Predicting article agglutination in Mauritian

Olivier Bonami[1] Fabiola Henri[2]

[1]U. Paris-Sorbonne & IUF & LLF
olivier.bonami@paris-sorbonne.fr

[2]U. Paris 7 & LLF
henrifabiola@gmail.com

Formal Approaches to Creole Studies
Lisbonne, November 2012

Outline

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.

French determiners		noun	example	trans.
la	⊕	tabl	latab	'table'
le	⊕	tuʁ	letuɚ	'turn'
	⊕	li	lili	'bed'
l	⊕	amuʁ	lamuɚ	'love'
dy	⊕	te	dite	'tea'
dəl	⊕	o	delo	'water'
ma	⊕	tãt	matãt	'aunt'
mɔ̃	⊕	pɛʁ	mɔ̃peɚ	'father'
ind.-n	⊕	ɛspɛs	nespes	'species'
plur.-z	⊕	animo	zanimo	'animal'

  - 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

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** `http://www.cnrtl.fr/`

Sources

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

Etymon	Size	example	trans.
French	**7325**	latab	'table'
		lestoma	'stomach'
Other	1169	zugader	'player'
		larurut	'arrow-root'
Creole innovation	306	koze	'talk'
		sɑ̃te	'song'
TOTAL	**8800**		

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

Etymon	Status	example	trans.
English	undisputable	larurut	'arrow-root'
Arabic	disputable	larak	'arak'
Hindi	disputable	lamaɔ	'rice water'

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 occuring 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

French article		noun	example	trans.	size
la	⊕	tabl	latab	'table'	457
le	⊕	tuʁ	letuʁ	'turn'	49
	⊕	li	lili	'bed'	11
l	⊕	amuʁ	lamuʁ	'love'	723
dy	⊕	te	dite	'tea'	37
dəl	⊕	o	delo	'water'	3
ma	⊕	tãt	matãt	'aunt'	3
mɔ̃	⊕	pɛʁ	mɔ̃peʁ	'father'	1
ind.-n	⊕	ɛspɛs	nespes	'species'	4
plur.-z	⊕	animo	zanimo	'animal'	62
TOTAL					**1350**

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

Outline

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 developped 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 simultaneaously
  - ☞ Both old and new imports have been found to be alternating

Bare form	aggl. form	age	trans.
koloni	lakoloni	1308	'colony'
aeropor	laeropor	1922	'airport'
ãtet	lãtet	1838	'heading'
dãs	ladãs	1172	'dance'

- 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

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. Susbtratic 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. ləky > liki (the ass > female genitals)
b. ləʃjɛ̃ > lisjɛ̃ (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 lekuyɔ̃, for instance, do not harmonize

Factors contributing to agglutination

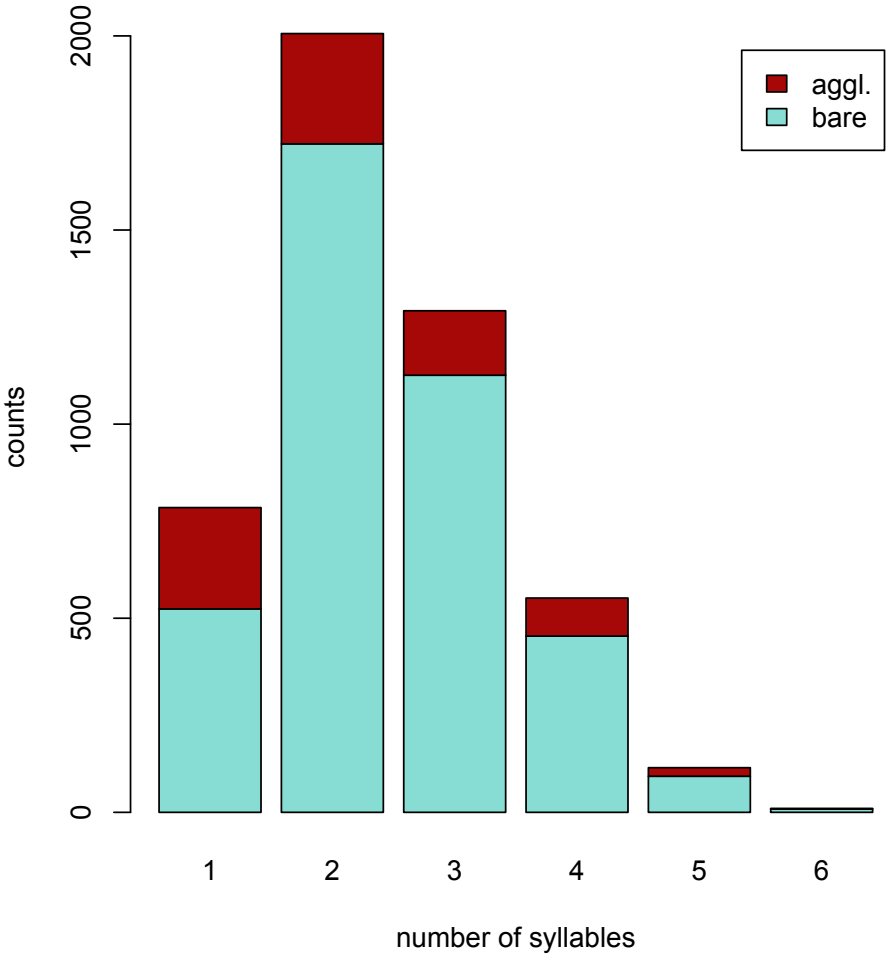
- We further add to the above, the following factors which are generally relevant when it comes to word formation (e.g. Plénat, 2000)

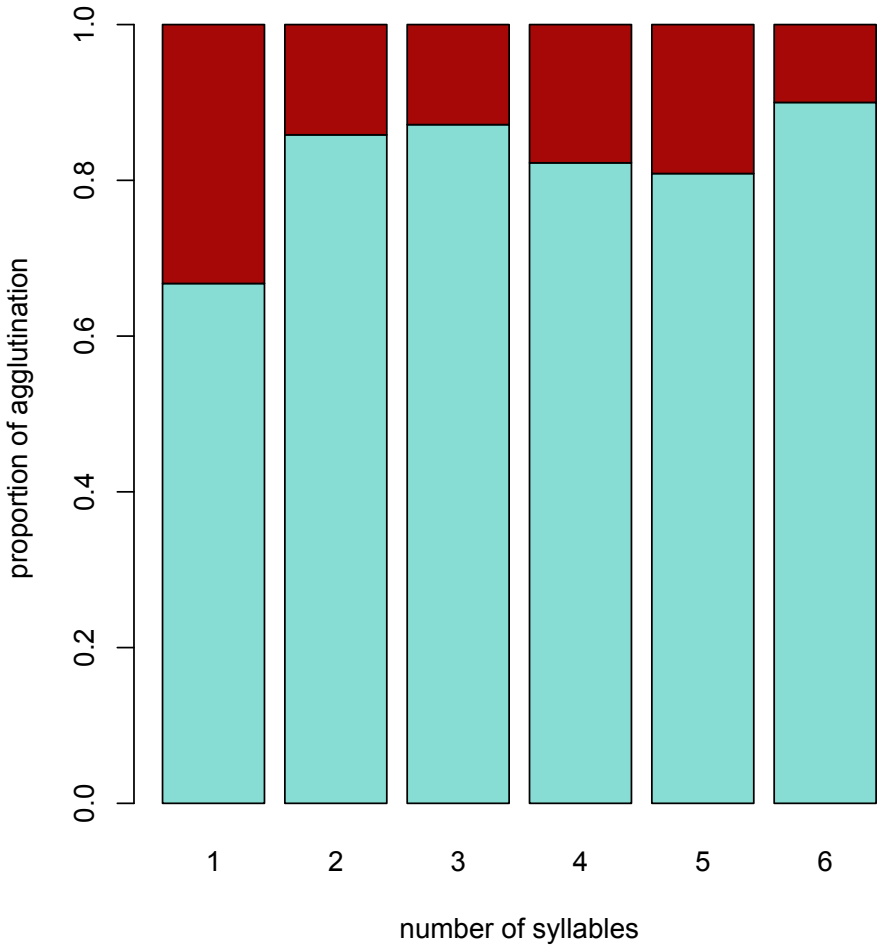
(4)
a. Gender
b. Dissimilation effects
c. Phonotactic preferences
d. Raw frequency

- 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

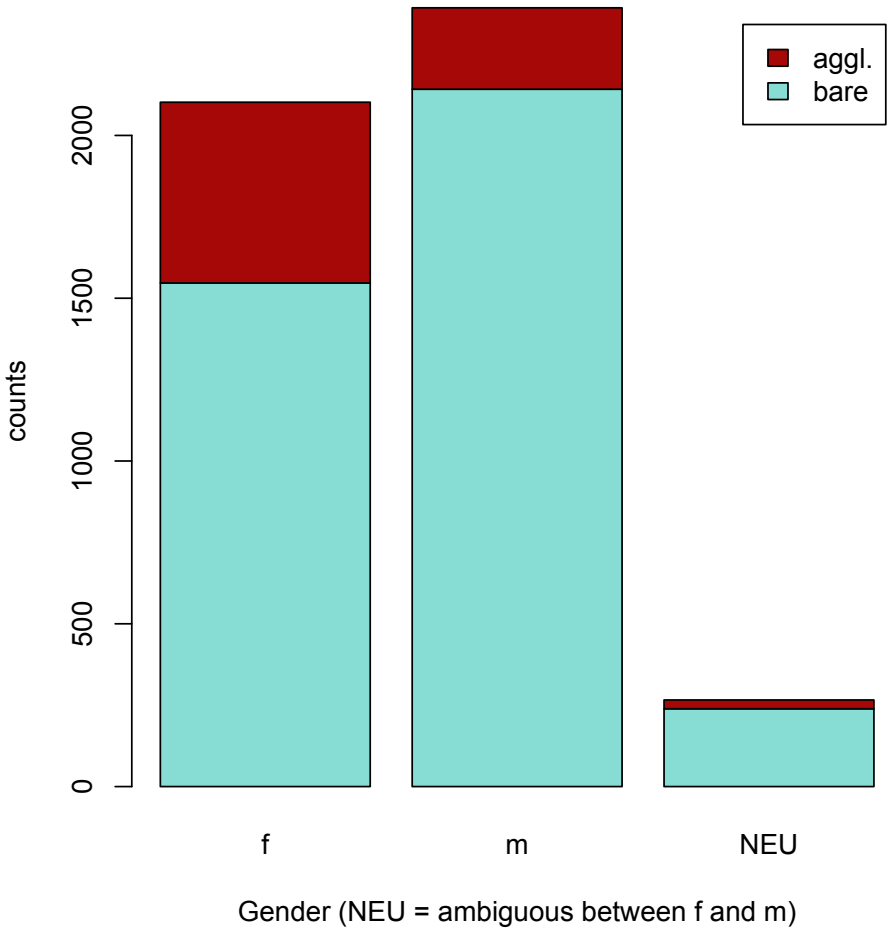
Relevant factors: length

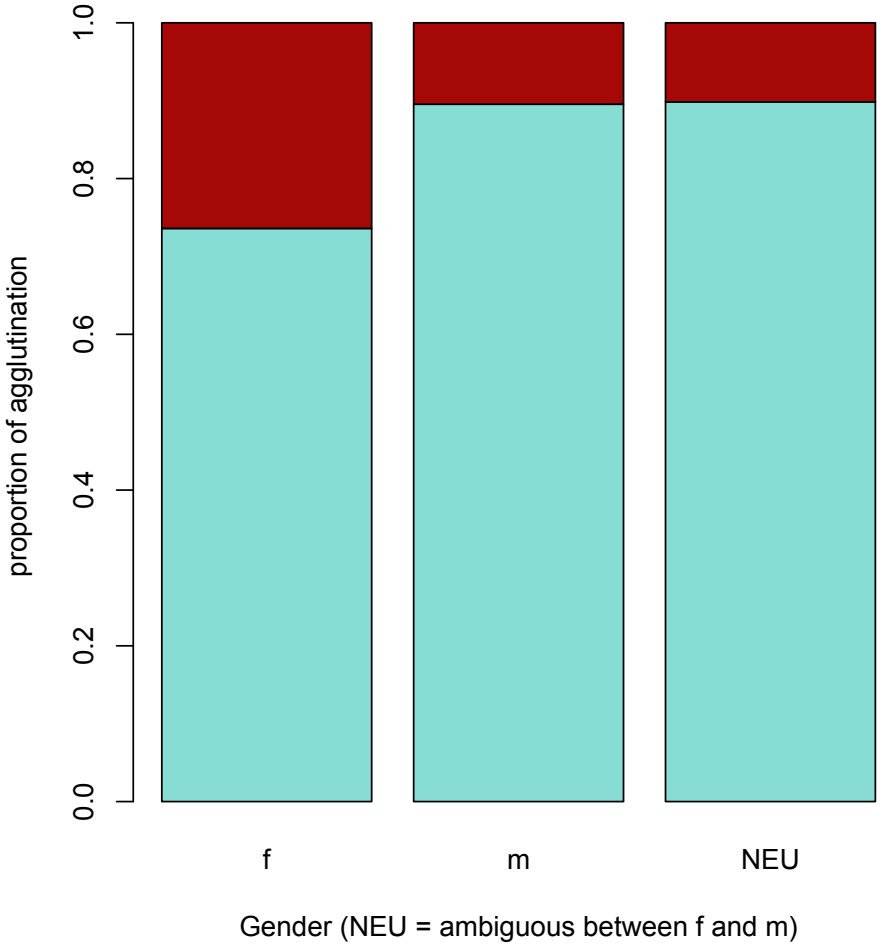




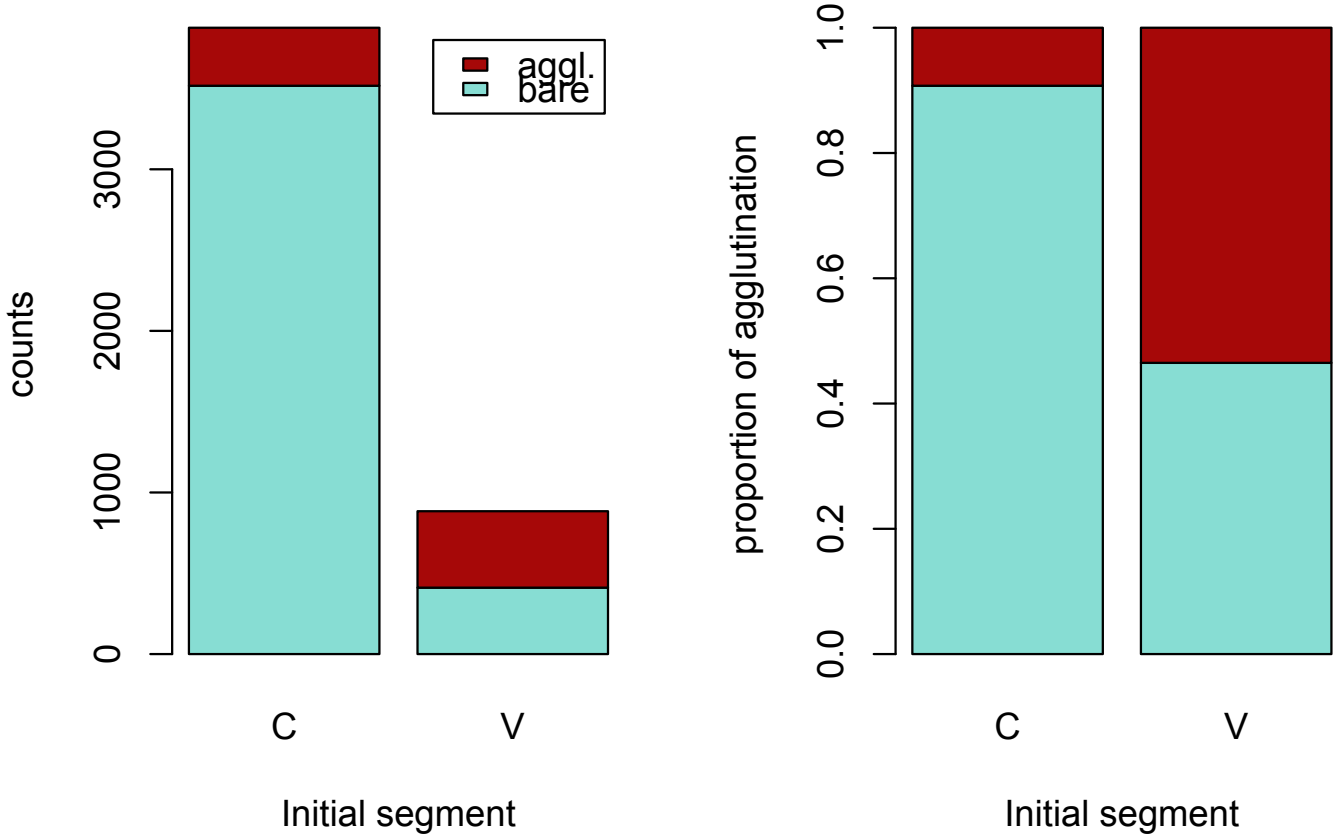
- The difference between monosyllabic and polysyllabic is highly significant (χ^2 test $p < 2 \times 10^{-16}$)
- For polysyllabic words, length is barely significant (χ^2 test ,$p \approx 0.04$)

Relevant factors: gender



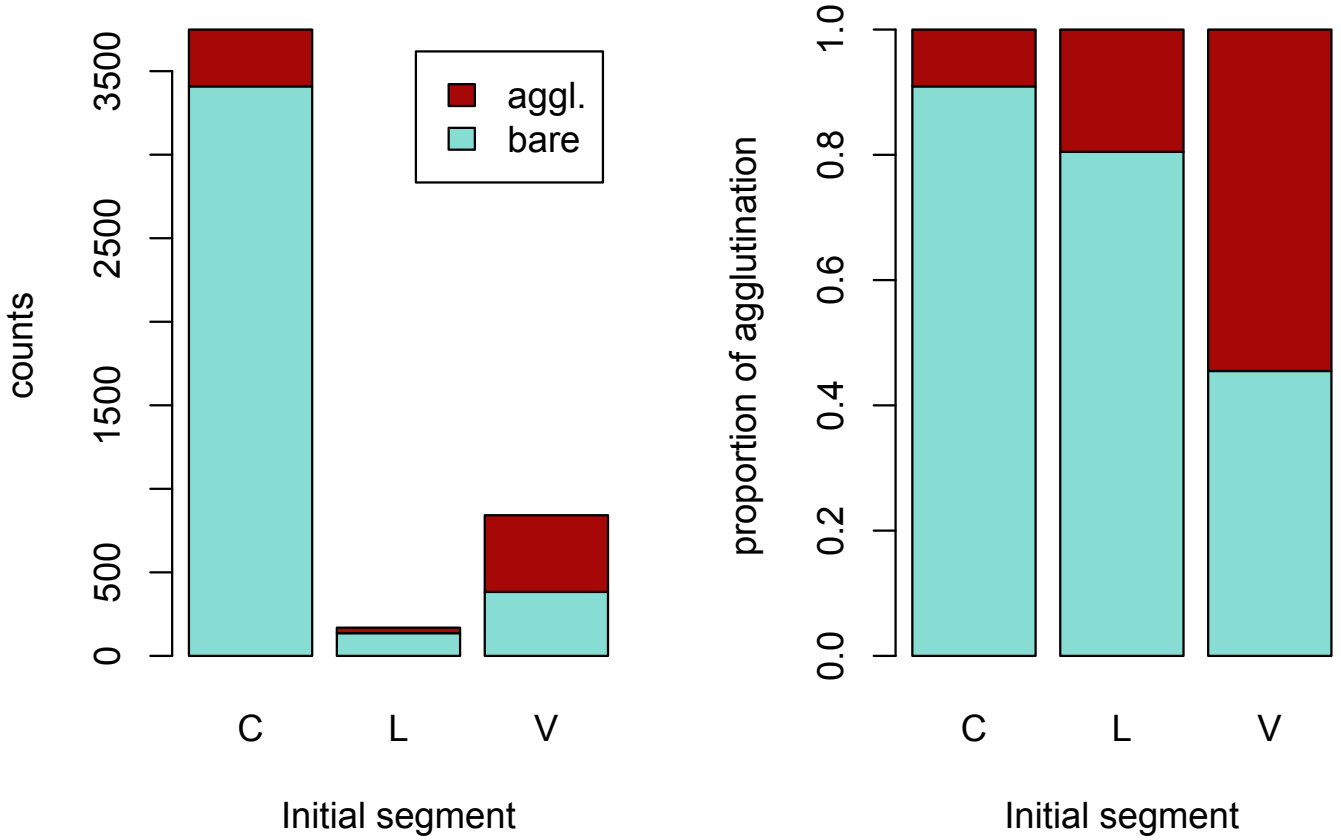


Relevant factors: initial segment type



Irrelevant factors: dissimilation

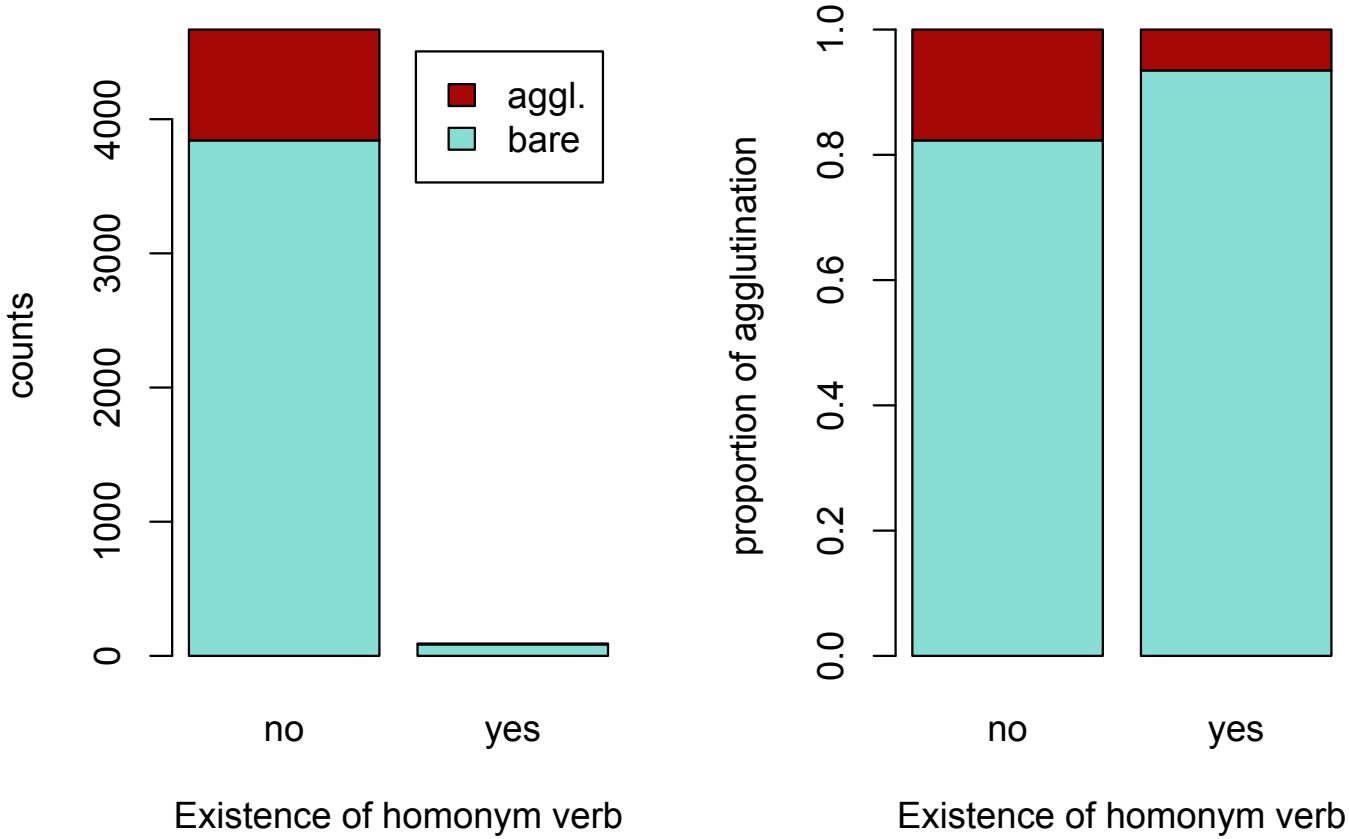
- We expect a dissimilation effect: etymons beginning in *l* should disfavor agglutination.



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

Irrelevant factors: homophony

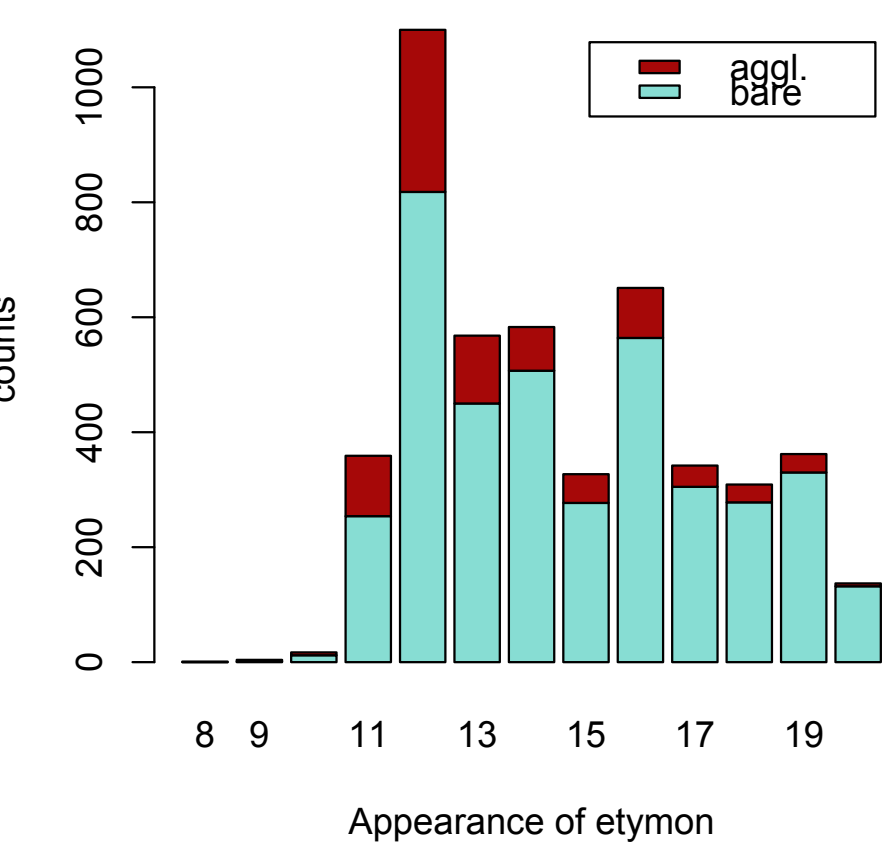
- We expect a homophony avoidance effect: existence of a verb homonym should favor agglutination.

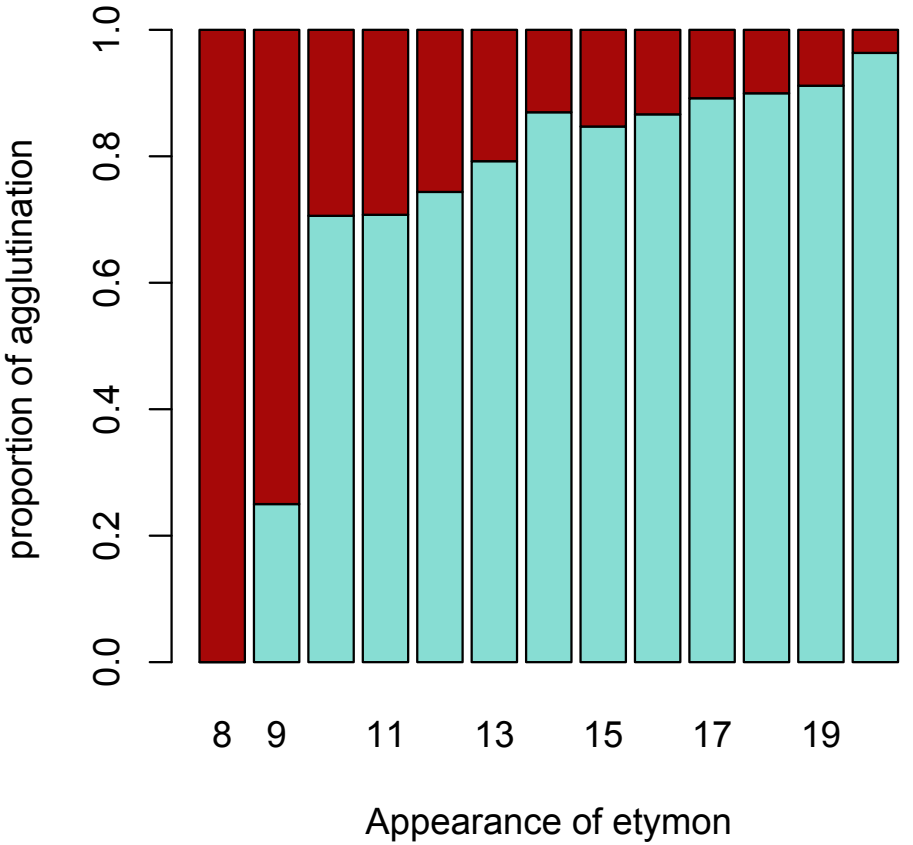


- We do find an effect, but in the opposite direction (χ^2 test $p \approx 0.008$)
- Anyway, not enough relevant data for this to be important.

Relevant factor: age

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

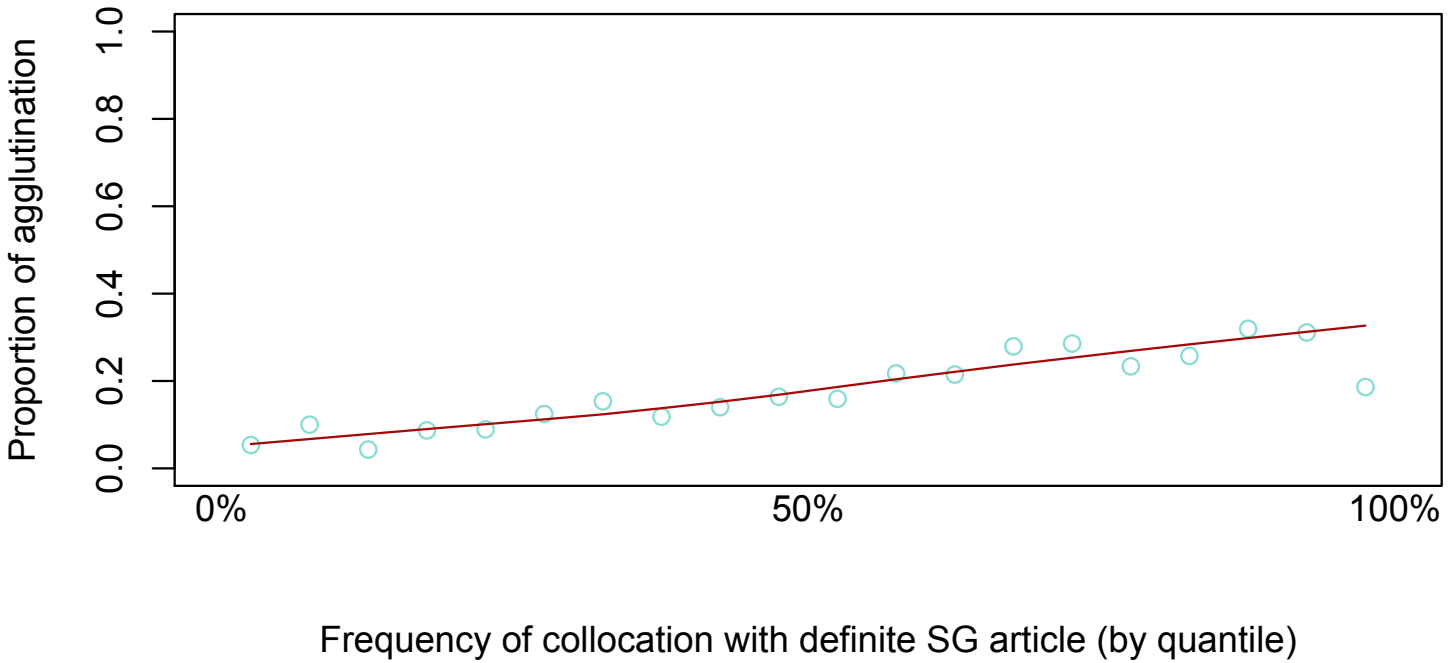




- Words with more recent French etymons agglutinate less. (logistic regression likelihood ratio: χ^2 test $p < 0.0001$)
- Probable combination of factors: frequency, date of Entry in the Mauritian lexicon.

Relevant factor: collocation with SG definite article

- Agglutination grows when collocation with the definite article grows.

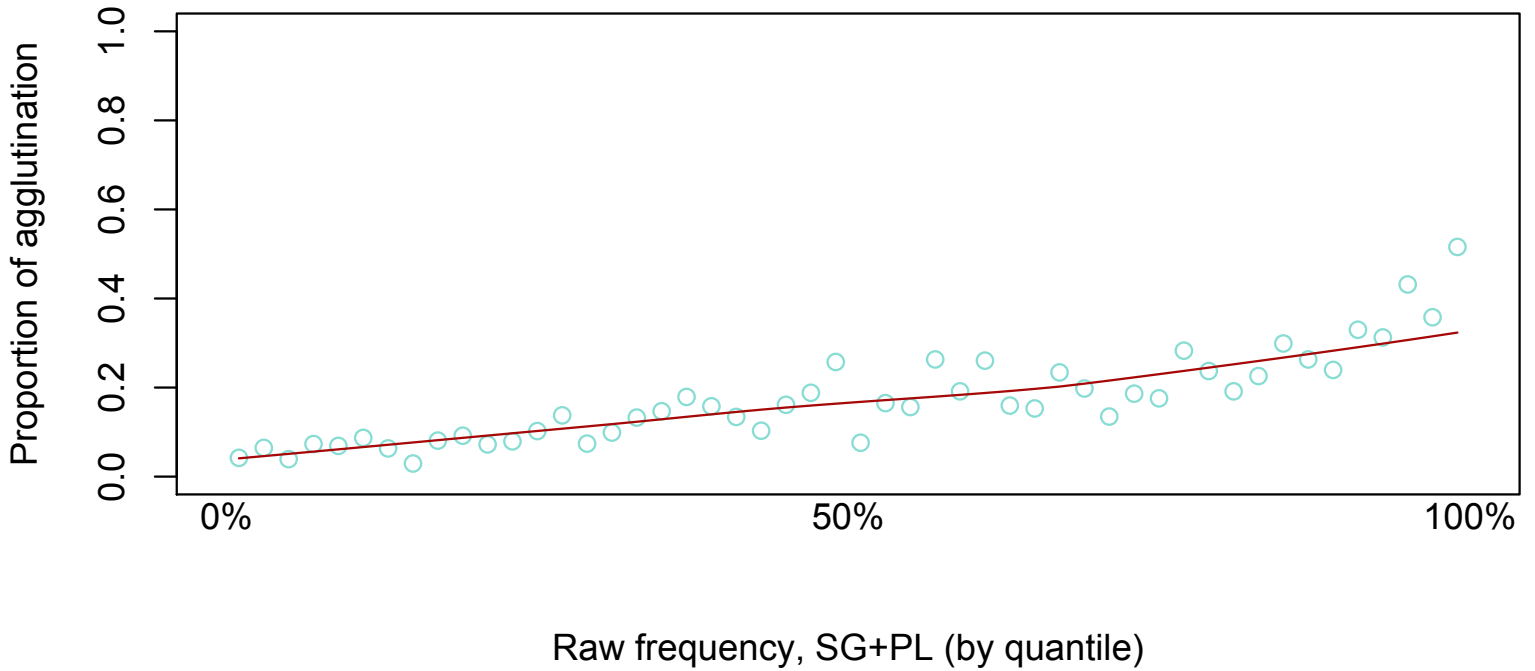


- Highly significant. (logistic regression likelihood ratio: χ^2 test $p < 0.0001$)

☞ 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: χ^2 test $p < 0.0001$)

Summing up

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

Predictor	Accuracy	Acc. increase	$D_{X,y}$
Monosyllabicity or Polysyllabicity	0.8252101	0	0.180
Initial segment type	0.8382353	0.0130252	0.464
Gender	0.8252101	0	0.274
Frequency of coll. with DEF.SG	0.8252101	0	0.274
Raw frequency	0.8254202	0.0002101	0.349
Age	0.8252101	0	0.277
Baseline (no predictor)	0.8252101		

- 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

A multiple logistic regression model

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

$$P(\texttt{agglutination}|\vec{X}) = \frac{e^{\alpha+\vec{\beta}\vec{X}}}{1+e^{\alpha+\vec{\beta}\vec{X}}}$$

Where:

- $\alpha = -4.4573$
- $\beta_1 X_1 = 1.9814 \times \texttt{monosyllabic}$
- $\beta_2 X_2 = 3.8591 \times \texttt{vowel\_initial}$
- $\beta_3 X_3 = 0.6507 \times \log \texttt{frequency}$
- $\beta_4 X_4 = 0.8935 \times \texttt{def.sg\_rel\_frequency}$
- $\beta_5 X_5 = 1.2750 \times \texttt{feminine}$
- $\beta_6 X_6 = -0.3156 \times \texttt{century\_of\_attestation}$

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

Predictors	Accuracy	Acc. increase	$D_{X,y}$
all	0.8798319	0.0546218	0.846
Baseline (no predictor)	0.8252101		

  - ☞ This is an unsubtle measure because it does not make a difference between a probability of 0.51 and a probability of 0.99.

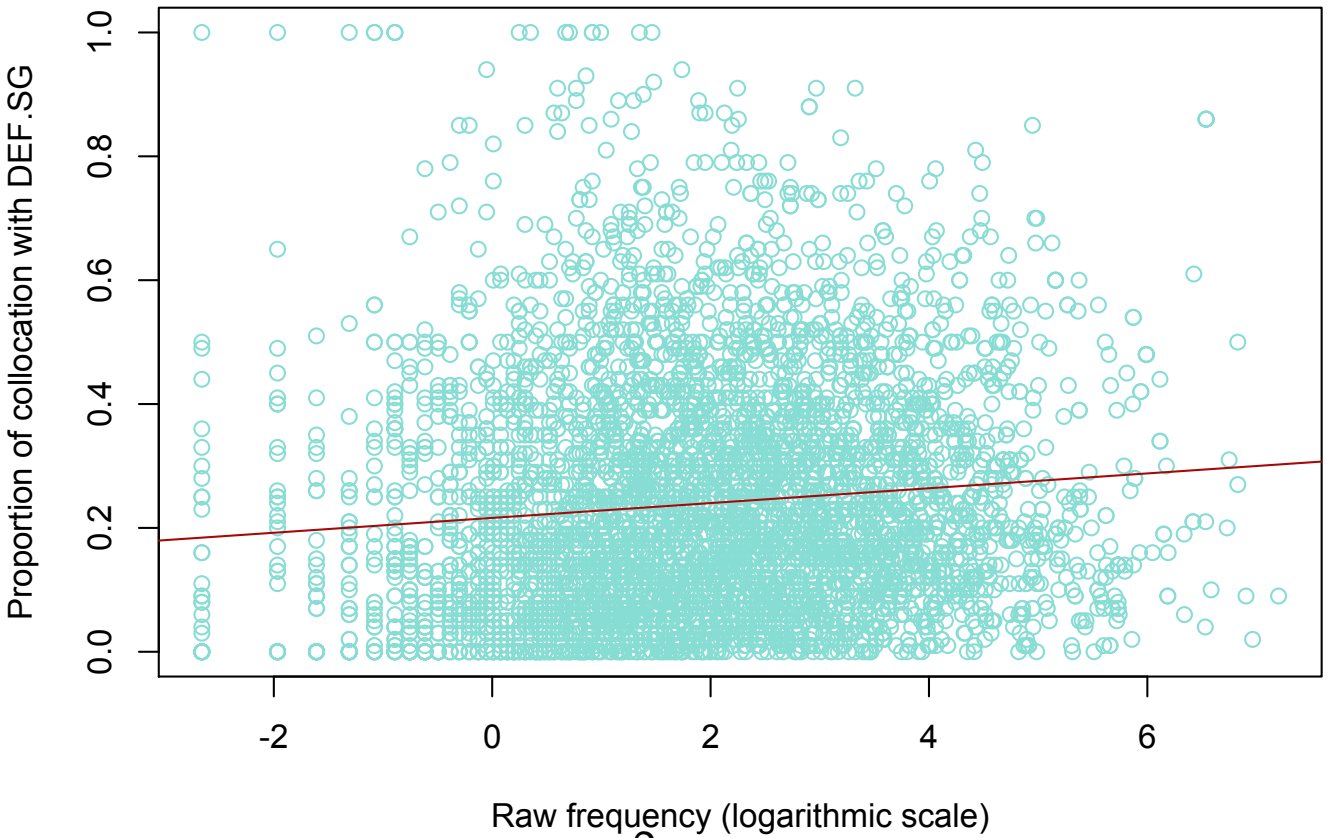
Assessing the model: the gory details

- Good likelihood: χ^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

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



☞ 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

	Older nouns		Newer nouns	
predictors	coefficients	p values	coefficients	p values
`monosyllabic`	2.0083	< 0.0001	0.9618	0.2313
`vowel_initial`	3.8709	< 0.0001	3.5815	< 0.0001
`frequency`	0.6230	< 0.0001	0.6997	0.0030
`def.sg_rel_freq.`	0.8836	< 0.0001	0.9503	< 0.0001
`feminine`	1.3261	< 0.0001	-0.3621	0.4583
`century`	-0.2604	< 0.0001	-0.3835	0.2311
	$D_{x,y} = 0.839$		$D_{x,y} = 0.890$	

Gender and age, continued

- 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 *l*.
  - Full adoption of the creole leads to the disappearance of Bantu influence, and hence to the disappearance of a preference for feminine agglutination.

Outline

Conclusions

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).

Baker, P. (1984). 'Agglutinated french articles in creole french: Their evolutionary significance'. *TeReo*.

Carpooran, A. (2011). *Diksioner Morisien*. Sainte Croix (Mauritius): Koleksion Text Kreol, 2nd Edition.

Grant, A. P. (1995). 'Article agglutination in creole french: a wider perspective'. In P. Baker and C. L. R. Group (eds.), *From contact to Creole and beyond*, Westminster Creolistics series. University of Westminster Press.

McWhorter, J. and Parkvall, M. (2002). 'Pas tout à fait du français, une étude cr/'eole'. *Études Créoles*, XXV-1:179–231.

New, B., Brysbaert, M., Veronis, J., and Pallier, C. (2007). 'The use of film subtitles to estimate word frequencies'. *Applied Psycholinguistics*, 28:661–677.

New, B. and Spinelli, E. (2012). 'Diphones-fr: A french database of diphones positional frequency'. In revision.

Plénat, M. (2000). 'Quelques thèmes de recherche actuels en morphophonologie française'. *Cahiers de lexicologie*, 77:27–62.

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

Strandquist, R. E. (2003). *Article Incorporation in Mauritian Creole, Master Thesis*. Ph.D. thesis, University of Victoria.