High-Definition Phonotactics Reflect Linguistic Pasts

Jayden L. Macklin-Cordes and Erich R. Round
Ancient Language Lab, School of Languages and Cultures
University of Queensland, Brisbane, Australia
j.macklincordes@uq.edu.au | e.round@uq.edu.au

***Abstract*—Typological datasets for quantitative historical-linguistic inquiry are growing in breadth, but a challenge is also to increase their depth, since advanced methods often ideally require many hundreds of traits per language. Using biphone transition probabilities from phonemicized vocabulary data, we extract several hundred high-definition phonotactic traits per language, for 17 languages in the Ngumpin-Yapa and Yolngu subgroups of the Pama-Nyungan family, Australia. We detect phylogenetic signal at a significant level ($p < 0.001$ for both subgroups), measured against a reference phylogeny inferred from basic vocabulary cognacy data. This contrasts with simpler, binary coding of biphones' occurrence, which provides insufficient detail for the detection of phylogenetic signal. Thus, we demonstrate the viability of a new method in quantitative historical linguistics, and emphasize the inferential power to be harnessed from high-definition, trait-rich datasets for comparative research.**

***Keywords*—Historical linguistics, Phonology, Phonotactics, Phylogenetic signal, Pama-Nyungan, Ngumpin-Yapa, Yolngu.**

I. Introduction

A. Richer data; more traits per language

Quantitative datasets are increasingly available which span large numbers of languages, yet sophisticated statistical methods often demand high numbers of *traits*. We investigate the potential of extracting many hundreds of phonotactic traits per language, from phonemicized vocabularies, and test these traits for phylogenetic signal. To set the bar high, we test our method on two language families of Australia. Australian languages are known for the homogeneity of their phonological systems [1]. This ought to provide a barrier to the recovery of phylogenetic signal, and thus, if our methods succeed with this data, we may be optimistic about wider applicability.

B. Phonotactic traits

All languages permit certain, but not other, sequences of their phonemes. Taking the most basic case, languages may be compared in terms of which two-segment sequences, $a+b$, they permit. For a set of phonemes in a language $\{p_1 \ldots p_n\}$ this yields an $n \times n$ matrix of binary 'biphone permissibility' traits. Such data is often provided in descriptive grammars, or can be extracted from phonemicized vocabularies. However, permissibility data is rather coarse. Higher-definition data can be obtained from facts of frequency. For example, a Markov chain (forward) transition probability of $a+b$, can be calculated as the frequency of occurrence of $a+b$ relative to all sequences $a+X$ in a vocabulary [2], [3]. This yields an $n \times n$ matrix of continuous traits.

This research has been supported by ARC grant DE150101024 to E. Round and NSF grant 1423711 to C. Bowern.

C. Trait inheritance in language change

Phonotactic data may offer particular insight into vertical inheritance, since when languages borrow lexicon or coin new lexical items, the incoming items are most often fit into existing phonotactic patterns [4], allowing those patterns persist even under conditions of borrowing and innovation.

D. Homogeneity in Australian Phonological Systems

Australian languages display a conspicuously low level of phonological diversity, even across distinct language families and in the midst of considerable variation in other linguistic categories [1], [5]–[8]. Common characteristics of Australian phoneme inventories include:

- 4–6 places of articulation: labial; velar; 2–4 coronal.
- 1 series of stops, with no voicing or length contrast.
- No contrastive fricatives.
- Nasals at every place of articulation.
- 1–4 laterals.
- A triangular system of vowel qualities.

A 'typical' Australian inventory is depicted in Table I.

Permissible phonotactic sequences in Australian languages are also highly constrained and similar across the continent [1]. Nevertheless, Gasser & Bowern [9] recently demonstrate that higher-definition frequency data may reveal variation that is not apparent in binary, permissibility data. One contribution of the present study is the first quantification of the difference in phylogenetic signal between coarse, permissibility data and richer, frequency data.

II. Language Data

We study 17 languages in two subgroups of the large, Pama-Nyungan family: Ngumpin-Yapa [10], [11], which stretches across central Australia, and Yolngu [12], located

TABLE II. 'Typical' Australian Inventory *(after [11, p. 141])*

	Peripheral		Apical		Laminal	
	Bilabial	*Dorso-velar*	*Apico-alveolar*	*Apico-retroflex*	*Lamino-dental*	*Lamino-palatal*
Stop	p	k	t	ʈ	t̪	c
Nasal	m	ŋ	n	ɳ	n̪	ɲ
Lateral			l	ɭ	l̪	ʎ
Trill			r			
Glide	w			ɹ		j

	Front	Back
High	i, iː	u, uː
Low	a, aː	

discontinuously from the rest of Pama-Nyungan, in the north.

A. The Ngumpin-Yapa subgroup of Pama-Nyungan

We choose Ngumpin-Yapa for two reasons: Firstly, the phonological systems of Ngumpin-Yapa languages accord closely with the 'typical' characteristics of Australian phonologies (§1). Eight of ten Ngumpin-Yapa languages feature a single series of stops (Warlmanpa and Warumungu have a second stop series). All have stops and nasals at five places of articulation, three laterals, two rhotics (three in Warlpiri), two semi-vowels and a triangular vowel system. Secondly, Ngumpin-Yapa vocabularies exhibit high levels of historical borrowing. In particular, the Eastern Ngumpin branch shows some of the highest rates of lexical borrowing observed in the world [13], [14]. This challenges our method, and therefore makes for a robust case study.

B. The Yolngu subgroup of Pama-Nyungan

Yolngu languages contrast with Ngumpin-Yapa in that they possess two phonemic series of stops, as well as six superlaryngeal places of articulation (five in Djinang). The glottal stop also has a marginal phonemic status. There are six nasals (five in Djinang), three laterals, two rhotics and two semi-vowels. Vowels contrast three qualities, plus length (except in Djinang, where length is not contrastive).

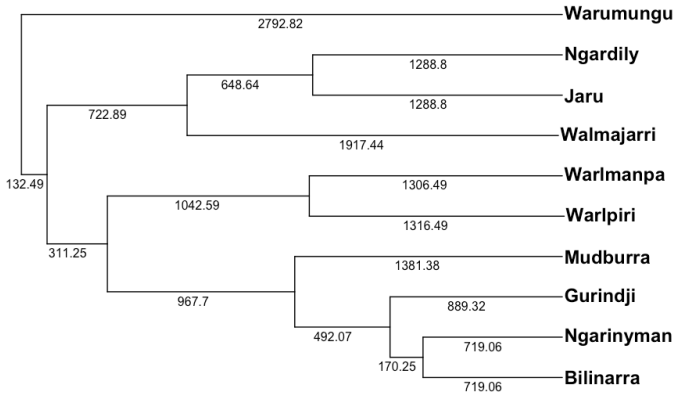

Fig. 1. Ngumpin-Yapa phylogeny and branch lengths.

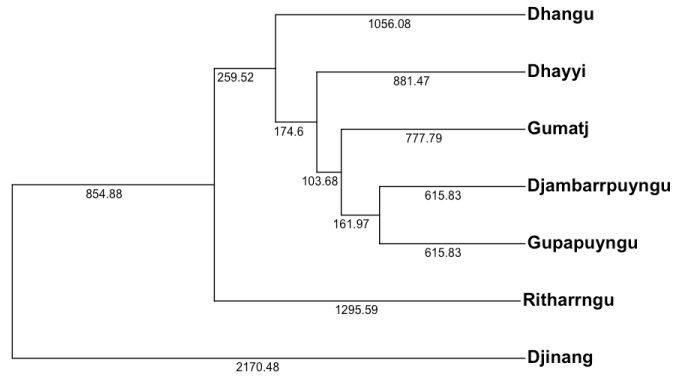

Fig. 2. Yolngu phylogeny and branch lengths.

C. Reference phylogenies

Our experiments below measure the phylogenetic signal in phonotactic data, relative to a reference phylogeny. Our reference phylogenies are from C. Bowern, inferred from basic vocabulary cognacy data, as expanded and updated from Bowern & Atkinson [11].

The reference phylogeny for Ngumpin-Yapa is in Fig. 1. This differs from prior, manual analysis [10] in that the Ngumpin clade is split, with Eastern Ngumpin languages grouping with the Yapa clade, and Western Ngumpin languages forming their own clade.

The reference phylogeny for Yolngu is in Fig. 2. This differs from prior scholarship in not containing the 'northern' clade proposed tentatively by Heath [15] and grouped tentatively with Heath's 'western' clade by Bowern [16].

D. Trait data

Phonotactic traits were extracted from orthographic wordlists drawn from the Australian lexical database [17]. After semi-automatic scrubbing, the 17 phonemicized lexicons contained from 535–8634 word forms (mean 2219). The permissibility of a biphone $a+b$ in a language was coded as a binary trait, and continuous traits encoded Markov chain (forward) transition probabilities.

III. Experiment one: Binary data

A. Method

Two experiments were performed. Experiment one tested binary, 'permissibility' data for phylogenetic signal using the D test of Fritz and Purvis [18], implemented in the *R* function *phylo.d* (package *caper*) [19]. The D statistic, like K below, is designed to be independent of tree size and shape [18], making it appropriate for comparison between different subgroups. The statistic is estimated for each trait, and the results examined as an ensemble. D sums differences between sister tips and sister internal nodes across the phylogeny, and deducts this from expected values, if the trait were distributed according to Brownian evolution. This quantity is then divided by the expected sum under Brownian evolution less the expected sum of a phylogenetically random distribution [18], as stated in equation 1:

$$D = (\Sigma d_{obs} - \Sigma d_o) / (\Sigma d_b - \Sigma d_r) \quad (1)$$

If tip values perfectly reflect the phylogeny, then D=0. If they are randomly distributed, D=1. The test can return values of D<0 if tip values are clumped more conservatively than expected under Brownian evolution and D>1 if they are dispersed more evenly than expected under a random distribution [18]. A potential drawback of D is that it loses stability and statistical power (i.e., the ability to discriminate true from false negatives) with small datasets (<50 taxa) [18]. This can elevate false discovery rates [20], which we keep in mind when interpreting the results.

B. Results

D tests require a trait to be valued for all languages, and to distinguish at least one language pair. Against these constraints, we extracted 184 coarse-grained, binary traits for

the Ngumpin-Yapa languages and 164 traits for Yolngu. Summary results are in Table II.

The *D* statistic is tested against two null hypotheses: that $D=0$, indicating that trait distributions perfectly fit the reference tree; and that $D=1$, where the trait distribution is random.

For Yolngu, both null hypotheses were rejected (Stouffer's combined $p < 0.001$). This suggests that the traits in our dataset are significantly more uniform than the lexical cognate traits on which the reference phylogeny is based. Given that phonotactics ought to be less prone to borrowing than is basic vocabulary (§1), we interpret this uniformity as reflecting a high level of conservation of phonotactic patterns in Yolngu.

For Ngumpin-Yapa, only the random distribution ($D=1$) null hypothesis was rejected (Stouffer's combined $p < 0.001$). This suggests that the binary trait distribution for Ngumpin-Yapa is non-random, and that possibly it contains a degree of phylogenetic signal, though given the above-mentioned limitations of *D* with small datasets, the failure to reject the second null hypothesis may be due to low statistical power.

IV. Experiment Two: Continuous Data

A. Method

The higher-resolution, continuous phonotactic probability datasets were tested using *K* [21], as implemented in the *multiPhylosignal* function in *R* (package *picante*) [22]. Like *D,* the *K* statistic is calculated for each trait and the results are examined as an ensemble. *K* uses phylogenetically independent contrasts (PICs), defined as the difference in trait values between two tips divided by the square root of the branch length distance between them [23]. The variances of all PICs for a given trait are taken as indication of how well the trait data fit the phylogeny—the lower the variance, the better the fit. The null hypothesis of no phylogenetic signal is rejected if the observed variances are less than the variances of a 95% threshold of random permutations. Dividing the mean variances of the tip data by the PIC variances across the phylogeny quantifies the magnitude of phylogenetic signal, and this number divided by its expectation given a Brownian model of evolution yields the test statistic, *K* [21]. $K=1$ indicates that the trait data perfectly fit their expectation under Brownian evolution, with $K=0$ indicating no phylogenetic signal. $K>1$ indicates that sister taxa and clades resemble each other more closely than expected under a Brownian motion model of evolution, or conversely, that distant taxa are more highly differentiated than expected.

B. Results

K requires each trait to distinguish at least one pair of languages, and tolerates missing values (as when a phoneme in sequence $a+b$ is entirely absent from a language, and thus its transition probability is undefined). Against these constraints, we extracted 451 traits for Ngumpin-Yapa and 541 for Yolngu.

Summary results are in Table III; we visualize *K* values of individual biphone traits in Fig. 3 and Fig. 4.

TABLE II. Results for coarse-grained, binary data (*D* test)

	n(traits)	Mean *D*	SD	MFDR-CI[a]
Ngumpin-Yapa	184	0.372	3.592	[−0.23, 0.97]
Yolngu	164	−1.486	4.269	[−2.97, −0.73]

TABLE III. For higher-defintition, continuous data (*K* test)

	n(traits)	Mean *K*	SD	MFDR-CI[a]
Ngumpin-Yapa	451	0.893	0.27	[0.86, 0.92]
Yolngu	541	1.206	0.595	[1.15, 1.26]

[a] Benjamini–Hochberg [20] mean false discovery rate adjusted CI

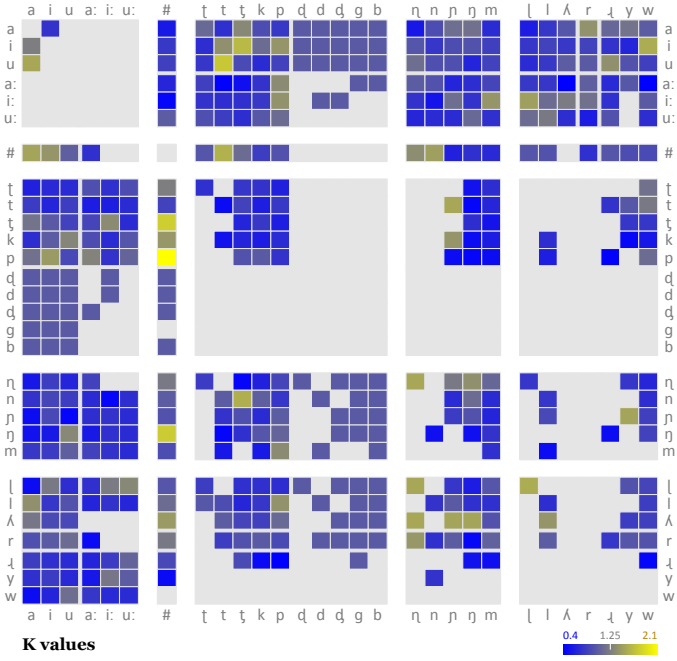

Fig. 3. *K* values of individual biphone traits in Ngumpin-Yapa.

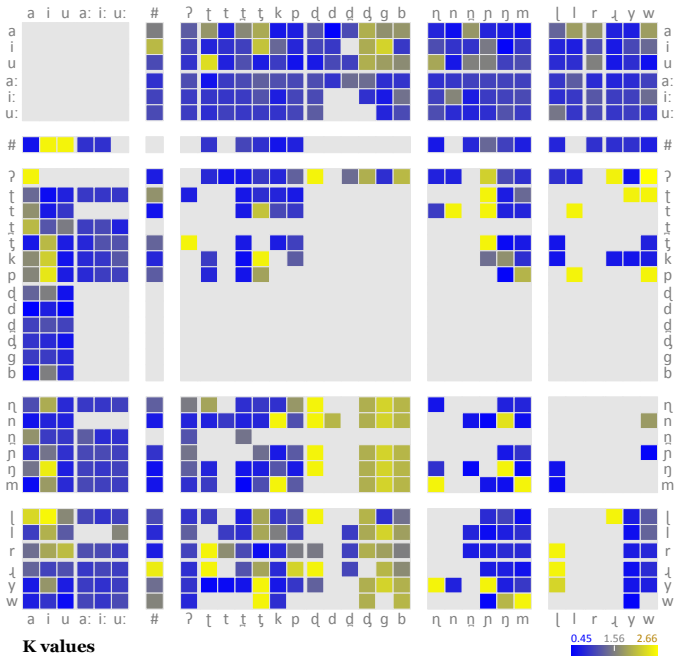

Fig. 4. *K* values of individual biphone traits in Yolngu.

The statistical significance of K is tested against a single null hypothesis, K=0, corresponding to a distribution of traits which is random relative to the reference phylogeny (i.e. no phylogenetic signal is present).

The null hypothesis is rejected for both Ngumpin-Yapa and Yolngu subgroups, suggesting that a statistically significant level of phylogenetic signal is present in both high-definition datasets (Stouffer's combined $p < 0.001$ for both subgroups). Mean K for the Ngumpin-Yapa trait data is 0.893 (CI [0.86, 0.92]), which indicates a very good match to the reference phylogeny. Mean K for Yolngu is 1.206 (CI [1.15, 1.26]). This indicates a good match to the reference phylogeny, while also suggesting that outer taxa are being discerned as particularly distinct. Indeed, in Fig 3(b), K values of individual traits are high (green-yellow) for biphones that are largely absent from all bar the outermost Yolngu language, Djinang (namely, the other languages collapse their stop series contrast in certain positions where in Djinang it is distinct). The conjecture that Djinang is driving the particularly high K result for Yolngu, prompted us to conduct our first additional test. A second additional test asks whether the improvement between experiment one, with coarse-grained data, and experiment two, with high-definition data, is driven solely by the increase in the number of biphones used, or whether the shift from coarse-grained to high-definition data was also important.

C. Additional test 1: Sensitivity to the outermost taxon

To test the conjecture that high values of K for the Yolngu subgroup were being driven by the particular distinctiveness of its phylogenetically outermost language taxon, K tests were run for each of the two subgroups with the outermost taxa removed. Without Warumungu, K for Ngumpin-Yapa changes only slightly (mean K = 0.844, CI [0.81, 0.87]). In contrast, and as predicted, once Djinang is removed, K drops for Yolngu, and suggests a near perfect match between the high-resolution phonotactic data and the reference phylogeny (mean K = 0.979, CI [0.96, 1]). We conclude that the distinctiveness of Djinang was acting to lift the K value for Yolngu in experiment two.

D. Additional testing 2: Size of the biphones set

To test whether the improvement in performance from experiment one to experiment two might be attributable to the increased size of our datasets, rather than to the shift to high-definition data, additional K tests were run using a dataset consisting of Markov chain transition frequencies (as in experiment two) but only for those biphones which were represented in experiment one, thus: 184 traits for Ngumpin-Yapa, down from 452; and 164 traits for Yolngu down from 541. The hypothesis is that if the improvement in experiment two was solely due to dataset size, then in this additional test, where we reduce the dataset in size, to the equivalent of experiment one, then the results obtained should be no better those of experiment one.

TABLE IV. *K* TEST BASED ON SAME, SMALLER BIPHONE SET AS IN EXP. 1

	n(traits)	Mean *K*	SD	MFDR-CI[a]
Ngumpin-Yapa	184	0.86	0.329	[0.80, 0.91]
Yolngu	164	1.309	0.787	[1.17, 1.45]

[a] Benjamini–Hochberg [20] mean false discovery rate adjusted CI

For Ngumpin-Yapa, with its reduced-size, high-definition dataset of 184 traits, mean K is 0.86 (CI [0.8, 0.91]), see also Table IV. Thus, even with many fewer traits, the null hypothesis of no phylogenetic signal is still rejected, as in experiment two but unlike experiment one. For Yolngu, with a reduced-size, high-definition dataset of 164 traits, mean K is 1.309 (CI [1.17, 1.45]), and thus the null hypothesis of no phylogenetic signal is again rejected. We conclude that the improvement in experiment two is not due to increased size of the dataset alone; it also due to the shift to high-definition data, enabling more powerful statistical inference.

Although the null hypotheses of no phylogenetic signal are rejected even with the reduced-size, high-definition dataset, the confidence intervals for the mean of K were wider. This suggests, as might be expected, that the extra trait data used in experiment two was informative, in that enabled a more precise estimate of mean K. (Examining the distributions of K values obtained in the additional tests and in experiment two, which are non-normal, we find a significant difference for Ngumpin-Yapa (Mann-Whitney $U(633) = 36088$, $Z = -2.578$, $p = 0.01$) but not for Yolngu ($U(703) = 45957$, $Z = -0.698$, $p = 0.485$).)

V. DISCUSSION

We have demonstrated that high-resolution phonotactic traits—specifically Markov chain transition probabilities for biphones—can be extracted in large numbers from phonemicized vocabularies. Thus, we contribute to the important task of developing methodologies which generate linguistic traits in high volumes. We then demonstrated that such data contains a degree of phylogenetic signal unlikely to arise by chance.

In contrast, the coarse-grained binary data, coding biphone permissibility, reveals little to no useful phylogenetic structure. While trait values are not randomly distributed with regards to the reference phylogeny in either subgroup, traits are too highly clumped to distinguish phylogenetic structure. This is unsurprising in light of the observation, noted in §1, that Australian languages are highly uniform when viewed in terms of permissible phonotactics.

Notwithstanding our main finding, two limitations of this study can be emphasized. First, our phonotactic datasets and reference phylogenies are not entirely independent. The vocabularies used to generate phonotactic data contain, as a small subset, the same basic vocabulary items from which lexical cognacy traits were inferred, and used to build the reference phylogenies. To ascertain whether this effect is significant, future studies should parameterize the inclusion/ exclusion of basic vocabulary from the phonotactic data. Second, the D and K tests of individual traits were treated in the present study as independent observations. This may be problematic since in language change individual phonemes frequently do not behave independently, but pattern according to natural classes. The fact that Djinang lacks an entire series of stops for example is not surprising, but commonplace. Dealing effectively with natural class effects among phonological traits (and equivalent dependencies in other domains of grammar) is

both challenging and essential, and thus is a high priority for methodological development.

Future research should aim to test larger phylogenies with more taxa, both to produce more reliable results and to test whether phylogenetic signal persists into greater time-depths.

VI. Conclusion

As linguists attempt to up-scale efforts in quantitative historical linguistics, we have demonstrated the power of phonotactic data, even at the relatively simple level of biphone transition probabilities. Our approach permitted the ready extraction of several hundred high-definition traits per language, which revealed phylogenetic signal in two subgroups of Australian languages, despite superficially extreme phonological uniformity and high rates of borrowing.

Acknowledgments

The authors thank C. Bowern, T. M. Ellison, M. Wieling and a second reviewer for valuable feedback and suggestions. J. L. M-C thanks audiences at the UQ SLC Honours Work-in-Progress Seminar and the Rhizomes IX Conference, for their thoughts and suggestions. This research has received support from Australian Research Council grant DE150101024 to E.R. This support is gratefully acknowledged.

References

[1] P.J. Hamilton, "Phonetic constraints and markedness in the phonotactics of Australian languages," Ph.D. dissertation, University of Toronto, 1996.

[2] W.K. Ching and M.K. Ng, *Markov Chains: Models, Algorithms and Applications*. New York: Springer, 2006.

[3] J.R. Norris, *Markov Chains*. New York: Cambridge University Press, 1997.

[4] L. Hyman, "The role of borrowings in the justification of phonological grammars," *Studies in African Linguistics* vol. 1, pp. 1–48, 1970.

[5] A. Capell, *A New Approach to Australian Linguistics*. Sydney: University of Sydney, 1956.

[6] R.M.W. Dixon, *The Languages of Australia*, Cambridge: Cambridge University Press, 1980.

[7] P. Busby, "The distribution of phonemes in Australian Aboriginal languages," *Papers in Australian Linguistics,* No. 14 (Pacific Linguistics Series A-60). Canberra: Australian National University, 1980.

[8] B. Baker, "Word structure in Australian languages," in *The Languages and Linguistics of Australia: A comprehensive guide*, H. Koch and R. Nordlinger, Eds. Berlin: De Gruyter Mouton, 2014, pp. 139-214.

[9] E. Gasser and C. Bowern, "Revisiting phonotactic generalizations in Australian languages," Proc. Annu. Meeting Phonology, 2014.

[10] P. McConvell and M. Laughren, "The Ngumpin-Yapa subgroup," in *Australian Languages: Classification and the comparative method*, C. Bowern and H. Koch, Eds. Amsterdam: John Benjamins, 2004, pp. 151-177.

[11] C. Bowern and Q.D. Atkinson, "Computational phylogenetics and the internal structure of Pama-Nyungan," *Language*, vol. 88, no. 4, pp. 817-845, 2012.

[12] Schebeck, Bernhard *Dialect and Social Groupings in North East Arnhem Land*, typescript, Australian Institute of Aboriginal and Torres Strait Islander Studies Library, Canberra, 1968.

[13] P. McConvell, "Loanwords in Gurindji, a Pama-Nyungan language of Australia," in *Loanwords in the world's languages: A comparative handbook*, M. Haspelmath and U. Tadmore, Eds. Berlin: Mouton de Gruyter, 2009, pp. 790–822.

[14] C. Bowern, et al., "Does lateral transmission obscure inheritance in hunter-gatherer languages?" *PLoS One*, 2011: e25195.

[15] J. Heath. *Basic materials in Ritharngu: Grammar, texts and dictionary*. Canberra: Australian Institute of Aboriginal Studies, 1980.

[16] C. Bowern, *The Yolngu subgroup of Pama-Nyungan*, guest lecture, Australian National University, 2007.

[17] C. Bowern. *Australian lexical database*. Yale Univeristy. nd.

[18] S.A. Fritz and A. Purvis, "Selectivity in mammalian extinction risk and threat types: A new measure of phylogenetic signal strength in binary traits," *Conserv. Biol.*, vol. 24, no. 4., pp. 1042-1051, 2010.

[19] D. Orme, et al., "caper: Comparative analyses of phylogenetics and evolution in R," R package version 0.5.2, http://CRAN.R-project.org/package=caper, 2013.

[20] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *J. R. Stat. Soc. Series B (Stat. Methodol.)*, vol. 57, no. 1, pp. 289-300, 1995.

[21] S.P. Blomberg, T. Garland and A.R. Ives, "Testing for phylogenetic signal in comparative data: Behavioural traits are more labile," *Evolution*, vol. 57, no. 4, pp. 717-745, 2003.

[22] S.W. Kembel, et al., "Picante: R tools for integrating phylogenies and ecology," Bioinformatics, vol. 26, pp. 1463-1464, 2010.

[23] J. Felsenstein, "Phylogenies and the comparative method," American Naturalist, pp. 1-15, 1985.