Joint predictiveness in inflectional paradigms

Olivier Bonami & Sacha Beniamine
Université Paris Diderot

1 Introduction

Inflectional paradigms[1] have what Wurzel (1984/1989) calls IMPLICATIVE STRUCTURE:

> The inflectional paradigms are, as it were, kept together by implications. There are no paradigms (except highly extreme cases of suppletion) that are not based on implications valid beyond the individual word, so that we are quite justified in saying that inflectional paradigms generally have an implicative structure, regardless of deviations in the individual cases.
>
> Wurzel (1989, 114)

While Wurzel does not give a precise definition of the notion of implication at hand, the examples he discusses are simple enough to clarify what he has in mind. Wurzel's implications take the form of conditional statements whose antecedent mentions phonological characteristics of one or more paradigm cells and the consequent phonological caracteristic of a single, disjoint cell. Examples for English conjugation are given in (1).

(1)
- (a) If X is the base form of a verb, then its present participle form is Xɪŋ.
- (b) If X is the base form of a verb, then its past form is Xd.
- (c) If the past form of a verb is Xd, then its past participle form is Xd.
- (d) If the base form of a verb is Xɪŋ and its past form is Xæŋ, then its past participle form is Xʌŋ.
- (e) If the base form of a verb is Xɪŋ and its past participle form is Xʌŋ, then its past form is Xæŋ.

The statements in (1) differ in various ways. (1a) and (1b) have maximally general antecedents and hence are relevant for all verbs, while (1c) and (1d) have more specific antecedents and hence are relevant only to a subset of verbs. We define the COVERAGE of an implicative statement as the proportion of lexemes to which the statement applies non-trivially, because the lexeme satisfies the statement's antecedent: hence (1a) and (1b) have a coverage of 1, but (1c), (1d) and (1e) have smaller coverage. (1a) and (1d) are categorically true statements about English conjugation: all lexemes that satisfy the

[1] This paper builds heavily on previous work by the first author in collaboration with Gilles Boyé, Ana R. Luís and Delphine Tribout, whose help is gratefully acknowledged; and on preliminary work he presented at the 8th *Décembrettes* conference (Bordeaux, 2012). The research reported here was presented at the final NetWordS conference (Pisa, 2015) and at a workshop at Université Paris Diderot (2015). We thank for their comments the audiences at these events, and notably Jim Blevins, Vito Pirrelli, Ingo Plag, and Benoît Sagot. We also benefited from various discussions with Farrell Ackerman and Rob Malouf over the last few years. This work was partially supported 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).

antecedent also satisfy the consequent. By contrast, (1b), (1c) and (1e) are not categorically true (consider exceptions such as SING for (1b), SHAVE for (1c), STRING for (1e)). We define the ACCURACY of an implicative statement as the proportion of lexemes satisfying the consequent among those that satisfy the antecedent; an implication is categorical if it has an accuracy of exactly 1. Finally, (1a), (1b), and (1c) mention one paradigm cell in their antecedent, while (1d) and (1e) mention two. We call statements such as (1a) UNARY implications, and statements such as (1d) BINARY implications; in general, we call n-ARY an implication mentioning n paradigm cells in its antecedent.

Recent years have witnessed a rise of interest in the study and modelling of the implicative structure of paradigms, with two main trends.[2] One trend has been concerned with the PARADIGM CELL FILLING PROBLEM (henceforth PCFP;Ackerman et al. 2009; Ackerman & Malouf 2013; Bonami & Boyé 2014; Bonami & Luís 2014; Sims 2015):

> What licenses reliable inferences about the inflected (and derived) surface forms of a lexical item?
>
> Ackerman et al. (2009, 54)

Research in this area has focused on assessing quantitatively the predictive value of paradigm cells using information-theoretic measures. Because gradient predictivity is the focus, non-categorical implications are central to the enterprise. On the other hand, most work centers on unary implications: while the existence of nontrivial binary implications was noted as early as Thymé et al. (1994), no study addresses them heads-on.

Another important trend is found in the work of Raphael Finkel and Greg Stump on formalizing and deploying the notion of a SYSTEM OF PRINCIPAL PARTS (Finkel & Stump, 2007, 2009; Stump & Finkel, 2013). Principal part systems originate in language pedagogy. In classical terms,[3] a system of principal parts is a set of paradigm cells such that for all lexemes, every other cell in the paradigm can be deduced with certainty from knowledge of that lexeme's principal parts.[4] There is a clear relation between systems of principal parts and paradigmatic implications: for any inflection system, $C = \{c_1, \dots c_n\}$ is a system of principal parts only if there exists a collection of categorical n-ary implications relating those paradigm cells in C with those paradigm cells that are not in C. In Finkel and Stump's view, principal part systems should be categorical to be of pedagogical value: they should help the language learner by focusing on memorization of the minimal information leading to certainty about the inflection of a lexeme.

The present paper is an attempt to bridge the gap between these two lines of research. Our goal is to assess the importance of the joint knowledge of multiple paradigm cells in addressing the PCFP. After justifying the relevance of joint predictiveness in section 2, we propose in section 3 a strategy for applying the type of information-theoretic measure of predictiveness used by Ackerman et al. (2009) and subsequent work in the case of n-ary implicative relations. We then apply this strategy in section 4 to two large datasets

[2] This followed 15 years of enthusiasm over the study of stem allomorphy, which very often took the form of families of implicative statements over stem alternants (see among others Maiden (1992); Aronoff (1994); Brown (1998); Hippisley (1998); Bonami & Boyé (2002); Blevins (2003), and the papers collected in Bonami (2012)). Blevins (2005, 2006) must be credited for shifting the perspective from stems to words. See Montermini & Bonami (2013) for a comparison of stem-based and word-based implicative approaches.

Albright (2002); Albright & Hayes (2003) present an early proposal for a quantitative study of implicative relations between words from a different perspective that had a strong influence on the work reported here.

[3] Finkel and Stump distinguish three conceptualizations of principal part systems: STATIC, ADAPTIVE and DYNAMIC systems. We focus here on static systems, which correspond to the traditional notion.

[4] This is not to say that principal part systems used in practice in pedagogical grammars actually match Finkel and Stump's ideals: very often, their predictive value does not extend to the most frequent, highly irregular lexemes.

of European French and Portuguese, and show that on average knowledge of multiple paradigm cells is dramatically more predictive than knowledge of a single cell. Finally, we discuss the consequences of these results for the study of principal part systems.

2 The relevance of joint predictiveness

In this section we present evidence that joint predictiveness is relevant for speakers. The evidence comes in three forms. First, we argue from the statistical distribution of inflected forms in corpora that speakers typically have access to multiple forms of lexemes without being exposed to the full paradigm: hence, joint knowledge of multiple forms is possible, and may be used to infer appropriate unknown forms. Second, we present linguistic evidence that there are important generalizations to be drawn from joint knowledge of multiple forms. Third and finally, we argue that there is limited but convincing behavioral evidence that speakers actually do build on their joint knowledge of multiple forms of the same lexeme to decide what an unknown form should be.

2.1 Corpus evidence

In this section, we use corpus evidence to assess whether the PCFP is indeed a concrete problem for speakers, and whether speakers have access to knowledge of multiple paradigm cells to address it.

Chan (2008, chap. 4) presents relevant evidence shown in Table 1. Chan defines the paradigm SATURATION of a corpus for a particular part of speech as the proportion of the paradigm that is realized in the corpus for the lexeme which maximizes that proportion. The table clearly establishes that in languages with sizable paradigms, at corpus sizes of the order of a million word, no lexeme is attested in all its forms. Chan argues that this is because morphological data is doubly sparse: not only do lexemes have a Zipfian distribution, with few lexemes being very frequent and many lexemes very infrequent, but the distribution of each lexeme's forms is highly skewed, with some paradigm cells being much more frequent than others.

Corpus	size (in millions)	max. number of distinct forms	saturation
English	1.2	6	1.000
Swedish	1.0	21	0.667
Basque	0.6	22	0.727
Slovene	2.4	32	0.750
Hebrew	2.5	33	0.697
Catalan	1.7	45	0.733
Spanish	2.6	51	0.667
Italian	1.4	55	0.855
Czech	2.0	72	0.569
Hungarian	1.2	76	0.632
Greek	2.8	83	0.542
Finnish	2.1	365	0.403

Tab. 1: Paradigm saturation for verbs in a collection of corpora (adapted from (Chan, 2008, 79))

While Chan's study is useful, it does not quite establish the reality of the PCFP. First, 0.6 to 2.8 million word tokens are rather small figures. A recent estimation (Gilkerson

& Richards, 2009)[5] shows that American children below the age of 4 are exposed to a daily average of about 6,000 (10th percentile) to 20,000 (90th percentile) word tokens of adult speech. Thus even children with low linguistic exposure will have heard about 9 million tokens by age 4, which is considerably more than the size of Chan's datasets: it might be that saturation gets reached for all languages if more realistically-sized datasets are considered. Second, saturation is a rather unsubtle measure, since it focuses only on the one most diversely used lexeme: it might tell us very little on central tendency. Third, showing that speakers have only partial knowledge of paradigms does not entail that they need to predict unknown cells: in many languages, some paradigm cells are so rarely used that speakers are unlikely to need them.

To compensate for this, we conducted a study on the basis of *FrWaC* (Baroni et al., 2009), a web corpus of 1.6 billion word tokens. Note that at this corpus size, we are clearly dealing with a quantity of data commensurate to the exposure of an adult speaker.[6] Because the corpus contains numerous segmentation and tagging errors, we restricted our attention to verbs documented in the *Lefff* lexicon (Sagot, 2010); as a consequence, recent neologisms are not taken into consideration.

Figure 1 shows how the average number of distinct forms per lexeme evolves when we progressively walk through the corpus. Note that we are counting distinct orthographic forms rather than distinct paradigm cells. French verbs have 51 paradigm cells, but because of widespread syncretism, there are only about 36 distinct forms per verb on average documented in the *Lefff* lexicon.[7] Two observations are in order. First, as the size of the corpus grows, the average number of forms per lexeme rises. Second, even at very large corpus sizes, the average number of forms per lexemes attested in the corpus is well below the maximal number of such forms.[8] These two observations strongly suggest that speakers do not saturate their vocabulary of useful inflected forms, and hence are likely to occasionally need to produce unknown forms of a known lexeme.

Figure 2 shows the proportion of lexemes that are found in varying numbers of forms in the corpus, as the size of the corpus grows. The number of lexemes found it at least two forms reaches a maximal value of more than 90% at a corpus size of about 100 million tokens, and then stays constant. This indicates that speakers are massively exposed to multiple forms of most lexemes: their linguistic experience gives them the means of predicting from multiple forms of the same lexeme. The same observation holds for larger sets of forms, although both the growth rate and maximal value decrease with the number of forms.

In closing, it should be noted that Blevins et al. (inpress) report slightly different results on the basis of a study of German nouns in the *SdeWaC* corpus (Faaß& Eckart, 2013): in their study, the average number of forms per lexeme actually decreases when

[5] Gilkerson & Richards's study improves on classical results by Hart & Risley (1995), being based on day-long recordings of both child-directed speech and speech overheard by children. Our thanks to Henny Yeung for pointing us to the LENA study.

[6] Let us suppose for the sake of the argument that an adult engaging in high quantities of linguistic interaction is exposed to 10 times more data than a child in Gilkerson & Richards's (2009) 90th percentile. Under this assumption, 1.6 billion words correspond to about 22 years of linguistic exposure.

[7] There is currently no morphological tagger for French that reliably disambiguates homographic forms of the same lexeme and that would allow us to make the same calculations for paradigm cells. Note that syncretism affects different inflection classes in different ways, leading to inter-lexeme variation in the number of distinct forms.

[8] Note that the exclusion of undocumented verbs is likely to lead to an overestimation of the average number of forms per lexemes, especially at large corpus sizes. Also note that there are 12 paradigm cells that have more or less fallen out of usage in contemporary French, i.e. the indicative simple past and the past subjunctive. One may argue that they should be excluded from the calculations. Even if they are, one still ends up with an average of only about 18 attested forms per lexeme out of 24.

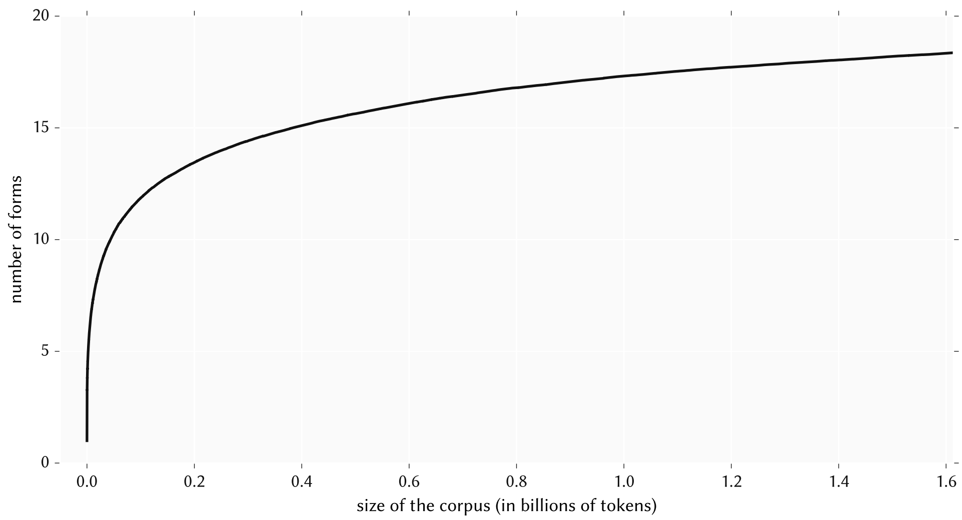

Fig. 1: Average number of forms per lexeme as a function of vocabulary size in the *FrWaC* corpus.

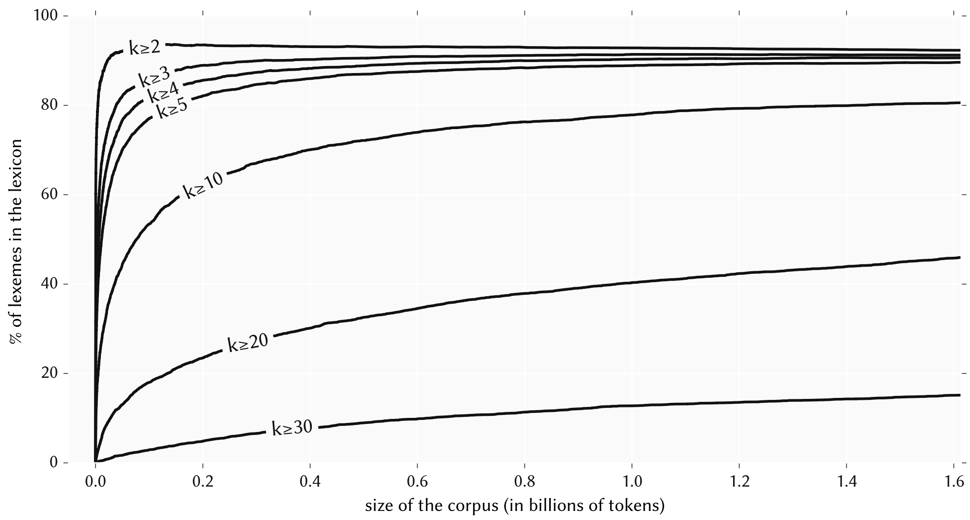

Fig. 2: Proportion of lexemes attested in at least k forms as a function of vocabulary size in the *FrWaC* corpus, for various values of k.

corpus size increases.[9] We can only speculate that this contrast is due to a combination of differences in the actual inflection system (in particular the very high prevalence of syncretism in German nominal paradigms) and differences in methodology.[10] Be that as it may, the results of their study are in no way contradictory with our conclusion that, in large corpora, many paradigms both contain more than one form and are incomplete.

[9] Blevins et al. (inpress) report sample sizes in numbers of types rather than tokens, and start at a corpus size of 1 million types.

[10] *SdeWaC* is the subset of the Wacky corpus for German that is parsable by the IMS Stutgart's FSPar parser. This amounts to a type of normalization that is quite different from filtering through a lexicon, presumably leading to higher recall but lower precision in the identification of well-formed words of a given part of speech.

2.2 Linguistic evidence

The previous subsection showed that speakers do have access to multiple forms of lexemes whose full paradigm they don't know. This does not entail, though, that this information is useful: it might be that knowledge of a second form of a lexeme does not improve the likelihood of being able to infer correctly an unknown form. It is the main goal of this paper to give quantitative evidence that joint knowledge does improve quality of prediction. In this subsection, we give preliminary evidence on the existence of relevant generalizations in select datasets.

A telling type of situation occurs when joint knowledge allows for categorical prediction when knowledge of just one form does not. Consider the dataset in Table 2, which illustrates the major alternation patterns relating the infinitive, present indicative, and past participle in French verbs.[11]

Class	Size	Sample lexeme	INF	PRS.3SG	PRS.3PL	PST.PTCP
(i)	4108	LIVRER 'deliver'	livʁe	livʁ	livʁ	livʁe
(ii)	210	RELIER 'link'	ʁəlje	ʁəli	ʁəli	ʁəlje
(iii)	22	RATISSER 'rake'	ʁatise	ʁatis	ʁatis	ʁatise
(iv)	327	BÂTIR 'build'	batiʁ	bati	batis	bati
(v)	37	TENIR 'hold'	təniʁ	tjɛ̃	tjɛn	təny
(vi)	8	OUVRIR 'open'	uvʁiʁ	uvʁ	uvʁ	uvɛʁ
(vii)	1	MOURIR 'die'	muʁiʁ	mœʁ	mœʁ	mɔʁ

Tab. 2: Exemplary subparadigms of French verbs

As this table illustrates, the past participle can be predicted categorically from the infinitive for verbs whose infinitives end in -e. However, things are not so simple for verbs with infinitives in -iʁ: while there is a dominant pattern corresponding to class (iv), the second conjugation in traditional descriptions of French, there are minority patterns of significant size. In the terms defined in the introduction, the implication in (2) has an accuracy of only 0.90.

(2) If a verb's infinitive is Xiʁ, then its past participle is Xi.

When considering other possible predictors of the past participle, it turns out that no cell gives rise to categorical prediction in all situations. In the present 3SG, the main difficulty comes from forms ending in -i, which could belong to class (ii), with a past participle in -je, or to class (iv), with a past participle in -i. In the present 3PL, it comes from forms ending in -is, which could belong to class (iii) and have a participle in -ise or to class (iv) and have a participle in -i. Overall, even when taking into account the remaining 47 paradigm cells not shown in this table, no single cell is a categorical predictor of the past participle.

Things are different however if we consider combinations of cells. Suppose a speaker has joint knowledge of the infinitive and present SG of a verb. If the infinitive ends in -e, then the past participle is certain to end in -e too. If it ends in -iʁ, then the present form disambiguates between classes (iv), (v), (vi) and (vii). In particular, the implication in

[11] Data from the *Flexique* database (Bonami et al., 2014). The patterns illustrated here correspond to what happens with 95% of the verbs. Excluded from consideration are verbs with infinitive endings other than -e or -iʁ or with stem-internal vowel alternations. All French and European Portuguese data is presented in phonemic transcription.

(3) singling out class (iv) verbs is categorical.[12]

(3) If a verb's infinitive is *X*iʁ and its present 3SG is *X*i, then its past participle is *X*i.

The preceding example shows that non-trivial joint prediction does occur in inflection systems. We now turn to an example from European Portuguese (originally due to Bonami & Luís 2014) that shows such non-trivial joint prediction to be systemic. Table 3 illustrates the contrasts between the three main conjugations for verbs with no stem-internal alternations.

	INF	1SG	2SG	3SG	1PL	2PL	3PL
FICAR 'become'	fiˈkaɾ	ˈfiku	ˈfikɐʃ	ˈfikɐ	fiˈkɐmuʃ	fiˈkaiʃ	ˈfikɐ̃ũ
VIVER 'live'	viˈveɾ	ˈvivu	ˈvivəʃ	ˈvivə	viˈvemuʃ	viˈvɐiʃ	ˈvivɐ̃ĩ
IMPRIMIR 'print'	ĩpriˈmiɾ	ĩˈprimu	ĩˈpriməʃ	ĩˈprimə	ĩpriˈmimuʃ	ĩpriˈmiʃ	ĩˈprimɐ̃ĩ

Tab. 3: Theme vowel alternations in European Portuguese (infinitive and present indicative)

The three conjugations are distinguished by contrasting thematic endings, which take the form of theme vowels a, e and i in the infinitive, and manifest themselves by other contrasts in the present 1PL and 2PL. In other paradigm cells however, these contrasts are partially or totally neutralized: there is no contrast between the second and third conjugation in the 2SG, 3SG and 3PL; and there is no contrast at all in the 1SG. These neutralizations between conjugations have clear implications for predictiveness: there can be no categorical prediction from a neutralized cell to a non-neutralized cell. The situation is particularly striking when trying to predict the infinitive from the present 1SG, as the present form gives no indication as to what theme vowel should be used in the infinitive.

Independently of this, European Portuguese is subject to a process of vowel reduction in unstressed syllables, giving rise to neutralizations of some vocalic contrasts, as indicated in (4).

(4) In unstressed syllables,
 (a) e and ɛ neutralize to ə.
 (b) a and ɐ neutralize to ɐ.
 (c) u, o and ɔ neutralize to u.

In conjugation, stress sometimes falls on the theme vowel, and sometimes on the last vowel of the stem, which we call the PRETHEMATIC VOWEL. This has the effect that prethematic vowels are subject to regular alternations, and that some contrasts are neutralized. This is illustrated with a sample of first conjugation verbs in Table 4: the vocalic contrasts manifest in the singular and 3PL forms are neutralized in the infinitive, 1PL and 3PL forms. Because of this, predicting a form with stress on the stem from a form with stress on the ending can be hard.

Now, the interesting observation is that neutralizations of conjugation class distinction and neutralizations of theme vowels affect complementary parts of the paradigm, because they are both dependent on stress placement: cells making a 3-way distinction

[12] With this particular dataset, joint knowledge of the infinitive and present 3PL or the two present forms also leads to categorical predictibility for the past participle. None of these pairs of cells is categorically predictive in the full French datasets however, because of opacities arising with verbs having infinitives in -Cʁ.

	INF	1SG	2SG	3SG	1PL	2PL	3PL
CHEGAR ‘reach’	ʃəˈgaɾ	ˈʃegu	ˈʃegɐʃ	ˈʃegɐ	ʃəˈgɐmuʃ	ʃəˈgaiʃ	ˈʃegɐ̃ũ
COMEÇAR ‘start’	kuməˈsaɾ	kuˈmɛsu	kuˈmɛsɐʃ	kuˈmɛsɐ	kuməˈsɐmuʃ	kuməˈsaiʃ	kuˈmɛsɐ̃ũ
PAGAR ‘pay’	pɐˈgaɾ	ˈpagu	ˈpagɐʃ	ˈpagɐ	pɐˈgɐmuʃ	pɐˈgaiʃ	ˈpagɐ̃ũ
CHAMAR ‘call’	ʃɐˈmaɾ	ˈʃɐmu	ˈʃɐmɐʃ	ˈʃɐmɐ	ʃɐˈmɐmuʃ	ʃɐˈmaiʃ	ˈʃɐmɐ̃ũ
RETOMAR ‘resume’	rətuˈmaɾ	rəˈtomu	rəˈtomɐʃ	rəˈtomɐ	rətuˈmɐmuʃ	rətuˈmaiʃ	rəˈtomɐ̃ũ
JOGAR ‘throw’	ʒuˈgaɾ	ˈʒɔgu	ˈʒɔgɐʃ	ˈʒɔgɐ	ʒuˈgɐmuʃ	ʒuˈgaiʃ	ˈʒɔgɐ̃ũ
MUDAR ‘change’	muˈdaɾ	ˈmudu	ˈmudɐʃ	ˈmudɐ	muˈdɐmuʃ	muˈdaiʃ	ˈmudɐ̃ũ

Tab. 4: Prethematic vowel alternations in the European Portuguese first conjugation (infinitive and present indicative)

in thematic endings are precisely those with an unstressed prethematic vowel. As a result, no cell in the paradigm can be a good predictor of all others. On the other hand, any pair of cells combining a cell stressed on the ending (say, the infinitive) and a cell stressed on the stem (say, the present 1SG) will allow for accurate prediction of the rest of the paradigm.

We conclude that some inflection systems, such as European Portuguese conjugation, give rise to accurate joint predictiveness despite poor single cell predictiveness for systemic reasons. This gives strong motivation for taking joint predictiveness at face value.

2.3 Behavioral evidence

In the previous two sections we established that joint knowledge of paradigm cells is available and potentially useful to speakers. Ideally we would now like to establish that binary implications play a role in the actual linguistic behavior of speakers. Unfortunately, to the best of our knowledge, the psycholinguistic literature on the acquisition and use of inflection systems has not taken up that issue up to now. There is some evidence that speakers do manifest sensitivity to paradigmatic properties of lexemes in the form of aspects of the statistical distribution of their forms (e.g. Milin et al., 2009), and that speakers do build on unary implications to predict unknown forms (e.g. Seyfarth et al., 2014), but we do not know of any experimental work testing directly whether and how speakers build on exposure to multiple forms to draw inferences on unknown forms.

In the absence of experimental work, we resort to more circumstantial evidence in the form of observations of inflection errors. Specifically, we are concerned with REGULARIZATION errors, where some lexeme instantiating a rare pattern is erroneously inflected according to a more frequent pattern (Marcus et al., 1992).[13] Regularization errors are known to occur commonly in native language acquisition, second language acquisition, and, with lower frequency, in the speech of adult native speakers, and provide evidence on the fact that speakers are indeed applying the pattern under consideration, rather than relying on memorized forms.

Table 5 gives some examples of common regularizations in French conjugation, chosen in the lists established by Kilani-Schoch & Dressler (2005).

The examples of DIRE and FAIRE highlight the workings of regularization. These are the 3rd and 4th most frequent verbs in the *Lexique* database (New et al., 2007), and the only verbs except *être* ‘be’ to have a PRS.2PL in -t rather than -e. There is no documented case of converse conjugation errors where -t would be used instead of -e. Hence type fre-

[13] This argument was first presented in a joint presentation by Olivier Bonami anf Gilles Boyé at the 2006 International Morphology Meeting, Budapest.

	Lexeme	Paradigm cell	Correct form	Regularization
(i)	DIRE 'say'	PRS.2PL	dit	dize
(ii)	FAIRE 'do'	PRS.2PL	fɛt	fəze
(iii)	PRÉVOIR 'plan'	PST.PTCP	pʁevy	pʁevwaje
(iv)	OUVRIR 'open'	PST.PTCP	uvɛʁ	uvʁi
(v)	PRENDRE 'take'	PST.PTCP	pʁi	pʁɑ̃dy
(vi)	PEINDRE 'paint'	PST.PTCP	pɛ̃	pɛ̃dy
(vii)	MOURIR 'die'	PST.PTCP	mɔʁ	muʁy

Tab. 5: Common overregularization errors in French conjugation

quency rather than token frequency is the main determinant of regularization errors: one regularizes to patterns used by many lexemes, not to patterns used by frequent lexemes.

Let us now focus on regularization errors in the past participle. Most of them may be analyzed as regularization to a pattern relating two cells in the paradigm: example (iii) aligns PRÉVOIR with first conjugation verbs with present forms in -wa(j), rather than the very small class of verbs exhibiting a -wa∼-y alternation between present and participle. Example (iv) aligns OUVRIR with second conjugation verbs, which likewise have an infinitive in -iʁ, rather than with the handful of verbs with an -iʁ∼-ɛʁ alternation between infinitive and participle. Example (v) and (vi) analogize PRENDRE and PEINDRE to the rather common pattern of -Cʁ∼-Cy alternation between infinitive and participle already discussed in Table 2.

The more intriguing case is the last one. MOURIR is unique in having a past participle in -ɔʁ, which makes it a likely candidate for regularization. But the expected regularized form would be, depending on which paradigm cell is used for analogy, muʁi,[14] muʁe or mœʁe. The only paradigm cells that would predict muʁy are those of the simple past and past subjunctive, but these are very unlikely bases for analogy, having almost fallen out of usage. The more probable explanation for the use of muʁy as a regularized form comes from the existence of reliable binary implications: the vast majority of verbs with an infinitive in -iʁ but no -i in the present have a past participle in -y.

Although this is at best circumstantial evidence, we submit that it supports the idea that speakers are attuned to binary implications just as they are to unary implications.

3 Quantifying joint predictiveness

In this section we outline a method for assessing joint predictiveness quantitatively. We start from the methodology proposed by Ackerman et al. (2009) and Ackerman & Malouf (2013), which uses the information-theoretic notion of CONDITIONAL ENTROPY to assess the usefulness of implicative structure in paradigms. We argue however that, to have full generality, implicative structure should be assessed by examining the distribution of ALTERNATION PATTERNS between pairs of cells, rather than that of the shapes filling different cells. This leads us to proposing an algorithm for inferring alternation patterns extending the one used by Albright's (2002) *Minimal Generalization Learner*. Finally, we show how a measure of joint predictiveness can be deduced from the observation of the distribution of alternation patterns between two cells.

[14] Kilani-Schoch & Dressler (2005) report that older sources document uses of muʁi. However we have been unable to confirm this. Note that the *FrWac* corpus contains 115 occurrences of *mouru* (pronounced muʁy), but none of *mouri* (pronounced muʁi).

3.1 Paradigm entropy

In this subsection we illustrate how Ackerman, Blevins and Malouf use entropy to quantify implicative structure. Let us consider the small Finnish dataset in Table 6, and consider how easy it is to predict the genitive singular from the nominative singular, assuming for illustrative purposes that this table is representative of the language as a whole.

	NOM.SG	GEN.SG	
(i)	lasi	lasin	'glass'
(ii)	nalle	nallen	'teddy'
(iii)	ovi	oven	'door'
(iv)	kirje	kirjeen	'letter'

Tab. 6: Partial paradigms of a few Finnish nouns with stems in -i or -e (from Ackerman et al. (2009))

It is manifest that nominative singular nouns fall in two groups depending on whether they end in -e or -i, and that genitive singulars fall in three groups, ending in -in, -en, or -een. Importantly, there is a dependency between what happens in the nominative and genitive: where the nominative is in -i, there are only two possibilities in the genitive, namely -in or -en. Likewise, where the nominative is in -e, there are only two possibilities in the genitive, namely -en and -een. Hence, assuming that all classes have an equal size, whatever the shape of the nominative is, there is the same amount of uncertainty as to what the corresponding genitive is: either possibility has $\frac{1}{2}$ probability of being true.

The notion of conditional entropy allows one to capture formally what happens here. The general definition of the entropy of a random variable X is reminded in (5).[15]

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

Suppose we see the arbitrary choice of a nominative singular noun in the Finnish lexicon as a random experiment, and we define a random variable 'NOM.SG' over the outcomes of that experiment with value '-i' when the noun ends in -i and '-e' when the noun ends in -e. Then the entropy of NOM.SG is as given in (6), assuming that classes (i) to (iv) are equally populated—hence, there is an equal chance of drawing a NOM.SG noun in -e or in -i. We find as expected that $H(\text{NOM.SG}) = 1$.

(6)
$$\begin{aligned} H(\text{NOM.SG}) &= -\begin{pmatrix} P(\text{NOM.SG} = \text{-i}) \log_2 P(\text{NOM.SG} = \text{-i}) \\ +P(\text{NOM.SG} = \text{-e}) \log_2 P(\text{NOM.SG} = \text{-e}) \end{pmatrix} \\ &= -(\tfrac{1}{2} \log_2 \tfrac{1}{2} + \tfrac{1}{2} \log_2 \tfrac{1}{2}) = -(\tfrac{1}{2} \times -1 + \tfrac{1}{2} \times -1) = 1 \end{aligned}$$

The same calculation (based on the same assumption of equally populated classes) leads to a higher entropy value for GEN.SG. This is clearly due to the fact that there is more diversity in the possible shapes of GEN.SG forms.

(7) $H(\text{GEN.SG}) = -(\frac{1}{2} \log_2 \frac{1}{2} + \frac{1}{4} \log_2 \frac{1}{4} + \frac{1}{4} \log_2 \frac{1}{4}) = -(\frac{1}{2} \times -1 + \frac{1}{4} \times -2 + \frac{1}{4} \times -2) = 1.5$

[15] Entropy measures how much uncertainty there is in a random variable getting a particular value. Entropy is 0 where there is no uncertainty, and can be an arbitrary large number when uncertainty raises. All other things being equal, entropy is higher for random variables with more values, and for random variables associated with more balanced probability distributions. The use of binary logarithms calibrates entropy so that a choice between two equally probable values leads to an entropy of 1.

Conditional entropy is meant to capture the dependency between two random variables: it measures the amount of uncertainty that remains on the value of variable Y once the value of X is known.

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

We can now evaluate how informative the NOM.SG is on the shape of the GEN.SG by computing the conditional entropy of GEN.SG given NOM.SG. The two possible values for NOM.SG are equiprobable, and, given the choice of one or the other value, there are two equiprobable choices for the GEN.SG; hence we find that the conditional entropy is 1.

(9) $H(\text{GEN.SG} \mid \text{NOM.SG}) = -\big(\frac{1}{2}(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}) + \frac{1}{2}(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2})\big) = \frac{1}{2} + \frac{1}{2} = 1$

We get a more interesting result if we now compute the conditional entropy of the NOM.SG given the GEN.SG. Remember that there are 3 possible shapes for the GEN.SG, which are not equally probable: -en is found in half of the data, while -in and -een are each found in a quarter of the data. Note in addition that there is no uncertainty as to the shape of the NOM.SG when the GEN.SG ends in -in or -een. This is reflected in the calculation of conditional entropy, where the terms corresponding to these two outcomes are each zero. As a result, the conditional entropy of the NOM.SG given the GEN.SG is lower than the conditional entropy of the GEN.SG given the NOM.SG.

(10) $H(\text{NOM.SG} \mid \text{GEN.SG}) = -\big(\frac{1}{2}(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}) + \frac{1}{4}\log_2 1 + \frac{1}{4}\log_2 1\big) = \frac{1}{2} + 0 + 0 = \frac{1}{2}$

This toy example illustrates how conditional entropy can be used to assess how predictive one paradigm cell is of another paradigm cell.[16] Ackerman & Malouf (2013) build on this to define the PARADIGM ENTROPY of an inflection system as the average conditional entropy of all pairs of paradigm cells: this provides an assessment of the overall predictiveness of the system, by evaluating how hard it is to predict some randomly selected cell of a lexeme from another randomly selected cell.[17] The central result of Ackerman & Malouf (2013) is that, for a variety of languages, paradigm entropy is a lot lower than would be expected given the diversity of inflectional realizations in each cell.

Our goal in this paper is to extend the notion of paradigm entropy to the case of prediction from multiple paradigm cells. Before we do this, however, there is one methodological issue that needs attention.

3.2 Classifying shapes or alternations?

The calculation of paradigm entropy as defined by Ackerman and Malouf relies on the use of a particular random variable over possible forms filling each paradigm cell. In effect, this means that the evaluation of entropy is dependent on a particular classification of the shapes found in each paradigm cell. In the previous example, we admitted without discussion that it was reasonable to split NOM.SG shapes in two classes depending on their final vowel, and GEN.SG shapes in three classes. This is a significant methodological problem, for two reasons. First, more than one classification may be reasonable for the same dataset: hence, some principled way of choosing one classification would be useful.

[16] Realistic applications rely on large inflected lexica, often derived from corpora, to provide reasonable assessments of the probability distributions of inflectional realizations for each paradigm cell (Bonami & Boyé, 2014; Bonami & Luís, 2014; Sims, 2015).

[17] Again, as Ackerman and Malouf note, using a weighted average taking into account the relative frequency of paradigm cells in a corpus would lead to a more accurate estimation.

Second and more importantly, it is very tempting for the analyst to rely on an independently established morphological analysis of the inflectional system at hand, typically in terms of stems and exponents, to design their classification. But segmentation into stems and exponents relies on comparison of forms across the paradigm; thus knowing what is the exponent in one cell for a particular lexeme often amounts to knowing something about the forms taken by that lexeme in other cells. For instance, consider again the French data in Table 2. Conventional wisdom holds that the exponent of the infinitive of French class (iv) verbs form is -ʁ, whereas it is -iʁ for classes (v) to (vii). While this makes perfect sense when one considers the whole paradigm of these verbs, there is no contrast between the phonological shapes of the infinitives in all four classes: all end in -iʁ. Only that information is available to a speaker trying to infer other forms of a lexeme from knowledge of its infinitive. Therefore, a classification of infinitives distinguishing (iv) and (v) relies on knowledge that is unavailable to a speaker attempting to solve the PCFP.

Bonami & Boyé (2014) discuss this issue and propose the following amendment to Ackerman and colleague's methodology. First, for any pair of cells (A, B), lexemes are classified according to which alternation they instantiate for that pair of cell. We note the corresponding random variable '$A \sim B$'. Second, for the purposes of assessing the predictive value of paradigm cell A, an opportunistic classification of the predictor cell is deduced based on the set of alternations it could conceivably enter into. We note the corresponding random variable '$A_{A\sim B}$'.

To understand the workings of this method, let us go back to the toy Finnish dataset. Table 7 shows the classification of alternations between the NOM.SG and GEN.SG deduced from the most naive pattern matching, and the associated classifications of NOM.SG (for the purposes of predicting GEN.SG).

	NOM.SG	GEN.SG	NOM.SG ∼ GEN.SG	$\text{NOM.SG}_{\text{NOM.SG}\sim\text{GEN.SG}}$
(i)	lasi	lasin	$X \sim X$n	$\{X \sim X\text{n}, X\text{i} \sim X\text{en}, X \sim X\text{en}\}$
(ii)	nalle	nallen	$X \sim X$n	$\{X \sim X\text{n}, X \sim X\text{en}\}$
(iii)	ovi	oven	Xi $\sim X$en	$\{X \sim X\text{n}, X\text{i} \sim X\text{en}, X \sim X\text{en}\}$
(iv)	kirje	kirjeen	$X \sim X$en	$\{X \sim X\text{n}, X \sim X\text{en}\}$

Tab. 7: Random variables for predicting GEN.SG from NOM.SG for the nouns in Table 6

As the table shows, we correctly deduce that the distinction between -e and -i endings in the NOM.SG is relevant for predicting the GEN.SG. However, this is now not stipulated, but deduced from the distribution of alternations: the difference between the two classes of NOM.SG forms is that the pattern 'Xi $\sim X$en' could not possibly apply to forms in -e. Notice that, although comparison of NOM.SG and GEN.SG forms plays a role in the computation of values for $\text{NOM.SG}_{\text{NOM.SG}\sim\text{GEN.SG}}$, what is encoded is general information about the shape of the system rather than particular information about one lexeme. Hence we are properly defining a random variable over possible NOM.SG shapes.

We thus take the problem of predicting paradigm cell B from cell A as the problem of determining which class of alternation relates B to A given which class A falls in. Accordingly, we assess the PCFP on the basis of the following conditional entropy, which we call the UNARY IMPLICATIVE ENTROPY FROM A TO B.

(11) $$H(A \Rightarrow B) = H(A \sim B \mid A_{A\sim B})$$

In the case at hand, we find that the implicative entropy is, again, of 1, which is in line with the conditional entropy obtained by Ackerman et al. (2009): we find again two equally sized classes of NOM.SG forms each of which is compatible with two equiprobable patterns.

(12) $$\begin{aligned} H(\text{NOM.SG} \Rightarrow \text{GEN.SG}) &= -\left(\tfrac{1}{2}(\tfrac{1}{2}\log_2\tfrac{1}{2}+\tfrac{1}{2}\log_2\tfrac{1}{2})+\tfrac{1}{2}(\tfrac{1}{2}\log_2\tfrac{1}{2}+\tfrac{1}{2}\log_2\tfrac{1}{2})\right) \\ &= \tfrac{1}{2}+\tfrac{1}{2}=1 \end{aligned}$$

Things are different however if we consider prediction in the other direction, as indicated in Table 8.

	NOM.SG	GEN.SG	NOM.SG ∼ GEN.SG	$\text{GEN.SG}_{\text{NOM.SG}\sim\text{GEN.SG}}$
(i)	lasi	lasin	$X \sim X$n	$\{X \sim X\text{n}\}$
(ii)	nalle	nallen	$X \sim X$n	$\{X \sim X\text{n}, X\text{i} \sim X\text{en}, X \sim X\text{en}\}$
(iii)	ovi	oven	Xi $\sim X$en	$\{X \sim X\text{n}, X\text{i} \sim X\text{en}, X \sim X\text{en}\}$
(iv)	kirje	kirjeen	$X \sim X$en	$\{X \sim X\text{n}, X\text{i} \sim X\text{en}, X \sim X\text{en}\}$

Tab. 8: Random variables for predicting NOM.SG from GEN.SG for the nouns in Table 6

When classifying GEN.SG shapes in terms of the alternations they could give rise to, the crucial difference is between class (i) and all other cases: because class (i) nouns do not have an -e- in their GEN.SG form, only one pattern can apply to them. We are thus now dealing with two unbalanced classes of GEN.SG, one of which is compatible with three patterns.

(13) $$\begin{aligned} H(\text{NOM.SG} \Rightarrow \text{GEN.SG}) &= -\left(\tfrac{1}{4}\log_2 1+\tfrac{3}{4}(\tfrac{1}{3}\log_2\tfrac{1}{3}+\tfrac{1}{3}\log_2\tfrac{1}{3}+\tfrac{1}{3}\log_2\tfrac{1}{3})\right) \\ &= 0+\tfrac{3}{4}\log_2 3 \approx 1.19 \end{aligned}$$

In this particular example, it turns out that implicative entropy assesses prediction of NOM.SG from GEN.SG to be harder than prediction of GEN.SG from NOM.SG. We argue that this is as it should: at least in the absence of further knowledge, there is no way for a speaker witnessing a new GEN.SG form in -en to decide whether it belongs to class (ii), (iii), or (iv). The calculation in (10) underestimated the difficulty by building into the classification of kirjeen information that really stems from examination of its NOM.SG.

We submit then that implicative entropy is a superior measure of the difficulty of predicting from one cell to another cell.

Although the computation of implicative entropy as detailed in the previous subsection does not rely on an *a priori* segmentation, it does depend on the choice of an algorithm producing a partition of any set of pair of forms into mutually exclusive alternation patterns. The design and application of a typologically unbiased such algorithm is beyond the scope of the present paper. We are thus opportunistically relying on an algorithm that is specifically fit to the types of morphophonological alternations found in the Romance languages we are focusing on: mainly suffixation, sometimes accompanied by alternations in the final or penultimate syllable. In the interest of space, we will skip the details, and just illustrate the main features of the algorithm on some of the European Portuguese data from Table 4.

- For any pair of strings (ϕ, ψ), we first left align ϕ and ψ and identify constant and variable substrings. This results in an elementary pattern consisting of three strings containing placeholders, written as $l \rightleftharpoons r/c$, where ϕ (respectively ψ) is the result of substituting the elements in c into the placeholders in l (respectively r). For instance, the pair (ʃəgar, ʃegu) leads to the elementary pattern in (14).

(14) …ə…aɾ ⇌ …e…u/ʃ…g…

- For any pair of cells, we then fuse the elementary patterns that share the same left-hand and righthand sequences, by generalizing over their contexts of use, through a variant of Albright's (2002) minimal generalization strategy. For instance, in our European Portuguese dataset, the pattern in (14) relating infinitive and present 1SG forms is generalized to (15). This captures the fact that all pairs in the dataset start with an arbitrary sequence X ending in a consonant and have a single non-lateral consonant between the two alternating vowels.

(15) …ə…aɾ ⇌ …e…u/X`[+cons]`…`[+cons,-lat]`…

The resulting contextualized patterns are then used in the computation of unary implicative entropy. The use of contextualized patterns allows for a more accurate estimation of pattern applicability, and hence to a reduction in the estimation of implicative entropy.

3.3 Binary implicative entropy

In the previous subsection we have defined and motivated the use of unary implicative entropy. We now turn to binary case.

Given that unary implicative entropy has been defined on the basis of alternation patterns between pairs of forms, the intuitively most satisfactory way of extending the notion to prediction from n cells would be to start from a notion of alternation patterns between $n+1$ forms. While this is certainly a conceivable strategy, it raises various issues, mostly linked to the lack of previous work on such alternations that could be used to provide intuitions as to the viability of different algorithms. We will thus not pursue that possibility. Instead, we propose to identify what can be deduced on prediction from multiple cells on the basis of simple, binary patterns.

We define the implicative entropy from A and B to C as follows:

(16) $H(A,B \Rightarrow C) = H(A \sim C, B \sim C \mid A_{A\sim C}, B_{B\sim C}, A \sim B)$

The intuition behind this definition is the following. We are trying to assess how difficult it is to predict C from knowledge of A and B. This is equivalent to assessing how difficult it is to know which pattern relates A to C and which pattern relates B to C given knowledge of A and B. Thus what we are trying to predict is the value of the joint random variable '$A \sim C, B \sim C$' given a random variable providing an appropriate joint classification of A and B. The joint random variable '$A_{A\sim C}, B_{B\sim C}, A \sim B$' is such a classification collecting all information that can easily be deduced using binary patterns of alternation.

The effects of definition (16) are best evaluated by looking at a concrete example. Let us go back to the French data in Table 2, and focus on prediction of the past participle from joint knowledge of the infinitive and present 3PL. Table 9 indicates values for the relevant variables INF ∼ PTCP, $\text{INF}_{\text{INF}\sim\text{PTCP}}$, 3PL ∼ PTCP, $\text{3PL}_{\text{3PL}\sim\text{PTCP}}$, and INF ∼ 3PL. As the reader can easily check, to every combination of values for the last three variables corresponds a single combination of value for the two first variables; in other terms, every possible predictor value gives rise to a single possible predicted value. As a result, there is no uncertainty, and the implicative entropy is 0.[18] This contrasts with what happens

[18] In this particular example, the final variable INF ∼ 3PL seems unnecessary, since the following holds.

$$H(A \sim C, B \sim C \mid A_{A\sim C}, B_{B\sim C}, A \sim B) = H(A \sim C, B \sim C \mid A_{A\sim C}, B_{B\sim C})$$

Class	Size	Sample lexeme	INF ∼ PTCP	3PL ∼ PTCP	$\text{INF}_{\text{INF}\sim\text{PTCP}}$	$\text{3PL}_{\text{3PL}\sim\text{PTCP}}$	INF ∼ 3PL
(i)	4108	LIVRER 'deliver' (livʁe,livʁ,livʁe)	$p_1 =$ '… ⇌ …/Xe'	$p_6 =$ '… ⇌ …e/X…'	$\{p_1\}$	$\{p_6\}$	$p_{12} =$ '…e ⇌ …/X…'
(ii)	210	RELIER 'link' (ʁəlje,ʁəli,ʁəlje)	p_1	$p_7 =$ '…i ⇌ …je/X…'	$\{p_1\}$	$\{p_6, p_7\}$	$p_{13} =$ '…je ⇌ …i/X…'
(iii)	22	RATISSER 'rake' (ʁatise,ʁatis,ʁatise)	p_1	p_6	$\{p_1\}$	$\{p_6, p_8\}$	p_{12}
(iv)	327	BÂTIR 'build' (batiʁ,batis,bati)	$p_2 =$ '…ʁ ⇌ …/Xi…'	$p_8 =$ '…s ⇌ …/Xi…'	$\{p_2, p_3\}$	$\{p_6, p_8\}$	$p_{14} =$ '…ʁ ⇌ …s/Xi…'
(v)	37	TENIR 'hold' (təniʁ,tjɛn,təny)	$p_3 =$ '…iʁ ⇌ …y/X…'	$p_9 =$ '…jɛn ⇌ …əny/Xn…'	$\{p_2, p_3\}$	$\{p_6, p_9\}$	$p_{15} =$ '…əniʁ ⇌ …jɛn/X…'
(vi)	8	OUVRIR 'open' (uvʁiʁ,uvʁ,uvɛʁ)	$p_4 =$ '…ʁiʁ ⇌ …ɛʁ/X…'	$p_{10} =$ '…ʁ ⇌ …ɛʁ/X…'	$\{p_2, p_3, p_4\}$	$\{p_6, p_{10}\}$	$p_{16} =$ '…iʁ ⇌ …/X…'
(vii)	1	MOURIR 'die' (muʁiʁ,mœʁ,mɔʁ)	$p_5 =$ '…u…iʁ ⇌ …ɔ…/m…ʁ'	$p_{11} =$ '…œ… ⇌ …ɔ…/Xm…ʁ'	$\{p_2, p_3, p_4, p_5\}$	$\{p_6, p_{11}\}$	$p_{17} =$ '…u…iʁ ⇌ …œ…/m…ʁ'

Tab. 9: Random variables for predicting PST.PTCP from INF and PRS.3PL in Table 2

when predicting the participle from a single cell. When predicting from the infinitive, the random variable over INF shapes groups together classes (iv) and (v) which instantiate different patterns. This leads to a non-null unary implicative entropy $H(\text{INF} \Rightarrow \text{PTCP}) \approx 0.0064$. Likewise, when predicting from the 3PL, the random variable over 3PL shapes groups together classes (iii) and (iv) which instantiate different patterns, leading to a non-null unary implicative entropy $H(\text{3PL} \Rightarrow \text{PTCP}) \approx 0.0251$.

We thus find that our definitions of implicative entropy capture our initial observations on the French dataset. In this example, the binary implicative entropy is 0, whereas the unary implicative entropy from either predictor cell is strictly positive. In general, (17) is a trivial consequence of our definitions:

(17) $H(A, B \Rightarrow C) \leq \min\{H(A \Rightarrow C), H(B \Rightarrow C)\}$

This captures a basic intuition about predictiveness: knowledge of two forms of one lexeme cannot be less predictive than knowledge of only one of these two forms. This does not entail however that knowledge of a second form always improves predictiveness. Thus it is an empirical question how useful is joint knowledge in full inflectional systems.

Let us finally note that the definition of binary implicative entropy easily extends to the general case of prediction from n cells. Given n paradigm cells $A^1, \ldots, A^n$, we note $[A^1 \sim \cdots \sim A^n]$ the joint of all random variables $A^i \sim A^j$ for $i, j \in \{1,,n\}$. We then define n-ary implicative entropy as follows.

(18) $H(A^1, \ldots, A^n \Rightarrow B) = H\left(A^1 \sim B, \ldots, A^n \sim B \mid A^1_{A^1 \sim B}, \ldots, A^n_{A^n \sim B}, [A^1 \sim \cdots \sim A^n]\right)$

In this section we have presented a general measure of the predictiveness of collections of paradigm cells that captures crucial intuitions on joint prediction. This measure builds heavily on proposals from Ackerman et al. (2009), as it reduces predictiveness to a form of conditional entropy of one random variable capturing properties of a predicted cell given a random variable capturing properties of predictor cells; it differs in deriving from explicitly constructed alternation patterns, and thereby avoiding methodological worries with non-formalized morphological classifications. The method has been fully implemented[19] and can thus easily be applied to realistically-sized datasets, so as to assess how much of a drop in implicative entropy results from joint prediction.

4 Empirical results

4.1 Joint predictiveness is helpful

The method above was applied to two datasets, respectively of French and European Portuguese conjugation. The French dataset contains all non-defective verbs in the *Flexique* lexicon (Bonami et al., 2014). *Flexique* is an inflectional lexicon providing full paradigms in (hand-checked) phonemic transcription for all nouns, verbs and adjectives documented in the *Lexique* database (New et al., 2007), whose nomenclature was itself derived from a written literature corpus and a subtitles corpus. The European Portuguese dataset was derived from the Coimbra Pronunciation Dictionary (Veiga et al., 2012) and contains

This is not the case in general, however. In any situation where A is a perfect predictor of C but B is not, $H(A \sim C, B \sim C \mid A_{A\sim C}, B_{B\sim C}, A \sim B) = 0$ whereas $H(A \sim C, B \sim C \mid A_{A\sim C}, B_{B\sim C}) = H(B \sim C \mid B_{B\sim C}) > 0$.

[19] The programs written for this study are freely available and can be downloaded from the website of the *Laboratoire de linguistique formelle*.

the 2000 most frequent verbs in the CETEMPúblico journalistic corpus (Santos & Rocha, 2001).[20] Table 10 provides more information on the two datasets.

	French	Portuguese
Number of lexemes	4951	1995
Number of paradigm cells	51	68

Tab. 10: Description of the French and European Portuguese datasets

In principle, we are interested in computing the average n-ary implicative entropy for the two datasets for all values of n. In practice however, the number of distinct combinations of n predictor cells with 1 predicted cell grows very quickly,[21] making calculations unpractical for large values of n. We thus stopped at $n = 4$. Table 11 presents the raw results for both languages.

n	French	Portuguese
1	0.174	0.205
2	0.054	0.106
3	0.021	0.076
4	0.009	—

Tab. 11: Average n-ary implicative entropy in French and European Portuguese

The answer to our main question is thus very clear: on average, in both languages, predicting from two paradigm cells is considerably easier than predicting from one, and predicting from three is still much easier than predicting from two. In each case, adding one predictor divides the implicative entropy by about two. We take this to establish beyond doubt that joint prediction is useful in solving the PCFP.

Ackerman & Malouf (2013) make a distinction between the ENUMERATIVE COMPLEXITY of an inflection system, roughly corresponding to the size of the system (size of paradigms, number of morphs per word, number of inflection classes, etc.) and its INTEGRATIVE COMPLEXITY, corresponding to the difficulty of using the system once it has been learned. They argue that while systems vary widely in enumerative complexity, their integrative complexity is uniformly low; this is assessed by showing that paradigm entropy, the predecessor of our notion of unary implicative entropy, is low. The present paper provides more support for that conclusion by showing that entropy gets a lot lower when taking into account more information available to speakers, including phonotactic knowledge on contexts for different alternations, statistical knowledge on the prevalence of different inflection strategies, and joint knowledge of multiple paradigm cells.

4.2 Joint predictiveness and principal part systems

We finally turn to the question of principal part systems (Hockett, 1967; Matthews, 1972; Finkel & Stump, 2007; Stump & Finkel, 2013) and their relation with implicative entropy. As we suggested in the introduction, a categorical principal part system is a set

[20] We are grateful to Fernando Perdigão for providing the European Portuguese dataset, checking the transcriptions, and allowing us to use it.

[21] In general, for n paradigm cells and k predictors, there are $\frac{n!}{k!(n-k-1)!}$ such combinations. With 3 predictors, there are already $11,745,300$ cases to consider for French and $52,120,640$ for Portuguese, in each of which we are iterating over a lexicon of thousands of lexemes.

of paradigm cells knowledge of which is sufficient to predict categorically all remaining cells. Clearly, categorical principal part systems can be deduced from implicative entropy. Given an inflectional system with paradigm cells $\mathscr{C}$, we define the RESIDUAL UNCERTAINTY of any subset $\mathscr{A} = \{A^1, \ldots, A^n\}$ of $\mathscr{C}$ as the average of all $H(A^1, \ldots, A^n \Rightarrow B)$ for $B \in \mathscr{C} \setminus \mathscr{A}$. A categorical principal part system is thus a set of paradigm cells with a residual uncertainty of exactly 0.

Table 12 indicates, for all values of n we could document, the number of distinct principal part systems of cardinality n.

n	French	Portuguese
1	0	0
2	0	184
3	0	7884
4	1344	—

Tab. 12: Number of principal part systems of cardinality n in French and European Portuguese

European Portuguese has a number of principal part systems of cardinality 2—in fact about 4% of all pairs of cells form a principal part system. Such a result can largely be explained by our observations in section 2.2. Remember that there are two main sources of uncertainty in Portuguese conjugation: neutralizations of theme vowel contrasts, and neutralizations of vocalic contrasts in unstressed prethematic vowels. Given this, any set of principal parts must contain at least one cell giving rise to a three-way thematic distinction and one cell with a stressed (and thus not neutralized) prethematic vowel. There are 506 distinct such combinations; the 184 observed binary principal part systems correspond to about one third of this candidate set. The remaining two thirds do not constitute principal part systems because of other complexities of the system not discussed in this paper.

The fact that we find no principal part system of cardinality $n < 4$ for French is not surprising, given the high prevalence of arbitrarily distributed stem allomorphy in that system (Bonami & Boyé, 2002). Note also that Stump & Finkel (2013) find no system with cardinality smaller than 5. Although we use different methodology and datasets, it would have been surprising to find a radically smaller result.[22] However, we submit that this result does not tell us much on the predictibility of the French system.

Categorical principal part systems are based on absolute certainty. But presumably, absolute certainty is of little concern to the language user. From a psycholinguistic perspective, it is clear that speakers do not manifest perfect mastery of the inflection system: even fully competent native speakers occasionally make inflection errors, of the same general kind as both L1 and L2 learners. This suggests that the use of the system is driven by reliable but not necessarily categorical inference. Given this, even from the pedagogical perspective that is the initial motivation for principal part systems, it seems worthless to focus on collections of cells leading to absolute certainty. From both perspectives, the important question is to determine which sets of paradigm cells lead to highly reliable, rather than perfectly categorical, inference.

With this in mind, let us examine the distribution of residual uncertainty values for all 2550 pairs of cells in the French system, as indicated in Figure 3. While a few pairs of cells give rise to high residual uncertainty, many of them are well below the average

[22] Bonami & Beniamine (2015) erroneously reports the existence of principal part systems of cardinality 3. This was due to now corrected errors in the *Flexique* dataset.

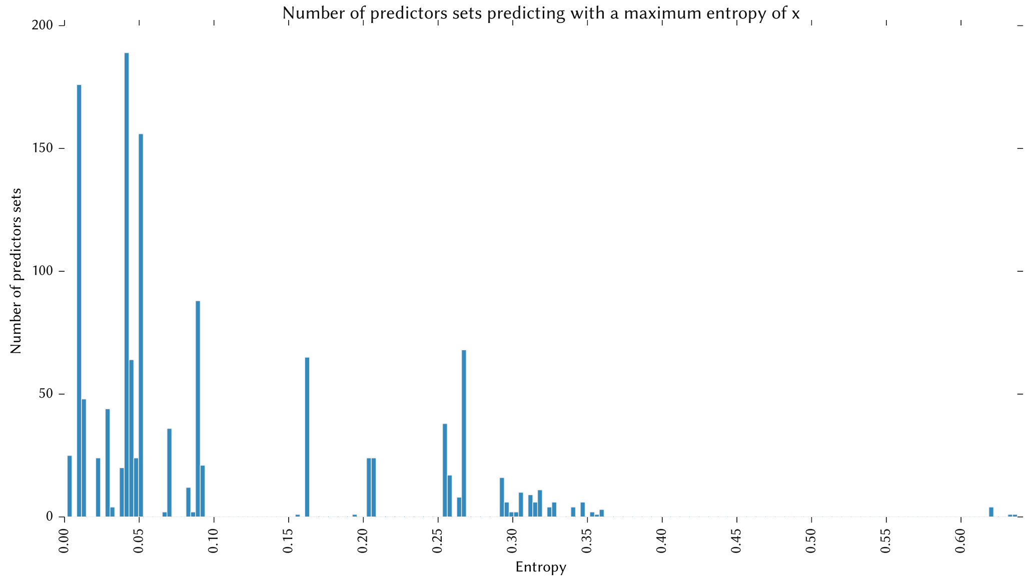

Fig. 3: Distribution of residual uncertainty for pairs of cells in French

of 0.056. In fact, there are 25 pairs of cells with a residual uncertainty lower than 0.005. We submit that such a low residual uncertainty makes these pairs of cells reliable enough for realistic purposes. Because the reader may not be used to evaluating uncertainty in terms of entropy, it is worth noting that an entropy of 0.005 corresponds roughly to the amount of uncertainty associated with an event of probability 0.9995.[23] Clearly, this is a level of certainty that speakers cannot be expected to exceed.

As a final observation, it is telling to observe the difference between the prevalence of systems of strict, categorical principal part systems with that of 'near' principal part systems. As Table 13 illustrates, European Portuguese has nearly as many 'near' as categorical principal part systems, while for French there is a very important difference. This observation suggests that a focus on categorical systems can be misleading for language comparison, and more generally, typology: our study suggests that the implicative structure of French and European Portuguese are more or less equally reliable, whether this is evaluated in terms of n-ary implicative entropy or cardinality of 'near' principal part systems.

n	French	Portuguese
1	0	0
2	25	184
3	5821	7912
4	130904	—

Tab. 13: Number of 'near' principal part systems of cardinality n in French and European Portuguese (residual uncertainty lower than 0.005)

[23] If X is a binary random variable one of whose outcomes has a probability of 0.9995, $H(X) = -(0.9995 \times \log_2 0.9995 + 0.0005 \times \log_2 0.0005) \approx 0.0062$.

5 Conclusion

The goal of this paper was to establish the relevance of joint information on multiple paradigm cells to address the Paradigm Cell Filling Problem. We have shown on the basis of corpus evidence that speakers indeed are likely to frequently find themselves in the situation of having to infer the form filling one cell while knowing the forms filling multiple other cells. We have then proposed a general method for assessing quantitatively the difficulty of the PCFP when predicting from multiple cells, and applied that method to realistically sized datasets in two languages; the clear conclusion is that, at least in these two languages, knowing more paradigm cells makes the PCFP notably easier to solve.

It should be stressed that this paper only established that speakers are exposed to relevant information and that this information is helpful; the next step, of course, is to establish experimentally that speakers do indeed rely on joint prediction when addressing predicting the form of unknown words. We are not aware of any study addressing exactly that issue. Until such studies available, though, there is no reason to doubt that speakers make use of what information is available to them.

References

Ackerman, Farrell, James P. Blevins & Robert Malouf. 2009. Parts and wholes: implicative patterns in inflectional paradigms. In James P. Blevins & Juliette Blevins (eds.), *Analogy in grammar*, 54–82. Oxford: Oxford University Press.

Ackerman, Farrell & Robert Malouf. 2013. Morphological organization: the low conditional entropy conjecture. *Language* 89. 429–464.

Albright, Adam C. 2002. *The identification of bases in morphological paradigms*: University of California, Los Angeles dissertation.

Albright, Adam C. & Bruce P. Hayes. 2003. Rules vs. analogy in english past tenses: A computational/experimental study. *Cognition* 90. 119–161.

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

Baroni, Marco, Silvia Bernardini, Adriano Ferraresi & Eros Zanchetta. 2009. The wacky wide web: A collection of very large linguistically processed web-crawled corpora. In *Language resources and evaluation*, vol. 43, 209–226.

Blevins, James. 2003. Stems and paradigms. *Language* 79. 737–767.

Blevins, James P. 2005. Word-based declensions in Estonian. In Geert E. Booij & Jaap van Marle (eds.), *Yearbook of morphology 2005*, 1–25. Springer.

Blevins, James P. 2006. Word-based morphology. *Journal of Linguistics* 42. 531–573.

Blevins, James P., Petar Milin & Michael Ramscar. inpress. The Zipfian Paradigm Cell Filling Problem. In Ferenc Kiefer, James P. Blevins & Huba Bartos (eds.), *Morphological paradigms and functions*, Leiden: Brill.

Bonami, Olivier (ed.). 2012. *Word structure 5.1: Stems in inflection and lexeme formation.* Edinburgh: Edinburgh University Press.

Bonami, Olivier & Sarah Beniamine. 2015. Implicative structure and joint predictiveness. In Vito Pirelli, Claudia Marzi & Marcello Ferro (eds.), *Word structure and word usage. proceedings of the networds final conference*, .

Bonami, Olivier & Gilles Boyé. 2002. Suppletion and stem dependency in inflectional morphology. In Franck Van Eynde, Lars Hellan & Dorothee Beerman (eds.), *The proceedings of the hpsg '01 conference*, 51–70. Stanford: CSLI Publications.

Bonami, Olivier & Gilles Boyé. 2014. De formes en thèmes. In Florence Villoing, Sarah Leroy & Sophie David (eds.), *Foisonnements morphologiques. etudes en hommage à françoise kerleroux*, 17–45. Presses Universitaires de Paris Ouest.

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

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

Brown, Dunstan. 1998. Stem lndexing and morphonological selection in the Russian verb: a network morphology account. In R. Fabri, A. Ortmann & T. Parodi (eds.), *Models of inflection*, 196–224. Niemeyer.

Chan, Erwin. 2008. *Structures and distributions in morphological learning*: University of Pennsylvania dissertation.

Faaß, Gertrud & Kerstin Eckart. 2013. Sdewac — a corpus of parsable sentences from the web. In Irina Gurevych, Chris Biemann & Torsten Zesch (eds.), *Language processing and knowledge in the Web*, 61–68. Heidelberg: Springer.

Finkel, Raphael & Gregory T. Stump. 2007. Principal parts and morphological typology. *Morphology* 17. 39–75.

Finkel, Raphael & Gregory T. Stump. 2009. Principal parts and degrees of paradigmatic transparency. In James P. Blevins & Juliette Blevins (eds.), *Analogy in grammar*, 13–54. Cambridge: Cambridge University Press.

Gilkerson, Jill & Jeffrey A. Richards. 2009. The power of talk, 2nd edition. Tech. rep. LENA Foundation. `https://www.lenafoundation.org/wp-content/uploads/2014/10/LTR-01-2_PowerOfTalk.pdf`.

Hart, Betty & Todd R. Risley. 1995. *Meaningful differences in the everyday experience of young american children*. Baltimore: Paul H Brookes Publishing.

Hippisley, Andrew. 1998. Indexed stems and Russian word formation: a Network Morphology account of Russian personal nouns. *Linguistics* 36. 1039–1124.

Hockett, Charles F. 1967. The Yawelmani basic verb. *Language* 43. 208–222.

Kilani-Schoch, Marianne & Wolfgang Dressler. 2005. *Morphologie naturelle et flexion du verbe français*. Tübingen: Gunter Narr Verlag.

Maiden, Martin. 1992. Irregularity as a determinant of morphological change. *Journal of Linguistics* 28. 285–312.

Marcus, Gary F., Steven Pinker, Michael Ullman, Michelle Hollander, T. John Rosen, Fei Xu & Harald Clahsen. 1992. *Overregularization in language acquisition*. Wiley.

Matthews, P. H. 1972. *Inflectional morphology. a theoretical study based on aspects of latin verb conjugation*. Cambridge: Cambridge University Press.

Milin, Petar, Dušica Filipović Đurđević & Fermin Moscoso del Prado Martín. 2009. The simultaneous effects of inflectional paradigms and classes on lexical recognition: Evidence from Serbian. *Journal of Memory and Language* 60. 50–64.

Montermini, Fabio & Olivier Bonami. 2013. Stem spaces and predictibility in verbal inflection. *Lingue e Linguaggio* 12. 171–190.

New, Boris, Marc Brysbaert, Jean Veronis & Christophe Pallier. 2007. The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics* 28. 661–677.

Sagot, Benoît. 2010. The *Lefff*, a freely available and large-coverage morphological and syntactic lexicon for French. In *Proceedings of lrec 2010*, .

Santos, Diana & Paulo Rocha. 2001. Evaluating cetempúblico, a free resource for Portuguese. In *Proceedings of the 39th annual meeting of the association for computational linguistics*, 442–449.

Seyfarth, Scott, Farrell Ackerman & Rob Malouf. 2014. Implicative organization facilitates morphological learning. In *Proceedings of the 40th annual meeting of the berkeley linguistics society*, 480–494.

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

Stump, Gregory T. & Raphael Finkel. 2013. *Morphological typology: From word to paradigm*. Cambridge: Cambridge University Press.

Thymé, Ann, Farrell Ackerman & Jeff Elman. 1994. Finnish nominal inflection: Paradigmatic patterns and token analogy. In Susan D. Lima, Roberta Corrigan & Gregory K. Iverson (eds.), *The reality of linguistic rules*, John Benjamins.

Veiga, Arlindo Oliveira da, Sara Candeias & Fernando Perdigão. 2012. Generating a pronunciation dictionary for european portuguese using a joint-sequence model with embedded stress assignment. *Journal of the Brazilian Computer Society* 88.

Wurzel, Wolfgang Ulrich. 1984. *Flexionsmorphologie und Natürlichkeit. ein Beitrag zur morphologischen Theoriebildung*. Berlin: Akademie-Verlag. Translated as Wurzel (1989).

Wurzel, Wolfgang Ulrich. 1989. *Inflectional morphology and naturalness*. Dordrecht: Kluwer.