Title
Exploring phonological areality in the circum-Andean region using a Naive Bayes Classifier

Journal
Language Dynamics and Change, 4(1)

Author
Michael, Lev David

Publication Date
2013-12-04

Peer reviewed

Exploring phonological areality in the circum-Andean region using a naive Bayes classifier

Lev Michael, Will Chang, and Tammy Stark

December 4, 2013

Abstract

This paper describes the Core and Periphery technique, a quantitative method for exploring areality that uses a naive Bayes classifier, a statistical tool for inferring class membership based on training sets assembled from members of those classes. The Core and Periphery technique is applied to the exploration of phonological areality in the Andes and surrounding lowland regions, based on the South American Phonological Inventory database (SAPhon 1.1.3; Michael, Stark, and Chang 2013). Evidence is found for a phonological area centering on the Andean highlands, and extending to parts of the northern and central Andean foothills regions, the Chaco, and Patagonia. Evidence is also found for Southern and North-Central phonological sub-areas within this larger phonological area.

Keywords: naive Bayes classifier; linguistic areality; Andean languages; South American languages

1 Introduction

The goals of this paper are to describe the Core and Periphery technique, an intuitively appealing quantitative method for exploring large linguistic datasets for evidence of linguistic areality, and to illustrate the utility of this technique by applying it to a dataset of South American phonological inventories, focusing on the evidence of phonological areality in the Andes and surrounding lowland areas.

Core and Periphery is a method that uses as a starting point linguists' knowledge of the languages and history of a region to generate initial hypotheses regarding 'cores': sets of languages that constitute possible linguistic areas (Campbell, Kaufman, and Stark 1986, Thomason 2000, Muysken 2008), or parts of ones. These hypotheses serve as the seed for the application of a statistical technique, naive Bayes classification (NBC), which determines what features, if any, distinguish the core languages from other languages in the region, and also to what degree languages outside the proposed core resemble the core languages. Those languages deemed core-like, together with the proposed core, constitute a candidate linguistic area, to be evaluated against pertinent sociohistorical and geographical facts. If the languages deemed core-like fail to make sense geographically, then the Core and Periphery technique has failed to identify a linguistic area around the proposed core.

The Core and Periphery technique improves on conventional practices of 'eyeballing' areas in three ways. First, it provides a quantitative evaluation of the degree to which the languages of a proposed area in fact exhibit features that distinguish them from the languages of the larger region of which the proposed area forms a part. Second, it provides a quantitative measure of similarity between languages that can be applied to large datasets, allowing linguists to locate unexpected similarities that help identify new areas or redefine accepted ones. And third, quantitative measures of similarity also make it possible to visualize and cogently discuss the structure of linguistic areas whose boundaries are gradient in nature. Note, however, that Core and Periphery is not strictly speaking a statistical test of areality, a point we return to in §6.

In this paper carry out two different Core and Periphery explorations of phonological areality in the circum-Andean region, first treating the entire Andean highlands from northern Chile to northern Ecuador as a single core, and then treating the Andean highlands as being constituted of two cores, a Southern Andean core and a North-Central Andean core. The dividing line between the latter two cores runs through the southern Peruvian Andes, grouping Cuzco-Collao Quechua and Jaqaru with the Southern Andean core, while the remaining Quechuan languages constitute the North-Central core. This dual core analysis is motivated by the qualitative observation that the Southern Andean languages, delimited in this way, share a number of phonological characteristics otherwise rare in South America, including a three-way contrast between plain, aspirated, and ejective stops.

The single core analysis reveals several clusters of languages in the Andean foothills and adjacent lowland regions that pattern more strongly with the languages of the Andean core than other lowland languages, including an Ecuadorean Andean foothills cluster, a Huallaga River valley cluster, a cluster of Arawak languages of the southern Peruvian Andean foothills, and a cluster of Chacoan and Patagonian languages. These results support the existence of a large South American phonological area that encompasses the Andean highlands and parts of the Andean foothills regions, with a tongue that extends from the Southern Andes into the Chaco and Patagonia.

The dual core analysis builds on the single core analysis by revealing a finer structure to this area, showing that the non-Andean languages which exhibit similarity to Andean languages generally resemble those of the core to which they are most proximally located, with the relevant Chacoan and Patagonian languages resembling the those of the Southern core, and the relevant languages of Peru and Ecuador resembling those of the North-Central core.

This paper is organized as follows: §2 presents a qualitative overview of the Core and Periphery technique, and §3 presents the data to which this technique is applied, as well as the overall goals of the analysis. A more technical description of the statistical technique underlying the Core and Periphery technique, the naive Bayes classifer, is provided in §4, with additional details provided in §A.2–A.4. The results of single and dual core analyses are presented and examined in §5, and §6 evaluates the Core and Periphery technique, discussing its strengths and weaknesses.

2 The Core and Periphery technique: A qualitative overview

The basic strategy for exploring phonological areality implemented by the Core and Periphery technique is to use a measure of inter-language similarity to bootstrap from a given set of geographically clustered and phonologically similar languages (the 'proposed core') to a larger set of similar languages (the 'core and periphery') that are deemed to form a quantitatively consistent linguistic area.

In a one-core analysis, the first step is to divide the languages of a region (South America, in our case) into three sets: a proposed core, a control class, and an equivocal class. The proposed core is a set of languages that are hypothesized to form a part of a larger linguistic area. The control set consists of languages that are unlikely to have been in contact with the core languages, and are therefore deemed unlikely to belong to the core or periphery *ex hypothesi*.[1] The equivocal class is that about which nothing is claimed in advance.

Motivations for choosing a proposed core may include ethnographic or historical observations that suggest the existence of a culture area, intuitions regarding areality based on 'eyeballing' the linguistic data, or even previous proposals that the core constitutes a linguistic area. As will become clear, the original rationale for selecting a particular core is unimportant for the operation of the quantitative analysis described below, since the results of that analysis will indicate whether the proposed core in fact constitutes a distinctive and homogeneous sub-area of a larger linguistic area.

Choosing the control class entails identifying a set of languages that are unlikely to have been influenced by contact with the core languages. The ultimate choice of non-core or 'control' languages depends a great deal on the analyst's knowledge of the history and geography of the region, but we have generally allowed the possibility of quite distant linguistic influence, leading us to select control regions that are quite distant from the cores. In the case of the single Andean core that we discuss in §5.1, for example, we define the control languages as consisting of all languages further than 1500 kilometers from the Andean core.[2]

After choosing these three sets, a naive Bayes classifier is trained on the proposed core and the control class. These two classes serve to exemplify the opposite ends of an axis along which the classifier will score languages. The classifier is then used to score *all* languages, include those from the proposed core and the control class. The highest-scoring languages constitute a refined hypothesis for a linguistic core, which likely includes most or all of the proposed core, providing it was well chosen to begin with. At the opposite end of the spectrum there will be languages with very low scores, most of which will be non-core languages, if the proposed core was well chosen. Finally, in some analyses such as ours, there will be languages with intermediate scores that are geographically clustered near the proposed core. These constitute the periphery.

With the NBC analysis complete, the final step of the Core and Periphery technique is to evaluate whether the languages with relatively high NBC scores were ever plausibly

[1] As one reviewer suggested, even languages on another continent could serve as control languages.

[2] The Core and Periphery results actually suggest that in most cases, the range of phonological influence of the Andes into the surrounding lowlands does not exceed a few hundred kilometers, but by choosing so distant a control class, we allow for the possibility of more distant influence.

involved in a donor-donee relationship with core languages, in light of available geographical, ethnohistorical, and archeological data. If such a relationship is plausible, we attribute the high NBC score to 'linguistic admixture', i.e. the diffusion of linguistic features between one or more of the core languages and the high-scoring language, with the result that it exhibits a mixture of core and non-core features. If the distribution of high-scoring languages makes no sense geographically or otherwise, then Core and Periphery essentially fails to support the proposed linguistic area. Note that even when Core and Periphery is successful, the probabilistic nature of NBC, and the limitations of using phonological inventories as evidence for contact, may yield 'false positives', i.e. languages that exhibit high NBC scores despite there being no plausible basis for contact between those languages and core languages. Such languages should be discarded, leaving a phonological area that is defensible both quantitatively and qualitatively.

A two-core analysis, in contrast, produces a four-way division of languages (Core 1, Core 2, the control class, and the equivocal class). The naive Bayes classifier is trained on each core and the control set, and a three-way classification is then performed, yielding three scores for each language, which indicate the similarity of every language to each of the cores and to the control class. Those languages that obtain high scores for either of the two cores are then evaluated for plausibly having been in contact with a core language.

3 Dataset and Analytical Goals

3.1 SAPhon

The quantitative exploration of phonological areality presented in this paper is based on the analysis of the phonological inventories found in the South American Phonological Inventory Database, version 1.1.3 (SAPhon 1.1.3; Michael, Stark, and Chang 2013).[3] In this section we briefly describe the structure of the database and discuss particular decisions that we made in populating the database and preparing it for quantitative analysis.

SAPhon 1.1.3 incorporates 359 phonological inventories that have been harvested from published sources, or contributed by linguists currently working on the languages in question. This represents over 95% coverage of South American languages for which phonological descriptions are known to exist in one form or another.[4] The vast majority of inventories in the SAPhon database belong to living languages, but SAPhon also includes inventories from recently extinct languages, such as Chamicuro (Parker, 1991), as well as inventories based on the careful interpretation and re-analysis of older resources, as in the case of Cholón (Alexander-Bakkerus, 2005).

To facilitate quantitative analysis, the phonological inventory of each language is coded in a comprehensive phonological feature matrix, with languages along the y-axis and features along the x-axis,[5] with a column for every phoneme and contrastive supersegmental feature

[3]Available online: http://linguistics.berkeley.edu/∼saphon

[4]This estimate is based on Fabre's (2005) extensive bibliography of publications on South American languages, from which our list of languages is largely drawn.

[5]In this article, *feature* always refers to a feature of a language as a whole (such as the presence or absence of a particular phoneme in the phonological inventory) rather than to phonological features such as *labial* or *unrounded*.

(e.g. nasal harmony) attested in a South American language. Each phonological inventory is coded as a row of ones and zeros in the table, where the presence of a given segment for a given language is coded as '1' in the appropriate column, and absence coded as '0'. Exhaustively coding the inventories in this fashion relieves us of having to decide in advance which segments or contrasts are relevant to the exploration of areality.[6]

We now turn to a number of methodological and analytical issues posed by the nature of the data on which SAPhon is based. Since SAPhon draws data from a considerable range of published and unpublished sources, issues of heterogeneity in those sources pose challenges for development of the database, and for the analytical purposes to which we put that data.

The first type of heterogeneity we must contend with is the existence of multiple, sometimes incompatible, phonological descriptions for a given language. Since allowing multiple inventories for a given language poses significant analytical difficulties, we typically select one inventory from among the various proposed for a given language, preferring those given in work that present considerable supporting data and analytical detail, and prepared by authors with the substantial linguistic training. We also typically prefer inventories based on more recent work, on the grounds that recent work takes into account both previous analyses and new data. To improve the quality of our judgments in evaluating conflicting analyses we also consulted specialists in particular languages, language families, and known linguistic areas in South America. In cases where there is compelling evidence that the differences between inventories proposed for a given language are due to dialectal differences, we include both dialects in the database.

The second type of heterogeneity we contend with stems from the divergent ways in which different linguists treat the same empirical phenomena. In particular, different representational choices can lead to differences in the inventories given for different languages that do not reflect significant empirical or analytical differences between the inventories in question. To remove these spurious differences, we subject the coded inventories to phonological regularization prior to quantitive analysis (while leaving the original coding intact in SAPhon).

To understand the motivation for phonological regularization, and to demonstrate how it is carried out, it is useful to consider some concrete examples. We first discuss the treatment of non-high front vowels in Tupí-Guaraní (TG) languages. All TG languages exhibit two contrastive front vowels, given in descriptions as /i/ and either /e/ or /ɛ/ (and in one case, /ɪ/). In some of the languages where the symbol chosen to represent the front mid vowel phoneme is /e/, the description explictly indicates that this vowel is phonetically realized as [ɛ] (e.g. Kamaiurá; Seki 2000), and in other TG languages the symbol chosen for the mid front vowel phoneme is /ɛ/ (e.g. Nhandeva; Costa 2003). In addition, there are several TG languages where the symbol used to represent the non-high front vowel phoneme is /e/, but no information is provided as to its phonetic realization. Crucially, no TG language exhibits two contrastive front mid vowels: we never encounter a contrast between /e/ and /ɛ/.

For purposes of the analysis presented in this paper, we treat all TG languages as having the same two front vowels phonologically: a high front vowel /i/ and a mid front vowel /e/. We implement this regularization by recoding the phonemes given as /e/ or /ɛ/ in these languages as {e} (leaving the phonemes in the underlying database untouched). The result

[6]We thank Mark Donohue for sharing this very useful coding technique with us.

of this normalization is to recast the inventories of TG languages as exhibiting no difference in their front vowels for the purposes of our quantitative analysis. We extend our treatment of vowel systems of these types to all languages in our dataset. We treat all languages that exhibit only /i, e/ or /i, ɛ/ in their inventory of front vowels as exhibiting /i, {e}/. Of course, in languages in which /e/ and /ɛ/ do contrast, as in the majority of Macro-Ge languages, no regularization of these segments is carried out.

The preceding motivation for regularization stems from the fact that linguists vary in their choices of symbol to represent a given phoneme, but there are also methodological and typological motivations for regularization. First, given the phonetic similarity of [e] and [ɛ] it is likely that not all field linguists systematically distinguish the two phones in languages in which they do not contrast. Moreover, one would expect to often find non-contrastive variation between these two phones *within* such languages, based on a variety of phonetic and sociolinguistic factors. This means that using both /e/ and /ɛ/ to represent the single mid front vowel present in different languages suggests a greater degree of phonetic precision than is probably warranted.

Second, it is clear that in cases like that of Kamaiurá, mentioned above, linguists choose the phoneme label that represents not the precise phonetic value of it basic allophone (i.e. [ɛ]), but the typologically expected phoneme in that area of the phonemic space (i.e. /e/), as delimited by the phonemes with which it contrasts. As such, phoneme representations of this sort are not directly comparable to those which opt for a representation that is more phonetically faithful to the basic allophone of the phoneme (i.e. /ɛ/). Regularization resolves the discrepancy between these two principles for choosing phoneme symbols by converting all 'phonetically faithful' phoneme symbols to 'typologically unmarked' ones.

A second phenomenon that illustrates a more analytically profound motivation for regularization comes from the treatment of contrastive nasality in Southern American languages, as exemplified by the treatment of surface nasal vowels in Tukanoan languages. Briefly, surface nasal vowels are accounted for in two ways in these languages: as the surface realization of underlying nasal vowels, or as vowels that have undergone nasalization due to a morpheme-level nasalization feature that spreads nasalization onto the vowels in question (see, e.g. Gomez-Imbert 1993 and Stenzel 2004). The former analysis tends to be common in earlier works on languages of this family, and the morpheme-level nasal spreading analysis is typical of more recents works. In general, these appear to be two different ways to analyze materially similar distributions of nasal features, and we regularize the phonological systems in question by including the nasal counterparts of all oral vowels in the phonological inventories of languages that have been analyzed as exhibiting morpheme-level nasal spreading.

We list the regularization rules and discuss how they are applied to the SAPhon dataset in §A.1.

3.2 Applying Core and Periphery to Andean languages

In this paper we illustrate the Core and Periphery technique by using it to explore the Andean phonological area, and two phonological sub-areas within this larger area, the Southern Andean phonological area, and North-Central phonological area. In doing so we exemplify how the technique works when selecting cores of varying degrees of initial insightfulness.

The choice of the Andean highlands as a candidate core is an obvious one for areal specialists. Büttner (1983: 179), for example, observed that Southern Andean languages exhibit similar phonological inventories, and observations by linguists like Dixon (1999) regarding the phonological distinctiveness of the Andean and Amazonian regions are generally deemed uncontroversial (even if detailed evidence for such claims is not presented). Similarly, the Andes is generally recognized as a culture area which has, at different points in time, been dominated by large empires or polities, including Wari, Tiwanaku, and the Inkas (Steward and Faron 1959: 5-16).

In our first Core and Periphery analysis, we operate on a proposed Andean core that consists of the 23 languages located in the contiguous mountainous region of western South America above 2,000 meters in elevation, from Patagonia in the south to Ecuadorean Andes in the north. The 2,000 meter limit clearly separates Amazonian groups whose territory extends into the Andean piedmont from Andean peoples, and the northern limit of the Ecuadorean Andes corresponds to the extent of the Andean culture area as defined by the northernmost limit of Quechuan expansion. In the control set we include the 113 languages of the region beginning at 1500 km from the nearest Andean language, extending to the furthest limits of the continent. The remaining 223 languages in the the 1,500 kilometer-wide strip between the core and control languages make up the equivocal set of languages about which we posit nothing in advance.

Our second Core and Periphery analysis is motivated by the observation that although all Andean languages share features that distinguish them from non-Andean languages, the Southern Andean languages exhibit distinctive features not found in most Central or Northern Andean languages (e.g. a series of ejective consonants) while the latter group of languages exhibits distinctive features not generally found in the former group (e.g. retroflex affricates). These facts suggest that that it may be useful to treat the Andean area as comprising two subcores: a Southern core and a North-Central core. There are also sociohistorical facts that suggest that it may be useful to distinguish two cores in this way, namely, the fact that the Southern core corresponds roughly to extensions of the Tiwanaku empire (corresponding roughly to modern highland Bolivia) and that the North-Central core corresponds roughly to the extension of the Wari horizon (Isbell 2008). For the purposes of this analysis, we posit a Southern Andean core of 10 Andean languages south of the line that separates languages with ejectives from those without ejectives, with the remaining 19 Andean languages constituting the North-Central core.

4 Exploring language contact with a naive Bayes classifier

4.1 Overview

A naive Bayes classifier is a probabilistic model that classifies objects into K classes. Such a classifier is first trained on many examples, each labeled by a human expert with the class to which it belongs. Thereafter, when presented with a novel object, the classifier will report

with what probability the object belongs to each of the K classes.[7]

A common application of this technology is spam filtering. An e-mail account may receive dozens of unwanted messages every day, but a typical classifier is smart enough to put almost all of them into a spam folder, saving the user the trouble of ever having to look at them. In this application there are two classes: spam and non-spam. The classifier is trained on messages that it knows to be spam (such as those the user manually flags) and those it knows to be non-spam (such as those that the user does not flag after reading). This continuously-trained classifier is applied to incoming messages, and usually works very well.[8]

A naive Bayes classifier analyzes each object in terms of features that characterize it. In the case of e-mail, the features are the words that a message contains. When an incoming message is analyzed, each word will push the classification toward spam or non-spam, depending on how strongly the word is associated with spam or non-spam in the messages on which the classifier has been trained. A word such as *Viagra* is a strong indicator of spam, whereas most low-frequency words (such as *analysis* or *linguistics*) are weak indicators of non-spam. The classifier combines the evidence from each word to reach a verdict about the message as a whole.

Adapting this technology to classifying languages is straightforward: we train a classifier on *training languages* from K classes of interest, and use it to probabilistically classify a *test language*. (If there are multiple test languages, the classifier is run once for each test language.) The analyst provides a featural specification for each language, and a class label for each training language. As explained in §3.1, the featural specification is an encoding of the phonological inventory in which each feature is a phoneme or a suprasegmental feature that is either present or absent in the language. During training, the classifier calculates how strongly each phoneme is associated with each class. Then, in order to classify a test language, the classifier combines the evidence from each feature and assigns K probabilities to the test language — these are the probabilities that the test language belongs to each of the K clusters.[9]

[7]The origin of the naive Bayes classifier is obscure. It is a straightforward but non-trivial application of Bayes Theorem, which dates from the 18th century. Widely-used texts such as Mitchell (1997), Manning & Schütze (1999), Bishop (2007), and Jurafsky & Martin (2009) discuss it without commenting on its origin. Gale et al. (1992), cited in Manning & Schütze (1999), applied a naive Bayes classifier to the problem of word-sense disambiguation in natural language processing, without referring to it as such. This paper in turn cited Mosteller & Wallace (1963), a famous paper that used a naive Bayes classifier (also not referred to as such) to determine the authorship of twelve of the Federalist Papers. We suspect that naive Bayes classifiers had been used in diverse settings before the name itself caught on.

[8]The first academic papers to discuss Bayesian spam classifiers appeared in 1998 (Pantel & Lin; Sahami et al.) but it was an essay from 2002 titled *A Plan for Spam* that popularized the concept and made specific proposals to lower the rate of false positives to the point where the technology became usable (Graham, 2008).

[9]When we were divising the Core and Periphery technique, we tried using other kinds of classifiers besides NBC, such as support vector machines and logistic regression. The latter two are most often presented as classifying objects into two classes, but multiclass versions exist. All three classifiers are supervised learners, in that they classify based on examples provided by the analyst. In practice, NBC worked better than the other two methods, perhaps because it is a generative model, whereas the other two are discriminative models. Generative models tend to work better when the number of data points in the training data is relatively small, and the dimensionality of the data is large (Ng and Jordan 2001).

As for unsupervised analyses such as principal components analysis or multidimensional scaling, these are certainly useful as exploratory data analyses, and they may even identify potentially interesting linguistic

4.2 Two-way classification

A two-way classifier is a special case of a general K-way classifier that can be explained in simpler terms, so it will be discussed first. Training a classifier with two classes entails calculating a *feature weight* for each feature that expresses how strongly each feature is associated with each class. The weight for feature l is

$$u_l = \log\left(\frac{N_{1l}}{N_{2l}} \div \frac{N_1}{N_2}\right) \quad [provisional].$$

N_{1l} is the number of training languages in class 1 that have feature l, and N_1 is the total number of training languages in class 1. N_{2l} and N_2 are analogous quantities for class 2. The first ratio N_{1l}/N_{2l} is a comparison of the counts of feature l in the two classes. This is counterweighted by the second ratio N_1/N_2, which expresses the relative sizes of the two classes. The logarithm has the effect of causing the weight to be zero when the feature is neutral, positive when it is associated with class 1, and negative when associated with class 2.

One problem with this formula is that when any of the counts are zero, the feature weight u_l ends up at either positive or negative infinity. To prevent this, we inflate the counts by a small amount in order to regularize the result:

$$u_l = \log\left(\frac{\alpha + N_{1l}}{\alpha + N_{2l}} \div \frac{\alpha + \beta + N_1}{\alpha + \beta + N_2}\right).$$

For many applications it suffices to set $\alpha = \beta = 1/2$, but in our analyses we fit these parameters to the data, as explained in §A.4.

Strictly speaking, the above expression gives the feature weight for the presence of a feature. It is also necessary to calculate weights for the absence of a feature, via

$$v_l = \log\left(\frac{\beta + N_1 - N_{1l}}{\beta + N_2 - N_{2l}} \div \frac{\alpha + \beta + N_1}{\alpha + \beta + N_2}\right).$$

The main difference is that counts for the presence of a feature N_{1l} and N_{2l} have been replaced by counts for the absence of the feature $N_1 - N_{1l}$ and $N_2 - N_{2l}$. Once feature weights (for both present and absence features) have been calculated, the classifier is ready to classify.

For the test language, the classifier produces a score

$$s = \sum_{l=1}^{L} \begin{cases} u_l \text{ if feature } l \text{ is present in the test language,} \\ v_l \text{ if feature } l \text{ is absent in the test language.} \end{cases} \tag{1}$$

This score is a summation over all features (numbered from 1 to L) of feature weights, using u_l if feature l is in the test language, or v_l if feature l is not. The interpretation of the score is similar to that of the weights. A score of zero means that the test language is equally likely to belong to either class; a positive score means that it is more likely to belong to class 1; and a negative score means that it is more likely to belong to class 2.

areas. But since they are unsupervised, they cannot be directed by an analyst to examine an areal hypothesis that the analyst is specifically interested in. We thus omit mention of these analyses in discussing the Core and Periphery technique.

4.3 Underlying model and K-way classification

The previous section discussed naive Bayes classification from a procedural perspective. Now we engage in a brief discussion of the model that underpins the procedures. The model posits that our data, which comprise the training languages, the test language, and the labels for the training languages, were generated via a set of random events, which are as follows.[10]

- Randomly generate a *feature frequency* θ_{kl} for each feature l and each class k. This is the probability that a language in class k will have feature l. Feature frequencies are unobserved.
- Assign each language, including the test language, to one of K classes with probability $1/K$.[11] The assignments of the training languages are observed. The assignment of the test language is unobserved.
- For each language, endow it with feature l with probability θ_{kl}, where k is the class of the language. Each feature is generated independently of the others, conditional on k. The features that a language has are all observed.

With this as the premise, the classifier seeks to infer the class of the test language. It calculates, for each class k, the probability $f(k)$ that the test language would be generated by the feature frequencies $\theta_{k1}, \ldots, \theta_{kL}$ of class k. From this it infers that the test language belongs to class k with probability

$$p_k = \frac{f(k)}{f(1) + f(2) + \cdots + f(K)}. \tag{2}$$

If the feature frequencies were known, the formula for $f(k)$ would be straightforward:

$$f(k) = \prod_{l=1}^{L} \begin{cases} \theta_{kl} & \text{if feature } l \text{ is in the test language,} \\ 1 - \theta_{kl} & \text{if feature } l \text{ is not in the test language.} \end{cases} \qquad [\textit{provisional}]$$

The classifier is essentially calculating the likelihood of each choice $f(k)$ by taking the product of the probability of generating each feature value (present or absent) in the test language. We do not know what these feature frequencies are, but we can obtain some insight (albeit not exactly the right answer) by estimating the feature frequencies directly from the data via the formula $\theta_{kl} = N_{kl}/N_k$, where N_{kl} is the number of times feature l exists among training languages of class k, and N_k is the total number of training languages of class k. We get:

$$f(k) = \prod_{l=1}^{L} \begin{cases} \frac{N_{kl}}{N_k} & \text{if feature } l \text{ is in the test language,} \\ \frac{N_k - N_{kl}}{N_k} & \text{if feature } l \text{ is not in the test language.} \end{cases} \qquad [\textit{provisional}]$$

[10] When thinking about such models, W. C. finds it helpful to imagine a deity generating the data according to the procedure given, with some of the deity's choices hidden from view. What is not hidden comprises the data. On the basis of this data, we infer some of the hidden things.

[11] In a more sophisticated variant of this model, each language is assigned to class k with some probability π_k. The random variable π_k is not observed, and must be inferred from the data. In two-way classification, this adds a term such as $\log[N_1/N_2]$ to the score of the test language. When the number of training languages is fixed (as in our analyses) this term moves all scores up or down by a fixed amount, and does not alter any conclusions.

The correct equation, obtained by integrating over all possible values for all feature frequencies, is similar. If we posit a beta distribution prior for each feature frequency $\theta_{kl} \sim \text{Beta}(\alpha, \beta)$ we get the following expression for the likelihood:

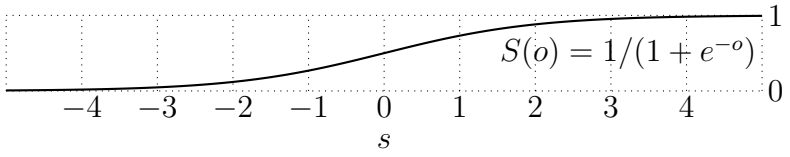
$$f(k) = \prod_{l=1}^{L} \begin{cases} \frac{\alpha+N_{kl}}{\alpha+\beta+N_k} & \text{if feature } l \text{ is in the test language,} \\ \frac{\beta+N_k-N_{kl}}{\alpha+\beta+N_k} & \text{if feature } l \text{ is not in the test language.} \end{cases} \tag{3}$$

This, along with Eq. 2, yields the probabilities $p_1, \ldots, p_K$ for K-way classification. Sections A.2 and A.3 restate the contents of this section more formally, and expand on it.

4.4 Probabilistic interpretation of NBC weights and scores

Eq. 1 in §4.2 describes how to compute an NBC score, which indicates how a test language is classified when $K = 2$. However, Eq. 2 in §4.3 derives a different indicator of classification: p_k, the probability with which the test language belongs to class k. How do these two kinds of indicators relate to each other?

It turns out that when $K = 2$, there is a straightforward mapping between the score s and p_1 (the probability that a test language belongs in class 1). They are related by the function $S(o) = 1/(1 + e^{-o})$. This function is plotted here.



In general, this sigmoid function translates from values that have a range of $[-\infty, \infty]$, to a probability, which has a range of $[0, 1]$. The argument o is a *log-odds*, so-called because it is the log of an odds ratio.

When $K = 2$, s is the log-odds that corresponds to the probability p_1, i.e. $S(s) = p_1$. Conversely we can apply the inverse function $S^{-1}(p) = \log p/(1-p)$ to p_1 to get s. According to Eq. 2, the probability that the test language belongs to class 1 has the form $p_1 = f(1)/[f(1)+f(2)]$. Converting this probability to a log-odds yields a score $s = \log f(1)/f(2)$, which expands to

$$s = \sum_{l=1}^{L} \log \begin{cases} \frac{\alpha+N_{1l}}{\alpha+N_{2l}} \div \frac{1+\alpha+\beta+N_1}{1+\alpha+\beta+N_2} & \text{if feature } l \text{ is in the test language,} \\ \frac{\beta+N_1-N_{1l}}{\beta+N_2-N_{2l}} \div \frac{1+\alpha+\beta+N_1}{1+\alpha+\beta+N_2} & \text{if feature } l \text{ is not in the test language.} \end{cases}$$

We see here that each feature value (present or absent) contributes in an additive way to the score. Comparing this to Eq. 1 shows how the feature weights were derived.

When $K > 2$, the structure of the computation in §4.3 does not result in additive feature weights for each feature, and since we do not compute feature weights before computing $p_1, \ldots, p_K$ for the test language, there is no distinct training stage. Also, since the classification results in more than two probabilities, it is no longer possible to indicate the classification of the test language with a single score. We can, however, convert each p_k into a log-odds and indicate the classification with K scores. When reporting the results of 3-way classification in §A.5.2, this is what we do.

4.5 Feature non-independence and the interpretation of results

The naive Bayes classifier's name derives from the naive assumption that the features in a language are generated independently, given the class of the language. In reality, however, the existence of one phoneme in an inventory is often strongly correlated with the existence of other phonemes in that inventory. For example, a language with /e/ often tends to have /o/, and vice-versa. Similarly, a language with an ejective stop at one place of articulation also tends to ones at other places of articulation. In this respect, there is a sense in which having multiple mid-vowels or having multiple ejective stops is a single 'fact' about a language, but a naive Bayes classifier will treat each fact of this sort as a set of multiple, independent facts. That is, the presence of mid-vowels is treated as two facts: the presence of /e/ and the presence of /o/. Similarly, the presence of ejective stops is treated as multiple facts about the presence of ejective stops at each place of articulation. This kind of multiple counting results in inflated scores, producing an effect of exaggerated certainty in classifying a language. All languages will suffer from this effect to some extent when undergoing classification, since feature non-independence (or, more colloquially, feature clumping) occurs frequently. Vowels of a given height, nasal vowels, long vowels, voiced stops, aspirated stops, ejective stops, etc.: each of these classes of sounds tend to be a clump. The presence or absence, in a test language, of any of these clumps exaggerates classification probabilities, rendering a literal probabilistic interpretation problematic. In our analyses, we sidestep this problem by disregarding the literal interpretation of the classification probabilities and reinterpreting them as measures of linguistic admixture. This interpretive leap calls for a careful explanation of admixture and how it is that admixture is not directly modeled by a naive Bayes classifier, to which we now turn.

By admixture, we refer to the phenomenon where the features of a language derive from two or more sources. This is analogous on some level to genetic admixture, where a person inherits certain genes from one parent and certain genes from the other; or, more abstractly, where a person inherits features from each of the K distinct ancestral populations in his or her ancestry. If we were to posit admixture for circum-Andean languages, one way to do this would be to posit two sources, one for the Andean core and one for the control class, described in §2. Each source is a hypothetical ancestral population in which there is a certain amount of linguistic diversity. A source does not have to be an actual set of precursor languages, though this is a good way to conceptualize it.[12] Each modern language descends from one or more sources.

A pure language derives its features from just one source. If, for example, all of the languages in the ancestral population have /p/, then a descendant of that source will also have /p/. If 60% of the languages in the ancestral population have /x/, then a descendant of that source will have /x/ with 60% probability. In general, the probability that a descendant has a feature matches the probability that a randomly-chosen constituent of the ancestral population has it.[13] Since there is some diversity in any ancestral population, one pure

[12]Formally, a source is represented by a bank of feature frequencies, one for each feature. Source k is represented by feature frequencies $(\theta_{k1}, \ldots, \theta_{kL})$, where θ_{kl} is the frequency of feature l among the languages of ancestral population k. This is formally identical to how a class is modeled in NBC; see first bullet in §4.3.

[13]This is formally identical to how languages are generated in NBC; see third bullet in §4.3.

descendant does not have to be identical to another, but it will in almost all cases be classified as descending from that population with little ambiguity, when all features are taken into account.

A mixed language derives its features from more than once source. If, for example, two ancestral populations are involved, then a certain fraction of the mixed language's features may derive from one, while the rest derive from the other.[14] It is often much more reasonable to posit that a language is mixed rather than pure. For instance, if a language has many distinctively Andean features and also many distinctively non-Andean features, then it is, on an intuitive level, best to posit admixture. (Just as, if a dog has many poodle features and many labrador features, one surmises that it is a mixed breed.) When a language is mixed, it is often possible to infer the extent to which it drew from each ancestral population. For circum-Andean languages such a statistic would indicate how core-like or control-like a language is.

However, as previously mentioned, the naive Bayes classifier is not a model of admixture. Rather unrealistically, every test language is assumed to be a pure language. Classification involves determining not *to what extent* the language descended from each ancestral population, but *with what probability*. Our interpretive leap is to use the latter as an indicator of the former. Unfortunately, the coarseness of this method of interpretation does not allow us to infer the absolute proportions of admixture in a language. If the model reports that a language belongs to class k with probability 0.7, that is by no means the same as indicating that 70% of the phonemes of the language are from the source identified with class k. We can only conclude that if p_k is higher for language X than for language Y, then X probably derives more of its phonemes from the source corresponding to class k than Y. This relativistic interpretive strategy, whatever its drawbacks, has the benefit that it allows us to work around the fact that feature clumping exaggerates classification probabilities and deprives them of their usual interpretation.

4.6 Details in applying the model

4.6.1 Feature culling

It was our assumption previously that feature clumps tend to be of limited size, so that there is a limit to how much a single clump can affect classification probabilities. In general this seems to be true, but there is a notable exception: rare features. Rare features tend to occur together in very large clumps. For instance, there are 112 features in our dataset that occur in exactly one language, but twelve of them occur in the same language, Paez, causing the classification of Paez to be greatly exaggerated. To prevent outcomes of this sort, we have discarded all features that occur five or fewer times in the training languages from our analyses. This amounted to discarding 225 of the 304 features in the dataset, leaving 79.

To be consistent with culling rare features, we have also culled near-universal features on the theory that when absences are rare, the absences can clump together just like rare features. We have discarded any feature that is present in all but five or fewer training languages. This resulted in discarding /t/, /k/, /i/, and /a/ from our analyses, leaving 75 features.

[14] For an example of a model that implements admixture in exactly this way, see Pritchard et al (2000).

4.6.2 Measuring admixture in a training language

When using a naive Bayes classifier to measure admixture, we should not exempt the training languages from scrutiny. However, it would prejudice the model for a test language to be a training language too. When we wish to apply the classifier to a training language in class k, we remove it from the set of training languages first. This lowers the count N_k in Eq. 3 by one, and lowers N_{kl} by one for each feature l present in the language. After this adjustment, classification proceeds as before.

4.6.3 Feature deltas

In a two-way classifier, the feature weights u_l and v_l give measures of the association between class 1 and, respectively, the presence or the absence of feature l. Having two weights for each feature is cumbersome if all we wish to know is the degree of association between a feature and a class. Using the formulas in §4.2 we define a measure called delta:

$$\delta_l = u_l - v_l = \log\left(\frac{\alpha + N_{1l}}{\beta + N_1 - N_{1l}}\right) - \log\left(\frac{\alpha + N_{2l}}{\beta + N_2 - N_{2l}}\right).$$

This measure is zero if the feature is neutral, positive if it is associated with class 1, and negative if it is associated with class 2. We can generalize delta to K-way classification by defining a set of K deltas for each feature:

$$\delta_{kl} = \log\left(\frac{h_{kl}}{1 - h_{kl}}\right) - \log\left(\frac{\sum_{j\neq k} h_{jl}}{\sum_{j\neq k} 1 - h_{jl}}\right),$$

where $h_{kl} = (\alpha + N_{kl})/(\alpha + \beta + N_k)$. The summations are from 1 to K, excluding k. The element δ_{kl} is zero if feature l is neutral with respect to class k, and positive or negative if feature l is positively or negatively associated with class k, respectively. A feature that is neutral with respect to all K classes will have zeros for all K deltas.

5 Results

5.1 Single Andean Core

The feature deltas (henceforth 'deltas') resulting from the NBC analysis of the Andean core are given in Figure 1. Positive deltas contribute to the classification of the languages that bear them as Andean, while negative deltas contribute to the classification of the languages that bear them as non-Andean. The presence of phonemes like /q/ and /ʎ/ in the inventory of a given language thus strongly contribute its classification as Andean, while the presence of /ɨ/ or /ã/ strongly contribute to its classification as non-Andean.

The deltas given in Figure 1 yield the distinctive phonological profile for the Andean core given in Tables 1-2. In these tables we (somewhat arbitrarily) select a delta of ± 2 ($p = 0.88$) or as the cutoff for segments whose presence or absence is strongly characteristic of the Andean core, and deltas between 1 and 2 ($0.73 < p < 0.88$) and -1 and -2 as the range for segments whose presence or absence, respectively, are moderately characteristic

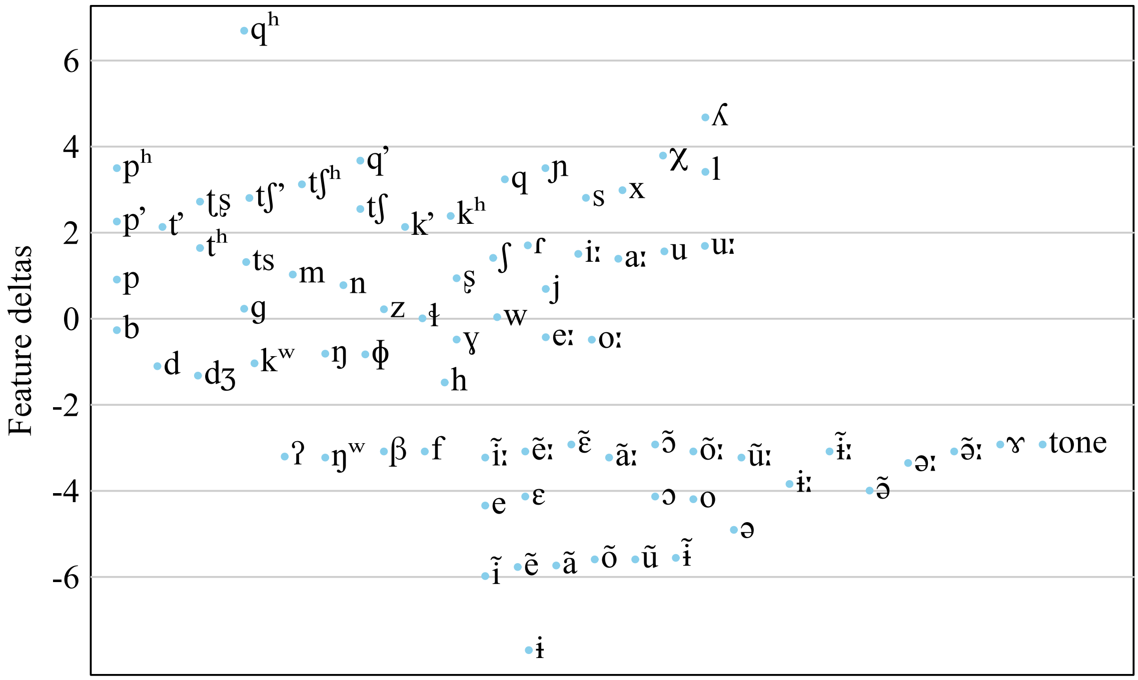


Figure 1: Feature deltas for single core Andean analysis

pʰ p'	tʰ **t'**				**kʰ k'**	**q qʰ q'**
		tʃ tʃʰ tʃ'	**ʈʂ**			
	s	ʃ			**x**	**χ**
				ɲ		
	l ɾ			**ʎ**		

iː		u uː
	aː	

Table 1: Distinctive phonemes of the Andean core languages (positive feature deltas)

	d	kʷ	**ʔ**
	dʒ		
β f			h
		ŋʷ	

ĩ ĩː	**ɨ ɨː ɨ̃ ɨ̃ː**	**ũ ũː**	**tone**
e ẽ ẽː ɛ ɛ̃	**ə əː ə̃ ə̃ː**	**o õ õː ɔ ɔ̃ ɤ**	
	ã		

Table 2: Distinctive absences in the Andean core languages (negative feature deltas)

of the Andean core. Strongly characteristic segments are printed in bold, while moderately characteristic ones are printed in normal weight.

The distinctive phonological profile of the Andean core languages, i.e. the set of segments that distinguish the Andean core languages from control languages in terms of either their presence and their absence, is large. The size of this distinctive phonological profile strongly suggests that the chosen core forms part of a phonological area distinguishable from the set of control languages.

The distinctive Andean consonantal profile can be positively characterized as exhibiting contrastive aspirated and ejective stops (a contrast found also in the postalveolar affricate), as well as a comparatively large number of affricates, fricatives, and liquids. Less common places of articulation that contribute positively to the profile include palatal (nasal and liquid) and uvular (stop and fricative). The consonantal profile can be negatively characterized as excluding the voiced alveolar stop and affricate, the labialized velar voiceless stop and nasal, voiced bilabial and voiceless labiodental fricatives, and the glottal stop and fricative. The distinctive Andean vocalic profile is positively characterized by /u/ and /iː, uː, aː/, but negatively by the absence of mid vowels, non-low central vowels, nasal vowels, and long versions of many of these vowels.

The NBC score of each language is given in Appendix A.5.1 and is plotted on a map in Figure 2, where the orange line is a smoothed version of the 2000 meter elevation contour. Languages with NBC scores near zero, and hence, difficult to classify as either Andean or non-Andean, appear in light gray. Higher NBC scores for a language correspond to greater red saturation, while the lower (i.e. negative) NBC scores correspond to greater blue saturation.

Inspection of Figure 2 reveals that a penumbra of languages with high NBC scores surrounds the posited Andean core, which is dense with languages with very high NBC scores. Following our discussion of the interpretation of NBC scores in §4.5, the high NBC scores of many of the languages in the circum-Andean peripheral region indicate that their phonological inventories much more closely resemble those of core Andean languages than those of the control languages, suggesting phonological admixture with Andean languages.

Inspection of Figure 2 also reveals that the NBC score tapers gradually with distance from the Andean core. The periphery of this phonological area is thus diffuse, and lacks clear boundary separating peripheral languages that are unambiguously members of the phonological area, such as Yanesha' [ame], from those that are clearly not, such as Aguaruna [agr]. If we consider any language with an NBC score greater than zero to be a candidate for membership in the area, and (somewhat arbitrarily) any language with an NBC score in the 95th percentile or greater to be a strong candidate for membership in the area, we obtain a partitioning of the periphery into 'strong' and 'weak' members of the linguistic area. These peripheral members of the Andean core mostly cluster geographically, as indicated below, and displayed in the more detailed maps in Figures 3-5.

ECUADOREAN FOOTHILLS

Strong: Cha'palaa [cbi] (Barbacoan)

Weak: Kamsá [kbh] (isolate)

HUALLAGA VALLEY

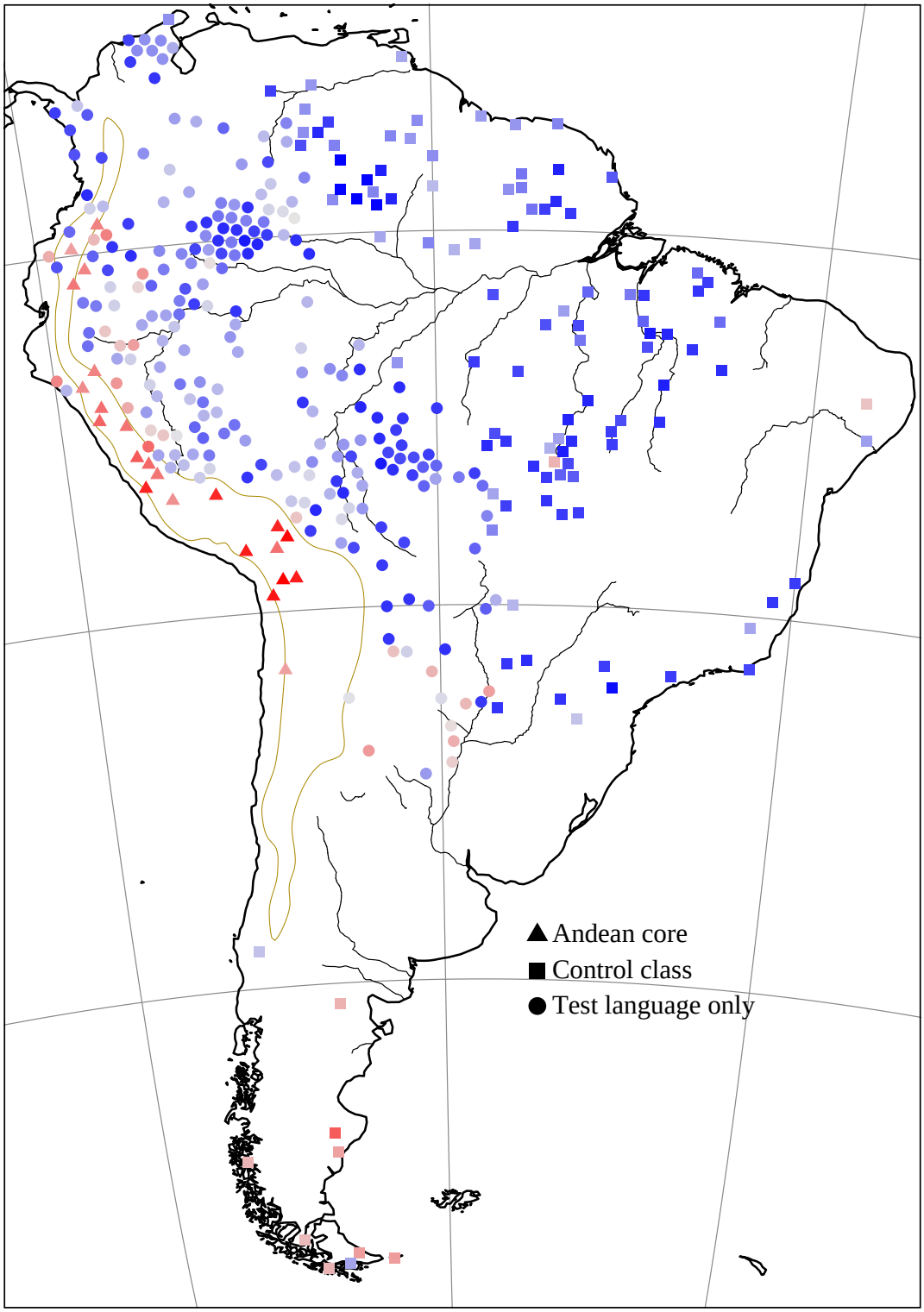


Figure 2: Languages of South America (two-way Andean core NBC scores)

Strong: Chamicuro [ccc] (Arawak), Cholón [cht] (isolate)

Weak: Shiwilu [jeb] (Cahuapanan), Candoshi [cbu] (isolate)

SOUTHERN PERUVIAN FOOTHILLS

Strong: Yanesha' [ame] (Arawak)

Weak: Ashéninka (Apurucayali [cpc] and Pichis [cpu] dialects) (Arawak)

CHACO

Strong: Vilela [vil] (isolate), Maká [mca], Chulupí [cag] (both Matacoan)

Weak: Wichí [mtp] (Matacoan), Toba Takshek [tob_tks], Toba Lañagashik [tob_lng], Mocoví [moc] (both Guaicuruan)

PATAGONIA

Strong: Ona [ona], Haush [ona_mtr], Puelche [pue], Tehuelche teh (

Weak: Northern Alacalufan [alc_nth], Central Alacalufan [alc_cnt], and Southern Alacalufan [alc_sth] (Alacalufan)

MISCELLANEOUS

Weak: Arabela (Zaparoan), Leko [lec] (isolate)

LOWLAND QUECHUAN LANGUAGES

Strong: Ferreñafe Quechua [quf], Inga (Jungle dialect) [inj], Napo Quichua [qvo], San Martín Quechua [qvs], Santiago del Estero Quechua [qus]

In several of these regions, such as the Ecuadorean foothills, the Huallaga River valley region, and the Southern Peruvian Foothills regions, significant contact between speakers of Andean languages and the relevant non-Andean languages is either known to have taken place (see, e.g. Adelaar and Muysken 2004: 411-413, Payne 1990: 1-10), or such contact is generally plausible, due to geographical proximity and the ubiquity of trade between adjacent highland and lowland regions.

Somewhat more surprising is the fact that Patagonia and the Chaco constitute an essentially contiguous phonological area with the southern Andes. Although there is evidence of trade between the Tiwanaku polity and the inhabitants of the Chaco between approximately 100 AD and 1100 AD (Angelo and Capriles 2000, Lecoq 2001, Torres 2006), it is unclear whether those relations would have been sufficiently intense to produce the kind of convergence we see between the southern Andean languages. Nevertheless, one Chacoan linguistic isolate (Vilela) and several Chacoan languages of the Matacoan and Guaicuruan families exhibit features strongly statistically associated with the Andean highlands, including ejectives, uvular consonants, and the palatal lateral. Evidence of contact between Patagonian and southern Andean peoples is even sparser, but the former languages likewise exhibit features characteristic of the Andean core languages. It should be noted that in Pre-Colombian times, the territory occupied by speakers of Patagonian languages was contiguous with that

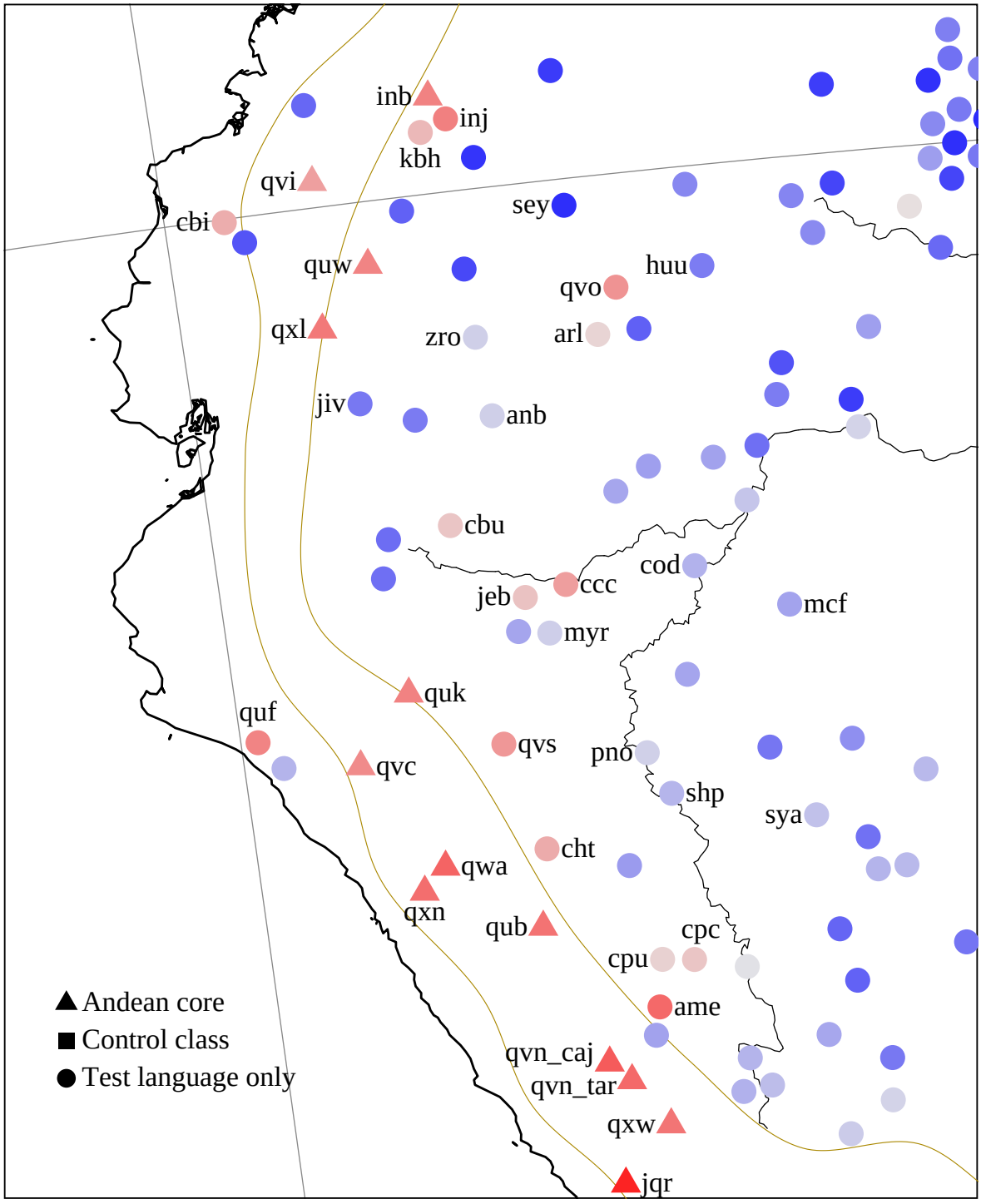


Figure 3: Languages of North Andes and Circum-Andean regions (two-way NBC scores)

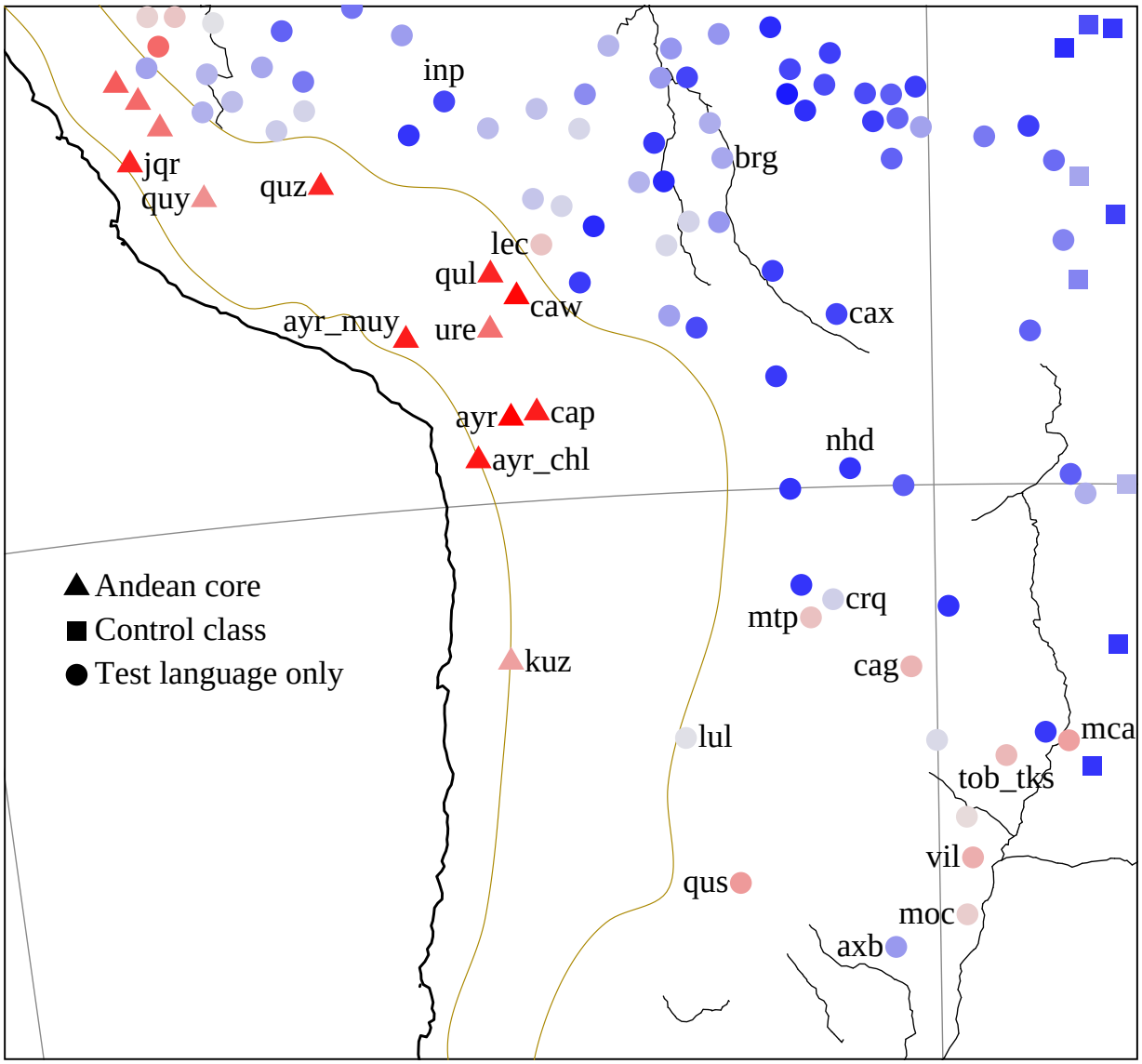


Figure 4: Languages of Central Andes and Circum-Andean regions (two-way NBC scores)

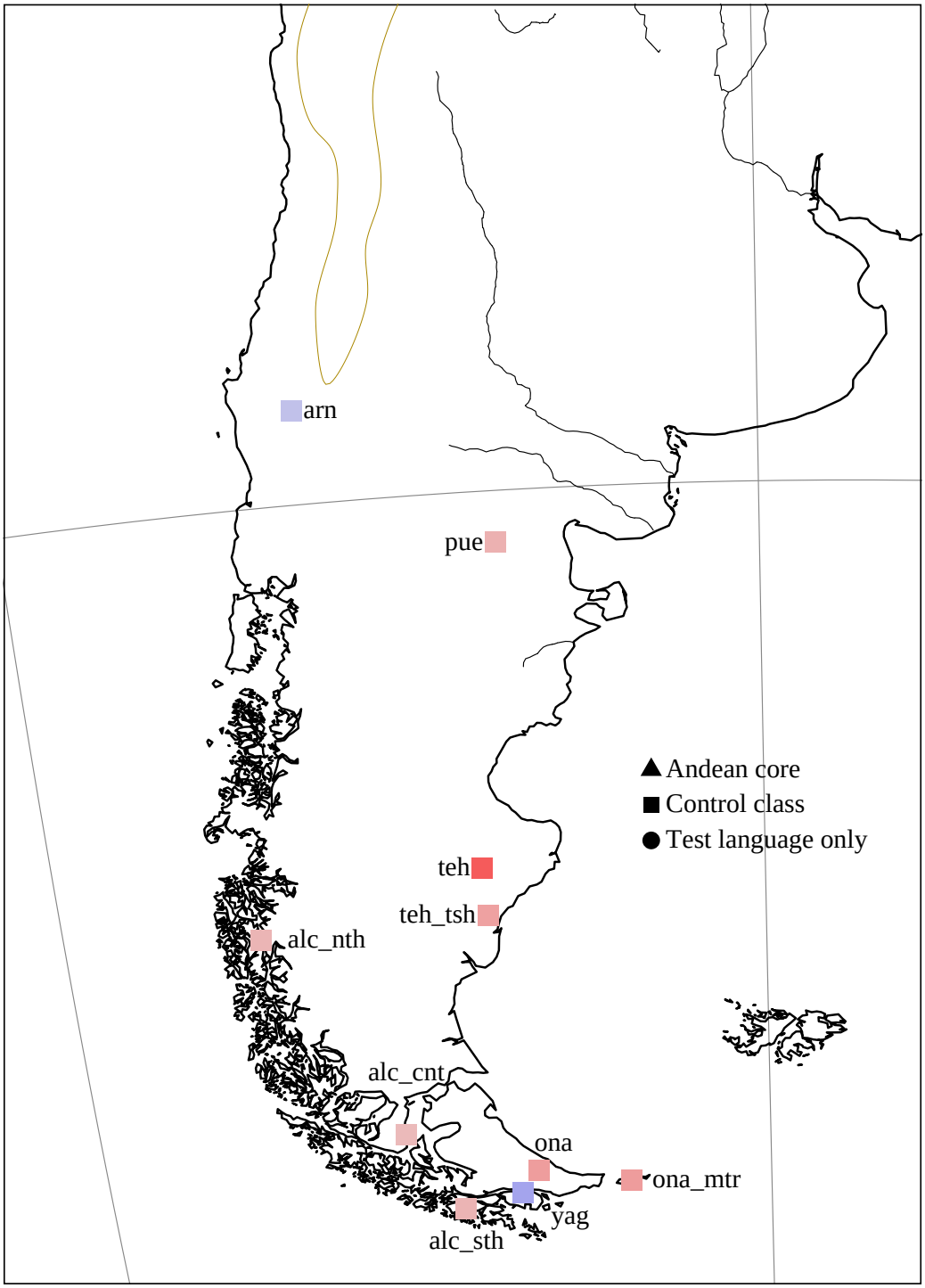


Figure 5: Languages of Patagonia (two-way NBC scores)

occupied by Chacoan peoples (Viegas 2005: 30), raising the possibility that the similarity between Andean and Patagonian languages arose not from direct contact between the languages of these two regions, but was mediated by Chacoan languages.

Admixture between circum-Andean languages and more northern languages of the Andean core appear to involve relatively local and recent convergence of these peripheral languages to Andean core ones, but the phonological convergence evident among Chacoan, Patagonian, and southern Andean languages does not exhibit clear directionality. The circumstances that led to this broader areal convergence are less clear, suggesting that much older, possibly multilateral, processes of phonological borrowing are responsible for the large scale phonological areality we see in the South American Cone.

In addition to the languages enumerated above, which comprise an essentially contiguous region with the Andean highlands, we find three other languages with positive NBC scores whose participation in the Andean and circum-Andean phonological area is dubious. These languages, listed below as OUTLIERS, obtain their high NBC scores due, in large part, to having aspirated stops and/or a palatal lateral in their phonological inventories. Given the probabilistic nature of NBC results, and the great distance of these languages from the Andean core, which renders historical contact with the Andean core languages extremely unlikely, we conclude that these languages simply bear a chance resemblance to the languages of the Andean core.

OUTLIERS:

Strong: Yawalapití [yaw] (Arawak)

Weak: Yucuna [ycn] (Arawak), Yaathe [fun] (Macro-Ge)

5.2 Southern and North-Central Cores

Although there are sound reasons for positing a single Andean core, there are also linguistic and socio-historical reasons to suspect that the Andean highlands exhibit linguistically-distinguishable sub-areas. For example, simple inspection of Andean phonological inventories reveals that southern Andean languages exhibit a three way aspirated/ejective/plain stop contrast and uvular consonants, whereas these features are rare or entirely absent in central or northern Andean languages. The social histories of the two regions are also quite different, with the southern Andes historically dominated first by the Tiwanaku polity and then by Aymaran peoples, who only partially penetrated into the central Andes (Adelaar 2012: 578). The central and northern Andes, in contrast, were dominated first by the Wari horizon and later by Quechuan peoples, who penetrated into the southern Andean region only shortly before the arrival of Europeans.

These observations motivate a dual core analysis that distinguishes Southern and North-Central cores, where the division is defined by a line that groups Jaqaru and Cuzco-Collao Quechua with all Andean languages to their south, and Ayacucho Quechua with all Andean languages to its north.[15] The deltas for the Southern core are given in Figure 6, and its

[15]This line was chosen to group together the Andean languages with a three-way contrast between plain, aspirated, and ejective stops.

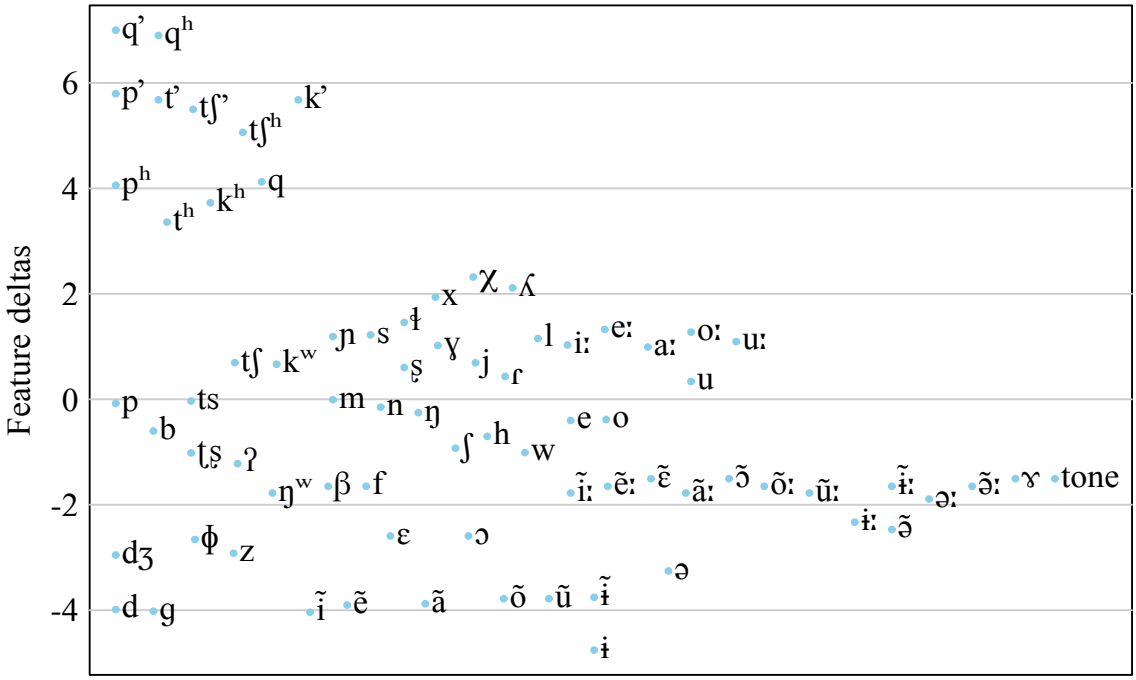


Figure 6: Southern Andean core feature deltas

distinctive phonological profile is given in Tables 3-4. The deltas for the North-Central core are given in Figure 7, and its distinctive phonological profile is given in Tables 5-6.

The deltas and distinctive profiles for the two cores exhibit significant differences, while sharing some charactersitics that distinguish them both from languages outside either cores. Consonants that positively characterize the distinctive phonological profiles of both Andean cores include /s l ʎ ɲ/, and those that negatively characterize both cores include /β f ŋʷ/. Vowels that positively characterize both cores include /iː uː aː/, while those that negatively

pʰ	**pʼ**		**tʰ**	**tʼ**					**kʰ**	**kʼ**		**q**	**qʰ**	**qʼ**
					tʃʰ	**tʃʼ**								
			s						x			**χ**		
			l	ɬ			**ʎ**							
							ɲ							
					iː		uː							
					eː		oː							
					aː									

Table 3: Distinctive phonemes of the Southern Andean core languages (positive feature deltas)

d				**g**	ʔ
		dʒ	ʈʂ		
ɸ **β** **f**	**z**	ʃ			
				ŋʷ	
w					
ĩ ĩː		**ɨ** **ɨː** **ɨ̃** ɨ̃ː		**ũ** ũː	tone
eː **ẽ** ẽː **ɛ** ɛ̃		**ə** əː **ə̃** ə̃ː		**õ** õː **ɔ** ɔ̃ ɤ	
		ã ãː			

Table 4: Distinctive absences in the Southern Andean core languages (negative feature deltas)

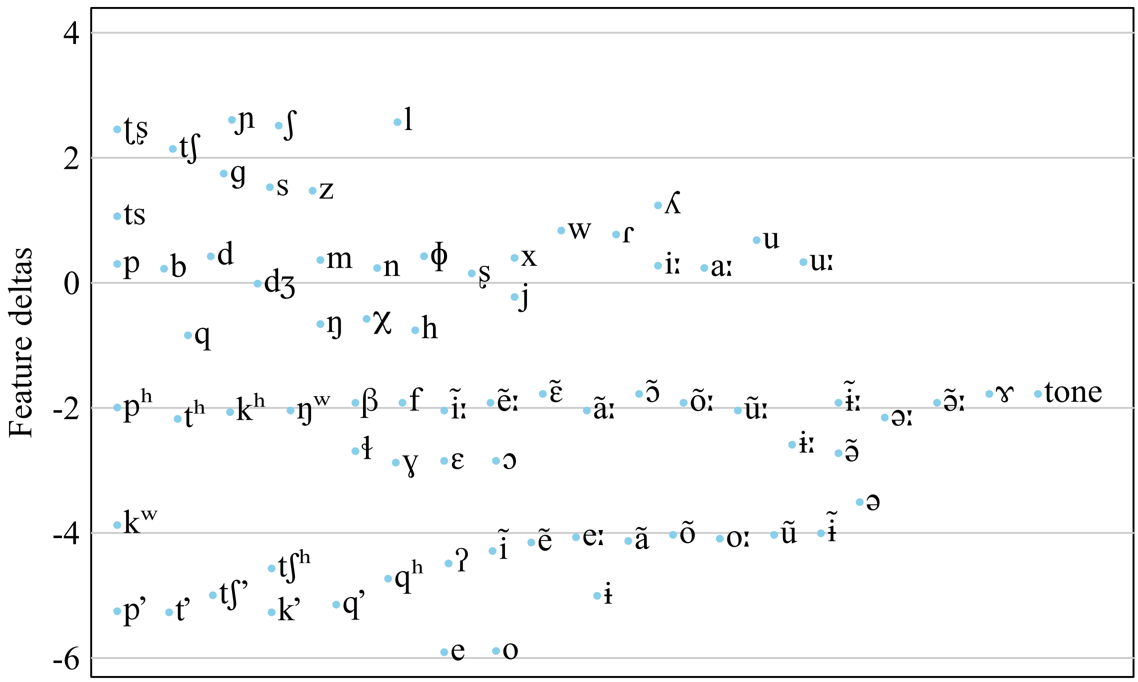


Figure 7: North-Central Andean core feature deltas

				g
ts	**tʃ**	**ʈʂ**		
s z	ʃ			
l			λ	
			ɲ	
	iː		uː	
		aː		

Table 5: Distinctive phonemes of the North-Central Andean core languages (positive feature deltas)

pʰ	**p’**	**tʰ**	**t’**				**kʰ**	**k’**	**kʷ**	q	**qʰ**	**q’**	**ʔ**
		tʃʰ	**tʃ’**										
β	f							**ɣ**					
			ɬ										
								ŋʷ					
		ĩ	**ĩː**		**ɨ**	**ɨː**	**ɨ̃**	**ɨ̃ː**			**ũ**	**ũː**	
	e	**eː**	**ẽ**	**ẽː**	**ɛ**	**ɛ̃**	**ə**	**əː**	**ə̃**	**ə̃ː**	**o**	**oː**	**õ**	**õː**	**ɔ**	**ɔ̃**	**ɤ**
							ã	**ãː**									

Table 6: Distinctive absences in the North-Central Andean core languages (negative feature deltas)

characterize them include the absence of mid vowels, non-low central vowels, and nasal vowels. Both cores also lack tone.

Other features yield large positive deltas for one core but negative ones for the other, serving to distinguish the cores not only from control languages, but from each other. Ejective and aspirated consonants yield positive deltas for the Southern Andean core, as do uvular stops and the lateral fricative /ɬ/, but negative deltas for the North-Central Andean core, whereas the converse holds for /tʂ g z ʃ/.

Yet other features yield large positive or negative deltas for one core, but do not yield a large deltas for the other. For the Southern core these include /x χ/ and the absence of /d ɸ w ɤ/. In contrast /ts/ is positively associated with the North-Central core profile and /ɣ/ negatively with it, but neither are salient for the Southern core. Turning to the vowels, both cores are negatively associated with central vowels, but the North-Central core exhibits a stronger negative association with short mid-vowels, as /e o/ are not significantly negatively associated with the languages of the Southern core.

The three-way NBC scores are plotted on a map in Figure 8. Whereas the two-way, single core results provide a one-dimensional measure of how core-like or control-like a given language is, the three-way, dual core results indicate to what degree a given language resembles the languages of either of the two cores, as well as the non-core languages. This we interpret as different degrees of admixture between the Northern-Central core, the Southern core, and the control class. The amount of yellow, red, and blue in the color of each dot encodes the proportion of those three components, respectively. In point of fact, there are no instances of significant admixture between just the two Andean cores, and all cases of significant admixture involve sizeable non-core components.

The qualitatively most significant result of the dual core analysis is that the majority of the languages of the Andean periphery identified in the single core analysis do in fact align with one of the two sub-cores, and do so in a geographically plausible manner. Languages which exhibit high Southern Core NBC scores are generally closer to the Southern Core than the North-Central Core, and conversely for languages with high North-Central NBC scores. The fact that Andean-like languages in the peripheral region pattern with the nearest core, rather than being randomly associated with either sub-core, indicates that

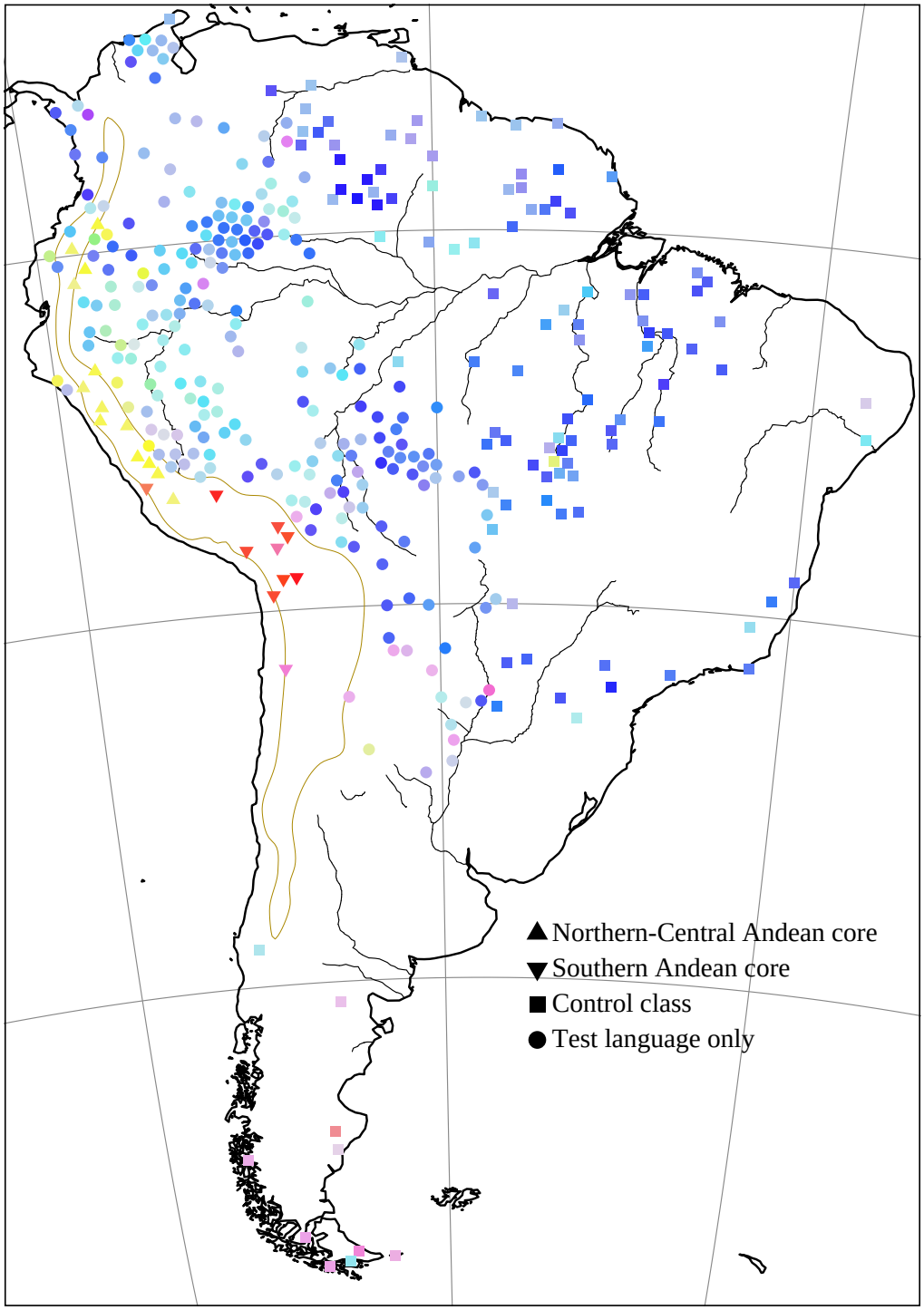


Figure 8: Languages of South America (three-way NBC scores)

convergence between circum-Andean languages and Andean languages is a relatively local effect attributable to language contact between the Andean languages of each sub-core and their circum-Andean neighbors.

The principal ways in which the results of the dual core analyses differ from the results of the corresponding single core analysis are to: 1) include more geographically proximal languages of this peripheral region in the phonological areas associated with the Andes; and 2) increase how strongly these peripheral languages pattern with the core languages, as listed below and displayed in Figure 9.[16] Kamsá [kbh], Shiwilu [jeb], Candoshi [cbu], for example, have gone from being weak members of the area to being strong members, and Panobo [pno] has gone from not being even a weak member of the area to being a strong member. Likewise, Andoa [anb], Sápara [zro], and Muniche [myr] went from being non-members to being weak members.

Languages of the North-Central Andean Periphery:

ECUADOREAN FOOTHILLS

Strong: Kamsá [kbh], Cha'palaa [cbi]

Weak: Andoa [anb], Sápara [zro]

HUALLAGA VALLEY

Strong: Shiwilu [jeb], Cholón [cht], Candoshi [cbu]

Weak: Muniche [myr]

SOUTHERN PERUVIAN FOOTHILLS

Strong: Yanesha' [ame], Panobo [pno]

Some languages, however, have ended up being excluded from membership in the North-Central area as a result of the dual core analysis, including Chamicuro [ccc], which was formerly a strong member of the (single core) Andean area, and Ashéninka (Apurucayali [cpc] and Pichis [cpu] dialects) and Arabela [arl], which were formerly weak members of the area. In the case of the Ashéninka varieties, they appear in the region of the tri-polar plot that suggests admixture of non-core features with both Southern and North-Central features, a result that is consistent with their location near the boundary of the Southern and North-Central cores. Chamicuro, in turn, just barely misses being a weak member of the North-Central area; although it exhibits strongly positive North-Central features like the retroflex affricate, and less strong ones, like the palatal lateral, it also exhibits mid-vowels and glottal stop, which are strongly negatively weighted for this core.

The languages in the periphery that pattern with the Southern sub-core are given below; plots of the languages of Patagonia and the Southern Andes and adjacent regions are given in Figures 10&11.

[16] That is, the NBC scores reflecting membership in the relevant cores increase for these languages in going from a single core to a two core analysis.

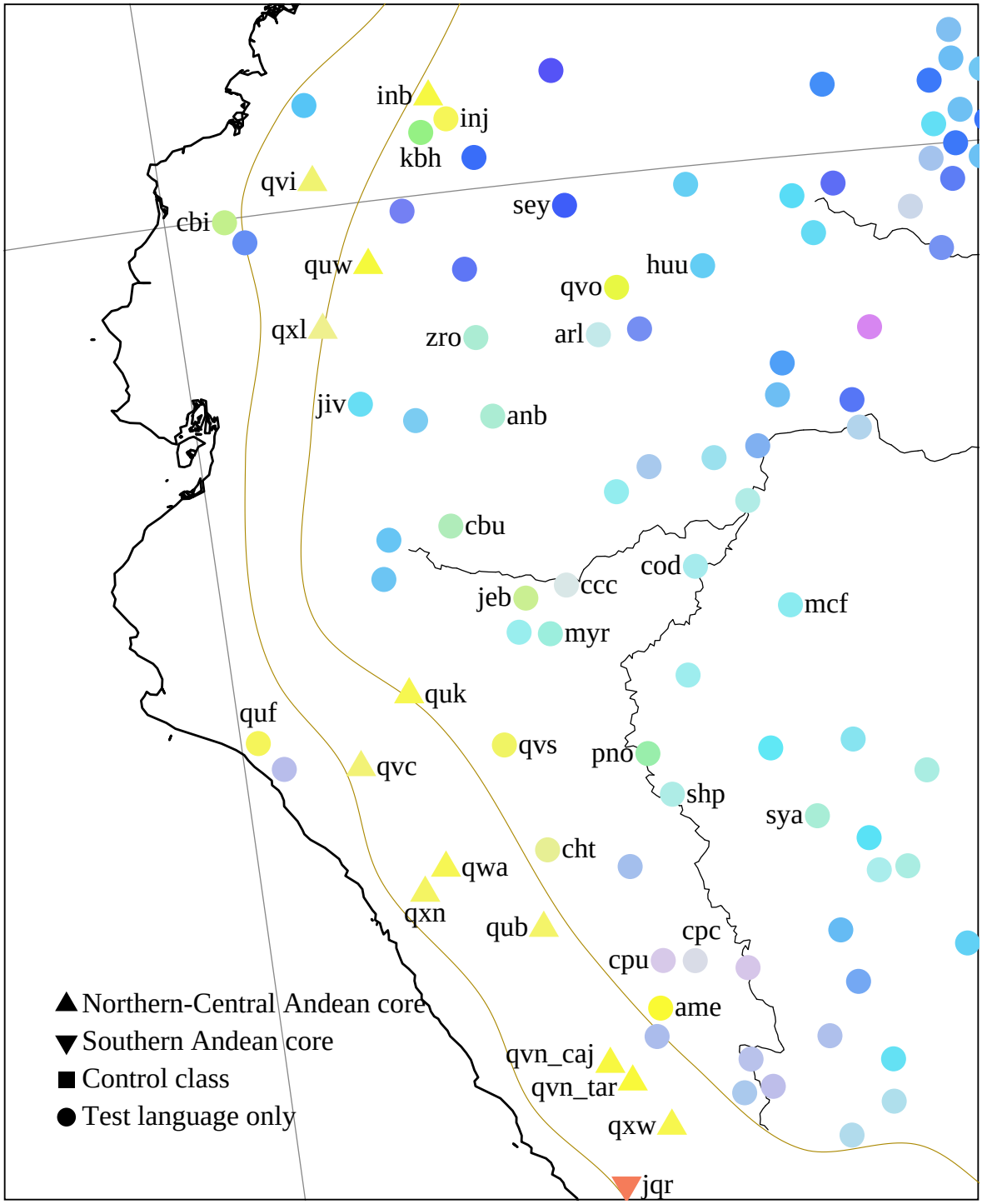


Figure 9: Languages of the Northern Andes and Circum-Andean regions (three-way NBC scores)

Languages of the Southern Andean Periphery:

CHACO

Strong: Maka [mca], Vilela [vil], Wichí (Mission de la Paz dialect) [mtp]

Weak (non-core admixture): Chulupí [cag]

PATAGONIA

Strong: Northern Alacalufan [alc_nth], Central Alacalufan [alc_cen], and Southern Alacalufan [alc_sth], Ona [ona], Haush [ona_mtr], Tehuelche [teh]

As in the case of the North-Central core, several languages have gone from being weak members of the single Andean core to being strong members of the Southern sub-core, including Wichí [mtp] and the Alacalufan languages [alc_nth, alc_cen, alc_sth]. Chulupí [cag] has experienced the opposite fate, and Puelche [pue] has gone from being a strong member of the area to being excluded, while Leko [lec], Toba Takshek [tob_tks], Toba Lañagashik [tob_lng], and Mocoví [moc] have gone from being weak members to being excluded. All of the languages excluded under a strict interpretation of NBC scores do, however, occupy regions near the zero log-odds line, a point we return to in the discussion in §6.

Convergence of Quechuan languages to the non-Quechuan languages of the Southern core is also suggested by the very high Southern Andean NBC scores obtained for Bolivian Quechua [qul_quh] and Cuzco-Collao Quechua [quz]. Other Quechuan languages have negative NBC scores for this core, indicating that Bolivian and Cuzco-Collao Quechua have been so significantly affected by contact with non-Quechuan Southern core languages that their phonological inventories pattern with those of these latter languages, rather than the Quechuan languages to which they are genetically related. Santiago del Estero Quechua is the next more non-North-Central-like Quechuan language, presumably due to the fact that its speakers migrated to the Argentinean pampas during the latest stages of the expansion of the Inka empire (Adelaar and Muysken 2004). At the same time, Jaqarú [jqr], a language belonging to the Aymaran language family, although still solidly patterning with Southern core languages, more closely resembles North-Central core languages than any of the Aymaran languages to which it is genetically related. Since Jaqarú is far to the north of the other Aymaran languages, and adjacent to Quechan languages of the North-Central core, its greater similarity to these languages reflects a history of contact with these Quechuan languages.

6 Discussion

Having demonstrated how Core and Periphery operates in exploring a particular set of hypotheses about linguistic areality, it is important to clarify a number of properties of this method. First, Core and Periphery is not a method which allows one to simply feed in data to an algorithm without any knowledge of the relevant languages or regions. Core and Periphery capitalizes on specialists' linguistic and non-linguistic knowledge both to generate fruitful initial hypotheses (i.e. the core and control language sets) and to interpret the

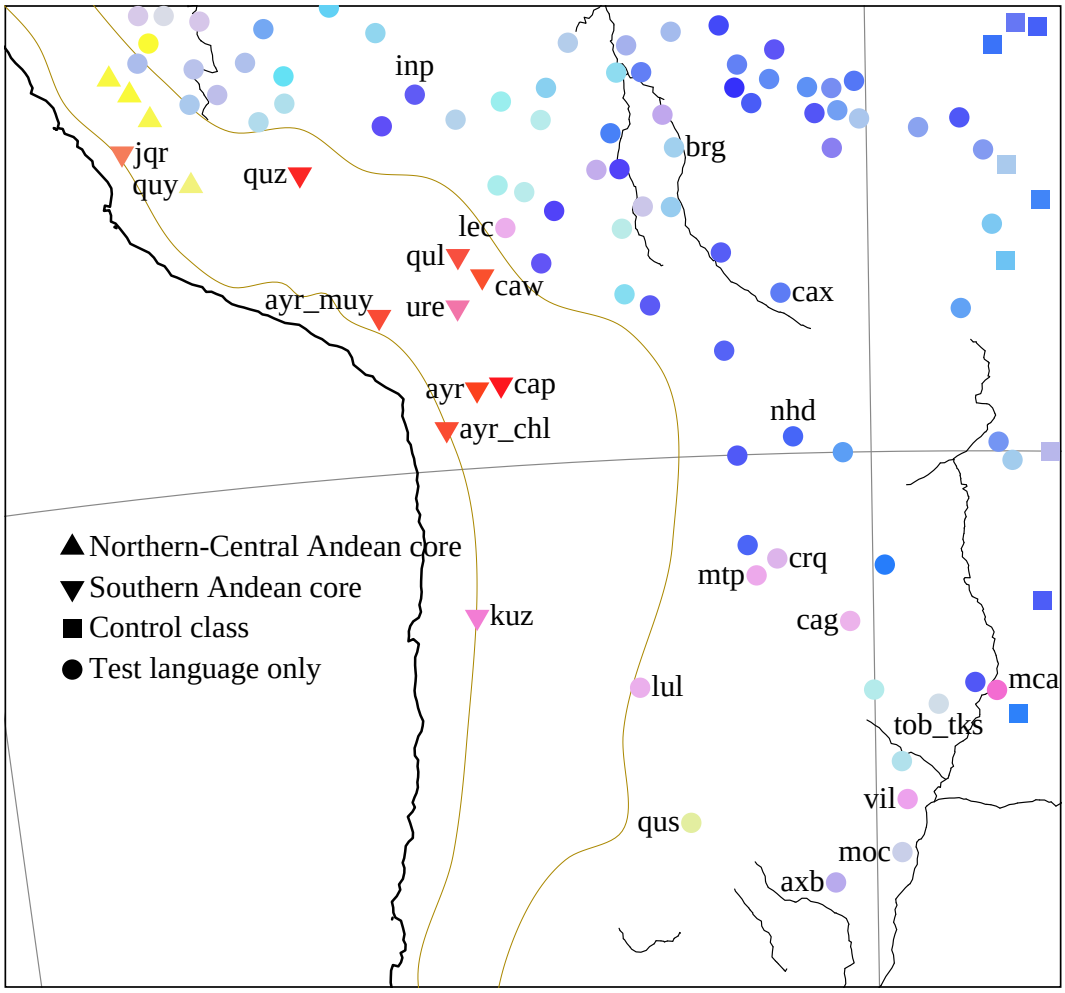


Figure 10: Languages of the Central Andes and Circum-Andean regions (three-way NBC scores)

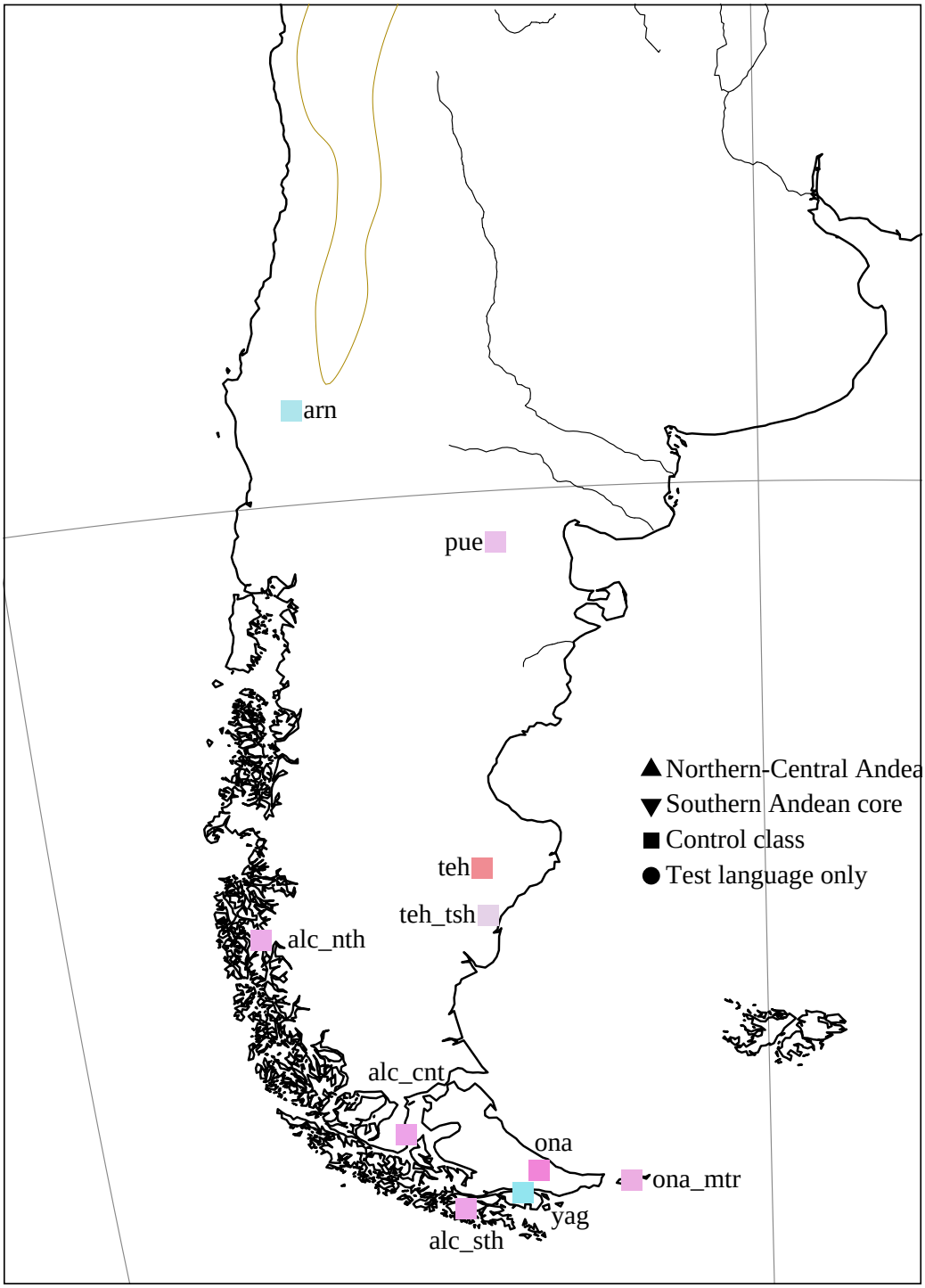


Figure 11: Languages of Patagonia (three-way NBC scores)

results, by evaluating the plausibility of language contact between core and peripheral languages with high NBC scores, and by filtering out languages whose high NBC scores can be attributed to genetic relatedness to core languages. What Core and Periphery provides is an intuitively straightforward means to explore large datasets for evidence of areality, by identifying features that distinctively characterize given groups of languages, and by providing a quantitatively explicit measure of similarity between a language and selected sets of languages.

It is important to note that despite the methodological priority of the proposed cores (i.e. that they are selected first), Core and Periphery makes no claims about the directionality of borrowing between core and periphery. There is no reason that, in principle, peripheral languages cannot be the original historical source of the segments that characterize a given phonological area. In these respects, then, Core and Periphery is a somewhat blunt tool: it is useful for identifying areality characterized by broad similarity among phonological inventories, but does not indicate the sources of the segments borrowed in the development of the phonological area.

Another crucial characteristic of Core and Periphery is that the NBC scores it generates are gradient, rather than categorial. On the one hand, this characteristic is a strength of the technique, since this feature makes the technique well-suited to analyzing areas with diffuse peripheries (see below). On the other hand, the gradient nature of NBC scores introduces a degree of arbitrariness if we seek to choose an NBC score to serve as a cutoff for identifying a languages as either core-like or control-like for the purposes of evaluating areality. To see the somewhat arbitrary nature of the cutoff, consider the NBC score cutoff of zero that we chose in this paper to distinguish core-like from control-like languages. This cutoff is actually somewhat conservative, as can be appreciated by considering languages that exhibit small negative NBC scores. These turn out to be predominantly located close to the edge of the Andean core (as defined by the 2000m contour line), an unexpected result if all languages with negative NBC scores were indeed wholly unaffected by contact with the relevant core languages. The significant clustering of circum-Andean languages in the region of small negative NBC scores would be explained, however, if these language were sufficiently affected by contact with Andean languages to raise their NBC scores from their 'pre-contact' scores, but not quite enough to give them positive NBC scores.[17] In short, contact with core Andean languages raises the NBC scores of the non-core languages in the circum-Andean periphery, but in some cases, not sufficiently for them to pass the zero NBC cutoff. The zero NBC cutoff is thus a relatively stringent criterion, in that it effectively excludes languages from the Andean core that were plausibly affected by contact with languages of the Andean core.

The issue of diffuseness of the periphery just raised suggests that it would be useful to consider the strengths and weaknesses of the Core and Periphery technique in terms of the areas it is well suited to studying. In qualitative terms, there are three important dimensions along which linguistic areas may vary: distinctiveness, core-homogeneity, and diffuseness. An area exhibits **distinctiveness** if some of its features occur at significantly higher or lower

[17]It would be useful for purposes of identifying the effects of contact in instances like this to have a measure of the degree to which a language diverges from related languages in the direction of neighboring languages to which it is not genetically related. The Relaxed Admixture Model, discussed in Chang and Michael (this issue), essentially provides a measure of this sort.

frequencies than the larger region of which it is a part. Distinctiveness is sometimes held to be definitional of a linguistic area (Aikhenvald 2006), and while distinctiveness does make an area conspicuous, it is clear that pairwise or multilateral borrowing of features may result in an area that does not stand out as having distinctive feature values (Thomason 2000). An area exhibits **core-homogeneity** if it is possible to identify a relatively contiguous set of languages within the area that are all very similar in the linguistic features being examined, either due to common descent or to long-term multilateral contact that has led to convergence to a shared linguistic profile. An area exhibits **diffuseness** if the area has a fuzzy boundary, i.e. a sizeble zone surrounding the core over which the concentration of core features gradually tapers off.

In light of these parameters, we can first observe that since the NBC scores that the Core and Periphery technique generates are continuous, it is well suited for the study of linguistic areas with diffuse boundaries. However, it should be noted that non-diffuse boundaries pose no problems for the method, and that such areas are amenable to being characterized by NBC score cutoffs that arises naturally from the data, rather than being completely arbitrary.

Next, Core and Periphery is significantly affected by the degree of homogeneity of a posited core. An important lesson from comparing the single core results and the dual core results is that combining two relatively homogeneous cores into a single less homogeneous core can reduce the efficacy of the NBC in identifying what are probably legitimate members of linguistic areas. In the single core analysis, we found that ejective and aspirated consonants had positive deltas; and that short mid vowels, the retroflex affricate, and the post-alveolar fricative had negative deltas. This means that the profile of the unified core resembled that of the Southern core more than the Northern-Central core. This explains the exclusion from the Andean periphery of some languages in the foothills and lowland regions proximal to the northern and central Andes, since languages in this region converged not to the profile of the unified core, but to that of the Northern-Central core, where ejective consonants, aspirated consonants, and mid vowels are uncommon, but the retroflex affricate and the post-alveolar fricative are common.

Finally, it is clear that Core and Periphery will only serve to identify *distinctive* areas, since in order for NBC to be able to classify languages as core-like or control-like, core and control languages have to be sufficiently different. However, as Chang & Michael (this issue) show with a different technique, there do in fact exist 'mosaic areas' in which language contact has led to borrowing among languages, without resulting in a core with distinctive phonemes or even a high degree of homogeneity.

The Core and Periphery technique is thus effective for identifying and delimiting linguistic areas that are distinctive and relatively core-homogeneous, and easily handles diffuse areas (though non-diffuseness poses no difficulties). And as we saw, the two Andean subcores, along with their respective proximal circum-Andean regions, each constitute distinctive, diffuse, and relatively core-homogeneous phonological areas.

Note that there is no intrinsic geographical or scale-based limitation to the technique. Due the nature of the dataset and the empirical questions that animated our interest in areality, this paper examined areality within a continent, selecting a core in a subregion of the continent and a set of control languages in another sub-region of the continent. These practical considerations are incidental, however, and nothing other than the rarity of suitable datasets prevents Core and Periphery to be applied to considerably larger regions (e.g. with

entire continents serving as cores), or using languages from considerably more distant regions (e.g. on the opposite sides of major oceans) as control languages.

7 Conclusion

We have presented in this paper a method for exploring linguistic areality that makes use of a naive Bayes classifier to quantify the similarity between candidates members of a linguistic area and a posited set of 'core' members of the area, with respect to the features that distinguish those core languages from a set of 'control' languages deemed extremely unlikely to be members of the area being explored. Versions of this 'Core and Periphery' technique were developed for both single core and multiple core analyses and applied to a concrete empirical test case: phonological areality in the South American Andes and surrounding lowland areas. This application resulted in the identification of several areas in which non-Andean languages show convergence with Andean languages generally (in the case of the single core analysis) and more locally to Southern Andean and North-Central subcores (in the case of the two core analysis). These results confirm that Core and Periphery is a useful exploratory tool, since they are generally plausible, in light of our knowledge of contact between Andean and neighboring non-Andean societies, yet at the same time identify instances of convergence which have previously gone unnoticed.

This initial application of Core and Periphery suggests a number of directions for the future development of this approach. First, since the similarity measure used by Core and Periphery relies on abstract features that impose few restrictions on the kind of linguistic traits that serve as the basis of classification, this method can be extended to morphological, syntactic, or even pragmatic, traits, as long as clumps of non-independent traits are small relative to the total number of traits. The application of Core and Periphery to non-phonological datasets is an obvious next step.

Second, as discussed in §4.5, statistical non-independence between phonological segments frustrates attempts to interpret NBC scores as simple probabilities of core vs. control group membership, and consequently hamstrings our ability to employ Core and Periphery as a quantitative test for areality. Though this property does not pose problems for its use as an exploratory tool, the power of Core and Periphery would be significantly enhanced by directly tackling the non-independence problem, suggesting another area for future research.

The spatial dimension of Core and Periphery, which at this point manifests only informally in linguists' identification of core and periphery areas, and through qualitative observations about the spatial distribution of NBC scores, is another area in which the technique could be usefully enriched. The analysis of NBC scores can be coupled to spatial statistical measures, for example, to yield a quantitative perspective on the gradience of fuzzy-edged linguistic areas. And rather than charactering distance solely in idealized Euclidean terms, it could be cast in more realistic movement cost measures, reflecting an understanding of space more finely attuned to human interactions with the environment.

8 References

Adelaar, Willem. 2012. Languages of the Middle Andes in areal-typological perspective: Emphasis on Quechuan and Aymaran. In Lyle Campbell and Veronica Grondona (eds.), The Indigenous Languages of South America: A Comprehensive Guide, pp. 575-624. Berlin: Walter de Gruyter.

Adelaar, Willem and Pieter Muysken. 2004. Languages of the Andes. Cambridge University Press.

Aikhenvald, Alexandra. 2006. Grammars in contact: A cross-linguistic perspective. In Alexandra Aikhenvald and R.M.W. Dixon (eds.), Grammars in contact: A cross-linguistic perspective, pp. 1-66. Oxford University Press.

Angelo, Dante and José Mariano Capriles. 2000. La importancia de las plantas psicotrópicas para la economia de intercambio y relaciones de interacción en el altiplano sur Andino. Complutum 11: 275-284.

Bishop, Christopher M. 2007. *Pattern Recognition and Machine Learning*. Springer.

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

Campbell, Lyle, Terrence Kaufman, and Thomas C. Smith-Stark. 1986. Meso-America as a linguistic area. Language 62: 530570.

Dixon, R.M.W. 1999. Introduction. In R.M.W. Dixon and Alexandra Aikhenvald (eds.), The Amazonian Languages. Cambridge University Press.

Fabre, Alain. 2005. Diccionario etnoling"uístico y guía bibliográfica de los pueblos indígenas suramericanos. Accessed: http://www.ling.fi/Diccionario%20etnoling.htm (June 1, 2011)

Gale, William A., Kenneth W. Church & David Yarowsky. 1992. A method for disambiguating word senses in a large corpus. *Computers and the Humanities* 26(5-6). 415–439.

Gomez-Imbert. 1993. Problemas en torno a la comparación de las lenguas tucano-orientales. In María Luisa Rodriguez de Montes (ed.), Estado actual de la clasificación de las lenguas indígenas de Colombia, pp. 235-267. Bogotá : Instituto Caro y Cuervo.

Graham, Paul. 2008. *Hackers & Painters: Big Ideas from the Computer Age*. O'Reilly Media.

Isbell, William. 2008. Wari and Tiwanaku: International identities in the Central Andean Middle Horizon. In Helaine Silverman and William Isbell (eds.), Handbooks of South American Archeology, pp. 731-760. Springer.

Jurafsky, Dan & James H. Martin. 2009. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech*. Pearson Prentice Hall 2nd edn.

Lecoq, Pierre. 1991. Sel et archeologie en Bolivie. De quelques problèmes relatifs à la occupation pr´hispanique de la cordillère Intersalar (Sud-Ouest bolivien). PhD dissertation: Universite de Paris 1.

Manning, Christopher D. & Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press.

Michael, Lev, Will Chang, and Tammy Stark (compilers). 2013. South American Phonological Inventory Database v.1.1.3. Available online: http://linguistics.berkeley.edu/ saphon/en/

Mitchell, Tom M. 1997. *Machine Learning*. McGraw Hill.

Mosteller, Frederick & David L. Wallace. 1963. Inference in an authorship problem. *Journal of the American Statistical Association* 58(302). 275–309.

Muysken, Pieter. 2008. Conceptual and methodological issues in areal linguistics. In Pieter Muysken (ed.), From linguistic areas to areal linguistics, pp. 1-24. Amsterdam: John Benjamins

Ng, Andrew Y. & Michael I. Jordan. 2001. On Discriminative vs. Generative Classifiers: A Comparison of Logistic Regression and Naive Bayes. In: Dietterich TG, Becker S, Ghahramani Z (eds) Proceedings of the Conference on Neural Information Processing Systems. MIT Press, MA, pp. 841848

Pantel, Patrick & Dekang Lin. 1998. Spamcop: A spam classification & organization program. In *Papers from the AAAI Workshop: Learning for Text Categorization (Technical Report WS-98-05)*, 98–105. AAAI.

Payne, Doris L. 1990. Introduction. In Doris L Payne, Amazonian linguistics: Studies in lowland South American languages, pp. 1-10. Austin: University of Texas Press.

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

Raiffa, Howard & Robert Schlaifer. 1961. *Applied Statistical Decision Theory*. Harvard University Press.

Sahami, Mehran, Susan Dumais, David Heckerman & Eric Horvitz. 1998. A Bayesian approach to filtering junk e-mail. In *Papers from the AAAI Workshop: Learning for Text Categorization (Technical Report WS-98-05)*, 98–105. AAAI.

Stenzel, Kristin. 2004. A reference grammar of Wanano. University of Colorado, PhD dissertation.

Steward, Julian and Louis Faron. 1959. Native peoples of South America. New York: McGraw-Hill.

Thibaux, Romain & Michael I. Jordan. 2007. Hierarchical Beta Processes and the Indian Buffet Process. In Marina Meila & Xiaotong Shen (eds.), *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics (AISTATS-07)*, vol. 2, 564–571. Journal of Machine Learning Research.

Thomason, Sarah. 2000). Linguistic areas and language history. In Dicky Gilbers, John Nerbonne, Jos Schaeken (eds.) Languages in Contact, pp. 311-327. Amsterdam: Rodopi.

Torres, Constantino Manuel. 2006. Anadenanthera: Visionary plant of ancient South America. Routledge.

Viegas, J. Pedro Barros. 2005. Voces en el viento: Raíces lingüísticas de la Patagonia. Buenos Aires: Mondragon Ediciones.

A Appendix

A.1 Inventory regularization rules

We regularize phonological inventories in a procedural way, with a series of replacement rules, as listed below. With consonants, we replace every sound on the left with the sound on the right, unless the sound on the right is already in the inventory.

$$\text{ⱡʲ} \rightarrow \text{ʎ}$$
$$\text{ʝ} \rightarrow \text{ʒ}$$
$$\text{ʋ} \rightarrow \text{β}$$

With vowels, the procedure is similar, but more elaborate. As with consonants, we try to replace each sound on the left with the sound on the right, but we also try to replace any phoneme that *contains* the sound on the left by replacing the matched character with the sound on the right. For example, the rule ʉ → ɨ will cause us also to try ʉi → ɨi, ʉː → ɨː, ʉ̃ → ɨ̃, etc. These replacements are not carried out if the sound on the right already exists in the inventory, either as itself or as part of another phoneme. The vowel replacement rules are as follows.

$$\text{ʉ} \rightarrow \text{ɨ}$$
$$\text{ɯɪ} \rightarrow \text{ɨ}$$
$$\text{ə} \rightarrow \text{ɨ}$$
$$\text{ʌ} \rightarrow \text{ɨ}$$
$$\text{ʊ} \rightarrow \text{u}$$
$$\text{ɔ} \rightarrow \text{o}$$
$$\text{ɤ} \rightarrow \text{o}$$
$$\text{ɪ} \rightarrow \text{i}$$
$$\text{e} \rightarrow \text{i}$$
$$\text{ɛ} \rightarrow \text{e}$$
$$\text{ɑ} \rightarrow \text{a}$$

These rules are applied in order. For example, if ʉ has been replaced with ɨ via the first rule, it is then no longer possible for ɯɪ to be replaced with ɨ via the second rule.

Finally, if the language has nasal harmony, we add the nasal version of each oral vowel to the inventory.

A.2 Model and data

A naive Bayes classifier is underpinned by a probabilistic generative model. What follows is a formal description of the model and data, as used in our analyses. The data can be viewed as having three parts.

- An $N \times L$ binary feature matrix X, where X_{nl} denotes the absence (0) or presence (1) of feature l in the phonological inventory of training language n. N is the number of training languages and L is the number of features.
- A set of labels for each training language $Y = (Y_1, \ldots, Y_N)$, where $Y_n \in \{1, \ldots, K\}$ denotes the class of language n. These labels are supplied by the analyst before analysis begins, as are the total number of classes K.
- A set of binary features for the test language $X_0 = (X_{01}, \ldots, X_{0L})$, where X_{0l} denotes the absence (0) or presence (1) of feature l in the phonological inventory of the test language.

The generative procedure that underlies the data is summarized as follows.

- For each class k and each feature l, generate a feature frequency via a beta distribution: $\theta_{kl} \sim \text{Beta}(\alpha, \beta)$.
- For each language pick a label from a categorical distribution: $Y = k$ with probability π_k. This is done for all training languages and the test langauge as well, but the outcome is observed for just the training languages.
- For each language n, generate each feature l via a weighted coin toss: $X_{nl} \sim \text{Bernoulli}(\theta_{Y_n l})$. The subscript on θ refers to the class denoted by Y_n (the class of language n) and the feature denoted by l.

We posit that θ_{kl} is generated by a beta distribution for mathematical convenience: the beta distribution is the *conjugate prior distribution* (Raiffa and Schlaifer 1961) of the Bernoulli distribution. Also favoring this choice is the fact that feature frequencies seem empirically to form such a distribution.

In our analyses, we set $\pi_k = 1/K$, but in other contexts it may make more sense to set $\pi_k = N_k/N$, where N_k is the number of training languages in class k. Settings for α and β will be discussed in §A.4.

A.3 Inference

The purpose of the model in §A.2 is to make it possible to infer the class of the test language. This entails computing p_k, the probability that the test language is in each class k, conditioned on the data. In standard notation, this is $\mathbb{P}(Y_0 = k \mid X_0, X, Y)$. By Bayes' Theorem,

$$\mathbb{P}(Y_0 = k \mid X_0, X, Y) = \frac{\mathbb{P}(X_0 \mid Y_0 = k, X, Y)\ \mathbb{P}(Y_0 = k \mid X, Y)}{\mathbb{P}(X_0 \mid X, Y)}.$$

Writing $f(k)$ for $\mathbb{P}(X_0 \mid Y_0 = k, X, Y)$, this expands to

$$\mathbb{P}(Y_0 = k \mid X_0, X, Y) = \frac{f(k)\pi_k}{f(1)\pi_1 + \cdots + f(K)\pi_K}.$$

Since the L features of the test language are generated independently, conditioned on the class of the test language, we have $f(k) = \prod_{l=1}^{L} f_l(k)$, where $f_l(k) = \mathbb{P}(X_{0l} \mid Y_0 = k, X_{\cdot l}, Y)$ and $X_{\cdot l}$ denotes elements of the lth column of X. Then,

$$f_l(k) = \frac{\mathbb{P}(X_{0l}, X_{\cdot l} \mid Y_0 = k, Y)}{\mathbb{P}(X_{\cdot l} \mid Y)}.$$

Note that X_{ml} and X_{nl} are fully independent if language m and language n belong to different classes. Thus the denominator factorizes into

$$\mathbb{P}(X_{\cdot l} \mid Y) = \prod_{j=1}^{K} \mathbb{P}(X_{I_j l} \mid Y),$$

where I_j denotes the set of test languages belonging to class j, and $X_{I_j l}$ denotes the entries indexed by I_j in the lth column of X. The numerator is identical, except in the factor corresponding to class k. After casting out factors that are identical in the top and bottom, we are left with

$$f_l(k) = \frac{\mathbb{P}(X_{0l}, X_{I_k l} \mid Y_0 = k, Y)}{\mathbb{P}(X_{I_k l} \mid Y)}.$$

Writing $f_{\theta_{kl}}(z)$ for the density function of $\theta_{kl} \sim \text{Beta}(\alpha, \beta)$, this expands to

$$\begin{aligned} f_l(k) &= \frac{\int_0^1 \mathbb{P}(X_{0l}, X_{I_k l} \mid \theta_{kl} = z, Y_0 = k, Y) f_{\theta_{kl}}(z) dz}{\int_0^1 \mathbb{P}(X_{I_k l} \mid \theta_{kl} = z, Y) f_{\theta_{kl}}(z) dz} \\ &= \frac{\int_0^1 \left[\prod_{n \in \{0\} \cup I_k} z^{X_{nl}} (1-z)^{1-X_{nl}} \right] z^{\alpha} (1-z)^{\beta} dz}{\int_0^1 \left[\prod_{n \in I_k} z^{X_{nl}} (1-z)^{1-X_{nl}} \right] z^{\alpha} (1-z)^{\beta} dz} \\ &= \frac{\int_0^1 z^{\alpha + N_{kl} + X_{0l}} (1-z)^{\beta + N_k - N_{kl} + 1 - X_{0l}} dz}{\int_0^1 z^{\alpha + N_{kl}} (1-z)^{\beta + N_k - N_{kl}} dz}, \end{aligned}$$

where N_{kl} is the number of training languages in class k with feature l and N_k is the total number of training languages in class k. We state the results of these integrals in terms of gamma functions before simplifying:

$$\begin{aligned} f_l(k) &= \frac{\Gamma(\alpha + N_{kl} + X_{0l}) \Gamma(\beta + N_k - N_{kl} + 1 - X_{0l}) / \Gamma(\alpha + \beta + N_k + 1)}{\Gamma(\alpha + N_{kl}) \Gamma(\beta + N_k - N_{kl}) / \Gamma(\alpha + \beta + N_k)} \\ &= \begin{cases} (\alpha + N_{kl}) / (\alpha + \beta + N_k) & \text{if } X_{0l} = 1, \\ (\beta + N_k - N_{kl}) / (\alpha + \beta + N_k) & \text{if } X_{0l} = 0. \end{cases} \end{aligned}$$

A.4 Setting hyperparameters

In §4.2 we suggested setting $\alpha = \beta = 1/2$, which corresponds to drawing feature frequencies via $\theta \sim \text{Beta}(1/2, 1/2)$. It would be better to find values for α and β such that $\text{Beta}(\alpha, \beta)$

reflects how feature frequencies are actually distributed. Though feature frequencies are hidden variables, we can estimate them via $\hat{\theta}_l = N_l/N$, where N_l is the number of languages in the entire dataset with feature l, and N is the total number of languages. From these estimates we compute a sample mean $\mu = \sum_{l=1}^{L} \hat{\theta}_l/L$ and a sample variance $\sigma^2 = \sum_{l=1}^{L}(\hat{\theta}_l - \mu)^2/L$. We set these equal to the mean and variance of Beta(α, β):

$$\mu = \frac{\alpha}{\alpha + \beta},$$
$$\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)},$$

and solve for α and β to obtain

$$\alpha = \frac{\mu^2(1-\mu)}{\sigma^2} - \mu,$$
$$\beta = \frac{\mu(1-\mu)^2}{\sigma^2} - 1 + \mu.$$

In our analyses, we use this procedure to estimate α and β before culling rare features, and get $\alpha \approx 0.071$ and $\beta \approx 0.755$.

One problem with the foregoing is that there may exist many features that we do not observe. In our dataset there is a long tail of low-frequency features, and extrapolating from this, it is not unreasonable to suppose that there may be a larger, or even infinite, number of features that we do not observe, due to the fact that their feature frequencies are extremely low. A naive Bayes classifier, despite working well in practice, is simply unable to account for this possibility. For a new model that explicitly addresses this issue, please see §7 of Thibaux & Jordan (2007).

A.5 NBC scores

A.5.1 NBC scores for single core analysis

This section lists each language along with its score from the single core analysis in §5.1. Languages are given by language codes, which can be looked up in §A.6. They are ordered by the score, which represents how highland-like the language is. Training languages (those that define the classes on which the classifier was trained) are marked with A (Andean core) or C (control class).

A	ayr	56.64	A	qvn_caj	26.62	A	qxw	19.32
A	caw	54.43	C	teh	26.48	A	qxl	18.30
A	ayr_chl	49.00	A	qwa	23.87		inj	16.69
A	ayr_muy	46.82		ame	22.61	A	quk	16.25
A	cap	46.43	A	qvn_tar	22.53	A	inb	16.16
A	jqr	44.05	A	qxn	21.24	A	quw	15.77
A	qul	43.78	A	ure	20.55		quf	15.71
A	quz	43.01	A	qub	19.97	A	qvc	14.06

A	quy	12.83		pno	-2.43		pid	-10.39
	qvo	12.30		anb	-2.56		pbg	-10.94
	qvs	11.42		zro	-2.56		mpq	-11.49
	qus	10.64		crq	-2.57		cul	-11.93
C	ona_mtr	10.31		gum	-2.63	C	wba	-11.93
C	ona	10.02		myr	-2.86		brg	-11.95
	ccc	9.87		dny	-3.01		pib	-12.09
A	qvi	9.59		mcb	-3.35		guo	-12.09
	mca	9.51		yvt	-4.18	C	xir	-12.30
A	kuz	9.39		aca	-4.37		ura	-12.45
C	teh_tsh	9.18		tna	-4.53	C	pab	-12.49
	cht	7.30		omg	-4.58		kaq	-12.54
	cbi	6.93		kvn	-4.73		cbt	-12.63
	vil	6.74	C	kgp	-4.90	C	prr	-12.66
C	pue	6.21		sya	-5.33	C	waw	-12.73
C	yaw	5.79	C	arn	-5.38	C	yag	-12.77
	cag	5.72		aro	-5.72		mcf	-13.06
C	alc_sth	5.69		cot	-6.36		sae	-13.37
C	alc_nth	5.47	C	wap	-6.80		prq	-13.48
	tob_tks	5.11		pbb	-6.82		iqu	-13.51
	kbh	5.03		ese_per	-6.88		yuz	-14.24
C	alc_cnt	4.76		cui	-7.02		boa	-14.41
	mtp	3.61		kpc	-7.03		trr	-14.45
	jeb	3.53		knt	-7.17	C	txi	-14.53
	lec	3.35		ywn	-7.17		cbb	-14.60
C	fun	3.20		bwi_rng	-7.39	C	kzw_dzu	-14.79
	cpc	3.06		kav	-8.03	C	aap	-14.80
	cbu	2.99		shp	-8.12		mpd	-15.13
	moc	1.97	C	ter	-8.15		tuf	-15.14
	cpu	1.51		ese_bol	-8.20	C	arw	-15.36
	arl	1.18		pad	-8.26		cbr	-15.53
	tob_lng	0.54		yrl	-8.32		pcp	-16.13
	ycn	0.26		omc	-8.56		axb	-16.13
	bae	0.06		cni	-8.56		pio	-16.24
	cpb	-0.29		sha_ywn	-8.58		jaa_jmm	-16.51
	lul	-0.31		ktx	-8.64	C	aoc_tar	-16.52
	gae	-0.97	C	hix	-8.78		srq	-16.79
	plg	-1.10		mzp	-8.81		cbv	-17.12
	ign	-1.41		cod	-9.12	C	car_ven	-17.19
	cav	-1.56		not	-9.24	C	car_esr	-17.19
	rey	-1.85		tit	-9.28		pav	-17.21
	trn	-1.91	C	tpy	-9.70		orw	-17.51
	rgr	-1.94		kbc	-9.93	C	myp	-17.68
	bwi_cen	-1.99		mbn	-10.26		xor	-17.90
	cox	-2.01		ito	-10.38		yup_mac	-18.86

C	pbh	-19.21	C	pak	-28.96		pui	-47.63
	mbp	-19.29		cbs	-29.47		auc	-47.90
C	kbb	-19.34	C	way	-29.54		emp	-47.95
C	guc	-19.57		psx	-30.49	C	kui	-48.03
	nuc	-19.61		adc	-30.65		mpu	-48.63
	chb	-19.68		mcd	-30.70		tnc	-49.11
	yup_irp	-19.72		agr	-31.13		ano	-49.16
C	yar	-20.06		yme	-31.18		kxo	-49.68
	jaa_jrw	-20.28		hub	-31.90		inp	-49.89
C	mbc	-20.30	C	gub	-32.77	C	tpn	-49.91
C	guu	-20.43		ltn	-33.03	C	kyr	-49.99
	cao	-20.76	C	tqb	-33.41		cax	-50.20
	arh	-20.98		boa_mrn	-33.60		kog	-51.20
C	ake	-21.01		kwi	-34.37		gvc	-52.05
	bmr	-21.15	C	xiy	-35.01	C	yae	-52.05
	jru	-21.20		yaa	-35.28	C	bkq_wst	-52.31
	cbd	-21.76		tba	-35.30		tpr	-52.62
	hto	-22.36		slc	-35.34		noa	-52.79
	swx	-22.83	C	gvp	-36.56	C	apy	-52.83
C	aoc_are	-23.14		wmd	-36.69	C	xet	-52.83
	noj	-23.17		amc	-36.79		gyr	-52.90
C	wca	-23.37		gta	-36.90		mbr	-52.93
	umo	-23.39		con	-36.96		irn	-52.97
	mcg	-23.45	C	mmh	-37.22		gvo	-53.33
	pev	-23.45		ash	-37.28		yad	-53.44
C	bor	-23.52		gqn	-38.18	C	avv	-53.45
C	car_frg	-23.55		nab_kth	-38.80	C	xsu	-53.56
C	atr	-24.28	C	wau	-38.90		xwa	-53.68
C	mch	-24.66		ayo	-38.91	C	xav	-53.72
	yui	-24.71		cto	-40.18	C	awe	-53.76
	des	-24.93	C	plu	-41.24	C	zkp	-53.83
	cub	-25.20		cof	-42.73		aqz	-53.85
	ore	-26.22		oca	-43.62	C	urb	-54.03
	huu	-26.28		bdc	-44.02	C	taf	-54.08
	acu	-26.68	C	jur	-44.26		coe	-54.12
C	tri	-26.73	C	mav	-45.22		cas_cov	-54.18
	unk	-27.48	C	guh	-45.27	C	opy	-54.20
C	asu	-27.50	C	sru	-45.48	C	myu	-54.21
	tav	-27.52	C	api	-45.76	C	mbl	-54.33
	tae	-27.72	C	kpj	-46.09		tuo	-54.60
	mts	-27.81		skf	-46.43		gui_chn	-54.70
	jiv	-27.89		jbt	-46.58	C	oym_jri	-55.42
	bsn	-28.21		cmi	-47.42		gug	-55.44
	pyn	-28.59		yuq	-47.56		cbg	-55.50
C	ako	-28.68	C	awt	-47.63		axg	-55.75

	cyb	-55.88		sri	-58.93		cas_tsi	-62.59
C	kqq	-56.55		tue	-58.93	C	yau	-63.36
	arr	-56.60	C	kay	-59.34	C	apn	-63.54
C	pta	-56.70	C	xer	-59.60		ktn	-64.60
C	kyz	-56.88		jua	-59.66		jup	-64.99
	tpj	-57.16		myy	-59.84	C	yrm_pac	-65.21
C	pto	-57.20		ark	-59.87	C	suy_tap	-65.29
	mot	-57.20		kwa	-60.01	C	txu	-65.38
	tca	-57.29		sey	-60.27	C	eme	-66.96
C	guq	-57.35	C	kgk	-60.50	C	shb	-67.99
C	gvj	-57.36	C	ram	-60.83	C	xra	-68.02
	snn	-57.38	C	suy	-61.45	C	kre	-69.16
	amr	-57.50		sja	-61.54	C	qpt	-70.17
C	asn	-57.56		bao	-61.74		wyr	-70.53
C	bkq_est	-57.59		mbj	-61.78		yab	-70.90
	nhd	-57.66	C	rkb	-61.88	C	xok	-77.43
C	xri	-57.79		adw	-61.92	C	wca_yma	-82.81
	apu	-57.85		pah	-61.92	C	wca_yae	-82.81
	gui_izo	-58.25		urz	-61.92	C	guu_ven	-82.81
C	gun	-58.38		pir	-62.13	C	xsu_kol	-83.10
	ceg	-58.91	C	oym_amp	-62.45	C	guu_par	-84.53
	cbc	-58.93		cas_msa	-62.59			

A.5.2 NBC scores for dual core analysis

This section lists each language along with its scores from the dual core analysis in §5.2. Languages are given by language codes, which can be looked up in §A.6. The three scores represent the resemblence that a language has to, respectively, the Northern-Central Andean core, the Southern Andean core, and the control class.[18] The languages are ordered by the third score. Training languages (those that define the classes on which the classifier was trained) are marked with N (Northern-Central Andean core), S (Southern Andean core), or C (control class).

S	ayr	-45.29	45.29	-65.21		ame	27.84	-28.31	-28.82
S	caw	-41.16	41.16	-60.77	N	qwa	20.76	-20.76	-28.33
S	ayr_chl	-42.32	42.32	-55.98	N	qvn_tar	26.55	-26.75	-28.24
S	ayr_muy	-42.00	42.00	-53.69	N	qxn	18.84	-18.84	-25.03
S	cap	-55.78	53.36	-53.45	N	qxw	23.49	-23.74	-24.98
S	quz	-50.27	50.04	-51.60	N	quw	23.87	-34.08	-23.87
S	qul	-40.66	40.66	-49.86	N	inb	23.60	-31.38	-23.60
S	jqr	-29.38	29.38	-47.34	N	qub	18.20	-18.20	-23.51
N	qvn_caj	24.18	-24.18	-30.82	N	quk	22.89	-25.47	-22.97

[18]The scores s_1, s_2, s_3 are the log-odds of a language belonging to class 1, 2, or 3. A log-odds s_k is related to the probability p_k of a language belonging to class k via the formula $s_k = \log p_k/(1 - p_k)$, as explained in §4.4.

S	ure	-36.94	22.94	-22.94
C	teh	-25.02	22.79	-22.91
	inj	21.93	-22.75	-22.51
	quf	21.39	-23.49	-21.52
	qvo	20.70	-36.27	-20.70
	qvs	17.92	-26.48	-17.92
N	qvc	17.24	-18.31	-17.65
	mca	-42.73	16.55	-16.55
N	qvi	16.34	-23.52	-16.34
N	quy	16.05	-17.97	-16.21
S	kuz	-36.79	14.28	-14.28
N	qxl	13.04	-16.03	-13.10
C	yaw	12.99	-24.19	-12.99
C	ona	-34.18	12.68	-12.68
	cht	10.34	-18.30	-10.34
	qus	8.45	-16.40	-8.45
	cbi	8.15	-22.65	-8.15
	jeb	7.90	-21.03	-7.90
C	ona_mtr	-20.34	6.55	-6.55
	kbh	6.10	-24.72	-6.10
C	alc_sth	-25.24	5.98	-5.98
C	alc_cnt	-25.50	5.88	-5.88
C	alc_nth	-22.20	4.91	-4.91
	mtp	-23.74	3.70	-3.70
	pno	3.44	-21.13	-3.44
	vil	-27.13	2.80	-2.80
	cbu	2.49	-14.70	-2.49
	cag	-20.39	1.61	-1.61
	anb	1.38	-20.59	-1.38
	zro	1.38	-20.59	-1.38
	sya	1.11	-22.64	-1.11
	gae	0.83	-25.04	-0.83
	myr	0.33	-28.03	-0.33
	kav	0.29	-26.89	-0.29
C	wap	-0.15	-29.78	0.15
	knt	-0.21	-23.27	0.21
	ywn	-0.21	-23.27	0.21
	ccc	-0.49	-4.98	0.46
	yvt	-0.86	-25.79	0.86
	shp	-0.86	-21.71	0.86
	omg	-0.88	-20.44	0.88
	lec	-22.86	-1.17	1.17
	ign	-1.28	-16.96	1.28
C	pue	-15.11	-1.90	1.90
	bwi_cen	-2.12	-18.14	2.12
	cav	-2.31	-18.06	2.31
C	teh_tsh	-7.68	-2.32	2.32
	lul	-22.02	-2.46	2.46
	sha_ywn	-2.81	-22.91	2.81
	rey	-3.24	-16.81	3.24
	tna	-3.44	-23.21	3.44
	bae	-3.59	-12.42	3.59
	arl	-3.79	-12.02	3.79
C	hix	-4.41	-30.02	4.41
	ktx	-4.44	-24.02	4.44
	plg	-4.56	-18.15	4.56
	cpc	-4.96	-5.79	4.59
	tob_tks	-5.61	-8.50	5.56
C	kgp	-5.66	-19.07	5.66
	aro	-6.18	-28.09	6.18
	dny	-6.43	-14.17	6.43
	gum	-7.20	-17.06	7.20
	cod	-7.31	-21.88	7.31
	kaq	-7.50	-24.97	7.50
C	fun	-12.09	-7.55	7.54
	crq	-19.26	-7.60	7.60
	cpu	-12.81	-7.75	7.75
C	arn	-7.97	-18.26	7.97
	cbt	-7.98	-26.99	7.98
	ycn	-8.08	-10.97	8.03
	bwi_rng	-8.12	-25.20	8.12
	cpb	-14.13	-8.32	8.31
	tob_lng	-8.36	-17.21	8.36
	cox	-9.06	-17.94	9.06
	ura	-9.27	-28.41	9.27
	pbb	-9.44	-13.17	9.42
	mcb	-9.99	-18.09	9.99
	kvn	-10.05	-18.10	10.05
	moc	-10.47	-12.14	10.30
	mpq	-10.32	-29.60	10.32
	pid	-37.96	-10.53	10.53
	yrl	-10.62	-25.44	10.62
	kpc	-10.76	-20.44	10.76
C	xir	-11.28	-30.52	11.28
	mcf	-11.32	-30.34	11.32
	iqu	-11.33	-24.25	11.33
	trn	-13.74	-11.86	11.71
C	yag	-11.78	-27.24	11.78
	rgr	-11.87	-19.02	11.87
	pad	-12.05	-27.20	12.05

C	waw	-12.32	-28.60	12.32
	ese_per	-12.39	-18.92	12.38
C	prr	-12.88	-26.03	12.88
	guo	-13.08	-30.49	13.08
	cui	-13.25	-18.93	13.25
	nuc	-13.56	-30.79	13.56
	boa	-36.86	-13.92	13.92
	ese_bol	-13.94	-19.80	13.94
	cbv	-13.97	-36.46	13.97
C	kzw_dzu	-14.00	-34.23	14.00
	pcp	-14.47	-28.89	14.47
	brg	-15.16	-24.92	15.16
C	myp	-15.22	-29.55	15.22
C	txi	-15.22	-30.75	15.22
	mpd	-15.36	-28.59	15.36
	yuz	-15.45	-31.66	15.45
	cot	-16.17	-16.44	15.61
	pio	-15.97	-30.43	15.97
C	pab	-16.02	-23.31	16.02
	not	-16.21	-23.36	16.21
	cni	-16.33	-18.74	16.24
C	aap	-16.26	-27.32	16.26
	kbc	-16.37	-25.45	16.37
	jaa_jmm	-16.42	-32.83	16.42
	trr	-16.60	-24.26	16.60
	mbp	-16.68	-45.01	16.68
	aca	-17.16	-19.45	17.06
	mzp	-25.54	-17.11	17.11
	sae	-17.48	-24.26	17.48
	omc	-17.67	-19.32	17.49
	mbn	-18.33	-18.13	17.53
	srq	-17.60	-28.68	17.60
	pbg	-17.76	-22.34	17.75
	pyn	-17.88	-43.67	17.88
C	wba	-17.92	-23.54	17.91
	cao	-17.98	-31.90	17.98
C	ter	-18.89	-18.55	18.02
	pib	-18.64	-23.23	18.63
	ito	-28.29	-18.63	18.63
	mts	-18.94	-42.01	18.94
C	arw	-19.01	-28.63	19.01
	cbb	-19.01	-26.34	19.01
	jiv	-19.16	-41.16	19.16
	swx	-19.63	-45.96	19.63
C	car_ven	-19.67	-28.90	19.67
C	car_esr	-19.67	-28.90	19.67
	cbd	-19.78	-42.81	19.78
	jaa_jrw	-19.97	-31.58	19.97
	cul	-20.35	-21.23	20.00
	yup_mac	-20.02	-34.23	20.02
	prq	-20.09	-24.65	20.08
	adc	-20.29	-45.59	20.29
	cbr	-20.35	-27.16	20.34
	bmr	-20.39	-42.26	20.39
	axb	-26.75	-20.90	20.90
	mcd	-20.97	-44.99	20.97
C	tpy	-24.03	-21.06	21.01
	acu	-21.02	-36.99	21.02
	tuf	-21.37	-26.68	21.36
	orw	-21.62	-27.55	21.62
	noj	-21.68	-46.11	21.68
C	yar	-21.74	-31.88	21.74
	umo	-22.10	-37.46	22.10
	arh	-22.24	-43.66	22.24
C	guc	-22.25	-31.75	22.25
	chb	-22.44	-31.45	22.44
	xor	-22.62	-28.80	22.62
	yup_irp	-22.66	-31.03	22.66
C	pbh	-22.73	-31.06	22.73
	tit	-22.81	-25.24	22.73
	hto	-22.92	-42.98	22.92
	cbs	-23.12	-44.17	23.12
C	guu	-23.22	-31.05	23.22
	des	-23.55	-40.53	23.55
	huu	-23.74	-43.91	23.74
C	kbb	-23.77	-31.12	23.77
	pav	-23.95	-27.20	23.91
	yui	-23.95	-41.15	23.95
	jru	-24.20	-32.53	24.20
	cub	-24.33	-38.01	24.33
C	aoc_tar	-28.70	-24.77	24.75
	agr	-24.91	-42.67	24.91
C	bor	-25.29	-42.71	25.29
	hub	-25.53	-44.19	25.53
C	wca	-25.62	-33.20	25.61
	tav	-26.19	-43.13	26.19
	bsn	-26.28	-43.87	26.28
	ore	-26.91	-43.77	26.91
C	aoc_are	-27.27	-33.21	27.27
	mcg	-27.34	-34.43	27.34

	pev	-27.34	-34.43	27.34
	psx	-27.45	-44.53	27.45
C	car_frg	-27.72	-33.61	27.72
C	mbc	-31.93	-27.76	27.74
	kwi	-27.85	-49.32	27.85
C	xiy	-28.16	-52.07	28.16
C	ake	-32.37	-28.69	28.66
	yaa	-28.94	-46.74	28.94
	yme	-29.39	-40.98	29.39
C	way	-29.83	-38.00	29.83
C	mch	-34.08	-31.51	31.43
C	atr	-31.64	-39.83	31.64
	unk	-32.02	-38.89	32.02
C	mmh	-32.07	-47.31	32.07
C	tri	-35.13	-33.47	33.30
	amc	-33.31	-45.99	33.31
C	asu	-33.56	-35.73	33.45
	cto	-66.22	-34.05	34.05
C	pak	-34.47	-36.39	34.33
C	ako	-38.34	-34.62	34.59
	tae	-35.89	-35.33	34.87
	ltn	-35.25	-41.60	35.25
C	wau	-35.27	-47.50	35.27
	slc	-35.43	-51.69	35.43
	tba	-36.37	-47.75	36.37
	gta	-37.40	-52.84	37.40
C	gvp	-38.03	-43.63	38.03
C	gub	-38.05	-43.29	38.04
C	tqb	-38.85	-44.80	38.85
	wmd	-43.80	-39.01	39.01
	gqn	-39.24	-48.03	39.24
	ayo	-39.39	-55.43	39.39
	boa_mrn	-39.75	-47.66	39.75
C	plu	-40.42	-56.53	40.42
	nab_kth	-40.48	-46.57	40.47
	oca	-40.56	-58.98	40.56
	ash	-40.70	-47.55	40.70
C	jur	-41.25	-61.61	41.25
C	kpj	-41.86	-54.95	41.86
C	sru	-42.17	-63.64	42.17
	cof	-43.34	-54.45	43.34
	skf	-43.53	-56.78	43.53
	con	-43.87	-47.34	43.84
	jbt	-44.94	-56.37	44.94
	cmi	-45.96	-58.20	45.96
	emp	-47.04	-61.87	47.04
	mpu	-47.10	-55.48	47.10
	bdc	-47.25	-50.79	47.22
	mbr	-47.34	-65.65	47.34
C	kui	-47.63	-55.46	47.63
	kxo	-47.72	-55.82	47.72
C	awt	-48.09	-57.68	48.09
C	api	-48.47	-53.34	48.46
C	kyr	-48.65	-63.49	48.65
	pui	-48.77	-52.89	48.76
	kog	-49.02	-67.79	49.02
	cax	-49.08	-57.55	49.08
	tnc	-49.12	-57.84	49.12
C	tpn	-49.61	-58.51	49.61
	axg	-49.98	-71.80	49.98
C	mav	-51.54	-50.38	50.11
	auc	-50.32	-57.10	50.31
C	zkp	-50.40	-61.99	50.40
C	bkq_wst	-50.66	-67.73	50.66
	cyb	-50.99	-65.58	50.99
	gvo	-51.05	-65.35	51.05
C	guh	-52.00	-53.51	51.80
	yad	-51.85	-60.52	51.85
	xwa	-52.01	-61.68	52.01
C	bkq_est	-52.03	-74.93	52.03
	ano	-52.20	-56.54	52.19
C	myu	-52.33	-66.92	52.33
C	apy	-52.43	-62.84	52.43
	tca	-52.43	-72.54	52.43
C	xsu	-52.55	-58.74	52.55
	mot	-53.33	-65.25	53.33
C	opy	-53.34	-67.77	53.34
C	xet	-53.55	-61.64	53.55
C	mbl	-54.18	-58.67	54.17
C	kqq	-54.25	-71.82	54.25
C	guq	-54.26	-74.69	54.26
	inp	-56.84	-54.50	54.41
C	xav	-54.71	-63.45	54.71
	yuq	-55.70	-55.35	54.82
C	avv	-54.84	-59.42	54.83
	cbc	-54.90	-70.63	54.90
	sri	-54.90	-70.63	54.90
	tue	-54.90	-70.63	54.90
	myy	-55.24	-70.71	55.24
	kwa	-55.39	-70.83	55.39

	cas_cov	-61.73	-55.76	55.75
	gui_chn	-55.81	-59.87	55.80
	gvc	-55.88	-59.12	55.84
	gyr	-56.35	-58.20	56.20
	ceg	-56.25	-77.41	56.25
C	xer	-56.31	-71.12	56.31
C	awe	-56.40	-59.39	56.35
C	xri	-56.54	-60.91	56.53
	cbg	-62.13	-56.61	56.61
	tpr	-60.40	-56.66	56.63
	arr	-56.69	-65.84	56.69
C	yae	-60.98	-57.06	57.04
C	rkb	-57.14	-72.04	57.14
	bao	-57.14	-72.61	57.14
C	urb	-57.36	-60.36	57.31
	amr	-64.58	-57.36	57.35
C	taf	-57.70	-59.19	57.50
	coe	-58.70	-57.96	57.57
	tpj	-57.71	-64.85	57.71
	aqz	-58.29	-58.89	57.85
	noa	-58.06	-60.64	57.99
	mbj	-58.08	-73.86	58.08
	gug	-58.38	-60.05	58.21
	gui_izo	-58.37	-61.00	58.30
	nhd	-58.38	-67.36	58.38
C	pto	-58.51	-68.83	58.51
C	pta	-58.58	-63.17	58.57
	tuo	-58.74	-60.79	58.61
	snn	-58.63	-72.48	58.63
	irn	-59.03	-61.29	58.94
	apu	-59.31	-60.22	58.97
C	kgk	-59.24	-68.07	59.24
C	oym_jri	-59.94	-60.51	59.49
C	suy	-59.58	-68.83	59.58
C	ram	-59.64	-65.41	59.63
	ark	-59.91	-65.28	59.91
	jua	-59.95	-68.53	59.95
C	kyz	-60.00	-67.60	60.00
C	asn	-60.05	-66.61	60.05
	pir	-60.30	-72.99	60.30
C	gun	-60.57	-63.67	60.52
C	gvj	-60.85	-62.65	60.70
C	yau	-61.85	-68.79	61.85
C	txu	-61.93	-79.05	61.93
	sey	-61.94	-70.51	61.94
C	apn	-62.30	-68.13	62.30
	cas_msa	-69.05	-62.58	62.58
	cas_tsi	-69.05	-62.58	62.58
C	kay	-63.05	-64.04	62.73
	sja	-71.23	-63.63	63.63
	adw	-63.87	-65.53	63.70
	pah	-63.87	-65.53	63.70
	urz	-63.87	-65.53	63.70
C	xra	-70.88	-64.04	64.04
C	suy_tap	-64.23	-68.97	64.22
C	yrm_pac	-66.87	-64.51	64.42
	ktn	-67.05	-65.20	65.05
	yab	-65.10	-78.96	65.10
C	oym_amp	-65.82	-65.94	65.18
C	eme	-65.98	-81.60	65.98
	jup	-66.41	-70.51	66.39
C	shb	-70.88	-66.59	66.58
C	kre	-68.99	-72.93	68.97
C	qpt	-69.42	-73.10	69.40
	wyr	-73.40	-71.49	71.35
C	xok	-76.23	-77.24	75.92
C	xsu_kol	-84.37	-77.31	77.31
C	wca_yma	-84.37	-78.15	78.15
C	wca_yae	-84.37	-78.15	78.15
C	guu_ven	-84.37	-78.15	78.15
C	guu_par	-84.10	-81.43	81.36

A.6 Language codes

This section lists the language codes used in this paper in alphabetical order. The vast majority of these codes are standard ISO-639 language codes, but they have been supplemented where necessary by three-letter extensions of related languages, or in a small number of cases, wholly new codes. Note that the use of a three-letter extension does not constitute a claim regarding the status of the variety so denoted as a dialect of the variety denoted by the first three letters, if indeed it even denotes a variety (this is not the case for [alc], for

example, since no single 'Alacalufan' language exists).

aap	Arára, Pará
aca	Achagua
acu	Achuar-Shiwiar
adc	Arara do Acre
adw	Amundava
agr	Aguaruna
ake	Ingarikó
ako	Akurio
alc_cnt	Alacalufe (Central)
alc_nth	Kawesqar
alc_sth	Alacalufe (Southern)
amc	Amahuaca
ame	Yánesha
amr	Amarakaeri
anb	Andoa
ano	Andoke
aoc_are	Pemon (Arekuna dialect)
aoc_tar	Pemon (Tarepang dialect)
api	Apiaká
apn	Apinayé
apu	Apurinã
apy	Apalaí
aqz	Akuntsú
arh	Ika
ark	Arikapú
arl	Arabela
arn	Mapudungun
aro	Araona
arr	Karo
arw	Lokono
ash	Aʔiwa
asn	Asurini do Xingú
asu	Asuriní do Tocantins
atr	Waimiri-Atroarí
auc	Waorani
avv	Avá-Canoeiro
awe	Awetí
awt	Araweté
axb	Abipon
axg	Arára do Mato Grosso
ayo	Ayoreo
ayr	Aymara (Central dialect)
ayr_chl	Aymara (Chilean dialect)

ayr_muy	Muylaq' Aymara
bae	Baré
bao	Waimaha
bdc	Emberá-Baudó
bkq_est	Bakairí (Eastern dialect)
bkq_wst	Bakairí (Western dialect)
bmr	Muinane
boa	Bora
boa_mrn	Miraña
bor	Borôro
brg	Baure
bsn	Barasana-Eduria
bwi_cen	Baniwa (Central)
bwi_rng	Baniwa (Rio Negro)
cag	Chulupí
cao	Chácobo
cap	Chipaya
car_esr	Carib (Suriname dialect)
car_frg	Carib (French Guiana dialect)
car_ven	Carib (Venezuela dialect)
cas_cov	Mosetén de Covendo
cas_msa	Mosetén de Santa Ana
cas_tsi	Tsimané
cav	Cavineña
caw	Callawaya
cax	Bésɨro
cbb	Cabiyarí
cbc	Karapanã
cbd	Carijona
cbg	Chimila
cbi	Cha'palaa
cbr	Cashibo-Cacataibo
cbs	Kashinawa
cbt	Shawi
cbu	Candoshi-Shapra
cbv	Kakua
ccc	Chamicuro
ceg	Chamacoco
chb	Muisca
cht	Cholón
cmi	Emberá-Chamí
cni	Asháninka
cod	Kokama-Kokamilla
coe	Koreguaje
cof	Tsáfiki

con	Cofán
cot	Caquinte
cox	Nanti
cpb	Ashéninka (Ucayali-Yurúa dialect)
cpc	Ashéninka (Apurucayali dialect)
cpu	Ashéninka (Pichis dialect)
crq	Chorote (Iyo'wujwa and Iyowa'ja dialects)
cto	Emberá-Catío
cub	Kubeo
cui	Cuiba
cul	Kulina
cyb	Cayubaba
des	Desano
dny	Dení
eme	Emerillon
emp	Northern Emberá
ese_bol	Ese Ejja (Bolivia)
ese_per	Ese Eja (Peru)
fun	Yaathe
gae	Warekena
gqn	Kinikinao
gta	Guató
gub	Guajajára
guc	Wayúu
gug	Paraguayan Guaraní
guh	Guahibo
gui_chn	Chiriguano (Chané dialect)
gui_izo	Chiriguano (Izoceño dialect)
gum	Guambiano
gun	Mbyá
guo	Guayabero
guq	Aché
guu	Yanomamö
guu_par	Yanomamɨ of Parawau
guu_ven	Yanomamɨ of Venezuela
gvc	Wanano
gvj	Guajá
gvo	Gavião do Jiparaná
gvp	Gavião do Pará
gyr	Guarayu
hix	Hixkaryána
hto	Huitoto, Minica
hub	Huambisa
huu	Huitoto, Murui
ign	Ignaciano

inb	Inga (Highland dialect)
inj	Inga (Jungle dialect)
inp	Iñapari
iqu	Iquito
irn	Myky
ito	Itonama
jaa_jmm	Jamamadí
jaa_jrw	Jarawara
jbt	Jabutí
jeb	Shiwilu
jiv	Shuar
jqr	Jaqaru
jru	Japreria
jua	Júma
jup	Hup
jur	Jurúna
kaq	Capanahua
kav	Katukína
kay	Kamayurá
kbb	Kaxuiâna
kbc	Kadiwéu
kbh	Camsá
kgk	Kaiwá
kgp	Kaingang
knt	Katukína (Panoan)
kog	Kogi
kpc	Curripaco
kpj	Karajá
kqq	Krenak
kre	Panará
ktn	Karitiâna
ktx	Kaxararí
kui	Kuikúro-Kalapálo
kuz	Kunza
kvn	Border Kuna
kwa	Dâw
kwi	Awa-Cuaiquer
kxo	Kanoé
kyr	Kuruáya
kyz	Kayabí
kzw_dzu	Karirí-Xocó (Dzubukua dialect)
lec	Leco
ltn	Latunde
lul	Lule
mav	Sateré-Mawé

mbc	Macushi
mbj	Nadëb
mbl	Maxakalí
mbn	Macaguán
mbp	Damana
mbr	Nukak
mca	Maka
mcb	Matsigenka
mcd	Sharanawa
mcf	Matsés
mcg	Mapoyo
mch	Yekwana
mmh	Mehináku
moc	Mocoví
mot	Barí
mpd	Manchinere
mpq	Matís
mpu	Makuráp
mtp	Wichí (Mision la Paz dialect)
mts	Yora
myp	Pirahã
myr	Muniche
myu	Mundurukú
myy	Macuna
mzp	Movima
nab_kth	Kithaulhu
nhd	Nhandeva
noa	Woun Meu
noj	Nonuya
not	Nomatsigenga
nuc	Nukini
oca	Ocaina
omc	Mochica
omg	Omagua
ona	Ona
ona_mtr	Haush
opy	Ofayé
ore	Máíhɨki
orw	Oro Win
oym_amp	Wayampi (Ampari dialect)
oym_jri	Wayampi (Alto Jarí dialect)
pab	Paresí
pad	Paumarí
pah	Tenharim
pak	Parakanã

pav	Wari'
pbb	Páez
pbg	Paraujano
pbh	Panare
pcp	Pacahuara
pev	Pémono
pib	Yine
pid	Piaroa
pio	Piapoco
pir	Piratapuyo
plg	Pilagá
plu	Palikúr
pno	Panobo
prq	Ashéninka (Perené dialect)
prr	Puri
psx	Pisamira
pta	Pai Tavytera
pto	Zo'é
pue	Puelche
pui	Puinave
pyn	Poyanáwa
qpt	Parkateje
qub	Huallaga Huánuco Quechua
quf	Ferreñafe Quechua
quk	Chachapoyas Quechua
qul	Bolivian Quechua (Northern and Southern dialects)
qus	Santiago del Estero Quechua
quw	Tena Quechua
quy	Ayacucho Quechua
quz	Cuzco-Collao Quechua
qvc	Cajamarca Quechua
qvi	Imbabura Quichua
qvn_caj	North Junín Quechua (San Pedro de Cajas dialect)
qvn_tar	North Junín Quechua (Tarma dialect)
qvo	Napo Quichua
qvs	San Martin Quechua
qwa	Ancash Quechua (Sihuas and Corongo dialects)
qxl	Salasca Quechua
qxn	Huaylas-Conchucos Quechua
qxw	Jauja-Huanca Quechua
ram	Canela
rey	Reyesano
rgr	Resígaro
rkb	Rikbaktsa
sae	Sabanê

sey	Secoya
sha_ywn	Shanenawa
shb	Ninam of Ericó
shp	Shipibo
sja	Epena
skf	Sakirabiá
slc	Sáliba
snn	Siona
sri	Siriano
srq	Sirionó
sru	Suruí
suy	Suyá
suy_tap	Tapayuna
swx	Suruahá
sya	Saynawa
tae	Tariana
taf	Tapirapé
tav	Tatuyo
tba	Aikanã
tca	Ticuna
teh	Tehuelche
teh_tsh	Teushen
ter	Terêna
tit	Tinigua
tna	Tacana
tnc	Tanimuca-Retuarã
tob_lng	Toba (Lañagashik dialect)
tob_tks	Toba (Takshek dialect)
tpj	Tapieté
tpn	Tupinambá
tpr	Tuparí
tpy	Trumai
tqb	Tembé
tri	Trió
trn	Trinitario
trr	Taushiro
tue	Tuyuca
tuf	Tunebo (Central dialect)
tuo	Tucano
txi	Ikpeng
txu	Mebengokre
umo	Umotína
unk	Enawené-Nawé
ura	Urarina
urb	Kaapor

ure	Uru
urz	Uru-Eu-Wau-Wau
vil	Vilela
wap	Wapichana
wau	Waurá
waw	Waiwai
way	Wayana
wba	Warao
wca	Yanomámi
wca_yae	Yanomae of Demini/Tototopi
wca_yma	Yanomama of Papiu
wmd	Mamaindé
wyr	Wayoró
xav	Xavánte
xer	Xerénte
xet	Xetá
xir	Xiriâna
xiy	Xipaya
xok	Xokleng
xor	Korubo
xra	Krahô
xri	Krinkati-Timbira
xsu	Sanumá
xsu_kol	Sanömá of Kolulu
xwa	Kwaza
yaa	Yaminawa
yab	Yuhup
yad	Yagua
yae	Pumé
yag	Yahgan
yar	Yabarana
yau	Hoti
yaw	Yawalapití
ycn	Yucuna
yme	Yameo
yrl	Nheengatú
yrm_pac	Yãroamë of Serra do Pacu/Ajarani
yui	Yurutí
yup_irp	Yukpa (de Irapa)
yup_mac	Yukpa (Macoíta)
yuq	Yuqui
yuz	Yurakaré
yvt	Yavitero
ywn	Yawanawa
zkp	Kaingáng, São Paulo

zro Sápara