Measuring inflectional complexity: French and Mauritian

Olivier Bonami[1] Gilles Boyé[2] Fabiola Henri[3]

[1]U. Paris-Sorbonne & Institut Universitaire de France

[2]U. de Bordeaux

[3]U. Sorbonne Nouvelle

QMMMD
San Diego, January 15, 2011

The inflectional complexity of Creoles

- Long history of claims on the morphology of Creole languages:
  - Creoles have no morphology (e.g. Seuren and Wekker, 1986)
  - Creoles have simple morphology (e.g. McWhorter, 2001)
  - Creoles have simpler inflection than their lexifier (e.g. Plag, 2006)
- Belongs to a larger family of claims on the simplicity of Creole languages (e.g. Bickerton, 1988)
- ☞ As (Robinson, 2008) notes, such claims on Creoles need to be substantiated by quantitative analysis.
- Here we adress the issue by comparing the complexity of Mauritian Creole conjugation with that of French conjugation.
- There are many dimensions of complexity. Here we focus on just one aspect.

The PCFP and a strategy for adressing it

- Ackerman et al. (2009); Malouf and Ackerman (2010) argue that an important aspect of inflectional complexity is the Paradigm Cell Filling Problem:
  - Given exposure to an inflected wordform of a novel lexeme, what licenses reliable inferences about the other wordforms in its inflectional family?

(Malouf and Ackerman, 2010, 6)

- Their strategy:
  - Knowledge of implicative patterns relating cells in a paradigm is relevant
  - This knowledge is best characterized in information-theoretic terms
  - ☞ The reliability of implicative patterns relating paradigm cell A to paradigm cell B is measured by the conditional entropy of cell B knowing cell A.

The goal of this paper

- We apply systematically Ackerman et al.'s strategy to the full assessment of two inflectional systems
- This involves looking at realistic datasets
  - Lexicon of 6440 French verb lexemes with 48 paradigm cells, adapted from the BDLEX database (de Calmès and Pérennou, 1998)
  - Lexicon of 2079 Mauritian verb lexemes, compiled from (Carpooran, 2009)'s dictionary
- Surprising conclusion: doing this is hard linguistic work (although it is computationally rather trivial).
- Our observations do not affect (Ackerman et al., 2009)'s general point on the fruitfulness of information theory as a tool for morphological theorizing.
- Rather, they show that interesting new questions arise when looking at large datasets

Outline

A toy example

- We illustrate the reasoning used by (Ackerman et al., 2009; Sims, 2010; Malouf and Ackerman, 2010)
- Looking at French infinitives and past imperfectives:
  - Assume there are just 5 conjugation classes in French
  - Assume all classes are equiprobable

IC	INF	IPFV.3SG	lexeme	trans.
1	sɔʁtiʁ	sɔʁtɛ	sortir	'go out'
2	amɔʁtiʁ	amɔʁtisɛ	amortir	'cushion'
3	lave	lavɛ	laver	'wash'
4	vulwaʁ	vulɛ	vouloir	'want'
5	batʁ	batɛ	battre	'fight'

- $H(\text{IPFV}|\text{INF} = \text{stem} \oplus \text{ʁ}) = 1\text{bit}$
- $H(\text{IPFV}|\text{INF} \neq \text{stem} \oplus \text{iʁ}) = 0\text{bit}$
- $H(\text{IPFV}|\text{INF}) = \frac{2}{5} \times 1 + \frac{3}{5} \times 0 = 0.4\text{bit}$

Discussion

- The claim: this way of evaluating $H(\text{IPFV}|\text{INF})$ provides a rough measure of the difficulty of the PCFP for INF↦IPFV in French.
  - Other factors (phonotactic knowledge on the makeup of the lexicon, knowledge of morphosemantic correlations, etc.) reduce the entropy; but arguably the current reasoning focuses on the specifically morphological aspect.
  - Because of the equiprobability assumption, what is computed is really an upper bound.
  - The reasoning relies on a preexisting classification of the patterns of alternations between forms. In a way, what we are measuring is the quality of that classification.

☞ When scaling up to a large data set, a number of methodological issues arise. We discuss 4.

Outline

Back to Ackerman, Blevins & Malouf

- (Ackerman et al., 2009; Malouf and Ackerman, 2010) construct a number of arguments on paradigm entropy on the basis of datasets with no type frequency information.
- Reasoning: by assuming that all inflection classes are equiprobable, one provides an upper bound on the actual paradigm entropy.
- This makes sense as long as the goal is simply to show that entropy is lower than in could be without any constraints on paradigm economy.
- However the resulting numbers can be very misleading.

A toy example

- Assume an inflection system with
  - 2 paradigm cells
  - 2 exponents for cell A
  - 4 exponents for cell B
  - A strong preference of one exponent in cell B

IC	A	B	type freq.
1	-i	-a	497
2	-i	-e	1
3	-i	-u	1
4	-i	-y	1
5	-o	-a	497
6	-o	-e	1
7	-o	-u	1
8	-o	-y	1

- Results:

	A	B
A	—	2
B	1	—

$H(\text{col}|\text{row})$, without frequency

	A	B
A	—	0.0624
B	1	—

$H(\text{col}|\text{row})$, with frequency

Discussion

- In the absence of type frequency information, one may conclude on:
  - The existence of an upper bound on conditional entropy
  - The existence of categorical implicative relations
- However no meaningful comparisons can be made between the computed entropy values
  - ☞ Upper bound can be very close to or very far from the actual value
- In this context, it is relevant to notice that entropy is commonly close to 0 without being null.
  - ☞ Among the 2256 pairs of cells in French verbal paradigms, 18% have an entropy below 0.1bit, while only 12% have null entropy.
- Thus type frequency information is necessary as soon as we want to be able to make comparative claims, even within a single language.

Outline

The problem

- Extant inflectional classifications are generally not directly usable.
- Example: for French, it is traditional to distinguish
  - 4 infinitival suffixes -e, -iʁ, -waʁ, -ʁ
  - Two types of imperfectives: with or without the augment -s-

IC	INF	IPFV.3SG	orth.	trans.
1	sɔʁtiʁ	sɔʁtɛ	sortir	go out
2	amɔʁtiʁ	amɔʁtisɛ	amortir	cushion
3	lave	lavɛ	laver	wash
4	vulwaʁ	vulɛ	vouloir	want
5	batʁ	batɛ	battre	fight

- Observation: the choice of the infinitive suffix fully determines the form of the imperfective, except when the suffix is -ʁ.
- For instance, $H(\text{IPFV} \mid \text{INF} = \text{stem} \oplus \text{iʁ}) = 0$

The problem

- The fact that $H(\text{IPFV} \mid \text{INF} = \text{stem} \oplus \text{iʁ}) = 0$ is of no use for solving the PCFP: when an infinitive ends in iʁ, there are really two possible outcomes.

IC	INF	IPFV.3SG	lexeme	trans.
1	sɔʁt-iʁ	sɔʁtɛ	sortir	go out
2	amɔʁti-ʁ	amɔʁtisɛ	amortir	cushion

☞ Speakers don't see morph boundaries

- So if we want to reason about implicative relations, we should be thinking of the entropy of the IPFV given some knowledge of what the final segments of the infinitive are, not of what the suffix is.

This is a general issue

- Traditional classifications usually rely on the identification of exponents

☞ Yet exponents presuppose bases (which the exponents modify).

  - Not compatible with a fully word-based, 'abstractive' (Blevins, 2006) view of inflection.
  - Even under a constructive view, there is uncertainty in the identification of bases.

- In practical terms, we can not rely on this type of classification when studying implicational relations.

☞ We should really be looking at patterns of alternation between two forms of each individual lexeme, not patterns of alternation between paradigmatic classes of forms.

Outline

Phonology masking morphological distinctions

- Perfectly predictable and regular phonological alternations can give rise to inflectional opacity
- Example in French: suffix -j in the IPFV.PL
  - j ⟶ ij / BranchingOnset __

IPFV.1SG	IPFV.1PL	lexeme	trans.
lavɛ	lavjɔ̃	LAVER	'wash'
pɔʁtɛ	pɔʁtjɔ̃	PORTER	'carry'
kɔ̃tʁɛ	kɔ̃tʁijɔ̃	CONTRER	'counter'
pwavʁɛ	pwavʁijɔ̃	POIVRER	'pepper'

  - j ⟶ ∅ / j __

IPFV.1SG	IPFV.1PL	lexeme	trans.
kajɛ	kajɔ̃	CAILLER	'curdle'
pijɛ	pijɔ̃	PILLER	'plunder'
kadʁijɛ	kadʁijɔ̃	QUADRILLER	'cover'
vʁijɛ	vʁijɔ̃	VRILLER	'pierce'

The problem

- This results in uncertainty when predicting the IPFV.SG from IPFV.1PL

IPFV.1PL	IPFV.1SG	lexeme	trans.
kɔ̃tʁijɔ̃	kɔ̃tʁɛ	CONTRER	‘counter’
pwavʁijɔ̃	pwavʁɛ	POIVRER	‘pepper’
kadʁijɔ̃	kadʁijɛ	QUADRILLER	‘cover’
vʁijɔ̃	vʁijɛ	VRILLER	‘pierce’

- Not a small phenomenon: 294 IPFV.1PL in -ijɔ̃ in our dataset
- Problem: this is often abstracted away from transcriptions

lexeme	IPFV.1PL	surface transcription	BDLEX transcription
POIVRER	*poivrions*	pwavʁijɔ̃	pwavʁjɔ̃
VRILLER	*vrillions*	vʁijɔ̃	vʁijɔ̃

What we learned

- As morphologists we are used to working on relatively abstract phonological transcriptions
- Thus simple phonological alternations are often abstracted away from our datasets
- This can result in artificially lowering the uncertainty in predicting one form from another: by undoing phonology, we in effect precode inflection class information.

☞ Phonological issues can not be ignored; the dataset should be as surface-true as possible

- In our case, tedious hand-editing of the BDLEX dataset

Outline

Use standardized classifications

- The preceding discussion shows that extant inflectional classifications cannot be trusted for this type of work.

☞ New, linguistically well-thought out classifications of patterns of alternation need to be designed.

- Yet, writing these by hand is not an option
  - In the case of French there are 2550 ordered pairs of cells, each of which is in need of its own clasification.
  - Although many of these are trivial, there are at least 132 hard cases

  ☞ 12 zones of interpredictibility ('alliances of forms') identified by (Bonami and Boyé, 2002)

☞ We need implemented algorithms for infering classifications

- Should be simple enough that descriptive linguists have an intuition as to their adequacy

☞ If we want to make meaningful comparison between languages, we need the descriptive linguist to check that the algorithm does not bias the comparison

Outline

The intuition

- Assume we have a reasonable, agreed-upon way of describing the patterns of alternation for going from cell A to cell B.
  1. We start by identifying, for each lexeme, which pattern maps its A form to its B form.
  2. We then identify, for each A form, the set of patterns that could have been used to generate a B form.
- Step 1 gives us a random variable over patterns of alternation between A and B. We note this A→B
- Step 2 gives us a random variable over A, which classifies A forms according to those phonological properties that are relevant to the determination of the B form.
- We submit that $H(\text{A}{\rightarrow}\text{B} \mid \text{A})$ is a reasonable estimate of the difficulty of predicting cell B from cell A.
- We call this the Implicational entropy from A to B.

An simple example

- Suppose we decide to classify our French data by assuming a maximally long, word-initial stem.

IC	INF	IPFV.3SG	pattern	classification of INF
1	sɔʁtiʁ	sɔʁtɛ	Xiʁ → Xɛ	$A = \{X$iʁ → Xɛ, Xʁ → Xsɛ, Xʁ → Xɛ$\}$
2	amɔʁtiʁ	amɔʁtisɛ	Xʁ → Xsɛ	$A = \{X$iʁ → Xɛ, Xʁ → Xsɛ, Xʁ → Xɛ$\}$
3	lave	lavɛ	Xe → Xɛ	$B = \{X$e → Xɛ$\}$
4	vulwaʁ	vulɛ	Xwaʁ → Xɛ	$C = \{X$waʁ → Xɛ, Xʁ → Xsɛ, Xʁ → Xɛ$\}$
5	batʁ	batɛ	Xʁ → Xɛ	$D = \{X$ʁ → Xsɛ, Xʁ → Xɛ$\}$

- If all classes were equiprobable:
  - $H(\text{INF} \rightarrow \text{IPFV.3SG} \mid \text{INF} \in A) = 1\text{bit}$
  - $H(\text{INF} \rightarrow \text{IPFV.3SG} \mid \text{INF} \notin A) = 0\text{bit}$
  - $H(\text{INF} \rightarrow \text{IPFV.3SG} \mid \text{INF}) = 0.4\text{bit}$

☞ Notice how classes of INF record information on the form of INF that might be relevant to the determination of the pattern.

A crucial caveat

- The algorithm used to classify patterns of alternation matters a lot.
  - Example A: stem maximization, purely suffixal
    For each pair $\langle x, y \rangle$, identify the longest σ such that $x = \sigma \oplus s_1$ and $y = \sigma \oplus s_2$. The pattern exemplified by $\langle x, y \rangle$ is replacement of s_1 by s_2.
  - Example B: 1 lexeme, 1 class
    For each pair $\langle x, y \rangle$, the pattern it exemplifies is replacement of x by y.

☞ Algorithm B will give rise to much smaller implicational entropy values (0 bit in most cases) than algorithm A. This does not make it a good choice.

- There are plenty of good possibilities to consider
- No universal solution is forthcoming. Thus we should focus on a solution that is adequate to the comparison at hand.

☞ For French and Mauritian, algorithm A will do for now

Outline

Introduction

- Our goal: assess empirically the claim that creole languages have a simpler inflectional system than their lexifier (e.g. Plag, 2006)
- To this end, we compare the complexity of Mauritian Creole conjugation with that of French conjugation
- There are many dimensions to inflectional complexity:
  1. Size and structure of the paradigm
  2. Number of exponents per word (number of rule blocks)
  3. Morphosyntactic opacity of the paradigm (presence of morphomic phenomena)
  4. Number of inflectional classes
  5. ...
  6. Difficulty of the PCFP
- Mauritian is undisputably simpler than French in dimensions 1 and 2. Henri (2010) argues that they are on a par with respect to dimension 3. Here we focus on dimension 6.

Outline

French paradigms

- 51 cells, analyzed in terms of 6 features
  - 3 suffixal rule blocks (Bonami and Boyé, 2007a)

Finite forms

TAM	1SG	2SG	3SG	1PL	2PL	3PL
PRS.IND	lav	lav	lav	lav-ɔ̃	lav-e	lav
PST.IND.IPFV	lav-ɛ	lav-ɛ	lav-ɛ	lav-j-ɔ̃	lav-j-e	lav-ɛ
PST.PFV	lavɛ	lava	lava	lava-m	lava-t	lavɛ-ʁ
FUT.IND	lavə-ʁ-ɛ	lavə-ʁ-a	lavə-ʁ-a	lavə-ʁ-ɔ̃	lavə-ʁ-e	lavə-ʁ-ɔ̃
PRS.SBJV	lav	lav	lav	lav-j-ɔ̃	lav-j-e	lav
PST.SBJV	lava-s	lava-s	lava	lava-s-j-ɔ̃	lava-s-j-e	lava-s
COND	lavə-ʁ-ɛ	lavə-ʁ-ɛ	lavə-ʁ-ɛ	lavə-ʁ-j-ɔ̃	lavə-ʁ-j-e	lavə-ʁ-ɛ
IMP	---	lav	---	lav-ɔ̃	lav-e	---

Nonfinite forms

INF	PRS.PTCP	PST.PTCP			
		M.SG	F.SG	M.PL	F.PL
lave	lav-ɑ̃	lave	lave	lave	lave

Morphomic stem alternations

- Cf. (Bonami and Boyé, 2002, 2003, 2007b):
  - no inflection classes of exponents
  - Intricate system of stem allomorphy relying on morphomic patterns

Finite forms

TAM	1SG	2SG	3SG	1PL	2PL	3PL
PRS.IND	$stem_3$	$stem_3$	$stem_3$	$stem_1$-ɔ̃	$stem_1$-e	$stem_2$
PST.IND.IPFV	$stem_1$-ɛ	$stem_1$-ɛ	$stem_1$-ɛ	$stem_1$-jɔ̃	$stem_1$-je	$stem_1$-ɛ
PST.PFV	$stem_{11}$	$stem_{11}$	$stem_{11}$	$stem_{11}$-m	$stem_{11}$-t	$stem_{11}$-r
FUT.IND	$stem_{10}$-ʁɛ	$stem_{10}$-ʁa	$stem_{10}$-ʁa	$stem_{10}$-ʁɔ̃	$stem_{10}$-ʁe	$stem_{10}$-ʁɔ̃
PRS.SBJV	$stem_7$	$stem_7$	$stem_7$	$stem_8$-jɔ̃	$stem_8$-je	$stem_7$
PST.SBJV	$stem_{11}$-s	$stem_{11}$-s	$stem_{11}$	$stem_{11}$-sjɔ̃	$stem_{11}$-sje	$stem_{11}$-s
COND	$stem_{10}$-ʁɛ	$stem_{10}$-ʁɛ	$stem_{10}$-ʁɛ	$stem_{10}$-ʁjɔ̃	$stem_{10}$-ʁje	$stem_{10}$-ʁɛ
IMP	---	$stem_5$	---	$stem_6$-ɔ̃	$stem_6$-e	---

Nonfinite forms

INF	PRS.PTCP	PST.PTCP			
		M.SG	F.SG	M.PL	F.PL
$stem_9$	$stem_4$-ɑ̃	$stem_{12}$	$stem_{12}$	$stem_{12}$	$stem_{12}$

Outline

Sources of the Mauritian lexicon

- Most of the language's vocabulary has been inherited from French with a few phonological adaptations.

French⟶Mauritian	example		trans.
ʃ⟶s	detaʃe	⟶detase	'detach'
ʒ⟶z	mɑ̃ʒe	⟶mɑ̃ze	'eat'
ʁ⟶ɚ/__[$_\sigma$	paʁti	⟶paɚti	'leave'
y⟶i	fyme	⟶fime	'smoke'
ə⟶e/#C__	ʁədone	⟶ʁedone	'give again'
ɛ⟶e	fɛʁ	⟶feɚ	'do'
ɔ⟶o	sɔʁti	⟶soɚti	'go out'

- A minority of lexemes borrowed from English, Hindi/Bhojpuri, Malagasy, (etc.)

Verb form alternations

- Most Mauritian verbs have two forms: the long form (LF) and the short form (SF).

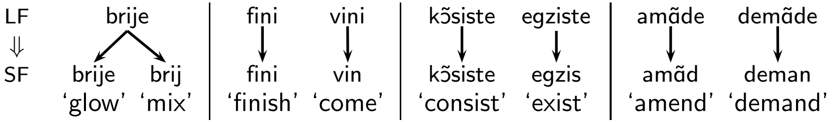
LF	brize	vɑ̃de	amɑ̃de	ɛgziste	vini	brije	kɔ̃siste	fini
SF	briz	van	amɑ̃d	ɛgzis	vin	brije	kɔ̃siste	fini
trans.	'break'	'sell'	'amend'	'exist'	'come'	'mix'	'consist'	'finish'

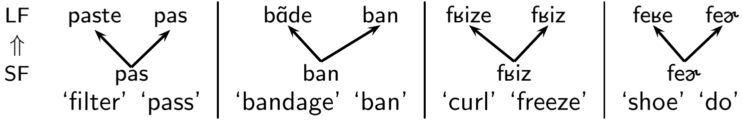
- ☞ The LF almost always derives from the Fr. infinitive(Veenstra, 2004)
- ☞ The SF often resembles a Fr. present singular

- The alternation probably started out as a sandhi rule (Corne, 1982): drop verb final e in appropriate contexts
  - Almost all alternating verbs are verbs ending in e
  - No verb drops e after a branching onset
    - ☞ Mauritian, (unlike French; Dell, 1995), disallows word-final branching onsets

Why Morphology?

- However today the alternation is not phonologically predictable





Distribution of long and short forms

- The division of labor between LF and SF is morphomic (Henri, 2010)

	Distribution	SF	LF
	Syntax		
No Verum Focus	V with nonclausal complements (NPs,APs,ADVPs,VPs,PPs)	yes	no
	V with no complements	no	yes
	V with clausal complements	no	yes
	only extracted complements	no	yes
	Verum Focus	no	yes
	Morphology		
	reduplicant	yes	no
	base	yes	yes

Table: Constraints on verb form alternation

Outline

Application to Mauritian

- We collected the 2079 distinct Mauritian verbs listed in Carpooran (2009), and coded their LF and SF.
- Using token frequency information from the lexique database (New et al., 2001) we extracted from BDLEX the paradigms of the 2079 most frequent nondefective verbs of French.
- We implemented a a stem maximization algorithm for finding patterns of alternation, and used it to compute the implicational entropy for all pairs of cells in both languages.
- Overall paradigm entropy:

Mauritian	0.744 bit
French	0.446 bit

☞ Notice that this is precisely contrary to expectations!

Variations

- This result seems quite robust:
  - If we now just compare the LF $\sim$ SF relation just to the INF $\sim$ PRS.3SG relation (to compare what is most directly comparable):

(Mauritian) LF $\mapsto$ SF	(French) INF $\mapsto$ PRS	(Mauritian) SF $\mapsto$ LF	(French) PRS $\mapsto$ INF
0.563	0.232	0.925	0.578

  - One might argue that type frequency information is information about the structure of the lexicon, not morphology.
  - If we leave out this information (take all classes to be equiprobable):

Mauritian	1.316
French	0.684

Why this result?

- In Mauritian, we find 11 patterns giving rise to 10 classes.

class	patterns	example		# of lex.	entropy
1	$\{X\text{e} \rightarrow X, X \rightarrow X\}$	kwafe	kwaf	1138	0.565
2	$\{X\text{te} \rightarrow X, X\text{e} \rightarrow X, X \rightarrow X\}$	gʁijote	gʁijot	268	0.845
3	$\{X \rightarrow X\}$	sufeɚ	sufeɚ	225	0.0
4	$\{X\text{ʁe} \rightarrow X\text{ɚ}, X\text{ʁe} \rightarrow X, X\text{e} \rightarrow X, X \rightarrow X\}$	kofʁe	kofʁe	159	0.835
5	$\{X\text{le} \rightarrow X, X\text{e} \rightarrow X, X \rightarrow X\}$	dekole	dekol	138	0.927
6	$\{X\text{i} \rightarrow X, X \rightarrow X\}$	fini	fini	116	0.173
7	$\{X\text{ãde} \rightarrow X\text{an}, X\text{e} \rightarrow X, X \rightarrow X\}$	ʁãde	ʁan	15	0.567
8	$\{X\text{ble} \rightarrow X\text{m}, X\text{le} \rightarrow X, X\text{e} \rightarrow X, X \rightarrow X\}$	ʁeduble	ʁeduble	13	0.391
9	$\{X\text{ɔ̃be} \rightarrow X\text{ɔm}, X\text{e} \rightarrow X, X \rightarrow X\}$	plɔ̃be	plɔ̃b	3	0.918
10	$\{X\text{õde} \rightarrow X\text{on}, X\text{e} \rightarrow X, X \rightarrow X\}$	fekõde	fekõd	4	0.811

Classification of Mauritian LFs on the basis of their possible relatedness with the SF

- Three well populated classes with a high entropy (# 2, 4, 5)
- ☞ For verbs whose LF ends in -te, -ʁe or -le, the SF is quite unpredictable
- Even for the remaining verbs in -e the predictibility is far from being total

Why this result?

- Compare the French situation:

class	patterns	example		# of lex.	entropy
1	{Xe → X}	asyme	asym	1279	0.0
2	{Xje → Xi, Xje → X, Xe → X}	pije	pij	171	1.515
3	{Xle → vX, Xe → X}	ale	va	153	0.057
4	{Xiʁ → X, Xʁ → X}	finiʁ	fini	142	0.313
5	{Xdʁ → X, Xʁ → X}	kudʁ	ku	55	0.0
6	{Xtiʁ → X, Xiʁ → X, Xʁ → X}	paʁtiʁ	paʁ	33	0.994
7	{Xtʁ → X, Xʁ → X}	konɛtʁ	konɛ	32	0.0
8	{Xɥe → Xy, Xe → X}	tɥe	ty	31	0.0
9	{Xəniʁ → X, Xjɛ̃ → X, Xʁ → X}	vəniʁ	vj ɛ	22	0.0
10	{Xʁ → X}	fɛʁ	fE	21	0.0
...	...	...	...	...	...
	(22 other classes with less than 20 members)				

Classification of French INFs on the basis of their possible relatedness with the PRS.3SG

- The infinitive is an excellent predictor of the present, except for verbs ending in -je or in -tir
- For the vast majority of verbs (73% of the 2079 most frequent) there is no uncertainty at all

Outline

A puzzling contrast

- These results were initially puzzling to us because of a previous study (Bonami and Henri, 2010).
- We trained Albright's (2002) *Minimal Generalization Learner* on French and Mauritian, to see how good it was at inferring the form of particular verbs
- The MGL is known for capturing efficiently morphophonological generalizations on the lexicon, in a way that correlates with experimental studies (Albright, 2003; Albright & Hayes, 2003).
- Results:

(Mauritian) LF $\mapsto$ SF	(French) INF $\mapsto$ PRS	(Mauritian) SF $\mapsto$ LF	(French) PRS $\mapsto$ INF
96.82%	94.77%	93.18%	91.86%

- According to the results of the MGL, the PCFP looks slightly more simple in Mauritian than in French.

☞ Why this difference?

A possible explanation for the contrast

- A possible explanation for the difference: the MGL uses a better method for classifying patterns of alternation
- Mauritian example:

LF	SF	pattern	set of possible patterns
egziste	egzis	$X\text{te} \to X$	$\{X\text{te} \to X, X\text{e} \to X, X \to X\}$
reste	res	$X\text{te} \to X$	$\{X\text{te} \to X, X\text{e} \to X, X \to X\}$
aʁete	aʁet	$X\text{e} \to X$	$\{X\text{te} \to X, X\text{e} \to X, X \to X\}$
eʁite	eʁit	$X\text{e} \to X$	$\{X\text{te} \to X, X\text{e} \to X, X \to X\}$
kõsiste	kõsiste	$X \to X$	$\{X\text{te} \to X, X\text{e} \to X, X \to X\}$
afekte	afekte	$X \to X$	$\{X\text{te} \to X, X\text{e} \to X, X \to X\}$

  - All verbs exhibiting the pattern pattern $X\text{te} \to X$ end in ste
  - The stem maximization algorithm does not see this; as a result egziste and aʁete end up in the same class, with non-null entropy
  - By contrast, the MGL derives a rule $[\text{te} \longrightarrow \emptyset / \text{s}\_\_\#]$ that does not apply to aʁete

☞ Does this explain the contrast between the two results?

The explanation fails

- To test this, for each relevant pair of cells A and B:
  1. We extracted from the output of the MGL the set of most general rules it generated
  2. We used this set of rules to classify the input cell
  3. We then computed $H(\text{A}\to\text{B} \mid \text{A})$ as usual.
- Results:

classification method	(Mauritian) LF ↦ SF	(French) INF ↦ PRS	(Mauritian) SF ↦ LF	(French) PRS ↦ INF
stem max.	0.563	0.232	0.925	0.578
MGL rules	0.476	0.079	0.846	0.296

- Conclusions
  - The pattern classification algorithm embedded in the MGL is more fine-grained, giving rise to lowered entropy in all cases.
  - However this does not explain the difference between the MGL results and the entropy results. If anything, the contrast between French and Mauritian is stronger when using context-dependent phonological rules.

An alternative explanation

- The MGL embodies an assumption that the degree of similarity of a novel input form to known input forms plays a role in determining the best candidate output.
- Example:
  - In Mauritian the pattern $X \to X$ is used 31% of the time
  - However for verbs in Cje it is always used
  - The MGL uses this information to decide that
    - copje→copje is more likely than copje→copj
    - stope→stop is more likely than stope→stope
- No such assumption in the current approach.

☞ There is still some of information in the input forms that

  - has an effect on the PCFP
  - is not taken into account by the current approach

- Naive question: should we care? Is this morphological information?

Outline

Conclusions

1. On Creole complexity:
   - Although there is less morphology in Mauritian than in French, it does not follow that the system is simpler.
   - ☞ the PCFP seems to be more complex in Mauritian.
   - To the extent that claims on Creole complexity are taken seriously, they should be assessed quantitatively.
2. On evaluating the PCFP:
   - We confirm on a large-scale study the fruitfulness of an information-theoretic measure of the difficulty of the PCFP.
   - The methods used for classifying patterns of alternation have crucial consequences.
   - ☞ Assessing the quality and the adequacy of these methods should be taken much more seriously.

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.

Bickerton, D. (1988). 'Creole languages and the bioprogram'. In F. Newmeyer (ed.), *Linguistic Theory: Extensions and Implications*, vol. 2 of *The Cambridge Survey*. Cambridge University Press, 268–284.

Blevins, J. P. (2006). 'Word-based morphology'. *Journal of Linguistics*, 42:531–573.

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.

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

——— (2007a). 'French pronominal clitics and the design of Paradigm Function Morphology'. In *Proceedings of the fifth Mediterranean Morphology Meeting*. 291–322.

——— (2007b). 'Remarques sur les bases de la conjugaison'. In E. Delais-Roussarie and L. Labrune (eds.), *Des sons et des sens*. Paris: Hermès, 77–90. Ms, Université Paris 4 & Université Bordeaux 3.

Bonami, O. and Henri, F. (2010). 'How complex is creole inflectional morphology? the case of mauritian'. Poster presented at the 14th International Morphology Meeting.

Carpooran, A. (2009). *Diksioner Morisien*. Sainte Croix (Mauritius): Koleksion Text Kreol.

Corne, C. (1982). 'The predicate in Isle de France Creole.' In P. Baker and C. Corne (eds.), *Isle de France Creole. Affinities and Origins*. Ann Arbor: Karoma, 31–48.

de Calmès, M. and Pérennou, G. (1998). 'BDLEX : a lexicon for spoken and written french'. In *Proceedings of the First International Conference on Language Resources and Evaluation*. Granada: ERLA, 1129–1136.

Dell, F. (1995). 'Consonant clusters and phonological syllables in french'. *Lingua*, 95:5–26.

Henri, F. (2010). *A Constraint-Based Approach to verbal constructions in Mauritian*. Ph.D. thesis, University of Mauritius and Université Paris Diderot.

Malouf, R. and Ackerman, F. (2010). 'Paradigms: The low entropy conjecture'. Paper presented at the Workshop on Morphology and Formal Grammar, Paris.

McWhorter, J. (2001). 'The world's simplest grammars are creole grammars'. *Linguistic Typology*, 5:125–166.

New, B., Pallier, C., Ferrand, L., and Matos, R. (2001). 'Une base de données lexicales du français contemporain sur internet: Lexique'. *L'Année Psychologique*, 101:447–462.

Plag, I. (2006). 'Morphology in Pidgins and Creoles'. In K. Brown (ed.), *Encyclopedia of Language and Linguistics, 2nd Edition*, vol. 8. 304–308.

Robinson, S. (2008). 'Why pidgin and creole linguistics need the statistician'. *Journal of Pidgin and Creole Languages*, 23:141–146.

Seuren, P. and Wekker, H. (1986). 'Semantic transparency as a factor in creole genesis'. In P. Muysken and N. J. Smith (eds.), *Substrata versus Universals in Creole Genesis*. Amsterdam: Benjamins, 57–70.

Sims, A. D. (2010). 'Probabilistic paradigmatics'. Paper presented at the 14th International Morphology Meeting, Budapest.

Veenstra, T. (2004). 'What verbal morphology can tell us about Creole genesis: the case of French-related Creoles'. In I. Plag (ed.), *Phonology and Morphology of Creole Languages*, no. 478 in Linguistische Arbeiten. Max Niemeyer Verlag Gmbh.