Predicting article agglutination in Mauritian

Olivier Bonami\textsuperscript{1} \quad Fabiola Henri\textsuperscript{2}

\textsuperscript{1}U. Paris-Sorbonne & IUF & LLF
olivier.bonami@paris-sorbonne.fr

\textsuperscript{2}U. Paris 7 & LLF
henrifabiola@gmail.com

Formal Approaches to Creole Studies
Lisbonne, November 2012
Outline

Introduction

A semi-automatic exploration of article agglutination
  The dataset
  Alternating forms
  Factors to consider

Statistical analysis and modeling
  Identifying correlations
  Statistical modeling
  Towards explaining unexpected correlations

Conclusions

References
The issue

- Creole words often agglutinate the phonology of two words from the lexifier
  - In French-based Creoles, inherited nouns come from the reanalysis of the sequence det + noun in the lexifier.

<table>
<thead>
<tr>
<th>French determiners</th>
<th>noun</th>
<th>example</th>
<th>trans.</th>
</tr>
</thead>
<tbody>
<tr>
<td>la ⊕</td>
<td>tabl</td>
<td>latab</td>
<td>‘table’</td>
</tr>
<tr>
<td>le ⊕</td>
<td>tuʁ</td>
<td>letuʁ</td>
<td>‘turn’</td>
</tr>
<tr>
<td>⊕ li</td>
<td>lili</td>
<td>‘bed’</td>
<td></td>
</tr>
<tr>
<td>l ⊕ amuʁ</td>
<td>lamuʁ</td>
<td>‘love’</td>
<td></td>
</tr>
<tr>
<td>dy ⊕</td>
<td>te</td>
<td>dite</td>
<td>‘tea’</td>
</tr>
<tr>
<td>dɛl ⊕</td>
<td>o</td>
<td>delo</td>
<td>‘water’</td>
</tr>
<tr>
<td>ma ⊕</td>
<td>tãt</td>
<td>matãt</td>
<td>‘aunt’</td>
</tr>
<tr>
<td>mɔ̃ ⊕</td>
<td>peʁ</td>
<td>mɔ̃peʁ</td>
<td>‘father’</td>
</tr>
<tr>
<td>ind.-n ⊕</td>
<td>espɛs</td>
<td>nespɛs</td>
<td>‘species’</td>
</tr>
<tr>
<td>plur.-z ⊕</td>
<td>animo</td>
<td>zanimo</td>
<td>‘animal’</td>
</tr>
</tbody>
</table>

- Also known as article ‘incorporation’ but is a misnomer
  - We avoid the term incorporation because of its use in morphology and syntax
Introduction

- Previous work on article agglutination has focused on
  - Why there is substantial agglutination in Mauritian compared to other French-based Creoles (Baker, 1984; Grant, 1995)
    - Different question: what favors agglutination?
    - However, our approach is compatible with an initial substratic influence
  - Predicting variation in the form of the agglutinated string (Strandquist, 2003)
  - Reanalysis of the sequence as a case of interrupted transmission (McWhorter and Parkvall, 2002)

- We examine visible factors in the lexicon of contemporary French which correlate with article agglutination

- Interpretation of this correlation
  - Substantiation or not of the previous observations
  - Addition of new factors that we think correlate with the phenomenon
  - We rely thoroughly on quantitative data both on the lexifier and on the creole
Outline

Introduction

A semi-automatic exploration of article agglutination
The dataset
  Alternating forms
  Factors to consider

Statistical analysis and modeling
  Identifying correlations
  Statistical modeling
  Towards explaining unexpected correlations

Conclusions

References
Sources

- **Database of 8800** nouns compiled on the basis of forms available in Carpooran (2011)
  - We coded their etyma and language of origin and
  - their agglutinated forms together with information regarding alternation
- For those **7325** nouns which are undisputably from French origin, we coded their
  - frequency of the etyma obtained from *Lexique 3* (New et al., 2007)
  - gender obtained from *Lexique 3* (New et al., 2007)
  - frequency with the definite compiled from the French subtitles corpus (New and Spinelli, 2012)
  - attestation dates obtained from the CNRTL
Sources

- We examine only a subset of the collected data, i.e. those of undisputably French origin

<table>
<thead>
<tr>
<th>Etymology</th>
<th>Size</th>
<th>Example</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>French</td>
<td>7325</td>
<td>latab</td>
<td>‘table’</td>
</tr>
<tr>
<td></td>
<td></td>
<td>lestoma</td>
<td>‘stomach’</td>
</tr>
<tr>
<td>Other</td>
<td>1169</td>
<td>zugader</td>
<td>‘player’</td>
</tr>
<tr>
<td></td>
<td></td>
<td>larurut</td>
<td>‘arrow-root’</td>
</tr>
<tr>
<td>Creole innovation</td>
<td>306</td>
<td>koze</td>
<td>‘talk’</td>
</tr>
<tr>
<td></td>
<td></td>
<td>săte</td>
<td>‘song’</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td>8800</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Agglutination of the French article can be erratically found with inherited nouns of other origin

<table>
<thead>
<tr>
<th>Etymology</th>
<th>Status</th>
<th>Example</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>undisputable</td>
<td>larurut</td>
<td>‘arrow-root’</td>
</tr>
<tr>
<td>Arabic</td>
<td>disputable</td>
<td>larak</td>
<td>‘arak’</td>
</tr>
<tr>
<td>Hindi</td>
<td>disputable</td>
<td>lamaø</td>
<td>‘rice water’</td>
</tr>
</tbody>
</table>
Phonetizing the subset

- We phonetized the Mauritian nouns based on their orthography.
- With phonetized etymons obtained from flexique, we conducted a semi-automatic reconstruction of phonetic changes from French to Mauritian.
  - Work in Progress: Semi-automatic reconstruction of phonetic changes occurring with inherited words.
- With these informations, we are able to provide a matching between the phonetic pattern of a noun and that of its etymon.
  - Currently available for 4881 of the remaining 7325 nouns.
Focussing on the definite

- We further focus on the subset which involves agglutination with the definite article
  - The other types are either rare or ambiguous

<table>
<thead>
<tr>
<th>French article</th>
<th>noun</th>
<th>example</th>
<th>trans.</th>
<th>size</th>
</tr>
</thead>
<tbody>
<tr>
<td>la +</td>
<td>tabl</td>
<td>latab</td>
<td>'table'</td>
<td>457</td>
</tr>
<tr>
<td>le +</td>
<td>tuʁ</td>
<td>letuʁ</td>
<td>'turn'</td>
<td>49</td>
</tr>
<tr>
<td>+ li</td>
<td>lili</td>
<td>'bed'</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>l + amuʁ</td>
<td>lamuʁ</td>
<td>'love'</td>
<td>723</td>
<td></td>
</tr>
<tr>
<td>dy +</td>
<td>te</td>
<td>dite</td>
<td>'tea'</td>
<td>37</td>
</tr>
<tr>
<td>dəl + o</td>
<td>delo</td>
<td>'water'</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>ma + tãt</td>
<td>matãt</td>
<td>'aunt'</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>mõ + rɛʁ</td>
<td>mõpeʁ</td>
<td>'father'</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>ind.-n + espɛs</td>
<td>nespes</td>
<td>'species'</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>plur.-z + animo</td>
<td>zanimo</td>
<td>'animal'</td>
<td>62</td>
<td></td>
</tr>
</tbody>
</table>

**TOTAL** 1350

- Final count of dataset examined is **4760**
Outline

Introduction

A semi-automatic exploration of article agglutination
  The dataset
  Alternating forms
  Factors to consider

Statistical analysis and modeling
  Identifying correlations
  Statistical modeling
  Towards explaining unexpected correlations

Conclusions

References
Distribution

- Mauritian particular wrt. to agglutination since it has by far reanalyzed the article-noun sequence as a single noun (Grant, 1995)
  - Interestingly, it has also developed a subset of alternating forms

(1) a. *Donn mwa enn liv pwason.*
  give.SF 1SG.STF IND pound fish
  ‘Give me a pound of fish.’

  b. *Komie ou dir laliv?*
  how _much 2SG.FOR say.SF pound
  ‘How much do you say the pound?*

- Among the 1240 nouns with an agglutinated form, 275 are alternating according to Carpooran (2011)
Alternating forms in the count

- We don’t know whether both forms appeared simultaneously
  - Both old and new imports have been found to be alternating

<table>
<thead>
<tr>
<th>Bare form</th>
<th>aggl. form</th>
<th>age</th>
<th>trans.</th>
</tr>
</thead>
<tbody>
<tr>
<td>koloni</td>
<td>lakoloni</td>
<td>1308</td>
<td>‘colony’</td>
</tr>
<tr>
<td>aeropor</td>
<td>laeropor</td>
<td>1922</td>
<td>‘airport’</td>
</tr>
<tr>
<td>ātet</td>
<td>lātet</td>
<td>1838</td>
<td>‘heading’</td>
</tr>
<tr>
<td>dās</td>
<td>ladās</td>
<td>1172</td>
<td>‘dance’</td>
</tr>
</tbody>
</table>

- Accounting for the distribution of alternating forms is a complicated task for various reasons
  - It seems that alternating forms are not randomly distributed but are however subject to dialectal variation
  - Agglutinated forms may have acquired complex properties that should be confirmed by corpus and experimental data

- We hence limit our study to predicting agglutination for the reasons mentionned above
  - Alternating forms are assessed together with agglutinated nonalternating ones
Outline

Introduction

A semi-automatic exploration of article agglutination
  The dataset
  Alternating forms
  Factors to consider

Statistical analysis and modeling
  Identifying correlations
  Statistical modeling
  Towards explaining unexpected correlations

Conclusions

References
Factors contributing to agglutination

- Previous studies have claimed that a number of factors correlate with agglutination in Mauritian

(2)  
  a. Frequency of collocation  
  b. Homophony  
  c. Substratic influence  
  d. Number of syllables  
  e. Vowel harmony

- All these studies are based on handpicked small samples and are in need of quantitative substantiation
Vowel harmony as a factor

- Following the idea that article agglutination is modelled on noun class prefixes from Bantu languages, Strandquist (2003) argue that vowel harmony, also a characteristic of the same languages, is also determining.
  - It crucially has an effect on those nouns that will be agglutinated and those that won’t
  - If the article’s vowel is in harmony with the noun, it will be agglutinated. If not then either the vowel harmonizes or agglutination does not occur

(3) a. lǝky > liki (the ass > female genitals)
   b. lǝʃjẽ > lisjẽ (the dog > dog)

- We claim (Contra Strandquist, 2003) that the correlation between vowel harmony and agglutination (le vs li) is real (Fisher’s exact test, \( p < 0.0001 \)), but it is not categorical and epiphenomenal (60 candidates, 1.3% of the data)
  - Nouns like letur, ledwa or lekuyš, for instance, do not harmonize
Factors contributing to agglutination

- We further add to the above, the following factors which are generally relevant when it comes to word formation (e.g. Plénat, 2000)

  (4)  
  a. Gender  
  b. Dissimilation effects  
  c. Phonotactic preferences  
  d. Raw frequency

- Although there can be other factors to be considered (syntactic, semantic, ...), we limit our study to the above

  Our choice is also limited by the data available
Outline

Introduction

A semi-automatic exploration of article agglutination
  The dataset
  Alternating forms
  Factors to consider

Statistical analysis and modeling
  Identifying correlations
  Statistical modeling
  Towards explaining unexpected correlations

Conclusions

References
The difference between monosyllabic and polysyllabic is highly significant ($\chi^2$ test $p < 2 \times 10^{-16}$)

For polysyllabic words, length is barely significant ($\chi^2$ test, $p \approx 0.04$)
Relevant factors: gender

Gender (NEU = ambiguous between f and m)

Counts

Proportion of agglutination

Bonami & Henri (Paris/LLF)
Statistical analysis and modeling

Identifying correlations

Relevant factors: initial segment type

Bonami & Henri (Paris/LLF)
Irrelevant factors: dissimilation

- We expect a dissimilation effect: etymons beginning in /l should disfavor agglutination.

- We find no such effect. Small effect to the contrary, but barely significant ($\chi^2$ test $p \approx 0.036$)
Irrelevant factors: homophony

- We expect a homophony avoidance effect: existence of a verb homonym should favor agglutination.

We do find an effect, but in the opposite direction ($\chi^2$ test $p \approx 0.008$)

Anyway, not enough relevant data for this to be important.
Relevant factor: age

- We plot the date of first attestation of an etymon, grouped by century:

- Words with more recent French etymons agglutinate less. (logistic regression likelihood ratio: $\chi^2$ test $p < 0.0001$)

- Probable combination of factors: frequency, date of Entry in the Mauritian lexicon.
Relevant factor: collocation with SG definite article

- Agglutination grows when collocation with the definite article grows.

- Highly significant. (logistic regression likelihood ratio: $\chi^2$ test $p < 0.0001$)

- Surprisingly so, given that we are estimating on the basis of late 20th century frequency information.
Relevant factor: raw frequency of etymon

- Agglutination grows when frequency of the etymon.

Highly significant. (logistic regression likelihood ratio: \( \chi^2 \) test \( p < 0.0001 \))
Summing up

We have seen that the following features of the etymon all contribute individually to predicting article incorporation:

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Accuracy</th>
<th>Acc. increase</th>
<th>$D_{x,y}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monosyllabic or Polysyllabic</td>
<td>0.8252101</td>
<td>0</td>
<td>0.180</td>
</tr>
<tr>
<td>Initial segment type</td>
<td>0.8382353</td>
<td><strong>0.0130252</strong></td>
<td>0.464</td>
</tr>
<tr>
<td>Gender</td>
<td>0.8252101</td>
<td>0</td>
<td>0.274</td>
</tr>
<tr>
<td>Frequency of coll. with DEF.SG</td>
<td>0.8252101</td>
<td>0</td>
<td>0.274</td>
</tr>
<tr>
<td>Raw frequency</td>
<td>0.8254202</td>
<td><strong>0.0002101</strong></td>
<td>0.349</td>
</tr>
<tr>
<td>Age</td>
<td>0.8252101</td>
<td>0</td>
<td>0.277</td>
</tr>
<tr>
<td>Baseline (no predictor)</td>
<td><strong>0.8252101</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

However each is a pretty bad predictor on its own.

A series of simple logistic regressions confirm this: no predictor leads to a sizable increase in accuracy.
Outline

Introduction

A semi-automatic exploration of article agglutination
  The dataset
  Alternating forms
  Factors to consider

Statistical analysis and modeling
  Identifying correlations
  Statistical modeling
  Towards explaining unexpected correlations

Conclusions

References
To assess the joint predictive character of the predictors, we run a multiple logistic regression.

Model:

\[
P(\text{agglutination}|\vec{X}) = \frac{e^{\alpha + \beta \vec{X}}}{1 + e^{\alpha + \beta \vec{X}}}
\]

Where:

- \( \alpha = -4.4573 \)
- \( \beta_1 X_1 = 1.9814 \times \text{monosyllabic} \)
- \( \beta_2 X_2 = 3.8591 \times \text{vowel}_\text{initial} \)
- \( \beta_3 X_3 = 0.6507 \times \text{log}_\text{frequency} \)
- \( \beta_4 X_4 = 0.8935 \times \text{def}_\text{sg}_\text{rel}_\text{frequency} \)
- \( \beta_5 X_5 = 1.2750 \times \text{feminine} \)
- \( \beta_6 X_6 = -0.3156 \times \text{century}_\text{of}_\text{attestation} \)
Assessing the model

- Simple but unsubtle measure of the quality of the model:
  - How often does the model correctly assign a probability above 0.5 for agglutinating forms and below 0.5 for bare forms?
  - I.e., how often does it make the right prediction?
- Answer: 88.5% of the time

<table>
<thead>
<tr>
<th>Predictors</th>
<th>Accuracy</th>
<th>Acc. increase</th>
<th>$D_{x,y}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>all</td>
<td>0.8798319</td>
<td>0.0546218</td>
<td>0.846</td>
</tr>
<tr>
<td>Baseline (no predictor)</td>
<td>0.8252101</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

This is an unsubtle measure because it does not make a difference between a probability of 0.51 and a probability of 0.99.
Assessing the model: the gory details

- **Good likelihood:** $\chi^2$ test $p < 0.0001$
  - It is very unlikely that the stated predictors do not jointly explain the probability of agglutination
- **Good predictors:** all predictors have a Wald Z value with $p < 0.0001$.
  - All predictors do make some contribution to the model
- **Good prediction:** rank correlation $D_{x,y} = 0.846$
  - The model predicts quite reliably the probability of agglutination for any combination of predictor values in the dataset
- **Little overfitting:** mean rank correlation for 200 bootstrap samples $D_{x,y} = 0.845$
  - The model’s accuracy does not change when it is trained on a slightly different dataset.
Outline

Introduction

A semi-automatic exploration of article agglutination
  The dataset
  Alternating forms
  Factors to consider

Statistical analysis and modeling
  Identifying correlations
  Statistical modeling
    Towards explaining unexpected correlations

Conclusions

References
Are raw frequency and DEF.SG frequency correlated?

- There is no obvious explanation for frequency favoring agglutination
- One hypothesis to refute: little multicolinearity between raw frequency and proportion of collocation with the DEF.SG article

![Scatter plot showing the relationship between raw frequency and proportion of collocation with DEF.SG. The linear regression slope is 0.011976, with $R^2 < 0.02$.](image)

Linear regression slope $0.011976$, $R^2 < 0.02$
Gender and age

- The remaining two surprising predictors are gender and age.
- Hypothesis: age might be relevant in the sense that words that belong to the creole from its formation behave with respect to agglutination differently from words borrowed from French by native creole speakers.
- To assess this, we ran separate regressions on newer (post-1800) and older (pre-1800) nouns.
Gender and age, continued

<table>
<thead>
<tr>
<th>predictors</th>
<th>Older nouns coefficients</th>
<th>Older nouns p values</th>
<th>Newer nouns coefficients</th>
<th>Newer nouns p values</th>
</tr>
</thead>
<tbody>
<tr>
<td>monosyllabic</td>
<td>2.0083</td>
<td>&lt; 0.0001</td>
<td>0.9618</td>
<td>0.2313</td>
</tr>
<tr>
<td>vowel_initial</td>
<td>3.8709</td>
<td>&lt; 0.0001</td>
<td>3.5815</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>frequency</td>
<td>0.6230</td>
<td>&lt; 0.0001</td>
<td>0.6997</td>
<td>0.0030</td>
</tr>
<tr>
<td>def.sg_rel_freq.</td>
<td>0.8836</td>
<td>&lt; 0.0001</td>
<td>0.9503</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>feminine</td>
<td>1.3261</td>
<td>&lt; 0.0001</td>
<td>-0.3621</td>
<td>0.4583</td>
</tr>
<tr>
<td>century</td>
<td>-0.2604</td>
<td>&lt; 0.0001</td>
<td>-0.3835</td>
<td>0.2311</td>
</tr>
</tbody>
</table>

\[ D_{x,y} = 0.839 \quad D_{x,y} = 0.890 \]
Conclusion: in the post-creolization period, monosyllabicity, gender and age probably stopped playing a role.

This might be interpreted as supporting the hypothesis of a substratic influence (Baker, 1984; Grant, 1995):

- The substrate influence hypothesis assumes that the French article was analogized by creole speakers to a nominal class marker.
- The feminine article makes for a much better class marker than the masculine:
  - The feminine has a full vowel, the masculine a droppable schwa.
  - Because of schwa drop, the masculine is often identical to the gender neutral prevocalic /,
- Full adoption of the creole leads to the disappearance of Bantu influence, and hence to the disappearance of a preference for feminine agglutination.
Outline

Introduction

A semi-automatic exploration of article agglutination
  The dataset
  Alternating forms
  Factors to consider

Statistical analysis and modeling
  Identifying correlations
  Statistical modeling
  Towards explaining unexpected correlations

Conclusions

References
Using statistical modeling over large datasets, we showed that:

- Article agglutination happens throughout the history of Mauritian, up to this day.

- Agglutination comes in two varieties:
  - Lexically fixed: one form per lexical item, either agglutinated or not.
  - Variable: two forms, with syntactic/semantic conditioning over their respective use
    - Further research will determine whether this should be considered plain inflectional morphology.

- Various causes contribute to the distribution of agglutination, in a non categorical fashion.

- The causes may have varied over the course of the history of the language.

- Where relevant data is available, creole linguistics has some use for the statistician (Robinson, 2008).


