cells is obtained by adding gender information – although their values of "cell predictiveness" and "cell predictability" are not based on entropy, but rather on the number of cells that can predict or be predicted from a given cell without any uncertainty.

Secondly, Stump & Finkel (2013: 145) observe that the four static principal parts that are needed to infer the whole paradigm of a Sanskrit lexeme can be reduced to three if the gender of the nominal is assumed to be known. It is interesting to check if a similar improvement can be obtained in our case too. If only categorical principal parts are considered, it emerges from our results (compare Table 25 with the results obtained without information on gender shown in Table 16 above) that this is not the case: we find the same number of principal part sets of different cardinalities even if we add information on gender. However, Table 26 shows that information on gender is revealed to be helpful in this respect even with our method, since it at least reduces the number of near principal parts that are necessary to predict the other paradigm cells with a minimal level of uncertainty: we do find a near principal part set of cardinality 2 already at the lowest of our thresholds ($H < 0.001$) when gender is added to the picture, whereas there were no principal part sets of that cardinality without such information; the number of possible near principal part sets at other cardinalities increases too if gender information is taken into account – although not dramatically so.

Table 25: Categorical principal part sets with gender information

cardinality	n.	%
2	0	0
3	3	15.8%
4	27	12.9%
5	73	29%%

Table 26: Near principal part sets (H < 0.001) of different cardinalities with and without gender information

cardinality	without gender		with gender	
	n.	%	n.	%
2	0	0	1	2.2%
3	10	8.33%	21	17.5%
4	65	31%	98	46.7%
5	112	44.4%	151	59.9%

5.3.4 Discussion

In previous works, the question has been raised of the possible function(s) of gender: why should different nouns require different agreement markers in targets according to a classification that appears to be at least partly arbitrary?

More generally, all agreement phenomena can be considered as redundant, since, by definition, some features that are already present in the controller are repeated in the target. However, redundancy is not necessarily a problem; although it is clearly a complication as far as production is concerned, it can be helpful in processing. Let us consider the Latin example (10), where the controller noun *aetas* and the target adjective *ultima* agree in gender, number and case, as do *carminis* and *Cumaei*. Although having to code all the values of these categories twice is clearly a burden from the point of view of the speaker, it is also very helpful from the point of view of the listener, who is allowed to know which elements belong together, despite their distance in linear order.

(10)

ultima *Cumaei* *venit* *iam* *carminis* *aetas*

last:NOM.F.SG Cumae:GEN.N.SG come:PRF.ACT.IND.3SG now song(N):GEN.SG age(f):NOM.SG

'The last age of the Cumaean song has now arrived'

(Verg. *Ecl*. 4, 4)

Another proposal that has been frequently put forward regarding the function of gender is that it can be helpful in facilitating reference tracking, as can be seen from the German example in (11), where the referent of the anaphoric pronoun *er* cannot but be *der Krug*, since it is masculine, while the other potentially available antecedent is feminine – *die Schale*.

(11)[14]

der Krug fiel in die Schale, aber er zerbrach nicht

the jug(M) fell into the bowl(F) but it:M broke not

'The jug fell into the bowl, but it didn't break'

If that is one of the functions of gender, nouns that are similar in meaning are expected to be assigned to different genders. As is noted by Corbett (1991: 321), this is only partly true, since even in formal assignment systems where semantics does not play a systematic role, there tend to be clusters of nouns with similar meanings and the same gender.

A more elaborate proposal has been recently put forward by Dye et al. (2017), that claim that gender markers play a role in redistributing the entropy that estimates the uncertainty in predicting the next element in discourse. For instance, Dye et al. (2017) compare the English sentence in (12a) to its German equivalent in (12b), focusing on the uncertainty in guessing the noun that follows the article (the one that is underlined in (12a-b).

(12) a. Yesterday I visited the <u>doctor</u>

b. Gestern besuchte ich den <u>Arzt</u>

[14] Example from Zubin & Köpke (1986: 174).

Although uncertainty about what the noun will be – as measured by entropy – will be very high in both cases, given the great number of available candidates due to the very limited contextual clues provided by the sentences, the presence of a gendered article in German strongly reduces the number of candidates, since only masculine nouns in the singular are left. On the other hand, the entropy of inferring the form of the article will be higher in German, since its gender and number need to be guessed, differently than in English. Now, since there is much more uncertainty in guessing nouns than there is in guessing articles to begin with (there are much more nouns than articles, and thus many more possible outcomes in the former case), gender allows for a more balanced distribution of entropy in discourse. Since it is shown that in language use, speakers tend to manage the distribution of information exactly in this way, by avoiding peaks and troughs in entropy in their messages, gender can be considered as one of the ways in which this purpose can be reached. The results of Dye et al. (2017) are also interesting in that they show that generally speaking nouns that are semantically similar tend indeed to be assigned to the same gender in German, but this is true especially for low-frequency items, while high-frequency nouns with similar meanings tend to belong to different genders. This makes perfect sense if one of the functions of gender is discriminating between different potential outcomes, since discriminating between frequent outcomes is much more useful than doing so for rare outcomes. This also makes sense in terms of learnability: learners will be exposed early to high-frequency nouns and thus they will be able to assign them to the correct gender by rote, while semantic regularities will be useful in making the assignment of a gender to low-frequency items more predictable.

Dye et al. (2017) also observe that gender can play different roles in different languages: its function does not need to be the same in German and in Latin. Now, the results that we have presented here can be interpreted as evidence for an additional function that a classification of nouns in different genders fulfils, at least in Latin. It reduces uncertainty in the PCFP, enhancing both the predictability and the predictiveness of paradigm cells, and increasing the number of sets of cells that function as near principal parts at different cardinalities. Our results can thus be

considered as giving a quantitative confirmation to the findings of previous studies claiming that a lexeme's gender can be useful to predict the inflectional behaviour of a lexeme: cf. e.g. Aronoff (1994: Chapter 3) on Spanish, Russian and Latin itself, Thornton (2001) on Italian. However, it should be remembered that other studies have found a relation in the opposite direction, i.e. cases where the inflectional behaviour of a lexeme can be used to predict its gender: see e.g. Corbett (1982) and Fraser & Corbett (1995) on Russian. In Latin too relations in this direction certainly play a role. An interesting aspect that could not be investigated in the present work and that we leave for further research is thus the uncertainty in predicting a lexeme's gender knowing one (or more than one) of its wordforms – rather than predicting the content of other paradigm cells from one (ore more than one) of its wordforms and its gender, as we did in this section.

Lastly, there is a caveat that should be made concerning our results, relating to the actual availability of gender information for Latin nouns. Although it was shown in §5.3.1 that there are cues from which a speaker can infer the gender of a newly encountered noun (for instance, its meaning and the agreement markers it displays), such cues are not equally available in the same way for all cells and for all lexemes. For instance, in DAT.PL – that was observed to be the cell whose predictiveness was enhanced more strikingly when gender was added to the picture, cf. above §5.3.3 – the agreement markers are not at all informative on gender, since in that cell all the targets display the same inflectional realizations whatever the gender of the controller noun, as shown in (13): therefore, the only way in which one can have information on the gender of the involved nouns is by means of semantic cues, that, however, are only relevant for nouns denoting humans and superior animates.

(13) a. *illīs* *puerīs*
that:DAT.M.PL boy(M):DAT.PL

b. *illīs* *puellīs*
that: DAT.F.PL girl(F): DAT.PL

c. *illīs* *mālīs*
that: DAT.N.PL apple(N): DAT.PL

Furthermore, even when agreement is at least potentially a relevant source of information, as was shown to be the case for NOM.SG (cf. above, Table 17), it is possible that the speaker was exposed to the noun in a context where there was no agreement target, and in that case the speaker would be left with semantic factors alone. This means that in order to be able to have a more precise idea of the relevance of gender information in reducing uncertainty in the PCFP, we should also be able to evaluate the likelihood of having reliable information on the gender of a lexeme when exposed to one (or more) of its inflected wordforms, a question that we leave aside for further research because of its complexity.

5.4 Conclusion

In this chapter, after having detailed the facts of Latin noun inflection in §5.1, in §5.2 we presented a quantitative assessment of the reliability of implicative relations in noun paradigms, by applying a procedure similar to the one that was followed for verb paradigms. We have shown that when the tabular paradigm is substituted by the cell paradigm, no further simplification of paradigm structure can be achieved based on fully reliable bidirectional implicative relations – i.e., no distillation can be obtained with a number of cells smaller in size than the cell paradigm, that can be considered to be itself the distillation of Latin noun paradigms. This is very different than what was observed with verb paradigms, but this fact is not so surprising given the very different size of the cell paradigm of verbs (152 cells) and the one of nouns (10 cells). However, as in the cell paradigm of nouns there are no mutually interpredictable pairs of cells, perhaps it should rather be compared with the distillation of Latin verb paradigms, that constitutes exactly what is left of the complexity of verb inflection when abstracting away from similar patterns of full interpredictability between inflected words. Indeed, the size of the cell paradigm of nouns (10 cells) is very close to the size of the distillation of verb paradigms (15 zones of interpredictability). Therefore, paradigms that are very different in terms of sheer size prove to be far more similar in the size of their distillation – since, as we saw, the cell paradigm of Latin nouns can be considered to be also its distillation. This result can be interpreted as pointing in the same

direction of Ackerman et al. (2009)'s Low Entropy Conjecture (see above, §2.2), showing that paradigms that appear to be very complex can be proven to be strongly simplified by patterns of interpredictability between full inflected wordforms, posing limits to the difficulty of the PCFP, as can be measured by means of implicative entropy.

As far as principal parts are concerned, the smaller set that emerges from our entropy-based analysis is composed of three cells, and not of two like one would expect based on traditional descriptions. However, there are at least near principal part sets composed of two cells, even when setting the threshold at very low entropy values. The pair of cells that is usually used in Latin grammars and dictionaries – NOM.SG and GEN.SG – indeed constitutes a near principal part set, allowing us to predict the remaining paradigm cells with an average entropy of about 0.003. As we saw, not even fluent native speakers reach complete accuracy in their inflectional behaviour – i.e., they sometimes have doubts on what the inflected wordform that fills the paradigm cell of a lexeme should be. Therefore, it is reasonable to abstract away from such very small entropy values, as we do with near principal parts. Thus, our results can be taken as being consistent with the account of traditional descriptions in this respect.

In §5.3, we have described how the picture changes when another piece of information – beside the phonotactic shape of the wordforms – is assumed to be known, namely the gender of the involved lexeme. Our results show that a relevant reduction in entropy values can be obtained in this way. We have discussed the relevance of such result to the debate on the functions of an apparently redundant grammatical feature like gender, suggesting that it can actually be useful for speakers by allowing them to reduce uncertainty in inflectional predictions – along with the other functions that were already proposed in previous research.

Chapter 6. The impact of derivational relatedness on inflectional predictions

In this chapter, we will investigate how the picture of interpredictability between paradigm cells changes when taking into account not only the phonotactic shape of inflected wordforms, but also additional information of a different kind, namely the derivational relatedness of the lexemes involved. In §6.1, we will use some examples to illustrate the question and to stress the potential relevance of knowing whether two lexemes are ultimately derived from a same base on the one hand or whether they are formed by means of the same derivational process on the other hand, proposing a method to take this information into account and briefly discussing the difference with the standard procedure for entropy computation. We will start from verbal lexemes that ultimately derive from the same base in §6.2, proposing a working definition of the notion of derivational-inflectional family and showing how our data were coded so as to include a classification in such families in §6.2.1, briefly providing a qualitative picture of the inflectional behaviour of verbs that belong to the same derivational-inflectional family in §6.2.2 and presenting our results in §6.2.3. The same line of reasoning will be followed in §6.3 for nouns that are formed by means of the same derivational process – i.e., that belong to the same derivational-inflectional series, as defined in §6.2.1. Lastly, in §6.4, we will discuss the theoretical and methodological implications of our results, also highlighting some problems that we leave to further research.

6.1 The question

The results presented in the previous chapters have been obtained with the methodology described in Chapter 2, which starts from the assumption that only the phonotactic shape of the inflected wordforms is known. Only in §5.3 the role played by additional information of a different kind – namely, the gender of a noun – has been investigated. In this chapter, we will focus on another aspect that is potentially available to speakers when facing the PCFP, and that can have an impact on the

interpredictability of paradigm cells: information on the derivational relatedness of lexemes. To see the potential impact of this kind of information on uncertainty in inflectional predictions, let us consider the reduced dataset given in Table 1 below.

Table 1: Predicting PRS.ACT.IND.3SG from PRS.ACT.IND.1SG without information on derivational relatedness (reduced dataset)

lexeme (meaning)	PRS.ACT.IND.1SG	PRS.ACT.IND.3SG	pattern
'to dedicate'	*dicō*	*dicat*	1. X*ō* ↔ X*at*
'to take a little'	*lībō*	*lībat*	1. X*ō* ↔ X*at*
'to fold'	*plicō*	*plicat*	1. X*ō* ↔ X*at*
'to fold back'	*replicō*	*replicat*	1. X*ō* ↔ X*at*
'to unfold'	*explicō*	*explicat*	1. X*ō* ↔ X*at*
'to say'	*dīco*	*dīcit*	2. X*ō* ↔ X*it*
'to drink'	*bibō*	*bibit*	2. X*ō* ↔ X*it*
'to write'	*scrībō*	*scrībit*	2. X*ō* ↔ X*it*
'to write in'	*inscrībō*	*inscrībit*	2. X*ō* ↔ X*it*
'to write after'	*postscrībō*	*postscrībit*	2. X*ō* ↔ X*it*
'to write back'	*rescrībō*	?	P(1)=5/5 P(2)=5/5
'to call'	*vocō*	?	P(1)=5/5 P(2)=5/5

If the Latin lexicon only comprised these 10 verbs, there would be two different alternation patterns, each of them attested in the same number of verbs: in 1st conjugation verbs like LIBO, the PRS.ACT.IND.3SG is obtained by replacing the *-ō* of the PRS.ACT.IND.1SG with *-at*, while in 3rd conjugation verbs like BIBO it is replaced by *-it*. As usual, for the sake of simplicity we omit the context of application of the alternation patterns, but it is important to observe that in this reduced dataset this factor would not be decisive in constraining their applicability: different patterns are applied to verbs whose PRS.ACT.IND.1SG are maximally similar in phonotactic shape, like *dicō* 'I dedicate' (with PRS.ACT.IND.3SG *dicat*, applying pattern 1) and *dīcō* 'to say' (with PRS.ACT.IND.3SG *dīcit*, applying pattern 2), and not even vowel length – which is the only phonetic difference between these two wordforms – can be taken as decisive, as is shown by the PRS.ACT.IND.3SG forms *lībat* from PRS.ACT.IND.1SG *lībō* – with application of pattern 1, and not of pattern 2 like in

dīcō – and *bibit* from PRS.ACT.IND.1SG *bibō* – with application of pattern 2, and not of pattern 1 like in *dicō*.

Therefore, by computing entropy in the same way we did in the previous chapters, we would have two equiprobable outcomes, and consequently an entropy value of 1 bit, as shown in (1) below.

(1) $\mathrm{H} = -\left[\left(\frac{5}{10} \cdot \log 2 \frac{5}{10}\right) + \left(\frac{5}{10} \cdot \log 2 \frac{5}{10}\right)\right] = 1 \text{ bit}$

However, this result is based on the simplifying assumption that speakers do not have any information on the derivational relatedness of the verbs in the sample. To fully understand the impact of this fact on the computation, let us pretend that we actually need to predict the PRS.ACT.IND.3SG of two verbs that are not in the initial dataset, RESCRIBO and VOCO (given in the last two lines of Table 1). With this assumption, we are computing entropy as if the prediction task were equally difficult for these two verbs: in both cases, the probability of applying pattern 1 and the probability of applying pattern 2 would be the same.

However, when trying to predict the PRS.ACT.IND.3SG of RESCRIBO, speakers might notice that its meaning is similar to the one of SCRIBO, INSCRIBO and POSTSCRIBO – they all mean something related to 'write' – and that those verbs also have some portion of form in common – they all contain the sequence <*scrīb*> [skri:b]. Furthermore, the sequence <*re*> [re] also appears in REPLICO, with a meaning similar to the one that is found in RESCRIBO. Therefore, it is not unreasonable to suppose that speakers do notice all these facts, taking advantage of them to conclude that RESCRIBO is morphologically analysable as being composed of the preverb *re-* attached to the base lexeme SCRIBO. If that is the case, then by computing entropy in the standard way we would be considerably overestimating the difficulty of this prediction task, since complex verbs like this one usually display the same inflectional behaviour as the base they come from: in this case, therefore, we would expect to find pattern 2 like in SCRIBO, rather than pattern 1, although both of them would be applicable on the basis of the phonotactic shape of the wordform alone.

It is therefore interesting to see what happens if we start from the opposite assumption that speakers do have completely reliable information on the

derivational relatedness of the verbs in the sample. In the last column of Table 2, we thus code complex verbs according to their ancestor – i.e., the base from which they ultimately derive.[1]

Table 2: Predicting PRS.ACT.IND.3SG from PRS.ACT.IND.1SG with information on derivational relatedness: verbs with the same ancestor (reduced dataset)

lexeme (meaning)	PRS.ACT. IND.1SG	PRS.ACT. IND.3SG	pattern	derivational relatedness
'to dedicate'	*dicō*	*dicat*	1. X*ō* ↔ X*at*	simple
'to take a little'	*lībō*	*lībat*	1. X*ō* ↔ X*at*	simple
'to fold'	*plicō*	*plicat*	1. X*ō* ↔ X*at*	simple
'to fold back'	*replicō*	*replicat*	1. X*ō* ↔ X*at*	< PLICO
'to unfold'	*explicō*	*explicat*	1. X*ō* ↔ X*at*	< PLICO
'to say'	*dīco*	*dīcit*	2. X*ō* ↔ X*it*	simple
'to drink'	*bibō*	*bibit*	2. X*ō* ↔ X*it*	simple
'to write'	*scrībō*	*scrībit*	2. X*ō* ↔ X*it*	simple
'to write in'	*inscrībō*	*inscrībit*	2. X*ō* ↔ X*it*	< SCRIBO
'to write after'	*postscrībō*	*postscrībit*	2. X*ō* ↔ X*it*	< SCRIBO
'to write back'	*rescrībō*	?	P(1)=0 P(2)=2/2	< SCRIBO
'to call'	*vocō*	?	P(1)=3/6 P(2)=3/6	simple

To take this information into account in our entropy calculations, we can exploit the same functionality of the Qumin toolkit that was used to evaluate the impact of the classification of nouns into different genders, by computing entropy values separately for lexemes that belong to different classes. In this case, we would have one class composed of six morphologically simple lexemes, and two separate classes for complex lexemes: one class of verbs that come from PLICO (namely, REPLICO and EXPLICO) and one of verbs that come from SCRIBO (namely, INSCRIBO and POSTSCRIBO).

In this way, we can capture the intuition that when guessing the PRS.ACT.IND.3SG of a verb like RESCRIBO, the prediction task is actually much simpler: since in verbs that belong to the relevant class – i.e., in verbs that come from SCRIBO – only pattern 2 is attested in our dataset, there would be no uncertainty whatsoever on the

[1] See §6.2.1 below for a definition of the notion of "ancestor" and for a discussion of its usefulness for our purposes.

PRS.ACT.IND.3SG of that verb. On the other hand, there would still be uncertainty in guessing the PRS.ACT.IND.3SG of a verb like VOCO, since in this case we would look at the patterns that are used in morphologically simple lexemes, where we would still have two equiprobable outcomes, with patterns 1 and 2 applied in three verbs each.

If we compute the overall entropy estimating the uncertainty in guessing PRS.ACT.IND.3SG from PRS.ACT.IND.1SG for this 10-verb lexicon, we would thus have an entropy of 1 bit in the class of simple verbs, but in the two classes of complex verbs – the ones derived from SCRIBO and the ones derived from PLICO – entropy would be 0, yielding an overall entropy value of 0.6 bit, considerably lower than the one obtained without taking derivational information into account. The details of the computation are given in (2)-(5).

(2) Entropy in the class of simple lexemes

$$\mathrm{H} = -\left[\left(\frac{3}{6}\cdot \log 2\frac{3}{6}\right) + \left(\frac{3}{6}\cdot \log 2\frac{3}{6}\right)\right] = 1 \text{ bit}$$

(3) Entropy in the class of lexemes that derive from SCRIBO

$$\mathrm{H} = -\left(\frac{2}{2}\cdot \log 2\frac{2}{2}\right) = 0 \text{ bit}$$

(4) Entropy in the class of lexemes that derive from PLICO

$$\mathrm{H} = -\left(\frac{2}{2}\cdot \log 2\frac{2}{2}\right) = 0 \text{ bit}$$

(5) Overall entropy value

$$\mathrm{H} = \left[\left(\frac{6}{10}\cdot 1\right) + \left(\frac{4}{10}\cdot 0\right)\right] = 0.6 \text{ bit}$$

As is shown in Table 3, an analogous reduction in uncertainty can be obtained when in the sample there are verbs that are formed by means of the same derivational suffix.

Table 3: Predicting PRS.ACT.IND.3SG from PRS.ACT.IND.1SG with information on derivational relatedness: verbs that are formed by means of the same derivational suffix (reduced dataset)

lexeme (meaning)	PRS.ACT. IND.1SG	PRS.ACT. IND.3SG	pattern	derivational relatedness
'to dedicate'	*dicō*	*dicat*	1. X*ō* ↔ X*at*	simple
'to take a little'	*lībō*	*lībat*	1. X*ō* ↔ X*at*	simple
'to fold'	*plicō*	*plicat*	1. X*ō* ↔ X*at*	simple
'to cry out aloud'	*clāmitō*	*clāmitat*	1. X*ō* ↔ X*at*	suffix *-it-*
'to flee eagerly'	*fugitō*	*fugitat*	1. X*ō* ↔ X*at*	suffix *-it-*
'to say'	*dīco*	*dīcit*	2. X*ō* ↔ X*it*	simple
'to drink'	*bibō*	*bibit*	2. X*ō* ↔ X*it*	simple
'to write'	*scrībō*	*scrībit*	2. X*ō* ↔ X*it*	simple
'to become ill'	*aegrēscō*	*aegrēscit*	2. X*ō* ↔ X*it*	suffix *-sc-*
'to become white'	*albēscō*	*albēscit*	2. X*ō* ↔ X*it*	suffix *-sc-*
'to seek earnestly'	*quaeritō*	?	P(1)=2/2 P(2)=0	suffix *-it-*
'to call'	*vocō*	?	P(1)=3/6 P(2)=3/6	simple

If we take derivational information into account, when guessing the PRS.ACT.IND.3SG of a verb that contain the iterative/intensive suffix *-it-*, like QUAERITO, there is no uncertainty, since we only look at the patterns used in other verbs that display the same suffix, and they all apply pattern 1 in our dataset, as they do in the whole Latin lexicon. On the other hand, there still is uncertainty in inflectional predictions about verbs that do not contain any derivational suffix. Therefore, we obtain the same reduction in overall entropy that was obtained with the dataset of Table 2, since again entropy would be 1 in the 6-verb class of simple lexemes and 0 in the two classes of complex lexemes (the ones with the iterative/intensive suffix *-it-* and the ones with the inchoative suffix *-sc-*), each with

two members: the computation will thus be exactly the same as the one shown in examples (2)-(5), to which the reader is referred.

It is important to notice that the prediction task that is modelled in the examples of Table 2 and Table 3 is slightly different than the one that is tackled in the standard approach – the one of e.g. Bonami & Boyé (2014) and Beniamine (2018), used also in this work to obtain the results of Chapter 4 and of the first part of Chapter 5. In the standard procedure, information of the same kind is uniformly used for all verbs in the lexicon: the task can be summarized as in (6).

(6) For a lexeme L, predict cell B from cell A knowing the distribution of alternation patterns between A and B in the whole lexicon

On the other hand, in the task that is modelled in the examples of Table 2 and Table 3, the prediction is based on different information depending on the lexeme involved: if it is a morphologically simple lexeme, then we look at the distribution of alternation patterns in other simple lexemes; if it is complex, then we look at the distribution of alternation patterns either in lexemes that derive from the same ancestor – as in Table 2, cf. (7) – or in lexemes that are formed by means of the same suffix – as in Table 3, cf. (8).

(7) for a lexeme L:
   a. if L is a complex lexeme that has L_x as ancestor, predict cell B from cell A knowing the distribution of alternation patterns between A and B in the set of complex lexemes that have L_x as ancestor
   b. if L is a simple lexeme, predict cell B from cell A knowing the distribution of alternation patterns between A and B in the set of simple lexemes

(8) for a lexeme L:
   c. if L is formed by means of the derivational suffix X, predict cell B from cell A knowing the distribution of alternation patterns between A and B

in the set of complex lexemes that are formed by means of the derivational suffix X

d. if L is a simple lexeme, predict cell B from cell A knowing the distribution of alternation patterns between A and B in the set of simple lexemes

Given this sort of split in the modified procedure, it is interesting to analyse not only the overall entropy value on the whole lexicon, but also the results that can be obtained separately for simple lexemes on the one hand and for complex lexemes on the other one. This is exactly what we will do in the remainder of this chapter, where we will evaluate the impact of information on the derivational relatedness of lexemes on the much larger dataset of LatInfLexi, by grouping together complex lexemes that share the same ancestor in verb paradigms (§6.2), and complex lexemes that are formed by means of the same suffixes in noun paradigms (§6.3), because of the different quantitative relevance of the two classifications in different lexical categories, as we will detail in the following sections.

6.2 Verbs that derive from the same ancestor: derivational-inflectional families

In this section, we will evaluate how our results change if entropy is computed separately for different classes containing complex lexemes that share the same ancestor. We will focus on verb inflection, since in Latin there are a lot of verbs that can be analysed as being created by adding one of a set of frequently used prefixes to a base lexeme: therefore, there are a lot of verbs that share the same ancestor, and where the ancestor remains the locus of inflection, as required by definition (9) (cf. §6.2.1 below). On the other hand, prefixation plays a much less relevant role in nominal inflection, where suffixation is much more frequently used: therefore, the role of families is much less relevant to inflection there.

6.2.1 Coding derivational-inflectional families

Our first step was obtaining information on the derivational relatedness of the complex lexemes of our dataset in a systematic fashion. To do so, we have again exploited the morphological analyser Lemlat 3.0 (cf. above, Chapter 3), that incorporates information on derivational relatedness from the "Word Formation Latin" project[2] (Litta et al. 2016). In the analyses performed by Lemlat, for each wordform information is provided on the base lexeme(s) from which it derives and on the affix that is applied to the base to obtain it (if any), as is shown in Table 4 and Table 5: a derived verb like REMITTO is analysed as formed by applying the prefix *re-* to the base MITTO (cf. Table 4); on the other hand, a compound verb like MANUMITTO is decomposed in its two constituents MANUS and MITTO.

Table 4: Derivational information in Lemlat: derived lexemes

complex lexeme	affix	base lexeme
REMITTO 'to send back'	*re-*	MITTO 'to send'

Table 5: Derivational information in Lemla: compound lexemes

complex lexeme	constituent 1	constituent 2
MANUMITTO 'to release from one's power (*manus*)'	MANUS 'hand'	MITTO 'to send'

For our purposes, however, there is no need to distinguish between derived and compound verbs: the only thing that matters is that in both cases the inflectional behaviour of the derived lexeme is inherited from the verb MITTO: in Table 6 it can be observed that the – not fully predictable – inflectional patterns of these two lexemes are the same that are found in the base lexeme.

[2] The data are available online at http://wfl.marginalia.it/.

Table 6: The inflectional behaviour of MITTO and verbs that come from it: PRS.ACT.IND.1SG and PRF.ACT.IND.1SG

lexeme	PRS.ACT. IND.1SG	PRF.ACT. IND.1SG	pattern
MITTO 'to send'	*mittō*	*mīsī*	*_ittō* ↔ *_īsī*
REMITTO 'to send back'	*remittō*	*remīsī*	*_ittō* ↔ *_īsī*
MANUMITTO 'to release from one's power'	*manūmittō*	*manūmīsī*	*_ittō* ↔ *_īsī*

Therefore, in our classification we group these two lexemes together, as shown in Table 7. This is because they can be said to belong to the same **derivational-inflectional family**, which we define as in (9) below. In its turn, definition (9) is based on the notion of ancestor, whose definition is given in (10).

Table 7: Coding derivational-inflectional families

complex lexeme	derivational-inflectional family
REMITTO	MITTO
MANUMITTO	MITTO

(9) Two complex lexemes CLa and CLb belong to the same derivational-inflectional family if they both have the same simple lexeme SL as their ancestor, and the ancestor is the locus of inflection.

(10) The ancestor of a complex lexeme is the simple lexeme from which it ultimately derives

It is very important to stress that the notion of derivational-inflectional family that we adopt in this work is partly different from the notion of derivational family as is normally used in the literature (cf. e.g. Hathout 2011, Baayen 2014). For instance, the derivational family of MITTO is the same as that of REMITTO in the standard usage of the term in morphological literature, but MITTO is assigned to a different derivational-inflectional family in our classification, because it is a simple lexeme, and definition (9) only involves complex lexemes: all simple lexemes are simply lumped in a same group in our classification. Furthermore, definition (9) excludes from the derivational-inflectional family lexemes where the ancestor is the same,

but it is not the locus of inflection: for instance, INVOCO ‘to invoke’ and REVOCO ‘to call back’ belong to the same derivational-inflectional family, since the locus of inflection is the base lexeme VOCO ‘to call’, but VOCITO, ‘to call loudly’ does not, since the locus of inflection is rather the suffix *-it-* (see below, §6.3). Therefore, our results have been obtained by using a classification that is closely related to the notion of derivational family in its standard usage, but not equivalent to it.

While in the example of Table 7 it was possible to directly use the information provided by Lemlat, sometimes things are not so straightforward. For instance, let us have a look at the verbs of Table 8.

Table 8: Coding derivational relatedness: differences between Lemlat and our dataset

complex lexeme	base lexeme according to Lemlat	derivational-inflectional family in our dataset
STUPEFACIO ‘to stupefy’	FACIO	FACIO
OBSTUPEFACIO ‘to astonish’	STUPEFACIO	FACIO

The lexeme OBSTUPEFACIO ‘to astonish’ is correctly analysed by Lemlat as being derived by applying the prefix *ob-* to the base STUPEFACIO ‘to stupefy’. However, by relying solely on this analysis, we would miss the fact that STUPEFACIO is itself a compound verb, with STUPEO ‘to be stunned’ as first constituent and FACIO ‘to make’ as second constituent: the base from which OBSTUPEFACIO ultimately derives is thus FACIO, rather than STUPEFACIO. Therefore, our coding systematically differs from Lemlat’s one in such cases: OBSTUPEFACIO and STUPEFACIO are grouped together in our dataset, on the basis of definitions (9) and (10): they both have the simple lexeme FACIO as ancestor, and the ancestor is the locus of inflection

Table 9: The inflectional behaviour of FACIO, STUPEFACIO and OBSTUPEFACIO

lexeme	PRS.ACT. IND.1SG	PRF.ACT. IND.1SG	pattern
FACIO ‘to make’	*faciō*	*fēcī*	*_a_iō ↔ _ē_ī*
STUPEFACIO ‘to stupefy’	*stupefaciō*	*fēcī*	*_a_iō ↔ _ē_ī*
OBSTUPEFACIO ‘to astonish’	*obstupefaciō*	*obstupefēcī*	*_a_iō ↔ _ē_ī*

Conversely, a compound verb like AEDIFICO ‘to erect a building’ is analysed by Lemlat as being derived from FACIO ‘to make’: therefore, it would belong to the same derivational family as e.g. STUPEFACIO according to the usual definition. However, in our data we treat these verbs as belonging to a separate derivational-inflectional family, as shown in Table 10.

Table 10: Coding derivational relatedness: differences between Lemlat and our dataset

complex lexeme	base lexeme according to Lemlat	derivational-inflectional family in our dataset
AEDIFICO ‘to erect a building’	FACIO	*FICO
STUPEFACIO ‘to stupefy’	FACIO	FACIO

This choice is justified by the fact that verbs like AEDIFICO are actually formed by means of a different derivational process.[3] Therefore, the inflectional behaviour of such verbs (that are inflected like regular 1st conjugation verbs, as one can see in Table 11) cannot in any way be considered as inherited from the ancestor FACIO (that belongs to the mixed conjugation), as is required by definition (9): therefore, they do not belong to the same derivational-inflectional family in our classification.

Table 11: The inflectional behaviour of FACIO and AEDIFICO

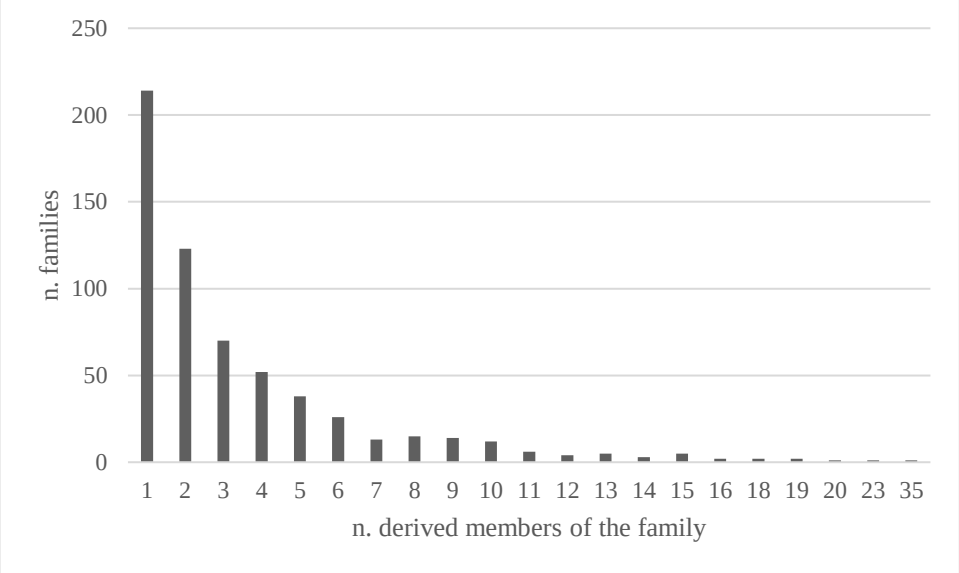
lexeme	PRS.ACT. IND.1SG	PRF.ACT. IND.1SG	pattern
FACIO ‘to make’	*faciō*	*fēcī*	*_a_iō ↔ _ē_ī*
AEDIFICO ‘to erect a building’	*aedificō*	*aedificāvī*	*_ō ↔ _āvī*

Once we have coded all the verbs of our sample in this way, we have on the one hand a very large class comprising simple lexemes and the few complex lexemes that happen to be the only member of their derivational family – for instance, PERDEPSO ‘to knead over’ which is present in LatInfLexi, unlike its base DEPSO ‘to knead’. This large class contains 1,187 verbs. On the other hand, there are many (609) small classes of complex lexemes, one for each different ancestor: each class thus corresponds to a derivational-inflectional family. The largest family is the one

[3] See e.g. Brucale (2012: 112), where compounds in *-ficō* and compounds in *-faciō* are listed as two separate sub-classes.

of verbs that come from FACIO, with 35 members. Most families, however, are very small. The shape of the curve in the plot of Figure 1 indicates a Zipfian[4] distribution, with many small families and a few large families: on the left side, we can see that there are more than two hundred derivational-inflectional families with only one member[5], and more than one hundred families with only two; on the opposite side, we have only one family with 35 members, one with 23, one with 20, and so on.

Figure 1: Derivational-inflectional families with different number of members



Given this classification and the procedure outlined above in §6.1, we will have on the one hand a single entropy value estimating the uncertainty of the PCFP for simple lexemes; on the other hand, for complex lexemes we will obtain many entropy values estimating the uncertainty of the PCFP in each derivational-inflectional family. Therefore, it is useful to have a qualitative look at the actual inflectional behaviour of verbs that belong to the same derivational-inflectional family: this is what we will do in the next sub-section.

[4] Cf. the well-known distribution of the so-called "Zipf's law" (cf. Zipf 1935), stating that the frequency of a word in a given corpus is inversely proportional to its rank in the list of words sorted by frequency – i.e., that in a given corpus there tend to be few frequent words and many rare words.
[5] On the basis of definition (9), the ancestor is not included in the derivational-inflectional family. Therefore, in this class we have e.g. the family of verbs that come from AMO 'love', that has only ADAMO 'to love truly' as member, AMO itself being left out because it is the ancestor.

6.2.2 The inflectional behaviour of verbs in the same family

In many cases, complex verbs that belong to the same derivational-inflectional family systematically display the same inflectional behaviour in all cells, inheriting all the alternation patterns that appear in their shared ancestor. For instance, DICO is a 3rd conjugation verb (cf. the PRS.ACT.INF *dīcere*) whose perfect stem and 3rd stem are obtained by suffixation of *-s-* (cf. the PRF.ACT.IND.1SG *dīx-ī* [di:ksi:]) and *-t-* (cf. the SUP.ACC *dict-um*), respectively: the same holds true for all the verbs that come from DICO, as illustrated in Table 12.

Table 12: The inflectional behaviour of DICO and verbs that derive from it

lexeme	PRS.ACT. IND.1SG	PRS.ACT. INF	PRF.ACT. IND.1SG	SUP.ACC
DICO ‘to say’	*dīcō*	*dīcere*	*dīxī*	*dīctum*
EDICO ‘to declare’	*ēdīcō*	*ēdīcere*	*ēdīxī*	*ēdīctum*
BENEDICO ‘to praise’	*benedīcō*	*benedīcere*	*benedīxī*	*benedīctum*
MALEDICO ‘to curse’	*maledīcō*	*maledīcere*	*maledīxī*	*maledīctum*

In other cases, the inflectional behaviour of complex verbs that share the same ancestor is at least partly different from the one of the ancestor itself, as in the example of Table 13. Verbs that derive from RAPIO, like e.g. ABRIPIO and PRORIPIO, belong to the mixed conjugation like their shared ancestor, with which they also share the suffixation of *-u-* to form the perfect stem and of *-t-* to form the third stem. However, the alternation patterns are not always exactly the same, because the complex verbs display *-i-* (cf. PRS.ACT.IND.1SG *abripiō*) or *-e-* (cf. SUP.ACC *abreptum*) where the ancestor had *-a-* (PRS.ACT.IND.SG *rapiō*, SUP.ACC *raptum*).

Table 13: The inflectional behaviour of RAPIO and verbs that derive from it

lexeme	PRS.ACT. IND.1SG	PRS.ACT. INF	PRF.ACT. IND.1SG	SUP.ACC
RAPIO 'to drag'	*rapiō*	*rapere*	*rapuī*	*raptum*
ABRIPIO 'to drag away'	*abripiō*	*abripere*	*abripuī*	*abreptum*
PRORIPIO 'to snatch forth'	*prōripiō*	*prōripere*	*prōripuī*	*prōreptum*

This is due to the effect of a phonological process of weakening of short vowels in non-initial syllables (cf. Oniga 1990, Weiss 2009: 116 ff.) that was active in Latin at a very early stage (cf. Cser 2016: 88 ff.), yielding /i/ in open syllable and /e/ in closed syllable, as illustrated in (11a-b).

(11) a. /ab/ + /ra.pi.o:/ → /ab.ri.pi.o:/ (open syllable)

b. /ab/ + /rap.tum/ → /ab.rep.tum/ (closed syllable)

However, this phonological process is no longer active in Classical Latin, as is demonstrated by the presence of complex verbs where the vowel of the ancestor is transparently preserved, like CALEFACIO in Table 14, unlike INFICIO, that shows the effects of the aforementioned process of vowel weakening.[6]

Table 14: The inflectional behaviour of FACIO and verbs that derive from it

lexeme	PRS.ACT. IND.1SG	PRS.ACT. INF	PRF.ACT. IND.1SG	SUP.ACC
FACIO 'to make'	*faciō*	*facere*	*fēcī*	*factum*
CALEFACIO 'to make warm'	*calefaciō*	*calefacere*	*calefēcī*	*calefactum*
INFICIO 'to put into'	*inficiō*	*inficere*	*infēcī*	*infectum*

Therefore, although CALEFACIO and INFICIO still fit the requirements of the definition in (9) and can therefore be taken as belonging to the same derivational-

[6] Verbs in *-faciō* also differ from verbs in *-ficiō* in other respects, for instance the presence of some (limited) degree of separability (cf. e.g. Cato *Agr*. 157, 9, ***ferve** bene **facito***) and freedom in the order of the two constituents (cf. e.g. Varr. *Rust.* 1, 41, 1, *facit putre*), as shown in Fruyt (2001).

inflectional family, since they both have the simple lexeme FACIO as ancestor and as the locus of inflection, their inflectional behaviour is not exactly the same. As a consequence, the entropy values estimating the uncertainty in the PCFP for verbs of this derivational-inflectional family will not be null: for instance, there is uncertainty when predicting PRS.ACT.IND.1SG from PRF.ACT.IND.1SG, because two different patterns are in competition, as shown in Table 15.

Table 15: Competition between different patterns in the derivational-inflectional family of verbs that derive from FACIO: predicting PRS.ACT.IND.1SG from PRF.ACT.IND.1SG

lexeme	family	PRF.ACT. IND.1SG	PRS.ACT. IND.1SG	pattern
CALEFACIO 'to make warm'	FACIO	*calefēcī*	*calefaciō*	*_ē_ī ↔ _a_iō*
INFICIO 'to put into'	FACIO	*infēcī*	*inficiō*	*_ē_ī ↔ _i_iō*

Another – less direct – effect of the same phonological process of vowel weakening is illustrated in Table 16, where we show some verbs derived from DO.

Table 16: The inflectional behaviour of DO and verbs that derive from it

lexeme	PRS.ACT. IND.1SG	PRS.ACT. IND.1PL	PRS.ACT. IND.3PL
DO 'to give'	*dō*	*damus*	*dant*
CIRCUMDO 'to put around'	*circumdō*	*circumdamus*	*circumdant*
OBDO 'to put against'	*obdō*	*obdimus*	*obdunt*

Among them, alongside verbs like CIRCUMDO that are transparently inflected exactly like the ancestor, we also find verbs that behave differently, like OBDO. In such verbs, forms like PRS.ACT.IND.3PL *obdunt* should be explained as due to analogical levelling on the model of 3rd conjugation verbs, on the basis of a proportional analogy like the one in (12) starting from forms like PRS.ACT.IND.1PL *obdimus*, whose ending coincides with the one of 3rd conjugation verbs simply

because of the application of the phonological process in (10) (cf. Ernout 1914: §261).

(12) *legimus : legunt = obdimus :* ***obdunt***

In this case too, such differences in the inflectional behaviour of verbs that belong to the same derivational-inflectional family will produce some uncertainty when computing entropy inside the involved family: knowing a cell like PRS.ACT.IND.1SG, where the same ending is used both in CIRCUMDO and in OBDO, the entropy of guessing other cells where the inflectional behaviour of these verbs is different will not be null.

Of course, the actual entropy value will depend on the number of verbs that display each of the different possible inflectional behaviours. For instance, regarding verbs that derive from DO (cf. Table 16), the lexemes that transparently exhibit the exact same behaviour of the ancestor are much rarer in LatInfLexi than the ones that display the effects of the phonological process in (11) and of the analogical levelling in (12), as is shown in Table 18. Therefore, the patterns that correspond to the latter type will be more likely to be applied, and the corresponding entropy value will not be very high. Conversely, in verbs that derive from FACIO (cf. Table 14), the verbs that are transparently inflected like the ancestor are more numerous, although the situation is much more balanced. (cf. Table 17). In the next section, a similar quantitative evaluation will be performed systematically for all derivational-inflectional families.

Table 17: The inflectional behaviour of verbs that derive from FACIO

inflectional behaviour	n. verbs
≈ CALEFACIO	22
≈ INFICIO	13

Table 18: The inflectional behaviour of verbs that derive from DO

inflectional behaviour	n. verbs
≈ CIRCUMDO	2
≈ OBDO	18

6.2.3 Results

With the qualitative picture sketched in the previous section in mind, we are finally in a position to move to the results that have been automatically obtained by means of the Qumin toolkit, focusing on the 15-cell distillation of Latin verb paradigms obtained in §4.5: in Table 19, we show the values of average cell predictability and predictiveness of one cell for each zone of interpredictability, with and without the classification in derivational-inflectional families.

Table 19: Average cell predictability and predictiveness in verb inflection with and without information on derivational-inflectional families

cell	no derivational information		information on derivational-inflectional families	
	average cell predictability	average cell predictiveness	average cell predictability	average cell predictiveness
Z1	0.229066	0.3122	0.05725	0.099836
Z2	0.342413	0.130643	0.106123	0.024179
Z3	0.315026	0.13107	0.087478	0.025986
Z4	0.231378	0.916636	0.059399	0.309907
Z5	0.315126	0.166161	0.091519	0.031317
Z6	0.269721	0.355214	0.078095	0.097872
Z7	0.311636	0.089352	0.087097	0.02415
Z8	0.309003	0.079394	0.085048	0.022384
Z9	0.302484	0.127819	0.086256	0.028663
Z10	0.343957	0.442993	0.097959	0.142516
Z11	0.244131	0.189036	0.070395	0.030202
Z12	0.240871	0.348084	0.059446	0.111953
Z13	0.208271	0.370468	0.053944	0.112975
Z14	0.255304	0.266108	0.079159	0.063405
Z15	0.263901	0.257111	0.085751	0.059574
Average implicative entropy	0.278819		0.078995	

These results provide a first answer to the empirical question regarding the impact of information on derivational relatedness on uncertainty in inflectional predictions:

it emerges very clearly that if the classification of verbs into different derivational-inflectional families is assumed to be known, the difficulty of the PCFP in Latin verb inflection is greatly reduced. The reduction in uncertainty is on the one hand quantitatively very relevant, on the other hand consistent across all the cells of the distillation.

Given the different information on which the computation is based in the modified task, as summarized above in (7), it is interesting to go beyond this overall result, by looking separately at the entropy value that refers to the class comprising simple lexemes (and complex lexemes that happen to be the only members of their morphological family, cf. §6.2.1 above) on the one hand, and to the entropy value that is found in the many classes corresponding to the various derivational-inflectional families, on the other hand: this is what we do in Table 20 and Table 21 below. In Table 22, we compare the values of average implicative entropy.

Table 20: Simple and complex verbs: average cell predictability

cell	no derivational information	information on derivational-inflectional families	
		simple lexemes only	complex lexemes only
Z1	0.229066	0.149572	0.006559
Z2	0.342413	0.2788	0.010579
Z3	0.315026	0.231711	0.007974
Z4	0.231378	0.157475	0.005492
Z5	0.315126	0.243021	0.007949
Z6	0.269721	0.209294	0.005738
Z7	0.311636	0.231628	0.007542
Z8	0.309003	0.225732	0.007446
Z9	0.302484	0.228976	0.007472
Z10	0.343957	0.242879	0.020711
Z11	0.244131	0.187238	0.007982
Z12	0.240871	0.156423	0.006666
Z13	0.208271	0.143236	0.004886
Z14	0.255304	0.205141	0.012417
Z15	0.263901	0.211056	0.019875

Table 21: Simple verbs and complex verbs: average cell predictiveness

cell	no derivational information	information on derivational-inflectional families	
		simple lexemes only	complex lexemes only
Z1	0.3122	0.268636	0.006874
Z2	0.130643	0.059701	0.004969
Z3	0.13107	0.063113	0.005932
Z4	0.916636	0.833136	0.021301
Z5	0.166161	0.077997	0.005928
Z6	0.355214	0.264829	0.006534
Z7	0.089352	0.059932	0.004803
Z8	0.079394	0.05576	0.004436
Z9	0.127819	0.073339	0.004421
Z10	0.442993	0.351879	0.030327
Z11	0.189036	0.078581	0.004402
Z12	0.348084	0.299866	0.008615
Z13	0.370468	0.298806	0.010226
Z14	0.266108	0.163506	0.010376
Z15	0.257111	0.153098	0.010142

Table 22: Simple and complex verbs: average implicative entropy

	no derivational information	information on derivational-inflectional families	
		simple lexemes only	complex lexemes only
Average implicative entropy	0.278819	0.206812	0.009286

These results show that there is a dramatic reduction of entropy values in complex lexemes. This is not at all unexpected: verbs in the same derivational-inflectional family are usually inflected in the same way, and cases where different inflectional behaviours are displayed by verbs that share the same ancestor – like in the examples of Table 14 and Table 16 – are not quantitatively very relevant. Therefore, entropy unsurprisingly approaches 0 if only complex lexemes are considered.

On the other hand, it can be observed that we also find a reduction in uncertainty in the class of simple lexemes, if compared with the entropy values that were obtained on all lexemes with no information on derivational relatedness. This difference, although quantitively less relevant, is nevertheless interesting, and calls for an explanation.

It is likely that the reason for this fact is to be found in the correlation of the token frequency of a verb with the size of its derivational-inflectional family on the one hand and with the presence of rare alternation patterns on the other hand. Verbs with a large morphological family usually have a high token frequency,[7] and it is well known that it is exactly in verbs with high token frequency that rare alternation patterns tend to appear. Therefore, counting each member of a large derivational-inflectional family as a separate type – as happens when computing entropy on the overall lexicon with no information on derivational relatedness – might lead to an overestimation of the weight of such rare alternation patterns, and therefore to a less skewed distribution and to more uncertainty, reflected in higher entropy values. For instance, the alternation pattern displayed by MITTO between PRS.ACT.IND.1SG and PRF.ACT.IND.1SG (cf. Table 23) is considered to have a relatively high type frequency, since it also appears in all the 19 derived verbs of its derivational-inflectional family (e.g. ADMITTO, CIRCUMMITTO, MANUMITTO, etc.).

Table 23: PRS.ACT.IND.1SG and PRF.ACT.IND.1SG: the alternation pattern displayed by MITTO and verbs that derive from it

lexeme (meaning)	PRS.ACT.IND.1SG	PRF.ACT.IND.1SG	alternation pattern
'send'	*mittō*	*mīsī*	*_ittō ↔ _īsī*

We will elaborate on this point below in §6.4, where, besides giving a more careful explanation, we will also discuss its theoretical and methodological consequences.

[7] If we consider the number of lexemes in each of the derivational-inflectional families of our sample, it can be observed that in 75% of the families whose value is above the 90th percentile, the ancestor of the family also has a token frequency value above the 90th percentile of the distribution of frequencies of lexemes according to the data of Delatte et al. (1981) – i.e., they occur very often in texts, displaying a token frequency of 110 or more. The percentage arrives at 86% if we look at the frequency of the most frequent member of the family, rather than at the ancestor: for instance, in the derivational-inflectional family of verbs that come from IACIO 'throw' although the ancestor has a token frequency of 105, and thus it is not above the 90th percentile of the distribution of frequencies of Delatte et al. (1981), there is another member of the family, namely ADICIO 'throw to', who is above the 90th percentile, with frequency 281.

6.3 Nouns that are formed by means of the same derivational suffix: derivational-inflectional series

In the previous section, to investigate the impact of derivational-inflectional families, we have focused on verb paradigms. This was justified by the fact that complex verbs formed by adding a prefix to an existing verbal base are very numerous in Latin. Conversely, prefixation is much less productive as a word-formation strategy for nouns, where it is suffixation that has the lion's share. Therefore, in this section we will exploit noun paradigms to investigate the role of a different kind of derivational relatedness, due to the fact that two lexemes are formed by means of the same suffixal derivational process.

6.3.1 Derivational-inflectional series: coding and inflectional behaviour

Let us start from an operational definition of the notion of **derivational-inflectional series**, parallel to the one provided in (9) for derivational-inflectional families.

(12) Two complex lexemes CLa and CLb belong to the same derivational-inflectional series if they are both formed by means of the same derivational process, that provides the instruction to obtain all the relevant inflected wordforms.

Nouns that contain the same prefix, besides being very rare, would not fall into the scope of definition (12), since prefixes do not provide the instruction to obtain the inflected wordforms of the lexemes formed by means of them, differently than suffixes, and in our methodology derivational relatedness is only taken into account when there is a correlation with the inflectional behaviour of lexemes – cf. also the definition given in (9) for derivational-inflectional families.
On the other hand, in cases of conversion, the absence of a formal marker of the derivational process beside the different inflectional behaviour would raise serious methodological questions. Therefore, it seemed safer to exclude such cases from our account of the impact of derivational information.

The source of information used to classify nouns in different derivational-inflectional series according to this definition was again WFL (cf. §6.2.1 above): this time, rather than lexemes with the same ancestor, as we did for verbs, we have grouped together nouns that contain the same suffix according to WFL. In this way, 14 distinct derivational-inflectional series were found in our data: they are listed in Table 24, where for each series we also provide one example and the number of lexemes it includes. Globally, the complex lexemes in our dataset are 234, the remaining 804 being simple.

Table 24: Derivational-inflectional series in the nouns of LatInfLexi

derivational-inflectional series[8]	n. nouns	example	
		complex noun	base
-(t)io	47	MUTATIO 'change'	$MUTO_V$ 'to change'
-ia	42	MISERIA 'misery'	$MISER_A$ 'miserable'
-tās	42	CELERITAS 'quickness'	$CELER_A$ 'quick'
-or	21	CLAMOR 'shout'	$CLAMO_V$ 'to shout'
-(t)or	19	ORATOR 'speaker'	ORO_V 'to speak'
-mentum	15	ALIMENTUM 'nourishment'	ALO_V 'to nourish'
-tudo	15	CLARITUDO 'clearness'	$CLARUS_A$ 'clear'
-itia	12	AMICITIA 'friendship'	$AMICUS_A$ 'friendly'
-men	5	CERTAMEN 'contest'	$CERTO_V$ 'to contend'
-ll(us/-a/-um)	4	LIBELLUS 'little book'	$LIBER_N$ 'book'
-(t)ura	4	NATURA 'nature'	$NASCOR_V$ 'to be born'
-trum	4	CLAUSTRUM 'lock'	$CLAUDO_V$ 'to shut'
-crum	2	SIMULACRUM 'representation'	$SIMULO_V$ 'to imitate'
-ēla	2	QUERELA 'complaint'	$QUEROR_V$ 'complain'

Regarding the inflectional behaviour of lexemes that belong to the same derivational-inflectional series, the situation is usually very straightforward: if two nouns are formed by means of the same suffixal derivational process, then they almost always display the same alternation patterns in all paradigm cells. Trivially, this happens because suffixes in Latin, besides assigning a specific gender to the nouns they form (as we saw above in §5.3.1, Footnote 10), also include the instruction on how to obtain all the paradigm cells of the derivative: the locus of inflection is the suffix in such complex nouns. For instance, the quality noun

[8] As a label for the series, we use the form of the suffix as it appears in the nominative singular.

forming suffix *-ia* assigns the lexemes that it creates to the feminine gender and to the 1st declension, while the action noun forming suffix *-mentum* is neuter and belongs to the 2nd declension, and nouns containing the agentive suffix *-tor* are masculine and are assigned to the 3rd declension.

The only exception to this generalization is constituted by diminutive suffixes like *-ll-* in Table 24. This suffix is transparent to the gender of the base to which it is applied, and nouns that are created by means of this derivational process are inflected as 1st declension nouns if they come from a feminine noun and as 2nd declension nouns if they come from a masculine or neuter noun. As is shown in Table 25, of the 4 nouns[9] displaying this suffix contained in LatInfLexi, one derives from a feminine noun and is therefore inflected like 1st declension nouns, while the other ones are inflected like 2nd declension nouns of different gender-based sub-classes: the one of masculine nouns in the two cases where the base itself is masculine, the one of neuter nouns when the base is neuter.

Table 25: Diminutives in *-ll-* in LatInfLexi

base	complex lexeme	inflectional behaviour
TABULA$_F$ ‘board’	TABELLA$_F$ ‘little board’	1st declension
OCULUS$_M$[10] ‘eye’	OCELLUS$_M$ ‘little eye’	2nd declension (masculine sub-class)
LIBER$_M$ ‘book’	LIBELLUS$_M$ ‘little book’	2nd declension (masculine sub-class)
FLAGRUM$_N$ ‘whip’	FLAGELLUM$_N$ ‘whip’	2nd declension (neuter sub-class)

[9] Of course, it would be preferable to have more than 4 nouns to be able to estimate more realistically the probability of application of the different inflectional patterns, but on the limitations of our noun sample see Chapter 3 above.

[10] At least from a diachronic point of view, OCULUS itself can be considered as a diminutive formed by means of the suffix *-ul(us/-a/-um)*: however, the base *OCUS from which both OCULUS and OCELLUS are derived is not attested. The same situation is found in the two diminutives TABULA and TABELLA. In our classification, we follow WFL and Lemlat in considering TABULA and OCULUS as the bases of TABELLA and OCELLUS. In any event, what matters in this context is the presence of the suffix, rather than the possibility to clearly identify a base.

6.3.2 Results

Regarding noun inflection, we will compare the results obtained with and without the classification in derivational-inflectional series directly on the full cell paradigm, since we have seen in §5.2.2 that no smaller distillation can be found based on mutual interpredictability between wordforms. The comparison of the results is given in Table 26 below.

Table 26: Average cell predictability and predictiveness in noun inflection with and without information on derivational-inflectional series

cell	no derivational information		information on derivational-inflectional series	
	average cell predictability	average cell predictiveness	average cell predictability	average cell predictiveness
NOM.SG	0.405303	0.311722	0.35942	0.297267
GEN.SG	0.246033	0.256772	0.203911	0.227088
DAT.SG	0.310147	0.27068	0.268091	0.241039
ACC.SG	0.23403	0.479689	0.180428	0.451156
VOC.SG	0.445589	0.199603	0.400022	0.187816
ABL.SG	0.301563	0.265813	0.253256	0.236968
NOM.PL	0.490967	0.190802	0.417078	0.169711
GEN.PL	0.474244	0.463356	0.397222	0.290467
DAT.PL	0.175355	0.982444	0.148133	0.806233
ACC.PL	0.535256	0.197606	0.458671	0.178488
Average implicative entropy	0.361849		0.308623	

The difference in the impact of information on derivational relatedness with the picture that was sketched for derivational-inflectional families in verb inflection in §6.2.3 is striking: the reduction in uncertainty that is achieved by adding information on the derivational-inflectional series of nouns is much less relevant than the one that was achieved by adding information on the derivational-inflectional families of verbs. In the latter case, the entropy values approach 0 (cf. Table 20, Table 21 and Table 22 above), while in the former they are not so different from the ones computed without derivational information.

This difference might be at least partly due to the different size of the datasets. Both our samples are frequency based, but the verb lexicon contains 3,348 entries, the

noun lexicon only 1,038: of course, the quantitative weight of complex lexemes is expected to be more relevant if there are many rare lexemes.

Therefore, it is useful to compare the results regarding families to the ones regarding series by using a sample of comparable size for verbs, including only the 1,094 verbal lexemes whose token frequency is more than 25 in the data of Delatte et al. (1981). Table 27 shows the number and percentage of derived lexemes in the different samples – the one of nouns, the one with all verbs, the one with 1,095 verbs only. It can be observed that the proportion of derived lexemes is indeed smaller in the reduced sample of verbs than in the complete one, although it is still remarkably greater than the proportion of derived lexemes in the sample of nouns.

As for the impact on entropy values, the results presented in Table 28 show that the reduction in uncertainty achieved by adding information on derivational-inflectional families with the reduced dataset of verbs is only slightly smaller than the one obtained using the whole lexicon, but it is still more relevant than the one obtained for noun inflection by means of the classification in derivational-inflectional series.

Table 27: Number and percentage of derived lexemes in different samples

sample	n. derived lexemes	% derived lexemes
verb paradigms (3,348 lexemes)	2,160	64.52 %
verb paradigms (1,094 lexemes)	612	55.94 %
noun paradigms (1,038 lexemes)	234	22.54 %

Table 28: Average implicative entropy with and without information on derivational relatedness: noun sample vs. reduced verb sample

noun paradigms (1,038 lexemes)		verb paradigms (1,094 lexemes)	
without derivational information	with derivational information	without derivational information	with derivational information
0.361849	0.308623	0.286686	0.106461

If we now look at the results regarding simple and complex lexemes separately also for derivational-inflectional series (cf. Table 29, Table 30 and Table 31 below), as we did for families (cf. §6.2.3 above), another interesting difference emerges.

Table 29: Simple and complex nouns: average cell predictability

cell	no derivational information	information on derivational-inflectional series			
		simple lexemes only		complex lexemes only	
NOM.SG	0.405303	0.462684		0.009432	
GEN.SG	0.246033	0.263689		0.001548	
DAT.SG	0.310147	0.346716		0.001548	
ACC.SG	0.23403	0.233206		0.001548	
VOC.SG	0.445589	0.515511		0.009432	
ABL.SG	0.301563	0.327504		0.001548	
NOM.PL	0.490967	0.536967		0.009412	
GEN.PL	0.474244	0.513389		0.001541	
DAT.PL	0.175355	0.191822		0	
ACC.PL	0.535256	0.590756		0.009412	

Table 30: Simple and complex nouns: average cell predictiveness

cell	no derivational information	information on derivational-inflectional series			
		simple lexemes only		complex lexemes only	
NOM.SG	0.311722	0.385		0	
GEN.SG	0.256772	0.292452		0.005256	
DAT.SG	0.27068	0.310699		0.005256	
ACC.SG	0.479689	0.582733		0.005256	
VOC.SG	0.199603	0.243273		0	
ABL.SG	0.265813	0.305286		0.005256	
NOM.PL	0.190802	0.219663		0	
GEN.PL	0.463356	0.374333		0.005244	
DAT.PL	0.982444	1.0378		0.019152	
ACC.PL	0.197606	0.231003		0	

Table 31: Simple and complex nouns: average implicative entropy

	no derivational information	information on derivational-inflectional series	
		simple lexemes only	complex lexemes only
Average implicative entropy	0.361849	0.398224	0.004542

It can be observed that while uncertainty is (almost) completely removed in the classes of complex nouns (as is expected, given the observations on their inflectional behaviour made in §6.3.1), in simple nouns the numbers are similar to the ones obtained on the whole lexicon without derivational information, and we

actually find a small increase in entropy values. Therefore, in this respect our results regarding the role of derivational-inflectional series are different than the ones regarding derivational-inflectional families, where – as we have seen above in §6.2.3 – there is a non-negligible reduction in uncertainty in the results obtained by considering only simple lexemes, if compared to the ones obtained on the whole lexicon (cf. Table 20 and Table 21 above).

Arguably, this is due to the fact that, as far as derivational-inflectional series are concerned, there is no correlation comparable to the one that was discussed above in §6.2.3 between the token frequency of lexemes, their inflectional behaviour and the size of their derivational-inflectional family. Furthermore, there is variation in the declension to which nouns formed by means of different derivational processes are assigned, as was noted above (cf. §6.3.1). Lastly, the derivational-inflectional series shown above in Table 24 often contain nouns whose inflectional behaviour is already highly predictable, even without taking derivational information into account: therefore, if complex nouns are left out, uncertainty is not necessarily expected to decrease.

6.4 Discussion

The results presented in the previous sections raise an interesting general methodological question regarding the way in which it is more reasonable to compute entropy values. As we have seen, in this work – like in previous studies conducted within this framework, e.g. Bonami & Boyé 2014 and Beniamine 2018 – as an estimate of the probability of application of different patterns, their type frequency is used, counting the number of lexemes in which they occur in the lexicon, disregarding token frequency, i.e. the number of occurrences of the inflected wordforms of such lexemes in texts. This choice is perfectly reasonable, and it is in line with Bybee's (1995: 433 ff.) observation that it is exactly type frequency that correlates with the productivity of morphological patterns, rather than token frequency: on the contrary, if the wordforms involved in the pattern have a high frequency in texts in terms of tokens, then it is more likely that they will be stored as such, rather than obtained from one another by means of the application

of the alternation pattern. Therefore, high token frequency actually detracts from the strength of a given alternation pattern, and it is preferable to rely on type frequency as an estimate of probability of application in entropy computations.

We have observed in §6.1 that in the standard procedure that is followed to compute entropy, derived lexemes that share the same ancestor are counted as different types. Thus, a minor alternation pattern like the one displayed by MITTO between the PRS.ACT.IND.1SG *mittō* and the PRF.ACT.IND.1SG *mīsī* (cf. Table 23 above) is considered to have a relatively high type frequency, since it also appears in all the 19 derived verbs of its derivational-inflectional family (e.g. ADMITTO, CIRCUMMITTO, MANUMITTO, etc.). However, the presence of many derived verbs that have MITTO as ancestor appears to be a by-product of its high token frequency: since MITTO is a common verb, it is also more easily available as a base for different word formation processes – hence the largeness of its derivational-inflectional family. To put it simply, in a sense in all the derived verbs of the derivational-inflectional family of MITTO what undergoes inflectional modifications is always the base MITTO, and therefore counting each verb in the family as a separate type could lead to an overestimation of the impact of the given alternation pattern on the whole lexicon. This is not an issue anymore in the modified procedure proposed in this chapter, where entropy is computed separately for simple lexemes (and derived lexemes that happen to be the only members of their derivational family) on the one hand, and for the various families of derived lexemes that have the same ancestor as the locus of inflection on the other hand.

As we have already pointed out in §6.1, what is reflected by the entropy values obtained for simple and derived lexemes is not exactly the same: it is only in simple lexemes that entropy estimates the uncertainty on which of the many alternation patterns available in the whole inflectional system will be applied, while for lexemes that belong to the same family it only gives us an idea of how likely it is that they will be inflected like their base. Now, it could be interesting to apply this modified procedure on a larger scale, for the purposes of typological generalizations on the inflectional complexity of languages in terms of uncertainty in the PCFP, as has been done in Beniamine 2018. The split between simple and derived lexemes would allow to separate the truly inflectional aspect, as reflected by simple lexemes,

from the (at least partly) different question of the inflectional behaviour of derived lexemes.

A similar line of reasoning can be applied to lexemes that belong to the same derivational-inflectional series, whose inflectional behaviour can be considered as being encoded in the suffixal derivational process by which such lexemes are formed: therefore, counting each derived lexeme as a separate type can lead to an overestimation of the weight of the alternation pattern involved in such lexemes.

However, it should be pointed out that our modified procedure also raises some additional problems, that could be addressed in future research.

Firstly, as was already hinted above in §6.3.2, the impact of derivational information on uncertainty in inflectional predictions is related to the size of the lexicon: the bigger the sample, the more derived lexemes will be in it, with obvious consequences on the overall uncertainty in inflectional predictions. Now, entropy can only provide a static picture of the uncertainty in the overall system, but it would be interesting to adopt an incremental approach to see how the situation evolves, testing the quality of actual inflectional predictions on novel verbs at different sample sizes.

Secondly, we have seen that the standard way of computing entropy is based on the simplified assumption that no information on derivational relatedness is available, but it should be noticed that also the opposite assumption that speakers have a completely reliable information on the derivational relatedness of all the lexemes of the lexicon is equally simplified: of course, what would be ideal is rather a dynamic approach allowing to take into account the formal and semantic factors that actually make two lexemes identifiable as derivationally related, although this would be much more difficult to operationalize.

6.5 Conclusion

In this chapter, we have investigated the impact of information on derivational relatedness on uncertainty in inflectional predictions. The classification into derivational-inflectional families (§6.2) has been shown to be able to decrease very strongly the entropy values estimating the difficulty of the PCFP regarding verb

paradigms; furthermore, it has been noticed that the reduction in uncertainty does not concern only derived verbs, as is expected, but also simple verbs, where the overall entropy value is smaller than the one obtained on the whole lexicon without derivational information. Conversely, regarding the classification into derivational-inflectional series operated for noun paradigms (§6.3), there is a small reduction in uncertainty only regarding complex nouns, while in the class of simple nouns there is an increase in entropy value. This difference has been explained on the basis of the fact that there is a correlation between token frequency, marginality of inflectional patterns and size of the derivational-inflectional family. This correlation can result in an overestimation of the weight of the uncertainty generated by rare alternation patterns that are attested in frequent verbs with a large family. On the other hand, no such correlation can be found for the suffixal derivational-inflectional series of nouns.

We have then discussed the theoretical and methodological consequences of these results, suggesting that information on derivational relatedness should not be neglected even when drawing typological generalizations on a larger scale, in order to obtain a more accurate picture of the different impact on uncertainty of simple and complex lexemes. Lastly, we have highlighted some problems of our modified procedure, envisaging a more principled methodology capable of dynamically taking into account the various formal and semantic aspects that make lexemes identifiable as derivationally related, and incrementally evaluating the reduction in uncertainty that can be obtained on lexicons of different size (§6.4).

Conclusion

In this work, we have proposed a novel analysis of Latin inflectional morphology, focusing on verb and noun paradigms. As we saw in Chapter 1, our approach is abstractive, rather than constructive: full inflected wordforms, rather than morphemes, are considered to be the basic morphological units of analysis; sub-word units are no more viewed as the atoms that are assembled to obtain wordforms, but only as (possibly) extracted *a posteriori* on the basis of the alternation patterns between wordforms themselves. Another notable characteristic of our work is that it focuses on implicative relations between wordforms, rather than on exponence (i.e., the relation between a wordform and the morphosyntactic property set it expresses). Lastly, our approach is also quantitative, in that the type frequency of inflectional patterns is considered to be an important factor when evaluating the complexity of the inflectional morphology of a language.

The theoretical framework of this work has been implemented by using an information-theoretic methodology, based on the notion of conditional entropy as a measure of uncertainty. The details of the procedure have been summarized in Chapter 2. In particular, we have quantified the uncertainty in predicting the content of a paradigm cell of a lexeme given knowledge of one (unary implicative entropy) or more than one (*n*-ary implicative entropy) inflected wordform: therefore, from a general point of view our results can be considered as an estimate of the difficulty of the Paradigm Cell Filling Problem.

To make a similar quantitative, entropy-based analysis possible, it is necessary to have a large, representative lexicon listing the inflected wordforms of verbs and nouns. For this purpose, we have exploited the database of Lemlat 3.0 to create LatInfLexi, an inflected lexicon of Latin comprising 3,348 verbs and 1,038 nouns (cf. Chapter 3). In itself, this constitutes a first contribution that the present work does to the field of Latin linguistics, providing a freely available lexical resource that can be exploited for other purposes too. In the future, on the one hand we aim at making the lexical coverage of LatInfLexi more systematic, including all the nouns of Delatte et al. (1981), rather than only the ones with token frequency of 30 or more as we do for now, and extending the lexicon to adjectives too. On the other

hand, we plan to include our data in the knowledge base of the LiLa (Linking Latin) project, whose purpose is to connect and make interoperable the wealth of language resources and NLP tools already available for Latin (cf. Passarotti et al. 2019). Given the design of our resource, its inclusion will enrich the knowledge base with many complete paradigms, listing also wordforms that are not attested in texts.

In Chapter 4 and Chapter 5, we have shown the new descriptive insights that emerge from the results of our quantitative, entropy-based analysis of verb and noun paradigms, respectively.

Our first step was using unary implicative entropy to investigate the structure of Latin paradigms in term of interpredictability. To do so, we have mapped the paradigm in different zones comprising cells that can be predicted from one another with no uncertainty, i.e. with null entropy. This yielded a very relevant simplification in verb paradigms, whose many cells can be reconducted to a handful of zones of full interpredictability (only 15). In nominal paradigms, conversely, no such simplification can be obtained, since none of the 10 cells constituting the cell paradigm of nouns can be predicted from another cell with no uncertainty. Therefore, it is interesting to observe that the difference in complexity between verb and noun paradigms in terms of interpredictability is a lot less relevant than it could appear from the sheer number of cells: while the cells of the verbal paradigm are much more numerous than the ones of the nominal paradigm (254 in the tabular paradigm and 152 in the cell paradigm for verbs, 12 and 10 respectively for nouns) the number of zones in the paradigm is much closer (15 for verbs, 10 for nouns).

The next step was moving to predictions from more than one cell, as measured by *n*-ary implicative entropy. Besides showing the constant reduction in uncertainty that can be obtained by assuming knowledge of an increasing number of cells, this allowed to recover the traditional notion of principal parts – a set of cells that predict the inflectional behaviour of a lexeme – on a more solid ground. Indeed, *n*-ary implicative entropy has been used as a principled way to find principal part sets, that are simply sets of cells from which the rest of the paradigm can be inferred with no uncertainty, i.e. with null entropy. In this way, we have found that the smallest principal part set available for verbs is made up of 4 cells, as the one of many traditional grammars and dictionaries. On the other hand, we have seen that for

nouns we need at least 3 principal parts to remove any uncertainty in predicting the whole paradigm for our dataset, while traditional descriptions always use 2 cells, NOM.SG and GEN.SG.

However, in a quantitative approach like the one we defend here, it is reasonable not to limit the investigation to categorical predictions: therefore, we have also extracted near principal parts, i.e. sets of cells from which the rest of the paradigm can be inferred with a very low (but not null) entropy value. Indeed, from NOM.SG and GEN.SG it is possible to predict all the other inflected wordforms of the nominal paradigm with very little uncertainty ($H < 0.01$), making their use as principal parts in traditional descriptions reasonable. As for verbs, it is possible to find near principal part sets smaller than the one that is traditionally used: if the threshold is set at 0.001, 3 near principal parts suffice, and even 2 cells are enough with the threshold at 0.01.

Starting from §5.3, we have introduced a methodological innovation: while in previous entropy-based analyses only the phonotactic shape of wordforms was assumed to be known, other properties of a lexeme can be useful in reducing uncertainty in the Paradigm Cell Filling Problem. For instance, the gender of a noun is at least partly informative on its inflectional behaviour, as is recognised in traditional descriptions. Indeed, the results obtained with our methodology shows that if also the gender of a nominal lexeme is assumed to be known, uncertainty is remarkably reduced. This result can be interpreted as highlighting an additional function – beside the ones already mentioned in previous research – of an apparently redundant feature like gender, that proves to be helpful in reducing the difficulty of the task of guessing unknown inflected wordforms of a lexeme.

Lastly, in Chapter 6 we have focused on the impact of information of a different kind, namely the derivational relatedness of lexemes of our sample. Because of the varying quantitative impact of different derivational processes in different lexical categories, different aspects were taken into account for verbs and nouns. In the former case, we have introduced a classification of lexemes on the basis of what we called derivational-inflectional families, grouping together verbs that have the same ancestor as the locus of inflection (cf. §6.2). In the latter case, we have used derivational-inflectional series, grouping together nouns formed by means of the

same suffix (cf. §6.3). Not only have our results shown that the impact of such information is very relevant (especially for families in verb inflection), but their interpretation also raises serious theoretical questions on what should be counted as a type when weighing the relevance of inflectional patterns: we have suggested that different lexemes that have the same ancestor as the locus of inflection on the one hand, or that are formed by means of the same suffix on the other hand, are probably better viewed as belonging to a same type in such counts.

To sum up, the contribution of the present work is threefold: firstly, it provides scholars working on Latin with a freely available lexical resource listing the inflected wordforms of a representative selection of nouns and verbs; secondly, it offers new descriptive insights on Latin verb and noun inflection, allowing for a mapping of paradigms in zones of interpredictability and for a recovery of the traditional notion of principal parts on a more solid ground; thirdly, it explores the possibilities opened up by the methodological innovation of taking into account also other pieces of information beside the phonotactic shape of wordforms when predicting paradigm cells.

Additionally, it should be noticed that the present work opens up interesting possibilities for future research: for instance, it paves the way for a diachronic perspective, comparing inflectional predictability in Latin and in the Romance languages, especially regarding verb morphology, on which there already is a lot of research investigating aspects that are closely related to the issue of interpredictability that is tackled in our work (cf. the literature cited in §1.3.2 above). On verb inflection, for a few Romance languages, an entropy-based analysis similar to the one provided in this work for Latin is already available, namely French (cf. Bonami & Boyé 2014) and Portuguese (cf. Bonami & Luís 2014). In any event, the results of the present work can be used as the starting point of a diachronic account of any Romance language, as well as in order to draw generalizations whose reliability can be considered to concern Romance in general, when such an entropy-based analysis will be available for other Romance languages, too.

References

Ackerman, Farrell & Blevins, James P. & Malouf, Robert. 2009. Parts and wholes: Implicative patterns in inflectional paradigms. In Blevins & Blevins (eds.). 54-82.

Ackerman, Farrell & Malouf, Robert. 2013. Morphological organization: The low conditional entropy conjecture. *Language* 89(3). 429-464.

Albright, Adam C. 2002. *The identification of bases in morphological paradigms*. PhD Thesis. University of California, Los Angeles.

Albright, Adam C. & Hayes, Bruce. 2003. Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition* 90(2). 119-161.

Allen, Sidney W. 1965. *Vox Latina: A guide to the pronunciation of classical Latin*. Cambridge: Cambridge University Press.

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

Baayen, R. Harald. 2014. Polysemy in derivation. In Lieber, Rochelle & Štekauer, Pavol (eds.), *The Oxford handbook of derivational morphology*. Oxford: Oxford University Press. 97-117.

Beniamine, Sacha. 2018. *Classifications flexionnelles. Étude quantitative des structures de paradigmes*. PhD Thesis. Université Sorbonne Paris Cité-Université Paris Diderot (Paris 7).

Beniamine, Sacha & Bonami, Olivier. 2018. *The segmentation problem in inflection.* (Paper presented at the workshop "New approaches to the typology of inflectional systems", Paris, 16 November 2018.)

Bennett, Charles Edwin. 1908. *New Latin Grammar*. Boston: Allyn and Bacon.

Berger, Adam L. & Della Pietra, Vincent J. & Della Pietra, Stephen A. 1996. A Maximum Entropy approach to Natural Language Processing. *Computational linguistics* 22(1). 39-71.

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

Blevins, James P. 2013. Word-based morphology from Aristotle to modern WP (word and paradigm models). In Allan, Keith (ed.), *The Oxford handbook of the history of linguistics*. Oxford: Oxford University Press. 375-395.

Blevins, James P. 2016. *Word and paradigm morphology*. Oxford: Oxford University Press.

Blevins, James P. & Blevins, Juliette. 2009. *Analogy in grammar: Form and acquisition*. Oxford: Oxford University Press.

Blevins, James P. & Milin, Petar & Ramscar, Michael. 2017. The Zipfian paradigm cell filling problem. In Kiefer, Ferenc & Blevins, James P. & Bartos, Huba (eds.), *Perspectives on Morphological Organization: Data and Analyses*: Leiden-Boston: Brill. 141-158.

Bochner, Harry. 1993. *Simplicity in Generative Morphology*. Berlin-New York: Mouton de Gruyter.

Bonami, Olivier. 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).

Bonami, Olivier & Beniamine, Sacha. 2016. Joint predictiveness in inflectional paradigms. *Word Structure* 9(2). 156-182

Bonami, Olivier & Boyé, Gilles. 2003. Supplétion et classes flexionnelles. *Langages* 37(152). 102-126.

Bonami, Olivier & Boyé, Gilles. 2014. De formes en thèmes. In Villoing, Florence & Leroy, Sarah & David, Sophie (eds.), *Foisonnements morphologiques. Études en hommage à Françoise Kerleroux*. Paris: Presses Universitaires de Paris-Ouest. 17-45.

Bonami, Olivier & Boyé, Gilles & Henri, Fabiola. 2011. Measuring inflectional complexity: French and Mauritian. (Paper presented at the "Workshop on Quantitative Measures in Morphology and Morphological Development", San Diego, 15-16 January 2011).

Bonami, Olivier & Caron, Gauthier & Plancq, Clément. 2014. Construction d'un lexique flexionnel phonétisé libre du français. In *Congrès Mondial de Linguistique Frannçaise – CMLF 2014 SHS Web of Conferences*. EDP Sciences. 2583-2596.

Bonami, Olivier & Luís, Ana R. 2014. Sur la morphologie implicative dans la conjugaison du portugais: une étude quantitative. *Mémoires de la Société de Linguistique de Paris* 22. 111-151.

Bonami, Olivier & Stump, Gregory T. 2016. Paradigm Function Morphology. In Hippisley & Stump (eds.). 449-481.

Boyé, Gilles & Cabredo Hofherr, Patricia, 2006. The structure of allomorphy in spanish verbal inflection. *Cuadernos de Lingüística* 13. 9-24

Boyé, Gilles & Schalchli, Gauvain. 2016. The status of paradigms. In Hippisley & Stump (eds.). 206-234.

Brown, Dunstan & Chumakina, Marina & Corbett, Greville G. (eds.). 2012. *Canonical morphology and syntax*. Oxford: Oxford University Press.

Brown, Dunstan & Hippisley, Andrew. 2012. *Network morphology: A defaults-based theory of word structure*. Cambridge: Cambridge University Press.

Brown, Peter F. & Della Pietra, Vincent J. & Mercer, Robert L. & Della Pietra, Stephen A. & Lai, Jennifer C. 1992. An estimate of an upper bound for the entropy of English. *Computational Linguistics*, 18(1). 31-40.

Brucale, Luisa. 2012. Latin compounds. *Probus* 24. 93-117.

Budassi, Marco & Passaroti, Marco. 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*. 90-94.

Bybee, Joan. 1995. Regular morphology and the lexicon. *Language and cognitive processes* 10(5). 425-455.

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

Carstairs, Andrew. 1984. Paradigm economy in the Latin third declension. *Transactions of the Philological Society*, 82(1). 117-137.

Carstairs, Andrew. 1987. *Allomorphy in inflexion*. London: Croom Helm.

Carstairs-McCarthy, Andrew. 1994. Inflection Classes, Gender, and the Principle of Contrast. *Language* 70(4). 737-788.

Cecchini, Flavio Massimiliano & Passarotti, Marco & Ruffolo, Paolo & Testori, Marinella & Draetta, Lia & Fieromonte, Martina & Liano, Annarita & Marini, Costanza & Piantanida, Giovanni. 2018. Enhancing the Latin Morphological Analyser LEMLAT with a Medieval Latin Glossary. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*.

Chan, Erwin. 2008. *Structures and distributions in morphology learning*. PhD Thesis. University of Pennsylvania, Philadelphia.

Charniak, Eugene. 2000. A maximum-entropy-inspired parser. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*. Stroudsburg: Association for Computational Linguistics. 132-139.

Chomsky, Noam & Halle, Morris. 1968. *The sound pattern of English*. New York: Harper & Row.

Corbett, Greville G. 1982. Gender in Russian: an account of gender specification and its relationship to declension. *Russian Linguistics* 6. 197-232.

Corbett, Greville G. 1991. *Gender*. Cambridge: Cambridge University Press.

Corbett, Greville G. 2005. The canonical approach in typology. In Frajzingier, Zygmunt & Hodges, Adam & Rood, David S. (eds.), *Linguistic Diversity and Language Theories*. Amsterdam: John Benjamins. 25-29.

Corbett, Greville G. 2009. Canonical inflectional classes. In Montermini, Fabio & Boyé, Gilles & Tseng, Jesse (eds.), *Selected proceedings of the 6th Décembrettes: Morphology in Bordeaux*. Somerville, MA: Cascadilla Proceedings Project. 1-11.

Corbett, Greville G. & Fraser, Norman M. 1993. Network Morphology: a DATR account of Russian nominal inflection. *Journal of linguistics* 29(1). 113-142.

Cover, Thomas & King, Roger. 1978. A convergent gambling estimate of the entropy of English. *IEEE Transactions on Information Theory* 24(4). 413-421.

Cser, András, 2015. The nature of phonological conditioning in Latin inflectional allomorphy. *Acta Linguistica Hungarica* 62(1). 1-35.

Cser, András. 2016. *Aspects of the phonology and morphology of Classical Latin*. PhD Thesis. Pázmány Péter Katolikus Egyetem.

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

Dressler, Wolfgang U. 2002. Latin inflection classes. In Bolkestein, A. Machtelt & Kroon, Caroline H.M. & Pinkster, Harm & Remmelink, H. Wim & Risselada, Rodie (eds.), *Theory and Description in Latin Linguistics: Selected Papers from the XIth International Colloquium on Latin Linguistics*. Amsterdam: Gieben. 91-110.

Du Cange, Charles Du Fresne & De Saint-Maur, Bénédictins de Saint-Maur & Carpentier, Pierre & Henschel, G. Louis & Favre, Léopold. 1883-1887. *Glossarium mediae et infimae latinitatis*. Niort.

Dye, Melody & Milin, Petar & Futrell, Richard & Ramscar, Michael (2017). A functional theory of gender paradigms. In Kiefer, Ferenc & Blevins, James P. & Bartos, Huba (eds.), *Perspectives on Morphological Organization. Data and Analyses*. Leiden: Brill. 212-239.

Ernout, Alfred. 1914. *Morphologie historique du latin*. Paris: Klincksieck.

Ernout, Alfred & Thomas, François. 1951. *Syntaxe latine*. Paris: Klincksieck.

Finkel, Raphael & Stump, Gregory T. 2007. Principal parts and morphological typology. *Morphology* 17(1). 39-75.

Finkel, Raphael & Stump, Gregory T. 2009a. Principal parts and degrees of paradigmatic transparency. In Blevins & Blevins (eds.). 13-53.

Finkel, Raphael & Stump, Gregory T. 2009b. What your teacher told you is true: Latin verbs have four principal parts. *Digital Humanities Quarterly* 3(1).

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

Fraser, Norman M. & Corbett, Greville G. 1995. Gender, Animacy, and Declensional Class Assignment: A Unified Account for Russian. In Booij,

Geert & Van Marle, Jaap (eds.), *Yearbook of Morphology 1994*. Dordrecht: Springer. 123-150.

Fruyt, Michèle. 2001. Réflexions sur la notion de 'mot' en latin: les verbes du type *calefacio*, in: Moussy, Claude (ed.), *De lingua latina novae quaestiones. Actes du Xè Colloque International de linguistique Latine: Paris-Sèvres, 19-23 avril 1999*. Paris-Louvain: Peeters. 81-94.

Georges, Karl Ernst & Georges, Heinrich. 1913-1918. *Ausführliches lateinisch-deutsches Handwörterbuch*. Hannover: Hahn.

Glare, Peter G. W. 2012. *Oxford Latin Dictionary*. Oxford: Oxford University Press.

Halle, Morris & Marantz, Alec. 1993. Distributed morphology and the pieces of inflection. In Hale, Kenneth & Keyser, Samuel J. (eds.), *The view from the building 20: Linguistic essays in honor of Sylvan Bromberger*. Cambridge: MIT Press. 111-176.

Harris, Zellig S. 1942. Morpheme alternants in linguistic analysis. *Language* 18. 169-180.

Hathout, Nabil. Une approche topologique de la construction des mots: propositions théoriques et application à la préfixation en anti. In Roché, Michel & Boyé, Gilles & Hathout, Nabil & Lignon, stéphanie & Plénat, Marc (eds.), *Des unités morphologiques au lexique*. Paris: Hermès/Lavoisier. 251-318.

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

Hippisley, Andrew & Stump, Gregory T. (eds.). 2016. *The Cambridge Handbook of Morphology*. Cambridge: Cambridge University Press.

Hockett, Charles F. 1954. Two models of grammatical description. *Word* 10. 210-213.

Janson, Tore. 1971. The Latin third declension. *Glotta* 49(1/2). 111-142.

Kaye, Steven. 2015. *Conjugation class from Latin to Romance: heteroclisis in diachrony and synchrony*. PhD Thesis. University of Oxford.

Leumann, Manu & Hofmann, Johann B. & Szantyr, Anton 1977. *Lateinische Grammatik*. München: CH Beck.

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

Lieber, Rochelle. 1992. *Deconstructing morphology: Word formation in syntactic theory*. Chicago: University of Chicago Press.

Litta, Eleonora & Passarotti, Marco & Culy, Chris. 2016. Formatio formosa est. Building a Word Formation Lexicon for Latin. In Corazza, Anna & Montemagni, Simonetta & Semeraro, Giovanni (eds.), *Proceedings of the Third Italian Conference on Computational Linguistics (CLiC-it 2016). 5-6 December 2016*. Napoli: Accademia University Press. 185-189.

Loporcaro, Michele. 2012. Stems, endings and inflectional classes in Logudorese verb morphology. *Lingue e linguaggio* 11(1). 5-34.

Loporcaro, Michele. 2018. *Gender from Latin to Romance: History, geography, typology*. Oxford: Oxford University Press

Maiden, Martin. 1992. Irregularity as a determinant of morphological change. *Journal of Linguistics* 28. 285-312.

Maiden, Martin. 1995. *A linguistic history of Italian*. Harlow: Longman.

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

Malouf, Robert. 2017. Abstractive morphological learning with a recurrent neural network. *Morphology* 27(4). 431-458.

Matthews, Peter H. 1972. *Inflectional morphology: A theoretical study based on aspects of Latin verb conjugation*. Cambridge: Cambridge University Press.

Matthews, Peter H. 1974. *Morphology*. Cambridge: Cambridge University Press.

Matthews, Peter H. 1991. *Morphology*. 2nd edn. Cambridge: Cambridge University Press.

McCullagh, Matthew. 2011. The Sounds of Latin: Phonology. In Clackson, James (ed.), *A companion to the Latin language*. Hoboken: Wiley-Blackwell. 83-91.

Milin, Petar & Kuperman, Victor & Kostić, Aleksandar & Baayen, R. Harald. 2009. Paradigms bit by bit: An information theoretic approach to the processing of paradigmatic structure in inflection and derivation. In Blevins & Blevins (eds.). 214-252.

Milizia, Paolo. 2013. *L'equilibrio nella codifica morfologica*. Roma: Carocci.

Montermini, Fabio & Bonami, Olivier. 2013. Stem spaces and predictability in verbal inflection. *Lingue e linguaggio* 12(2). 171-190.

Montermini, Fabio & Boyé, Gilles. 2012. Stem relations and inflection class assignment in Italian. *Word Structure* 5(1). 69-87.

Moscoso del Prado Martín, Fermín & Kostić, Aleksandar & Baayen, R. Harald. 2004. Putting the bits together: An information theoretical perspective on morphological processing. *Cognition* 94(1). 1-18.

Nyman, Martti. 1987. Is the paradigm economy principle relevant?. *Journal of Linguistics* 23(2). 251-267.

Oniga, Renato. 1990. L'apofonia nei composti e l'ipotesi dell'intensità iniziale in latino (con alcune conseguenze per la teoria dell'ictus metrico). In Danese, Roberto M., Gori, Franco & Questa, Cesare (eds.), *Metrica classica e linguistica. Atti del colloquio. Urbino 3-6 ottobre 1988*. Urbino: QuattroVenti. 195-236.

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

Passarotti, Marco & Cecchini, Flavio Massimiliano & Franzini, Greta & Litta, Eleonora & Mambrini, Francesco & Ruffolo, Paolo. 2019. The LiLa Knowledge Base of Linguistic Resources and NLP Tools for Latin. *2nd Conference on Language, Data and Knowledge. LDK 2019, May 20–23, 2019, Leipzig, Germany*.

Pellegrini, Matteo & Passarotti, Marco. 2018. LatInfLexi: An Inflected Lexicon of Latin Verbs. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*.

Pirrelli, Vito. 2000. *Paradigmi in morfologia: un approccio interdisciplinare alla flessione verbale dell'italiano*. Pisa-Roma: Istituti editoriali e poligrafici internazionali.

Pirrelli, Vito, and Marco Battista. 2000. The paradigmatic dimension of stem allomorphy in Italian verb inflection. *Italian Journal of Linguistics* 12(2). 307-380.

Ricca, Davide. 2017. Morfomi, allomorfie, partizioni: uno sguardo ai paradigmi verbali del torinese. In D'Alessandro, Roberta & Iannaccaro, Gabriele & Passino, Diana & Thornton, Anna M. (eds.), *Di tutti i colori. Studi linguistici per Maria Grossmann*. Utrecht: Utrecht University Repository. 257-282.

Risch, Ernst. 1977. Das System der lateinischen Deklinationen. *Cahiers Ferdinand de Saussure* 31. 229-245.

Robins, Robert H. 1959. In defence of WP. *Transactions of the Philological Society* 58. 116-144.

Shannon, Claude E. 1948. A mathematical theory of communication. *Bell system technical journal* 27(3). 379-423.

Shannon, Claude E. 1951. Prediction and entropy of printed English. *Bell system technical journal* 30(1). 50-64.

Sims, Andrea D. & Parker, Jeff. 2016. How inflection class systems work: On the informativity of implicative structure. *Word Structure* 9(2). 215-239.

Petrov, Slav & Das, Dipanjan & McDonald, Ryan. 2011. A universal part-of-speech tagset. *Proceedings of the Eigth International Conference on Language Resources and Evaluation (LREC 2012)*. 2089–2096

Skut, Wojciech & Brants, Thorsten. 1998. A maximum-entropy partial parser for unrestricted text. In Charniak, Eugene (ed.), *Proceedings of the Sixth Workshop on Very Large Corpora*.

Spencer, Andrew. 2012. Identifying stems. *Word Structure* 5(1). 88-108.

Steele, Susan. 1995. Towards a theory of morphological information. *Language* 71. 260-309.

Stump, Gregory T. 2001. *Inflectional morphology: A theory of paradigm structure*. Cambridge: Cambridge University Press.

Stump, Gregory T. 2006. Heteroclisis and paradigm linkage. *Language* 82(2). 279-322.

Stump, Gregory T. 2016. *Inflectional paradigms: Content and form at the morphology-syntax interface*. Cambridge: Cambridge University Press.

Stump, Gregory T. & Finkel, Raphael A. 2013. *Morphological typology: From word to paradigm*. Cambridge: Cambridge University Press.

Thornton, Anna M. 2001. Some reflections on gender and inflection class assignment in Italian. In Schaner-Wolles, Chris & Rennison, John R. & Neubarth, Friedrich (eds.), *Naturally! Linguistic studies in honour of Wolfgang Ulrich Dressler presented on the occasion of his 60th birthday*, Torino: Rosenberg & Sellier. 479-487.

Thornton, Anna M. 2011. Overabundance (multiple forms realizing the same cell): A non-canonical phenomenon in Italian verb morphology. In Maiden, Martin & Smith, John Charles & Goldbach, Maria & Hinzelin, Marc–Olivier (eds.), *Morphological autonomy: Perspectives from Romance inflectional morphology*. Oxford: Oxford University Press. 358-381.

Thornton, Anna M. 2019. Overabundance: A canonical typology. In Rainer, Franz & Gardani, Francesco & Dressler, Wolfgang U. & Luschützky, Hans Christian (eds.), *Competition in morphology*. Berlin: Springer. 223-258.

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

Walther, Géraldine & Sagot, Benoît. 2011 Modélisation et implémentation de phénomènes flexionnels non-canoniques. *Traitement Automatique des Langues* 52(2). 91-122.

Weiss, Michael L. 2009. *Outline of the historical and comparative grammar of Latin*. Ann Arbor-New York: Beech Stave Press.

Wurzel, Wolfgang Ullrich. 1984. *Flexionsmorphologie und Natürlichkeit: ein Beitrag zur morphologischen Theoriebildung*. Berlin: Akademie-Verlag.

Zanchetta, Eros & Baroni, Marco. 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*.

Zipf, George Kingsley. 1935. *The psycho-biology of language*. Oxford: Houghton-Mifflin.

Zubin, David & Köpke, Klaus Michael. 1986. Gender and folk taxonomy: the indexical relation between grammatical and lexical categorization. In Craig, Colette (ed.), *Noun Classes and Categorization. Proceedings of a Symposium on Categorization and Noun Classification. Eugene, Oregon, October 1983*. Amsterdam-Philadelphia: John Benjamins. 139-180.

Zwicky, Arnold M. 1985. How to describe inflection. In Niepokuj, Mary & Van Clay, Mary & Nikiforidou, Vassiliki & Feder, Deborah (eds.), *Proceedings of the Eleventh Annual Meeting of the Berkeley Linguistics Society*. Berkeley: Berkeley Linguistics Society. 372-386.