Jayden L. Macklin-Cordes* and Erich R. Round

Challenges of sampling and how phylogenetic comparative methods help

With a case study of the Pama-Nyungan laminal contrast

Received 27 May 2021; accepted 29 December 2021

Abstract: Phylogenetic comparative methods are new in our field and are shrouded, for most linguists, in at least a little mystery. Yet the path that led to their discovery in comparative biology is so similar to the methodological history of balanced sampling, that it is only an accident of history that they were not discovered by a typologist. Here we clarify the essential logic behind phylogenetic comparative methods and their fundamental relatedness to a deep intellectual tradition focussed on sampling. Then we introduce concepts, methods and tools which will enable typologists to use these methods in everyday typological research. The key commonality of phylogenetic comparative methods and balanced sampling is that they attempt to deal with statistical non-independence due to genealogy. Whereas sampling can never achieve independence and requires most comparative data to be discarded, phylogenetic comparative methods achieve independence while retaining and using all data. We discuss the essential notions of phylogenetic signal; uncertainty about trees; typological averages and proportions that are sensitive to genealogy; comparison across language families; and the effects of areality. Extensive supplementary materials illustrate computational tools for practical analysis and we illustrate the methods discussed with a typological case study of the laminal contrast in Pama-Nyungan.

Keywords: Phylogenetic comparative methods; Balanced sampling; Genealogy; Phylogenetic autocorrelation; Phylogenetic signal; Genealogically-sensitive averages; Mass comparison; Areality

***Corresponding author: Jayden L. Macklin-Cordes,** Laboratoire Dynamique Du Langage (UMR 5596), CNRS / Université Lyon 2; Ancient Language Lab, The University of Queensland. E-mail: jayden.macklin-cordes@cnrs.fr

Erich R. Round, Surrey Morphology Group, University of Surrey; Ancient Language Lab, The University of Queensland; Max Planck Institute for the Science of Human History.

1 Introduction

Linguistic typology examines the known diversity of languages with the aim of uncovering insights into the nature of human language itself. The task of cross-linguistic comparison is complicated, however, by the interwoven patterns of historical descent and contact between languages. These patterns of historical relatedness can manifest in shared forms and features in languages today. Consequently, there is widespread recognition that shared histories must be taken into account in typological analysis (see Section 2.2), and there is an abiding concern that the methods used in typology be attuned to the complications of genealogy to the best extent possible.

The non-independence of synchronic observations due to histories of shared descent is a fundamental concept not only in linguistics, but also in other fields where entities share common paths of descent, such as biology and anthropology. Nevertheless, there are a variety of lines of thought and responses that have developed in different fields over the course of a century of scholarship. Consequently, we begin our paper by considering this well-worn discussion within a cross-disciplinary scope. We find that all fields share, in origin, similar lines of development in the elaboration of sampling methodologies for producing phylogenetically independent samples. During this common phase, many independent developments in linguistics and biology have been uncannily parallel. However, biology is now pursuing a different set of solutions to challenges that we have long faced in common. It is instructive, therefore, to understand why a discipline that mirrored linguistic typology for so long has now shifted its approach, and to see how the factors that motivated the change in biology also exist in linguistics.

The paper proceeds as follows: Section 2 reviews literature on *phylogenetic autocorrelation*—the tendency of languages to show similarities due to phylogenetic relatedness—and the methodological responses to it in linguistic typology and cognate fields (comparative biology, in particular). Section 3 then introduces the concept of phylogenetic signal, the degree of phylogenetic autocorrelation that is present in a comparative dataset, and describes statistical tools for quantifying it. Section 4 addresses the topic of uncertainty in linguistic genealogies, and discusses ways in which phylogenetic comparative methods enable a nuanced, explicit examination of how inferences that are drawn from cross-linguistic data are affected by hypotheses about genealogy. In Section 5, because two of the most common types of scientific finding in typology are cross-linguistic averages of typological variables and proportions of languages that have particular properties, we describe phylogenetic methods for the calculation of averages and proportions that take genealogy into account. In Section 6 we present a typological case study of the laminal places of articulation in the Pama-Nyungan languages of Australia. Here we illustrate both the principles and methods introduced earlier, and produce some new insights about this facet of Australian phonological typology that are obtainable only with phylogenetic comparative tools. To discuss and conclude, Section 7 returns to the topics of mass comparison and deep-time language relateness, and language contact and areality, in the light of the foregoing discussions, and in Section 8 we offer a concluding outlook.

2 Phylogenetic autocorrelation: The consequences of relatedness

Phylogenetic autocorrelation is common to many comparative fields of science. It is a potential problem for comparative study, because shared phylogenetic histories limit the independence of observations in a comparative dataset. Observations from more closely related entities will tend to show less variation than more distantly related entities, because they share a longer period of common history and have had less time to diverge since the splitting up of their most recent common ancestor. If this tendency towards similarity due to shared phylogenetic history is not taken into account, it will introduce bias into the dataset and consequently affect statistical analysis. This section discusses phylogenetic autocorrelation and the history of responses to it in different fields. We emphasise some remarkable parallels across disciplines in their independent lines of thinking, especially around the issue of data sampling. However, we also highlight a significant distinction that has emerged since the uptake of quantitative phylogenetic comparative methods in comparative biology. We begin with some cross-disciplinary background (Section 2.1) then focus in particular on linguistics (Section 2.2) and biology (Section 2.3). We unpack the key methodological breakthrough that lies behind phylogenetic comparative methods (Section 2.4) and then discuss its uptake in disciplines beyond biology (Section 2.5).

2.1 Phylogenetic autocorrelation across the sciences

Different fields have their own lines of literature grappling with phylogenetic autocorrelation extending back many decades. In comparative anthropology, this issue was noted as early as 1889 by Sir Francis Galton in the context of cross-cultural datasets, which lack independence due to shared histories of cultural innovation and exchange between societies (Naroll 1961: 15). This phenomenon, known as *Galton's Problem*, is now more precisely understood as a form of statistical *autocorrelation*, i.e., similarity between observations that correlates with their proximity, in this case, their proximity in evolutionary time. The same phenomenon has been recognised in comparative biology too. A seminal study concerning comparative studies of phenotypes, Felsenstein (1985) demonstrates that data from species cannot be assumed to be independently drawn from the same distribution, because species are related to one another via a branching, hierarchical phylogeny, thus, statistical methods that assume independent, identically-distributed observations will inflate the significance of the test (discussed further in Section 2.3 below). Linguists, it was argued, had been somewhat slower than those in other fields to acknowledge exposure to Galton's problem, or phylogenetic autocorrelation (Perkins 1989: 293). However, this is a central concern of Dryer (1989: 259) and has been addressed in a considerable body of linguistic typological literature since then.

Statistical non-independence due to shared history is thus no new revelation, not in comparative anthropology, not in comparative biology, nor in linguistic typology. However, there are many possible approaches to dealing with its challenges and a sizeable body of literature on the topic. As we will see, although precise strategies are varied, a notable commonality to all fields is a history of first attempting to address phylogenetic autocorrelation through the development of sampling methods for the creation of phylogenetically independent—or phylogenetically balanced—samples. The most striking differences between disciplines emerges only later, following the uptake in comparative biology of phylogenetic comparative methods.

2.2 Phylogenetic autocorrelation in linguistics

In linguistic typology, the use of phylogenetically balanced language samples remains the predominant way of accounting for phylogenetic autocorrelation and literature on this topic extends back several decades. Bell (1978: 145–149) argues that common strategies which simply ensure equally-weighted representation of "all major families" or all continents is inadequate due to differing rates of divergence among families. He estimates the number of language groups separated by more than 3,500 years of divergence and uses it as a heuristic for estimating genealogical biases in a selection of proposed language samples. He concludes that European languages tended to be overrepresented and Indo-Pacific languages underrepresented in typological language samples at his time of writing. He attributes this to a corresponding over/under-representation among quality language resources, which is a persistent problem for comparative linguistics.

Perkins (1980, 1988) creates a sample of 50 languages, later adapted by Bybee (1985), which attempts to account for both genealogical and areal biases by selecting no more than one language from each language phylum following Voegelin & Voegelin (1966) and no more than one language from each cultural and geographic area following work in comparative anthropology (Kenny 1975, Murdock 1967). This method attempts to account for non-independence due to areal spread, unlike Bell's heuristic measure which accounts only for genealogical bias, however it does not account for differing ages of divergence and size of language phyla in the way Bell does.

Balanced sampling methods seek to produce linguistic samples that are independent, by selectively excluding the vast majority of attested languages, as necessitated by their extensive, inherent non-independence. As typologists have developed these methods, they have confronted two main complications.

The first complication is that it may be difficult to find criteria for the inclusion/exclusion of languages which truly remove all dependencies, or which are uncontroversial. Dryer (1989: 261) refers to the example of the inclusion of three languages in Perkins' sample (Ingassana, Maasai and Songhai) which potentially are related as part of the Nilo-Saharan family, and thus non-independent, although these relationships are remote and subject to debate. One aspect of this problem is that the maximal extent of presently established language families is partially a product of the extent of adequate documentation and scholarly attention, rather than a reflection of the fullest extent to which the family may be reconstructed (Levinson et al. 2011). Two languages which are presently understood to be unrelated, and therefore statistically independent, may in fact belong to a shared larger grouping, which has not yet been identified due to poor documentation or lack of historical-comparative study. A second aspect is that language families undoubtedly share deep-time relationships that are currently beyond the reach of the comparative method, even if all extant languages were documented and compared completely. Both challenges can lead to languages being deemed as independent when in reality they are not. Dryer (1989: 263) raises a related concern, which is that languages selected on the basis of genealogical independence may nonetheless share characteristics due to non-genealogical processes—language contact and borrowing. This motivates the use of areal criteria in addition to genealogical ones when constructing an independent sample. As Dryer (1989: 284) acknowledges however, linguistic areas may be also subject to the same concerns about undetected historical non-independence and it is possible that the whole world may, in effect, function as a single linguistic area, such that the distribution of certain linguistic features may reflect extremely remote areal or genealogical patterns rather than some true tendency of human language.

The second complication is that once all genealogical and areal criteria are adhered to, the resulting sample may be too small for use in statistical analysis (Cysouw 2005, Jaeger et al. 2011, Piantadosi & Gibson 2014). In response, linguists have proposed various procedures for constructing samples which, if not fully independent, at least have a high degree of independence. Dryer's proposed solution is to build a sample of languages of approximately equal relative independence (at the level of major subfamilies within Indo-European, such as Romance, Germanic, and so on) for each of five large linguistic areas which are assumed to be independent, or at least sufficiently independent for statistical purposes. Any statistical test can then be applied to each of the five areas and only if the same result is replicated in all five areas is it considered statistically significant. If the same result is replicated in four of five areas, this falls short of statistical significance, although Dryer (1989: 272–273) considers such cases to be evidence of a "trend". Nichols (1992: 41) uses Dryer's area-by-area testing method as part of a three-pronged approach. For any given question, Nichols first conducts a chi-square test of the world sample and then re-tests the significance of the finding using either Dryer's method or by running the same test on only the sample of "New World" languages (comprising North, Central and South America). Rijkhoff et al. (1993) and Rijkhoff & Bakker (1998) develop another approach to account for the possibility of non-independence across large linguistic areas and large, as-yet-undetected families. They permit multiple languages within a family to be included but develop a measure, based on the density of nodes in a known language phylogeny, to determine how many languages should be included. In this way, they also aim to account for the fact that some language families will have greater internal diversity than others (see also Bakker 2011, Miestamo, Bakker & Arppe 2016).

Another approach is to include/exclude languages based on their typological profile. Following the logic that historical relatedness and interactions tend to result in elevated similarity, these methods bias their sample in favour of typological diversity, as a proxy for independence. Dryer & Haspelmath (2013) propose setting a minimum threshold of typological distance between languages, calculated from the *World Atlas of Language Structures* (WALS), such that languages must be sufficiently typologically distinct from others in the sample to warrant inclusion. Bickel (2009) develops an alternative algorithm based on Dryer (1989), which allows all uniquely-valued data points within a family to be included in the sample, but then reduces the weighting of data points in the final analysis where a particular value is over-represented within a family. In other words, if all the languages in a particular family share the same value for a typological variable of interest, those observations may be reduced to a single data point.

In these ways, developments in typological methodology have treated historical non-independence between languages as a challenge to be addressed through sampling. Earlier researchers sought to maintain the independence of their sample by maximising the genealogical distance between the languages in their sample, such that no two languages were known to belong to the same family. Later, with subsequent acknowledgement of the possibility of non-independence from very large language families, as well as large-scale areal diffusion and effects from as-yet undetected or unconfirmed historical relations, it became apparent that it may be impossible to create a sample which is simultaneously independent and sufficiently large to generate statistical significance. As discussed above, typologists have primarily responded to this dilemma by developing a variety of robustness checks, even bootstrapping-like processes, whereby languages are sampled at an approximately equal relative level of independence and the sample is then subdivided in some way and a statistical test replicated over each subdivision. More recent years have seen the continued evolution of statistics and robustness checking methods (for an overview, see Roberts 2018), although balanced sampling remains a common element of modern, large-scale comparative linguistic studies (for example, Everett, Blasi & Roberts 2015, Everett 2017, Blasi, Michaelis & Haspelmath 2017).

Before turning to biology, it is worth underscoring how linguistic typology has arrived at its current mode of response to phylogenetic autocorrelation. The starting point is that many conventional statistical methods require observations that are independent, yet languages are non-independent. For four decades, the response has been to change the dataset, by means of balanced sampling, so that it better corresponds to the requirements of the statistics. Doing so requires excluding the vast majority of documented languages from the dataset and hence from the analysis, and even then, the result is still not truly independent. In the next section, we will see that biology initially followed the same path. The key breakthrough, though, was to invert the response to the original problem that phylogenetic autocorrelation posed: to change not the dataset to suit the statistics, but the statistics to suit the dataset. Those changed statistics are phylogenetic comparative methods.

2.3 Phylogenetic autocorrelation in comparative biology

Comparative biology faces the same issue of phylogenetic autocorrelation as comparative linguistics. Many conventional statistical methods assume that observations are independent, which is problematic since observations come from species, which are related to one another through shared evolutionary histories.

Earlier approaches to phylogenetic autocorrelation in biology are in a similar vein to the sampling methods in linguistic typology discussed in the previous section. Harvey & Mace (1982: 346–347) seek to find a taxonomic level to sample from, which strikes the right balance in terms of being sufficiently statistically independent without being so conservative that sample sizes become prohibitively small, an aim similar to Dryer (1989). Their proposed solution is to identify and sample from the lowest taxonomic level which can be "justified on statistical grounds". One method of doing this is suggested by Clutton-Brock & Harvey (1977: 6–8), who conduct a nested analysis of variance and then select the taxonomic level containing the greatest level of variation. Similar to the methods of Dryer & Haspelmath (2013) and Bickel (2009), this approach makes reference to diversity in the traits of the species (cf. diversity in typological traits) to guide the sampling procedure.[1]

As in linguistics, areality is also an issue in biology. Geographical and ecological proximity can lead to similarities in taxa (i.e., species or languages) which is causally separate from the effects of genealogy. Two distinct, causal scenarios can be distinguished. In the first scenario, material is passed directly between taxa, such as lateral transfer of genetic material between species, especially but not exclusively in prokaryotic life forms such as bacteria (Keeling & Palmer 2008), or borrowing between languages. In the second scenario there is no direct transfer of material, rather a shared environment leads to similar developments in taxa, such as parallel dwarfism on islands or, in some cases more contentiously, parallel conditioning of language by its environment (Everett, Blasi & Roberts 2015, Everett 2017, Blasi, Michaelis & Haspelmath 2017, Everett 2021). In both kinds of scenario, there is a causal, areally-correlated contribution to similarity which is separate from the contribution due to shared genealogy. While it is true that modern, genomic studies can circumvent some of the difficulties due to the second scenario in biology, it should be noted that phylogenetic comparative methods in biology predated the emergence of widespread genomic sequencing, and for many species including those attested only as fossils, genetic data is still

[1] Once Clutton-Brock & Harvey (1977) identify their taxonomic level of interest, they average out data for all species within a given genus for which they have data. In other words, the unit of analysis has shifted from individual species to genera, and each data point represents a genus in the form of an averaged representation of all the species within the genus. This genus-level averaging process is in contrast to balanced sampling methods discussed in the previous section, where an unaltered observation from a single exemplar language is taken as representative of its given family, subfamily or other defined grouping, though has affinities with Bickel (2009), which also reduces with-family observations to a smaller number of data points (albeit of a different kind to an average).

not available. Consequently, the problem of convergent evolution due to areality was and still is a genuine, hard problem that comparative biology has faced, and should not be misunderstood as a problem specific to linguistics. In an approach with strong conceptual similarities to the area-by-area robustness checking of Dryer (1989) and Nichols (1992), Baker & Parker (1979: 85–86) discussed how the causal effects of ecological areas might be addressed while constructing a sample which is genealogically balanced. To do so, Baker & Parker (1979) replicate their analysis within individual families as well as within different ecological areas, with the assumption that if the same associations are observed within different areas as across the dataset as a whole, then one can discount the possibility that the full analysis is simply picking up differences between different families or different ecological areas.

In essence, both linguistics and biology face the same phenomenon of phylogenetic autocorrelation including the complication of areality, and for several decades explored strikingly similar methodological responses based on sampling. However, in recent decades the primary methods in linguistic typology and biology have diverged as biology has undergone a fundamental shift. While typologists continue to focus on sampling procedures as the response to phylogenetic autocorrelation, comparative biologists have moved to a more direct, statistical solution. Since the solution addresses phylogenetic autocorrelation, not areality, our focus will narrow now to the genealogical aspects of taxon relatedness. We return to the separate and additional problem of areality in Section 7.2.

2.4 Phylogenetically independent contrasts

Felsenstein (1985) demonstrates that it is possible to account for phylogenetic non-independence in a statistical model without the need to remove data or compromise the unit of analysis (for example, by collapsing or averaging observations within a subgroup). Felsenstein's breakthrough insight is that this can be achieved not by directly comparing non-independent observations but by comparing *phylogenetically independent contrasts* (PICs) between observations. His method has become, by one estimate, the most widespread in comparative biology (Nunn 2011: p. 162). The essential insight is relatively straightforward. Consider the tree in Figure 1. Any traits of A and B will be non-independent observations, since much of their evolutionary history is shared: all of the evolutionary change between points I and H, and between H and G, has contributed equally to both A and B. However, any differences (or in biological parlance, *contrasts*) between A and B have the particular status that they must have arisen after the split at point G. That period of development, after split G until the

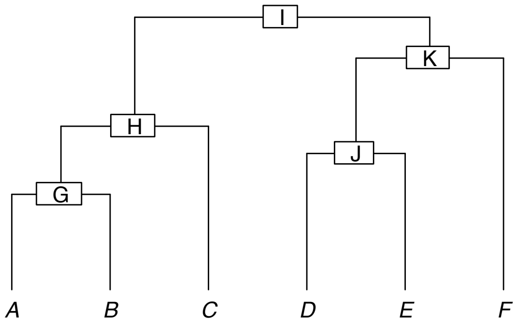

Fig. 1: A phylogeny of six species or languages

modern species (or languages) A and B is not shared with any other part of the tree. It is independent. Felsenstein's insight is that by examining phylogenetic contrasts such as this, one can obtain observations that truly are independent. It is then possible to apply standard statistical tests to the phylogenetically independent contrasts (rather than directly to observed values) without phylogenetic autocorrelation introducing bias into the results.

In the remainder of this subsection we discuss some finer technical points of Felsenstein's notion for readers who are interested. Others may wish to skip ahead directly to the next subsection.

In order to calculate PICs not only between sister tips of a tree such as A and B, but also between sister interior nodes such as H and K, or node-tip sisters such as G and C, one requires in addition to a phylogeny, a model according to which the variable evolves. As a starting point, Felsenstein assumes a *Brownian motion* model of evolution, since Brownian motion is one of the simplest and most fundamental of all stochastic processes. In a Brownian motion model, an evolving quantitative trait can wander positively or negatively with equal probability, and each new time step is independent from the last, with the resulting effect that displacement of the variable over time will be drawn from a normal distribution with a mean of zero and variance proportional to the amount of elapsed time (Felsenstein 1985: p. 8). An observed contrast can be scaled by dividing it by the standard deviation of its expected variance. This gives a statistically independent contrast of expectation zero and unit variance (i.e. variance equal to 1). This process can be repeated for all adjacent tips in the tree. Contrasts can then be extracted from adjacent nodes in the tree, where the value of the node is an average of the observed values of the tips below it. In the end, there will be a collection of phylogenetically independent contrasts, all of expectation zero and unit variance, to which statistical analysis can be applied.

One drawback of Felsenstein's initial method is the reliance on the assumption of Brownian motion as a model of variable evolution. Grafen (1989) subsequently devises a similar method, *the phylogenetic regression*, which has the flexibility to incorporate models of evolution other than Brownian motion. Further, Grafen's method is able to be applied in situations where phylogenetic information is incomplete (for example, where the phylogeny is an incomplete work-in-progress rather than an accepted gold-standard). This method is a phylogenetic adaptation of *generalised least squares* (GLS). In this model, the value of a dependent variable, y_i, is predicted by the equation $y_i = \alpha + \beta x_i + \epsilon$, where α is the intercept, β is the regression slope, x is the independent variable and ϵ is an error term (Nunn 2011: p. 164). Phylogenetic information can be incorporated into the error term, in the form of a variance-covariance matrix of phylogenetic distances between tips in a tree. PICs and GLS are mathematically equivalent when a Brownian motion evolutionary model is assumed and the reference tree is fully bifurcated, so PICs are essentially a special case of GLS where these assumptions are met (Nunn 2011).

2.5 Phylogenetic comparative methods beyond biology

Linguistic typology and comparative anthropology have long faced the same essential problem of phylogenetic autocorrelation that comparative biology contends with. Initially, all three disciplines followed similar trajectories, responding to phylogenetic autocorrelation through the development of increasingly elaborate methods of balanced sampling. By historical accident it was in biology that the breakthrough of examining PICs occurred, but the breakthrough is a solution to an inherent problem that transcends disciplinary boundaries. Anthropologists, recognising the same problem in kind, followed this breakthrough in biology with their own uptake of phylogenetic comparative methods around 10–20 years later (e.g Mace et al. 1994, Holden & Mace 2003, 2009, Jordan et al. 2009, Nunn 2011), and recently there has been growing interest in the application of phylogenetic comparative methods in linguistics (e.g Maslova 2000a,b, Dunn et al. 2011, Maurits & Griffiths 2014, Verkerk 2014, Birchall 2015, Zhou & Bowern 2015, Calude & Verkerk 2016, Dunn et al. 2017, Verkerk 2017, Bentz et al. 2018, Cathcart et al. 2020, Macklin-Cordes, Bowern & Round 2021, Jäger & Wahle 2021).

One of our motivations for this paper, however, is that despite the increasing uptake of phylogenetic comparative methods in linguistics, there has been little attempt until now to explain why phylogenetic comparative methods can best be understood as a continuation of a tradition of inquiry that typology is greatly invested in. Previously, that tradition of inquiry, whether in comparative biology, comparative anthropology or linguistics, had led to methods of balanced sampling. Like balanced sampling methods, phylogenetic comparative methods are a response to phylogenetic autocorrelation, one of the central and most persistent problems of linguistic typology. Methodologists working on balanced sampling have striven to generate samples that come as close as possible to phylogenetic independence, but the goal cannot be fully attained even with the most elaborate sampling procedures, and in the meantime procedures of balanced sampling require the exclusion of the vast majority of documented languages from the dataset and hence from the analysis. As it turns out, the solution is to be found not in phylogenetically independent samples, but in phylogenetically independent contrasts (PICs). By focussing on PICs, Felsenstein unlocked a method for obtaining truly independent observations, without excluding data. This is why typologists have every reason to be keenly interested in phylogenetic comparative methods: they solve a problem which has stood at the centre of our discipline for decades.

In the sections that remain, we shift our focus away from theory and onto practicality: how can typologists begin making use of phylogenetic comparative methods? In Sections 3–5 we introduce key phylogenetic concepts and techniques that typologists can employ, followed by a phylogenetic typological case study in Section 6. In Section S1 of the Supplementary Materials, we provide an extended practical introduction to a suite of computational tools that have been designed with the typologist in mind (Round 2021a,b), enabling phylogenetic comparative methods to be used in everyday typological research. In Section 7 we return to the topic of areality.

3 Phylogenetic signal: The extent to which synchronic distributions mirror genealogy

As discussed in Section 2, phylogenetic comparative methods are applicable in linguistic typology when phylogeny is a causal factor that has shaped the distribution of a linguistic variable. The previous section described the means by which phylogenetic comparative methods are able to take such a phylogeny into account in statistical analysis. However, some variables may not evolve through descent with modification and consequently may not pattern phylogenetically. Others may be subject not only to descent with modification, but to other causal factors in addition such as areality, and thus may pattern phylogenetically only weakly. How, then, does one determine for a variable of interest whether a phylogeny may have contributed to the cross-linguistic distribution of diversity? In the last twenty years, an advance in this area has been the advent of methods for explicitly quantifying the degree of *phylogenetic signal* in comparative data (Freckleton, Harvey & Pagel 2002, Blomberg, Garland & Ives 2003). Phylogenetic signal refers to the tendency of phylogenetically-related entities to resemble one another (Blomberg & Garland 2002, Blomberg, Garland & Ives 2003: p. 717). This resemblance is more technically defined as statistical non-independence among observation values due to phylogenetic relatedness between taxa (Revell et al. 2008: p. 591). This concept of phylogenetic signal has important applications in comparative linguistics. Here we argue that for many purposes, measuring phylogenetic signal should be considered as a first step in a phylogenetically aware comparative methodology, since it can determine empirically whether phylogenetic comparative methods are required or whether regular statistical methods may suffice (as in Irschick et al. 1997).[2] Further, the result of a phylogenetic signal test can contribute to evolutionary hypotheses in its own right, as we will see in the case study in Section 6.

This section describes fundamental methods for measuring phylogenetic signal in variables with continuous values (Section 3.1) and with discrete binary values (Section 3.2). The discussion below will get technical, but we have included it because we expect that some readers will be interested in the details and the underlying logic. For others, who may prefer to skim over the denser technical passages here or skip directly to Section 4, it will suffice to make note of the core message, that testing for phylogenetic signal provides insight into how strongly genealogy may be shaping the data. This is useful knowledge in itself and it enables a more nuanced, judicious use of other phylogenetic comparative methods. For these reasons, testing for phylogenetic signal as part of a research workflow is good practice and is widely employed in phylogenetic studies.

3.1 Phylogenetic signal in continuous variables

Blomberg, Garland & Ives (2003) provide a suite of tools for quantifying phylogenetic signal, which have become somewhat of a standard in the field (cited 3780 times as of September 2021, according to Google Scholar).[3] Recent comparative studies using these tools include Balisi, Casey & Valkenburgh (2018), Hutchinson, Gaiarsa & Stouffer (2018) and Macklin-Cordes, Bowern & Round

[2] Note, however, that the absence of phylogenetic signal does not necessarily indicate that non-phylogenetic statistical methods are appropriate in all cases, in particular for phylogenetic generalised least squares (PGLS) (Revell 2010, Symonds & Blomberg 2014).

[3] In the R statistical programming language (R Core Team 2021) the tests described here are implemented in the `phylosig` function of the *phytools* package (Revell 2012).

(2021). Blomberg, Garland & Ives (2003) present a descriptive statistic, K, which is generalizable across phylogenies of different sizes and shapes. In addition, they provide a randomisation test for checking whether the degree of phylogenetic signal for a given dataset is statistically significant. K can be calculated using either phylogenetically independent contrasts (PICs) (Felsenstein 1985) or generalised least squares (GLS) (Grafen 1989) (see Section 2.4). In a Brownian motion model, where variable values can wander up and down with equal probability through time, PIC variances are expected to be proportional to elapsed time. Among more closely related languages, where there has been less divergence time for variable values to wander, the variance of PICs is expected to be low. The randomisation test works by comparing whether observed PICs are lower than the PIC values obtained by randomly permuting the data across the tips of the tree. The process of permuting data across tree tips at random is repeated many times over. If the real variances, with data in their correct positions on the tree, are lower than 95% of the randomly permuted datasets, then the null hypothesis of no phylogenetic signal can be rejected at the conventional 95% confidence level. In other words, closely related languages resemble one another to a statistically significantly greater degree than would be expected by chance.

The descriptive statistic, K, quantifies the strength of phylogenetic signal. As with the randomisation procedure above, the input is a set of observed values, where each observation is associated with a tip of the reference tree. Blomberg, Garland & Ives (2003: 722) give an explanation of the calculation of the K statistic. To summarise briefly, K is calculated by, firstly, taking the mean squared error (MSE_0), as measured from a phylogenetic mean,[4] and dividing it by the mean squared error (MSE) calculated using a variance-covariance matrix of phylogenetic distances between tips in the reference tree (the same variance-covariance matrix of phylogenetic distances incorporated into the error term in GLS-based phylogenetic regression, as discussed in the previous section). This latter value, MSE, will be small when the pattern of covariance in the data matches what would be expected given the phylogenetic distances in the reference tree, leading to a high MSE_0/MSE ratio and vice versa. Thus, a high MSE_0/MSE ratio indicates higher phylogenetic signal. Finally, the observed ratio can be scaled according to its expectation under the assumption of Brownian motion evolution along the tree. This gives a K score which can be compared directly between analyses using different tree sizes and shapes. Where $K = 1$,

[4] We discuss the phylogenetic mean further in Section 5 below. Simply taking a non-phylogenetic mean of a variable would be misleading in cases where members of a particularly large clade happen to share similar values at an extreme end of the range.

this suggests a perfect match between the covariance observed in the data and what would be expected given the reference tree and the assumption of Brownian motion evolution. Where $K < 1$, close relatives in the tree bear less resemblance in the data than would be expected under the Brownian motion assumption. $K > 1$ is also possible—this occurs where there is less variance in the data than expected, given the Brownian motion assumption and divergence times suggested by the reference tree. In other words, close relatives bear greater resemblance than would be expected, given the overall phylogenetic diversity.

As discussed, the assumption of a Brownian motion model of evolution, where a variable is free to wander up or down, with equal probability, as time passes, is central to quantification of phylogenetic signal with the K statistic. Blomberg, Garland & Ives (2003: 726–727) extend their approach to cover two different modes of evolution as well. This is achieved by incorporating extra parameters into the variance-covariance matrix to reflect different evolutionary processes. The first evolutionary model alternative is the Ornstein-Uhlenbeck (OU) model (Felsenstein 1988, Garland et al. 1993, Hansen & Martins 1996, Lavin et al. 2008) whereby variables are still free to wander up or down at random, but there is a central pulling force towards some optimum value. The second alternative is an acceleration-deceleration (ACDC) model, developed by Blomberg, Garland & Ives (2003) where a variable value moves up or down with equal probability (like Brownian motion) but the rate of evolution will either accelerate or decelerate over time.

Other statistics for quantifying phylogenetic signal have been proposed and warrant mention. Freckleton, Harvey & Pagel (2002) propose using the λ (lambda) statistic, based on earlier work by Pagel (1999). As for Blomberg, Garland & Ives (2003), this approach works with a variance-covariance matrix showing the amount of shared evolutionary history between any two tips in the tree (the diagonal of the matrix, the variances, will indicate the total height of the tree; the off-diagonals, the covariances, will indicate the amount of shared evolutionary history between two given entities, before they diverge in the tree). The statistic, λ is a scaling parameter which can be applied to this variance-covariance matrix. Scaling the values in the matrix by λ transforms the branch lengths of the tree, from $\lambda = 1$, where branch lengths are left unscaled, to $\lambda = 0$, where all covariances in the matrix will be zero, in other words, no covariance through shared evolutionary history is indicated between any tips, thus all tips will be joined at the root by branches of equal length (a star phylogeny). Freckleton, Harvey & Pagel (2002) present a method for finding the λ parameter that maximises the likelihood of a set of observations arising, given a Brownian motion model of evolution. If λ is close to 1, this indicates high phylogenetic signal, where the data closely fit expectation given the shared evolutionary histories in the tree and a Brownian motion model of evolution. Further measures which have been proposed are I (Moran 1950), a spatial autocorrelation measure which was adapted for phylogenetic analyses by Gittleman & Kot (1990), and C_{mean} (Abouheif 1999), which is a test for serial independence (for an overview, see Münkemüller et al. 2012). In an evaluation of different methods Münkemüller et al. (2012) find that, assuming a Brownian motion model of evolution, C_{mean} and λ generally outperform K and I. However, C_{mean} considers only the topology of the reference tree (i.e., the order of the branches from top to bottom), but not branch length information, and the value of the C_{mean} statistic is partially dependent on tree size and shape, so it lacks comparability between different studies. In addition, λ shows some unreliability with small sample sizes (trees with <20 tips).

3.2 Phylogenetic signal in binary variables

The methods so far described concern continuously-valued data. Other methods have been proposed for quantifying phylogenetic signal in binary and categorical variables too. Abouheif (1999) presents a simulation-based approach for testing whether discrete values along the tips of a phylogeny are distributed in a phylogenetically non-random way. Although this method is useful for testing whether the phylogenetic signal in a set of discretely-valued data is statistically significant, it does not provide a quantification of the level of phylogenetic signal which is comparable between different datasets. Although specific to binary data only, Fritz & Purvis (2010) present a statistic, D, which quantifies the strength of phylogenetic signal for binary variables.

The D statistic is based on the sum of differences between sister tips and sister clades, Σd. To summarise, following Fritz & Purvis (2010), differences between values at the tips of the tree are summed first (all tips will either share the same value, 0 or 1, with 0 difference; or one will be 0 and the other will be 1, for a difference of 0.5). Nodes immediately above the tips are valued as an average of the two tips below (either 0, 0.5 or 1) and the differences between sister nodes is summed. This process is repeated for all nodes in the tree, until a total sum of differences, Σd, is reached. At two extremes, data may be maximally clumped, such that all 1s are grouped together in the same clade in the tree and likewise for all 0s, or data may be maximally dispersed, such that no two sister tips share the same value (every pair of sisters contains a 1 and a 0, leading to a maximal sum of differences). Lying somewhere in between will be both a phylogenetically random distribution and a distribution that is clumped to a degree expected under a Brownian motion model of evolution. A distribution of sums of differences following a phylogenetically random pattern, Σd_r, is obtained by shuffling variable values among tree tips many times over. A distribution of sums of differences following a Brownian motion pattern, Σd_b is obtained by simulating the evolution of a continuous trait along the tree, following a Brownian motion process, many times over. Resulting values at the tips above a threshold are converted to 1, values below the threshold are converted to 0. The threshold is set to whatever level is required to obtain the same proportion of 1s and 0s as observed in the real data. Finally, D is determined by scaling the observed sum of differences to the means of the two reference distributions (the expected sums of differences under a phylogenetically random pattern and under a Brownian motion pattern).

$$D = \frac{\Sigma d_{obs} - mean\left(\Sigma d_b\right)}{mean\left(\Sigma d_r\right) - mean\left(\Sigma d_b\right)} \tag{1}$$

Scaling D in this way provides a standardised statistic which can be compared between different sets of data, with trees of different sizes and shapes, as with K for continuous variables. One disadvantage of D, however, is that it requires quite large sample sizes (>50), below which it loses statistical power, increasing the chance of a false positive result (type I error).

Although we have restricted our focus to continuous and binary data here, some recent developments in testing for phylogenetic signal in other kinds of data warrant brief mention also. For example, Borges et al. (2019) have developed a statistic, δ, for quantifying phylogenetic signal in multivalued categorical variables. Other developments concern multivariate and multidimensional data. Zheng et al. (2009) present a multivariate version of the K statistic discussed in Section 3.1, for measuring phylogenetic signal in groups of related variables. Their statistic also incorporates measurement error. Finally, Adams (2014) presents K_{mult}, a statistic for detecting phylogenetic signal in multivariate traits, i.e. conceptually unitary evolutionary traits that are defined by multiple values (e.g. in biology, a set of measurements that together define skull shape).

In this section we have introduced the fundamental notion of phylogenetic signal—the degree to which the distribution of synchronic diversity reflects the shape of a phylogeny—and some key methods for estimating it. Of course, doing this requires a phylogeny to begin with, and typologists may have questions about the suitability of current linguistic trees for such purposes. It is to this important topic that we turn next.

4 Approaches to uncertainty in linguistic trees

A reasonable concern that typologists may have is whether currently available language trees are of sufficient quality to support the use of quantitative phylogenetic methods. Fortunately, there is a clear, technically sound response to this concern. However, the response is not necessarily intuitive, so here we examine it through both logical argumentation and an example.

Not by accident, a parallel concern about the quality of available phylogenies was raised directly by Felsenstein (1985: 14) in his seminal work on phylogenetic comparative biology.[5] In response to this concern, Felsenstein stresses that logically, because genealogies are fundamental to comparative biology (as they are to comparative linguistics), they are also inescapable: "there is no doing [comparison] without taking them into account". No matter what methods we choose to use, if we make comparisons in biology or linguistics, we will inevitably implicate some genealogy, because genealogies are an inherent component of the real-world causal structure that underlies the data. The question, then, will always be not whether to use trees, but which trees to use. Methods of comparison which purport to operate independently of genealogies actually will implicate a phylogeny covertly.

To take a concrete example, consider a situation where the true phylogenetic history of six languages is as shown in Figure 2a, but that currently, this true history is only partially understood. Such is the case for almost any language family. Linguists may possess only a preliminary hypothesis of subgrouping, as in Figure 2b, with little certainty about how deep in time the major splits are. Phylogeny 2b is therefore a sub-optimal representation of 2a and understandably, concern may arise over using it. However, using the tree in Figure 2b would still be preferable to using no tree at all. Technically speaking, it is not possible to use 'no tree'. When phylogeny is ignored entirely, then all languages are set on equal footing, which is equivalent to hypothesising a star tree, also called a rake tree, as in Figure 2c (Purvis & Garland 1993). Consequently, the choice between using the tree in Figure 2b and 'no tree' is in fact a choice between two trees: Figure 2b or 2c, and the former is almost certainly the better approximation of the true phylogeny, Figure 2a. Evaluative studies have shown that even when phylogenies are incomplete, lacking branch length information, or subject to a degree of error, phylogenetic comparative methods still typically out-perform equivalent non-phylogenetic comparative methods, which effectively assume a

[5] It should be remembered that phylogenetic comparative methods arose in biology *before* the widespread availability of high-quality phylogenies based on genome sequencing.

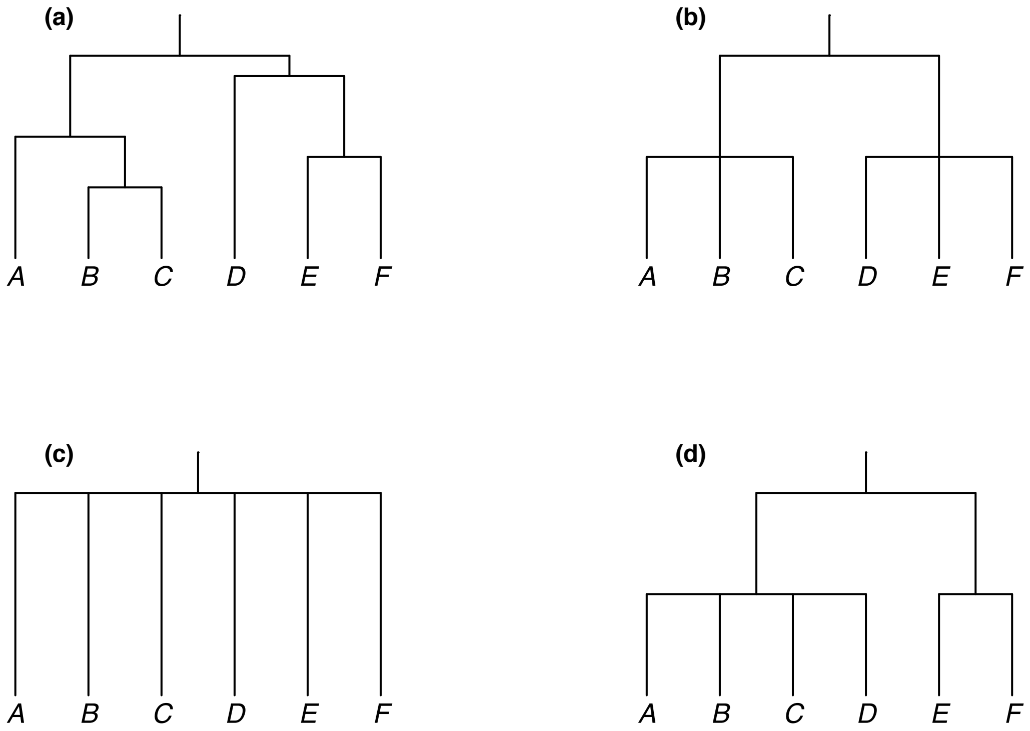

Fig. 2: Four phylogenies of six languages: (a) with detailed branch lengths and topology (nesting structure), (b) with less detail, (c) a star phylogeny (rake phylogeny), (d) an alternate phylogeny with little detail

star phylogeny in this way (Grafen 1989, Purvis, Gittleman & Luh 1994, Symonds & Page 2002, Rohlf 2006). By using Figure 2b with phylogenetic methods, it is possible to derive results that are 'state-of-the-art' in the sense that they reflect the best of current knowledge; this is not true when using a star phylogeny.

Once it is recognised that using 'no tree' is technically not possible, the question still remains of which tree to use. Linguistic trees are often subject to ongoing debate. For instance, different expert analyses may group six languages not only as Figure 2b, but also as Figure 2d. Expert debates such as this are reflective of the *phylogenetic uncertainty* that currently exists about the details of the tree. In these cases, phylogenetic methods can be applied to multiple, alternative trees and the result interpreted critically. Applying phylogenetic methods to multiple trees enables us to move beyond merely disagreeing over phylogenetic hypotheses, towards clarifying what the implications are of adopting different genealogical hypotheses: some results may pivot crucially upon which phylogeny is assumed, while others are largely independent of the choice. Because modern phylogenetic methods are principally computational, there is little practical impediment to examining multiple, alternative tree hypotheses whenever the methods are used. Modern methods of tree inference (e.g. Bouckaert et al. 2012, Chang et al. 2015, Kolipakam et al. 2018, Bouckaert, Bowern & Atkinson 2018) produce large sets termed *tree samples*, of alternative, highly-likely trees, all of which can be used.[6] In our case study in Section 6 below, we demonstrate this approach by using a tree sample of 100 highly-likely phylogenies to investigate the typology of laminal place of articulation contrasts in Pama-Nyungan languages.

In this section on phylogenetic uncertainty, we have framed our discussion primarily in terms of the kind of uncertainty that can surround the tree of a single language family. However, in linguistics we currently possess many separate trees, for many separate language families. The question arises, how can phylogenetic comparative methods be applied across multiple, distinct language families when there is no known, deep-time tree that links them together? We return to this issue in Section 7.1, however the reader may already discern what the response will be, considering that our lack of a global linguistic tree is itself a matter of uncertainty: very likely, many if not all known language families in reality are genealogically linked. If this is true, then even though we are highly uncertain about what their deep-time genealogical links are, it will technically not be possible to use 'no tree' when comparing across them, since in reality their genealogical relationships are an inherent component of the real-world causal structure behind the global typological diversity that we wish to analyse. We return to this matter in Section 7.1.

5 Genealogically-sensitive averages and proportions

A perennial task in typology is the characterisation of frequencies of traits of interest among the world's languages. The scientific interest of such questions typically lies not merely in the contingent facts of today's particular languages and language families, rather the goal is to characterise the nature of human language in general, using today's contingent empirical data as evidence. Because of this, we are striving ideally for an answer that takes into account the unequal representation of different families and subgroups. Phylogenetic comparative methods can assist in achieving this recurrent and indispensable objective of typological research. In this section we describe methods for deriving genealogically-sensitive averages and proportions.

[6] Even if only one phylogeny appears in a published diagram, studies of this kind will almost certainly have produced a full tree sample.

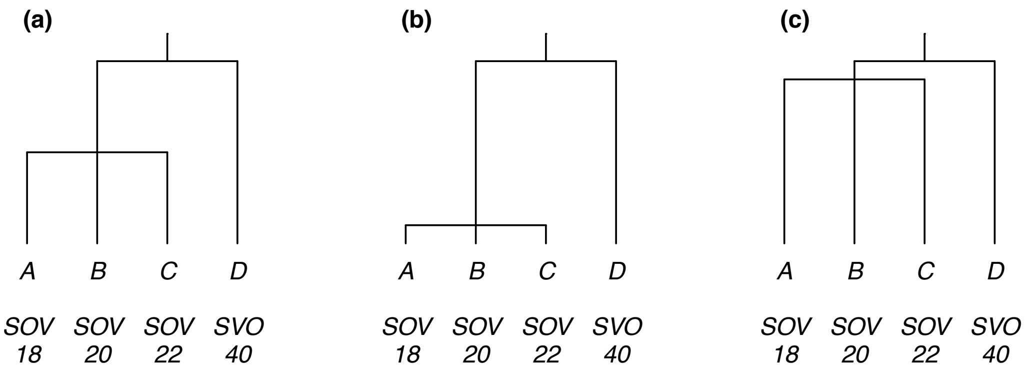

Fig. 3: Three minimally different phylogenies of the same four languages, indicating their dominant word order and number of consonant phonemes

The essential challenge of formulating meaningful averages and proportions when languages are related will be well familiar to typologists. Figure 3 shows three, minimally different phylogenies for a set of four languages, together with the languages' dominant word order pattern and their number of consonant phonemes. If asked what proportion of these languages are SOV, a literal reply would be 75%. However, that answer will strike us as less than satisfactory because languages A–C are more closely related to one another than to D. Merely tallying up the languages allows one of the two major branches in the tree to count three times more than the other. Moreover, the degree to which this answer seems unsatisfactory can vary between phylogenies 3a,b,c. For instance, the answer '75%', which is unsatisfactory for Figure 3a, is arguably worse for Figure 3b, since now A–C are very closely related indeed. Conversely, a reply of 75% for Figure 3c is still imperfect but arguably less unsatisfactory, since although A–C are more closely related to one another than to D, the difference is only slight. This example illustrates the fact that when quantifying the proportion of languages that have some property, any satisfactory method will need to take into account at least two facts about the phylogeny: its topology (i.e., the hierarchical embedding of subgroups) and its branch lengths (note that differing branch lengths are all that distinguish Figures 3a,b,c). The same issues arise if we are seeking not a proportion but an average, such as the 'average' size of the consonant inventories in these languages. The literal mean, $(18+20+22+40)/4 = 25$, is unsatisfactory for the same reason, that it accords much more weight to one major branch than the other. And similarly, it is even more unsatisfactory for Figure 3b than for Figure 3a, though less so for Figure 3c.

There already exists a substantial literature on how to obtain principled values for proportions and averages that are sensitive to genealogy. Here we present two of the methods that have been developed. Before we do, it is useful to recall that even within non-phylogenetic statistics, there are multiple ways of formulating and defining an average, including means, medians, modes, harmonic means, geometric means, and so forth. Each of these operationalises a slightly different concept of the 'representative middle value', or *central tendency*, of some set of observations. Different averages have different properties which may prove advantageous or not, depending on the objectives and datasets at hand. For instance, means can be sensitive to outliers while medians are less so. It should be no surprise, then, that comparable issues arise in the formulation of phylogenetic averages, and the technical literature has discussed them at length (Altschul, Carroll & Lipman 1989, Vingron & Sibbald 1993, Stone & Sidow 2007, De Maio et al. 2020). Here we will emphasise important properties of phylogenetic averages, in relation to the tasks that typologists face.

One way of construing different kinds of averages is in terms of the relative weight they accord to each observation. For instance, a simple mean accords every observation the same weight. Other kinds of averages can be expressed in terms of the slightly different weights they accord to each data point. This approach, of describing averages in terms of a list of weights for each observation, has also been used in the literature on phylogenetic averages, and we will adopt it here. We can also note that a proportion can be re-expressed as an average. Asking for the proportion of languages that are SOV is equivalent to asking for the mean of x, where $x = 1$ if a language is SOV and $x = 0$ if it is not. Correspondingly, a method for constructing weighted averages will extend directly to the construction of weighted proportions. To take an example, suppose we assigned the four languages in Figure 3a the weights $\{0.2, 0.2, 0.2, 0.4\}$, which sum to 1. The weighted average of the consonant inventory sizes would then be $(0.2 \times 18 + 0.2 \times 20 + 0.2 \times 22 + 0.4 \times 40)/(0.2 + 0.2 + 0.2 + 0.4) = 28$. The correspondingly weighted proportion of SOV languages would be $(0.2 \times 1 + 0.2 \times 1 + 0.2 \times 1 + 0.4 \times 0)/(0.2 + 0.2 + 0.2 + 0.4) = 0.6$ or 60%. Any method which can assign weights to a set of languages in a phylogenetically judicious manner will therefore enable us to calculate genealogically-sensitive averages and proportions.

The nearest phylogenetic equivalent to a simple mean is obtained by what is known as the 'ACL' method presented by Altschul, Carroll & Lipman (1989). This kind of genealogically-sensitive average is often referred to as the *phylogenetic mean*. It provides an unbiased estimate of the central tendency of a set of observations, taking into account tree topology and branch lengths. Nevertheless, the ACL method, like non-phylogenetic means, is known to be sensitive to outliers (Stone & Sidow 2007). In a phylogeny, an outlier is a language (or subgroup) located on an early branch, only distantly related to the rest of the tree, such as language E in Figure 4. Because the ACL method accords a high weight to outliers, its results can be particularly sensitive to the highest-level structure in a phylogeny. This can be of concern when confidence in the highest-order

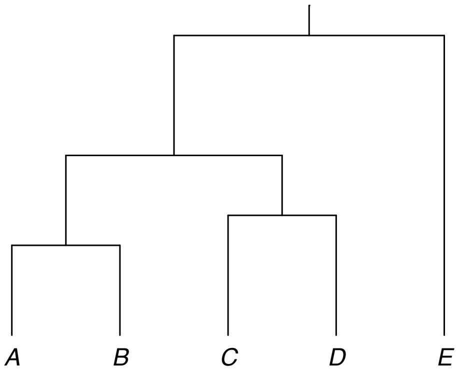

Fig. 4: A phylogeny in which E is an outlier

branching of the tree is low, as is often the case in linguistics, where the deepest splits in a family's history are also the murkiest or most contested by scholars. For that reason, it is prudent to consider another phylogenetic average, which was designed with this problem in mind.

The BranchManager (BM) method of Stone & Sidow (2007) is also an unbiased estimate of the central tendency of a set of observations, taking into account tree topology and branch lengths. However, it is mathematically formulated to accord less extreme weight to high-order branching, in comparison to the ACL method. Arguably, this makes it a more conservative choice in cases where a phylogeny is especially uncertain at its greatest time depths. Moreover, it is possible to use both the ACL method and the BM method to estimate phylogenetically-sensitive proportions and averages, and then to compare them. The comparison will offer an indication of how the implied central tendency of the dataset changes, as we invest a greater or lesser degree of confidence in the correctness of the deepest levels of the tree structure. We make use of this approach in our case study, to which we now turn.

6 A phylogenetic comparative case study: Laminal contrasts in Pama-Nyungan

Phonemic systems are inherited with modification from ancestral languages into their descendants. Consequently, they are expected to contain considerable phylogenetic signal. In Australia, however, for one aspect of phonemic systems it has long been supposed that this is not the case. Australian languages contrast between four and six superlaryngeal places of articulation (Evans 1995, Round 2022): bilabial, dorsal-velar and either one or two apical places (articulated with

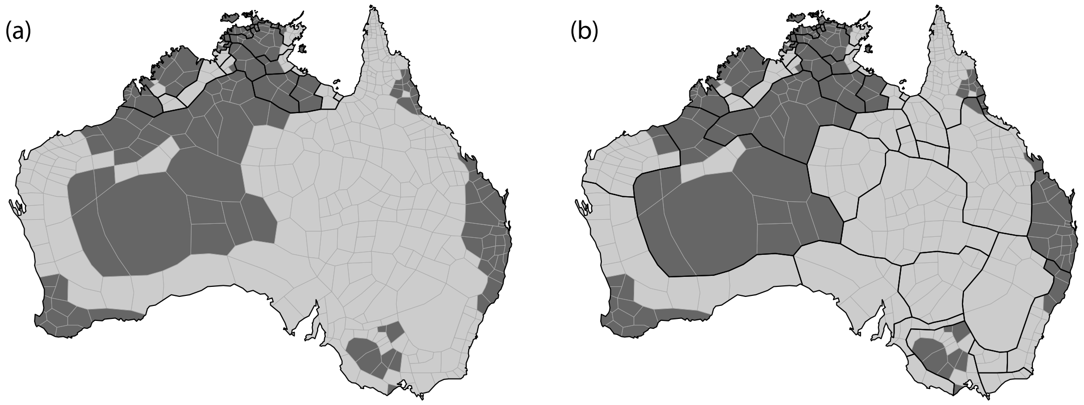

Fig. 5: The distribution of the presence (light) and absence (dark) of a laminal place of articulation contrast in Australian languages. Dark lines indicate (a) language family boundaries, (b) major subgroups of Pama-Nyungan also.

the tongue tip) and either one or two laminal places (articulated with the tongue blade). In this case study we focus on the laminals, and whether languages possess a contrast between two laminal places – laminal dentals and laminal pre-palatals – or just one. We introduce some long-standing claims about the distribution of this contrast across the continent, and then apply the kinds of analyses introduced in Sections 3–5 above.

If we express the figure as a simple proportion, then around 62% of Australian languages have a laminal contrast, according to data in Round (2019). The geographic distribution of the contrast is shown in Figure 5a, along with the boundaries of Australia's 25 mainland language families. The geographic distribution covers large contiguous swathes of the continent and can appear to exhibit little regard for the boundaries of language families. Understandably, this striking aspect of the distribution has been emphasised repeatedly in the literature on Australian phonological typology (Dixon 1970, 1980, Evans 1995). However, here we ask, does this distribution also contain phylogenetic signal?

We begin by adding some additional information to our map. Figure 5b shows the same information as Figure 5a, but adds the boundaries of major subgroups of the Pama-Nyungan language family which dominates the continent. The reader may find that the effect of the map has changed: the distribution of the laminal contrast is largely organised neatly within the major phylogenetic units across the continent. Inspecting maps in this fashion can suggest potential conclusions about phylogenetic signal, but a more secure line of analysis is to use quantitative methods. Here we will focus on Pama-Nyungan. Within Pama-Nyungan, 73% of languages have a laminal contrast, expressed as a simple proportion. In the remainder of the section, we first estimate the degree of phylogenetic signal in the

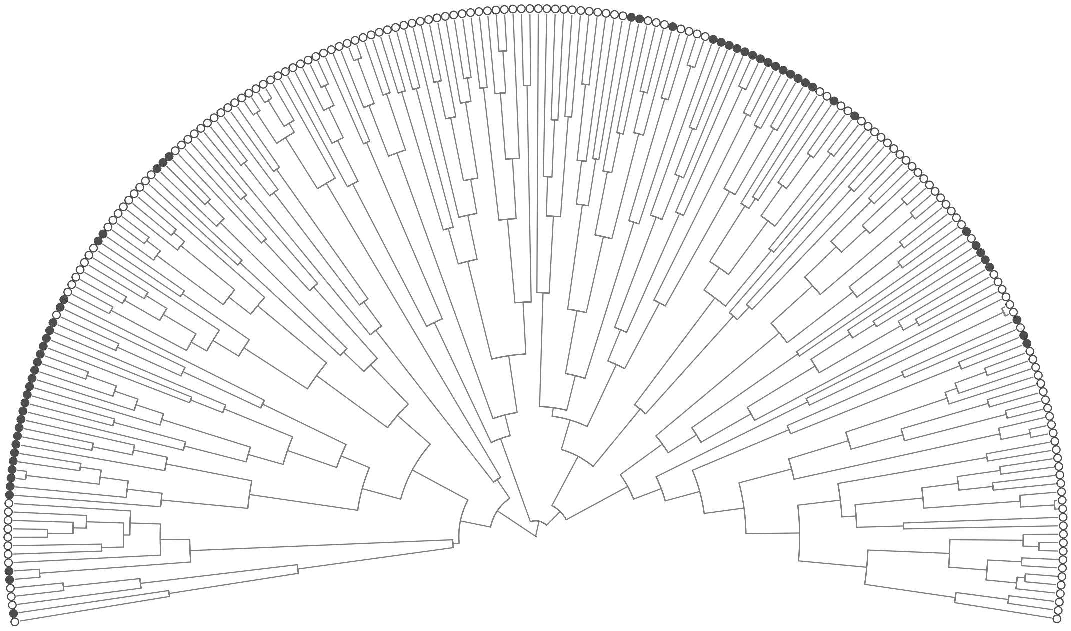

Fig. 6: The distribution of the presence (light) and absence (dark) of a laminal place of articulation contrast across Pama-Nyungan, displayed on a maximum clade credibility (MCC) tree. An MCC tree is a single tree within a tree sample which most adequately represents the highest-probability subgroups in the trees of the sample. This MCC tree is taken from the sample of 100 highly-likely Paman-Nyungan phylogenies used in the current study.

distribution of the laminal contrast using the D statistic we introduced in Section 3.2, which measures phylogenetic signal in binary variables. We then turn to some more fine grained phonotactic data, to which we apply the K statistic introduced in Section 3.1, which measures phylogenetic signal in continuous variables. Having ascertained the level of phylogenetic signal in the Pama-Nyungan laminals, we then estimate the phylogenetically-weighted proportion of languages with a laminal contrast in Pama-Nyungan using the ACL and BM methods. To account for phylogenetic uncertainty, we consider results using a set of 100 Pama-Nyungan trees inferred by Bowern (2015) and described in Macklin-Cordes, Bowern & Round (2021).

6.1 Phylogenetic signal in the binary laminal contrast

In Figure 5b, we saw that the distribution of the laminal contrast in Pama-Nyungan hews closely to major subgroup boundaries, so we will not be surprised if a D test returns a strong confirmation of phylogenetic signal. Figure 6, which

Tab. 1: Phylogenetic signal in the binary presence/absence of a phonemic laminal contrast in 216 Pama-Nyungan languages. D statistic using a sample of 100 reference trees, and p values for the hypotheses of randomness (rejected) and phylogenetic signal (not rejected).

D statistic	p (randomness)	p (phylogenetic signal)
-0.439 (SD 0.019)	0.000 (SD 0.000)	0.987 (SD 0.005)

plots the presence and absence of a laminal contrast against the Pama-Nyungan tree, reinforces this expectation. We tested a set of 216 Pama-Nyungan languages (Round 2019), each coded for the binary presence/absence of the phonemic laminal contrast. To account for phylogenetic uncertainty, the statistic is calculated using 100 individual reference phylogenies.

The 100 results are summarised in Table 1. The mean D statistic obtained is low, at -0.439, indicating that the data is phylogenetically clumped to an even greater degree than expected under a Brownian motion model of evolution. Results like this can emerge when the variable under study has changed only rarely, and the changes have mostly been deep within the tree. This is the case in Pama-Nyungan, where variation in the presence/absence of the laminal contrast is mainly *between* major subgroups rather than within them. Returning to the statistical results, the hypothesis of randomness is rejected ($p < 0.001$) and the hypothesis of phylogenetic signal is not rejected ($p = 0.987 \pm 0.005$). The values of the D statistic have a small standard deviation (0.019), indicating that a similar result is obtained for all 100 reference trees. In sum, the D test results confirm, in a quantitative manner and taking into account our uncertainty in the Pama-Nyungan phylogenetic tree, what our inspection of the map in Figure 5b could only suggest: that the binary presence/absence of the laminal contrast in Pama-Nyungan has strong phylogenetic signal.

6.2 Phylogenetic signal in continuously-valued phonotactic variables

Languages vary not only in what contrastive segments they have but also in how frequently they use them (Frisch, Pierrehumbert & Broe 2004, Hall 2009, Wedel, Kaplan & Jackson 2013, Macklin-Cordes & Round 2020). For example, Pitta Pitta (Blake 1990) and Burduna (Burgman 2007) are similar in that they both contrast laminal stops, nasals and laterals in word-initial position. However, a closer examination reveals notable differences. In word-initial position before /u/, 29% of the consonantal laminals in Pitta Pitta are pre-palatal while

71% are dental, whereas in Burduna the frequencies are reversed, with 68% pre-palatal and just 32% dental. Frequency measures such as these can be viewed as continuous variables that can be investigated for phylogenetic signal (Macklin-Cordes, Bowern & Round 2021). In this section we examine continuous variables of this kind, which describe the relative predominance of pre-palatals versus dentals in nine phonotactic positions, across 76 languages that possess the contrast. Data is from a phonemicised lexical database of Australian languages, which is under development (Round 2017), and which extends and enhances the Chirila database (Bowern 2016). Raw data tables and details of the primary language documentation sources are provided in Section S2 of the Supplementary Materials.

Our choice of nine variables is informed by the typological literature on Australian phonology. One long-established characteristic of Australian laminals is that their relative frequencies are sensitive to the quality of neighbouring vowels (Dixon 1970, 1980).[7] Most Australian languages have three contrastive vowel qualities (Round 2022), with /i/ contexts favouring the laminal pre-palatal, /u/ contexts favouring the dental, and /a/ contexts somewhere in between. Here we examine the relative predominance of pre-palatals in word-initial position before /i,a,u/ and in intervocalic position before /i,a,u/ and after /i,a,u/.[8] We apply the randomisation test described in Section 3.1 and then calculate a K statistic. As in our D test, we address phylogenetic uncertainty by applying the statistical tests using a sample of 100 reference trees.

Results are summarised in Table 2. The randomisation test finds phylogenetic signal to be statistically significant ($p < 0.05$) in all 9 variables and 100 reference trees except in two cases: these were the a_V and V_a contexts, for the same, one tree. Given that both contexts are judged to have significant phylogenetic signal in all other 99 trees in the 100-tree sample, we conclude that phylogenetic signal is present at a stastically significant level in all nine phonotactic variables.

The findings for the K statistic differ among the variables. For the word-initial variables, K is high, ranging from 0.783 to 1.322, whereas for the intervocalic variables it is uniformly lower, ranging from 0.337 to 0.696. In all cases, the standard deviation is low, indicating that similar results are obtained for all 100 reference trees. To put these K values in perspective, Blomberg, Garland & Ives (2003) examined 121 biological traits of a wide variety of plant and animal

[7] The palatal semi-vowel /j/ patterns more freely. In this section we set it aside and examine the consonantal laminals, i.e., laterals, nasals and obstruents.

[8] To minimise error in the values of the variables, we include observations only from those languages in whose lexicons at least 20 consonantal laminals are attested in the relevant phonotactic context (see further, Section S2 the Supplementary Materials).

Tab. 2: Phylogenetic signal in nine continuous variables describing the proportion of laminals which are pre-palatal, in specific phonotactic contexts. K statistic using a sample of 100 reference trees, and p values for the hypothesis of randomness (rejected in all cases).

Context	K	p (randomness)
#_a	0.827 (SD 0.052)	0.001 (SD 0.000)
#_i	1.322 (SD 0.055)	0.001 (SD 0.000)
#_u	0.783 (SD 0.040)	0.001 (SD 0.000)
a_V	0.480 (SD 0.038)	0.002 (SD 0.009)
i_V	0.536 (SD 0.031)	0.002 (SD 0.001)
u_V	0.615 (SD 0.018)	0.001 (SD 0.000)
V_a	0.337 (SD 0.019)	0.015 (SD 0.011)
V_i	0.696 (SD 0.025)	0.001 (SD 0.000)
V_u	0.620 (SD 0.019)	0.003 (SD 0.002)

organisms, finding mean K of 0.35 for behavioral traits, 0.54 for physiology and 0.83 for traits related to body size. Macklin-Cordes, Bowern & Round (2021) estimated K for biphones (sequences of two adjacent phonemes) in Pama-Nyungan and found mean K of 0.52 for biphones of individual segments, and K of 0.63 when segments are binned into groups by place or manner of articulation. This suggests that our laminal phonotactic variables exhibit a level of phylogenetic signal at least as high as many evolved, biological traits, as well as the Pama-Nyungan biphone variables investigated in Macklin-Cordes, Bowern & Round (2021).

The highest K value, at 1.322, is for laminals in word-initial position before /i/. A K value well above 1 is consistent with a scenario in which a linguistic property varies between deep branches of the tree, but much less so within the subgroups below those branches. This is true of Pama-Nyungan laminals word-initially before /i/. In the western half of the family, this position favours pre-palatals, reflecting a typical effect of the neighbouring vowel, whereas in the eastern half, the initial position in a word is one which favours dentals, irrespective of the following vowel.

A novel and consistent finding was that laminals exhibit stronger phylogenetic signal in word-initial position than intervocalically. There are many reasons why this might be so, and here we consider just one. *Pertinacity* (Dresher & Lahiri 2005) refers the perpetuation of linguistic patterns even as the items that instantiate them change. For instance, though a borrowed word may be new, its phonology is often reshaped to conform to the existing patterns in the recipient language (Hyman 1970), which then perpetuates the phonological patterns even as the set of items instantiating them changes. Similarly, if neologisms conform to existing statistical patterns in the lexicon, they too will contribute to pertinacity. Because our phonotactic variables are based on whole lexicons, and not merely a basic vocabulary list, lexical turnover will have been an important contributor to their historical dynamics. If it is the case that word-initial laminals have been subject to more-pertinacious changes than intervocalic laminals, such as more reshaping of borrowed words, or neologism which more closely replicates existing statistical patterns in the lexicon, then this could potentially lead to the difference in phylogenetic signal that we find. Whether there is additional evidence to support this hypothesis remains a question for future research, however the fact that such a hypothesis is able to emerge, illustrates how phylogenetic analysis can supplement the typologist's existing toolkit for generating theoretically interesting hypotheses from the analysis of cross-linguistic data.

6.3 Genealogically-sensitive proportions of languages with a laminal contrast

We turn now to examine the phylogenetically-weighted proportion of Pama-Nyungan languages that have a laminal contrast. We know already, just by counting, that the simple proportion of Pama-Nyungan languages with a laminal contrast is 157/216 = 0.727. Our question here is, what is the proportion when genealogy is taken into account? As discussed in Section 5, there are different methods available for calculating this phylogenetic quantity, just as there are different kinds of non-phylogenetic averages. Here we compare the ACL and BM methods introduced earlier. We account for phylogenetic uncertainty by calculating them with respect to a sample of 100 reference trees. Table 3 reports the results. In this case the answer is broadly similar according to all three methods: the simple proportion is 0.727, the ACL-weighted proportion is somewhat higher, at 0.761 (SD 0.009) and the BM-weighted proportion marginally lower, at 0.705 (SD 0.003). The standard deviations of the phylogenetically weighted proportions are low, indicating that a similar result is obtained for all 100 reference trees. As mentioned in Section 5, an ACL proportion is more sensitive to genealogical structure deep within the tree than the BM method is, thus if we wish to remain conservative about our confidence in deep tree structure, we could conclude that a figure of around 71% (but perhaps as high as 76%) provides a good representation of the proportion of Pama-Nyungan languages that possess a laminal contrast. Note that unlike for balanced sampling, we did not need to discard any data, meaning that our results provide a faithful reflection of the evidence provided by all 216 languages and they do so while taking

Tab. 3: Genealogically sensitive proportions of Pama-Nyungan languages with a laminal contrast.

Simple proportion	ACL weighting	BM weighting
0.727	0.761 (SD 0.009)	0.705 (SD 0.003)

phylogenetic autocorrelation, including our uncertainty about Pama-Nyungan genealogy, into account.

Our case study has illustrated the application of methods and principles introduced in earlier sections. We have confirmed that the presence/absence of a laminal contrast in Pama-Nyungan has significant phylogenetic signal, notwithstanding a long history in the literature of emphasising its apparent areality. An examination of phylogenetic signal in continuously-valued phonotactic variables prompted us to notice a major east-west split in the treatment of word-initial laminals before /i/ and suggested a potential difference in the pertinacity of laminals and their statistical frequencies in word-initial versus intervocalic positions. Finally, having first confirmed the presence of phylogenetic signal, we then calculated genealogically-weighted proportions of the Pama-Nyungan languages which have the laminal contrast. This was done taking into account phylogenetic uncertainty in the Pama-Nyungan tree, and using two weighing methods which allow us to compare the consequences of investing a more conservative or less conservative degree of confidence in the deep-time branching structure of the trees.

7 Discussion

Phylogenetic autocorrelation has long challenged the analysis of comparative data both in linguistics and in other comparative sciences, such as comparative anthropology and comparative biology. The core problem is that many statistical methods require observations that are independent, yet languages, cultures and species are inherently non-independent owing to the way they develop historically. For several decades, comparative fields explored methodological approaches which were broadly parallel, focussed on balanced sampling. Obvious drawbacks of such approaches are that the vast majority of available comparative data must be ignored, and that even then, complete independence remains elusive. In 1985, Felsenstein showed that by focussing on phylogenetically independent contrasts it is possible even under conditions of phylogenetic autocorrelation to extract truly independent observations for subsequent analysis. We have argued that it is nothing more than historical accident that this breakthrough occurred in biology and not in linguistic typology or anthropology, since it is the solution to a problem that is shared across disciplinary boundaries. One of the motivations behind this article, is that while phylogenetic comparative methods have been gaining currency in linguistics, their essential relationship to balanced sampling in linguistic typology has not been clearly articulated, and we hope to have achieved that here.

In Sections 3–6 we introduced concepts and related methods for reckoning with phylogenetic signal, phylogenetic uncertainty and genealogically-sensitive averages. A leitmotif running through that presentation was that phylogenetic comparative methods do not lock the typologist into any single assumption about a phylogeny. On the contrary, because these methods require a precise statement of one's hypothesised phylogeny, it is possible to compare multiple hypotheses and explicitly examine their impacts on the analysis. In this section we expand on some of our earlier points in relation to two topics of central importance in typology: comparison across families and areality.

7.1 Comparison across families and deep-time genealogy

Throughout our paper, we have discussed phylogenetic comparative methods primarily within the scope of a single family. In this single-family, single-tree context we have examined phylogenetic uncertainty, testing for phylogenetic signal and the estimation of genealogically-sensitive averages and proportions. However, in Section 4 at the end of our discussion of uncertainty in phylogentic trees, we mentioned the problem of comparing across language families. We noted that logically, if it is believed that multiple families ultimately are related genealogically, then it is not possible to compare them without implicating a grand phylogeny that links them all. Methods which place all families on equal footing merely do this by positing a rake tree. Thus, as radical as it may sound to say that we must hypothesise a deep-time tree which links currently-distinct families together, this is in fact something linguists have been doing for decades, covertly. Consequently, the question is not whether to use a grand, supra-familial tree but instead, which grand tree to use. Until now, linguists have generally declined to engage in positing grand trees that span beyond the reach of the comparative method, for the eminently good reason, that such trees cannot be demonstrated to be correct. However, as we have emphasised, trees do not need to be verifiably correct to be gainfully used with phylogenetic comparative methods. Instead, trees are hypotheses. Even if we do not, or cannot, know what the correct tree is, we surely can distinguish between more or less plausible hypotheses. Once we view the creation of grand trees as a matter of *hypothesis generation*, then there is every reason to begin working with them earnestly. For readers who find themselves still skeptical, consider the issue presented in the form of this question: Is a rake tree truly the best hypothesis that linguists could come up with about deep-time relatedness, entailing that every language family everywhere in the world is exactly equally related to every other? If our answer is anything other than an unequivocal yes, then we are effectively, tacitly entertaining the existence of other, more plausible grand trees.

To summarise so far, in order to apply phylogenetic comparative methods not only within but also across known families, we join the families in a grand tree. If the grand tree is a rake, then we are effectively continuing current practice in supra-familial language sampling. If the grand tree is otherwise, then we are beginning to explore alternative hypotheses for deep-time relatedness. As with the examples discussed earlier in the paper, phylogenetic comparative methods can be applied to multiple, alternative grand trees in order to reflect phylogenetic uncertainty and to investigate its implications.

Given this state of affairs, it strikes us that an important task for linguistic typology in coming years will be to establish an inventory of deep-time genealogical hypotheses, represented as phylogenies, as key ingredients for phylogenetic typological research, much in the way that the field in previous decades developed a variety of sampling techniques. Hypotheses within this inventory might come from many sources, whether from detailed interdisciplinary studies such as Matsumae et al. (2021) or novel linguistic attempts such as Jäger (2018), or more prosaically in the form of random samples of plausible hypotheses that meet certain constraining assumptions. There is ample scope for innovation. In Section S1 of the Supplementary Materials, we provide an extended description of a set of tools (Round 2021a) designed specifically with linguists in mind, for generating hypotheses about linguistic genealogy either within or across families, by creating and adjusting explicit linguistic phylogenies.

7.2 Areality

In scientific discussions with colleagues, we have encountered the concern that phylogenetic comparative methods cannot work, because they do not take into account the effects of areality (similarly, in published work see e.g. Blench 2015, François 2014). We believe that this concern may follow from a partial misapprehension about what phylogenetic comparative methods ought to be able to achieve. By way of comparison, it would be amiss to argue that a good model of gender should not be incorporated into a sociolinguistic analysis, merely because it does not account for geography. One could argue with good justification that we also desire an account of geography, but that is not the same thing as rejecting the successful model of gender. Similarly, we should not dismiss the breakthrough that Felsenstein achieved, dealing with genealogy far more effectively than in previous methods, merely because areality remains as difficult a problem as it always was. Here we briefly discuss why areality remains a hard problem and what can be done about it.

Viewed in mathematical and statistical terms, phylogenies are rather simple geometric objects. One consequence of their simplicity is that PICs can also be defined in a simple and effective manner. In contrast, the relationships implied by thousands of years of areality, including interactions with languages that have left no direct descendants, are significantly more complex. As mentioned in Section 4, comparative biology is also confronted with similarities shaped by areality, including in high-stakes fields such as bacteriology. Thus it is not for lack of motivation or interest that mathematical biologists are yet to produce methodological solutions to areality that match the solutions for phylogeny. The work is well underway, but the mathematics of *historical networks*, which such phenomena imply, is truly challenging (Elworth et al. 2019).

In this context, it is imperative for typologists to continue grappling with the problem of areality, though not by rejecting phylogenetic comparative methods, but instead by supplementing them. Recent methodological work that addresses areality in concert with phylogenetic comparative methods includes Cathcart et al. (2018) on areality in grammatical change, and Verkerk (2019) on estimating areality effects in relation to phylogenetic uncertainty. Similarly, it will be important to continue to learn more about the empirical facts of areality and its typological implications, to better understand its expected quantitative impact on the performance of phylogenetic comparative methods. For example, in the domain of lexical phylogenetic inference, Bowern et al. (2011) clarified empirical levels of lexical borrowing among hunter-gatherer and small-scale agriculturalist societies, providing crucial empirical knowledge about areality which could then be compared with the results of robustness studies (Greenhill, Currie & Gray 2009), to suggest that at known empirical rates of borrowing, quantitative inference of phylogenies from lexical data should not suffer from significant impairment.

In all likelihood, areality will remain a tough challenge for linguistic typology, as it is for comparative biology, for some time to come. The problems that areality presents are different to and more complex than phylogeny. However, the mere fact that areality is hard is no sound reason to reject the advances offered by phylogenetic comparative methods. Instead, as always, the best available methods for handling genealogy must be supplemented with the current best attempts at handling areality.

8 Conclusions

Typologists are deeply invested in the methodology of balanced sampling, because traditionally it has been our best response to the fundamental challenge of phylogenetic autocorrelation. However, phylogenetic comparative methods provide a better solution to the same problem. The fact that these methods were invented in biology is an accident of history; they could just as well have been invented in linguistics. While phylogenetic comparative methods do not solve all of the problems of typological analysis, they do solve the core challenge of phylogeny. For this reason, we see little reason not to adopt them, apart from inertia and perhaps a little professional envy (given that a linguist did not, in fact, discover them). To assist typologists who are interested in exploring these methods, here we introduced some fundamental concepts and methodological tools, and provided an illustration of their application in a typological case study. In Section S1 of the Supplementary Materials, we introduce computational tools for converting genealogical hypotheses into trees, and using the trees to calculate genealogically-sensitive averages. See also footnotes in Section 3 for references to other, free computational tools for examining phylogenetic signal. Phylogenetic comparative methods will enable typologists for the first time to use all available documentary data when drawing inferences about the diversity of human language, and to begin a far richer discussion on how competing hypotheses about linguistic genealogy—whether in shallow or in deep time—can alter the inferences we draw about the nature of human language from the empirical evidence granted us by today’s seven thousand tongues.

Data availability statement

Data and results files are available on Zenodo at https://doi.org/10.5281/zenodo.5602216. Documentation and code for performing the analysis is available in Supplementary Materials Section S2. The R packages *glottoTrees* (Round 2021a) and *phyloWeights* (Round 2021b) referred to in Supplementary Materials Section S1 are available at https://github.com/erichround/glottoTrees and https://github.com/erichround/phyloWeights.

References

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

Adams, Dean C. 2014. A generalized K statistic for estimating phylogenetic signal from shape and other high-dimensional multivariate data. *Systematic Biology* 63(5). 685–697. https://doi.org/10.1093/sysbio/syu030.

Altschul, Stephen F., Raymond J. Carroll & David J. Lipman. 1989. Weights for data related by a tree. *Journal of Molecular Biology* 207(4). 647–653.

Baker, R. R. & G. A. Parker. 1979. The evolution of bird coloration. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 287(1018). 63–130. https://doi.org/10.1098/rstb.1979.0053.

Bakker, Dik. 2011. Language sampling. In Jae Jung Song (ed.), *The Oxford handbook of linguistic typology*, 100–127. Oxford: Oxford University Press.

Balisi, Mairin, Corinna Casey & Blaire van Valkenburgh. 2018. Dietary specialization is linked to reduced species durations in North American fossil canids. *Royal Society Open Science* 5(4). 171861. https://doi.org/10.1098/rsos.171861.

Bell, Alan. 1978. Language samples. In Joseph H. Greenberg (ed.), *Universals of human language*, vol. 1, 123–156. Stanford, California: Stanford University Press.

Bentz, Christian, Dan Dediu, Annemarie Verkerk & Gerhard Jäger. 2018. The evolution of language families is shaped by the environment beyond neutral drift. *Nature Human Behaviour* 2(11). 816–821. https://doi.org/10.1038/s41562-018-0457-6.

Bickel, Balthasar. 2009. A refined sampling procedure for genealogical control. *STUF - Language Typology and Universals (Sprachtypologie und Universalienforschung)* 61(3). 221. https://doi.org/10.1524/stuf.2008.0022.

Birchall, Joshua. 2015. A comparison of verbal person marking across Tupian languages. *Boletim do Museu Paraense Emílio Goeldi. Ciências Humanas* 10(2). 325–345. https://doi.org/10.1590/1981-81222015000200007.

Blake, Barry J. 1990. Pitta Pitta wordlist. Australian Institute of Aboriginal and Torres Strait Islander Studies, Australian Indigenous Languages Collection. ASEDA 0275. Canberra.

Blasi, Damián E., Susanne Maria Michaelis & Martin Haspelmath. 2017. Grammars are robustly transmitted even during the emergence of creole languages. *Nature Human Behaviour* 1(10). 723–729. https://doi.org/10.1038/s41562-017-0192-4.

Blench, Roger. 2015. *'New mathematical methods' in linguistics constitute the greatest intellectual fraud in the discipline since Chomsky*. Nijmegen, Netherlands.

Blomberg, Simon P., Theodore Garland & Anthony R. Ives. 2003. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution* 57(4). 717–745. https://doi.org/doi:10.1111/j.0014-3820.2003.tb00285.x..

Blomberg, Simon. P. & Theodore Garland. 2002. Tempo and mode in evolution: phylogenetic inertia, adaptation and comparative methods. *Journal of Evolutionary Biology* 15(6). 899–910. https://doi.org/10.1046/j.1420-9101.2002.00472.x.

Borges, Rui, João Paulo Machado, Cidália Gomes, Ana Paula Rocha & Agostinho Antunes. 2019. Measuring phylogenetic signal between categorical traits and phylogenies. *Bioinformatics* 35(11). 1862–1869. https://doi.org/10.1093/bioinformatics/bty800.

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.

Bouckaert, Remco R., Claire Bowern & Quentin D. Atkinson. 2018. The origin and expansion of Pama–Nyungan languages across Australia. *Nature Ecology & Evolution* 2(4). 741–749. https://doi.org/10.1038/s41559-018-0489-3.

Bowern, Claire. 2015. Pama-Nyungan phylogenetics and beyond [plenary address]. In *Lorentz center workshop on phylogenetic methods in linguistics*. Leiden University, Leiden, Netherlands. https://doi.org/10.5281/zenodo.3032846.

Bowern, Claire. 2016. Chirila: Contemporary and historical resources for the Indigenous languages of Australia. *Language Documentation and Conservation* 10. http://hdl.handle.net/10125/24685.

Bowern, Claire, Patience Epps, Russell Gray, Jane Hill, Keith Hunley, Patrick McConvell & Jason Zentz. 2011. Does lateral transmission obscure inheritance in hunter-gatherer languages? *PLoS ONE* 6(9). e25195. https://doi.org/10.1371/journal.pone.0025195.

Burgman, Albert. 2007. *Burduna dictionary: English-Burduna wordlist and thematic wordlist*. South Hedland, Western Australia: Wangka Maya Pilbara Aboriginal Language Centre.

Bybee, Joan L. 1985. *Morphology: A study of the relation between meaning and form* (Typological Studies in Language 9). Amsterdam: John Benjamins Publishing. https://doi.org/10.1075/tsl.9.

Calude, Andreea S. & Annemarie Verkerk. 2016. The typology and diachrony of higher numerals in Indo-European: A phylogenetic comparative study. *Journal of Language Evolution* 1(2). 91–108. https://doi.org/10.1093/jole/lzw003.

Cathcart, Chundra, Gerd Carling, Filip Larsson, Niklas Johansson & Erich R. Round. 2018. Areal pressure in grammatical evolution. *Diachronica* 35(1). 1–34.

Cathcart, Chundra, 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. 1–53.

Chang, Will, David Hall, Chundra Cathcart & Andrew Garrett. 2015. Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. *Language*. 194–244.

Clutton-Brock, T. H. & Paul H. Harvey. 1977. Primate ecology and social organization. *Journal of Zoology* 183(1). 1–39. https://doi.org/10.1111/j.1469-7998.1977.tb04171.x.

Cysouw, Michael. 2005. Quantitative methods in typology = Quantitative Methoden in Der Typologie. In Reinhard Köhler, Gabriel Altman & Rajmund G. Piotrowski (eds.), *Quantitative Linguistik: Ein Internationales Handbuch*, 554–578. Berlin: Walter de Gruyter.

De Maio, Nicola, Alexander V. Alekseyenko, William J. Coleman-Smith, Fabio Pardi, Marc A. Suchard, Asif U. Tamuri, Jakub Truszkowski & Nick Goldman. 2020. Phylogenetic novelty scores: A new approach for weighting genetic sequences. *bioRxiv*. https://doi.org/10.1101/2020.12.03.410100.

Dixon, R. M. W. 1970. Proto-Australian laminals. *Oceanic Linguistics* 9(2). 79–103.

Dixon, R. M. W. 1980. *The languages of Australia*. Cambridge: Cambridge University Press.

Dresher, B. Elan & Aditi Lahiri. 2005. Main stress left in Early Middle English. In Michael Fortescue, Jens Erik Mogensen & Lene Schøsler (eds.), *Historical linguistics 2003:*

Selected papers from the 16th international conference on historical linguistics, 76–85. Amsterdam: John Benjamins. https://doi.org/10.1075/cilt.257.

Dryer, Matthew S. 1989. Large linguistic areas and language sampling. *Studies in Language* 13(2). 257–292. https://doi.org/10.1075/sl.13.2.03dry.

Dryer, Matthew S. & Martin Haspelmath (eds.). 2013. *WALS online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wals.info/.

Dunn, Michael, Tonya Kim Dewey, Carlee Arnett, Thórhallur Eythórsson & Jóhanna Barðdal. 2017. Dative sickness: A phylogenetic analysis of argument structure evolution in Germanic. *Language* 93(1). e1–e22. https://doi.org/10.1353/lan.2017.0012.

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. https://doi.org/10.1038/nature09923.

Elworth, R. A. Leo, Huw A. Ogilvie, Jiafan Zhu & Luay Nakhleh. 2019. Advances in computational methods for phylogenetic networks in the presence of hybridization. In Tandy Warnow (ed.), *Bioinformatics and phylogenetics*, 317–360. Cham, Switzerland: Springer.

Evans, Nicholas. 1995. Current issues in the phonology of Australian languages. In John A. Goldsmith (ed.), *The handbook of phonological theory*, 723–761. Cambridge, MA: Blackwell.

Everett, Caleb. 2017. Languages in drier climates use fewer vowels. *Frontiers in Psychology* 8. https://doi.org/10.3389/fpsyg.2017.01285.

Everett, Caleb. 2021. The sound systems of languages adapt, but to what extent? *Cadernos de Linguística* 2(1). 01–23. https://doi.org/10.25189/2675-4916.2021.v2.n1.id342.

Everett, Caleb, Damián E. Blasi & Seán 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. https://doi.org/10.1073/pnas.1417413112.

Felsenstein, Joseph. 1985. Phylogenies and the comparative method. *The American Naturalist* 125(1). 1–15. https://doi.org/10.1086/284325.

Felsenstein, Joseph. 1988. Phylogenies and quantitative characters. *Annual Review of Ecology and Systematics* 19. 445–471. https://doi.org/10.1146/annurev.es.19.110188.002305.

François, Alexandre. 2014. Trees, Waves, and Linkages: Models of Language Diversification. In Claire Bowern & Bethwyn Evans (eds.), *The Routledge Handbook of Historical Linguistics*, 161–189. Abingdon, UK: Routledge.

Freckleton, Robert P., Paul H. Harvey & Mark Pagel. 2002. Phylogenetic analysis and comparative data: A test and review of evidence. *The American Naturalist* 160(6). 712–726. https://doi.org/10.1086/343873.

Frisch, Stefan A., Janet B. Pierrehumbert & Michael B. Broe. 2004. Similarity avoidance and the OCP. *Natural Language & Linguistic Theory* 22(1). 179–228. https://doi.org/10.1023/B:NALA.0000005557.78535.3c.

Fritz, Susanne A. & Andy Purvis. 2010. Selectivity in mammalian extinction risk and threat types: A new measure of phylogenetic signal strength in binary traits. *Conservation Biology* 24(4). 1042–1051. https://doi.org/10.1111/j.1523-1739.2010.01455.x.

Garland, Theodore, Allan W. Dickerman, Christine M. Janis & Jason A. Jones. 1993. Phylogenetic analysis of covariance by computer simulation. *Systematic Biology* 42(3). 265–292. https://doi.org/10.1093/sysbio/42.3.265.

Gittleman, John L. & Mark Kot. 1990. Adaptation: Statistics and a null model for estimating phylogenetic effects. *Systematic Biology* 39(3). 227–241. https://doi.org/10.2307/2992183.

Grafen, A. 1989. The phylogenetic regression. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 326(1233). 119–157.

Greenhill, Simon J., Thomas E. Currie & Russell D. Gray. 2009. Does horizontal transmission invalidate cultural phylogenies? en. *Proceedings of the Royal Society of London B: Biological Sciences* 276(1665). 2299–2306. https://doi.org/10.1098/rspb.2008.1944.

Hall, Kathleen Currie. 2009. *A probabilistic model of phonological relationships from contrast to allophony*. The Ohio State University Ph.D. Dissertation.

Hansen, Thomas F. & Emília P. Martins. 1996. Translating between microevolutionary process and macroevolutionary patterns: The correlation structure of interspecific data. *Evolution* 50(4). 1404–1417. https://doi.org/10.1111/j.1558-5646.1996.tb03914.x.

Harvey, Paul H. & Georgina M. Mace. 1982. Comparisons between taxa and adaptive trends: Problems of methods. In King's College Sociobiology Group, Cambridge (ed.), *Current problems in sociobiology*, 343–361. Cambridge: Cambridge University Press.

Holden, Clare J. & Ruth Mace. 2003. Spread of cattle led to the loss of matrilineal descent in Africa: A coevolutionary analysis. *Proceedings of the Royal Society of London. Series B: Biological Sciences* 270(1532). 2425–2433. https://doi.org/10.1098/rspb.2003.2535.

Holden, Clare J. & Ruth Mace. 2009. Phylogenetic analysis of the evolution of lactose digestion in adults. *Human Biology* 81(5/6). 597–619. https://doi.org/10.3378/027.081.0609.

Hutchinson, Matthew C., Marília P. Gaiarsa & Daniel B. Stouffer. 2018. Contemporary ecological interactions improve models of past trait evolution. *Systematic Biology* 67(5). 1–13. https://doi.org/10.1093/sysbio/syy012.

Hyman, Larry M. 1970. The role of borrowing in the justification of phonological grammars. *Studies in African Linguistics* 1(1). 1–48. https://journals.linguisticsociety.org/elanguage/sal/article/view/927.html.

Irschick, Duncan J., Laurie J. Vitt, Peter A. Zani & Jonathan B. Losos. 1997. A comparison of evolutionary radiations in mainland and Caribbean anolis lizards. *Ecology* 78(7). 2191–2203. https://doi.org/10.1890/0012-9658(1997)078[2191:ACOERI]2.0.CO;2.

Jaeger, T. Florian, Peter Graff, William Croft & Daniel Pontillo. 2011. Mixed effect models for genetic and areal dependencies in linguistic typology. 15(2). 281–319. https://doi.org/10.1515/lity.2011.021.

Jäger, Gerhard. 2018. Global-scale phylogenetic linguistic inference from lexical resources. *Scientific Data* 5(1). 180189. https://doi.org/10.1038/sdata.2018.189.

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

Jordan, Fiona M., Russell D. Gray, Simon J. Greenhill & Ruth Mace. 2009. Matrilocal residence is ancestral in Austronesian societies. *Proceedings of the Royal Society B: Biological Sciences* 276(1664). 1957–1964. https://doi.org/10.1098/rspb.2009.0088.

Keeling, Patrick J & Jeffrey D Palmer. 2008. Horizontal gene transfer in eukaryotic evolution. *Nature Reviews Genetics* 9(8). 605–618.

Kenny, James Andrew. 1975. *A numerical taxonomy of ethnic units using Murdock's 1967 world sample*. Bloomington: Indiana University Ph.D. thesis.

Kolipakam, Vishnupriya, Fiona M. Jordan, Michael Dunn, Simon J. Greenhill, Remco Bouckaert, Russell D. Gray & Annemarie Verkerk. 2018. A bayesian phylogenetic study of the dravidian language family. *Royal Society open science* 5(3). 171504.

Lavin, Shana R., William H. Karasov, Anthony R. Ives, Kevin M. Middleton & Theodore Garland. 2008. Morphometrics of the avian small intestine compared with that of nonflying mammals: A phylogenetic approach. *Physiological and Biochemical Zoology* 81(5). 526–550. https://doi.org/10.1086/590395.

Levinson, Stephen C., Simon J. Greenhill, Russell D. Gray & Michael Dunn. 2011. Universal typological dependencies should be detectable in the history of language families. 15(2). 509–534. https://doi.org/10.1515/lity.2011.034.

Mace, Ruth, Mark Pagel, John R. Bowen, Keith F. Otterbein, Mark Ridley, Thomas Schweizer & Eckart Voland. 1994. The comparative method in anthropology [and comments and reply]. *Current Anthropology* 35(5). 549–564. https://doi.org/10.1086/204317.

Macklin-Cordes, Jayden L., Claire Bowern & Erich R. Round. 2021. Phylogenetic signal in phonotactics. *Diachronica*. https://doi.org/10.1075/dia.20004.mac.

Macklin-Cordes, Jayden L. & Erich R. Round. 2020. Re-evaluating phoneme frequencies. *Frontiers in psychology* 11. 3181.

Maslova, Elena. 2000a. A dynamic approach to the verification of distributional universals. en. 4(3). Publisher: De Gruyter Mouton Section: Linguistic Typology, 307–333. https://doi.org/10.1515/lity.2000.4.3.307. http://www.degruyter.com/document/doi/10.1515/lity.2000.4.3.307/html (9 September, 2021).

Maslova, Elena. 2000b. Stochastic models in typology: Obstacle or prerequisite? *Linguistic Typology* 4(3). 357–364.

Matsumae, Hiromi, Peter Ranacher, Patrick E. Savage, Damián E. Blasi, Thomas E. Currie, Kae Koganebuchi, Nao Nishida, Takehiro Sato, Hideyuki Tanabe, Atsushi Tajima, Steven Brown, Mark Stoneking, Kentaro K. Shimizu, Hiroki Oota & Balthasar Bickel. 2021. Exploring correlations in genetic and cultural variation across language families in northeast asia. *Science Advances* 7(34). eabd9223. https://doi.org/10.1126/sciadv.abd9223.

Maurits, Luke & Thomas L. Griffiths. 2014. Tracing the roots of syntax with Bayesian phylogenetics. *Proceedings of the National Academy of Sciences* 111(37). 13576–13581. https://doi.org/10.1073/pnas.1319042111.

Miestamo, Matti, Dik Bakker & Antti Arppe. 2016. Sampling for variety. *Linguistic Typology* 20(2). 233–296. https://doi.org/10.1515/lingty-2016-0006.

Moran, P. A. P. 1950. Notes on continuous stochastic phenomena. *Biometrika* 37(1/2). 17–23. https://doi.org/10.2307/2332142.

Münkemüller, Tamara, Sébastien Lavergne, Bruno Bzeznik, Stéphane Dray, Thibaut Jombart, Katja Schiffers & Wilfried Thuiller. 2012. How to measure and test phylogenetic signal. *Methods in Ecology and Evolution* 3(4). 743–756. https://doi.org/10.1111/j.2041-210X.2012.00196.x.

Murdock, George Peter. 1967. *Ethnographic atlas*. Pittsburgh: University of Pittsburgh Press.

Naroll, Raoul. 1961. Two solutions to Galton's problem. *Philosophy of Science* 28(1). 15–39. https://doi.org/10.1086/287778.

Nichols, Johanna. 1992. *Linguistic diversity in space and time*. Chicago: University of Chicago Press. 388 pp.

Nunn, Charles L. 2011. *The comparative approach in evolutionary anthropology and biology*. Chicago: University of Chicago Press.

Pagel, Mark. 1999. Inferring the historical patterns of biological evolution. *Nature* 401(6756). 877–884. https://doi.org/10.1038/44766.

Perkins, Revere D. 1980. *The evolution of culture and grammar*. Buffalo, New York: State University of New York Ph.D. thesis.

Perkins, Revere D. 1988. The covariation of culture and grammar. In Michael Hammond, Edith A. Moravcsik & Jessica R. Wirth (eds.), *Studies in syntactic typology*. Amsterdam: John Benjamins Publishing.

Perkins, Revere D. 1989. Statistical techniques for determining language sample size. *Studies in Language* 13(2). 293–315. https://doi.org/10.1075/sl.13.2.04per.

Piantadosi, Steven T. & Edward Gibson. 2014. Quantitative standards for absolute linguistic universals. *Cognitive Science* 38(4). 736–756. https://doi.org/10.1111/cogs.12088.

Purvis, Andy & Theodore Garland. 1993. Polytomies in comparative analyses of continuous characters. *Systematic Biology* 42(4). 569–575. https://doi.org/10.2307/2992489.

Purvis, Andy, John L. Gittleman & Hang-Kwang Luh. 1994. Truth or consequences: Effects of phylogenetic accuracy on two comparative methods. *Journal of Theoretical Biology* 167(3). 293–300. https://doi.org/10.1006/jtbi.1994.1071.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/.

Revell, Liam J. 2010. Phylogenetic signal and linear regression on species data. *Methods in Ecology and Evolution* 1(4). 319–329. https://doi.org/10.1111/j.2041-210X.2010.00044.x.

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

Revell, Liam J., Luke J. Harmon, David C. Collar & Todd Oakley. 2008. Phylogenetic signal, evolutionary process, and rate. *Systematic Biology* 57(4). 591–601. https://doi.org/10.1080/10635150802302427.

Rijkhoff, Jan & Dik Bakker. 1998. Language sampling. *Linguistic Typology* 2(3). 263–314.

Rijkhoff, Jan, Dik Bakker, Kees Hengeveld & Peter Kahrel. 1993. A method of language sampling. *Studies in Language* 17(1). 169–203. https://doi.org/10.1075/sl.17.1.07rij.

Roberts, Seán G. 2018. Robust, causal, and incremental approaches to investigating linguistic adaptation. *Frontiers in Psychology* 9. https://doi.org/10.3389/fpsyg.2018.00166.

Rohlf, F. James. 2006. A comment on phylogenetic correction. *Evolution* 60(7). 1509–1515. https://doi.org/10.1111/j.0014-3820.2006.tb01229.x.

Round, Erich R. 2017. The AusPhon-lexicon project: 2 million normalized segments across 300 Australian languages. In *47th poznań linguistic meeting*. Poznań, Poland.

Round, Erich R. 2019. Phonemic inventories of Australia [Database of 392 languages]. In Steven Moran & Daniel McCloy (eds.), *PHOIBLE 2.0*. Jena: Max Planck Institute for the Science of Human History.

Round, Erich R. 2021a. *glottoTrees: Phylogenetic trees in Linguistics.* R package version 0.1. https://github.com/erichround/glottoTrees.

Round, Erich R. 2021b. *phyloWeights: Calculation of Genealogically-sensitive Proportions and Averages*. R package version 0.3. https://github.com/erichround/phyloWeights.

Round, Erich R. 2022. Segment inventories. In Claire Bowern (ed.), *Oxford Guide to Australian languages*. Oxford: Oxford University Press.

Stone, Eric A. & Arend Sidow. 2007. Constructing a meaningful evolutionary average at the phylogenetic center of mass. *BMC bioinformatics* 8(1). 222.

Symonds, Matthew R. E. & Simon P. Blomberg. 2014. A primer on phylogenetic generalised least squares. In László Zsolt Garamszegi (ed.), *Modern phylogenetic comparative methods and their application in evolutionary biology: Concepts and practice*, 105–130. Berlin: Springer. https://doi.org/10.1007/978-3-662-43550-2_5.

Symonds, Matthew R. E. & Rod Page. 2002. The effects of topological inaccuracy in evolutionary trees on the phylogenetic comparative method of independent contrasts. *Systematic Biology* 51(4). 541–553. https://doi.org/10.1080/10635150290069977.

Verkerk, Annemarie. 2014. Diachronic change in Indo-European motion event encoding. *Journal of Historical Linguistics* 4(1). 40–83. https://doi.org/10.1075/jhl.4.1.02ver.

Verkerk, Annemarie. 2017. Phylogenetic comparative methods for typologists (Focusing on families and regions: A plea for using phylogenetic comparative methods in linguistic typology). In *Quantitative Analysis in Typology: The logic of choice among methods (workshop at the 12th Conference of the Association for Linguistic Typology*. Australian National University, Canberra, Australia.

Verkerk, Annemarie. 2019. Detecting non-tree-like signal using multiple tree topologies. *Journal of Historical Linguistics* 9(1). 9–69. https://doi.org/10.1075/jhl.17009.ver. (1 January, 2021).

Vingron, Martin & Peter R. Sibbald. 1993. Weighting in sequence space: A comparison of methods in terms of generalized sequences. *Proceedings of the National Academy of Sciences* 90(19). 8777–8781.

Voegelin, C. F. & F. M. Voegelin. 1966. Index of languages of the world. *Anthropological Linguistics* 8(6). 1–222. http://www.jstor.org/stable/30029439.

Wedel, Andrew, Abby Kaplan & Scott Jackson. 2013. High functional load inhibits phonological contrast loss: A corpus study. *Cognition* 128(2). 179–186.

Zheng, Li, Anthony R. Ives, Theodore Garland, Bret R. Larget, Yang Yu & Kunfang Cao. 2009. New multivariate tests for phylogenetic signal and trait correlations applied to ecophysiological phenotypes of nine Manglietia species. *Functional Ecology* 23(6). 1059–1069. https://doi.org/10.1111/j.1365-2435.2009.01596.x.

Zhou, Kevin & Claire Bowern. 2015. Quantifying uncertainty in the phylogenetics of Australian numeral systems. *Proceedings of the Royal Society B* 282(1815). 20151278. https://doi.org/10.1098/rspb.2015.1278.