Title
A relaxed admixture model of contact

Journal
Language Dynamics and Change, 4(1)

Author
Michael, Lev David

Publication Date
2014

Peer reviewed

A relaxed-admixture model of language contact

Will Chang and Lev Michael
University of California, Berkeley

Under conditions of language contact, a language may gain features from its neighbors that it is unlikely to have gotten endogenously. We describe a method for evaluating pairs of languages for potential contact by comparing a null hypothesis in which a target language obtained all its features by inheritance, with an alternative hypothesis in which the target language obtained its features via inheritance *and* via contact with a proposed donor language. Under the alternative hypothesis the donor may influence the target to gain features, but not to lose features. When applied to a database of phonological characters in South American languages, this method proves useful for detecting the effects of relatively mild and recent contact, and for highlighting several potential linguistic areas in South America.

Keywords: probabilistic generative model; language contact; linguistic areality; Upper Xingú; South America; phonological inventory.

1. Introduction

Tukano	pʰ	p	b	tʰ	t	d			kʰ	k	g	ʔ					
Tariana	pʰ	p	b	tʰ	t	d	dʰ	tʃ	kʰ	k							
Arawak	4	38	17	8	42	14	0	30	4	41	10	16					
Tukano							s	h	w		j	ɾ					
Tariana	m	mʰ	n	nʰ	ɲ	ɲʰ	s	h	w	wʰ	j	ɾ	l				
Arawak	41	0	42	0	24	0	33	37	31	0	39	36	14				
Tukano	i	ĩ		e	ẽ		a	ã		o	õ	u	ũ		ɨ	ɨ̃	
Tariana	i	ĩ	iː	e	ẽ	eː	a	ã	aː	o	õ	u	ũ	uː	ɨ		
Arawak	42	7	22	38	7	20	42	7	22	28	5	26	5	10	13	4	

Table 1: The phonemes of Tukano and Tariana; how often each occurs in 42 Arawak languages, not including Tariana.

Tukano is a Tukanoan language spoken in northwest Amazonia. Tariana, a linguistic neighbor, is an Arawak language. Did Tariana gain phonemes as a result of contact with Tukano? Table 1 shows the phonemes of both languages, along with counts of how often each occurs in 42 Arawak languages. Arawak is geographically widespread and has fairly diverse phonological inventories. But aspirated voiceless stops (pʰ tʰ kʰ), nasal vowels (ĩ ẽ ã õ ũ), and the unrounded high central vowel (ɨ) are rare. The fact that Tariana — and Tukano — have all of these sounds points to borrowing as the right explanation. Upon closer inspection, we find that the aspirated voiceless stops are shared by Tariana and other Arawak languages in the region, and thus may not

have been borrowed. However, the case for Tukano-Tariana influence is still intact, with multiple possible causes, such as the fact that speakers from both communities practice linguistic exogamy (where one inherits one's language from one's father, and may not marry those that have inherited the same language) or the fact that Tariana speakers have been switching to Tukano, which has been promoted as a lingua franca by missionaries and civil authorities (Aikhenvald, 2003).

This abbreviated case study of phoneme borrowing had both quantitative and qualitative elements. In this article we describe a statistical test for performing the main quantitative task: measuring the extent to which borrowing from a proposed donor language is integral to explaning the phonological inventory of a target language. Just as in the case study, this purely quantitative measure of the plausibility of borrowing must be synthesized with sociohistorical and geographical considerations to yield a complete picture. But even by itself, the test can, given a reliable linguistic database, yield a panoptic view of how languages interact on a continental scale; and this can direct the linguist to phenomena that may merit further attention.

For reasons that will become clear below, we call this statistical test a RAM test, where RAM stands for *relaxed admixture model*. As a probabilistic model of admixture in languages, RAM has at least two antecedents. One is STRUCTURE, which was originally designed to cluster biological specimens by positing a small number of ancestral populations from which they descend, with the possibility for some specimens to be classified as deriving from multiple ancestral populations (Pritchard et al., 2000). STRUCTURE has been applied to linguistic data as well: Reesink et al. (2009) examined the distribution of typological features in languages of Maritime Southeast Asia and Australia, and Bowern (2012) evaluated the integrity of word list data from extinct Tasmanian languages as preparation for classifying the languages. Another antecedent of RAM is a model by Daumé (2009) in which a language's features are treated as an admixture of phylogenetically inherited features and areal features. In this model, linguistic phylogeny, the borrowability of each linguistic feature, and the locations of linguistic areas are all reified underlying variables.

RAM differs from its antecedents in two significant ways. Both STRUCTURE and Daumé's model are global models, in the sense that they seek a coherent explanation for the entire dataset. RAM is a local model. It evaluates the possibility of borrowing between a pair of languages, without regard to other languages. Despite the crudeness of this approach, we find that it suffices to generate interesting areal hypotheses and to answer basic questions such as which features were borrowed. RAM's simplicity also yields dividends in computational speed: it allows for fast, exact inference in the main calculation (see §4.3, §A.2).

The second way in which RAM differs from its antecedents is in how admixture is actually modeled. In both STRUCTURE and Daumé's model, every feature is assigned one source. Admixture is modeled by allowing different features in the same language to be assigned to different sources.[1] In RAM, a feature may have two sources, and the sources are additive. Each feature can be inherited with some frequency (first source), but failing that, the feature can still be borrowed from a donor (second source). In effect, the presence of a feature can be borrowed, but the absence of a feature cannot be. We

1 In STRUCTURE, the source is one of K ancestral populations. In Daumé's model, the source is either the phylogeny or the area. In both models, there is a latent matrix variable (written as Z in both cases) that designates the source for each of a language's features. The value of Z_{il} determines the source for feature l in language i. This source is used to look up the feature frequency for feature l, which is then used to generate the feature value via a Bernoulli distribution (i.e. tossing a biased coin).

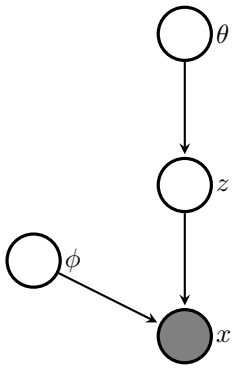


Figure 1: Diagram of a probabilistic generative model.

term this mechanism *relaxed admixture*. It is this mechanism that allows the model to detect more superficial contact, which we believe tends to be additive in nature.

In this paper we apply the RAM test to a database of the phonological inventories of South American languages, described in §2. Some statistical concepts undergirding this test are briefly discussed in §3. The test itself and the RAM model are presented in §4. Analysis results are discussed in §5 along with cultural and linguistic areas proposed by others. Finally §6 examines one such area more closely. The Upper Xingú, we argue, is a linguistic area, but it is hard to demonstrate this by other quantitative methods.

2. Dataset

Our analyses operate on phonological inventories obtained from SAPhon (Michael et al., 2013), which aims to be a high-quality, exhaustive database of the phonological inventories of the languages of South America. For each of 359 languages, SAPhon encodes its phonological inventory as a binary vector, with each element indicating the presence or absence of a particular phoneme in the phonological inventory. There are also a small number of elements in this vector that indicate more general things about the phonology of the language, such as whether it has tone or nasal harmony. In this article we will refer to the vector as a feature vector, and to each element as a linguistic feature. These features are not to be confused with phonological features such as *continuant* or *unrounded*, which are not features of languages but of phonemes.

Some regularization has been done on the phonological inventories, to make them easier to compare. For example, /ɛ/ has been replaced by /e/ whenever /e/ doesn't already exist, since in this case the choice between /e/ and /ɛ/ is fairly arbitrary. After regularization, the database has 304 features. Other information such as language family and geography are discarded during analysis, but are used in plotting results.

For more details on the dataset or the regularization procedure, please see the article by Michael et al. in this volume.

3. Probabilistic generative models

This work employs probabilistic generative models, which can be used to construct expressive models for diverse physical phenomena. Such models are often surprisingly tractable, thanks to a rich set of mathematical formalisms (Jordan, 2004). The term *gen-*

erative means that the data we seek to explain are modeled as having been generated via a set of hidden or underlying variables; and *probabilistic* means that variables are related by probabilistic laws, as opposed to deterministically.

Such models are often represented graphically as in Fig. 1, in which each variable is a node. By convention, an observed variable (i.e. data) is represented by a filled node. Thus, x is data and ϕ, θ, and z are unobserved, underlying variables.[2] Causal relationships between the variables are shown by arrows, with the direction of the arrow showing the direction of causation. Here, θ generates z; and ϕ and z together generate x: the model defines the conditional distributions $p(z \mid \theta)$ and $p(x \mid \phi, z)$. Variables such as ϕ and θ that lack arrows that lead to them are generated *ex nihilo* by drawing from respective prior distributions $p(\phi)$ and $p(\theta)$. These distributions encode our beliefs about what ϕ and θ could be, before we see the data.

It is important to note that the model as a whole is a description of how the data x is generated, and that the model assigns a probability to the data. There are typically many ways that the data could be generated, in the sense that the underlying variables could assume many different values and still generate the data with varying probabilities. But if we sum (or integrate) over all the possible values for the underlying variables, we get the absolute (i.e. marginal) probability of the data. More formally, the model provides that

$$p(x, z, \phi, \theta) = p(x \mid z, \phi)\ p(\phi)\ p(z \mid \theta)\ p(\theta).$$

Suppose that ϕ and θ are continuous variables, and that x and z are discrete. We can integrate over the continuous underlying variables and sum over the discrete underlying variables to get the marginal probability

$$p(x) = \int_\theta \int_\phi \sum_z p(x \mid z, \phi)\ p(\phi)\ p(z \mid \theta)\ p(\theta)\ d\phi\, d\theta.$$

We will interpret this probability as a measure of the aptness of the model. In this context, the marginal probability of the data is known as the marginal likelihood of the model. In the following section we will build two competing models for explaining the same data, calculate their marginal likelihoods, and use the ratio as a measure for their relative aptness.

4. RAM test

The RAM test is set up as a statistical hypothesis test. The analyst picks a target language and a donor language — these are treated as givens. Then we ask the question: is the inventory of the target language better explained as a product of inheritance from its linguistic family alone; or is it better explained as a joint product of inheritance and borrowing from the donor? These two hypotheses are fleshed out by two models: the inheritance only model $\mathcal{M}_0$, which we treat as a null hypothesis; and the relaxed admixture model $\mathcal{M}_1$, which we treat as the alternative hypothesis.

2 We will write x for both a random variable and for particular values of that random variable. We write $p(x)$ for the mass function of x if it is a discrete random variable, and the same for its density function if x is continuous. In expressions such as $x \sim \text{Beta}(1/2, 1/2)$ or $E[x] = \int xp(x)dx$, context should suffice to indicate that the first x in each expression is a random variable, and the other instances of x are bound values.

Universal frequency of feature l.	$\mu_l \in (0, 1)$.
Generality of universal feature frequency μ_l.	$\lambda_l \in (0, \infty)$.
Frequency of feature l in the target's family.	$\theta_l \sim \text{Beta}(\lambda_l \mu_l, \lambda_l(1 - \mu_l))$, $\theta_l \in (0, 1)$.
Feature l in target language	$x_{0l} \sim \text{Bernoulli}(\theta_l)$, $x_{0l} \in \{0, 1\}$.
Feature l in language n, which is in the target's family.	$x_{nl} \sim \text{Bernoulli}(\theta_l)$, $x_{nl} \in \{0, 1\}$.

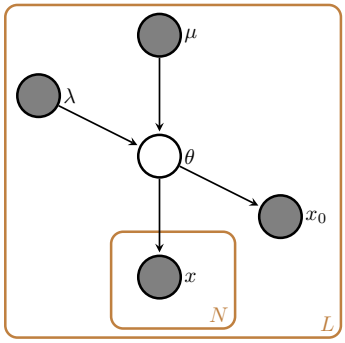


Figure 2: Inheritance-only model $\mathcal{M}_0$.

4.1. Model $\mathcal{M}_0$: Inheritance only

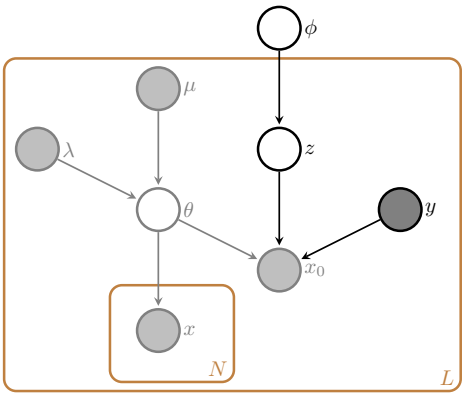
The inheritance only model is depicted in Fig. 2. The rounded rectangles are *plates*. They convey that the variables contained by them are arrayed. For example, θ is a vector with L elements, and x is an $N \times L$ matrix. Arrows that cross into a plate denote that each element of the downstream variable is independently generated and identically distributed. For example, the arrow from θ to x crosses a plate, denoting that for each l, the elements $x_{1l}, x_{2l}, \ldots, x_{Nl}$ are independently generated from θ_l and are identically distributed.

The inheritance-only model works by characterizing each language family as a vector of feature frequencies $\theta = (\theta_1, \ldots, \theta_L)$, one for each feature. Each language of the language family, including the target language, is modeled as being generated by these feature frequencies. The variable $x_0 = (x_{01}, x_{02}, \ldots, x_{0L})$ is a feature vector encoding the phonological inventory and other phonological characteristics of the target language, with x_{0l} encoding the presence (1) or absence (0) of feature l. N is the number of languages in the family of the target language, not counting the target language. The variable x is an $N \times L$ binary matrix that encodes the inventories of the other languages in the family. For each language n and feature l, x_{nl} is generated from θ_l. It is present ($x_{nl} = 1$) with probability θ_l or absent ($x_{nl} = 0$) otherwise. The feature frequency θ_l is generated by drawing from a beta distribution whose parameters are a function of μ_l and λ_l (see figure for details). The vector $\mu = (\mu_1, \ldots, \mu_L)$ contains "universal frequencies" for each feature and $\lambda = (\lambda_1, \ldots, \lambda_L)$ describes how universal these universal frequencies are. When λ_l is high, the feature frequency of l in each language family closely resembles the universal frequency μ_l, and the opposite is true when λ_l is low. These parameters become very significant when the target is an isolate, or when its family is small. There is not enough data to infer these parameters, so they are set before any RAM tests are run by estimating them from the entire dataset, as described in §A.1.

4.2. Model $\mathcal{M}_1$: Relaxed admixture model

The relaxed admixture model (RAM) is depicted in Fig. 3. Under relaxed admixture, the presence of a sound can be borrowed, but the absence of a sound cannot be.

The parts that are similar to the inheritance-only model have been grayed out. The new elements model the possibility for the target language to obtain features from a donor, whose feature inventory is denoted by feature vector y. The underlying variable



$\phi \sim \text{Beta}(\frac{1}{2}, \frac{1}{2})$, $\phi \in (0, 1)$.	Admixture parameter.
$z_l \sim \text{Bernoulli}(\phi)$, $z_l \in \{0, 1\}$.	Can feature l be borrowed?
$y_l \in \{0, 1\}$.	Feature l in donor language.
$\begin{cases} x_{0l} = 1 \text{ if } z_l = 1 \text{ and } y_l = 1, \\ x_{0l} \sim \text{Bernoulli}(\theta_l) \text{ otherwise.} \end{cases}$	Feature l in target language.
$x_{nl} \sim \text{Bernoulli}(\theta_l)$, $x_{nl} \in \{0, 1\}$.	Feature l in language n, which is in the target's family.

Figure 3: Alternative hypothesis model $\mathcal{M}_1$.

z is a vector, of which each element z_l encodes whether the target will attempt to borrow feature l from the donor. The target language has two chances to gain a feature. If the attempt is made ($z_l = 1$) and the donor has the feature ($y_l = 1$), it gets feature l from the donor. Otherwise, it may still get feature l via inheritance, with probability θ_l. The admixture parameter ϕ is used to generate z. Each element z_l will be one with probability ϕ. Since any feature will be borrowed with this probability *a priori*, the admixture parameter serves to denote the fraction of donor's features that are given to the target.

How realistic is RAM as a model of phoneme borrowing? It is easy to find examples where one language has influenced another to lose phonemes. Yánesha, an Arawak language, lacks mid vowels due to influence from Quechua (Wise, 1976). Nukak, a Kakua-Nukak language, lacks phonemic nasal stops due to influence from nearby Tukanoan languages (Epps, 2011). Yet it is our (unquantified) impression that the borrowing of phonemes is much more common than the borrowing of absences. We also proceed on the assumption that gaining a sound can easily happen in instances of superficial contact, but that losing a sound generally entails a deeper structural change in the phonology of the language, which necessitates more intense contact. We felt that being unable to model the latter phenomenon was a reasonable price to pay for having a simple model that was sensitive to the former phenomenon, which we posit to be more common.

A more general respect in which RAM is unrealistic is that each feature is modeled as largely independent of the others. Common sense (along with examples such as Tariana) suggests that many phonemes are borrowed as a clump, as in the case of aspirated voiceless stops or nasal vowels. Properly construed, what are borrowed are phonological features such as aspiration or nasality, which manifest as new sets of phonemes. The model, however, counts each phoneme as being borrowed on its own.

The way endogenous features are modeled is naive in the same way. Both $\mathcal{M}_0$ and $\mathcal{M}_1$ model inherited phonemes as being generated independently. Since our method relies on comparing two models, and the two models are naive in the same way, the clumping of features does not bias the result in favor of either model. However, clumps will cause the magnitude of the difference between the models to be exaggerated, since each clump of phonemes, in essence a unitary phenomenon, gets counted many times. Moreover, the borrowing of a large clump, such as nasal vowels, will have an outsized effect compared to the borrowing of a phoneme such as /ɨ/ which tends not to partici-

pate in clumps. This may cause the sensitivity of the RAM test to degrade substantially, but we have no way to address this problem.

4.3. Borrowing score

In order to quantify the aptness of $\mathcal{M}_1$ with respect to $\mathcal{M}_0$ for explaining the data x_0, we compute the Bayes factor, which is the ratio of the marginal likelihood of each model:[3]

$$\mathcal{K} = \frac{p(x_0 \mid x, y, \mathcal{M}_1)}{p(x_0 \mid x, \mathcal{M}_0)}.$$

We expand this so as to be explicit about which underlying variables are being summed or integrated over:

$$\mathcal{K} = \frac{\sum_z \int_\phi \int_\theta p(x_0, \theta, z, \phi \mid x, y, \mathcal{M}_1)\, d\theta}{\int_\theta p(x_0, \theta \mid x, \mathcal{M}_0)\, d\theta\, d\phi}.$$

Both models are simple enough that the Bayes factor can be computed exactly (see §A.2 for details). We use the log of the Bayes factor as a *borrowing score* for each donor-target pair. When the borrowing score is greater than zero, $\mathcal{M}_1$ is favored over $\mathcal{M}_0$, and one can conclude that it is more likely that the target borrowed features from the donor, than that all its features were inherited. In our analyses, we will look for borrowing by applying the RAM test to all pairs of languages in the dataset.

4.4. Caveats

No tractable mathematical model can account for every detail of a phenomenon as complex as language. This is all the more true of models as rudimentary as $\mathcal{M}_0$ and $\mathcal{M}_1$, which are, by design, just complex enough to yield useful results. Below we list the more conspicuous ways in which the models are unrealistic, and describe how this influences our interpretation of the borrowing score.

1. It bears repeating that the phenomenon of feature clumping, discussed in §4.2, causes the borrowing score to be exaggerated. Consequently it is ill-advised to interpret the borrowing score as the logarithm of a Bayes factor, despite the fact that formally, that is what it is. Applying the well-known interpretive scale proposed by Jeffreys (1961) would yield highly misleading results. Instead, we recommend evaluating a borrowing score by comparing it to what might be obtained from pairs of languages chosen at random (see analysis in §5) and by choosing a threshold that is high enough to filter out language pairs that are deemed unlikely, on extralinguistic grounds, to have interacted. It should be noted that more sophisticated models of admixture such as STRUCTURE and Daumé's model (§1) also assume that linguistic features are conditionally independent, and are susceptible to the same kind of problem. Short of

3 To avoid clutter we write $p(x_0 \mid x, \mathcal{M}_0)$ rather than $p(x_0 \mid \lambda, \mu, x, \mathcal{M}_0)$, and $p(x_0 \mid x, y, \mathcal{M}_1)$ rather than $p(x_0 \mid \lambda, \mu, x, y, \mathcal{M}_1)$, since λ and μ are fixed parameters. When the model being referred to is clear from context, we may write just $p(x_0 \mid x)$ or $p(x_0 \mid x, y)$.

building in a model for featural non-independence (this is a hard) the only way to ameliorate this problem is to choose datasets with features that are less non-independent.

2. $\mathcal{M}_0$ and $\mathcal{M}_1$ are not phylogenetic models. They do not model cladal innovations within language families. Consequently the borrowing score is unreliable as an indicator of intrafamily borrowing, as it will tend to ascribe to borrowing the effects of shared innovations. For example, a subset of Arawak languages in Peru have /tʲ/, which could well be a shared innovation. But since this phoneme is rare in Arawak as a whole, the RAM test will tend to treat it as having been borrowed between these Peruvian Arawak languages.

3. It must be emphasized that $\mathcal{M}_1$ is a local model of borrowing, in the sense that if another language is a more suitable choice as a donor than the donor in $\mathcal{M}_1$, that does not affect the borrowing score. In our analyses, some targets will have high borrowing scores with a large number of donors, but there is probably just one actual donor.

4. Borrowing score is often a poor indicator of the direction of borrowing. In theory it is asymmetric: if we switch the target and donor in a RAM test, the borrowing score will be different. In practice, however, the real direction of borrowing may correspond to the configuration that yields the lower score. How does this happen? Suppose language A_1 from family A has given features to languages B_1, B_2, and B_3, which make up family B. Since every language in B has features borrowed from A_1, the model will believe that these features are endogenous to family B, and may even believe that they were given to A_1. In some of our analyses we use a symmetric version of the borrowing score by computing the borrowing score for two languages both ways, and taking the larger score to be the score between the two languages.

5. Borrowing score is mildly transitive. If A gives features to B that B gives to C, the borrowing score between A and C may be high. This means that a high borrowing score is not necessarily indicative of direct contact. In practice this is not a serious problem: transitivity leads to a small amount of clutter in the results, but it is easy to identify and discount it.

6. Borrowing score does not respect distance. If one had wanted to model the effect of distance, one could, in $\mathcal{M}_1$, adjust the prior for the admixture parameter ϕ to have a lower mean for greater distances, but we did not do this, as we were unable to think of how to do it in a principled way. We opted instead to account for distance *post hoc*, as discussed in the next section.

5. Results

The plots in Fig. 4 show, for each pair of languages not in the same family, the higher of the two borrowing scores involving the pair, plotted against the distance between the

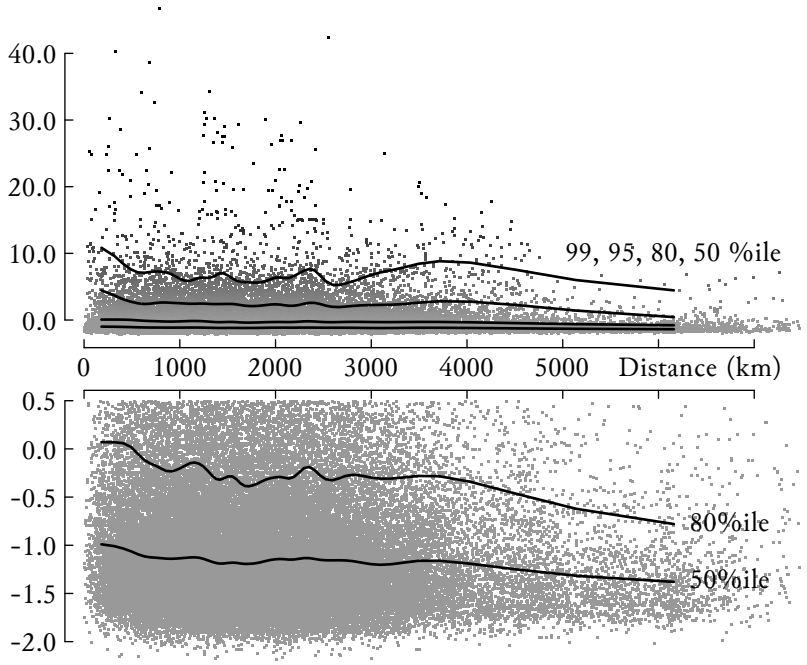


Figure 4: Interfamily borrowing scores by distance. The second plot is an enlargement of the first.

pair in kilometers. Also shown are quantile lines, conditioned on distance.[4] There are three things about this plot that accord with our intuitions about how a useful indicator of borrowing should behave.

- The median line is well below zero. The probability that two languages not from the same family, chosen at random, have a positive score is 0.17.
- The median line is relatively flat, indicating that borrowing score is not merely a function of proximity.
- The higher-quantile lines get higher as the distance decreases. The RAM test finds that the closer two languages are, the more profitable it is to posit borrowing. The bulge at 3800 km is due to how Andean and Patagonian languages have similar phonological profiles.

It is desirable to plot these results on a map, to show the language pairs that are likely to have had exchanged features. One possibility is to plot each pair with a score higher than zero, by drawing a line between the languages on the map. But this plot would be far too cluttered to be informative. In Fig. 5 we opt to plot just the pairs that are less than 400 km apart, that score higher than an arbitrary threshold of 3. At distances greater than

4 The median (50 percentile) line is drawn so that half the data points at any given distance are below it. The 80 percentile line is drawn so that 80% of the data points at any given distance are below it. Etc.

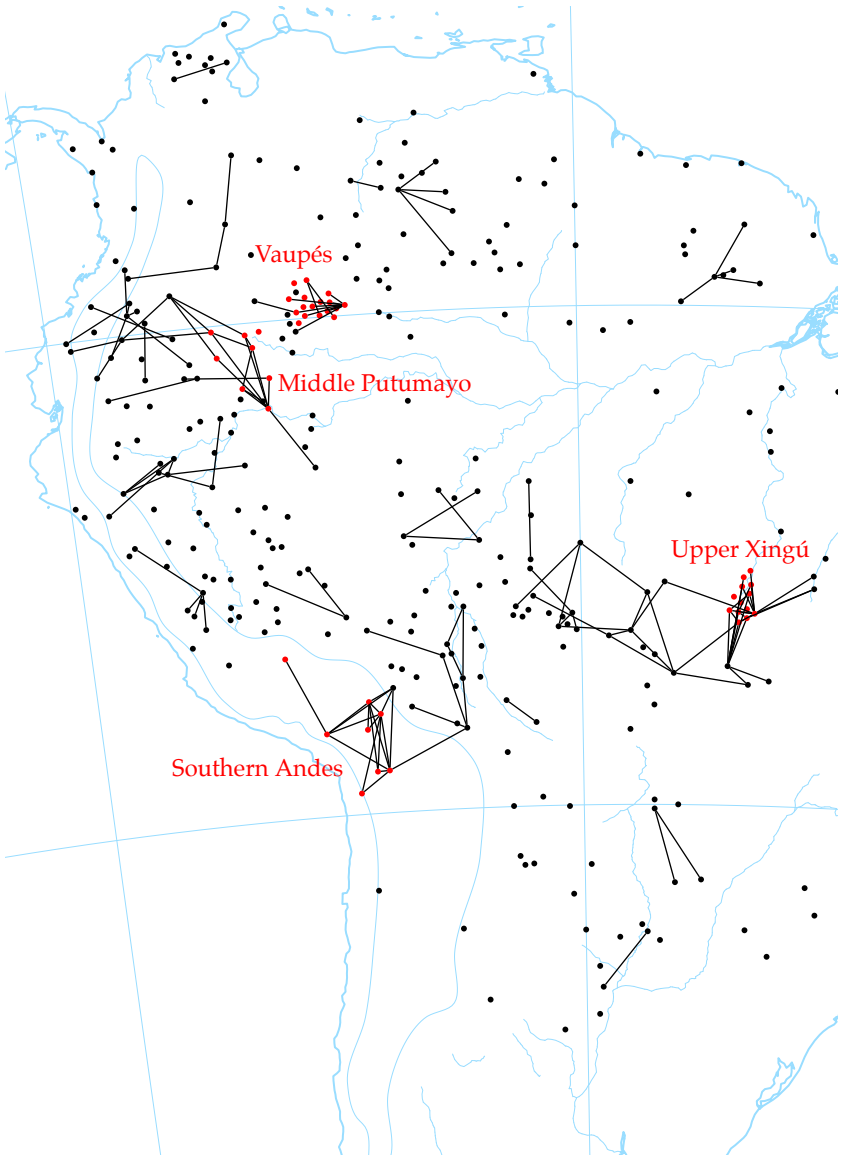


Figure 5: Line plot showing languages pairs with high borrowing scores. Some languages have been nudged apart for clarity. Languages in red belong to a proposed culture or linguistic area. The supplement contains a larger diagram with all languages labeled.

400 km, even borrowing scores in the 99 percentile range are mostly spurious, due to coincidental resemblances and to how feature clumping exaggerates borrowing scores.

Fig. 5 has 143 line segments, each representing a hypothesis of contact. Only by careful sociohistorical and geographical considerations can each of these hypotheses be confirmed, but it is very suggestive that there are some places on the map where the connections are especially dense. Four of these correspond to proposed culture areas

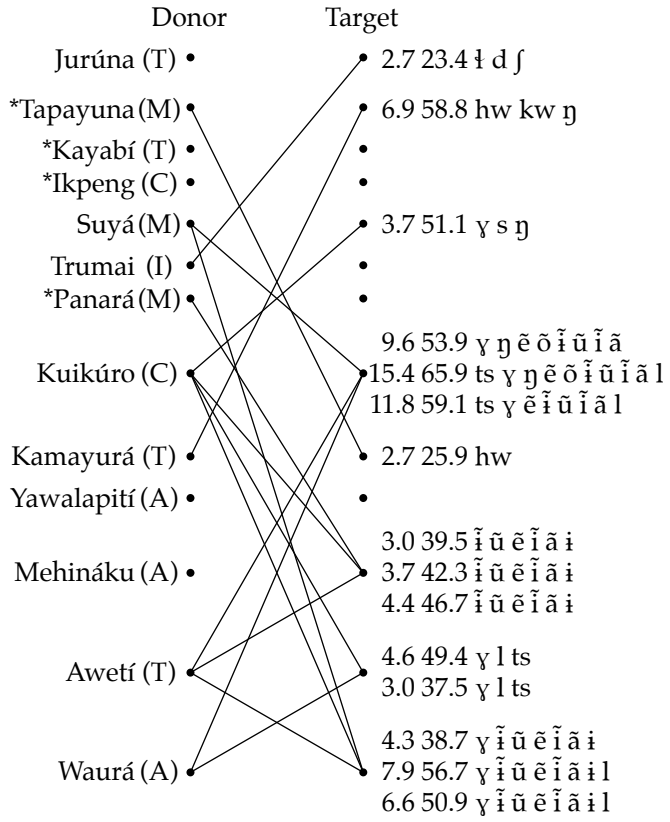


Figure 6: Languages of the Upper Xingú culture area, ordered from north (downriver) to south (upriver). Asterisk marks those that entered the area after 1950. Line segments show pairs with borrowing scores higher than 2. (Nodes on the left represent languages as donors; nodes on the right respresent them as targets.) For each such pair, we report the borrowing score, the mean percentage of the donor's segments that were borrowed, and the segments identified as likely borrowed, contingent on contact.

found in the literature on South American languages. The constituent languages are colored red.

In two of these areas, phonological diffusion has already been documented: the Vaupés (Aikhenvald, 2002) and the Southern Andes (Büttner, 1983). In the other two, our findings of phonological diffusion are novel, but plausible. The middle reaches of the Putumayo River and its tributaries constitute a culture area in lowland South America, whose constituent ethnolinguistic groups are known as the People of the Center (Seifart, 2011). More well-known is the Upper Xingú, a very strong culture area located along an upriver (southern) section of the Xingú River.

6. Upper Xingú

6.1. Phonological diffusion in the Upper Xingú

In this section we take a closer look at RAM test results for the languages of the Upper Xingú. Our main references for the linguistic situation in this culture area are articles by Lucy Seki (1999; 2011). Since the time of the colonization of Brazil, the Upper Xingú has served as a refugium for Indian tribes. As a culture area it is young — just 150 or 200 years in age — with a complex history of having other tribes pushed into its orbit by settlers or transferred there by the Brazilian government. Intermarriage is common, and on some occasions entire tribes that speak different languages came to live together. Of the 16 languages or dialects that Seki (1999) lists as being in or associated with the area, our dataset has phonological inventories for the 13 listed in Fig. 6. Their order in the diagram corresponds to their order along the banks of the Xingú River. The languages that entered the area after 1950 are marked with stars.

Except for Trumai, the Xinguan languages belong to four large families (Arawak, Carib, Macro-Ge, and Tupí) that all happen to be widely dispersed in South America. This makes it possible to infer the direction of borrowing. Accordingly, we distinguish between donor and target, and draw line segments for the pairs that score higher than 2. We have simplified the diagram somewhat by showing, for each target, only the highest-scoring donor from each language family. For each line segment, we report two numbers and a list of features. The first number is the borrowing score. The second is the estimated mean of ϕ in $\mathcal{M}_1$ scaled up by 100. This is the percentage of the donor's features given to the target.[5] The features that follow are the features that have the highest estimated means for z_l in $\mathcal{M}_1$. These are the features that were most likely to have been given to the target. (See §A.3 for the details of these calculations.)

We note three patterns in how phonemes diffuse in the Upper Xingú.

- Arawak and Carib languages are the recipients of nasal vowels.
- Arawak languages are the recipients of /ɨ/
- Carib and Tupí languages are the recipients of /ts/.

Table 2 gives some intuition for how the model arrived at these conclusions, by listing the continent-wide frequencies of these features in Arawak, Carib, Macro-Ge, and Tupí. We see that in each case, the recipient of a feature belongs to a family in which the received feature occurs less frequently. Table 2 also serves to show that the features in question have diffused widely, at least among the languages that are older to the area.

5 One reviewer remarked that these borrowing percentages are huge — as much as 65.9% in the case of Awetí to Kuikúro. Such a high figure is partly an effect of relaxed admixture, but also not surprising after considering the particulars of these languages.

With a standard model of admixture, a borrowing percentage of 100% means that the target ends up identical to the donor, since it borrows both presence and absence of each of the donor's feature values. Under relaxed admixture, even a borrowing percentage of 100% does not prevent the target from receiving features by inheritance. It would only mean that the target receives every feature that the donor has. The flexibility of relaxed admixture implies higher borrowing percentages.

As for Awetí and Kuikúro, each has 26 sounds, of which they share 22. Of the sounds that are in less than a third of all Carib languages, Kuikúro has /ts dʲ ŋ ɣ l ĩ ẽ ã õ ũ ɨ̃/ and Awetí has all of these except /dʲ/. As a donor, Awetí thus accounts nicely for 10 of Kuikúro's 26 sounds, and has only 4 sounds that are not in Kuikuŕo: /ʔ ʐ̩ j ɾ/.

	/ɨ/	/ts/	/ã/, etc.
Arawak	65	19	30
Carib	3	14	100
Macro-Ge	7	79	93
Tupí	15	85	96
Jurúna (T)	Y		Y
Suyá (M)	Y		Y
Trumai (I)	Y	Y	
Kuikúro (C)	Y	Y	Y
Kamayurá (T)	Y	Y	Y
Yawalapití (A)	Y	Y	
Mehináku (A)	Y	Y	Y
Awetí (T)	Y	Y	Y
Waurá (A)	Y	Y	Y

Table 2: Feature frequencies in four South American language families, normalized to 100; and whether those features are present in the nine languages from Figure 6 that are older to the Upper Xingú (those that predate 1950). The last column is for any nasalized vowel.

Our analysis often suggests several candidate donors for each target, but there is often no obvious reason to prefer one over another. It may even be the case that the actual donor is now extinct. On the other hand, the identity of recipients is less equivocal, since that is inferred from the fact that they have features that are unlikely to be endogenous, and that possible donors exist. It is worth noting that of the five languages not identified as recipients, three are recent arrivals.[6]

6.2. A linguistic area without distinctive features?

The Upper Xingú is not documented as having many distinctive linguistic features. We can show that this is actually the case for phonemes by attempting to train a naive Bayes classifier to discriminate between Xinguan languages and other Amazonian languages (see Michael et al., this volume, for details on this method). Let us posit an areal core consisting of the nine languages in Fig. 6 that were present in the Upper Xingú before 1950. We also construct a control class of languages to serve as a background to the core languages. These consist of the 43 languages in SAPhon that are at a distance of between 400 km and 1,000 km from the nearest language of the core. These two sets of languages are depicted via dots of different shapes in Fig. 7. They are fed into a naive Bayes classifier, which calculates *feature deltas*, which are the strength of the association between each feature and each training class (Fig. 8). The classifier also calculates *language scores*, which denote the probability of membership of each language in each training class (Fig. 7).

6 The other two are Yawalapití and Trumai. Seki (1999: 426) refers to a reconstruction of Proto-Arawak to suggest that Yawalapití /ɨ/ may be a diffused Xinguan feature; but since /ɨ/ is present in 30% of Arawak languages, the RAM test, with its coarse model of inheritance, could not conclude the same. As for Trumai, it is hard for the RAM test to decide if any of its features are exogenous because it is an isolate.

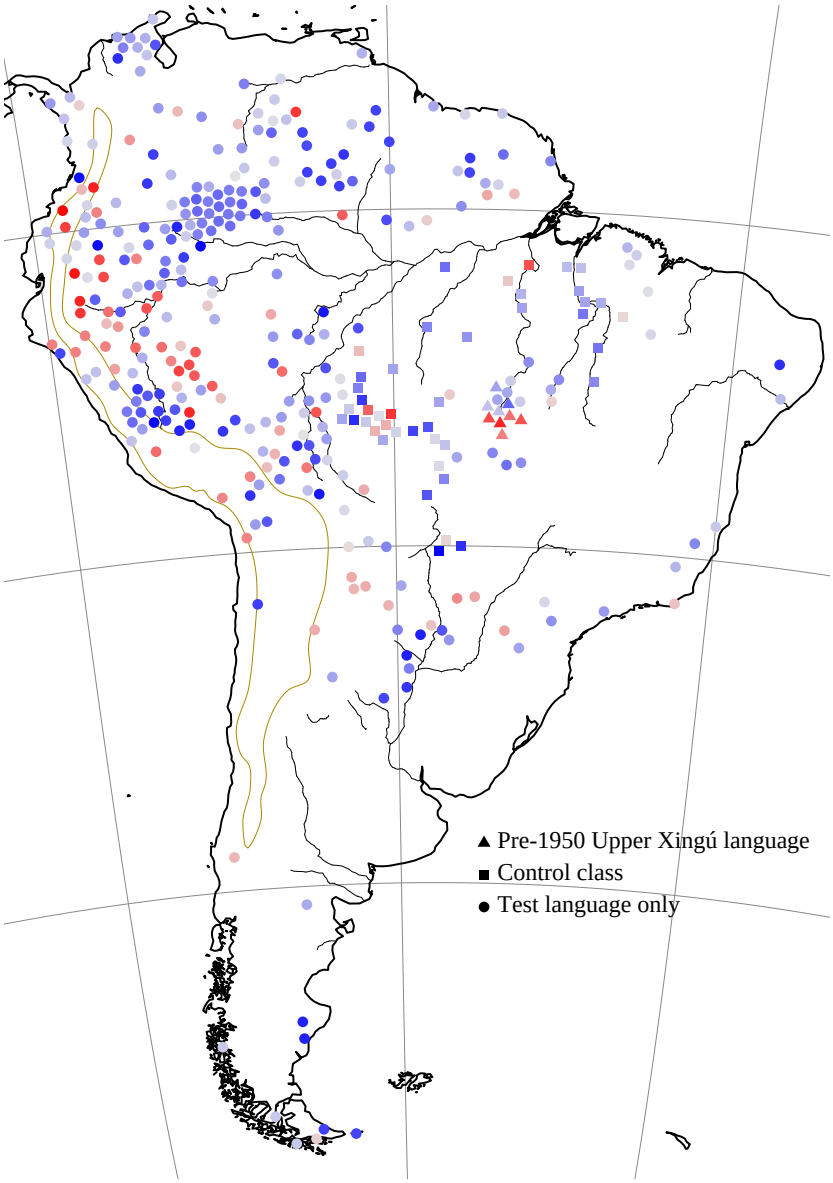


Figure 7: Language scores assigned by a naive Bayes classifier. Redness (or blueness) denotes the probability of membership in the Upper Xingú core (or the control).

As can be seen from Fig. 7, the classifier is unsuccessful in discriminating between core languages and other languages. There are languages in the core that it believes, based on their features, are not core-like (the blue triangles). And there are a very large number of languages in both the control set and farther away that the classifier believes should be in the core (red squares and circles). This suggests that the core languages lack distinctive features. Fig. 8 provides more direct evidence of this. There are just a few features with deltas greater than 2 or less than -2: these are relatively strong indicators of the membership of a language. The analysis tells us that a Xinguan language is

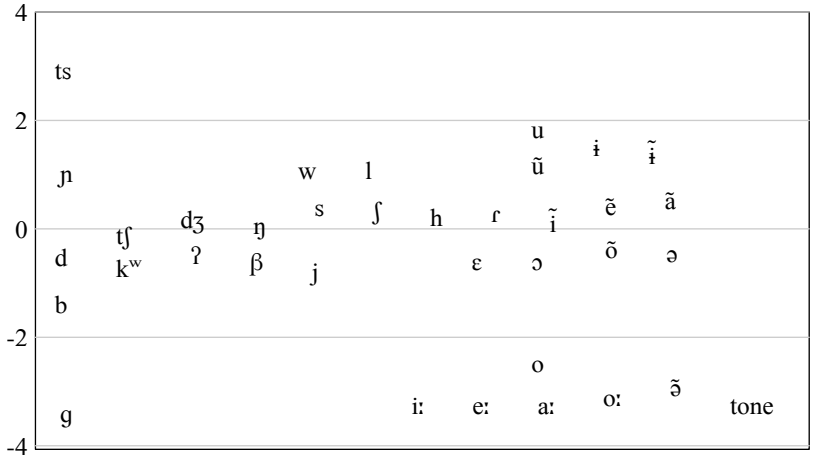


Figure 8: Feature deltas assigned by a naive Bayes classifier. Positivity (or negativity) along the y-axis denotes the strength of the association between a feature and the Upper Xingú core (or the control).

distinguished from its Amazonian neighbors by the presence of /ts/ and the absence of /ɡ/, /o/, /ə̃/, tone, and long vowels. Clearly it is easy enough for this sort of language to arise by chance, as it has in many other parts of South America.

We thus conclude that the Upper Xingú lacks truly distinctive phonemes. But could it still be considered a linguistic area? Seki has called it an *incipient* linguistic area, noting that as a culture area, it is young, and that there are relatively few features that have diffused throughout the area. Our analysis favors the less nuanced conclusion that it is a full-fledged linguistic (or, more precisely, phonological) area. The Upper Xingú was one of the areas to be picked out by the RAM test as exhibiting a high density of potential pairwise borrowing. We looked into $\mathcal{M}_1$ to see which languages and sounds were involved, and found that the inferred targets of borrowing and the sounds inferred as transfered were plausible. We also saw that some of these sounds were widespread among the more longstanding members of the area. Despite that the sounds are not distinctively Xinguan, we were led to the conclusion that they diffused throughout the area. What remains is for other linguistic features (lexical, typological) to be investigated, and other investigations to be carefully synthesized with ours, before an unequivocal conclusion can be reached.

References

Aikhenvald, Alexandra Y. 2002. *Language Contact in Amazonia*. Oxford University Press.

——. 2003. *A Grammar of Tariana, from Northwest Amazonia*. Cambridge University Press.

Bowern, Claire. 2012. The riddle of Tasmanian languages. *Proceedings of the Royal Society B: Biological Sciences* 279: 4590–4595.

Büttner, Thomas Th. 1983. *Las lenguas de los Andes centrales: Estudios sobre la clasificación genética, areal y tipológica*. Ediciones Cultura Hispánica del Instituto de Cooperación Iberoamericana, Madrid.

Daumé, Hal, III. 2009. Non-parametric Bayesian areal linguistics. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, 593–601.

Epps, Patience. 2011. Phonological diffusion in the Amazonian Vaupés. Talk presented at the CUNY Phonology Forum Conference on the Phonology of Endangered Languages.

Hastings, W.K. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57: 97–109.

Jeffreys, Harold. 1961. *The Theory of Probability*. Oxford University Press, 3rd ed.

Jordan, Michael I. 2004. Graphical models. *Statistical Science* 19: 140–155.

Michael, Lev, Tammy Stark, and Will Chang (compilers). 2013. South American Phonological Inventory Database v1.1.3. Survey of California and Other Indian Languages Digital Resource. Berkeley: University of California.

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

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

Seifart, Frank. 2011. *Bora loans in Resígaro: Massive morphological and little lexical borrowing in a moribund Arawakan language.* Série Monografias, 2, Cadernos de Etnolingüística.

Seki, Lucy. 1999. The Upper Xingú as an incipient linguistic area. In Dixon, R.M.W. and Alexandra Y. Aikhenvald (eds.) *The Amazonian Languages*, Cambridge University Press.

——. 2011. Alto Xingu: uma área linguística? In Franchetto, Bruna (ed.) *Alto Xingu: uma sociedade multilíngue*, Rio de Janeiro: Museu do Índio - Funai, 57–84.

Wise, Mary Ruth. 1976. Apuntes sobre la influencia inca entre los amuesha: Factor que oscurece la clasificación de su idioma. *Revista del Museo Nacional* 42: 355–66.

A. Appendix

A.1. Universal feature frequencies

Both the inheritance-only model $\mathcal{M}_0$ and the relaxed admixture model $\mathcal{M}_1$ require reasonable settings for the universal feature frequencies μ and the generality parameter λ, both of which are vectors of L elements, where L is the number of features in the analysis. In order to set μ and λ, we extend $\mathcal{M}_0$ to include all languages in the dataset (Fig. 9) and estimate the mean of μ and λ in this extended model using Markov chain Monte Carlo sampling. We set μ and λ in $\mathcal{M}_0$ and $\mathcal{M}_1$ to these estimated means.

The extension of $\mathcal{M}_0$ is defined as follows. We write K for the number of language families, and N_k for the number of languages in family k. The data x is organized as a vector of binary matrices $(x_1, \ldots, x_K)$, where x_k is a matrix of size $N_k \times L$. Each family k is characterized by a bank of feature frequencies $\theta_k = (\theta_{k1}, \ldots, \theta_{kL})$, one for each feature. Feature l in language n of family k is present ($x_{knl} = 1$) with probability θ_{kl} or absent ($x_{knl} = 0$) otherwise. Feature frequency θ_{kl} is generated by drawing from a beta distribution whose shape is determined by λ_l and μ_l, and whose mean is μ_l. Each μ_l is drawn from a beta distribution parameterized by ρ_1 and ρ_0. Each λ_l is drawn from a gamma distribution with shape parameter α and rate parameter β. Tying the elements of μ, λ, and θ together in this way is a form of data smoothing. It prevents them from being too extreme with features that are very common or very rare, as would be the case if they

	$\rho_1, \rho_0 \in (0,1)$, drawn uniformly.
	$\alpha, \beta \in (0,\infty)$, drawn uniformly.
Universal frequency of feature l.	$\mu_l \sim \text{Beta}(\rho_1, \rho_0)$, $\mu_l \in (0,1)$.
Generality of universal feature frequency μ_l.	$\lambda_l \sim \text{Gamma}(\alpha, \beta)$, $\lambda_l \in (0,\infty)$.
Frequency of feature l in family k.	$\theta_{kl} \sim \text{Beta}(\lambda_l\mu_l, \lambda_l(1-\mu_l))$, $\theta_{kl} \in (0,1)$.
Feature l in language n of family k.	$x_{knl} \sim \text{Bernoulli}(\theta_{kl})$, $x_{knl} \in \{0,1\}$.

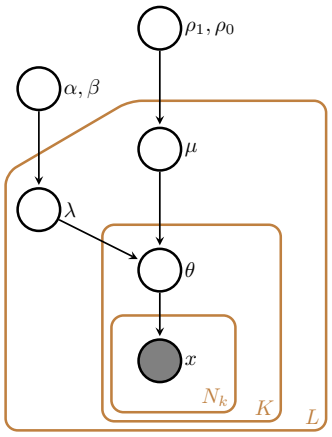


Figure 9: Model for estimating μ and λ.

were estimated in a less structured way (e.g., assigning to θ_{kl} the number of occurrences of feature l in cluster k, divided by the number of languages in cluster k).

During inference we collapse θ and work directly with the relationship between x, λ and μ. If we define $N_{kl} = x_{k1l} + x_{k2l} + \cdots + x_{kN_kl}$, then for all $k \in \{1, \ldots, K\}$, conditioned on λ_l and μ_l, N_{kl} has a beta-binomial distribution, and

$$\begin{aligned} p(x_{k1l}, \ldots, x_{kN_kl} \mid \lambda_l, \mu_l) &= \binom{N_k}{N_{kl}}^{-1} p(N_{kl} \mid \lambda_l, \mu_l) \\ &= \frac{B(N_{kl} + \lambda_l\mu_l, N_k - N_{kl} + \lambda_l(1-\mu_l))}{B(\lambda_l\mu_l, \lambda_l(1-\mu_l))}, \end{aligned}$$

where $B(\cdot,\cdot)$ is the beta function $B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}dt$. We sampled each of the uncollapsed variables α, β, ρ_1, ρ_0, λ, and μ from their posterior distributions using the Metropolis-Hastings algorithm (Hastings, 1970).[7] We write $\hat{\alpha}$ for the posterior mean of α, etc. We obtained these posterior means for the following hyperparameters.

$$\hat{\rho}_1 \approx 0.34 \qquad \hat{\rho}_0 \approx 2.24 \qquad \hat{\alpha} \approx 3.89 \qquad \hat{\beta} \approx 1.75$$

It is interesting that of the features for which $\hat{\mu}_l > 0.1$, the ones with the highest $\hat{\lambda}_l$ values are /h/, /dʒ/, and /p/, which have $\hat{\lambda}_l$ values of 3.45, 3.31, and 3.20 and $\hat{\mu}_l$ values of 0.68, 0.14, and 0.91. These phonemes are inferred to have similar feature frequencies in all language families. The ones with the lowest $\hat{\lambda}_l$ values are /kʼ/, /tʼ/, and /q/, which have $\hat{\lambda}_l$ values of 0.50, 0.54, and 0.77, and $\hat{\mu}_l$ values of 0.14, 0.16, and 0.10. These are sounds that appear either at high or low frequencies, depending on the language family, but seldom at frequencies close to $\hat{\mu}_l$.

7 In our MCMC sample chain, we resampled each uncollapsed variable in a fixed order, and did this 100,000 times. Each element of λ and μ was resampled individually. We discarded the first half of the sample chain and used the second half as the posterior sample.

A.2. Marginal likelihoods

A.2.1. Marginal likelihood of $\mathcal{M}_0$. To compute the marginal likelihood of $\mathcal{M}_0$, we observe that it factorizes:

$$p(x_0 \mid x) = \prod_{l=1}^{L} p(x_{0l} \mid x_{\cdot l}),$$

with $x_{\cdot l}$ being a shorthand for the vector $(x_{1l}, \ldots, x_{Nl})$. We capitalize on the fact that the beta distribution is the conjugate prior of the Bernoulli distribution, and observe that

$$\theta_l \mid x_{\cdot l} \sim \text{Beta}(N_l + \lambda_l \mu_l, N - N_l + \lambda_l(1 - \mu_l)),$$

where $N_l = x_{1l} + \cdots + x_{Nl}$. This implies that

$$p(x_{0l} \mid x_{\cdot l}) = \begin{cases} \frac{N_l + \lambda_l \mu_l}{N + \lambda_l} \text{ if } x_{0l} = 1, \\ 1 - \frac{N_l + \lambda_l \mu_l}{N + \lambda_l} \text{ if } x_{0l} = 0. \end{cases} \tag{A.1}$$

A.2.2. Marginal likelihood of $\mathcal{M}_1$. To compute the marginal likelihood of $\mathcal{M}_1$, it is useful to think of this model as a mixture of $\mathcal{M}_0$-like models, with each possible value of z yielding a component of the mixture:

$$p(x_0 \mid x, y) = \sum_z p(z)\, p(x_0 \mid z, x, y). \tag{A.2}$$

We write $H(z)$ for $z_1 + \cdots + z_L$ and observe that $H(z)$ has a beta-binomial distribution, and thus

$$p(z) = \frac{B(H(z) + \frac{1}{2})}{B(L + 1)}.$$

The conditional probability of x_0, like the marginal probability of x_0 under $\mathcal{M}_0$, factorizes:

$$p(x_0 \mid z, x, y) = \prod_{l=1}^{L} p(x_{0l} \mid z_l, x_{\cdot l}, y_l).$$

When $z_l = 0$, the lth factor is identical to the quantity computed in Eq. A.1:

$$p(x_{0l} \mid z_l = 0, x_{\cdot l}, y_l) = p(x_{0l} \mid x_{\cdot l}, \mathcal{M}_0)$$

When $z_l = 1$, it is

$$p(x_{0l} \mid z_l = 1, x_{\cdot l}, y_l) = \begin{cases} \frac{N_l + \lambda_l \mu_l}{N + \lambda_l} & \text{if } x_{0l} = 1 \text{ and } y_l = 0, \\ 1 - \frac{N_l + \lambda_l \mu_l}{N + \lambda_l} & \text{if } x_{0l} = 0 \text{ and } y_l = 0, \\ 1 & \text{if } x_{0l} = 1 \text{ and } y_l = 1, \\ 0 & \text{if } x_{0l} = 0 \text{ and } y_l = 1. \end{cases} \tag{A.3}$$

Note that this equation is what establishes relaxed admixture. In a more convention model of admixture, the right hand side would simply be one when $x_{0l} = y_l$ and zero otherwise.

Now that we explained the elements of Eq. A.2, we turn to the question of how to evaluate it. Since z has 2^L possible values, simply summing over all terms is computationally infeasible. Our solution exploits a recurrence relation to evaluate Eq. A.2 in $O(L^2)$ arithmetic operations. For notational convenience, we define:

$$\begin{aligned} a_l &= p(x_{0l} \mid z_l = 0, x_{\cdot l}, y_l), \\ b_l &= p(x_{0l} \mid z_l = 1, x_{\cdot l}, y_l) \\ &\qquad \text{for } l = 1, \ldots, L. \end{aligned}$$

and

$$\begin{aligned} q_h &= p(z) \text{ when } H(z) = h, \\ s_h &= \sum_{z:H(z)=h} p(x_0 \mid z, x, y) \\ &\qquad \text{for } h = 0, \ldots, L. \end{aligned}$$

To be explicit, the summation for s_h is over all values of z that contain h ones. We can rearrange the terms of Eq. A.2 thus:

$$p(x_0 \mid x, y) = \sum_{h=0}^{L} q_h s_h. \tag{A.4}$$

What remains is to compute s_h efficiently. We define the recurrence relation

$$S_h^l = \begin{cases} 1 & \text{if } h = 0 \text{ and } l = 0, \\ 0 & \text{if } h > l \text{ or } h < 0, \\ S_h^{l-1} a_l + S_{h-1}^{l-1} b_l & \text{otherwise,} \end{cases}$$

whence $s_h = S_h^L$ for $h = 0, \ldots, L$. To make it easy for the reader to verify this recurrence, we write out S_h^l for small values of l and h:

	$h=0$	$h=1$	$h=2$	$h=3$
$l=0$	1	0	0	0
$l=1$	a_1	b_1	0	0
$l=2$	a_1a_2	$a_1b_2 + b_1a_2$	b_1b_2	0
$l=3$	$a_1a_2a_3$	$a_1a_2b_3 + a_1b_2a_3 + b_1a_2a_3$	$a_1b_2b_3 + b_1a_2b_3 + b_1b_2a_3$	$b_1b_2b_3$

Note that S_h^l is what s_h would be if just the first l features were part of the model. In our routines for computing S_h^l we represent it on a log scale, to avoid problems of floating-point underflow.

A.3. Borrowed features

A.3.1. Fraction of features transferred. In the relaxed admixture model $\mathcal{M}_1$, the parameter ϕ describes the fraction of the donor's features that are transferred to the target. We describe how to infer the posterior mean of ϕ. Variables defined in our calculation of the marginal likelihood (§A.2.2) apply here. As in that calculation, we treat $\mathcal{M}_1$ as a mixture.

$$E(\phi \mid x_0, x, y) = \sum_z E(\phi \mid z) p(z \mid x_0, x, y).$$

By Bayes' theorem,

$$E(\phi \mid x_0, x, y) = \frac{\sum_z E(\phi \mid z) p(z) p(x_0 \mid z, x, y)}{p(x_0 \mid x, y)}. \tag{A.5}$$

The denominator is the marginal likelihood. The numerator is similar in form to the marginal likelihood as stated in Eq. A.2, but each term has an extra factor of $E(\phi \mid z)$. We capitalize on the fact that we are working with conjugate distributions, and observe that

$$\phi \mid z \sim \text{Beta}(H(z) + \tfrac{1}{2}, L - H(z) + \tfrac{1}{2}),$$

whence

$$E(\phi \mid z) = \frac{H(z) + \frac{1}{2}}{L + 1}.$$

Since this is a function of $H(z)$, we can evaluate the numerator of Eq. A.5 efficiently. We write r_h for $E(\phi \mid z)$ when $H(z) = h$. Then,

$$\sum_z E(\phi \mid z) p(z) p(x_0 \mid z, x, y) = \sum_{h=0}^{L} r_h q_h s_h.$$

A.3.2. Loan probability of a feature. We now describe how to compute the *transfer probability* of feature l, i.e. the probability that feature l was transferred from the donor to the target under model $\mathcal{M}_1$. Variables defined in §A.2.2 apply here. The transfer probability of feature l is

$$\Pr(z_l = 1 \mid x_0, x, y) = \frac{\sum_{z : z_l = 1} p(z) p(x_0 \mid z, x, y)}{p(x_0 \mid x, y)}.$$

The denominator is the marginal likelihood (§A.2.2). The numerator can be computed in a similar way as the marginal likelihood. For feature $l = L$, the numerator could be restated as

$$\sum_{z : z_L = 1} p(z) p(x_0 \mid z, x, y) = \sum_{h=1}^{L} q_h S_{h-1}^{L-1} b_L.$$

To calculate the transfer probability of other features, we reorder the features so that the feature of interest is in position L. Concomitantly we must recompute S_{h-1}^{L-1} for $h = 1, \ldots, L$, necessitating $O(L^2)$ operations.[8]

8 There is a way to compute the numerator for any feature l in $O(L)$ operations, but it is numerically very unstable. We describe it here in hopes that an interested reader could invent a numerically stable version of it. We restate the numerator as $\sum_{h=0}^{L} q_h t_h$, where

$$t_h = \sum_{\substack{z:H(z)=h, \\ z_l=1}} p(x_0 \mid z, x, y).$$

By this definition, $t_0 = 0$. Other values in the sequence can be derived via the recurrence relation $t_h = (b_l/a_l)(s_{h-1} - t_{h-1})$.