ISSN 2835-4138 (digital) Isogloss 2021, 7/9
https://doi.org/10.5565/rev/isogloss.109 1-35

The fine implicative structure of European Portuguese conjugation

Sacha Beniamine[1], Olivier Bonami[2], and Ana R. Luís[3]
[1]University of Surrey, SMG, s.beniamine@surrey.ac.uk
[2]Université de Paris, LLF, CNRS, olivier.bonami@u-paris.fr
[3]Universidade de Coimbra, CELGA-ILTEC, aluis@fl.uc.pt

Received: 26-01-2021
Accepted: 20-07-2021
Published: 03-09-2021

How to cite Beniamine, Sacha, Bonami, Olivier & Luís Ana R. 2021. The fine implicative structure of European Portuguese conjugation. Isogloss. Open Journal of Romance Linguistics 7, 9: 1-35.
DOI: https://doi.org/10.5565/rev/isogloss.109

Abstract

Recent literature has highlighted the extent to which inflectional paradigms are organised into systems of implications allowing speakers to make full use of the inflection system on the basis of exposure to only a few forms of each word. The present paper contributes to this line of research by investigating in detail the implicative structure of European Portuguese verbal paradigms. After outlining the computational methods we use to that effect, we deploy these methods on a lexicon of about 5000 verbs, and show how the morphological and phonological properties of European Portuguese verbs lead to the observed patterns of predictability.

Keywords: European Portuguese; Conjugation; Implicative structure; Quantitative morphology

1. Introduction

In the last fifteen years, the study of predictability relations in inflectional paradigms has become one of the central issues in theoretical morphology. Building on previous work on the distribution of stem allomorphs (Aronoff, 1994; Maiden, 1992; Stump, 2001, and many others) and on the predictive structure of affixal inflection classes (Carstairs-McCarthy, 1994; Wurzel, 1984, and many others), Ackerman, Blevins, and Malouf (2009) showed that inflectional systems are structured in terms of *implications* between the surface forms filling paradigms, and that the reliability of such implications can and should be assessed quantitatively.

	1sg	2sg	3sg	1pl	2pl	3pl
utilizar ‘use’	utilˈizu	utilˈizɐʃ	utilˈizɐ	utilizˈɐmuʃ	utilizˈajʃ	utilˈizɐ̃w
aprender ‘learn’	ɐpɾˈẽdu	ɐpɾˈẽdəʃ	ɐpɾˈẽdə	ɐpɾẽdˈemuʃ	ɐpɾẽdˈɐjʃ	ɐpɾˈẽdɐ̃j
imprimir ‘print’	ĩpɾˈimu	ĩpɾˈiməʃ	ĩpɾˈimə	ĩpɾimˈimuʃ	ĩpɾimˈiʃ	ĩpɾˈimɐ̃j

Table 1: The present indicative of three verbs, each representative of one of the three main inflectional classes

For example, the present indicative forms of regular verbs in the three European Portuguese conjugations, illustrated in Table 1,[1] displays complex implicative relations: the 2sg fully predicts the 3sg through a simple and general implication (the 3sg is identical to the 2sg minus the final /ʃ/); it also fully predicts the 3pl, although this time two more local generalizations must be observed (/ɐʃ/ corresponds to /ɐ̃w/ while /əʃ/ corresponds to /ɐ̃j/); it fails to fully predict the 1pl (the /əʃ/ ending is ambiguous between second and third conjugation) but still has some predictive value, as the opposition between /ɐʃ/ and /əʃ/ discriminates between the first and the two other conjugations. Crucially, the 1sg is an even worse predictor of the 1pl, completely failing to distinguish the three conjugations. Finally note that predictability is an asymmetric property: while the 2sg fails to fully predict the 1pl, it is fully predicted by it (/ɐmuʃ/ in the 1pl corresponds to /ɐʃ/ in the 2sg, /emuʃ/ to /əʃ/, /imuʃ/ to /əʃ/).

Ackerman, Blevins, and Malouf (ibid.) argue that such networks of implicative relations are key to solving the Paradigm Cell Filling Problem, that is, the problem of assessing what information speakers of a language may rely on to infer unknown forms of a lexeme from forms they have already encountered. In later work, Ackerman and Malouf (2013) argued that the study of implicative structure is of high theoretical importance: the implicative organization of paradigms reveal an aspect of morphological complexity that is orthogonal to “enumerative” indicators such as Greenberg’s (1954) indices of synthesis and agglutination: while languages vary widely in the amount of information that is conveyed by inflected words and how it is conveyed, the inflectional strategies are organized in such a fashion that the forms filling a paradigm are not too hard to predict from one another. In parallel, Stump and Finkel (2013) demonstrated the feasability of evaluating implicative structure on a large scale using computational methods applied to large inflected lexica. These publications led to an expanding literature discussing empirical and

[1] Here and throughout we use phonemic transcription rather than standard orthography, as Portuguese orthography is markedly misleading on the very issues that are the topic of this paper.

computational refinements to the assessment of implicative structure (see e.g. Bonami and Beniamine, 2016; Boyé and Schalchli, 2019; Cotterell et al., 2019; Sims, 2015) as well as empirical studies of the distribution and sources of implicative structure in various systems (see notably Bonami and Boyé 2014 on French; Bonami and Luís 2014 on European Portuguese; Mansfield 2016 on Murrinhpatha; Sims 2015 on Modern Greek; Guzmán Naranjo 2020 on Russian; Pellegrini 2021 on Latin; Wilmoth and Mansfield 2021 on Pitjantjatjara).

The present study has the double goal of contributing to both of these lines of research. From a methodological standpoint, we elaborate on Bonami and Boyé (2014) and later literature in assuming a strictly word-based approach (Blevins, 2006) to implicative structure, where the shape of an unknown word is inferred from the surface shape of a known word, without any prior information on how that word could be segmented into stems and exponents. To do so, we improve on all the studies cited above by relying on an algorithm for inferring alternations between surface forms that is not biased towards a particular morphological type, and hence is equally applicable to any language (Beniamine, 2018). From an empirical standpoint, we provide an in depth exploration of the morphological and morphophonological sources of unpredictability in European Portuguese verbal paradigms: using computationally identified predictability values as our guideposts, we identify which properties of the system lead to such values. In this area we improve dramatically on Bonami and Luís (2014) by relying on a much larger lexicon and properly taking into account stress-conditioned vowel alternations.

The structure of the paper is as follows. In section 2, we briefly outline the structure of European Portuguese verbal paradigms, and present the dataset used in this paper. Section 3 guides the reader through the particular method for assessing predictability relations used in this paper: we show how pairs of forms can be classified into alternation types, how these alternation types can be used to assess the probability of the form in a predicted cell from the form in a predictor cell, and how conditional entropy can then be used as a measure of average predictability. Section 4 applies this method to our dataset, and identifies two main sources and a few ancillary factors leading to low predictability in European Portuguese conjugation. Section 5 concludes the paper.

2. The dataset

2.1 *The verbal morphology of European Portuguese*

Table 2 presents the general form of a verbal paradigm of European Portuguese. There are 5 tenses within the indicative mood (present, imperfect, simple past, pluperfect, future), and 3 tenses within the subjunctive mood (past, present, future), a conditional, and an imperative with only second person forms.[2] As to the nonfinite forms, there is the ordinary infinitive, in addition to a personal infinitive which agrees in person with the subject and whose 3SG form is syncretic with the ordinary infinitive;[3] there are also the gerund and past participle forms, the latter further inflecting in gender and number. Some transitive verbs distinguish two forms of past participle, e.g. ENCARREGAR 'put in charge' has participles

[2] We leave aside the 3SG, 1PL and 3PL forms that are sometimes listed in the imperative but are really present subjunctive forms that can be used in contexts similar to that of the imperative.

[3] The personal infinitive is most often syncretic with the future subjunctive, but some verbs distinguish them by stem alternation, for example, *ir-es* 'GO.INF-2SG' vs. *for-es* 'GO.FUT.SBJV-2SG'.

encarregue and *encarregado*, and tradition holds that the two forms are used respectively for the formation of perfect and passive periphrases. It is however unclear whether this is a situation of overdifferentiation (Brown, 2007; Corbett, 2007), where the normal contexts of use of the participle are split and the two forms are in complementary distribution, rather than overabundance (Thornton, 2012), with an overlap in the distribution of the two forms. We leave this issue to future research.

	1SG	2SG	3SG	1PL	2PL	3PL
PRS.IND	fˈalu	fˈalɐʃ	fˈalɐ	fɐlˈɐmuʃ	fɐlˈajʃ	fˈalɐ̃w
PST.IMPF.IND	fɐlˈavɐ	fɐlˈavɐʃ	fɐlˈavɐ	fɐlˈavɐmuʃ	fɐlˈavɐjʃ	fɐlˈavɐ̃w
PST.PFV.IND	fɐlˈɐj	fɐlˈaʃtə	fɐlˈo	fɐlˈamuʃ	fɐlˈaʃtəʃ	fɐlˈaɾɐ̃w
PST.PERF.IND	fɐlˈaɾɐ	fɐlˈaɾɐʃ	fɐlˈaɾɐ	fɐlˈaɾɐmuʃ	fɐlˈaɾɐjʃ	fɐlˈaɾɐ̃w
FUT.IND	fɐlɐɾˈɐj	fɐlɐɾˈaʃ	fɐlɐɾˈa	fɐlɐɾˈemuʃ	fɐlɐɾˈɐjʃ	fɐlɐɾˈɐ̃w
COND	fɐlɐɾˈiɐ	fɐlɐɾˈiɐʃ	fɐlɐɾˈiɐ	fɐlɐɾˈiɐmuʃ	fɐlɐɾˈiɐjʃ	fɐlɐɾˈiɐ̃w
PRS.SBJV	fˈalə	fˈaləʃ	fˈalə	fɐlˈemuʃ	fɐlˈɐjʃ	fˈalɐ̃j
PST.SBJV	fɐlˈasə	fɐlˈasəʃ	fɐlˈasə	fɐlˈasəmuʃ	fɐlˈasɐjʃ	fɐlˈasɐ̃j
FUT.SBJV	fɐlˈaɾ	fɐlˈaɾəʃ	fɐlˈaɾ	fɐlˈaɾmuʃ	fɐlˈaɾdəʃ	fɐlˈaɾɐ̃j
IMP		fˈalɐ			fɐlˈaj	
PER.INF	fɐlˈaɾ	fɐlˈaɾəʃ	fɐlˈaɾ	fɐlˈaɾmuʃ	fɐlˈaɾdəʃ	fɐlˈaɾɐ̃j
			PST.PTCP			
	INF	GER	M.SG	M.PL	F.SG	F.PL
	fɐlˈaɾ	fɐlˈɐ̃du	fɐlˈadu	fɐlˈadɐ	fɐlˈaduʃ	fɐlˈadɐʃ

Table 2: The verbal paradigm of FALAR ‘speak’

Traditionally, European Portuguese verbs are grouped into three classes distinguished by the theme vowels -a, -e or -i (visible in the infinitive forms). Examples of this were already shown in Table 1 with the forms of the present indicative of three representative verbs[4], where four distinct person markers can be identified, independent of class: 1SG /u/, 2SG /ʃ/, 1PL /muʃ/, 2PL /ʃ/.[5] The theme vowel is present in its distinctive form in the 1PL and 2PL, but can manifest itself in a reduced form depending on the case (/ɐ/ vs. /ai/, /e/ vs. /ɐi/). In the 2SG and 3SG, the distinction between the theme vowels of the second and third

[5] For the purposes of this section we assume the following segmentation for 2PL present forms:

(i) First conjugation: X+ai+ʃ

(ii) Second conjugation: X+ɐi+ʃ

(iii) Third conjugation: X+i+ʃ

This segmentation, which follows Boyé's (2000) segmentation guidelines for Romance verbs, makes it possible to establish a consistent marker for 2PL forms, in the form of /ʃ/, including the few irregular verbs that have a 2PL form in /dəʃ/ (such as VIR in Table 1). It is worth noting that this segmentation is not to be taken as the only possible one: a plausible alternative would be to posit the contrasting theme vowels /a/ vs. /ɐ/ vs. /i/, followed by a marker /iʃ/ with coalescence of the two /i/ in the 3rd conjugation. Such alternative segmentation would treat the 2PL form of verbs as a suppletive inflected form (Boyé and Cabredo Hofherr, 2006a). However, deciding on the segmentation is of little consequence for this analysis: one of the virtues of having a word-based (rather than a stem-based) approach here is precisely that it does not require the choice of uniform segmentation. See Section 4 below on this point.

conjugations is neutralised. In the 3PL, the theme vowel merges[6] with the person marker, giving us the realisation /ɐ̃ũ/ for the first conjugation and /ɐ̃ĩ/ for the other two. Finally, 1SG does not realise the theme vowel at all.

The situation found in the present tense is characteristic of the whole paradigm, where there are other cases of alternations in vowel quality of the theme vowel, partial neutralization between classes 2 and 3, and fusional morphology. Overall, it is clear that the system of theme vowels contributes to increasing predictability of patterning, while their neutralisation (or even their absence in the PRS.IND.1SG) is a source of opacity: if a speaker has only been exposed to the PRS.IND.1SG form of a verb, they cannot reliably infer to which class this verb belongs. On the other hand, if they are exposed to a 2SG in /əʃ/, they can deduce unequivocally that the verb is not of the first conjugation (although uncertainty remains as to whether it belongs to the second or third).

The three major inflectional classes do not exhaustively model all conjugation patterns found in European Portuguese. Table 3 gives some additional examples of verbs in the present indicative that highlight a number of relevant phenomena: vowel alternations between stressed and unstressed vowels (SECAR, REZAR), stressed vowel alternations (DIVERTIR, DORMIR), allomorphic or stem alternations (TRAZER, OUVIR, VIR), and irregular inflection (VIR, ESTAR). Once again, the present indicative is representative of the rest of the paradigm, which exhibits other examples of phenomena of the same type.

	1SG	2SG	3SG	1PL	2PL	3PL
SECAR ‘dry’	sˈɛku	sˈɛkɐʃ	sˈɛkɐ	səkˈɐmuʃ	səkˈajʃ	sˈɛkɐ̃w
REZAR ‘pray’	rˈɛzu	rˈɛzɐʃ	rˈɛzɐ	rəzˈɐmuʃ	rəzˈajʃ	rˈɛzɐ̃w
DIVERTIR ‘amuse’	divˈiɾtu	divˈɛɾtəʃ	divˈɛɾtə	divəɾtˈimuʃ	divəɾtˈiʃ	divˈɛɾtɐ̃j
DORMIR ‘sleep’	dˈuɾmu	dˈɔɾməʃ	dˈɔɾmə	duɾmˈimuʃ	duɾmˈiʃ	dˈɔɾmɐ̃j
TRAZER ‘bring’	tɾˈagu	tɾˈazəʃ	tɾˈaʃ	tɾɐzˈemuʃ	tɾɐzˈɐjʃ	tɾˈazɐ̃j
OUVIR ‘listen’	ˈosu	ˈovəʃ	ˈovə	ovˈimuʃ	ovˈiʃ	ˈovɐ̃j
VIR ‘come’	vˈɐɲu	vˈɐ̃jʃ	vˈɐ̃j	vˈimuʃ	vˈĩdəʃ	vˈɐ̃jɐ̃j
ESTAR ‘be’	əʃtˈo	əʃtˈaʃ	əʃtˈa	əʃtˈɐmuʃ	əʃtˈajʃ	əʃtˈɐ̃w

Table 3: Present indicative of verbs with deviant patterns

2.2 *The lexicon employed*

Many studies of inflectional systems rely on a first classification of data in line with the pedagogical tradition, and often only take into consideration one well-behaved paradigm for each identified inflectional pattern (see for example Bonami and Boyé 2003, Ackerman and Malouf 2013, Stump and Finkel 2013). This approach, while having the advantage of employing easily accessible data, has three major disadvantages for the study of implicative relations between forms. First, it is common for the pedagogical tradition to gloss over some subtle contrasts in the data that can have important ramifications. Second, the

[6] Mateus and d’Andrade (2000, pp. 74–75) analyse the 3PL form as the combination of a theme vowel with a nasal autosegment. This classical analysis assumes that there is a theme vowel adjacent to the stem. Within the word-based approach adopted in this paper, however, we assume that the distinctive part of the ending is the last segment and not the first.

analysis of only exemplary lexemes alone does not allow one to build up an image of the frequency distribution of the phenomena in question, which can lead to inadequate estimations of what should be considered regular. Third and finally, it prevents any detailed study of the predictive value of phonotactic properties of lexemes that exemplify an inflectional pattern (Albright and Hayes, 2002; Guzman Naranjo, 2019).

The present work is therefore based on an extensive lexicon,[7] consisting of the full paradigm of about 5000 lexemes. We first used frequency lists provided by the AC/DC project[8] to select the 5000 most frequent verb lexemes in the CETEMPúblico corpus (Santos and Rocha, 2001). Fernando Perdigão then kindly provided us with paradigms for these verbs in phonemic transcriptions. These were obtained using pronunciation dictionaries and text to speech tools developed at the University of Coimbra (Candeias, Veiga, and Perdigão, 2015; Marquiafável et al., 2014) and corrected by hand. In the process, a handful of verbs had to be excluded.

The transcriptions used are surface-oriented and standardised,[9] as there is of course variability in the pronunciation of European Portuguese words, conditioned on variety, register and speech rate, which affects in particular the realisation of unstressed vowels. The transcription used here corresponds to a possible realisation in a formal context with a relatively slow speech rate, which minimises the instances of fusion or coarticulation of vocalic sounds.

An important feature of the lexicon we used is that only one form is given per paradigm cell of each lexeme: overabundance (Thornton, 2012) is not taken into account. Notably, the lexicon records a single form of the past participle, in this case the most ”regular” according to traditional description, and does not take into account variation in gender and number. The final lexicon consists of 4991 paradigms with 65 cells each, leading to a total of 324,415 inflected forms.

3. Methods

3.1 *Identifying patterns of alternation*

We now turn to the question of studying the implicative relations between paradigm cells. For paradigms with 65 word forms, there are 4160 pairs of cells to examine.[10] It is of course out of the question to work through them manually (let alone through a 4991 verb lexicon), and it is therefore necessary to automate the process.

The starting point of the process is to infer patterns of alternation from a set of pairs of

[7] The lexicon can be consulted or downloaded at `https://sbeniamine.gitlab.io/europeanportugueseverbs` and is permanently recorded on Zenodo under the DOI `10.5281/zenodo.5121543`.

[8] `https://www.linguateca.pt/acesso/contabilizacao.php\#listaPosCETEMPUBLICO`, accessed on March 20, 2021

[9] The transcriptions adopted are similar to those of Mateus and d’Andrade (2000), with four differences: (i) semi-vowels are not distinguished from high vowels; (ii) the non-low central vowel is transcribed [ə] rather than [ɨ]; (iii) stress is marked by using the IPA symbol placed immediately before the stressed vowel ; (iv) diphthongs are written using the glides [j] and [w] after the initial vowel.

[10] The study of implicative relations involves looking at each of the cases where one of the 65 cells predicts one of the 64 other cells, and $65 \times 64 = 4160$. Since patterns of alternation are bidirectional (the same pattern relating cell c to cell c' also relates cell c' to cell c), we can divide the size of the problem by two when looking for patterns, but not when examining asymmetric predictability relations based on these patterns.

forms. The algorithm we use to describe these relations is implemented as part of the free software toolkit Quantitative Modelling of Inflection (Qumín)[11] and described in detail in Beniamine (2018). This algorithm was inspired by Bonami and Boyé (2014) and Albright and Hayes (2002, 2006). A concrete example is presented in Table 4, which shows the automatically detected alternations for the 20 most frequent lexemes of the corpus for the bidirectional relation between the 1SG present indicative and the 1SG present subjunctive. Note that the identified patterns are often (but not always) compatible with a segmental analysis that splits up stem and suffix. For verbs showing stem allomorphy (e.g. SABER 'know'), the difference between the two stems is encoded in the pattern of alternation: this situation is not problematic, since the purpose of this procedure is not to determine morpheme boundaries but to identify sets of lexemes that inflect in the same way.

Lexeme	PRS.IND.1SG	PRS.SBJV.1SG	Alternation
SER 'be'	sˈo	sˈɐjʒɐ	_ˈo ⇌ _ˈɐjʒɐ
TER 'have'	tˈɐɲu	tˈɐɲɐ	_u ⇌ _ɐ
ESTAR 'be'	əʃtˈo	əʃtˈɐjʒɐ	_ˈo ⇌ _ˈɐjʒɐ
FAZER 'do'	fˈasu	fˈasɐ	_u ⇌ _ɐ
PODER 'be able to'	pˈɔsu	pˈɔsɐ	_u ⇌ _ɐ
IR 'go'	vˈo	vˈa	_ˈo ⇌ _ˈa
DIZER 'say'	dˈigu	dˈigɐ	_u ⇌ _ɐ
HAVER 'there to be'	ˈɐj	ˈaʒɐ	_ˈɐj ⇌ _ˈaʒɐ
DEVER 'must'	dˈevu	dˈevɐ	_u ⇌ _ɐ
DAR 'give'	dˈo	dˈe	_ˈo ⇌ _ˈe
VER 'see'	vˈɐjʒu	vˈɐjʒɐ	_u ⇌ _ɐ
PASSAR 'pass'	pˈasu	pˈasə	_u ⇌ _ə
FICAR 'stay'	fˈiku	fˈikə	_u ⇌ _ə
VIR 'come'	vˈɐɲu	vˈɐɲɐ	_u ⇌ _ɐ
QUERER 'want'	kˈɛɾu	kˈɐjɾɐ	_ˈɛ_u ⇌ _ˈɐj_ɐ
SABER 'know'	sˈɐj	sˈajbɐ	_ˈɐj ⇌ _ˈajbɐ
CHEGAR 'arrive'	ʃˈegu	ʃˈegə	_u ⇌ _ə
AFIRMAR 'assert'	ɐfˈiɾmu	ɐfˈiɾmə	_u ⇌ _ə
ENCONTRAR 'meet'	ẽkˈõtɾu	ẽkˈõtɾə	_u ⇌ _ə
CONSIDERAR 'consider'	kõsidˈɛɾu	kõsidˈɛɾə	_u ⇌ _ə

Table 4: Basic alternations relating the PRS.IND.1SG to the PRS.SBJV.1SG for the 20 most frequent verbs in the dataset

The algorithm starts from a lexicon tabulating full paradigms, a table providing decompositions of each phoneme into minimal features. It then calculates alternation patterns which describe bidirectional implicative relations between all pairs of cells, using the format shown in Figure 1, which reads: /u/ alternates with /ɐ/ at the end of the word, in a context consisting of an unbounded sequence of segments followed by a non-nasal, a non-nasal sonorant, and a non-lateral.[12]

[11] Qumín is a python toolkit distributed under GPLv.3., and can be accessed at `https://github.com/XachaB/Qumin`.

[12] The details of the syntax of patterns are as follows. A pattern consists of a description of the *alternation* between the two forms combined with a description of the *context* in which this alternation takes place. Both

$$\underbrace{\text{\_u} \rightleftharpoons \text{\_ɐ}}_{\substack{\textbf{alternation:}\\ \text{/u/ becomes /ɐ/}\\ \text{and reciprocally}}} \;/\; \underbrace{\text{X*[-nas][+son -nas][-lat]\_}}_{\substack{\textbf{context:}\\ \text{suffixal change, with phonotactic restriction}\\ \text{on preceding segments}}}$$

Figure 1: Structure of the alternation pattern instantiated by lexemes such as *ter* in Table 5

Notice that Qumín adapts notation from classical generative phonological rules (Albright and Hayes, 2002; Chomsky and Halle, 1968); but, unlike classical rules, alternation patterns are bidirectional and can describe changes in more than one location in the word. The algorithm is agnostic as to the language or the type of morphological alternations it displays.

There are always many possible ways of describing how to transform one sequence of phonemes into another, which makes the task of inferring patterns difficult. To create good descriptions, the Qumín algorithm follows two criteria: patterns should be formally simple and they should be as general as possible. The algorithm attempts to satisfy them through three steps, as shown below, repeated for all pairs of cells (α, β) in the paradigms:

(1) a. For each lexeme, formulate a set of formally simple hypotheses to describe the differences between the pair of forms in cells (α, β)

b. Calculate phonotactic contexts for these hypotheses using a variant of minimal generalisation (Albright and Hayes, 2002).

c. Choose the best hypotheses according to their generality.

Table 5 illustrates the algorithm for the verbs from Table 4. Step one consists in generating one hypothetical alternation patterns for each pair of forms (see the column labeled "Fine-grained pattern" in Table 4). This is done by aligning the phonological segments of each pair of forms in order to identify variable and stable material. Optimal alignments are chosen using a phonologically weighted scoring scheme[13] and the Needleman and Wunsch (1970) dynamic algorithm (also known as the Wagner–Fischer algorithm). As shown in Table 6 for the lexeme ENTRAR 'enter' between the PST.PFV.IND.1SG and the PER.INF.3PL, often there can be more than one optimal alignment for a given pair of forms. Since they all constitute possible hypotheses, they are all retrieved. In the case of the alternations from Table 4, none of the alignments are ambiguous, and the 20 pairs of forms lead to 20 fine-grained patterns in Table 5.

the alternation and the context contain placeholders in the form of underscore characters ("_") indicating how the context combines with the alternation: in the present instance, the alternation happens at the end of the context. Descriptions can make use of single characters representing phonological segments (u and ɐ in Figure 1), feature matrices representing natural classes of segments (e.g. [+son –lat]), sets of segments (e.g. {s, t}), variables over segments (e.g. X), and the Kleene star ("X*") and Kleene plus operator ("X+"), to represent unbounded sequence of segments.

[13] The cost of a substitution between two phonemes a and b is set to $1 - -\text{sim}(a, b)$, where 'sim' is the similarity between two phonemes, calculated using the Jaccard similarity over their sets of natural classes, following Frisch, Pierrehumbert, and Broe (2004): for any two phonemes a and b, $\text{sim}(a, b) = \dfrac{\text{C}(a) \bigcap \text{C}(b)}{\text{C}(a) \bigcup \text{C}(b)}$, where $\text{C}(x)$ is the set of natural classes to which x belongs. The cost of an insertion (or deletion) is a constant set to 0.4 times the median of all similarity costs among our inventory of phonemes.

Lexeme	Fine-grained pattern	Generalized pattern	Count
SER	ˈo ⇌ ˈɐjʒɐ / s_	ˈo ⇌ ˈɐjʒɐ / X*{s,t}_	2
ESTAR	ˈo ⇌ ˈɐjʒɐ / əʃt_		
TER	u ⇌ ɐ / tˈɐɲ_	u ⇌ ɐ / [-son -dors][-nas +stress][-approx]_	7
VER	u ⇌ ɐ / vˈɐjʒ_		
VIR	u ⇌ ɐ / vˈɐɲ_		
FAZER	u ⇌ ɐ / fˈas_		
PODER	u ⇌ ɐ / pˈɔs_		
DIZER	u ⇌ ɐ / dˈig_		
DEVER	u ⇌ ɐ / dˈev_		
IR	ˈo ⇌ ˈa / v_	ˈo ⇌ ˈa / v_	1
HAVER	ˈɐj ⇌ ˈaʒɐ / _	ˈɐj ⇌ ˈaʒɐ / _	1
DAR	ˈo ⇌ ˈe / d_	ˈo ⇌ ˈe / d_	1
PASSAR	u ⇌ ə / pˈas_	u ⇌ ə / X+[+cons -lab -dent -lat]_	6
FICAR	u ⇌ ə / fˈik_		
CHEGAR	u ⇌ ə / ʃˈeg_		
AFIRMAR	u ⇌ ə / ɐfˈiɾm_		
ENCONTRAR	u ⇌ ə / ẽkˈõtɾ_		
CONSIDERAR	u ⇌ ə / kõsidˈɛɾ_		
QUERER	ˈɛ_u ⇌ ˈɐj_ɐ / k_ɾ_	ˈɛ_u ⇌ ˈɐj_ɐ / k_ɾ_	1
SABER	ˈɐj ⇌ ˈajbɐ / s_	ˈɐj ⇌ ˈajbɐ / s_	1

Table 5: Inferring patterns in three steps for the cells the PRS.IND.1SG and PRS.SBJV.1SG of the 20 most frequent lexemes.

PST.PFV.IND.1SG	ẽ	t	–	–	ɾ	ˈɐj
PER.INF.3PL	ẽ	t	ɾ	ˈa	ɾ	ɐ̃j

PST.PFV.IND.1SG	ẽ	t	ɾ	–	–	ˈɐj
PER.INF.3PL	ẽ	t	ɾ	ˈa	ɾ	ɐ̃j

Table 6: Two alignments of the PST.PFV.IND.1SG and PER.INF.3PL forms of ENTRAR ‘enter’ with identical cost.

The goal of the second step (column “Generalised pattern” in Table 5) is to merge contexts across lexemes to capture phonotactic constraints on the distribution of patterns. These are obtained using the Minimal Generalisation approach defined by Albright and Hayes (2002). First, all sets of patterns which share the same structural alternation are grouped together. Second, the most specific description which applies to all of these patterns is inferred, segment by segment. This differs from Albright & Hayes in two main ways: all patterns are merged at once, rather than merging them two by two. Second, no intermediate results are memorised. This is crucial to keeping generalization computationally tractable and hence applicable at a large scale.

As Table 5 shows, in the case of 1st conjugation regular verbs such as passar, we find a rather general pattern: the pattern constrains the forms to end with a consonant with a limited choice of place and manner of articulation, preceded by any nonempty sequence (noted ‘X+’). For verbs of the 2nd and 3rd conjugations such as ter (the two conjugations do not differ with regards to the two forms of interest here), we can identify phonotactic restrictions based on the 7 examples present: all end with a sequence consisting of a non-dorsal obstruent, a stressed oral vowel, and an non-approximant. However, if we look at more data, neither of these restrictions hold anymore.

Citation form	Sample lexeme prs.ind.1sg	 prs.sbjv.1sg	Generalized pattern	Count
passar ‘pass’	pˈasu	pˈasə	u ⇌ ə / X+_	4166
ter ‘have’	tˈɐɲu	tˈɐɲɐ	u ⇌ ɐ / X+_	816
ser ‘be’	sˈo	sˈɐjʒɐ	ˈo ⇌ ˈɐjʒɐ / X*{s,t}_	2
querer ‘want’	kˈɛɾu	kˈɐjɾɐ	ˈɛ_u ⇌ ˈɐj_ɐ / [+son]*k_ɾ_	2
saber ‘know’	sˈɐj	sˈajbɐ	ˈɐj ⇌ ˈajbɐ / [+nas]*s_	2
ir ‘go’	vˈo	vˈa	ˈo ⇌ ˈa / v_	1
haver ‘there to be’	ˈɐj	ˈaʒɐ	ˈɐj ⇌ ˈaʒɐ / _	1
dar ‘give’	dˈo	dˈe	ˈo ⇌ ˈe / d_	1

Table 7: Patterns of alternation between the prs.ind.1sg and prs.sbjv.1sg, with minimal generalisation over contextual phonotactic properties.

Table 7 reports the list of patterns computed over the whole lexicon. We can see that neither of the [u ⇌ ə] and [u ⇌ ɐ] alternations are associated anymore with phonotactic conditions. In addition, an interesting generalisation emerges concerning the prs.ind.1sg of verbs in /o/ : the two verbs in question (ser ‘be’ and estar ‘be’) have stems ending in a voiceless anterior coronal obstruent. Finally, Table 5 and 7 both document some very specific patterns applicable to a single verb (e.g. ir) or a very small class of closely related verbs (e.g. querer and its derivative malquerer). Such patterns with very low type frequency for high-frequency verbs are the hallmark of irregular inflection in the traditional sense.

During this step, the program also attempts to generalize over alternations by recognising simple phonological functions. For illustration of this process, we turn to the alternation between the gerund and the indicative future 3sg. As exemplified in Table 8, this pair of cells leads to parallel alternations: [ˈĩ ⇌ i] for 397 lexemes, [ˈɐ̃ ⇌ ɐ] for 168 others. There is a clear generalisation here, that stressed nasal vowels alternate with their unstressed oral counterparts. This is captured by the generalised alternation shown in the table.

Lexeme	Surface alternation	Generalised alternation
PARTIR ‘break’	ˈĩdu ⇌ iɾˈa	[-nas -stress]ɾˈa ⇌ [+nas +stress]du / X*_
PASSAR ‘pass’	ˈɐ̃du ⇌ ɐɾˈa	

Table 8: Generalisation for phonological alternations (GER ~ FUT.IND.3SG)

Step three chooses the most general among the competing hypotheses. The generality of a pattern is assessed as its (type) frequency, that is, the number of lexemes to which it can be applied. Then for each lexeme, the most general applicable pattern is kept. In the present example, there are no competing patterns, so this step is trivial.[14]

This procedure allows Qumín to find alternation patterns which are both simple and general. While the results do not reproduce usual morphemic segmentations, they are obtained by systematically applying the same principles over the entire lexicon. They are easy to examine, analyze and evaluate, and can be readily obtained from large lexicons.

3.2 *Measuring predictability*

Regardless of the exact method used to classify patterns of alternation between forms, one form can be informative in predicting the other. Let us consider again the data in Table 7. There are clearly several generalisations. First, the three patterns compatible with an PRS.IND.1SG ending in stressed /o/ place complementary requirements on the phonotactic context: one requires the last consonant to be /s/ or /t/, the second /v/, the third /d/. Faced with a form in /o/, one can therefore categorically predict what the PRS.SBJV.1SG will be. Second, if a verb ends in /u/, but doesn’t end in either /ˈɐjʒu/ nor in /kˈɛɾu/ in the PRS.IND.1SG, two endings are possible at PRS.SBJV.1SG: /ə/ or /ɐ/. As a consequence, we cannot categorically predict the ending of the subjunctive on the basis of the indicative. However, the two patterns contrast in terms of frequency, which gives us an indication of the probability of each alternative: given an unknown verb ending in /u/, but not in /ˈɐjʒu/ or /kˈɛɾu/, in the PRS.IND.1SG, the probability that its PRS.SBJV.1SG is in /ə/ is about 4166/(4166+816) ≈ 84%.

We can therefore see that knowledge of patterns of alternation and their distribution provides crucial information for prediting one form from another. However, there is relevant information that is not explicit in Table 7. Let us take the case of verbs that end in /ˈɛɾu/ in the PRS.IND.1SG. In principle, three patterns could be satisfied by such verbs: the first or second pattern, which only requires that the form end in /u/; or the fourth pattern, which is compatible only with those forms ending more specifically in /ˈkɛɾu/. However, to assess how much uncertainty there is, we need to check what proportion of the verbs in the lexicon instantiate each of these possibilites.[15]

[14] Though it is not the case in this example, generalised patterns can apply successfully to lexemes which did not lead to their discovery. This makes the algorithm more robust, as it can discover very good patterns even for pairs of forms for which the alignment step went down a wrong path.

[15] As it turns out, there are 34 verbs in /ˈɛɾu/ instantiating the first pattern (e.g. GERAR ‘generate’, PRS.IND.1SG /ʒˈɛɾu/, PRS.SBJV.1SG /ʒˈɛɾə/), and two verbs instantiating que fourth (QUERER and MALQUERER), but no verb in /ˈɛɾu/ instantiating the second.

To obtain this information, lexemes must now be classified not according to the pattern of alternation they display, but according to the patterns which *could* be satisfied given the phonological characteristics of their PRS.IND.1SG form. The lexicon must be examined once more, to identify all verbs compatible with the fourth pattern – i.e. all verbs with the PRS.IND.1SG form ending in /kˈɛɾu/. The result of this second search is that, while a total of 70 verbs end in /ˈɛɾu/, only 2 of those end in /kˈɛɾu/, namely the two verbs QUERER 'want' and MALQUERER 'wish ill'; hence there is in fact no uncertainty as to the subjunctive of a verb whose PRS.IND.1SG is in /kˈɛɾu/.

Table 9 outlines the details of this search taking all lexemes in our dataset into account. The lexemes are categorized into classes based on the phonological shape of the PRS.IND.1SG form, that is, on the basis of the inventory of patterns that they could theoretically participate in.[16] The number of verbs in each class is recorded, as well as the number of lexemes that instantiate each of the possible patterns. In this example, only class c_1 leads to uncertainty. In all of the other classes, all lexemes instantiate the same pattern. It is important to remember that, even though the classification is entirely based on phonological form, the resulting categories are not necessarily phonologically natural: for example, class c_1 is the class of verbs whose PRS.IND.1SG form ends in /u/, but not specifically in /kˈɛɾu/. This lack of phonological naturalness of the classes is expected: the categories are automatically determined according to the inventory of the patterns of alternation encountered, and not according to preconceptions on the sensitivity of the alternation to any phonological property. It also indicates that the classification in question firmly belongs to the domain of morphology, and cannot be reduced to purely phonological generalisations.

Class	Size	Sample lexeme	Pattern	Example forms	Count
c_1	4982	PASSAR 'pass'	u ⇌ ə / X+_	pˈasu ⇌ pˈasə	4166
		TER 'have'	u ⇌ ɐ / X+_	tˈɐɲu ⇌ tˈɐɲɐ	816
c_2	2	QUERER 'want'	ˈɛ_u ⇌ ˈɐj_ɐ / [+son]*k_ɾ_	kˈɛɾu ⇌ kˈɐjɾɐ	2
			u ⇌ ə / X+_		0
			u ⇌ ɐ / X+_		0
c_3	2	SER 'be'	ˈo ⇌ ˈɐjʒɐ / X*s,t_	sˈo ⇌ sˈɐjʒɐ	2
c_4	2	SABER 'know'	ˈɐj ⇌ ˈajbɐ / [+nas]*s_	sˈɐj ⇌ sˈajbɐ	2
c_5	1	IR 'go'	ˈo ⇌ ˈa / v_	vˈo ⇌ vˈa	1
c_6	1	DAR 'give'	ˈo ⇌ ˈe / d_	dˈo ⇌ dˈe	1
c_7	1	HAVER 'there to be'	ˈɐj ⇌ ˈaʒɐ / _	ˈɐj ⇌ ˈaʒɐ	1

Table 9: Classes of PRS.IND.1SG for prediction of PRS.SBJV.1SG

Having established and exemplified the methodology of the algorithm, let us examine the results of its application for some well-known examples. First, it is important to remember that the implicative relations between two forms are generally directional: what is predictable in one direction may be unpredictable in the other, and vice versa. This is apparent in the task of predicting the PRS.IND.1SG from the PRS.SBJV.1SG. As shown in

[16] The classes in question are specific to a pair of a predictor and a predicted paradigm cell. Hence they are *not* inflection classes in the traditional sense. In Tables 11 to 12, classes are numbered arbitrarily, from the largest to the smallest. There is typically no correspondence between the members of e.g. class c_3 in one table and the next, and no interpretation to the fact that two classes in different tables have the same label.

Table 10,[17] the uncertainty present in the opposite direction is absent here: it is not possible to predict the PRS.SBJV.1SG's theme vowel from the athematic form of PRS.IND.1SG, but conversely, removing the theme vowel from the PRS.SBJV.1SG to form the PRS.IND.1SG can be done with almost perfect certainty. It follows that PRS.IND.1SG is almost categorically predictable from PRS.SBJV.1SG.

Class	Size	Sample lexeme	Pattern	Example forms	Count
c_1	4166	PASSAR 'pass'	ə ⇌ u / X+_	pˈasə ⇌ pˈasu	4166
c_2	814	TER 'have'	ɐ ⇌ u / X+_	tˈɐɲɐ ⇌ tˈɐɲu	814
c_3	3	QUERER 'want'	ˈɐj_ɐ ⇌ ˈɛ_u / [+son]*k_ɾ_	kˈɐjɾɐ ⇌ kˈɛɾu	2
		REQUERER 'require'	ɐ ⇌ u / X+_	ʀəkˈɐjɾɐ ⇌ ʀəkˈɐjɾu	1
c_4	2	SER 'be'	ˈɐjʒɐ ⇌ ˈo / X*s,t_	sˈɐjʒɐ ⇌ sˈo	2
			ɐ ⇌ u / X+_		0
c_5	2	SABER 'know'	ˈajbɐ ⇌ ˈɐj / [+nas]*s_	sˈajbɐ ⇌ sˈɐj	2
			ɐ ⇌ u / X+_		0
c_6	2	HAVER 'there to be'	ˈaʒɐ ⇌ ˈɐj / _	ˈaʒɐ ⇌ ˈɐj	1
		AGIR 'act'	ɐ ⇌ u / X+_	ˈaʒɐ ⇌ ˈaʒu	1
c_7	1	IR 'go'	ˈa ⇌ ˈo / v_	vˈa ⇌ vˈo	1
c_8	1	DAR 'give'	ˈe ⇌ ˈo / d_	dˈe ⇌ dˈo	1

Table 10: Classes of PRS.SBJV.1SG for prediction of PRS.IND.1SG

In some extreme cases, there is perfect predictability between two cells in the paradigm in both directions. An example of such a situation is found between the infinitive and the PRS.IND.1PL form. Tables 11 and 12 show the patterns of alternation and their distribution in both directions. It can be seen that in both cases, there is no uncertainty: the first two patterns have mutually exclusive contexts of application while the other classes each contain a single verb.

Class	Size	Sample lexeme	Pattern	Example forms	Count
c_1	4168	ESTAR 'be'	ˈaɾ ⇌ ˈɐmuʃ / X+_	əʃtˈaɾ ⇌ əʃtˈɐmuʃ	4168
c_2	821	TER 'have'	ɾ ⇌ muʃ / X+{ˈe,ˈi,ˈo,ˈu}_	tˈeɾ ⇌ tˈemuʃ	821
c_3	1	SER 'be'	ˈeɾ ⇌ ˈomuʃ / s_	sˈeɾ ⇌ sˈomuʃ	1
			ɾ ⇌ muʃ / X+{ˈe,ˈi,ˈo,ˈu}_		0
c_4	1	IR 'go'	ˈiɾ ⇌ vˈɐmuʃ / _	ˈiɾ ⇌ vˈɐmuʃ	1

Table 11: Classes of INF for predicition of PRS.IND.1PL

The examples above have been chosen for their simplicity. On average, two paradigm cells X and Y are subject to a large number of patterns of alternation in either direction, along with a significant degree of unpredictability. Because of this, manually examining the patterns of alternation and their distribution one by one is a long and tedious task, which should only be undertaken if strictly necessary.

[17] For ease of readability we display different patterns in Tables 9 and 10, with the predictor form described on the left hand side of the double arrow, although technically the same patterns are used in both cases, wiht a change of directionality.

Class	Size	Sample lexeme	Pattern	Example forms	Count
c_1	4168	estar ‘be’	ˈɐmuʃ ⇌ ˈaɾ / X+_	əʃtˈɐmuʃ ⇌ əʃtˈaɾ	4168
c_2	821	ter ‘have’	muʃ ⇌ ɾ / X+{ˈe,ˈi,ˈo,ˈu}_	tˈemuʃ ⇌ tˈeɾ	821
c_3	1	ser ‘be’	ˈomuʃ ⇌ ˈeɾ / s_	sˈomuʃ ⇌ sˈeɾ	1
			muʃ ⇌ ɾ / X+{ˈe,ˈi,ˈo,ˈu}_		0
c_4	1	ir ‘go’	vˈɐmuʃ ⇌ ˈiɾ / _	vˈɐmuʃ ⇌ ˈiɾ	1
			ˈɐmuʃ ⇌ ˈaɾ / X+_		0

Table 12: Classes of prs.ind.1pl for prediction of inf

3.3 *Conditional entropy as an overall measure of predictability*

The distribution tables of patterns of alternation presented above give a fairly detailed picture of the implicative structure of only a small sample of the conjugation of European Portuguese. To assess the implicative structure of the system as a whole, item-by-item examination of these tables is not a viable method: with a 65 cell paradigm, one would need to examine $65 \times 64 = 4160$ tables. It is therefore essential to have quantitative measures that summarize the information in these tables. Following Ackerman, Blevins, and Malouf (2009), we use conditional entropy as a quantitative measure to obtain a global picture of the predictability of a cell in one paradigm from another. The rationale for choosing this measure goes beyond what can be elaborated upon in this article. We will therefore simply give an overview that will allow the reader to understand how conditional entropy is calculated.

The initial idea is to give an explicitly probabilistic interpretation to the distributions of the patterns of alternation studied above. Let us first consider how likely it is that a random verb partake in a particular alternation pattern relating its prs.ind.1sg and prs.sbjv.1sg forms. Table 7 gives us type frequency counts for all the possible patterns, which we can use to estimate probabilities, as illustrated in (2).

(2) a. $P(\text{ind.prs.1sg} \sim \text{subj.prs.1sg} : [\text{u} \rightleftharpoons \text{ə}]) \approx \frac{4166}{4991} = 0.8347$

b. $P(\text{ind.prs.1sg} \sim \text{subj.prs.1sg} : [\text{u} \rightleftharpoons \text{ɐ}]) \approx \frac{816}{4991} = 0.1635$

c. $P(\text{ind.prs.1sg} \sim \text{subj.prs.1sg} : [\text{ˈɛ\_u} \rightleftharpoons \text{ˈɐj\_ɐ}]\ ) \approx \frac{2}{4991} = 0.0004$

d. $P(\text{ind.prs.1sg} \sim \text{subj.prs.1sg} : [\text{ˈɐj} \rightleftharpoons \text{ˈajbɐ}]\ ) \approx \frac{2}{4991} = 0.0004$

e. $P(\text{ind.prs.1sg} \sim \text{subj.prs.1sg} : [\text{ˈo} \rightleftharpoons \text{ˈɐjʒɐ}]\ ) \approx \frac{2}{4991} = 0.0004$

f. $P(\text{ind.prs.1sg} \sim \text{subj.prs.1sg} : [\text{ˈɐj} \rightleftharpoons \text{ˈaʒɐ}]\ ) \approx \frac{1}{4991} = 0.0002$

g. $P(\text{ind.prs.1sg} \sim \text{subj.prs.1sg} : [\text{ˈo} \rightleftharpoons \text{ˈa}]\ ) \approx \frac{1}{4991} = 0.0002$

h. $P(\text{ind.prs.1sg} \sim \text{subj.prs.1sg} : [\text{ˈo} \rightleftharpoons \text{ˈe}]\ ) \approx \frac{1}{4991} = 0.0002$

Let us then consider the size of the classes based on the prs.ind.1sg form (cf. the “Class size” column of Table 9). These allow us to estimate the probability that, given the prs.ind.1sg of a random lexeme, it belongs to each of the classes in question.

(3) a. $P(\text{IND.PRS.1SG} \in c_1) \approx \dfrac{4982}{4991} = 0.9982$

b. $P(\text{IND.PRS.1SG} \in c_2) \approx \dfrac{2}{4991} = 0.0004$

c. $P(\text{IND.PRS.1SG} \in c_3) \approx \dfrac{2}{4991} = 0.0004$

d. $P(\text{IND.PRS.1SG} \in c_4) \approx \dfrac{2}{4991} = 0.0004$

e. $P(\text{IND.PRS.1SG} \in c_5) \approx \dfrac{1}{4991} = 0.0002$

f. $P(\text{IND.PRS.1SG} \in c_6) \approx \dfrac{1}{4991} = 0.0002$

g. $P(\text{IND.PRS.1SG} \in c_7) \approx \dfrac{1}{4991} = 0.0002$

Finally, the numbers in the "Pattern frequency" column of Table 9 can be interpreted in probabilistic terms. These correspond to conditional probabilities: for example, if we restrict ourselves to class c_1, 4166 out of 4982 $\approx$ 84% of verbs instantiate the pattern [u ⇌ ə / X+_] (first row on the table). Based on this percentage, we can infer that, knowing a verb belongs to class c$_1$, the conditional probability for a random verb from class c_1 to instantiate the pattern [u ⇌ ə / X+_] is about 0.84.

Table 13 shows the conditional probabilities that can be deduced in this way from Table 9. It should be noted that a zero value for conditional probability may occur for two reasons: either the pattern, although applicable in principle to this class, is not instantiated by any member of the class; or the pattern is simply not applicable to members of the class. The cells corresponding to the latter situation are greyed out in the table.

Pattern	c_1	c_2	c_3	c_4	c_5	c_6	c_7
u ⇌ ə / X+_	0.843	0	0	0	0	0	0
u ⇌ ɐ / X+_	0.164	0	0	0	0	0	0
ˈɛ_u ⇌ ˈɐj_ɐ / [+son]*k_ɾ_	0	1	0	0	0	0	0
ˈo ⇌ ˈɐjʒɐ / X*{s,t}_	0	0	1	0	0	0	0
ˈɐj ⇌ ˈajbɐ / [+nas]*s_	0	0	0	1	0	0	0
ˈo ⇌ ˈa / v_	0	0	0	0	1	0	0
ˈo ⇌ ˈe / d_	0	0	0	0	0	1	0
ˈɐj ⇌ ˈaʒɐ / _	0	0	0	0	0	0	1

Table 13: Conditional probability of each pattern relating PRS.IND.1SG to PRS.SBJV.1SG, given the category of the PRS.IND.1SG form

Table 13 provides an explicit probabilistic interpretation of the observations on predictability that we have deduced from Table 9. This interpretation has the advantage of allowing the use of tools from information theory to summarize the predictability of the relation between two paradigm cells. Given any random variable X, the entropy of this random variable, noted $H(X)$, is defined as (4).

(4) $H(X) = -\sum\limits_{x \in X} P(x) \log_2 P(x)$

Intuitively, entropy captures the uncertainty inherent in selecting a value for a random variable. Imagine a situation where a variable X_1 has two equally probable possible values. In this case, as shown in the calculation in (5), the entropy is 1.

(5) $H(X_1) = -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) = -\log_2\frac{1}{2} = \log_2 2 = 1$

If one of the possible values is more likely than the other, the entropy decreases, as seen in (6): if we have two possible values, one is three times more likely than the other, the entropy drops below 1.

(6) $H(X_2) = -\left(\frac{3}{4}\log_2\frac{3}{4} + \frac{1}{4}\log_2\frac{1}{4}\right) \approx 0.81$

An extreme case is illustrated in (7): if one of the two values, although theoretically possible, is never attested in practice, the entropy falls to 0.[18] An entropy value of zero therefore corresponds to a situation of certainty.

(7) $H(X_3) = -1 \times \log_2 1 = 0$

If the number of possible values for the random variable increases, all other things being equal, the entropy also increases: for example, if there are 4 equally probable values, as illustrated in (8), the entropy is 2.

(8) $H(X_1) = -\left(\frac{1}{4}\log_2\frac{1}{4} + \frac{1}{4}\log_2\frac{1}{4} + \frac{1}{4}\log_2\frac{1}{4} + \frac{1}{4}\log_2\frac{1}{4}\right) = -\log_2\frac{1}{4} = \log_2 4 = 2$

Let us look now look at a concrete example: from the data in (3), let us determine the uncertainty in ascertaining the class of a previously unseen verb. If we trust the information in our dataset, there are 5 possibilities (one for each of the 7 classes). The calculation in (9) quantifies the entropy as being very close to zero, which corresponds to the intuition that given a random verb, there is very little uncertainty about the kind of shape its PRS.IND.1SG could have.

(9) $$H(\textsc{prs.ind.1sg}) = -\begin{pmatrix} & 0.9982 \times \log_2 0.9982 \\ + & 0.0004 \times \log_2 0.0004 \\ + & 0.0004 \times \log_2 0.0004 \\ + & 0.0004 \times \log_2 0.0004 \\ + & 0.0002 \times \log_2 0.0002 \\ + & 0.0002 \times \log_2 0.0002 \\ + & 0.0002 \times \log_2 0.0002 \end{pmatrix} \approx 0.0306$$

We now turn to conditional entropy. Given two random variables X and Y, the conditional entropy of Y knowing X, written as $H(Y \mid X)$, is defined in (10).

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

Conditional entropy captures the dependence between the uncertainty associated with two random variables: if the value of X is strongly predictive of the value of Y, conditional entropy will be low; if the value of X only slightly predicts the value of Y, entropy will be higher. The calculation takes the form of a weighted sum of the entropies calculated for

[18] Classes with a probability of exactly 0 are not taken into account in entropy calculations, since log 0 is undefined. Note though that this is innocuous, since $\lim_{x \to 0} x \log_2 x = 0$.

each possible value of the variable X. In the extreme case where the values of X and Y are completely independent of each other, $H(Y \mid X) = H(Y)$. Conversely, if the value of X unequivocally determines the value of Y, $H(Y \mid X) = 0$.

We are now in a position to use entropy to provide a global measure of predictability within paradigms. Specifically, we can compute the conditional entropy of choosing a pattern of alternation given class membership of the predictor form. For example, from (3) and Table 9, we compute the conditional entropy of the patterns relating PRS.IND.1SG and PRS.SBJV.1SG given the class of the PRS.IND.1SG form:

(11)
$$H(\textsc{prs.ind.}1\textsc{sg} \sim \textsc{prs.sbjv.}1\textsc{sg} \mid \textsc{prs.sbjv.}1\textsc{sg}) = \\ -\left(\begin{array}{rl} & 0.9982 \times (0.836 \times \log_2 0.836 + 0.164 \times \log_2 0.164) \\ + & 0.0004 \times (1 \times \log_2 1) \\ + & 0.0004 \times (1 \times \log_2 1) \\ + & 0.0004 \times (1 \times \log_2 1) \\ + & 0.0002 \times (1 \times \log_2 1) \\ + & 0.0002 \times (1 \times \log_2 1) \\ + & 0.0002 \times (1 \times \log_2 1) \end{array}\right) \approx 0.6421$$

Applying the same calculation to the data in Table 10 results in a conditional entropy of zero for prediction of the PRS.IND.1SG from the PRS.SBJV.1SG): since each class contains only one pattern with a non-zero probability, each of the terms in the sum is zero. The difference between total and partial predictiveness is therefore well captured, and degrees of partial predictiveness can be differentiated quantitatively.

It is important to keep in mind that entropy can only provide a summary of a conditional probability distribution and, like any summary, it provides less information than the original. Looking at the conditional entropy values is therefore always a shortcut, and it is necessary to examine the full probability tables to understand the linguistic factors responsible for a certain value. This will be the approach of the following section: starting from a macroscopic view of the paradigms of European Portuguese induced by the examination of the distribution of conditional entropy values, we will examine in more detail specific implicative relations between pairs of cells and determine which properties of the system lead to a particularly high or particularly low predictability.

4. Empirical results

In this section we examine the results produced by the automatic analysis of the implicative relations between the 4160 pairs of cells in the European Portuguese paradigms. Since it is, of course, not possible to examine all cases one by one, we will be selective in the information we present.

4.1 *Categorical predictability*

Let us start by looking at examples of categorical implication, that is, implications with no exceptions. A first important observation is that out of the 4160 ordered pairs of cells, entropy is null in 818 cases, or about 19% of the cases. In other words, if two cells are randomly selected from the paradigm, in one out of five cases the shape of the second cell is completely predictable on the basis of that of the first. Likewise, there is categorical predictability in both directions for 382 out of 2080 unordered pairs of cells, or about 18% of the cases. This situation is markedly different from full interpredictability over the whole paradigm.

If we examine the situations that favour total interpredictability, it is clear that it is most often due to a situation that resembles canonical inflectional morphology (Corbett, 2007). In (12) we identify some sufficient conditions:

(12) Sufficient conditions for a relation of absolute interpredictability
  a. The two paradigm cells have the stress in the same position, and
  b. there is regular demarcation throughout the lexicon between a constant string (a pseudo-stem) and a variable string (a pseudo-ending),[19] and
  c. theme vowels, where they are present, vary predictably from one cell to the other.

A closer look at the forms of the imperfect (Table 14) illustrates this point: stress is on the theme vowel for all forms. The theme vowel's quality is not constant, but it is nevertheless predictable: lexemes that have /a/ in the 1sg have constant vowel quality throughout the paradigm, and those that have /iɐ/ keep it throughout except in 3pl where they have a /iɐ̃/. And lastly, the forms show no stem allomorphy.

	1sg	2sg	3sg	1pl	2pl	3pl
utilizar 'use'	utiliz'**a**vɐ	utiliz'**a**vɐʃ	utiliz'**a**vɐ	utiliz'**a**vɐmuʃ	utiliz'**a**vɐjʃ	utiliz'**a**vɐ̃w
aprender 'learn'	ɐpɾẽd'**iɐ**	ɐpɾẽd'**iɐ**ʃ	ɐpɾẽd'**iɐ**	ɐpɾẽd'**iɐ**muʃ	ɐpɾẽd'**iɐj**ʃ	ɐpɾẽd'**iɐ̃w**
imprimir 'print'	ĩpɾim'**iɐ**	ĩpɾim'**iɐ**ʃ	ĩpɾim'**iɐ**	ĩpɾim'**iɐ**muʃ	ĩpɾim'**iɐj**ʃ	ĩpɾim'**iɐ̃w**

Table 14: Representative examples of the imperfect paradigm

Not all cases of absolute interpredictability adhere to the conditions in (12). A particularly enlightening exception is found in the perfective preterite (Table 15). Regular verbs have 5 directly analogous forms in this subparadigm, with stress on the theme vowel and a three-way vowel quality contrast. The 1sg is an exception, in that it neutralizes the distinction between the second and third conjugations; it is therefore not a good predictor of the rest of the sub-paradigm. If we examine irregular verbs however, both the 1sg and 3sg forms reveal a number of verbs that are exceptions to (13): the expected ending and/or stem are not used. It is interesting to note that these cases do not in fact reduce predictiveness because of their distinctive phonotactic properties. All these forms have a property that makes it possible to distinguish them from the standard inflectional pattern: an unusual ending in the 1sg and 3sg, the theme vowel /ɛ/ in the 4 remaining cells.

Another example of the same type can be found in the present indicative (Table 16). In this subparadigm, 2sg and 3sg are interpredictable. For these two cells, the vast majority of verbs satisfy the conditions in (12): stress is on the pre-theme vowel, and there is a two-way theme vowel contrast (/ɐ/ vs. /ə/) between first conjugation and other verbs, before invariant endings (/ʃ/ or zero). However, there is a small number of irregular verbs that use no suffix in the 3sg: 42 verbs of the regular 2nd and 3rd conjugations with an stem ending

[19] We talk of pseudo-stem and pseudo-ending because the segmentation used only makes sense for the two cells in question within the paradigm, and can't be extended to other forms, even in cases in which, from a descriptive perspective, we wouldn't want to assert that there is stem allomorphy.

	1SG	2SG	3SG	1PL	2PL	3PL
UTILIZAR ‘use’	utilizˈɐj	utilizˈaʃtə	utilizˈo	utilizˈamuʃ	utilizˈaʃtəʃ	utilizˈarɐ̃w
APRENDER ‘learn’	ɐprẽdˈi	ɐprẽdˈeʃtə	ɐprẽdˈew	ɐprẽdˈemuʃ	ɐprẽdˈeʃtəʃ	ɐprẽdˈerɐ̃w
IMPRIMIR ‘print’	ĩprimˈi	ĩprimˈiʃtə	ĩprimˈiw	ĩprimˈimuʃ	ĩprimˈiʃtəʃ	ĩprimˈirɐ̃w
FAZER ‘do’	f'iʃ	fizˈɛʃtə	f'eʃ	fizˈɛmuʃ	fizˈɛʃtəʃ	fizˈɛrɐ̃w
QUERER ‘want’	kˈiʃ	kizˈɛʃtə	kˈiʃ	kizˈɛmuʃ	kizˈɛʃtəʃ	kizˈɛrɐ̃w
TRAZER ‘bring’	trˈosə	trosˈɛʃtə	trˈosə	trosˈɛmuʃ	trosˈɛʃtəʃ	trosˈɛrɐ̃w
opor ‘oppose’	opˈuʃ	opuzˈɛʃtə	opˈoʃ	opuzˈɛmuʃ	opuzˈɛʃtəʃ	opuzˈɛrɐ̃w
VIR ‘come’	vˈĩ	viˈɛʃtə	vˈɐju	viˈɛmuʃ	viˈɛʃtəʃ	viˈɛrɐ̃w

Table 15: Representative examples of the perfective preterite paradigm

in /z/ (the /z/ is palatalised and devoiced word-finally, a regular phonological process), as well as QUERER ‘want’ and its derivatives, which are the only verbs with a stem ending in /kˈɛr/. In both cases, the specific phonotactic characteristics of the subclass avoid any ambiguity: no verb ends in /ʃ/ or /kˈɛr/ in the 3SG if it does not fall under one of these cases; the same applies to verbs ending in /zəʃ/ or /kˈɛrəʃ/ in the 2SG.

	1SG	2SG	3SG	1PL	2PL	3PL
UTILIZAR ‘use’	utilˈizu	utilˈizɐʃ	utilˈizɐ	utilizˈɐmuʃ	utilizˈajʃ	utilˈizɐ̃w
APRENDER ‘learn’	ɐprˈẽdu	ɐprˈẽdəʃ	ɐprˈẽdə	ɐprẽdˈemuʃ	ɐprẽdˈɐjʃ	ɐprˈẽdɐ̃j
IMPRIMIR ‘print’	ĩprˈimu	ĩprˈiməʃ	ĩprˈimə	ĩprimˈimuʃ	ĩprimˈiʃ	ĩprˈimɐ̃j
TRADUZIR ‘translate’	trɐdˈuzu	trɐdˈuzəʃ	trɐdˈuʃ	trɐduzˈimuʃ	trɐduzˈiʃ	trɐdˈuzɐ̃j
QUERER ‘want’	kˈɛru	kˈɛrəʃ	kˈɛr	kərˈemuʃ	kərˈɐjʃ	kˈɛrɐ̃j

Table 16: The indicative present paradigm, showing interpredictability between the 2SG and the 3SG

4.2 *Distilling the paradigm*

From the observation of mutual interpredictability relations, it is possible to identify sets of paradigm cells that mutually predict each other. By choosing the largest possible sets, we end up with a partition of the paradigm into zones of perfect mutual interpredictability, corresponding to what Ackerman, Blevins, and Malouf (2009) call “alliances of forms”. [20] In Table 17, the paradigm space has been partitioned into 12 sets of cells or “zones of interpredictability”, each labelled with a different label (a label appears multiple times when the zone it labels does not correspond to a contiguous area in the table).

The boundaries of Table 17 highlight the morphomic character (as defined by Aronoff, 1994) of zones of interpredictability. Some areas correspond to a morphosyntactically natural class of forms: this is trivially the case for single-cell zones (Z1, Z5, Z7, and

[20] See also the related concepts of *overall distribution schema* (Pirrelli and Battista, 2000). These zones are related to Bonami and Boyé’s 2002 stem spaces, with important caveats. First, two cells may be interpredictable without sharing a stem, if the stem allomorphy is fully predictable. Second, Bonami and Boyé’s approach allowed for destructive phonological operations in the derivation of surface wordforms from stems; hence in some cases wordforms may share a stem without being interpredictable. See Bonami and Boyé (2014) for a spectacular example of that situation in French.

<table>
<tr><th></th><th>1SG</th><th>2SG</th><th>3SG</th><th>1PL</th><th>2PL</th><th>3PL</th></tr>
<tr><td>PRS.IND</td><td>Z1</td><td colspan="2">Z2</td><td>Z3</td><td>Z4</td><td>Z5</td></tr>
<tr><td>PST.IMPF.IND</td><td colspan="6">Z6</td></tr>
<tr><td>PST.PFV.IND</td><td>Z7</td><td colspan="5" rowspan="2">Z8</td></tr>
<tr><td>PST.PERF.IND</td><td></td></tr>
<tr><td>FUT.IND</td><td colspan="6" rowspan="2">Z9</td></tr>
<tr><td>COND</td></tr>
<tr><td>PRS.SBJV</td><td colspan="3">Z10</td><td colspan="2">Z11</td><td>Z10</td></tr>
<tr><td>PST.SBJV</td><td colspan="6" rowspan="2">Z8</td></tr>
<tr><td>FUT.SBJV</td></tr>
<tr><td>IMP</td><td>—</td><td>Z2</td><td>—</td><td>—</td><td>Z4</td><td>—</td></tr>
<tr><td>PER.INF</td><td colspan="6">Z3</td></tr>
</table>

<table>
<tr><th>INF</th><th>GER</th><th>PST.PTCP.M.SG</th></tr>
<tr><td colspan="2">Z3</td><td>Z12</td></tr>
</table>

Table 17: Partition into zones of interpredictability of the verbal paradigm of European Portuguese

Z11), as well as Z6 (indicative past imperfective) and one could argue that this is also true for Z9 (assuming that future and conditional have semantic properties in common). On the other hand, the inventories of forms corresponding to the remaining zones are clearly disjunctive, as illustrated by the statement in (13): in each case, there is no way to describe the set of cells of the paradigm concerned with a descriptive statement that would only involve conjunctions (and not disjunctions or negations).

(13) a. Z2: (2SG or 3SG) and (PRS.IND or IMP.2SG)
b. Z3: INF or GER or PRS.IND.1PL
c. Z4: (PRS.IND or IMP) and 2PL
d. Z8: (PST.IND and (not IMPF) and (not PFV.1SG)) or (SBJV and (not PRS))
e. Z10: PRS.SBJV and (SG or 3PL)
f. Z11: PRS.SBJV and (1PL or 2PL)

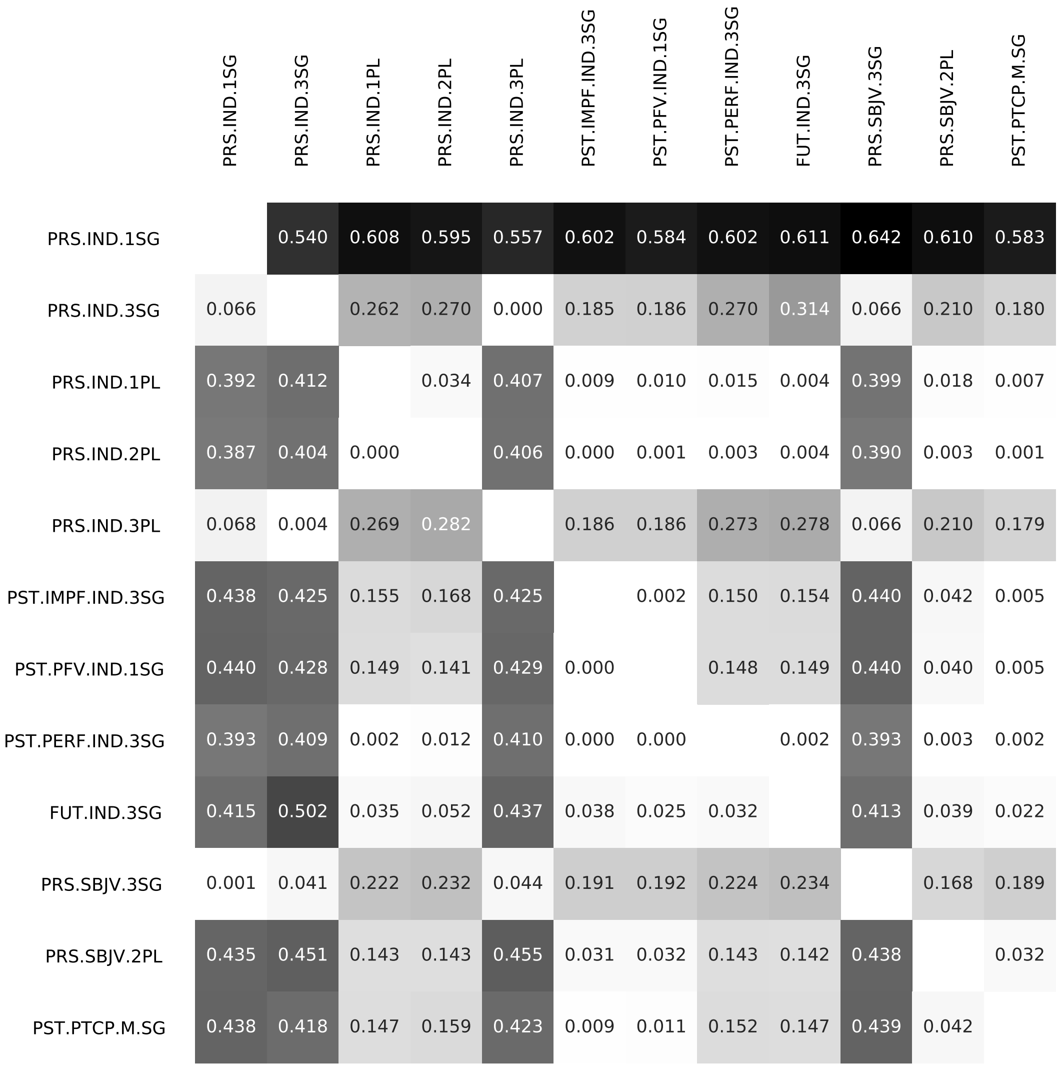
From the partition of Table 17 we can construct what Stump and Finkel (2013) call a *distillation* of the paradigm. A distillation is a set of cells in which each member corresponds to a distinct interpretability zone. The purpose of a distillation is to allow the study of a sub-paradigm of a much more reasonable size without loss of information: given two zones *Z* and *Z'*, we expect that any implicative relation of a cell of *Z* to a cell of *Z'* has an exact correspondent which connects any other *Z* cell to any other *Z'* cell. In principle, the choice of the cell representing each zone is inconsequential; among the hundreds of thousands of possibilities,[21]we choose the one indicated in Table 18.

Focusing on a distillation, it is possible to examine in detail the conditional entropy values linking the cells of the paradigm. Figure 2 provides detailed figures for the selected

[21] The choice made between the cells of each zone is independent of those made for the other zones. For a partition in n zones Z_1, …, Z_n there are therefore $|Z_1|\times$□$\times|Z_n|$ different partitions ; in the case at hand, $1 \times 3 \times 9 \times 2 \times 1 \times 6 \times 1 \times 23 \times 12 \times 4 \times 2 \times 1 = 715,392$.

Zone	Cell	Zone	Cell	Zone	Cell
Z1	PRS.IND.1SG	Z5	PRS.IND.3PL	Z9	FUT.IND.3SG
Z2	PRS.IND.3SG	Z6	PST.IMPF.IND.3SG	Z10	PRS.SBJV.3SG
Z3	PRS.IND.1PL	Z7	PST.PFV.IND.1SG	Z11	PRS.SBJV.2PL
Z4	PRS.IND.2PL	Z8	PST.PERF.IND.3SG	Z12	PST.PTCP

Table 18: A distillation of the European Portuguese conjugation paradigm



	PRS.IND.1SG	PRS.IND.3SG	PRS.IND.1PL	PRS.IND.2PL	PRS.IND.3PL	PST.IMPF.IND.3SG	PST.PFV.IND.1SG	PST.PERF.IND.3SG	FUT.IND.3SG	PRS.SBJV.3SG	PRS.SBJV.2PL	PST.PTCP.M.SG
PRS.IND.1SG		0.540	0.608	0.595	0.557	0.602	0.584	0.602	0.611	0.642	0.610	0.583
PRS.IND.3SG	0.066		0.262	0.270	0.000	0.185	0.186	0.270	0.314	0.066	0.210	0.180
PRS.IND.1PL	0.392	0.412		0.034	0.407	0.009	0.010	0.015	0.004	0.399	0.018	0.007
PRS.IND.2PL	0.387	0.404	0.000		0.406	0.000	0.001	0.003	0.004	0.390	0.003	0.001
PRS.IND.3PL	0.068	0.004	0.269	0.282		0.186	0.186	0.273	0.278	0.066	0.210	0.179
PST.IMPF.IND.3SG	0.438	0.425	0.155	0.168	0.425		0.002	0.150	0.154	0.440	0.042	0.005
PST.PFV.IND.1SG	0.440	0.428	0.149	0.141	0.429	0.000		0.148	0.149	0.440	0.040	0.005
PST.PERF.IND.3SG	0.393	0.409	0.002	0.012	0.410	0.000	0.000		0.002	0.393	0.003	0.002
FUT.IND.3SG	0.415	0.502	0.035	0.052	0.437	0.038	0.025	0.032		0.413	0.039	0.022
PRS.SBJV.3SG	0.001	0.041	0.222	0.232	0.044	0.191	0.192	0.224	0.234		0.168	0.189
PRS.SBJV.2PL	0.435	0.451	0.143	0.143	0.455	0.031	0.032	0.143	0.142	0.438		0.032
PST.PTCP.M.SG	0.438	0.418	0.147	0.159	0.423	0.009	0.011	0.152	0.147	0.439	0.042	

Figure 2: Conditional entropy within the distillation from Table 18

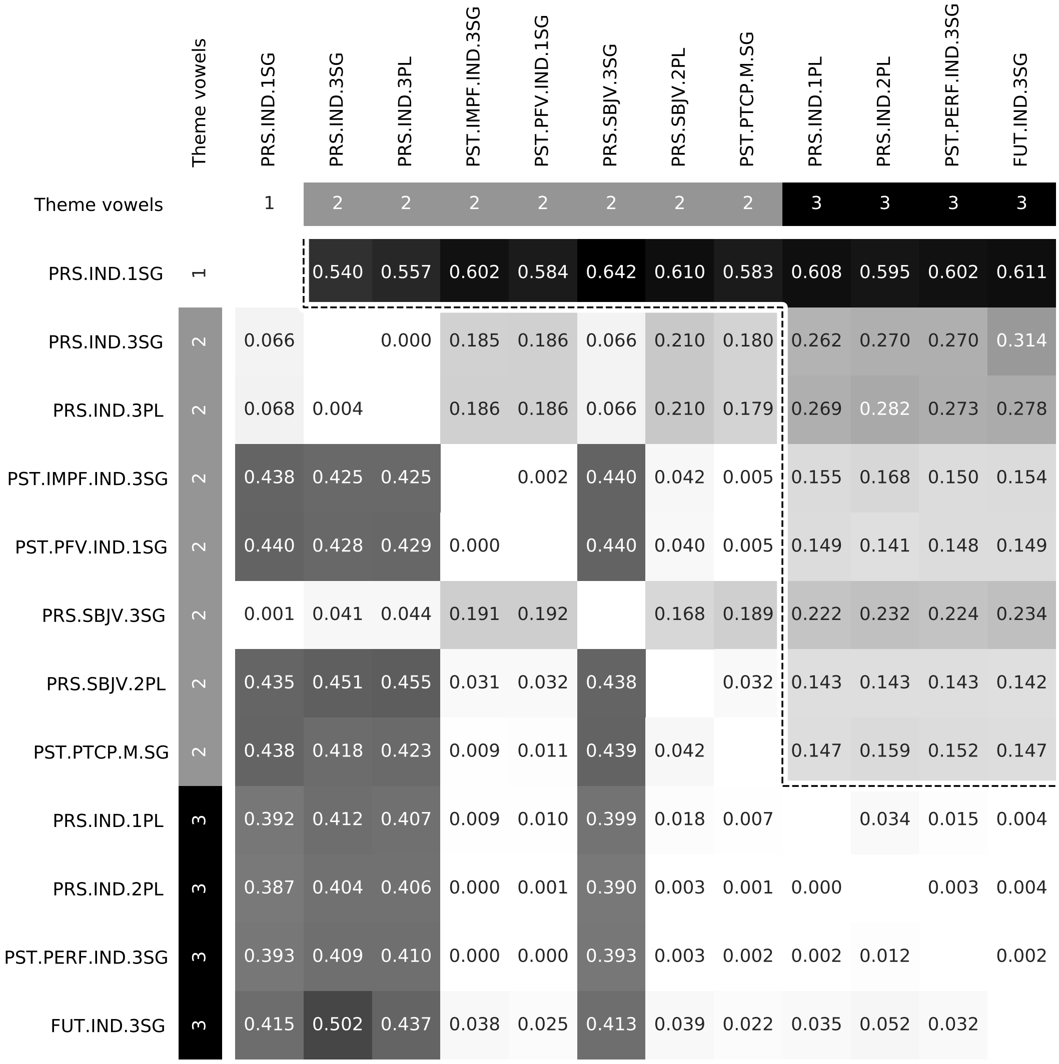
distillation. The remainder of this section will be devoted to identifying the what regularities in the dataset lead to the predictability contrasts that this figure highlights. In this figure, the shade of the background of each cell shows the relative height of the entropy value found by the algorithm: the darker the background, the higher the value. A simple examination of the rows of Figure 2 reveals that the PRS.IND.1SG is the worse predictor for the rest of the paradigm: the values in the first row are higher than all the other values in the table. Things are a lot more gradient at the other end of the spectrum, with the best predictors being the PRS.IND.2PL (average predictiveness 0.058) and PST.PERF.IND.3SG (0.057). If we now look at the columns of Figure 2, the structure is not so clear-cut, although we can see that predictiveness is unevenly distributed: depending on the predictor form, low predictability is related to different parts of the paradigm. In the following paragraphs, we identify four main sources of prediction difficulty and show how these factors lead to the numbers in Figure 2.

4.3 *Neutralisation of inflectional classes*

A main source of unpredictability in paradigms is clearly the neutralisation of distinctions between conjugations in certain cells. Let us look again at the present indicative of regular verbs (Table 1). The 1PL and 2PL cells are the only ones to present a distinctive conjugation marker, in the form of an opposition between 3 theme vowels: /ɐ/ vs. /e/ vs. /i/ in the 1PL, /ai/ vs. /ɐi/ vs. /i/ in the 2PL. Three other cells neutralize the distinction between 2nd and 3rd conjugation: we have /ɐ/ vs. /ə/ only in the 2SG and 3SG forms, and /ɐ̃ũ/ vs. /ɐ̃ĩ/ only in the 3PL forms; this leads to a two-way instead of a three-way contrast. It follows from this observation that predicting the 1PL form from the 3SG, for example, is a problem: if the 2SG ends with /ə/, it is not clear whether /e/ or /i/ should surface in the 1PL. In the opposite direction, predicting the theme vowel is not a problem. The 1SG form is even more problematic since it neutralizes the distinctions between all three conjugations. According to the same reasoning, we therefore expect to have difficulty predicting the other forms: to predict the 3SG form, we do not know whether the unreduced form of the vowel is /ɐ/ or /ə/, and for the 1PL, whether it is /ɐ/, /e/ or /i/.

All other things being equal, these observations lead to the expectation that predicting a cell with more theme vowel contrasts from a cell with fewer theme vowel contrasts should be harder than the other way around. In Figure 3, we present again the same numbers from Figure 2, but rearranging rows and columns to highlight the expected contrasts: the area set apart by a dashed line in the upper left part of the table is where we expect higher numbers. As the reader can check, our expectations are only partially met. All predictions from the fully-neutralised PRS.IND.1SG are maximally hard, as expected. Predictions from cells with a two-way contrast to cells with a three-way contrast are non-trivial: entropy values range from 0.141 to 0.314, and are an order of magnitude higher than those for pairs of cells that both exhibit a three-way contrast. (bottom right corner). However, these values are still lower than those found in some areas of the lower left part of the figure, where on the basis of the behaviour of theme vowels alone we would expect very low numbers.

We thus provisionally conclude that, while theme vowels alternations do contribute to explaining some contrasts in predictability in European Portuguese paradigms, they cannot be the only relevant factor.



	Theme vowels	PRS.IND.1SG	PRS.IND.3SG	PRS.IND.3PL	PST.IMPF.IND.3SG	PST.PFV.IND.1SG	PRS.SBJV.3SG	PRS.SBJV.2PL	PST.PTCP.M.SG	PRS.IND.1PL	PRS.IND.2PL	PST.PERF.IND.3SG	FUT.IND.3SG
Theme vowels		1	2	2	2	2	2	2	2	3	3	3	3
PRS.IND.1SG	1		0.540	0.557	0.602	0.584	0.642	0.610	0.583	0.608	0.595	0.602	0.611
PRS.IND.3SG	2	0.066		0.000	0.185	0.186	0.066	0.210	0.180	0.262	0.270	0.270	0.314
PRS.IND.3PL	2	0.068	0.004		0.186	0.186	0.066	0.210	0.179	0.269	0.282	0.273	0.278
PST.IMPF.IND.3SG	2	0.438	0.425	0.425		0.002	0.440	0.042	0.005	0.155	0.168	0.150	0.154
PST.PFV.IND.1SG	2	0.440	0.428	0.429	0.000		0.440	0.040	0.005	0.149	0.141	0.148	0.149
PRS.SBJV.3SG	2	0.001	0.041	0.044	0.191	0.192		0.168	0.189	0.222	0.232	0.224	0.234
PRS.SBJV.2PL	2	0.435	0.451	0.455	0.031	0.032	0.438		0.032	0.143	0.143	0.143	0.142
PST.PTCP.M.SG	2	0.438	0.418	0.423	0.009	0.011	0.439	0.042		0.147	0.159	0.152	0.147
PRS.IND.1PL	3	0.392	0.412	0.407	0.009	0.010	0.399	0.018	0.007		0.034	0.015	0.004
PRS.IND.2PL	3	0.387	0.404	0.406	0.000	0.001	0.390	0.003	0.001	0.000		0.003	0.004
PST.PERF.IND.3SG	3	0.393	0.409	0.410	0.000	0.000	0.393	0.003	0.002	0.002	0.012		0.002
FUT.IND.3SG	3	0.415	0.502	0.437	0.038	0.025	0.413	0.039	0.022	0.035	0.052	0.032	

Figure 3: The influence of theme vowel distinctions on conditional entropy

4.4 *Vowel alternations*

While the number of theme vowel contrasts in each cell is the most obvious factor having an influence on predictability in European Portuguese, it is not the only one. Another major factor is the reduction of unstressed vowels. In European Portuguese, unstressed oral vowels are typically reduced according to a pattern that induces the neutralisation of distinctions visible only in stressed syllables (see for example Mateus & d'Andrade, 2000: 17-23).[22]Table 19 lists the alternations in question.

stressed	i	e ɛ	a ɐ	ɔ o u
unstressed	i	ə	ɐ	u

Table 19: Unstressed vowel reduction in European Portuguese

Although this reduction process is lexically-conditioned and not systematic (Vigário 2003: 68-73, see also Mateus & d'Andrade, 2000: 136), it is highly prevalent, and leads to neutralisations of contrasts with important consequences on the predictability relations between forms. Once again, let us focus on the present indicative to start. In the 1sg, 2sg, 3sg and 3pl, stress is on the prethematic vowel. In the 1pl and 2pl, it falls on the theme vowel. As a result, the prethematic vowel is not stressed, leading to neutralisation, which is illustrated in Table 20 with verbs from the first conjugation.

	1sg	2sg	3sg	1pl	2pl	3pl
utilizar 'use'	utilˈ**iz**u	utilˈ**iz**ɐʃ	utilˈ**iz**ɐ	util**iz**ˈɐmuʃ	util**iz**ˈajʃ	utilˈ**iz**ɐ̃w
encenar 'stage'	ẽsˈ**en**u	ẽsˈ**en**ɐʃ	ẽsˈ**en**ɐ	ẽs**ən**ˈɐmuʃ	ẽs**ən**ˈajʃ	ẽsˈ**en**ɐ̃w
encetar 'start'	ẽsˈ**ɛt**u	ẽsˈ**ɛt**ɐʃ	ẽsˈ**ɛt**ɐ	ẽs**ət**ˈɐmuʃ	ẽs**ət**ˈajʃ	ẽsˈ**ɛt**ɐ̃w
matar 'kill'	mˈ**at**u	mˈ**at**ɐʃ	mˈ**at**ɐ	m**ɐt**ˈɐmuʃ	m**ɐt**ˈajʃ	mˈ**at**ɐ̃w
chamar 'call'	ʃˈ**ɐ**mu	ʃˈ**ɐ**mɐʃ	ʃˈ**ɐ**mɐ	ʃ**ɐ**mˈɐmuʃ	ʃ**ɐ**mˈajʃ	ʃˈ**ɐ**mɐ̃w
convocar 'summon'	kõvˈ**ɔk**u	kõvˈ**ɔk**ɐʃ	kõvˈ**ɔk**ɐ	kõv**uk**ˈɐmuʃ	kõv**uk**ˈajʃ	kõvˈ**ɔk**ɐ̃w
funcionar 'function'	fũsiˈ**on**u	fũsiˈ**on**ɐʃ	fũsiˈ**on**ɐ	fũsi**un**ˈɐmuʃ	fũsi**un**ˈajʃ	fũsiˈ**on**ɐ̃w
educar 'educate'	edˈ**uk**u	edˈ**uk**ɐʃ	edˈ**uk**ɐ	ed**uk**ˈɐmuʃ	ed**uk**ˈajʃ	edˈ**uk**ɐ̃w

Table 20: Examples of the effects of vowel reduction on the present indicative

Because it results in opacity, vowel reduction increases unpredictability: when faced with a 1pl or 2pl form with the prethematic vowel /ə/, there is uncertainty about the quality of the prethematic vowel in related forms, as it can surface as either one of /e/ or /ɛ/. The same issue arises for the 1pl and 2pl forms with prethematic /ɐ/ and /u/.

To assess the importance of this phenomenon, we examined systematically vowel alternations linked to stress shift in the relationship between prs.ind.1sg and prs.ind.1pl and found that 4909 of our verbs have the potential for such alternations.[23] Table 21 shows, for each unstressed prethematic vowel, what the options are for its stressed counterpart,

[22] Depending on the dialect and speech rate, some unstressed vowels may not be realised at all. Taking into account such graded variation phenomena is beyond the scope of the methods and data used in this article.

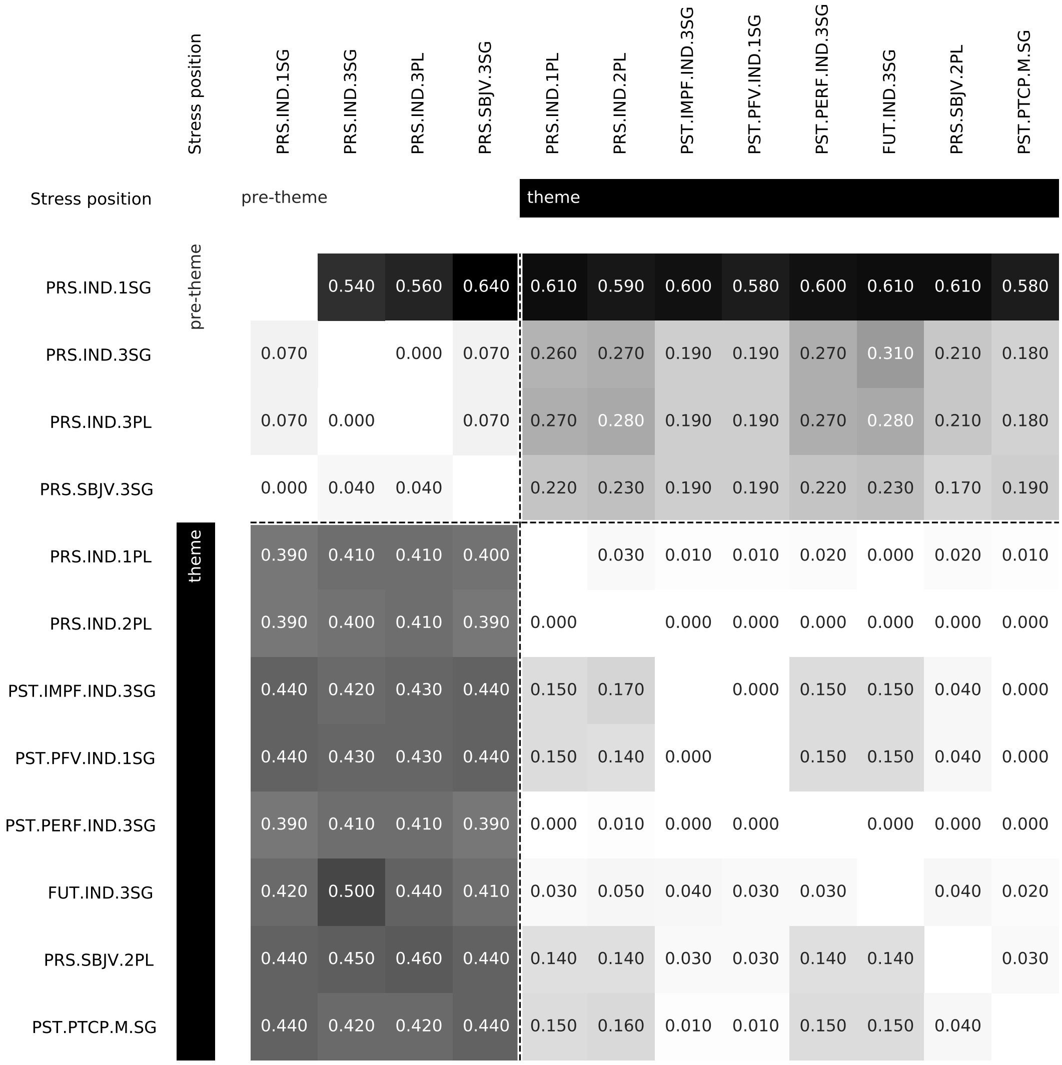
[23] The remaining 82 verbs either exhibit no stress shift, involve diphthongation, or implement some non-trivial stem allomorphy such that it is not obvious which unstressed vowel in the plural form should count as the counterpart of the stressed vowel in the singular form.

and how many verbs in the sample instantiate each possibility. For brevity, the last row lumps together all unstressed vowels that have a single stressed counterpart, so that stress shift cannot lead to opacity.

Unstressed	Stressed	Count
e	ˈe	4
	ˈew	2
	ˈɐ	1
	ˈɛ	2
i	ˈi	1469
	ˈɐj	177
o	ˈo	27
	ˈɔ	10
u	ˈo	202
	ˈu	544
	ˈɔ	367
ɐ	ˈa	533
	ˈaj	18
	ˈɐ	80
	ˈɐj	3
ɔ	ˈo	11
	ˈɔ	24
ə	ˈe	216
	ˈi	67
	ˈɐ	9
	ˈɐj	62
	ˈɛ	402
ɛ	ˈe	6
	ˈi	3
	ˈɛ	50
ẽ	ˈĩ	7
	ˈẽ	193
No opacity		420

Table 21: Prediction of stressed from unstressed prethematic vowels

As the table shows, the well-documented vowel reductions lead to a sizable amount of unpredictability; in addition, more situations of vowel alternation not listed in Table 19 emerge, that are less familiar but nevertheless contribute to making prediction hard. Overall then, vowel alternations are expected to lead to relatively high entropy values whenever a cell in which stress falls on the prethematic vowel, is predicted from a cell in which stress falls on the theme vowel. Figure 4 shows that this expectation is clearly met. We are again looking at the same numbers, but rows and columns have been reorganised to group together cells that exhibit the same stress pattern. As expected, the numbers in the lower left corner are all relatively high, between 0.39 and 0.5. In fact, these are the highest entropy values in the table if we leave out the prediction from the PRS.IND.1SG.



	Stress position	PRS.IND.1SG	PRS.IND.3SG	PRS.IND.3PL	PRS.SBJV.3SG	PRS.IND.1PL	PRS.IND.2PL	PST.IMPF.IND.3SG	PST.PFV.IND.1SG	PST.PERF.IND.3SG	FUT.IND.3SG	PRS.SBJV.2PL	PST.PTCP.M.SG
Stress position		pre-theme				theme							
PRS.IND.1SG	pre-theme		0.540	0.560	0.640	0.610	0.590	0.600	0.580	0.600	0.610	0.610	0.580
PRS.IND.3SG		0.070		0.000	0.070	0.260	0.270	0.190	0.190	0.270	0.310	0.210	0.180
PRS.IND.3PL		0.070	0.000		0.070	0.270	0.280	0.190	0.190	0.270	0.280	0.210	0.180
PRS.SBJV.3SG		0.000	0.040	0.040		0.220	0.230	0.190	0.190	0.220	0.230	0.170	0.190
PRS.IND.1PL	theme	0.390	0.410	0.410	0.400		0.030	0.010	0.010	0.020	0.000	0.020	0.010
PRS.IND.2PL		0.390	0.400	0.410	0.390	0.000		0.000	0.000	0.000	0.000	0.000	0.000
PST.IMPF.IND.3SG		0.440	0.420	0.430	0.440	0.150	0.170		0.000	0.150	0.150	0.040	0.000
PST.PFV.IND.1SG		0.440	0.430	0.430	0.440	0.150	0.140	0.000		0.150	0.150	0.040	0.000
PST.PERF.IND.3SG		0.390	0.410	0.410	0.390	0.000	0.010	0.000	0.000		0.000	0.000	0.000
FUT.IND.3SG		0.420	0.500	0.440	0.410	0.030	0.050	0.040	0.030	0.030		0.040	0.020
PRS.SBJV.2PL		0.440	0.450	0.460	0.440	0.140	0.140	0.030	0.030	0.140	0.140		0.030
PST.PTCP.M.SG		0.440	0.420	0.420	0.440	0.150	0.160	0.010	0.010	0.150	0.150	0.040	

Figure 4: The influence of stress placement on predictability

Another striking observation when looking at Figure 4 is that the numbers in the upper right corner, in which a word with stress on the theme is predicted from a word with stress on the prethematic vowel, are relatively high: while they are lower than those in the lower left corner, they are all higher than those in cells in the upper left corner and lower right corners. Why is that so? Again, we can point to stress-conditioned vowel alternation as the source. We already discussed the fact that vowel alternations led to opacity when going from an unstressed to a stressed prethematic vowel. We now observe in Table 22 that the converse is also true: because there are exceptions to vowel reduction (Vigário, 2003), the stressed version of the vowel in the stem does not always fully determine the unstressed version. However, the amount of uncertainty is significantly lower (where there is opacity, one of the options is much more prevalent than the others), which is why the numbers in the upper right corner of Figure 4 are still much lower than those in the lower left corner.

The interaction between the two sources of unpredictability identified so far gives rise

Stressed	Unstressed	Count
ˈa	a	50
	ɐ	533
ˈaj	aj	25
	ɐ	18
ˈe	e	4
	ə	216
	ɛ	6
ˈi	i	1469
	ə	67
	ɛ	3
ˈo	o	27
	u	202
	ɔ	11
ˈĩ	ĩ	66
	ẽ	7
ˈɐ	e	1
	ɐ	80
	ə	9
ˈɐj	i	177
	ɐ	3
	ɐj	60
	ə	62
ˈɔ	o	10
	u	367
	ɔ	24
ˈɛ	e	2
	ə	402
	ɛ	50
No opacity		958

Table 22: Prediction of unstressed from stressed prethematic vowels

to an interesting situation. There is a correlation between the locus of stress and the number of theme vowel distinctions: the forms that distinguish 3 theme vowels are those that have stress on the theme vowel position (or on the post-thematic in the case of the future and the conditional). These are therefore also those that have an unstressed prethematic vowel. As a result, the set of pairs of cells in which uncertainty is introduced by theme vowels doesn't overlap with the set of cell pairs in which uncertainty is introduced by alternating prethematic vowels.

This situation has implications for the possibility of defining a system of principal parts for the conjugation of European Portuguese. Following Finkel and Stump (2007), a system of (static) principal parts is defined as a set of cells in the paradigm from which all other cells can be categorically inferred. Ideally, an inflectional system has a single principal part (see Albright's (2002) Uniform Base Hypothesis). This single principal part is a good candidate to be used as a citation form, as its knowledge is sufficient to derive the whole paradigm. In practice, however, as Stump and Finkel (2013) show in detail, it is rare that a single principal part is sufficient, if only because some irregular lexemes have unpredictable inflection.

In European Portuguese, the interaction between conjugation neutralisation and vowel reduction has the consequence that even if we limit ourselves to the subsystem of regular verbs of the three traditional conjugations, at least two principal parts are necessary: any form with stress on the prethematic vowel will be a poor predictor of forms with stress on the theme or post-thematic vowel, and vice versa.

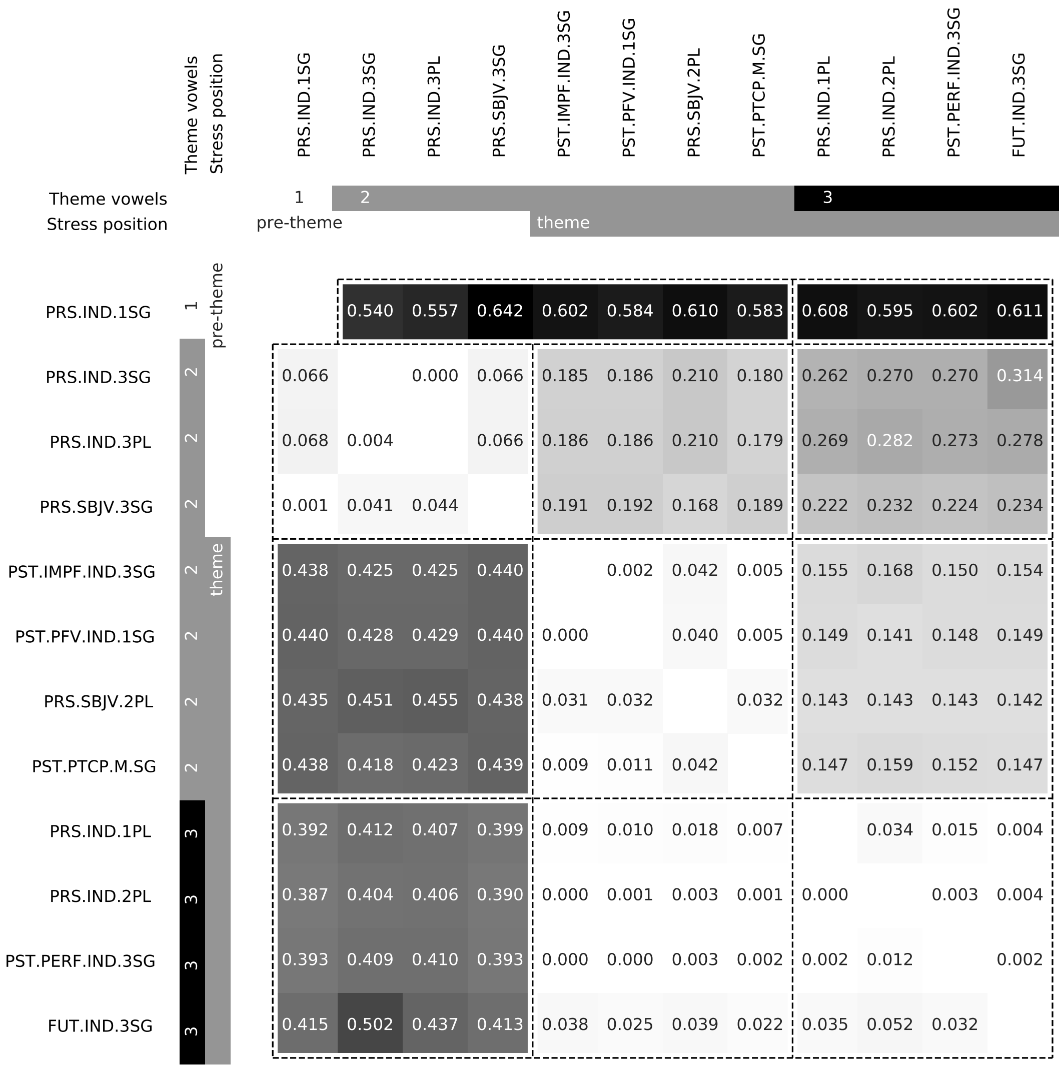
4.5 *Local irregularities*

We have now identified the two main sources of unpredictability in the system. This is shown visually in Figure 5, where rows and columns have been sorted by both stress placement and number of contrasts in theme vowels, and areas where identified determinants of predictability have the same effect have been materialised by dashed lines. It is striking here that, as expected, we have very low entropy (lower than 0.1) for pairs of cells in the four areas where (i) there is no contrast in stress placement between predictor and predictee, and (ii) the predictor exhibits as many or more theme vowel contrasts than the predictee.

What has been left unaccounted for is the remaining variability within each of the areas: if we had identified all sources of unpredictability, the numbers within an area would all be exactly the same, and all four whiteish areas would contain only zeroes.

A clear remaining source of uncertainty is the existence of irregular allomorphy. Once again, the general situation can be illustrated by taking examples from the present indicative. Consider the pair relationship between the 1PL and the 2PL. All patterns encountered are shown in Table 23. We see that there is a small set of 63 verbs that do not follow any of the three regular patterns of the three conjugations. It is natural to interpret this situation as a case of stem allomorphy: the strategy that uses the same stem for the 1PL and the 2PL is in opposition to two other strategies, which use the 2PL with a stem augment /də/, with or without nasalisation of the theme vowel. The high frequency irregulars SER and IR present isolated alternations.

This situation of stem allomorphy leads to some uncertainty, insofar as some patterns of alternation are applicable in the same phonological contexts. Table 24 lists these situ-



	Theme vowels	Stress position	PRS.IND.1SG	PRS.IND.3SG	PRS.IND.3PL	PRS.SBJV.3SG	PST.IMPF.IND.3SG	PST.PFV.IND.1SG	PRS.SBJV.2PL	PST.PTCP.M.SG	PRS.IND.1PL	PRS.IND.2PL	PST.PERF.IND.3SG	FUT.IND.3SG
Theme vowels			1	2							3			
Stress position			pre-theme				theme							
PRS.IND.1SG	1	pre-theme		0.540	0.557	0.642	0.602	0.584	0.610	0.583	0.608	0.595	0.602	0.611
PRS.IND.3SG	2		0.066		0.000	0.066	0.185	0.186	0.210	0.180	0.262	0.270	0.270	0.314
PRS.IND.3PL	2		0.068	0.004		0.066	0.186	0.186	0.210	0.179	0.269	0.282	0.273	0.278
PRS.SBJV.3SG	2		0.001	0.041	0.044		0.191	0.192	0.168	0.189	0.222	0.232	0.224	0.234
PST.IMPF.IND.3SG	2	theme	0.438	0.425	0.425	0.440		0.002	0.042	0.005	0.155	0.168	0.150	0.154
PST.PFV.IND.1SG	2		0.440	0.428	0.429	0.440	0.000		0.040	0.005	0.149	0.141	0.148	0.149
PRS.SBJV.2PL	2		0.435	0.451	0.455	0.438	0.031	0.032		0.032	0.143	0.143	0.143	0.142
PST.PTCP.M.SG	2		0.438	0.418	0.423	0.439	0.009	0.011	0.042		0.147	0.159	0.152	0.147
PRS.IND.1PL	3		0.392	0.412	0.407	0.399	0.009	0.010	0.018	0.007		0.034	0.015	0.004
PRS.IND.2PL	3		0.387	0.404	0.406	0.390	0.000	0.001	0.003	0.001	0.000		0.003	0.004
PST.PERF.IND.3SG	3		0.393	0.409	0.410	0.393	0.000	0.000	0.003	0.002	0.002	0.012		0.002
FUT.IND.3SG	3		0.415	0.502	0.437	0.413	0.038	0.025	0.039	0.022	0.035	0.052	0.032	

Figure 5: Combined sources of low predictability

ations. However, the relevant classes (c_3, c_6, c_7, c_8, c_9) make up only 10% of the lexicon, and the distribution of patterns within these classes is highly skewed. Remarkably, stem allomorphy does not lead to uncertainty in the other direction, because the form of the shorter stem is entirely predictable on the basis of that of the longer one.

The prediction of PRS.IND.3SG from PRS.IND.3PL illustrates a different type of local irregularity, this time due to irregular endings and not to stem allomorphy. The verbs of the 2nd and 3rd conjugation normally end in /ɐ̃ĩ/ in the 3PL and /ə/ in the 3SG. However, there are some verbs that do not have the expected ending in the 3SG, which we have already noted when commenting on Table 3. In most cases there is no uncertainty, because the class of verbs concerned has a distinctive phonological characteristic. For instance, all and only verbs with a 3PL ending in /jɐ̃ĩ/ have a 3SG form in /j/ instead of the expected /jə/; see e.g. VIR in Table 3. Nonetheless, the small set of unpredictable endings in the 3SG does contribute some very limited uncertainty (conditional entropy 0.007).

While commenting on the causes of each of the entropy values found for the 132 pairs of cells in the distillation individually may offer more insights about predictability,

Pattern	Frequency	Example
ˈɐmu ⇌ ˈaj / X+_ʃ	4168	ESTAR ‘be’
mu ⇌ / X+ˈi_ʃ	385	CONSEGUIR ‘obtain’
ˈemu ⇌ ˈɐj / X+_ʃ	375	FAZER ‘do’
[-nas]mu ⇌ [+nas]də / X*[-son -dors]_ʃ	48	TER ‘have’
mu ⇌ də / X*{l,v,z,ɾ,ʀ,ʎ,ʒ} {ˈe,ˈi}_ʃ	13	VER ‘see’
ˈomu ⇌ ˈoj / s_ʃ	1	SER ‘be’
vˈɐmu ⇌ ˈidə / _ʃ	1	IR ‘go’

Table 23: List of patterns of alternation relating PRS.IND.1PL and PRS.IND.2PL

Class	Size	Sample lexeme	Pattern	Example forms	Count
c_3	250	PODER	ˈemu ⇌ ˈɐj / X+_ʃ	pudˈemuʃ ⇌ pudˈɐjʃ	240
		TER	[-nas]mu ⇌ [+nas]də / X*[-son -dors]_ʃ	tˈemuʃ ⇌ tˈẽdəʃ	10
c_6	100	FAZER	ˈemu ⇌ ˈɐj / X+_ʃ	fɐzˈemuʃ ⇌ fɐzˈɐjʃ	93
		VER	mu ⇌ də / X*{l,v,z,ɾ,ʀ,ʎ,ʒ} {ˈe,ˈi}_ʃ	vˈemuʃ ⇌ vˈedəʃ	7
			[-nas]mu ⇌ [+nas]də / X*[-son -dors]_ʃ		0
c_7	80	SURGIR	mu ⇌ / X+ˈi_ʃ	suɾʒˈimuʃ ⇌ suɾʒˈiʃ	71
		VIR	[-nas]mu ⇌ [+nas]də / X*[-son -dors]_ʃ	vˈimuʃ ⇌ vˈĩdəʃ	9
			mu ⇌ də / X*{l,v,z,ɾ,ʀ,ʎ,ʒ} {ˈe,ˈi}_ʃ		0
c_8	69	REFERIR	mu ⇌ / X+ˈi_ʃ	ʀəfəɾˈimuʃ ⇌ ʀəfəɾˈiʃ	67
		RIR	mu ⇌ də / X*{l,v,z,ɾ,ʀ,ʎ,ʒ} {ˈe,ˈi}_ʃ	ʀˈimuʃ ⇌ ʀˈidəʃ	2
c_9	31	QUERER	ˈemu ⇌ ˈɐj / X+_ʃ	kəɾˈemuʃ ⇌ kəɾˈɐjʃ	27
		LER	mu ⇌ də / X*{l,v,z,ɾ,ʀ,ʎ,ʒ} {ˈe,ˈi}_ʃ	lˈemuʃ ⇌ lˈedəʃ	4

Table 24: Classes of PRS.IND.1PL relevant for the prediction of PRS.IND.2PL in which verbs can belong to either of two patterns.

it would extend our paper significantly. We believe we have illustrated the kind of factors that can determine entropy, and that other pairs of cells implement various combinations of the factors already identified.

5. Conclusions

This paper had both methodological and empirical objectives. On the methodological side, we presented an algorithmic solution to address the Paradigm Cell Filling Problem strictly inductively. The Qumín package takes as input raw lexical data with no morphological information other than an organization into lexemes and paradigm cell. Provided with information on the interpretation of phonemic symbols in terms of distinctive features, it allows for the automatic extraction of patterns of alternation between forms. Important design features of the algorithm are that it assumes a very general format for alternations, allowing for multiple points of variation, and does not encode any bias as to the direction of alignment between forms—hence any mode of affixation (prefixation, suffixation, circumfixation, infixation, or even root-and-pattern) can be discovered from the data automatically. In the present study, these features are crucial to capturing the interplay of suffixation, theme vowel alternations, and stress-conditioned vowel alternations that is the hallmark of European Portuguese conjugation.

The classification of pairs of forms in terms of alternation patterns was then used to infer relevant probability distributions describing dependencies between the shape found in some predictor cell and possible shapes in another predicted cell. We used these probability distributions to compute conditional entropy values assessing the average predictability of one cell from another, but insisted that attention to the full distribution is necessary to get a more fine-grained picture of the system. This proved very useful in our exploration of the sources of (un)predictability in European Portuguese conjugation: detailed examination of the distribution of patterns allowed us to go beyond a broad assessment of the difficulty of the Paradigm Cell Filling Problem and into a detailed discussion of the specific challenges posed by the fine structure of the European Portuguese system.

Moving on to empirical results, we showed that the predictive structure of European Portuguese verbal paradigms is conditioned by four factors which each lead to situations of opacity in the prediction of one cell from another. First, cells that neutralize the three-way distinction between theme vowels to a two-way or no distinction at all are poor predictors, because the speaker has to guess what the theme vowel is. Second, neutralizations of vocalic contrasts in unstressed positions lead to forms with an unstressed prethematic vowel being poor predictors of the stressed variant found in other forms. Third, although they are less prevalent than in other Romance languages, stem alternations do lead to unpredictability, when there is uncertainty as to whether the form in the unknown cell uses the same stem as the predictor form or a different stem. Fourth, irregular suffixal exponence is a minor source of unpredictability.

There are many directions in which the research reported here can and should be extended. In the area of Romance morphology, the last three decades have witnessed an exhuberance of research on stem alternations, both in terms of detailed descriptions of synchronic systems (see e.g. Bonami and Boyé 2003 on French, Boyé and Cabredo Hofherr 2006b on Spanish, Guerrero 2011 on Catalan, Montermini and Bonami 2013 on Italian) and of diachronic work on the emergence and maintenance of such systems (see Maiden

2018 for a recent and detailed assessment and guide to the literature). Stem alternations are an important contributor to unpredictability, but, as the present paper shows, by no means the only one. In addition, as Bonami and Boyé's (2014) self-criticism highlights, large-scale work on stem alternations forces the analyst into poorly justifiable segmentation decisions. The approach illustrated here avoids such decisions entirely, and hence would allow for a large-scale reassessment of predictability in Romance conjugation on a quantitative basis.

From a theoretical standpoint, it is worth noting that the present paper adopts a somewhat ambiguous posture. On the one hand, the research is grounded in a purely inductive, word-based approach to morphological structure, where no assumption is made as to a stable segmentation of words in subword constructs such as stems, theme vowels, and affixes. On the other hand, these concepts were crucial to our investigation of the factors influencing predictability, where we reverted to the very traditional descriptive vocabulary. We do not think there is a contradiction here, although there is a gap in current research that needs to be filled. The research tradition this paper participates in highlights the importance of an approach to the relations between forms in a paradigm that is both quantitative and inductive; and we insist that, when addressing prediction from form to form, we should avoid basing such predictions on a preconceived segmentation that speakers trying to solve the Paradigm Cell Filling Problem cannot be aware of. But that in no way entails that there is no other aspect to morphological structure. Quite to the contrary, we assume that the Paradigm Cell Filling Problem is paralleled by the *Inflected Word Recognition Problem*, the problem speakers face when trying to assess, from the shape of an unknown word, which paradigm cell of which lexeme that word belongs to (Bonami and Beniamine, 2021). We submit that traditional subword units capture some of those properties of words that play a role in addressing this problem, but do so in an approximate fashion, by insisting on discrete boundaries between exponents. To avoid this, such properties could be investigated with the same type of inductive, quantitative methods deployed in this paper. This still needs to be demonstrated in practice though. A full picture of the structure of European Portuguese conjugation awaits that demonstration.

Acknowledgements

This paper grew out of previous research on Portuguese conjugation by Bonami and Luís (2014) and the computational tools for studying implicative structure by Beniamine (2018). It could not have seen the light of day without the help of Fernando Perdigão, who provided most of the primary data we worked from and painstakingly checked part of it by hand. We also thank Maria Copot for her help translating Bonami and Luís (2014) into English, which was the initial step leading to the present project. Finally, we thank our three anonymous reviewers as well as our editor at Isogloss for insightful remarks that led to significant improvements. Any remaining errors are our own responsibility. This work was partially funded by a public grant overseen by the French National Research Agency (ANR) as part of the "Investissements d'Avenir" program (reference: ANR-10-LABX-0083) and by a British Academy International Newton Fellowship (reference: NIF23\100218).

References

Ackerman, Farrell, James P. Blevins, and Robert Malouf (2009). "Parts and wholes: implicative patterns in inflectional paradigms". In: *Analogy in Grammar*. Ed. by James P. Blevins and Juliette Blevins. Oxford: Oxford University Press, pp. 54–82.

Ackerman, Farrell and Robert Malouf (2013). "Morphological organization: the low conditional entropy conjecture". In: *Language* 89, pp. 429–464.

Albright, Adam (2002). "The Identification of Bases in Morphological Paradigms". PhD thesis. University of California, Los Angeles.

Albright, Adam and Bruce Hayes (2002). "Modeling English Past Tense Intuitions with Minimal Generalization". In: *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning - Volume 6*. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 58–69. doi: 10.3115/1118647.1118654.

— (2006). "Modeling productivity with the Gradual Learning Algorithm: the problem of accidentally exceptionless generalizations". In: *Gradience in Grammar: Generative Perspectives*. Ed. by Gisbert Fanselow et al. Oxford: Oxford University Press, pp. 185–204.

Aronoff, Mark (1994). *Morphology by itself.* Cambridge: MIT Press.

Beniamine, Sacha (2018). "Typologie quantitative des systèmes de classes flexionnelles". PhD thesis. Université Paris Diderot.

Blevins, James P. (2006). "Word-based morphology". In: *Journal of Linguistics* 42, pp. 531–573.

Bonami, Olivier and Sacha Beniamine (2016). "Joint predictiveness in inflectional paradigms". In: *Word Structure* 9.2, pp. 156–182.

— (2021). "Leaving the stem by itself". In: *All Things Morphology*. Ed. by Marcia Haag et al. Amsterdam: John Benjamins, pp. 81–98.

Bonami, Olivier and Gilles Boyé (2002). "Suppletion and stem dependency in inflectional morphology". In: *The Proceedings of the HPSG '01 Conference*. Ed. by Franck Van Eynde, Lars Hellan, and Dorothee Beerman. Stanford: CSLI Publications, pp. 51–70.

— (2003). "Supplétion et classes flexionnelles dans la conjugaison du français". In: *Langages* 152, pp. 102–126.

— (2014). "De formes en thèmes". In: *Foisonnements morphologiques. Etudes en hommage à Françoise Kerleroux*. Ed. by Florence Villoing, Sarah Leroy, and Sophie David. Presses Universitaires de Paris Ouest, pp. 17–45.

Bonami, Olivier and Ana R. Luís (2014). "Sur la morphologie implicative dans la conjugaison du portugais : une étude quantitative". In: *Morphologie flexionnelle et dialectologie romane. Typologie(s) et modélisation(s).* Ed. by Jean-Léonard Léonard. Mémoires de la Société de Linguistique de Paris 22. Leuven: Peeters, pp. 111–151.

Boyé, Gilles (2000). "Problèmes de morpho-phonologie verbale en français, espagnol et italien". PhD thesis. Université Paris 7.

Boyé, Gilles and Patricia Cabredo Hofherr (2006a). "The structure of allomorphy in Spanish verbal inflection". In: *Cuadernos de Lingüística*. Vol. 13. Instituto Universitario Ortega y Gasset, pp. 9–24.

— (2006b). "The structure of allomorphy in spanish verbal inflection." In: *Cuadernos de Lingüística* 13, pp. 9–24.

Boyé, Gilles and Gauvain Schalchli (2019). “Realistic data and paradigms: the paradigm cell finding problem”. In: *Morphology* 29, pp. 199–248.

Brown, Dunstan (2007). “Peripheral functions and overdifferentiation: The Russian second locative”. In: *Russian Linguistics* 31, pp. 61–76.

Candeias, Sara, Arlindo Veiga, and Fernando Perdigão (2015). *Pronunciação de Verbos Portugueses - Guia Prático*. LIDEL.

Carstairs-McCarthy, Andrew (1994). “Inflection Classes, Gender, and the Principle of Contrast”. In: *Language* 70, pp. 737–788.

Chomsky, Noam and Morris Halle (1968). *The sound pattern of English*. Harper and Row.

Corbett, Greville G. (2007). “Canonical typology, suppletion and possible words”. In: *Language* 83, pp. 8–42.

Cotterell, Ryan et al. (2019). “On the Complexity and Typology of Inflectional Morphological Systems”. In: *Transactions of the Association for Computational Linguistics* 7, pp. 327–342.

Finkel, Raphael and Gregory T. Stump (2007). “Principal parts and morphological typology”. In: *Morphology* 17, pp. 39–75.

Frisch, Stefan A., Janet B. Pierrehumbert, and Michael B. Broe (2004). “Similarity avoidance and the OCP”. In: *Natural Language and Linguistic Theory* 22, pp. 179–228.

Greenberg, Joseph H. (1954). “A quantitative approach to the morphological typology of language”. In: *Method and Perspective in Anthropology: Papers in Honor of Wilson D. Wallis*. Ed. by Robert F. Spencer. University of Minnesota Press.

Guerrero, Aurélie (2011). “Verbal inflection in Central Catalan: a realisational analysis”. In: *Lingue e linguaggio* 10, pp. 265–282.

Guzman Naranjo, Matías (2019). *Analogical classification in formal grammar*. Berlin: Language Science Press.

Guzmán Naranjo, Matías (2020). “Analogy, complexity and predictability in the Russian nominal inflection system”. In: *Morphology* 30.3, pp. 219–262.

Maiden, Martin (1992). “Irregularity as a determinant of morphological change”. In: *Journal of Linguistics* 28, pp. 285–312.

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

Mansfield, John (2016). “Intersecting formatives and inflectional predictability: How do speakers and learners predict the correct form of Murrinhpatha verbs?” In: *Word Structure* 9, pp. 183–214.

Marquiafável, Vanessa et al. (Oct. 2014). “Rule-Based Algorithms for Automatic Pronunciation of Portuguese Verbal Inflections”. In: *International Conf. on Computational Processing of Portuguese - PROPOR*. Vol. 8775, pp. 36–47. DOI: 10.1007/978-3-319-09761-9_4.

Mateus, Maria Helena and Ernesto d’Andrade (2000). *ThePhonology of Portuguese*. Oxford: Oxford University Press.

Montermini, Fabio and Olivier Bonami (2013). “Stem Spaces and Predictibility in Verbal Inflection”. In: *Lingue e Linguaggio* 12, pp. 171–190.

Needleman, Saul B. and Christian D. Wunsch (1970). “A general method applicable to the search for similarities in the amino acid sequence of two proteins”. In: *Journal of Molecular Biology* 48.3, pp. 443–453. ISSN: 0022-2836. DOI: 10.1016/0022-

2836(70)90057-4. URL: http://www.sciencedirect.com/science/article/pii/0022283670900574.

Pellegrini, Matteo (2021). “Patterns of interpredictability and principal parts in Latin verb paradigms: an entropy-based approach”. In: *Journal of Latin Linguistics* 20.1.

Pirrelli, Vito and Marco Battista (2000). “The Paradigmatic Dimension of Stem Allomorphy in Italian Verb Inflection”. In: *Rivista di Linguistica* 12.

Santos, Diana and Paulo Rocha (July 2001). “Evaluating CETEMPúblico, a Free Resource for Portuguese”. In: *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*. Toulouse, France: Association for Computational Linguistics, pp. 450–457. DOI: 10.3115/1073012.1073070. URL: https://aclanthology.org/P01-1058.

Sims, Andrea (2015). *Inflectional defectiveness*. Cambridge: Cambridge University Press.

Stump, Gregory T. (2001). *Inflectional Morphology. A Theory of Paradigm Structure*. Cambridge Studies in Linguistics 93. Cambridge: Cambridge University Press.

Stump, Gregory T. and Raphael Finkel (2013). *Morphological Typology: From Word to Paradigm*. Cambridge: Cambridge University Press.

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

Vigário, Marina (2003). *The Prosodic Word in European Portuguese*. Berlin, Boston: De Gruyter Mouton. ISBN: 978-3-11-090092-7. DOI: 10.1515/9783110900927.

Wilmoth, Sasha and John Mansfield (2021). “Inflectional predictability and prosodic morphology in Pitjantjatjara and Yankunytjatjara”. In: *Morphology* 31.4. DOI: 10.1007/s11525-021-09380-y.

Wurzel, Wolfgang Ulrich (1984). *Flexionsmorphologie und Natürlichkeit. Ein Beitrag zur morphologischen Theoriebildung*. Translated as **Wurzel89**. Berlin: Akademie-Verlag.