Measuring inflectional complexity: French and Mauritian

Olivier Bonami\textsuperscript{1} Gilles Boyé\textsuperscript{2} Fabiola Henri\textsuperscript{3}

\textsuperscript{1}U. Paris-Sorbonne & Institut Universitaire de France
\textsuperscript{2}U. de Bordeaux
\textsuperscript{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 address 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 addressing 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.
Introduction

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
Methodological issues

Ackerman et al.’s strategy

Outline

Introduction

Methodological issues

Acknowledgments

References

Conclusions

A final puzzle

Assessing the relative complexity of the two systems

An outline of Mauritian conjugation

An outline of French conjugation

Application

A modified methodology

Issue 4: choosing the right classification

Issue 3: beware of phonology

Issue 2: don’t trust inflection classes

Issue 1: watch out for type frequency

Ackerman et al.’s strategy
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

<table>
<thead>
<tr>
<th>IC</th>
<th>INF</th>
<th>IPFV.3SG</th>
<th>lexeme</th>
<th>trans.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>sortir i</td>
<td>sortir e</td>
<td>sortir</td>
<td>‘go out’</td>
</tr>
<tr>
<td>2</td>
<td>amortir i</td>
<td>amortir is e</td>
<td>amortir</td>
<td>‘cushion’</td>
</tr>
<tr>
<td>3</td>
<td>laver</td>
<td>laver e</td>
<td>laver</td>
<td>‘wash’</td>
</tr>
<tr>
<td>4</td>
<td>vouloir</td>
<td>vouloir e</td>
<td>vouloir</td>
<td>‘want’</td>
</tr>
<tr>
<td>5</td>
<td>battre</td>
<td>battre e</td>
<td>battre</td>
<td>‘fight’</td>
</tr>
</tbody>
</table>

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

- The claim: this way of evaluating $H(\text{IPFV}|\text{INF})$ provides a rough measure of the difficulty of the PCFP for $\text{INF} \to \text{IPFV}$ 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.
Introduction

Methodological issues

Ackerman et al.’s strategy

Issue 1: watch out for type frequency
Issue 2: don’t trust inflection classes
Issue 3: beware of phonology
Issue 4: choosing the right classification

A modified methodology

Application

An outline of French conjugation
An outline of Mauritian conjugation
Assessing the relative complexity of the two systems
A final puzzle

Conclusions

References
(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

<table>
<thead>
<tr>
<th></th>
<th>IC</th>
<th>A</th>
<th>B</th>
<th>type freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>-i</td>
<td>-a</td>
<td>497</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>-i</td>
<td>-e</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>-i</td>
<td>-u</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>-i</td>
<td>-y</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>-o</td>
<td>-a</td>
<td>497</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>-o</td>
<td>-e</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>-o</td>
<td>-u</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>-o</td>
<td>-y</td>
<td>1</td>
</tr>
</tbody>
</table>

- Results:

\[
\begin{array}{cc}
\text{A} & \text{B} \\
\hline
\text{A} & 2 \\
\text{B} & 1 \\
\end{array}
\]

\[
\text{IC} = \text{H}(\text{col}|\text{row}), \text{without frequency}
\]

\[
\begin{array}{cc}
\text{A} & \text{B} \\
\hline
\text{A} & 0.0624 \\
\text{B} & 1 \\
\end{array}
\]

\[
\text{IC} = \text{H}(\text{col}|\text{row}), \text{with frequency}
\]
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 0.1 bit, 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

Introduction

**Methodological issues**

- Ackerman et al.’s strategy
- Issue 1: watch out for type frequency
- **Issue 2: don’t trust inflection classes**
- Issue 3: beware of phonology
- Issue 4: choosing the right classification

A modified methodology

Application

- An outline of French conjugation
- An outline of Mauritian conjugation
- Assessing the relative complexity of the two systems
- A final puzzle

Conclusions

References
The problem

- Extant inflectional classifications are generally not directly usable.
- Example: for French, it is traditional to distinguish
  - 4 infinitival suffixes -e, -i\text{\`e}, -waw, -\text{\`a}
  - Two types of imperfectives: with or without the augment -s-

<table>
<thead>
<tr>
<th>IC</th>
<th>INF</th>
<th>IPFV.3 SG</th>
<th>orth.</th>
<th>trans.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>sc\text{`e}t\text{i`e}</td>
<td>sc\text{`e}t\text{`e}</td>
<td>sortir</td>
<td>go out</td>
</tr>
<tr>
<td>2</td>
<td>am\text{`e}t\text{i`e}</td>
<td>am\text{`e}t\text{`i`e}s\text{`e}</td>
<td>amortir</td>
<td>cushion</td>
</tr>
<tr>
<td>3</td>
<td>lave</td>
<td>lave</td>
<td>laver</td>
<td>wash</td>
</tr>
<tr>
<td>4</td>
<td>vulwaw</td>
<td>vul\text{`e}</td>
<td>vouloir</td>
<td>want</td>
</tr>
<tr>
<td>5</td>
<td>bat\text{`a}</td>
<td>bat\text{`e}</td>
<td>battre</td>
<td>fight</td>
</tr>
</tbody>
</table>

- Observation: the choice of the infinitive suffix fully determines the form of the imperfective, except when the suffix is -\text{\`a}.
- For instance, \( H(\text{IPFV} \mid \text{INF} = \text{stem} \oplus \text{\`i\`e}) = 0 \)
The problem

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

<table>
<thead>
<tr>
<th>IC</th>
<th>INF</th>
<th>IPFV.3SG</th>
<th>lexeme</th>
<th>trans.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>sortir-ib</td>
<td>sortiré</td>
<td>sortir</td>
<td>go out</td>
</tr>
<tr>
<td>2</td>
<td>amortir-i</td>
<td>amortiré</td>
<td>amortir</td>
<td>cushion</td>
</tr>
</tbody>
</table>

- 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

Introduction

Methodological issues

Ackerman et al.’s strategy
Issue 1: watch out for type frequency
Issue 2: don’t trust inflection classes

Issue 3: beware of phonology

Issue 4: choosing the right classification

A modified methodology

Application

An outline of French conjugation
An outline of Mauritian conjugation
Assessing the relative complexity of the two systems

A final puzzle

Conclusions

References
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

<table>
<thead>
<tr>
<th>IPFV.1SG</th>
<th>IPFV.1PL</th>
<th>lexeme</th>
<th>trans.</th>
</tr>
</thead>
<tbody>
<tr>
<td>lavε</td>
<td>lavjō</td>
<td>LAVER</td>
<td>‘wash’</td>
</tr>
<tr>
<td>pɔʁtε</td>
<td>pɔʁtjɔ</td>
<td>PORTER</td>
<td>‘carry’</td>
</tr>
<tr>
<td>kɔʁtε</td>
<td>kɔʁtjɔ</td>
<td>CONTRER</td>
<td>‘counter’</td>
</tr>
<tr>
<td>pwavε</td>
<td>pwavjį</td>
<td>POIVRER</td>
<td>‘pepper’</td>
</tr>
</tbody>
</table>

- j → ∅ / j

<table>
<thead>
<tr>
<th>IPFV.1SG</th>
<th>IPFV.1PL</th>
<th>lexeme</th>
<th>trans.</th>
</tr>
</thead>
<tbody>
<tr>
<td>kajε</td>
<td>kajį</td>
<td>CAILLER</td>
<td>‘curdle’</td>
</tr>
<tr>
<td>pijε</td>
<td>pijį</td>
<td>PILLER</td>
<td>‘plunder’</td>
</tr>
<tr>
<td>kad pijε</td>
<td>kad pijį</td>
<td>QUADRILLER</td>
<td>‘cover’</td>
</tr>
<tr>
<td>vbijε</td>
<td>vbijį</td>
<td>VRILLER</td>
<td>‘pierce’</td>
</tr>
</tbody>
</table>
The problem

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

<table>
<thead>
<tr>
<th>IPFV.1PL</th>
<th>IPFV.1SG</th>
<th>lexeme</th>
<th>trans.</th>
</tr>
</thead>
<tbody>
<tr>
<td>kɔtɔiɔ̃</td>
<td>kɔtɔe</td>
<td>CONTRER</td>
<td>‘counter’</td>
</tr>
<tr>
<td>pwavɔiɔ̃</td>
<td>pwavɔe</td>
<td>POIVRER</td>
<td>‘pepper’</td>
</tr>
<tr>
<td>kadɔiɔ̃</td>
<td>kadɔiɔ̃</td>
<td>QUADRILLER</td>
<td>‘cover’</td>
</tr>
<tr>
<td>νɔiɔ̃</td>
<td>νɔiɔ̃</td>
<td>VRILLER</td>
<td>‘pierce’</td>
</tr>
</tbody>
</table>

- Not a small phenomenon: 294 IPFV.1PL in -iɔ̃ in our dataset

- Problem: this is often abstracted away from transcriptions

<table>
<thead>
<tr>
<th>lexeme</th>
<th>IPFV.1PL</th>
<th>surface transcription</th>
<th>BDLEX transcription</th>
</tr>
</thead>
<tbody>
<tr>
<td>POIVRER</td>
<td>poivrions</td>
<td>pwavɔiɔ̃</td>
<td>pwavɔiɔ̃</td>
</tr>
<tr>
<td>VRILLER</td>
<td>vrillions</td>
<td>νɔiɔ̃</td>
<td>νɔiɔ̃</td>
</tr>
</tbody>
</table>
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
Methodological issues

Issue 1: watch out for type frequency
Issue 2: don’t trust inflection classes
Issue 3: beware of phonology
Issue 4: choosing the right classification
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 classification.
  - Although many of these are trivial, there are at least 132 hard cases

- We need implemented algorithms for inferring 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

Introduction

Methodological issues

  Ackerman et al.’s strategy
  Issue 1: watch out for type frequency
  Issue 2: don’t trust inflection classes
  Issue 3: beware of phonology
  Issue 4: choosing the right classification

A modified methodology

Application

  An outline of French conjugation
  An outline of Mauritian conjugation
  Assessing the relative complexity of the two systems
  A final puzzle

Conclusions

References
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 $H(A \rightarrow B \mid A)$ 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.

<table>
<thead>
<tr>
<th>IC</th>
<th>INF</th>
<th>IPFV.3SG</th>
<th>pattern</th>
<th>classification of INF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>s écrit</td>
<td>s écrit</td>
<td>Xib → Xε</td>
<td>A = {Xib → Xε, Xβ → Xsε, Xβ → Xε}</td>
</tr>
<tr>
<td>2</td>
<td>am écrit</td>
<td>am écrit</td>
<td>Xβ → Xsε</td>
<td>A = {Xib → Xε, Xβ → Xsε, Xβ → Xε}</td>
</tr>
<tr>
<td>3</td>
<td>lav</td>
<td>lav ε</td>
<td>Xe → Xε</td>
<td>B = {Xε → Xe}</td>
</tr>
<tr>
<td>4</td>
<td>vulvab</td>
<td>vul ε</td>
<td>Xvunsab → Xε</td>
<td>C = {Xvunsab → Xε, Xβ → Xsε, Xβ → Xε}</td>
</tr>
<tr>
<td>5</td>
<td>bat</td>
<td>bat ε</td>
<td>Xβ → Xε</td>
<td>D = {Xβ → Xsε, Xβ → Xε}</td>
</tr>
</tbody>
</table>

- 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 $\sigma$ 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

Methodological issues
  Ackerman et al.’s strategy
  Issue 1: watch out for type frequency
  Issue 2: don’t trust inflection classes
  Issue 3: beware of phonology
  Issue 4: choosing the right classification

A modified methodology

Application
  An outline of French conjugation
  An outline of Mauritian conjugation
  Assessing the relative complexity of the two systems
  A final puzzle

Conclusions

References
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

Introduction

Methodological issues

  Ackerman et al.’s strategy
  Issue 1: watch out for type frequency
  Issue 2: don’t trust inflection classes
  Issue 3: beware of phonology
  Issue 4: choosing the right classification

A modified methodology

Application

  An outline of French conjugation
  An outline of Mauritian conjugation
  Assessing the relative complexity of the two systems
  A final puzzle

Conclusions

References
French paradigms

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

### Finite forms

<table>
<thead>
<tr>
<th>TAM</th>
<th>1SG</th>
<th>2SG</th>
<th>3SG</th>
<th>1PL</th>
<th>2PL</th>
<th>3PL</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRS.IND</td>
<td>lav</td>
<td>lav</td>
<td>lav</td>
<td>lav-ŷ</td>
<td>lav-e</td>
<td>lav</td>
</tr>
<tr>
<td>PST.IND.IP</td>
<td>lav-ε</td>
<td>lav-ε</td>
<td>lav-ε</td>
<td>lav-j-ŷ</td>
<td>lav-j-e</td>
<td>lav-ε</td>
</tr>
<tr>
<td>PST.PFV</td>
<td>lavε</td>
<td>lava</td>
<td>lava</td>
<td>lava-m</td>
<td>lava-t</td>
<td>lavε-ѵ</td>
</tr>
<tr>
<td>FUT.IND</td>
<td>lav génér-ε</td>
<td>lav génér-ε</td>
<td>lav génér-ε</td>
<td>lav génér-ŷ</td>
<td>lav génér-ε</td>
<td>lav génér-ε</td>
</tr>
<tr>
<td>PRS.SBJV</td>
<td>lav</td>
<td>lav</td>
<td>lav</td>
<td>lav-j-ŷ</td>
<td>lav-j-e</td>
<td>lav</td>
</tr>
<tr>
<td>PST.SBJV</td>
<td>lava-s</td>
<td>lava-s</td>
<td>lava</td>
<td>lava-s-j-ŷ</td>
<td>lava-s-j-e</td>
<td>lava-s</td>
</tr>
<tr>
<td>COND</td>
<td>lav génér-ε</td>
<td>lav génér-ε</td>
<td>lav génér-ε</td>
<td>lav génér-j-ŷ</td>
<td>lav génér-ε</td>
<td>lav génér-ε</td>
</tr>
<tr>
<td>IMP</td>
<td>---</td>
<td>lav</td>
<td>---</td>
<td>lav-ŷ</td>
<td>lav-e</td>
<td>---</td>
</tr>
</tbody>
</table>

### Nonfinite forms

<table>
<thead>
<tr>
<th>INF</th>
<th>PRS.PTCP</th>
<th>PST.PTCP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M.SG</td>
<td>F.SG</td>
</tr>
<tr>
<td>lave</td>
<td>lav-ɻ</td>
<td>lave</td>
</tr>
</tbody>
</table>
Morphomic stem alternations

  - no inflection classes of exponents
  - Intricate system of stem allomorphy relying on morphomic patterns

### Finite forms

<table>
<thead>
<tr>
<th>TAM</th>
<th>1SG</th>
<th>2SG</th>
<th>3SG</th>
<th>1PL</th>
<th>2PL</th>
<th>3PL</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRS.IND</td>
<td>stem₃</td>
<td>stem₃</td>
<td>stem₃</td>
<td>stem₁-ǐ</td>
<td>stem₁-e</td>
<td>stem₂</td>
</tr>
<tr>
<td>PST.IND.IPVF</td>
<td>stem₁-ε</td>
<td>stem₁-ε</td>
<td>stem₁-ε</td>
<td>stem₁-jǐ</td>
<td>stem₁-je</td>
<td>stem₁-ε</td>
</tr>
<tr>
<td>PST.PFV</td>
<td>stem_{11}</td>
<td>stem_{11}</td>
<td>stem_{11}</td>
<td>stem_{11}-m</td>
<td>stem_{11}-t</td>
<td>stem_{11}-r</td>
</tr>
<tr>
<td>FUT.IND</td>
<td>stem_{10}-κε</td>
<td>stem_{10}-κε</td>
<td>stem_{10}-κε</td>
<td>stem_{10}-κǐ</td>
<td>stem_{10}-κε</td>
<td>stem_{10}-κǐ</td>
</tr>
<tr>
<td>PRS.SBJV</td>
<td>stem₇</td>
<td>stem₇</td>
<td>stem₇</td>
<td>stem₈-ǰǐ</td>
<td>stem₈-je</td>
<td>stem₇</td>
</tr>
<tr>
<td>PST.SBJV</td>
<td>stem_{11}-s</td>
<td>stem_{11}-s</td>
<td>stem_{11}</td>
<td>stem_{11}-sį</td>
<td>stem_{11}-sje</td>
<td>stem_{11}-s</td>
</tr>
<tr>
<td>COND</td>
<td>stem_{10}-κε</td>
<td>stem_{10}-κε</td>
<td>stem_{10}-κε</td>
<td>stem_{10}-κį</td>
<td>stem_{10}-κje</td>
<td>stem_{10}-κε</td>
</tr>
<tr>
<td>IMP</td>
<td>---</td>
<td>stem₅</td>
<td>---</td>
<td>stem₆-į</td>
<td>stem₆-e</td>
<td>---</td>
</tr>
</tbody>
</table>

### Nonfinite forms

<table>
<thead>
<tr>
<th>INF</th>
<th>PRS.PTCP</th>
<th>PST.PTCP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M.SG</td>
<td>F.SG</td>
</tr>
<tr>
<td>stem₉</td>
<td>stem₄-ā</td>
<td>stem₁₂</td>
</tr>
</tbody>
</table>

Bonami, Boye & Henri () Measuring inflectional complexity January 15, 2011
Outline

Introduction

Methodological issues

- Ackerman et al.’s strategy
- Issue 1: watch out for type frequency
- Issue 2: don’t trust inflection classes
- Issue 3: beware of phonology
- Issue 4: choosing the right classification

A modified methodology

Application

- An outline of French conjugation
- An outline of Mauritian conjugation
- Assessing the relative complexity of the two systems
- A final puzzle

Conclusions

References
Sources of the Mauritian lexicon

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

<table>
<thead>
<tr>
<th>French</th>
<th>Mauritain</th>
<th>example</th>
<th>trans.</th>
</tr>
</thead>
<tbody>
<tr>
<td>s</td>
<td>detase</td>
<td>detaše</td>
<td>‘detach’</td>
</tr>
<tr>
<td>z</td>
<td>māze</td>
<td>māže</td>
<td>‘eat’</td>
</tr>
<tr>
<td>ι</td>
<td>pāti</td>
<td>pāti</td>
<td>‘leave’</td>
</tr>
<tr>
<td>y</td>
<td>fyme</td>
<td>fime</td>
<td>‘smoke’</td>
</tr>
<tr>
<td>η</td>
<td>bədone</td>
<td>bədone</td>
<td>‘give again’</td>
</tr>
<tr>
<td>ε</td>
<td>feə</td>
<td>feə</td>
<td>‘do’</td>
</tr>
<tr>
<td>ι</td>
<td>soəti</td>
<td>soəti</td>
<td>‘go out’</td>
</tr>
</tbody>
</table>

- A minority of lexemes borrowed from English, Hindi/Bhojpuri, Malagasy, (etc.)
Most Mauritian verbs have two forms: the long form (LF) and the short form (SF).

<table>
<thead>
<tr>
<th>LF</th>
<th>brize</th>
<th>vãde</th>
<th>amãde</th>
<th>ĕgziste</th>
<th>vini</th>
<th>brije</th>
<th>kõsiste</th>
<th>fini</th>
</tr>
</thead>
<tbody>
<tr>
<td>SF</td>
<td>briz</td>
<td>van</td>
<td>amãd</td>
<td>ĕgzis</td>
<td>vin</td>
<td>brije</td>
<td>kõsiste</td>
<td>fini</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>LF</th>
<th>brije</th>
<th>fini</th>
<th>vini</th>
<th>kōsiste</th>
<th>egziste</th>
<th>amāde</th>
<th>demāde</th>
</tr>
</thead>
<tbody>
<tr>
<td>SF</td>
<td>brije</td>
<td>brij</td>
<td>fini</td>
<td>vin</td>
<td>kōsiste</td>
<td>egzis</td>
<td>amād</td>
</tr>
<tr>
<td></td>
<td>‘glow’</td>
<td>‘mix’</td>
<td>‘finish’</td>
<td>‘come’</td>
<td>‘consist’</td>
<td>‘exist’</td>
<td>‘amend’</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LF</th>
<th>paste</th>
<th>pas</th>
<th>bāde</th>
<th>ban</th>
<th>fəiz</th>
<th>fəiz</th>
<th>febe</th>
<th>feər</th>
</tr>
</thead>
<tbody>
<tr>
<td>SF</td>
<td>pas</td>
<td>ban</td>
<td>fəiz</td>
<td>fəiz</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>‘filter’</td>
<td>‘pass’</td>
<td>‘bandage’</td>
<td>‘ban’</td>
<td>‘curl’</td>
<td>‘freeze’</td>
<td>‘shoe’</td>
<td>‘do’</td>
</tr>
</tbody>
</table>
Distribution of long and short forms

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

<table>
<thead>
<tr>
<th>Verum Focus</th>
<th>Syntax</th>
<th>SF</th>
<th>LF</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>V with nonclausal complements (NPs, APs, ADVPs, VPs, PPs)</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td></td>
<td>V with no complements</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td></td>
<td>V with clausal complements</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td></td>
<td>only extracted complements</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td></td>
<td>Verum Focus</td>
<td>no</td>
<td>yes</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Morphology</th>
</tr>
</thead>
<tbody>
<tr>
<td>reduplicant</td>
</tr>
<tr>
<td>base</td>
</tr>
</tbody>
</table>

Table: Constraints on verb form alternation
Outline

Introduction

Methodological issues

  Ackerman et al.’s strategy
  Issue 1: watch out for type frequency
  Issue 2: don’t trust inflection classes
  Issue 3: beware of phonology
  Issue 4: choosing the right classification

A modified methodology

Application

  An outline of French conjugation
  An outline of Mauritian conjugation

  Assessing the relative complexity of the two systems

A final puzzle

Conclusions

References
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 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:

<table>
<thead>
<tr>
<th>Language</th>
<th>Entropy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mauritian</td>
<td>0.744 bit</td>
</tr>
<tr>
<td>French</td>
<td>0.446 bit</td>
</tr>
</tbody>
</table>

☞ Notice that this is precisely contrary to expectations!
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):

<table>
<thead>
<tr>
<th>(Mauritian)</th>
<th>(French)</th>
<th>(Mauritian)</th>
<th>(French)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LF (\leftrightarrow) SF</td>
<td>INF (\leftrightarrow) PRS</td>
<td>SF (\leftrightarrow) LF</td>
<td>PRS (\leftrightarrow) INF</td>
</tr>
<tr>
<td>0.563</td>
<td>0.232</td>
<td>0.925</td>
<td>0.578</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>class</th>
<th>patterns</th>
<th>example</th>
<th># of lex.</th>
<th>entropy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>{Xe \rightarrow X, X \rightarrow X}</td>
<td>kwafe</td>
<td>1138</td>
<td>0.565</td>
</tr>
<tr>
<td>2</td>
<td>{Xte \rightarrow X, Xe \rightarrow X, X \rightarrow X}</td>
<td>gwijote</td>
<td>268</td>
<td>0.845</td>
</tr>
<tr>
<td>3</td>
<td>{X \rightarrow X}</td>
<td>sufe(\text{\textacute{e}})</td>
<td>225</td>
<td>0.0</td>
</tr>
<tr>
<td>4</td>
<td>{X(\text{\textacute{e}})e \rightarrow X(\text{\textacute{e}}), X(\text{\textacute{e}})e \rightarrow X, Xe \rightarrow X, X \rightarrow X}</td>
<td>kof(\text{\textacute{e}})e</td>
<td>159</td>
<td>0.835</td>
</tr>
<tr>
<td>5</td>
<td>{Xle \rightarrow X, Xe \rightarrow X, X \rightarrow X}</td>
<td>dekole</td>
<td>138</td>
<td>0.927</td>
</tr>
<tr>
<td>6</td>
<td>{Xi \rightarrow X, X \rightarrow X}</td>
<td>fini</td>
<td>116</td>
<td>0.173</td>
</tr>
<tr>
<td>7</td>
<td>{X(\text{\textacute{a}})de \rightarrow X(\text{\textacute{a}})n, Xe \rightarrow X, X \rightarrow X}</td>
<td>(\text{\textacute{a}})de</td>
<td>15</td>
<td>0.567</td>
</tr>
<tr>
<td>8</td>
<td>{Xble \rightarrow Xm, Xle \rightarrow X, Xe \rightarrow X, X \rightarrow X}</td>
<td>beduble</td>
<td>13</td>
<td>0.391</td>
</tr>
<tr>
<td>9</td>
<td>{X(\text{\textacute{e}})be \rightarrow X(\text{\textacute{e}})m, Xe \rightarrow X, X \rightarrow X}</td>
<td>plöbe</td>
<td>3</td>
<td>0.918</td>
</tr>
<tr>
<td>10</td>
<td>{X(\text{\textacute{e}})de \rightarrow X(\text{\textacute{e}})n, Xe \rightarrow X, X \rightarrow X}</td>
<td>fek(\text{\textacute{e}})de</td>
<td>4</td>
<td>0.811</td>
</tr>
</tbody>
</table>

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, -\(\text{\textacute{e}}\)e 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:

<table>
<thead>
<tr>
<th>class</th>
<th>patterns</th>
<th>example</th>
<th># of lex.</th>
<th>entropy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>{Xe \rightarrow X}</td>
<td>asyme asym</td>
<td>1279</td>
<td>0.0</td>
</tr>
<tr>
<td>2</td>
<td>{Xje \rightarrow Xi, Xje \rightarrow X, Xe \rightarrow X}</td>
<td>pij pij</td>
<td>171</td>
<td>1.515</td>
</tr>
<tr>
<td>3</td>
<td>{Xle \rightarrow vX, Xe \rightarrow X}</td>
<td>ale va</td>
<td>153</td>
<td>0.057</td>
</tr>
<tr>
<td>4</td>
<td>{Xīb \rightarrow X, Xīb \rightarrow X}</td>
<td>finīb fini</td>
<td>142</td>
<td>0.313</td>
</tr>
<tr>
<td>5</td>
<td>{Xd̄b \rightarrow X, Xīb \rightarrow X}</td>
<td>kud̄b ku</td>
<td>55</td>
<td>0.0</td>
</tr>
<tr>
<td>6</td>
<td>{Xt̄ib \rightarrow X, Xīb \rightarrow X, Xīb \rightarrow X}</td>
<td>pautīb paut</td>
<td>33</td>
<td>0.994</td>
</tr>
<tr>
<td>7</td>
<td>{Xt̄b \rightarrow X, Xīb \rightarrow X}</td>
<td>kon̄ēt̄b kon̄e</td>
<td>32</td>
<td>0.0</td>
</tr>
<tr>
<td>8</td>
<td>{Xye \rightarrow Xy, Xe \rightarrow X}</td>
<td>tye ty</td>
<td>31</td>
<td>0.0</td>
</tr>
<tr>
<td>9</td>
<td>{Xēn̄ib \rightarrow X, Xj̄ē \rightarrow X, Xīb \rightarrow X}</td>
<td>v̄n̄ēn̄ib v̄j̄ ē</td>
<td>22</td>
<td>0.0</td>
</tr>
<tr>
<td>10</td>
<td>{Xīb \rightarrow X}</td>
<td>f̄ēb fE</td>
<td>21</td>
<td>0.0</td>
</tr>
</tbody>
</table>

... ... ... ... ... ...

(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

Introduction

Methodological issues
  Ackerman et al.’s strategy
  Issue 1: watch out for type frequency
  Issue 2: don’t trust inflection classes
  Issue 3: beware of phonology
  Issue 4: choosing the right classification

A modified methodology

Application
  An outline of French conjugation
  An outline of Mauritian conjugation
  Assessing the relative complexity of the two systems
  A final puzzle

Conclusions

References
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:

<table>
<thead>
<tr>
<th>(Mauritian)</th>
<th>(French)</th>
<th>(Mauritian)</th>
<th>(French)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LF $\mapsto$ SF</td>
<td>INF $\mapsto$ PRS</td>
<td>SF $\mapsto$ LF</td>
<td>PRS $\mapsto$ INF</td>
</tr>
<tr>
<td>96.82%</td>
<td>94.77%</td>
<td>93.18%</td>
<td>91.86%</td>
</tr>
</tbody>
</table>

- 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:

<table>
<thead>
<tr>
<th>LF</th>
<th>SF</th>
<th>pattern</th>
<th>set of possible patterns</th>
</tr>
</thead>
<tbody>
<tr>
<td>egziste</td>
<td>egzis</td>
<td>Xte → X</td>
<td>{Xte → X, Xe → X, X → X}</td>
</tr>
<tr>
<td>reste</td>
<td>res</td>
<td>Xte → X</td>
<td>{Xte → X, Xe → X, X → X}</td>
</tr>
<tr>
<td>așete</td>
<td>așet</td>
<td>Xe → X</td>
<td>{Xte → X, Xe → X, X → X}</td>
</tr>
<tr>
<td>eșite</td>
<td>eșit</td>
<td>Xe → X</td>
<td>{Xte → X, Xe → X, X → X}</td>
</tr>
<tr>
<td>kősiste</td>
<td>kősiste</td>
<td>X → X</td>
<td>{Xte → X, Xe → X, X → X}</td>
</tr>
<tr>
<td>afekte</td>
<td>afekte</td>
<td>X → X</td>
<td>{Xte → X, Xe → X, X → X}</td>
</tr>
</tbody>
</table>

- All verbs exhibiting the pattern pattern Xte → X end in ste
- The stem maximization algorithm does not see this; as a result egziste and așete end up in the same class, with non-null entropy
- By contrast, the MGL derives a rule [te → ∅/s_#] that does not apply to așete

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 $H(A \rightarrow B | A)$ as usual.

Results:

<table>
<thead>
<tr>
<th>classification method</th>
<th>(Mauritian)</th>
<th>(French)</th>
<th>(Mauritian)</th>
<th>(French)</th>
</tr>
</thead>
<tbody>
<tr>
<td>stem max.</td>
<td>0.563</td>
<td>0.232</td>
<td>0.925</td>
<td>0.578</td>
</tr>
<tr>
<td>MGL rules</td>
<td>0.476</td>
<td>0.079</td>
<td>0.846</td>
<td>0.296</td>
</tr>
</tbody>
</table>

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 $X \rightarrow X$ is used 31% of the time
  - However for verbs in Cje it is always used
  - The MGL uses this information to decide that
    - copje$\rightarrow$copje is more likely than copje$\rightarrow$copj
    - stope$\rightarrow$stop is more likely than stope$\rightarrow$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

Introduction

Methodological issues
- Ackerman et al.’s strategy
- Issue 1: watch out for type frequency
- Issue 2: don’t trust inflection classes
- Issue 3: beware of phonology
- Issue 4: choosing the right classification

A modified methodology

Application
- An outline of French conjugation
- An outline of Mauritian conjugation
- Assessing the relative complexity of the two systems
- A final puzzle

Conclusions

References
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.


