The evolution of similarity avoidance: a phylogenetic approach to phonotactic change

Chundra A. Cathcart
Department of Comparative Language Science, University of Zurich
DFG Center for Advanced Studies – Words, Bones, Genes, Tools, University of Tübingen

1 Introduction

The worlds languages vary dramatically in terms of the phonotactic patterns they allow, as well as the frequencies of different patterns they display. For instance, hypothetical **bnick* is not well-formed in English, but similar forms are valid in other languages, e.g., Moroccan Arabic *bniqa* 'closet' (Gorman, 2011). Furthermore, lineage-specific phonotactic patterns exhibit high diachronic stability and phylogenetic signal (Macklin-Cordes et al., 2021).

At the same time, a number of recurrent trends have been identified in large numbers of genetically diverse languages with respect to the sound patterns they display. For instance, robust sound-meaning correspondences, partly rooted in perceptuomotor analogies, have been detected cross-linguistically (Blasi et al., 2016; Johansson et al., 2020). Additionally, recent work points to statistical underrepresentation of voiced velar plosives relative to other sounds, measured according to type frequencies in word lists (Everett, 2018). Finally, a large body of research speaks to the statistical underrepresentation of consonants sharing a place of articulation within lexical items, documented in a diverse sample of languages, a phenomenon known as similar place avoidance.

All of the patterns mentioned above are conceivably rooted in communicative pressures; at the same time, the mechanisms involved in their emergence and maintenance remain poorly understood. In this paper, I investigate the factors that underlie the persistence of similar place avoidance in a large sample of Austronesian languages using phylogenetic methods, a popular methodology imported from computational biology for use in typology. I find that forms violating constraints on similar place avoidance (by exhibiting identical adjacent consonants) are more likely to die out than forms adhering to such constraints. Attempts to investigate how constraints on similar place avoidance interact with more and less frequent vocabulary items produce more ambiguous results. I discuss the value of such tools for exploring the evolution of sound patterns, and also discuss some limitations of the implementation used in the paper to be improved upon.

2 Similar place avoidance

Categorical and gradient constraints on similar or identical adjacent elements are documented at a number of linguistic levels (Nevins, 2012). Repetition of formally identical elements tends to be dispreferred within English words (e.g., **sillily*, **friendlily* etc.). Additionally, some languages do not allow identical case markers to appear on adjacent words (Allen, 1984), exhibiting identical element avoidance at the sentence level. The extent to which similarity avoidance obtains in non-human communication systems has not been fully investigated. Analyses of *pyow-hack* call sequences in Putty-nosed monkeys reveal a high degree of tolerance for adjacent identical elements (Arnold & Zuberbühler, 2012; Schlenker et al., 2016). However, while Black-and-white Colobus monkeys can produce adjacent roars, adjacent snorts cannot occur without intervening pauses (Schel et al., 2009).

In phonology, the statistical avoidance of adjacent consonants agreeing in place of articulation is reported in a variety of languages from different genetic groups (Greenberg 1950; Buckley 1997; Berkley 2000; Frisch et al. 2004; Pozdniakov & Segerer 2007; Coetzee & Pater 2008; Rácz et al. 2016; Grotberg 2022, a.o.). This phenomenon is thought to represent a gradient instantiation of the obligatory contour principle (OCP), a constraint against adjacent identical elements (McCarthy, 1986). The majority of the literature on this subject is concerned with measuring the strength of avoidance constraints across different places of articulation as

Proceedings of *AMP 2022*

well as non-place features. The effect has been quantified in different ways, e.g., via the ratio of the number of observed consonant pairs versus the number expected under chance (Frisch et al., 2004), log-linear models (Wilson & Obdeyn 2009; see also Stanton & Stanton 2022), and logistic regression (Graff & Jaeger, 2009). Constraints on co-occurrence vary across place of articulation and interact with non-place features (Coetzee & Pater, 2008), though the nature of these interactions is debated (Wilson & Obdeyn, 2009).

Experimental literature provides support for the idea that identical consonants within words are problematic from the perspective of processing and production. Participants in lexical decision tasks are slower to accept words and faster to reject non-words containing identical consonants at any distance (van de Weijer, 2005). Additionally, while repeated syllables are easier for children to produce and learn, adults exhibit a faster speech rate for sequences of different syllables (Lancheros et al., 2020); this is all the more interesting in that identical consonants are common in nursery words (e.g., *mama*, *cookie*, etc.).

While similar place avoidance is an uncontroversial property of human languages and is plausibly rooted in constraints on processing and/or production, little is understood regarding the specific diachronic mechanisms that bring about this synchronic state of affairs. A possibility is that sound change operates in some capacity to ensure that sequences of consonants with identical place of articulation arise infrequently. However, compendia of sound changes do not indicate that dissimilation in place of articulation, a sound change that would directly result in similar place avoidance, is more frequent than assimilation in place of articulation (Kümmel 2007; see also Pozdniakov & Segerer 2007); this hypothesis is also at odds with a view of sound change as a non-teleological process that does not directly serve to bring about synchronically optimal configurations (Bach & Harms, 1972; Ohala, 1993), though at the same time there is some evidence that "sporadic" sound changes can operate under some circumstances, e.g., for the purpose of homophony avoidance (Blevins & Wedel, 2009); it is therefore not inconceivable that sporadic sound change or some sort of related process (e.g., analogical change) could serve to bring about similar place avoidance in certain lexical items. Another view found in the literature but as yet untested empirically on a large scale hypothesizes that words containing adjacent consonants with identical place of articulation are rare due to dynamics of lexical usage: items containing phonotactically suboptimal patterns are unlikely to be coined or borrowed into a language, and when present, are likely to be phased out of use (Frisch et al., 2004; Martin, 2007; Pozdniakov & Segerer, 2007). This paper explores the diachronic factors that give rise to similar place avoidance. I use a phylogenetic model of lexical evolution that quantifies the support for different diachronic scenarios. Specifically, we can test the prediction that if a dispreferred phonotactic pattern arises in a lexical item (e.g., via regular sound change), it is likely to become more marginal in its usage, losing out to potential competitors. Additionally, we can assess whether adjacent consonants sharing a place of articulation are less likely to arise in less marginal vocabulary items than in more marginal ones, and more likely to be lost; this might indicate that some force other than lexical usage is responsible for mediating the patterns we observe synchronically.

3 Phylogenetic modeling

Attempts to quantify cross-linguistic diversity encounter the need to control for genetic sources of non-independence, a phenomenon known as Galton's problem (Naroll, 1961). Standard approaches for dealing with this problem include stratified sampling, limiting analyses to one language per genetic grouping (Dryer, 2000) and mixed-effects regression, controlling for historical and spatial relatedness when estimating cross-linguistic preferences for a feature or the effect of one feature on another (Jaeger et al., 2011; Naranjo & Becker, 2022). These approaches can be contrasted with phylogenetic models that explicitly specify the transmission and diffusion processes thought to give rise to the diversity we observe, rather than treating them as nuisance factors (Cathcart, 2018).

Phylogenetic methods in linguistics are taken over from computational biology, where they are used to infer phylogenies of organisms and investigate the evolutionary dynamics of different traits or features. Many of these models assume that discrete features evolve according to a continuous-time Markov (CTM) chain, a stochastic process parameterized according to transition rates governing the frequency of changes between different feature values. In linguistics, these methods have been used to infer chronologically detailed phylogenetic representations of different language families using lexical cognacy data (Gray & Atkinson, 2003; Gray et al., 2009; Bouckaert et al., 2012; Honkola et al., 2013; Chang et al., 2015; Sagart et al., 2019). This work involves the use of Bayesian methods which relax the assumption of uniform lexical

replacement thought to be fatal to the enterprise of glottochronology (Bergsland & Vogt, 1962), leading to more realistic chronologies. An equally fruitful area of research involves the application of phylogenetic comparative methods to questions in linguistics. Given an existing phylogenetic representation of a sample of languages, this diverse family of models facilitates hypothesis testing regarding a wide range of facets of language change. A popular method is the Discrete model of correlated evolution (Pagel, 1994), which tests whether a feature is more likely to arise when another feature is already present (or absent); this and related models have been used to investigate dependencies between features in language change (Dunn et al., 2011; Haynie & Bowern, 2016; Cathcart et al., 2020).

Phylogenetic models are not without their limitations. Under most models, change can only be modeled between states that are attested within the language sample. Furthermore, these models' use is generally restricted to large, well-studied phylogenies in order to generate reliable parameter estimates (though see Jäger & Wahle 2021, which extends these methods to small families and isolates). Standard phylogenetic models are incapable of teasing apart genetic and areal pressures (though see Kelly & Nicholls 2017; Neureiter et al. 2022). Additionally, some degree of simplification of feature representations is often required for model tractability. At the same time, these methods provide a powerful means of explicitly testing hypotheses regarding diachronic change, and can be evaluated via careful model criticism (including comparing ancestral state reconstructions produced by the model with received wisdom from historical linguistics).

4 Data

Phylogenetic methods provide a means of investigating the evolution of cognate lexical items over a phylogeny, taking into account whether they are more likely to grow more marginal in usage and die out in particular lineages if a dispreferred phonotactic pattern arises. Questions of this sort call for data sets that code REFLEX words in related languages according to the ancestral ETYMA from which they descend regardless of their meaning. Resources of this sort are in fact quite rare: while there exist a number of lexical databases that provide cognacy judgments for words sharing a basic meaning (Greenhill et al., 2008; Kaiping & Klamer, 2018), there are relatively few computationally tractable databases that code cross-semantic cognacy. An obvious reason for this discrepancy is that cognacy judgments within concept slots (e.g., asking whether French *chien* and English *dog* are cognate) are easier for analysts with a moderate degree of expertise to carry out quickly and can even be automated to some extent (List, 2012; Jäger, 2013; Rama, 2016), while resources organized around etyma require considerable specialist knowledge of the languages under study as well as the historical developments affecting them (e.g., to know that Latin *crabro* 'hornet' and Sanskrit *śiras* 'head' are cognate); even under optimal circumstances, certain etymological connections affecting obscure forms may go undetected.

I made use of data from the Austronesian Comparative Dictionary (ACD; Blust & Trussel 2013), available via Lexibank (List et al. 2022a; data were downloaded from `https://github.com/lexibank/acd/tree/main/cldf` on 4 September, 2022), which organizes words in 1019 Austronesian languages according to the etyma in Proto-Austronesian and intermediate reconstructed languages from which they descend. Cognacy is coded at both the root and word level; e.g., Acehnese *lakòë* 'husband', Tagalog *laláke* 'man, an adult male; male, masculine' and Malagasy *laláhy* 'man (provincial)' are all cognates at the root level, reflecting Proto-Malayo-Polynesian *laki, but only the latter two forms are cognates at the word level, descending from derived Proto-Western-Malayo-Polynesian *la-laki).

Transcriptions in the ACD are not normalized across languages, making it difficult to easily extract phonological features for individual segments. For this reason, I restricted the scope of my study to pairs of IDENTICAL CONSONANTS SEPARATED BY A SINGLE VOWEL (IC), a feature that was straightforward to extract, with a few caveats. For example, in many cases, it is difficult to distinguish between tautosyllabic long vowels and broken long vowels (Zuraw, 2018), so this distinction was not taken into consideration.

An issue that arose in the process of detecting identical consonants involved the way in which morpheme boundaries are represented in the ACD. Most formulations of constraints on similar place avoidance make reference to co-occurring consonants within uninflected simplex morphological forms, though work on phonotactic generalizations also investigates constraints within complex forms and sentences (Martin, 2011; Breiss & Hayes, 2020). In Austronesian languages in particular, co-occurrence rates of consonants with identical place of articulation differ across tautomorphemic and heteromorphemic contexts, given the frequent

occurrence of reduplication and infixation processes that create identical adjacent consonants in derived forms (Rácz et al., 2016; Zuraw & Lu, 2009). Accordingly, models may infer different degrees of diachronic tolerance for identical consonants, depending on whether only tautomorphemic sequences are taken into consideration. An obstacle to considering only tautomorphemic sequences is the fact that the ACD marks affix and infix boundaries that were active in ancestral forms but not necessarily active in the reflexes where they are marked. As an example, the ACD gives the Aklanon word for 'woman' as *ba-bayi* on the basis of reduplicated Proto-Austronesian *ba-bahi, even though a morpheme boundary is not marked in the source from which the word is taken (Zorc, 1969) and the form is presumably synchronically tautomorphemic. Coding only the presence of identical consonants within hyphen-delimited forms after stripping out infixes runs the risk of severely under-counting tautomorphemic violations of IC avoidance. A potential solution to this issue, not undertaken here, would be to treat hyphens in the data as representing synchronically active morpheme boundaries in a language only if a base form exists along the putatively derived form in the same language and the two forms are in a semantically transparent relationship. For the purposes of this paper, I generated two data sets, one coding the occurrence of adjacent identical consonants at the MORPH level (i.e, in hyphen-delimited environments) versus at the WORD level (i.e., in whitespace-delimited environments) in order to see whether results hold across both data sets.

The literature described in previous sections predicts that if identical adjacent consonants arise in a word, the word is likely to become more marginal in use. Ideally, we would use frequency as a proxy for a word's marginality, but word frequency values are no available for all languages in the ACD. As an alternative, I coded forms in the ACD according to whether or not they were present in the Austronesian Basic Vocabulary Database (ABVD; Greenhill et al. 2008), available via Lexibank (data were downloaded from `https://github.com/lexibank/abvd/raw/master/cldf` on 4 September 2022). Basic vocabulary items, e.g., those found in Swadesh lists and related taxonomies, tend to be more frequent than non-basic items in the world's languages (Calude & Pagel, 2011), and semantic shifts between basic and non-basic meaning are common, e.g., Latin *pellis* 'pelt, hide' (non-basic) > French *peau* 'skin' (basic). This distinction provides a means, albeit a coarse one, of exploring differences in the evolution of similarity avoidance in both high-frequency and low-frequency vocabulary items that descend from the same etyma.

Phylogenetic comparative methods like the one used here require the use of a phylogenetic representation of the languages under study. I used the phylogeny of Gray et al. (2009), which provides a dated Bayesian tree sample for several hundred Austronesian languages.

The data processing workflow I employed was as follows: first, I used an iterative algorithm (Needleman & Wunsch, 1970; Jäger, 2013) to align each reconstructed etymon with the portion of each corresponding entry most likely to descend from it. The purpose of this was to minimize the risk, particularly in multi-word expressions, of extracting the presence of identical consonants in an element not homologous with the etymon whose evolution is being tracked. Alignment was carried out at both the word level and the morph level. For each form, a script was used to automatically extract whether identical consonants separated by a single vowel were present in the space-delimited word and hyphen-delimited morph aligned with the etymon ancestral to the form. In subsequent processing steps, I retained only forms corresponding to an etymon if both presence and absence of identical consonants were attested among the reflexes of the etymon under consideration at either the word or morph level, as etyma lacking variation of this sort are uninformative with respect to the diachronic dynamics of this feature. Following the detection of IC presence/absence, a script was used to detect which forms in the data set were present in the ABVD and thus part of the basic vocabulary of the language containing them.

To generate data sets for analyses, I retained data from 106 languages that were present in the phylogeny of Gray et al. (2009) and were attested at least 100 times in the data set described above. The rationale here was to use data from well-studied, thoroughly etymologized languages, since the absence of a descendant form of an etymon in a poorly studied language may be due either to its actual absence or because its etymology has not yet been determined by historical linguists.

In total, four data sets were generated: two data sets containing reflexes of etyma showing PRESENCE/ABSENCE of IC at the word level and the morph level, and two datasets containing reflexes of etyma attesting all possible combinations of ±IC, ±BASIC at the word and morph level. Reflexes of etyma that were present in at least 10% of languages were retained. In the PRESENT data sets, etyma were represented by word IDs, since the evolution of derived elements is in a sense more tangible than that of more abstract roots; however, the BASIC data sets took root IDs to represent etyma, since there were relatively few word

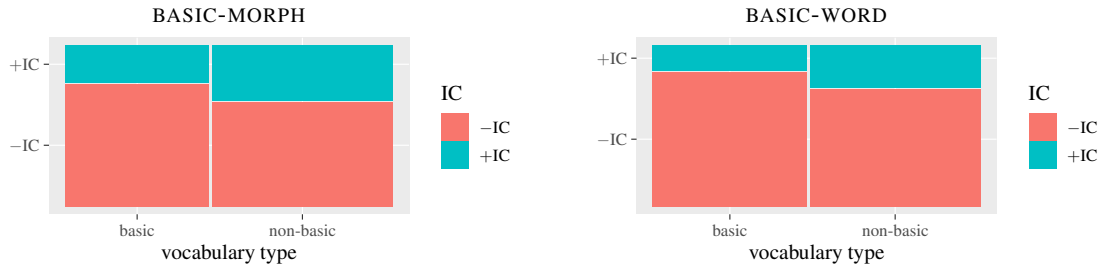

Figure 1: Joint distributions of the presence of IC and vocabulary type for BASIC-MORPH and BASIC-WORD data sets.

IDs giving rise to all possible combinations of ±IC, ±BASIC. The number of etyma in each data set were 149 (PRESENT-WORD), 55 (PRESENT-MORPH), 138 (BASIC-WORD), and 56 (BASIC-MORPH).

A final processing step for phylogenetic analysis was to convert the data sets into likelihood matrices, setting values attested for a given etymon in a given language to 1 and all unattested values to 0 (values consist of {ABSENT, −IC, +IC} for the PRESENT data set, and {ABSENT, (−IC,−BASIC), (−IC,+BASIC), (+IC,−BASIC), (+IC,+BASIC)} for the BASIC data set). Since languages often attest more than one value for a given etymon, some languages had multiple likelihoods set to one for different etyma. It is worth highlighting that this is a method for dealing with data ambiguity in cladistics rather than actual polymorphism (Felsenstein, 2004).

Mosaic plots displaying the joint distributions of the presence of IC and vocabulary type for BASIC-MORPH and BASIC-WORD data sets are found in Figure 1. I carried out mixed-effects logistic regression analyses of the BASIC data sets using the R package lme4 (Bates et al., 2015) to assess the effect of basic/nonbasic vocabulary type on the presence of IC while controlling for group-level idiosyncrasies at the language and etymon level via random intercepts. Global intercepts were negative and significant, indicating an overall dispreference for IC (BASIC-MORPH: $\beta_0 = -0.7232, p < 0.001$; BASIC-WORD: $\beta_0 = -1.16195$, $p < 0.001$); the effect of basic vocabulary on the presence of IC is negative for both data sets, indicating that IC are less likely to be found in basic than in non-basic vocabulary, but significant in only one (BASIC-MORPH: $\beta = -0.1846, p = 0.138$; BASIC-WORD: $\beta = -0.53186, p < 0.001$); the effect of vocabulary type on IC presence is significant for the BASIC-MORPH data set if the random intercept by etymon ID is omitted.

5 Method

Under a CTM process, etyma undergo transitions between the state ABSENT and the various states that cognate sets can express during their evolution. Computational biology provides a number of techniques for modeling anatomically dependent traits such as tail color, which is relevant only if tails are present in an organism (Maddison, 1993; Tarasov, 2019); the question of whether IC are present in a cognate lexical item in a language is only applicable if the etymological item in question has survived into the language. Unlike tails and other anatomical characters, once a cognate is lost in a linguistic lineage, it cannot arise again (in the absence of philologically informed revitalization efforts for which there is little evidence prior to the contemporary period), though it could in theory be borrowed from a geographically and genetically proximate language.

Figure 2 provides schemata of the two CTM processes assumed in this paper for the PRESENT and BASIC data sets. Under these models, a lexical item corresponding to an etymon in the ACD is born once, and transitions between different states (represented by non-gray nodes) over the course of its evolution before dying. Dotted lines represent birth events, which can happen at most once during a etymon's history. The state DEAD, represented by a node with a doubled boundary, is an absorbing state; once an etymon dies out, it cannot be reborn. For the BASIC model, I assume that state transitions are incremental, and that simultaneous changes for the features ±IC and ±BASIC do not occur. I additionally assume that basic vocabulary items do not die out directly, but spend some time as non-basic items before falling out of use.

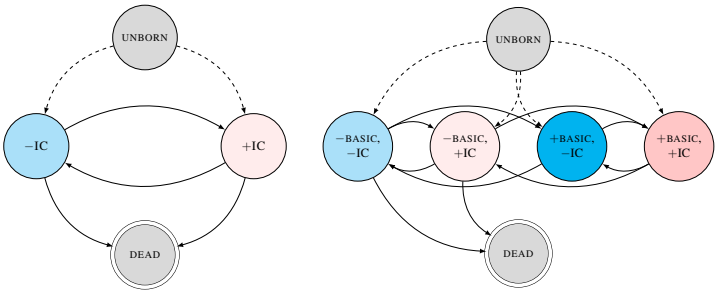

Figure 2: Evolutionary models employed in this paper, representing states which an etymological lexical item can visit according to a CTM process. Dotted lines represent transitions that can happen maximally once in an item's history; a doubled border around a node indicates an absorbing state, i.e., one that cannot be exited.

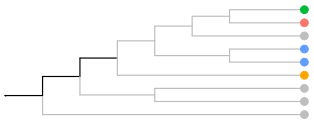

Figure 3: Visualization of birth loci on a phylogenetic tree according to the SDC model. Births can only occur once in a trait's history, and can only occur on branches that are ancestral to all languages where the trait is present (black branches; gray nodes indicate languages where the trait is absent).

The Stochastic Dollo character (SDC) model (Nicholls & Gray, 2006) of character evolution enforces a single-birth criterion (see Figure 3 for a visualization). Originally designed to model the evolution of linguistic root-meaning traits (which code whether a particular concept belongs to some cognate class in a given language), SDC has been shown to be inappropriate for modeling this particular data type, since root-meaning traits can arise in parallel (Chang et al., 2015). At the same time, the SDC model may be appropriate for representing the diachronic development of etymological items, since they cannot be reborn after dying, except in highly exceptional situations.

Alekseyenko et al. (2008) describe a version of the SDC for multistate traits that allows death rates to vary across states, providing an efficient way to compute a quantity proportional to the likelihood of the model under the observed data and a phylogeny. One disadvantage of this model is that it requires the initial state of a trait to be drawn from the stationary distribution of the CTM process characterizing its evolution in order for the likelihood to have an analytical solution. The authors imply that an analytical solution may be available if the initial state is assumed to have a specific value, but do not provide a formula for the likelihood under such a scenario. If true, we could fix the initial state to follow specialist reconstructions (e.g., if a proto-form is reconstructed with identical consonants, we could assume that the initial state of the etymon had this pattern); an alternative to using the SDC model might assume that an etymon is born on the branch directly ancestral to the proto-language where it is reconstructed by specialists.

Eight models were fitted in total. For each of the four data sets (PRESENT/BASIC × WORD/MORPH), I carried out posterior inference for two models. The first of these is a hierarchical Bayesian model, where transitions between states take place according to global transition rates that are allowed to vary in a constrained manner across individual features (i.e., etyma in the data set) according to rate multipliers. The second of these is a flat Bayesian model, which assumes that transitions between individual pairs of states occur at the same rate across individual etyma. More detailed model specification and details regarding inference can be found in the Appendix. Model comparison could ideally be used to assess whether the hierarchical or flat model provides a better fit to each data set (Vehtari et al., 2017; Yao et al., 2017) but these methods require a normalized likelihood, which I do not use due to the computationally intensive nature of computing the normalization constant.

6 Results

Posterior samples of rate parameters inferred during model fitting were used to generate quantities of interest to hypotheses regarding the evolution of IC. We can quantify whether one transition takes place more frequently than another by taking the difference between transition rates across posterior samples and analyzing the distribution of the differences. A difference in rates is taken to be decisive if the 95% highest density interval (HDI) of the difference excludes zero.

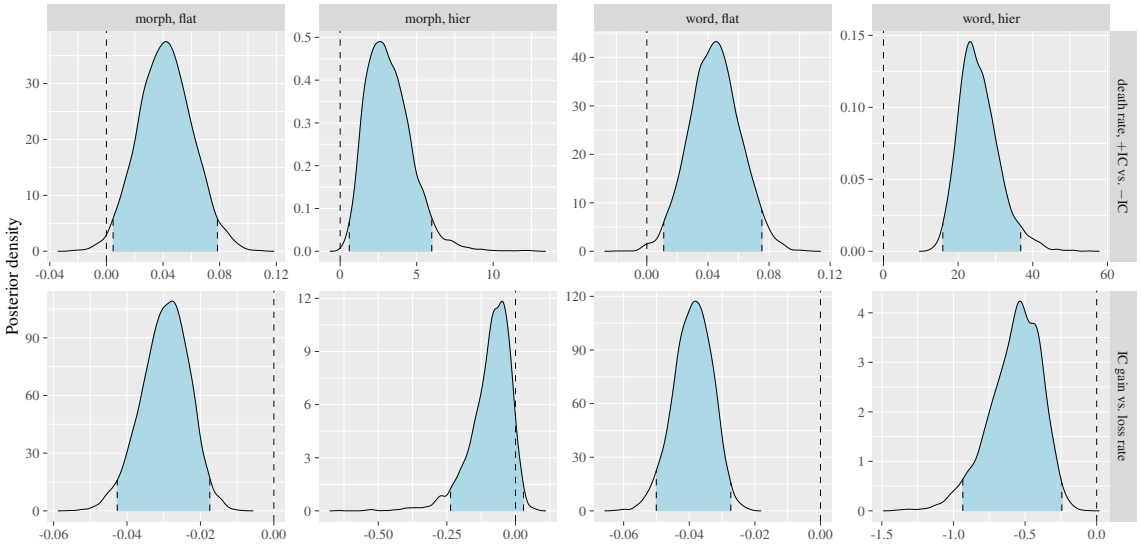

Figure 4: Posterior rate differences of interest under different variations of the PRESENT model. The top panels show the difference between the rate at which forms with +IC die and the rate at which forms with −IC die. Positive values indicate that forms with +IC die more frequently than those without. The bottom panels show the difference between the gain rate and loss rate of +IC; negative values indicate that transitions of the type +IC → −IC are more frequent than those of the type −IC → +IC. Shaded areas indicate 95% HDIs.

Figures 4 and 5 display density curves for posterior differences of interest under different versions of the PRESENT and BASIC models, respectively, with the 95% HDI shaded. Results for all versions of the PRESENT model indicate that forms with IC are more likely to die out than those without, but not all versions support the notion that IC are lost more frequently than gained. Both the flat and hierarchical BASIC-WORD models failed to reach convergence, so I display only results from the BASIC-MORPH models, given the difficulty of interpreting multimodal posterior distributions. This failure to converge points to model misspecification and/or problems with the data used; a full investigation into the factors involved is outside the scope of this paper. Results from the BASIC-MORPH models are also difficult to interpret in that the models do not agree in their overall behavior. The flat model indicates that counter to our expectations, the death rate for the state (+IC,−BASIC) is decisively *lower* than the death rate for the state (−IC,−BASIC). At the same time, results from this model suggest that IC are gained with decisively lower frequency in basic than in non-basic vocabulary items, which might account for their underrepresentation. However, these effects do not obtain in the hierarchical model, suggesting that there is too much variation at the etymon level for a clear trend to be detected.

7 Discussion and future directions

Phylogenetic models analyzing the evolution of presence/absence of IC in etymologically related cognate sets found that forms with IC die out more frequently than forms without IC, and showed partial support for the idea that IC are gained in general more frequently than they are lost. Phylogenetic models investigating diachronic interactions between IC and vocabulary type provided inconclusive results. It is possible that the basic/nonbasic distinction is too coarse a proxy for frequent vs. infrequent vocabulary items, and that it fails to capture the dynamics of forms in different frequency bins in a meaningful manner. Barring the collection of word frequencies for languages in a sample as large as that of the ACD, if glosses were reconciled with the Concepticon taxonomy (List et al., 2022b), the basic/non-basic vocabulary distinction can be abandoned in favor of a more granular metric correlated with frequency such as concept rank (Dellert & Buch, 2018).

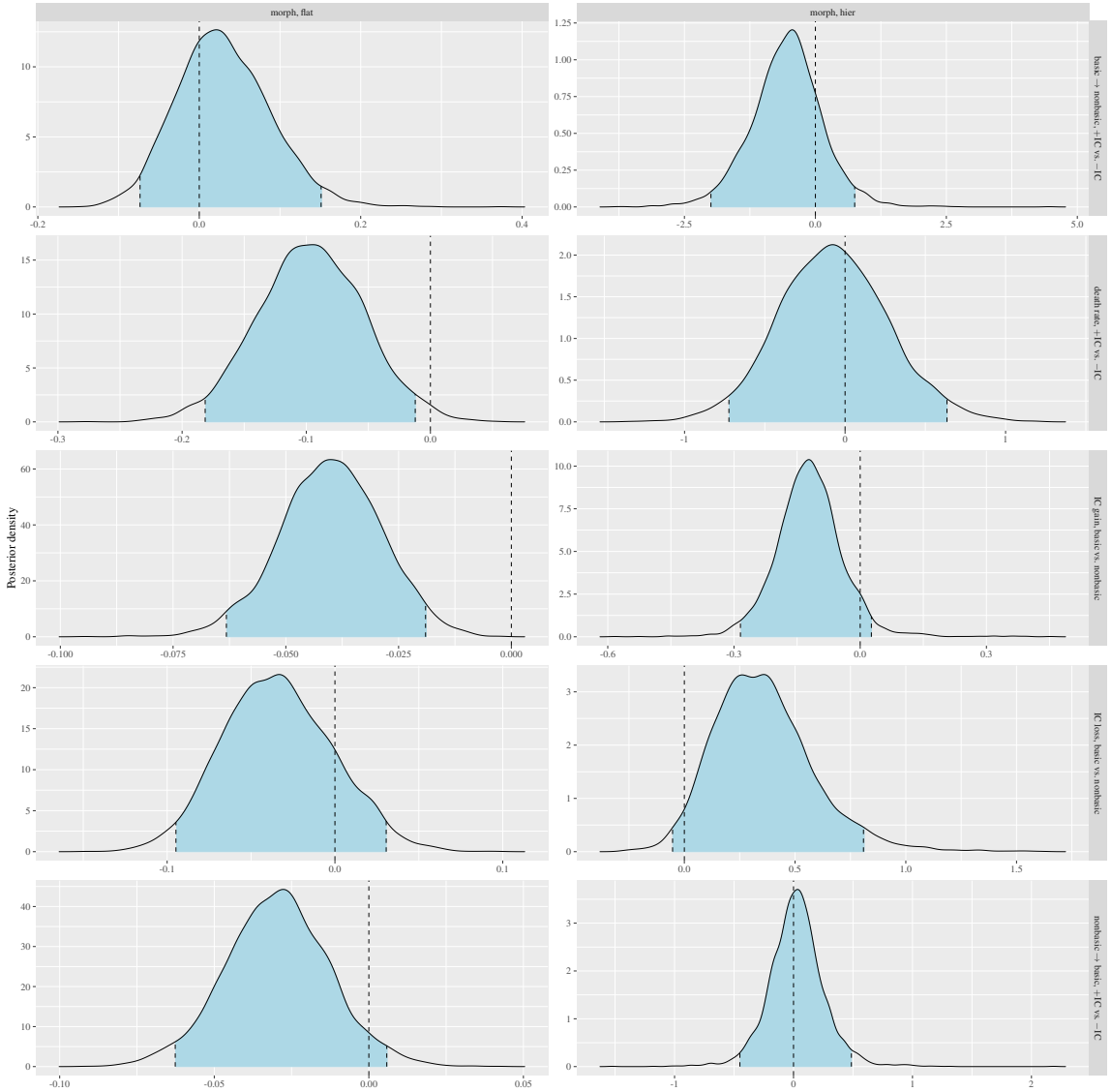

Figure 5: Posterior rate differences of interest under different variations of the BASIC model. Panels show, from top to bottom, (1) the rate of moving from basic to non-basic vocabulary when IC are present versus absent, (2) the death rate for non-basic forms with IC versus without, (3), the gain rate for IC in basic versus non-basic vocabulary, (4) the loss rate for IC in basic versus non-basic vocabulary, and (5) the rate of moving from non-basic to basic vocabulary when IC are present versus absent. Shaded areas indicate the 95% highest density interval.

PRESENT/ABSENT		
+IDCC	→	−IDCC
−IDCC	→	+IDCC
+IDCC	→	DEAD
−IDCC	→	DEAD

BASIC/NON-BASIC		
−BASIC, −IDCC	→	−BASIC, +IDCC
−BASIC, −IDCC	→	+BASIC, −IDCC
−BASIC, +IDCC	→	−BASIC, −IDCC
−BASIC, +IDCC	→	+BASIC, −IDCC
+BASIC, −IDCC	→	+BASIC, +IDCC
+BASIC, −IDCC	→	−BASIC, −IDCC
+BASIC, +IDCC	→	−BASIC, +IDCC
+BASIC, +IDCC	→	+BASIC, −IDCC
−BASIC, −IDCC	→	DEAD
−BASIC, +IDCC	→	DEAD

Table 1: State transitions under the different models used in this paper

This paper used an SDC model of character evolution for the data under study, under which etyma are born and move between different states before dying. SDC constrains traits to be born only once. An etymon should arise only once on one branch of a phylogeny, unless it enters another lineage (presumably a geographically proximate one) due to contact. The SDC model used here did not account for the latter possibility, and is perhaps too strict in that regard. The SDC model has other disadvantages, among them the assumption that the initial state of an etymon, upon being born, is drawn from the stationary process of the CTM process characterizing its evolution. It would be ideal to allow the initial state of the system to be independent from the CTM process under which the system changes, as this could help to address questions as to whether etyma are more likely to be born without IC than with. Of course, if we assume that specialist reconstructions are correct (and there is no good reason not to), we can simply calculate this probability by observing reconstructions; at the same time, sophisticated hierarchical models may allow us to make inferences regarding unascertained etyma that have never been observed.

An issue not fully unpacked in this paper concerns the way in which we should interpret gains and losses of IC under a phylogenetic model. A number of developments, including regular sound change, analogical change, and derivational processes can serve to bring about or remove IDCC in a form. The model used here is not an explicit model of sound change (Hruschka et al., 2013; Bouchard-Côté et al., 2013; Jäger, 2019), but is simply sensitive to the presence of a pattern in a cognate set (cf. Blasi et al., 2019). The fact that IC may arise and disappear at different rates across different vocabulary types does not necessarily point to irregular sound change, but may simply be a coincidence. Though difficult to implement, a more explicit model of sound change might be able to tell us, for example, whether some putatively regular sound change would be blocked if it were to give rise to IC in a frequent form.

Phylogenetic models are particularly good at modeling lexical history, and have generated phylogenies dovetailing with received wisdom on the basis of models of lexical character evolution. Though some refinement is needed, approaches of this sort can potentially shed light on multiple aspects of the diachrony of phonotactic constraints.

Appendix

State transitions allowed under the PRESENT/ABSENT and BASIC/NON-BASIC models are given in Table 1. For each model, let R denote the number of transition types and D the number of etyma in the data sets used. For the hierarchical version of each model, I set the transition rate for each state transition type for each etymon $\tau_{d,r} : r \in \{1, ..., R\}, d \in \{1, ..., D\}$ for each etymon to be equal to $\rho_r \sigma_{d,r}$. The parameter ρ_r is a global transition rate, while $\sigma_{d,r}$ is an etymon-level rate multiplier. We place the priors Gamma(α, α) and Gamma(β_d, β_d) over ρ_r and $\sigma_{d,r}$ respectively, where $\alpha, \beta_d : d \in \{1, ..., D\} \sim$ Exp(1.5) are hyperparameters. I use the shape-rate parameterization of the gamma distribution; all of the gamma priors used here have a mean of 1, but the dispersion of the distribution varies according to different hyperparameter values. This

allows transition rates to deviate in a restricted manner across etyma from the global rates, allowing us to make inferences regarding global dynamics of change on the basis of the global rates. For the flat models, $\tau_{d,r} = \rho_r$.

Alekseyenko et al. (2008) derive a quantity proportional to the likelihood of observations under a multistate SDC model, allowing us to infer posterior distributions for each model's rate parameters. Data were processed using the R packages phytools (version 0.6-99, Revell 2012) and phangorn (version 2.5.5, Schliep et al. 2017). Models were fitted using a maximum clade credibility tree of the Gray et al. (2009) tree sample. Model fitting was carried out using RStan (version 2.26.13, Carpenter et al. 2017). Model convergence was assessed via the potential scale reduction factor (Gelman & Rubin, 1992), with values under 1.1 taken to indicate convergence. All code and data used in this paper can be found at `https://github.com/chundrac/amp-2022-proc`.

Acknowledgements

I gratefully acknowledge the support of the NCCR Evolving Language (Swiss National Science Foundation Agreement Nr. 51NF40_180888) and the DFG Center for Advanced Studies – Words, Bones, Genes, Tools.

References

Alekseyenko, Alexander V., Christopher J. Lee & Marc A. Suchard (2008). Wagner and Dollo: A Stochastic Duet by Composing Two Parsimonious Solos. *Systematic Biology* 57:5, 772–784, URL `https://doi.org/10.1080/10635150802434394`.

Allen, W Sidney (1984). On certain case constraints and their interpretation. *Lingua* 63:1, 1–15.

Arnold, Kate & Klaus Zuberbühler (2012). Call combinations in monkeys: compositional or idiomatic expressions? *Brain and language* 120:3, 303–309.

Bach, Emmon & Robert T. Harms (1972). How do languages get crazy rules? Stockwell, R. & R. Macauley (eds.), *Linguistic Change and Generative Theory*, Bloomington, 1–21.

Bates, Douglas, Martin Mächler, Ben Bolker & Steve Walker (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67:1, 1–48.

Bergsland, Knut & Hans Vogt (1962). On the validity of glottochronology. *Current Anthropology* 3:2, 115–153.

Berkley, Deborah Milam (2000). *Gradient obligatory contour principle effects*. Ph.D. thesis, Northwestern University.

Blasi, Damián E, Søren Wichmann, Harald Hammarström, Peter F Stadler & Morten H Christiansen (2016). Sound–meaning association biases evidenced across thousands of languages. *Proceedings of the National Academy of Sciences* 113:39, 10818–10823.

Blasi, Damián E, Steven Moran, Scott R Moisik, Paul Widmer, Dan Dediu & Balthasar Bickel (2019). Human sound systems are shaped by post-Neolithic changes in bite configuration. *Science* 363(6432):6432.

Blevins, Juliette & Andrew Wedel (2009). Inhibited sound change: An evolutionary approach to lexical competition. *Diachronica* 26:2, 143–183.

Blust, Robert & Stephen Trussel (2013). The Austronesian comparative dictionary: a work in progress. *Oceanic Linguistics* 52:2, 493–523.

Bouchard-Côté, Alexandre, David Hall, Thomas L Griffiths & Dan Klein (2013). Automated reconstruction of ancient languages using probabilistic models of sound change. *Proceedings of the National Academy of Sciences* 110:11, 4224–4229.

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

Breiss, Canaan & Bruce Hayes (2020). Phonological markedness effects in sentence formation. *Language* 96:2, 338–370.

Buckley, Eugene (1997). Tigrinya root consonants and the ocp. *University of Pennsylvania Working Papers in Linguistics* 4:3, 19–51.

Calude, Andreea S & Mark Pagel (2011). How do we use language? shared patterns in the frequency of word use across 17 world languages. *Philosophical Transactions of the Royal Society B: Biological Sciences* 366:1567, 1101–1107.

Carpenter, Bob, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li & Allen Riddell (2017). Stan: A probabilistic programming language. *Journal of statistical software* 76(1):1, 1–32.

Cathcart, Chundra A. (2018). Modeling linguistic evolution: a look under the hood. *Linguistics Vanguard* 1.

Cathcart, Chundra A, Andreas Hölzl, Gerhard Jäger, Paul Widmer & Balthasar Bickel (2020). Numeral classifiers and number marking in indo-iranian: A phylogenetic approach. *Language Dynamics and Change* 1:aop, 1–53.

Chang, William, Chundra Cathcart, David Hall & Andrew Garrett (2015). Ancestry-constrained phylogenetic analysis supports the Indo-European Steppe Hypothesis. *Language* 91:1, 194–244.

Coetzee, Andries W & Joe Pater (2008). Weighted constraints and gradient restrictions on place co-occurrence in muna and arabic. *Natural Language & Linguistic Theory* 26:2, 289–337.

Dellert, Johannes & Armin Buch (2018). A new approach to concept basicness and stability as a window to the robustness of concept list rankings. *Language Dynamics and Change* 8:2, 157–181.

Dryer, Matthew S (2000). Counting genera vs. counting languages. *Linguistic Typology* 4:3, 334–356.

Dunn, Michael, Simon J Greenhill, Stephen C Levinson & Russell D Gray (2011). Evolved structure of language shows lineage-specific trends in word-order universals. *Nature* 473:7345, 79–82.

Everett, Caleb (2018). The global dispreference for posterior voiced obstruents: A quantitative assessment of word-list data. *Language* 94:4, e311–e323.

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

Frisch, Stefan A, Janet B Pierrehumbert & Michael B Broe (2004). Similarity avoidance and the OCP. *Natural Language & Linguistic Theory* 22:1, 179–228.

Gelman, Andrew & Donald B. Rubin (1992). Inference from iterative simulation using multiple sequences. *Statistical Science* 7, 457–511.

Gorman, Kyle (2011). A program for phonotactic theory. *Proceedings of the Annual Meeting of the Chicago Linguistic Society* 47:1, 79–93.

Graff, Peter & T Jaeger (2009). Locality and feature specificity in OCP effects: Evidence from Aymara, Dutch, and Javanese. *Proceedings from the annual meeting of the Chicago linguistic society*, Chicago Linguistic Society, vol. 45, 127–141.

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

Gray, Russell D, Alexei J Drummond & Simon J Greenhill (2009). Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* 323:5913, 479–483.

Greenberg, Joseph H (1950). The patterning of root morphemes in semitic. *Word* 6:2, 162–181.

Greenhill, Simon J, Robert Blust & Russell D Gray (2008). The Austronesian basic vocabulary database: from bioinformatics to lexomics. *Evolutionary Bioinformatics* 4, EBO–S893.

Grotberg, April Lynn (2022). *Quantifying Phonological Feature Co-occurrence*. Ph.D. thesis, Purdue University Graduate School.

Haynie, Hannah J & Claire Bowern (2016). Phylogenetic approach to the evolution of color term systems. *Proceedings of the National Academy of Sciences* 113:48, 13666–13671.

Honkola, T., O. Vesakoski, K. Korhonen, J. Lehtinen, K. Syrjänen & N. Wahlberg (2013). Cultural and climatic changes shape the evolutionary history of the uralic languages. *Journal of Evolutionary Biology* 26:6, p. 12441253.

Hruschka, Daniel J., Simon Branford, Eric D. Smith, Jon F. Wilkins, Andrew Meade, Mark Pagel & Tanmoy Bhattacharya (2013). Detecting regular sound changes in linguistics as events of concerted evolution. *Current Biology* 25, 1–9.

Jaeger, T Florian, Peter Graff, William Croft & Daniel Pontillo (2011). Mixed effect models for genetic and areal dependencies in linguistic typology. *Linguistic Typology* 15, 281–320.

Jäger, Gerhard (2013). Phylogenetic inference from word lists using weighted alignment with empirical determined weights. *Language Dynamics and Change* 3, 245–291.

Jäger, Gerhard (2019). Computational historical linguistics. *Theoretical Linguistics* 45:3-4, 151–182.

Jäger, Gerhard & Johannes Wahle (2021). Phylogenetic typology. *arXiv preprint arXiv:2103.10198* .

Johansson, Niklas Erben, Andrey Anikin, Gerd Carling & Arthur Holmer (2020). The typology of sound symbolism: Defining macro-concepts via their semantic and phonetic features. *Linguistic Typology* 24:2, 253–310.

Kaiping, Gereon A & Marian Klamer (2018). Lexirumah: An online lexical database of the lesser sunda islands. *PloS one* 13:10, p. e0205250.

Kelly, Luke J & Geoff K Nicholls (2017). Lateral transfer in stochastic dollo models. *The Annals of Applied Statistics* 11:2, 1146–1168.

Kümmel, Martin (2007). *Konsonantenwandel*. Dr. Ludwig Reichert Verlag, Wiesbaden.

Lancheros, M, AL Jouen & M Laganaro (2020). Neural dynamics of speech and non-speech motor planning. *Brain and Language* 203, p. 104742.

List, Johann-Mattis (2012). Lexstat: Automatic detection of cognates in multilingual wordlists. *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, Association for Computational Linguistics, 117–125.

List, Johann-Mattis, Robert Forkel, Simon J Greenhill, Christoph Rzymski, Johannes Englisch & Russell D Gray (2022a). Lexibank, a public repository of standardized wordlists with computed phonological and lexical features. *Scientific Data* 9:1, 1–16.

List, Johann Mattis, Annika Tjuka, Christoph Rzymski, Simon Greenhill & Robert Forkel (eds.) (2022b). *CLLD Concepticon 3.0.0*. Max Planck Institute for Evolutionary Anthropology, Leipzig, URL `https://concepticon.clld.org/`.

Macklin-Cordes, Jayden L, Claire Bowern & Erich R Round (2021). Phylogenetic signal in phonotactics. *Diachronica* 38:2, 210–258.

Maddison, Wayne P. (1993). Missing data versus missing characters in phylogenetic analysis. *Systematic Biology* 42:4, 576–581, URL http://www.jstor.org/stable/2992490.

Martin, Andrew (2011). Grammars leak: Modeling how phonotactic generalizations interact within the grammar. *Language* 87:4, 751–770.

Martin, Andrew Thomas (2007). *The evolving lexicon*. Ph.D. thesis, University of California, Los Angeles.

McCarthy, John J (1986). Ocp effects: Gemination and antigemination. *Linguistic inquiry* 17:2, 207–263.

Naranjo, Matías Guzmán & Laura Becker (2022). Statistical bias control in typology. *Linguistic Typology* 26:3, 605–670.

Naroll, Raoul (1961). Two solutions to galton's problem. *Philosophy of Science* 28:1, 15–39.

Needleman, Saul B. & Christian D. Wunsch (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* 48, 443–53.

Neureiter, Nico, Peter Ranacher, Nour Efrat-Kowalsky, Gereon A Kaiping, Robert Weibel, Paul Widmer & Remco R Bouckaert (2022). Detecting contact in language trees: a bayesian phylogenetic model with horizontal transfer. *Humanities and Social Sciences Communications* 9:1, 1–14.

Nevins, Andrew (2012). Haplological dissimilation at distinct stages of exponence. *The morphology and phonology of exponence* 84–116.

Nicholls, Geoff K. & Russell D. Gray (2006). Quantifying uncertainty in a stochastic Dollo model of vocabulary evolution. Forster, Peter & Colin Renfrew (eds.), *Phylogenetic methods and the prehistory of languages*, McDonald Institute for Archaeological Research, Cambridge, 161–71.

Ohala, John J (1993). The phonetics of sound change. Longman, London, p. 237278.

Pagel, Mark (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.

Pozdniakov, Konstantin & Guillaume Segerer (2007). Similar place avoidance: A statistical universal .

Rácz, Péter, Jennifer Hay, Jeremy Needle, Jeanette King & Janet B Pierrehumbert (2016). Gradient māori phonotactics. *Te Reo* 59.

Rama, Taraka (2016). Siamese convolutional networks for cognate identification. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, The COLING 2016 Organizing Committee, Osaka, Japan, 1018–1027, URL https://www.aclweb.org/anthology/C16-1097.

Revell, Liam J. (2012). phytools: An r package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution* 3, 217–223.

Sagart, Laurent, Guillaume Jacques, Yunfan Lai, Robin J Ryder, Valentin Thouzeau, Simon J Greenhill & Johann-Mattis List (2019). Dated language phylogenies shed light on the ancestry of sino-tibetan. *Proceedings of the National Academy of Sciences* 116:21, 10317–10322.

Schel, Anne Marijke, Sandra Tranquilli & Klaus Zuberbühler (2009). The alarm call system of two species of black-and-white colobus monkeys (colobus polykomos and colobus guereza). *Journal of Comparative Psychology* 123:2, p. 136.

Schlenker, Philippe, Emmanuel Chemla, Kate Arnold & Klaus Zuberbühler (2016). Pyow-hack revisited: Two analyses of putty-nosed monkey alarm calls. *Lingua* 171, 1–23.

Schliep, Klaus, Potts, Alastair J., Morrison, David A., Grimm & Guido W. (2017). Intertwining phylogenetic trees and networks. *Methods in Ecology and Evolution* 8:10, 1212–1220, URL http://dx.doi.org/10.1111/2041-210X.12760.

Stanton, Juliet & John F Stanton (2022). In defense of o/e. URL https://ling.auf.net/lingbuzz/006391.

Tarasov, Sergei (2019). Integration of anatomy ontologies and evo-devo using structured markov models suggests a new framework for modeling discrete phenotypic traits. *Systematic biology* 68:5, 698–716.

van de Weijer, Joost (2005). Listeners' sensitivity to consonant variation within words. *Lund University Working Papers in Linguistics* 51, 225–238.

Vehtari, Aki, Andrew Gelman & Jonah Gabry (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and computing* 27:5, 1413–1432.

Wilson, Colin & Marieke Obdeyn (2009). Simplifying subsidiary theory: statistical evidence from arabic, muna, shona, and wargamay. *Ms, Johns Hopkins University* .

Yao, Yuling, Aki Vehtari, Daniel Simpson & Andrew Gelman (2017). Using stacking to average Bayesian predictive distributions. *Bayesian Analysis* .

Zorc, R. David (1969). A study of the Aklanon dialect, volume two: Dictionary (of root words and derivations), Aklanon to English .

Zuraw, Kie (2018). Beyond trochaic shortening: A survey of central pacific languages. *Language* 94:1, e1–e42.

Zuraw, Kie & Yu-An Lu (2009). Diverse repairs for multiple labial consonants. *Natural Language & Linguistic Theory* 27:1, 197–224.