Implicative structure and joint predictiveness

Olivier Bonami
Université Paris-Sorbonne
Laboratoire de linguistique formelle
(U. Paris Diderot & CNRS)
olivier.bonami@paris-sorbonne.fr

Sarah Beniamine
Université Paris Diderot
Laboratoire de linguistique formelle &
Alpage (Inria & U. Paris Diderot)
sarah.beniamine@inria.fr

1 Introduction

(Ackerman et al., 2009) define the PARADIGM CELL FILLING PROBLEM (PCFP), which we paraphrase in (1), as the cornerstone of the study of inflectional paradigms.

(1) How do speakers know how to inflect the full paradigm of a lexeme on the basis of exposure to only some of its forms?

(Ackerman et al., 2009) go on to argue that speakers rely on knowledge of the IMPLICATIVE STRUCTURE of paradigms (Wurzel, 1984): paradigms are structured in such a way that there are reliable correlations between the form filling one paradigm cell A and the form filling another cell B. The reliability of these correlations depends on the particular pair of cells A and B under scrutiny; it can be assessed quantitatively by examining the statistical distribution of operations required to go from A to B in the lexicon.

This presentation focuses on one particular aspect of implicative structure, which we call JOINT PREDICTIVENESS. In some situations, joint knowledge of two paradigm cells A and B provides more information on cell C than could be inferred from knowledge of either A or B. Table 1 below provides a simple example from French, using lexemes illustrating 7 patterns corresponding to of 95% of the verbs documented in the *Flexique* phoneticized lexicon (Bonami et al., 2014). In French conjugation, predicting the past participle from the infinitive is hard, because of the opacity between second conjugation infinitives, such as BÂTIR, and some third conjugation infinitives, such as TENIR, OUVRIR, MOURIR. Predicting the past participle from present SG forms is also hard, this time because some first conjugation verbs with a stem in *-i* (e.g. RELIER) are not distinguished from second conjugation verbs. A different subset of first conjugation verbs (e.g. RATISSER) raises similar problems for PL forms.

Overall, no other cell in the paradigm is a very good predictor of the past participle. However, joint knowledge of some pairs of paradigm cells radically improves the quality of prediction. For instance, joint knowledge of the infinitive and some present plural form removes all uncertainty in the sample in Table 1: knowledge of the infinitive form partitions the set of lexemes in two classes within which the PRS.3PL is fully predictive of the past participle.

Although the existence of joint predictiveness is acknowledged in the literature (Matthews, 1972; Thymé et al., 1994; Ackerman et al., 2009; Stump and Finkel, 2013; Blevins, forthcoming; Sims, forthcoming), little attention has been given to quantifying its importance. In this paper we first give further arguments that joint predictiveness is a crucial aspect of implicative structure, and that a careful empirical examination of joint predictiveness is essential to both linguistic and psycholinguistic assessment of the PCFP and related issues. We then propose and illustrate a method for the quantitative evaluation of joint predictiveness. We end with a discussion of principal part systems.

2 The relevance of joint predictiveness

We start by establishing that speakers do have the opportunity to use joint predictiveness. Figure 1 plots how the number of forms per lemma evolves when walking through the 1.6 billion words of the *FrWaC* web corpus (Baroni et al., 2009), restricting attention to the 6847 verbs documented in the *Lefff* lexicon (Sagot, 2010) to compensate for tagging errors.[1] Note that 1.6 billion words is

[1] Note that this restriction leads to overestimating the average number of forms per lemma, as neologisms, very rare words and hapaxes not present in the lexical resource are not included. We are counting distinct forms rather than distinct paradigm cells, as there is currently no tagger for French that reliably disambiguates homographic forms of the same lexeme. French verbs have 51 paradigm cells, and the average number of distinct forms per verb in the *Lefff* lexicon is 35.8.

In Vito Pirrelli, Claudia Marzi, Marcello Ferro (eds.): *Word Structure and Word Usage*. Proceedings of the NetWordS Final Conference, Pisa, March 30-April 1, 2015, published at http://ceur-ws.org

Lexeme	INF	PRS.3SG	PRS.3PL	PST.PTCP	#
LIVRER 'deliver'	livʁe	livʁ	livʁ	livʁe	4108
RELIER 'link'	ʁəlje	ʁəli	ʁəli	ʁəlje	210
RATISSER 'rake'	ʁatise	ʁatis	ʁatis	ʁatise	22
BÂTIR 'build'	batiʁ	bati	batis	bati	327
TENIR 'hold'	təniʁ	tjɛ̃	tjɛn	təny	37
OUVRIR 'open'	uvʁiʁ	uvʁ	uvʁ	uvɛʁ	8
MOURIR 'die'	muʁiʁ	mœʁ	mœʁ	mɔʁ	1

Table 1: Exemplary paradigms for inflection patterns for 4-cell subparadigms of French verbs (data from *Flexique* — 5% of the lemmas illustrating minor patterns have been excluded)

in the order of magnitude of the overall linguistic exposure of an adult speaker. The distribution strongly suggests that, as speakers get exposed to more words, paradigms fill slowly on average, so that predicting unknown forms stays relevant; at the same time, speakers are massively exposed to multiple forms of the same lexemes, which makes knowledge of joint predictiveness relevant to addressing the PCFP.

A second relevant observation is that speakers do manifest knowledge of joint predictiveness. Although this topic deserves dedicated experimental studies that are beyond the scope of this paper, circumstantial evidence from speech errors is easy to find. One common conjugation error in French (Kilani-Schoch and Dressler, 2005) is to use *mouru* as the past participle of MOURIR, whereas *mouri* is almost never used (140 relevant occurrences of *mouru* in the full *FrWaC* corpus, 0 or *mouri*). This would be surprising if speakers were analogizing from a single paradigm cell: given knowledge of the sole infinitive, *mouri* would be the most likely regularization; given knowledge of some present form, *mour* or *meur* would be expected.[2] Thus the property speakers seem to be sensitive to is the existence of an allomorphic relation between the infinitive and the present stem—hence, employing joint predictiveness from two cells to infer the likely form of the participle.

The final observation is that there are important linguistic generalizations that can only be obtained by looking at joint predictiveness. To supplement the French data presented in the introduction, let us consider a spectacular example from European Portuguese, concerning the prediction of the form of the infinitive from those of the present singular. Table 2 presents relevant data. Because it does not contain a theme vowel, the present 1SG is a bad predictor of the infinitive: a priori, any present 1SG could correspond to a first, second or third conjugation verb. 2SG and 3SG forms are slightly better predictors, as they distinguish first conjugation endings (-ɐʃ,-ɐ) from second/third conjugation endings (-əʃ,-ə); the distinction between the two last conjugations is still neutralized. However, if a verb has a mid prethematic vowel in the 2SG and 3SG, the shape of that vowel is raised to high-mid in the 1SG in the second conjugation (witness RECEBER, RECORRER), and to high in the third conjugation (witness SEGUIR, SUBIR). Whether one sees this phenomenon as the result of a synchronic vowel harmony in the 1SG operating prior to theme vowel deletion (Mateus and d'Andrade, 2000) or as a historical remnant with no synchronic motivation, it remains that on the surface, for verbs with a mid prethematic vowel in the 2SG and 3SG, knowledge of the 1SG disambiguates whether the verb belongs to the second or third conjugation and thus helps predict the infinitive.

3 Quantifying joint predictiveness

To assess the importance of joint predictiveness, we build on previous proposals by (Bonami and Boyé, 2014) and (Bonami and Luís, 2014) on the evaluation of predictiveness from a single paradigm cell, themselves improving on (Acker-

[2] A reviewer points out that if speech errors are due to analogy to the nearest (frequent) neighbor, *mouru* is unsurprising, as *courir* (past participle *couru*) is the most frequent of the verbs whose infinitive is at a minimal edit distance from *mourir*. This assumption however is not plausible. Witness the case of the verb *dire*, whose present 2PL *dites* is very commonly overregularized to *disez*. The most frequent phonological neighbor of *dire* is *lire*; however, according to the *lexique* database (New et al., 2007), *dire* is 8 times more frequent than *lire* in written French, and 17 times in spoken French. It is thus not plausible that analogical regularization is driven by the closest neighbor; rather, it is driven by general patterns applying across lexemes—for instance, *dire* is one of a handful of exceptions to the regular $Xons \sim Xez$ alternation between 1PL and 2PL, that is overwhelmingly prevalent both in type and token frequency.

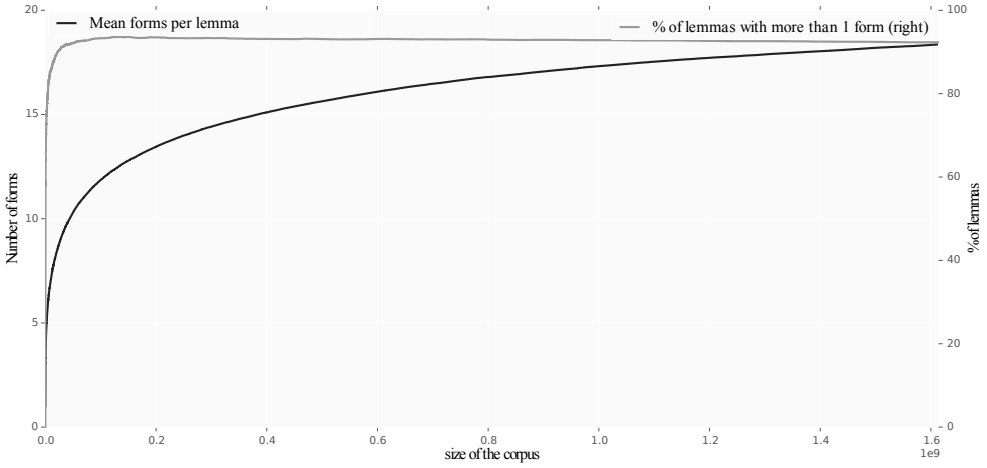

Figure 1: Mean number of forms per lemma and proportion of lemmas with multiple forms as a function of vocabulary size (*FrWaC* corpus)

	INF	1SG	2SG	3SG	1PL	2PL	3PL
LEVAR	lə'vaʁ	'lɛvu	'lɛvɐʃ	'lɛvɐ	lə'vɐmuʃ	lə'vaiʃ	'lɛvɐ̃u
NOTAR	nu'taʁ	'nɔtu	'nɔtɐʃ	'nɔtɐ	nu'tɐmuʃ	nu'taiʃ	'nɔtɐ̃u
RECEBER	rəsə'beʁ	rə'sebu	rə'sɛbəʃ	rə'sɛbə	rəsə'bemuʃ	rəsə'bɐiʃ	rə'sɛbɐ̃i
RECORRER	rəku'reʁ	rə'koru	rə'kɔrəʃ	rə'kɔrə	rəku'remuʃ	rəku'rɐiʃ	rə'kɔrɐ̃i
SEGUIR	sə'giʁ	'sigu	'sɛgəʃ	'sɛgə	sə'gimuʃ	sə'giʃ	'sɛgɐ̃i
SUBIR	su'biʁ	'subu	'sɔbəʃ	'sɔbə	su'bimuʃ	su'biʃ	'sɔbɐ̃i

Table 2: Selected European Portuguese verbs in the infinitive and present indicative

man et al., 2009) and (Ackerman and Malouf, 2013). Specifically, for every pair of paradigm cells A and B, we infer a classification of patterns of alternation relating these two cells. These patterns are then used to define a random variable $A{\sim}B$ over pairs of forms corresponding to the distribution of patterns, and a random variable $A_{A\sim B}$ classifying possible form for A on the basis of the patterns they could possibly instantiate. For instance, going back to the data in Table 1, INF∼PST.PTCP partitions the set of pairs in 5 subsets corresponding to the patterns Xe∼Xe, Xiʁ∼Xi, Xiʁ∼Xy, Xʁiʁ∼Xɛʁ and Xuʁiʁ∼Xɔʁ, while $\text{INF}_{\text{INF}\sim\text{PST.PTCP}}$ partitions the set of infinitive forms in 4 sets, depending on whether they end in -e, -uʁiʁ, -Vʁiʁ with $V \neq$ u, or -Xiʁ with $X \neq$ ʁ.

$H(A{\sim}B \mid A_{A\sim B})$, the conditional entropy of the pattern relating A and B given relevant features of the form filling A, evaluates how well A predicts B.

Crucial to this computation is the choice of a strategy of exhaustive classification of patterns of alternation between pairs of forms. Since the design of an algorithm finding an optimal such classification from raw data is an open research question,[3] we opportunistically use the algorithm sketched in (2) that we know to give satisfactory results for the languages at hand.

(2) a. For any pair of strings $\langle\phi_1, \phi_2\rangle$, find strings $\alpha, \gamma, \beta_1, \beta_2, \delta_1$ and δ_2 such that $\phi_1 = \alpha\beta_1\gamma\delta_1$ and $\phi_2 = \alpha\beta_2\gamma\delta_2$, where β_1 and β_2 have the same length; segments in β_1 and β_2 (resp. δ_1 and δ_2) match in category (vowel vs. consonant), starting from the left; and the length of α is maximal. Classify the pair as instantiating pattern $[X\beta_1 Y\delta_1{\sim}X\beta_2 Y\delta_2 \;/\; \alpha\_\gamma\_]$.

b. For all patterns instantiating the same alternation $[x{\sim}y \,/\, \alpha_1\_\gamma_1\_], \ldots, [x{\sim}y \,/\, \alpha_n\_\gamma_n\_]$, determine maximally specific feature descriptions of sets of strings $\{\alpha_1, \ldots, \alpha_n\}$

[3] The problem can be presented as that of finding, for any set of pairs of forms, a minimal set of subsequential finite-state transducers such that one of the transducers maps each input form to the correct output. Even if that problem were solved, it is entirely possible for there to be more than one such minimal set, leading to competing classifications of the pairs and thus to different assessments of predictiveness.

and $\{\gamma_1, \ldots, \gamma_n\}$, using (Albright, 2002)'s Minimal Generalization strategy.

Joint predictiveness can then be assessed looking at joint random variables: predicting C from A and B is evaluated by (3): we assess the uncertainty associated with predicting both the pattern relating A to C and the pattern relating B to C, given knowledge of relevant properties of A, relevant properties of B, and the pattern relating A and B. Notice that this easily generalizes to prediction given joint knowledge of n different cells.

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

Table 3 shows the average entropy from 1 or 2 cells for 5000 French verbs and 2000 European Portuguese verbs respectively.[4] In both languages, knowing a second cell significantly reduces uncertainty on average.

# of predictor cells	French	Portuguese
1	0.1670	0.1649
2	0.0540	0.0818

Table 3: Average conditional entropy when predicting from 1 or 2 cells

4 Principal part systems

A system of principal parts is a set of paradigm cells such that knowledge of the forms filling these cells is sufficient to derive the rest of the paradigm (Hockett, 1967; Matthews, 1972; Finkel and Stump, 2007; Stump and Finkel, 2013).[5] The validity of a principal part system thus rests on the existence of systematic categorical joint predictiveness; and the evaluation method outlined in the preceding section may be used to infer sets of principal parts.

Exploring this issue on the European Portuguese dataset, we find that there are 177 such systems for Portuguese. All these systems include a form with a 3-way contrast of theme vowels, such as the infinitive, and a form with stress on the prethematic vowel, such as the present 3SG. This corresponds to the observation in (Bonami and Luís, 2014) that such pairs of cells have complementary predictive power. The sheer number of alternative principal part systems highlights the arbitrariness of the choice of a particular set of principal parts (Matthews, 1972; Ackerman et al., 2009; Blevins, forthcoming).

Turning to French, we found no set of principal parts of cardinality 2, as already observed by (Stump and Finkel, 2013). This is testament to the prevalence of erratic stem allomorphy in French conjugation, leading to numerous situations of unpredictibility local to a small subpart of the paradigm (Bonami and Boyé, 2002). However, this observation should be modalized in two ways.

First, our method yields 396 sets of principal parts of cardinality 3, whereas (Stump and Finkel, 2013) found no set of cardinality smaller than 5. This difference seems to be due to the fact that, under the methodology used here, the applicability of a pattern of alternation is sensitive to phonotactic properties of the stem (thanks to the use of the Minimal Generalization strategy in (2b)), whereas (Stump and Finkel, 2013) only look at exponence. Arguably then, the present method provides a superior evaluation of the diagnostic value of paradigm cells.

Second, although there is no pair of cells with categorical diagnostic value, some come very close. There are 25 pairs of cells (among which pairs of very frequent cells such as the present 3PL and the infinitive) such that predicting any other cell from this pair yields an entropy below 0.005. This means that given knowledge of these two cells, trying to guess any other cell will be about as hard as predicting an event with a 99.95% probability of occurrence.[6] This casts doubts both on the pedagogical value of categorical principal part systems and on the usefulness of principal part systems, as opposed to graded evaluations of joint predictiveness, for the study of morphological competence.

Acknowledgments

This work was partially supported by a public grant overseen by the French National Research

[4] The French dataset was extracted from *Flexique* (Bonami et al., 2014). The Portuguese dataset was derived from the University of Coimbra pronunciation dictionary (Veiga et al., 2012) for the purpose of (Bonami and Luís, 2013).

[5] We focus here on traditional static principal part systems. See (Bonami and Boyé, 2007; Finkel and Stump, 2007; Stump and Finkel, 2013) for alternative formulations of the notion of principal part where different sets of paradigm cells serve as predictor depending on the lexeme.

[6] If X is a binary random variable one of whose values has a probability of 0.9995, $H(X) > 0.0062$.

Agency (ANR) as part of the "Investissements d'Avenir" program (reference: ANR-10-LABX-0083).

References

[Ackerman and Malouf2013] Farrell Ackerman and Robert Malouf. 2013. Morphological organization: the low conditional entropy conjecture. *Language*, 89:429–464.

[Ackerman et al.2009] Farrell Ackerman, James P. Blevins, and Robert Malouf. 2009. Parts and wholes: implicative patterns in inflectional paradigms. In James P. Blevins and Juliette Blevins, editors, *Analogy in Grammar*, pages 54–82. Oxford University Press, Oxford.

[Albright2002] Adam C. Albright. 2002. *The Identification of Bases in Morphological Paradigms*. Ph.D. thesis, University of California, Los Angeles.

[Baroni et al.2009] Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: A collection of very large linguistically processed web-crawled corpora. In *Language Resources and Evaluation*, volume 43, pages 209–226.

[Blevinsforthcoming] James P. Blevins. forthcoming. *Word and Paradigm Morphology*. Oxford University Press, Oxford.

[Bonami and Boyé2002] Olivier Bonami and Gilles Boyé. 2002. Suppletion and stem dependency in inflectional morphology. In Franck Van Eynde, Lars Hellan, and Dorothee Beerman, editors, *The Proceedings of the HPSG '01 Conference*, pages 51–70. CSLI Publications, Stanford.

[Bonami and Boyé2007] Olivier Bonami and Gilles Boyé. 2007. Remarques sur les bases de la conjugaison. In Elisabeth Delais-Roussarie and Laurence Labrune, editors, *Des sons et des sens*, pages 77–90. Hermès, Paris.

[Bonami and Boyé2014] Olivier Bonami and Gilles Boyé. 2014. De formes en thèmes. In Florence Villoing, Sarah Leroy, and Sophie David, editors, *Foisonnements morphologiques. Etudes en hommage à Françoise Kerleroux*, pages 17–45. Presses Universitaires de Paris Ouest.

[Bonami and Luís2013] Olivier Bonami and Ana R. Luís. 2013. Causes and consequences of complexity in portuguese verbal paradigms. In *9th Mediterranean Morphology Meeting*, Dubrovnik, septembre.

[Bonami and Luís2014] Olivier Bonami and Ana R. Luís. 2014. Sur la morphologie implicative dans la conjugaison du portugais : une étude quantitative. *Mémoires de la Société de Linguistique de Paris*, 22:111–151.

[Bonami et al.2014] Olivier Bonami, Gauthier Caron, and Clément Plancq. 2014. Construction d'un lexique flexionnel phonétisé libre du français. In Franck Neveu, Peter Blumenthal, Linda Hriba, Annette Gerstenberg, Judith Meinschaefer, and Sophie Prévost, editors, *Actes du quatrième Congrès Mondial de Linguistique Française*, pages 2583–2596.

[Finkel and Stump2007] Raphael Finkel and Gregory T. Stump. 2007. Principal parts and morphological typology. *Morphology*, 17:39–75.

[Hockett1967] Charles F. Hockett. 1967. The Yawelmani basic verb. *Language*, 43:208–222.

[Kilani-Schoch and Dressler2005] Marianne Kilani-Schoch and Wolfgang Dressler. 2005. *Morphologie naturelle et flexion du verbe français*. Gunter Narr Verlag, Tübingen.

[Mateus and d'Andrade2000] Maria Helena Mateus and Ernesto d'Andrade. 2000. *ThePhonology of Portuguese*. Oxford University Press, Oxford.

[Matthews1972] P. H. Matthews. 1972. *Inflectional Morphology. A Theoretical Study Based on Aspects of Latin Verb Conjugation*. Cambridge University Press, Cambridge.

[New et al.2007] Boris New, Marc Brysbaert, Jean Veronis, and Christophe Pallier. 2007. The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics*, 28:661–677.

[Sagot2010] Benoît Sagot. 2010. The *Lefff*, a freely available and large-coverage morphological and syntactic lexicon for French. In *Proceedings of LREC 2010*.

[Simsforthcoming] Andrea Sims. forthcoming. *Inflectional defectiveness*. Cambridge University Press, Cambridge.

[Stump and Finkel2013] Gregory T. Stump and Raphael Finkel. 2013. *Morphological Typology: From Word to Paradigm*. Cambridge University Press, Cambridge.

[Thymé et al.1994] Ann Thymé, Farrell Ackerman, and Jeff Elman. 1994. Finnish nominal inflection: Paradigmatic patterns and token analogy. In Susan D. Lima, Roberta Corrigan, and Gregory K. Iverson, editors, *The Reality of Linguistic Rules*. John Benjamins.

[Veiga et al.2012] Arlindo Oliveira da Veiga, Sara Candeias, and 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.

[Wurzel1984] Wolfgang Ulrich Wurzel. 1984. *Flexionsmorphologie und Natürlichkeit. Ein Beitrag zur morphologischen Theoriebildung*. Akademie-Verlag, Berlin. Translated as (Wurzel, 1989).

[Wurzel1989] Wolfgang Ulrich Wurzel. 1989. *Inflectional Morphology and Naturalness*. Kluwer, Dordrecht.