

Statistical Modeling of Creole Genesis

MURAWAKI Yūgo
Kyoto University



Multilingual processing?

- Technically yes, but I do not process *text* at all
- Each language is represented as a set of **typological** features

Saramaccan:

0	1	4	...	0	2
---	---	---	-----	---	---

Haitian Creole:

0	0	3	...	1	1
---	---	---	-----	---	---

Feature 81A
Order of SOV

- 0: SOV
- 1: SVO
- 2: VSO
- ⋮

- Not an ordinary NLP task, but *literally* computational linguistics

Why outlier?

- Recent advances in statistical methods come from computational biology
 - Genome sequences → linguistic features
- But I believe that the ACL community can contribute to this field

Why linguistic typology?

- Last resort for Japanese, for which cognate-based methods are not applicable
- Linguistic typology allows us to compare an arbitrary pair of languages

Japanese:

1	1	2	...	0	4
---	---	---	-----	---	---

Korean:

2	2	1	...	0	4
---	---	---	-----	---	---

Ainu:

0	1	1	...	0	4
---	---	---	-----	---	---

Feature 81A
Order of SOV

- 0: SOV
- 1: SVO
- 2: VSO
- ⋮

Why creole genesis?

- I am by no means an expert of creole studies
- Some argue that Japanese was a creole [Kawamoto, 1974,1990][Akiba-Reynolds, 1980]
 - The mixed language hypothesis first proposed by Polivanov (1924)
 - Japanese as a hybrid of Altaic (northern) and Austronesian (southern) elements
- No large dataset was available to test this hypothesis

APiCS enables us to take a data-driven approach to the problem



The screenshot shows the top portion of the APiCS Online website. At the top left is a logo of a palm tree. To its right is the text 'THE ATLAS OF PIDGIN AND CREOLE LANGUAGE STRUCTURES ONLINE'. Further right is a world map with red dots indicating the locations of the languages included in the atlas. Below the header is a dark red navigation bar with links for 'Home', 'Languages', 'Features', 'WALS-APiCS', 'Examples', 'Sources', and 'Authors'. Below that is a light yellow navigation bar with links for 'Credits', 'Legal', 'Download', 'Contact', and 'Help'. The main content area has a heading 'Welcome to APiCS Online' and a featured image of a busy fishing harbor. To the right of the image is a paragraph of text describing the website's content. At the bottom of the main content area is another paragraph of text.

THE ATLAS OF PIDGIN AND CREOLE LANGUAGE STRUCTURES ONLINE

Home Languages Features WALS-APiCS Examples Sources Authors

Credits Legal Download Contact Help

Welcome to *APiCS Online*



Fishing boats: Fishermen selling their catch at Abandze, Ghana, the site of the first British trading station on the Gold Coast, Fort Kormantin,

This web site contains supporting electronic material for the *Atlas of Pidgin and Creole Language Structures (APiCS)*, a publication of Oxford University Press. APiCS shows comparable synchronic data on the grammatical and lexical structures of 76 pidgin and creole languages. The language set contains not only the most widely studied Atlantic and Indian Ocean creoles, but also less well known pidgins and creoles from Africa, South Asia, Southeast Asia, Melanesia and Australia, including some extinct varieties, and several mixed languages.

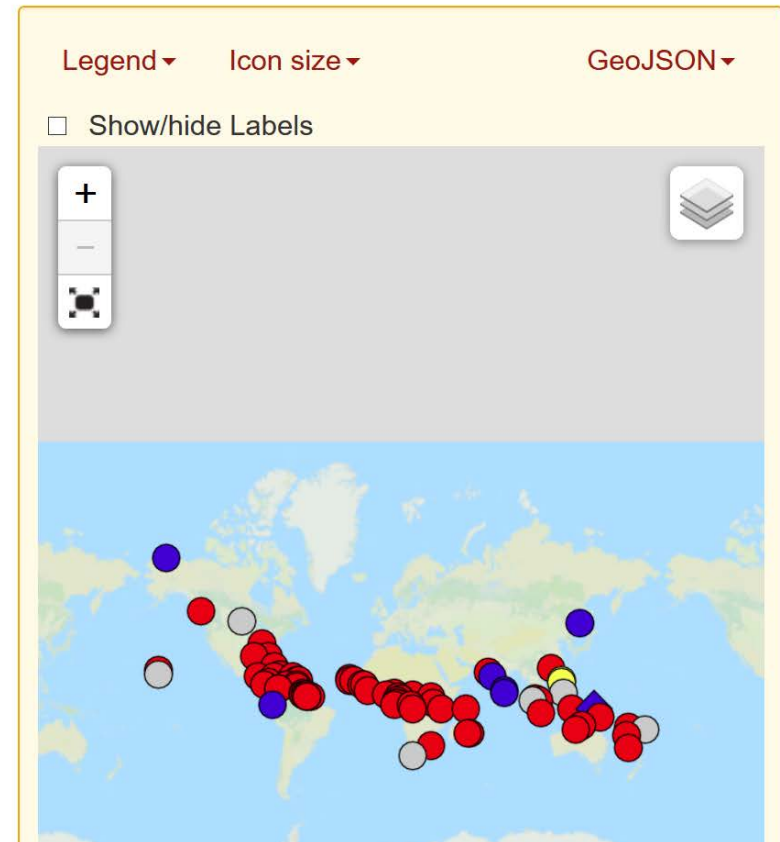
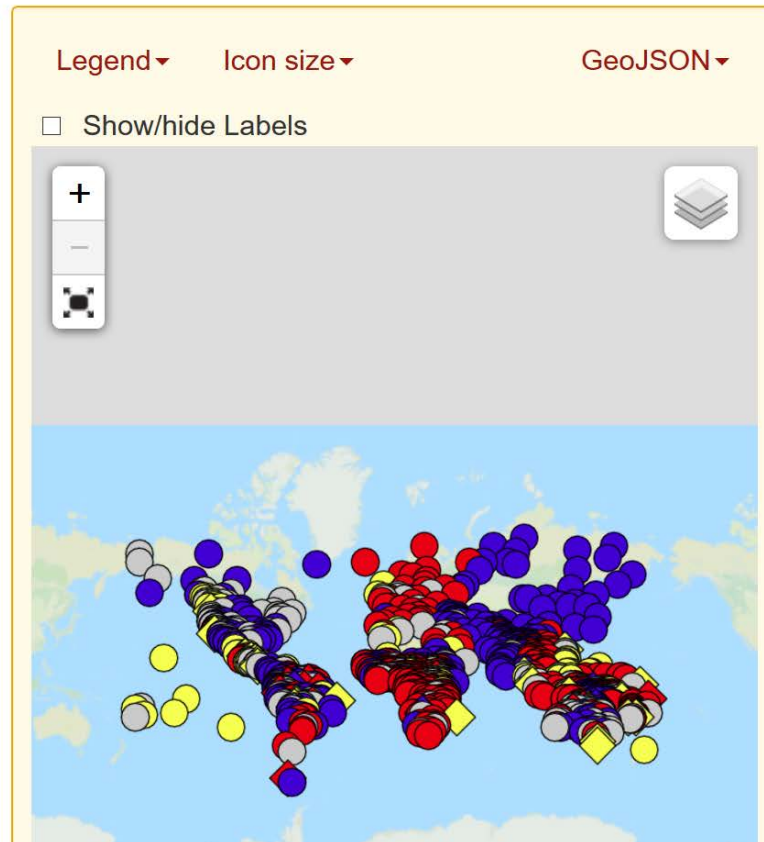
APiCS Online is a separate publication, edited by Susanne Maria Michaelis, Philippe Maurer, Martin Haspelmath, and Magnus Huber. It was made possible by support from the Deutsche

APiCS features are mapped to WALS, the standard typological database

WALS map:  **81A Order of Subject, Object and Verb**

by Matthew S. Dryer

WALS-like APiCS map: **1 Order of subject, object, and verb**



Research questions

Two related research questions found in the literature [Bakker+, 2011][Daval-Markussen+, 2012]

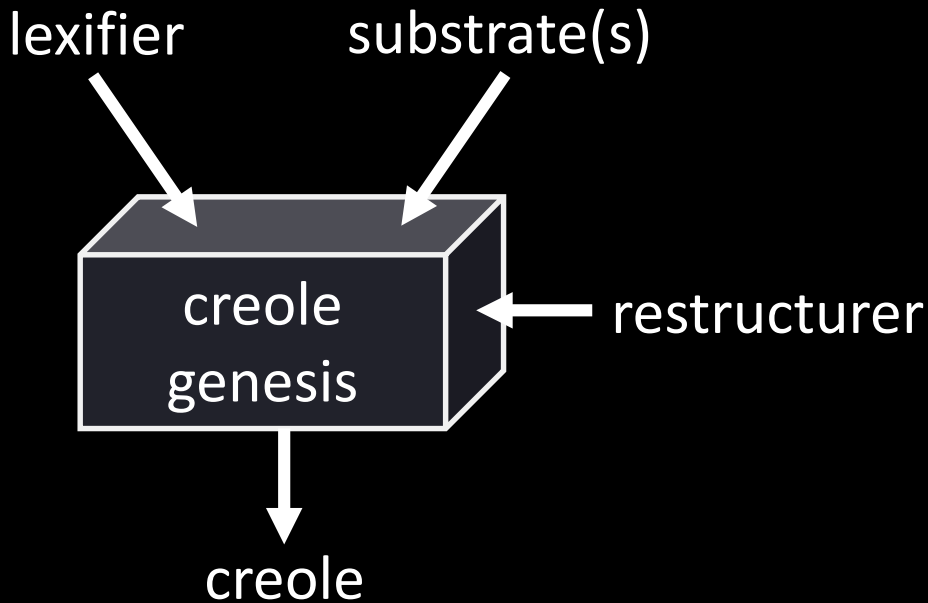
1. Are creoles distinct from non-creoles?

- Casted as a binary classification problem
- The two distributions are very different but nevertheless overlap
- Japanese is far from the creole cluster
- (See paper for details)

2. Which source plays a major role in creole genesis?

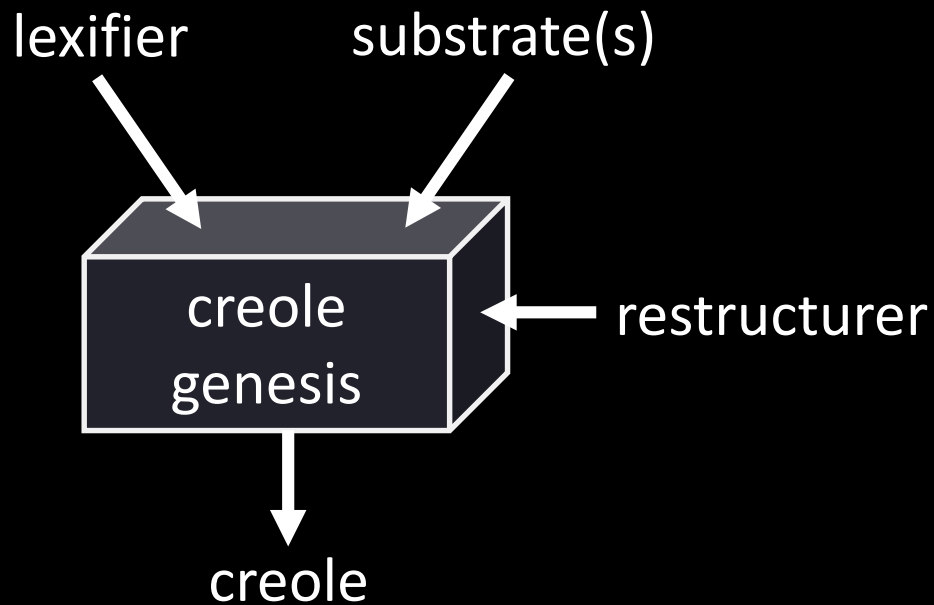
- (Next slide)

Question: Which source plays a major role in creole genesis?



1. *Superstratist*:
Lexifier
2. *Substratist*:
Substrate(s)
3. *Feature pool*:
Both L & S
4. *Universalist*:
Restructuring
universals

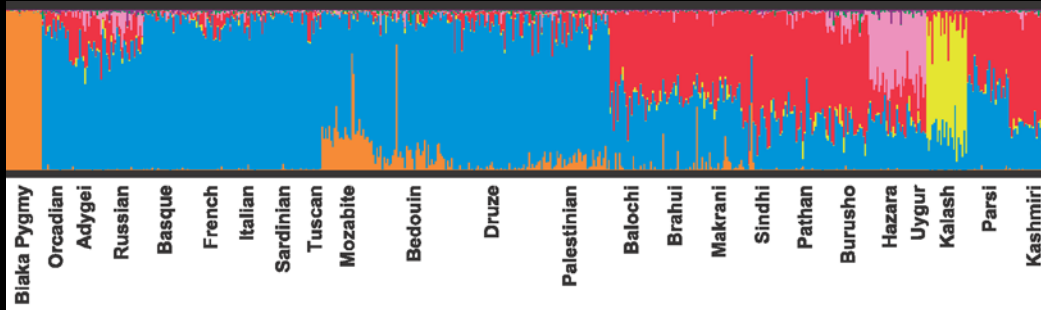
Solution: LDA-like mixture model



- Assumption: each creole is generated by stochastically mixing three types of sources
- Infer mixing proportions from observed data

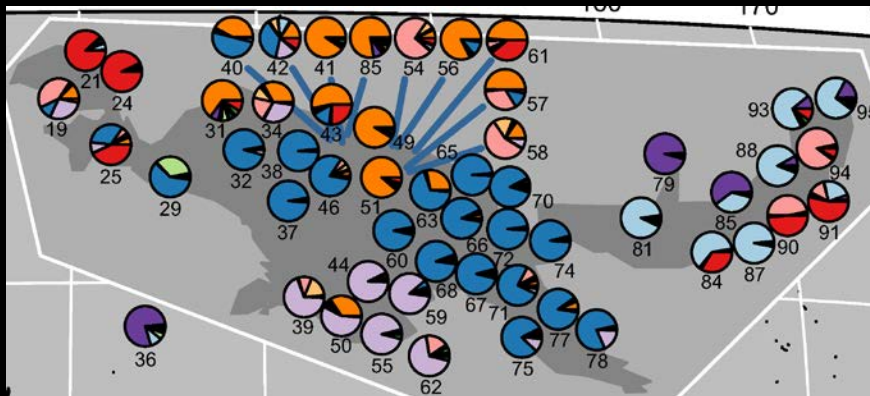
Related work

- Admixture analysis in population genetics [Pritchard+, 2000]
 - Each individual as a mixture of K ancestral components (topics)



[Rosenberg+, 2006]

- An application to linguistic typology
 - Each language as a mixture of K ancestral components (topics)



[Reesink+, 2009]

Mixture model

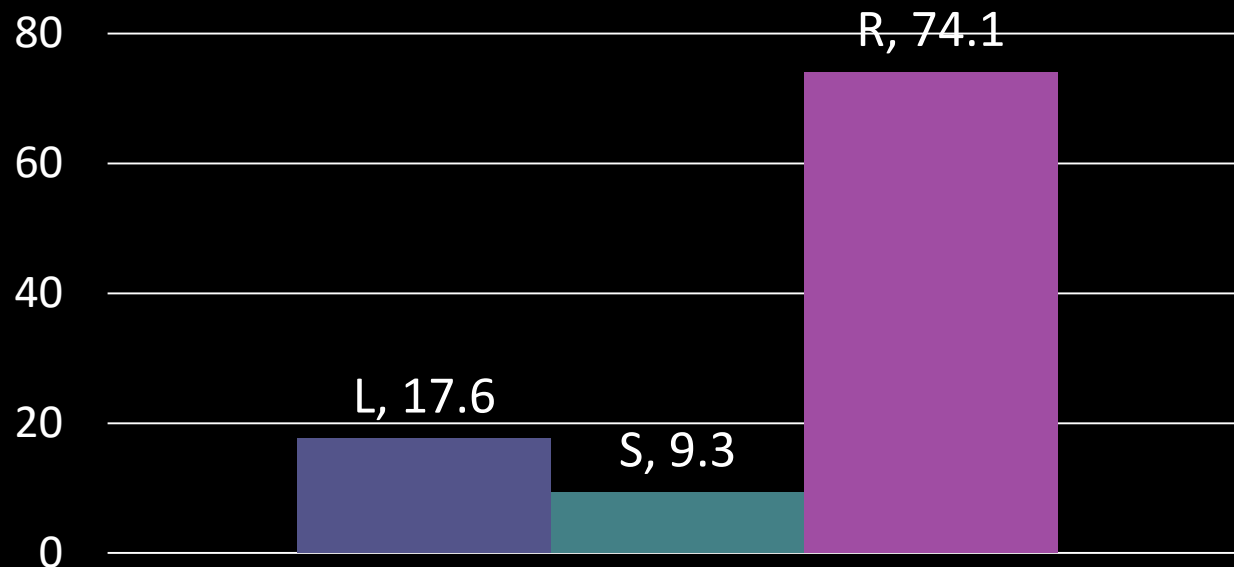
- Assume a creole is a stochastic mixture of L, S, and R

$$\theta_{i,j,k} = \frac{\exp(m_{j,k} + n_{i,k})}{\sum_k \exp(m_{j,k} + n_{i,k})}$$

- $\theta_{i,j,k}$: prob. of deriving creole i 's feature j from source k
 - $m_{j,k}$: per-feature factor (deriving feature j from source k)
 - $n_{i,k}$: per-creole factor (deriving creole i from source k)
- $z_{i,j} \sim \text{Categorical}(\theta_{i,j,L}, \theta_{i,j,S}, \theta_{i,j,R})$
- $x_{i,j} \sim \begin{cases} \delta(y_{i,j,L}) & \text{If } z_{i,j} = L, \text{ copy the value of } i\text{'s lexifier} \\ \delta(y_{i,j,S}) & \text{If } z_{i,j} = S, \text{ copy the value of } i\text{'s substrate} \\ \text{Categorical}(\phi_j) & \text{If } z_{i,j} = R, \text{ draw from } R\text{'s feature distribution} \end{cases}$

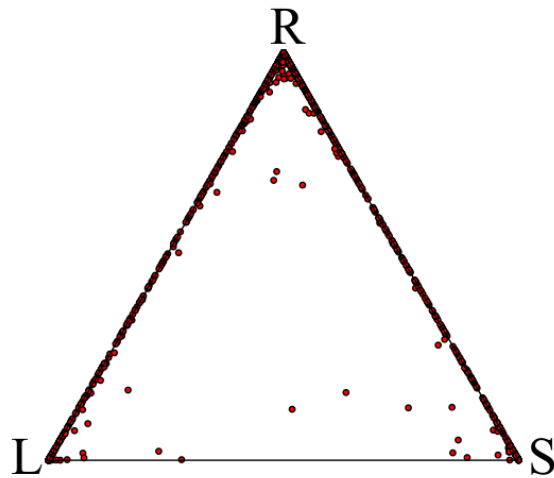
Results: Mixing proportions

- Support the universalist hypothesis
- Negative implications for tree-based phylogenetic inference

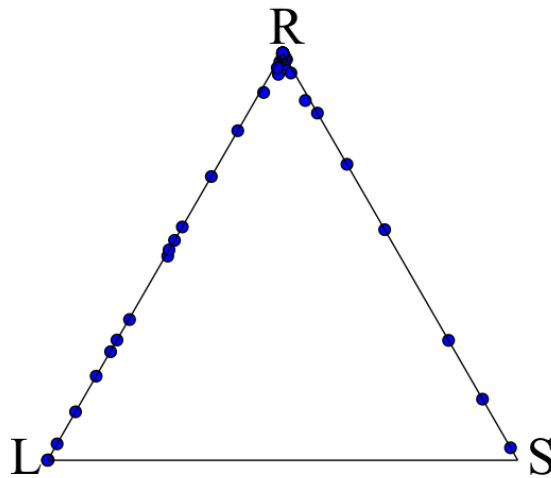


Results: Mixing proportions

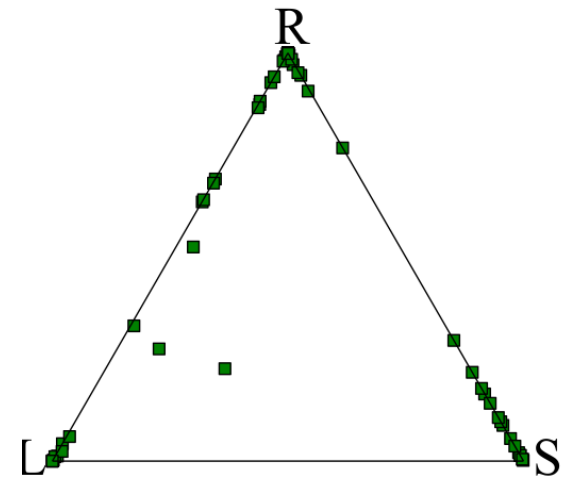
$\theta_{i,j}$



$m_{j,k}$ (per-feature)



$n_{i,k}$ (per-creole)



Results: Top 10 features derived from the restructurer

%	Feature type: value	jpn
91.2%	Numeral Classifiers: Absent	✗
74.3%	Gender Distinctions in Independent Personal Pronouns: No gender distinctions	✗
72.3%	Negative Indefinite Pronouns and Predicate Negation: Predicate negation also present	✓
70.5%	Occurrence of Nominal Plurality: All nouns, always optional	✗
69.7%	Intensifiers and Reflexive Pronouns: Identical	✓
68.4%	Distributive Numerals: No distributive numerals	✗
67.2%	Expression of Pronominal Subjects: Obligatory pronouns in subject position	✗
66.9%	Politeness Distinctions in Pronouns: No politeness distinction	✗
66.6%	Alignment of Case Marking of Pronouns: Nominative - accusative (standard)	✓
66.3%	Order of Numeral and Noun: Numeral-Noun	✓

Conclusion

- Proposed mixture models as a natural choice to analyze creole genesis
- Experiments supported the universalist hypothesis
- Japanese is not like a creole
- Feature work
 - Beyond a proof-of-concept demonstration
 - Data and models for less drastic contact phenomena

