

Phylogenetic signal in phonotactics

Jayden L. Macklin-Cordes,¹ Claire Bower² and
Erich R. Round^{1,3,4}

¹ The University of Queensland | ² Yale University | ³ University of Surrey
| ⁴ Max Planck Institute for the Science of Human History

Phylogenetic methods have broad potential in linguistics beyond tree inference. Here, we show how a phylogenetic approach opens the possibility of gaining historical insights from entirely new kinds of linguistic data – in this instance, statistical phonotactics. We extract phonotactic data from 112 Pama-Nyungan vocabularies and apply tests for *phylogenetic signal*, quantifying the degree to which the data reflect phylogenetic history. We test three datasets: (1) binary variables recording the presence or absence of *biphones* (two-segment sequences) in a lexicon (2) frequencies of transitions between segments, and (3) frequencies of transitions between natural sound classes. Australian languages have been characterized as having a high degree of phonotactic homogeneity. Nevertheless, we detect phylogenetic signal in all datasets. Phylogenetic signal is greater in finer-grained frequency data than in binary data, and greatest in natural-class-based data. These results demonstrate the viability of employing a new source of readily extractable data in historical and comparative linguistics.

Keywords: historical signal, phylogenetic comparative methods, historical linguistics, phonology, comparative linguistics, linguistic phylogenetics, Pama-Nyungan, Australian languages

1. Introduction

A defining methodological development in 21st century historical linguistics has been the adoption of computational phylogenetic methods for inferring phylogenetic trees of languages (Steiner et al. 2011; Bower 2018a; Jäger 2019). The computational implementation of these methods means that it is possible to analyse large samples of languages, thereby inferring the phylogeny (evolutionary tree) of large language families at a scale and level of internal detail that would be difficult, if not impossible, to ascertain manually by a human researcher (Bower &

Atkinson 2012: 827). There is more to phylogenetics than building trees, and there exists untapped potential to explore the language sciences and human history with a phylogenetic approach. For example, in linguistics, phylogenetic methods have been integrated with geography to infer population movements (Walker & Ribeiro 2011; Bouckaert et al. 2018). In comparative biology, phylogenetic methods have been applied profitably to investigations of community ecology (Webb et al. 2002), ecological niche conservatism (Losos 2008), paleobiology (Sallan & Friedman 2012) and quantitative genetics (Villemereuil & Nakagawa 2014). At the heart of these methods, however, is a sound understanding of the evolutionary dynamics of comparative structures. In this paper, we present a foundational step by detecting *phylogenetic signal*, the tendency of related species (in our case, language varieties) to share greater-than-chance resemblances (Blomberg & Garland 2002), in quantitative phonotactic variation.

Throughout recent advances in linguistic phylogenetics, less attention has been paid to methodological development at the stage of data preparation. Large-scale linguistic phylogenetic studies (e.g., Chang et al. 2015; Bouckaert et al. 2018; Kolipakam et al. 2018) continue, by-and-large, to rely on lexical data which have been manually coded according to the principles of the *comparative method* (as described by Meillet 1925; Durie & Ross 1996; Campbell 2004; Weiss 2014) – the comparative method being the long-standing gold-standard of historical linguistic methodology. This article demonstrates that phonotactics can also present a source of historical information. We find that, for a sample of 112 Pama-Nyungan language varieties, collections of relatively simple and semi-automatically-extracted phonotactic variables (termed *characters* throughout) contain phylogenetic signal. This has positive implications for the utility of such phonotactic data in linguistic phylogenetic inquiry, but also introduces methodological considerations for phonological typology.

In §1–2, we discuss the motivations for looking at phonotactics as a source of historical signal, and we give some broader scientific context that motivates the methodological approach we take later on. In §3–6, we present tests for phylogenetic signal in phonotactic characters extracted from wordlists for 112 Pama-Nyungan language varieties. Section 3 details the materials used and reference phylogeny. Section 4 tests for phylogenetic signal in binary characters that code the presence or absence of biphones (two-segment sequences) in each wordlist, capturing information on the permissibility of certain sequences in a language. Section 5 also tests for phylogenetic signal in biphones, but extracts a finer-grained level of variation by taking into account the relative frequencies of transitions between segments. Section 6 groups segments into natural sound classes and tests for phylogenetic signal in characters coding the relative frequencies of transitions between different classes. We finish in §7–8 with discussion of

the limitations of the study design, implications of the results and directions for future research.

1.1 Motivations

There are at least two reasons why consideration of alternative data sources could be fruitful in historical linguistics. The first is that a bottleneck persists in linguistic phylogenetics when it comes to data processing. The data for most linguistic phylogenetic studies are lexical cognate data – typically binary characters marking the presence or absence of a cognate word in the lexicon of each language – which have been assembled from the manual judgements of expert linguists using the traditional comparative method of historical linguistics (e.g., Weiss 2014). Although data assembled in this way is likely to remain the benchmark in historical linguistics for the foreseeable future, it nevertheless constitutes slow and painstaking work (notwithstanding efforts to automate parts of the process; see List et al. 2017; Rama et al. 2018; List et al. 2018). This restricts the pool of languages that can be included in phylogenetic research to those that have been more thoroughly documented, introducing the risk of a sampling bias, where relatively well-studied regions of the global linguistic landscape are over-represented in historical and comparative work.

The second motivation for considering alternative historical data sources is that there are inherent limitations associated with lexical data. Undetected semantic shifts and borrowed lexical items erode patterns of vertical inheritance in the lexicon of a language. Put another way, these changes create noise in the historical signal in the lexicon of a language. Chance resemblances between non-historically cognate words are another source of noise in lexical data. Eventually, semantic shifts, borrowings and chance resemblances will accumulate to a point where genuine historical signal is indistinguishable from noise. This imposes a maximal cap on the time-depth to which the comparative method can be applied, which is typically assumed to sit somewhere around 10,000 years BP, based on the approximate age of the Afro-Asiatic family (Nichols 1997: 135). Some phylogenetic studies have attempted to push back the time-depth limitations of lexical data by using characters that code for a range of grammatical features, under the rationale that a language's grammatical structures should be more historically stable than its lexicon (Dunn et al. 2005; Rexová et al. 2006). However, contrary to expectation, a recent study suggests that grammatical characters evolve faster than lexical data (Greenhill et al. 2017). Differing rates of evolution are also found in phonology, specifically the rates of change in vowel inventories versus consonant inventories (Moran & Verkerk 2018; Moran et al. 2020). An additional issue with grammatical characters is that the space of possibilities for a grammatical variable is

often restricted. This means that chance similarities due to *homoplasy* (parallel historical changes) will be much more frequent (cf. Chang et al. 2015). For example, many unrelated languages will share the same basic word order by chance, because there is a logical limit on the number of basic word order categories.

1.2 Phonotactics as a source of historical signal

The motivation for considering a language's phonotactics as a potential source of historical information is based partly on practical and partly on theoretical observations. From a practical perspective, it is possible to extract phonotactic data with relative ease, at scale, from otherwise resource-poor languages. This is because the bulk of a language's phonotactic system can be extracted directly from phonemicized wordlists. As long as there is a wordlist of suitable length (Dockum & Bowerman 2019) and a phonological analysis of the language, phonotactic information can be deduced and coded from the sequences of segments found in the wordlist with a high degree of automation. This modest minimum requirement with regards to language resources is a valuable property in less documented linguistic regions of the world. We detail the process of data extraction for this study in §3 below.

An additional benefit of extracting phonotactic data from wordlists is the potential for expanding the depth of comparative datasets. Although macro-scale studies, including hundreds or even thousands of the world's languages (in other words, *broader* datasets), are increasingly common in comparative linguistics, less attention has been paid to the number of characters per language (dataset *depth*). It is quite a different situation in evolutionary biology, where there has been tremendous growth in whole genome sequencing, thanks to technological advances and falling costs (Delsuc et al. 2005; Wortley et al. 2005). This, consequently, has led to tremendous growth in the depth of biological datasets. This is an important consideration because the quantity of characters required by modern computational phylogenetic methods can be substantial (Wortley et al. 2005; Marin et al. 2018). Certainly, phonotactic data is unlikely to approach the scale of large genomic datasets in biology, but it could effectively deepen historical linguistic datasets.

From a theoretical perspective, there is reason to suspect that the phonotactics of a language preserve a degree of historical signal. There is some evidence that when a borrowed word enters the lexicon of a language, speakers tend to adapt it to suit the phonotactic patterns of that language (Hyman 1970; Silverman 1992; Crawford 2009; Kang 2011). Consequently, in such a case the historical phonotactic structure of the lexicon remains largely intact, even as particular ancestral words are lost and replaced (a property termed *pertinacity* by Drescher &

Lahiri 2005). Similarly, historical phonotactic properties of a language will remain in the phonotactics of a language in the case of an undetected semantic shift.

Laboratory evidence shows that speakers have a high degree of sensitivity to the statistical distribution of phonological segments and structures when producing novel words. Examples of such studies include Coleman & Pierrehumbert (1997); Albright & Hayes (2003) and Hayes & Londe (2006), among others (see Gordon 2016: 20–21). Lexical innovation then, should have a relatively conservative impact on the frequency distributions of phonotactic characters. Every new word that enters a language's lexicon will have a minute impact on the frequencies of segments and particular sequences of segments in that language. But, over time, the cumulative effect of a new lexicon entering a language on phonological and phonotactic frequency distributions will be more modest than if speakers generated new words with no regard for existing frequencies. Thus, there is reason to expect that quantitative phonotactic characters are likely to be conservative.

This is not to say that a language's phonotactic system remains completely immobile over time. Phonotactic systems are affected by sound changes and are not totally immune to borrowing. As mentioned above, frequencies of phonotactic characters will shift, however gradually, with the accumulation of lexical innovations. We make no strong claim about phonotactics being the key to a language's history. We merely note there are grounds to expect that phonotactic data will often be historically conservative, relative to cognate data which contains noise from lexical innovation, borrowing and semantic shift. Correspondingly, our hypothesis is that phonotactic data will contain relatively strong historical signal, which we test in §4–6 below.

Many kinds of phonotactic structures exist, which could be studied phylogenetically. Here, because we wish to adhere to the basic methodological principle of studying maximally simple and clear cases first before progressing to more complex ones, we limit ourselves to the simplest of phonotactic structures, namely biphones. That being said, there is every reason to expect our results would generalize, perhaps with interesting variations, to other phonotactic structures. Moreover, many of those structures would have the same benefits as our biphones, in terms of their being readily generated in an automated fashion from wordlists. This will be a promising direction for future investigation.

2. Phylogenetic signal

The concept of *phylogenetic signal* (Blomberg & Garland 2002; Blomberg et al. 2003: 717) originates in comparative biology, where it refers to the tendency of phylogenetically related species to resemble one another to a greater degree than

would otherwise be expected by chance. This expectation derives from the evolutionary history shared between species. Two closely-related species, which share a relatively recent common ancestor, have had less time in which to diverge evolutionarily. We expect more distantly-related species, whose most recent common ancestor lies much further in the past, to tend to be more different, since they have spent longer on separate evolutionary paths.

Phylogenetic signal manifests itself as *phylogenetic autocorrelation* in comparative studies. That is, species observations in a comparative dataset tend not to behave as independent data points, but rather pattern as a function of the amount of shared evolutionary history between species. For many statistical methods that assume data are independent and identically distributed (i.i.d.), this is a problem. Phylogenetic autocorrelation has long been recognized as an issue in linguistic typology and comparative biology, and both fields share comparable histories of developing sampling methodologies that attempt to correct for or offset phylogenetic relatedness in some way. More recent times have seen the rise of *phylogenetic comparative methods*, statistical methods that directly account for phylogenetic autocorrelation, rather than offsetting it, beginning with foundational works by Felsenstein (1985) and Grafen (1989).¹ Although now practically ubiquitous in comparative biology, uptake of phylogenetic comparative methods has been slower in comparative linguistics (notwithstanding studies such as Dunn et al. 2011; Maurits & Griffiths 2014; Verkerk 2014; Birchall 2015; Zhou & Bowerman 2015; Calude & Verkerk 2016; Dunn et al. 2017; Verkerk 2017; Widmer et al. 2017; Blasi et al. 2019).

Since the turn of the century, methods have been developed for explicitly quantifying the degree of phylogenetic signal in a dataset (Revell et al. 2008: 591). Measuring phylogenetic signal can be the first step of a comparative study, to test whether there is sufficient phylogenetic signal to necessitate implementation of a phylogenetic comparative method in a later stage of analysis, or to establish the suitability of standard statistical methods if no phylogenetic signal is detected. Measures of phylogenetic signal can also be used to re-evaluate the validity of older results that pre-date modern phylogenetic comparative methods, as in Freckleton et al. (2002). In other instances, the presence or absence of phylogenetic signal in certain data may be an interesting result in itself. In this study, we present a novel source of linguistic data which traditionally has not been considered a salient source of historical signal for historical linguistic study (indeed, given descriptions of Australian languages, it may have been considered a particularly unlikely source of historical signal; see §3.1). We use measures of phyloge-

1. See Nunn (2011) for discussion.

netic signal to test the hypothesis that our data contain historical information and, therefore, could contribute to future historical linguistic study.

Blomberg et al. (2003) provide a set of statistics for measuring phylogenetic signal, which remains prevalent today (for example, Balisi et al. 2018; Hutchinson et al. 2018; Leff et al. 2018). We use one of these statistics, K . The K statistic has the desirable property of being independent of the size and shape of the phylogenetic tree being investigated, which means that studies with different sample sizes can be compared directly. Briefly (following Blomberg et al. 2003: 722), the calculation of K requires three components: (i) character data (i.e., observations for the variable of interest); (ii) a *reference phylogeny*, a phylogenetic tree which has been generated independently from the character data; and (iii) a *Brownian motion* model of evolution.² These components entail two assumptions of the method: the assumption that the reference phylogeny is an accurate representation of the phylogenetic history of the populations being studied and the assumption that Brownian motion accurately models the evolution of the character data. In practice, the reference phylogeny will be subject to uncertainty. We return to this point in §7 and evaluate the robustness of our results against phylogenetic uncertainty. Similarly, in practice, the Brownian motion model may not be realistic. Nevertheless, it is a simple model and straightforward to implement, and thus commonly used as a starting point before exploring more complex models of evolution later on. We discuss this further in §7 and outline possible extensions to the model for future study, taking sound change processes into account. To the extent that Brownian motion fails to model the evolution of phonotactic characters, this should make it more difficult to detect phylogenetic signal.

The K statistic is calculated by, firstly, taking the mean squared error of the data (MSE_o), as measured from a *phylogenetic mean*,³ and dividing it by the mean squared error of the data (MSE), calculated using a variance-covariance matrix of phylogenetic distances between tips in the reference tree (see Blomberg et al. 2003). 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_o/MSE ratio and vice versa. Thus, a

2. A *Brownian motion* model of evolution describes a model of character evolution where the character can move up or down with equal probability as it evolves through time. Under this model of evolution, variance in character values throughout a phylogeny will increase proportionally as time elapses.

3. Simply taking the mean of some 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. A *phylogenetic mean* is an estimate of the mean which takes into account any overrepresentation by larger subclades (see, for example, Garland & Díaz-Uriarte 1999).

high MSE_o/MSE ratio indicates higher phylogenetic signal. Finally, the observed MSE_o/MSE ratio can be scaled according to the expected MSE_o/MSE ratio given a Brownian motion model of evolution. This gives a statistic, K , which can be compared directly between studies using different trees. When $K=1$, 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. When $K<1$, 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 closer resemblance than would be expected if the variable evolved along the tree following a Brownian motion model of evolution.

Blomberg et al. (2003) also present a *randomization procedure* for testing whether the degree of phylogenetic signal in a dataset is statistically significant. The randomization procedure utilizes Felsenstein's (1985) *phylogenetic independent contrasts* (PICs) method. Felsenstein's insight is that, although two character values (x and y) from two sister taxa cannot be considered independent due to phylogenetic autocorrelation, the contrast between them ($x-y$) is phylogenetically independent, since these values can only diverge in the time since the two sisters split from their most recent common ancestor. Given a set of character data and a phylogenetic tree, Felsenstein (1985) presents a method for harvesting a whole set of phylogenetically independent data points, PICs, which can be used for statistical analysis in lieu of the raw set of observations. Blomberg et al. (2003) take advantage of the expectation that, given a Brownian motion model of evolution, PIC variance is expected to be proportional to time. PICs among more closely-related taxa will tend to be lower than more distant relatives, since they have had less time to diverge from common ancestors. The randomization procedure first extracts PICs for a given character and records the variance. Then, it extracts PICs and records the variance after randomly shuffling character data among taxa (thereby destroying phylogenetic signal). PIC variance is recorded typically for many thousands of such random permutations. If the true PIC variance (for original, unshuffled data) is lower than the variance of PICs in $> 95\%$ of random permutations, the null hypothesis of no phylogenetic signal can be rejected at the conventional 95% confidence level.

In this study, we also use a second statistic, D , which was developed to measure phylogenetic signal in binary data. The D statistic is described by Fritz & Purvis (2010). To summarize briefly, the D statistic is based on the sum of differences between sister tips and sister clades, $\sum d$. First, differences between values at the tips of the tree are summed. Since D concerns binary variables, each taxon

will either have a 0 or 1 value. At the level of the tips, then, all sister tips will either share the same value (in which case, the difference=0) or one tip will have a 0 value and the other will have a 1 value (in which case, the difference=1). Nodes immediately above the tree tips are given the average value of their daughter tips below (which, in a fully bifurcating phylogeny, will either be 0, 0.5 or 1). 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 (i) a distribution that is entirely random relative to phylogenetic structure and (ii) a distribution that is clumped exactly to the degree expected if the character evolved along the tree following a Brownian motion model of evolution. Two permutation procedures are used to determine where these two points lie for a given dataset and phylogenetic tree. Firstly, like Blomberg et al.'s permutation test described above, character values are shuffled at random among tips of the tree many times over, thereby destroying phylogenetic signal. The sums of differences are taken from each random permutation to obtain a distribution of sums of differences, given phylogenetic randomness: Σd_r . Then, to obtain a contrasting distribution of sums of differences, the process of character evolution along the tree following a Brownian motion model is simulated many times over. Since Brownian motion is a model of evolution of continuous characters, and what we need here is a distribution of binary character values, the permutation test simulates the evolution of a continuous-valued character and then simply binarizes the tip values to 0 or 1 by observing whether they fall above or below a threshold value. This threshold is set to whatever level will produce the same proportion of 1s and 0s as observed in the real data. The sums of differences are then taken from each simulation, giving a distribution where phylogenetic signal is present: Σd_b . Finally, the D statistic is determined by scaling the observed sum of differences relative to the means of the two reference distributions just described:

$$(1) \quad D = \frac{\Sigma d_{obs} - \text{mean}(\Sigma d_b)}{\text{mean}(\Sigma d_r) - \text{mean}(\Sigma d_b)}$$

Scaling D in this way provides a standardized statistic with the desirable property that it can be compared between different sets of data, with trees of different sizes and shapes, as with K for continuous characters. One disadvantage of D , however, is that it requires quite large sample sizes (>50), below which it loses statistical power.

Two p values determine the statistical significance of D , one each for the null hypotheses that $D=0$ (phylogenetic signal present) and $D=1$ (the character is distributed randomly relative to phylogenetic structure). These p values are obtained by comparing the observed sum of sister tip/clade differences (Σd) to the two distributions of simulated sums of sister tip/clade differences described above (Σd_r and Σd_b). The fraction of randomly simulated Σd_r scores smaller than the observed Σd is taken as the p value for $H_0(D=1)$. Conversely, the p value for $H_0(D=0)$ is defined as the proportion of the simulated Σd_b scores greater than the observed Σd value.

To summarize, phylogenetic signal is the tendency of closely related species to share closer character resemblances than more distantly related species, due to their relatively greater shared evolutionary history. Given a reference phylogeny and an evolutionary model, it is possible to quantify phylogenetic signal for a character of interest by comparing the contrasts between closely related species to species selected at random from the tree. In this study, we quantify phylogenetic signal in phonotactic data with the aid of a phylogenetic reference tree of Pama-Nyungan languages and the two statistical implementations described above. We use the K statistic developed by Blomberg et al. (2003) to quantify phylogenetic signal in continuous, frequency-based character data and we use the accompanying randomization procedure to test the statistical significance of that signal. Likewise, we use the D statistic developed by Fritz & Purvis (2010) to quantify phylogenetic signal in binary character data and we use its corresponding randomization procedure to test for statistical significance as well – with the randomization procedure, in this latter instance, testing against two null hypotheses: one of phylogenetic signal, and one of random noise.

3. Materials

Our study measures phylogenetic signal in a variety of types of phonotactic characters, extracted using semi-automated methods from wordlists within the Pama-Nyungan family (Australia). Throughout, we take the *doculect* to be our unit of study. A doculect is a language variety as documented in a given resource (Cysouw & Good 2007; Good & Cysouw 2013). That is to say, we treat each wordlist as its own unit of study, without making any claims about the status of the documented language variety as a language or dialect. This agnosticism is advantageous in phylogenetic studies, since the terms ‘language’ and ‘dialect’ imply something about the relationship of a documented language variety to other documented language varieties, and a commitment to one term or the other therefore represents a phylogenetic assumption.

3.1 Language sample

Pama-Nyungan is by far the largest language family on the Australian continent, covering nearly 90% of its landmass (everywhere except for three areas: part of the Top End, part of the Kimberley, and the whole of Tasmania) and encompassing around two-thirds of the languages present at the time of European settlement (Bown & Atkinson 2012: 817). Pama-Nyungan was first proposed and named by Kenneth Hale (Wurm 1963: 136) and it has been the subject of considerable historical linguistic study since this time. Although the family has presented some challenges for historical linguistics, the phylogenetic unity of Pama-Nyungan has been established on traditional historical linguistic grounds (Alpher 2004) with many subgroups identified within (for example, O'Grady et al. 1966; Wurm 1972; Austin 1981). For an overview of the history of Pama-Nyungan classification, see Bown & Koch (2004: Chapter 1–5) and Koch (2014). Bown & Atkinson (2012) perform a computational phylogenetic analysis of Pama-Nyungan using lexical data from 194 language varieties, providing for the first time a fully bifurcating phylogeny of the entire Pama-Nyungan family. Bouckaert et al. (2018) subsequently perform a phylogeographic analysis using the same dataset, but refined and expanded to 306 language varieties and including a geographic element to estimate the point of origin and spread pattern of the family through time and space.

The Pama-Nyungan family provides an excellent test case for the present study. It holds practical advantages which make the task of phonological comparison easier, but it also provides us with a deliberately high bar to clear from a theoretical perspective. Both of these features are a result of the unusual degree of phonological homogeneity observed among Australian languages. Australian languages have long been noted for a degree of similarity between phonological inventories of contrastive segments that is exceptional and unexpected in light of the phylogenetic and geographical breadth of the family, the level of diversity observed in vocabulary and aspects of grammar, and the level of phonological diversity found in comparably-sized families of languages elsewhere in the world. This has been noted as early as Schmidt (1919) and in more recent times by Capell (1956); Voegelin et al. (1963); Dixon (1980); Busby (1982); Hamilton (1996); Baker (2014); Bown (2017) and Round (2021a), among others. This curious level of homogeneity extends to phonotactics too (Dixon 1980; Hamilton 1996; Baker 2014; Round 2021b).

On one hand, the abundance of similar phonological inventories makes the task of comparison between them easier, because it limits the problem of *dataset sparsity*. Consider a character coding the frequency of some sequence of two segments *xy* in a language: This character can only be compared between languages

that contain both x and y segments in their inventories. If a language lacks either segment in its inventory, then the character will be coded as absent or missing (as distinct from 0, where a language possesses both segments but never permits them in sequence). We expect fewer missing values in Australia, where languages tend to share a large proportion of directly comparable segments, when compared to other parts of the world where we would expect to see many more missing values.

On the other hand, an ostensibly high degree of phonological homogeneity, in spite of considerable phylo-genetic diversity, presents challenges for historical linguistics. Baker (2014: 141) and Alpher (2004: 103) have both written on the difficulties for historical reconstruction in Australia because of this. Moreover, a phylogeny implies some degree of historical divergence, but in the case of Australian languages, there would appear to be little by way of phonological divergence, let alone divergences which are phylogenetically patterned. We therefore choose to study an Australian language family as a deliberately difficult test case, where we expect the bar to be set high with respect to detecting phylogenetic signal.

Gasser & Bower (2014) counter prevailing views on Australian phonological homogeneity. They find that common assumptions, of the kind discussed above and commonly found repeated in reference grammars, mask a degree of variation which is otherwise revealed by, firstly, extracting data on segmental inventories directly from wordlists and, secondly, considering segmental frequencies extracted from wordlists. This result motivates our current approach; here, we are also concerned with matters of frequency, extracted directly from language wordlists. However, we look at different kinds of characters, pertaining not to single segments but to biphones, and consider them with respect to their phylogenetic implications.

3.2 Wordlists

Our Pama-Nyungan phonotactic data is extracted from 112 wordlists which are part of a database under development by the last author (Round 2017b), extending the CHIRILA resources for Australian languages (Bower 2016). In this study we restrict our attention to the most accurate sources available, and use only lexical data that is compiled by trained linguists and for which the underlying dataset is available in published or archived form. Additionally, we restrict our sample to wordlists containing a minimum of 250 words. We include this cut-off since measurement accuracy is a concern for smaller wordlists. A documented wordlist is necessarily only a subset of the complete lexicon of a language and it is unclear how big a wordlist must be before frequency statistics begin to stabilize around a sufficient level of accuracy. There is some work in this space concerning frequencies of single segments (Dockum & Bower 2019), suggesting a

rapid decline in the accuracy of phoneme frequencies as wordlists drop below 250 words. Longer wordlists will always be better, however we select 250 words as a reasonable compromise which maintains a generally broad coverage of Pama-Nyungan languages. The 112 language varieties in our sample represent 26 of the 36 major Pama-Nyungan subgroups and 3 isolates listed in Bown (2018b). We return to the subject of wordlist sizes and potential implications for our results in §7.1.

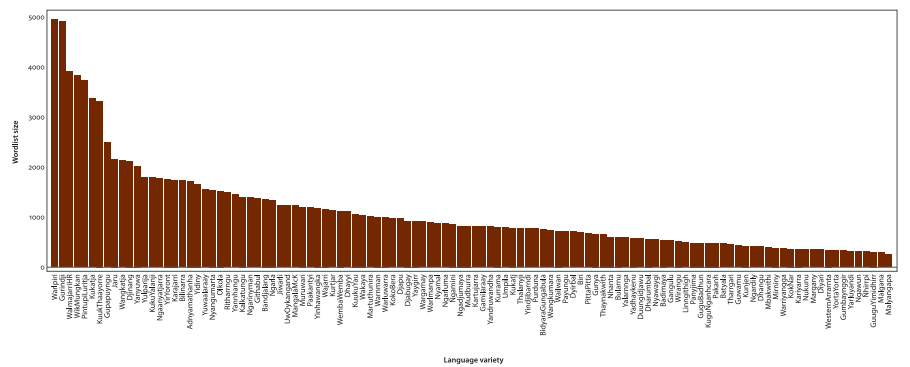


Figure 1. Lexicon sizes.

Bibliographic details for all underlying data are available in §S2 of the Supplementary Information (<https://doi.org/10.1075/dia.20004.mac.additional>). Owing to differences in the length of primary sources, there is considerable diversity in the size of the lexicons we use. As shown in Figure 1, the difference from smallest to largest is over an order of magnitude (min. 250, max. 4955), with the middle fifty percent between 518 and 1364 items. Mean lexicon size is 1109 (*SD* 913).

Original source data, which is typically orthographic and, if digital, is sometimes mixed with metadata or other extraneous material, has undergone extensive data scrubbing, conversion to phonemic form using language-specific orthography profiles (Moran & Cysouw 2018), and additional automated and manual error checking. These procedures ensure basic data cleanliness. Separately however, it has long been recognized that the segmental-phonological analysis of languages is a non-deterministic process (Chao 1934; Hockett 1963; Hyman 2008; Drescher 2009). Two linguists faced with the same data may produce different analyses, not due to error but due to different applications of the very many analytic criteria that figure into any analysis of segments. Consequently, the cross-linguistic phonological record varies not only according to language facts per se, but due also to variation in the practice of linguistic analysis. Recent literature (Lass 1984; Hyman 2008; Van der Hulst 2017; Round 2017a; Kiparsky 2018) emphasizes the

value of normalizing source descriptions prior to the analysis of cross-linguistic phonological datasets. This is not an information destroying process – it does not ‘standardize’ languages – but it may shift information from one part of the representation (e.g., contrast between individual symbols) to another (e.g., contrasts between sequences of symbols), in order that information is located in a comparable way across the languages in the dataset, and therefore is more amenable to comparative analysis. Our wordlist data is normalized in this sense. Complex segments are split into simple sequences (e.g., prenasalized stops are split into a homorganic nasal + stop sequence); long vowels are represented as a sequence of identical short vowels, and vowel-glide-vowel sequences in which the glide is homorganic with either vowel are normalized to vowel-vowel; fortis consonants are represented as a sequence of identical short consonants, and positionally neutralized fortis/lenis stops as singletons; laminal consonants which do not figure in a pre-palatal versus dental opposition are represented as palatal, and rhotic glides which do not figure in an alveolar versus post-alveolar opposition are represented as post-alveolar (see also Round 2019a; Round 2019b). The phonotactic character sets used in this study were extracted from these normalized, comparably segmented wordlists.

3.3 Reference phylogeny

The reference phylogeny we use is a maximum clade credibility tree⁴ of 285 Pama-Nyungan language varieties inferred using lexical cognate characters by the second author (Figure S1, Supplementary Information). It was inferred independently of this study, prior to this study’s conception and without the involvement of the first and third authors. It was inferred using the same Stochastic Dollo model as Bownern & Atkinson (2012), but with an expanded and refined dataset. Further details of the model and phylogeny construction are described in Bownern & Atkinson (2012); Bownern (2015) and Bouckaert et al. (2018). The cognate dataset used to infer the reference phylogeny is available on Zenodo (Bownern 2018b). See §S1 of the Supplementary Information for more information on the reference phylogeny.

4. Bayesian phylogenetic methods return not a single phylogenetic tree but a posterior distribution of many possible trees. These trees can be summarized into a single maximum clade credibility tree with confidence levels for each node in the tree, corresponding to how frequently that node appears in the posterior sample. It is the maximum clade credibility tree that we use for a reference phylogeny in this study. See Bownern & Atkinson (2012) and Bouckaert et al. (2018) for a full explanation of the methods used to infer the phylogenies considered in this section.

We considered a reference tree from a newer phylogeographic analysis of Pama-Nyungan based on largely the same data plus further expansion to 304 doculects and continued refinement (Bouckaert et al. 2018), however, we opted against its use for this particular study. The reason for this is that, although Bayesian inference of phylogenetic tree topology is considered generally robust to the levels of lexical borrowing observed among Pama-Nyungan languages (Greenhill et al. 2009; Bowern et al. 2011), borrowing still has the effect of reducing branch lengths across the tree (Greenhill et al. 2009). This effect, and consequently the accuracy of branch length estimates, is equally applicable to both trees considered here. However, the geographic element in the phylogeographic study uses, in part, branch lengths to model geographic dispersal. The posterior distribution of trees, which is jointly informed by cognate data and geography, may therefore show a bias towards geographically proximal languages whose apparent divergence times have been reduced by high rates of borrowing. Thus, although branch length estimates will be impacted by borrowing in any phylogenetic study of Pama-Nyungan, there is more chance of borrowing affecting topology in the phylogeographic study.

We consider it unlikely that the overall conclusions of the study would be altered by the choice of which version of the Pama-Nyungan phylogeny we use as a reference tree. Each of the studies referenced above produced highly congruent Pama-Nyungan phylogenies (for a detailed comparison, see Bouckaert et al. 2018). Furthermore, Bouckaert et al. (2018) features fixed clade priors based on subgroups identified in earlier studies, so topological differences are constrained to some extent by design. Nevertheless, the accuracy of the reference tree is a key assumption of the methods we use in this study, and thus phylogenetic uncertainty is an important consideration. We return to this point in §7.1 and evaluate the overall robustness of our results to phylogenetic uncertainty by replicating a subset of phylogenetic signal tests over a posterior sample of trees.

As discussed above, we treat each wordlist in our study as its own doculect. The reference tree in this study was inferred using a similar approach, while remaining less-committal about the particular status of the unit of analysis. Resources were sometimes combined for a particular language, but they are also frequently broken up into separate units, particularly when the resources come from different authors and different time periods. We have taken care to match the wordlists in this study to their exact or best corresponding tip in the reference tree. In most cases, the wordlists we use here are the same as those used to generate the cognacy judgements used to infer the reference phylogeny. In other cases, we use a different source to the one used for the reference phylogeny but there is, nevertheless, a straightforward one-to-one mapping between the language variety our wordlist represents and a corresponding tip in the tree. In one case, our

Mudburra source (Nash et al. 1988) matches neither of the sources for the two Mudburra tips in the reference phylogeny. However, the two varieties in the reference phylogeny have the same date. This entails that when either of them is removed from the tree, the exact same result is obtained in terms of tree geometry, which is what is significant for our investigation. Accordingly, we remove one and match our source to the other.

4. Phylogenetic signal in binary phonotactic data

In the simplest case, the phonotactics of different languages may be compared in terms of which sequences of two segments (*biphones*) they permit and which they do not. If claims about the relative homogeneity of phonotactic constraints in Australian languages hold, then we would expect this kind of comparison to yield little, if any, phylogenetic signal.

In this test, we construct the dataset as follows: We automatically extract from all wordlists every unique sequence of two segments – or more accurately, sequences of xy where each of x and y is either a phonological segment or a word boundary '#'. Each sequence becomes a character (variable) in the dataset, for which every language receives a binary value: 1 if the sequence xy is found in the language's wordlist (even if only once); 0 if the language contains both segments in its inventory but the sequence xy never appears in its wordlist; or, NA (not applicable, missing) if the language does not contain one (or both) of either segment x or y in its inventory (and therefore cannot contain the sequence xy a priori). Binary data of this kind represents sequence permissibility: Where a language contains both segments in its inventory, it will either permit them to appear together in sequence or it will not. In this respect, the information encoded by these characters is similar to what one might find in the phonotactics section of a descriptive grammar, where one often encounters a description in prose and/or a basic tabulation of which segments are permitted and where, within syllable and word structures. However, this kind of information is also, in a sense, quite coarsegrained, since there are only two possible values. A sequence which is very common in one language will be coded in exactly the same way as a sequence which only appears a handful of times in another language.

We apply the D test individually to each character in the dataset that meets two conditions: at least 50 non-missing values (due to the aforementioned reliability issue with sample sizes smaller than this) and at least one instance of variation (we do not test characters where all languages share identical 1 or 0 values). Given the extensive history of description of Australian languages as phonotactically homogenous, our prior expectation is that testing binary data will fail to

yield significant phylogenetic signal. Indeed, we might expect that D will fall significantly below 0, indicating that values are clumped among tips on the reference phylogeny even more conservatively than would be expected if they had evolved in the same phylogenetic pattern as lexical data.

To evaluate the statistical significance of D for any given character, a p value is estimated for each of two null hypotheses: The null hypothesis that $D=1$ ($H_0(D=1)$, character values are distributed randomly with regards to phylogeny) and the null hypothesis that $D=0$ ($H_0(D=0)$, character values are distributed as could be expected if the character has evolved along the phylogeny according to a Brownian motion model). Each p value is calculated using the permutation procedure described at the end of §2. Our distributions of randomly simulated and phylogenetically simulated sums of sister tip/clade differences (Σd_r and Σd_b) are obtained via 10,000 permutations per biphone. The conventional cutoff for statistical significance is $p=0.05$. Here, we use the corresponding Bonferroni-corrected cutoff of 0.025.⁵ For any given character, there are six possible results:

- D is significantly below 0. Character values are even more tightly clumped among close relatives than Brownian motion alone would lead us to expect.
- D is significantly below 1 and not significantly different from 0. The data patterns phylogenetically, i.e., there is phylogenetic signal.
- D is significantly above 0 and below 1. The data is neither clearly random nor clearly phylogenetic.
- D is significantly above 0 and not significantly different from 1. It is consistent with randomness, not phylogeny.
- D is significantly above 1. It is even more dispersed than expected via a random process.
- D is not significantly distinct from 0 nor 1. The patterning of the data is indeterminate, and cannot be distinguished from randomness nor from Brownian phylogenetic evolution.⁶

5. Bonferroni correction is used because the conventional threshold for statistical significance, 0.05, represents the expected chance of a false discovery (false positive). This figure is known as the type I error rate (α). The chance of a false discovery is multiplied when multiple tests are carried out. Bonferroni correction, which involves dividing the threshold for statistical significance by the number of tests being conducted, ensures that the chance of observing a false positive in any of the set of tests remains at the conventional rate, $\alpha=0.05$. In our case, two null hypotheses are tested for each character, hence we divide the threshold for statistical significance by two, ensuring the chance of a false positive for any particular character is 0.05.

6. Technically, an indeterminate result can also arise in the seemingly contradictory instance that D is significantly below 0 but not significantly distinct from 1. This can happen when the distributions of phylogenetic and random simulations are highly overlapping and the random

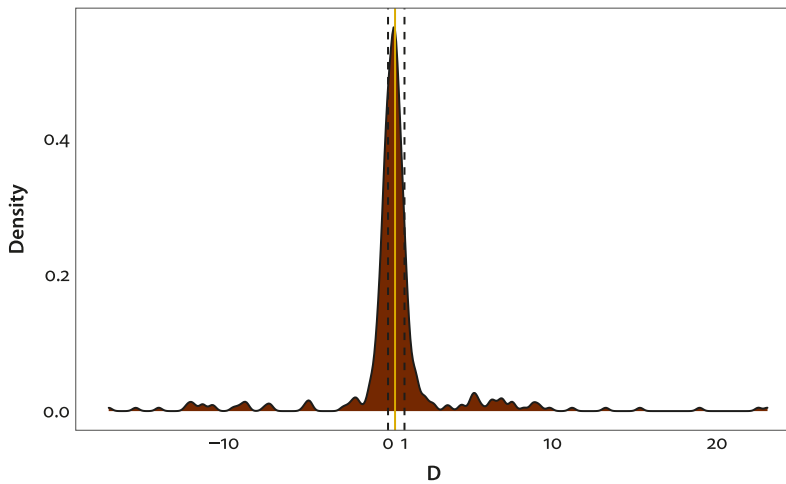


Figure 2. Density of D estimates for binary biphone characters. Dotted lines mark $D=0$, the phylogenetic expectation, and $D=1$, the random expectation. Mean D for all characters is 0.43, marked in red.

To summarize, the testing procedure proceeds as follows. For each binary biphone character, if the character has at least 50 non-NA values *and* the character has at least one ‘i’ and one ‘o’ value (i.e., not every value is identical), then test as follows (otherwise discard):

- Calculate D .
- Conduct randomization procedure to calculate p for $H_o(D=0)$.
- Conduct randomization procedure to calculate p for $H_o(D=1)$.

A result is interpreted from the combination of D and two p values.

distribution is wider and flatter than the phylogenetic distribution. We observed one instance of this for a biphone with a single ‘i’ value and 72 ‘o’ values. As one would expect when every language but one shares the same value, this biphone is more clumped than in almost all simulations – 99.25% of phylogenetic simulations and 98.57% of random simulations. This makes the biphone statistically significantly more clumped than the phylogenetic expectation but fractionally misses the threshold for statistical significance (98.75% for a Bonferroni-corrected, two-tailed p test) in determining whether it is significantly different from the random expectation. This rare anomaly can only occur in cases of extremely skewed data (for example, when all languages share an identical value except for one outlier language), where phylogenetic and random simulations tend to overlap. We discuss the effect of highly skewed data and the effect of removing such data in §4.2 below.

4.1 Results for binary phonotactic data

We estimate D for 415 biphone characters using a script based on the *phylo.d* function in the *caper* package (Orme et al. 2013), implemented in the statistical software *R* (R Core Team 2017).⁷ The 415 D values cluster centrally around a mean of 0.43. The distribution is leptokurtic (kurtosis=11.42), meaning there are more outliers relative to a normal distribution, making the distribution appear as a tall, narrow peak with long tails (Figure 2). The standard deviation is large (3.74).

The D test was of indeterminate significance for half of all characters (234 characters, 56% of the dataset). The D scores for 160 characters (39% of the total dataset) show evidence of phylogenetic signal. Just 17 characters (4%) show the opposite result, where the character is consistent with randomness and there is no phylogenetic signal present. Both null hypotheses are rejected for the remaining 4 characters, all of which are more clumped than their phylogenetic expectation. None fall somewhere between phylogenetic and random expectations (where D is significantly above 0 but significantly below 1), nor are any significantly more dispersed than the random expectation. The distribution of these results among different biphone characters is plotted in Figure 3. Note that we expect around 5% of null hypothesis rejections (approximately 9 of 185 rejected null hypotheses) to be false discoveries. Nevertheless, when considering the whole dataset as an ensemble rather than each character individually, a general result can be discerned. The clearest conclusion is that binary, permissibility-based characters tend to be low yielding in information, giving a statistically significant outcome in fewer than half of cases. Nevertheless, where a significant result can be determined, phylogenetic signal does tend to be present – to a degree that is perhaps surprising in light of previous literature describing the relative homogeneity of Australian phonotactic restrictions and their lack of utility in historical endeavours. This result suggests there may be a greater degree of historical information contained in Pama-Nyungan phonotactics than previously thought. However, it may be that a finer-grained approach to data extraction is needed in order to detect it.

These results can be compared to two earlier studies performing the same test on much smaller samples of languages. In the first, Macklin-Cordes & Round (2015) find no evidence for phylogenetic signal in the Yolngu subgroup of Pama-Nyungan – rather, data are significantly over-clumped, suggesting a higher degree of conservatism in phonotactic restrictions relative to lexical data. They fail to

7. The dataset for this and subsequent tests are available on Zenodo at <http://doi.org/10.5281/zenodo.3988775>. This repository also includes a full table of results and the R scripts used to perform the analysis and produce figures for the paper. See §S3 of the Supplementary Information (<https://doi.org/10.1075/dia.20004.mac.additional>) for usage instructions and a full description of these materials.

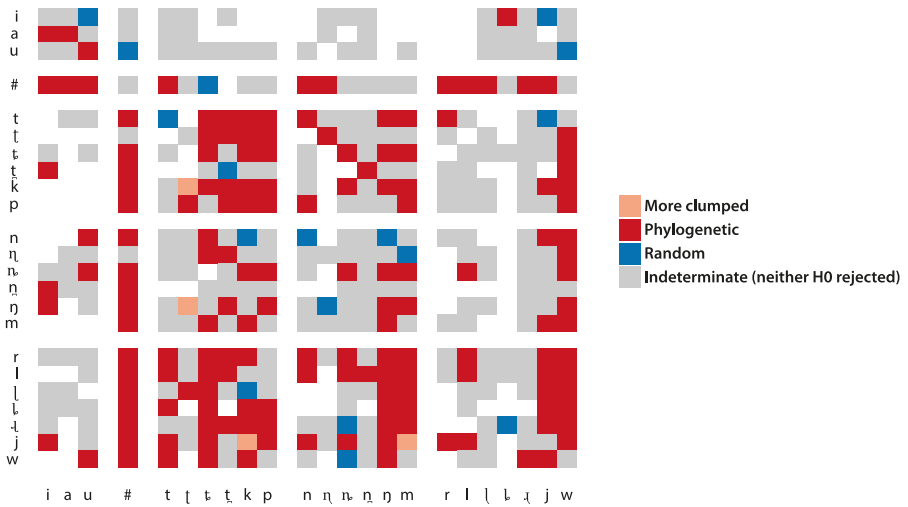


Figure 3. Phylogenetic signal significance testing for binary biphone characters. This grid colour-codes each biphone character according to the results of its respective significance tests. The grid is arranged such that the vertical axis represents the first segment of the biphone and the horizontal axis represents the second segment. Besides a tendency for phonotactic restrictions at word boundaries to show phylogenetic signal, few patterns stand out.

reject a null hypothesis of $D=0$ for Ngumpin-Yapa, suggesting there may be a degree of phylogenetic signal in the Ngumpin-Yapa dataset. However, the pilot study results should be treated with caution – particularly the failure to reject the $D=0$ null hypothesis in the case of Ngumpin-Yapa – due to the small sample sizes (10 languages for Ngumpin-Yapa, 7 for Yolngu), well below the minimum of 50 taxa recommended by Fritz & Purvis (2010). Dockum (2018) performs the same analysis using biphone characters from 20 Tai language varieties of the Kra-Dai family. In contrast to Macklin-Cordes & Round (2015), Dockum finds some evidence of phylogenetic signal in the Tai data and suggests perhaps the earlier result was due to insufficient variation in that particular language sample rather than a limitation of binary biphone characters per se. Although the low information yield from binary data is to be expected, our results here appear to support Dockum's conclusion.

4.2 Robustness checks

Given a sufficient number of taxa for which data are available (>50), D scores should reflect a degree of phylogenetic signal present in the data, independently

of tree size (the number of taxa) and shape (branching patterns). To check this, in Figure 4(A) for each character we plot its D score against the number of doculects for which it had non-missing values. Irrespective of the number of doculects that supply non-missing values, the D scores appear to cluster centrally around mean D for the dataset, suggesting that D is not being unduly affected by missing values for particular characters. A second check leads to rather different results, however. We check whether skewed distributions of character values affects D scores (Figure 4(B)). Here, we consider the distribution of 1s and 0s for each character and plot D against how skewed the distribution of character values is towards a particular value. For example, a character where there are 108 ‘1’ values in the dataset and only 4 ‘0’ values will have a skewing rating of 0.96 (the count of ‘1’ values, 108, divided by 112 total observations). Here, we find that when the ratio of 1s and 0s for a character is highly unequal, estimates of D tend towards extreme magnitudes while also being unrevealing, that is, statistically indistinguishable from 0 and 1.⁸ As described above, Australian languages are known for homogeneity in phonological inventories and phonotactic restrictions, so consequently there are many characters with skewed distributions affecting the results. In Figures 5–6, we plot a subset of the D test results, restricted to characters with less skewing.

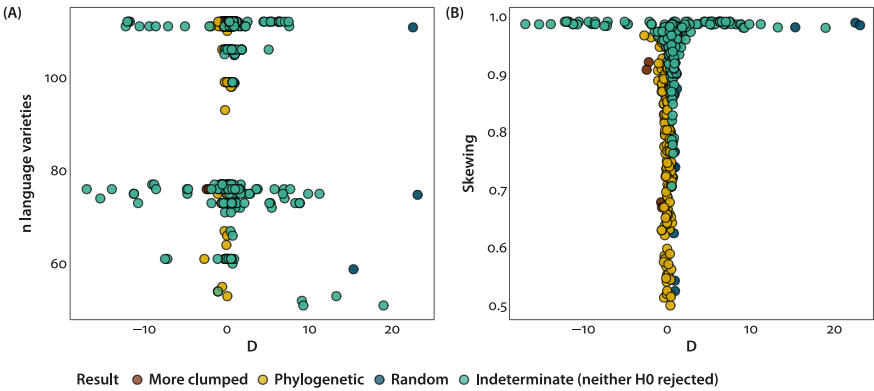


Figure 4. Scatterplot of D scores against (A) the number of doculects with non-missing values for each character and (B) the skewing of the distribution of 1s and 0s for each character. D clusters evenly around the mean regardless of the number of missing values. Variation in D increases greatly among characters where all but 1 or a few doculects share the same value, but the results are overwhelmingly not significant.

8. Note that it is a desirable feature of the D test that it should return a lack of significance when there is a near-complete lack of variability in the data for the test to evaluate. This is an issue with the data, not the test.

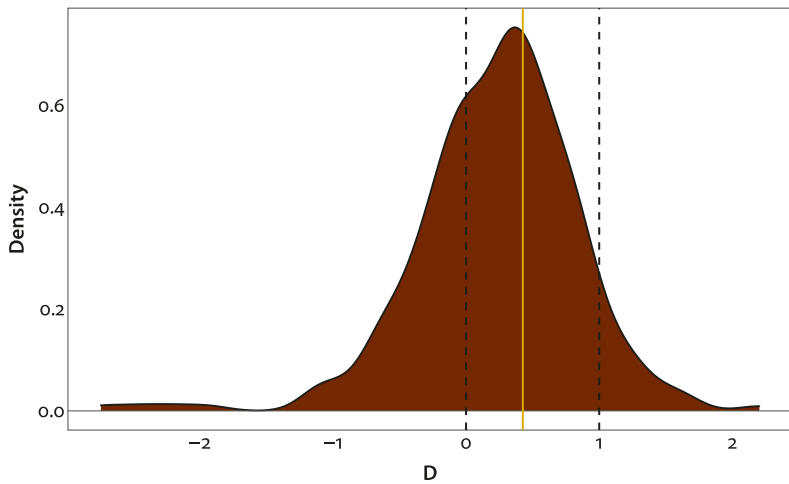


Figure 5. Density of D estimates for binary biphone characters where character values are skewed less than 97:3.

5. Phylogenetic signal in continuous phonotactic data

We test whether a higher degree of phylogenetic signal is detectable in continuous-valued biphone characters. As in the previous test, we take every possible sequence of two segments, or biphones, in our sample of 112 Pama-Nyungan wordlists. In this case, however, rather than simply coding for the presence of a biphone in a language's lexicon, we consider the relative frequencies of transitions between the segments in that biphone across the language's lexicon. For each biphone character, we take two values: The Markov chain forward transition probability – that is, for a biphone xy , the probability of x being followed by y , normalized over all instances of x . This captures, if only in a basic way, our awareness that words do not consist of strings of independent segments, but rather the probability of observing some segment is very much dependent on what came before it. Secondly, we take Markov chain backward transition probabilities – that is, for the biphone xy , the probability of y being preceded by x , normalized over all instances of y in the lexicon. The frequency characters we extract come from wordlists. This is advantageous in that they are somewhat independent of word frequency effects since each word is counted only once, in contrast to frequencies extracted from language corpora. On the other hand, speakers show sensitivity to phoneme frequencies in language use (for example, when coining novel words) (Coleman & Pierrehumbert 1997; Zuraw 2000; Ernestus & Baayen 2003; Albright & Hayes 2003; Eddington 2004; Hayes & Londe 2006; Gordon 2016) so word frequency will likely have some effect on phoneme and biphone frequency even in

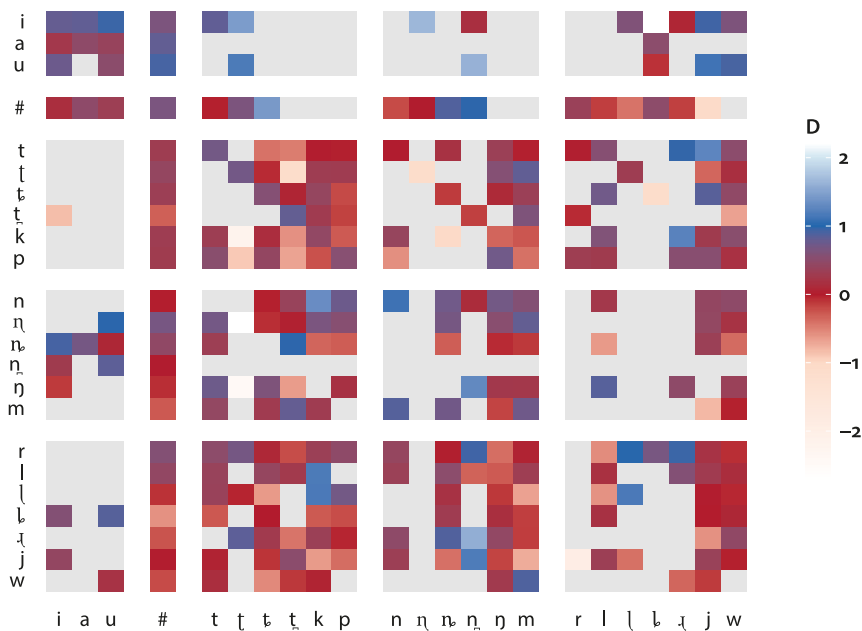


Figure 6. *D* scores for binary biphone characters where character values are skewed less than 97:3. Red shades show a character’s proximity to $D=0$ – darker red indicates stronger phylogenetic signal. Blue shades show a character’s proximity to $D=1$, where character values are distributed randomly. This heat grid shows the proximity of each individual character’s *D* value to $D=0$, its expectation if the character evolved along the phylogeny following a Brownian motion process. The vertical axis represents the first segment of the biphone, the horizontal axis represents the second segment.

a wordlist. Investigation of phylogenetic signal in frequency characters extracted from corpora versus wordlists may be a possibility for future study.

We quantify phylogenetic signal by estimating *K* (Blomberg et al. 2003) individually for each character, using the *multiPhylosignal* function, in the *picante* package (Kembel et al. 2010), in *R* statistical software. The *K* test has somewhat greater statistical power than the *D* test, enabling us to apply the test to characters with as few as 20 non-missing values. Calculation of *K* works with non-zero values only, so zero values (where the language contains both segments *x* and *y* but *x* is never followed by *y*, or vice versa) are considered not applicable and removed from calculation. A total of 490 characters (245 biphone forward transition probabilities and 245 backward transition probabilities) meet the criterion of at least 20 languages with non-missing and non-zero values for testing. Subsequently, to evaluate whether the level of phylogenetic signal is significant for a given character,

we conduct Blomberg, Garland and Ives' (2003) randomization procedure with 10,000 random permutations per character.

To summarize, this testing procedure proceeds as follows. For each biphone frequency character, if the character has at least 20 non-NA values *and* the character has at least two unique frequency values (i.e., not every character value is the same), then test as follows (otherwise discard):

- Calculate K .
- Conduct randomization procedure to calculate p for $H_0(K=0)$.

Mean K for all 490 characters is 0.54 (SD 0.21) (Figure 7). This is comparable to certain physiological traits presented as examples of biological traits with a high degree of phylogenetic signal by Blomberg et al. (2003), for example, $K=0.55$ for log body mass of primates. Using the Blomberg et al. (2003) randomization procedure, we find a statistically significant degree of phylogenetic signal for 351 of 490 characters (178 forward transition characters, 173 backward transition characters), or 72% of the total dataset.

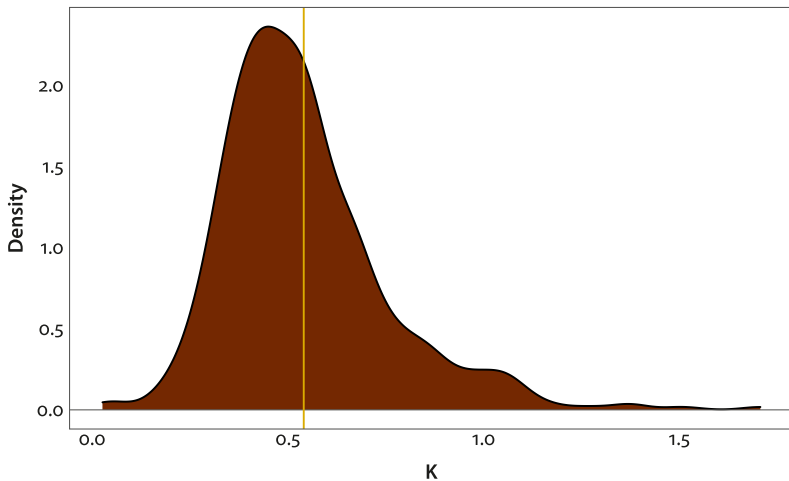


Figure 7. Density of K scores for all frequency-based biphone character (frequencies of both forward and backward transitions between segments).

We consider whether phylogenetic signal is higher or lower in certain kinds of biphone characters. Figure 8 shows a matrix of K scores for forward transition characters, with rows and columns arranged by phonological natural class. No clear pattern stands out.

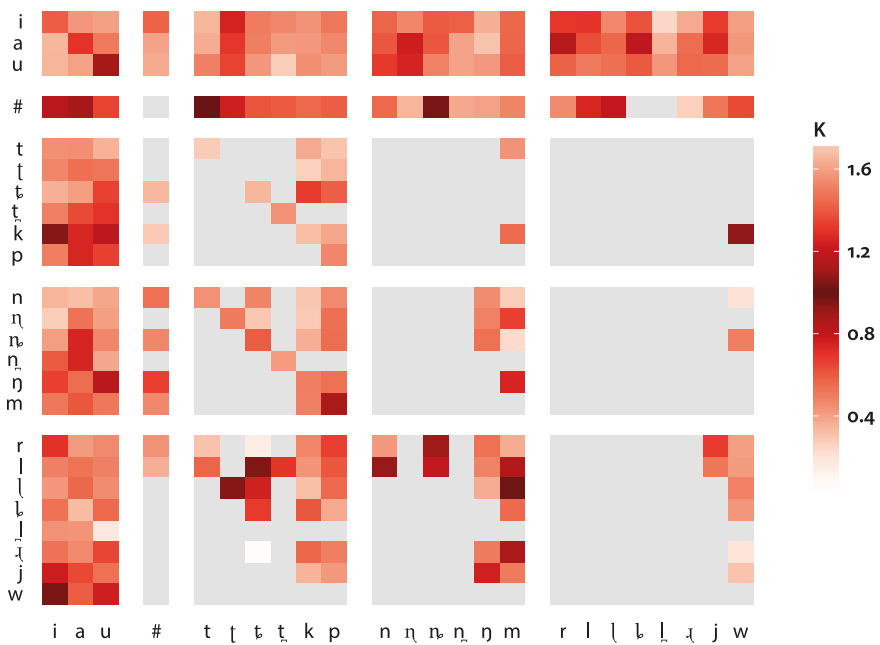


Figure 8. Phylogenetic signal for forward transition frequencies. This heat grid shows K scores for biphone characters (forward transition frequencies only). Each square represents a biphone (where the first segment is listed on the vertical axis and second segment on the horizontal axis). Data points are taken from the frequencies of each biphone, xy , over the total frequency of segment x in each language, and then phylogenetic signal K is measured for each biphone. Darker red shades indicate a stronger degree of phylogenetic signal. As with the D test, no clear pattern of high versus low K scores stands out, although there is a high degree of phylogenetic signal in the dataset overall. The symbols / t , η , l / represent alveo-palatals.

5.1 Robustness checks

Although K is intended to be a measure of phylogenetic signal that is independent of tree size and shape, tree size and shape can have some effect on results in practice (Münkemüller et al. 2012). We wish to check that the Pama-Nyungan tree does not contain any unusual properties that could cause either the K statistic or the randomization procedure to perform unexpectedly. To do this, we allow simulated characters to evolve specifically along the Pama-Nyungan reference tree. We vary the model of evolution, between perfect Brownian motion along the entire tree and pure randomness generated directly at the tips of the tree, by mixing different strengths of Brownian phylogenetic signal and non-phylogenetic noise. 1000 traits are simulated at each percentage point interval for

100,000 total simulated traits (in other words, 1000 traits simulated with 100% Brownian motion, then 1000 traits simulated with 99% Brownian motion and 1% randomness, and so on, until the traits evolve 100% at random). Each simulated trait is then tested for statistical significance using the randomization procedure described in §2, with 1000 repetitions to determine a p value. In a robust testing scenario, K will scale appropriately between 0 and 1 according to the level of Brownian motion and random noise being simulated, and the randomization procedure will distinguish between traits with and without a significant degree of phylogenetic signal with a satisfactory amount of Type I (false positive) and Type II (false negative) errors.

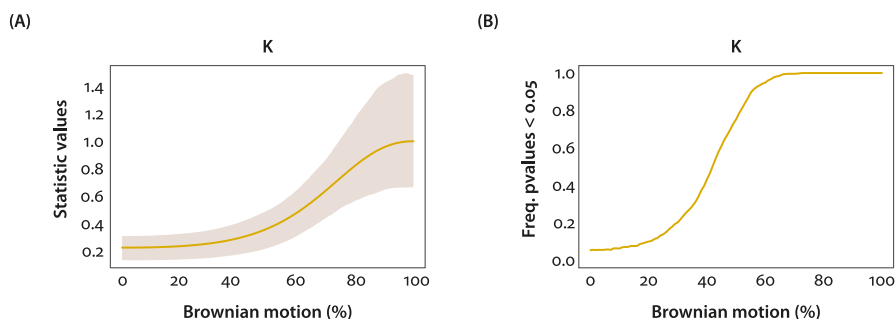


Figure 9. Behaviour of the K statistic and randomization procedure with the Pama-Nyungan reference phylogeny. Artificial characters are simulated evolving along the phylogeny with varying levels of non-Brownian noise. When a pure Brownian motion process operates, K averages around 1 as expected. Where there is no Brownian process at all (and therefore no phylogenetic signal) K is elevated to around 0.2 – likely an artefact of this particular tree size and shape.

The results are plotted in Figure 9. The K statistic shows a considerable degree of variability but, in the absence of substantial random noise, centres slightly below $K=1$ which suggests the statistic is behaving as expected (if not slightly conservatively) when phylogenetic signal is present. For characters whose simulated evolution is near-random, the baseline of K seems to be elevated a little by our particular reference tree, with non-phylogenetic simulated characters ranging from the expected $K=0$ to around $K=0.3$. This should be kept in mind when interpreting K scores across our results. As for the randomization procedure, Figure 9(B) shows the percentage of simulated traits that were identified as having a significant degree of phylogenetic signal ($p < 0.05$) at a given level of Brownian motion mixed with random noise. Above around 65% Brownian motion, there are no Type II errors. The ability to detect significant phylogenetic signal drops as the level of random noise increases beyond 35%, though overall the test's sensitivity

seems acceptable. At the opposite extreme, where characters are simulated completely at random (and, therefore, there is no phylogenetic signal to detect) the randomization procedure falsely detects phylogenetic signal 5.4% of the time, very close to the expected false discovery rate of 5% (given the conventional threshold for statistical significance of $\alpha=0.05$). On the basis of these simulations, we are satisfied that randomization procedure is sufficiently robust, given the particular size and shape of the Pama-Nyungan reference phylogeny.

As a final check, we consider whether the K statistic might be affected by the quantity of missing or ‘not applicable’ values for a given character. We inspect this visually by plotting, for all biphone characters, the relation between a biphone’s K score and the number of language varieties with non-missing data points on which K was calculated (Figure 10). When K is calculated on fewer than around 40 non-missing values, the statistic shows a wider degree of variability. In addition, phylogenetic signal is deemed statistically significant for fewer characters in this range, suggesting that the quantity of missing values is affecting the statistical power of the test. However, all K scores cluster centrally around the mean regardless of the number of languages with non-missing values, suggesting that the mean K we observe for the dataset overall is not significantly affected by missing data.

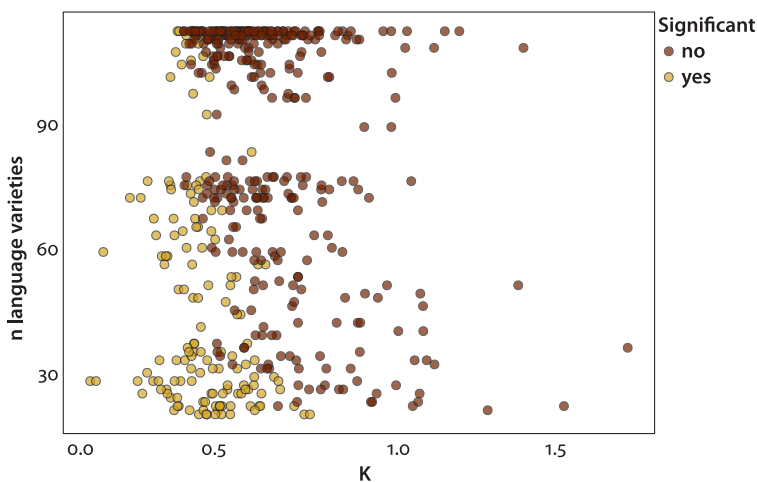


Figure 10. Estimates of K for forward and backward transition frequency characters plotted against the number of doculects with non-missing values for each character. Although some statistical power is lost and variability increases among characters with more missing values, K scores cluster evenly around mean K (0.52).

5.2 Forward transitions versus backward transitions

We find no significant difference in the means of K for forward transition characters (mean $K=0.544$) and backward transition characters (mean $K=0.537$) ($t=0.33$, $df=487.97$, $p=0.74$, 95% CI $[-0.03, 0.04]$).

The distributions of K scores for forward and backward transitions are plotted in Figure 11, showing a high degree of overlap between the two.

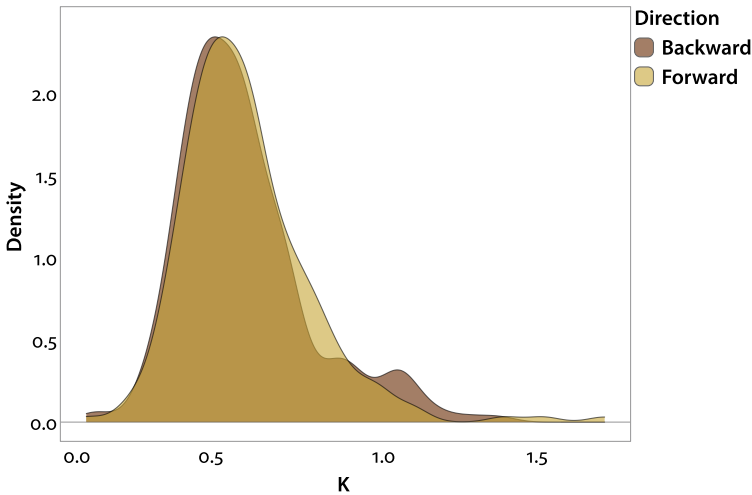


Figure 11. Distribution of K scores for forward transition frequencies versus backward transition frequencies. We find no significant difference between these character types.

5.3 Normalization of character values

Visual inspection of the density plots for each character shows there is a tendency for character data to be negatively skewed (the weight of the distribution is left-of-centre), although this is not universally the case. To test whether the particular, heavy-tailed nature of the data has an effect on tests for phylogenetic signal, we apply Tukey's Ladder of Powers transformation to each character in the dataset and re-run both the K test and randomization procedure. This is a power transformation, which makes the data fit a normal distribution as closely as possible. It does this by finding the power transformation value, λ , that maximizes the W statistic of the Shapiro-Wilk test for normality for each character individually. For our purposes, this transformation is effectively a change in the evolutionary model: A Brownian motion process is still assumed – a character value may wander up or down with equal probability – but, in this model, character values shift up or down along a transformed scale.

Mean K for normalized character data is 0.55 (SD 0.2). Of 490 characters, 405 or 83% (206 forward transitions, 199 backward transitions) contain phylogenetic signal significantly above the random expectation. There is no statistically significant difference between mean K for untransformed data (0.54) versus mean K for normalized data ($t=0.49$, $df=972.26$, $p=0.622$, 95% CI $[-0.02, 0.03]$) – see Figure 12.

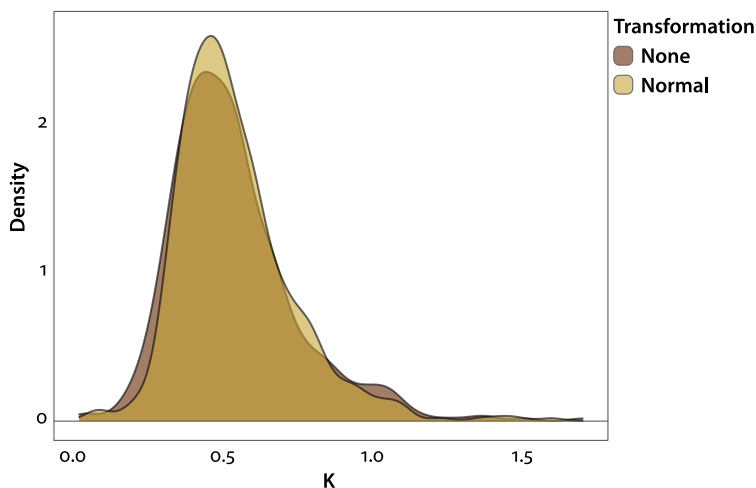


Figure 12. Distributions of K for untransformed character values and their normalized counterparts. We find no significant difference between these distributions.

6. Phylogenetic signal in natural-class-based characters

One limitation of analysing phylogenetic signal in biphone characters is the assumption that every biphone character is a statistically independent observation. In historical linguistic processes, however, phonological segments rarely behave independently. Rather, sound changes are applied to whole sound classes, thereby affecting any one of various cross-cutting sets of phonological segments (and, therefore, biphone characters we have used in this study).

To account for this non-independence and more faithfully model what we know about how phonotactic systems operate in a language, we extract forward and backward transition probabilities for sequences of phonological features. For the purposes of this experiment, word boundaries are counted as a class and vowels are reduced to a single ‘vowel’ class. Three sets of characters are extracted: forward and backward transition probabilities between natural classes based on place of articulation (segments belonging to the following classes: word boundary,

labial, dental, alveolar, retroflex, palatal, velar, glottal, vowel); forward and backward transition probabilities between natural classes based on major places of articulation, where coronal contrasts have been collapsed (word boundary, labial, apical, laminal, velar, vowel); and natural classes based on manner of articulation (word boundary, obstruent, nasal, vibrant, lateral, glide, rhotic glide, vowel). The choice of natural classes is based on well-established principles of organization among segments in Australian languages (Dixon 1980; Hamilton 1996; Baker 2014; Round 2021a; Round 2021b).

Table 1. Summary of *K* analysis for forward and backward transition frequencies between different natural classes. The two rightmost columns indicate the total number of characters analysed and the percentage of those characters with a significant degree of phylogenetic signal according to the randomization procedure.

Classes	Mean <i>K</i>	n characters	significant (%)
Place	0.61	126	94
Major place	0.63	96	75
Manner	0.59	88	66

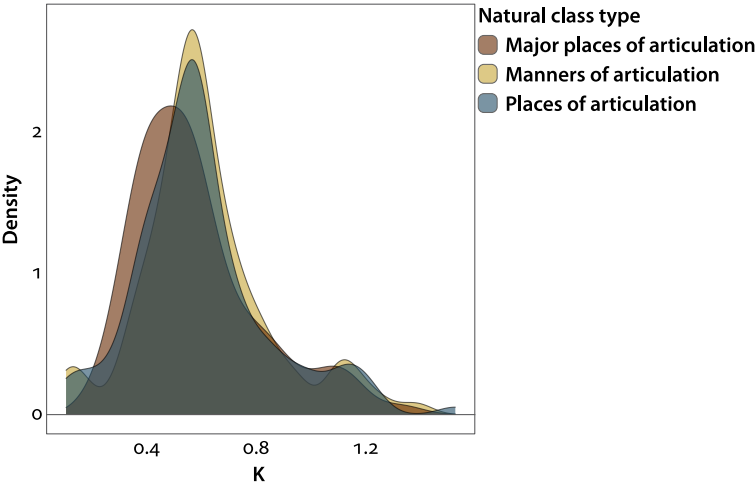


Figure 13. Comparison of *K* scores for transitions between different kinds of natural classes. The differences between all three distributions are not statistically significant.

Table 1 presents mean *K* and the proportion of significant characters for each of these three natural-class-based datasets. All show highly similar distributions (Figure 13). There is no statistically significant difference in the means of *K* for the three feature types, according to a one-way ANOVA ($F(2,307)=0.49$,

$p=0.616$). An Anderson-Darling k -sample test, which tests the hypothesis that k independent samples come from a common, unspecified distribution (i.e., no prior assumption about normality) also finds no significant difference in the distributions of K scores for the three natural-class-based datasets ($AD=2.19$, $T.AD=0.18$, $p=0.329$).

6.1 Natural-class-based characters versus biphones

We compare the degree of phylogenetic signal in the biphone data tested in §5 to the natural-class-based data tested here (see Figure 14). Mean K for all natural-class-based characters is 0.61, which is significantly higher than the mean K for biphone data, 0.54 ($t=4.38$, $df=622.47$, $p<0.001$, 95% CI [0.04, 0.1]). The Kolmogorov-Smirnov test, which is more sensitive to the overall shape of the distribution than a t -test comparison of means, also finds a significant distinction between K for biphone characters and K for natural-class-based characters ($D=0.21$, $p<0.001$).

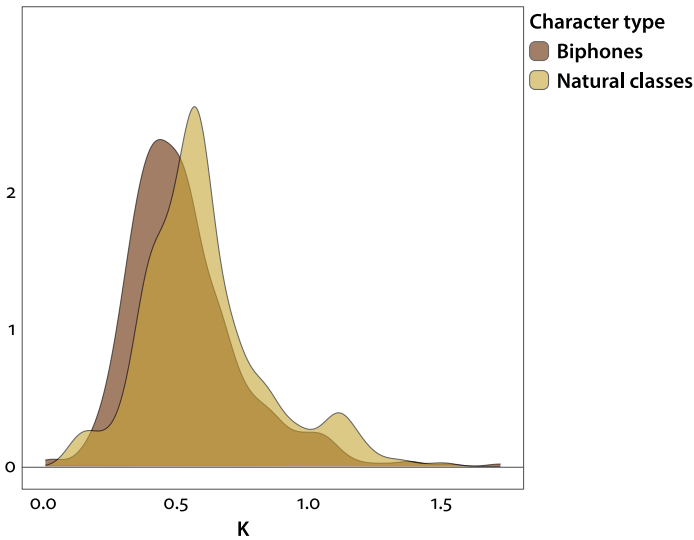


Figure 14. Distributions of K scores for biphone characters, coding the relative frequencies of transitions between phonological segments, and natural-class-based characters, coding the relative frequencies of transitions between natural classes of segments. Phylogenetic signal is higher overall in the natural-class-based dataset than the biphone-based dataset.

7. Discussion

Computational phylogenetic methods are increasingly commonplace in historical linguistics. However, there has been relatively less consideration of the range of data types that might profitably be used with computational phylogenetic methods, beyond traditional, manually-assembled sets of lexical cognate data. In this study, we have considered the potential utility of quantitative phonotactic data for historical linguistics, for the reasons that quantitative phonotactic data is (i) readily extractable from basic wordlists, and (ii) may show certain kinds of historical conservatism, where the historical signal in more traditional lexical data would be affected by borrowing and lexical innovation.

We extracted frequencies of transitions between phonological segments in scrubbed and comparably segmented wordlists representing 112 Pama-Nyungan language varieties. As points of comparison, we extracted two additional datasets: Firstly, a binarized version of the dataset, which simply records whether or not particular two-segment sequences, a.k.a. biphones, are present in a language's wordlist. This is to emulate, in a simple sense, the kind of information which is often recorded in the phonology section of published descriptive grammars. We also extracted frequencies for transitions between natural classes of sounds. This is to account (at least, to some partial degree) for the fact that phonological segments tend not to evolve independently but pattern into natural classes, thereby limiting the independence of biphone-based variables.

To test whether historical information is preserved in our phonotactic datasets, we tested for phylogenetic signal, that is, the degree to which variance in the data reflects the evolutionary history of the 112 language varieties. We took an independent phylogenetic tree, inferred by the second author using lexical cognate data, and assumed a simple Brownian motion model of evolution. Our first key finding is that a significant degree of phylogenetic signal is detected in all three datasets – binary, segment-based and sound-class-based. Finding phylogenetic signal in the binary dataset is somewhat surprising, given previous descriptions of homogeneity in the phonotactics of Australian languages. Our second key finding is that phylogenetic signal is significantly stronger in the higher-definition, frequency-based datasets than it is in the binary dataset. In turn, phylogenetic signal in the sound-class-based dataset is significantly stronger than the segment-based datasets. We took a closer look at certain comparisons within datasets, namely, whether there is a difference between forward and backward transitions, different kinds of sound classes, or between original (heavy-tailed) data and transformed data (to more closely fit a normal distribution). In all three cases, no significant differences were found.

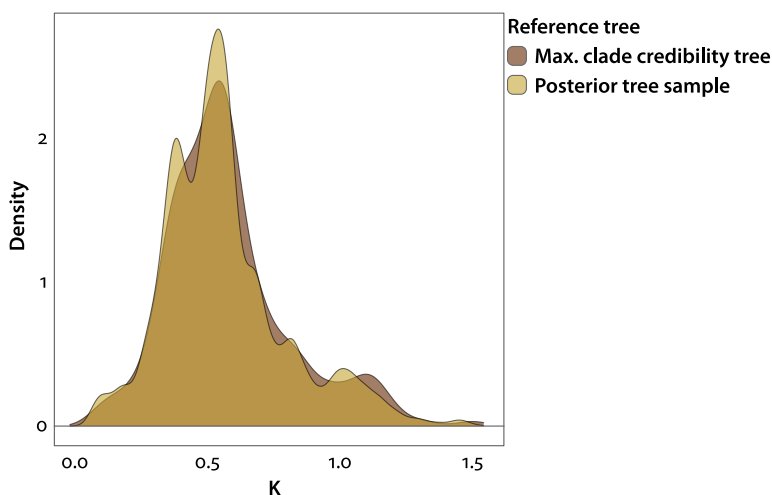


Figure 15. Comparison of K statistics using a 100-tree posterior sample versus the maximum clade credibility tree alone.

7.1 Overall robustness

One important assumption of our method is that the tree being used as a reference phylogeny is an accurate depiction of the phylogeny for the languages in question. It is impossible to verify if this assumption is being satisfied with complete certainty, since it is impossible to observe the past and record a phylogeny directly. Instead, we must rely on the best available data and methods to infer a best estimate of the true historical phylogeny. As described in §3.3, the Pama-Nyungan phylogeny we use in this study was inferred by the second author using lexical cognate data and a Bayesian computational method that produces a large posterior sample of trees. The posterior sample contains a large number of trees, all of which have a relatively high plausibility, and relatively equal plausibility, of being correct. However, we require a single tree. We select it using the *maximum clade credibility* method, which runs as follows. All trees contain hierarchies of clades. For any two trees, we can compare how many clades they share. In our posterior sample, we can ask which one tree shares the greatest number of clades with all others (we can also ask which trees rank second, third, fourth, etc. according to the same metric). Thus in §6 we used the maximum clade credibility tree drawn from our posterior sample of trees. This is a commonly used approach in the phylogenetic sciences for summarising many trees, using a single tree that is maximally representative of the whole sample. However, naturally there is no guarantee that the maximum clade credibility tree is the true tree. It may be the case that a better representation of the true Pama-Nyungan phylogeny (assuming

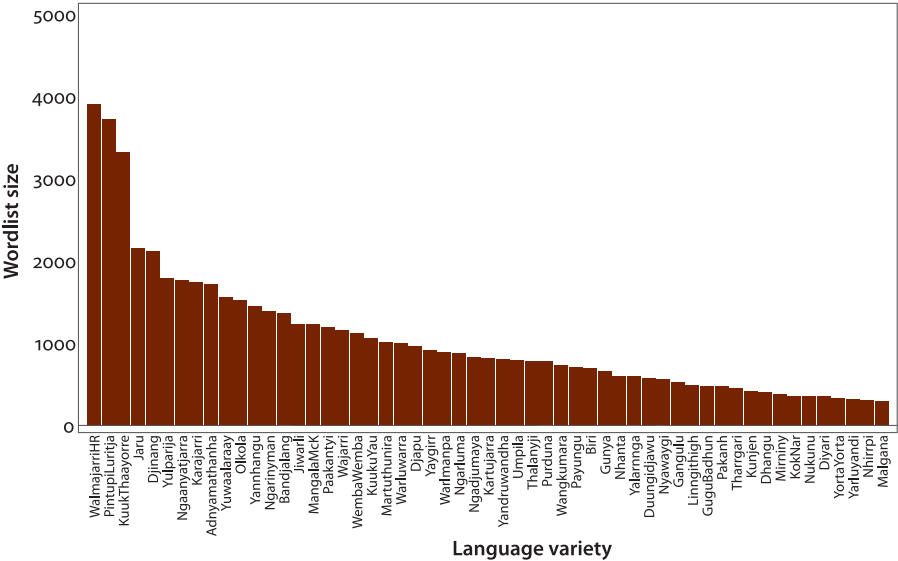
it is best represented by a single tree) exists among the many other trees the posterior sample. This kind of uncertainty is known as phylogenetic uncertainty.

To evaluate the robustness of our results against phylogenetic uncertainty, we repeat the K test for phylogenetic signal using a set of alternative trees. Namely, we test a subset of sound-class-based characters against each of 100 trees that we select from the posterior sample. The subset of characters includes forward and backward transitions between place features, and forward and backward transitions between manner features. The trees selected are the 100 top-ranked trees in the posterior sample according to the maximum clade credibility metric. In this way, we incorporate a degree of uncertainty in the topology and branch lengths of the Pama-Nyungan phylogeny, while restricting our attention to a subset of the most representative alternatives in the posterior sample. Note that overall the Pama-Nyungan tree is quite well resolved, so these trees differ from one another only minimally.

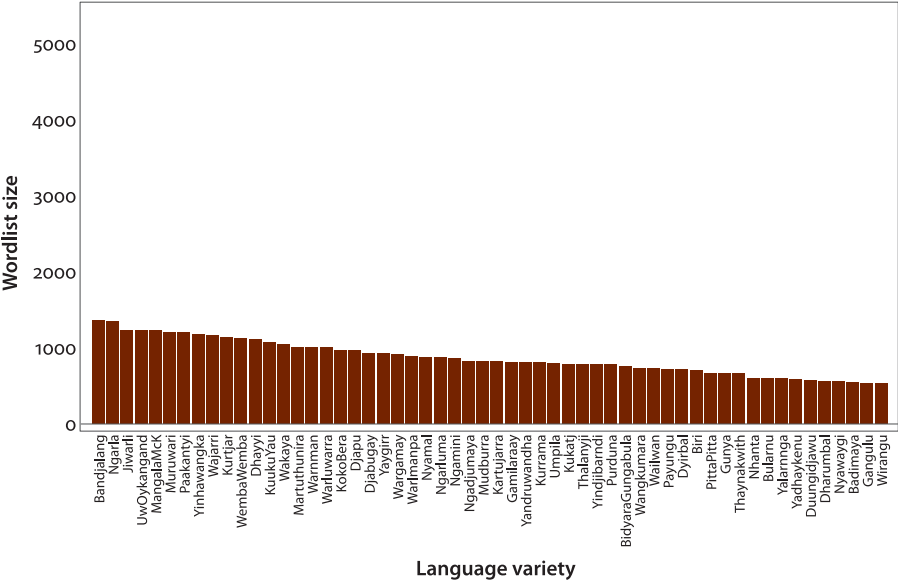
214 sound class characters in total are tested, giving 21,400 K statistics in total (each character multiplied by 100 trees). The mean of these K scores is 0.59, which compares to a mean K of 0.6 for the same characters applied only to the maximum clade credibility tree as in §6. This difference is not statistically significant ($t = -0.92$, $df = 217.21$, $p = 0.359$, 95% CI $[-0.05, 0.02]$). The distributions of these K scores are illustrated in Figure 15. This result suggests our results are robust to phylogenetic uncertainty associated with our reference tree. The maximum clade credibility tree we use is not significantly inflating or deflating the K statistic.

A reviewer points out that we have only demonstrated robustness to phylogenetic uncertainty for a Pama-Nyungan phylogeny generated with a particular lexical dataset. We have not demonstrated robustness to potential further refinement of the Pama-Nyungan phylogeny that may come about with future expansion and/or refinement of the lexical dataset, and/or inclusion of data from other linguistic domains such as morphology. Future revisions of the Pama-Nyungan phylogeny will necessitate further robustness testing. The effect of such revisions may be to strengthen phylogenetic signal in phonotactic data (if limitations of the current phylogeny manifest as noise in the signal), weaken phylogenetic signal (if, for example, phonotactic data were to correlate more strongly with lexical data than morphological or other kinds of data), or be neutral in effect.

In §3.2 we mentioned that our wordlists vary significantly in size. The length of a wordlist can correlate with many other linguistic properties that it has. For instance, in our data, the number of entries in a wordlist and the mean phonemic length of those entries have a Pearson's correlation coefficient of $r = 0.71$ ($p < 0.001$). That is, longer lists tend to contain longer words (presumably since shorter lists are weighted towards more basic, shorter vocabulary items). The existence of correlations like this means that it is not possible simply to 'counterbal-



A.



B.

Figure 16. Wordlists ranked by size. (A) shows every second wordlist in the language sample. (B) shows the middle 50% of wordlists. Each subset contains the same number of wordlists but discrepancy in their size is greatly reduced in (B).

ance' the length of wordlists by sub-sampling or resampling their items so that the resampled lists all have the same length. For example, the resampled lists that derive from longer underlying lists would still contain longer words. Nevertheless, it would still be desirable to know whether our results in §4–6 are unduly influenced by the disparity in our wordlist lengths, by manipulating wordlist length in a controlled fashion. To do this, we extracted two different language sub-samples from our dataset. For the first, we ranked all wordlists by size and selected every second language, producing a sample half the original size but with the same disparity in lengths. For the second, we ranked all wordlists by size and selected the middle 50% of the ranking, again producing a sample of half the size but this time with heavily reduced disparity, as shown in Figure 16. Wordlists in the middle 50% range in length from 528 to 1361 (mean 870).

For these samples, we ran K tests on the same subset of place and manner natural class characters as for the tree uncertainty robustness test above. Our reasoning is that if wordlist disparity strongly affects the estimation of phylogenetic signal, then we should see a clear difference in the results. Mean K for the middle 50% of wordlists is 0.78, which is somewhat greater than the mean for every second wordlist, 0.69. This difference is small but statistically significant ($t=3.6$, $df=425.76$, $p=0$, 95% CI [0.04, 0.13]). This suggests that disparities in wordlist length are somewhat degrading the phylogenetic signal in our data, although there remains broad similarity between them, as pictured in Figure 17.

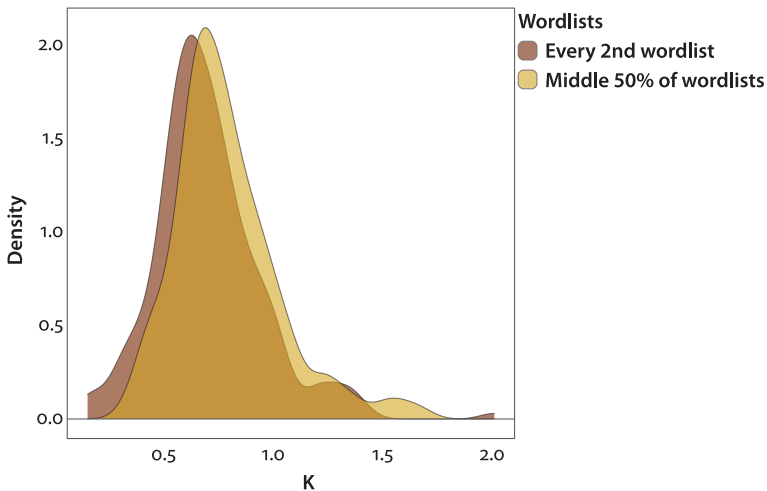


Figure 17. Comparison of K statistics using only wordlists falling within the 25th and 75th quantiles (middle 50%) for wordlist size, versus a sample of every 2nd wordlist (when ranked in order of size).

One attributing factor for this small degradation in phylogenetic signal may be measurement error among the bottom quartile of small wordlists. Throughout this study, we have assumed that all character values are accurate and do not account for measurement error. Accounting for measurement error when testing for phylogenetic signal is an area of active development in comparative biology (Zheng et al. 2009). In future studies, this degradation in phylogenetic signal could be investigated by relating variation in K statistics carefully to various linguistic properties that correlate with the length of wordlists.

7.2 Limitations

Any investigation of phylogenetic signal in essence is an investigation of cross-linguistic (dis)similarity. Accordingly, whenever our representations of linguistic facts are altered, then the phylogenetic signal detectable in them will almost certainly change to some degree. In this paper, our focus has been trained on the initial, fundamental question of whether phylogenetic signal is detectable in maximally simple phonotactic characters. However, as work like ours increases, one priority will be to examine how investigators' choices about how data is represented affects results.

Relevant for the current study, in §3.2 we described a process of normalization. The motivation for this was to attempt to minimize certain aspects of variation in phonological representation that can arise from variation in how different linguists analyse the same essential facts (Chao 1934; Hockett 1963; Hyman 2008; Drescher 2009). While normalization per se ought to improve the quality of cross-linguistic comparisons that the data enables, there is still the question of which targets one ought to normalize the data towards, and what effect that choice can have. For example, a reviewer asks whether our choice to split up complex segments might amplify phylogenetic signal if it leads to certain phylogenetically distributed complex segments counting instead as biphones. This can be answered in three ways. First, in the general case, since splitting segments changes representations, it will alter aspects of (dis)similarity in the data, and so is very likely to affect phylogenetic signal in some manner. Second, in this particular case, the reviewer is likely to be correct, due to details of our method. Any cross-linguistically rare, complex segment would likely get excluded from our dataset. This is because, although it would figure in certain biphones, we have made use only of biphones that reach a minimal level of recurrence across our language sample, and thus the biphones containing the rare segment quite likely would not qualify. However, if we split this complex segment into two segments, thus into a biphone, the resulting biphone may well have sufficient cross-linguistic recurrence to qualify for inclusion in our dataset, and subsequently may

contribute to raising phylogenetic signal. A final observation on this point is that such questions, about how choices in data representation interact with the results from corresponding quantitative analysis, are made tractable by our method of data preparation. Unlike many state-of-the-art cross-linguistic datasets, in which values for each language are hand-coded and thus incapable of being ‘recalculated’ under altered assumptions, our phonotactic characters are generated algorithmically from an underlying, very rich dataset. With a change to algorithmic parameters, we can systematically split segments or glue them together, neutralize them or keep them distinct, and document what we have done and how. As mentioned, in this paper our focus is on the simple existence of phylogenetic signal. Our methods, though, naturally extend to enable comprehensive checking of such interactions between data choices and results. Ultimately, as a discipline, we would like this to be true for all typological research, not just phylogenetics (Round 2017a). An advantage of our general approach is that it opens the doors to this rigorous mode of inquiry.

A further limitation of this study relates to the assumption that the data being tested for phylogenetic signal are independent of the data used to infer the reference phylogeny. In this study, the wordlists from which we extracted phonotactic characters contain, as a small subset, the basic vocabulary items from which lexical cognate characters were inferred and subsequently used to build the reference phylogeny. It is unclear exactly to what degree this inclusion of basic vocabulary compromises the independence of our reference tree and phonotactic data. A reviewer points out that cognate data and phonotactic data are still somewhat independent, even when extracted from identical wordlists, since phonotactic attributes are not directly called on to make cognacy judgements. Nevertheless, sound change affects both phonotactics and cognate identification, so some degree of non-independence is to be expected. To ascertain whether this effect is significant, future studies could parameterize the inclusion/exclusion of basic vocabulary from the phonotactic data.

One reviewer raises the correlation between phylogeny and geography. A noted limitation of phylogenetic comparative methods is the inability to account for geography as a possible confound (Sookias et al. 2018) and this limitation applies to this study. Although we leave it as a priority for future work, the task of disentangling phylogeny and geography is not intractable. For example, Freckleton & Jetz (2009) present a method for quantifying the relative degree of spatial versus phylogenetic effects in comparative characters.

One final point to note is that recent research suggests that phylogenetic signal can be inflated when character values evolve according to a Levy process, where a character value can wander as per a Brownian motion process, but with the addition of discontinuous paths (i.e., sudden jumps in the character’s value)

(Uyeda et al. 2018). This is a realistic concern in the linguistic context, where segment frequencies are subject to sudden shifts caused by phonological mergers and splits. The possibility of Levy-like evolutionary processes is a matter of concern also in comparative biology and methods to investigate it are subject to active development in that field (Uyeda et al. 2018).

8. Conclusion

Historical and synchronic comparative linguistics are increasingly making use of phylogenetic methods for the same reasons that led biologists to switch to them several decades ago. Our central contention has been that phylogenetic methods not only give us new ways of studying existing comparative data sets, but open up the possibility to derive insights from new kinds of data. Here we demonstrate the potential for phylogenetically investigating phonotactic data, by showing that it indeed contains the kind of phylogenetic signal which is the prerequisite for a whole spectrum of phylogenetic analyses.

We find significant phylogenetic signal for several hundred phonotactic characters extracted semi-automatically from 112 Pama-Nyungan wordlists, demonstrating that historical information is detectable in phonotactic data, even at the relatively simple level of biphones and despite ostensibly high phonological uniformity. Contrary to the prevailing view in literature on Australian languages, and contrary to the findings of an earlier pilot study on a much smaller language sample, we find that binary characters marking the presence or absence of biphones in a doculect contain enough phylogenetically-patterned variation to detect phylogenetic signal. However, we find that statistical power is relatively low when operating with coarse-grained binary data and quantification of the degree of phylogenetic signal is affected by a large number of low-variation characters, where all but one or a few doculects share the same value. We find stronger phylogenetic signal in biphone characters of forward and backward transition frequencies. This reaffirms the results of earlier work, for the first time on a sample of languages spanning an entire large family and the vast majority of a continent. It also reaffirms earlier findings that Australian phonologies show a greater level of variation than traditionally has been appreciated, once matters of frequency are taken into account. We find a significantly greater level of phylogenetic signal again in characters based on the frequencies of forward and backward transitions between natural sound classes. The sound-class-based approach reduces the quantity of characters available to test, but limits sparsity in the dataset and accounts somewhat better for the role of sound classes in the evolutionary processes that affect phonotactic patterns in human language. Interestingly,

although there exists considerable variation in the level of phylogenetic signal found in individual characters, we find no observable pattern to this variation in segment-based biphone characters nor between the mean levels of phylogenetic signal observed for different kinds of sound classes (e.g., characters concerning place versus manner of articulation).

This work has implications for comparative linguistics, both typological and historical. Firstly, we recommend the use of phylogenetic comparative methods in typological work where the phylogenetic independence of a language sample (or lack thereof) is paramount. In the immediate term, this should be the case for any typological work concerning phonotactics, even in parts of the world such as Australia where phonotactics traditionally have been assumed to be relatively independent of phylogeny. Beyond phonotactics, however, explicit measurements of phylogenetic signal can be made for any set of cross-linguistic data and this can be built into statistical analysis, even in the presence of gaps and uncertainty in phylogenetic knowledge. In two decades of quantitative development in historical linguistics, there has still been relatively limited consideration of the kinds of characters used for inferring linguistic histories. The phonotactic characters presented here can be extracted relatively simply and in large quantities from wordlists, even where a full descriptive grammar is not available. Here, we test only the degree to which patterns of variation in our data match our independent, pre-existing knowledge of the phylogenetic history of the Pama-Nyungan family. However, the results suggest that phonotactic data of this kind could be used where the phylogeny is less certain, either by incorporating phonotactic data into phylogenetic inference directly or by constraining parts of the tree where lexical data on its own returns some doubt.

Funding

Research funded by Australian Government Research Training Program (RTP) (Scholarship) to Jayden L. Macklin-Cordes, by Australian Research Council (DE150101024) to Erich R. Round, by Australian Research Council (ARC) Centre of Excellence for the Dynamics of Human Language (TIG322015) to Erich R. Round and by National Science Foundation (NSF BCS-1423711) to Claire Bowern.

Acknowledgements

We wish to thank David Nash and two anonymous reviewers for their generous and constructive feedback throughout preparation of this manuscript.

Author contribution statement

JM-C and ER jointly conceived the study and developed the experimental design. CB compiled the CHIRILA lexical database, produced the reference phylogeny and provided valuable guidance and theoretical input throughout. ER contributed AusPhon lexical data (extending on CHIRILA), sections of text and code and further theoretical guidance. JM-C wrote the code (with contributions from ER), conducted the analysis and created the figures. JM-C prepared the manuscript and supplementary materials with feedback and revisions from ER and CB. All authors approved the final manuscript.

Since the second author, Claire Bowern, is *Diachronica's* Executive Editor, the review process for this paper was handled entirely by former editor, Joseph Salmons. CB was not involved in any aspect of the review process.

Supplementary materials

Datasets are available on Zenodo at <http://doi.org/10.5281/zenodo.3988775>. This repository also includes a full table of results and the R scripts used to perform the analysis and produce figures for the paper. Usage instructions and a full description of these materials is available with this article at <https://doi.org/10.1075/dia.20004.mac.additional>.

References

- Albright, Adam & Bruce Hayes. 2003. Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition* 90(2). 119–161.
[https://doi.org/10.1016/S0010-0277\(03\)00146-X](https://doi.org/10.1016/S0010-0277(03)00146-X)
- Alpher, Barry J. 2004. Pama-Nyungan: Phonological reconstruction and status as a phylogenetic group. In Claire Bowern & Harold Koch (eds.), *Australian languages: Classification and the comparative method* (Current Issues in Linguistic Theory 249), 93–126. Amsterdam: John Benjamins Publishing Company.
<https://doi.org/10.1075/cilt.249.09alp>
- Austin, Peter. 1981. Proto-Kanyara and Proto-Mantharta historical phonology. *Lingua* 54(4). 295–333. [https://doi.org/10.1016/0024-3841\(81\)90009-7](https://doi.org/10.1016/0024-3841(81)90009-7)
- Baker, Brett. 2014. Word structure in Australian languages. In Harold Koch & Rachel Nordlinger (eds.), *The languages and linguistics of Australia: A comprehensive guide*, 139–214. Berlin: De Gruyter Mouton. <https://doi.org/10.1515/9783110279771.139>
- 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>
- 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>
- Blasi, Damián E., Steven Moran, Scott R. Moisik, Paul Widmer, Dan Dediu & Balthasar Bickel. 2019. Human sound systems are shaped by post-Neolithic changes in bite configuration. *Science* 363(6432). <https://doi.org/10.1126/science.aav3218>

- Blomberg, Simon P. & Theodore Garland Jr. 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>
- Blomberg, Simon P., Theodore Garland Jr. & Anthony R. Ives. 2003. Testing for phylogenetic signal in comparative data: Behavioral traits are more labile. *Evolution* 57(4). 717–745. <https://doi.org/10.1111/j.0014-3820.2003.tb00285.x>
- 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/CitetononCRdoi: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. 2017. Standard Average Australian? In *Association for Linguistic Typology (ALT)*. Australian National University, Canberra, Australia. <https://doi.org/CitetononCRdoi:10.5281/zenodo.1104222>
- Bowern, Claire. 2018a. Computational phylogenetics. *Annual Review of Linguistics* 4(1). 281–296. <https://doi.org/10.1146/annurev-linguistics-011516-034142>
- Bowern, Claire. 2018b. Pama-Nyungan cognate judgements: 285 languages. <https://doi.org/CitetononCRdoi:10.5281/zenodo.1318310>
- Bowern, Claire & Quentin D. Atkinson. 2012. Computational phylogenetics and the internal structure of Pama-Nyungan. *Language* 88(4). 817–845. <https://doi.org/10.1353/lan.2012.0081>
- 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>
- Bowern, Claire & Harold Koch (eds.). 2004. *Australian languages: Classification and the comparative method* (Current Issues in Linguistic Theory 249). Amsterdam: John Benjamins. <https://doi.org/10.1075/cilt.249>
- Busby, Peter A. 1982. The distribution of phonemes in Australian Aboriginal languages. *Pacific Linguistics. Series A.* (60). 73–139.
- 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>
- Campbell, Lyle. 2004. *Historical linguistics: An introduction*. Cambridge, MA: MIT Press.
- Capell, Arthur. 1956. *A new approach to Australian linguistics*. Sydney: University of Sydney.
- Chang, Will, Chundra Cathcart, David Hall & Andrew Garrett. 2015. Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. *Language* 91(1). 194–244. <https://doi.org/10.1353/lan.2015.0005>
- Chao, Yuen-Ren. 1934. The non-uniqueness of phonemic solutions of phonetic systems. *Bulletin of the Institute of History and Philology, Academia Sinica* 4(4). 363–397.
- Coleman, John & Janet Pierrehumbert. 1997. Stochastic phonological grammars and acceptability. In *Computational phonology: ACL special interest group in computational phonology*, 49–56. Somerset, NJ: Association for Computational Linguistics. <http://arxiv.org/abs/cmp-1g/9707017> (8 March, 2018).

- Crawford, Clifford J. 2009. *Adaptation and transmission in Japanese loanword phonology*. Cornell University thesis. <http://core.ac.uk/download/pdf/4912071.pdf>
- Cysouw, Michael & Jeff Good. 2007. Towards a comprehensive languoid catalog. In *Language catalogue meeting*. Leipsig, Germany. http://cysouw.de/home/presentations_files/cysouwCATALOGUE_slides.pdf (11 December, 2019).
- Delsuc, Frédéric, Henner Brinkmann & Hervé Philippe. 2005. Phylogenomics and the reconstruction of the tree of life. *Nature Reviews Genetics* 6(5). 361–375. <https://doi.org/10.1038/nrg1603>
- Dixon, R. M. W. 1980. *The languages of Australia* (Cambridge Language Surveys). Cambridge: Cambridge University Press.
- Dockum, Rikker. 2018. Phylogeny in phonology: How Tai sound systems encode their past. In *Annual meeting on phonology (AMP)*. New York: Linguistic Society of America. <https://doi.org/10.3765/amp.v51o.4238>
- Dockum, Rikker & Claire Bowern. 2019. Swadesh lists are not long enough: Drawing phonological generalizations from limited data. *Language Documentation and Description* 16. 35–54. <http://www.elpublishing.org/PID/168>
- Dresher, B. Elan. 2009. *The contrastive hierarchy in phonology*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511642005>
- Dresher, B. Elan & Aditi Lahiri. 2005. Main stress left in early Middle English. In Michael Fortescue, Jens E. Mogensen & Lene Schøsler (eds.), *International conference on historical linguistics (ICHL)*, 76–85. Amsterdam: John Benjamins. <https://doi.org/10.1075/cilt.257>
- 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>
- Dunn, Michael, Angela Terrill, Ger Reesink, Robert A. Foley & Stephen C. Levinson. 2005. Structural phylogenetics and the reconstruction of ancient language history. *Science* 309(5743). 2072–2075. <https://doi.org/10.1126/science.1114615>
- Durie, Mark & Malcolm Ross. 1996. *The comparative method reviewed: Regularity and irregularity in language change*. New York: Oxford University Press.
- Eddington, David. 2004. *Spanish phonology and morphology: Experimental and quantitative perspectives*. Amsterdam: John Benjamins. <https://doi.org/10.1075/sfsl.53>
- Ernestus, Mirjam T. C. & R. Harald Baayen. 2003. Predicting the unpredictable: Interpreting neutralized segments in Dutch. *Language* 79(1). 5–38. <https://doi.org/10.1353/lan.2003.0076>
- Felsenstein, Joseph. 1985. Phylogenies and the comparative method. *The American Naturalist* 125(1). 1–15. <https://doi.org/10.1086/284325>
- 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>
- Freckleton, Robert P. & Walter Jetz. 2009. Space versus phylogeny: Disentangling phylogenetic and spatial signals in comparative data. *Proceedings of the Royal Society B: Biological Sciences* 276(1654). 21–30. <https://doi.org/10.1098/rspb.2008.0905>
<https://doi.org/10.1098/rspb.2008.0905>

- 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, Jr. & Ramón Díaz-Uriarte. 1999. Polytomies and phylogenetically independent contrasts: Examination of the bounded degrees of freedom approach. *Systematic Biology* 48(3). 547–558. <https://doi.org/10.1080/106351599260139>
- Gasser, Emily & Claire Bowern. 2014. Revisiting phonotactic generalizations in Australian languages. In *Annual meeting on phonology (AMP)*. University of Massachusetts, Amherst: Linguistic Society of America. <https://doi.org/10.3765/amp.v1i1.17>
- Good, Jeff & Michael Cysouw. 2013. Languoid, doculect, and glossonym: Formalizing the notion ‘language’. *Language Documentation and Conservation* 7. 331–359. <http://hdl.handle.net/10125/4606>
- Gordon, Matthew K. 2016. *Phonological typology* (Oxford Surveys in Phonology and Phonetics 1). Oxford: Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199669004.001.0001>
- Grafen, Alan. 1989. The phylogenetic regression. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 326(1233). 119–157. <https://doi.org/10.1098/rstb.1989.0106>
- Greenhill, Simon J., Thomas E. Currie & Russell D. Gray. 2009. Does horizontal transmission invalidate cultural phylogenies? *Proceedings of the Royal Society of London B: Biological Sciences* 276(1665). 2299–2306. <https://doi.org/10.1098/rspb.2008.1944>
- Greenhill, Simon J., Chieh-Hsi Wu, Xia Hua, Michael Dunn, Stephen C. Levinson & Russell D. Gray. 2017. Evolutionary dynamics of language systems. *Proceedings of the National Academy of Sciences* 114(42). E8822–E8829. <https://doi.org/10.1073/pnas.1700388114>
- Hamilton, Philip J. 1996. *Phonetic constraints and markedness in the phonotactics of Australian Aboriginal languages*. Toronto: University of Toronto thesis.
- Hayes, Bruce & Zsuzsa C. Londe. 2006. Stochastic phonological knowledge: The case of Hungarian vowel harmony. *Phonology* 23(1). 59–104. <https://doi.org/10.1017/S0952675706000765>
- Hockett, Charles F. 1963. The problem of universals in language. In Joseph Greenberg (ed.), *Universals of language*, 1–29. Cambridge, MA: MIT Press.
- Hutchinson, Matthew C., Marília P. Gaiarsa & Daniel B. Stouffer. 2018. Contemporary ecological interactions improve models of past trait evolution. *Systematic Biology* 67(5). 861–872. <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/elaugue/sal/article/view/927.html>
- Hyman, Larry M. 2008. Universals in phonology. *The Linguistic Review* 25(1). 83–137. <https://doi.org/10.1515/TLIR.2008.003>
- Jäger, Gerhard. 2019. Computational historical linguistics. *Theoretical Linguistics* 45(3). 151–182. <https://doi.org/10.1515/tl-2019-0011>
- Kang, Yoonjung. 2011. Loanword phonology. In Marc Van Oostendorp, Colin J. Ewen, Elizabeth Hume & Keren Rice (eds.), *The Blackwell companion to phonology*, vol. IV, 2258–2282. Oxford: Wiley-Blackwell. <https://doi.org/10.1002/9781444335262.wbctp0095>

- Kembel, Steven W., Peter D. Cowan, Matthew R. Helmus, William K. Cornwell, Helene Morlon, David D. Ackerly, Simon P. Blomberg & Campbell O. Webb. 2010. Picante: R tools for integrating phylogenies and ecology. *Bioinformatics* 26(11). 1463–1464. <https://doi.org/10.1093/bioinformatics/btq166>
- Kiparsky, Paul. 2018. Formal and empirical issues in phonological typology. In Larry M. Hyman & Frans Plank (eds.), *Phonological typology*, 54–106. Berlin: De Gruyter Mouton. <https://doi.org/10.1515/9783110451931-003>
- Koch, Harold. 2014. Historical relations among the Australian languages: Genetic classification and contact-based diffusion. In Harold Koch & Rachel Nordlinger (eds.), *The languages and linguistics of Australia: A comprehensive guide*, 23–90. Berlin: De Gruyter Mouton. <https://doi.org/10.1515/9783110279771.23>
- Kolipakam, Vishnupriya, Fiona M. Jordan, Michael Dunn, Simon J. Greenhill, Remco R. Bouckaert, Russell D. Gray & Annemarie Verkerk. 2018. A Bayesian phylogenetic study of the Dravidian language family. *Royal Society Open Science* 5(3). 171504. <https://doi.org/10.1098/rsos.171504>
- Lass, Roger. 1984. Vowel system universals and typology: Prologue to theory. *Phonology Yearbook* 1. 75–111. <https://doi.org/10.1017/S0952675700000300>
- Leff, Jonathan W., Richard D. Bardgett, Anna Wilkinson, Benjamin G. Jackson, William J. Pritchard, Jonathan R. Long, Simon Oakley, et al. 2018. Predicting the structure of soil communities from plant community taxonomy, phylogeny, and traits. *The ISME Journal* 12. 1794–1805. <https://doi.org/10.1038/s41396-018-0089-x>
- List, Johann-Mattis, Simon J. Greenhill & Russell D. Gray. 2017. The potential of automatic word comparison for historical linguistics. *PLoS ONE* 12(1). e0170046. <https://doi.org/10.1371/journal.pone.0170046>
- List, Johann-Mattis, Mary Walworth, Simon J. Greenhill, Tiago Tresoldi & Robert Forkel. 2018. Sequence comparison in computational historical linguistics. *Journal of Language Evolution* 3(2). 130–144. <https://doi.org/10.1093/jole/lyz006>
- Losos, Jonathan B. 2008. Phylogenetic niche conservatism, phylogenetic signal and the relationship between phylogenetic relatedness and ecological similarity among species. *Ecology Letters* 11(10). 995–1003. <https://doi.org/10.1111/j.1461-0248.2008.01229.x>
- Macklin-Cordes, Jayden L. & Erich R. Round. 2015. High-definition phonotactics reflect linguistic pasts. In Johannes Wahle, Marisa Köllner, Harald Baayen, Gerhard Jäger & Tineke Baayen-Oudshoorn (eds.), *Quantitative investigations in theoretical linguistics (QITL-6)*. Tübingen: University of Tübingen. <https://doi.org/CitetononCRdoi:10.15496/publikation-8609>
- Marin, Julie, S. Blair Hedges & Koichiro Tamura. 2018. Undersampling genomes has biased time and rate estimates throughout the tree of life. *Molecular Biology and Evolution* 35(8). 2077–2084. <https://doi.org/10.1093/molbev/msy103>
- 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>
- Meillet, Antoine. 1925. *La méthode comparative en linguistique historique*. Paris: Honoré Champion.
- Moran, Steven & Michael Cysouw. 2018. *The Unicode cookbook for linguists: Managing writing systems using orthography profiles*. Berlin: Language Science Press.

- Moran, Steven, Eitan Grossman & Annemarie Verkerk. 2020. Investigating diachronic trends in phonological inventories using BDPROTO. *Language Resources and Evaluation*. <https://doi.org/10.1007/s10579-019-09483-3>
- Moran, Steven & Annemarie Verkerk. 2018. Differential rates of change in consonant and vowel systems. In C. Cuskley, M. Flaherty, H. Little, Luke McCrohon, A. Ravignani & T. Verhoef (eds.), *The evolution of language (EVOLANGXII)*. NCU Press. <https://doi.org/10.12775/3991-1.077>. <http://evolang.org/torun/proceedings/papertemplate.html?p=98>
- Münkemüller, Tamara, Sébastien Lavergne, Bruno Bzeznik, Stéphane Dray, Thibaut Jombart, Katja Schiffrers & 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>
- Nash, David, Patrick McConvell, Arthur Capell, Ken Hale, Peter Sutton, Deborah Bird Rose & Jim Wafer. 1988. *Mudburra wordlist*. *Word list*. Canberra: Australian Institute of Aboriginal and Torres Strait Islander Studies, Australian Indigenous Languages Collection, ms. http://aiatsis.gov.au/sites/default/files/catalogue_resources/0031_access.zip
- Nichols, Johanna. 1997. Sprung from two common sources: Sahul as a linguistic area. In Patrick McConvell & Nicholas Evans (eds.), *Archaeology and linguistics: Aboriginal Australia in global perspective*. Melbourne: Oxford University Press.
- Nunn, Charles L. 2011. *The comparative approach in evolutionary anthropology and biology*. Chicago: University of Chicago Press. <https://doi.org/10.7208/chicago/9780226090009.001.0001>
- O'Grady, Geoffrey N., Charles F. Voegelin & Florence M. Voegelin. 1966. Languages of the world: Indo-Pacific fascicle six. *Anthropological Linguistics* 1–197.
- Orme, David, Rob Freckleton, Gavin Thomas, Thomas Petzoldt, Susanne Fritz, Nick Isaac & Will Pearse. 2013. *caper: Comparative analyses of phylogenetics and evolution in R*. <https://CRAN.R-project.org/package=caper>
- Rama, Taraka, Johann-Mattis List, Johannes Wahle & Gerhard Jäger. 2018. Are automatic methods for cognate detection good enough for phylogenetic reconstruction in historical linguistics? In *North American chapter of the Association for Computational Linguistics (ACL): Human language technologies, volume 2 (short papers)*, 393–400. New Orleans: Association for Computational Linguistics. <https://doi.org/10.18653/v1/N18-2063>
- R Core Team. 2017. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. <https://www.R-project.org/>
- 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>
- Rexová, Kateřina, Yvonne Bastin & Daniel Frynta. 2006. Cladistic analysis of Bantu languages: A new tree based on combined lexical and grammatical data. *Naturwissenschaften* 93(4). 189–194. <https://doi.org/10.1007/s00114-006-0088-z>
- Round, Erich R. 2017a. Matthew K. Gordon: Phonological typology [book review]. *Folia Linguistica* 51(3). 745–755. <https://doi.org/10.1515/flin-2017-0027>
- Round, Erich R. 2017b. The AusPhon-Lexicon project: 2 million normalized segments across 300 Australian languages. In *Poznań linguistic meeting*. Poznań, Poland. http://wa.amu.edu.pl/plm_old/2017/files/abstracts/PLM2017_Abtract_Round.pdf

- Round, Erich R. 2019a. Phonemic inventories of Australia [database of 392 languages]. In Steven Moran & Daniel McCloy (eds.), *PHOIBLE 2.0*. Jena, Germany: Max Planck Institute for the Science of Human History.
- Round, Erich R. 2019b. *Australian phonemic inventories contributed to PHOIBLE 2.0: Essential explanatory notes*. Zenodo. <https://doi.org/CitetononCRdoi:10.5281/zenodo.3464333>
- Round, Erich R. 2021a. Segment inventories in Australian languages. In Claire Bown (ed.), *Oxford handbook of Australian languages*. Oxford: Oxford University Press.
- Round, Erich R. 2021b. Phonotactics in Australian languages. In Claire Bown (ed.), *Oxford handbook of Australian languages*. Oxford: Oxford University Press.
- Sallan, Lauren Cole & Matt Friedman. 2012. Heads or tails: Staged diversification in vertebrate evolutionary radiations. *Proceedings of the Royal Society B: Biological Sciences* 279(1735). 2025–2032. <https://doi.org/10.1098/rspb.2011.2454>
- Schmidt, Wilhelm. 1919. *Die Gliederung australischen Sprachen: Geographische, bibliographische, linguistische Grundzüge der Erforschung der australischen Sprachen*. Vienna: Druck und Verlag der Mechitharisten-Buchdruckerei.
- Silverman, Daniel. 1992. Multiple scansions in loanword phonology: Evidence from Cantonese. *Phonology* 9(2). 289–328. <https://doi.org/10.1017/S0952675700001627>
- Sookias, Roland B., Samuel Passmore & Quentin D. Atkinson. 2018. Deep cultural ancestry and human development indicators across nation states. *Royal Society Open Science* 5(4). 171411. <https://doi.org/10.1098/rsos.171411>
- Steiner, Lydia, Michael Cysouw & Peter Stadler. 2011. A pipeline for computational historical linguistics. *Language Dynamics and Change* 1(1). 89–127. <https://doi.org/10.1163/221058211X570358>
- Uyeda, Josef C., Rosana Zenil-Ferguson, Matthew W. Pennell & Nicholas Matzke. 2018. Rethinking phylogenetic comparative methods. *Systematic Biology* 67(6). 1091–1109. <https://doi.org/10.1093/sysbio/syy031>
- Van der Hulst, Harry. 2017. Phonological typology. In Alexandra Y. Aikhenvald & R. M. W. Dixon (eds.), *The Cambridge handbook of linguistic typology*, 39–77. Cambridge: Cambridge University Press. <https://doi.org/10.1017/9781316135716.002>
- 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)*. Canberra, Australia: Australian National University.
- Villemereuil, Pierre de & Shinichi Nakagawa. 2014. General quantitative genetic methods for comparative biology. In László Z. Garamszegi (ed.), *Modern phylogenetic comparative methods and their application in evolutionary biology: Concepts and practice*, 287–303. Berlin: Springer. https://doi.org/10.1007/978-3-662-43550-2_11
- Voegelin, Florence M., Stephen A. Wurm, Geoffrey O'Grady, Tokuichiro Matsuda & Charles F. Voegelin. 1963. Obtaining an index of phonological differentiation from the construction of non-existent minimax systems. *International Journal of American Linguistics* 29(1). 4–28. <https://doi.org/10.1086/464707>
- Walker, Robert S. & Lincoln A. Ribeiro. 2011. Bayesian phylogeography of the Arawak expansion in lowland South America. *Proceedings of the Royal Society B: Biological Sciences* 278(1718). 2562–2567. <https://doi.org/10.1098/rspb.2010.2579>

- Webb, Campbell O., David D. Ackerly, Mark A. McPeck & Michael J. Donoghue. 2002. Phylogenies and community ecology. *Annual Review of Ecology and Systematics* 33(1). 475–505. <https://doi.org/10.1146/annurev.ecolsys.33.010802.150448>
- Weiss, Michael. 2014. The comparative method. In Claire Bowern & Bethwyn Evans (eds.), *The Routledge handbook of historical linguistics* (Routledge Handbooks in Linguistics), 127–145. London: Routledge. <https://doi.org/10.4324/9781315794013.ch4>
- Widmer, Manuel, Sandra Auderset, Johanna Nichols, Paul Widmer & Balthasar Bickel. 2017. NP recursion over time: Evidence from Indo-European. *Language* 93(4). 799–826. <https://doi.org/10.1353/lan.2017.0058>
- Wortley, Alexandra H., Paula J. Rudall, David J. Harris, Robert W. Scotland & Peter Linder. 2005. How much data are needed to resolve a difficult phylogeny? Case study in lamiales. *Systematic Biology* 54(5). 697–709. <https://doi.org/10.1080/10635150500221028>
- Wurm, Stephen A. 1963. Aboriginal languages: The present state of knowledge. In Helen Shiels (ed.), *Australian Aboriginal studies: A symposium of papers presented at the 1961 research conference*, 127–148. Melbourne: Oxford University Press.
- Wurm, Stephen A. 1972. *Languages of Australia and Tasmania*. The Hague: Mouton. <https://doi.org/10.1515/9783110808292>
- Zheng, Li, Anthony R. Ives, Theodore Garland Jr., 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>
- Zuraw, Kie R. 2000. *Patterned exceptions in phonology*. Los Angeles: University of California dissertation.

Résumé

Les méthodes phylogénétiques ont le potentiel en linguistique pour faire de l'inférence des arbres. Ici, nous montrons comment une approche phylogénétique peut être utilisée avec de nouveaux types de données linguistiques – la phonotactique statistique. Nous quantifions le signal phylogénétique – le degré auquel les données reflètent la phylogénie – dans 112 vocabulaires de la famille Pama-Nyungan. Nous testons trois types de données: (1) des variables binaires enregistrant la présence ou l'absence de biphones (séquences de deux segments) (2) les fréquences de les transitions entre segments, et (3) les fréquences de transitions entre les classes sonores naturelles. Nous détectons un signal phylogénétique dans tous les ensembles de données. Le signal phylogénétique est plus élevé dans les données de fréquence plus détaillé que dans les données binaires, et le plus grand dans les données qui enregistrent des classes sonores. Ces résultats montrent que c'est possible d'employer une nouvelle source de données facilement extractibles en linguistique historique.


Zusammenfassung

Phylogenetische Methoden haben ein breites Potenzial in der Linguistik, das über die Inferenz von Bäumen hinausgeht. Hier zeigen wir, wie ein phylogenetischer Ansatz mit neuen Arten von Sprachdaten verwendet werden kann – in diesem Fall mit statistischer Phonotaktik. Wir quantifizieren das phylogenetische Signal – den Grad, in dem die Daten die Phylogenie widerspiegeln – in phonotaktischen Daten aus 112 Pama-Nyungan-Vokabularen. Wir testen drei Datensätze: (1) binäre Variablen, die das Vorhandensein oder Fehlen von Biphonen (Zwei-Segment-Sequenzen) aufzeichnen. (2) Häufigkeit von Übergängen zwischen Segmenten und (3) Häufigkeit von Übergängen zwischen natürlichen Lautklassen. Wir erkennen phylogenetische Signale in allen Datensätzen. Phylogenetische Signal ist in feinkörnigeren Frequenzdaten größer als in binären Daten und in Daten auf Basis natürlicher Klassen am größten. Diese Ergebnisse zeigen die Durchführbarkeit davon, eine neue Quelle leicht extrahierbarer Daten in der historischen und vergleichenden Linguistik einzusetzen.

Address for correspondence

Jayden L. Macklin-Cordes
School of Languages and Cultures
The University of Queensland
Room 320, Gordon Greenwood Building (32)
ST LUCIA QLD 4072
Australia


jayden.macklincordes@uqconnect.edu.au

 <https://orcid.org/0000-0003-1910-3969>

Co-author information


Claire Bownen
Department of Linguistics
Yale University

claire.bownen@yale.edu

 <https://orcid.org/0000-0002-9512-4393>

Erich R. Round
School of Languages and Cultures
The University of Queensland

e.round@uq.edu.au

 <https://orcid.org/0000-0002-7533-8052>

Publication history

Date received: 17 January 2020

Date accepted: 27 August 2020

Published online: 2 February 2021