Statistical Modeling of Creole Genesis

Yugo Murawaki
Graduate School of Informatics, Kyoto University
Yoshida-honmachi, Sakyo-ku, Kyoto, 606-8501, Japan
murawaki@i.kyoto-u.ac.jp

Abstract

Creole languages do not fit into the traditional tree model of evolutionary history because multiple languages are involved in their formation. In this paper, we present several statistical models to explore the nature of creole genesis. After reviewing quantitative studies on creole genesis, we first tackle the question of whether creoles are typologically distinct from non-creoles. By formalizing this question as a binary classification problem, we demonstrate that a linear classifier fails to separate creoles from non-creoles although the two groups have substantially different distributions in the feature space. We then model a creole language as a mixture of source languages plus a special *restructurer*. We find a pervasive influence of the restructurer in creole genesis and some statistical universals in it, paving the way for more elaborate statistical models.

1 Introduction

While most linguistic applications of computational phylogeny rely on lexical data (Gray and Atkinson, 2003; Bouckaert et al., 2012), there is a growing trend to make use of typological data (Tsunoda et al., 1995; Dunn et al., 2005; Teh et al., 2008; Longobardi and Guardiano, 2009; Murawaki, 2015). One advantage of typological features over lexical traits (cognates) is that they allow us to compare an arbitrary pair of languages even if they do not share enough cognates. For this reason, they have the potential of uncovering external relations involving language isolates and tiny language families such as Ainu, Basque, and Japanese.

However, our understanding of typological changes is far from satisfactory in at least two respects. First, typological changes are less intuitive than the birth and death of a lexical trait. Modeling word-order change with a single transition matrix (Maurits and Griffiths, 2014), for example, appears to be an oversimplification because some complex mechanisms must be hidden behind the changes (Murawaki, 2015).

The second point, the main focus of this paper, is that it is not clear whether typological data fit into the traditional tree model for a group of languages, which has long been used as the default choice to summarize evolutionary history (Schleicher, 1853). To be precise, regardless of whether typological features are involved, linguists have viewed the tree model with suspicion. A central problem of the tree model is its assumption that after a branching event, two resultant languages evolve completely independently. However, linguists have noted that horizontal contact is a constitutive part of evolutionary history. Various models for contact phenomena have been proposed to address this problem, including the wave theory (Schmidt, 1872) and the gravity model (Trudgill, 1974). As for linguistic typology, areal linguistics has worked on the diffusion of typological features across languages within a geographical area (Campbell, 2006).

In this paper, we study creole languages as an extreme case of non-tree-like evolution (Wardhaugh and Fuller, 2015). A creole is developed as a result of intense contact between multiple languages: typically one socioculturally dominant language (superstrate) and several low-prestige languages (*sub-*

Proceedings of NAACL-HLT 2016, pages 1329–1339,
San Diego, California, June 12-17, 2016.

strates). Superstrates are also known as *lexifiers* because the lexicon of a creole is largely derived from its superstrate. In spite of this, the grammar of a creole is drastically different from that of its lexifier. It is often said that creole grammars are simpler than non-creole ones although it is not easy to measure grammatical simplicity.

Creoles are not irrelevant to historical linguistics because some have speculated about the plausibility of creole status for Middle English (Bailey and Maroldt, 1977) and (pre-)Old Japanese (Kawamoto, 1974; Akiba-Reynolds, 1984; Kawamoto, 1990) while they have been criticized harshly by others.

This controversy can only be settled by fully understanding creole genesis, or the question of how creoles emerge, which also remains unresolved (Wardhaugh and Fuller, 2015). One theory called the gradualist hypothesis suggests their staged development from pidgin languages. A pidgin arises when speakers of different languages with no common language try to have a makeshift conversation, which results in a drastic simplification of grammar called *pidgin formation*. A pidgin is then acquired by children as their first language and is transformed into a fully functional language. This process of grammatical elaboration is known as *creole formation*.

We follow Bakker, Daval-Markussen and colleagues in taking data-driven approaches to this problem (Bakker et al., 2011; Daval-Markussen and Bakker, 2012; Daval-Markussen, 2013). They argue that creoles are typologically distinct from non-creoles. They compare four theories of creole genesis:

1. *Superstratist*. The lexifier plays a major role in creole genesis.

2. *Substratist*. Substrates do, instead of the lexifier.

3. *Feature pool*. Both the lexifier and substrates, but nothing else, provide a feature pool from which each feature value of a creole is selected.

4. *Universalist*. Innate ability of humans to create language is emphasized. The hypothetical linguistic universals for creole formation are called *restructuring universals*.

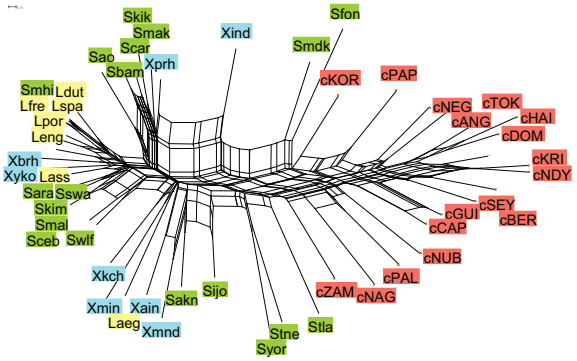
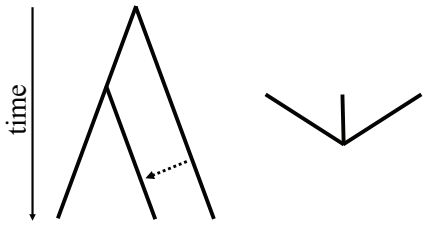


Figure 1: NeighborNet graph of creoles (c), lexifiers (L), substrates (S) and other non-creole languages (X). A language type (c/L/S/X) is followed by a three-letter language code. The bottom-up clustering (i.e., basically tree-building) method produced a cluster of creoles on the right. We reproduced Figure 5 of Daval-Markussen and Bakker (2012) using their online supplementary materials.

To test these explanations, they apply NeighborNet (Bryant and Moulton, 2004), a bottom-up, distance-based clustering method developed in the field of computational biology. By demonstrating that, as in Figure 1, creoles form a cluster distinguishable from lexifiers, substrates and other non-creole languages, they argue for the universalist position and for creole distinctiveness.

However, we find both theoretical and methodological problems in their discussion. Theoretically, the synchronic question of creole distinctiveness is confused with the diachronic question of creole genesis. If creoles are distinct from non-creoles, then something specific to creoles (e.g., restructuring universals) would play a role, but not vice versa. It is logically possible that even with restructuring universals, creoles are indistinguishable from (a subgroup of) non-creoles. Methodologically speaking, NeighborNet does not straightforwardly explain fundamentally non-tree-like evolution because it does assume tree-like evolution even though it represents conflicting signals with reticulations.

We begin by addressing the question of creole (non-)distinctiveness. Whether one is distinct from another is straightforwardly formalized as a binary classification problem. We show that an SVM classifier fails to separate creoles from non-creoles. Fol-



(a) Tree model. (b) Mixture model.

Figure 2: Schematic comparison of two approaches to creole genesis.

lowing a practice in population genetics, we visualize the data using principal component analysis (PCA). The result suggests that although creoles have a substantially different distribution from non-creoles, they nevertheless overlap.

Next, we propose to model creole genesis with mixture models. In this approach, a creole is stochastically generated by mixing its lexifier, substrate(s) plus a special *restructurer*. Conceptually, this is the opposite of the tree model, as illustrated in Figure 2. Specifically, we present two Bayesian models. The first one considers one mixing proportion per creole, and the other decomposes the proportions into per-feature and per-creole factors. Our experimental results suggest that the restructurer dominates creole genesis, dismissing the superstratist, substratist and feature pool theories. We also find some statistical universals in the restructurer although we refrain from identifying them as *restructuring* universals. In this way, we represent a first step toward understanding the complex process of creole genesis through statistical models.

2 Related Work

2.1 Population Genetics

Like Bakker, Daval-Markussen and colleagues (Bakker et al., 2011; Daval-Markussen and Bakker, 2012; Daval-Markussen, 2013), we borrow ideas from computational biology. For reasons unknown to us, they chose clustering models that basically assume tree-like evolution (Saitou and Nei, 1987; Bryant and Moulton, 2004). However, creole genesis is more comparable to models that explicitly take into account genetic admixture (i.e., contact phenomena). See Jones et al. (2015), for example, to take a look at standard practices in population genetics.

Population genetic analysis of genotype data (binary sequences comparable to sets of linguistic features) can be grouped into two types: population-level and individual-level analysis. Populations, such as Sardinian, Yoruba and Japanese, are predefined sets of individuals. Population-level analysis utilizes genetic variation within a population (Patterson et al., 2012). From a modeling point of view, languages are more comparable to individuals. Although a language is spoken by a population, no linguistic data available are comparable to *a set of* individuals.

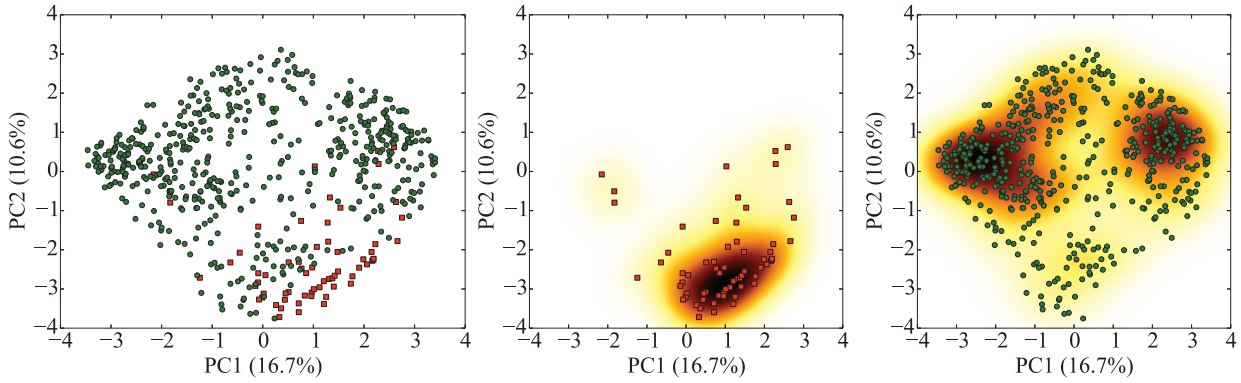
Individual-level analysis, where population labels are used only for the purpose of visualization, is often done using PCA and admixture analysis. PCA is used for dimensionality reduction: by selecting the first two principal components, high-dimensional sequences are projected onto an informative two-dimensional diagram (Patterson et al., 2006).

Admixture analysis (Pritchard et al., 2000; Alexander et al., 2009) closely resembles topic models, most notably Latent Dirichlet Allocation (LDA) (Blei et al., 2003), in NLP. It assumes that each individual is a mixture of K ancestral components (i.e., topics). One difference is that while each LDA topic is associated with a single word distribution (K distributions in total), each SNP (i.e., feature type) has its own distribution ($K \times J$ distributions in total for sequences with length J).

2.2 Linguistic Typology and Non-tree-like Evolution

Like lexical data, typological features are usually analyzed with a tree model, but Reesink et al. (2009) are a notable exception. They applied admixture analysis to Australian and Papuan languages, for which tree-building techniques had not been successful. They related inferred ancestral components to putative prehistoric dispersals and contacts.

Independently of biologically-inspired studies, Daumé III (2009) incorporated linguistic areas into a phylogenetic tree. In his Bayesian generative model, each feature of a language has a latent variable which determines whether it is derived from an areal cluster or the tree. Thus his model can be seen as a mixture model.



(a) Both. (b) Creoles. (c) Non-creoles.

Figure 3: PCA of creoles (red squares) and non-creoles (green circles) with explained variance in the labels of the axes. (a) Scatterplot of both types of languages. (b) Kernel density estimates (KDEs) of creoles. (c) KDEs of non-creoles.

3 Data and Preprocessing

We used the online edition[1] of the Atlas of Pidgin and Creole Language Structures (APiCS) (Michaelis et al., 2013), a database of pidgin and creole languages. It was larger than the datasets of Bakker et al. (2011). As of 2015, it contained 76 languages (104 varieties). It was essentially a pidgin-and-creole version of the online edition[2] of the World Atlas of Language Structures (WALS) (Haspelmath et al., 2005), but it contained sociolinguistic features, phonological inventories and example texts in addition to typological features.

As APiCS did not mark creoles, we used the sociolinguistic feature "Ongoing creolization of pidgins" as a criterion to select creoles. Specifically, we filtered out languages whose feature value was neither "Not applicable (because the language is not a pidgin)" nor "Widespread."

In APiCS, 48 out of 130 typological features were mapped to WALS features. We used these features to combine creoles from APiCS with non-creoles from WALS. Since the WALS database was sparse, we selected languages for which at least 30% of the features were present. As a result, we obtained 64 creoles and 541 non-creoles.

We imputed missing data using the `R` package *missMDA* (Josse et al., 2012). It handled missing values using multiple correspondence analysis (MCA). Specifically, we used the `imputeMCA` function to predict missing feature values.

When investigating creole distinctiveness, we used binary representations of features. Using a one-of-K encoding scheme, we transformed 48 categorical features into 220 binary features.

Our mixture models require each creole to be associated with a lexifier and substrate(s). Unfortunately, APiCS described these languages in an obscure way (and many of them are indeed not fully resolved). We had no choice but to manually select several modern languages as *proxies* for them. For simplicity, we chose only one substrate per creole, but it is not difficult to extend our model for multiple substrates. We are aware that these are oversimplification, but we believe they would be adequate for a proof-of-concept demonstration.

4 Creole Non-distinctiveness

4.1 Binary Classification

To determine whether creoles are distinct from non-creoles, we apply a linear SVM classifier to the typological data. Here, linearity is assumed for two reasons. First, since the supposed distinctiveness is explained by restructuring universals, there is no way for creoles to have an XOR-like distribution. Second, Daval-Markussen (2013) claims that as few as three features are sufficient to distinguish creoles from non-creoles. If this is correct, it is expected that given 48 categorical features, even a simple lin-

[1] `http://apics-online.info/`
[2] `http://wals.info/`

		System	
		C	**NC**
Reference	**C**	54	10
	NC	7	534

Table 1: Confusion matrix of binary classification. C stands for creoles and NC for non-creoles.

ear classifier can work nearly perfectly.

The classifier is trained to classify whether a given language, represented by binarized features, is a creole ($+1$) or non-creole (-1). We use 5-fold cross validation with grid search to tune hyperparameters.

In our experiments, the accuracy, recall, precision and F1-measure were 97.2%, 88.5%, 84.4% and 86.4%, respectively. Table 1 shows the confusion matrix. We can see that the classifier failed to separate creoles from non-creoles. Although the classifier worked well, borderline cases remained.

4.2 PCA

For exploratory analysis and visualization, we applied PCA to creoles and non-creoles, again represented by binarized features. Figure 3 depicts the scatterplot of the first two principal components. We can see that creoles were characterized by quite a different distribution from that of non-creoles. The creoles were concentrated on the lower center while most non-creoles belonged to one of two clusters in the middle. However, the distribution of creoles did overlap with that of non-creoles.

Having a closer look at the diagram, we found that Negerhollands (Dutch), Cape Verdean Creole of Brava (Portuguese) and Vincentian Creole (English) were among the most "typical" creoles (lexifiers in parentheses). Tok Pisin (English) and Bislama (English) were at the periphery of the cluster. The outliers on the upper left included Korlai (Portuguese) while Kikongo-Kituba (Bantu) lay on the upper right.

On the other hand, "creole-like" non-creoles included Chontal Maya (a Mayan language of Mexico), Mussau (an Oceanic language of Papua New Guinea),[3] Catalan and other European languages. The non-creole cluster on the middle left consisted of Japanese, Kannada, Maltese and others. Another non-creole cluster on the middle right included Swahili, Hawaiian and Khmer. The creoleless upper central area was occupied by Lalo (Sino-Tibetan), Maninka (Western) (Mande in West Africa), Salt-Yui (Trans-New Guinea) among others.

[3]Interestingly, Mussau is noted for contact-induced changes (Brownie, 2012).

5 Mixture Models for Creole Genesis

5.1 Basic Idea

Forsaking the quest for synchronic distinctiveness, we take a more direct approach to the diachronic question of creole genesis. Since multiple languages are involved in creole genesis, it is reasonable to apply a mixture model. We assume that a creole is stochastically generated by mixing three sources: (1) a lexifier, (2) a substrate and (3) a global *restructurer*. Under this assumption, the main question is with what proportions these sources are mixed.

An unusual property of our model as a mixture model is that not only outcomes (creoles) but most sources (lexifiers and substrates) are observed. We only need to infer the restructurer. Thus another question is what the restructurer looks like.

Note that our model is constructed such that it does not commit to a particular theory of creole genesis. If the superstratist theory is correct, then lexifiers would dominate the inferred mixing proportions. The same is true of the substratist theory. Similarly, the feature pool theory entails that the restructurer only occupies negligible portions. Also note that even if the restructurer plays a significant role, it does not necessarily imply the universalist position. The restructurer is a set of catch-all feature distributions for those which are explained neither by the lexifier nor by the substrate (that is why we avoid calling it restructuring universals). In order for it to be linguistic universals, it must show some consistent patterns in its distributions.

5.2 MONO Model

Our idea is materialized in two Bayesian generative models. The first one, called MONO, is similar to the STRUCTURE algorithm of admixture analysis (Pritchard et al., 2000).[4]

[4]As seen in Section 2.1, MONO is more similar to STRUCTURE than to LDA in that each feature type has its own distributions. The difference is that while STRUCTURE infers all K global components, MONO always has one global component

Every language in the model is represented by a sequence of categorical features. The number of possible values varies among feature types. For feature j of creole i, the latent assignment variable $z_{i,j}$ determines from which source the feature is derived, a lexifier (L), a substrate (S) or the restructurer (R). Each creole i is associated a priori with a lexifier and a substrate. Let $y_{i,j,\mathrm{L}}$ and $y_{i,j,\mathrm{S}}$ be the values of feature j of creole i's lexifier and substrate, respectively. If the source is the lexifier (or substrate), the creole simply copies $y_{i,j,\mathrm{L}}$ (or $y_{i,j,\mathrm{S}}$). For the sake of uniformity, we can think of a lexifier (or substrate) as a set of feature distributions each of which concentrates all probability mass on its observed value (i.e., the δ function). The remaining source, the restructurer, is a set of categorical feature distributions each of which is drawn from a Dirichlet prior.

The assignment variable $z_{i,j}$ is generated from θ_i, which in turn is generated from a Dirichlet prior. $\theta_i = (\theta_{i,\mathrm{L}}, \theta_{i,\mathrm{S}}, \theta_{i,\mathrm{R}})$ is the parameter of a categorical distribution which specifies the mixing proportion of the three sources for creole i.[5]

More concretely, the generative story of MONO is as follows:

1. For each feature type $j \in \{1, \cdots, J\}$ of the restructurer:
   (a) draw a distribution from a symmetric Dirichlet distribution $\phi_j \sim \mathrm{Dir}(\beta_j)$
2. For each creole $i \in \{1, \cdots, N\}$:
   (a) draw a mixing proportion from a symmetric Dirichlet distribution $\theta_i \sim \mathrm{Dir}(\alpha_i)$
   (b) then for each feature type $j \in \{1, \cdots, J\}$:
      i. draw a topic assignment $z_{i,j} \sim \mathrm{Categorical}(\theta_i)$
      ii. draw a feature value

$$x_{i,j} \sim \begin{cases} \delta(y_{i,j,\mathrm{L}}) & \text{if } z_{i,j} = \mathrm{L} \\ \delta(y_{i,j,\mathrm{S}}) & \text{if } z_{i,j} = \mathrm{S} \\ \mathrm{Categorical}(\phi_j) & \text{if } z_{i,j} = \mathrm{R} \end{cases}$$

As usual, we marginalize out ϕ_j and θ_i using conjugacy of Dirichlet and categorical distributions (Griffiths and Steyvers, 2004). We use Gibbs sampling to infer $z_{i,j}$, whose probability conditioned on the rest is proportional to

$$\begin{cases} \left(\alpha_i + c_{i,\mathrm{L}}^{-(i,j)}\right) \mathrm{I}(x_{i,j} = y_{i,j,\mathrm{L}}) & \text{if } z_{i,j} = \mathrm{L} \\ \left(\alpha_i + c_{i,\mathrm{S}}^{-(i,j)}\right) \mathrm{I}(x_{i,j} = y_{i,j,\mathrm{S}}) & \text{if } z_{i,j} = \mathrm{S} \\ \left(\alpha_i + c_{i,\mathrm{R}}^{-(i,j)}\right) \frac{\beta_j + c_{\mathrm{R},j,x_{i,j}}^{-(i,j)}}{B_j + c_{\mathrm{R},j,*}^{-(i,j)}} & \text{if } z_{i,j} = \mathrm{R} \end{cases} \quad (1)$$

where I is an indicator function, $B_j = \sum \beta_j$, $c_{i,k}^{-(i,j)}$ is the number of assignments for creole i, except $z_{i,j}$, whose values are k, and $c_{\mathrm{R},j,l}^{-(i,j)}$ is the number of observed features for feature type j, except $x_{i,j}$, that is derived from the restructurer and has l as its value. Intuitively, the first term gives priority to the source from which many other features of creole i are derived. The second term concerns how likely the source generates the feature value. For the lexifier or the substrate, it is 1 only if the source shares the same feature value with the creole; otherwise 0. To tune hyperparameters α_i and β_j, we set a vague gamma prior $\mathrm{Gamma}(1, 1)$ and sample these parameters using slice sampling (Neal, 2003).

5.3 FACT Model

It is said that some features are more easily borrowed than others (Matras, 2011). For creoles, some seems to reflect substrate influence on phonology while reduced inflections might be attributed to the restructurer. Inspired by these observations, we extend the MONO model such that some feature types can have strong connections to particular sources. We call this extended model FACT.

To do this, we decomposes the mixing proportions into per-feature and per-creole factors. We apply additive operations to these factors in log-space in a way similar to the Sparse Additive Generative model (Eisenstein et al., 2011). As a result of this extension, every feature j of creole i has its own mixing proportion, $\theta_{i,j} = (\theta_{i,j,\mathrm{L}}, \theta_{i,j,\mathrm{S}}, \theta_{i,j,\mathrm{R}})$:

$$\theta_{i,j,k} = \frac{\exp(m_{j,k} + n_{i,k})}{\sum_k \exp(m_{j,k} + n_{i,k})}, \quad (2)$$

where $m_{j,k}$ is a factor specific to feature type j and $n_{i,k}$ is the one specific to creole i. To penalize extreme values, we put Laplacian priors on $m_{j,k}$ and $n_{i,k}$, with mean 0 and scale γ.

To sum up, the generative story of FACT is as follows:

and two local, observed components.

[5] By letting another categorical distribution subdivide $\theta_{i,\mathrm{S}}$, we can incorporate multiple substrates into the model.

Model		Sources		
		L	S	R
MONO		16.6%	9.3%	74.1%
FACT	Combined	17.6%	6.0%	76.4%
	Per-feature	22.5%	6.8%	70.8%
	Per-creole	25.0%	20.4%	54.6%

Table 2: Summary of mixing proportions. The arithmetic mean of 50 samples after 5,000 iterations, with an interval of 100 iterations.

1. For each feature type $j \in \{1, \cdots, J\}$:
   (a) draw $\phi_j \sim \mathrm{Dir}(\beta_j)$
   (b) for each source $k \in \{\mathrm{L}, \mathrm{S}, \mathrm{R}\}$:
      i. draw $m_{j,k} \sim \mathrm{Laplace}(0, \gamma)$
2. For each creole $i \in \{1, \cdots, N\}$:
   (a) for each source $k \in \{\mathrm{L}, \mathrm{S}, \mathrm{R}\}$:
      i. draw $n_{i,k} \sim \mathrm{Laplace}(0, \gamma)$
   (b) then for each feature $j \in \{1, \cdots, J\}$:
      i. normalize $m_{j,k}$ and $n_{i,k}$ to obtain $\theta_{i,j}$ (Equation (2))
      ii. draw a topic assignment $z_{i,j} \sim \mathrm{Categorical}(\theta_{i,j})$
      iii. draw $x_{i,j}$ as in MONO

ϕ_j is integrated out as before, but the conjugacy no longer holds for $\theta_{i,j}$.

For inference, a modification is needed to infer $z_{i,j}$: the first term $\alpha_i + c_{i,k}^{-(i,j)}$ of Equation (1) is replaced with $\theta_{i,j,k}$. $m_{j,k}$ and $n_{i,k}$ are sampled using the Metropolis algorithm, with a Gaussian proposal distribution centered at the previous value. Hyperparameter γ is set to 10.

5.4 Results

Table 2 summarizes mixing proportions. For MONO and FACT (combined), we use a fraction of assignment variables pointing to a particular souce. Per-feature and per-creole factors are converted into probabilities as follows: per-feature proportions $\tilde{\phi}_j = (\tilde{\phi}_{j,\mathrm{L}}, \tilde{\phi}_{j,\mathrm{S}}, \tilde{\phi}_{j,\mathrm{R}})$, where $\tilde{\phi}_{j,k} = \frac{\exp(m_{j,k})}{\sum_k \exp(m_{j,k})}$. Similarly, per-creole proportions $\tilde{\theta}_i = (\tilde{\theta}_{j,\mathrm{L}}, \tilde{\theta}_{j,\mathrm{S}}, \tilde{\theta}_{j,\mathrm{R}})$, where $\tilde{\theta}_{i,k} = \frac{\exp(n_{i,k})}{\sum_k \exp(n_{i,k})}$.

We can see that the overwhelming majority of features were derived from the restructurer both in MONO and FACT (combined). The restructurer was followed by lexifiers, and substrates were the least influential.[6] These results can be interpreted as counter-evidence to the superstratist, substratist and feature pool theories.

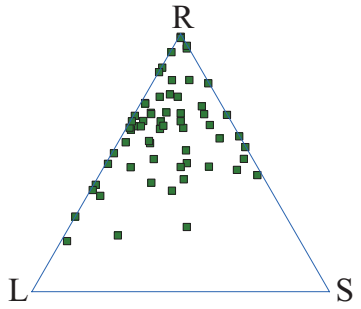


Figure 4: Mixing proportions of MONO projected onto a simplex. Each point denotes a creole. It is the parameter of the posterior predictive distribution of an assignment variable: $\tilde{\theta}_i = (\frac{\alpha_i + c_{i,\mathrm{L}}}{Z}, \frac{\alpha_i + c_{i,\mathrm{S}}}{Z}, \frac{\alpha_i + c_{i,\mathrm{R}}}{Z})$, where the normalizer $Z = \sum_k \alpha_i + c_{i,k}$. One sample after 10,000 iterations.

MONO and FACT (combined) exhibited similar patterns. When the mixing proportions are decomposed into per-feature and per-creole factors, per-creole factors exhibited less uneven distributions than per-feature factors. This implies heterogeneous behavior of features in creole genesis. Table 3 lists top-5 feature types for each source.

Figure 4 plots creoles on a simplex of mixing proportions in MONO. Creoles scattered across the simplex but leaned toward the restructurer. This implies that a lexifier cannot be mixed with substrates without interference from the restructurer.

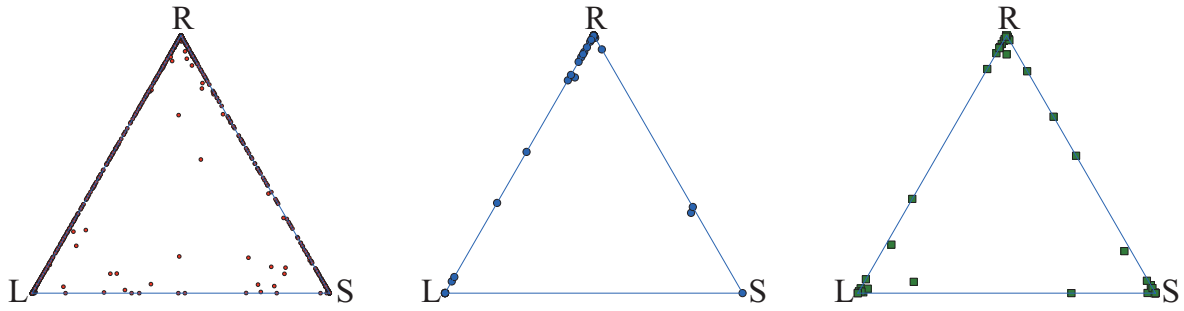
Compared with MONO, FACT tended to push points to the edges of the simplex. This can be confirmed in Figure 5. In particular, Figure 5(c) is directly comparable to Figure 4. It is possible that halfway points in MONO were artifacts of its limited expressive power.

Table 4 lists the top-10 feature type-value pairs that were derived from the restructurer. In other words, we stochastically removed the influence of the lexifiers and substrates from creole data. These features can be regarded as (statistical) universals

[6] The substrates would probably occupy a larger portion if multiple substrates are incorporated in future work.

Source	Ratio	Feature type
Lexifier	100.0%	Order of Adposition and Noun Phrase
	100.0%	Order of Relative Clause and Noun
	99.9%	Applicative Constructions
	99.5%	The Prohibitive
	98.2%	Alignment of Case Marking of Full Noun Phrases
Substrate	84.9%	Order of Genitive and Noun
	56.6%	Tone
	54.9%	Order of Subject, Object and Verb
	26.5%	Pronominal and Adnominal Demonstratives
	24.0%	Relativization on Subjects
Restructurer	100.0%	Intensifiers and Reflexive Pronouns
	100.0%	Numeral Classifiers
	100.0%	Suppletion According to Tense and Aspect
	100.0%	Expression of Pronominal Subjects
	100.0%	Polar Questions

Table 3: Top-5 feature types for each source according to per-feature factors of FACT. The arithmetic mean of 50 samples after 5,000 iterations, with an interval of 100 iterations.



(a) Combined. (b) Per-feature factors. (c) Per-creole factors.

Figure 5: Mixing proportions of FACT projected onto a simplex. One sample after 10,000 iterations. (a) $J \times N$ points for combined mixing proportions $\theta_{i,j}$. (b) J points for per-feature factors $\tilde{\phi}_j$ as in Table 2. (c) N points for per-creole factors $\tilde{\theta}_i$.

although our model leaves the possibility that they were not *restructuring* universals. To answer this question, we need to break down the restructurer by types of linguistic universals.

Among the 10 feature type-value pairs, only four apply to Japanese (Negative Indefinite Pronouns and Predicate Negation, Intensifiers and Reflexive Pronouns, Alignment of Case Marking of Pronouns, and Order of Numeral and Noun). For reference, English has seven. Combined with the PCA analysis in Section 4.2, this suggests that Japanese is a very non-creole-like language. However, we are unsure if the possibility of creole status for (pre-)Old Japanese is completely rejected. This question might be answered if we figure out how long it takes to make creole-like traits disappeared.

It is often said that creoles have SVO word order. According to APiCS, the number of creoles with SVO order was 61 (exclusive) and 71 (exclusive plus shared) in the 76 language dataset. However, this feature value only gained the ratio of 67.3%. This is mainly because SVO is the word order of most lexifiers, but it can also be attributed to data representation: since WALS did not allow multi-valued features (e.g., SVO *and* SOV), some creoles with multiple word orders were mapped to a separate category "No dominant order," underestimating the influence of SVO.

Ratio	Feature type	Feature value
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

Table 4: Top-10 features derived from the restructurer in FACT. The ratio of the feature type-value pair (j, l) is defined as $|\{(i \mid x_{i,j} = l, z_{i,j} = \mathrm{R}\}| \ / \ N$. The arithmetic mean of 50 samples after 5,000 iterations, with an interval of 100 iterations.

5.5 Discussion

The main contribution of our work is the introduction of mixture models to creole studies. This is, however, only the first step toward understanding the complex process of creole genesis by means of statistical modeling. Better data are needed with respect to proxies for substrates, missing values, multi-valued features among others.

With better data, more elaborate models could uncover the detailed process of creole genesis. Our models mix several sources in one step, but we may want to model the staged development of pidgin formation and creole formation. As a result of continued influence from its superstrate, a creole might undergo *decreolization*. It is argued that pidgins themselves have several development stages, from each of which creoles can emerge (Mühlhäusler, 1997). Hopefully, these hypotheses could be tested with statistical models.

Our finding that the restructurer plays a dominant role in creole genesis has a negative implication for tree-based inference of language relationships. If most features of a language come from nowhere, we are unable to trace its origin back into the deep past. In the meanwhile, it has been argued that creole genesis only occurred in modern and early-modern, exceptional circumstances and cannot be responsible for most historical changes. Thus identifying the social conditions under which creoles arise (Tria et al., 2015) is another research direction to be explored.

6 Conclusion

In this paper, we present several statistical models of linguistic typology to answer questions concerning creole genesis. First, we formalized creole (non-)distinctiveness as a binary classification problem. Second, we propose to model creole genesis with mixture models, which makes more sense than tree-building techniques.

Recent studies on linguistic applications of computational phylogeny have been heavily influenced from computational biology. They often depend on ready-to-use software packages developed in that field. We observe that, as a result, linguistic phenomena that lack exact counterparts in biology tend to be left untouched. In this paper, we have hopefully demonstrated that computational linguists could fill the gap.

Acknowledgment

This work was partly supported by JSPS KAKENHI Grant Number 26730122.

References

Katsue Akiba-Reynolds. 1984. Internal reconstruction in pre-Japanese syntax. In Jacek Fisiak, editor, *Historical Syntax*, pages 1–23. Walter de Gruyter.

David H. Alexander, John Novembre, and Kenneth Lange. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19(9):1655–1664.

Charles J. Bailey and Karl Maroldt. 1977. The French lineage of English. In Jürgen M. Meisel, editor, *Langues en contact – Pidgins – Creoles*, pages 21–53. Narr.

Peter Bakker, Aymeric Daval-Markussen, Mikael Parkvall, and Ingo Plag. 2011. Creoles are typologically distinct from non-creoles. *Journal of Pidgin and Creole Languages*, 26(1):5–42.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Remco Bouckaert, Philippe Lemey, Michael Dunn, Simon J. Greenhill, Alexander V. Alekseyenko, Alexei J. Drummond, Russell D. Gray, Marc A. Suchard, and Quentin D. Atkinson. 2012. Mapping the origins and expansion of the Indo-European language family. *Science*, 337(6097):957–960.

John Brownie. 2012. Multilingualism and identity on Mussau. *International Journal of the Sociology of Language*, 2012(214).

David Bryant and Vincent Moulton. 2004. Neighbor-Net: An agglomerative method for the construction of phylogenetic networks. *Molecular Biology and Evolution*, 21(2):255–265.

Lyle Campbell. 2006. Areal linguistics. In *Encyclopedia of Language and Linguistics, Second Edition*, pages 454–460. Elsevier.

Hal Daumé III. 2009. Non-parametric Bayesian areal linguistics. In *HLT-NAACL*, pages 593–601.

Aymeric Daval-Markussen and Peter Bakker. 2012. Explorations in creole research with phylogenetic tools. In *Proc. of LINGVIS & UNCLH*, pages 89–97.

Aymeric Daval-Markussen. 2013. First steps towards a typological profile of creoles. *Acta Linguistica Hafniensia*, 45(2):274–295.

Michael Dunn, Angela Terrill, Ger Reesink, Robert A. Foley, and Stephen C. Levinson. 2005. Structural phylogenetics and the reconstruction of ancient language history. *Science*, 309(5743):2072–2075.

Jacob Eisenstein, Amr Ahmed, and Eric P. Xing. 2011. Sparse additive generative models of text. In *Proc. of ICML*, pages 1041–1048.

Russell D. Gray and Quentin D. Atkinson. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature*, 426(6965):435–439.

Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. *PNAS*, 101:5228–5235.

Martin Haspelmath, Matthew Dryer, David Gil, and Bernard Comrie, editors. 2005. *The World Atlas of Language Structures*. Oxford University Press.

Eppie R. Jones, Gloria Gonzalez-Fortes, Sarah Connell, Veronika Siska, Anders Eriksson, Rui Martiniano, Russell L. McLaughlin, Marcos Gallego Llorente, Lara M. Cassidy, Cristina Gamba, Tengiz Meshveliani, Ofer Bar-Yosef, Werner Muller, Anna Belfer-Cohen, Zinovi Matskevich, Nino Jakeli, Thomas F. G. Higham, Mathias Currat, David Lordkipanidze, Michael Hofreiter, Andrea Manica, Ron Pinhasi, and Daniel G. Bradley. 2015. Upper Palaeolithic genomes reveal deep roots of modern Eurasians. *Nature Communications*, 6.

Julie Josse, Marie Chavent, Benot Liquet, and François Husson. 2012. Handling missing values with regularized iterative multiple correspondence analysis. *Journal of Classification*, 29(1):91–116.

Takao Kawamoto. 1974. Agreements and disagreements in morphology between Japanese and Austronesian (chiefly Melanesian) languages. *The Japanese Journal of Ethnology*, 39(2):113–129. (in Japanese).

Takao Kawamoto. 1990. Pijin kureōru-ka to Nihongo no keisei [Pidginization-creolization and the formation of Japanese]. In Osamu Sakiyama, editor, *Nihongo no Keisei [Formation of Japanese]*, pages 130–168. Sanseido. (in Japanese).

Giuseppe Longobardi and Cristina Guardiano. 2009. Evidence for syntax as a signal of historical relatedness. *Lingua*, 119(11):1679–1706.

Yaron Matras. 2011. Universals of structural borrowing. In Peter Siemund, editor, *Linguistic Universals and Language Variation*, pages 204–233. Walter de Gruyter.

Luke Maurits and Thomas L. Griffiths. 2014. Tracing the roots of syntax with Bayesian phylogenetics. *PNAS*, 111(37):13576–13581.

Susanne Maria Michaelis, Philippe Maurer, Martin Haspelmath, and Magnus Huber, editors. 2013. *APiCS Online*. Max Planck Institute for Evolutionary Anthropology.

Peter Mühlhäusler. 1997. *Pidgin and Creole Linguistics: Expanded and revised Edition*. University of Westminster Press.

Yugo Murawaki. 2015. Continuous space representations of linguistic typology and their application to phylogenetic inference. In *Proc. of NAACL-HLT*, pages 324–334.

Radford M. Neal. 2003. Slice sampling. *Annals of Statistics*, 31(3):705–767.

Nick Patterson, Alkes L. Price, and David Reich. 2006. Population structure and eigenanalysis. *PLoS Genetics*, 2(12):e190, 12.

Nick Patterson, Priya Moorjani, Yontao Luo, Swapan Mallick, Nadin Rohland, Yiping Zhan, Teri Genschoreck, Teresa Webster, and David Reich. 2012. Ancient admixture in human history. *Genetics*, 192(3):1065–1093.

Jonathan K. Pritchard, Matthew Stephens, and Peter Donnelly. 2000. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959.

Ger Reesink, Ruth Singer, and Michael Dunn. 2009. Explaining the linguistic diversity of Sahul using population models. *PLoS Biology*, 7(11).

Naruya Saitou and Masatoshi Nei. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425.

August Schleicher. 1853. Die ersten Spaltungen des indogermanischen Urvolkes. *Allgemeine Monatsschrift für Wissenschaft und Literatur*, 3:786–787. (in German).

Johannes Schmidt. 1872. *Die Verwandtschaftsverhältnisse der indogermanischen Sprachen*. Hermann Böhlau. (in German).

Yee Whye Teh, Hal Daumé III, and Daniel Roy. 2008. Bayesian agglomerative clustering with coalescents. In *NIPS*, pages 1473–1480.

Francesca Tria, Vito D.P. Servedio, Salikoko S. Mufwene, and Vittorio Loreto. 2015. Modeling the emergence of contact languages. *PLoS ONE*, 10(4):e0120771, 04.

Peter Trudgill. 1974. Linguistic change and diffusion: Description and explanation in sociolinguistic dialect geography. *Language in Society*, 3:215–246.

Tasaku Tsunoda, Sumie Ueda, and Yoshiaki Itoh. 1995. Adpositions in word-order typology. *Linguistics*, 33(4):741–762.

Ronald Wardhaugh and Janet M. Fuller. 2015. *An Introduction to Sociolinguistics, 7th Edition*. John Wiley & Sons.