Linguistic evolution in time and space: addressing the methodological challenges

Chundra Cathcart[1,2] and Balthasar Bickel[1,2]

[1]Department of Comparative Language Science, University of Zurich
[2]Center for the Interdisciplinary Study of Language Evolution, University of Zurich

Abstract

One of the most specific features of the human language faculty is its intrinsic spatio-temporal dynamic, as reflected in languages' characteristic mode of extra-genomic evolution. Understanding what has emerged in the hominin lineage therefore requires capturing this dynamic and the diversity of languages and structures that it generates. In this chapter, we review the state of the art in probabilistic models designed for this task. Current approaches tend to excel either in capturing temporal or spatial processes. We summarize these and describe avenues for integrating them, illustrated by a case study on the evolution of sound inventories over time and space.

1 Introduction

The study of language evolution is an ambiguous project, concerning itself with the biological evolution of the language faculty as well as with the spatio-temporal dynamics of languages over space and time, i.e., what is sometimes called more specifically "linguistic evolution" (Bickel et al. 2023). Attempts to connect these two foci propose that at least some aspects of the evolution of the language faculty were driven by the same extra-genomic transmission processes that also drive linguistic evolution (e.g. Kirby 2017). In other words, currently observable language change to some extent recapitulates the evolution of the language faculty in the hominin lineage. The central methodological toolkit for this enterprise are experiments on social learning and micro-evolutionary studies on how linguistic structures emerge in iterated populations of users.

Here we focus on an alternative bridge, with different methodological implications (Bickel 2015, Bickel et al. 2023). We probe linguistic evolution as a specific trait of the human phenotype, an intrinsic property of the language faculty. Language change reveals some of the key mechanisms that underlie this faculty, and it therefore points to what exactly must have come together when the faculty emerged — specific social processes of convergence and divergence and specific neural processes of learning, producing and comprehending, both conditioned on laws of mathematics (e.g. information theory) and (bio)physics (e.g. articulator movement). Examining these mechanisms through evidence from language change expands the purview of language origins research from the static possibilities (such as basic compositionality) to the dynamic probabilities of the language faculty (such as the preference for minimizing dependencies during language change, or the preference for accelerated divergence when languages split).

The methodological toolkit needed for this enterprise consists primarily of models that capture language change and diversification processes. Interest in questions of this sort directly connects with a long history in the fields of historical linguistics and linguistic typology. In past decades, the advent of probabilistic modeling of linguistic questions has made it possible to formulate explicit testable hypotheses regarding language change. This chapter focuses on Bayesian computational models designed for this purpose. Models of this sort aim to shed light on the processes that give rise to extant linguistic data, taking into account neurobiological biases of learning and processing along with social and demographic phenomena such as language contact, as well as interactions between these conditions on linguistic evolution.

In this chapter, we provide a review of statistical modeling of linguistic evolution, focusing on two powerful families of Bayesian models adopted from computational biology that are suited to complementary aims, explicitly characterizing on one hand evolutionary pressures in language change and the stability of features, and on the other contact and admixture between languages. While some attempts have been made to reconcile these two approaches and integrate them with each other in models of vertical transmission and horizontal transfer, these models are beset by a number of limitations.

We introduce the nascent framework of distributional phylogenetic modeling, a flexible family of phylogenetic modeling that allows for the exploration of the effects of certain evolving variables on the patterns of change in others. We outline some ways in which distributional phylogenetic models can incorporate phylogeographic information in order to test hypotheses regarding the relationship between population history and patterns of change. We present a case study that assesses the role of patterns of geographic dispersal in the evolution of phonological inventories in the Austronesian languages.

We find decisive evidence that longer migrations from the Austronesian homeland coincide with faster rates of linguistic change and decreases in inventory size, partly but not entirely in line with predictions derived from the literature on sociolinguistic pressures in language change. We discuss the broader implications of these results, highlighting some pitfalls of relying on geospatial information alone as proxies for language contact and stable multilingualism. We discuss ways in which this framework can be further enriched and expanded to shed light on the joint evolution of linguistic and extralinguistic traits.

2 Background

Linguistic typology seeks an understanding of the processes giving rise to distributions of linguistic features, with an eye to associations between linguistic features as well as associations between extralinguistic and linguistic features (Bickel 2007). In recent decades, the availability of large linguistic databases (see Ivani and Bickel 2024 for a review of syntactic databeses; ASJP (Wichmann et al. 2018) for a lexical database, and PHOIBLE (Moran and McCloy 2019) or LAPsyd (Maddieson et al. 2013) for phonological databases) has made it possible to conduct data-driven investigations into the distribution of linguistic features, quantifying the effects of both historical contingencies ("event-based" factors) and of general principles of language use, learning, production and comprehension ("functional" factors), as well as the interaction of these (Bickel 2015). Linking large linguistic databases with non-linguistic metadata pertaining to the languages in them has facilitated investigations into a large number of associations between linguistic and extra-linguistic features that bear on evolutionary patterns (Nettle 1999, Atkinson 2011, Everett 2013, Bickel et al. 2015, Everett et al. 2015, Blasi et al. 2019, Roberts 2018, Barbieri et al. 2022).

As in other fields, a challenge to modeling the relationship between linguistic features and other linguistic or extralinguistic features is the presence of non-independence among the data points observed due to phylogenetic relatedness, an issue known as Galton's Problem (Narroll 1961). A group of related languages may display a correlation between two linguistic features because there is some functional benefit to their cooccurrence, or because they have simply inherited both features from a common ancestor, in which the features came to cooccur by chance. Properly addressing this issue is key to understanding the role of different pressures in language change and explaining which constellations of features are due to functional pressures and which are due to historical contingencies, or both.

Galton's Problem has been addressed in a variety of ways. The commonest strategy takes relatedness as a nuisance factor in need or control. The classical implementation of this (Dryer 1989) was to restrict samples so as to minimize relatedness, e.g. by sampling only one language per genus, a taxonomic level defined as groups roughly comparable to Indo-European subgroups. This approach effectively enforces phylogenetic non-independence within a data set, but results in massive under-use of available data, which can result in reduced statistical power. An more recent alternative are hierarchical (alternatively "mixed-effects" or "multilevel") regression models (Jaeger et al. 2011). Models of this sort involve partial pooling, assessing the influence of effects of interest on a response variable while accounting for variation at the group level ("random effects") that should not be conflated with the population level effect of interest ("fixed effects""), e.g., by capturing family- and area-specific biases that might influence fixed effects in models with complete pooling (Gelman and Hill 2007).

A drawback of hierarchical models in which coefficients vary at the family level is that they treat families as discrete units and relatedness among languages as a binary variable. French, Italian, and Pashto are all Indo-European languages, but French and Italian are demonstrably more closely related to each other than they are to Pashto, as they both descend from Latin. This property of relatedness can be captured by hierarchical models in which the propensity for a feature which a language displays due to its phylogenetic relatedness to other languages is generated by a Gaussian process (Guzmán Naranjo and Becker 2022); language-level feature biases are distributed according to a multivariate normal distribution with a mean of zero and a covariance matrix based on the phylogenetic relatedness of languages in the sample. The same process can be used to quantify the expected similarity among languages with respect to some propensity for a feature due to their geographic proximity.

In short, models in this tradition ensure that all potential nuisance factors are controlled for when investigating the relationship between two variables. The benefit is a control for false positives. Controling for relatedness in hierarchical models guards against mistaking a pattern for independent evidence when it exist just because languages descend from the same ancestor. At the same time, however, these models risks false negatives. When related languages maintain the same pattern since they split from each other, this could just as well point to some external pressure (functional or event-based) that maintains the pattern despite the split (Bickel 2015). Hierarchical models absorb any such signal as part of the group variance (random effects), not as effects of interest at the population (fixed-effect) level. Therefore, we might miss signals, and the probability of this increases with phylogenetic depth in the groups we control for.

An alternative approach that has gained considerable track over the past 20 years is that of phylogenetic linguistics (or "phylolinguistics", Greenhill et al. 2021), which borrows phylogenetic comparative methods (PCMs) from computational biology to the study of linguistic evolution. For the most part, these approaches explicitly model the transmission and diffusion processes thought to give rise to the diversity we observe, rather than treating them as nuisance factors. Another approach to analyzing linguistic diversity in this tradition is the use of models which explicitly represent the role of areal admixture in shaping cross-linguistic distributions of features. We describe these families of models in the following section.

3 Explicit models of linguistic evolution

3.1 Bayesian modeling

Recent years have seen the rise of Bayesian modeling in a number of subfields of linguistics. Bayesian methodologies allow practitioners to create models characterizing stochastic processes capable of generating observed data, expressed in the form of probability distributions (e.g., Normal, Binomial, etc.). This generative process usually results in a tractable likelihood (on models with intractable likelihoods, see Sisson and Fan 2011) of the data under different values of model parameters (unobserved quantities involved in the generative model), written $P(\text{data}|\theta)$. Model likelihoods can be used to estimate parameter values. In a maximum likelihood framework, this involves finding a single point estimate for θ which maximizes the model likelihood. Bayesian modeling incorporates prior beliefs regarding parameter values to estimate posterior parameter distributions, which peak around values with high posterior probabilities, according to Bayes' theorem:

$$P(\theta|\text{data}) = \frac{P(\text{data}|\theta)P(\theta)}{\int P(\text{data}|\theta)P(\theta)d\theta}$$

Parameter estimation allocates credibility to different hypotheses regarding the data generation process, as certain parameter values will be compatible with certain hypotheses but not others.

For some simple models, posterior parameter distributions have an analytic solution. In most cases, however, posterior distributions involve an intractable integral and require some form of estimation. Markov chain Monte Carlo (MCMC) is a popular family of sampling-based algorithms for inference of posterior parameters. Parameters are initialized with random values, which are stochastically updated according to different algorithms designed to steer values toward regions of higher posterior probability (Geman and Geman 1984, Robert et al. 2010) for a large number of iterations, until posterior samples are drawn from the stationary distribution of the chain. For a more detailed introduction to Bayesian inference, see Gelman et al. 2013, Stone 2013, Kruschke

2014, McElreath 2020.

3.2 Phylogenetic modeling

A prominent use of phylogenetic methods in linguistics has involved inferring phylogenies of language families. While the traditional comparative method of historical linguistics is capable of establishing secure subgroups within language families on the basis of shared innovations, it cannot always resolve higher order subgrouping within families as well as questions regarding the absolute chronology of divergences. Bayesian phylogenetic inference provides a means of allocating credibility to different candidate tree topologies and estimating divergence dates while incorporating a degree of uncertainty. In many cases, these methods have produced results that dovetail with received wisdom regarding the dispersal of language families, but in other cases, have produced conflicting results (Bouckaert et al. 2012, Chang et al. 2015, Heggarty et al. 2023).

Phylogenetic comparative methods (PCMs) are a family of methods that model the evolutionary dynamics of specific features instead of entire lineages. In general, these methods rely on an existing phylogeny of the organisms or languages under study, on which dynamics of change are conditioned. Likelihood-based PCMs have a number of applications. A popular set of methods quantify the phylogenetic stability of features (Blomberg et al. 2003, Abouheif 1999, Borges et al. 2019) and compute phylogenetically corrected means. These have enjoyed a degree of use in linguistics (Macklin-Cordes et al. 2021, Carling and Cathcart 2021a).

PCMs can be applied to categorical and continuous data types. Most applications in linguistics have so far focused on categorical data because the relevant features are typically discrete, or modelled as discrete (e.g., verb-object vs object-verb order, labiodental vs bilabial stops, or descendents of *sem- vs *Hoy- as the label for ‘one’ in Indo-European). A popular model for the evolution of this kind of data assumes that features undergo state changes over a phylogeny according to a continuous-time Markov (CTM) process, a stochastic process under which transitions between different states (or feature values) in a system occur according to transition rates, representing the frequency of different changes. Estimated rate values can be used to address hypotheses regarding the dynamics of change of the features in question. The likelihood of a set of rate values in a phylogenetic tree, $P(\text{data}|\text{rates}, \text{tree})$ can be efficiently computed by the PRUNING ALGORITHM (Felsenstein 1981, 2004), a dynamic program that exploits the independence structure of directed acyclic graphs (see Figure 1 and Section 8.8 in Harmon (2019) for a full walk-through). Posterior distributions for rates can be estimated via an MCMC algorithm. Following inference of rates, character histories can be simulated over the phylogeny using a technique known as stochastic character mapping (Huelsenbeck et al. 2003, Revell 2013, Bollback 2006). This allows

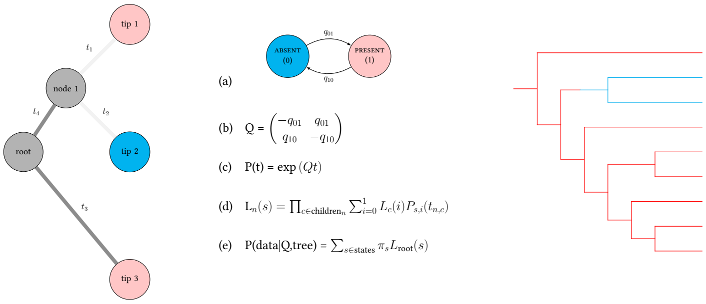

Figure 1: Illustration of the computation of rate likelihoods under the pruning algorithm (Felsenstein 1981) for a binary character (lefthand figure) and a hypothetical history of a binary character, visualized as a stochastic character map (righthand figure). Transitions between states occur according to a gain rate q_{01} and loss rate q_{10} (a). The rate matrix Q (b) can be exponentiated in order to generate a matrix of state transition probabilities $P(t)$ over a given interval of time t (c). For a given node in the tree n, the likelihood of a state s conditioned on the state values of all nodes descending from node n (d) can be computed using these transition probabilities, along with likelihoods at descendant states — e.g. the likelihood of node 1 in state 1 (red), $L_1(1)$, is the product of the probability of ending in state 1, $L_{c=\text{tip }1}(1)$, in the t_1-branch, $P_{1,1}(t_1)$, and the probability of ending in state 0, $L_{c=\text{tip }2}(0)$, in the t_2-branch, $P_{1,0}(t_2)$, with $L_c(i)$ normally evaluating to 0 or 1 at the tips (though see Jing et al. 2023 for generalization). These likelihoods are computed recursively in post-order traversal, with parent nodes of lighter branches visited prior to parent nodes of darker branches. The likelihood for the entire tree (e) is equal to the dot product of state likelihoods at the root and $\boldsymbol{\pi}$, the vector of prior root state probabilities, e.g., drawn from a uniform or the stationary distribution (see Felsenstein 2004, FitzJohn et al. 2009, Beaulieu and O'Meara 2014, Maddison et al. 2007).

graphical visualizations of how probabilities of states gradually change over time (Widmer et al. 2017, Blasi et al. 2019, Cathcart et al. 2020, Jing et al. 2023).

The CTM model can be used to address a wide range of questions regarding language change. A CTM chain has a STATIONARY DISTRIBUTION, which represents the proportion of time that the system is expected to be in a particular feature state as time approaches infinity. When the CTM has reached stationarity, this means that there will always be the same proportion of languages in different states even though individual languages keep changing between states. Accordingly, the proportions are independent of any original or current state, i.e., they are in a sense "timeless". Assume there are 100 languages in state A and 100 in state B, and transition probabilities of P(A⟶B) = .2 and P(B⟶A) = .8. A will gain 80 and lose 20; and B will gain 20 and lose 80, resulting in 160 A and 40 B languages. The same result will be achieved when starting from, say, 150 A and 50 B languages (A: $150 - .2 \times 150 + .8 \times 50$; B: $50 - .8 \times 50 + .2 \times 150$), and in fact from any distribution of frequencies. Accordingly, CTM models can be used to estimate long-term preferences for feature states that are independent of time (Maslova 2000, Cysouw 2011, Bickel et al. 2015, Widmer et al. 2017, Jäger and Wahle 2021).

Another use of these models is to investigate evolutionary interdependence between features. A popular model assesses whether changes in the state of one feature depend on values of a second feature (Pagel 1994); this model has enjoyed wide use in linguistics in order explore the dynamics of change between features that are synchronically highly correlated (Dunn et al. 2011, Cathcart et al. 2020, Jäger and Wahle 2021). Models of this sort have been used to investigate constrained pathways of change (Dunn et al. 2017, Shirtz et al. 2021). Additionally, the CTM model of character evolution is capable of carrying out ancestral state reconstruction to internal nodes of a phylogeny representing unnattested proto-languages (Maurits and Griffiths 2014, Carling and Cathcart 2021b, Widmer et al. 2017, Blasi et al. 2019, Zhou and Bowern 2015). While these reconstuctions typically fit well with traditional qualitative reconstructions, they have the advantage that they come with uncertainty estimates and can reach much deeper time depth than qualitative methods that rely on the documentation of earlier stages (ancient languages).

Despite their productive use in linguistics, CTM models are not without their limitations, at least in their most commonly used implementations. One drawback is that the transitions that can be modeled involve only attested states; for practical purposes, a CTM model can only make inferences regarding transitions between feature values observed at the tips of a phylogeny, and therefore, would not be capable of reconstructing a presently unobserved value. Sometimes this problem can be resolved by by re-coding features; for example a word order variable containing values such as SOV, SVO, etc. can be recoded into orthogonal binary features with dynamics that can be inferred independently (SV vs. VO). This provides the opportunities for a wider range of feature combinations to be reconstructed.

Another issue is that fitting CTM models is generally restricted to large language families with some branch length estimates in terms of calibrated calendar years or in terms of amounts of change (substitutions). There have been only few attempts to generalize the approach to small families and isolates, all seeking to combine information from small and large families in a single model to the maximum extent possible. One uses stationary probabilities from large families as priors for modeling the probability distributions in isolates and small families synchronically (Bickel et al. 2015, Bickel 2015). Another approach assumes that an isolate is not only an extant language but at the same time the root of a degenerate (zero time) phylogeny. Its likelihood is then estimated together with the likelihoods at the root from larger language families (Step (e) in Figure 1) (Jäger and Wahle 2021). Yet another approach might assume the time depth of isolates to lie between 5000 years (after which one would detect phylogenetic relations) and a maximum based on major human migration events (before which lineages would stem from the same population). These estimates could then be combined into a global supertree (Bouckaert et al. 2022, Shcherbakova et al. 2023), or, alternatively, into a forest of trees on which a model is fitted.

A further issue is that CTM models work best when targetting a single variable which takes one value per language (and this is how typological databases tend to code languages, apart from few exception; see Ivani and Bickel 2024 for a survey). But languages are complex systems with rampant POLYMORPHISM: languages are very often split in terms of word order (e.g. by clause type or information structural condition), case systems (e.g. by tense or aspect, or person), morphological syncretisms (e.g. by which cells they span, or by conjugation class) or even syntactic possibilities (e.g. allowing recursive noun phrases with prepositions but not with adjectivization). There have been various attempts of modelling such splits, tailored to specific data and research questions. Some have treated sub-systems as uncertainties, or as if they descended from a language like dialects (Bickel et al. 2015). A more explicit approach models sub-systems independently, each on their own, and than captures interactions and dependencies by aggregating posterior probabilities per time interval through stochastic character mapping (Widmer et al. 2017). An approach that models the dependencies between subsystems directly is to include split states into the CTM, so that languages can transition from, say, general non-ergative to a split ergative/accusative to a general accusative system, with or without additionally allowing direct transitions (Hong et al. in prep.). Alternatively, one can treat the split condition as a predictor variable for the presence of a sub-system (e.g. different cell collections that each predict the presence of syncretism), and then model the conditional presence in the CTM (cf. Cathcart et al. 2022 and Section 4 below).

Splits often come with probability distributions that shift between generations, and this represents yet another challenge for classical CTM approaches. From a micro-evolutionary perspective, a “gain” or “loss” of a linguistic feature represents the point in time when the inflection point of the S-shaped growth curve that characterizes the rise of linguistic variants (Greenberg et al. 1954, Bailey 1973, Niyogi et al. 1997, Yang 2000) is surpassed. What is therefore needed is models that go beyond discrete grammar features and tackle the probabilistic nature of language directly. One recent approach treats variation in word order as the product of probabilistic grammars, and then takes corpus-based probabilities for the likelihood estimation at the tip in the CTM model (Figure 1) (Jing et al. 2023). Another approach replaces CTM models by models for continuous measures, such as Brownian Motion or Ornstein-Uhlenbeck models (Witzlack-Makarevich et al. 2016, Hahn and Xu 2022).

Finally, CTM models do not explicitly model contact dynamics; if under a CTM model a feature is likely to have arisen frequently within a phylogeny, a CTM model cannot distinguish which births are due to functional pressures (e.g., adaptation to some processing principle) and event-based triggers where the feature has entered a language due to social conformity. Post-hoc analyses can give some insights, for example by comparing whether the evolutionary dynamics is similar between families in the same contact area vs. those in different areas (Bickel et al. 2015, Bickel 2019). Direct inference of contact effects is rare but there are promising developments in the phylogenetic inference of language trees. Some models are succesful in inferring lateral transfer between related languages based on patterns of cooccurrence (Kelly and Nicholls 2017, Neureiter et al. 2022), but these do not (yet) generalize to PCMs and do not take geography or proposed areal hypotheses into account.

3.3 Areal modeling

Other research explains variation in cross-linguistic patterns via models that explicitly represent areal influence and admixture between languages. A significant portion of this work makes use of mixture models. Mixture models partition data points in a sample into clusters depending on the values for features exhibited by individual data points, with data points sharing similar values belonging to similar clusters. As such, mixture models provide a means of dimensionality reduction, the resulting representation of which may have an interpretation involving language contact. Some, but not all of this work injects explicit spatial information into the probabilistic model used.

For linguistic purposes, clustering can take place at a number of levels. Languages themselves can be clustered on the basis of the features they display (Michael et al. 2014). Alternatively, the features within languages can be partitioned among different clusters. This is the objective of a subtype of mixture model known as admixture or mixed membership models. Mixed membership models allow for features within a language, words within a document (Blei et al. 2003), or single-nucleotide polymorphisms within genomic data (Pritchard et al. 2000) to be generated by one of a number of latent components. This methodology has been used to model admixture within languages corresponding to different areal sources (Reesink et al. 2009, Bowern 2012, Syrjänen et al. 2016, Cathcart 2020, 2022). Mixed membership models have also been used to ask which of a language's features are due to genetic inheritance as opposed to contact or universal preferences (Chang and Michael 2014), and can thus be used to quantify the borrowability and propensity for inheritance of different features. A recent approach, sBayes (Ranacher et al. 2021) infers the presence of geographical groupings across languages. Under the generative process for this model, a language can be assigned to a geographical group according to a spatially informed prior distribution over group membership (the other models cited above do not provide any explicit geographic information to their models). Within a language, each feature is assumed to come from three different sources: inheritance (depending on the distribution of the feature in the family to which the language belongs), location (depending on the distribution of the feature in the geographical group to which the language is assigned), and universal preferences (depending on the global distribution of the feature). Locations can reflect recent contact or the residue of earlier distributions that deviate from current global patterns. The three distributions are multinomial probabilities expressing the preference for a given feature value in a given linguistic taxonomy.

While mixture model-based approaches can explicitly model geography in a way that standard phylogenetic methods do not, they lack phylogenetic models' explicit, fine-grained represention of time depth and diachrony. The standard admixture model strongly embodies the Hardy-Weinberg principle in that is assumes that no mutations have taken place since the time of admixture between populations. This assumption may not be valid for all linguistic features. For instance, two languages could share a word order due to a relatively recent change, yet an admixture model could potentially impute this resemblance to deeper contact. One way to allay this concern is to work with features thought to be relatively stable or invariant, such as the operation of sound changes demonstrated to have taken place at an early date (Cathcart 2020, 2022). At least one population genetic model, overlooked in the linguistics literature, allows for a limited number of mutation events representing changes occurring between the time of admixture and the present day (Shringarpure and Xing 2009). Even with this relaxation, these models' representation of chronology is simplistic and not as expressive as that of timed trees. This temporal invariance is also seen in the representation of phylogenetic featural propensities found in sBayes. In sBayes, the distribution of variants of a feature (e.g., possible word orders) in a family such as Indo-European is represented by a single multinomial probability distribution generated by a Dirichlet prior distribution. This flat representation of family-level feature distributions has no sensitivity to the fact that Welsh and Irish are more likely to share a word order given their shared phylogenetic history than Welsh and Pashto. The main benefit of this representation is its contribution to the model's computational tractability. While this can be accomodated to some extent by varying the taxonmic level of family relations, the model still underuses phylogenetic information such as branch length and mutation rates. An alternative would be to assume that the log probabilities of different features evolve according to Brownian motion, and that feature probabilities themselves follow the logistic normal distribution, for which there exists a conjugate prior for the multinomial distribution which is however somewhat involved to implement (Mimno et al. 2008).

A number of methods attempt to synthesize phylogenetic and areal models somewhat more directly, although they do not necessarily capture the full evolutionary dynamics of transmission and diffusion. Daumé III (2009) constructs a mixture model designed to tease apart the roles of contact and descent in shaping typological distributions. Under this model, a feature in a language is generated according to either the areal group to which the language is assigned, if it is assigned to one, or via a coalescent process. The author shows that this model's spatial awareness results in the inference of more accurate phylogenies than a model without spatial awareness. At the same time, this model would be difficult to adapt to questions regarding prehistoric contact, etc. In a similar vein, Murawaki and Yamauchi (2018) employ an autologistic model to infer vertical stability and horizontal transmissibility. While these models advance our understanding of the interplay of different pressures in linguistic evolution, adapting and expanding these models to take into account different sources of variation (and to accommodate different prior distributions) may be a challenge, given their reliance on certain inference algorithms. More critically, information regarding geography tends to be based on present-day distributions, with no readily available way of shedding light on sociolinguistic pressures active at a greater time depth.

3.4 Issues regarding implementation

This brings us to yet another point. Coding up inference algorithms for complex probabilistic models from scratch is time consuming and error-prone, and may be beyond the remit of many linguists. Fortunately, a number of accessible probabilistic programming languages on the market make it possible for analysts to specify the data generating process, while the software does the heavy lifting of posterior parameter inference. A popular language is Stan (Carpenter et al. 2017), which is well documented and has a large online community of users. A limitation of these software packages, which generally make use of gradient-based methods to steer parameter estimates toward regions of high posterior probability, is that they cannot estimate discrete posterior distributions (as they are not differentiable), which must be marginalized out by summing.

This places constraints on the models that can be fitted using such programs, and in some cases requires a bit of rethinking of such models from a standpoint of tractability. For instance, it would not be possible to directly implement sBayes in Stan. sBayes samples a discrete variable representing the membership for a language with index l in a geographical group, which we can call $z_l \in \{1, ..., K\}$, and for each feature with index $d \in \{1, ..., D\}$ in the language, samples a discrete variable indicating whether the feature in question is inherited, borrowed (or otherwise geographically preserved), or due to universal preferences, which we can call $s_{l,d} \in \{I, B, U\}$. The likelihood of a configuration for a single language could be written as follows, with simplified notation (θ refers to all model parameters; $P(y_{l,d}|s_{l,d})$ is the probability of the observed value of a feature in a language conditioned on its source):

$$P(z_l|\theta) \prod_{d=1}^{D} P(s_{l,d}|\theta) P(y_{l,d}|s_{l,d})$$

In the context of Gibbs sampling with a conjugate prior, the different discrete variables (z_l; $\boldsymbol{s_{l,\cdot}}$) can be sampled conditioned on the variables on which they depend, which include current values for other discrete parameters. If configurations of discrete variables must be marginalized out, we are faced with the following sum:

$$\sum_{k=1}^{K} \sum_{x_1 \in \{I,B,U\}} \dots \sum_{x_D \in \{I,B,U\}} P(z_l = k) P(s_{l,d} = x_1|\theta) P(y_{l,d}|s_{l,d} = x_1) \dots P(s_{l,d} = x_D|\theta) P(y_{l,d}|s_{l,d} = x_D)$$

This sum is intractable as it enumerates $K3^D$ different configurations of values for discrete parameters. The issue is that because $\boldsymbol{s_{l,\cdot}}$ depends on z_l, all possible combinations of this collection of discrete variables must be taken into account. This issue would be alleviated if individual features in a language could belong to different linguistic areas $z_{l,d} : d \in \{1, \dots, D\}$, altering the above likelihood as follows:

$$\prod_{d=1}^{D} P(z_{l,d}|\theta) P(s_{l,d}|\theta) P(y_{l,d}|s_{l,d})$$

Each term in this product represents the likelihood of a feature in a language, $y_{l,d}$. Summing out the discrete parameters for all of these likelihoods in a single language requires only $3DK$ operations. Ultimately, whether or not they are justified, sacrifices like this one open the door to greater ease of model fitting with more flexibility, including use of diverse, not necessarily conjugate priors.

It is likely that in the near future the field will see an explosion of models designed in this spirit. We believe that progress in this direction will be facilitated by easily adaptable and expandable models cast in probabilistic programming languages that serve as lingue franche for analysts in the field. In the following sections, we outline a program for flexible phylogenetic models implemented in RStan, inspired by advances in hierarchical Bayesian modeling made by the brms framework (Bürkner 2017).

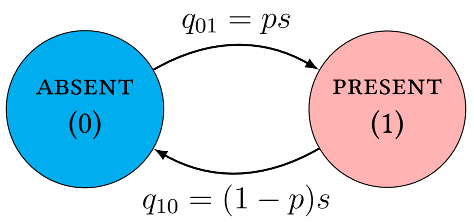

Figure 2: Binary continuous-time Markov process representing changes between presence and absence of a feature. Transitions are annotated according to the gain rate (q_{01}) and loss rate (q_{10}) of the feature, with alternative parameterizations according to speed of change (s) and stationary probability of feature presence (p).

4 Distributional phylogenetic modeling

In this section, we present a flexible method for analyzing multifaceted aspects of linguistic evolution in a nuanced manner. We term this program DISTRIBUTIONAL PHYLOGENETIC MODELING. We take as our point of departure two unrelated families of statistical models. The first group consists of phylogenetic models and phylogenetic comparative methods which model rate variation in the evolutionary histories of traits. The second group consists of distributional regression models, an increasingly popular Bayesian modeling framework which allows both the location and scale of a regression model to vary as a function of predictor variables (Bürkner 2017), thus relaxing a number of assumptions found in classical regression, such as homoskedasticity. We demonstrate a novel use of this modeling strategy, highlighting its ability to integrate spatial and temporal aspects of linguistic evolution and shed light on poorly understood aspects of change.

PCMs for discrete features generally assume that a feature moves between different values or states according to a continuous-time Markov process parameterized by transition rates (Figure 1). Under the standard view of a continuous-time Markov process for binary data, a feature arises and is lost according to a gain rate and a loss rate. Assuming a speed of change s (loss *or* gain) and stationary probability p, the gain rate and loss rate can be rewritten as ps and $(1-p)s$, respectively (Figure 2). This is the binary case of a general time-reversible model (Tavaré 1986), which parameterizes changes between multiple states in a continuous-time Markov chain according to stationary probabilities of state presence and exchange rates or speeds of change between each pair of states. Figure 3 shows simulated trajectories of change under binary CTM processes with different speeds and stationary probabilities.

Classical CTM models make global estimates regarding the dynamics of evolution of a feature, but they do not always deal with the impact of population history and the variation in dynamics

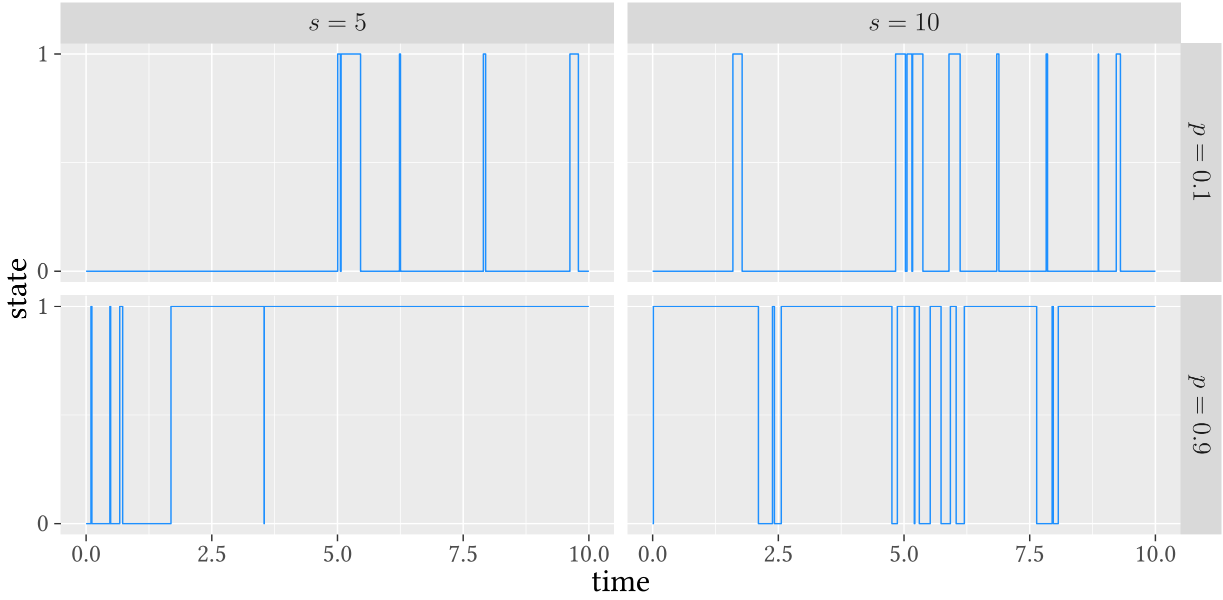

Figure 3: Simulated CTM processes showing transitions between states of a binary feature under different speeds ($s \in \{5, 10\}$) and stationary probabilities ($p \in \{0.1, 0.9\}$).

that external factors may bring about. Some linguistic work (Greenhill et al. 2017) has made use of modeling techniques that allow transition rates to vary across features and branch levels (Huelsenbeck and Suchard 2007), but this allows only post-hoc interpretations regarding which branches in a phylogeny are most affected by factors such as language contact, population density, etc. We expand this variable-rates (or "heterotachy") approach to allow for the regressing of various predictors, extralinguistic and linguistic, on the evolutionary parameters of the CTM process, profiting from advances made in biology and anthropology in assessing correlations between evolving variables (see Lartillot and Poujol 2011, Lartillot 2013, Horvilleur and Lartillot 2014, Ross et al. 2016, Ringen et al. 2021 and Cathcart forthcoming for a review).

We furthemore extend the models to allow both the speed of change and the stationary probability of a feature to vary according to one or more predictors. We do this in analogy with so-called DISTRIBUTIONAL MODELS, which allow both the predicted measure of central tendency and the dispersion of a response in a regression model to vary as a function of predictor variables (Rigby and Stasinopoulos 2005, Bürkner 2017).

A more refined understanding of whether certain predictors of change affect speed versus stationary biases toward a particular feature is crucial for understanding linguistic evolution, in particular if the goal is to characterize the intrinsic dynamic of the language faculty independently of the vagaries of population history. For instance, it helps answer question on how complexity evolves, for example in morphology ("polysynthesis"). While the decrease of (some aspects of) complexity has been succesfully linked to contact and admixture of second language learners (Lupyan and Dale 2010, Bentz and Winter 2013, Widmer et al. 2021), the apparent increase of complexity in regions of greater social isolation is unexplained. Is this increase driven by a stationary long-term preference in the language faculty, or is it a side effect from more rapid change under isolation that lets grammaticalization accumulate in repeated waves?

Predictors can be either static or evolve together with the response, they can be discrete or continuous, and, unlike classical models of correlated evolution (Pagel 1994), the design is not limited to feature pairs. Cathcart et al. (2022) use this framework to investigate whether the evolution of irregular stem alternation types in Romance verbs is a byproduct of a general instability in lower-frequency verbs, or driven by a long-term preference for regular paradigms in lower-frequency verbs and irregular ones in higher-frequency ones. This was tested by assessing the effect of frequency on both speed of pattern change as well as stationary probability (i.e., long-term preference) for irregular patterns. Results indicated an effect of frequency on long-term preference for irregular patterns, but not on speed of change, indicating that frequency does not play a role in a verb's stability of inflectional patterns, but instead impacts timeless biases toward or against irregular patterns.

5 A case study

Here, we focus on the phylogenetic and geospatial dimensions of the development of sound systems, probing the extent to which dynamics of geographic dispersal can be taken as a proxy for variation in sociolinguistic milieux that would have different impacts on linguistic evolution. We chose as an example the Austronesian languages, a family exhibiting considerable variation in terms of phonological inventories across vastly different language ecologies. Different regions of the Austronesian-speaking world exhibit disparities in the number of contrastive segments that the languages in them display, and underwent different processes of language change during their history. The Western Pacific, in particular archipelagos like Vanuatu, has been characterized by small languages in stable multilingual environments in which there are strong language ideologies and pressures toward schismogenesis and esoterogeny. Polynesia on the other hand is characterized by greater linguistic homogeneity and less pronounced contact between speakers of different languages.

Geographic dispersals have a number of potential impacts on linguistic change which are not fully understood and partly at odds with each other. Rapid dispersals have the potential to bring speakers into contact with and absorb populations speaking a different language, which can lead to rapid change; this may not have been the case in Austronesian, however, as rapid dispersals brought Austronesian speakers to uninhabited areas, and the Western Pacific, a region of longer settlement, bears witness to scenarios in which languages remaining closer to the Austronesian homeland for longer periods of time work to make themselves maximally divergent, in particular in terms of lexical items and phonological characteristics (Wurm 1967, François 2010, 2011, Rangelov et al. 2023). Similarly, language families often exhibit higher amounts of diversity (i.e., larger numbers of languages) closer to their homelands (Sapir 1916), reflecting greater degrees of schismogenesis. By contrast, languages further from the homeland, spoken in the vicinty of fewer closely related languages, may not exhibit as much schismogenetic effects.

Assuming that sound systems are prime indices of group identity in dense linguistic ecological niches (Evans 2019), the question arises whether schismogenetic effects primarily affect speed of segment replacement or the (stationary) probabilities of segment presence, or both together. This question can only be addressed with a distributional phylogenetic model. Concretely, we ask the following: (i) Do lineages closer to the homeland undergo faster speeds of change in segmental inventories? (ii) Do they display a longer-term preference for more contrastive and non-contrastive segments than lineages that migrated further away and are more isolated from other languages. Or do we find both effects together?

In what follows we explore the direct impact of patterns of geographic dispersal on linguistic evolution, jointly estimating the parameters of a phylogeographic model of language diffusion and a CTM model of changes in phonological inventories.

5.1 Data

We use data from PHOIBLE v. 2.0 (Moran and McCloy 2019), a large cross-linguistic database of segmental inventories which contains language-level metadata such as geographic coordinates at which a language is spoken. The data in PHOIBLE generally follows the analytical decisions made in the source from which it is extracted, and often contains more than one record of an inventory for a single language. We randomly sampled a record for languages where more than one record was available, and retained only segments found in the Cross-Linguistic Transcription Systems database (CLTS; Anderson et al. 2018) that were found more than once across all languages, in order to maximize cross-linguistic comparability.

We use the Austronesian phylogeny of Gray et al. (2009), which consists of a Bayesian tree sample of several hundred languages, to conduct our phylogenetic comparative analyses. Branch lengths were scaled to one time unit per millennium, so that a speed of, say, $s = .5$ means two

($1/s$) expected changes (gains or losses) per millenium. We discard languages in the tree not found in PHOIBLE. For languages in PHOIBLE but not in the tree (such as Cocos Malay), we use a semi-automated script in order to graft languages in PHOIBLE but not in the phylogeny onto the most appropriate loci in each tree in the tree sample. In total, our data consist of 192 binary segmental features in 110 languages.

5.2 Models

We build on the relaxed random walk model (RRW) of phylogeography, under which geospatial diffusion takes place according to a process of Brownian motion, the scale of which exhibits branch-level variation: on some branches of a phylogeny, geographic dispersal will take place at faster rates than on others. This model and extensions serve as the standard for modeling linguistic migration in continuous space (Lemey et al. 2010, Gill et al. 2017), and accurately recapitulate language families' spreads when dispersal involves expansion from a given point of origin, but not necessarily when wholesale migration from the point of origin has taken place (Neureiter et al. 2021); accordingly, the RRW and its extensions may not be appropriate for all of the world's language families. The basic RRW employed in this paper is not sensitive to environmental features in the way that more sophisticated discrete-space models are (Bouckaert et al. 2012, 2018, Koile et al. 2022). We choose this model to illustrate ease of implementabilty in standard statistical programming languages.

Concretely, the latitude and longitude (written ℓ for brevity) at the tips of a phylogeny are assume to follow a multivariate normal distribution (cf. O'Meara et al. 2006):

$$\boldsymbol{\ell}^t \sim \text{MultiNormal}(\boldsymbol{\ell}^r, \Sigma),$$

where $\boldsymbol{\ell}^t$ is the vector of observed values at the tips of the tree, and $\boldsymbol{\ell}^r$ is a vector the same length as $\boldsymbol{\ell}^t$ which repeats the value at the root, which is unobserved; Σ represents the phylogenetic covariance between tips in the phylogeny. Under time-homogeneous Brownian motion, the phylogenetic covariance between languages i and j, Σ_{ij}, is equal to the sum of the lengths of the branches leading from the root of phylogeny to the most recent common ancestor (MRCA) of the two languages, multiplied by a parameter σ representing the scale of the diffusion process. Under a RRW model, Σ_{ij} is equal to $\sigma \boldsymbol{\rho}^\top \boldsymbol{\zeta}_{\text{MRCA}_{ij}}$, where σ is the global scale of the process' diffusion, $\boldsymbol{\rho}$ is a vector of scale multipliers for each branch in the phylogeny and $\boldsymbol{\zeta}_{\text{MRCA}_{ij}}$ is a vector that contains the lengths of the branches that intervene between the root of the phylogeny and the most recent common ancestor of nodes i and j, with zeroes corresponding to branches that do not intervene. We assume that for each phylogeny, lon^t and lat^t are distributed as above, with separate scale parameters σ^{lon}, σ^{lat} and $\boldsymbol{\rho}^{\text{lon}}$, $\boldsymbol{\rho}^{\text{lat}}$ across both dimensions of diffusion. For simplicity, we do not model correlation between the two dimensions, and do not account for measurement error in longitude and latitude values recorded for languages. We transform longitude and latitude coordinates to an equal earth projection (Šavrič et al. 2019), which minimizes distortions in distance found at different latitude values and accounts for the fact that coordinates on either side of the prime meridian are relatively close to each other (despite having highly disparate longitude values).

Using the parameters $\text{lon}^r, \text{lat}^r, \sigma^{\text{lon}}, \sigma^{\text{lat}}, \boldsymbol{\rho}^{\text{lon}}, \boldsymbol{\rho}^{\text{lat}}$, longitude and latitude values for internal nodes of the tree can be sampled from a conditional multivariate normal distribution. For a given sample, it is then possible to compute the great-circle distance between the root node of the tree and and the descendant node of every branch in the tree, yielding values for branch-level displacements, $\delta_b : b \in \{1, ..., B\}$, i.e., the distance that the speech community represented by the descendant node has migrated from the linguistic homeland.

We link the quantity $\boldsymbol{\delta}$ to the CTM process of segmental inventory evolution. For simplicity and tractability we assume that each segment in each language can be gained or lost independently according to a binary CTM process. While this is not entirely justified given chain shifts and other dependencies, it has served an approximation with reasonable performance in previous work (e.g. Nikolaev and Grossman 2020, Guzmán Naranjo and Mertner 2022).

In line with our distributional modeling approach we parameterize change in segmental inventories according to a speed of change and an expected stationary probability. On a given branch $b \in \{1, ..., B\}$, a segment $d \in \{1, ..., D\}$ evolves according to a binary CTM process with speed $s_{d,b}\tau$ (τ being the global speed of change in the tree) and stationary probability $p_{d,b}$. We force the speed multiplier $s_{d,b}$ and stationary probability $p_{d,b}$ into the interval $(0, 1)$ by using a logit transformation and assume the following form in a hierarchical (multilevel) model:

$$\text{logit}\, s_{d,b} = \alpha_0^s + \alpha_d^s + (\beta_0^s + \beta_d^s) \log \delta_b + \gamma_{b,0}^s \sigma_0^s + \gamma_{b,d}^s \sigma_d^s$$

$$\text{logit}\, p_{d,b} = \alpha_0^p + \alpha_d^p + (\beta_0^p + \beta_d^p) \log \delta_b + \gamma_{b,0}^p \sigma_0^p + \gamma_{b,d}^p \sigma_d^p$$

In plain terms, the logit of the speed multiplier for segment d on branch b of the phylogeny is a function of an intercept-like baseline (α_0^s); a deflection to this baseline for segment d (α_d^s); coefficients representing the effects of displacement in each branch b (δ_b) on speed (β_0^s) and a deflection to this coefficient for segment d (β_d^s); as well as coefficients representing variation in branch-level speeds not explained by fluctuations in dispersal values, globally ($\gamma_{b,0}^s$) and at the

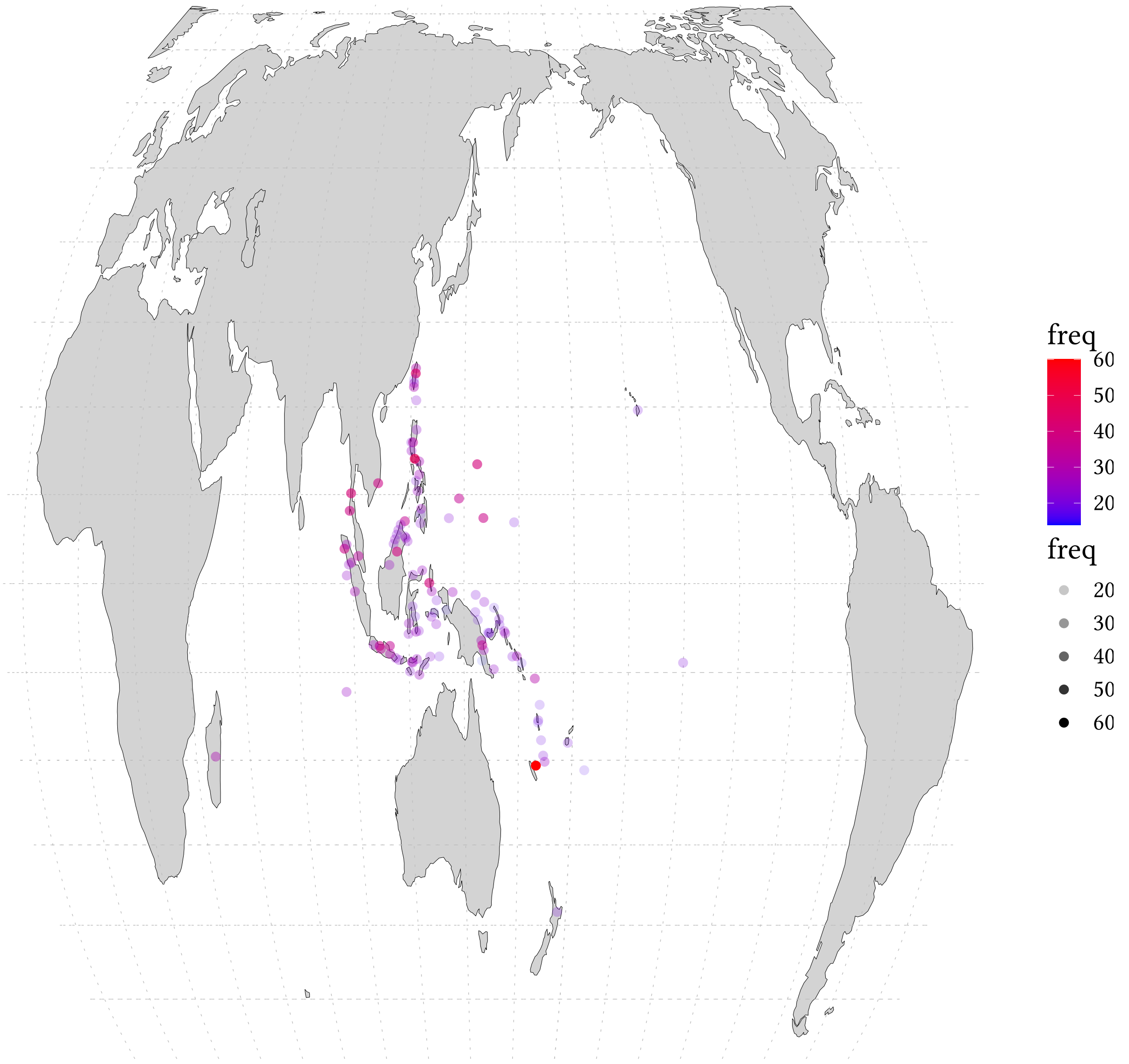

Figure 4: Austronesian languages in PHOIBLE. Color and transparency indicate the number of segments in the inventory selected for the language.

segment level ($\gamma^s_{b,d}$); $\boldsymbol{\sigma}^s$ represents scale parameters allowing these coefficients to capture as much variance as is justified. The parameters pertaining to the stationary probability of segment d's presence for branch b have the same interpretation. The likelihood of the observed distributions of each segment under the model parameters and a phylogeny can be computed via Felsenstein's pruning algorithm (Figure 1).

The model described above can be fitted in a variety of ways. One possibility is to jointly infer all parameters, phylogeographic and phylogenetic, in one model, which can be computationally costly. Another option is to first infer the phylogeographic parameters, most importantly $\boldsymbol{\delta}$, separate from the phylogenetic parameters. Subsequently, posterior estimates of this quantity can be used as an input to the phylogenetic model described above as a variable with measurement error (Bürkner 2017, Driver et al. 2017, Bürkner et al. 2019). This is justified for models which posit the directed influence of phylogeography on linguistic evolution and not vice versa. Here, we infer all parameters jointly, and leave a comparison of the merits of different methods for future work.

We place Normal$(0, 1)$ priors on the parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$, HalfNormal$(0, 1)$ priors on the parameters $\boldsymbol{\rho}$ and $\boldsymbol{\sigma}$, and a Uniform$(0, 20)$ prior on τ, which ensures that changes will not occur more frequently than once per 50 years (1/20 per millenium), roughly over a small number of generations. To incorporate phylogenetic uncertainty, we fit the model on 25 trees from the tree sample and concatenate the resulting posterior samples. We use RStan (Carpenter et al. 2017) for model fitting and make all data and code available at `https://github.com/chundrac/disp-phon`).

5.3 Results

Figure 5 provides a visualization of dispersal trajectories inferred by our model; blue branches are closer to the inferred homeland. Regarding the accuracy of the inferred homeland, there is perhaps room for improvement: while Taiwan as a homeland is supported by this posterior distribution, the northern Philippines are as well, and this latter homeland is not in line with received wisdom. It is possible that the relatively small number of Austronesian taxa in PHOIBLE is responsible for this relatively high uncertainty, and that more precise results will be reached with a greater number of languages. Our model allocates credibility to potential homeland locations over water. This is behavior potentially displayed by the RRW and other Gaussian process-based geospatial models. Implementing spatial constraints directly within programming languages like Stan (i.e., assigning zero prior probability to homeland locations that do not overlap with landmasses) is not straightforward, and at this point in time we are are not fully aware of the practical issues surrounding the implementation of discrete, feature-aware models of migration (in general these involve massive transition matrices between points on the globe, and dynamic solutions for storing pre-computed values). One post-hoc solution to this issue (which we do not employ here) is to discard samples after inference for which the root location is over water, which effectively produces results that would be obtained with a spatially informed prior.

Figure 6 shows the 95% and 85% posterior credible intervals (CIs) for β_0^s and β_0^p, model parameters representing the overall effect of dispersal on speed of change and long-term preference for segments. A commonly used criterion holds that distributions where the 95% CI (i.e., the interval spanning from the 2.5% to 97.5% percentiles) does not contain zero represent decisive evidence for an effect of a predictor on a response (Kruschke 2021). The results answer question negatively (i): We find a decisive positive effect of dispersal distance on speed of change, i.e. slower, rather than faster change closer to the homeland. But they answer question (ii) in the affirmative: We find a decisive negative effect of dispersal distance on stationary probability of segment presence, i.e., languages show more segments closer to the homeland This indicates that branches undergoing more extensive migration undergo faster speeds of phonemic change, but display a long-term trend towards having fewer consonants than branches closer to the point of origin of the family.

To further understand the reasons for this result, we inspect the branch-level phylogeographic diffusion patterns for one of the 25 tree samples over which we run our model. Figure 7 colors branches of the tree according to the distance between the root of the tree and the child nodes of each branch. Blue branches have undergone less displacement. Polynesian languages (at the top of the tree) are conspicuous in having smaller inventories and being geographically distant from the Austronesian homeland. Larger inventories can be found in languages that are closer to the homeland (e.g., in the bottom half of the tree). Our model may have inferred that these larger inventories are at least partially the product of shared retentions preserved by lower speeds of change, rather than brought on by rapid complexification of inventories due to extreme schismogenetic pressures.

Inspection of additional model parameters also allows us to assess the degree to which dynamics of change vary at the segment level. Figure 8 shows coefficients representing segment-level deflections from the overall effect of diffusion rate on speed of change and stationary probability of presence. The left-hand graphic shows 85% and 95% CIs for coefficients reflecting variation in the effect of dispersal distance on speed of change. For visibility, we show only segments with the 30 lowest and 30 highest median values. Here, all 95% CIs (and most 85% CIs) overlap with zero, indicating that there are not many decisive segment-level deviations from the overall effect of dispersal on speed of change. The right-hand graph arranges segments according to the degree

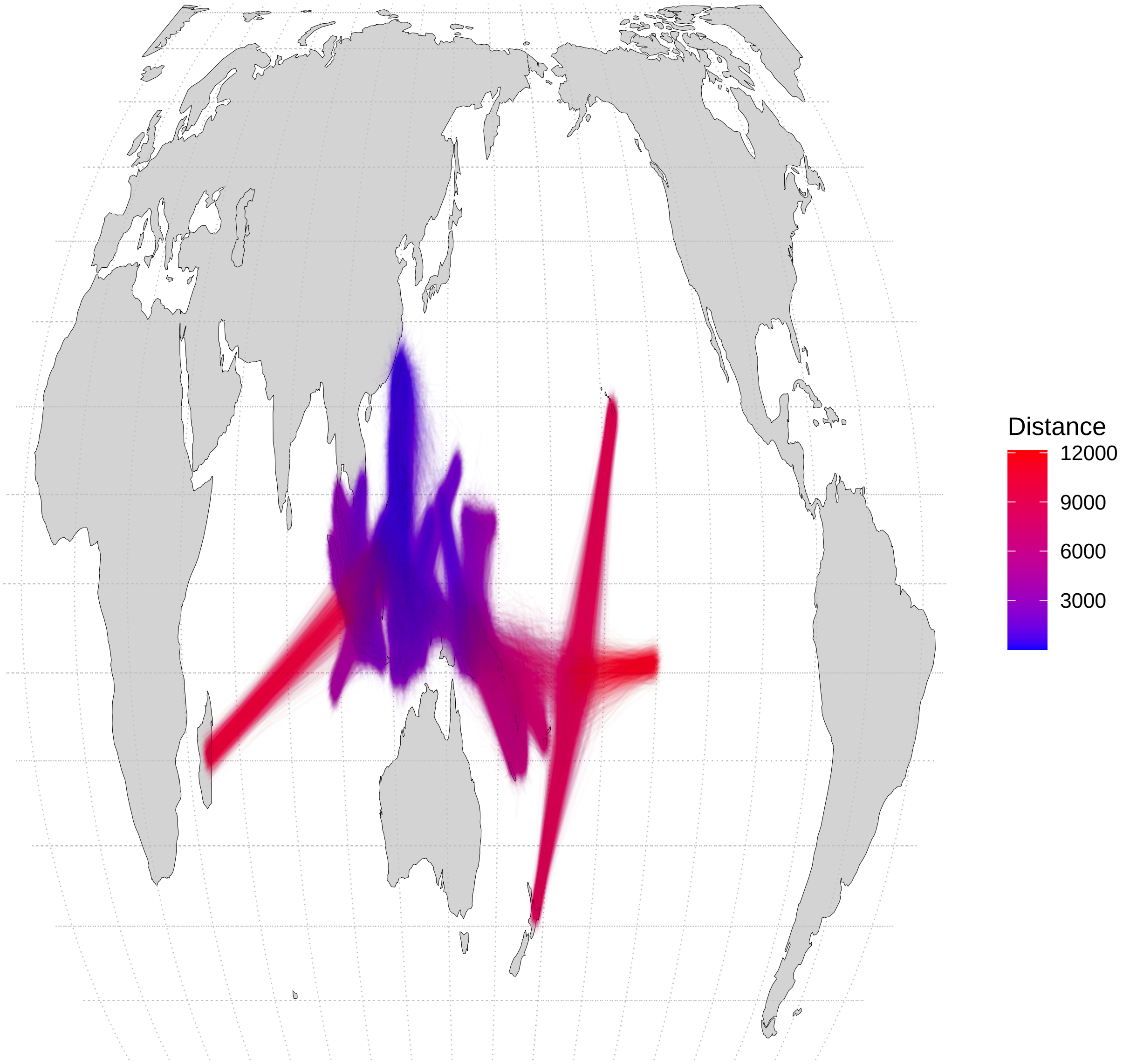

Figure 5: Posterior samples of phylogeographic trajectories inferred by our model. Colors represent distances (in km) from root location of tree for child of each branch.

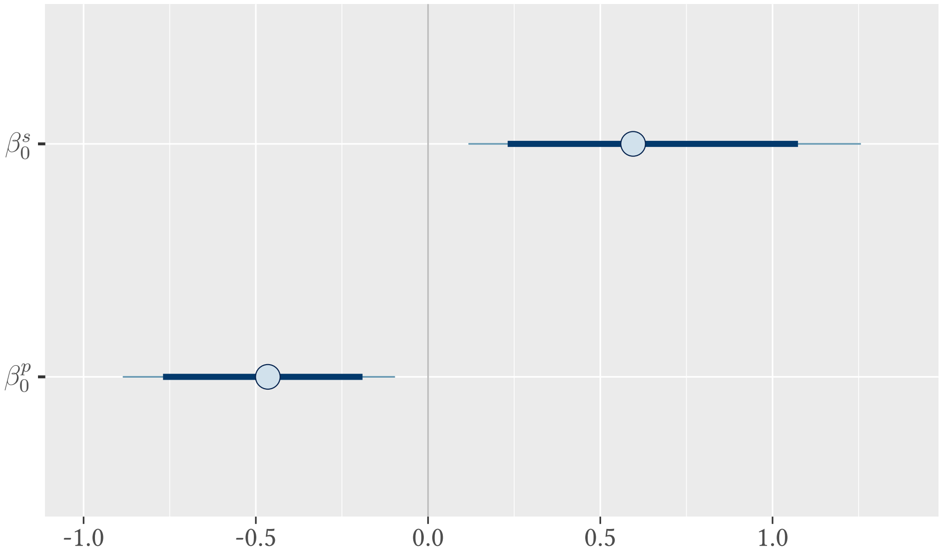

Figure 6: 95% and 85% posterior credible intervals (CIs) for β_0^s and β_0^p. Distributions where the 95% CI does not contain zero represent decisive evidence for an effect of a predictor on a response.

to which their long-term probability of presence is affected by dispersal distance (as above, with only the 30 lowest and 30 highest median values). Here, we see a larger number of 85% and 95% CIs that do not overlap with zero, indicating that segments vary considerably more in terms of their resistance to the effects of dispersal. At the top of the graph, we see segments for which we expect a decisively lower-than-baseline stationary probability of presence as distance from the homeland increases; the segments /ʐ/, /q/, and /ħ/ are restricted to languages spoken on Taiwan, close to the inferred homeland. At the other end of the spectrum is /m/, found in all languages; the stationary probability of /m/ will be higher than expected for increases in distance from the homeland, as it is a segment that universally preferred. In general, segments making up a core of most languages' inventories appear to be less sensitive to the effects of dispersal. These include nasals like /m/ as well as relatively basic segments like /k/, /b/, /d/, etc. In general, the sounds most resistant to changes coinciding with dispersal tend to include consonants that are thought to be more "primal" and basic in inventories (Bybee and Easterday 2022).

5.4 Discussion

The coefficients for speed and stationary probability provide a way of assessing the degree to which a particular linguistic feature is resistant to the pressures of contact (if our model of phylogeographic dispersal can in fact be taken as a proxy for contact proneness). Some idiosyncrasies

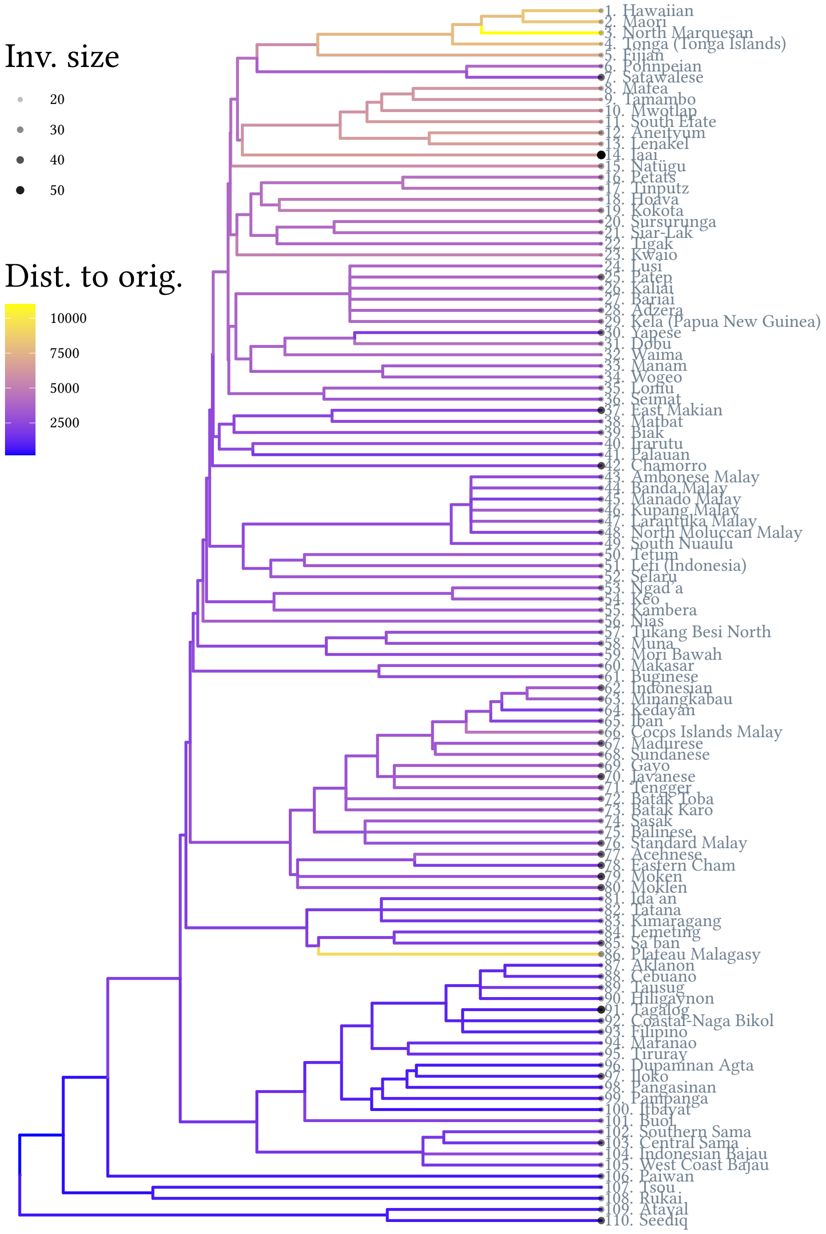

Figure 7: Median branch-level dispersals (great-circle distances in km between the inferred root location and the inferred/observed locations of the child nodes of each branch) for a tree from the tree sample, visualized using ggtree (Yu et al. 2018). Blue branches have undergone less dispersal than yellow ones. Dot sizes at tips of the tree represent inventory sizes for each language.

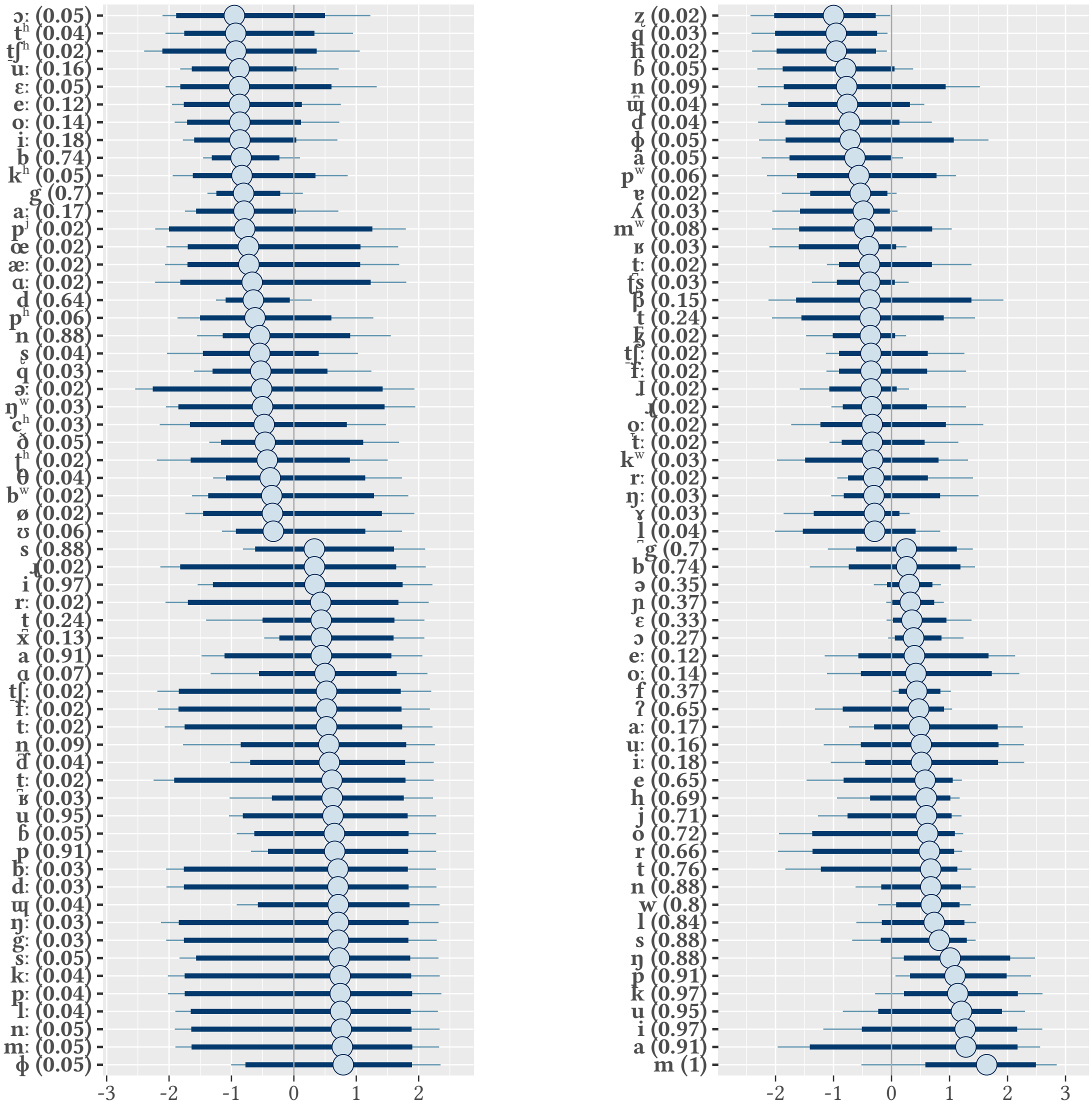

Figure 8: 95% and 85% posterior credible intervals for segment-level deflections to the global effect of dispersal on speed of change (left) and stationary probability of presence (right). The relative frequency of each segment in the data set is given in parentheses.

seen here may be artifacts of using PHOIBLE's segmental representations directly. An alternative is to use features (e.g., LABIAL, NASAL, etc.), if care is taken to code the data such that interactions of interest (e.g., VELAR PLOSIVE) are taken into account.

Given the large amount of literature arguing for the effect of longstanding stable multilingualism on patterns of schismogenesis, it is surprising to see that lineages that undergo more geographic dispersal, rather than those with longer histories of local settlement, show higher rates of phonemic change. Further inspection of these rates showed that branches exhibiting greater dispersals tend to lead to languages varying considerably in the size of their inventories, explaining why we see this effect at a large family-wide scale.

At the stage, we have only tentative explanations for this result. As mentioned previously, in any other scenario, we might interpret rapid language change accompanying large-scale migration as a signature of contact-induced changes brought on by the absorption of other populations, but this does not hold for Austronesian, as longer migrations largely brought speakers to uninhabited locations. It is possible that the faster linguistic changes we detect coincided with a sort of societal complexification (perhaps involving dialect leveling) that accompanied the development of technological advances needed to carry out large-scale migrations. Padilla-Iglesias et al. (2020) link geographic isolation of the sort found in the Eastern Pacific to acceleration of linguistic change, in the form of word loss.

Our results, obtained based on a wide phylogenetic scale, do not invalidate the idea that stable multilingualism fosters pressures to differentiate oneself from other social groups. It may be the case that using phylogeographic variables as a proxy for the social pressures of interest is not appropriate, and other variables need to be included. Recent attempts to characterize distributions of languages and linguistic features purely according to geographic predictors have produced mixed results (Urban 2021, Shcherbakova et al. 2023). In fact, a larger number of continuous coevolving predictors representing demographic information may need to be integrated into such models (cf. Ringen et al. 2021). It may also be the case that using segmental inventories as the response variable of interest does not capture fine-grained patterns of schismogenesis. Finally, it could be that the relationship between phylogeographic diffusion and linguistic change varies at cladistic levels smaller than the family, such as within individual subgroups.

We are confident that the distributional phylogenetic approach will facilitate tackling such questions with more specific and informative predictors, disentangling the various forces that shape the evolutionary dynamics of language.

6 Outlook

Bayesian models of linguistic evolution present a number of intriguing avenues for understanding the intrinsic dynamics of the language faculty. Considerable challenges lay ahead in terms of the integration of phylogenetic models, which incorporate flexible modeling of language change, with areal models showing an explicit representation of geospatial neighborhoods. Here, we discussed some potential rapprochements between these two families of models, and presented the results of a case study assessing the effect of phylogeography on linguistic evolution, couched in the nascent framework of distributional phylogenetic modeling. A benefit of this approach is that it can help recover deep historical pressures that cannot be detected solely from synchronic distributions of linguistic and extralinguistic features. A careful inspection of model parameters shed light on why our result obtained, and we outline additional ways in which models of this sort can be expanded and bring more linguistic insights into consideration. Model expansion in this direction will be key to validating a number of hypotheses regarding linguistic evolution. Many tasks remain in fully understanding the forces that shape linguistic evolution. We hope that with an increase of flexible approaches like the one outlined here, accompanied by careful model criticism, we will move closer to this goal.

References

Abouheif, E. (1999). A method for testing the assumption of phylogenetic independence in comparative data. *Evolutionary Ecology Research 1*(8), 895–909.

Anderson, C., T. Tresoldi, T. Chacon, A.-M. Fehn, M. Walworth, R. Forkel, and J.-M. List (2018). A cross-linguistic database of phonetic transcription systems. In *Yearbook of the Poznan Linguistic Meeting*, Volume 4, pp. 21–53. De Gruyter Open.

Atkinson, Q. D. (2011). Phonemic diversity supports a serial founder effect model of language expansion from africa. *Science 332*(6027), 346–349.

Bailey, C.-J. N. (1973). *Variation and linguistic theory.* Washington, D.C.: Center for Applied Linguistics.

Barbieri, C., D. E. Blasi, E. Arango-Isaza, A. G. Sotoripoulos, H. Hammarström, S. Wichmann, S. J. Greenhill, R. D. Gray, R. Forkel, B. Bickel, and K. K. Shimizu (2022). A global analysis of matches and mismatches between human genetic and linguistic histories. *Proceedings of National Academy of Science 119*(47), e2122084119.

Beaulieu, J. M. and B. C. O'Meara (2014). Hidden Markov models for studying the evolution of binary morphological characters. In L. Z. Garamszegi (Ed.), *Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology: Concepts and Practice*, pp. 395–408. Heidelberg, New York, Dordrecht, London: Springer.

Bentz, C. and B. Winter (2013). Languages with more second language learners tend to lose nominal case. *Language Dynamics and Change 3*, 1–27.

Bickel, B. (2007). Typology in the 21st century: Major current developments. *Linguistic Typology 11*, 239–251.

Bickel, B. (2015). Distributional typology: Statistical inquiries into the dynamics of linguistic diversity. In B. Heine and H. Narrog (Eds.), *The Oxford Handbook of Linguistic Analysis, 2nd Edition*, pp. 901–923. Oxford: Oxford University Press.

Bickel, B. (2019). Large and ancient linguistic areas. In M. Crevels, J.-M. Hombert, and P. Muysken (Eds.), *Language dispersal, diversification, and contact: a global perspective.* Oxford: Oxford University Press [pre-print available at `http://www.ivs.uzh.ch/bickel-files/papers/Ancient_areasBickel2015.pdf`].

Bickel, B., A.-L. Giraud, K. Zuberbühler, and C. P. Van Schaik (2023). Language follows a unique mode of extra-genomic evolution. *Evolutionary Human Sciences (under review).*

Bickel, B., A. Witzlack-Makarevich, K. K. Choudhary, M. Schlesewsky, and I. Bornkessel-Schlesewsky (2015). The neurophysiology of language processing shapes the evolution of grammar: Evidence from case marking. *PLoS ONE 10*, e0132819.

Blasi, D. E., S. Moran, S. R. Moisik, P. Widmer, D. Dediu, and B. Bickel (2019). Human sound systems are shaped by post-Neolithic changes in bite configuration. *Science 363(6432)*(6432).

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). Latent dirichlet allocation. *Journal of Machine Learning research 3*, 993–1022.

Blomberg, S. P., T. Garland Jr., and A. R. Ives (2003). Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution 57*(4), 717–45.

Bollback, J. P. (2006). SIMMAP: Stochastic character mapping of discrete traits on phylogenies. *BMC Bioinformatics 7*, 88.

Borges, R., J. P. Machado, C. Gomes, A. P. Rocha, and A. Antunes (2019). Measuring phylogenetic signal between categorical traits and phylogenies. *Bioinformatics 35*(11), 1862–1869.

Bouckaert, R., P. Lemey, M. Dunn, S. J. Greenhill, A. V. Alekseyenko, A. J. Drummond, R. D. Gray, M. A. Suchard, and Q. D. Atkinson (2012). Mapping the origins and expansion of the Indo-European language family. *Science 337*(6097), 957–960.

Bouckaert, R., D. Redding, O. Sheehan, T. Kyritsis, R. Gray, K. E. Jones, and Q. Atkinson (2022, July). Global language diversification is linked to socio-ecology and threat status.

Bouckaert, R. R., C. Bowern, and Q. D. Atkinson (2018). The origin and expansion of Pama-Nyungan languages across australia. *Nature ecology & evolution 2*(4), 741–749.

Bowern, C. (2012). The riddle of Tasmanian languages. *Proceedings of the Royal Society B: Biological Sciences 279*(1747), 4590–4595.

Bürkner, P.-C. (2017). Bayesian distributional non-linear multilevel modeling with the r package brms. *arXiv preprint arXiv:1705.11123*.

Bürkner, P.-C., N. Schulte, and H. Holling (2019). On the statistical and practical limitations of thurstonian irt models. *Educational and Psychological Measurement 79*(5), 827–854.

Bybee, J. and S. Easterday (2022). Primal consonants and the evolution of consonant inventories. *Language Dynamics and Change*, 1 – 33.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software 80*(1), 1–28.

Carling, G. and C. Cathcart (2021a). Evolutionary dynamics of Indo-European alignment patterns. *Diachronica 38*, 358–412.

Carling, G. and C. Cathcart (2021b). Reconstructing the evolution of Indo-European Grammar. *Language 97*, 1–38.

Carpenter, B., A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell (2017). Stan: A probabilistic programming language. *Journal of statistical software 76*(1).

Cathcart, C. (to appear). Rate variation in language change: toward distributional phylogenetic modeling. In A. Karakostis and G. Jäger (Eds.), *Biocultural Implications: An Agenda for Integrative Approaches*. Tübingen: Kerns.

Cathcart, C., B. Herce, and B. Bickel (2022). Decoupling speed of change and long-term preference in language evolution: insights from romance verb stem alternations.

Cathcart, C. A. (2020). A probabilistic assessment of the Indo-Aryan Inner-Outer hypothesis. *Journal of Historical Linguistics 10*, 42–86.

Cathcart, C. A. (2022). Dialectal layers in West Iranian: a Hierarchical Dirichlet Process approach to linguistic relationships. *Transactions of the Philological Society 120*(1), 1–31.

Cathcart, C. A., A. Hölzl, G. Jäger, P. Widmer, and B. Bickel (2020). Numeral classifiers and number marking in indo-iranian: A phylogenetic approach. *Language Dynamics and Change 1*(aop), 1–53.

Chang, W., C. Cathcart, D. Hall, and A. Garrett (2015). Ancestry-constrained phylogenetic analysis supports the Indo-European Steppe Hypothesis. *Language 91*(1), 194–244.

Chang, W. and L. Michael (2014). A relaxed admixture model of language contact. *Language Dynamics and Change 4*(1), 1–26.

Cysouw, M. (2011). Understanding transition probabilities. *Linguistic Typology 15*, 415–431.

Daumé III, H. (2009). Non-parametric Bayesian areal linguistics. In *Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL*, Boulder, CO, pp. 593–601. Association for Computational Linguistics.

Driver, C. C., J. H. Oud, and M. C. Voelkle (2017). Continuous time structural equation modeling with r package ctsem. *Journal of Statistical Software 77*, 1–35.

Dryer, M. S. (1989). Large linguistic areas and language sampling. *Studies in Language 13*(2), 257–292.

Dunn, M., T. K. Dewey, C. Arnett, T. Eythórsson, and J. Barðdal (2017). Dative sickness: A phylogenetic analysis of argument structure evolution in Germanic. *Language 93*(1), e1–e22.

Dunn, M., S. J. Greenhill, S. C. Levinson, and R. D. Gray (2011). Evolved structure of language shows lineage-specific trends in word-order universals. *Nature 473*(7345), 79–82.

Evans, N. (2019). Linguistic divergence under contact. In *Historical Linguistics 2015*, pp. 564–591. John Benjamins.

Everett, C. (2013). Evidence for direct geographic influences on linguistic sounds: The case of ejectives. *PloS one 8*(6), e65275.

Everett, C., D. E. Blasi, and S. G. Roberts (2015). Climate, vocal folds, and tonal languages: Connecting the physiological and geographic dots. *Proceedings of the National Academy of Sciences 112*(5), 1322–1327.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution 17*(6), 367–76.

Felsenstein, J. (2004). *Inferring phylogenies.* Sunderland, Mass.: Sinauer Associates.

FitzJohn, R. G., W. P. Maddison, and S. P. Otto (2009). Estimating trait-dependent speciation and extinction rates from incompletely resolved phylogenies. *Systematic biology 58*(6), 595–611.

François, A. (2010). The languages of Vanikoro: three lexicons and one grammar. In B. Evans (Ed.), *Discovering history through language: Papers in honour of Malcolm Ross*, Volume 605 of *Pacific Linguistics.* Canberra: Research School of Pacific and Asian Studies, Australian National University.

François, A. (2011). Social ecology and language history in the northern Vanuatu linkage: A tale of divergence and convergence. *Journal of Historical Linguistics 1*(2), 175–246.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2013). *Bayesian data analysis.* Boca Raton, FL: CRC press.

Gelman, A. and J. Hill (2007). *Data analysis using regression and multilevel/hierarchical models.* Cambridge university press.

Geman, S. and D. Geman (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence 6*, 721–741.

Gill, M. S., L. S. Tung Ho, G. Baele, P. Lemey, and M. A. Suchard (2017). A relaxed directional random walk model for phylogenetic trait evolution. *Systematic biology 66*(3), 299–319.

Gray, R. D., A. J. Drummond, and S. J. Greenhill (2009). Language phylogenies reveal expansion pulses and pauses in pacific settlement. *science 323*(5913), 479–483.

Greenberg, J. H., C. E. Osgood, and S. Saporta (1954). Language change. *Psycholinguistics: A survey of theory and research problems*, 146–63.

Greenhill, S. J., P. Heggarty, and R. D. Gray (2021). Bayesian phylolinguistics. In *The Handbook of Historical Linguistics*, pp. 226–253. London: Wiley-Blackwell.

Greenhill, S. J., C.-H. Wu, X. Hua, M. Dunn, S. C. Levinson, and R. D. Gray (2017). Evolutionary dynamics of language systems. *Proceedings of the National Academy of Sciences 114*(42), E8822–E8829.

Guzmán Naranjo, M. and L. Becker (2022). Statistical bias control in typology. *Linguistic Typology 26*(3), 605–670.

Guzmán Naranjo, M. and M. Mertner (2022). Estimating areal effects in typology: a case study of african phoneme inventories. *Linguistic Typology* (0).

Hahn, M. and Y. Xu (2022, June). Crosslinguistic word order variation reflects evolutionary pressures of dependency and information locality. *Proceedings of the National Academy of Sciences 119*(24), e2122604119.

Harmon, L. J. (2019). *Phylogenetic Comparative Methods.* GitHub: https://lukejharmon.github.io/pcm/.

Heggarty, P., C. Anderson, M. Scarborough, B. King, R. Bouckaert, L. Jocz, M. J. Kümmel, T. Jügel, B. Irslinger, R. Pooth, H. Liljegren, R. F. Strand, G. Haig, M. Macák, R. I. Kim, E. Anonby, T. Pronk, O. Belyaev, T. K. Dewey-Findell, M. Boutilier, C. Freiberg, R. Tegethoff, M. Serangeli, N. Liosis, K. Stroński, K. Schulte, G. K. Gupta, W. Haak, J. Krause, Q. D. Atkinson, S. J. Greenhill, D. Kühnert, and R. D. Gray (2023, July). Language trees with sampled ancestors support a hybrid model for the origin of Indo-European languages. *Science 381*(6656), eabg0818.

Horvilleur, B. and N. Lartillot (2014). Monte carlo algorithms for brownian phylogenetic models. *Bioinformatics 30*(21), 3020–3028.

Huelsenbeck, J. P., R. Nielsen, and J. P. Bollback (2003). Stochastic mapping of morphological characters. *Systematic Biology 52*, 131–158.

Huelsenbeck, J. P. and M. A. Suchard (2007). A nonparametric method for accommodating and testing across-site rate variation. *Systematic Biology 56*(6), 975–987.

Ivani, J. and B. Bickel (2024). Databases for comparative syntactic research. In S. Barbiers, N. Corver, and M. Polinsky (Eds.), *The Cambridge Handbook Comparative Syntax*. Cambridge: Cambridge University Press.

Jaeger, T. F., P. Graff, W. Croft, and D. Pontillo (2011). Mixed effect models for genetic and areal dependencies in linguistic typology. *Linguistic Typology 15*, 281–320.

Jing, Y., P. Widmer, and B. Bickel (2023). Word order evolves at similar rates in main and subordinate clauses. *Diachronica 40*, 1–32.

Jäger, G. and J. Wahle (2021). Phylogenetic typology. *Frontiers in Psychology 12*.

Kelly, L. J. and G. K. Nicholls (2017). Lateral transfer in stochastic dollo models. *The Annals of Applied Statistics 11*(2), 1146–1168.

Kirby, S. (2017). Culture and biology in the origins of linguistic structure. *Psychonomic Bulletin & Review 24*, 118–137.

Koile, E., S. J. Greenhill, D. E. Blasi, R. Bouckaert, and R. D. Gray (2022). Phylogeographic analysis of the bantu language expansion supports a rainforest route. *Proceedings of the National Academy of Sciences 119*(32), e2112853119.

Kruschke, J. (2014). Doing bayesian data analysis: A tutorial with r, jags, and stan.

Kruschke, J. K. (2021). Bayesian analysis reporting guidelines. *Nature Human Behaviour 5*(10), 1282–1291.

Lartillot, N. (2013). Phylogenetic patterns of gc-biased gene conversion in placental mammals and the evolutionary dynamics of recombination landscapes. *Molecular biology and evolution 30*(3), 489–502.

Lartillot, N. and R. Poujol (2011). A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters. *Molecular biology and evolution 28*(1), 729–744.

Lemey, P., A. Rambaut, J. J. Welch, and M. A. Suchard (2010). Phylogeography takes a relaxed random walk in continuous space and time. *Molecular Biology and Evolution 27*(8), 1877–1885.

Lupyan, G. and R. Dale (2010). Language structure is partly determined by social structure. *PLoS ONE 5*, e8559.

Macklin-Cordes, J. L., C. Bowern, and E. R. Round (2021). Phylogenetic signal in phonotactics. *Diachronica 38*(2), 210–258.

Maddieson, I., S. Flavier, E. Marsico, C. Coupé, and F. Pellegrino (2013). LAPSyd: Lyon-Albuquerque phonological systems database. *INTERSPEECH 14*.

Maddison, W. P., P. E. Midford, and S. P. Otto (2007). Estimating a Binary Character's Effect on Speciation and Extinction. *Systematic Biology 56*(5), 701–710.

Maslova, E. (2000). A dynamic approach to the verification of distributional universals. *Linguistic Typology 4*, 307–333.

Maurits, L. and T. L. Griffiths (2014). Tracing the roots of syntax with Bayesian phylogenetics. *Proceedings of the National Academy of Sciences 111*(37), 13576–13581.

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan.* CRC press.

Michael, L., W. Chang, and T. Stark (2014). Exploring phonological areality in the circum-andean region using a naive bayes classifier. *Language Dynamics and Change 4*(1), 27–86.

Mimno, D., H. Wallach, and A. McCallum (2008). Gibbs sampling for logistic normal topic models with graph-based priors. In *NIPS Workshop on Analyzing Graphs*, Volume 61.

Moran, S. and D. McCloy (Eds.) (2019). *PHOIBLE 2.0.* Jena: Max Planck Institute for the Science of Human History.

Murawaki, Y. and K. Yamauchi (2018). A statistical model for the joint inference of vertical stability and horizontal diffusibility of typological features. *Journal of Language Evolution 3*(1), 13–25.

Narroll, R. (1961). Two solutions to Galton's Problem. *Philosophy of Science 28*, 15–29.

Nettle, D. (1999). Is the rate of linguistic change constant? *Lingua 108*(2-3), 119–136.

Neureiter, N., P. Ranacher, N. Efrat-Kowalsky, G. A. Kaiping, R. Weibel, P. Widmer, and R. R. Bouckaert (2022). Detecting contact in language trees: a bayesian phylogenetic model with horizontal transfer. *Humanities and Social Sciences Communications 9*(1), 1–14.

Neureiter, N., P. Ranacher, R. van Gijn, B. Bickel, and R. Weibel (2021). Can bayesian phylogeography reconstruct migrations and expansions in linguistic evolution? *Royal Society open science 8*(1), 201079.

Nikolaev, D. and E. Grossman (2020). Consonant co-occurrence classes and the feature-economy principle. *Phonology 37*(3), 419–451.

Niyogi, P., R. C. Berwick, et al. (1997). A dynamical systems model for language change. *Complex Systems 11*(3), 161–204.

O'Meara, B. C., C. Ané, M. J. Sanderson, and P. C. Wainwright (2006). Testing for different rates of continuous trait evolution using likelihood. *Evolution 60*(5), 922–933.

Padilla-Iglesias, C., E. Gjesfjeld, and L. Vinicius (2020). Geographical and social isolation drive the evolution of austronesian languages. *PloS one 15*(12), e0243171.

Pagel, M. (1994). Detecting correlated evolution on phylogenies: A general method for the comparative analysis of discrete characters. *Proceedings of the Royal Society of London B 255*, 37–45.

Pritchard, J. K., M. Stephens, and P. Donnelly (2000). Inference of population structure using multilocus genotype data. *Genetics 155*(2), 945–959.

Ranacher, P., N. Neureiter, R. van Gijn, B. Sonnenhauser, A. Escher, R. Weibel, P. Muysken, and B. Bickel (2021). Contact-tracing in cultural evolution: a bayesian mixture model to detect geographic areas of language contact. *Journal of The Royal Society Interface 18*(181), 20201031.

Rangelov, T., M. Walworth, and J. Barbour (2023). A multifaceted approach to understanding unexpected sound change: The bilabial trills of vanuatu's malekula island. *Diachronica.*

Reesink, G., R. Singer, and M. Dunn (2009). Explaining the linguistic diversity of Sahul using population models. *PLoS Biology 7*, e1000241.

Revell, L. J. (2013). Two new graphical methods for mapping trait evolution on phylogenies. *Methods in Ecology and Evolution 4*, 754–759.

Rigby, R. A. and D. M. Stasinopoulos (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society Series C: Applied Statistics 54*(3), 507–554.

Ringen, E., J. S. Martin, and A. Jaeggi (2021). Novel phylogenetic methods reveal that resource-use intensification drives the evolution of "complex" societies. *EcoEvoRxiv.*.

Robert, C., G. Casella, C. P. Robert, and G. Casella (2010). Metropolis–hastings algorithms. *Introducing Monte Carlo Methods with R*, 167–197.

Roberts, S. G. (2018). Robust, causal, and incremental approaches to investigating linguistic adaptation. *Frontiers in psychology 9*, 166.

Ross, C. T., P. Strimling, K. P. Ericksen, P. Lindenfors, and M. B. Mulder (2016). The origins and maintenance of female genital modification across africa: Bayesian phylogenetic modeling of cultural evolution under the influence of selection. *Human Nature 27*, 173–200.

Sapir, E. (1916). *Time perspective in aboriginal American culture, a study in method*, Volume 90 of *Geological Survey Memoir.* Ottawa, CA: Government Printing Bureau.

Šavrič, B., T. Patterson, and B. Jenny (2019). The equal earth map projection. *International Journal of Geographical Information Science 33*(3), 454–465.

Shcherbakova, O., S. M. Michaelis, H. J. Haynie, S. Passmore, V. Gast, R. D. Gray, S. J. Greenhill, D. E. Blasi, and H. Skirgård (2023). Societies of strangers do not speak less complex languages. *Science Advances 9*(33), eadf7704.

Shirtz, S., L. Talamo, and A. Verkerk (2021). The evolutionary dynamics of negative existentials in indo-european. *Front. Commun. 6: 661862. doi: 10.3389/fcomm.*

Shringarpure, S. and E. P. Xing (2009). mStruct: inference of population structure in light of both genetic admixing and allele mutations. *Genetics 182*(2), 575–593.

Sisson, S. A. and Y. Fan (2011). Likelihood-free mcmc. *Handbook of Markov Chain Monte Carlo*, 313–335.

Stone, J. V. (2013). Bayes' rule: a tutorial introduction to bayesian analysis.

Syrjänen, K., T. Honkola, J. Lehtinen, A. Leino, and O. Vesakoski (2016). Applying population genetic approaches within languages: Finnish dialects as linguistic populations. *Language Dynamics and Change 6*, 235–283.

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of dna sequences. *Lectures on Mathematics in the Life Sciences 17*(2), 57–86.

Urban, M. (2021). The geography and development of language isolates. *Royal Society open science 8*(4), 202232.

Wichmann, S., E. W. Holman, and C. H. Brown (2018). The ASJP database (version 18). http://asjp.clld.org/.

Widmer, M., S. Auderset, P. Widmer, J. Nichols, and B. Bickel (2017). NP recursion over time: Evidence from Indo-European. *Language 93*, 1–36.

Widmer, M., M. Jenny, W. Behr, and B. Bickel (2021). Morphological structure can escape reduction effects from mass admixture of second language speakers: Evidence from Sino-Tibetan. *Studies in Language 45*, 707–752.

Witzlack-Makarevich, A., T. Zakharko, L. Bierkandt, F. Zúñiga, and B. Bickel (2016). Decomposing hierarchical alignment: Co-arguments as conditions on alignment and the limits of referential hierarchies as explanations in verb agreement. *Linguistics 54*, 531–561.

Wurm, S. A. (1967). Linguistics and the prehistory of the south-western pacific. *The journal of Pacific history 2*(1), 25–38.

Yang, C. D. (2000, October). Internal and external forces in language change. *Language Variation and Change 12*(3), 231–250.

Yu, G., T. T.-Y. Lam, H. Zhu, and Y. Guan (2018). Two methods for mapping and visualizing associated data on phylogeny using ggtree. *Molecular Biology and Evolution 35*, 3041–3043.

Zhou, K. and C. Bowern (2015). Quantifying uncertainty in the phylogenetics of Australian numeral systems. *Proceedings of the Royal Society B: Biological Sciences 282*, 20151278.