Causes and consequences of complexity in Portuguese verbal paradigms

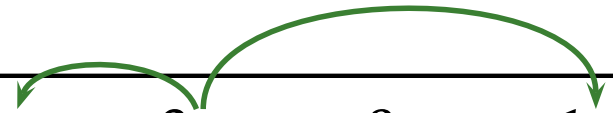
Olivier Bonami[1] & Ana R. Luís[2]

[1]U. Paris-Sorbonne, IUF, Laboratoire de Linguistique Formelle
[2]U. de Coimbra, CELGA

Ninth Mediterranean Morphology Meeting, Dubrovnik, September 2013

Introduction

- Implicative structure of paradigms (Wurzel, 1984): the form filling a cell in the paradigm provides information on the forms filling other cells.



	1SG	2SG	3SG	1PL	2PL	3PL
FICAR	fˈiku	fˈikɐʃ	fˈikɐ	fikˈɐmuʃ	fikˈaiʃ	fˈikɐ̃ũ
VIVER	vˈivu	vˈivəʃ	vˈivə	vivˈemuʃ	vivˈɐiʃ	vˈivɐ̃ĩ
IMPRIMIR	ĩpɾˈimu	ĩpɾˈiməʃ	ĩpɾˈimə	ĩpɾimˈimuʃ	ĩpɾimˈiʃ	ĩpɾˈimɐ̃ĩ

Indicative present of 3 European Portuguese fully regular verbs

- Two basic questions about implicative structure:
  - How much information is provided?
  - What aspects of the form provide that information?
  - ☞ In some cases, segmentable morphs
  - ☞ In other cases, other types of systematic covariation

Instrumented Item and Pattern morphology

- Item and pattern (IPa) morphology (Blevins, to appear):
  - Focuses on patterns of alternation among forms filling cells in a paradigm
  - Not morph-centric: while patterns may consist of morph insertion/deletion/substitution, this is not necessary to their identification and use.
- *Quantitative* IPa
  - Heavy use of quantitative methods, including information-theoretic measures
    - ★ (Ackerman et al., 2009) and later work, e.g. Sims (2010); Bonami et al. (2011, 2012); Ackerman and Malouf (in press); Blevins (to appear)
- *Instrumented* IPa:
  - Based on large scale inflected lexica
  - Automatic inference and analysis of patterns using simple, opportunistic methods
  - Focus on coverage and precision of empirical generalizations

Structure

The dataset

- Full paradigms of the 2000 most frequent verbs in the CETEMPúblico corpus (Santos and Rocha, 2001)
- Fully transcribed in IPA on the basis of the U. of Coimbra pronunciation dictionary (Veiga et al., 2012)
  - ☞ Unique transcription for each paradigm cell of each lexeme, which entails a certain amount of idealization.
  - ☞ The transcription corresponds most closely to slow, careful speech in central Portugal.

Finding patterns

The general problem

- A basic building block for the kind of investigation at hand is a method for identifying patterns of alternation.
- Which method one uses has dramatic effects on the ensuing analyses.
- A general, language-independent method is hard to define and computationally expensive.
  - For a large (>1000) set of pairs of forms, find the smallest set of subsequential finite-state transducers relating these pairs.
- Opportunistic strategy: we use prior knowledge of the system to decide on a reasonably simple method that we suspect won't miss important patterns.
- For Portuguese: We know that inflection is suffixal, but that the last vowel of the stem (the prethematic vowel) often alternates.

An opportunistic solution

- Over 2000 pairs of cells:
  1. Identify 'quasi-suffixes': what remains if one drops the longest identical initial substrings
  2. Fuse patterns with covarying central consonant cluster
  3. Record of common phonotactic properties of the nonalternating parts, using a Minimal Generalization strategy (Albright, 2002)

lexeme	PRS.1SG	PRS.1PL	step 1	step 2
FICAR	fˈiku	fikˈɐmuʃ	Xu ⇌ Xɐmuʃ	Xu ⇌ Xɐmuʃ
PASSAR	pˈasu	pɐsˈɐmuʃ	Xasu ⇌ Xɐsɐmuʃ	Xa Yu ⇌ Xɐ Yɐmuʃ
PAGAR	pˈagu	pɐgˈɐmuʃ	Xagu ⇌ Xɐgɐmuʃ	Xa Yu ⇌ Xɐ Yɐmuʃ
CHEGAR	ʃegu	ʃəgˈɐmuʃ	Xegu ⇌ Xəgɐmuʃ	Xe Yu ⇌ Xə Yɐmuʃ
MOSTRAR	mˈɔʃtɾu	muʃtɾˈɐmuʃ	Xɔ Yu ⇌ Xu Yɐmuʃ	Xɔ Yu ⇌ Xu Yɐmuʃ

- Sample phonotactic condition:
  XC_1eC_2u ⇌ XC_1əC_2ɐmuʃ,
  where X is any sequence, C_1 : [+cons, −voc], C_2 : [+cons, −voc, −lat]

Using patterns

Using patterns

Induction of inflection classes

Inflection classes as vectors of patterns

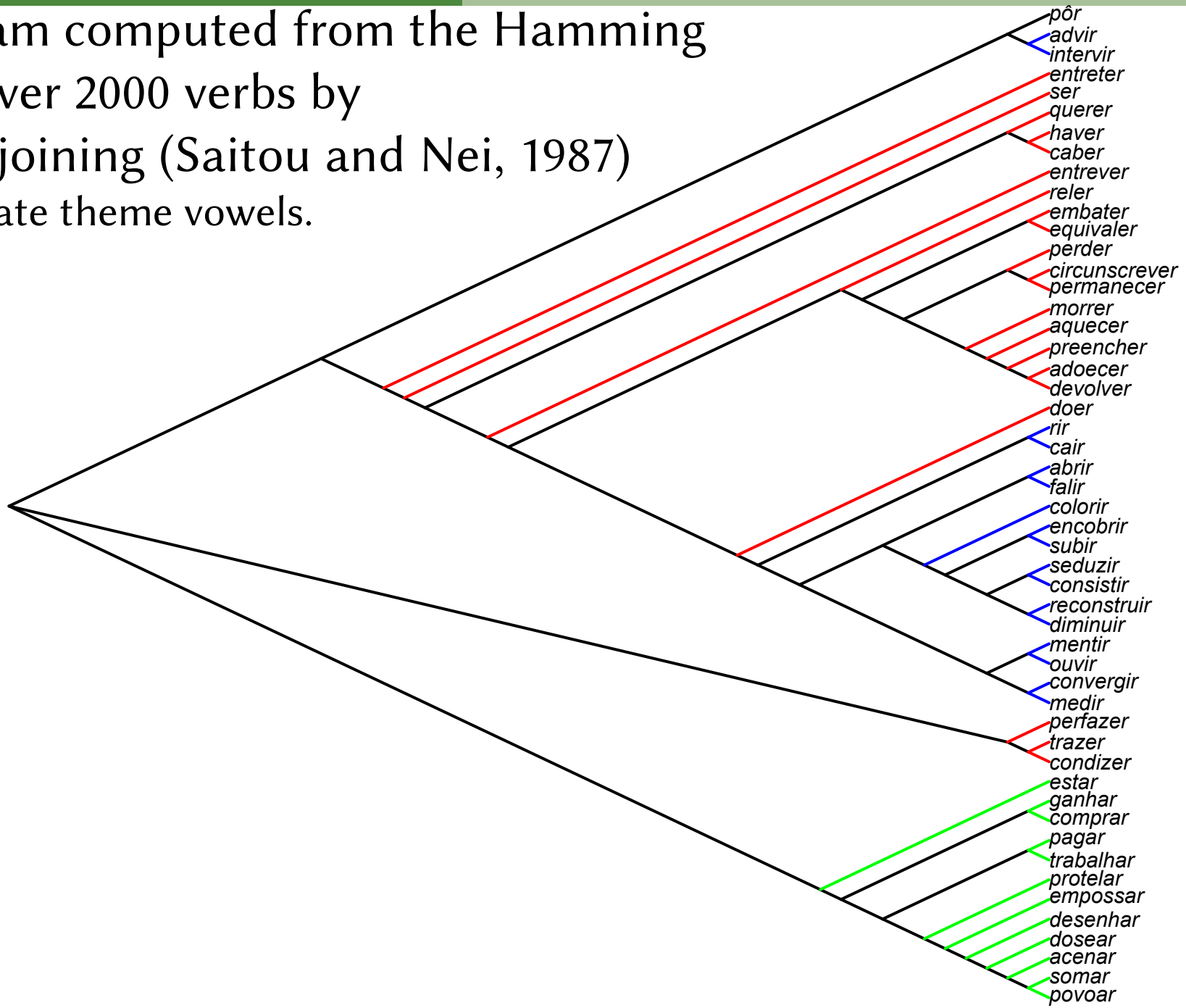
- Each lexeme is now characterized by the vector of patterns it uses to relate each pair of cells in the paradigm

lexeme	⟨1SG,2SG⟩	⟨1SG,3SG⟩	⟨1SG,1PL⟩	⟨1SG,2PL⟩	⟨1SG,3PL⟩	⟨2SG,3SG⟩	...
ficar	X u ⇌ X ɐ ʃ	X u ⇌ X ɐ	X u ⇌ X ɐ m u ʃ	X u ⇌ X a i	X u ⇌ X ɐ̃ ũ	X ɐ ʃ ⇌ X ɐ	...
viver	X u ⇌ X ə ʃ	X u ⇌ X ə	X u ⇌ X e m u ʃ	X u ⇌ X ɐ i	X u ⇌ X ɐ̃ ĩ	X ə ʃ ⇌ X ə	...
imprimir	X u ⇌ X ə ʃ	X u ⇌ X ə	X u ⇌ X i m u ʃ	X u ⇌ X i	X u ⇌ X ɐ̃ ĩ	X ə ʃ ⇌ X ə	...
...	...	...	...	...	...	...	...

- Gives us a very fine-grained definition of inflection class: if two lexemes have the exact same vector of patterns, then they definitely belong to the same inflection class.
- Gives us a very simple notion of distance between inflection classes: the Hamming distance between the two vectors
  - ☞ The number of pairs of cells for which the two vectors differ
- This distance can then be used with off-the-shelf clustering algorithms to produce groupings in superclasses

Dendrogram computed from the Hamming distance over 2000 verbs by Neighbor-joining (Saitou and Nei, 1987)

Colors indicate theme vowels.



The virtues of the method

- We get a classification of overall inflection patterns that is
  - Entirely automated
  - Not dependent on fine decisions of segmentation
  - Easily criticizable
- Similar in spirit (but not in execution) to Brown and Evans (2012)
  - Where Brown and Evans (2012) use compression distance, we use distance between (vectors of) patterns
    - ☞ Two lexemes with identical inflection have a distance of zero
  - Where Brown and Evans (2012) use a sophisticated clustering method, we use a very simple one
    - ★ Easier to understand what the clustering method really does

Using patterns

Partitioning paradigms

Fully interpredictable cells

- Two patterns relating cells c and c' are mutually exclusive if they impose incompatible constraints on both c and c'.

PRS.1SG ⇌ PRS.1PL
*X*u ⇌ *X*ɐmuʃ
*X*u ⇌ *X*imuʃ
not exclusive

PRS.3SG ⇌ PRS.2PL
*X*ɐ ⇌ *X*aiʃ
*X*a*C*ɐ ⇌ *X*ɐ*C*aiʃ
not exclusive

PRS.3SG ⇌ PRS.3PL
Xɐ ⇌ Xɐ̃ũ
Xə ⇌ Xɐ̃ĩ
exclusive

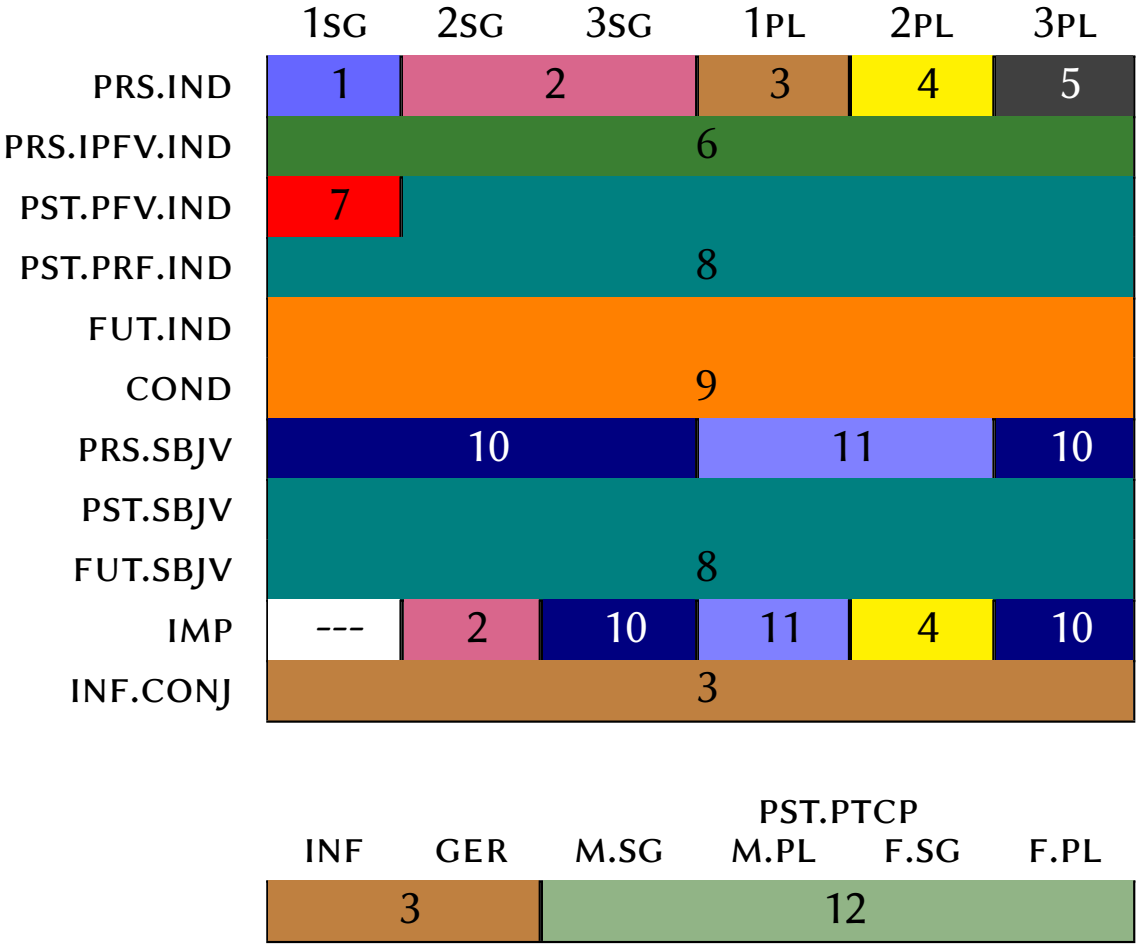
- NB: the existence of non-exclusive patterns sometimes leads to genuine ambiguity, even given perfect knowledge of the lexicon

lexeme	PRS.1SG	PRS.1PL
girar	ʒiɾu	ʒiɾɐmuʃ
gerir	ʒiɾu	ʒəɾimuʃ

- If all patterns used to relate those two cells are pairwise mutually exclusive, then these two cells are fully interpredictable.

Partitioning the paradigm

- We can now partition the paradigm into zones of perfect interpredictibility (Ackerman et al.'s (2009) 'alliances of forms')



Discussion

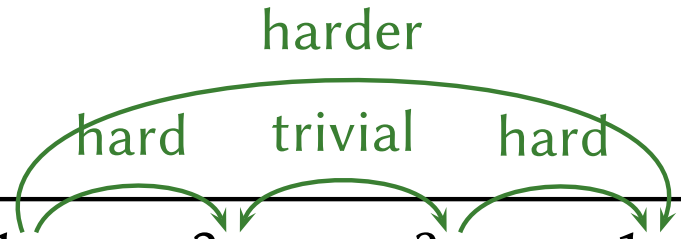
- The partition highlights morphomic patterns
- Reminiscent of Pirelli and Battista's (2000) 'Overall distribution schema' or Bonami and Boyé's (2002) 'Stem spaces'
  - ☞ See Bonami and Boyé (to appear) for systematic discussion of differences
- However here predictibility is defined purely on the basis of full forms,
  - Does not presuppose any disputable decision on segmentation into stems and exponents (Boyé, 2000; Spencer, 2012; Stump and Finkel, 2013)
- A practical consequence of the identification of a partition is that we can focus on a distillation of the paradigm (Stump and Finkel, 2013): just pick one cell from each cell in the partition, and forget about the others.

Using patterns

Gradient predictiveness

Gradient predictiveness

- Predictiveness is clearly a gradient property: some predictions are categorical, others are very reliable, others still have little reliability.

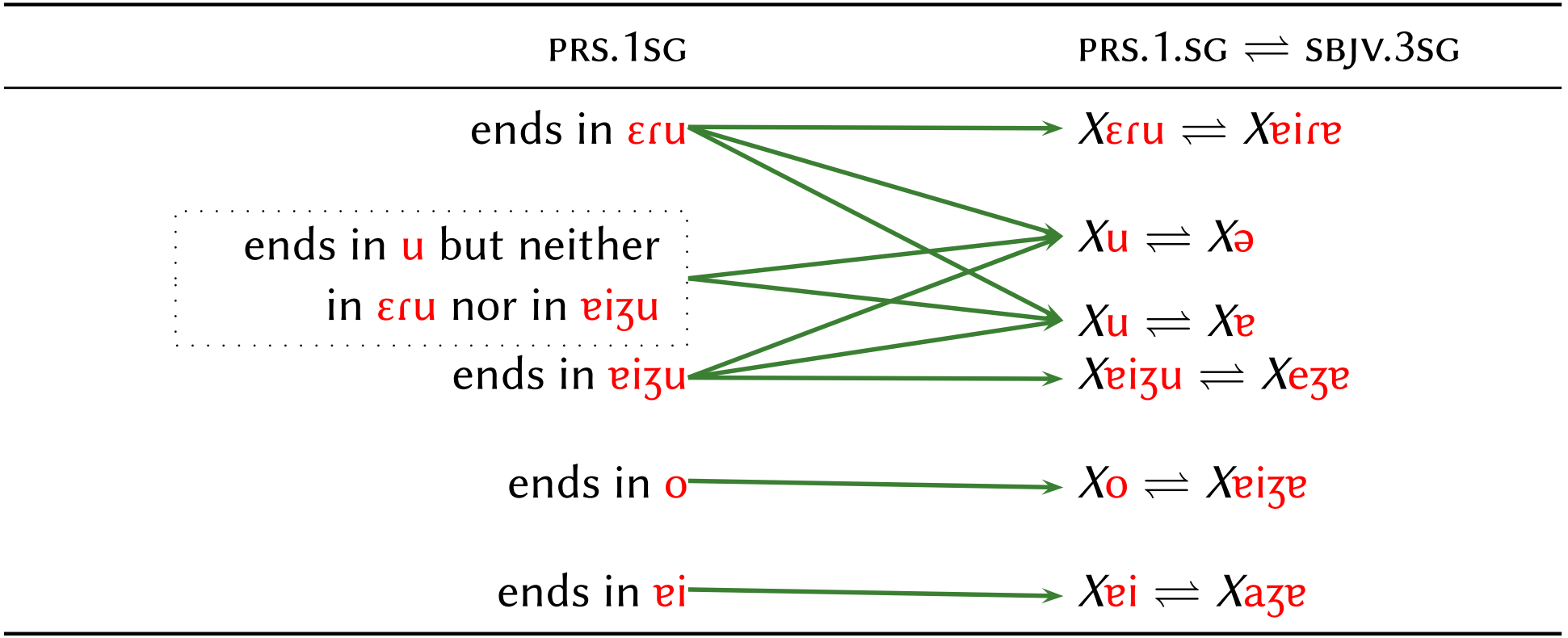


	INF	1SG	2SG	3SG	1PL	2PL	3PL
FICAR	fikˈaɾ	fˈiku	fˈikɐʃ	fˈikɐ	fikˈɐmuʃ	fikˈaiʃ	fˈikɐ̃ũ
VIVER	vivˈeɾ	vˈivu	vˈivəʃ	vˈivə	vivˈemuʃ	vivˈɐiʃ	vˈivɐ̃ĩ
IMPRIMIR	ĩpɾimˈiɾ	ĩpɾˈimu	ĩpɾˈiməʃ	ĩpɾˈimə	ĩpɾimˈimuʃ	ĩpɾimˈiʃ	ĩpɾˈimɐ̃ĩ

- Capturing this gradience motivates the use of conditional entropy to model implicative relations in paradigms (Ackerman et al., 2009).

Measuring predictiveness

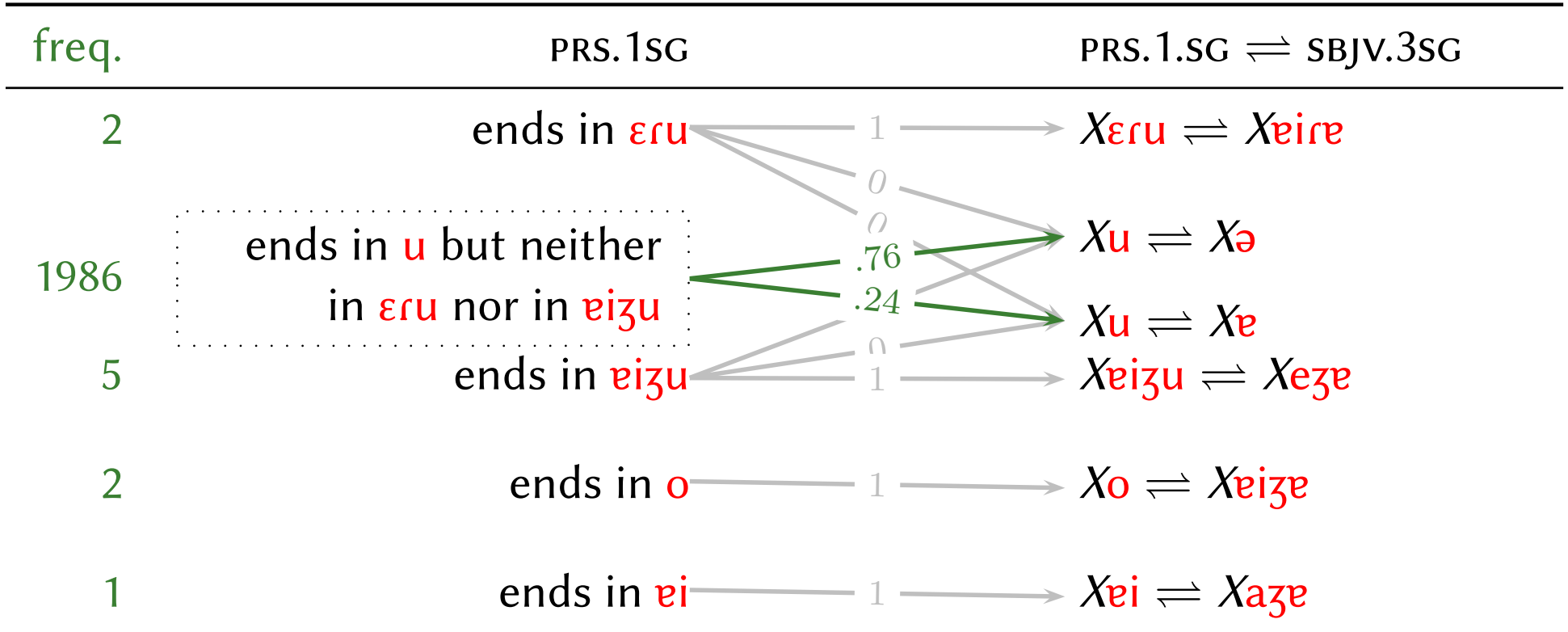
- To evaluate how c predicts c', we need to identify, for each possible form filling c, the set of patterns that could relate c to c'.



Distribution of patterns relating PRS.1SG to SBJV.3SG for 1996 verbs

Measuring predictiveness

- To quantify predictiveness, we want to evaluate the likelihood of each possibility.
  - Approximate probabilities on the basis of type frequency
  - Use conditional entropy of a pattern as a measure of predictiveness
  - $H(\text{pattern} \mid \textsc{prs.1sg}) \approx -\frac{1986}{1996}(0.76 \log_2 0.76 + 0.24 \log_2 0.24) \approx 0.7855$



Distribution of patterns relating PRS.1SG to SBJV.3SG for 1996 verbs

Raw results

- Systematic application of this method to a distillation of the paradigm:

	INF	PRS.IND.1SG	PRS.IND.3SG	PRS.IND.2PL	PRS.IND.3PL	PST.IPFV.IND.3SG	PST.PFV.IND.1SG	PST.PFV.IND.3SG	FUT.IND.3SG	PRS.SBJV.3SG	PRS.SBJV.2PL	PST.PTCP
INF	0	0.3427	0.3032	0.0541	0.3706	0.0163	0.0163	0.0263	0	0.3427	0.0295	0.0121
PRS.IND.1SG	0.6990	0	0.6366	0.6990	0.6594	0.6832	0.6761	0.6990	0.6990	0.7855	0.6821	0.6678
PRS.IND.3SG	0.2044	0.0819	0	0.2044	0.0041	0.0856	0.0856	0.2042	0.2382	0.0848	0.1461	0.0837
PRS.IND.2PL	0.0316	0.3422	0.3574	0	0.3605	0.0316	0.0312	0.0312	0.0312	0.3422	0.0307	0.0311
PRS.IND.3PL	0.2124	0.1012	0.0059	0.2124	0	0.0859	0.0856	0.2084	0.2102	0.0936	0.1469	0.0838
PST.IPFV.IND.3SG	0.2184	0.4136	0.3755	0.2300	0.3812	0	0	0.2120	0.2011	0.4136	0.0609	0.0094
PST.PFV.IND.1SG	0.2594	0.4102	0.3720	0.2525	0.3773	0.0471	0	0.2592	0.2464	0.4102	0.1062	0.0563
PST.PFV.IND.3SG	0.0030	0.3316	0.3498	0.0136	0.3521	0	0	0	0.0030	0.3316	0.0016	0.0030
FUT.IND.3SG	0.0333	0.3441	0.3776	0.0650	0.3699	0.0533	0.0245	0.0345	0	0.3441	0.0444	0.0203
PRS.SBJV.3SG	0.1894	0.0000	0.0657	0.1894	0.0632	0.1350	0.1350	0.1894	0.1894	0	0.0917	0.1332
PRS.SBJV.2PL	0.2049	0.3912	0.4138	0.2049	0.4187	0.0483	0.0483	0.2049	0.1836	0.3912	0	0.0483
PST.PTCP	0.2109	0.4218	0.3431	0.2133	0.3806	0.0191	0.0191	0.2209	0.1970	0.4218	0.0657	0

- We can now look for patterns in this table and look for causes of particular entropy values in the log

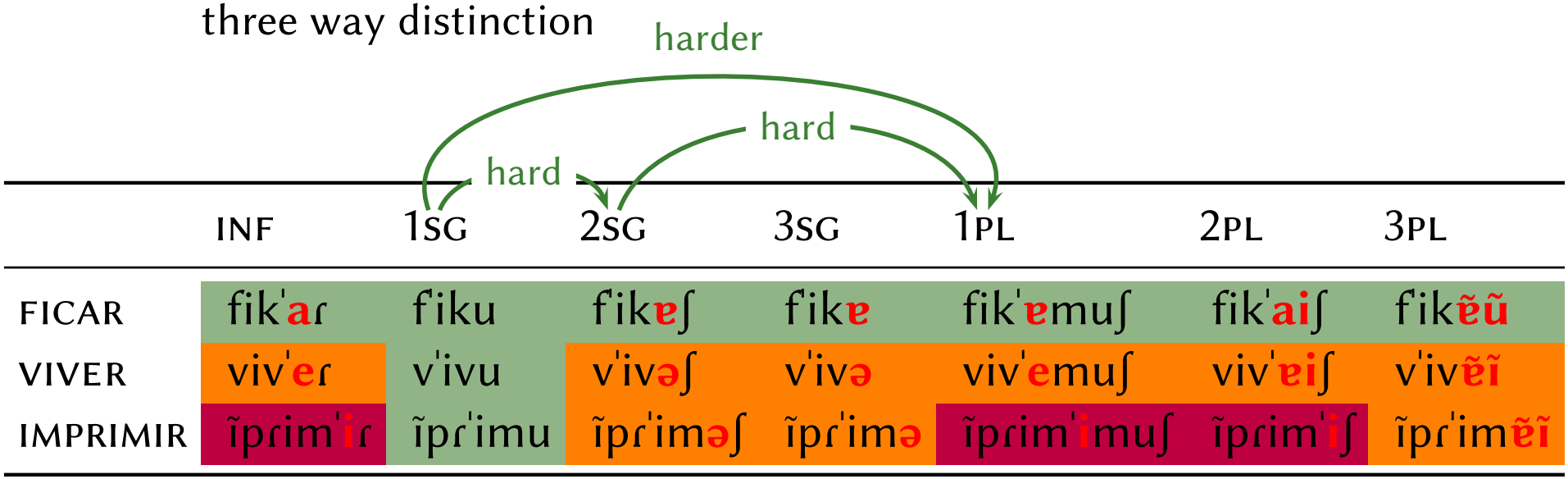
```
****************************************
        PresConj3 ==> PresIndic1
****************************************
Inferring rules...
ə --> u / X[p,t,k,b,d,g,f,s,ʃ,v,z,ʒ,m,n,ɲ,l,ʎ,l~,r,ɾ,i,ɨ,e,ɛ,ə,o,ɔ,u,ĩ,ẽ,õ,ũ] ___ # 1516
ɐ --> u / X[p,t,k,b,d,g,f,s,ʃ,v,z,ʒ,m,n,ɲ,l,ʎ,l~,r,ɾ,i,ɨ,e,ɛ,ə,o,ɔ,u,ĩ,ẽ,õ,ũ] ___ # 470
ɐiɾɐ --> ɛɾu / Xk ___ # 2
eʒɐ --> ɐiʒu / Xv ___ # 5
ɐiʒɐ --> o / X[t,s] ___ # 2
aʒɐ --> ɐi / # ___ # 1
done.

class 1 ( rəkɐiɾɐ ~ rəkɛɾu ):  2 members
ɐ --> u / X[p,t,k,b,d,g,f,s,ʃ,v,z,ʒ,m,n,ɲ,l,ʎ,l~,r,ɾ,i,ɨ,e,ɛ,ə,o,ɔ,u,ĩ,ẽ,õ,ũ] ___ # : 0
ɐiɾɐ --> ɛɾu / Xk ___ # : 2 (requerer, etc.)
local conditional entropy: -0.0
-----
class 2 ( aʒɐ ~ ɐi ):  1 members
ɐ --> u / X[p,t,k,b,d,g,f,s,ʃ,v,z,ʒ,m,n,ɲ,l,ʎ,l~,r,ɾ,i,ɨ,e,ɛ,ə,o,ɔ,u,ĩ,ẽ,õ,ũ] ___ # : 0
aʒɐ --> ɐi / # ___ # : 1 (haver, etc.)
local conditional entropy: -0.0
-----
class 3 ( mənuʃpɾɛzə ~ mənuʃpɾɛzu ):  1516 members
ə --> u / X[p,t,k,b,d,g,f,s,ʃ,v,z,ʒ,m,n,ɲ,l,ʎ,l~,r,ɾ,i,ɨ,e,ɛ,ə,o,ɔ,u,ĩ,ẽ,õ,ũ] ___ # : 1516 (menos
local conditional entropy: -0.0
-----
```

Causes of low predictiveness

Theme vowels

- The fate of Latin theme vowels in Portuguese is variable:
  - In the PRS.1SG, complete loss of theme vowel distinctions
  - Many paradigm cells carry some suffixal material coding a two-way distinction between first and other conjugations
  - Still other cells keep a 3 way distinction, although that does not entail that the vowels are unaltered
- Consequences:
  - The PRS.1SG is a bad predictor of all other paradigm cells
  - Cells making a two way distinction are bad predictors of cells making a three way distinction



Theme vowels

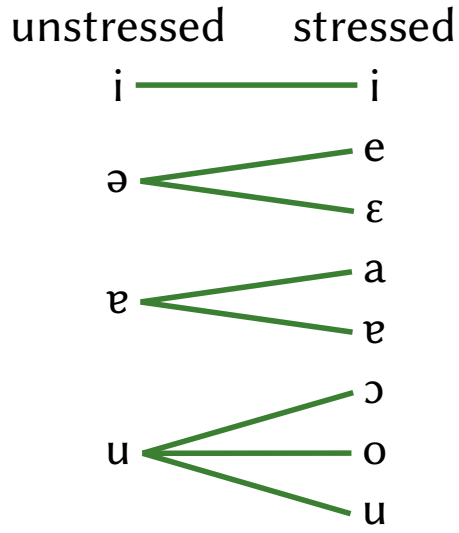
	INF	PRS.IND.1SG	PRS.IND.3SG	PRS.IND.2PL	PRS.IND.3PL	PST.IPFV.IND.3SG	PST.PFV.IND.1SG	PST.PFV.IND.3SG	FUT.IND.3SG	PRS.SBJV.3SG	PRS.SBJV.2PL	PST.PTCP
INF	0	0.3427	0.3032	0.0541	0.3706	0.0163	0.0163	0.0263	0	0.3427	0.0295	0.0121
PRS.IND.1SG	0.6990	0	0.6366	0.6990	0.6594	0.6832	0.6761	0.6990	0.6990	0.7855	0.6821	0.6678
PRS.IND.3SG	0.2044	0.0819	0	0.2044	0.0041	0.0856	0.0856	0.2042	0.2382	0.0848	0.1461	0.0837
PRS.IND.2PL	0.0316	0.3422	0.3574	0	0.3605	0.0316	0.0312	0.0312	0.0312	0.3422	0.0307	0.0311
PRS.IND.3PL	0.2124	0.1012	0.0059	0.2124	0	0.0859	0.0856	0.2084	0.2102	0.0936	0.1469	0.0838
PST.IPFV.IND.3SG	0.2184	0.4136	0.3755	0.2300	0.3812	0	0	0.2120	0.2011	0.4136	0.0609	0.0094
PST.PFV.IND.1SG	0.2594	0.4102	0.3720	0.2525	0.3773	0.0471	0	0.2592	0.2464	0.4102	0.1062	0.0563
PST.PFV.IND.3SG	0.0030	0.3316	0.3498	0.0136	0.3521	0	0	0	0.0030	0.3316	0.0016	0.0030
FUT.IND.3SG	0.0333	0.3441	0.3776	0.0650	0.3699	0.0533	0.0245	0.0345	0	0.3441	0.0444	0.0203
PRS.SBJV.3SG	0.1894	0.0000	0.0657	0.1894	0.0632	0.1350	0.1350	0.1894	0.1894	0	0.0917	0.1332
PRS.SBJV.2PL	0.2049	0.3912	0.4138	0.2049	0.4187	0.0483	0.0483	0.2049	0.1836	0.3912	0	0.0483
PST.PTCP	0.2109	0.4218	0.3431	0.2133	0.3806	0.0191	0.0191	0.2209	0.1970	0.4218	0.0657	0

- The entropy associated with the prediction of 3-way distinctions from 2-way distinction is not very high

☞ Explanation: 1st conjugation verbs make up 76% of our data

Prethematic vowels

- Portuguese oral vowels exhibit stress-conditioned alternations
- While some cells have a stressed prethematic vowel, in other cells stress falls elswhere, typically on the theme vowel.
- This causes uncertainty when trying to predict stressed vowels from unstressed ones.



	INF	1SG	2SG	3SG	1PL	2PL	3PL
CHEGAR	ʃəgˈar	ʃˈegu	ʃˈegɐʃ	ʃˈegɐ	ʃəgˈɐmuʃ	ʃəgˈaiʃ	ʃˈegɐ̃ũ
COMEÇAR	kumə̱sˈar	kumˈɛ̱su	kumˈɛ̱sɐʃ	kumˈɛ̱sɐ	kumə̱sˈɐmuʃ	kumə̱sˈaiʃ	kumˈɛ̱sɐ̃ũ
PAGAR	pɐ̱gˈar	pˈa̱gu	pˈa̱gɐʃ	pˈa̱gɐ	pɐ̱gˈɐmuʃ	pɐ̱gˈaiʃ	pˈa̱gɐ̃ũ
CHAMAR	ʃɐ̱mˈar	ʃˈɐ̱mu	ʃˈɐ̱mɐʃ	ʃˈɐ̱mɐ	ʃɐ̱mˈɐmuʃ	ʃɐ̱mˈaiʃ	ʃˈɐ̱mɐ̃ũ
JOGAR	ʒu̱gˈar	ʒˈɔ̱gu	ʒˈɔ̱gɐʃ	ʒˈɔ̱gɐ	ʒu̱gˈɐmuʃ	ʒu̱gˈaiʃ	ʒˈɔ̱gɐ̃ũ
MUDAR	mu̱dˈar	mˈu̱du	mˈu̱dɐʃ	mˈu̱dɐ	mu̱dˈɐmuʃ	mu̱dˈaiʃ	mˈu̱dɐ̃ũ

Prethematic vowels

- The difficulty of predicting the quality of stressed prethematic vowels is the second highest contributor of entropy

	INF	PRS.IND.1SG	PRS.IND.3SG	PRS.IND.2PL	PRS.IND.3PL	PST.IPFV.IND.3SG	PST.PFV.IND.1SG	PST.PFV.IND.3SG	FUT.IND.3SG	PRS.SBJV.3SG	PRS.SBJV.2PL	PST.PTCP
INF	0	0.3427	0.3032	0.0541	0.3706	0.0163	0.0163	0.0263	0	0.3427	0.0295	0.0121
PRS.IND.1SG	*0.6990*	0	*0.6366*	*0.6990*	*0.6594*	*0.6832*	*0.6761*	*0.6990*	*0.6990*	*0.7855*	*0.6821*	*0.6678*
PRS.IND.3SG	*0.2044*	0.0819	0	*0.2044*	0.0041	0.0856	0.0856	*0.2042*	*0.2382*	0.0848	0.1461	0.0837
PRS.IND.2PL	0.0316	0.3422	0.3574	0	0.3605	0.0316	0.0312	0.0312	0.0312	0.3422	0.0307	0.0311
PRS.IND.3PL	*0.2124*	0.1012	0.0059	*0.2124*	0	0.0859	0.0856	*0.2084*	*0.2102*	0.0936	0.1469	0.0838
PST.IPFV.IND.3SG	*0.2184*	0.4136	0.3755	*0.2300*	0.3812	0	0	*0.2120*	*0.2011*	0.4136	0.0609	0.0094
PST.PFV.IND.1SG	*0.2594*	0.4102	0.3720	*0.2525*	0.3773	0.0471	0	*0.2592*	*0.2464*	0.4102	0.1062	0.0563
PST.PFV.IND.3SG	0.0030	0.3316	0.3498	0.0136	0.3521	0	0	0	0.0030	0.3316	0.0016	0.0030
FUT.IND.3SG	0.0333	0.3441	0.3776	0.0650	0.3699	0.0533	0.0245	0.0345	0	0.3441	0.0444	0.0203
PRS.SBJV.3SG	*0.1894*	0.0000	0.0657	*0.1894*	0.0632	0.1350	0.1350	*0.1894*	*0.1894*	0	0.0917	0.1332
PRS.SBJV.2PL	*0.2049*	0.3912	0.4138	*0.2049*	0.4187	0.0483	0.0483	*0.2049*	*0.1836*	0.3912	0	0.0483
PST.PTCP	*0.2109*	0.4218	0.3431	*0.2133*	0.3806	0.0191	0.0191	*0.2209*	*0.1970*	0.4218	0.0657	0

Exceptions to vowel reduction

- While vowel reduction in unstressed position is the default behavior, it is not systematic.
- This causes more uncertainty now in the opposite direction: when predicting a cell with an unstressed prethematic vowel from a cell with a stressed prethematic vowel, uncertain whether reduction will take place.

	INF	1SG	2SG	3SG	1PL	2PL	3PL
achar	**ɐ**ʃˈaɾ	ˈ**a**ʃu	ˈ**a**ʃɐʃ	ˈ**a**ʃɐ	**ɐ**ʃˈɐmuʃ	**ɐ**ʃˈaiʃ	ˈ**a**ʃɐ̃ũ
relaxar	rəl**a**ʃˈaɾ	rəlˈ**a**ʃu	rəlˈ**a**ʃɐʃ	rəlˈ**a**ʃɐ	rəl**a**ʃˈɐmuʃ	rəl**a**ʃˈaiʃ	rəlˈ**a**ʃɐ̃ũ
vetar	vɛtˈaɾ	vˈɛtu	vˈɛtɐʃ	vˈɛtɐ	vɛtˈɐmuʃ	vɛtˈaiʃ	vˈɛtɐ̃ũ
encetar	ẽsətˈaɾ	ẽsˈɛtu	ẽsˈɛtɐʃ	ẽsˈɛtɐ	ẽsətˈɐmuʃ	ẽsətˈaiʃ	ẽsˈɛtɐ̃ũ

Exceptions to vowel reduction

- The effects of exceptions to vowel reduction are subtle.
- However in a few cases they are the main cause of uncertainty.

	INF	PRS.IND.1SG	PRS.IND.3SG	PRS.IND.2PL	PRS.IND.3PL	PST.IPFV.IND.3SG	PST.PFV.IND.1SG	PST.PFV.IND.3SG	FUT.IND.3SG	PRS.SBJV.3SG	PRS.SBJV.2PL	PST.PTCP
INF	0	0.3427	0.3032	0.0541	0.3706	0.0163	0.0163	0.0263	0	0.3427	0.0295	0.0121
PRS.IND.1SG	0.6990	0	0.6366	0.6990	0.6594	0.6832	0.6761	0.6990	0.6990	0.7855	0.6821	0.6678
PRS.IND.3SG	0.2044	0.0819	0	0.2044	0.0041	0.0856	0.0856	0.2042	0.2382	0.0848	0.1461	0.0837
PRS.IND.2PL	0.0316	0.3422	0.3574	0	0.3605	0.0316	0.0312	0.0312	0.0312	0.3422	0.0307	0.0311
PRS.IND.3PL	0.2124	0.1012	0.0059	0.2124	0	0.0859	0.0856	0.2084	0.2102	0.0936	0.1469	0.0838
PST.IPFV.IND.3SG	0.2184	0.4136	0.3755	0.2300	0.3812	0	0	0.2120	0.2011	0.4136	0.0609	0.0094
PST.PFV.IND.1SG	0.2594	0.4102	0.3720	0.2525	0.3773	0.0471	0	0.2592	0.2464	0.4102	0.1062	0.0563
PST.PFV.IND.3SG	0.0030	0.3316	0.3498	0.0136	0.3521	0	0	0	0.0030	0.3316	0.0016	0.0030
FUT.IND.3SG	0.0333	0.3441	0.3776	0.0650	0.3699	0.0533	0.0245	0.0345	0	0.3441	0.0444	0.0203
PRS.SBJV.3SG	0.1894	0.0000	0.0657	0.1894	0.0632	0.1350	0.1350	0.1894	0.1894	0	0.0917	0.1332
PRS.SBJV.2PL	0.2049	0.3912	0.4138	0.2049	0.4187	0.0483	0.0483	0.2049	0.1836	0.3912	0	0.0483
PST.PTCP	0.2109	0.4218	0.3431	0.2133	0.3806	0.0191	0.0191	0.2209	0.1970	0.4218	0.0657	0

Other causes of low predictiveness

Irregular endings

	1SG	2SG	3SG	1PL	2PL	3PL	#
CEDER	sˈedu	sˈɛdəʃ	sˈɛdə	sədˈemuʃ	sədˈeiʃ	sˈɛdẽĩ	99
QUERER	kˈɛru	kˈɛrəʃ	kˈɛɾ	kəɾˈemuʃ	kəɾˈeiʃ	kˈɛɾẽĩ	2

Stem allomorphy

	1SG	2SG	3SG	1PL	2PL	3PL	#
RANGER	rˈẽʒu	rˈẽʒəʃ	rˈẽʒə	rẽʒˈemuʃ	rẽʒˈeiʃ	rˈẽʒẽĩ	196
MANTER	mẽtˈɐɲu	mẽtˈẽĩʃ	mẽtˈẽĩ	mẽtˈemuʃ	mẽtˈẽdəʃ	mẽtˈẽĩẽĩ	30

Suppletion

	1SG	2SG	3SG	1PL	2PL	3PL
IND.PRS	sˈo	ˈɛʃ	ˈɛ	sˈomuʃ	sˈoiʃ	sˈẽũ
IND.PST.IPFV	ˈɛɾɐ	ˈɛɾɐʃ	ˈɛɾɐ	ˈɛɾɐmuʃ	ˈɛɾɐiʃ	ˈɛɾẽũ
IND.PST.PFV	fˈui	fˈoʃtə	fˈoi	fˈomuʃ	fˈoʃtəʃ	fˈoɾẽũ

- However the prevalence of these phenomena in Portuguese is low and hence makes only a small contribution to uncertainty.

Joint predictiveness

- Joint predictiveness: prediction from knowledge of two or more cells
  - ☞ Stump and Finkel (2013) on principal part systems
- The interplay between vowel alternations and theme vowel reductions entails that no single cell can be a good predictor of the whole paradigm.
  - ☞ No cell with stress on the prethematic vowel makes a three way distinction of theme vowels.
  - ☞ Many pairs combining
    - ★ a cell with a three-way distinction in endings
    - ★ a cell with pre-thematic stress

    are perfect overall predictors of the paradigm (i.e., constitute a set of principal parts).
- Some other pairs of cells have surprisingly good joint predictiveness.
- A case in point: PRS.1SG and PRS.3SG

Joint predictiveness

- 2nd and 3rd conjugation verbs have raised prethematic mid-vowels in the PRS.1SG

	INF	1SG	2SG	3SG	1PL	2PL	3PL
LEVAR	ləvˈaɾ	lˈɛvu	lˈɛvɐʃ	lˈɛvɐ	ləvˈɐmuʃ	ləvˈaiʃ	lˈɛvɐ̃ũ
NOTAR	nutˈaɾ	nˈɔtu	nˈɔtɐʃ	nˈɔtɐ	nutˈɐmuʃ	nutˈaiʃ	nˈɔtɐ̃ũ
RECEBER	rəsəbˈeɾ	rəsˈebu	rəsˈɛbəʃ	rəsˈɛbə	rəsəbˈemuʃ	rəsəbˈɐiʃ	rəsˈɛbɐ̃ĩ
RECORRER	rəkurˈeɾ	rəkˈoru	rəkˈɔrəʃ	rəkˈɔrə	rəkurˈemuʃ	rəkurˈɐiʃ	rəkˈɔrɐ̃ĩ
SEGUIR	səgˈiɾ	sˈigu	sˈɛgəʃ	sˈɛgə	səgˈimuʃ	səgˈiʃ	sˈɛgɐ̃ĩ
SUBIR	subˈiɾ	sˈubu	sˈɔbəʃ	sˈɔbə	subˈimuʃ	subˈiʃ	sˈɔbɐ̃ĩ

- As a result, the PRS.1SG sometimes disambiguates between 2nd and 3rd conjugation

⟨PRS.1SG,PRS.3SG⟩ → INF	
⟨*X*i*C*u , *X*ɛ*C*ə⟩	→ *X*ə*C*iɾ
⟨*X*e*C*u , *X*ɛ*C*ə⟩	→ *X*ə*C*eɾ

⟨PRS.1SG,PRS.3SG⟩ → INF	
⟨*X*u*C*u , *X*ɔ*C*ə⟩	→ *X*u*C*iɾ
⟨*X*o*C*u , *X*ɔ*C*ə⟩	→ *X*u*C*eɾ

Conclusions

- We have illustrated the use of IPa methods for the practical description of inflection systems
  - Inflectional classification (clustering of lexemes)
  - Paradigm partition (clustering of cells)
  - Cell predictiveness
    - ★ Elaboration of (Ackerman et al., 2009) etc.
- Striking result: in Portuguese conjugation, most complexity results from historically motivated but now morphologized patterns of vowel alternation.
- Thus patterns orthogonal to exponence are crucial to an understanding of the system.
- While there are ways of accounting for such patterns (e.g. morphomic stem alternants, morphophonological rules), one virtue of the current method is to help us find the patterns and evaluate their relevance.

References I

Ackerman, F., Blevins, J. P., and Malouf, R. (2009). `Parts and wholes: implicative patterns in inflectional paradigms'. In J. P. Blevins and J. Blevins (eds.), *Analogy in Grammar*. Oxford: Oxford University Press, 54--82.

Ackerman, F. and Malouf, R. (in press). `Morphological organization: the low conditional entropy conjecture'. *Language*, 89.

Albright, A. C. (2002). *The Identification of Bases in Morphological Paradigms*. Ph.D. thesis, University of California, Los Angeles.

Blevins, J. P. (to appear). *Word and Paradigm Morphology*. Oxford: Oxford University Press.

Bonami, O. and Boyé, G. (2002). `Suppletion and stem dependency in inflectional morphology'. In F. Van Eynde, L. Hellan, and D. Beerman (eds.), *The Proceedings of the HPSG '01 Conference*. Stanford: CSLI Publications.

--------- (to appear). `De formes en thèmes'. In F. Villoing, S. Leroy, and S. David (eds.), *Du régulier et du minutieux*. Presses Universitaires de Paris Ouest.

Bonami, O., Boyé, G., and Henri, F. (2011). `Measuring inflectional complexity: French and mauritian'. In *Quantitative Measures in Morphology and Morphological Development*. San Diego: University of California.

Bonami, O., Henri, F., and Luís, A. R. (2012). `Tracing the origins of inflection in Creoles: a quantitative analysis'. Paper presented at the 9th Creolistics Workshop, Aarhus, Denmark.

References II

Boyé, G. (2000). *Problèmes de morpho-phonologie verbale en français, espagnol et italien*. Ph.D. thesis, Université Paris 7.

Brown, D. and Evans, R. (2012). `Morphological complexity and unsupervised learning: validating Russian inflectional classes using high frequency data'. In F. Kiefer, M. Ladányi, and P. Siptár (eds.), *Current Issues in Morphological Theory: (Ir)regularity, analogy and frequency*. Amsterdam: John Benjamins, 135--162.

Pirelli, V. and Battista, M. (2000). `The paradigmatic dimension of stem allomorphy in italian verb inflection'. *Rivista di Linguistica*, 12.

Saitou, N. and Nei, M. (1987). `The neighbor-joining method: a new method for reconstructing phylogenetic trees'. *Molecular biology and evolution*, 4:406--425.

Santos, D. and Rocha, P. (2001). `Evaluating cetempúblico, a free resource for Portuguese'. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*. 442--449.

Sims, A. (2010). `Probabilistic paradigmatics: Principal parts, predictability and (other) possible particular pieces of the puzzle'. Paper presentend at the Fourteenth International Morphology Meeting, Budapest.

Spencer, A. (2012). `Identifying stems'. *Word Structure*, 5:88--108.

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

References III

Veiga, A. O. d., Candeias, S., and Perdigão, F. (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, W. U. (1984). *Flexionsmorphologie und Natürlichkeit. Ein Beitrag zur morphologischen Theoriebildung*. Berlin: Akademie-Verlag. Translated as (Wurzel, 1989).

--------- (1989). *Inflectional Morphology and Naturalness*. Dordrecht: Kluwer.