THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

Phonotactics in historical linguistics:

Quantitative interrogation of a novel data source

Jayden Macklin-Cordes

BA (Hons)

0000-0003-1910-3969

A thesis submitted for the degree of Doctor of Philosophy at

The University of Queensland in 2021

School of Languages and Cultures

Abstract

Historical linguistics has been uncovering historical patterns of shared linguistic descent with considerable success for nearly two centuries. In recent decades, historical linguistics has been augmented with quantitative phylogenetic methods, increasing the scope and scale of linguistic phylogenetic trees that can be inferred and, by extension, the kinds of historical questions that can be investigated. Throughout this period of rapid methodological development, lexical cognate data has remained the main currency of interest. However, there is an immense array of structural variation throughout the world's languages—the outcome of thousands of years of linguistic evolution—raising the prospect of additional sources of historical signal which could help infer shared linguistic pasts.

This thesis investigates the prospect of making linguistic phylogenetic inferences using quantitative phonotactic data, semi-automatically extracted from wordlists. The primary research question motivating the study is as follows. Can a linguistic phylogeny be inferred with greater confidence when lexical cognate data is combined with phonotactic data? Underlying this is the more general question of how to approach and evaluate novel empirical data sources in linguistic phylogenetic research.

I begin with a literature review, situating current linguistic phylogenetic research within historical linguistics broadly and elucidating the motivations for looking further into phonotactics as a data source. I also introduce Australian languages and their phonological profiles, as these provide the language sample for the rest of the thesis.

Four papers follow. The first re-evaluates phoneme frequencies. The deceptively simple question of which statistical distribution best characterises the frequencies of phonemes in a language is crucial, because particular distribution types can be suggestive of particular causal processes that generated them. Using a maximum likelihood framework, I find modest evidence that more frequent phonemes are characterised by a power law distribution while less frequent phonemes are characterised by an exponential distribution. This is followed by discussion of theoretical implications.

The second paper reviews the place of phylogenetic methods in linguistics. It considers the scientific task of comparison and the potential confound of phylogenetic autocorrelation. I discuss comparative approaches from within linguistics and other domains of science and I advocate for the increased uptake of methods which control for phylogeny directly over various balanced sampling methods commonly employed in linguistic typology.

The third paper draws on the phylogenetic comparative methods discussed above to quantify the degree of phylogenetic signal in phonotactic data. Three simple data representations of phonotactics are extracted from wordlists: binary biphone data coding whether or not any given sequence of two phonological segments is present, biphone frequency data, coding the relative frequencies of a segment being preceded or followed by another segment, and sound class biphone data, coding the relative frequencies of a natural sound class being preceded or followed by another sound class. I find some degree of phylogenetic signal in all datasets, with the strongest signal in the sound class dataset.

Finally, the fourth paper tests the primary research question of this thesis. Using a Bayesian computational approach, I infer and compare phylogenies of 44 western Pama-Nyungan languages

using two models: one in which the tree is inferred using binary biphone data, sound class data and pre-existing cognate data together, versus one in which two trees are inferred, one using the phonotactic data and the other using the cognate dataset alone. Within the bounds of current computational limits, I do not find evidence that phonotactic data strengthen support for the resulting phylogeny. Although marginal likelihood estimates seem to offer prima facie support for the hypothesis, there is evidence that the evolutionary model for phonotactic data is insufficient and the resulting calculations are unreliable.

I conclude the thesis with a discussion, drawing together the previous papers and detailing a methodological scaffold for investigating novel sources of linguistic data in quantitative historical linguistics in future. There are clear steps for future research to pursue with phonotactics. More generally, however, I present an empirical, step-wise process for methodological development in quantitative historical linguistics, from the initial link between language change processes and observable data distributions, to implementation within complex, expressive phylogenetic models.

Declaration by author

This thesis is composed of my original work, and contains no material previously published or written by another person except where due reference has been made in the text. I have clearly stated the contribution by others to jointly-authored works that I have included in my thesis.

I have clearly stated the contribution of others to my thesis as a whole, including statistical assistance, survey design, data analysis, significant technical procedures, professional editorial advice, financial support and any other original research work used or reported in my thesis. The content of my thesis is the result of work I have carried out since the commencement of my higher degree by research candidature and does not include a substantial part of work that has been submitted to qualify for the award of any other degree or diploma in any university or other tertiary institution. I have clearly stated which parts of my thesis, if any, have been submitted to qualify for another award.

I acknowledge that an electronic copy of my thesis must be lodged with the University Library and, subject to the policy and procedures of The University of Queensland, the thesis be made available for research and study in accordance with the Copyright Act 1968 unless a period of embargo has been approved by the Dean of the Graduate School.

I acknowledge that copyright of all material contained in my thesis resides with the copyright holder(s) of that material. Where appropriate I have obtained copyright permission from the copyright holder to reproduce material in this thesis and have sought permission from co-authors for any jointly authored works included in the thesis.

Publications included in this thesis

Macklin-Cordes, Jayden L., Claire Bowern & Erich R. Round. 2021. Phylogenetic signal in phonotactics. *Diachronica* 38. https://doi.org/10.1075/dia.20004.mac.

Macklin-Cordes, Jayden L. & Erich R. Round. 2020a. Re-evaluating phoneme frequencies. *Frontiers in Psychology* 11. https://doi.org/10.3389/fpsyg.2020.570895.

Submitted manuscripts included in this thesis

Macklin-Cordes, Jayden L. & Erich R. Round. Submitted. Challenges of sampling and how phylogenetic comparative methods help: With a case study of the Pama-Nyungan laminal contrast. Preprint. https://doi.org/10.5281/zenodo.4796981.

Other publications during candidature

Round, Erich R., T. Mark Ellison, Jayden L. Macklin-Cordes & Sacha Beniamine. 2020. Automated parsing of interlinear glossed text from page images of grammatical descriptions. In *12th language resources and evaluation conference (LREC-2020)*, 2871–2876. Marseille, France: European Language Resources Association. https://www.aclweb.org/anthology/2020.lrec-1.350.

Conference presentations

Hollis, Jordan, Genevieve C. Richards, Jayden L. Macklin-Cordes & Erich R. Round. 2016. The Cape York lexical records of Bruce Sommer. In *Australian Linguistics Society (ALS) annual conference*. Monash University, Caulfield, Australia. https://doi.org/10.6084/m9.figshare.4299377.

Macklin-Cordes, Jayden L. 2017. High-definition phonotactic typology in Sahul: Choosing the data-rich approach. In *Association of linguistic typology (ALT-12)*. Australian National University, Canberra, Australia. https://doi.org/10.6084/m9.figshare.5861067.

Macklin-Cordes, Jayden L., Nathaniel L. Blackbourne, Thomas J. Bott, Jacqueline Cook, T. Mark Ellison, Jordan Hollis, Edith E. Kirlew, Genevieve C. Richards, Sanle Zhao & Erich R. Round. 2017. Robots who read grammars [poster]. In *CoEDL fest*. Alexandra Park Conference Centre, Alexandra Headlands, QLD, Australia. https://doi.org/10.6084/m9.figshare.4625248.

Macklin-Cordes, Jayden L., Steven P. Moran & Erich R. Round. 2016. Evaluating information loss from phonological dimensionality reduction. In *Speech science and technology (SST) satellite symposium: The role of predictability in shaping human language sound patterns*. Western Sydney University, Australia. https://doi.org/10.6084/m9.figshare.4465976.

Macklin-Cordes, Jayden L. & Erich R. Round. 2016a. High-definition phonotactic data contain phylogenetic signal [poster]. In *90th annual meeting of the Linguistic Society of America (LSA)*. Washington, DC. https://doi.org/10.6084/m9.figshare.4534238.

Macklin-Cordes, Jayden L. & Erich R. Round. 2016b. High-definition phonotactics reflect linguistic pasts. Yale Linguistics Friday Lunch Talk. Yale University, New Haven, CT.

Macklin-Cordes, Jayden L. & Erich R. Round. 2016c. Reflections of linguistic history in quantitative phonotactics. In *Australian Linguistics Society (ALS) annual conference*. Monash University, Caulfield, Australia. https://doi.org/10.6084/m9.figshare.4299365.

Macklin-Cordes, Jayden L. & Erich R. Round. 2017. Fine-grained phonotactics in Sahul: Data-rich comparison for deep-time analysis [poster]. LSA Linguistic Institute Poster Session. University of Kentucky, Lexington.

Macklin-Cordes, Jayden L. & Erich R. Round. 2018. Phylogeny in phonotactics: A multivariate approach for stronger historical signal. In *Australian Linguistics Society (ALS) annual conference*. University of South Australia, Adelaide.

Macklin-Cordes, Jayden L. & Erich R. Round. 2019. Historical signal beyond the cognate: Phonotactics in linguistic phylogenetics. In *International conference on historical linguistics (ICHL24)*. Australian National University, Canberra, Australia. https://cloudstor.aarnet.edu.au/plus/s/zN5AbJ4fXoVR0ax.

Macklin-Cordes, Jayden L. & Erich R. Round. 2020b. Zipf's law? Re-evaluating phoneme frequencies. In *Australian linguistics society (ALS) annual conference*. [online conference].

Round, Erich R., Jacqueline Cook, T. Mark Ellison, Jordan Hollis, Jayden L. Macklin-Cordes & Genevieve C. Richards. 2016. The grammar harvester. In *Australian Linguistics Society (ALS) annual conference*. Monash University, Caulfield, Australia.

Round, Erich R. & Jayden L. Macklin-Cordes. 2019. Clouded vision: Why insights improve when uncertainty about data is made explicit. In *International conference on historical linguistics (ICHL24)*. Australian National University, Canberra, Australia. https://cloudstor.aarnet.edu.au/plus/s/WOmUM73NVJQFqf6.

Contributions by others to the thesis

Data for this research comes from the Ausphon Lexicon database (Round 2017c), which is designed and maintained by Erich Round and extends the CHIRILA database (Bowern 2016).

Additional data, methodological advice and feedback on manuscripts come from Claire Bowern, Simon Greenhill, Gabriel Hassler, David Nash, Caleb Everett, Steven Moran and two anonymous reviewers. These contributions are detailed further at the beginning of each chapter, where applicable.

Statement of parts of the thesis submitted to qualify for the award of another degree

No works submitted towards another degree have been included in this thesis.

Research involving human or animal subjects

No animal or human subjects were involved in this research.

Acknowledgments

First and foremost, I wish to thank my phenomenal supervisor, colleague and friend, Erich Round. To Erich, thank you so much for your invaluable mentorship, all the tea times, and your willingness to indulge my occasional thought bubbles. Thank you also to fellow supervisor, Simon Greenhill, for your guidance and expertise and for the incredible opportunities afforded to me. I feel so fortunate to have worked with two such exceptional people. Thank you also to Ilana Mushin and Barbara Hanna for your support and feedback, and keeping my candidature on track. Naturally, any outstanding deficiencies in my work remain my own.

Undoubtedly, one of the most rewarding aspects of this PhD program has been the wonderful friends and colleagues I've spent time with along the way. Thank you to Mitch Browne, Brian Collins, Jacqui Cook, Vivien Dunn, Tom Ennever, Claire Gourlay, Amanda Hamilton-Holloway, Jordan Hollis, Celeste Humphris, Edith Kirlew, David Osgarby, Martin Schweinberger, Janet Watts and Ruihua Yin. Thank you also to Rikker Dockum, Vanya Kapitonov, Nay San, Sasha Wilmoth, and fellow friends and classmates from the 2016 Quantitative Methods Spring School, 2017 LSA Summer Institute, CoEDL Summer Schools, and various conferences. Inevitably, there will be more people that I've missed, but I truly appreciate every one.

I owe eternal thanks to my family for their boundless love and support, without which none of this would be possible. Credit must also go also to my friends outside of academia for their role in keeping me healthy, happy and grounded. This thesis could not have happened without a break for parkrun and pancakes on a Saturday morning.

Final thanks are reserved for my wife, Brittany. Thank you for your heroic support, patience and everything else.

Financial support

This research was supported by an Australian Government Research Training Program Scholarship.

Summer Institute attendance, lab visits and conference travel were supported by the UQ School of Languages and Cultures, the Max Planck Institute for the Science of Human History, the Linguistic Society of America, the ARC Centre of Excellence for the Dynamics of Language and the Yale University Linguistics Department.

I gratefully acknowledge all the generous support that made this thesis possible.

Keywords

Australian languages, Pama-Nyungan, linguistic phylogenetics, phoneme inventories, frequency distributions, phylogenetic comparative methods, phylogenetic signal, quantitative methods, phylogenetic tree inference, phonology

Australian and New Zealand Standard Research Classifications (ANZSRC)

ANZSRC code: 200406, Language in Time and Space, 60%
ANZSRC code: 200408, Linguistic Structures, 30%
ANZSRC code: 200402, Computational Linguistics, 10%

Fields of Research (FoR) Classification

FoR code: 2004, Linguistics, 100%

For GGD.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Phonotactics and phylogeny

Can the minutiae of sound sequences in human language encode something of human history? And how can we tell? This thesis presents a methodology for interrogating novel data sources in quantitative historical linguistics and finds that the answer may be yes, though extracting it is a complex challenge.

Quantitative advances in historical linguistics mean that researchers are able to extract fine-grained historical patterns from large datasets and infer increasingly detailed phylogenies of the world's language families (see Dunn 2015, Bowern 2018a). However, quantitative historical linguistics remains predominantly reliant on lexical cognate data. Lexical cognate identification has been a cornerstone of historical linguistic methodology for over a century and a half, but requires extensive human expertise and manual labour, creating a bottleneck in current quantitative historical linguistic practice. It also represents but one subpart of the human linguistic system. There has been some interest in phylogenetic tree inference using structural features of languages (e.g. Dunn et al. 2005, 2008, Sicoli & Holton 2014) but the limited parameter space and apparent fast rate of change of many grammatical features can cloud historical signal and make phylogenetic inference difficult (Reesink & Dunn 2012, Greenhill et al. 2017).

This thesis presents an empirical evaluation of the inferential capacity of phonotactics—the system of constraints defining how phonemes may fit together in linear sequences in a language—in quantitative historical linguistics. The motivation to consider phonotactics as a data source in historical linguistic study is two-fold. Firstly, there is a practical element. Phonotactic information can be extracted relatively simply by examining the sequences of segments that appear in a language's lexicon. Given a set of comparably segmented wordlists from a variety of languages, a comparative dataset can be harvested rapidly using automated means. In the simplest case, this could be a dataset simply coding the presence or absence of sequences of two segments, or *biphones*. Frequencies of sequences can also be extracted, which are shown to contain finer grained, informative variation in Chapter 5. It would also be possible to extract automatically more sophisticated forms of data, encoding syllable structure for example, though this is beyond the scope of this thesis (though see Chapter 7 for further discussion

of future directions). This relatively low resource requirement is advantageous in under-described parts of the world, such as Australia, which is the focus of this thesis. The second motivation is more theoretical. A prerequisite for reconstructing phylogeny from some novel data source is, naturally, that the data source contains historical signal (Dunn 2015). The hypothesis that phonotactics might potentially be a source of historical signal is based on evidence that languages often adapt loanwords to fit within the existing phonotactic system such that novel words will reflect phonotactic tendencies that are present in the already-existing lexicon (Hyman 1970, Silverman 1992, Crawford 2009, Kang 2011), and thus a language's phonotactic system could be expected to remain historically conservative even in instances of lexical borrowing, which is a source of noise in lexical data. A language's phonotactic system is also unaffected by instances of undetected semantic shift, which is another source of noise in lexical data. Finally, with regards to phonotactic frequencies specifically, there is evidence that speakers are sensitive to phonological frequencies generally, and this has a role in lexical innovation (Coleman & Pierrehumbert 1997, Albright & Hayes 2003, Hayes & Londe 2006, see also Gordon 2016: pp. 20–21).

The question this thesis seeks to address, in simple terms, is this: Does phonotactic data help or hinder the task of phylogenetic tree inference? As just described, there are reasons to expect that it should help. But this immediately raises the question: What does it mean to 'help' phylogenetic tree inference, and how can we know whether it helps or not? Unlike, say, the evaluation of a novel drug or vaccine, where it is possible to link a test condition to an observable outcome, it is not as immediately clear how to measure the effect of a novel methodology in linguistic phylogenetics, where the target outcome (i.e. the true historical phylogeny) is in the non-observable past. In service of the main aims, this thesis presents a step-by-step process for empirically interrogating new data in a linguistic phylogenetic context, moving from theoretical speculation that the data contains phylogenetic signal to mathematical operationalisation of language change processes, implementation in phylogenetic models and, finally, evaluation using model comparison methods.

This thesis concentrates on languages of Australia, starting with an Australia-wide sample of 166 languages in Chapter 3 and narrowing to 44 languages of the western branch of the Pama-Nyungan family in 6. The methodological reasons for shrinking the language sample will be elaborated on in the relevant chapters ahead. The Pama-Nyungan family, far and away the largest and most geographically expansive language family in Australia, has been well established on traditional historical linguistic grounds and its phylogenetic structure has been inferred using computational phylogenetic methods and lexical cognate data (Bowern & Atkinson 2012, Bouckaert, Bowern & Atkinson 2018). However, deep-time relationships between Pama-Nyungan and the various smaller Australian language families in the north are not well established. Even more ancient relationships between Australian languages and the rest of Sahul—the continent of Australia and New Guinea, which has constituted a single landmass for most of human history—is subject to intrigue and conjecture. I discuss this background further in Chapter 2 and argue that linguistics is an important component within the multidisiciplinary efforts that will be required to further complete this picture. Returning to the present study, though, the primary goal here is not to increase our understanding of the phylogeny of Pama-Nyungan or

Australian languages per se. Rather, Australian languages and Pama-Nyungan languages in particular serve as an excellent test case since data availability is good (lexical data in Bowern 2016, Round 2017c) and established phylogenies exist to serve as a yardstick or point of comparison (Bowern & Atkinson 2012, Bouckaert, Bowern & Atkinson 2018). Theoretically too, Australian languages serve as an interesting test case. Australian languages have been described frequently as being remarkably homogeneous in their phonologies and phonotactics (Round 2021b,a) (see 2 for more discussion and references). This should make for a deliberately difficult test case, since such homogeneity would be expected to cloud phylogenetic signal.

This thesis finds that phonotactics, extracted at the relatively simple level of binary and frequency-based biphone characters, do show evidence of containing phylogenetic signal. However, a more complex story unfolds amidst the challenge of extracting it. On the main question of whether phonotactic data 'helps' phylogenetic inference, the results at this stage are indeterminate. The phylogenetic models with phonotactic data that I test largely fail to converge, and produce marginal likelihood estimates (which indicate model fit) that are unreliable. There are two possible interpretations of this, which are not mutually exclusive. One is that the phonotactic data are genuinely non-tree-like and introduce noise to the model. The other is that the implementation of an evolutionary model for phonotactic data is inadequate. The latter seems possible, even likely, since the evolutionary model had to be compromised a great deal in an effort to keep the project computationally feasible. The challenge of designing an evolutionary model that is simultaneously faithful to the real-world evolutionary processes of phonotactics while remaining computationally realistic is difficult though perhaps not entirely intractable. I outline further directions on this topic in Chapter 7.

This thesis nevertheless contributes to the field a paradigm for empirical data innovation in linguistic phylogenetics. The set of steps employed in the following chapters can be generalised to any novel kind of linguistic data for which there is a reasoned hypothesis that historical signal might be present. This starts with a sound statistical (re-)evaluation of how the data is distributed and how this fits causal linguistic processes. Next, phylogenetic signal needs to be quantified in an area where an independent reference phylogeny already exists as a yardstick. The detection of phylogenetic signal does not in itself indicate the data can be used for phylogenetic tree inference. But, in combination with a well-founded hypothesis on causal evolutionary processes, it would justify further testing. Next, the causal processes that shape the data need to be operationalised in an evolutionary model, then phylogenetic tree inference can take place. The overall fit of a phylogenetic tree model which includes the novel data can be compared to the fit of an equivalent model with separate tree elements for novel and existing data, giving an indication of whether the novel data strengthens phylogenetic tree inference.

1.2 Research question

The primary research question this thesis seeks to address is whether or not phonotactic data, in conjunction with lexical cognate data, can strengthen the inference of linguistic phylogenies. However,

since it is not possible to directly observe the past and see whether a phylogenetic tree inferred with phonotactic data more closely approaches the true historical phylogeny, this presents a challenge of testability. What does it mean to strengthen phylogenetic tree inference? In this thesis, I operationalise this question in terms of model fit. The main research question can then be restated more precisely as follows. Does the fit of a phylogenetic model improve when the phylogenetic tree is inferred from cognate data and phonotactic data together, compared to an equivalent model in which trees are inferred from cognate data and phonotactic data separately? Model fit, in this context, is inferred by estimating marginal likelihoods and, secondarily, comparing support values for clades in each tree.

Underlying this is the more general question of how to approach and evaluate empirically novel data sources in linguistic phylogenetic research, in particular continuous-valued characters such as frequencies. If one has a well-reasoned hunch that some data may contain informative phylogenetic information, how should one test this hypothesis? How should one progress from this hypothesis to implementation in a phylogenetic model and eventual tree inference, if no one has inferred trees from this kind of data before? In the chapters that follow, I illustrate this progression. In the course of doing so, I address some specific research questions along the way: (1) What mathematical distribution best characterises the observed frequencies of a language's phonemes? And what links, if any, can be inferred between phoneme frequency distributions and causal sound/language change processes that shaped them? (2) What methods exist for identifying phylogenetic information (or, more precisely, *phylogenetic signal*) in comparative datasets? (3) Does phonotactic frequency data contain phylogenetic signal? And, finally, (4) do marginal likelihoods and clade support values support a phylogeny inferred from lexical cognate data and phonotactic frequency data jointly over a comparable model in which trees are inferred from cognate data and phonotactic data separately? What is the effect of adding phonotactic frequency data to a phylogenetic tree model?

1.3 Chapter outline

This thesis presents four research papers, completed by a literature review chapter at the start, and discussion and conclusion chapters at the end. Each of the four research chapters are either published, accepted for publication or in preparation for publication independently from one another. Together, however, these papers demonstrate a step-wise process of empirically evaluating and implementing a novel source of data in quantitative historical linguistics. The discussion chapter ties these parts together and outlines a future research paradigm for data innovation in quantitative historical linguistics. I briefly introduce each remaining chapter below.

Chapter 2 places the thesis within its scientific domain by reviewing the history of the historical linguistics field, particularly with regards to methodological development and the rise of phylogenetic linguistic methods in recent decades. Subsequently, I give a background overview of Australian languages and their phonological systems. The final section widens the scope to frame the previous discussion within the context of the deep-time human history of Sahul—the continent of Australia and New Guinea. This is a fascinating area of multidisciplinary endeavour, and it will be shown that

historical linguistic investigation of Australian languages is an essential component. The chapter concludes with a discussion of how present limitations of existing works motivate the research aims of this thesis.

Chapter 3 presents a re-evaluation of phoneme frequencies in the lexicons of 166 Australian languages. The motivation for this re-evaluation is two-fold. Firstly, phoneme frequency distributions are the stationary outcomes of diachronic, causal processes. Developing a sound understanding of these distributions can, therefore, improve our understanding of the processes that shaped them. Secondly, the identification of putative power law distributions, which includes the famous Zipfian distribution, has undergone major revision across the sciences in recent times as traditional methods for power law identification were shown to be unreliable (Clauset, Shalizi & Newman 2009). I use a more reliable maximum likelihood statistical framework to evaluate whether phoneme frequencies are fitted best by a power law distribution or one of several alternative, heavy-skewed distributions (lognormal, exponential and Poisson). The results largely confirm an earlier study of phoneme frequencies in mostly European languages (Tambovtsev & Martindale 2007) but also add qualification and nuance. This is followed by reasoned discussion on potential links between these results and causal processes in phonological evolution.

Chapter 4 reviews literature on the topic of phylogenetic autocorrelation—similarity due to phylogenetic relatedness—in linguistics and comparative biology. The paper compares and contrasts methodological approaches to phylogenetic autocorrelation in both fields, introduces phylogenetic comparative methods and discusses the concept of phylogenetic signal—the tendency of closely related languages/species to appear similar to one another to a greater degree than expected by chance. I argue for the continued uptake of phylogenetic comparative methods in comparative fields of linguistics and foreshadow the methodological approach that follows in Chapter 5.

Chapter 5 measures phylogenetic signal in several kinds of phonotactic data harvested from the lexicons of 112 Pama-Nyungan languages. Three datasets are tested using diagnostic statistics developed by Blomberg, Garland & Ives (2003): (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. The results indicate that a statistically significant phylogenetic signal is present in all three datasets, but in varying degrees. The binary dataset (unsurprisingly) contains the weakest phylogenetic signal. The finer-grained frequency datasets contain stronger signal, and signal in the sound class dataset in particular is strongest. The presence of non-random, phylogenetic tree-like structure within these relatively simple representations of phonotactic systems gives support to the notion that phonotactic data could be used profitably for novel phylogenetic tree inference in linguistics.

Chapter 6 puts into practice the findings of previous chapters and attempts to test the main research question of this thesis. I attempt to infer a phylogeny of 44 western Pama-Nyungan languages using a combination of existing lexical cognate data from Bouckaert, Bowern & Atkinson (2018), binary biphone data and sound class transition frequency data. Two phylogenetic models are evaluated using a Bayesian Markov chain Monte Carlo process: (1) A ‘linked’ model in which a single tree is inferred

from all three datasets together, and (2) a 'separate' model in which two trees are inferred—one from cognate data only and the other from the two phonotactic datasets only. The two models are compared firstly by comparing marginal likelihood estimates, which give an overall indication of model fit, and secondly by comparing maximum clade credibility trees and posterior clade support values. The present study returns results that are indeterminate overall. Inferring trees using continuous valued characters at scale is computationally non-trivial and I am unable to infer consistent, stable results within current constraints of the model. I provide a breakdown of these limitations and detail the steps required to investigate the role of phonotactics in phylogenetics further.

Tying together each of the previous chapters, Chapter 7 proposes a future research program for the empirical interrogation of novel data sources in linguistic phylogenetics. Beyond the immediate steps necessary to further investigate phonotactics specifically, I suggest there is a more generalisable set of principles which could be applied to linguistic frequency data of other kinds. This would start with a well-reasoned hypothesis for how the frequency characters would change, tested empirically through statistical evaluation of the shape of frequency distributions. The next prerequisite is a well-reasoned hypothesis for why the frequency characters are expected to be historically conservative, followed by a statistical evaluation of whether this is the case via phylogenetic comparative methods. Finally, knowledge of how the frequency characters shift through time needs to be built into a phylogenetic model for eventual implementation in phylogenetic tree inference.

The thesis concludes with a brief summary in Chapter 8.

Appendix A contains an illustration of the reference phylogeny used in Chapter 5 for convenience. It is also available online with the rest of the supplementary materials associated with that chapter.

Appendix B contains the bibliographic details of all the original sources for language wordlists used throughout the thesis. Additional supplementary information is available online, including guides for navigating datasets and running code, and ancillary information on methods. Details of these online repositories are given in the applicable chapters. All the data, code and raw results files are also available online. Each study is designed to be fully reproducible and uses only free, open source software throughout.

Chapter 2

Literature Review

This chapter surveys existing literature, firstly to illuminate the motivation for this thesis and position the research questions of the following chapters in their scientific context, and secondly to situate this study in a broader context of deep-time human and linguistic history in the continent of Sahul. The chapter proceeds in two parts, addressing each of these goals in turn. Section 2.1 starts by tracing the origin and development of historical linguistics as a field. Particular focus is given to rapid developments in the past two decades, which increasingly enable integration of historical linguistics with other aspects of human history, rapidly shifting the kinds of questions that can be investigated with comparative linguistic data. This topic is subsequently taken up again in Chapter 4 with a particular concentration on the fundamental task of comparison in historical linguistics and beyond. With this context in mind, Section 2.3 turns to a particular world domain, namely the continent of Sahul. The section begins with an overview of the geospatial and ecological context of Sahul and our current best understanding of the earliest human history in the continent. The section proceeds with a picture of the current linguistic diversity in Sahul and the historical linguistic hypothesis posited for how the present linguistic ecology came to be. I then discuss how the current historical linguistic picture reconciles with our general understanding of human history in Sahul as discussed previously. Finally, a general phonological profile of the languages of Australia is provided. As the studies presented in the following chapters use data from Australian languages only, Australian languages will be the focus of these sections of the literature review.

2.1 Phylogenetic thinking in historical linguistics

Charting patterns of descent through time is a task common to several fields across the biological and cultural sciences. Occasionally, this is possible through direct observation (e.g. examination of historical or archaeological records) or controlled laboratory experiments (e.g. observing microbial evolutionary processes in a lab setting). Often, however, questions relate to a historical past that cannot be observed directly (without time travel, at least). In these cases, the best that can be done is indirect inference. The main method of historical inference is comparison. Extant entities, be they organisms,

languages, populations, etc., are compared and historical relationships between entities are inferred. To this end, a countless array of comparative data types and theoretical models have been employed, summoning what knowledge we have about how historical processes operate. Some studies may have the benefit of a partial historical record as a scaffold (e.g. biological data from modern species combined with a partial fossil record). Indirect historical inference via comparison is particularly familiar in linguistics, given the ephemeral nature of spoken language. Clearly, the global body of written language is extensive. But even where written language extends back farthest it remains a relatively recent human invention, covering around 5,000 years (Weiss 2014), which represents at most around 10% of the timespan of human language generally (and probably much less. Pagel 2017). In many parts of the world, including all of Sahul, written language has only been introduced in the past few centuries, and there remain some languages with no written records at all. Likewise, audio and audio-visual records of human language extend back only 150 years or so.

Phylogenetics is a particular subtype of historical science, focusing on questions of phylogeny. That is, the tree-like evolutionary relationships that develop from the assumption that some entities are descended from a shared common ancestor. Biologists have charted patterns of descent in phylogenetic trees since Darwin, and so too have linguists for languages for approximately as long. A phylogenetic tree is not the only way schematise historical relationships, nor is a phylogenetic schematisation necessarily mutually exclusive with other, non-tree-like historical approaches. It has, however, proved a highly successful tool for furthering understanding of all sorts of historical questions for over a century and a half.

In the last two decades, as computational phylogenetic methods have entered the historical linguistic mainstream, there has been a resurgence of discussion on the supposed analogies between linguistic evolution and biological evolution (e.g. Atkinson & Gray 2005). But how analogous is linguistic evolution to biological evolution really? And can modern computational phylogenetic methods in biology be applied to linguistics through analogy? Here, I argue for continued uptake and development of computational phylogenetic methods in linguistics but through a shifted paradigm that focusses less on direct analogy between linguistic and biological evolution and instead concentrates on the underlying mathematical elements that may or may not be common to both.

Discussions comparing biological evolution to linguistic evolution typically begin with Darwin, who seems to be the first to propose an explicit analogy between biological evolution and linguistic evolution (Darwin 1859: p. 422). Later, Darwin (1871: p. 57) states that “the formation of different languages and of distinct species, and the proofs that both have been developed through a gradual process, are curiously the same.” And this oft-quoted passage is a common starting point in discussion of linguistic and biological evolution today (e.g. Atkinson & Gray 2005, Bromham 2017, Mesoudi 2017) In a footnote to this passage, Darwin refers to Lyell (1863: ch. 13), who compares the descent of languages from generation to generation, with modification via processes such as lexical innovation, to “the force of inheritance in the organic world” (Lyell 1863: p. 457). Phylogenetic concepts were being actively developed in linguistics around the same time. Schleicher (1853) is credited with creating the earliest phylogenetic tree-like model of languages some years prior to Darwin’s famous

tree illustrations in *The Origin of Species*, and highlighted the similarity of his earlier 'Stammbaum' model to Darwin's trees (Schleicher 1863).

This early period of relatively shared phylogenetic thinking subsequently diverged into two quite independent streams of thought in the 20th century. Although biology and linguistics share more similarities in their methodological histories than is perhaps appreciated (as I discuss in Chapter 4), rapid advances in genetics and biology's subsequent quantitative turn did lead to a considerable gulf developing between the fields in the second half of the 20th century in particular. With new-found access to large volumes of genetic data and expanding computational capacity, biologists increasingly turned to quantitative algorithms to infer phylogenetic relationships and test evolutionary hypotheses (Atkinson & Gray 2005). In contrast, historical linguistics tended to remain reliant on manual, expert linguistic judgements for data acquisition (Nunn 2011). The linguistic *Comparative Method*[1] remained the methodological bedrock of historical linguistics throughout the entire 20th century and arguably remains the "gold standard" of the field today (Dunn et al. 2008: p. 712). Following Thomason & Kaufman (1988) and Campbell (2004), the Comparative Method can be summarised briefly as follows. Firstly, sets of likely cognate word forms are assembled. First pass cognacy judgements will likely be based on a fairly superficial level of similarity, with subsequent refinement later on. Known borrowings and other chance resemblances should be removed as much as possible. From these cognate sets, it is possible to identify sound correspondences between languages (i.e. where segment *x* in language A and segment *y* in language B consistently recur in the same context). Proto-phonemes can be deduced by reconciling sound correspondences with sound change processes (e.g. various categorical split and merger processes, and substantive changes such as lenition, elision etc.) and, ultimately, a lexicon of proto-word forms can be established. Whether or not it is explicitly stated, reconstruction of proto-phonemes and proto-word forms effectively involves the assumption of a kind of *weighted parsimony* evolutionary model. That is, the most parsimonious reconstructions are favoured, both in terms of the absolute number of hypothesised sound changes and the physiological and typological plausibility of those sound changes (e.g. the lenition of a voiceless stop to its voiced counterpart is a highly common, plausible sound change across the world, but a sound change from **t* to *k*, although famously attested in several Polynesian languages (Blust 2004), is rare and unusual). The Comparative Method has been applied successfully the world over and boasts a robust track record over a century and a half. However, it remains, for the most part, a time-consuming manual process that requires considerable expertise in the particular languages being studied. There has been a substantial amount of development on automating step 1—the assemblage of candidate cognate sets on the basis of similarity (List, Greenhill & Gray 2017, Rama et al. 2018, List et al. 2018). But identification of spurious matches due to borrowing and missed matches due to semantic shifts still requires manual intervention, ideally with localised knowledge of sociolinguistic context, migration patterns, trade practices and so on. A recent paper attempts to identify borrowings computationally, but with mixed results (Miller et al. 2020). With regards to identification of sound correspondences and sound change,

[1] Throughout, I capitalise the term 'Comparative Method' to distinguish the specific historical linguistic methodology that goes by this name from the generic 'comparative method', which could refer to any of the various methodologies in comparative fields of science.

there is some work towards automation in this space as well (Steiner, Cysouw & Stadler 2011, Brown, Holman & Wichmann 2013, Bouchard-Côté et al. 2013, Hruschka et al. 2015, List 2018), but this task also remains at a relatively youthful stage of development. One limitation is that, although we know in general a good deal about which kinds of sound changes are common and which are rare, there is no definitive, prior reference that quantifies exactly the likelihood of particular sound changes in the world's languages. Wu et al. (2020) present a current best practice "computer-assisted framework" (as distinct from a fully automated or fully manual one) for the Comparative Method. Nevertheless, the Comparative Method remains largely a labour- and resource-intensive process.

The eventual output of the Comparative Method is phylogenetic in nature. To classify languages into families and subgroups (or classify an individual language as a family-level or subgroup-level isolate on its own), optionally with higher-level phyla or various intermediate groupings, is to define a phylogenetic tree structure. The difference between these kinds of more traditional language classifications, obtained via the comparative method, and a fully bifurcating phylogenetic tree, as familiar in biology or computational phylogenetic studies such as Bowern & Atkinson (2012), is that more traditional linguistic classifications only infer branching events at particular levels (the level of the family, the subgroup, etc.) and lack the very detailed internal structure of a fully resolved, bifurcating phylogeny. Accordingly, multiple languages will branch out from a single common ancestor node in a rake-like pattern. The tree structure in the Glottolog database (Hammarström et al. 2020), which reflects traditional subgroupings, gives a good illustration of this.

One early, quantitative departure from the Comparative Method in historical linguistics is the method of *lexicostatistics*, developed by Swadesh (1952). The lexicostatistical method involves the inference of historical relatedness by identifying cognates and quantifying pairwise percentages of shared vocabulary between languages. Simple clustering algorithms are then used to identify families and subgroups. Swadesh (1955) later presented a method termed *glottochronology* for dating linguistic divergence events. Glottochronology effectively involves an evolutionary *strict clock* model, which assumes a constant, universal rate of lexical replacement. By assuming a particular rate of lexical replacement, phylogenetic branch lengths can be calculated and dates can be calibrated. However, glottochronology and the lexicostatistical approach fell into disfavour following a rejection of Swadesh's *universal constant* theory of lexical replacement (Blust 2000) and the unrealistic corollary that similarity necessarily equals relatedness (Bowern 2018a). Another weakness was an inability to quantify uncertainty in the results (Atkinson & Gray 2005). Nevertheless, the lexicostatistical method did provide some utility in Australia, by enabling a rough, overall classification of language families and subgroups that could serve as a preliminary sketch until such time that more robust, future comparative work could be conducted. This seems to be the spirit in which O'Grady, Voegelin & Voegelin (1966) present their lexicostatistical analysis of Australia.

The turn of the 21st century brought with it the beginnings of a new quantitative era in historical linguistics, with a surge of interest in adapting computational phylogenetic methods from evolutionary biology to infer answers to linguistic questions. Two decades into the century, these methods are increasingly commonplace in the mainstream historical linguistics scene and *linguistic phylogenetics*

arguably constitutes a subfield of study in its own right[2]. This surge towards computational phylogenetics has been driven by a multitude of factors. Some factors are technological. Computational resources have increased in power and also decreased in cost, leading to greater accessibility. Improvements in Internet connectivity and greater awareness of open science principles have led to a proliferation of large scale, open datasets. An additional motivation is the desire for greater empirical rigour in the field (e.g. testability, quantification of uncertainty, reproducibility) (Atkinson & Gray 2005, McMahon & McMahon 2003, Nunn 2011).

Much of linguistic phylogenetic work has focused on the fundamental task of phylogenetic tree inference. Several early studies focused on tree inference in the relatively well-studied confines of Indo-European (Gray & Atkinson 2003, Atkinson et al. 2005, Nicholls & Gray 2008) followed later by Ryder & Nicholls (2011), Bouckaert et al. (2012b) and Chang et al. (2015). Other early examples include studies of Austronesian (Gray & Jordan 2000) and Bantu (Holden 2002, Holden & Gray 2006). Computational phylogenetic tree inference remains a priority. More recent examples include (but are not limited to) studies of the following families: Aslian (Dunn, Burenhult, et al. 2011), Arawak (Walker & Ribeiro 2011), Pama-Nyungan (Bowern & Atkinson 2012, Bouckaert, Bowern & Atkinson 2018), Tupí-Guaraní (Michael et al. 2015), Chapacuran (Birchall, Dunn & Greenhill 2016), Dravidian (Kolipakam et al. 2018), Sino-Tibetan (Sagart et al. 2019), Dene-Yeniseian (Sicoli & Holton 2014, but c.f. Yanovich 2020) and Turkic (Savelyev & Robbeets 2020). Presently, the maturation of linguistic phylogenetics and of a broader field of evolutionary linguistics generally (Dediu & de Boer 2016, Nölle, Hartmann & Tinits 2020) has brought a widened focus on the diversity of linguistic hypotheses that can be tested with an arsenal of computational phylogenetic methods. Typological features can be studied with the use of a phylogenetic tree and phylogenetic comparative methods (e.g. Dunn, Greenhill, et al. 2011) and this will be the topic of Chapter 4. Linguistic phylogenetic trees and phylogenetic methods can be combined with other kinds of data (such as geospatial information) and evidence from other fields such as archaeology to investigate ancient human migration movements (e.g. Gray, Drummond & Greenhill 2009, Bouckaert et al. 2012b, Bouckaert, Bowern & Atkinson 2018). Other studies investigate evolutionary questions in a computational way, without explicitly including phylogenetic trees or methods, e.g. the evolution of sound systems with respect to ancient changes in human physiology (Blasi et al. 2019), the evolution of sound systems with respect to ecological environment (Everett, Blasi & Roberts 2015), sound-meaning association biases (Blasi et al. 2016), and emotion semantics (Jackson et al. 2019). Although one can draw a distinction between studies of language evolution and linguistic phylogenetics, or as Haspelmath (2020) puts it, the "evolution of linguisticality" (i.e. the biological capacity for language) and the "evolution of languages", the two can also be tied and can inform one another (Blasi et al. 2016, Nölle, Hartmann & Tinits 2020). On this basis, Segovia-Martín & Balari (2020) argues for a unified *eco-evo-devo* framework.

The uptake of computational phylogenetic methods in linguistics has not been accepted uncritically in all quarters. Atkinson & Gray (2005: p. 520) reports a "curious" aversion to phylogenetic methods in

[2]Illustrating this point, the International Conference on Historical Linguistics in July 2019, Canberra, Australia, featured a dedicated session on "Computational and phylogenetic historical linguistics".

historical linguistics, speculating that perhaps some are "haunted ... by the ghost of glottochronology past" (see below). Bowern (2018a) groups criticism of phylogenetic methods in linguistics into three categories. The first of these is the contention that an analogy between linguistic evolution and biological evolution is invalid; languages do not evolve like species and, therefore, methods for inferring the evolution of species cannot be applied justly to linguistic data (Andersen 2006, Blench 2015)[3]. It is true, of course, that languages do not evolve in perfectly analogous ways to species and many phylogenetic methods applied in biology carry in-built assumptions that do not hold in linguistics. In Chapter 4, however, I make the case that many of the supposed incompatibilities of linguistic data with phylogenetic methods, e.g. horizontal transmission in language contact situations, non-independence between correlated or logically dependent features, and uncertainty around clock models or various other parameters, are not as unique to linguistics as often assumed and there is often either existing literature or active development on addressing particular challenges applicable to linguistics without abandoning phylogenetics or quantitative methods altogether. It is useful to take a step back from direct analogising between linguistics and evolutionary biology, and instead focus on selecting the right tools for the right job by linking historical processes to basic mathematical functions. Naturally, there will still be limitations and uncertainty but these can be made explicit and evaluated empirically. In the remaining chapters of this thesis, I demonstrate this approach.

A second line of criticism relates to the validity of applying quantitative methods to linguistic data, stemming from (arguably erroneous) associations to lexicostatistics and glottochronology (Eska & Ringe 2004, Holm 2007). The problems with lexicostatistics and glottochronology are well described and these methods have been discredited for several decades (see Bowern 2018a: pp. 285–286). The history of computational phylogenetic methods, though, is essentially independent from lexicostatistics and glottochronology and the methods share little in common besides their generally quantitative nature. If lexicostatistics and glottochronology are rejected, it does not follow that computational phylogenetic methods must be rejected as well. To give some specific examples, computational phylogenetic methods need not carry the assumption that there is a constant rate of lexical replacement, as assumed in glottochronology. An array of clock models exist. These can allow, for example, different parts of the lexicon to evolve at faster or slower rates, and different rates of change in different subparts of a phylogenetic tree. Further, different clock models can be evaluated and compared empirically. Furthermore, phylogenetic methods rely on vastly more sophisticated evolutionary models to determine branching events, and do not simply cluster languages based on a similarity metric.

A third line of criticism concerns a reliance on lexical data within phylogenetic methods, rather than reconstructed sound changes or morphological changes. Despite some efforts to infer phylogenies with datasets of various typological features (e.g. Dunn et al. 2005, 2008, Sicoli & Holton 2014), the vast bulk of phylogenetic studies in linguistics do rely on lexical cognate data. The extent to which this is a problem is unclear. Bowern (2018a) argues that there is no evidence to suggest that lexical data are subject to an especially different evolutionary process than other parts of language. She also questions the primacy of sound change evidence over evidence from other parts of language

[3] See also Alexander Lehrman, quoted by Balter (2004: p. 1326).

in historical linguistics, and points out that lexical data and sound change data are not completely independent, since sound change reconstruction relies on the assemblage of lexical cognate data in the first place. I would argue that, to address these questions, a priority for the methodological development of linguistic phylogenetics is to evaluate empirically the evolutionary dynamics of linguistic features beyond lexical data. Greenhill et al. (2017) give a particularly clear-eyed example of this, evaluating evolutionary rates of change of grammatical features versus lexical ones and finding, contrary to prior expectations, that the grammatical features evolved faster. In a similar vein, I present an evaluation of the phylogenetic patterning of phonotactic data in Chapter 5.

A fourth point of criticism, recurring particularly frequently, is the apparent limitation or inadequacy of a family-tree-based model which only shows vertical patterns of inheritance and not horizontal diffusion through, for example, linguistic contact and borrowing (Bateman et al. 1990, Donohue, Denham & Oppenheimer 2012, Gould 1987). Countering this concern, Bowern (2010) claims that this is confusing the family tree, which is essentially a visualisation tool, with the methods themselves. Further, while there is still much work to be done (and not only in linguistics), phylogenetic methods are capable of modelling assumptions about horizontal admixture, not to mention quantifying uncertainty. In addition, Greenhill, Currie & Gray (2009) conduct a simulation study to test whether phylogenetic methods are invalidated by horizontal diffusion. They find that, to the contrary, their Bayesian phylogenetic method of choice is quite robust to areal borrowing between languages (albeit with some caveats relating to dating phylogenetic branches). Robustness to borrowing is also extensively evaluated in Bouckaert, Bowern & Atkinson (2018).

A final source of criticism relates to reduction of hugely complex linguistic systems into neat datasets of numerical or binarised characters for phylogenetic analysis, without excessively compromising the dataset's ability to be meaningful and informative in a linguistic sense. Further, there is the question of whether linguistic variables can be assumed to be independent and equivalent to the degree required of phylogenetic algorithms, given the complex interplay between linguistic features and non-random, directional patterns of linguistic change that can be observed (Heggarty 2006, Round & Macklin-Cordes 2015). I would argue that these are not insurmountable issues. Reducing complex information systems into statistical models is a challenge for all fields (Round & Corbett 2020) and this does not uniquely invalidate phylogenetic methods in linguistics. These are, however, exactly the kinds of questions that linguistic phylogeneticists should consider. A lack of independence between phonotactic characters is a genuine limitation of the studies presented in the following chapters of this thesis. I return to this topic in the Discussion chapter.

Before moving on, I will make the point that, although the Comparative Method carries an unparalleled legacy of success in historical linguistics, no method is perfect and the Comparative Method has its own limitations and flaws (see Durie & Ross 1996). The neogrammarian assumption of the regularity of sound change (Osthoff & Brugmann 1878), which is a crucial assumption of the Comparative Method (Campbell 2004: p. 166), is an assumption about an evolutionary process that has been subject to long-running debate (Labov 2020) and apparent exceptions (see Chen & Wang 1975, Wang 1977, Durie & Ross 1996, Miceli & Round n.d.). Furthermore, the Comparative Method

can lack transparency. Analytical decisions during cognate identification can be opaque and difficult for others to evaluate without an extensive background knowledge in the languages being studied. There is presently no straightforward way to incorporate uncertainty into a manual analysis using the Comparative Method. Complete accuracy with regards to identifying borrowed lexicon cannot be guaranteed. In contrast, phylogenetic model choices can be declared explicitly for evaluation by others and statistical uncertainty in the results can be quantified. This is not to make the case that the Comparative Method is flawed or that computational phylogenetic methods are superior (I would argue they serve varied functions and can coexist, rather than stand as direct competitors), only that no method is free of assumptions, limitations or problematic aspects, and this should be kept in mind before dismissing any particular method outright.

To summarise this section, analogies have been drawn between linguistic evolution and biological evolution, in both directions, for practically as long as phylogenetic tree diagrams have existed. The early history of these academic fields in particular is somewhat intertwined. Although methods in biology and linguistics diverged substantially in the 20th century, phylogenetic thinking, in essence, never really left historical linguistics. The Comparative Method infers patterns of the fundamental evolutionary process of descent with modification and classifies languages in a phylogenetic structure. In recent decades, historical linguistics has taken a quantitative turn and embraced some of the computational phylogenetic methods pioneered by evolutionary biology. Linguistic phylogenetics is still a young subfield and there exists huge scope for methodological development, of which this thesis is a part. Debate continues over the validity of applying phylogenetic methods to linguistic data, partly due to mistaken associations with the tarnished legacy of lexicostatistics and glottochronology. But there are also valid questions around linguistic data structures and their application in evolutionary models. This discussion is important, and need not result in dichotomised 'for' and 'against' camps of thought. In the chapters that follow, I engage with some of these questions through an empirical evaluation of phonotactic data in phylogenetic methods. In this way, I follow the suggestion of Greenhill, Currie & Gray (2009: p. 2299), who recommend that rather than engaging in "armchair speculation" on the suitability of phylogenetic methods "on *a priori* grounds", it is more beneficial to engage with these methods and linguistic theory, crunch the numbers, and quantify how useful given methods are for testing various historical hypotheses.

2.2 Languages of Australia

This section gives a brief overview of the genealogical classifications and phonological systems of Australian languages, which make up the language samples in subsequent chapters of this thesis. This is followed in the next section with some contextual information on Sahul, the continent which includes Australia and the island of New Guinea. By doing so, I aim to place the studies in the following chapters in their broader continental context and foreshadow some points to be addressed in the Discussion chapter further on.

Somewhere between 250–300 (though perhaps more) distinct languages are understood to have

been present in Australia at the time of European colonisation in the 18th century, which can be subdivided further into around 700-800 language varieties when dialectal variation is considered (Koch & Nordlinger 2014). These linguistic communities are characterised by small populations, with a high degree of multilingualism facilitating communication between groups (Koch & Nordlinger 2014).

Colonisation has led to devastating shifts in the linguistic landscape. Of around 120 languages presently spoken in some capacity, only a dozen are estimated to be in a "relatively strong" position of sustainability with regards to childhood first language speakers, excluding new language varieties such as Kriol (Australian Government Department of Infrastructure, Transport, Regional Development and Communications et al. 2020: p. 42). The effects of colonisation and language loss on the affected communities is a deeply important and multifaceted topic demanding of its own dedicated discussion. I do not wish to give an unjustly inadequate summary of this subject here. For the purposes of this thesis, I make note only of the direct consequences of colonisation and language loss for studies such as this one. Fundamentally, the rapid loss of languages in Australia and lack of value placed on Indigenous languages by colonists has led to Australian languages being under-resourced (with regard to documentation, grammatical description, dictionaries, research) and under-represented in global cross-linguistic research. For many languages, existing resources are old and of variable quality (e.g. wordlists compiled by early settlers with no linguistic training). This limits what can be achieved with methods that demand high fidelity data and/or a high quantity of data. It can mean a relative lack of statistical representation in otherwise global typological datasets, which then affects our ability to make global scale generalisations about language. Consequently, overcoming these challenges is a priority for methodological development in large-scale comparative linguistics.

A broad understanding of the phylogenetics of Australian languages has been established through several decades of historical linguistic work using the Comparative Method. By far the largest family is Pama-Nyungan, which encompasses around two thirds of Australia's pre-colonial linguistic diversity and nearly 90% of its landmass (Bowern & Atkinson 2012: p. 817). The Pama-Nyungan family was first formally proposed by O'Grady, Voegelin & Voegelin (1966) as part of a continent-wide classification of Australian languages on lexicostatistical grounds. However, Hale (1964) generally is credited with coining the language family's name, which is a compound of words for 'person' in the southwestern and northeastern extremes of the continent (see also Wurm 1963). This work was preceded by various earlier, approximately analogous groupings identified on the basis of lexical similarities (Grey 1841: pp. 207–212, Moorhouse 1846: p. iv) and lexical cognacy (more systematically established than in earlier work) with a few phonological and grammatical features (Schmidt 1919, Kroeber 1923). Capell's (Capell 1937, 1940) primarily typological classification of Australian languages can be included in these earlier precedents too, if one interprets Capell's description of 'Common Australian' in the same way Koch (2004) does—that is, a proto-language ancestral to most but not all Australian language groups, and distinct from any hypothetical proto-Australian language, which Capell terms 'Original Australian' (Koch's interpretation seems valid, based on Capell's (1979) later work). Despite some skepticism on the validity of the Comparative Method in Australia (Dixon 2002, but see O'Grady & Hale 2004: for a counter-argument), there has been an abundance of work establishing subgroups

within Pama-Nyungan via the Comparative Method in the decades since O'Grady, Voegelin & Voegelin (1966), Bowern & Koch (see an overview and several examples in 2004). There are presently around 36 identified subgroups within Pama-Nyungan plus several isolates (following Bowern 2018b), however, not all of these have been established thoroughly with the Comparative Method and the precise number may differ depending on the analyst and the standard of evidence they accept. In recent times, computational phylogenetic methods have been applied to discern higher order groupings above the level of these subgroups, as well as an estimated age and an estimated geographic point of origin for the Pama-Nyungan family. Bowern & Atkinson (2012) presents computational phylogenetic evidence for four high-level divisions, clustering geographically into southeastern, northern, central and western groups. Bouckaert, Bowern & Atkinson (2018) recovers these same high-level groupings and further estimates a geographic point of origin in the Gulf of Carpentaria region, dated to the mid-Holocene period, between approximately 4,500–7,000 years ago.

The little over 10% of the remaining Australian mainland once Pama-Nyungan is accounted for is home to around 27 language families and isolates (Evans 2003). These are located in two discontinuous areas of Australia's north—one in the Kimberley region and one in the Top End. In addition, Tasmania is home to at least one and possibly more independent language families, though the precise number of languages present in Tasmania at the time of colonisation and their family structure is difficult to discern, owing to the particularly devastating effects of colonisation and paucity of linguistic documentation on the island (Bowern 2012). In the north, there is scope for discrepancies on the exact number of families depending on certain classification choices. For example, Evans' figure of 27 families includes an 'unclassified' group containing likely at least two family-level isolates (Evans 2003: p. 11). Other factors affecting this figure include the treatment of Daly languages, assumed to be a single family by Tryon (1974) but which have subsequently been argued to consist of up to four separate families (Evans 2003: ch. 4–7); the reclassification of several languages into an expanded Gunwinyguan family and possibly even a larger 'Arnhem' family (Evans 2003: p. 14–15); and the place of the Tangkic languages and possible neighbouring isolate Minkin (Evans 1990, Memmott et al. 2016), the Garrwan languages, and Yanyuwa, all of which have been argued variously as either within or outside of Pama-Nyungan (Evans 2003: p. 12). Whatever their precise classifications, it is clear that an exceptional degree of high order linguistic phylogenetic diversity is packed into a particularly small geographic area, in stark contrast to the geographic extensiveness of Pama-Nyungan. It has not been conclusively demonstrated via the Comparative Method how the various non-Pama-Nyungan families relate to one another nor how they relate to Pama-Nyungan. Nevertheless, curiously similar phonologies are present across Pama-Nyungan and non-Pama-Nyungan areas (discussed below in Section 2.2.1) and it is commonly assumed that all Australian language families group together under a putative 'proto-Australian' ancestor, to the exclusion of all other world languages (e.g.@dixon_languages_1980; Hamilton 1996). In addition, Harvey & Mailhammer (2017) argue for proto-Australian on the basis of evidence from nominal prefixes. Evans (2003) discusses some possibilities for family relationships underneath a proto-Australian root. Proposals include the following: A 'rake model', in which all 27 or so non-Pama-Nyungan families and Pama-

Nyungan descend equally from the same root with no intervening higher-order structure (O'Grady, Voegelin & Voegelin 1966). A 'binary model', in which proto-Australian splits into two sisters, proto-Pama-Nyungan and proto-non-Pama-Nyungan (Heath 1978a). This is approximately equivalent to Capell's (1956) typological division between prefixing and non-prefixing languages (though there are examples of prefixing Pama-Nyungan languages and non-prefixing non-Pama-Nyungan languages). A 'Pama-Nyungan offshoot model', in which Pama-Nyungan, Garrwan, Tangkic and Gunwinyguan families are contained beneath a higher order 'proto-macro-Pama-Nyungan' ancestor, and thereby more closely related to one another than to other non-Pama-Nyungan families (O'Grady 1979, Evans 1997, Evans & McConvell 1998). Several decades of progress notwithstanding, relating non-Pama-Nyungan languages to one another, to Pama-Nyungan, and to a putative proto-Australian root remains an active priority for historical linguistics research in Australia.

This thesis concentrates primarily on the Pama-Nyungan family. Chapters 5–6 use lexical data from 112 exclusively Pama-Nyungan language varieties. These varieties represent 26 of the 36 Pama-Nyungan subgroups and 3 isolates identified in Bowern (2018b). In Chapter 3, which does not require a pre-existing phylogenetic reference tree, an additional 54 languages are included representing all major non-Pama-Nyungan families.

2.2.1 Australian language phonologies

Australian language phonologies are typologically distinctive. Descriptions of the nature of language phonologies in Australia make reference to a uniquely constrained set of characteristics and the phonologies of individual Australian languages are frequently described in terms of a specific, narrow set of parameters of variation that follow from the broadly understood profile of what Australian phonologies look like (e.g. Goddard & Goddard 1985: p. 11, Evans 1995: p. 52, Breen & Barry J. Blake 2007: p. 7, Gaby 2006: p. 23). This apparent homogeneity of phonological systems across Australia stands out as unusual for such a geographically expansive and phylogenetically diverse area, when compared to the phonological diversity observed in other parts of the world of similar areal and genealogical scope (Maddieson 1984, Mielke 2008). The first instance of this apparently constrained nature of Australian phonological variation appears to be Schmidt (1919). Subsequent references include Capell (1956), Voegelin et al. (1963), Dixon (1980), Busby (1982), Hamilton (1996), Butcher (2006), Dixon (2002), and Baker (2014), among others. There is, however, some effort to push back against this consensus and illuminate Australian phonological variation to a degree that is perhaps underappreciated within the field. Bowern (2017) makes a case for treating generalised claims about "Standard Average Australian" language characteristics with caution, finding that many such characteristics described in the literature suffer from a lack of testability. In this light, Gasser & Bowern (2014) attempt to re-examine previous typological claims about Australian langauge phonologies in a more quantifiable way, using lexical data from 120 Australian language varieties. They concede the homogeneity of Australian segmental inventories, but demonstrate that there is considerably greater variation between languages with regards to the frequencies of phonological segments, rather than

simply the segmental inventories themselves. Most recently, Round (2021b) quantifies some of the key characteristics traditionally said to define Australian language phonologies within six major divisions of the Pama-Nyungan family plus 13 other Australian language families. He finds that five parameters of variation are sufficient to describe the consonantal phoneme inventories of around two thirds of the language sample. These parameters are:

- One versus two contrastive apical places of articulation.
- One versus two contrastive laminal places of articulation.
- The presence or absence of a contrastive glottal stop.
- Whether there is a single lateral phoneme, a maximal number of laterals equal to the number of contrastive coronal places of articulation, or some 'submaximal' number of laterals in between.
- One versus two contrastive series of plosives.

For the approximately two thirds of Australian languages that fit within these five parameters of variation, the common characteristics that define the rest of the phoneme inventory are as follows:

- A contrastive labial and dorsal place of articulation (which are often analysed as forming their own 'peripheral' natural class).
- A maximal set of contrastive nasals, one at each supralaryngealplace of articulation.
- A labio-velar and palatal glide.
- Two contrastive 'rhotic' phonemes—an alveolar tap/trill and a retroflex glide.
- No contrastive fricatives (notable in light of the cross-linguistic abundance of fricatives elsewhere in the world).

Quantification of these parameters reveals some interesting nuance to previous characterisations of Australian phonologies. For example, although Australian languages seem roughly evenly split on the first two points (one versus two apical places and one versus two laminal places) in raw terms, only the laminals parameter is split relatively evenly at the family level (or level of major subclades, in the case of Pama-Nyungan). Languages with a single apical series are present in relatively few language families but happen to be common among two of the largest subclades of Pama-Nyungan. Thus, although there appears to be a relatively even split between individual languages with one versus two apical series, Round (2021b) demonstrates that the lack of an apical distinction is relatively rare at the genealogical level. With regards to the other parameters of variation, glottal stops are restricted genealogically and in gross terms, being present only in a few relatively small language families and two Pama-Nyungan subgroups, all in the north of the country. The subject of plosive contrasts in Australian languages is more complex. Languages vary in the phonetic factors by which plosive contrasts are maintained, and this contrast is sometimes phonologically analysed as a singleton/geminate distinction rather than a voiced/voiceless distinction or some other fortis/lenis distinction (for a more detailed overview, see Round 2021b).

A variety of other segments are found in the remaining one third of Australian languages that do not fit this paradigm. Languages with contrastive fricatives are found in languages in the Cape York

and Daly River regions, the former of which have been analysed as historically deriving from stops (Dixon 1980: p. 199). Other exceptions are a small number of languages (around 2% for each) that are either missing a nasal phoneme at a supralaryngeal place of articulation or lack one of the rhotic tap/trill or glide phonemes (Round 2021b). In addition, various complex segment types have been argued to exist in a number of languages, not always with uniform agreement (see Round 2021b). A classic example is the series of labialised consonants proposed for Arandic languages (Wilkins 1989, Breen 2001).

With regards to vowels, Australian languages are most typically described as featuring a triangular three-vowel system, optionally with a length distinction. Other common variations are the addition of mid vowels and the addition of a non-low central vowel. Rarer exceptions are nasalised vowels in Anguthimri (Crowley 1981), contrastive rounding distinctions between vowels of identical height and frontness in particular Daly, Giimbiyu and Paman languages (Round 2021b), and phonemic diphthongs (more typically analysed as vowel-glide-vowel sequences in Australian languages) in Mbarrumbathama (Verstraete 2019).

Beyond segmental inventories, remarkable consistency among Australian languages has been described in phonetics (Dixon 1980) and phonotactics (Dixon 1980, Hamilton 1996, Baker 2014). Turning now to phonotactics, Hamilton (1996) describes an ostensibly considerable degree of variation between languages. However, he finds that this variation is "systematic and highly constrained" (Hamilton 1996: p. 29). Hamilton refers to this underlying similarity, in which the superficial phonotactic variation that does exist can be understood within the tightly prescribed bounds of a uniform, continent-wide system, as *concert*. Round (2021a) quantifies the phonotactic characteristics used to describe Australian languages and the systematic parameters of variation identified by Hamilton (1996), in a similar vein to his quantification of phonemic generalisations described above.

One notable feature of Australian phonotactic systems is that they cannot be described easily using the syllable as the fundamental unit of organisation, where phonotactic rules and constraints are localised to prosodic positions within the syllable (e.g. Itō 1988, Goldsmith 1990). Australian phonotactic systems are better served by a disyllabic structure as the fundamental unit of organisation and, resultingly, few Australian languages permit words of fewer than two syllables (Dixon 1980). Hamilton (1996: p. 75) presents two disyllabic templates, following Dixon (1980), presented in (2.1) and (2.2) below, where C_{init} is a word-initial consonant, C_{inter} is an intervocalic consonant, C_{fin} is an optional word-final consonant, C_1 is a pre-consonantal consonant and C_2 is a post-consonantal consonant.

$$C_{init}VC_{inter}V(C_{fin}) \tag{2.1}$$

$$C_{init}VC_1.C_2V(C_{fin}) \tag{2.2}$$

The consonantal slots in (2.1) and (2.2) are characterised by the following salient points, as described by Dixon (1980), Hamilton (1996) and Baker (2014).

- Of all consonant positions, C_{inter} is generally the only one in which all consonants may appear.
- Apicals are typically restricted from appearing in C_{init} position, and peripherals are preferred over laminals. Obstruents are preferred over sonorants for both C_{init} and C_2 positions.
- Conversely, sonorants are preferred over obstruents in C_{fin} and C_1 positions. Coronal consonants are also preferred in this position over peripherals in these positions.
- Many languages do not include the C_{fin} position at all, with words necessarily being vowel-final, e.g. Anindhilyakwa (Van Egmond 2012), Kayardild (Evans 1995), and Diyari (Austin 1981a).

Homorganic nasal+stop clusters are almost universal, although Thargari (Klokeid 1969) is an unusual exception, where homorganic nasal+stop sequences have developed into a series of fortis stop phonemes, so almost all the remaining nasal+stop clusters are heterorganic (Austin 1981a). One particularly distinctive feature of Australian language phonotactics is the commonality of a rich array of heterorganic nasal+stop clusters (Baker 2014, Fletcher & Butcher 2014, Round 2021a). These are typically restricted in frequency and type to a greater degree than their homorganic counterparts (Hamilton 1996: pp. 78–82), but are nevertheless noteworthy given the rarity and susceptibility of heterorganic nasal+stop clusters to assimilation in other parts of the world (Baker 2014: p. 144). Lateral+stop clusters are also common, and in this case, heterorganic combinations are more common while homorganic clusters are dispreferred or restricted all together.

In examining the overall typology of Australian language phonologies and phonotactics, one point to consider is the generally non-deterministic nature of phonological analysis 5, there are many criteria that are factored into the determination of a language's segmental inventory and, consequently, multiple possible solutions depending on the application and analytic order of a particular linguist. This has implications for the broad typology of Australian phonology and phonotactics, with particular regards to the impression of homogeneity in these systems. One such criterion for segmental-phonological analysis in an individual language is phonotactic parsimony. For example, by one analysis, Alawa features a contrastive series of prenaslised stop segments (Sharpe 1972) and, likewise, the same has been described for Tiwi (Lee 1987). The motivation for these analyses is that nasal+stop sequences would otherwise be found in the C_1 position of the disyllabic templates above. These so-called complex segments enable the preservation of a parsimonious phonotactic analysis and a phonotactic analysis that is in concert with fellow Australian languages. However, this is at the expense of creating expanded segmental inventories for Tiwi and Alawa that diverge from the standard average Australian characteristics described above. An alternative analysis of Alawa and Tiwi phonologies might place a lower value on phonotactic criteria and, for example, place a relatively greater emphasis on finding minimally contrastive pairs. Word-initial nasal+stop sequences could be analysed as separate segments, producing smaller, simpler segmental inventories that are more in keeping with the typological standard. This, however, would require acceptance of a greater degree of phonotactic variation in Australia and revision of disyllabic templates (2.1) and (2.2) if these templates are to be held as applicable Australia-wide. Round (2021a) gives an extended analysis of prenasalised stops and also prestopped laterals and prestopped nasals that have been proposed in various languages around the country via systematic assessment of the distribution of complex segments in each word position and comparison

to the permissibility of other comparable clusters in each language. He finds that, for the languages for which stop+lateral sequences have been analysed as unitary prestopped lateral segments, these sequences are licensed to appear in the same positions as other bisegmental clusters and, therefore, are equally consistent with being analysed as either single segments or bisegmental clusters on phonotactic grounds. The same is true for stop+nasal sequences, which have been analysed as prestopped nasals, in languages of the Thura-Yura and Karnic subgroups. However, several other Pama-Nyungan subgroups (Arandic, Paman, Kulin, plus Djinang of the Yolngu subgroup) permit stop+nasal sequences in positions in which single nasal segments are permitted but not other clusters, thus stop+nasal segments pattern phonotactically more like single segments in these languages. Together, these examples serve to illustrate the point that, although Australian phonologies and phonotactics are often held to be homogeneous to a considerable degree, various exceptions to the norm do exist and, furthermore, where and how variation manifests is, in part, a product of the analytic decisions of the individual typologist and/or the individual linguists who document languages. I take these issues into account explicitly in the comparative studies that follow in subsequent chapters.

2.3 Australian languages within the continent of Sahul

Sahul is the continent that prinicipally includes the Australian mainland, Tasmania and the island of New Guinea. These form a single landmass when sea levels are low, as was the case throughout most of Sahul's human history. The current separation of Tasmania and New Guinea from the Australian mainland is the result of rapidly rising sea levels around the beginning of the Holocene era. Specifically, this includes the inundation of the Arafura Sill (a land bridge connecting Arnhem Land to New Guinea) and Lake Carpentaria (an enormous fresh water body situated in the area of the present-day Gulf of Carpentaria) around 13 ka; the inundation of Bass Strait, separating Tasmania from the Australian mainland approximately 11 ka; and finally, the inundation of Torres Strait, separating Cape York from New Guinea approximately 8 ka (Reeves et al. 2008, Williams et al. 2018). These dates are of interest here in a couple of respects. Firstly, for at least five sixths of the history of occupation, Sahul was a single landmass (see below for discussion of human migration). Secondly, this trio of inundation events tantalisingly overlap with the maximal limit for linguistic reconstruction via the Comparative Method, which is commonly cited as around 10,000 years (Nichols 1997: p. 135).

The initial peopling of Sahul is an active and rapidly developing area of interest. The earliest date of human arrival in the continent has been revised substantially in recent years. The earliest archaeological evidence currently comes from the Madjedbebe rock shelter in Arnhem Land, dated to around 65 ka (Clarkson et al. 2017, see also Florin et al. 2020), which pushed back earlier estimates by 5,000–18,000 years at the time (O'Connell & Allen 2015, Clarkson et al. 2015). One of the challenges for dating the earliest migration to Sahul has been Australia's relatively fragmented archaeological record (Bradshaw et al. 2019) but, in addition to continuing archaeological work, the triangulation of evidence from ancient DNA, computational modelling and linguistic evidence is helping to clarify the picture. Another recent advance includes identification of Australia's first underwater archaeological

site (Benjamin et al. 2020). Genomic evidence generally points to an early intial colonisation of Sahul, consistent with archaeological evidence, and subsequent isolation from populations outside of Sahul in the tens of thousands of years that followed (Hudjashov et al. 2007, Nagle et al. 2017, Pedro et al. 2020). In addition, computational modelling has been used to simulate possible migration routes taken to reach Sahul from Sunda to the north. Bird et al. (2019) model historical coastlines and ocean drift and find that, even with low sea levels, humans would be unlikely to reach Sahul by random drift. Rather, Bird et al. (2019) conclude that the first migrants to Sahul likely had the capacity to undertake multi-day open sea voyages—remarkable, given the antiquity of such voyages and the implications for our understanding of technological capacity of early modern humans. Bradshaw et al. (2019) model the minimum founding population size that would be required for this initial migration and find that between 1,300 and 1,550 people would be required at a minimum. These individuals would either have to migrate all at once or in smaller voyages of at least 130 people successively over 700–900 years. There are two main proposed routes that these early seafarers might have taken to reach Sahul, proposed by Birdsell (1977) and debated ever since. These are a northern route into New Guinea via Sulawesi, or a southern route into north-western Australia via Timor and Bali. Bradshaw et al. (2019) determine that the northern route would be easier and therefore more likely to support a successful migration, requiring as few as three crossings, all of which could be made in 2–3 days (assuming good conditions) and all while maintaining sight of the destination island. Kealy, Louys & O'Connor (2018) similarly favours the northern route as the easier of the two, though, as Bradshaw et al. (2019) points out, more complex routes would not necessarily be beyond the first migrants depending on their level of seafaring skills and technology.

In a landmark study, Malaspinas et al. (2016) inferred more detailed migration movements within Sahul. They find an initial diversification of Papuans from Indigenous Australians at 25–40 ka, a common ancestor to all Indigenous Australians in their sample 10–32 ka, and, significantly, a population expansion out of Austrlia's north-east within the last 10,000 years, potentially consistent with the expansion of Pama-Nyungan languages. Malaspinas et al. (2016) compared their genomic evidence to a linguistic phylogeny inferred from lexical cognate data. Curiously, the topologies of these trees are highly congruent but apply on different timescales, the genomic tree being at least twice as old as the linguistic one. The explanation for this discrepancy in timing is uncertain. A subsequent study of mitochondrial DNA by Tobler et al. (2017) similarly finds evidence for an initial ancient migration followed by rapid spread around Australia following the coasts. However, unlike Malaspinas et al. (2016), Tobler et al. (2017) find that populations remained remarkably geographically static for the 50,000 years or so following initial migration and settlement. They dispute the evidence for a more recent Holocene-era expansion, saying the "genetic signal remains ambiguous at best" (Tobler et al. 2017: p. 183).

This discussion draws back importantly to the discussion of the Pama-Nyungan family of languages in Section 2.2 above. The unresolved question of whether or not genetic evidence favours a rapid Holocene expansion relates to the linguistic question of how the Pama-Nyungan family came to dominate so much of the Australian conteninent. Note that there is no *a priori* reason why a rapid

linguistic expansion should necessarily be accompanied by a rapid genetic expansion as well, but this would have implications for the method of linguistic expansion. It could be the case that some kind of social context arises in which there is rapid language shift but speakers nevertheless remain in their ancestral homelands, as opposed to, say, a conquest situation that would leave a genetic signature. In the case of Pama-Nyungan, several theories have been put forward to explain its expansion, such as a rapid mid-Holocene spread of small tool technology and/or social/ceremonial restructuring (Evans & McConvell 1998), or expansion from refugia from the Last Glacial Maximum as climactic conditions became more favourable in the early Holocene (Bouckaert, Bowern & Atkinson 2018). Bouckaert, Bowern & Atkinson (2018) conducts a phylogeographic study using lexical cognate data and finds support for a geographic point of origin around the Gulf of Carpentaria. This point of origin is consistent with the Holocene expansion identified by Malaspinas et al. (2016), but, once again, the inferred age of the linguistic phylogeny is much younger than the age inferred from genomic evidence in Malaspinas et al. (2016). The age of Pama-Nyungan inferred by Bouckaert, Bowern & Atkinson (2018) is more consistent with a rapid mid-Holocene expansion faciliated through technological and cultural advantages.

Taking a step back, the linguistic prehistory of Sahul remains enigmatic in historical linguistics. The question of if and how Australian languages might relate to their Papuan neighbours to the north has been the subject of speculation for many years (e.g. O'Grady, Voegelin & Voegelin 1966, Wurm 1975). But, besides a handful of tentative cognates (Foley 1986) and shared structural features (Nichols 1997, Reesink, Singer & Dunn 2009), conclusive evidence for a connection remains elusive. Of course, there has been much successful genealogical classification within Sahul—many small families plus the very large Pama-Nyungan family have been well established in Australia and, similarly, many small families plus the very large Trans-New Guinea family have been identified in New Guinea. Nevertheless, the diagnosis of deeper historical relations is handicapped by several factors in this part of the world. These limitations include a lack of adequate description (Bowern & Atkinson 2012), an absence of pre-colonial written sources (Foley 1986), apparent homogeneity of phonological systems (Baker 2014, Round 2021b,a) and millennia of extensive horizontal diffusion (Foley 1986, Dixon 2002) (although the extent to which this horizontal diffusion exists may be disputed. c.f. Bowern et al. 2011). Even in ideal circumstances, the maximal time-depth of the Comparative Method is usually assumed to be around 10,000 years, as noted above. Although this overlaps with the land bridge between Cape York and New Guinea, Arnhem Land had already been separated by the Gulf of Carpentaria and Tasmania had already been separated from the Australian mainland by 10 ka.

Recent methodological advances, as discussed in Section 2.1, give cause for optimism that we may be able to extend the time-depth at which historical linguistics can operate and unravel something of the prehistory of Sahul. As for data, and the possible erosion of historical signal after 10,000 years in lexical data specifically, two observations motivate the approach of the present study: Firstly, linguistics has well and truly entered 'the age of big data' and computational methods now enable us to extract minute threads of significance from large volumes of data. This enables us to compare more data points across wider groups of languages quickly and efficiently, and identify trends and

correlations which would otherwise escape the attention of even the most highly trained human eye. This thesis evaluates the profitability of complementing lexical data (which continues to form the backbone of most historical linguistic work) with other kinds of data too—namely, phonotactics.

Amid present uncertainty and conflicting findings, it seems clear that linguistics has an important place within the wider science of human history. It offers a unique perspective and level of granularity within a particular time-depth that is not necessarily captured by genomic evidence alone. As seen in the case of the expansion of Pama-Nyungan, the combination of linguistic evidence with genomic evidence offers the potential to infer not just who was where and when, but also to give a more nuanced view of why, since the combination of evidence can indicate the method of expansion in a way that either piece of evidence in isolation could not. Continuing advances in both fields will help to resolve current ambiguities and the future of this multidisciplinary endeavour looks bright.

2.4 Conclusion

This chapter began with a walk through the early history of historical linguistics as a discipline, and the development of phylogenetic thinking within. The phylogenetic tree, a graphical devices for charting past relationships of shared descent with modification, shares its early origin in linguistics as well as biology. Methodologically, both fields divided substantially in the 20th century, with the Comparative Method the enduring gold standard in historical linguistics. Meanwhile, advances in genetics led biologists to embrace quantitative historical methods. In recent times, computational phylogenetics has taken an increasingly mainstream place within historical linguistics. Section 2.1 listed numerous examples of linguistic phylogenies inferred through computational phylogenetic methods. Beyond tree inference, however, Section 2.3 shows how linguistic phylogenetics is playing an increasingly valuable role in multidisciplinary investigations of deep-time human history.

Section 2.2 and Section 2.2.1 give an overview of the Australian languages that are the focus of this thesis and their phonologies. The most numerically and geographically expansive family, Pama-Nyungan, is of particular interest since it is the basis of the language samples in Chapters 3–6. This ties into the discussion of Sahul's human history in Section 2.3, since the precise nature and timing of Pama-Nyungan's dramatic expansion across the Australian mainland remains uncertain and subject to active investigation. Continuing linguistic resesarch will be essential to advancing our understanding on this topic. Linguists should not fall into the self-depreciating assumption that geneticists will 'come in and solve everything'.[4] Firstly, as seen in Section 2.3, ancient DNA research produces uncertainty and conflicting evidence of its own and has its own limitations (e.g. expense and difficulty of data collection). Moreover, the linguistic record will have its own story to tell, which may or may not be correlated with other sources of evidence, offering nuanced insights into human history.

The goal of this thesis is to contribute to the methodological development of linguistic phylogenetics by investigating a novel source of data—quantitative phonotactic information extracted from Australian language wordlists. In doing so, however, I aim not simply to analogise between biological evolution

[4]This may be an obvious point to some but I will admit it was not immediately obvious to me.

and linguistic evolution but to develop a model of phylogenetic phonotactics critically, and where necessary from the ground up, using presently available quantitative tools. The first step in this process is a re-evaluation of individual phoneme frequencies in Australian languages, which follows in the next chapter.

The following publication has been incorporated as Chapter 3:

Jayden L. Macklin-Cordes & Erich R. Round. 2020a. Re-evaluating phoneme frequencies. *Frontiers in Psychology* 11. https://doi.org/10.3389/fpsyg.2020.570895

Contributor	Statement of contribution	%
Jayden Macklin-Cordes	initial concept	100
	scripting, analysis	80
	theoretical derivations	60
	writing of text	75
	preparation of figures	90
	proof-reading	20
Erich Round	supplying data	100
	scripting, analysis	20
	theoretical derivations	40
	supervision, guidance	100
	writing of text	25
	preparation of figures	10
	proof-reading	80

Data is sourced from the Ausphon Lexicon database, created and maintained by Erich Round.

Macklin-Cordes came up with the initial concept for the study, wrote scripts and performed all the initial analysis. He wrote and subsequently revised the initial draft of the paper. Macklin-Cordes also wrote the phoneme frequency viewer app contained in the supplementary materials.

Round provided guidance throughout. He revised significant portions of the text and added to the theoretical framing and contribution of the paper. He also revised scripts for efficiency and extended likelihood ratio testing to include x_{min} parameters (described in 3.3.7).

Chapter 3

Re-evaluating Phoneme Frequencies

Causal processes can give rise to distinctive distributions in the linguistic variables that they affect. Consequently, a secure understanding of a variable's distribution can hold a key to understanding the forces that have causally shaped it. A storied distribution in linguistics has been Zipf's law, a kind of power law. In the wake of a major debate in the sciences around power-law hypotheses and the unreliability of earlier methods of evaluating them, here we re-evaluate the distributions claimed to characterise phoneme frequencies. We infer the fit of power laws and three alternative distributions to 166 Australian languages, using a maximum likelihood framework. We find evidence supporting earlier results, but also nuancing them and increasing our understanding of them. Most notably, phonemic inventories appear to have a Zipfian-like frequency structure among their most-frequent members (though perhaps also a lognormal structure) but a geometric (or exponential) structure among the least-frequent. We compare these new insights the kinds of causal processes that affect the evolution of phonemic inventories over time, and identify a potential account for why, despite there being an important role for phonetic substance in phonemic change, we could still expect inventories with highly diverse phonetic content to share similar distributions of phoneme frequencies. We conclude with priorities for future work in this promising program of research.

3.1 Introduction

Linguistic theorists seek to reveal causal mechanisms which explain the observable diversity of human language. Good causal hypotheses are often suggested by the mathematical distribution that a linguistic variable is described by, owing to the fact that the distribution can be understood as an emergent outcome of some underlying causal process, and that a given mathematical distribution will be consistent with only certain mathematical kinds of underlying processes. Consequently, it is important for the development of theory that proposed claims about distributions be as sound as possible. For instance, one of the most famous distributions in linguistics is the Zipfian distribution, which technically speaking is a kind of power law. Recently, however, the evaluation of putative power laws across the sciences have come under intense scrutiny and often been found wanting. In response,

methodologists have developed more rigorous and secure methods for diagnosing power laws and for distinguishing them from similar but significantly different distributions. For linguists, this creates an opportunity, to re-examine our own putative power law distributions, and by doing so to improve the pathway to sound explanatory theorising.

Here, we re-evaluate the status of mathematical distributions for characterising phoneme frequencies. Previous studies have proposed that phoneme frequencies follow a particular member of the power law family, the Yule-Simon distribution (Martindale et al. 1996, Tambovtsev & Martindale 2007). But in the wake of a recent, major debate across the sciences regarding power laws, a reconsideration of this earlier research is timely. In this paper, we apply state-of-the-art maximum likelihood methods for the detection and assessment of power laws, to derive a better understanding of the distributions that do and do not describe phoneme frequencies well. Our results clear the way for more informed research into the ultimate, processual causes behind the frequency patterns of phonemes in human language.

Our focus here is on distributions and their potential for shedding light on language. Distributions are properties of variables. A variable can be defined as the set of values that characterise something, be it a *sample* (e.g. a set of languages), or a *real population* that the sample is drawn from (e.g. the set of all current languages), or even an *idealised population* which the real population is believed to approximate (e.g. the set of all possible languages). Often, the ultimate object of scientific interest is an idealised population, and thus its distribution. In empirical work, we cannot directly access this ultimate object of interest, and so we rely on real populations, or very often, a sample. Consequently, although we may have direct access only to the literal distribution of a sample, with its many idiosyncrasies, we tend to be more concerned with an overall pattern which we believe it approximates, one which often is elegantly characterised by a distribution which mathematically is relatively simple.

With these motivations in mind, how can an investigator decide that a certain distribution characterises a variable satisfactorily? One method is visual inspection. This typically involves observing a close match between two histogram plots: one of the data and one of the candidate distribution, or plotting a regression of the data against the candidate distribution. In work on phoneme frequencies, visual inspection has been the primary method of assessing candidate distributions. A more rigorous alternative is to apply quantitative, statistical tests to evaluate how well a particular distribution model (such as the normal distribution) fits a data sample. The purpose of these tests is not to prove definitively that some variable follows a particular distribution, but rather to quantify the degree to which a sample's distribution is consistent with its having been drawn from a population of a particular distribution.[1] Some of these tests will be the subject of the sections that follow.

A strong motivation for testing the consistency of observed data with a particular distribution, is that many distributions can be described as the *outcome* of certain kinds of processes (Frank 2014). Thus there is a direct link between the quality of our evaluation of distributions, and the reasonableness

[1] Within frequentist statistics, there are tests which test against the null hypothesis that data are drawn from a particular distribution (Shapiro-Wilk, Kolmogorov-Smirnov, Pearson's chi-squared test, to name a few). Within a Bayesian framework (Spiegelhalter 1980, Farrell & Rogers-Stewart 2006), the approach is to calculate the likelihood of observing the data given a distribution and a set of the distribution's parameters. Bayes factors can be computed to compare the relative likelihoods of observing the data given competing kinds of distributions and parameter sets.

of the causal, explanatory hypotheses we subsequently entertain. Consider for example a so-called *preferential attachment process*, also known as a Yule process or rich-get-richer process. This can be imagined as having a set of urns, into which balls are added one at a time. Specifically, the urn to which a new ball is added is selected with a probability proportional to the number of balls already in the urn. This simple process has an interesting outcome. Initially, each urn is equally likely to be selected, but the distribution will soon skew, as urns with more balls accumulate additional balls faster than the others. If we *rank* the urns in terms of which has the most balls, then with time, the relationship between an urn's rank and how many balls it contains will come to obey a *power law*. A power law is a mathematical a relationship between two quantities where one varies as a power of the other. Consequently, if a variable can be shown to be consistent with a power law distribution, then this is consistent with there existing a preferential attachment process as the causal mechanism underlying the behaviour of variable. Yule (1925, see also 2011) made this connection a century ago. Yule showed that among flowering plants the level of species richness within a genus follows a power law, and linked that observation to a preferential attachment mechanism.

Since Yule's first demonstration of the link between power law distributions and preferential attachment processes, power laws have been used to characterise the distributions of a diverse array of phenomena in the natural and physical world and in human society (Clauset, Shalizi & Newman 2009: p. 661). City populations (Gabaix 1999, Levy 2009, Malevergne, Pisarenko & Sornette 2011), authorship of scientific publications (Simon 1955), income distribution (Simon 1955), the superstar phenomenon in the music industry (Chung & Cox 1994) and the network topology of the Internet (Faloutsos, Faloutsos & Faloutsos 1999) are but a few of the phenomena for which power laws have been proposed (see Newman (2005), pp. 327–329 for further examples). And in linguistics, the Zipfian distribution has been used to characterise word frequencies in text corpora (Estoup 1916, Zipf 1932, 1949).

Given the apparent pervasiveness of power laws in a diverse range of unrelated contexts, it is little surprise that there is a rich vein of literature dedicated to evaluating power laws (Clauset, Shalizi & Newman 2009: p. 662) as well as a century of theorising on the mechanistic processes by which they arise (see Newman 2005: pp. 336–348, Mitzenmacher 2004: pp. 230–243). However, verifying the presence of a power law is not a straightforward task (Stumpf & Porter 2012: p. 666), and validation of earlier power law proposals using increasingly robust and powerful statistical methods is an active line of inquiry across many fields of science (Malevergne, Pisarenko & Sornette 2011).[2]

The traditional approach to power law validation, following Pareto's (1897) work on wealth distribution, was visual inspection. When visually inspecting a histogram plotted on log-log scales, a straight line would suggest the presence of a power law. The defining shape parameter (see below), α, could then be obtained by calculating the slope of the straight line using standard linear regression

[2]The question of whether certain phenomena are characterised best by a power law model or some other distribution can be contentious. See, for example, the debate between Eeckhout (2004) and Levy (2009) on the distribution of city population sizes (the former favouring a lognormal model, the latter favouring a power law). Another example concerns the distribution of computer file sizes, where Barford & Crovella (1998) and Barford et al. (1999) argue in favour of a power law model and Downey (2001) argues in favour of a lognormal model.

(Clauset, Shalizi & Newman 2009: p. 665, Urzúa 2011: p. 254) and the R^2 statistic could give an indication of the goodness of fit of the model. However, it has since been demonstrated that this traditional approach can be systematically unreliable (Clauset, Shalizi & Newman 2009: p. 665). A key assumption when estimating the standard error of the slope (shaded in grey) is that noise in the data is normally distributed, however, this is not the case for the logarithms of frequency data (Clauset, Shalizi & Newman 2009: p. 691). Further, the R^2 statistic commonly used to validate the presence of a power law (including by Tambovtsev & Martindale 2007), has low statistical power. That is, it often fails to distinguish between data truly drawn from a power law distribution and data drawn from other distribution types. This unreliability becomes particularly acute when there is a small number of observations, since the ability to distinguish a power law distribution from other similar distributions, including the log-normal, using R^2 is reduced (Clauset, Shalizi & Newman 2009: p. 691).

To remedy the shortcomings of earlier methods, Clauset, Shalizi & Newman (2009) developed power law validation procedures within a more rigorous, maximum likelihood framework. These procedures have since been adopted widely in the literature (for example, Touboul & Destexhe 2010, Cho, Myers & Leskovec 2011, Brzezinski 2014, and Lee & Kim 2018), but have not previously been applied to phoneme frequencies.

Some brief mathematical preliminaries will be useful here. When we refer to distributions, we are referring to mathematically defined *functions*, that relate one quantity to another. Those functions may in addition have *free parameters* which can be varied in order to produce a family of closely related distributions. A power law is a relationship between two quantities where one varies as a fixed power of the other, for example $y = x^3$, or $y = x^{-2}$ (which can also be written $y = 1/x^2$). For present purposes, where we will not be concerned with negative quantities or zeros, we will use a more narrow definition by Clauset, Shalizi & Newman (2009: p. 662), who define a power law as a relationship in which a quantity, x, is drawn from the distribution defined in Equation (3.1), where the free parameter α is greater than zero and the variable x likewise is greater than zero.[3] (The symbol '$\propto$' means 'is proportional to'.) For example, x might denote items' frequencies, while $p(x)$ is the probability that a given item has a frequency of x.

$$p(x) \propto \frac{1}{x^\alpha}, \quad x > 0 \tag{3.1}$$

In practice, Clauset, Shalizi & Newman (2009: p. 662) observe that the exponent (or 'scaling parameter'), α, typically, though not exclusively, falls in the range $2 < \alpha < 3$. They also observe that, in practice, many phenomena will not actually obey a power law for all values of x. Rather, the power law will apply to values only above some minimum threshold value, x_{min}. For example, in frequency data, it may be that only items whose frequencies meet or exceed a lower threshold will follow a power law. More generally, power law distributions come in a variety of specific forms, with different numbers of free parameters. We detail some of these in greater depth in the Discussion section below.

[3]For the area under the distribution curve to integrate properly to 1, the power function $1/x^\alpha$ must be multiplied by a normalisation constant (denoted C in the probability density function $p(x) = C/x^\alpha$). The normalisation constant will be calculated differently depending on the value of α and whether x is continuous or discrete. Clauset, Shalizi & Newman (2009: p. 664) give some examples.

A distinction can be made between power laws that apply to continuous variables and those that apply to discrete ones. Frequency data, including the phoneme frequencies used in this study, are typically discrete. Zipf's Law (3.2) applies to a discrete number of n observations whose values, x, are ranked by descending magnitude $x_1 \geq x_2 \geq \ldots \geq x_n$. For example, x may be the token frequency of n types, with x_k the frequency of the kth-ranked type. In (3.2), the quantity $p(x_k)$ is the relative frequency of the kth-ranked type (i.e., its frequency scaled such that the n relative frequencies sum to 1).[4]

$$p(x_k) \propto \frac{1}{k^\alpha}, \quad 1 \leq k \leq n \tag{3.2}$$

There is a long history of studying power laws in linguistics, however, the evaluation of statistical support for a power law relationship is far from straightforward and remains topical across a wide range of scientific fields (Stumpf & Porter 2012). Although several different models have been compared for their goodness-of-fit to the frequencies of phonological segments (Martindale et al. 1996, Tambovtsev & Martindale 2007), the method used to measure fit (using the R^2 statistic) has been shown to be systematically unreliable (Clauset, Shalizi & Newman 2009). Accordingly, the methodological limitations of previous studies and the renewed, general scientific interest in power law phenomena motivate the re-evaluation of a power law model with respect to phoneme frequencies. Our goal here is to verify the presence or absence of power law behaviour in the frequency distributions of phonological segments in the lexicons of Australian languages. It is, to the best of our knowledge, the first attempt to validate a power law model for phonological segments using a maximum likelihood framework as suggested by Clauset, Shalizi & Newman (2009).

3.2 Materials and methods

3.2.1 Data

As our data, we take phoneme frequencies in the lexicons of 166 language varieties of Australia, covering the 19 of Australia's language families for which phonemic lexical data is available, including all major branches of the Pama-Nyungan family which dominates the continent.[5] Our choice of dataset brings advantages and limitations. Prior studies of phoneme frequency distributions have overwhelming focused on languages of Eurasia, and it is valuable to confirm whether similar results emerge on other continents. Because Australian languages are known to have quite similar phonemic inventories across the continent (Capell 1956, Dixon 1980, Busby 1982, Hamilton 1996, Baker 2014, Round 2019b, 2021b), it is useful to confirm whether, even under conditions of highly constrained variation in the substance of the phoneme inventories, we still find recognisable variation in the distribution of frequencies of the phonemes. There are reasons expect this will be the case. Though phoneme

[4] This is equivalent to the probability that a token selected at random belongs to the kth-ranked type.

[5] The dataset closely approximates an exhaustive sample, containing around 80% of Australian languages for which phonemic lexical data is known to exist, and on the order of 40% of the varieties that were spoken at the onset of European colonisation.

inventories in Australia are similar, the phonemes themselves are known to exhibit considerable variation in their frequency distributions (Gasser & Bowern 2014, Macklin-Cordes & Round 2015). Likewise, phonemic bigram frequencies in the large Pama-Nyungan family exhibit diversity with a strong phylogenetic signal (Macklin-Cordes, Bowern & Round 2021), suggesting that variations in Australian phonological frequencies have evolved over a deep time span. The main limitations of examining an Australian dataset is that Australian languages will lack many phoneme types (such as ejective stops and front rounded vowels) that may be found elsewhere in the world, and may have different characteristic frequencies of some individual phonemes (such as high frequency velar nasals, or low frequency long vowels); on the other hand, this is equally true of the Eurasian languages which have been the dominant basis of prior research. On balance, we find it worthwhile to increase the scope of knowledge to a new continent, and as the results will show, Australian phoneme inventories behave in ways that are clearly recognisable from Eurasia, suggesting that there are important, fundamental dynamics at play in the shaping of phoneme system that transcend the specifics of which phonemes are involved.

The phonemic frequencies in this study are extracted from wordlists. A consequence of this is that in our data, all lexemes are counted just once and so are weighted equally.[6] This differs from earlier work which has studied the frequencies of phonemes in discourse. In discourse, lexemes have unequal frequencies and so are weighted unequally. Indeed, since lexemes are known to have a Zipfian frequency distribution in discourse, one might wonder whether our lexical phoneme frequency data, which is not affected by these discourse effects, should correspondingly be expected to follow a different distribution – perhaps a less Zipfian one. While the concern is understandable (we held it ourselves initially), the reality is that mathematically, the expected difference between phonemes' lexical and discourse frequencies *on average* is zero (see Section S1, Supplementary Materials for a more extended discussion). So, although lexical and discourse frequencies will differ, their distributions will not differ systematically purely as a result of word frequencies being distributed in any certain way in discourse. Having said that, there could still be systematic differences due to other factors. For instance, if words with high discourse frequency were especially likely to contain the language's least-frequent phonemes, this could give rise to systematically different distributions; note however, that this effect would not be due merely to words having unequal frequencies in discourse, but be due to an additional factor, which links words' discourse frequencies directly and exceptionally to phoneme frequencies. While we cannot rule out the possibility of such additional effects, it remains true that the neutral expectation is that lexical and discourse phonemes frequency will be distributed similarly. Moreover, our study reveals results that are consistent with this neutral expectation. Our confirmation that this similarity in distributions is expected to be the case, and actually is the case, sets up a useful context for future studies also. For most languages of the world, discourse frequency data is not yet available, but wordlists are. As research into frequency distributions extends in scope to cover more of the world's languages, it is valuable to have clarified how results based on lexical

[6]There is a significant body of research suggesting that frequencies defined in this manner are implicitly accessible to speakers and thus psychologically real (for example, Coleman & Pierrehumbert 1997, Zuraw 2000, Ernestus & Baayen 2003, Albright & Hayes 2003, Eddington 2004, Hayes & Londe 2006).

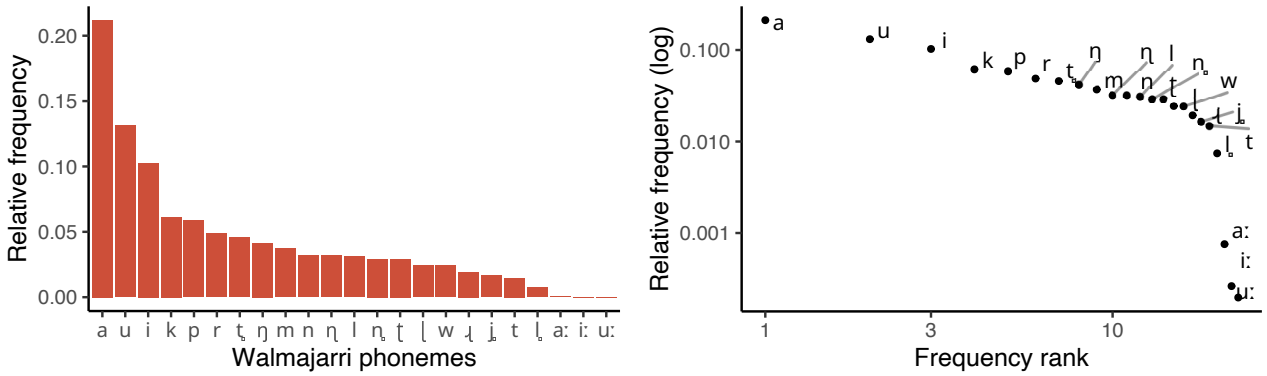

Figure 3.1: Frequency of phonemes in Walmajarri lexicon (Hudson & Richards 1993). Plot A, left, displays relative frequencies of each segment type. Plot B, right, shows the same frequencies on log-transformed *x* and *y* axes—the traditional visual device used to identify power laws.

frequencies and discourse frequencies relate to one another.

Our data comes from the Ausphon-Lexicon database, under development by the second author (Round 2017c). Ausphon-Lexicon extends the Chirila resources for Australian languages (Bowern 2016). It adds additional varieties and applies extensive data scrubbing, manual and automatic error-checking, and phonemic conversion using language-specific orthography profiles (Moran & Cysouw 2018). A standing challenge for typological phonemic research is the long-recognised fact that phonemic analysis itself is non-deterministic (Chao 1934, Hockett 1963, Hyman 2008, Dresher 2009). Presented with identical sets of language data, two linguists may produce differing phonological analyses, not due to any error on the part of the linguist but due to differing applications of the multitude of criteria by which decisions are made during the analysis of a phonemic system. As a consequence, cross-linguistic phonological variation can be attributed not only to language facts, but also to variation in linguistic practice. In cross-linguistic research, it is desirable for information to be represented in a comparable way throughout a dataset, and so recent phonological literature has emphasised the value of *normalising* source descriptions prior to cross-linguistic analysis (Lass 1984, Hyman 2008, Van der Hulst 2017, Round 2017b, Kiparsky 2018). Phonemic representations in Ausphon-Lexicon are normalised in this sense. Section S3, Supplementary Materials, details the normalisations applied, together with bibliographic details of original data sources.

To illustrate an example of the phoneme frequencies in our sample, Figure 3.1 plots the frequencies of phonological segments in the Walmajarri lexicon (Hudson & Richards 1993). Equivalent plots for every language in our sample can be viewed through an interactive visualisation app that we provide in Section S6 of the Supplementary Materials.

Phonological frequency data differs in some respects from the data types most commonly encountered in scientific power law studies, such as word frequencies or city populations. Typically, in order to understand a population (and some property of it), such as the cities in the United States (and their sizes), or the words of English (and their frequencies), it is impractical to examine every last member of the population, and so the study will examine a sample. Ensuring that a sample is of sufficient size is an important consideration, firstly in order to adequately represent the population and additionally

because a sufficiently large sample size is an important requirement in maximum likelihood estimation (Barndorff-Nielsen & Cox 1994, Newman 2005). In contrast, the phonemic inventory of any language is relatively small, and it is entirely feasible to examine exhaustive populations of phonemes.[7] An advantage of this is that the sample is highly representative of the population, but a disadvantage is that the number of observations is small and cannot be increased.

Given a sample of phonemes, we require an estimate, or measurement, of their frequencies. Measurement error is a potential concern in this study. Our segment frequencies are calculated from documented wordlists, which necessarily are limited representations of the complete vocabulary of the languages that the wordlists represent. One concern is that the particular morphology of a language's citation forms may cause certain segments in the language to be overrepresented in a wordlist which contains only citation forms. This would represent a bias, that is, a factor that pushes observations in a certain direction. We have attempted to control for this, by removing identifiable citation-form tense morphology from verbal words and noun-class prefixes from nominals. Another source of concern is that wordlists with a smaller number of words will necessarily entail a greater level of uncertainty in the observed segment frequencies. This will be a source of noise in the data. It does not push observed frequencies in any particular direction, but makes them generally less accurate. To address this, in our study, we restrict the language sample to language varieties with a minimum wordlist size of 250 lexical items. We selected 250 lexical items as a cut-off on the basis of Dockum & Bowern (2019), who investigate the effect of wordlist size on phonological segment frequencies. Dockum & Bowern (2019) report accelerating losses in the fidelity of segment frequency estimates as a wordlist drops below 250 items. While more words will always yield better frequency estimates, we select a minimum of 250 as a reasonable compromise. This gives us a sample of 166 Australian language varieties. Wordlist sizes range from 268 to 8742 (median 1072, mean 1438).

3.2.2 Statistical framework

We test for the presence or absence of a power law in the distributions of phonological segments following the maximum likelihood framework described by Clauset, Shalizi & Newman (2009). In brief, Clauset et al.'s (2009: p. 663) proposed procedure consists of three steps:

1. Estimate the parameters x_{min} and α of the power law model using the maximum likelihood method (Barndorff-Nielsen & Cox 1994, Newman 2005)[8].

2. Calculate the goodness-of-fit between the data and the power law using the Kolmogorov–Smirnov (KS) statistic, where a larger value corresponds to a worse fit. Using a Monte Carlo procedure,

[7] The probability that we have failed to observe some phoneme that exists in a language is small, though non-zero. In Section S5, Supplementary Materials we evaluate whether this is the case in our data. We find only three languages where it is. Even in such cases, the missing segment inevitably will be an especially low frequency type, and is unlikely to dramatically alter the overall frequency distribution of segments in the language.

[8] Maximum likelihood estimation (MLE) is a method for estimating the parameters in a statistical model, given some set of observations by finding the set of parameter values, $\hat{\theta}$, that maximise a likelihood function, $P(x \mid \hat{\theta})$, where x is a set of observations. In our case, the parameters, $\hat{\theta}$, to be estimated are those which define a particular distribution—for example, α and (optionally) x_{min} in a power law model.

a *bootstrapped p value* is calculated[9], and used to evaluate the plausibility of the power law. Namely, if this p value falls below a plausibility threshold of 0.1, the power law model is rejected.[10] Otherwise, the power law model remains an initially plausible hypothesis, and we proceed to step 3.

3. Compare the power law model with a set of models representing alternative hypotheses. For each alternative model, a bootstrapped p value is calculated as in steps 1 and 2 above. A likelihood ratio test is performed, comparing the fit of the alternatives with those of the power law model. If the calculated likelihood ratio is significantly different from zero, this indicates a significant difference in plausibility, and its sign (positive or negative) indicates which model is favoured (Clauset, Shalizi & Newman 2009: p. 680).

We use the *poweRlaw* package (Gillespie 2014) in *R* (R Core Team 2017a) to infer all maximum likelihood estimates and conduct bootstrapping to derive p values. We run 10,000 bootstrap iterations per language, per distribution type.[11]

As a brief point of comparison to prior work, we return to the Walmajarri example and plot the linear relationship between phoneme frequencies and rank on a log-log plot. Tambovtsev & Martindale (2007) find that a Zipfian distribution consistently underestimates the frequency of both high- and low-ranking segments while overestimating the frequency of those in the middle. The dashed black slope on Figure 3.2 shows a similar pattern. However, when the five lowest-frequency segments (i.e., those with the greatest statistical rank) are removed from the equation, the linear model fits much better (solid blue line). This is consistent with the observation by Clauset, Shalizi & Newman (2009) that, in practice, power laws are rarely observed across the whole distribution—rather, there is a threshold, the x_{min} parameter, below which the power law ceases to apply. Visual inspection of other languages in the dataset indicates that Walmajarri's pattern of phoneme frequencies is common, although there is a good deal of variation (and, consequently, variation in the fit of a linear model). However, given the known limitations of applying a linear model to a log-log plot, we now turn to more reliable methods for validating the presence of a power law, using the maximum likelihood method outlined above.

[9]This is a well-established statistical technique. A large number of simulated datasets are created, with data points drawn from the model power law distribution hypothesised in step 1. Each is then fitted to its own power law model and a KS statistic is calculated for the simulated dataset, relative to this model. The p value is defined as the fraction of these simulated KS distances larger than the actual, observed KS distance.

[10]Here we follow the method of Clauset, Shalizi, and Newman (2009), who suggest a threshold of 0.1. Note though, that even when p>0.1, we still do not necessarily accept that the power law is a good fit, rather there is a further round of evaluation (step 3). This use of a 'p value' differs from the more common use case where a null hypothesis is rejected when the p value is above a certain level. The reason for the difference lies in how the hypothesis of interest is related to the null hypothesis. Commonly, the hypothesis of interest is set up as the alternative hypothesis, and low p-values are required to reject the null hypothesis (not of interest). Here, the hypothesis of interest (power law is plausible) is set up as the null hypothesis. Accordingly, it too is rejected when the p-value is low. By allowing it to be rejected all the way up to 0.1 (rather than 0.05, for example), we are setting the bar relatively high. This approach may seem counterintuitive in the context of testing a single distribution hypothesis (where it might seem better to make the distribution of interest deliberately harder to accept than to reject). But in the context of testing which distribution fits the data best among multiple alternatives, it makes sense to make it deliberately hard to reject any particular distribution type.

[11]We find that 10,000 iterations are sufficient to obtain stable parameter estimates. Beyond 10,000 iterations, estimates will continue to fluctuate but in a tightly prescribed range. Plots of all bootstrapping runs can be viewed in the interactive visualisation app provided in Section S6.

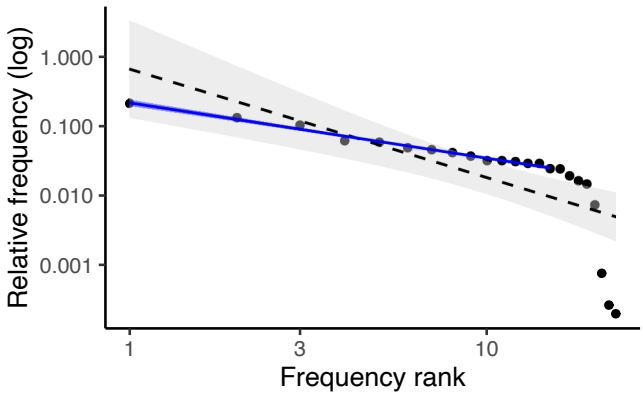

Figure 3.2: Log-log plot of frequencies versus frequency ranks in Walmajarri. When a linear model is fitted to the full distribution (dashed black), high and low frequency segments are overestimated and mid-rank segments are underestimated. When lowest-frequency segments are removed from the model (solid blue), the model appears to fit well.

Table 3.1: Power law (without x_{min}). Summary of α paramter, goodness-of-fit and p values for the power law distribution fitted to each language's full phonemic inventory.

	Mean	**SD**	**Min**	**Max**
α	1.38	0.17	1.16	2.18
goodness-of-fit	0.35	0.07	0.15	0.53
p	0.01	0.03	0.00	0.27

3.3 Results

We firstly infer the fit of a power law to the full distribution of phoneme frequencies for each language, without estimating an x_{min} parameter. In Table 3.1 we summarise the maximum likelihood estimates of the power law distribution's defining shape parameter, α, the goodness-of-fit of the estimated power law distribution to the observed distribution of phoneme frequencies, and bootstrapped p values for the null hypothesis that the data are plausibly drawn from a power law distribution.

Mean α is 1.38 (SD 0.17). As discussed earlier, the standard range of α is $2 < \alpha < 3$ (Clauset, Shalizi & Newman 2009: p. 662). α falls within this range for only 1 language. Furthermore, p values are very low. Just 2 of the 166 languages give a p value above the plausibility threshold.

Throughout this study, the possibility of type I error (false positives) must be taken into consideration. By setting our implausibility range at $p \leq 0.1$, we accept a one in ten chance of incorrectly rejecting a power law hypothesis which in fact is plausible—this can occur when the distribution's poor fit is due to chance fluctuation alone. Given 166 tests (one test per language), we would therefore expect to reject H_0 incorrectly in around 17 (10%) of those tests. In this instance though, we have rejected H_0 as implausible in 99% of the language sample. Thus it is clear that the power law distribution is being deemed implausible not merely by chance. It is genuinely a poor fit for the vast majority of languages. This result accords well with earlier work which has found that a simple, one-parameter form of the power law distribution poorly characterises phoneme frequencies (Sigurd 1968, Martindale et al. 1996,

Table 3.2: Power law distribution (with x_{min}). Summary of α paramter, goodness-of-fit and p values for the power law distribution fitted to a subset of more frequent phonemes in each language.

	Mean	**SD**	**Min**	**Max**
α	2.75	0.65	1.51	6.14
goodness-of-fit	0.14	0.03	0.08	0.22
p	0.62	0.26	0.01	1.00

Tambovtsev & Martindale 2007).

As discussed earlier, our dataset of phoneme frequencies is very likely to contain the complete population of phonemes in each language. At the same time, the number of observations per language is low—ranging from 16 to 37 segments in our language sample (mean 24.5, SD 3.8). Such a small set of observations can be a barrier to highly accurate maximum likelihood estimation. Clauset, Shalizi & Newman (2009: p. 669) suggest that a minimum sample size of around 50 is needed to get a maximum likelihood estimate of α accurate to at least 1%. This is simply not possible for most of the world's languages (including all languages in this study) due to the limited size of segment inventories. Thus, in phonemic studies such as ours there is likely to be an unavoidable uncertainty in the estimate of α.

3.3.1 Power law distribution with x_{min}

If the power law distribution, as inferred above, is inadequate for characterising phoneme frequencies, then what other options are there? There are a couple of approaches to this question. One is to add an additional parameter to improve the fit of the power law; the other is to consider alternative distribution types. In this and the following sections we explore both approaches.

Here, we infer the fit of a power law distribution with an additional x_{min} parameter, whose effect is to remove some of the least-frequent observations from the sample which is being fitted. As above, we use maximum likelihood to infer the best-fitting x_{min} threshold for each language. Results are summarised in Table 3.2.

After inferring an x_{min} parameter, the power law distribution is fitted to an average of only 13.9 segments, though there is a wide degree of variation (SD 4.1). In percentage terms, the power law distribution is fitted to an average of 57% of a language's segmental inventory (SD 16%). 117 languages (70%) fall within the normal 2–3 range for α. Having only a small number of included observations above the x_{min} threshold can drive unreasonably high estimates of the α scaling parameter. A sizeable portion of our sample (43 languages, 26%) fall in this high range with α above 3. At the other extreme, 6 languages (4%) have an unusually low α under 2. Mean α is 2.75 (SD 0.65).

When x_{min} is included, the power law hypothesis is accepted as plausible (though, to emphasise, not necessarily correct) in the 158 of 166 language varieties for which $p > 0.1$. p falls below the 0.1 plausibility threshold in the remaining 8 languages. The lowest p value for any language is 0.011. This puts the chance of incorrectly rejecting H_0 at around 1 in 100. The likelihood of a 1-in-100 event is high in a set of 166 tests. Overall, since the number of p values below 0.1 is considerably fewer than

the number we would expect to observe through chance, and since there is a reasonable possibility that the lowest p value, 0.011, is a type I error, we cannot confidently rule out the power law hypothesis for any language in our sample.

Although we have failed to rule out the power law distribution as implausible for any specific language, this still does not mean that the power law distribution is the optimal one for our data, and there are some important caveats to our results so far.

A distribution will always fit a set of data at least as well as the same distribution with one fewer parameter. Thus, the observation that the power law distribution fits better when x_{min} is added requires some interpretation. Of greatest interest in this respect is the striking degree of improvement in fit, such that the power law distribution shifts from a largely implausible fit against full phoneme inventories, to a largely plausible fit after we exclude the least-frequent observations from samples. This raises the obvious question of why this might be so. We consider this in our Discussion, after we have also examined distributional alternatives to power laws.

The inclusion of an x_{min} parameter when fitting power laws is common practice, but its use is most obviously motivated in contexts where there are very many possible observations. For example, Clauset et al. (2009: p. 684) fit a power law to frequencies of unique words in Moby Dick and find a best-fitting x_{min} of 7 (± 2). Words occurring fewer than 7 times can be disregarded and this still leaves nearly 3,000 unique words to which the power law distribution can be fitted. In contrast to this typical use case, where a large number of observations remain in play and do fit the power law, our use of x_{min} with phoneme datasets results in the exclusion of data points from an already small sample, leaving an even smaller set of data being fitted. As a general fact, it is inherently difficult to identify the most appropriate distribution for a small collection of observations. Correspondingly, it is not automatically an insightful finding, that a power law can be plausibly fitted to such small datasets. However, as mentioned just above, it is noteworthy that the same power law did not fit well to the slightly larger datasets that were being used without the x_{min} parameter. This suggests that it is not the merely small size of the dataset which is causing the good plausibility of the fit.

Small samples of observations can inflate p values, as is the case when investigating phonemes. We have good reason to suspect our p values are being inflated by the low number of observations per language, the evidence being that the number of p values we observe below 0.1 is considerably fewer than we would expect by chance. The difficulty we find in ruling out the power law distribution may reflect this.

3.3.2 Alternative distributions

In addition to considering the merits of adding extra parameters to a distribution, we must also consider whether a completely different distribution would provide an equally good or better fit to the data. We consider three alternative distributions, which are not part of the power law family and may suggest different underlying generative processes. These are the lognormal, exponential and Poisson distributions. Like the power law distribution, the shape of these distributions can have a sharp initial

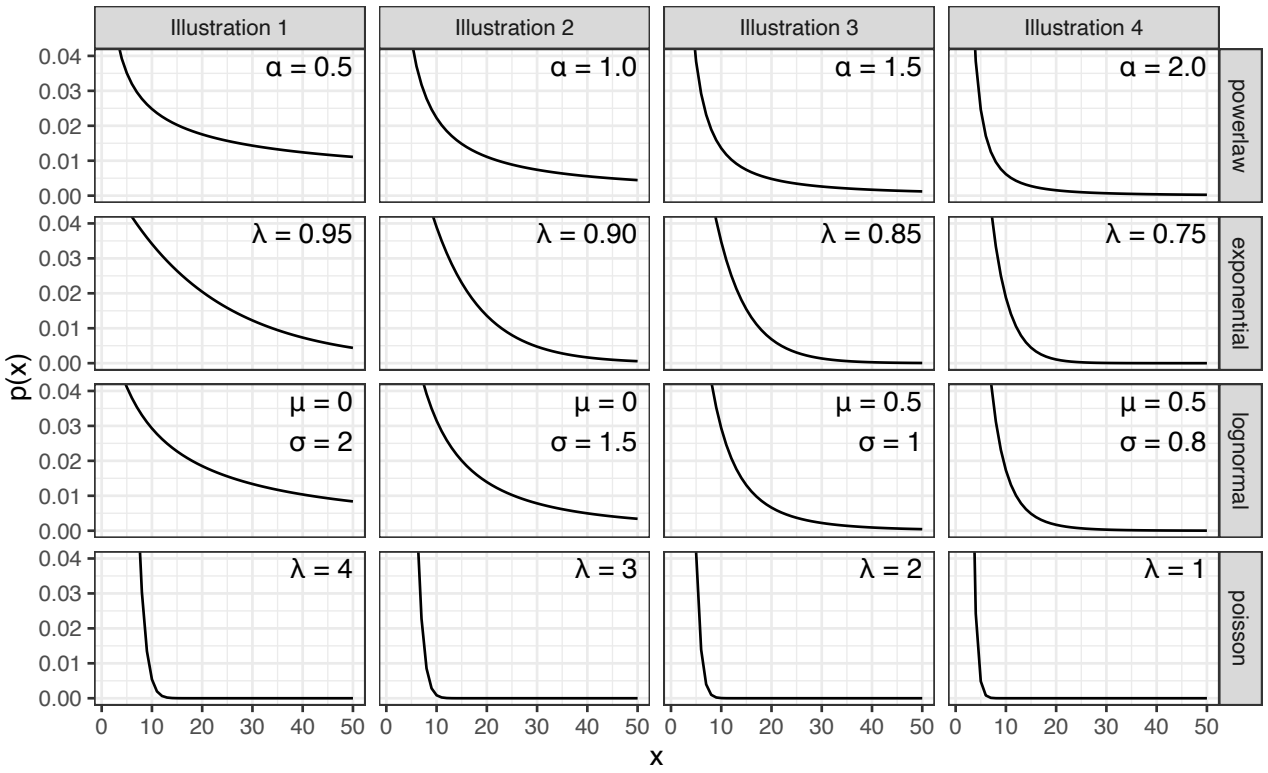

Figure 3.3: Four distribution types: power law, exponential, lognormal and Poisson, each illustrated with four parameterisations.

peak and a rapidly decaying tail, as illustrated in Figure 3.3.

3.3.3 Lognormal distribution

The lognormal distribution is one where the data form a normal distribution when transformed on a log scale. Once again, we use the *poweRlaw* package (Gillespie 2014) to estimate parameter values using maximum likelihood. In this instance, the parameters to be estimated are log mean and log standard deviation parameters—the log-scale equivalent of the two parameters that define a normal distribution. We fit the distribution to the whole set of segment frequencies for each language—we do not estimate an x_{min} parameter at this stage (though see below). The lognormal distribution narrowly construed is a continuous distribution, however the *poweRlaw* package contains a corresponding discretised version, appropriate to phoneme frequency data.

As for the power laws above, we calculate bootstrapped p values to assess the plausibility of the fit of the lognormal distribution for each language. The p values obtained are highly variable throughout the dataset. There are 73 languages (44% of the language sample) for which p falls in the range of implausibility, below 0.1. This is higher than we would expect if the lognormal distribution were plausible for all languages and $p \leq 0.1$ values were due to type I error alone. This result is a little difficult to interpret, given the previously discussed difficulties with small samples of observations per language. What seems clear is that, given the rate of $p \leq 0.1$ values is elevated beyond chance, we cannot say that the lognormal distribution plausibly characterises the segment frequencies of

all languages. Nevertheless, for many languages—56% of languages in our sample—we cannot confidently rule out the lognormal distribution. Overall, this makes the lognormal distribution with no x_{min} a better fit than the power law distribution with no x_{min}, which we ruled out for up to 99% of languages in the sample. One caveat to keep in mind is that the lognormal distribution is minimally defined by two parameters rather than one, which potentially puts it at an advantage compared to the single-parameter power law distribution.

3.3.4 Exponential distribution

An exponential distribution, and its discrete analogue, a geometric distribution, is one in which frequencies decay at a constant proportional rate. Thus for the frequencies of any two successively-ranked phonemes, x_k and x_{k+1}, the distribution is characterised by a rate parameter, λ, so that $x_{k+1} = \lambda . x_k$. Here, as above, we use maximum likelihood to estimate the parameter λ and the bootstrapping procedure to obtain a p value.[12]

Bootstrapped p values are above the 0.1 plausibility threshold for 147 of 166 languages. The number of languages for which $p \leq 0.1$ is 19, close to the 17 or so that we would expect from type I errors. This, on the face of it, seems to make the exponential distribution quite a plausible model for phonological segment frequencies more generally. It must be noted, however, that there are a few languages for which the exponential distribution is a very poor fit. The most extreme, Miriwoong, has a goodness-of-fit statistic of 0.27 and a p value of 0.036. The poor quality of fit is visually evident on a log-log plot (see Section S6, Supplementary Materials).

3.3.5 Poisson distribution

The final distribution we consider is the Poisson distribution, which is related to the exponential distribution. The Poisson distribution is typically used to model the frequency of an event within some interval of time or space. Our case is a bit different since we are modelling the relationship between the frequency of many different events (different phonological segments) and their frequency rank in a language's phonological inventory. As with the exponential distribution, we use maximum likelihood to estimate a single parameter, λ, and use bootstrapping to obtain a p value for the plausibility of the distribution.

The Poisson distribution is totally implausible for all languages in our language sample. Goodness-of-fit statistics range from 0.43 to 0.75 (mean 0.59, SD 0.07). We find p values indistinguishable from 0 in all cases.

[12] This generates a relationship where $p(x_k) \propto \lambda^k$. An alternative expression of the same relationship is $p(x_k) \propto e^{-\lambda' k}$, where $\lambda' = -log(\lambda)$. Results reported in Sections S5 and S6 of the Supplementary Material use this second definition, with λ'.

Table 3.3: Results summary. For each of the four distributions considered, this table lists the number of languages (and percentage of the total language sample) for which the distribution plausibly fits, as indicated by an uncorrected $p > 0.1$ value. 'Prop. fitted' gives the average proportion of each language's phoneme inventory above x_{min}.

	Without x_{min}	**With x_{min}**	**Prop. fitted**
Power law	2 (1%)	158 (95%)	56%
Lognormal	93 (56%)	155 (93%)	78%
Exponential	147 (89%)	146 (88%)	84%
Poisson	0 (0%)	43 (26%)	17%

3.3.6 Summary of results by individual distribution type

In Table 3.3, we summarise results for the four distribution types evaluated in this study. For each distribution type, we give the number of languages for which the distribution's fit was deemed plausible ($p > 0.1$). For completeness, we give results for the exponential, lognormal and Poisson distributions when x_{min} is included, just as we did for the power law distribution. (Note: for one language, the bootstrapped p value estimation procedure failed to converge for the lognormal distribution with x_{min}. This is the only distribution we tested which has three free parameters, and in this instance, the algorithmic procedure struggles to differentiate solutions with very similar likelihoods.) Perhaps most noteworthy among these results is the greatly increased inconclusiveness of the method when applied to the reduced set of data points lying above the x_{min} threshold. When the fitting task is restricted to a subset of only the most frequent segments in a language, it is possible to plausibly fit all but the Poisson distribution to any language, after type I error is factored in. One difference, which we nuance further in the next section, is that power law distributions with x_{min} are fitted on average to only 57% of a language's phonemes, whereas the lognormal and exponential distributions are fitted to closer to 80%.

3.3.7 Evaluation of comparative best fit

The third and final step in Clauset *et al.*'s (2009) framework is a likelihood ratio test. This third step may not always be necessary. If the bootstrapping procedure, above, were to show that only one distribution type plausibly fits the data, it would already have been shown to have the best fit among the candidates examined. However, bootstrapping may identify multiple distribution types as plausible. It will be recalled that just because a distribution is judged plausible via the bootstrapping process does not mean that it is the optimal one, since there may be other equally or more plausible distributions. Accordingly, when there are multiple plausible candidate distributions, Clauset, Shalizi & Newman (2009) recommend using Vuong's (1989) likelihood ratio test for model selection, to determine the best-fitting of any pair of competing models. Full results of all likelihood ratio tests described in this section are tabled in Section S5 of the Supplementary Materials.

Vuong's (1989) test uses the Kullback-Leibler Information Criterion (Kullback & Leibler 1951) to

calculate the log likelihood of observing the data given a distribution model, and compares this to the log likelihood of observing the same data given a competing distribution model. The test returns a test statistic, which gives an indication of how strongly one model is favoured over another, and a p value, indicating whether the difference in the support for each model is statistically significant.

We begin by comparing distributions without the x_{min} parameter. As summarised in Table 3.3, two of these distributions (the power law and Poisson distributions, without x_{min}) have already been rejected as implausible for all or nearly all languages. Accordingly, we conduct just one likelihood ratio test per language, comparing the fit of the exponential versus lognormal distributions. Overall, we find that Vuong's likelihood ratio test somewhat favours the exponential distribution. Likelihood ratios favour the exponential distribution for 122 languages, and the lognormal distribution for 44 languages. However after Bonferroni correction, the difference in the likelihood of exponential and lognormal models is statistically significant for only two languages, Thaynakwithi and Dalabon, both favouring the exponential distribution.

Turning to distributions with the x_{min} parameter, since we have already rejected the Poisson distribution, we conduct likelihood ratio tests pairwise among the remaining three distributions. In order to compare distributions with x_{min} parameters, it is necessary to set x_{min} to the same value in both distributions (Gillespie 2014). Thus, to make a pairwise comparison, we take the x_{min} value from distribution A and using it, re-estimate the other parameters of distribution B, and conduct one likelihood ratio test. Then we take x_{min} from B, use it and re-estimate the other parameters of distribution A, and conduct a second likelihood ratio test, giving two results for each pair of distributions.

Comparing the exponential and lognormal distributions, the likelihood ratios favour the lognormal distribution (139 languages to 27) using x_{min} from the lognormal fit, and favours the exponential distribution (103 languages to 63) using x_{min} from the exponential fit, however none of these comparisons reaches significance after Bonferroni correction.

Comparing the power law and lognormal distributions, likelihood ratios favour the lognormal distribution (144 languages to 22) using x_{min} from the power law fit, and all languages when using x_{min} from the lognormal fit, however only two of these comparisons reaches significance after Bonferroni correction. Yir Yoront favours the power law when using x_{min} from the power law fit and Malyangapa favours the lognormal distribution using x_{min} from the lognormal.

Comparing the power law and exponential distributions, the likelihood ratios favour the power law (133 languages to 33) when taking x_{min} from the power law fit, though no comparison reaches significance. They favour the exponential distribution 164 languages to 2 when x_{min} is taken from the exponential fit. Thirteen of those comparisons reach significance.

In sum, we found earlier that when parameterised without x_{min}, only the exponential and lognormal distributions were broadly plausible. Voung's likelihood ratio test marginally favours the exponential test over the lognormal when fitted against entire phonemic inventories, but the difference is at most slight. When parameterised with x_{min}, the power law distribution is fitted to around 60% of languages' phonemes on average, while the exponential and lognormal are fitted to around 80% (Table 3.3).

Pairwise likelihood ratio tests, which apply one distribution's x_{min} parameter to the other, provide slender evidence of the following. Even when fitted against the small phonemic subsets favoured by the power law, the lognormal distribution may weakly outperform the power law, but the exponential distribution does not. Fitted against the larger subsets favoured by the exponential and lognormal distributions, the power law is outperformed by the exponential and lognormal. The performance of the latter two distributions is indistinguishable.

3.4 Discussion

Here we contextualise the current study more fully in the history of research into power laws and phoneme frequencies, before drawing implications and conclusions from the new findings we have obtained.

3.4.1 Power laws in linguistics

Investigation of power laws in the linguistic sphere has a long history. One of the oldest and best-known examples of a power law in any discipline is the distribution of word frequencies in text corpora, first noted by Estoup (1916) and subsequently described by Zipf (1932, 1949). Zipf's Law, as it has come to be known, is a discrete power law distribution. Its exponent parameter, α, is typically very close to 1, in which case, the second ranked item will be approximately half as frequent as the first, the third ranked item will be one third as frequent as the first, and so on. Zipf's Law continues to garner considerable attention, for example in Kucera et al. (1967), Montemurro (2001), and more recently in Baayen (2001, 2008). Various modifications to Zipf's formula have been suggested (notably Mandelbrot 1954) and theoretical explanations put forward (Li 1992, Naranan & Balasubrahmanyan 1992, 1998).

Power laws have also been proposed to describe the distribution of phoneme frequencies. The use of Zipf's Law to model the frequencies of phonological segments initially appears to be an attractive prospect (Witten & Bell 1990: pp. 565–566). Nevertheless, a selection of alternative, non-power law distributions has also been suggested.

Sigurd (1968) is an early study evaluating the fit of a Zipfian distribution to phoneme frequencies, where the exponent, α, is set to 1. His evaluation method is a simple visual inspection, comparing observed phoneme frequencies in five languages (selected for their variety in segmental inventory size) with their expected frequencies assuming a Zipfian rank-frequency relationship. Sigurd (1968: p. 8) observes that the phoneme frequency distributions do not approximate a Zipfian curve, particularly for the most common segments. Rather, Sigurd (1968) finds better approximations using a geometric series equation, so that $x_{k+1} = \lambda . x_k$ for some rate parameter λ, giving the discrete distribution:

$$p(x_k) \propto \lambda^k \tag{3.3}$$

Good (1969: p.577) suggests an alternative method of approximation: following Whitworth (1901), Good calculates the expected frequencies of each phoneme given a process whereby a unit interval probability space $[0,1]$ is divided into n parts at random (where n is the number of phonemes in the language), following a uniform distribution. This is equivalent to a so-called stick-breaking process: imagine a stick, which represents the unit interval probability space. The stick is broken into n parts; the $n-1$ places along the stick at which a break is made are selected randomly and all at once, with any place along the stick equally likely to be selected as any other. When these parts are rearranged by size, from smallest to largest, their expectation follows the distribution:

$$\frac{1}{n^2}, \quad \frac{1}{n}\left(\frac{1}{n}+\frac{1}{n-1}\right), \quad \frac{1}{n}\left(\frac{1}{n}+\frac{1}{n-1}+\frac{1}{n-2}\right), \quad \ldots \tag{3.4}$$

Which is to say:

$$p(x_k) \propto \sum_{i=k}^{n} \frac{1}{i}, \quad 1 \leq k \leq n \tag{3.5}$$

In support of this model, Good (1969: p.577) provides a table of observed versus expected frequencies of both graphemes and phonemes in English, however the sample size is modest (1000 words) and does not extend to any other languages. Furthermore, there is no visual or statistical evaluation of the goodness-of-fit. Good (1969) intends for the results to be taken as a curious observation only, with no strong theoretical position or claim of generalisability.

As n in equation (3.5) grows large, the summation term $\sum_{i=k}^{n} \frac{1}{i}$ converges towards $-log(\frac{k}{n+1})$ (Loeb & Rota 1989: p. 12), meaning that Good's distribution can be considered an approximation of (3.6), a discretised negative logarithmic distribution.

$$p(x_k) \propto -\log \frac{k}{n+1}, \quad 1 \leq k \leq n \tag{3.6}$$

Gusein-Zade (1988) and Borodovsky & Gusein-Zade (1989) visually evaluate the fit of (3.6) to the graphemes of English, Estonian, Russian and Spanish.[13] They also use the equation to describe the distribution of DNA codons (Borodovsky & Gusein-Zade 1989). Witten & Bell (1990: pp. 563–566) examine the frequencies of single graphemes, graphemic bigrams and trigrams in the Brown Corpus and compare the fits of Good's distribution and Zipf's Law by comparing expected entropy values for each model to observed entropy scores. They find that the quality of the fit of Good's model declines with bigrams and trigrams compared to single graphemes, although the observed distribution curves are broadly of the same shape (and resemble the shape of Good's distribution rather than the Zipfian distribution). When assessed using metrics based on entropy, Good's distribution fits better than or around equally as well as the Zipfian distribution for all three datasets. Good's distribution and the

[13] Of course, the statistics of graphemes are different from the statistics of phonological segments. As Bloomfield (1935: pp. 136–137) rather emphatically points out: "If we take a large body of speech, we can count out the relative frequencies of phonemes and of combinations of phonemes. This task has been neglected by linguists and very imperfectly performed by amateurs, who confuse phonemes with printed letters." Nevertheless, the frequencies of graphemes has been of interest historically in many applications; for example, in traditional printing, the development of Morse code, and library cataloguing (Witten & Bell 1990: pp. 550–551).

negative logarithmic distribution it approximates also have the advantage of parsimony, since they are parameter-free: knowing how many unique items (phonemes, graphemes, bigrams, etc.), n, are in the dataset is sufficient to calculate their expected distribution of frequencies—there are no additional parameters to estimate such as α in (3.2) or λ in (3.3).

Martindale et al. (1996) compare the fit of four different distributions to frequencies of both graphemes and phonemes in text corpora from 18 languages. Using the R^2 statistic in a linear regression, they compare the fit of the parameter-free negative logarithmic equation of Borodovsky & Gusein-Zade (1989) in (3.6) to the Zipfian power-law distribution (3.2), Sigurd's geometric series distribution (3.3), and the Yule-Simon distribution (Yule 1925, Simon 1955), which can be written:

$$p(x_k) \propto \frac{1}{k^\alpha}.\lambda^k \tag{3.7}$$

The Yule-Simon equation in (3.7) is the product of the power law in (3.2) and the geometric equation in (3.3). Because of the differing rates at which the two parts of the equation decay as k increases, equation (3.7) produces a distribution which is more like a power law (3.2) for low values of k (and thus for high frequency items, for instance) and more like the geometric (3.3) for high values of k (low frequency items) (Simon 1955).

The Yule-Simon equation in (3.7) has not just one free parameter but two, the exponent α and the rate λ, and the Zipfian and Sigurd equations are effectively special cases of it, each with one parameter fewer. The Zipfian distribution is equivalent to (3.7) with λ set to 1 (so that $\lambda^k = 1$), while the geometric equation is equivalent to (3.7) with α set to zero (so that $1/k^\alpha = 1$). This is important, since as a general fact, if distribution A is a special case of distribution B, with fewer free parameters than it, then B will always perform at least as well as A when fitting the same set of data. Thus, the Yule-Simon distribution will necessarily fit the same set of data at least as well as the Zipfian distribution, and Sigurd's geometric distribution.

Martindale et al. (1996) find that the Yule-Simon distribution fits best, for both graphemes and phonemes. They find that the Zipfian distribution tends to overestimate both high and low frequency items, although the differences they observe between models are only small. On this basis, they conclude that it is "a matter of taste" whether one opts for the more precise Yule-Simon distribution or simpler models with fewer parameters to estimate (Martindale et al. 1996: p. 111). Tambovtsev & Martindale (2007) expand Martindale et al.'s (1996) study to include phoneme frequencies in 95 languages (90 of these are Eurasian; 2 are from Oceania and 1 each from Australia, Africa and South America). The sample is divided into four language groups (Indo-European, Altaic and Yukaghir-Uralic–plus a miscellaneous group) and a series of pairwise sign tests are conducted to test whether the difference in mean R^2 is significant between different distributions for each language group. Again, they find that the Yule-Simon distribution fits best overall.[14]

Obtaining a better fit by using a distribution with an additional parameter may be relatively trivial mathematically speaking, but this does not mean it is uninteresting. The extra parameter may work to capture a significant real-world nuance in an underlying causal process or describe the effect of one or

[14] Although in their statistical tests they do not adjust their significance levels to correct for multiple hypothesis testing.

more secondary processes. A compelling causal explanation of a complex distribution might therefore be formulated by identifying some real-world factor and explaining how its mathematical effect on the distribution is expected to match what we find. It is also important to consider the possibility of equifinality—the fact that multiple, different real-world phenomena may have equivalent mathematical effects. Tests of goodness-of-fit examine only the mathematical aspect, and cannot distinguish between different phenomena whose detectable mathematical contribution is equivalent.

Martindale et al. (1996: p. 111) and Tambovtsev & Martindale (2007: p. 9) note that a Zipfian distribution describes frequencies of phonological segments less well than it describes frequencies of words, in part because the highest-frequency phonemes are not frequent enough. They speculate that this may be so, because if the most-frequent phonemes did pattern in a Zipfian way, then perception problems could arise for language users owing to the small size of a phonological inventory. This speculation does not meet the criteria for a compelling causal explanation though. It is not clarified what the linguistic mechanism is, that acts to prevent such perceptual problems, and thus we do not have a real-world phenomenon whose mathematical properties could be interrogated. Nor is it explained why, if such a mechanism exists, its mathematical effect would be to contribute something like the extra geometric term λ^k that differentiates the Yule-Simon distribution (3.7) from the Zipfian (3.2).[15] This is not to say that an explanation in terms of perceptibility and confusability is implausible, but rather if our aim is for causal hypotheses that can be evaluated, then more steps are needed, a topic that we return to below. Next, however, we consider the history of investigation presented here, together with our new results.

3.4.2 Findings from a more reliable evaluation procedure

Power laws have attracted wide and sustained scientific interest. Recent debates on their validity have prompted the development and widespread adoption of novel evaluation methods that are more reliable than those used in the past (Clauset, Shalizi & Newman 2009, Stumpf & Porter 2012). In this study, we re-evaluated the plausibility of several distribution types as characterisations of phoneme frequencies using a maximum likelihood statistical framework presented by Clauset, Shalizi & Newman (2009) and a sample of 166 Australian language varieties.

Using a more reliable evaluation procedure than previous investigations, we have confirmed the finding that a basic power law distribution, with a single free parameter, is generally insufficient for characterising phoneme frequencies. Additionally, we reconfirm a result going back to Sigurd (1968), that an exponential (or geometric) distribution, with a single free parameter, is a good plausible fit for full phonemic inventories. Furthermore, we find that a lognormal distribution, with two free parameters, is an additional plausible fit, whereas a Poisson distribution, with a single free parameter, is implausible. We did not attempt to fit the two-parameter Yule-Simon distribution in (3.7), since to our

[15]The Yule-Simon equation, which Martindale et al. (1996) and Tambovtsev & Martindale (2007) find to be a superior fit, describes a distribution which is most similar to a power law for high frequency (low k) items, and most like the geometric for low frequency (high k). The claim that its superior fit is due to *non*-power-law-like behaviour of high frequency items is therefore hard to reconcile with the mathematics.

knowledge, there is no maximum likelihood estimation procedure currently available for estimating its parameters. However, we do return to the question of this distribution just below.

A second novel contribution was to consider the addition of an x_{min} parameter, a practice which is now common in power law research. Notably, while power laws are largely implausible fits for entire phoneme inventories, their plausibility is improved strikingly once a subset of the least-frequent phonemes is removed from the sample. This is despite that fact that the full inventories and the reduced ones share the property of comprising notably small samples. The subset removed in order to achieve maximum likelihood is on average large, at 43%. This result indicates that power laws constitute a plausible characterisation for the more-frequent portion of phonemic inventories, and explains why the upper end of a Yule-Simon distribution, which most closely approximates a power law, should be a reasonable fit. We note however, that the lognormal distribution also performs well in this same, high frequency region of phonemic inventories. Exponential (or geometric) distributions do not fit the higher-frequency portion of inventories as well the power law or lognormal do, but they are good fits for entire inventories, suggesting that they fit particularly well in lower-frequency portions. This would explain why the lower end of a Yule-Simon distribution, which most closely approximates a geometric distribution, should be a reasonable fit.

Using an evaluation procedure which has since been shown to be unreliable, Martindale et al. (1996) and Tambovtsev & Martindale (2007) concluded that the two-parameter Yule-Simon distribution fit whole inventories better than a power law or a geometric distribution. Implicitly, this is a conclusion in two parts: more-frequent phonemes are more power-law-like, and less-frequent are more geometric-like. Here we have not been able to directly evaluate the Yule-Simon distribution using the more reliable, maximum likelihood method. However, we have found evidence supporting a similar conclusion, that the more-frequent and less-frequent portions of phonemic inventories are characterised by different distributional properties. The more-frequent portion better matches a power law, though also a lognormal distribution. The less-frequent portion better matches a geometric distribution. These two findings serve to clarify and qualify the two implicit halves of the main finding of Martindale et al. (1996) and Tambovtsev & Martindale (2007), and here we have arrived at them by more reliable methods. Furthermore, by estimating x_{min} parameters, we have provided some estimates of where power-law-like behaviour starts to cut out within a phonemic inventory. To understand what these results entail for theory, we return to the question of causal processes.

3.4.3 Distributions: outcomes of stochastic processes linked to causal factors

Ultimately, linguistic theory seeks to explain the patterns that can be observed in human language. This endeavour will be aided by a sound knowledge of which mathematical distributions plausibly characterise a given variable x (such as phoneme frequency), since those distributions will be consistent with only certain mathematical kinds of underlying processes. In this paper, we have improved the certainty of our understanding of observed distributions of phoneme frequencies, using state-of-the-art statistical methods. This is a necessary step, but a first step only. A fruitful next step used widely in

other sciences is to explicitly consider mathematical families of stochastic processes, whose signature outcome distributions are consistent with those of our empirical observations, and then to ask in turn, what plausible, real-world causal processes could be consistent with these observation-matching stochastic processes.

It can be emphasised that in this model of progress, a good understanding of families of stochastic processes will play an important, enabling role. Accordingly, it will be useful to harness advances that have already been made elsewhere. For instance, many discrete systems can be profitably conceptualised in terms of urn processes with characteristically associated distributions. As Kuba & Panholzer (2012: p. 87) remark, "[u]rn models are simple, useful mathematical tools for describing many evolutionary processes in diverse fields of application". There exist well-studied urn processes which yield many kinds of distributions. It will be profitable in linguistic research to more clearly relate our own theories of change, including change in phonemic inventories, to these mathematically more generalised processes. By doing so, linguists will be able to tap into related, existing mathematical results (such as relating processes to distributions), that can assist us to further differentiate the theories that are more viable from those that are less so.

3.4.4 Casual processes of phoneme addition, removal and redistribution

Naturally, any observation-matching families of stochastic processes will still need to be related to linguistically plausible causal processes (Cysouw 2009). Though our paper has primarily focused on methods that will strengthen future discussions around links between observations and causal mechanisms, rather than the causal mechanisms themselves, here we offer some brief remarks on historical mechanisms affecting phoneme frequencies. As we do, we continue using the imagery of urns to represent phonemes types; the balls in them to represent their tokens in the lexicon; and processes that serve to add, remove and redistribute them. (For a like-minded review of potential mechanisms behind Zipf's law in the frequencies of *words*, see Piantadosi 2014)

Phoneme frequencies undergo constant modification due to changes occurring at multiple levels of linguistic structure, including the phonology proper, morphology, syntax and the grammar of discourse. Within the phonology itself, fundamental changes include deletions, insertions and changes of phonemic category, all of which will impact phoneme frequency distributions. In an urn model, historical deletions will result in balls being removed from a phoneme's corresponding urn while insertions will add balls. Adding a level of complexity, many phonological changes affect not just one phoneme category, but natural classes of sounds, thus deletions and insertions that apply to natural classes will remove or add balls simultaneously from a non-random, 'natural' set of urns.

Phonemic mergers (Hoenigswald 1965) are changes which collapse two erstwhile phoneme categories into one, effectively emptying all balls from one urn into another. Phonemic partial mergers, also known as primary splits, shift only some of a phoneme's instances into another, existing phonemic category, transferring some portion of an urn's balls to another. Phonemic (secondary) splits involve an existing phonemic category splitting into multiple new categories, entailing the creation of one or more

new urns, filled with some proportion of the balls from an existing urn, which itself remains non-empty. In both mergers and splits, natural classes may be involved, entailing simultaneous transfers between, and creation or loss of, 'natural' sets of urns.

A phonemic category will map onto multiple actual speech sounds, or phones, and the phonemic categorisation of a phone frequently depends on its contextual environment. Consequently, it is not uncommon for multiple types of phonemic changes to occur at once, owing to the fact that when one sound changes, so too does the contextual environment of its neighbours. Under these circumstances, the ubiquity of non-uniform distributions of sounds in various contexts will lead to non-accidental, correlational relationships among these coupled changes.

At the morphological level, changes in the frequency of certain formatives will cause concerted frequency changes in the set of phonemes comprising the formative, though in this case, there is no expectation that the set involved will form a natural class. If the object of study is the discourse frequencies of phonemes, then similar effects will arise when the usage frequencies of words undergo change. Furthermore, lexicons are not closed systems. Words can be borrowed from other languages. The effects of borrowing on phoneme frequencies in the recipient language is a complex matter (Boretzky 1991). Here we name just a few factors. The donor language will have its own phonemic repertoire and associated frequencies, which will bias what can be donated and at what relative rates. Phonemic borrowing is not direct, but is mediated by phonetic similarities and psychological equivalences drawn by speakers across languages (Flege 1987, Kang 2003). Correspondences between donor and recipient phonemes will therefore exhibit correspondences that are broadly natural (Paradis & LaCharité 1997), yet the process is frequently variable (Lev-Ari, San Giacomo & Peperkamp 2014) and may involve centuries of subsequent remodeling of borrowed material (Crawford 2009). Moreover, additional systematic mutations, deletions and insertions of phonemes may be motivated by constraints on the permissible phonotactic (contextual) arrangement of phonemes in the recipient language.

There are implications to be drawn from this, for the explanation of distributions of phoneme frequencies. We turn to these implications below.

3.4.5 Implications for explanatory accounts

In most of the historical changes noted above, it matters not only whether, and how many, balls are moving from one urn to another, but exactly which real-world phonemes the urns themselves represent. Phonemes form natural classes, which may change in tandem. Phonotactic arrangements, which, notwithstanding variation, exhibit strong similarities cross-linguistically, will impact how changes mediated by contexts are correlated. And borrowing too operates in reference to phonemes' actual substance. This importance of phonemic substance raises a host of questions, which can be addressed in new and potentially revealing ways within a stochastic-modelling research program. For instance, the frequencies with which various kinds of phonemes are present or absent across the world's languages varies greatly (Maddieson 1984, S. Moran 2012, Everett 2018b): what is the range of assumptions under which this result would emerge, within a stochastic model? Does it demand

models with strong effects corresponding to well-known articulatory, acoustic and perceptual factors that cross-linguistically favour certain phoneme types (Liljencrants & Lindblom 1972, Stevens 1989, Johnson, Ladefoged & Lindau 1993, Browman & Goldstein 1986, Proctor 2009, Becker-Kristal 2010, Everett 2018a) or can it be derived from weak ones? And can some effects arise entirely through indirect interactions? For instance, vowels typically stand adjacent only to consonants, while in most languages consonants may be adjacent to vowels or other consonants; can it be shown that this basic phonotactic asymmetry predicts differing rates of change among vowels and consonants (Moran & Verkerk 2018a)? Returning to the main topic of this paper, might we predict that languages can exhibit significantly different overall phoneme frequency distributions depending on which phonemes they contain? For example, might observations for one part of the world (such as the Australian data in this paper) therefore differ in an explicable manner from observations in others?

Even if we abstract away from issues related to phoneme substance, a successful stochastic model of phoneme frequency evolution may still require a mixture of sub-models, obeying different principles, given the existence of the multiple types of historical processes that affect phoneme frequency: deletion, insertion, splits, mergers, borrowings. Though it might be clear in outline what (some of) these sub-models must be, it is far less clear, empirically speaking, what their quantitative parameterisations are. In what precise mix do these sub-processes occur? How often do events occur independently or in concert? What empirical grounds do we have for estimating these parameters, and what are our levels of uncertainty? In these respects, there exists another interesting role to be played by stochastic modelling studies. We currently lack empirical answers to many of these quantitative questions, but can we reveal limits to what is plausibly compatible with observed phoneme distributions? A related open question, connected to issues raised in the previous paragraph, is to what extent a successful stochastic model will need to reflect empirical complexities, like phonemic substance, natural classes, phonotactics and concerted changes, as points of non-independence within and between sub-models, or can some of these be ignored with little impact on the outcomes predicted by the model? Whatever the answers may be, we see a productive and interesting field of inquiring lying directly ahead.

We conclude this section by returning to our empirical findings and situating them within the causal reasoning just outlined. Our empirical finding was that the more frequent phonemes in a language's inventory tend to be distributed more in line with a power law, and the less frequent more in line with an exponential distribution. As we mentioned earlier, power law distributions can be stochastically generated by preferential attachment processes. In addition to this, exponential distributions can be stochastically generated by a so-called birth-death process, in which entities arise at some characteristic birth rate λ, and disappear at a characteristic death rate μ. It may be—and here we are speculating, but in a reasoned manner—that as tokens of phones shift between phonemic categories during the kinds of sound change processes we mentioned above, different phonemic categories have different probabilistic propensities to be source categories (i.e., the erstwhile category of the changing phone) and destination categories (its new category after the change). This alone would constitute a birth-death regime (Cysouw 2009), and its causal roots would lie in phonemic substance. Taking this further, it may also be that there is a separate propensity for phones to migrate towards numerically stronger

categories. This would contribute a preferential attachment dynamic, whose causal roots would be something over and above mere phonemic substance. An interesting point to note, is that under this scenario, it is not important which specific phonemes an inventory contains: so long as phonemes have differing propensities to be source or destination categories during changes, and so long as numerically stronger categories are favoured as destinations, the same outcomes should be obtained. This would explain how our Australian results can resemble earlier, Eurasian, results so closely. However, what remains incomplete in this picture is firstly: When a preferential attachment process and a birth-death process are simultaneously active in one and the same system, under what parameterisations of the system does it fall out that preferential attachment affects the more-frequent categories more strongly than the less-frequent, in a fashion similar to what appears to be the case in phoneme inventories? And secondly: What factor(s) might underlie both the preferential attachment dynamic itself and the overall parameterisation of the system? There has been some work on this question, suggesting that climactic factors (namely the effect of humidity on vocal fold physiology) (Everett, Blasi & Roberts 2015), universal sound-meaning associations (Blasi et al. 2016) and/or relative physiological ease of articulation (Everett 2018b,a) might drive a preferential attachment dynamic, in which languages gravitate towards certain phonemic categories on the basis of these various factors. It may be that different hypotheses about possible causal mechanisms for the preferential attachment dynamic entail different predictions on these crucial points. These ideas are worth chasing up further.

3.4.6 Reasons to seek wider horizons

There are likely to be additional, valuable variables, and possibly more tractable variables, to study beyond just phoneme frequency distributions. In this paper we have focused on phoneme frequencies because they have occupied such a prominent place in the history of investigations of distributions. However, by doing so we do not wish to suggest that phonemes ought to continue to occupy such a prominent place. Firstly, we agree with Piantadosi (2014), that the focus of investigation should not be on distributions for their own sake, but on what a better understanding of distributions can tell us about language. It may be, that further study reveals that phoneme frequency distributions are not very powerful, when it comes to discriminating plausible from implausible causal models, whether this is because phoneme inventories are too small to allow their distributions to be characterised with sufficient precision or because those distributions themselves are consistent with too many competing explanations. However, the very same research may reveal other variables that are more valuable. To speculate, perhaps it will prove more interesting to ask how *sets* of phonemes, such as natural classes, pattern within their languages' frequency ranks across languages; or to examine not merely phonemes (of which there are only few categories per language), but phonemes within certain contextual environments (of which there are more). Moreover, some interesting results may emerge only within models and statistical analyses that grant a role to phylogeny (Macklin-Cordes, Bowern & Round 2021).

3.4.7 Conclusions

There are many branching paths of research ahead. Nevertheless, there will be some basic principles that we see remaining constant, and these have been the core focus of our contribution here. In this paper we have demonstrated and outlined a template for future work on distributions. Ideally, such work should begin with critical assessment of links that can be made between existing or new causal hypotheses, including diachronic processes, and from these, via stochastic models, to particular distributional outcomes. Subsequently, the fit of the hypothesised distributions to real-world data should be evaluated rigorously using robust statistical methods. Lastly, an attempt must be made to rule out competing distribution types and alternative generative mechanisms. As our investigation demonstrates, this may be challenging, given the inherent limitations of working with small sets of observations. Maintaining a clear-eyed view of these limitations, and using advances already made in allied fields, will help spur this field of inquiry to new, robust insights into the dynamics of phoneme inventories.

Acknowledgments

We wish to thank Caleb Everett and Steven Moran for their insightful comments on earlier drafts of the manuscript. Their feedback and guidance helped us to improve the paper considerably, and we appreciate the generosity of their time and effort.

Funding: JM is supported by an Australian Government Research Training Program Scholarship. Data compilation was funded by Australian Research Council grant DE150101024 to ER.

Supplemental Data

Supplementary information and materials for this paper are organised into 7 parts.

Text-based supplementary information is available online at `https://www.frontiersin.org/articles/10.3389/fpsyg.2020.570895/full#supplementary-material`. This document includes a discussion of neutral expectations about phonemes' lexical and discourse frequencies (S1), a guide to data and code (S2), a bibliographical list of original wordlist sources (S3), a comparison of phoneme inventories in Ausphon and Phoible 2.0 (S4), and full tables of results (S5).

Further supplementary materials are contained in a zip folder *S6-S7_suppmaterials*, hosted on Zenodo at `https://doi.org/10.5281/zenodo.4104116`. The *S6_data_viewer* directory contains an interactive data viewer. The interactive data viewer is also hosted in an accompanying R package on Github (`https://github.com/JaydenM-C/phonfreq`) for ease of installation and use (see the description in S2 for further details). The *S7_data_code_results* directory contains 3 subdirectories. The *data* subdirectory contains a tab-delimited format spreadsheet of phoneme frequency data. The *R* subdirectory contains code used to perform analysis and create figures. The *results* subdirectory contains the output of analysis saved as *.Rdata* format objects for use with R statistical software.

Text from Chapter 4 has been incorporated into the following manuscript submitted for publication: Jayden L. Macklin-Cordes & Erich R. Round. Submitted. Challenges of sampling and how phylogenetic comparative methods help: With a case study of the Pama-Nyungan laminal contrast. Preprint. https://doi.org/10.5281/zenodo.4796981

Contributor	Statement of contribution	%
Jayden Macklin-Cordes	initial concept	100
	analysis	100
	theoretical derivations	100
	writing of text	100
	proof-reading	50
Erich Round	supervision, guidance	100
	proof-reading	50

Content from following chapter forms part of a manuscript which has been submitted for publication. Macklin-Cordes contributed the background and theoretical component of the manuscript, which is what is presented here in this thesis. In the manuscript, Round contributes additional discussion and a case study of laminal contrasts in Australian languages. Note that the percentages given in the statement of contribution above relate to the thesis chapter only. Both authors contributed equally to the submitted manuscript.

Chapter 4

Phylogenetic Comparative Methods in Linguistics

Historical linguistics is increasingly making use of phylogenetic methods. However, phylogeny is consequential for all comparative study, including synchronic typology. Many fields of science, including linguistic typology, have a long history of considering phylogenetic relationships in data sampling. However, a family of statistical methods, *phylogenetic comparative methods*, enable the incorporation of phylogeny directly in a statistical model, with no loss of data. This paper clarifies the logic behind phylogenetic comparative methods, and argues for their applicability in linguistic typology. I make the case for phylogenetic comparative methods firstly by surveying responses to the issue of phylogenetic independence in linguistics and other fields of science. I find that all fields share, in origin, similar lines of development in sampling methodology to create phylogenetically independent samples. However, since embracing quantitative methods, comparative biologists have developed mathematical frameworks for identifying and modelling phylogenetic effects in statistical analysis. Linguists have the opportunity to incorporate this wealth of experience into comparative research design. Latter sections of the paper outline how this can be done, firstly by describing *phylogenetic signal* and methods for its measurement. Secondly, I survey existing applications of phylogenetic signal methods and phylogenetic comparative methods more generally in linguistics. Drawing together previous sections, I argue for the continued uptake of phylogenetic comparative methods in linguistics and describe some implications of this. A desirable outcome for future research would be the collation of high-resolution phylogenetic language trees inferred over the past two decades, brought together into a common data structure. Furthermore, whole-family language sampling in concert with phylogenetic comparative methods should be considered in linguistic typology, beyond language samples that aim for much more sparse, global coverage.

4.1 Introduction

The comparative language sciences are indispensable in the study of human language and offer a unique contribution to the study of human history. Besides being an intriguing academic pursuit in its own right, historical linguistics can be triangulated with other fields, such as genetics, archaeology and anthropology, to infer early population movements and interactions between cultures (e.g. Hunley et al. 2008, Gray, Drummond & Greenhill 2009, Bouckaert et al. 2012b, Malaspinas et al. 2016, Bouckaert, Bowern & Atkinson 2018). Similarly, linguistic typology has been combined with fields including ecology and physiology to produce some remarkable theories on the evolution of human language (e.g. Everett, Blasi & Roberts 2015, Everett 2017, Bentz et al. 2018, Blasi et al. 2019).

The fundamental task of comparison underlies both linguistic typology and historical linguistics. In this respect, they are synchronic and diachronic sides of the same coin. In the historical case, extant language data are compared to infer past relationships between languages that cannot be observed directly due to the passing of time. In typology, language data are compared as well, though with different aims. These aims principally concern the nature of human language itself—the limits of possibility and tendencies in language structures, for example. The task of cross-linguistic comparison is complicated, however, by the interwoven patterns of historical descent and language contact that inevitably bind languages to varying degrees and manifest in shared linguistic forms and features observable today. Consequently, shared histories must be taken into account before drawing any inferences from cross-linguistic datasets. The prevalence of a linguistic variable may give the appearance of a particular linguistic tendency if languages are considered independent, but it may also be explained by a single linguistic innovation which was subsequently inherited by a large number of descendant languages. Shared linguistic histories must also be considered in any statistical analysis, since independence of data points is a fundamental assumption of many statistical methods.

Non-independence due to shared histories through descent, termed *phylogenetic non-independence* or, more precisely, *phylogenetic autocorrelation*, is not an unfamiliar concept in linguistics, nor other fields where entities share common paths of descent, such as biology and anthropology. Over a century of thought and methodological development has been dedicated to the topic. However, divergences exist in the lines of thought and development of different fields. This paper considers this old discussion within a cross-disciplinary scope. There are many challenges associated with accounting for phylogenetic autocorrelation in comparative methods, however I find that certain challenges are not as unique to linguistics as often has been assumed. Comparative biologists continue to give a good deal of consideration to methodological challenges of interest to linguists as well. In particular, I find that a family of statistical methods, *phylogenetic comparative methods* (PCMs), are immediately applicable to comparative linguistics. In this paper, I elucidate how and why this is the case and demonstrate empirically the need to account for phylogenetic autocorrelation in variables which might have been assumed to be distributed independently of phylogeny, based on previous descriptions—these variables concern the typology of laminal consonants in the Pama-Nyungan languages of Australia.

This paper proceeds as follows. Section 4.2 reviews literature on *phylogenetic autocorrelation*—

the tendency of languages to show similarities due to phylogenetic relatedness—in linguistics and cognate fields (comparative biology, in particular). The aim is to provide a broad picture of the scientific context that motivates the methodologies discussed later on. While identifying historical signal in a source of data is of clear interest to historical linguistics, it is also of interest to linguistic typology, where phylogenetic autocorrelation is a source of bias that must be controlled. I therefore survey literature from an array of comparative fields of science in which phylogenetic autocorrelation occurs, comparing methodological approaches for identifying and accounting for patterns of historical relatedness among observations in comparative datasets. Section 4.3 outlines some statistical tools for quantifying *phylogenetic signal*, the degree of phylogenetic autocorrelation present in a comparative dataset. Quantifying phylogenetic signal is the first step in a phylogenetically informed comparative methodology—the presence or absence of phylogenetic signal determines the need for phylogenetic comparative methods in subsequent analysis. Then, in Section 4.4, I discuss examples of studies measuring phylogenetic signal in linguistics. For instance, I consider the example of laminal phonemes in Pama-Nyungan languages (Australia). In this example, Round (2017a) shows that, although place contrasts of laminal consonants are traditionally described as being areally distributed and independent of the Pama-Nyungan family's internal phylogeny, laminal consonants do show phylogenetic patterning when finer-grained variables are examined that capture matters of frequency. A finding like this emphasises the need to consider phylogenetic autocorrelation in linguistic typology, even when studying phenomena that might be considered 'safe' from the effects of phylogeny based on prior research. I conclude by advocating for the continued uptake of phylogenetic comparative methods in linguistics. I discuss the implications of this uptake for language sampling strategy in linguistic typology and sketch an outline for future research.

4.2 Phylogenetic autocorrelation

Phylogenetic autocorrelation is common to many comparative fields of science, and linguistics is no exception. Phylogenetic autocorrelation is a potential problem for comparative study, because shared phylogenetic histories limit the independence of observations in a comparative dataset. Observations from more closely related entities will tend to show less variation than more distantly related entities, because they share a longer period of evolutionary history prior to splitting off from their most recent common ancestor, and will have had less time to diverge evolutionarily. If this tendency towards similarity due to shared phylogenetic history is not taken into account, it will introduce bias into the dataset and consequently affect statistical analysis. Different fields have their own lines of literature grappling with this phenomenon extending back many decades. Although there are many similarities between fields, key differences emerge since the uptake of quantitative methods in comparative biology. This section discusses phylogenetic autocorrelation and the history of responses to it in different fields, focusing in particular on linguistics (Section 4.2.1) and biology (Section 4.2.2).

4.2.1 Phylogenetic autocorrelation in linguistics and other fields

Both historical linguistics and linguistic typology are comparative fields, in that they necessarily rely on cross-linguistic datasets and the task of making comparisons between different languages is inherent to both. The presence of historical signal in a dataset—where the set of values reflects something of the history of featured languages—will be of interest to any researcher working comparatively, whether they are directly interested in reconstructing the history of those languages (as in historical linguistics) or not (as in linguistic typology). In the case of typology, this is because historical signal represents a kind of statistical non-independence between languages. Statistical non-independence between languages due to shared history is no new revelation in linguistic typology, however there are many possible approaches to dealing with it and a sizeable body of literature on the topic.

We explore typological literature further below, but first I take up the topic of phylogenetic non-independence in more detail. In any cross-linguistic study, there will be some degree of statistical non-independence between languages. This is because languages do not evolve independently, but share various historical connections through time, resulting in shared lexical or grammatical material, whether through inheritance from a common ancestor or borrowing from neighboring languages. Historical relationships between languages are, therefore, an inescapable concern for linguistic typology. Likewise, typological comparison has a place in historical linguistics, as researchers have investigated historical questions by applying statistical methods to typological datasets. A seminal example of this is Nichols (1992), and more recent examples within linguistic phylogenetics include Dunn et al. (2005), Dunn et al. (2008), Rexová, Bastin & Frynta (2006), Reesink, Singer & Dunn (2009) Sicoli & Holton (2014) and Greenhill et al. (2017). Non-independence of languages due to historical relationships, and the consequences for cross-linguistic comparison, has direct and indirect implications for both fields.

As noted above, linguistics is far from the only field to face the challenge of phylogenetic non-independence. In comparative anthropology, this issue was noted as early as 1889 by Sir Francis Galton in the context of cross-cultural datasets, which lack independence due to shared histories of cultural innovation and exchange between societies (Naroll 1961: p. 15). This phenomenon, known as *Galton's Problem*, is now more precisely understood as a form of statistical autocorrelation (similarity between observations as a function of the time lag between them). The same phenomenon has been recognised in comparative biology too. A seminal study concerning comparative studies of phenotypes, Felsenstein (1985) demonstrates that data from species cannot be assumed to be independently drawn from the same distribution, because species are related to one another via a branching, hierarchical phylogeny, thus, statistical methods that assume independent, identically-distributed observations will inflate the significance of the test (discussed further in Section 4.2.2 below). Linguists, it has been argued, have been somewhat slower than those in other fields to acknowledge exposure to Galton's problem, or phylogenetic autocorrelation (Perkins 1989: p. 293). Nevertheless, this is a central concern of Dryer (1989: p. 259) and has been addressed in a considerable body of linguistic literature since then.

Although precise strategies are varied, common to all fields is a history of addressing phylogenetic autocorrelation at the data sampling stage. In linguistics, the use of sampling methods for creating a

phylogenetically independent or phylogenetically balanced language sample remains the predominant way of accounting for phylogenetic autocorrelation and literature on this topic extends back several decades. A. Bell (1978: pp. 145–149) argues that common strategies which simply ensure equally-weighted representation of "all major families" or all continents is inadequate due to differing rates of divergence among families. He estimates the number of language groups separated by more than 3,500 years of divergence and uses it as a heuristic for estimating genetic biases in a selection of proposed language samples. He concludes that European languages tended to be overrepresented and Indo-Pacific languages underrepresented and attributes this to a corresponding over/under-representation among quality language resources, which is a persistent problem for comparative linguistics. Perkins (1980, 1988) creates a sample of 50 languages, later adapted by Bybee (1985), which attempts to account for both genetic and areal biases by selecting no more than one language from each language phylum (following Voegelin & Voegelin 1966) and no more than one language from each cultural and geographic area (following Kenny 1975, Murdock 1967). This method attempts to account for nonindependence due to areal spread, unlike Bell's heuristic measure which accounts only for genetic bias, however it does not account for differing ages of divergence and size of language phyla in the way Bell does. Crucially for this discussion, these sampling methods inherently have what I refer to as a *lossy* quality. In information technology, lossy methods of data compression involve erasure of parts of the data to create a smaller approximation of the original file (for example, JPEG image files). Similarly, in the language sampling methods thus far described, a smaller approximation of a larger language sample is created by removing languages with certain historical connections that would compromise the independence of observations in the sample.

There are a number of shortfalls associated with lossy sampling strategies. At the crux of these is that it may not be possible to create an independent sample of sufficient size when distant genetic relationships and long-range areal phenomena are considered. There may be uncertainty and scope for disagreement on the independence of languages in a sample. Dryer (1989: p. 261) refers to the example of the inclusion of three languages in Perkins' sample (Ingassana, Maasai and Songhai) which may be related as part of the Nilo-Saharan family, although these relationships are remote and subject to debate. When the largest proposed areal and genetic groupings are considered, it may simply not be possible to create an independent sample of a sufficient size for generating statistically significant inferences. Another problem is that the maximal extent of presently established language families is partially a product of the extent of adequate documentation and scholarly attention, rather than a completely true reflection of the fullest extent to which the family may be reconstructed. That is to say, two languages which are presently understood to be unrelated, and therefore statistically independent, may in fact belong to a shared larger grouping, which has not yet been identified due to poor documentation or some other factor. Dryer (1989: p. 263) raises another concern, which is that languages selected on the basis of genetic independence may nonetheless share characteristics due to non-genetic processes—language contact and borrowing. When areal phenomena are considered, Dryer contends, the practicality of constructing a truly independent sample of sufficient size is further stretched. Dryer's proposed solution is to build a sample of languages of approximately equal relative

independence (at the level of major subfamilies within Indo-European, such as Romance, Germanic, and so on) for each of five large linguistic areas which are assumed to be independent, or at least sufficiently independent for statistical purposes. Any statistical test can then be applied to each of the five areas and only if the same result is replicated in all five areas is it considered statistically significant. If the same result is replicated in four of five areas, this falls short of statistical significance, although Dryer (1989: pp. 272–273) considers such cases to be evidence of a "trend". Even still, as Dryer (1989: p. 284) acknowledges, his five linguistic areas may be subject to the same concerns about undetected historical non-independence and it is possible that the whole world may, in effect, function as a single linguistic area, such that the distribution of certain linguistic features may reflect extremely remote areal or genealogical pattern rather than some true tendency of human language.

Nichols (1992: p. 41) uses Dryer's area-by-area testing method as part of a three-pronged approach. For any given question, Nichols first conducts a chi-square test of the world sample and then re-tests the significance of the finding using either Dryer's method or by running the same test on only the sample of "New World" languages (comprising North, Central and South America). Rijkhoff et al. (1993) and Rijkhoff & Bakker (1998) develop another approach to account for the possibility of non-independence across large linguistic areas and large, as-yet-undetected families. They permit multiple languages within a family to be included but develop a measure, based on the density of nodes in a known language phylogeny, to determine how many languages should be included. In this way, they also aim to account for the fact that some language families will have greater internal diversity than others (see also Bakker 2011, Miestamo, Bakker & Arppe 2016). Another proposed method is to set a minimum threshold of typological distance between languages, calculated from the *World Atlas of Language Structures* (WALS) (Dryer & Haspelmath 2013), such that languages must be sufficiently typologically distinct from others in the sample to warrant inclusion. Bickel (2009) develops an alternative algorithm based on Dryer (1989), which allows all uniquely-valued data points within a family to be included in the sample, but then reduces the weighting of data points in the final analysis where a particular value is over-represented within a family. In other words, if all the languages in a particular family share the same value for a variable of interest, those observations may be reduced to a single data point, since this homogeneity is likely the result of a shared retention or innovation.

All up, the developments in typological methodology that have been discussed here demonstrate that historical non-independence between languages has been treated predominantly as a sampling issue in linguistics. Earlier researchers sought to maintain the independence of their sample by maximising the genetic distance between the languages in their sample, such that no two languages were known to belong to the same family. Later, with subsequent acknowledgement of the possibility of non-independence from very large language families, as well as large-scale areal diffusion and effects from as-yet undetected or unconfirmed historical relations, it became apparent that it may be impossible to create a sample which is simultaneously independent and sufficiently large to generate statistical significance. The response to this has been a variety of robustness checks, even bootstrapping-like processes, whereby languages are sampled at an approximately equal relative level of independence

and the sample is then subdivided in some way and a statistical test replicated over each subdivision. More recent years have seen the continued evolution of statistics and robustness checking methods (for an overview, see Roberts 2018), although balanced sampling remains a common element of modern, large-scale comparative linguistic studies (for example, Everett, Blasi & Roberts 2015, Everett 2017, Blasi, Michaelis & Haspelmath 2017).

4.2.2 Phylogenetic autocorrelation in comparative biology

Comparative biology faces the same issue of phylogenetic autocorrelation as comparative linguistics. Many conventional statistical methods assume that observations are independent and identically distributed, which is problematic in biology since observations come from species, which are related to one another through shared evolutionary histories (as charted graphically in a tree diagram). Further, Felsenstein (1985: p. 4) shows that even non-parametric statistics are not immune to violations of this independence assumption. Although both fields face the same phenomenon in essence, linguistics and biology have diverged in their methodological response in recent decades. While linguistics continues to focus on sampling procedures, giving less attention to the methods of subsequent statistical analysis, comparative biologists have shifted towards more direct, statistical solutions.

Earlier approaches to phylogenetic autocorrelation in biology are in a similar vein to the sampling methods discussed in the previous section. Harvey & Mace (1982: pp. 346–347) seek to find a taxonomic level to sample from, which strikes the right balance in terms of being sufficiently statistically independent without being so conservative that sample sizes become prohibitively small. Their proposed solution is to identify and sample from the lowest taxonomic level which can be "justified on statistical grounds". One method of doing this is suggested by Clutton-Brock & Harvey (1977: pp. 6–8), who conduct a nested analysis of variance and then select taxonomic level containing the greatest level of variation. Once Clutton-Brock & Harvey (1977) identify their taxonomic level of interest, they average out data for all species within a given genus for which they have data. In other words, the unit of analysis has shifted from individual species to genera, and each data point represents a genus in the form of an averaged representation of all the species within the genus. Although a similar method in essence, this genus-level averaging process is in marked contrast to balanced sampling methods discussed in the previous section, where an unaltered observation from a single exemplar language is taken as representative of its given family, subfamily or other defined grouping.

Baker & Parker (1979: pp. 85–86) discuss the same problem. They use an approach not too dissimilar from the area-by-area robustness checking by Dryer (1989) and Nichols (1992). Baker & Parker (1979) replicate their analysis within individual families as well as within different ecological areas, with the assumption that if the same associations are observed within different groups as they are across the dataset as a whole, then one can discount the possibility that the full analysis is simply picking up differences between different families or different ecological groups. A contrasting approach, at least in instances where categorical data are of interest, is to reconstruct ancestral states throughout the phylogeny, enabling one to directly observe whether species which share a common

trait do so because of (non-independent) shared inheritance from a common ancestor or whether the traits have evolved independently. Gittleman (1981) uses a *parsimony model* to reconstruct states in this way. This is where the states of traits at ancestral nodes in a tree are reconstructed in such a way as to minimise the number of evolutionary changes that would need to take place to produce the observed data for extant species on the tips of the tree. There are a couple of shortfalls to this approach. One is that it is only a partial solution to the problem—it tells us which data points are independent and which are not, but besides potentially being used as a tool to identify a taxonomic level with the greatest level of diversity (proceeding in a similar way to Clutton-Brock & Harvey 1977), there is no indication of how to proceed with comparative study when non-independence has been identified. Further, there are biases in the parsimony method (Felsenstein 1985: p. 7) which are unlikely to be satisfying for linguists—for example, if a small group of related species (or languages) all share a trait, they will always be reconstructed as inheriting the shared trait from the nearest common ancestor, despite the possibility of parallel evolution (homoplasy), horizontal diffusion, environmental pressures, and so on.

Felsenstein (1985) contends that it is possible to account for phylogenetic non-independence in a statistical model without the need to remove non-independent data points or compromise the unit of analysis (by, for example, comparing averaged data points representing genera rather than individual species). Felsenstein's breakthrough insight is that this can be achieved not by directly comparing non-independent observations but by comparing *phylogenetically independent contrasts* (PICs) between them. His method has become, by one estimate, the most widespread in comparative biology (Nunn 2011: p. 162). Calculating phylogenetically independent contrasts is possible given a continuously-valued variable of interest, an assumed phylogeny and an assumed model of variable evolution. As a natural starting point, Felsenstein assumes a *Brownian motion* model of evolution. This is where an evolving trait can wander positively or negatively with equal probability, and each new time step is independent from the last, with the resulting effect that displacement of the variable over time will be drawn from a normal distribution with a mean of zero and variance proportional to the amount of elapsed time (Felsenstein 1985: p. 8). Consider two sister tips on a phylogenetic tree and two accompanying observations for a continuously-valued variable of interest. The two observations themselves cannot be considered statistically independent, since the two sister tips share much of their evolutionary history through a common point of origin. However, the *contrast* between the two values is independent, because any difference between the two sisters will be the consequence of evolutionary events occurring only along the two *separate* branches linking each sister to their last common ancestor. If the historical evolution of this variable follows a Brownian motion model as described above, then the contrast between the two sister tips will be drawn from a normal distribution with a mean of zero and variance proportional to the time that has elapsed since the two tips split in the tree. An observed contrast can be scaled by dividing it by the standard deviation of its expected variance. This gives a statistically independent contrast of expectation zero and unit variance. This process can be repeated for all adjacent tips in the tree. Contrasts can then be extracted from adjacent nodes in the tree, where the value of the node is an average of the observed values of the tips below it. In the end, there will be a collection of phylogenetic independent contrasts, all of expectation zero and unit variance. It is

then possible to apply standard statistical tests to the phylogenetic independent contrasts (rather than directly to observed values) without phylogenetic autocorrelation introducing bias into the results.

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

There has been some interest in applying phylogenetic comparative methods to cross-linguistic data (for example, Dunn, Greenhill, et al. 2011, Maurits & Griffiths 2014, Verkerk 2014, Birchall 2015, Zhou & Bowern 2015, Calude & Verkerk 2016, Dunn et al. 2017, Verkerk 2017, Bentz et al. 2018). In contrast to the lossy sampling methods discussed in the previous section, phylogenetic comparative methods are *lossless*. They make it possible to incorporate all available data by directly incorporating phylogenetic history into the statistical model, rather than removing data points in order to balance the sample. Another reason that phylogenetic comparative methods are important is that no comparative study in either biology or linguistics is phylogenetically neutral, no matter what balanced sampling procedure might have been used. One may assume that a set of sampled languages are all sufficiently independent from one another and, accordingly, treat data from those languages as independent observations in statistical analysis, but this is still a phylogenetic assumption: one in which all languages are equally distant from one another in a phylogeny (in other words, all languages are connected to a single node by branches of the same length, with no intermediary structure—a *star phylogeny*) (Purvis & Garland 1993). Phylogenetic comparative methods enable the direct inclusion of existing phylogenetic knowledge beyond a simple star phylogeny.

One limitation of phylogenetic comparative methods, and a likely source of criticism in linguistics, is that they rely on access to high-quality, fully-resolved phylogenies complete with branch lengths. This simply is not realistic for families of languages across many parts of the world. Further, phylogenetic comparative methods assume reference phylogenies are accurate, when, in practice, phylogenies are subject to uncertainty. While tree inference in modern biology benefits from advances in high quality, large scale genomic data, it would be natural to balk at this assumption in the context of historical linguistics, where so many phylogenetic relationships within and between language families remain uncertain o unknown. It may come as somewhat of a surprise then, that Felsenstein (1985: p. 14) expresses precisely the same concern about phylogenetic uncertainty in the context of comparative

biology. Nevertheless, he says "phylogenies are fundamental to comparative biology; there is no doing it without taking them into account", and this is unavoidably true of comparative linguistics as well. Further, even if a study does not explicitly consider phylogeny, it is still unavoidably making phylogenetic assumptions. Comfortingly, computational advances make it feasible to incorporate phylogenetic uncertainty into analysis explicitly (and this is an area of continuing active development). Also, there is evidence that even when phylogenies are incomplete, lacking branch length information, or even subject to a degree of error, phylogenetic comparative methods still typically out-perform equivalent (non-phylogentic) comparative methods, which effectively assume a star phylogeny (Grafen 1989, Purvis, Gittleman & Luh 1994, Symonds & Page 2002).

4.3 Phylogenetic signal

As discussed in Section 4.2, phylogenetic comparative methods are applicable in linguistic typology and any kind of comparative linguistic study where phylogeny is a confound. The previous section described phylogenetic comparative methods, which account for phylogeny. Some variables, however, may not evolve through descent with modification and may not pattern phylogenetically, or may do so only weakly. For example, some biological characteristics may reflect adaptations to a certain ecological niche. Ecological niche hypotheses have been proposed for some phonological variables too (Everett, Blasi & Roberts 2015, Everett 2017, Blasi, Michaelis & Haspelmath 2017). How does one determine, then, whether phylogeny is a confounding factor for a variable of interest? In the last 15 years, an advance in this area has been the advent of methods for quantifying explicitly the degree of *phylogenetic signal* in comparative data (Freckleton, Harvey & Pagel 2002, Blomberg, Garland & Ives 2003). Phylogenetic signal refers to the tendency of phylogenetically-related entities to resemble one another (Blomberg & Garland 2002, Blomberg, Garland & Ives 2003: p. 717). This resemblance is more technically defined as statistical non-independence among observation values due to phylogenetic relatedness between taxa (Revell et al. 2008: p. 591). This concept of phylogenetic signal has important applications in comparative linguistics. Here I argue that measuring phylogenetic signal should be considered as a first step in a phylogenetically aware comparative methodology, since it can determine empirically whether phylogenetic comparative methods are required or whether regular statistical methods may suffice. Further, the result of a phylogenetic signal test can contribute to evolutionary hypotheses in its own right, for example by giving evidence for or against a variable following certain modes of evolution.

Rather than assuming phylogenetic non-independence *a priori*, as the phylogenetic comparative methods discussed in Section 4.2.2 do, or lacking any statistical control for phylogeny and relying on balance sampling alone, measures of phylogenetic signal provide the advantage of being able to quantify explicitly the degree of phylogenetic non-independence in a dataset (Revell et al. 2008: p. 591). Phylogenetic signal may be expected to be strong in some cases or weak in others—given some data and a phylogenetic tree for reference, this can be tested empirically. If a typologist were to find a linguistic variable that is distributed significantly independently of phylogeny, this may be

a result of interest in itself—at the least, it will provide an empirical basis that justifies proceeding with regular statistical methods over phylogenetic comparative methods (as in Irschick et al. 1997). Measures of phylogenetic signal will also be of interest to historical linguistics, as a diagnostic tool for testing the degree and nature in which some data reflect the phylogenetic history of the languages from which they came. The following section describes some methods for measuring phylogenetic signal in different kinds of variables.

4.3.1 Quantifying phylogenetic signal in continuous variables

Blomberg, Garland & Ives (2003) provide a suite of tools for quantifying phylogenetic signal, which have become somewhat of a standard in the field (cited 3150 times on 31 May 2020, according to Google Scholar). Recent comparative studies using these tools include Balisi, Casey & Valkenburgh (2018), Hutchinson, Gaiarsa & Stouffer (2018) and Leff et al. (2018). Blomberg, Garland & Ives (2003) present a descriptive statistic, K, which is generalisable across phylogeneies of different sizes and shapes. In addition, they provide a randomisation test for checking whether the degree of phylogenetic signal for a given dataset is statistically significant. K can be calculated using either phylogenetic independent contrasts (PICs) (Felsenstein 1985) or generalised least squares (GLS) (Grafen 1989) (see Section 4.2.2). In a Brownian motion model, where variable values can wander up and down with equal probability through time, PIC variances are expected to be proportional to elapsed time. Among more closely related languages, where there has been less divergence time for variable values to wander, the variance of PICs is expected to be low. The randomisation test works by comparing whether observed PICs are lower than the PIC values obtained by randomly permuting the data across the tips of the tree. The process of permuting data across tree tips at random is repeated many times over. If the real variances, with data in their correct positions on the tree, are lower than 95% of the randomly permuted datasets, then the null hypothesis of no phylogenetic signal can be rejected at the conventional 95% confidence level. In other words, closely related languages resemble one another to a statistically significantly greater degree than would be expected by chance.

The descriptive statistic, K, quantifies the strength of phylogenetic signal. As with the randomisation procedure above, the input is a set of observed values, where each observation is associated with a tip of the reference tree. Blomberg, Garland & Ives (2003: p. 722) give an explanation of the calculation of the K statistic. To recap briefly, K is calculated by, firstly, taking the mean squared error of the data (MSE_0), as measured from a phylogenetically-corrected mean[1], 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 (the same variance-covariance matrix of phylogenetic distances incorporated into the error term in GLS-based phylogenetic regression, as discussed in the previous section). This latter value, MSE, will be small when the pattern of covariance in the data matches what would be expected given the phylogenetic distances in the reference tree, leading to a

[1] 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 has been corrected for overrepresentation by larger subclades (see Garland, Díaz-Uriarte & Westneat 1999).

high MSE_0/MSE ratio and vice versa. Thus, a high MSE_0/MSE ratio indicates higher phylogenetic signal. Finally, the observed ratio can be scaled according to its expectation under the assumption of Brownian motion evolution along the tree. This gives a K score which can be compared directly between analyses using different tree sizes and shapes. Where $K = 1$, this suggests a perfect match between the covariance observed in the data and what would be expected given the reference tree and the assumption of Brownian motion evolution. Where $K < 1$, close relatives in the tree bear less resemblance in the data than would be expected under the Brownian motion assumption. Notably, $K > 1$ 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, relatives bear greater resemblance than would be expected.

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

Other statistics for quantifying phylogenetic signal have been proposed and warrant mention. Freckleton, Harvey & Pagel (2002) propose using the λ (lambda) statistic, based on earlier work by Pagel (1999). As for Blomberg, Garland & Ives (2003), this approach works with a variance-covariance matrix showing the amount of shared evolutionary history between any two tips in the tree (the diagonal of the matrix, the variances, will indicate the total height of the tree; the off-diagonals, the covariances, will indicate the amount of shared evolutionary history between two given entities, before they diverge in the tree). The statistic, λ is a scaling parameter which can be applied to this variance-covariance matrix. Scaling the values in the matrix by λ transforms the branch lengths of the tree, from $\lambda = 1$, where branch lengths are left unscaled, to $\lambda = 0$, where all covariances in the matrix will be zero, in other words, no covariance through shared evolutionary history is indicated between any tips, thus all tips will be joined at the root by branches of equal length (a star phylogeny). Freckleton, Harvey & Pagel (2002) present a method for finding the λ parameter that maximises the likelihood of a set of observations arising, given a Brownian motion model of evolution. If λ is close to 1, this indicates high phylogenetic signal, where the data closely fit expectation given the shared evolutionary histories in the tree and a Brownian motion model of evolution. Further measures which have been proposed are I (P. A. P. Moran 1950), a spatial autocorrelation measure which was adapted for phylogenetic analyses by Gittleman & Kot (1990), and C_{mean} [abouheif_method_1999], which is a test for serial independence (for an overview, see Münkemüller et al. 2012). In an evaluation of different methods

Münkemüller et al. (2012) find that, assuming a Brownian motion model of evolution, C_{mean} and λ generally outperform K and I. However, C_{mean} considers only the topology of the reference tree (i.e., the order of the branches from top to bottom), but not branch length information, and the value of the C_{mean} statistic is partially dependent on tree size and shape, so it lacks comparability between different studies. In addition, λ shows some unreliability with small sample sizes (trees with ˜20 tips).

4.3.2 Quantifying phylogenetic signal in binary variables

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

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

$$D = \frac{\Sigma d_{obs} - mean\,(\Sigma d_b)}{mean\,(\Sigma d_r) - mean\,(\Sigma d_b)} \tag{4.1}$$

Scaling D in this way provides a standardised statistic which can be compared between different sets of data, with trees of different sizes and shapes, as with K for continuous variables. One disadvantage

of D, however, is that it requires quite large sample sizes (>50), below which it loses statistical power, increasing the chance of a false positive result (type I error).

4.3.3 Quantifying phylogenetic signal in multivariate and multidimensional data

A notable, more recent development concerns the generalisation of methods for quantifying phylogenetic signal in multivariate and multi-dimensional data. Methods discussed so far quantify phylogenetic signal for a single variable of interest. Zheng et al. (2009) present a generalisation of the Blomberg, Garland & Ives (2003) K statistic for jointly estimating the strength of phylogenetic signal in a collection of variables. In addition, their method allows the incorporation of measurement error or variation within the entities being studied (be they species, languages, etcetera) (see Ives et al. 2007). Both of these developments are expected to improve the statistical power of the test, which is an advantage particularly where small sample sizes are concerned, since the original K statistic requires a minimum sample size of around 20 and lacks sufficient statistical power below this level. This is achieved by standardising the values of each variable to have mean 0 and variance 1 then jointly measuring the MSE for all variables. Adams (2014) presents another generalisation of K for use with 'multivariate traits', K_{mult}. Whereas Zheng et al. (2009) estimate phylogenetic signal for a set of multiple *independent* variables simultaneously, a multivariate trait is conceptually a single evolutionary trait but has multiple values associated with it, which are mathematically interrelated and cannot be analyzed independently. The example Adams (2014) uses is head shape for a family of salamander species.

4.4 Phylogenetic signal and PCMs in linguistics

Macklin-Cordes, Bowern & Round (2021) presents an application of phylogenetic signal measuring methods in linguistics, specifically a dataset of binary and continuous-valued phonotactic variables from 111 Pama-Nyungan languages using the D and K tests described above. Contrary to expectation, given prior descriptions of Pama-Nyungan phonology and phonotactics, strong phylogenetic signal is detected in this phonotactic data. This work has been tentatively expanded by Macklin-Cordes & Round (2018). Since phonological processes and sound change affect natural classes of sounds rather than individual phonemes, phonotactic variables are subject to complex patterns of non-independence between them. Macklin-Cordes & Round (2018) attempts to account for this by implementing the K_{mult} method described above and shows that, at first pass, K_{mult} seems to detect stronger phylogenetic signal than testing several hundred phonotactic variables individually and averaging the results. This work requires further exploration, however. The evaluation of phylogenetic signal is, in itself, the end goal of these studies. They focus on the question of whether or not historical information is present in a novel dataset, which is more straightforwardly of interest for historical linguistic inquiry. However, the results hold secondary implications for the typological matter of non-independence between languages. The detection of phylogenetic signal in the phonotactics of Pama-Nyungan languages would need

to be incorporated into any future typological methodology. The need for phylogenetic comparative methods has been established in this particular domain.

In another phylogenetic signal detection example using the same Pama-Nyungan reference phylogeny, Round (2017a) re-examines assumptions about the laminal contrast in Australian languages. The distribution of Australian languages with a single contrastive laminal versus two contrastive laminal places at first seems not to correspond strongly with family and subgroup boundaries, leading researchers to propose that the distribution of one versus two laminal languages has been driven historically by areal diffusion (Dixon 1970, 1980, Breen 1997, Dixon 2002). Round (2017a), however, extracts a finer grained level of variation by considering not just the single, binary presence or absence of a laminal contrast but the frequencies at which those laminals appear in certain contexts before and after different vowels. Round (2017a) detects strong phylogenetic signal in these frequency variables, suggesting that, contrary to previous proposals, laminals do not appear to evolve in a distinctly non-phylogenetic way.

Without delving into detail, other recent examples of phylogenetic comparative methodologies include Bromham et al. (2015), which examines the relationship between rates of lexical change to population size in Austronesian languages; Verkerk (2015), investigating the development of manner verbs and path verbs in Indo-European languages; Calude & Verkerk (2016), reconstructing numerals systems in Indo-European languages; Bentz et al. (2018), investigating links between language evolution and ecological factors in a sample of nearly 7,000 languages across the world; and Moran, Grossman & Verkerk (2020), quantifying differential rates of change in consonants versus vowels in 8 language families across 6 continents.

there are a few limitations of PCMs to note. One limiting assumption is that the tree being used as a reference phylogeny is an accurate representation of the true historical phylogeny. Since the past cannot be observed directly, the best available yardstick is limited to the best available phylogeny that has been inferred independently. In linguistics, this can be a particular problem. The cross-linguistic coverage of linguistic phylogenetic studies has expanded rapidly in the past two decades and some families, such as Indo-European and Austronesian, have been studied extensively. However, high-resolution phylogenetic trees do not exist for large swathes of the globe's linguistic diversity. In the context of Sahul, there is good coverage of the Pama-Nyungan family (Bowern & Atkinson 2012, Bouckaert, Bowern & Atkinson 2018) but, as yet, few if any studies computationally inferring phylogenies of non-Pama-Nyungan families nor any Papuan language families in New Guinea to the north. There are several large databases of world language classifications and efforts to make them easily available for phylogenetic comparative study (Dediu 2018), but these can lack resolution and branch length information. One partial solution is to incorporate phylogenetic uncertainty into any PCM analysis explicitly by replicating the analysis over a posterior sample of trees rather than a single summary tree, replicating the analysis over separate, competing reference trees, or, in the case of a limited tree structure with lots of polytomies, randomly simulating bifurcating tree structures to simulate uncertainty.

A second key assumption that can prove problematic is that the data being tested are assumed to be

completely independent of the data that was used to infer the reference phylogeny. This can be tricky in comparative linguistic data, which tend to contain complex interdependencies. As one example, the phonotactic frequency datasets in Macklin-Cordes, Bowern & Round (2021) were extracted from wordlists which contained, as a subset, the words that were used to code cognate data, from which the reference tree was inferred.

One final point to note is that recent research suggests that phylogenetic signal can be inflated when variable values evolve according to a Lévy process, where a variable value can wander as per a Brownian motion process, but with the addition of discontinuous paths (i.e., sudden jumps in the variable's value) (Uyeda et al. 2018). This is particularly concerning for any frequency-based phonological data, which will be subject to sudden shifts caused by phonemic mergers, splits, and other regular sound changes. This is likewise a matter of concern in comparative biology and subject to active development in that field (Uyeda et al. 2018).

4.5 Conclusion

Historical and synchronic comparative linguistics are increasingly making use of phylogenetic methods for the same reasons that led biologist to switch to them several decades ago. The central contention of this research essay 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.

The following publication has been incorporated as Chapter 5:

Jayden L. Macklin-Cordes, Claire Bowern & Erich R. Round. 2021. Phylogenetic signal in phonotactics. *Diachronica* 38. https://doi.org/10.1075/dia.20004.mac

Contributor	Statement of contribution	%
Jayden Macklin-Cordes	initial concept	50
	scripting, analysis	80
	writing of text	90
	proof-reading	20
	preparation of figures	75
Claire Bowern	supplying data	10
	supervision, guidance	5
	proof-reading	10
Erich Round	initial concept	50
	supplying data	90
	scripting, analysis	20
	supervision, guidance	95
	writing of text	10
	proof-reading	70
	preparation of figures	25

Data for this paper comes from the Ausphon Lexicon database, created and maintained by Erich Round and extending on the CHIRILA database created and maintained by Claire Bowern. The Pama-Nyungan reference phylogeny was inferred and supplied by Claire Bowern.

Responsibility for the study's initial conceptualisation and development is shared approximately equally between Macklin-Cordes and Round, representing the culmination of innumerable discussions over a long period of time.

Experimental design, scripting and analysis was conducted by Macklin-Cordes (though note that the scripts depend considerably on functions developed by Round). Round contributed text to Section 5.3.2. Macklin-Cordes drafted the remainder of the text. The 'swatch' visual device (e.g. Figure 5.8) was created by Round and minimally adapted by Macklin-Cordes in this paper. Remaining figures were created by Macklin-Cordes with guidance from Round. Macklin-Cordes and Round revised text with feedback from three anonymous reviewers. Bowern assisted with proofreading.

Chapter 5

Phylogenetic signal in phonotactics

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 characterised 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.

5.1 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, Cysouw & Stadler 2011, Bowern 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 (Bowern & Atkinson 2012: p. 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, Bowern & Atkinson 2018). In comparative biology, phylogenetic methods have been applied profitably to investigations of community ecology (Webb et al. 2002), ecological niche conservatism (Losos 2008), paeleobiology (Sallan & Friedman 2012) and quantitative genetics (de 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, Bowern & Atkinson 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 Sections 5.1–5.2, 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 Sections 5.3–5.6, we present tests for phylogenetic signal in phonotactic characters extracted from wordlists for 112 Pama-Nyungan language varieties. Section 5.3 details the materials used and reference phylogeny. Section 5.4 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 5.5 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 5.6 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 Sections 5.7–5.8 with discussion of the limitations of the study design, implications of the results and directions for future research.

5.1.1 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, Greenhill & Gray 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: p. 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, Bastin & Frynta 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 2018b, Moran, Grossman & Verkerk 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.

5.1.2 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 phonemicised wordlists. As long as there is a wordlist of suitable length (Dockum & Bowern 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 Section 5.3 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, Brinkmann & Philippe 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, Hedges & Tamura 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 Dresher & 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: pp. 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 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 Sections 5.4–5.6 below.

Many kinds of phonotactic structure 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 generalise, 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.

5.2 Phylogenetic signal

The concept of *phylogenetic signal* (Blomberg & Garland 2002, Blomberg, Garland & Ives 2003: p. 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 recognised 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, Greenhill, et al. 2011, Maurits & Griffiths 2014, Verkerk 2014, Birchall 2015, Zhou & Bowern 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: p. 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, Harvey & Pagel (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 Section 5.3.1). We use measures of phylogenetic signal to test the hypothesis that our data contain historical information and, therefore, could contribute to future historical linguistic study.

[1] See Nunn (2011) for discussion.

Blomberg, Garland & Ives (2003) provide a set of statistics for measuring phylogenetic signal, which remains prevalent today (for example, Balisi, Casey & Valkenburgh 2018, Hutchinson, Gaiarsa & Stouffer 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, Garland & Ives 2003: p. 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 Section 5.7 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 Section 5.7 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_0), 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, Garland & Ives 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_0/MSE ratio and vice versa. Thus, a high MSE_0/MSE ratio indicates higher phylogenetic signal. Finally, the observed MSE_0/MSE ratio can be scaled according to the expected MSE_0/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.

[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).

Blomberg, Garland & Ives (2003) also present a *randomisation procedure* for testing whether the degree of phylogenetic signal in a dataset is statistically significant. The randomisation procedure utilises 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, Garland & Ives (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 randomisation 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 summarise briefly, the D statistic is based on the sum of differences between sister tips and sister clades, Σd. 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 binarises 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:

$$D = \frac{\Sigma d_{obs} - mean\,(\Sigma d_b)}{mean\,(\Sigma d_r) - mean\,(\Sigma d_b)} \tag{5.1}$$

Scaling D in this way provides a standardised 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 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 summarise, 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, Garland & Ives (2003) to quantify phylogenetic signal in continuous, frequency-based character data and we use the accompanying randomisation procedure to test the statistical significance of that signal. Likewise, we use the D statistic developed by Fritz & Purvis (2010) to quantify phylognetic signal in binary character data and we use its corresponding randomisation procedure to test for statistical signficance as well—with the randomisation procedure, in this latter instance, testing against two null hypotheses: one of phylogenetic signal, and one of random noise.

5.3 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's status 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.

5.3.1 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 (Bowern & Atkinson 2012: p. 817). Pama-Nyungan was first proposed and named by Kenneth Hale (Wurm 1963: p. 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, Voegelin & Voegelin 1966, Wurm 1972, Austin 1981b). For an overview of the history of Pama-Nyungan classification, see Bowern & Koch (2004: ch. 1–5) and Koch (2014). Bowern & 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, Bowern & Atkinson (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), Bowern (2017) and Round (2021b), among others. This curious level of homogeneity extends to phonotactics too (Dixon 1980, Hamilton 1996, Baker 2014, Round 2021a).

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 phylogenetic diversity, presents challenges for historical linguistics. Baker (2014: p. 141) and Alpher (2004: p. 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 & Bowern (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.

5.3.2 Wordlists

Our Pama-Nyungan phonotactic data is extracted from 112 wordlists which are part of a database under development by the last author (Round 2017c), extending the Chirila resources for Australian languages (Bowern 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 stabilise around a sufficient level of accuracy. There is some work in this space concerning frequencies of single segments (Dockum & Bowern 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 Bowern (2018b). We return to the subject of wordlist sizes and potential implications

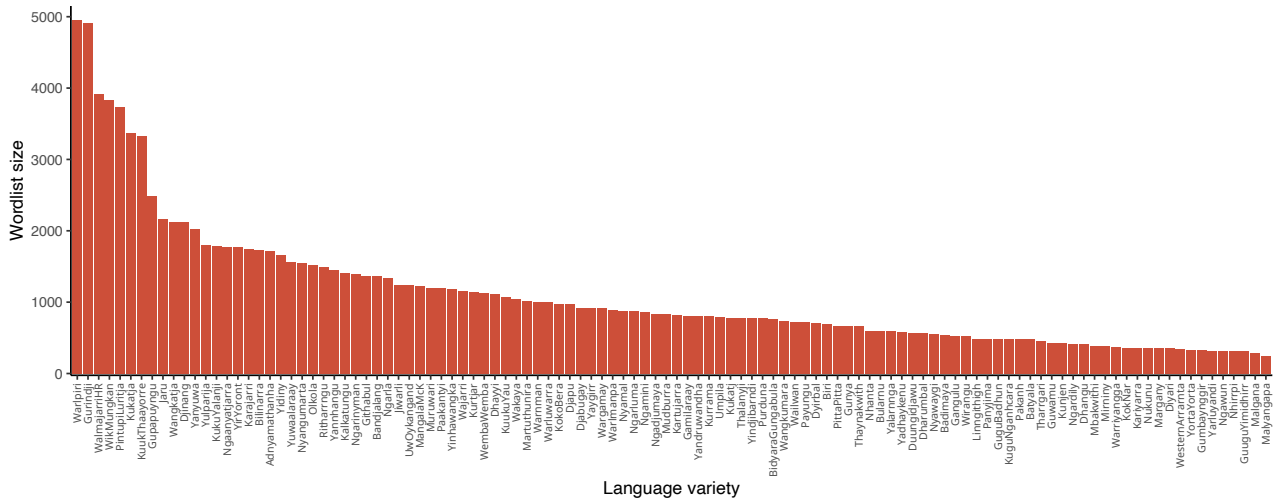

Figure 5.1: Lexicon sizes.

for our results in Section 5.7.1.

Bibliographic details for all underlying data is available in Section S2 of the Supplementary Information. 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 5.1, 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 recognised that the segmental-phonological analysis of languages is a non-deterministic process (Chao 1934, Hockett 1963, Hyman 2008, Dresher 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 2017b, Kiparsky 2018) emphasises the value of normalising source descriptions prior to the analysis of cross-linguistic phonological datasets. This is not an information destroying process—it does not 'standardise' 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 normalised is this sense. Complex segments are split into simple sequences (e.g., prenasalised 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 normalised to vowel-vowel; fortis consonants are represented as a sequence of identical short consonants, and positionally neutralised 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 2019b,a). The phonotactic character sets used in this study were extracted from these normalised, comparably segmented wordlists.

5.3.3 Reference phylogeny

The reference phylogeny we use is a maximum clade credibility tree[4] 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 Bowern & Atkinson (2012), but with an expanded and refined dataset. Further details of the model and phylogeny construction are described in Bowern & Atkinson (2012), Bowern (2015) and Bouckaert, Bowern & Atkinson (2018). The cognate dataset used to infer the reference phylogeny is available on Zenodo (Bowern 2018b). See Section S1 of the Supplementary Information for more information on the reference phylogeny.

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, Bowern & Atkinson 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, Currie & Gray 2009, Bowern et al. 2011), borrowing still has the effect of reducing branch lengths across the tree (Greenhill, Currie & Gray 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, Bowern & Atkinson 2018). Furthermore, Bouckaert, Bowern & Atkinson (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

[4]Bayesian phylogenetic methods return not a single phylogenetic tree but a posterior distribution of many possible trees. These trees can be summarised 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 Bowern & Atkinson (2012) and Bouckaert, Bowern & Atkinson (2018) for a full explanation of the methods used to infer the phylogenies considered in this section.

assumption of the methods we use in this study, and thus phylogenetic uncertainty is an important consideration. We return to this point in Section 5.7.1 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.

5.4 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 holds, 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 coarse-grained, 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 homogeneous, 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 Section 5.2. 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

[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.

cannot be distinguished from randomness nor from Brownian phylogenetic evolution.[6]

To summarise, 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 '1' and one '0' value (i.e., not every value is identical), then test as follows (otherwise discard):

- Calculate D.
- Conduct randomisation procedure to calculate p for $H_{0(D=0)}$.
- Conduct randomisation procedure to calculate p for $H_{0(D=1)}$.

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

5.4.1 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 2017b)[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 5.2). 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 5.3. 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

[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 distribution is wider and flatter than the phylogenetic distribution. We observed one instance of this for a biphone with a single '1' value and 72 '0' 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 Section 5.4.2 below.

[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 Section S3 of the Supplementary Information for usage instructions and a full description of these materials.

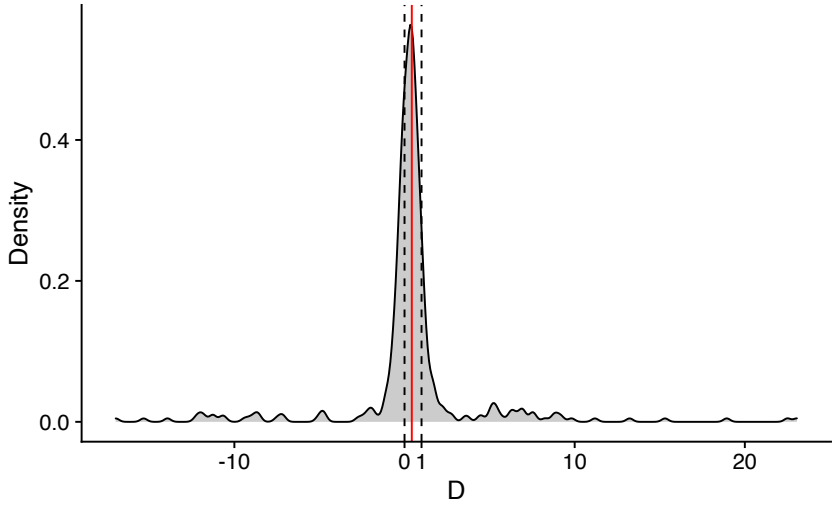

Figure 5.2: Density of D estimates for binary biphone characters. Dotted lines mark $D = 0$, the phylogenetic expectation, and $D = 1$, the random expectation. Mean D for all characters is 0.43, marked in red.

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 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.

5.4.2 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

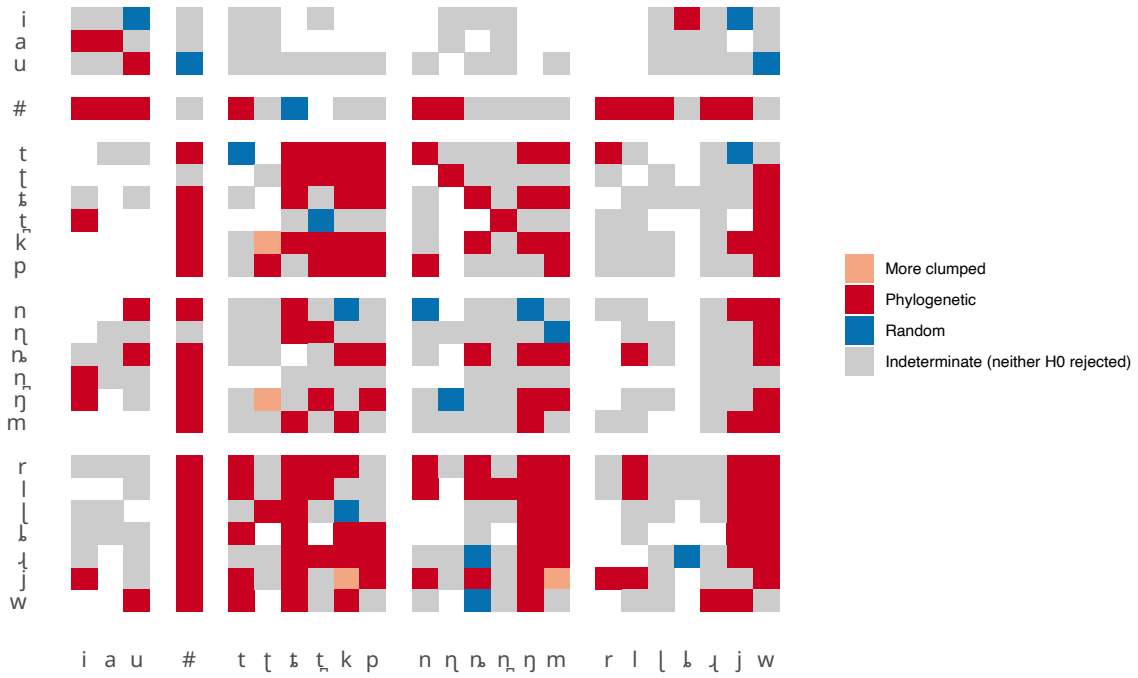

Figure 5.3: Phylogenetic signal significance testing for binary biphone characters. This grid colour-codes each biphone character according to the results of its respective significance tests. The grid is arranged such that the vertical axis represents the first segment of the biphone and the horizontal axis represents the second segment. Besides a tendency for phonotactic restrictions at word boundaries to show phylogenetic signal, few patterns stand out.

(branching patterns). To check this, in Figure 5.4(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 5.4(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, 107, 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 5.5–5.6, we plot a subset of the *D* test results, restricted to characters with less skewing.

[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.

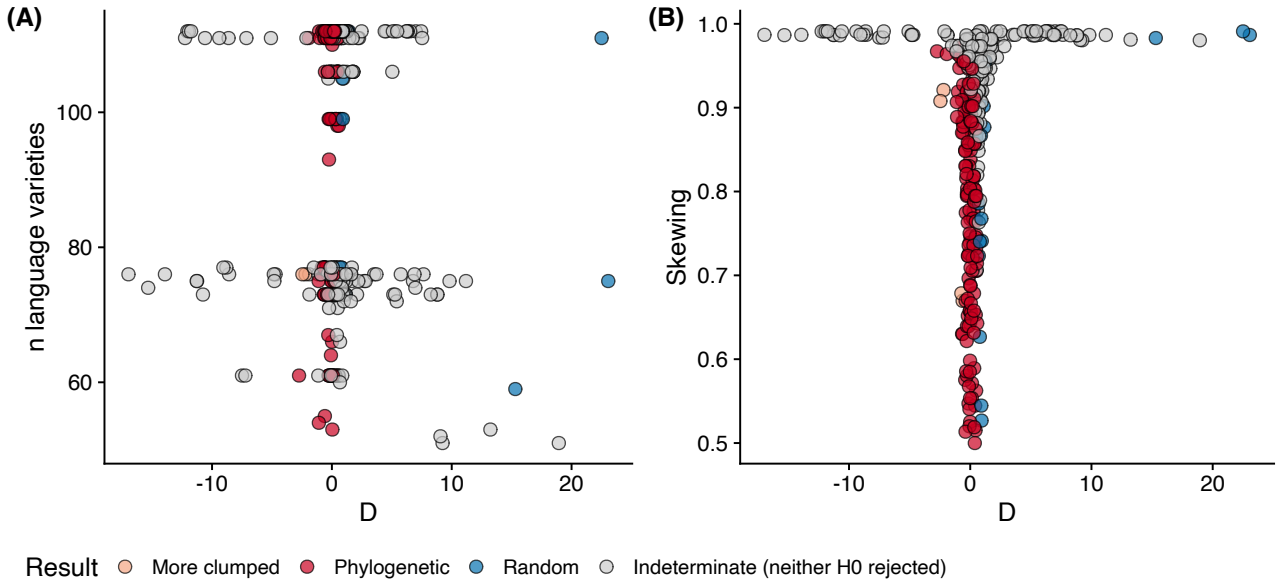

Figure 5.4: Scatterplot of *D* scores against (A) the number of doculects with non-missing values for each character and (B) the skewing of the distribution of 1s and 0s for each character. *D* clusters evenly around the mean regardless of the number of missing values. Variation in *D* increases greatly among characters where all but 1 or a few doculects share the same value, but the results are overwhelmingly not significant.

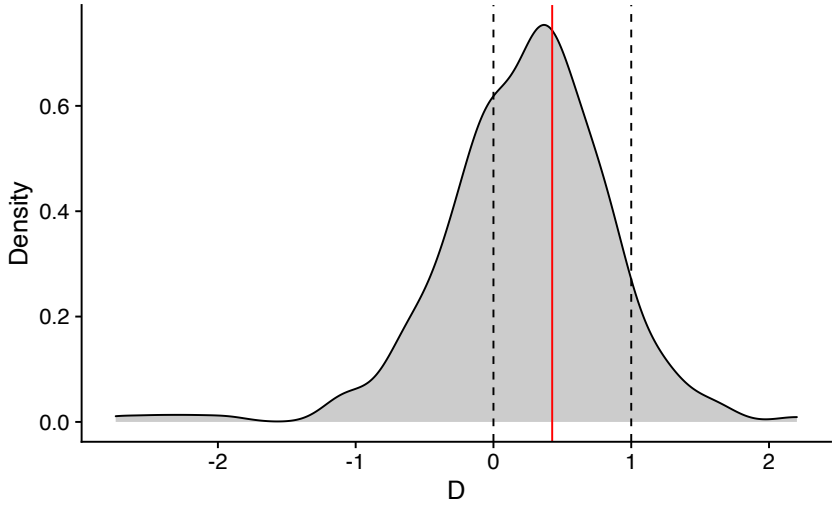

Figure 5.5: Density of *D* estimates for binary biphone characters where character values are skewed less than 97:3.

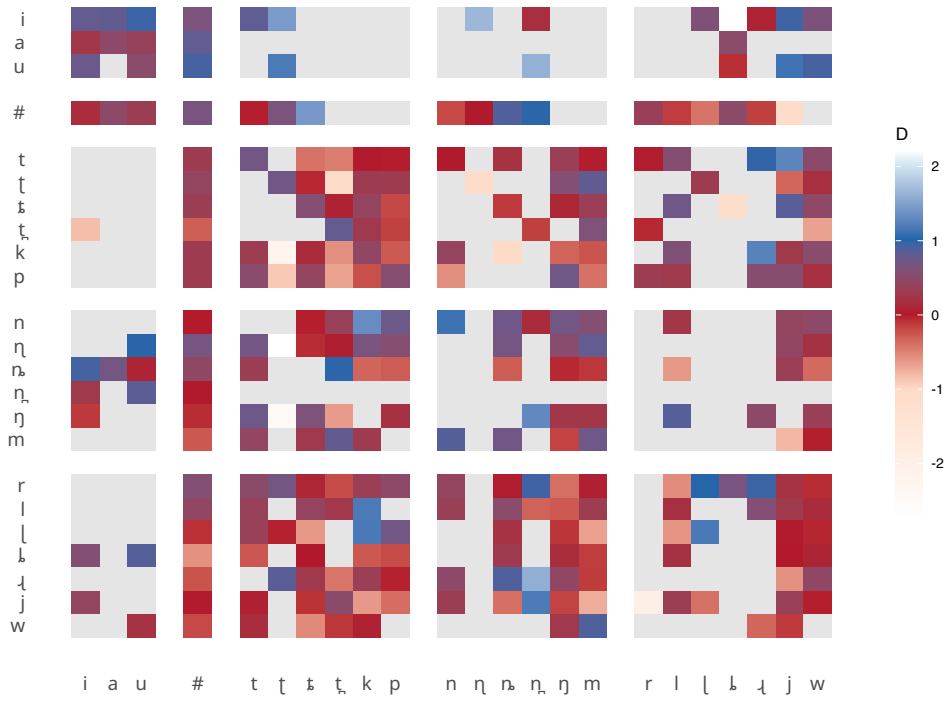

Figure 5.6: D scores for binary biphone characters where character values are skewed less than 97:3. Red shades show a character's proximity to $D = 0$—darker red indicates stronger phylogenetic signal. Blue shades show a character's proximity to $D = 1$, where character values are distributed randomly. 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.

5.5 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, normalised 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, normalised 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 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, Garland & Ives 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) randomisation procedure with 10,000 random permutations per character.

To summarise, 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 randomisation procedure to calculate p for $H_{0(K=0)}$.

Mean K for all 490 characters is 0.54 (*SD* 0.21) (Figure 5.7). This is comparable to certain physiological traits presented as examples of biological traits with a high degree of phylogenetic signal by Blomberg, Garland & Ives (2003), for example, $K = 0.55$ for log body mass of primates. Using the Blomberg, Garland & Ives (2003) randomisation 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.

We consider whether phylogenetic signal is higher or lower in certain kinds of biphone characters. Figure 5.8 shows a matrix of K scores for forward transition characters, with rows and columns arranged by phonological natural class. No clear pattern stands out.

5.5.1 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 randomisation 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

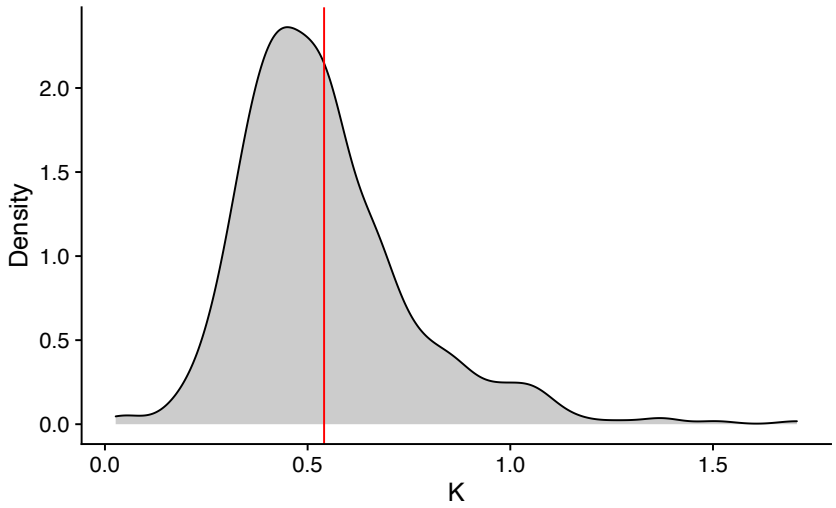

Figure 5.7: Density of K scores for all frequency-based biphone characters (frequencies of both forward and backward transitions between segments).

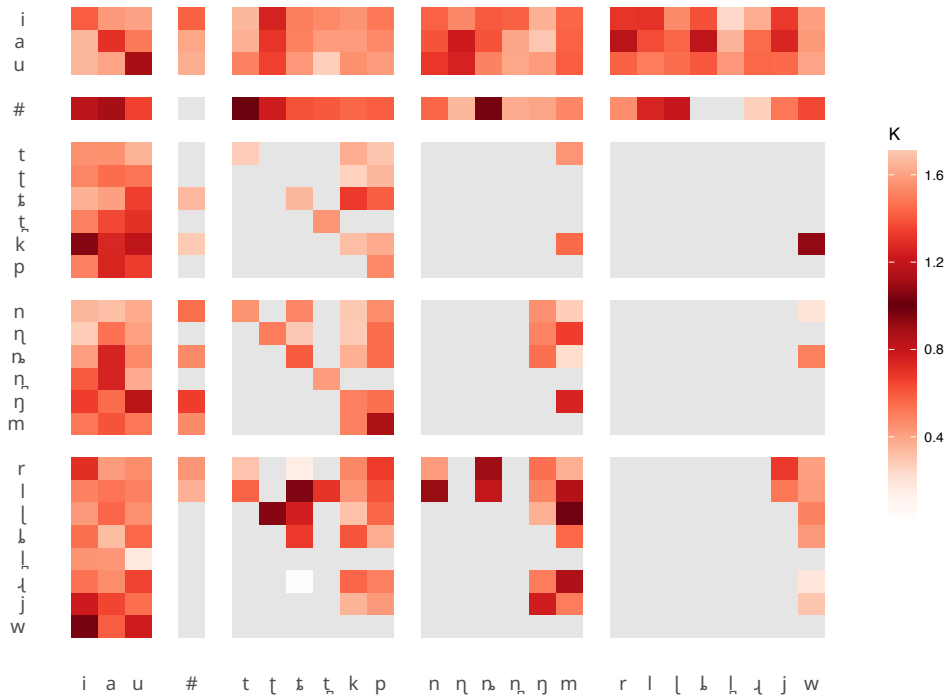

Figure 5.8: Phylogenetic signal for forward transition frequencies. This heat grid shows K scores for biphone characters (forward transition frequencies only). Each square represents a biphone (where the first segment is listed on the vertical axis and second segment on the horizontal axis). Data points 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.

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 randomisation procedure described in Section 5.2, 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 randomisation 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.

The results are plotted in Figure 5.9. 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 randomisation procedure, Figure 5.9(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 randomisation 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 randomisation 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 5.10). 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.

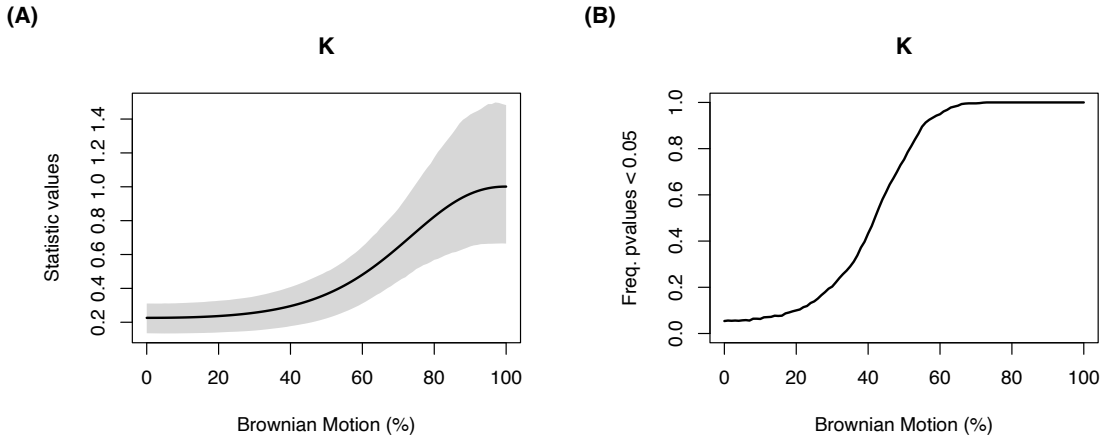

Figure 5.9: Behaviour of the *K* statistic and randomisation procedure with the Pama-Nyungan reference phylogeny. Artificial characters are simulated evolving along the phylogeny with varying levels of non-Brownian noise. When a pure Brownian motion process operates, *K* averages around 1 as expected. 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.

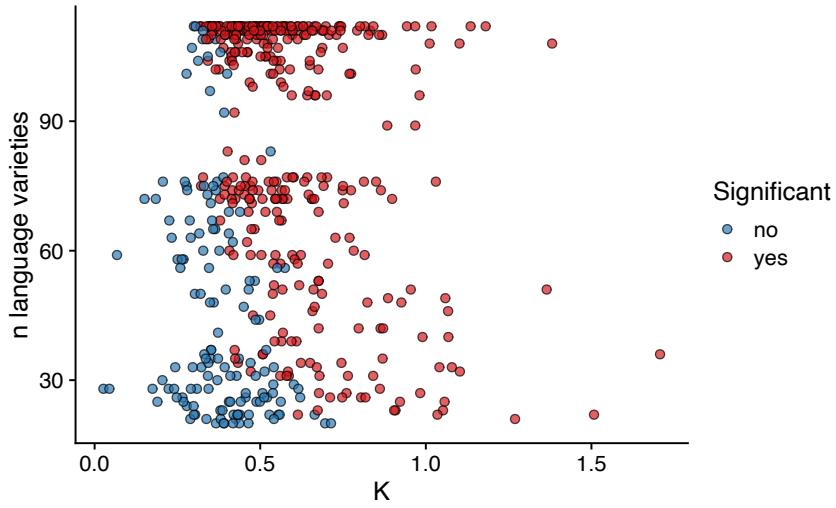

Figure 5.10: Estimates of *K* for forward and backward transition frequency characters plotted against the number of doculects with non-missing values for each character. Although some statistical power is lost and variability increases among characters with more missing values, *K* scores cluster evenly around mean *K* (0.52).

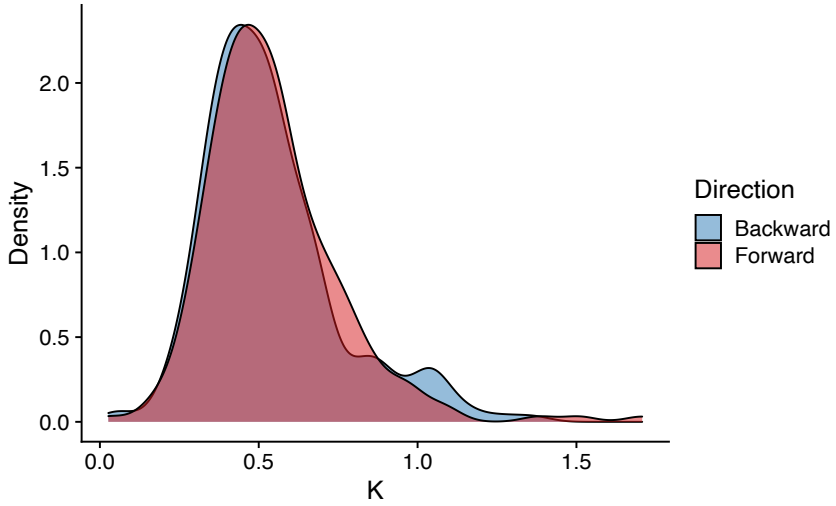

Figure 5.11: Distribution of K scores for forward transition frequencies versus backward transition frequencies. We find no significant difference between these character types.

5.5.2 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 5.11, showing a high degree of overlap between the two.

5.5.3 Normalisation 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 randomisation 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 maximises 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 normalised 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 normalised data ($t = 0.49$, $df = 972.26$, $p = 0.622$, 95% CI [-0.02, 0.03])—see Figure 5.12.

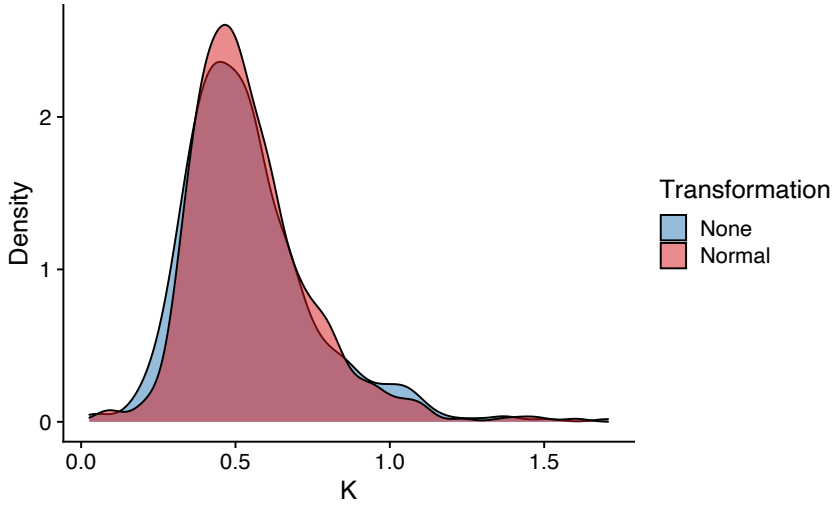

Figure 5.12: Distributions of K for untransformed character values and their normalised counterparts. We find no significant difference between these distributions.

5.6 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 organisation among segments in Australian languages (Dixon 1980, Hamilton 1996, Baker 2014, Round 2021b,a).

Table 5.1 presents mean K and the proportion of significant characters for each of these three natural-class-based datasets. All show highly similar distributions (Figure 5.13). 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

Table 5.1: 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 randomisation procedure.

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

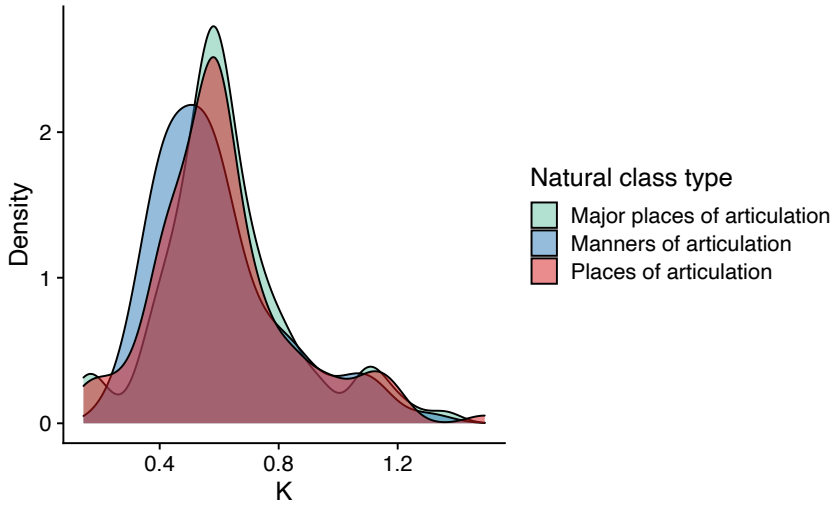

Figure 5.13: Comparison of K scores for transitions between different kinds of natural classes. The differences between all three distributions are not statistically significant.

natural-class-based datasets ($AD = 2.19$, $T.AD = 0.18$, $p = 0.329$).

5.6.1 Natural-class-based characters versus biphones

We compare the degree of phylogenetic signal in the biphone data tested in Section 5.5 to the natural-class-based data tested here (see Figure 5.14). 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 ($D = 0.21$, $p < 0.001$).

5.7 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

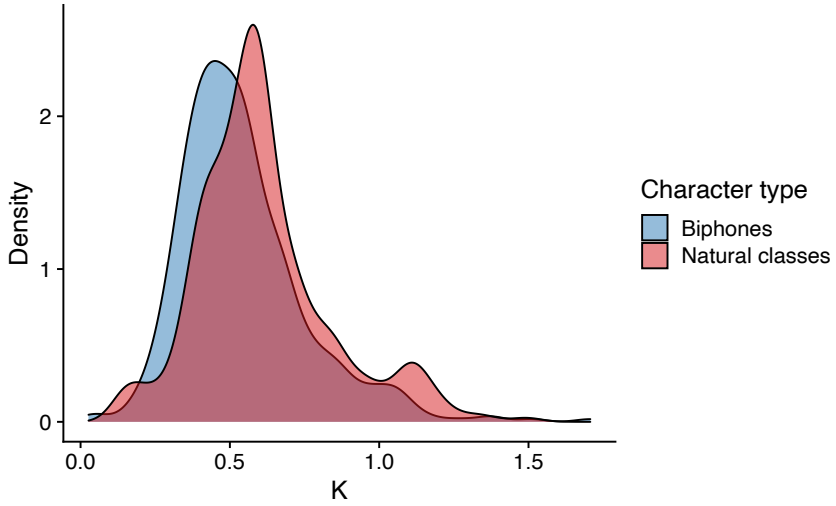

Figure 5.14: Distributions of K scores for biphone characters, coding the relative frequencies of transitions between phonological segments, and natural-class-based characters, coding the relative frequencies of transitions between natural classes of segments. Phylogenetic signal is higher overall in the natural-class-based dataset than the biphone-based dataset

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 binarised 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 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.

5.7.1 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 Section 5.3.3, 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 Section 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 Section 5.6. 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 5.15. This result suggests our results are

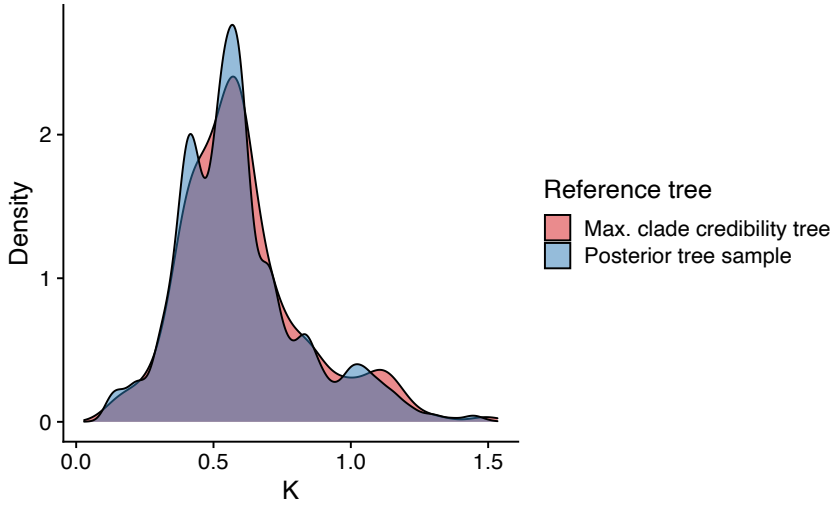

Figure 5.15: Comparison of K statistics using a 100-tree posterior sample versus the maximum clade credibility tree alone.

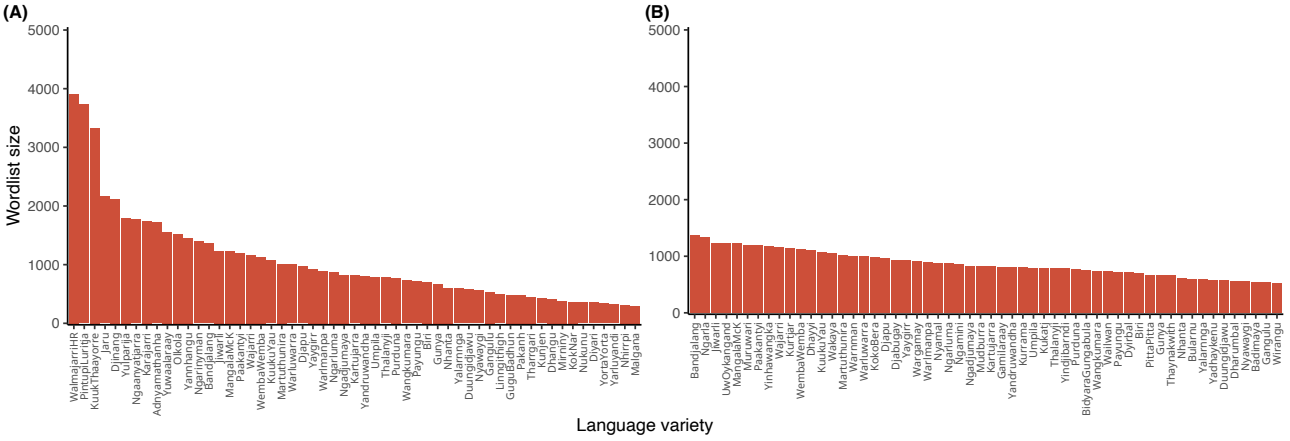

Figure 5.16: Wordlists ranked by size. (A) shows every second wordlist in the language sample. (B) shows the middle 50% of wordlists. Each subset contains the same number of wordlists but discrepancy in their size is greatly reduced in (B).

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 Section 5.3.2 we mentioned that our wordlists vary significantly is size. The length of a wordlist

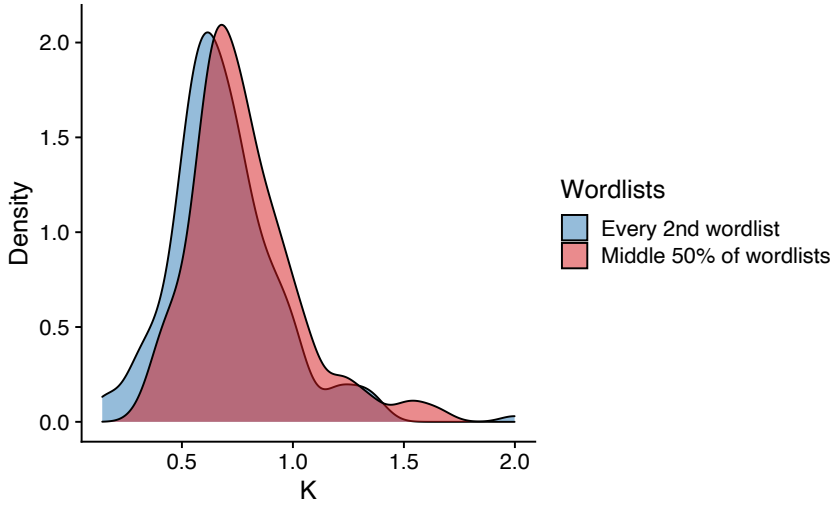

Figure 5.17: Comparison of K statistics using only wordlists falling within the 25th and 75th quantiles (middle 50%) for wordlist size, versus a sample of every 2nd wordlist (when ranked in order of size).

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 'counterbalance' 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 Sections 5.4–5.6 are unduly influenced by the disparity in our wordlist lengths, by manipulating it 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 5.16. 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 5.17.

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.

5.7.2 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 investigate how investigators' choices about how data is represented affects results.

Relevant for the current study, in Section 5.3.2 we described a process of normalisation. The motivation for this was to attempt to minimise certain aspects of variation in phonological representation that can arise from variation in how different linguists analyses the same essential facts (Chao 1934, Hockett 1963, Hyman 2008, Dresher 2009). While normalisation 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 normalise 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, neutralise 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 2017b). An advantage of our general approach, is that it open 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 that was 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 parameterise 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, Passmore & Atkinson 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 Lévy 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 Lévy-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).

5.8 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.

Acknowledgements

We wish to thank David Nash and two anonymous reviewers for their generous and constructive feedback throughout preparation of this manuscript. This research is funded by an Australian Government Research Training Program (RTP) Scholarship to JM-C, Australian Research Council (ARC) grant DE150101024 and ARC Centre of Excellence for the Dynamics of Language grant TIG322015 to ER and National Science Foundation (NSF) grant 1423711 to CB. We gratefully acknowledge this support.

The following manuscript has been incorporated as Chapter 6.

Jayden L. Macklin-Cordes, Gabriel Hassler & Erich R. Round. Manuscript. Pama-Nyungan tree inference with phonotactics.

Contributor	Statement of contribution	%
Jayden Macklin-Cordes	initial concept	60
	scripting, analysis	80
	writing of text	100
	proof-reading	10
	preparation of figures	100
Gabriel Hassler	methodological guidance	50
	scripting, analysis	20
Erich Round	initial concept	40
	supplying data	100
	supervision, guidance	50
	proof-reading	90

The following chapter is a manuscript intended for publication as a standalone article. Erich Round supplied lexical data from the Ausphon Lexicon database, proof-read the draft manuscript and provided general guidance throughout. Gabriel Hassler and Simon Greenhill provided guidance on the experimental design. Hassler provided important methodological guidance and implemented updates to BEAST software necessary to run the analysis. Macklin-Cordes developed the concept and methodology, wrote the manuscript, ran the analysis and created figures.

Supplementary code and original output files are deposited at The University of Queensland. The materials are open access and available at following link: https://doi.org/10.14264/4e5077e

Code and materials are also available on Github: https://github.com/JaydenM-C/PN_treebuilding. The study can be reproduced from the materials on Github; only the original output files are missing due to Github's file size constraints.

Chapter 6

Pama-Nyungan tree inference with phonotactics

Historical linguistics in the 21st century has been shaped by a quantitative turn. Computational phylogenetic methods have been used to infer detailed trees for an expansive array of language families around the world. However, data innovation remains a priority for methodological development in this field. Most linguistic phylogenies are inferred from lexical cognate data alone, leaving untapped sources of evidence from other parts of language. To this end, prior study suggests that phonotactics, including frequency data from biphone sequences in wordlists, could be a novel source of phylogenetic signal. In this paper, we evaluate the capacity to capture the phylogenetic signal in phonotactics for the purposes of tree inference. We test whether the inclusion of phonotactic data, in conjunction with lexical cognate data, can strengthen phylogenetic tree inference. To do this, we take 2,236 binary phonotactic variables and 133 frequency-based phonotactic variables and combine them with existing lexical cognate data from 44 western Pama-Nyungan languages. We run a Bayesian Markov chain Monte Carlo (MCMC) approach on two competing phylogenetic models. In one, the tree is inferred jointly using the phonotactic data and lexical cognate data. In the other, two separate trees are inferred, one from phonotactic data and the other from cognate data alone. The models are compared by estimating marginal likelihoods and calculating a Bayes factor. We also compare the competing tree topologies and clade support values. We return indeterminate results. We find that compromises to the evolutionary model for frequency data, which were necessary to make the study computationally feasible, had the undesirable effect of preventing consistently stable MCMC convergence and reliable marginal likelihood estimation. We discuss the reasons for this and provide an outline of future research steps to evaluate whether these findings support a true negative result or merely reflect current limitations of the methodology.

6.1 Introduction

Methodological development in 21st century historical linguistics has been defined by the uptake and increasing ubiquity of quantitative methods, and in particular Bayesian phylogenetic methods for inferring language trees. Language trees, or phylogenies, have been inferred using these methods for language families in Australia (Pama-Nyungan: Bowern & Atkinson 2012), Asia (Aslian: Dunn, Burenhult, et al. 2011, Dravidian: Kolipakam et al. 2018), Africa (Bantu: Whiteley, Xue & Wheeler 2019), Oceania (Austronesian: Gray, Drummond & Greenhill 2009), the Americas (Chapacuran: Birchall, Dunn & Greenhill 2016, Tupí-Guaraní: Michael et al. 2015), and Europe (Indo-European: Gray & Atkinson 2003, Chang et al. 2015) (to select just a few examples). The task of phylogenetic tree inference is intrinsically interesting to historical linguistics, which is centrally concerned with the genealogical classification of languages. But it is important beyond historical linguistics too. Access to highly detailed phylogenies is necessary for the continued uptake of phylogenetic comparative methods in typology (see 4), and phylogenies can be combined with other sources of evidence to investigate migration patterns and other questions of human history (e.g. Hunley et al. 2008, Bouckaert et al. 2012b, Malaspinas et al. 2016, Bouckaert, Bowern & Atkinson 2018).

With a few exceptions (e.g. Dunn et al. 2005, 2008, Reesink, Singer & Dunn 2009, Sicoli & Holton 2014), most quantitative historical linguistic studies have inferred trees exclusively from lexical cognate data—binary characters coding whether each language contains a member of a set of related word forms in its lexicon. Phylogenetic tree inference from lexical data has been a tremendously successful enterprise. Results have tended to align with traditional historical linguistic classifications where previous classifications exist, while adding detail to the tree and generating new insights (for example, the identification of major divisions within the Pama-Nyungan family, linking smaller, previously-identified subgroups, based on early phylogenetic splits inferred by Bowern & Atkinson 2012). Nevertheless, lexical cognate data in isolation has some limitations. It is taxing to acquire, requiring high quality language documentation and considerable expertise on the part of the linguist to identify and distinguish genuine cognate words from chance resemblances. The historical signal in lexical data also contains noise from instances of lexical borrowing and undetected semantic shifts.

It is pertinent to consider whether other forms of linguistic data could supplement lexical data in linguistic phylogenetics—in this instance, phonotactic data that is relatively easy to extract automatically from wordlists. Interestingly, a similar parallel discussion is taking place in other phylogenetic sciences. In biology, Parins-Fukuchi (2018) find that combining continuous-valued anatomical morphological characters with more traditional, categorically-valued genomic data can strengthen tree inference. In linguistics, Macklin-Cordes, Bowern & Round (2021) found phylogenetic signal in phonotactic data, suggesting that phonotactic data could be a profitable addition for inferring phylogenetic trees. That study, however, was limited to measuring phylogenetic signal by comparing the phonotactic data to an existing phylogenetic tree. This study goes further by putting phonotactics into the actual practice of phylogenetic tree inference and evaluating whether this produces stronger results than phylogenetic tree inference based on lexical cognate data alone.

In this study, we test the question of whether phylogenetic tree inference can be strengthened by inferring the phylogeny from a combination of lexical cognate data and phonotactic data, compared to more traditional phylogenetic tree inference using cognate data in isolation. To test this, we take a sample of 44 languages from the western branch of the Pama-Nyungan family, Australia. We take previously published lexical cognate data (Bouckaert, Bowern & Atkinson 2018) and extract phonotactic data from wordlists from the Ausphon Lexicon database (Round 2017c), a comparably segmented, phonemicised extension of the CHIRILA database (Bowern 2016). Two phonotactic datasets are extracted, following Macklin-Cordes, Bowern & Round (2021). One is binary, coding the presence or absence of two-segment sequences (biphones) in a given language. The other is a sound class transition frequency dataset which records the frequency of a sequence of two natural sound classes, *xy*, relative to all instances of the sound class *x* (i.e. for all instances of sound class *x*, how often is it followed by the sound class *y*?). We compare two phylogenetic models, one in which a single tree is inferred jointly from the cognate data and phonotactic data, and a second model containing two trees, one inferred from cognate data and the other inferred from phonotactic data separately. We compare the two using a standard, well-established method of model comparison in phylogenetics, which is to infer marginal likelihood estimates for each model and compare them by calculating the Bayes factor (Kass & Raftery 1995, Brown & Lemmon 2007). In this instance, our model comparison returns indeterminate results due to the failure of our phylogenetic models to converge consistently. The challenge we encounter is striking the right balance between a model that is rich and linguistically plausible, while simultaneously remaining within current computational constraints. This challenge, though difficult, is perhaps not entirely intractable, and we conclude the paper with a detailed discussion of what will be required in future research to evaluate phonotactics in phylogenetics more fully.

The paper proceeds with the methodology in Section 6.2, defining the research question, detailing the overarching structure of the experimental section that follows, and describing the construction of the language sample. There are four experimental sections that follow: A first preliminary test (Section 6.3), a second preliminary test (Section 6.4), the main test (Section 6.5) and a follow-up test (Section 6.6). The discussion (Section 6.7) and conclusion (Section 6.8) detail the study's limitations and future directions.

6.2 Methodology

6.2.1 Research question

This study tests the question of whether or not the addition of phonotactic data strengthens linguistic phylogenetic tree inference. To test this question, we infer a phylogeny of the western branch of Pama-Nyungan languages using lexical cognate data and phonotactic data and repeat this process twice—once in which the phylogeny is inferred jointly from cognate and phonotactic data and once in which separate trees are inferred from cognate data and phonotactic data. The strength of phylogenetic

inference in each instance is evaluated by estimating and comparing log marginal likelihoods for each, and comparing the topologies and posterior clade support values of maximum clade credibility trees.

6.2.2 Experimental design

There are four test components in this study: two preliminary test components and two main test components. Preliminary testing consists of (1) evaluating the best-fitting evolutionary model for binary biphone data, and (2) tree inference using cognate data only, essentially aiming to replicate Bouckaert, Bowern & Atkinson (2018). Model comparison is a regular step in linguistic phylogenetic studies, as typically a number of clock models, site models and other parameters are considered. In this study, the cognate data is sourced from a prior study (Bouckaert, Bowern & Atkinson 2018) which already evaluated the best-fitting model for this particular set of data. Accordingly, we pass over this evaluative process and replicate as much as possible the best supported model and priors from Bouckaert, Bowern & Atkinson (2018), which includes a covarion model with a relaxed clock and fixed rates across cognate classes. It is a different story, however, for the binary biphone data which has not been used previously for phylogenetic tree inference. The novelty of this data source necessitates a thorough evaluation of which evolutionary model can best be applied to it. This is the subject of the first preliminary test.

The second preliminary test involves inferring a phylogeny using only cognate data from Bouckaert, Bowern & Atkinson (2018). This test functions as a sanity check to ensure that our phylogenetic model and software implementation, which approximates but does not exactly replicate Bouckaert, Bowern & Atkinson (2018), produces suitably equivalent results.

These preliminary tests are followed by the main test evaluating the primary research question described above. We use the Bayesian Markov chain Monte Carlo (MCMC) method implemented in BEAST phylogenetic software to infer a phylogeny of western Pama-Nyungan using cognate data, binary biphone data and continuous phonotactic data consisting of sound class transition frequencies. In a second MCMC run, we include two tree models, one inferred with cognate data only and another with phonotactic data only (both binary and continuous). Marginal likelihoods are estimated using the stepping stone sampling method (Baele et al. 2013). For each tree model, we produce a maximum clade credibility tree from the posterior sample of trees that BEAST produces and then compare the topologies and clade support values of these.

We conduct one follow-up test after this. To test our suspicion that the binary phonotactic data was contributing undue weight to likelihood calculations without contributing much phylogenetic information, we re-ran the experiment with this data partition removed. In this instance, the second BEAST run contains two tree models, one inferred with cognate data only and one inferred using cognate data and continuous phonotactic data together.

There are a couple of additional points on software and model design that hold throughout all four tests. Firstly, regarding the evolutionary model for frequency-based phonotactic characters, we take a standard, lightweight, Brownian motion model in which frequency values can wander up or down

with equal probability through time. We are currently limited to this model by software constraints, but accept it as a reasonable starting point. Firstly, Brownian motion is a common starting point in comparable biological studies that jointly infer trees with continuous data. Secondly, it follows the model used in Macklin-Cordes, Bowern & Round (2021). One difference between Macklin-Cordes, Bowern & Round (2021) and this study is that Macklin-Cordes, Bowern & Round (2021) use raw frequency values whereas we use logit-transformed frequency values.

We use a Bayesian computational approach to infer linguistic phylogenies using BEAST phylogenetic software (v1.10.5) (Suchard et al. 2018). This is similar to earlier work on the Pama-Nyungan phylogeny (Bowern & Atkinson 2012, Bouckaert, Bowern & Atkinson 2018) which used BEAST2 (Bouckaert et al. 2019). We selected BEAST over BEAST2 because it offers the ability to infer trees with continuous characters. Throughout, we generally follow Bouckaert, Bowern & Atkinson (2018) as closely as possible. We follow Bouckaert, Bowern & Atkinson (2018) in constraining the tree topology using clade priors for well-established and commonly accepted Pama-Nyungan subgroups, as established by O'Grady, Voegelin & Voegelin (1966), Mühlhäusler, Tryon & Wurm (1996) and Koch (2014) and subsequently recovered in computational phylogenetic analysis by Bowern & Atkinson (2012). Dating the Pama-Nyungan tree is a central focus of Bouckaert, Bowern & Atkinson (2018), combining lexical cognate data with geographical data and archaeological calibration points to give a best-available estimate of the geographic and temporal point of origin of the family. Accordingly, we retain their calibration prior on the Wati subgroup, which places a 95% probability of the subgroup's origin dating between 3,000-5,000 years, with most of the probability density skewing towards the younger end of that range (a gamma distribution of $\alpha = 2$, $\beta = 359$, with 3,000 year offset) based on a synthesis of archaeological evidence (see Bouckaert, Bowern & Atkinson 2018: p. 746). We place a prior on the root age of the Pama-Nyungan family centred on a mean of 5,791 years B.P., following the findings of Bouckaert, Bowern & Atkinson (2018). 5,791 years is the mean root age of the posterior for their best supported hypothesis on Pama-Nyungan's origins. We model this as a normal distribution (SD = 730) approximating the 95% range of posterior root age estimates. One aspect in which we differ from Bouckaert, Bowern & Atkinson (2018) is tip dates. Bouckaert, Bowern & Atkinson (2018) use a birth-death skyline tree model which allows for tip dates to differ and includes a parameter corresponding to the proportion of total taxa sampled at a given point in time. This is reasonable since they use language sources span over 200 years. In contrast, we assume all tips are contemporaneous. In our case, since we restrict attention to relatively modern sources, any extra precision to be gained from including tip dates is not worth the reduced tree model choice in BEAST and extra computational expense.

6.2.3 Language sample

The target language sample, in the first instance, is the 306 Pama-Nyungan language varieties represented in Bouckaert, Bowern & Atkinson (2018). Of these, we restrict attention to languages meeting the following criteria. Firstly, the language must be represented in the Ausphon Lexicon database

(Round 2017c) from which we source wordlist data. Secondly, the original wordlist data source must have been compiled by a trained linguist from primary fieldwork with living speakers or a combination of fieldwork and archival materials (no sources reconstituted only from archival materials). Thirdly, each wordlist must contain at least 250 lexical items. This leaves a subset of 112 Pama-Nyungan languages. In Preliminary Test 1, where computational demands are relatively minimal, this is the language sample we use. Due to computational constraints associated with large amounts of continuous data and large phylogenies, in the main test we restrict attention to the western branch of the Pama-Nyungan family identified in Bowern & Atkinson (2012) and Bouckaert, Bowern & Atkinson (2018). This gives a sample of 44 western Pama-Nyungan languages covering nine Pama-Nyungan subgroups. A list of the bibliographic details of original language sources is available in Appendix B.

6.3 Preliminary test 1: Evolutionary model for binary biphone data

As the number of linguistic phylogenetic studies using lexical cognate data has grown, some consistent findings have emerged; for example, the covarion model now seems widely preferred (e.g. Bouckaert, Bowern & Atkinson 2018, Kolipakam et al. 2018, Savelyev & Robbeets 2020). However, this is to the best of our knowledge the first attempt at Bayesian phylogenetic tree inference with binary biphone characters (though see Jäger & Wichmann (2016) for a similar approach using a distance-based tree inferencing method). Accordingly, we start by embarking on the process of model testing and selection for this novel data type. The aim is to identify a sensible model and set of priors that we can specify for the binary biphone data in subsequent testing.

For each model specification, we run two independent MCMC chains of 25,000,000 iterations, with parameters logged every 10,000 iterations. Log marginal likelihoods are estimated using BEAST's path sampling/stepping stone sampling procedure (Baele et al. 2013) consisting of 50 path steps of 500,000 iterations, with parameters logged every 10,000 iterations, conducted on each chain then combined to get an overall marginal likelihood estimate (MLE). We conducted autocorrelation and convergence checks using Tracer v1.7.1 software (Rambaut et al. 2018). Note that the results here are a preliminary exploration of model parameters to determine the best parameter settings for the tree inference presented in Section 6.5 below. We do not anticipate that binary biphone characters will produce especially high quality or realistic language phylogenies on their own. The goal is to get a handle on how best to model the evolutionary dynamics of this dataset when used in combination with other sources of evidence.

We test a total of 16 models, consisting of each logically possible combination of site and clock model components. We describe each of these alternatives in turn below.

6.3.1 Binary biphone data

A language's phonotactic system consists of rules governing how phonemic segments may combine into larger syllables and words. To represent phonotactics, we extract data on the presence and frequencies of *biphones*, two-segment sequences, from language wordlists. From each wordlist, we extract data on the presence and absence of *biphones*, sequences of two segments (where each segment is either a phoneme or a word boundary). A biphone is marked '1' if it is present anywhere in a language's wordlist. If the biphone consists of two segments that are part of the language's phonemic inventory (and therefore the biphone could, in principle, occur in the language) but the biphone never occurs, it is marked '0' for absent. If one or both segments in the biphone are not part of the language's phonemic inventory, then it is marked as a gap '-' in the data. A total of 2236 binary biphone characters are extracted.

6.3.2 Site and clock settings

Site models describe how binary biphone characters evolve through time. The site model is defined by three parameters, giving eight possible combinations to test:

For this stage of evaluation, we fix the clock model to a strict clock (no variation in evolutionary rates between branches) and fix the tree model to a simple calibrated Yule tree model with a uniform birth rate prior (Yule tree models do not allow for extinction events). We then test all eight combinations of three site model parameters:

- A simple continuous time Markov chain (CTMC) model (which contains a single estimated parameter that specifies the frequencies with which biphones are gained and lost) versus a covarion model (which allows sites to switch between fast and slow states). The covarion model is the preferred model of lexical cognate evolution in Bouckaert et al. (2012a), Bouckaert, Bowern & Atkinson (2018) and Kolipakam et al. (2018), although Chang et al. (2015: p. 219) find little difference between them and opt for the increased simplicity of the former model.
- Empirical character state frequencies versus estimated character state frequencies.
- Site homogeneity (fixed evolutionary rates across all character sites) versus heterogeneity (estimated using four gamma distributed categories, following Kolipakam et al. (2018)). For cognate data, Bouckaert, Bowern & Atkinson (2018) find a better fit with homogeneous rates but Kolipakam et al. (2018) find a better fit with heterogenous ones.

We use Bayes factors (Kass & Raftery 1995) to determine the best supported site, clock and tree models. Bayes factors give an indication of the support for one model over another and are calculated by calculating the ratio of the log marginal likelihoods of each model. A Bayes factor of 5 to 20 is taken as substantial support, greater than 20 as strong support, and greater than 100 as decisive (Kass & Raftery 1995). We table Bayes factors comparing each combination of model settings in Table 6.1. The names of each model indicate site settings as follows: (S)imple CTMC versus (C)ovarion model, e(M)pirical versus e(S)timated character frequencies, (H)omogenous rates versus (G)amma-distributed

heterogenous rates. So, for example, the model termed "CMH" consists of a covarion model with empirical frequencies and homogeneous rates across all sites.

We test two clock models: A strict clock, in which a single evolutionary rate is fixed across all branches in the tree, and a lognormally-distributed, uncorrelated relaxed clock. This relaxed clock model generally has been found to outperform a strict clock when modelling lexical cognate evolution (Bouckaert, Bowern & Atkinson 2018, Kolipakam et al. 2018). Clock settings are denoted in Table 6.1 by 'S' for strict and 'R' for relaxed. So, for example, the model termed "SSG-S" is a simple site model with estimated character frequencies and gamma-distributed heterogenous rates, combined with a strict clock.

For the relaxed clock, we used an uncorrelated lognormal setting with a uniform prior [0,1] following Kolipakam et al. (2018). Bouckaert, Bowern & Atkinson (2018) constrain the upper bound to 1.0E-4 to reduce burn-in time since, in practice, the mean clock value never approaches even that level. We chose the less informative upper bound given the uncertainty of working with a novel data type.

A third Bayesian phylogenetic model component is the tree model which defines the speciation process. Two main alternatives appear frequently in linguistic phylogenetic research. These are birth-death speciation models, which allow for extinction events, and Yule speciation models, which allow birth events only (as preferred in Bowern & Atkinson 2012, Kolipakam et al. 2018). Bouckaert, Bowern & Atkinson (2018) use a Birth-Death Skyline model which enables tips to be sampled at different dates, but fix the death rate to zero. This is reasonable since they use language sources span over 200 years. Throughout this study, we follow this precedent by using a birth-death model with the death rate set to zero. However, instead of including different tip ages, we assume all tips are contemporaneous and set their ages to 0, which is not strictly accurate since original wordlist sources do vary in age. In our case, since we restrict attention to relatively modern sources, any extra precision to be gained from including tip dates is not worth the reduced tree model choice in BEAST and extra computational expense. One additional parameter that we do include instead is an incomplete sampling parameter to account for the fact that our language sample is only a subset of the full 306 languages included in Bouckaert, Bowern & Atkinson (2018). This incomplete sampling parameter is set to 0.366 in this preliminary test ($\frac{112}{306}$).

6.3.3 Results

Bayes factors of pairwise comparisons between candidate models are listed in Table 6.1. These results can be broken down into some key findings as follows:

- The covarion model overwhelmingly outperforms the simple CTMC model in all instances.
- Variable evolutionary rates are favoured over fixed rates. This will increase the computational demand of models in the main experiment later on, since it increases the number of parameters to be estimated. Models with heterogenous rates require 3–4 times as long as equivalent models with fixed rates.

- There is decisive support for estimating character state frequencies rather than simply taking the observed frequencies when the covarion model is used, although the opposite is true with a CTMC model.
- Using estimated frequencies seems more important than using heterogenous rates. A covarion model with estimated frequencies and disfavoured, homogeneous evolutionary rates will score higher than a covarion model with heterogenous rates but disfavoured empirical frequencies.

All up, we determine the best site model to be a covarion model with estimated frequencies and rate heterogeneity.

With regards to the clock model, site models paired with a relaxed clock tend to do better than their direct equivalents paired with a strict clock, though covarion models with homogeneous rates are exceptions. The best site model, denoted 'CSG' (covarion, estimated frequencies, gamma heterogenous rates), scores decisively higher than all others regardless of which clock model is used. The 'winning' model, which we proceed to apply to binary biphone data throughout the rest of this study, is the CSG model with a relaxed clock ('CSG-R').

One limitation to note is that we have not considered the stochastic Dollo model, which has been implemented with some success for cognate data in linguistics (Bowern & Atkinson 2012) (although the covarion model was subsequently found to be better in Bouckaert, Bowern & Atkinson (2018)). Stochastic Dollo only allows characters to spring into existence once and any losses are permanent. Such a model is perhaps a bit more realistic for cognates, since the state space of possible words is practically infinite (i.e. the chance of different people inventing the same word for the same thing independently is low, although of course it does happen sometimes)[1]. By contrast, there are only so many possible biphone combinations, many unrelated/distantly related languages share biphones (consider, for example, shared biphones between English and Pama-Nyungan languages) and it seems unreasonable to assume a single common point of origin for all of them. For this reason we disregarded the Dollo model for this study.

6.4 Preliminary test 2: Tree inference using cognate data only

This second preliminary test aims to replicate the results of Bouckaert, Bowern & Atkinson (2018) using the same cognate data and approximately the same evolutionary model. The goal is to ensure that our particular software implementation and slight differences in implementation of the evolutionary

[1] As an aside, the stochastic Dollo model is not particularly realistic for cognates either since it does not allow for borrowing, which manifests as two independent origin points when plotted on a phylogenetic tree. The presence of borrowing likely explains why the covarion model tends to work better than stochastic Dollo in linguistic phylogenetic studies. As an aside, an ideal phylogeographic model of cognate evolution would allow for independent points of origin at two rates of likelihood: a very low rate of likelihood of the cognate originating independently anywhere throughout the tree (capturing the likelihood of undetected chance resemblances coded as cognates), and a relatively high likelihood of a cognate independently originating in languages that are geographically adjacent to a cognate where the cognate is already present (to capture effectively borrowing events rather than genuine instances of convergent evolution). However, this would almost certainly be computationally expensive.

Table 6.1: Bayes factors for different site models. Each Bayes Factor represents the support for one model (vertical axis) against another (horizontal). A positive value indicates the first model (left) is supported, and conversely, a negative value indicates the second model (top) is supported. A value over 100 is considered decisive.

Site model	CMG-R	CMG-S	CMH-R	CMH-S	CSG-R	CSG-S	CSH-R	CSH-S	SMG-R	SMG-S	SMH-R	SMH-S	SSG-R	SSG-S	SSH-R	SSH-S
CMG-R	–	20,274	44,531	20,746	-33,840	-16,924	42,417	22,034	59,797	59,966	61,049	61,250	59,793	59,932	61,036	61,246
CMG-S	-20,274	–	24,257	472	-54,114	-37,198	22,143	1,760	39,523	39,692	40,775	40,976	39,519	39,658	40,762	40,972
CMH-R	-44,531	-24,257	–	-23,785	-78,371	-61,455	-2,114	-22,497	15,266	15,435	16,518	16,719	15,262	15,401	16,505	16,715
CMH-S	-20,746	-472	23,785	–	-54,586	-37,670	21,671	1,288	39,051	39,220	40,303	40,504	39,047	39,186	40,290	40,500
CSG-R	33,840	54,114	78,371	54,586	–	16,916	76,257	55,874	93,637	93,806	94,889	95,090	93,633	93,772	94,876	95,086
CSG-S	16,924	37,198	61,455	37,670	-16,916	–	59,341	38,958	76,721	76,890	77,973	78,174	76,717	76,856	77,960	78,170
CSH-R	-42,417	-22,143	2,114	-21,671	-76,257	-59,341	–	-20,383	17,380	17,549	18,632	18,833	17,376	17,515	18,619	18,829
CSH-S	-22,034	-1,760	22,497	-1,288	-55,874	-38,958	20,383	–	37,763	37,932	39,015	39,216	37,759	37,898	39,002	39,212
SMG-R	-59,797	-39,523	-15,266	-39,051	-93,637	-76,721	-17,380	-37,763	–	169	1,252	1,453	-4	135	1,239	1,449
SMG-S	-59,966	-39,692	-15,435	-39,220	-93,806	-76,890	-17,549	-37,932	-169	–	1,083	1,284	-173	-34	1,070	1,280
SMH-R	-61,049	-40,775	-16,518	-40,303	-94,889	-77,973	-18,632	-39,015	-1,252	-1,083	–	201	-1,256	-1,117	-13	197
SMH-S	-61,250	-40,976	-16,719	-40,504	-95,090	-78,174	-18,833	-39,216	-1,453	-1,284	-201	–	-1,457	-1,318	-214	-4
SSG-R	-59,793	-39,519	-15,262	-39,047	-93,633	-76,717	-17,376	-37,759	4	173	1,256	1,457	–	139	1,243	1,453
SSG-S	-59,932	-39,658	-15,401	-39,186	-93,772	-76,856	-17,515	-37,898	-135	34	1,117	1,318	-139	–	1,104	1,314
SSH-R	-61,036	-40,762	-16,505	-40,290	-94,876	-77,960	-18,619	-39,002	-1,239	-1,070	13	214	-1,243	-1,104	–	210
SSH-S	-61,246	-40,972	-16,715	-40,500	-95,086	-78,170	-18,829	-39,212	-1,449	-1,280	-197	4	-1,453	-1,314	-210	–

model do not unduly impact the results, which we then use as a baseline for comparison in the main test below.

Differences between the implementation of this test and Bouckaert, Bowern & Atkinson (2018) are as follows. Firstly, the language sample is greatly reduced. Owing to the computational demands of inferring a large phylogeny with a large quantity of continuous-valued data, the language sample for the main test is reduced to 44 western Pama-Nyungan languages and so this is the language sample we use here. Secondly, we include an incomplete sampling parameter of 0.537, to account for the fact that our language sample covers 44 of the 82 languages in the western branch of the Bouckaert, Bowern & Atkinson (2018) phylogeny. Another difference is that, as described above, we do not include divergent tip ages in our model, although the 44 languages we use are sourced from modern sources anyway. Finally, we run the model in BEAST software (v1.10.5) (Suchard et al. 2018) rather than BEAST2 (Bouckaert et al. 2019). We selected BEAST over BEAST2 because it offers the ability to infer trees with continuous characters (necessary for the main test below).

We follow Bouckaert, Bowern & Atkinson (2018) in constraining the tree topology using clade priors for well-established and commonly accepted Pama-Nyungan subgroups, as established by O'Grady, Voegelin & Voegelin (1966), Mühlhäusler, Tryon & Wurm (1996), Koch & Nordlinger (2014) and Bowern & Atkinson (2012). There are 9 of these subgroups represented in our language sample: Kanyara-Mantharta, Kartu, Marrngu, Ngayarta, Ngumpin-Yapa, South-West, Thura-Yura, Wati, and Arandic (though Arandic is represented by a solitary language in this study, Western Arrernte, so a clade constraint on this subgroup would have no effect). Dating the Pama-Nyungan tree is a central focus of Bouckaert, Bowern & Atkinson (2018), combining lexical cognate data with geographical data and archaeological calibration points to give a best-available estimate of the geographic and temporal point of origin of the family. Accordingly, we retain their calibration prior on the Wati subgroup, which places a 95% probability of the subgroup's origin dating between 3,000-5,000 years, with most of the probability density skewing towards the younger end of that range (a gamma distribution of $\alpha = 2$, $\beta = 359$, with 3,000 year offset) based on a synthesis of archaeological evidence (see Bouckaert, Bowern & Atkinson 2018: p. 746). We place a flat, uniform prior on the root age of the western Pama-Nyungan phylogeny of 3,000–7,000 years. This covers the time period from the youngest possible age of the Wati subgroup to the oldest end of the age range of the Pama-Nyungan family as a whole, under the best supported hypothesis of the origin of the Pama-Nyungan family in Bouckaert, Bowern & Atkinson (2018) (the mean root age in that study is 5,791 years). If the Wati age prior and best supported Pama-Nyungan root age in Bouckaert, Bowern & Atkinson (2018) are accepted, it follows that this is the logically maximal age range in which the origin of the western Pama-Nyungan branch must lie.

6.4.1 Results

We run four independent MCMC chains of 100 million iterations each and discard the initial 10% burn-in. Logs were inspected and combined in Tracer (v1.7.1) to confirm that each run converged

properly in the same probability space and that ESS values for all parameters were sufficiently high (>200). Chains 1–2 converge nicely in the same space, with all ESS values over 800. Chain 3 got stuck in a local optimum in the state space and chain 4 failed to converge, so these were discarded. Chains 1–2 were combined with 10% burn-in on each, giving a 180-million-state tree sample and combined ESS figures all greater than 1,900. A maximum clade credibility tree was produced using TreeAnnotator (v1.10.4). This is displayed in Figure 6.1 next to the same subset of languages from the maximum clade credibility tree in Bouckaert, Bowern & Atkinson (2018). We recover largely the same topology with some minor differences. These discrepancies seem approximately equivalent to the level of discrepancies between the Bouckaert, Bowern & Atkinson (2018) phylogeny and phylogenies from Bowern & Atkinson (2012) and Macklin-Cordes, Bowern & Round (2021).

Western Arrernte, the sole representative of the Arandic subgroup, shifts places among deeper nodes in the tree. The internal structure of the Ngumpin-Yapa subgroup also differs, which is interesting since the internal structure of Ngumpin-Yapa also shifts between Bowern & Atkinson (2012) and Bouckaert, Bowern & Atkinson (2018). In other words, in this instance divergences exist in a subgroup of the tree where there was already uncertainty between different analyses. Overall, notwithstanding these discrepancies, we are content that our model and method is working as it should with cognate data.

6.5 Main test: Tree inference with phonotactic data

6.5.1 Data

There are three main sources of data for this experiment. One is cognate data from Bouckaert, Bowern & Atkinson (2018), to which we apply an evolutionary model following Bouckaert, Bowern & Atkinson (2018), as evaluated in Preliminary Test 2 (Section 6.4). There are two sets of phonotactic data, one binary and one continuous. Binary biphone data, which codes the presence or absence of biphones—sequences of two phonemes—in each language, is included with the best supported evolutionary model evaluated in Preliminary Test 1 (Section 6.3). Finally, we use a dataset of continuous phonotactic characters coding frequencies of transitions between natural sound classes. The data extraction process is detailed further as follows. Each consonant phoneme appearing in our wordlist data is binned into two natural classes, one for place of articulation and one for manner of articulation. Place classes are labial, dental, alveolar, retroflex, palatal and velar. Manner classes are obstruent, nasal, vibrant, lateral, glide, and rhotic glide. For the purposes of this experiment, we group all vowels into a single 'vowel' class and also include word boundaries. The choice of place and manner natural classes is based on well-established principles of organisation among segments in Australian languages (Dixon 1980, Hamilton 1996, Baker 2014, Round 2021b,a). From each wordlist, we extract the frequency of a sound class x followed by a sound class y, relativised over all instances of sound class x. For use in BEAST, we logit-transform all the data to move from a 0–1 frequency interval to the real line. To include 0 and 1 frequency values in the logit-transformation, we follow a standard procedure of converting 0

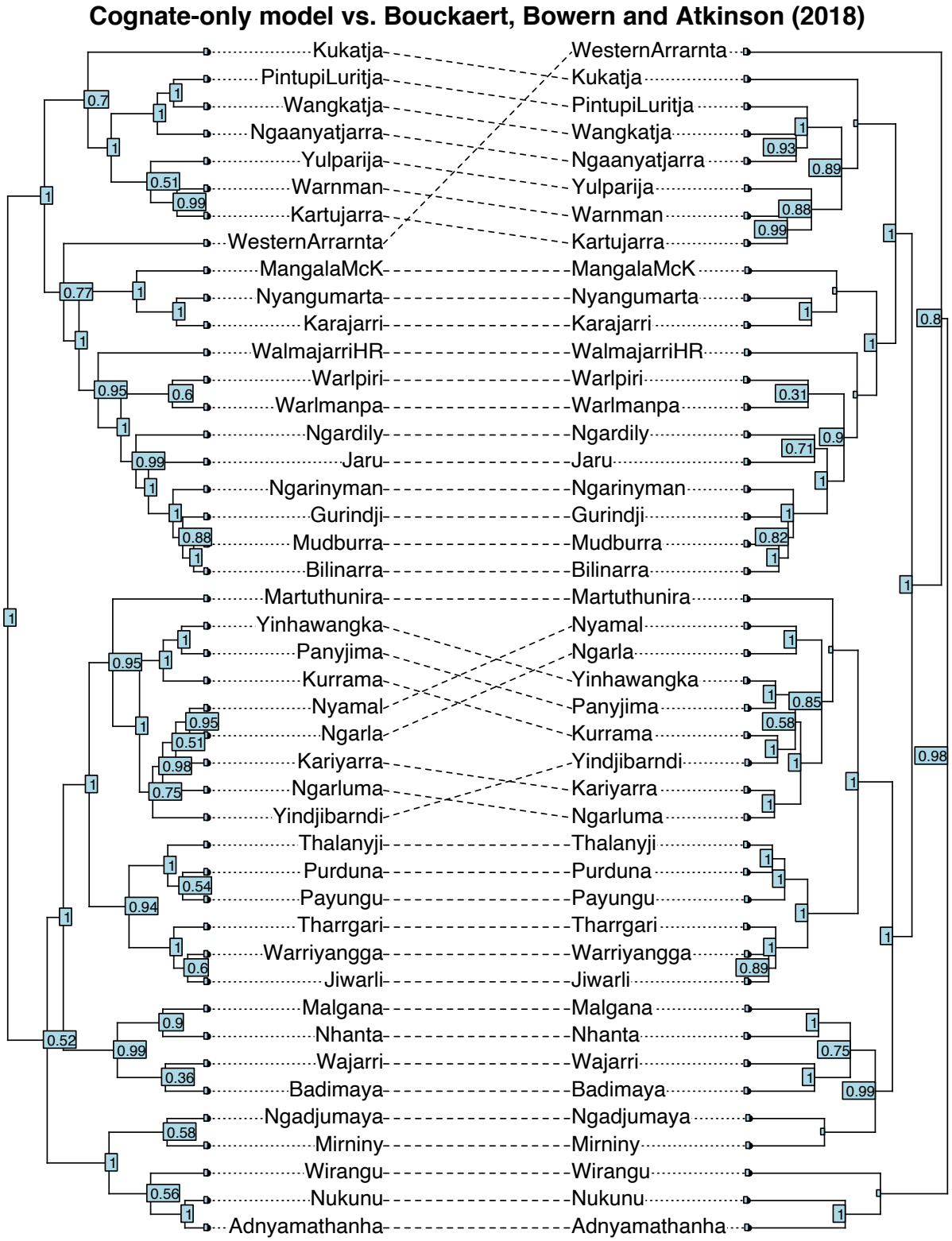

Figure 6.1: The maximum clade credibility trees of our cognates-only model (left) and the Bouckaert, Bowern & Atkinson (2018) phylogeny (right). Our model aims to replicate the BBA phylogeny as close as possible, although the language sample is restricted and there are some small model and software differences. Overall, the trees are mostly congruent, which gives reassurance that our baseline model is reliable.

values to $\frac{min}{2}$, i.e. half the smallest non-zero value, and converting 1 values to $0.5(1+max)$, i.e. half way between 1 and the maximum value less than 1. The evolutionary model applied to sound class transition frequency data in BEAST is a simple Brownian motion model which allows character values to wander up or down with equal possibility. We initially considered applying a single multivariate Brownian motion model to all continuous traits, however, the computational memory and time required by a multivariate Brownian motion model scales exponentially as the number of characters is increased. Instead, we fit a Brownian motion model to each character individually. This greatly increases the number of parameters in the overall model and potentially makes it less realistic, but it keeps the project computationally feasible. We return to this topic further in the discussion.

The choice of phonotactic data warrants brief further comment. We select sound class transition frequency data for a few reasons. One is that it partially, though not entirely, accounts for a lack of independence between biphone characters. This is because phonological rules, phonotactic restrictions and historical sound changes typically affect natural classes of phonemes rather than individual phonemes. This is not a perfect solution—sound classes themselves are independent from one another to varying degrees. A second reason is that Macklin-Cordes, Bowern & Round (2021) find a stronger degree of phylogenetic signal in sound class transition frequency data compared to transition frequencies between individual phonemes. A third reason is computational. A dataset of continuous characters equivalent in size to the binary biphone dataset would increase the computational time required to run BEAST by orders of magnitude, rendering the study impractical.

The decision to focus on frequency data in particular also warrants further explication. Notwithstanding the computational limitations we encounter, frequency data offers a number of possible advantages for phylogenetic inference. One is that it allows us to capture a finer grained level of information than binary data would allow. Binary data is more similar to the kind of phonotactic information one might find in a published language grammar, where a description of phonotactics that one would typically encounter involves a series of statements on the (binary) permissibility or otherwise of certain combinations of segments or natural classes of segments. This information does not, however, account for quantitative differences between common, high frequency sequences versus disfavoured sequences that rarely arise in a language's lexicon. There is considerable evidence to suggest that speakers are psychologically attuned to these kinds of phonological frequencies (Coleman & Pierrehumbert 1997, Zuraw 2000, Ernestus & Baayen 2003, Albright & Hayes 2003, Eddington 2004, Hayes & Londe 2006, Gordon 2016). The second reason is that the relatively rapid, semi-automated extraction of transition frequencies from wordlists captures structural variation between languages at a scale and degree of precision that would be difficult to attain from manual data coding methods (as preferred for the coding of lexical cognate data and grammatical data used in previous linguistic phylogenetic work). Macklin-Cordes, Bowern & Round (2021) show that this transition frequency dataset contains stronger phylogenetic signal than its binary equivalent. Increasing variation by shifting from categorical-valued characters to continuous characters has also been noted as advantageous for phylogenetics in biology, in the context of anatomical morphological data (Parins-Fukuchi 2018, Wright 2019). An additional advantage is that our method avoids observer bias. We don't have to rely on an expert picking and

choosing which parts of a grammatical or lexical system are interesting and worth coding. This has been noted as an advantage of large-scale extraction of continuous morphological characters in biology too (Wright 2019). Lastly, since we automatically extract frequencies for all possible biphones, we do not need to correct for acquisition bias or ascertainment bias (Leaché et al. 2015).

There is one limitation of the frequency data, which is that the Brownian motion model requires values above 0 and below 1. As described above, 0 values and 1 values are transformed to values very slightly above 0 and very slightly below 1 prior to logit-transformation. The inclusion of binary data in this study as well as frequency data is partly an effort to retain 0 and 1 values, which can be linguistically meaningful. One final limitation is that, throughout, our phonotactic data concerns linear sequences of exactly two segments. This is phonotactics in the simplest sense, and does not directly capture phonotactic restrictions that depend on sequences beyond two segments, syllable structure or morpheme boundaries. Nevertheless, Macklin-Cordes, Bowern & Round (2021) confirm that this simple level of phonotactic data is sufficient to detect strong phylogenetic signal.

As mentioned in Section 6.4, we reduce the language sample to 44 western Pama-Nyungan languages. This was necessary to keep the computational demands of the experiment reasonable. We selected the western branch of Pama-Nyungan due to it being relatively well-attested in modern sources compared to those in, for example, the south-east of Australia where language documentation is more often reliant on archival sources due to the disproportionate impact of colonialism in this part of the continent.

We run BEAST on two phylogenetic models: a 'linked' model, which contains a single tree likelihood term calculated using all cognate and phonotactic data, and a 'separate' model, which contains two tree likelihood terms, one calculated using cognate data only and the other using phonotactic data only (both binary and continuous). We run ten independent MCMC chains of 100 million iterations on each model, inspect and combine the results, estimate marginal likelihoods for each of the 'linked' and 'separate' models and generate three maximum clade credibility trees (one for the 'linked' model, and one each for the cognate-only and phonotactics-only elements of the 'separate' model).

6.5.2 Results

One first observation is that the MCMC process in BEAST tends to get stuck in local optima within the probability space and occasionally fails to converge. When this became clear, we increased the number of independent chains from an initial four to ten. Inspecting the joint probability traces for each of the linked model's ten runs, we find one clearly preferred area of the state space (Figure 6.2). We remove two clear outliers stuck in local optima above and below the others. Two additional runs (visible in Figure 6.2 as two overlapping, slightly bimodal distributions, just to the left of the bulk of other chains) are removed since, although they overlap considerably with the others, they too seem to represent a distinct local optimum in the state space (one of these also failed to reach stable convergence). Three more chains failed to reach stable convergence after a reasonable (10–20%) burn-in period and had

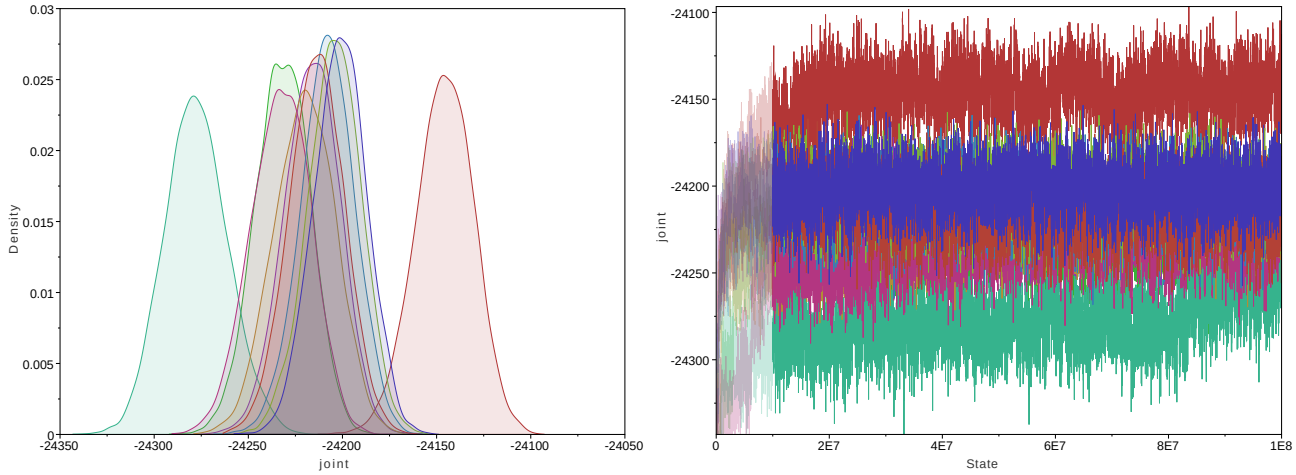

Figure 6.2: MCMC joint probability traces for the linked model. Although a preferred general area of the probability space emerges, this area has wide bounds. There are two clear outliers and there is a considerable degree of inconsistency between chains.

to be removed. This leaves three stable chains with suitably high ESS values and consistent areas of convergence (the three tallest, central peaks in Figure 6.2). The combined ESS for joint probability is 196. Over 200 would be preferable but we accept this as marginally acceptable. After discarding 10% burn-in, we are left with a combined 270-million-state analysis.

BEAST computes two marginal likelihood estimates, one via the path sampling method (Baele et al. 2012) and one via the stepping stone sampling method (Baele et al. 2013). In normal operation these two estimates should be roughly equivalent. In this instance, the log marginal likelihood estimates for the combined analysis are 598,885 (estimated via path sampling) and 988,231 (estimated via stepping stone sampling). Currently, it is unclear whether or not these log MLE figures are meaningful. Log MLEs for each individual chain vary between 459,380 and 672,466 which is considerably more variation than we would expect, given that a difference of 20 is generally considered significant. Given that all chains represent the same data and model and converge in approximately the same space, we would expect their log MLEs to be around the same and certainly not as divergent as this. We discuss this further below.

The separate model, in which two trees are sampled at each step, one inferred using only cognate data and one inferred using only phonotactic data, fails to reach stable convergence and gives unacceptably low ESS figures in most instances. With a burn-in of 15%, as per the linked model above, only 3 chains give a stable posterior sample with satisfactory ESS values. A fourth stabilises after a 20% burn-in, and three others eventually stabilise in similar areas of the probability space to the others, but only after burn-in values of 50%, 60% and 70%. We apply a consistent 20% burn-in to the four most stable chains and discard the rest. Two of these converge in the same area while the other two converge in spaces of slightly lower joint likelihood. On this evidence, it is not entirely clear which is the most correct posterior distribution. We accept the two higher likelihood chains space as it makes the study's null hypothesis deliberately harder to reject (the null hypothesis that the joint model is no

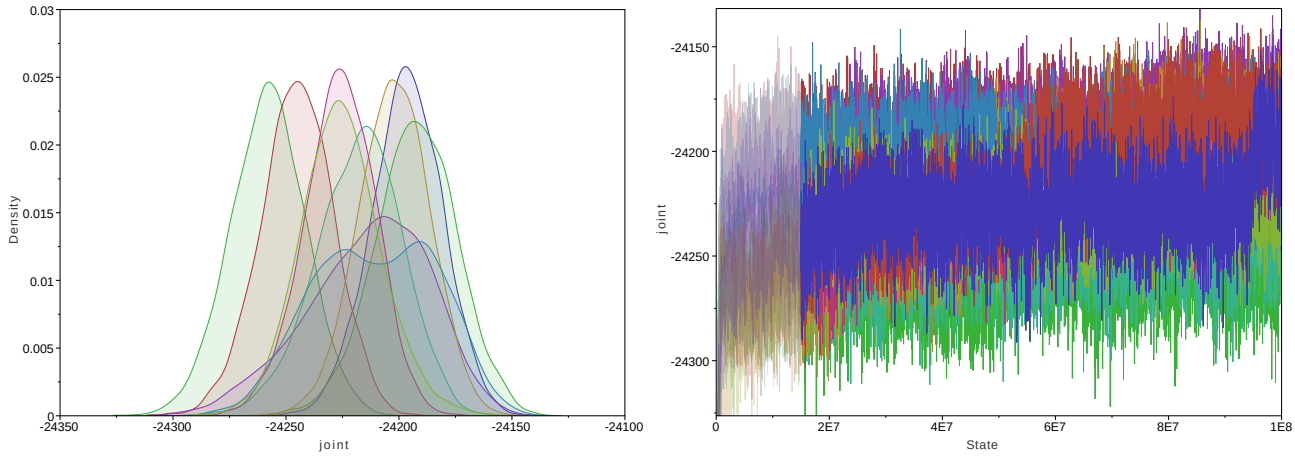

Figure 6.3: MCMC joint probability traces for the separate model. The lack of consistent (as seen on the left) and stable convergence (as seen on the right) is readily apparent.

Table 6.2: Marginal likelihood estimates (MLEs) estimated via path sampling and stepping stone sampling methods for linked and separate phylogenetic models.

Model	Path	Stepping stone
Linked	598,885	988,231
Separate	348,287	568,402

better than the separate model is harder to reject if we raise the bar set by the separate model). This gives a posterior sample of 160 million states and a marginally satisfactory ESS of 100 or greater for all parameters. Maximum clade credibility trees for each of the two tree elements—one inferred only with cognate data and the other inferred only with phonotactic data—are depicted in Figure 6.5.

The log MLEs for the separate model are 348,287 (estimated via path sampling) and 568,402 (estimated via stepping stone sampling). These MLEs can be compared with MLEs for the linked model in Table 6.2. The linked and separate models can be compared using Bayes factors, following Equation (6.1).

$$BF = log(MLE_{linked}) - log(MLE_{separate}) \tag{6.1}$$

The sign of the Bayes factor indicates which model is preferred. Bayes factors testing support for the linked model over the separate model are given in Table 6.3. Using both path sampling and stepping stone sampling methods, the Bayes factors seem to indicate decisive support for the linked model over the separate model. At face value, these figures are evidence for the hypothesis that including quantitative phonotactic data with more traditional cognate data can strengthen phylogenetic tree inference in linguistics. However, given the extreme divergences between MLE estimates from chain to chain and between MLE inferential methods, this evidence should be considered tentative at best.

Another way to evaluate these results is to consider the topologies and posterior support values

Table 6.3: Bayes factors indicating support for the linked model over the separate model.

Method	BF
Path sampling	250,598
Stepping stone	419,829

of maximum clade credibility trees. In Figure 6.4, we compare the maximum clade credibility tree from the linked model to the maximum clade credibility tree from the cognates-only model discussed in Section 6.4. A first observation is that the linked model posits Western Arrernte as the outermost taxon, branching off first from all other subgroups. This is different to the cognates-only model but in line with Bouckaert, Bowern & Atkinson (2018). Western Arrernte is notable for a couple of reasons. Firstly, in our language sample it is the sole representative of the whole Arandic subgroup. Secondly, in the case of the linked model, it is almost certainly being pushed to the outermost position in the phylogeny by its phonotactics. Arandic languages show evidence of major phonological change prior to proto-Arandic. In this instance, all the other subgroups are being grouped together based on their conservatism, driving Arandic into the outlier position. This is unfortunate, since subgroups should be identified from shared innovations, not shared retentions (as is happening here). This is potentially a weakness of phonotactic characters: Phylogenetic methods rely on there being many changes in order to overcome their inability to directly distinguish innovations from retentions. However, many phonological changes are so rare (relative to the phylogeny being studied) that there is not enough evidence to do this, resulting in this kind of undesirable grouping by shared retention.

The trees are largely congruent although there are some shifts in internal structure in some subgroups. The internal structure of Ngumpin-Yapa shifts once again—the Yapa branch (Warlpiri and Warlmanpa) gets split in the linked model. Overall, there is a general flattening of the tree structure in the linked model, approaching a star phylogeny. One possibility is that the tree structure is being flattened by the addition of the binary phonotactic data. As discussed, this sample of western Pama-Nyungan languages share largely homogeneous phoneme inventories and similar phonotactics in binary, permissibility terms. The binary data is relatively low resolution and might be masking phylogenetically informative variation between languages. That said, for the lower order topological structure that is present, the linked model tends to give good posterior support values, with a couple of exceptions. Perhaps unsurprisingly, Ngumpin-Yapa is an exception with very low support values. Another exception is the subgroup of Warnman, Kartujarra and Yulparija in the Wati clade. In other cases though, the addition of phonotactic data in the linked model seems to strengthen some support values, for example, within the Kanyara-Mantharta (languages listed Purduna–Tharrgari in Figure 6.4) and Kartu (Badimaya–Nhanta) subgroups.

The flattening effect of the phonotactic data is evident when MCC trees are calculated for each of the two tree elements of the separate model. Recall that the separate model includes two separate tree elements—one inferred with cognate data only and one inferred with phonotactic data only (both binary and continuous). Maximum clade credibility trees are inferred for each of these elements

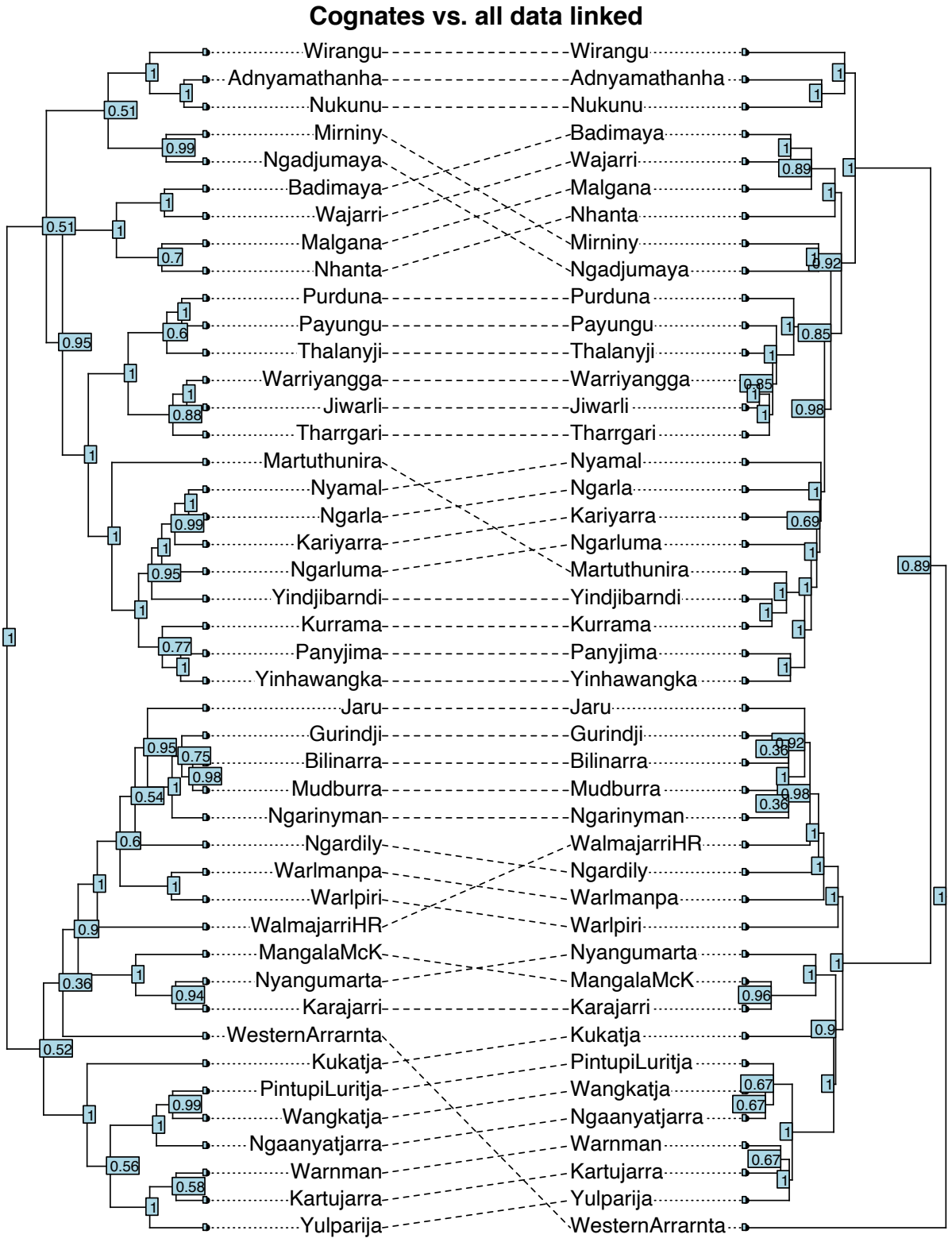

Figure 6.4: Maximum clade credibility trees for the cognates-only model from Preliminary Test 2 (left) and the linked model (right), which was inferred from cognate data and phonotactic data together.

and plotted in Figure 6.5. Side-by-side, the loss of intermediate tree-like structure is apparent in the phonotactics-only MCC tree. It is perhaps unsurprising then that the linked model, in which a single tree is inferred from all the data together, produces an intermediate level of tree-likeness, less tree-like than the cognates-only model but not as star-like as the phonotactics-only tree in the separate model.

6.6 Follow-up test: Removing binary biphone data

Upon reviewing results of the main test, one initial hypothesis was that the binary biphone data, which we expected would contain a little phylogenetic signal but to a relatively low degree, might be having an outsized influence on likelihood calculations in BEAST. This relatively homogeneous dataset, we reasoned, might be driving star-like tree signal in the tree topologies and increasing uncertainty in the analysis. In the likelihood calculations for our models, each data partition is effectively weighted proportionately to the amount of data within it (not due to a particular analytical decision but just because of the way BEAST is presently implemented). Our models contain 1,097 binary lexical cognate characters and 2,236 binary biphone characters, so the cognate data gets outweighed over 2-to-1 despite our prior expectation that the cognate data should be relatively more informative.

To test this hypothesis, we ran one follow-up test which is a replication of the linked model from the main test but with the binary biphone data removed. That is, the phylogeny is inferred from lexical cognate data and continuous phonotactic data only. We ran four independent 100-million-state MCMC chains in BEAST. Unfortunately, the new linked model does not seem to produce more stable or consistent convergences than the original linked model presented above. Of the four MCMC chains, two fail to reach stable convergence while the other two converge to slightly different optima in the state space. We err on the side of the chain in the lower likelihood space, which also has the most stable convergence and highest ESS values (1,269 for joint probability and $>$200 for all other parameters). We discard 10% burn-in and calculate an MCC tree for the remaining sample. This is plotted in Figure 6.6, side-by-side with the MCC tree for the main linked model above. As can be seen, the linked model produces practically identical tree topologies regardless of whether the binary biphone data is included or not. Some differences in clade support values can be observed, but in some instances these values increase with the removal of binary biphone data while in other instances they decrease. Overall, these differences in clade support values seem not to favour one particular tree over the other.

We conclude that the presence of a large partition of binary biphone data in the main study was not unduly impacting likelihood calculations and the MCMC process. The results are largely replicated whether or not this dataset is included in the study.

6.7 Discussion

In this study, we sought to test the question of whether phonotactic data could help strengthen the inference of phylogenetic signal, when included in a Bayesian computational phylogenetic model in conjunction with more traditional, lexical cognate data. We were motivated by earlier findings that

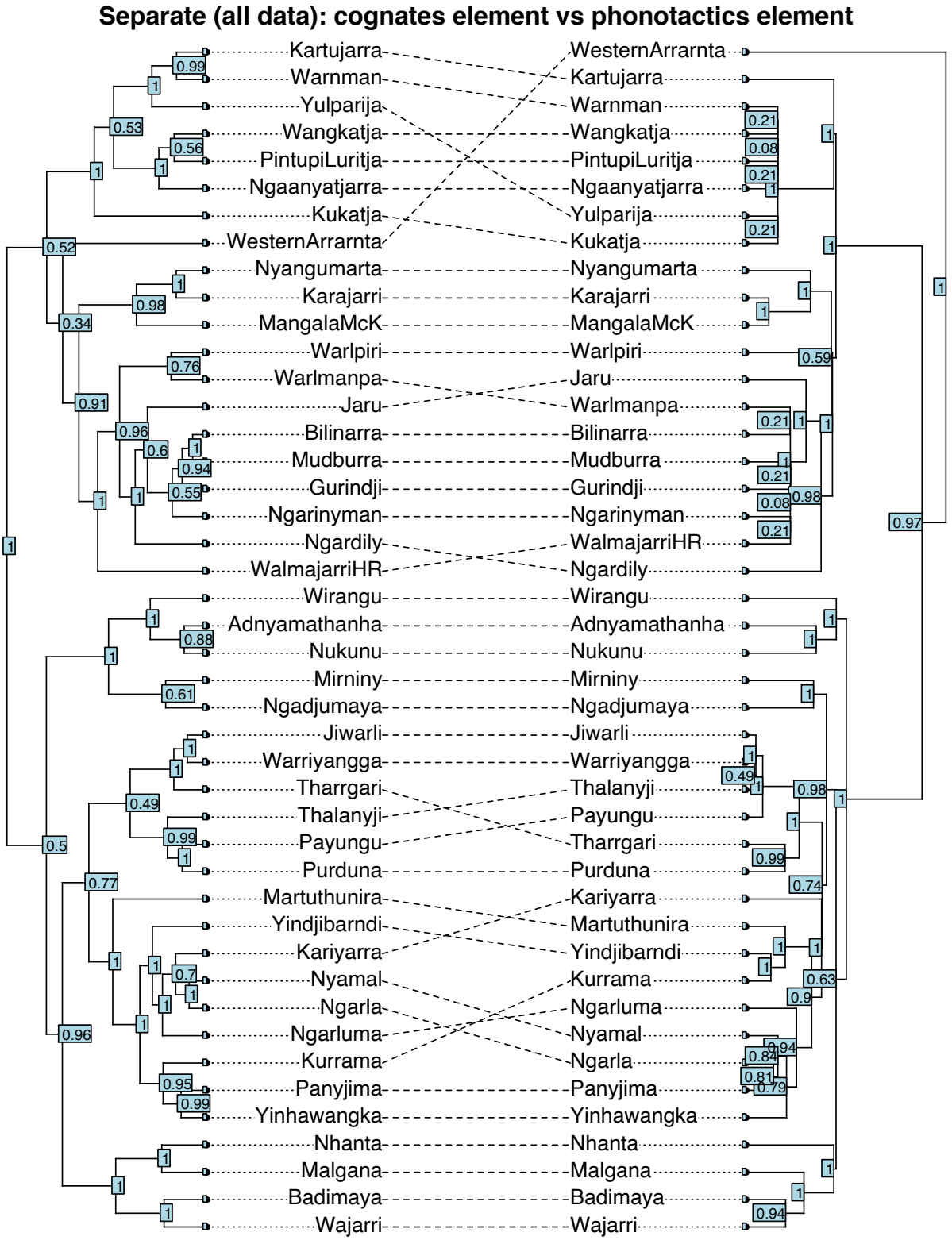

Figure 6.5: Maximum clade credibility trees from the separate model. The tree inferred from cognate data only is depicted on the left, the tree inferred from phonotactics only is on right. The flatter, star-like topology and low clade support values in the tree on the right is suggestive of non-treelike signal in the phonotactics data.

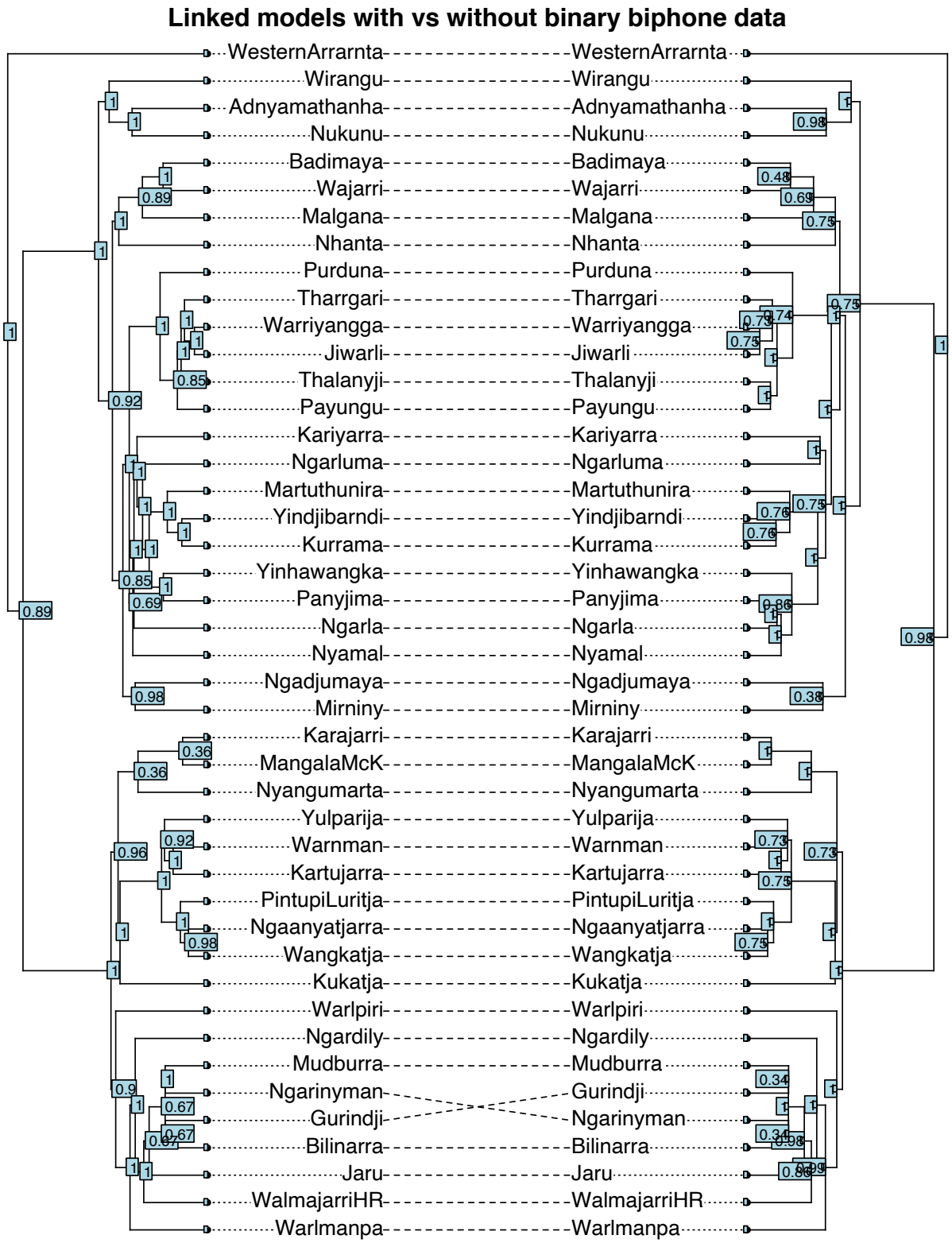

Figure 6.6: Comparison of maximum clade credibility trees from the linked model in the Main Test (left) and a follow-up model which includes cognate data and sound class frequency data, but excludes the binary biphone data. Both trees appear highly congruent, suggesting the binary biphone data had relaively little effect on the results of the Main Test.

phonotactic data, represented here by binary characters coding the presence or absence of biphones and continuous characters coding the frequency of natural sound classes being followed by other sound classes in biphone sequences, contain phylogenetic signal (Macklin-Cordes, Bowern & Round 2021). Our method of testing this hypothesis was to develop two phylogenetic models, a 'linked' model and a 'separate' model, and infer phylogenies for each using a Bayesian MCMC process implemented in BEAST. In the linked model, tree likelihoods are calculated at each step using existing lexical cognate data from Bouckaert, Bowern & Atkinson (2018) and both binary and continuous phonotactic data. In the separate model, two tree likelihoods are calculated, one using cognate data and the other using phonotactic data separately. Marginal likelihoods are estimated to give an indication of overall support for each model. Maximum clade credibility trees are calculated to compare tree topologies and support for particular subclades.

Our results do not support the hypothesis that a better supported phylogeny could be inferred with phonotactic data. To the contrary, we found that the addition of phonotactic data contributed instability and uncertainty into the MCMC process. We also found the MLE process to be unreliable, with extreme divergences between even chains that seemed to reach stable convergence in the same probability space. Inspecting the maximum clade credibility trees, the linked phylogeny seemed to produce an MCC tree broadly comparable to the equivalent phylogeny produced from cognate data only. However, the internal structure of some subgroups did get altered, the posterior clade support values did not, on the whole, improve, and there seemed to be a degree of star-like flattening of the tree shape. To the extent that the linked model recovered the topology of the cognates-only phylogeny, it might be said that it did so in spite of apparent star-like signal in the phonotactic data, rather than with the assistance of phonotactic data. The conclusion is that phonotactic data, in the form presented here, does not strengthen phylogenetic tree inference, at least within the present statistical and computational implementation.

The next question is to what extent these findings represent a genuine negative result, i.e. evidence for a more general conclusion that phonotactic data does not strengthen phylogenetic tree inference. If this is the case, then future linguistic phylogenetics research should abandon phonotactics as a potential data source. The alternative, however, is that the negative result is attributable to any combination of issues with data acquisition and/or experimental design. In this case, we cannot say the study supports any particular insight into the phylogenetic dynamics of phonotactic systems in human language. Rather, we would require further study, hopefully benefiting from the lessons of the present attempt, before making any conclusions either way. There are reasons, as discussed in the paragraphs that follow, to suspect that this study is an instance of the latter. If that is correct, then the current experiment is best interpreted not as evidence against any particular hypothesis, but rather as evidence for the need for further refinement along the lines we discuss below.

The following points of discussion fall under two main umbrellas. We start with more technical issues relating to experimental design and implementation, then follow with more general, theoretical matters. Firstly, we must consider the difficulties we observed achieving stable, consistent convergences between MCMC chains. It is possible that the development of relatively simple technical fixes might

help rectify this issue to some degree. Although 100 million states seems to be enough to reach normal, stable convergence for many chains in this study, others fail to converge or, in some instances, get stuck for a while before managing to escape a local optimum and converging somewhere else very late in the chain. The most obvious solution to this is to try running much longer chains. The difficulty here is operating within computational constraints, as continuous characters increase the running time of the MCMC process by orders of magnitude. The issue of computational feasibility is compounded again as the size of the phylogeny increases. It would be challenging, for example, to run even our present 100 million state chains if we extended our model to the 112-language Pama-Nyungan sample we used in Section 6.3. One second technical fix would be to adjust the operators in BEAST that govern how parameter values shift from state to state. Allowing larger jumps in the parameter space could help the MCMC chain to escape local optima.

A more complex technical issue is the implementation of an evolutionary model for continuous characters. When working with binary datasets in phylogenetic software, one would usually work with a single data partition containing all binary characters or a small number of data partitions. The problem with continuous characters is that the computational time and memory demands of a continuous data partition expand exponentially as the number of continuous characters increases. For example, the sound class biphone dataset that we use in this study would take several years rather than several days to run on identical computing hardware, if implemented as a single data partition. To get around this, we created an individual data partition for each individual character. This brought the study into the realm of reality from a practical standpoint, but at the cost of hugely increasing the number of parameters that form part of the model, because individual rates of evolution and likelihood scores had to be inferred for every character. Besides the question of whether treating each character as its own element in the model is realistic (which we address more below), having such a great array of parameters to estimate likely contributes to uncertainty in the analysis. Future study should consider some option in between the extremes of a single data partition and individual data partitions for all continuous characters. The question is how continuous characters should be partitioned and on what basis. One option is to pre-define data partitions on some linguistically meaningful basis (for example, perhaps grouping characters by the sound class of the first biphone segment). The other option is to infer how the data should be partitioned as part of the analysis. There is work in this space using phylogenetic factor analysis, under active development (Tolkoff et al. 2018, Hassler et al. 2020), and this could be explored in the context of large linguistic datasets of continuous variables in future work. Addressing this issue will have the added benefit of addressing the lack of independence issue.

There are other limitations to the evolutionary model for continuous characters that we used. For one, we use a simple Brownian motion model which allows characters to wander up or down with equal probability. This model has the advantages of being straightforward to implement with existing software and also aligns with earlier work by Macklin-Cordes, Bowern & Round (2021). But it also fails to fully encapsulate the way that biphone frequencies are likely to change over time. One question to consider is what a more realistic model of evolution would look like for frequency-based phonotactic characters. To answer this, we need to think about the causal forces of language change that shape

these frequencies. We note two such forces here. The first is the introduction of new vocabulary to a language via lexical innovation or borrowing. Each new word entering a lexicon will alter minutely the frequencies of biphone transitions in the language (similarly, transition frequencies will decline as words are replaced or fall out of usage). This is the kind of gradual accumulation of changes that we might expect to follow a Brownian motion-like pattern of evolution (although maybe the rates of going up and down are not equal). Further, since speakers show a preference for high frequency phonotactic sequences over low frequency sequences when coining new words, we might expect this accumulation of changes to follow some a kind of 'rich-get-richer' preferential attachment process which would result in the kind of skewed frequency distributions that we observe. Also, when languages borrow vocabulary, the trend is for foreign words with disfavoured phonotactic sequences to shift towards more natively preferred patterns (sometimes gradually over a long period of time, e.g. in some French borrowings in English, stress has shifted to English pattern but not yet in others; see also Crawford (2009) on Japanese), which would strengthen this kind of preferential attachment process and also keep phonotactic frequency data historically conservative. The second major force on biphone frequencies is regular sound change. We would expect regular sound changes to result in sudden jumps in the frequencies of affected biphones, sometimes to 0 or 1. For example, if a language includes heterorganic nasal+stop sequences that subsequently undergo place assimilation, this will result in a sudden jump, rather than a gradual drift in frequencies from heterorganic nasal+stop sound class characters to their homorganic counterparts. Our binary characters capture some of these effects to a limited extent. For example, in the place assimilation example above, certain biphones will switch from '1' to '0'. In other instances, sound changes could cause characters to switch from missing all together to some non-missing value or vice versa. For example, if a contrastive vowel length distinction emerges, certain biphones (those with long vowels) will go from being a gap to '1' in the binary biphone data for that language. In the case of a merger between short and long vowels, the opposite will be true. Our model, at present, does not obviously account well for regular sound change. One option for addressing this in future work may be to include a Lévy process in the model, which allows for gradual drift punctuated by sudden jumps (Uyeda et al. 2018). One final limitation to note is that the frequency data and evolutionary model in this study do not account for measurement error. Wordlists are necessarily a limited representation of the complete, true lexicon of a language, and so particularly in the case of shorter wordlists, there may be a margin of error between the true frequencies in the language and the frequencies we extract from its wordlist. There is evidence to suggest that accounting for measurement error could help improve results (Round & Macklin-Cordes 2019). Overall, future work will need to consider more carefully how a realistic evolutionary model for continuous characters should be designed and, if necessary, do the heavy lifting of implementing the model in phylogenetic software rather than relying on currently available implementations. Furthermore, more careful re-evaluation of the frequency distributions of phonemes in Macklin-Cordes & Round (2020a) should be replicated at the biphone level.

Any statistical model is necessarily an abstraction of reality. The challenge is striking the right balance between a model that is realistic enough to produce realistic outcomes and abstract enough

to be both practical and generalisable. In this study, we may have compromised on the realism of the model too much in order to keep the project computationally feasible with presently available software and hardware. Technological advances in central and graphical processing units and increased computational parallelisation (e.g. Holbrook et al. 2020) will mitigate this dilemma somewhat in future, but additional work will be required to create evolutionary models that are more faithful to the real-world diachronic processes that generate phonotactic frequencies.

A final point is that future research might consider whether phonotactic data has a role in phylogenetic tree inference or some other aspect of linguistic phylogenetics beyond tree inference itself. In biology, there is ongoing debate on whether morphological data is best suited to a "simultaneous" approach, where the tree is inferred jointly from morphological and genomic data, versus a "scaffolding" approach, where the tree is inferred from genomic data only, then morphological data is used to investigate other phylogenetic questions (e.g. dating branching events using morphological data from the fossil record) while being constrained to the genomic tree's topology (de Queiroz 1996, Lee & Palci 2015). We do not take up the topic further here, but an analogous discussion could be had in linguistics. It may be the case that phonotactic data or some other kinds of linguistic data are better suited to a scaffolding-type approach, playing a role in phylogenetic investigations but only through application to a phylogenetic tree topology that has already been inferred using cognate data.

6.8 Conclusion

Following earlier findings that phonotactic data contains phylogenetic signal, we sought to create and evaluate evolutionary models for two basic representations of phonotactic information for a sample of 112 Pama-Nyungan languages—binary biphone data coding the presence or absence of biphones in a lexical wordlist, and sound class biphone frequency data coding the frequencies of forward transitions between sound classes in a lexical wordlist. Subsequently, we attempted to infer a phylogeny of a subset of 44 Pama-Nyungan languages representing the western branch of the family, using phonotactic data in combination with existing lexical cognate data. Although marginal likelihood estimates on their face seem to support a linked model where the phylogeny is inferred jointly from phonotactic and cognate data, over a separate model where separate trees are inferred from phonotactic and cognate data separately, we find that the calculation of these figures is unreliable in this instance. Further, through comparison of maximum clade credibility trees for the linked model versus a cognates-only model, we find that the addition of phonotactic data does not seem to strengthen posterior clade support values. To the contrary, the phonotactic data seems to introduce a modest star-like signal to the phylogeny, leading to a slightly flatter tree structure.

These results, we argue, are likely the result of oversimplifying the model of continuous phonotactic character evolution and introducing too many free parameters to the model in an effort to achieve computational feasibility and operate within the bounds of existing phylogenetic software. Future research should develop the phonotactic evolutionary model further before re-evaluating its place in the task of phylogenetic tree inference. The first step of this process would be to evaluate empirically the

frequency distributions of biphone characters and attempt to link them to causal processes (e.g. mergers and splits of phonemic categories). Secondly, the Brownian motion model of character evolution would be replaced with a model that incorporates the identified frequency distribution from the previous step and allows character values to move in ways consistent with the causal processes. Thirdly, dependencies and correlated evolution between characters would be identified and the dataset partitioned accordingly. Fourthly, matters of the technical implementation and MCMC convergence process would be addressed, perhaps with advances in GPU parallelisation, with the aim of inferring reliable marginal likelihood estimates.

As linguistic phylogeneticists look beyond the cognate for sources of novel insight, we advocate a methodical progression and thorough methodological evaluation, from language change processes to large-scale linguistic evolution.

Chapter 7

Towards a phylogenetic phonotactic research program

The motivating research question posed at the start of this thesis was whether or not phylogenetic tree inference could be strengthened by complementing traditional, lexical cognate data with phonotactic data. Given the indeterminate results of the previous chapter, the answer to this question remains elusive. However, there is still much to be gleaned from this experience and general lessons that can be applied to future quantitative interrogations of novel data sources in historical linguistics. This chapter begins with a brief summary of each of the previous chapters, followed by an explanation of how the four separate studies presented cohere into a step-wise interrogation of a novel data source, phonotactics, in linguistic phylogenetics. I present the novel insights that can be drawn from this evaluation. Follow that, I address limitations of the thesis. I turn attention to future research directions and point to possible solutions to limitations that could be pursued in subsequent study. I note how the scope of this present work could be expanded to cover the Sahul continent and the challenges associated with this. Finally, I sketch an outline of the future of methodological development in linguistic phylogenetics and propose a research paradigm in which this could be done.

7.1 Summary so far

The preceding five chapters began with a literature review followed by four papers (three experimental studies and one research essay). The first paper presented a critical re-evaluation of frequency data in phonology, starting at the most basic level of individual phoneme frequencies. The second examined the role of phylogenetic comparative methods in linguistics and presented some use cases for measuring phylogenetic signal. The third, put those methods for measuring phylogenetic signal into practice, detecting significant phylogenetic signal in phonotactic data. The fourth inferred a Pama-Nyungan phylogeny using phonotactic data combined with and separately from previously published lexical cognate data. I now briefly recapitulate the journey so far.

Chapter 2 critically examined the history of phylogenetic thinking in linguistics and the burgeoning

subfield of linguistic phylogenetics. One limitation of linguistic phylogenetic tree inference presently is an over-reliance on lexical data. A relatively restricted number of studies have tried inferring trees with typological characters (e.g. Dunn et al. 2005, 2008, Sicoli & Holton 2014), but these tend to suffer from a limited amount of data (Yanovich 2020), logical dependencies between characters (Yanovich 2020) and rapid evolutionary rates and homology within a restricted state space (Greenhill et al. 2017). Linguistic phylogenetics stands to benefit from an increased volume of data, including data drawn from a more diverse selection of parts of human language. Separately, with regards to phylogenetic methods themselves, it is recognised that language evolution is not directly analogous to biological evolution and thus linguistic evolutionary processes warrant methodological consideration in their own right. Furthermore, as the discussion in Chapter 4 emphasises, linguistic phylogenetics commands its own position in evolutionary sciences. Linguists should not be seen to be rushing to catch up to biologists by taking off-the-shelf methods from a more mature field. There is a great deal of scope for adapting biological tools and methods for inferring linguistic evolution, since biological and linguistic evolution share certain similarities in evolutionary processes. But this should be done by firstly building a ground-up understanding of linguistic evolutionary processes. In this spirit, I seek to evaluate the utility of phonotactic data in linguistic phylogenetics in this thesis, in particular the frequencies of biphones—sequences of two phonological segments.

Chapter 3 presented the first of a four-part investigation. The output of a phylogenetic methodology is only as good as the data fed to it. Before simply pouring frequency statistics into a phylogenetic algorithm, I first examined the nature of phonological frequencies and the causal processes that the shape of those frequencies represents. Previous literature has looked at the frequencies of phonemes in particular languages and found that they tend to form heavily skewed distributions (such that there are a few high frequency phonemes and a long tail of many low frequency phonemes), though not skewed in a way that is well described by Zipf's law, which characterises word frequencies. I also found, however, that phoneme frequencies warranted a re-evaluation in light of advances in statistical methods for identifying distributions and demonstrations showing that previous methods were unreliable (Clauset, Shalizi & Newman 2009). In addition, previous studies tended to be dominated by European languages and Australian languages were either absent or dramatically underrepresented. I used the maximum likelihood framework suggested by Clauset, Shalizi & Newman (2009) to evaluate the distributions of phoneme frequencies in the lexicons of 166 Australian languages. I found weak support for a Zipfian-like power law structure among more frequent phonemes (though perhaps also a lognormal structure) and weak support for a geometric (or exponential) structure among less frequent phonemes. In the subsequent discussion in Chapter 3, I offer some tentative comments on causal processes. These comments remain speculative, because of the inherent uncertainty in fitting distributions to 20–30 observations (the number of unique phonemes in a given language). However, exponential distributions may be linked to birth-death processes and power law distributions may be linked to preferential attachment processes.

In Chapter 4, I break from discussions phonological data and turn attention to phylogenetic methodologies in comparative study. Phylogenetic methods might be most immediately associated

with phylogenetic tree inference in historical linguistics. However, phylogenetic thinking is essential in all comparative fields of linguistics. As Chapter 4 demonstrates, phylogenetic effects have long been a consideration in linguistic typology and also comparative biology where, similar to linguistic typologists, biologists are not immediately concerned with the task of tree inference but must consider historical relatedness as a confounding factor in their results. Both fields share similarities but also significant differences in how they deal with phylogeny as a confounding effect in comparative study, more properly termed phylogenetic autocorrelation. I argue in favour of the uptake of phylogenetic comparative methods in linguistics, in which phylogenetic information is incorporated directly into the statistical model. The critical advantage of this approach is that it enables a comparative study to include a maximal amount of data and the maximal extent of existing phylogenetic knowledge. This is in contrast to methods that involve discarding languages in order to create a phylogenetically balanced sample, and/or making limited phylogenetic assumptions, e.g. assuming equal phylogenetic distance between subgroups of languages in a sample (effectively assuming a star-like phylogenetic structure). Phylogenetic comparative methods require pre-existing phylogenies of languages to implement. Crucially, however, it is possible to use limited phylogenetic information (i.e. trees that are not fully resolved into bifurcating branches) and it is possible to incorporate phylogenetic uncertainty. And, as I argue, it is better to incorporate some phylogenetic information even if not perfect, rather than neglect phylogenetic autocorrelation altogether and effectively, perhaps unwittingly, assume a star phylogeny. One other use of phylogenetic comparative methods is to evaluate various questions of evolutionary dynamics, as I went on to demonstrate in Chapter 5.

Chapter 5 quantifies the phylogenetic information content in phonotactics, a novel data source which has hitherto not featured in linguistic phylogenetic tree inference. I implement offshoot methods of the phylogenetic comparative methods discussed in the previous chapter to measure phylogenetic signal—the tendency of more closely related languages to resemble one another more than more distantly related languages. I extract phonotactic data at the most simple, basic level, firstly by coding the presence or absence of biphones and secondly by extracting biphone transition frequencies from 112 lexicons of Pama-Nyungan languages, corresponding to 112 tips in Claire Bowern's Pama-Nyungan phylogeny. I found a statistically significant degree of phylogenetic signal in all datasets, but particularly strong signal in frequency data, which provides a finer-grained level of information than the relatively low-yield binary data. This result demonstrated the potential utility of phonotactic data in phylogenetic tree inference, which is perhaps contrary to what one might have expected given the ostensibly high degree of phonotactic homogeneity among Australian languages.

Following on from the conclusions of Chapter 5, in Chapter 6 I present the first instance of phylogenetic tree inference with the addition of the biphone data evaluated in the previous chapter. I combine the binary biphone dataset and biphone frequency transition dataset with a partition of lexical cognate data from Bouckaert, Bowern & Atkinson (2018). For an evolutionary model of lexical cognates, I reproduce as close as possible the model that Bouckaert, Bowern & Atkinson (2018) found to be best supported. I use a simple Brownian motion model of evolution for logit transformed biphone frequency data. For binary biphone data, I run an independent preliminary test of 16 different iterations

of several evolutionary parameters and then use the best supported iteration as the evolutionary model for this data partition in the main study. Finally, I calibrated the age of the Wati subgroup and overall age of the tree to the date ranges best supported in Bouckaert, Bowern & Atkinson (2018). I then inferred Pama-Nyungan phylogenies using a Markov Chain Monte Carlo (MCMC) process twice. In the first run, the Pama Nyungan tree was inferred jointly from cognate and phonotactic data. In the second run, trees were inferred from cognate data and phonotactic data separately. After discarding an initial burn-in period, a stepping stone process (Baele et al. 2013) was used to estimate an overall marginal likelihood for each analysis and, from this, a Bayes Factor was calculated to test whether the cognate-phonotactics model produced a significantly better supported tree. This would be evidence for the hypothesis that Bayesian computational inference of linguistic phylogenies can be improved by including partitions of phonotactic data with more traditional lexical cognate data. Unfortunately, these models fail to reach consistent convergence within current computational limitations, and the marginal likelihood estimates prove to be unreliable.

Altogether, the results of these papers demonstrate that large volumes of phonotactic information can be extracted from language wordlists. We hypothesise that the distribution of this phonotactic information is the outcome of diachronic processes and demonstrate that the that the data pattern phylogenetically. That would suggest that the phonotactic data could strengthen phylogenetic tree inference, however implementing it remains challenging. Besides interrogating phonotactics specifically, this thesis illustrates a step-wise procedure for evaluating the utility of other kinds of linguistic data from other parts of language in phylogenetic tree inference. This study is not an attempt to add novel data into a phylogenetic black box and see what happens. The risk of a black box approach is summed up by the old adage 'junk in, junk out'. In any results, it would be difficult to discern novel insights from artefacts of the data or unexpected interactions between the data and intricacies of the model and its assumptions. It is possible that this kind of approach may have contributed to a shortcoming of linguistic phylogenetics over the past two decades. That is, although phylogenetic studies have contributed a great deal of new information on questions that historical linguistics had not yet solved, e.g. detailed internal branching structure within families/subgroups, no one yet has been able to increase the time-depth of existing historical linguistic methods and infer novel macrofamilial relationships. Even as the field of linguistic phylogenetics matures, much of the present work focuses on demonstrating the validity of the method by reproducing existing historical linguistic results. This might be due, in part, to a lack of understanding of what it really means when the output of phylogenetic methods diverges from existing knowledge and whether this constitutes genuine novel insight or some kind of noise. In contrast to the above 'junk in, junk out' scenario, I have attempted in this thesis to present a careful, considered, ground-up evaluation of the data and methods before attempting to infer a phylogeny with phonotactic data in Chapter 6. This process is generalisable to any kind of linguistic data that would be new to phylogenetic tree inference and it can be summarised in an idealised sense as follows.

1. Consider the theoretical motivations for testing the novel data source. Simply testing everything because it is available invites the possibility of spurious correlations and false positives. Instead,

one should consider the linguistic theory that would motivate the novel data source's use in linguistic phylogenetic methods. What is known about the historical processes behind the data? Is there reason to suspect that this part of language is historically conservative or resistant to horizontal transmission between languages?

2. Consider the structure and nature of the data. How is the data distributed? Are there correlations or logical dependencies between variables? What does this say about the diachronic processes that produced the present distribution of observations?

3. Find a case study of languages for which there is a high quality pre-existing phylogeny, quantify the tree-like signal in the data, and evaluate its evolutionary dynamics. Is there phylogenetic signal? If the data is distributed randomly with regards to phylogeny (no phylogenetic signal) then some non-phylogenetic process is at play and the data will only contribute noise during phylogenetic tree inference. A negative result here would not mean the data is uninteresting—it could say something interesting about geography, ecology, language contact or something else—but it will not suit a tree model. Alternatively, if phylogenetic signal is detected, ideally this would be followed up with a subsequent evaluation of evolutionary dynamics. For example, is a Brownian motion evolutionary model appropriate or is a punctuated equilibrium model better supported?

4. At this point, there is well-founded, theoretically grounded evidence that the novel data patterns phylogenetically, a rigorous understanding of the data's distributional and logical structure, and a theoretical and empirical demonstration of its phylogenetic information content. The final step in this proof-of-concept is to attempt to infer a linguistic phylogeny with the addition of the novel data source. A Bayes factor can be calculated to test whether the addition of the novel data source adds sufficient phylogenetic information to infer the phylogeny with greater confidence.

If the novel data source passes each of these steps, then we can be more confident about using the same kind of data to infer linguistic phylogenies in other contexts where less is known, and potentially make new knowledge claims about the results. To take the example of Sahul, if it had been shown in Chapter 6 that phonotactic data strengthened tree inference within the confines of the relatively well-studied Pama-Nyungan family, then the most immediate avenue for future work would have been expansion to the non-Pama-Nyungan areas of Australia and eventually the rest of Sahul. However, there are further methodological considerations and subsequent evaluation that would need to occur before such expansion could take place. Notwithstanding the indeterminate results in Chapter 6, I outline the prospects for Sahul expansion below, which could nevertheless be helpful for future interrogation of phonotactics or other novel data types. I turn attention now to some of the limitations of the work presented in the previous chapters of this thesis and the outstanding questions that still need to be considered.

7.2 Current limitations and future work

By outlining a general framework for assessing the phylogenetic potential of novel linguistic data sources, I wish to convey that work on the phylogenetics of phonotactics has now begun in earnest, but is far from complete. There is a lot of work still to be done in the phylogenetic phonotactics space. I now discuss some of the limitations and unanswered questions from the studies presented in this thesis, interleaved with research priorities for future work which could address them. These research priorities form three main avenues. One concerns expansion of the language sample to the rest of Australia, Sahul, or elsewhere in the world. The second avenue concerns the expansion of phonotactic datasets. The third avenue involves developing an improved understanding of the evolutionary dynamics of phonotactics.

One analytical choice which has been questioned by reviewers pertains to the language sample, which was restricted to the Pama-Nyungan family in Chapters 5–6 and Australia broadly in Chapter 3. One response to this point is that historically an overrepresentation of European languages (especially western ones) in linguistics and under-documentation in other parts of the world has meant that linguistic generalisastions have been made often based on evidence from a single continent, or at least overrepresentation by a single continent. An example of this is Tambovtsev & Martindale (2007), which attempts to generalise about phoneme frequencies based on a dataset of 90 Eurasian languages, 2 Oceanic languages, and just a single language from each of Australia, Africa and South America. Another point to make is that although the studies in this thesis are genealogically and geographically restricted, within the Pama-Nyungan family (or Australia generally in the case of Chapter 3), I have been able to include a maximal sample of all the languages for which there is sufficient high quality lexical data available. There has been no need to discard data in order to balance the sample between subgroups or geographic regions. All of this is to say that the restricted scope of the language sample in this thesis may not be quite as large a caveat to the results as initially might be assumed. Nevertheless, future expansion of the study to other linguistic regions would be beneficial and I turn to the question of how to do that now.

I have sought as much as possible to make the studies in this thesis reproducible by making code and data available and relying only on open source software. An advantage of this is that expansion of this research to other parts of the world should be relatively straightforward in cases where there exists i) comparably segmented and phonemically standardised wordlists and ii) a high quality reference phylogeny, ideally for a sample of at least 30 languages. Reproduction in this way would be a useful validation of the results presented here. Beyond this, a priority for future work is to expand phylogenetic tree inference with phonotactics to the rest of Australia and potentially Sahul. The linguistic phylogeny of Pama-Nyungan is relatively well-studied now in the sense that we have large, high quality phylogeny for the family, with dates calibrated against archaeological evidence and even a phylogeographical component. however, this is only a fraction of Australia's linguistic diversity. No detailed phylogeny yet exists linking the many families and isolates in Australia's Top End, nor linking Australia's non-Pama-Nyungan families to Pama-Nyungan under a putative proto-Australian root.

Expansion to the rest of Australia is a particularly promising possibility for expanded phylogenetic studies with phonotactics, because although there is enormous lexical diversity between the languages, non-Pama-Nyungan languages tend to share similar phonemic inventories to their Pama-Nyungan counterparts and, therefore, there should be plenty of characters with non-missing data for most or all languages. There is also scope to improve our understanding of the phylogenetics of less well-studied subgroups or subgroups with ostensibly high rates of horizontal transfer within Pama-Nyungan. An example is Cape York, where some progress has been made processing archival materials from the linguist Bruce Sommer (Hollis et al. 2016). Cape York languages show more divergent phonology from the rest of Australia and so expanded coverage in this area would constitute an interesting challenge case for phylogenetic phonotactics.

Potential expansion of phylogenetic tree inference with phonotactics to Sahul deserves special consideration. The first challenge in this enterprise will be sourcing data from New Guinea. Two sources of lexical data are the *TransNewGuinea* database (Greenhill 2015) and a dataset of Mandang comparative wordlists (Z'Graggen 1980). The Mandang dataset holds promise, albeit within a restricted area (the Mandang Province). It contains good-sized wordlists (ranging from 242–300 lexical items per language) which are also phonemicised, for 98 languages. As for the *TransNewGuinea* database, the scope and scale of the database is enormous. It contains lexical data for 1,028 language varieties, mostly but not entirely belonging to the Trans-New Guinea family. There are two issues pertaining to the use of *TransNewGuinea* data in phylogenetic phonotactic studies. The first is that the wordlists tend to be too short for these kinds of studies, averaging 142 lexical items per language. There are many sufficiently longer wordlists that could be used, but most of the languages in the database would fail to meet the 250-item threshold used in this thesis. The second is that, although the wordlists are transcribed in phonemic form, it would take a considerable amount of work to ensure these forms are phonemically standardised such that each individual segment is comparable across languages. One possible subject for future study, then, is to build a comparative database of New Guinea phonologies, following the principles of AusPhon (Round 2017c). Towards this goal, there is some preliminary work on automated parsing of language grammars in New Guinea, however, this remains in an embryonic stage of development (Round et al. 2020). In addition, advances in computational linguistic tools for grapheme-to-phoneme conversion processes may help (see Salesky et al. 2020). Once lexical data from New Guinea has been acquired and segmented, the next major challenge will be dataset sparsity. This is because the phonemic inventories of languages in New Guinea are diverse. This results in a greater quantity of mismatches between languages, where a biphone character for a biphone present in one language is recorded as missing for another language because the other language lacks a segment from its inventory entirely. Biphone frequency transition data, in the form used throughout this thesis, might only give low level phylogenetic information between smaller families/subgroups that share similar phonemic inventories. Matches between more distantly related languages are more likely to be spurious due to the segments being generally common cross-linguistically. For example, just about every language will have the sequence /ma/, but whether the frequency of a transition from /m/ to /a/ is indicative of deep-time relationships between hitherto unrelated languages is untested (and

perhaps doubtful on the face of it). There is no straightforward way around this challenge, but there are a couple of possibilities. Firstly, future evaluation of the evolutionary dynamics of phonotactics should attempt to quantify empirically where in the tree phylogenetic signal is coming from (particular clades or particular levels of time-depth) rather than simply quantifying phylogenetic signal throughout the tree overall. Secondly, one potential way to overcome dataset sparsity is to include different kinds of phonotactic variables with less mismatch between languages. This approach would, however, likely come with the trade-off of reducing the size of the dataset and/or introducing new issues of independence between variables.

Turning now to the second major limitation of phylogenetic phonotactics, there is the issue that biphones are an extremely simple representation of the phonotactic system of a language. Biphone characters only capture the effects of phonotactic restrictions on immediately adjacent segments, they do not account for phonotactic restrictions on longer strings, e.g. clusters of three consonants, nor phonotactic rules concerning morpheme boundaries. This might be especially significant in Australia where, as discussed in Chapter 2, word structure has been traditionally described using a disyllabic template. Furthermore, there is no distinction between vowels and consonants in biphone characters, nor distinct consonant slots (e.g. C_1, C_2 and C_3 in the disyllabic template discussed in Chapter 2). In response to these limitations, future studies should consider alternative phonotactic variables. Perhaps the most straightforward response would be to consider trigrams, or even longer n-gram sequences. This would capture phonotactic dependencies across longer distances and would greatly expand the size of the dataset, but it would face the same challenges of dataset sparsity discussed above. Other possibilities include the extraction of transition frequencies between consonants, ignoring vowels. Vowel-to-vowel transitions, ignoring intervening consonants, would also be a possibility. Recording transitions across morpheme boundaries and syllable boundaries would be a possibility, although it would require syllabified lexical data. AusPhon lexical data would require syllabification, with all the challenges of determining precise syllable boundaries that that entails. Lastly, there is the option of incorporating natural sound class information in the dataset, such as the dataset of frequency transitions between place and manner sound classes tested in Chapters 5–6. This helps solve the problem of dataset sparsity, since languages are likely to share many of the same places and manners of articulation, even if they do not share many directly comparable phonemes. However, it also introduces issues associated with binning of data into larger categories. Phonemes share overlapping phonological features and, therefore, may be binned into different overlapping natural-class-based variables and counted different numbers of times. Again, there is no easy solution to this issue. However, further consideration of evolutionary models and how they apply to phonotactic datasets is likely to help. I turn to discussion of evolutionary models now.

The third limitation to discuss is the sole reliance on a Brownian motion model of biphone frequency evolution in this thesis. As discussed in Chapter 5, Brownian motion is a simple model to implement and makes a natural starting point for investigation. But future studies should consider whether the Brownian motion model or another model is more appropriate, with consideration of our existing knowledge of sound change. Brownian motion, in which a continuous character value can

wander up or down with equal probability, is well suited to shifts in biphone frequency associated with lexical changes. As languages gain or lose individual lexical items, the frequencies of transitions from segments in that lexical item will shift minutely. Over time, these minute shifts will accumulate, resulting in shifts in character values that resemble Brownian motion. The biggest problem for this model is that lexical replacement and innovation are not the only diachronic processes affecting biphone transition frequencies. Sound change processes will create sudden jumps in biphone character values, often to 0 or 1, as well as having the effect of both creating and eliminating new biphone data points in a language. Consider an assimilation process in a particular environment, for example, place assimilation of nasals to match a following stop. This will cause biphone transition frequencies for heterorganic nasal+stop sequences to jump to 0 and homorganic nasal+stop sequences to jump to 1. Relatedly, continuous models frequently cannot account for absolute 0 and 1 frequency values, which is problematic since 0 and 1 frequency values are meaningful in phonotactics, given binary phonotactic rules that place absolute restrictions on certain phonemes in certain environments or the kinds of assimilation processes that guarantee certain segments in certain environments as just described. With regards to sound change, phonemic splits and mergers will create new biphone data points and eliminate others completely for a language. In addition, sound change can cause erroneous matches and mismatches between biphone characters for different languages, owing to the direct matching of comparable phonemes. For example, a series of fricatives in language A might be historically related to a series of voiced stops in language B. In the methods used in this thesis, biphone characters involving fricatives in language A would not be matched to biphone characters involving voiced stops in language B, despite their common origin. The problem of sound change is genuinely difficult and deserves prioritisation in future work. Notwithstanding efforts at automated sequence alignment, a true computational model of sound change remains elusive. With regards to evolutionary models beyond Brownian motion, an immediate short-term priority is to test whether Brownian motion is best supported over other off-the-shelf models such as punctuated equilibrium or the Ornstein-Uhlenbeck model. Moreover, future work should incorporate a Lévy process which allows for sudden, discontinuous jumps. There is some preliminary work in this space (Landis, Schraiber & Liang 2012, Landis & Schraiber 2017, Blomberg, Rathnayake & Moreau 2020). One potential path for future evaluation would be to work with simulated data. That is, rather than working with natural language data, various sound change and lexical change processes could be simulated on a wordlist, and the resulting frequency distributions could be compared to actual observed frequency distributions using a maximum likelihood framework, as in Chapter 3, or similar.

A related concern for future research concerns the independence of phonotactic variables and inference of evolutionary parameters. Throughout this thesis, biphone characters have been treated as independent. This meant that phyogenetic signal was inferred for each character individually in Chapter 5 and, likewise, independent evolutionary processes were inferred for each individual character in Chapter 6. However, the independence of biphone transition frequencies is not a tenable assumption, due to the tendency of phonemes to form natural classes. Phonotactic rules tend to apply to natural classes of segments, rather than an individual segment, and will therefore manifest in the

frequencies of a whole group of biphone characters. Likewise, as discussed above, sound change processes will also tend to apply to whole classes of phonemes and therefore affect the frequencies of whole classes of biphone characters. One simple way of dealing with this is to bin phonemes into natural classes and calculate transition frequencies between classes rather than individual segments, as illustrated in Chapters 5–6. This is not entirely satisfying, however, due to the overlapping nature of natural classes and considerable reduction in dataset size. Another approach is to treat all biphone characters as part of a single, multidimensional system. Methods for doing this are being developed in the field of morphometrics, in which one often has to work with complex, multivariate datasets of interrelated measurements (e.g. various measurements capturing the shape of a skull). I have made one early attempt to measure phylogenetic signal using such a method (Macklin-Cordes & Round 2018), however, this reduces an entire dataset of phonotactic variation to a single distance matrix. Treating biphone data as a single, multivariate system also renders phylogenetic tree inference impractical. For example, a frequency dataset of the scale of the 2,236 binary biphone characters used in Chapter 6 would constitute a single multivariate diffusion model of 2,236 dimensions, which is too computationally intensive to be repeated millions of times over in a Markov Chain Monte Carlo process. Between these two extremes, there are options for partitioning biphone data into k independent groups. The issue here is how to decide on how the number of independent groups and how to delineate the boundaries between them. This is a challenge and an active area of development in evolutionary biology as well, particularly given the rise of morphometrics. One approach being developed currently is phylogenetic factor analysis (Tolkoff et al. 2018, Hassler et al. 2020), which infers groupings itself (removing the need for the analyst to partition the data beforehand) but still requires the analyst to define the k number of groups beforehand. Besides following the methodological development in this space, future work towards understanding the patterns of correlation and interaction between phonotactic characters will be useful.

7.3 Future directions

Given everything we know about phonology, phonotactics, sound change and language change generally, what would an ideal evolutionary model for phonotactics look like? Most straightforwardly, a more informed implementation of the expected frequency distributions of phonotactic frequency characters would be helpful, along with a Lévy model that allows frequency characters to make particular sudden jumps, as we would expect from certain sound change processes, rather than wander up or down only gradually. A more complex question, which will require extra statistical evaluation as well as theoretical contemplation to address, is the question of non-independence and correlated evolution between phonotactic characters, due to the non-independent behaviour of phoneme segments and overlapping phonological natural classes. Relatedly, the ideal data partitioning structure remains to be determined. The importance of data partitioning for inferring accurate tree topologies and reliable posterior probabilities is emphasised by (Brown & Lemmon 2007). It seems likely that the ideal phylogenetic model for phonotactic frequencies would partition the data into a number of smaller

groups of related characters, somewhere between the two extremes of treating every character as independent and treating all characters as a single, high-dimensional structure, neither of which is particularly linguistically realistic.

Ultimately, it is highly unlikely that simply adding phonotactic frequency data, even if paired with an ideal evolutionary model, will be the secret ingredient necessary to push back the time-depth of phylogenetic inferences and identify ancient linguistic relationships across Sahul. However, if the implementation of phonotactic frequency data in phylogenetic algorithms could be improved, the benefits could include more solid reconstructions of lower-level relationships, an additional line of evidence for solving conflicts and ambiguities in prior studies, and additional information from which branch lengths can be inferred. There is no guarantee on any of these points. It is equally possible that further study will find that phonotactic frequencies genuinely contain non-treelike noise that cannot be attributed to deficiencies of the evolutionary model. This extended evaluation of phonotactics in phylogenetics, addressing the limitations discussed in this chapter, would be justified, however. The potential pay-offs could help piece together a better picture of the linguistic phylogenetics of Australia, and ultimately contribute to the multidisciplinary investigations of the human history of Sahul.

Chapter 8

Conclusion

What do the minutiae of sound sequences in human language encode about the forces of human history that shaped them? This thesis has shown that the minutiae of sound sequences in human language do appear to contain a signal of human history. Presently, however, their integration in a successful phylogenetic model for the purposes of inferring phylogenetic trees remains elusive. Nevertheless, it is valuable to conclude by recalling the purpose of this investigation into phonotactics in linguistic phylogenetics, and to emphasise the lessons it holds for future data innovation in this field?

The first two decades of the 21st century in historical linguistics have been characterised by the mainstream adoption of quantitative methods, particularly Bayesian phylogenetic methods for inferring phylogenetic trees and testing other evolutionary questions (see Dunn 2015, Bowern 2018a). Notwithstanding the many successes of this quantitative turn, a rate-limiting step is the acquisition of lexical cognate data which is typically resource-intensive and at least partially, if not entirely, a manual process. A consequence, also, of the predominance of lexical cognate data in modern linguistic phylogenetics is that most of the resulting linguistic phylogenies inferred in recent years have been inferred from a single source of linguistic evidence representing only one sub-part of the linguistic system. Although it seems unlikely that lexical cognate data, which has been the bedrock of the historical linguistic Comparative Method for over 150 years, will be surpassed as the primary linguistic data source any time soon, it is timely to consider whether modern quantitative historical linguistics could benefit from data from other parts of language as well.

This thesis presented an interrogation of phonotactic data, in particular phonotactic frequency data extracted from Australian language wordlists, in linguistic phylogenetics. This focus on phonotactics stemmed from the observation that a language's phonotactic system tends to remain conservative in the face of language contact and semantic shift events which create noise in the historical signal in lexical data. Frequency data was considered, since Australian languages contain fine-grained variation at the frequency level that is not immediately apparent at a more categorical or binary level, and further, frequencies seem to have a psychological reality in the minds of speakers. One final motivation is the efficiency with which phonotactic data can be extracted from wordlists, which is advantageous in parts of the world where language documentation and research is relatively sparse.

Following the introduction and overview in Chapter 1, the thesis began with a literature review in Chapter 2. This covered the history of phylogenetic thinking in historical linguistics. Historical linguistics has been argued to be an inherently phylogenetic field (Dunn 2015) and has a history of phylogenetic tree inference extending back even further than Darwin's seminal work in biology. However, the uptake of computational phylogenetic methods in linguistics is relatively recent. Over the past 20 years, linguistic phylogenies have been inferred using these methods for numerous language families the world over and the subfield sits increasingly within the mainstream of historical linguistics. This discussion was followed by an overview of Australian languages in Section 2.2. Australia is a linguistically diverse continent, home to at least 250–300 distinct languages at the time of European arrival (of which around 120 remain presently with living speakers). By far the largest Australian language family is Pama-Nyungan, which covers nearly 90% of the continent. Some 27 other language families and isolates are packed into Australia's Top End and Kimberley regions, while the linguistic picture in Tasmania to the south, which suffered disproportionately from colonisation, remains unclear. One area in which Australian languages curiously have been described as lacking in diversity is phonology, extending to phonotactics. Australian languages have been described as sharing similar, consonant-heavy phonemic inventories, varying along certain restricted lines. Recent studies that consider frequency data both challenge and nuance these characterisations, however (Gasser & Bowern 2014, Round 2021b,a). The chapter concluded with Section 2.3, which placed Australian languages and linguistic phylogenetics within the context of broader interest in the human history of Sahul, the continent of Australia and New Guinea. Linguistic evidence will continue to be an important component for increasing our understanding of Sahul's human history, particularly from the early Holocene onwards.

The next chapter, Chapter 3, presented a re-evaluation of the shape of rank-frequency distributions of phonemes in a language, as extracted from the wordlists of 166 Australian languages. Phoneme frequencies were re-evaluated by using a maximum likelihood statistical framework (Clauset, Shalizi & Newman 2009) to test whether observed distributions of phoneme frequencies are consistent with several similar, competing mathematical distributions. This method has been shown to be more reliable than the statistical methods used to evaluate phoneme frequency distributions in the past. Individual languages in the sample show a high degree of variation and there is a wide range of uncertainty due to the small number of observations in each language (limited to the number of contrastive phonemes in the language). An overall tendency does emerge, however. The most frequent phonemes in a language tend to follow a Zipfian-like power law distribution (though perhaps also a lognormal distribution) while the least frequent phonemes tend to follow a geometric (or exponential) distribution. Mathematically, power laws are associated with preferential attachment processes and geometric distributions are associated with birth-death processes. The chapter concludes with reasoned speculation that phoneme frequency distributions are consistent with a variety of sound changes (e.g. phonemic mergers and splits), operationalised as birth-death stochastic processes, with a separate preferential attachment dynamic in which phones tend to shift into numerically stronger categories.

Chapter 4 followed by making the case for phylogenetic comparative methods in linguistics. Com-

parative linguistics, which includes historical linguistics but also typology, is necessarily phylogenetic due to the shared histories between languages descended from a common ancestor. This phenomenon, phylogenetic autocorrelation, has long been recognised in comparative linguistics. However, other comparative fields of science have diverged from linguistics in recent decades towards quantitative phylogenetic comparative methods for controlling for phylogenetic autocorrelation. By contrast, linguistics largely tends to continue controlling for phylogenetic autocorrelation indirectly through, for example, creating phylogenetically balanced language samples. I argued that with the proliferation of detailed linguistic phylogenies that have been inferred in the last 20 years, a priority for comparative linguistics should be a continued shift towards phylogenetic comparative methods. An offshoot of phylogenetic comparative methods is a selection of statistics for measuring phylogenetic signal (e.g. Blomberg, Garland & Ives 2003), which can quantify to what degree some observed variation is attributable to the phylogenetic distances between languages or species. This is useful for quantifying the degree to which phylogeny needs to be controlled for as a confound in a synchronic, comparative study. However, it is also useful as a diagnostic tool for determining whether a comparative dataset contains phylogenetic signal which could help infer phylogenetic trees. In this latter context, I employ some of these tools for measuring phylogenetic signal in Chapter 5.

In Chapter 5, I tested the hypothesis that phonotactics could contain phylogenetic signal. Phylogenetic signal was measured in three datasets coding the presence or absence of biphones, frequencies of transitions between segments and frequencies of transitions between natural sound classes. The datasets were extracted from 111 Pama-Nyungan languages and phylogenetic signal was measured using a pre-existing, independent reference phylogeny inferred from traditional lexical cognate data. Phylogenetic signal was detected in all three datasets. Not unexpectedly, the signal strength was relatively reduced in the binary data and greater in the finer-grained frequency data. The strongest phylogenetic signal was detected in the sound class transition frequency dataset. These results supported the hypothesis that sound sequences might encode phylogenetic information and motivated further evaluation of phonotactics in phylogenetic tree inference.

Chapter 6 presented an attempt to infer a phylogeny of 44 western Pama-Nyungan languages, using phonotactic data in conjunction with pre-existing lexical cognate data. The chapter started with a preliminary test to evaluate the best evolutionary model parameter settings for binary biphone characters and a second preliminary test to reproduce earlier work (Bouckaert, Bowern & Atkinson 2018) using only lexical cognate data. A simple Brownian motion model was applied to sound class transition frequency data. Independent evolutionary rates and likelihood scores were inferred for each frequency character due to the heavy computational demands of a multivariate continuous character model. A Bayesian MCMC process was run on two phylogenetic models. The first model, termed the 'linked' model, included a single tree inferred from the combination of lexical and phylogenetic data. The second, termed the 'separate' model, included two trees—one inferred from lexical data alone and the other inferred from phonotactic data separately. Marginal likelihood estimates, which give an overall indication of model fit and enable model comparison, were inferred for each model. However, the marginal likelihood estimates in this study proved to be unreliable. Some limited insight

could be gleaned from comparing the maximum clade credibility trees generated by the linked model and the cognate-only model from the preliminary section, suggesting that some non-treelike signal possibly was being introduced by the phonotactic data, but firm conclusions were difficult to draw. The difficulty I encountered is that it is very computationally expensive to infer a phylogenetic tree from a highly multidimensional partition of continuous-valued characters. Treating each continuous character as evolutionarily independent simplifies the model computationally but, firstly, introduces orders of magnitude more free parameters to estimate and, secondly, makes the model less realistic. The combined effect is that the models with frequency-based phonotactic characters fail to converge consistently and fail to produce reliable marginal likelihood estimates. Looking ahead, I suggested that future interrogations of phonotactics in phylogenetics should consider more carefully the frequency distributions of biphone-based characters (not just frequencies of single segments, as in Chapter 3) and re-evaluate whether the Brownian motion model of character evolution is ideal or should be modified. Secondly, patterns of co-evolution among phonotactic characters should be evaluated. We expect that phonotactic characters will not be completely independent from one another but the models in this study do not account for this. Between the extremes of a single highly dimensional partition of continuous characters and completely independent partitions for each continuous character, there may be a middle ground in which characters are partitioned into smaller groups that both make sense linguistically and keep the model computationally practical.

Finally, the previous chapter, Chapter 7, drew together the four previous chapters into an overall assessment of phonotactics in phylogenetics. Notwithstanding the unreliability of the results in Chapter 6, and the ambiguity it leaves us with, it is worth reconfirming what this foray into phonotactic phylogenetics has uncovered. There are two parts to this. Most immediately, the results highlight the non-triviality of computationally inferring a phylogeny with a large dataset of continuous-valued characters. A 'plug in and play' approach will simply not work from a practical standpoint, even if it were acceptable in a linguistic sense. I have detailed the future steps for model development that will be necessary to address this. More generally though, I hope to have made the following contribution: A set of steps for the empirical interrogation of linguistic data in phylogenetics, which is generalisable to any set of linguistic data, the application of which is novel in this context. These steps proceed as follows:

(1) Articulate a well-reasoned hypothesis on how the linguistic characters in question are expected to change through time.
(2) Operationalise the language change process as a stochastic mathematical process and statistically evaluate whether the characters' observed frequency distributions fit the associated mathematical distribution.
(3) Scan the data for phylogenetic signal, using an independent reference phylogeny.
(4) Create a phylogenetic model, paying attention to the modelling of both the evolution of individual characters and the relationships between characters.
(5) Evaluate whether the addition of the novel data strengthens phylogenetic tree inference, by comparing the model fit of two otherwise equivalent models, one in which the novel data is

combined with traditional lexical cognate data and the other in which the novel data and lexical data are kept separate.

For the sake of under-resourced parts of the linguistic world, or parts where the lexical historical signal is noisy, data innovation should remain a priority for linguistic phylogenetics. With this thesis, I hope to have demonstrated one empirical approach to this task.

Bibliography

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

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

Albert, James S., Henry J. Bart Jr. & Roberto E. Reis. 2011. Species richness and cladal diversity. In James S. Albert & Roberto E. Reis (eds.), *Historical biogeography of neotropical freshwater fishes*, 89–104. Berkeley, CA: University of California Press.

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.

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.

Andersen, Henning. 2006. Synchrony, diachrony, and evolution. In Ole Nedergaard Thomsen (ed.), *Competing models of linguistic change: Evolution and beyond*, 59–90. Amsterdam: John Benjamins Publishing.

Atkinson, Quentin D. & R. D. Gray. 2005. Curious parallels and curious connections: Phylogenetic thinking in biology and historical linguistics. *Systematic Biology* 54(4). 513–526. https://doi.org/Doi10.1080/10635150590950317.

Atkinson, Quentin D., Geoff Nicholls, David Welch & Russell Gray. 2005. From words to dates: Water into wine, mathemagic or phylogenetic inference? *Transactions of the Philological Society* 103(2). 193–219. https://doi.org/10.1111/j.1467-968X.2005.00151.x.

Austin, Peter K. 1981a. *A grammar of Diyari, South Australia* (Cambridge Studies in Linguistics 32). Cambridge; New York: Cambridge University Press. 269 pp.

Austin, Peter K. 1981b. Proto-Kanyara and Proto-Mantharta historical phonology. *Lingua* 54(4). 295–333. https://doi.org/10.1016/0024-3841(81)90009-7.

Australian Government Department of Infrastructure, Transport, Regional Development and Communications, Jacqueline Battin, Jason Lee, Douglas E. Marmion, Rhonda Smith, Tandee Wang, Yonatan Dinku, Janet Hunt, Francis Markham, Denise Angelo, Emma Browne, Inge Kral, Carmel

O'Shannessy, Jane Simpson & Hilary Smith. 2020. *National Indigenous Languages Report*. https://www.arts.gov.au/what-we-do/indigenous-arts-and-languages/national-indigenous-languages-report.

Baayen, Harald. 2008. *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Baayen, R. Harald. 2001. *Word frequency distributions* (Text, Speech and Language Technology). Dordrecht, Netherlands: Springer Science+Business Media.

Baele, Guy, Philippe Lemey, Trevor Bedford, Andrew Rambaut, Marc A. Suchard & Alexander V. Alekseyenko. 2012. Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Molecular Biology and Evolution* 29(9). 2157–2167. https://doi.org/10.1093/molbev/mss084.

Baele, Guy, Wai Lok Sibon Li, Alexei J. Drummond, Marc A. Suchard & Philippe Lemey. 2013. Accurate model selection of relaxed molecular clocks in Bayesian phylogenetics. *Molecular Biology and Evolution* 30(2). 239–243. https://doi.org/10.1093/molbev/mss243.

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.

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

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

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

Balter, Michael. 2004. Search for the Indo-Europeans. *Science* 303(5662). 1323–1326. https://doi.org/10.1126/science.303.5662.1323.

Barford, Paul, Azer Bestavros, Adam Bradley & Mark Crovella. 1999. Changes in web client access patterns: Characteristics and caching implications. *World Wide Web* 2(1). 15–28. https://doi.org/10.1023/A:1019236319752.

Barford, Paul & Mark Crovella. 1998. Generating representative web workloads for network and server performance evaluation. In *Measurement and modeling of computer systems (SIGMETRICS)*, 151–160. New York: Association for Computing Machinery (ACM). https://doi.org/10.1145/277851.277897.

Barndorff-Nielsen, Ole E. & David Roxbee Cox. 1994. *Inference and asymptotics* (Monographs on statistics and applied probability 52). London: Chapman & Hall. 360 pp.

Bateman, Richard, Ives Goddard, Richard O'Grady, V. A. Funk, Rich Mooi, W. John Kress & Peter Cannel. 1990. Speaking of forked tongues: The feasibility of reconciling human phylogeny and the history of language. *Current anthropology* 31(1). 1–24.

Becker-Kristal, Roy. 2010. *Acoustic typology of vowel inventories and Dispersion Theory: Insights from a large cross-linguistic corpus*. Los Angeles, CA: University of California Ph.D. Dissertation.

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

Benjamin, Jonathan, Michael O'Leary, Jo McDonald, Chelsea Wiseman, John McCarthy, Emma Beckett, Patrick Morrison, Francis Stankiewicz, Jerem Leach, Jorg Hacker, Paul Baggaley, Katarina Jerbić, Madeline Fowler, John Fairweather, Peter Jeffries, Sean Ulm & Geoff Bailey. 2020. Aboriginal artefacts on the continental shelf reveal ancient drowned cultural landscapes in northwest Australia. *PLOS ONE* 15(7). e0233912. https://doi.org/10.1371/journal.pone.0233912.

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

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

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

Birchall, Joshua, Michael Dunn & Simon J. Greenhill. 2016. A combined comparative and phylogenetic analysis of the Chapacuran language family. *International Journal of American Linguistics* 82(3). 255–284. https://doi.org/10.1086/687383.

Bird, Michael I., Scott A. Condie, Sue O'Connor, Damien O'Grady, Christian Reepmeyer, Sean Ulm, Mojca Zega, Frédérik Saltré & Corey J. A. Bradshaw. 2019. Early human settlement of Sahul was not an accident. *Scientific Reports* 9(1). 8220. https://doi.org/10.1038/s41598-019-42946-9.

Birdsell, J. B. 1977. The recalibration of a paradigm for the first peopling of greater Australia. In Jim Allen, Jack Golson & Rhys Jones (eds.), *Sunda and Sahul: Prehistoric studies in Southeast Asia, Melanesia and Australia*, 113–167. London: Academic Press.

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

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.

Blasi, Damián E., Søren Wichmann, Harald Hammarström, Peter F. Stadler & Morten H. Christiansen. 2016. Sound–meaning association biases evidenced across thousands of languages. *Proceedings of the National Academy of Sciences* 113(39). 10818–10823. https://doi.org/10.1073/pnas.1605782113.

Blench, Roger. 2015. 'New mathematical methods' in linguistics constitute the greatest intellectual fraud in the discipline since Chomsky. Nijmegen, Netherlands. https://www.academia.edu/13181187/_New_mathematical_methods_in_linguistics_constitute_the_greatest_intellectual_fraud_in_the_discipline_since_Chomsky.

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.

Blomberg, Simone P, Suren I Rathnayake & Cheyenne M Moreau. 2020. Beyond Brownian motion and the Ornstein-Uhlenbeck process: Stochastic diffusion models for the evolution of quantitative characters. *The American Naturalist* 195(2). 145–165.

Bloomfield, Leonard. 1935. *Language*. London: Allen & Unwin.

Blust, Robert. 2000. Why lexicostatistics doesn't work: The 'universal constant' hypothesis and the Austronesian languages. In Colin Renfrew, April M. S. McMahon & R. L. Trask (eds.), *Time depth in historical linguistics*, vol. 2, 31–332. Cambridge: McDonald Institute for Archaeological Research.

Blust, Robert. 2004. *t to k: An Austronesian sound change revisited. *Oceanic Linguistics* 43(2). 365–410. https://doi.org/10.1353/ol.2005.0001.

Boretzky, Norbert. 1991. Contact-induced sound change. *Diachronica* 8(1). 1–15.

Borodovsky, Mark Yu & S. M. Gusein-Zade. 1989. A general rule for ranged series of codon frequencies in different genomes. *Journal of Biomolecular Structure and Dynamics* 6(5). 1001–1012. https://doi.org/10.1080/07391102.1989.10506527.

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

Bouckaert, Remco 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.

Bouckaert, Remco R., Joseph Heled, Denise Kühnert, Tim Vaughan, Chieh-Hsi Wu, Dong Xie, Marc A. Suchard, Andrew Rambaut & Alexei J. Drummond. 2014. BEAST 2: A software platform for Bayesian evolutionary analysis. *PLoS Computational Biology* 10(4). e1003537. https://doi.org/10.1371/journal.pcbi.1003537.

Bouckaert, Remco R., Philippe Lemey, Michael Dunn, Simon J. Greenhill, Alexander V. Alekseyenko, Alexei J. Drummond, Russell D. Gray, Marc A. Suchard & Quentin D. Atkinson. 2012a. Corrections and clarifications. *Science* 342(6165). 1446. https://doi.org/10.1126/science.342.6165.1446-a.

Bouckaert, Remco R., Philippe Lemey, Michael Dunn, Simon J. Greenhill, Alexander V. Alekseyenko, Alexei J. Drummond, Russell D. Gray, Marc A. Suchard & Quentin D. Atkinson. 2012b. Mapping

the origins and expansion of the indo-european language family. *Science* 337(6097). 957–960. https://doi.org/10.1126/science.1219669.

Bouckaert, Remco R., Timothy G. Vaughan, Joëlle Barido-Sottani, Sebastián Duchêne, Mathieu Fourment, Alexandra Gavryushkina, Joseph Heled, Graham Jones, Denise Kühnert, Nicola De Maio, Michael Matschiner, Fábio K. Mendes, Nicola F. Müller, Huw A. Ogilvie, Louis du Plessis, Alex Popinga, Andrew Rambaut, David Rasmussen, Igor Siveroni, Marc A. Suchard, Chieh-Hsi Wu, Dong Xie, Chi Zhang, Tanja Stadler & Alexei J. Drummond. 2019. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLOS Computational Biology* 15(4). e1006650. https://doi.org/10.1371/journal.pcbi.1006650.

Bowe, Heather & Stephen Morey. 1999. *The Yorta Yorta (Bangerang) language of the Murray Goulburn including Yabula Yabula* (Pacific Linguistics Series C 154). Canberra: Pacific Linguistics. 286 pp.

Bowern, Claire. 2010. Historical linguistics in Australia: Trees, networks and their implications. *Philosophical Transactions of the Royal Society B-Biological Sciences* 365(1559). 3845–3854. https://doi.org/Doi10.1098/Rstb.2010.0013.

Bowern, Claire. 2012. The riddle of tasmanian languages. *Proceedings of the Royal Society B-Biological Sciences* 279(1747). 4590–4595. https://doi.org/Doi10.1098/Rspb.2012.1842.

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

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

Bowern, Claire. 2017. Standard Average Australian? In *Association for Linguistic Typology (ALT)*. Australian National University, Canberra, Australia. https://doi.org/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*. 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. 2004. *Australian languages: Classification and the comparative method* (Amsterdam studies in the theory and history of linguistic science Series IV, Current issues in linguistic theory, 249). Amsterdam, Philadelphia: John Benjamins.

Bradshaw, Corey J. A., Sean Ulm, Alan N. Williams, Michael I. Bird, Richard G. Roberts, Zenobia Jacobs, Fiona Laviano, Laura S. Weyrich, Tobias Friedrich, Kasih Norman & Frédérik Saltré. 2019. Minimum founding populations for the first peopling of Sahul. *Nature Ecology & Evolution* 3(7). 1057–1063. https://doi.org/10.1038/s41559-019-0902-6.

Breen, Gavan. 1981a. Margany and Gunya. In R. M. W. Dixon & Barry Blake (eds.), *Handbook of Australian languages*, vol. 2, 275–394. Amsterdam: John Benjamins.

Breen, Gavan. 1997. Taps, stops & trills. In Darrell T. Tryon & Michael Walsh (eds.), *Boundary rider: Essays in honour of Geoffrey O'Grady*, 71–93. Canberra: Pacific Linguistics.

Breen, Gavan. 2001. The wonders of Arandic phonology. In Jane Simpson, David Nash, Mary Laughren, Peter K. Austin & Barry J. Alpher (eds.), *Forty years on: Ken Hale and Australian languages* (Pacific Linguistics 512), 45–69. Canberra: Pacific Linguistics.

Breen, Gavan & Barry J. Blake. 2007. *The grammar of Yalarnnga: A language of western Queensland* (Pacific linguistics 584). Canberra, ACT: Pacific Linguistics.

Bromham, Lindell. 2017. Curiously the same: Swapping tools between linguistics and evolutionary biology. *Biology & Philosophy* 32(6). 855–886. https://doi.org/10.1007/s10539-017-9594-y.

Bromham, Lindell, Xia Hua, Thomas G. Fitzpatrick & Simon J. Greenhill. 2015. Rate of language evolution is affected by population size. *Proceedings of the National Academy of Sciences* 112(7). 2097–2102. https://doi.org/10.1073/pnas.1419704112.

Browman, Catherine P. & Louis M. Goldstein. 1986. Towards an articulatory phonology. *Phonology* 3(01). 219–252. https://doi.org/10.1017/S0952675700000658.

Brown, Cecil H., Eric W. Holman & Søren Wichmann. 2013. Sound correspondences in the world's languages. *Language* 89(1). 4–29. https://doi.org/10.1353/lan.2013.0009.

Brown, Jeremy M. & Alan R. Lemmon. 2007. The importance of data partitioning and the utility of Bayes factors in Bayesian phylogenetics. *Systematic Biology* 56(4). 643–655. https://doi.org/10.1080/10635150701546249.

Brzezinski, Michal. 2014. Power laws in citation distributions: Evidence from Scopus. *arXiv e-prints*, arXiv:1402.3890.

Busby, Peter A. 1982. *The distribution of phonemes in Australian Aboriginal languages* (Series A-60 14). Canberra: Pacific Linguistics.

Butcher, Andy. 2006. Australian Aboriginal languages: Consonant-salient phonologies and the place-of-articulation imperative. In Jonathan Harrington & Marija Tabain (eds.), *Speech production: Models, phonetic processes, and techniques*, 187–210. New York: Psychology Press.

Bybee, Joan L. 1985. *Morphology: A Study of the Relation Between Meaning and Form* (Typological Studies in Language 9). Amsterdam: John Benjamins Publishing.

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. 1937. The structure of Australian languages. *Oceania* 8(1). 27–61. https://doi.org/10.1002/j.1834-4461.1937.tb00405.x.

Capell, Arthur. 1940. The classification of languages in north and north-west Australia. *Oceania* 10(3). 241–272. https://doi.org/10.1002/j.1834-4461.1940.tb00292.x.

Capell, Arthur. 1956. *A new approach to Australian linguistics*. Sydney: University of Sydney.

Capell, Arthur. 1979. The history of Australian languages: A first approach. In Stephen A. Wurm (ed.), *Australian linguistic studies*, vol. C54, 419–619. Canberra, Australia: Pacific Linguistics.

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.

Chen, Matthew Y. & William S-Y. Wang. 1975. Sound change: Actuation and implementation. *Language* 51(2). 255–281. https://doi.org/10.2307/412854.

Cho, Eunjoon, Seth A. Myers & Jure Leskovec. 2011. Friendship and mobility: User movement in location-based social networks. In *Knowledge discovery and data mining (SIGKDD)*, 1082–1090. New York: Association for Computing Machinery (ACM). https://doi.org/10.1145/2020408.2020579.

Chung, Kee H. & Raymond A. K. Cox. 1994. A stochastic model of superstardom: An application of the yule distribution. *The Review of Economics and Statistics*. 771–775.

Clarkson, Chris, Zenobia Jacobs, Ben Marwick, Richard Fullagar, Lynley Wallis, Mike Smith, Richard G. Roberts, Elspeth Hayes, Kelsey Lowe, Xavier Carah, S. Anna Florin, Jessica McNeil, Delyth Cox, Lee J. Arnold, Quan Hua, Jillian Huntley, Helen E. A. Brand, Tiina Manne, Andrew Fairbairn, James Shulmeister, Lindsey Lyle, Makiah Salinas, Mara Page, Kate Connell, Gayoung Park, Kasih Norman, Tessa Murphy & Colin Pardoe. 2017. Human occupation of northern Australia by 65,000 years ago. *Nature* 547(7663). 306–310. https://doi.org/10.1038/nature22968.

Clarkson, Chris, Mike Smith, Ben Marwick, Richard Fullagar, Lynley A. Wallis, Patrick Faulkner, Tiina Manne, Elspeth Hayes, Richard G. Roberts, Zenobia Jacobs, Xavier Carah, Kelsey M. Lowe, Jacqueline Matthews & S. Anna Florin. 2015. The archaeology, chronology and stratigraphy of Madjedbebe (Malakunanja II): A site in northern Australia with early occupation. *Journal of Human Evolution* 83. 46–64. https://doi.org/10.1016/j.jhevol.2015.03.014.

Clauset, A., C. R. Shalizi & M. E. J. Newman. 2009. Power-law distributions in empirical data. *SIAM Review* 51(4). 661–703. https://doi.org/10.1137/070710111.

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

Coleman, John & Janet Pierrehumbert. 1997. Stochastic phonological grammars and acceptability. In *Computational phonology: The third meeting of the ACL special interest group in computational phonology*, 49–56. Somerset, NJ: Association for Computational Linguistics. http://arxiv.org/abs/cmp-lg/9707017.

Crawford, Clifford James. 2009. *Adaptation and transmission in Japanese loanword phonology*. Ithaca, NY: Cornell University Ph.D. Dissertation. https://hdl.handle.net/1813/13947.

Crowley, Terry. 1981. The Mpakwithi dialect of Anguthimri. In R. M. W. Dixon & Barry Blake (eds.), *Handbook of Australian languages*, vol. 2, 147–196. Amsterdam: John Benjamins.

Crowley, Terry. 1983. Uradhi. In R. M. W. Dixon & Barry J. Blake (eds.), *Handbook of Australian languages*, vol. 3, 5 vols., 307–428. Amsterdam: John Benjamins.

Cysouw, Michael. 2009. On the probability distribution of typological frequencies. In Christian Ebert, Gerhard Jäger & Jens Michaelis (eds.), *The mathematics of language*, 29–35. Berlin: Springer.

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.

Darwin, Charles. 1859. *On the origin of species by means of natural selection*. London: J. Murray.

Darwin, Charles. 1871. *The descent of man, and selection in relation to sex*. D. Appleton.

Dediu, Dan. 2018. Making genealogical language classifications available for phylogenetic analysis: Newick trees, unified identifiers, and branch length. *Language Dynamics and Change* 8(1). 1–21. https://doi.org/10.1163/22105832-00801001.

Dediu, Dan & Bart de Boer. 2016. Language evolution needs its own journal. *Journal of Language Evolution* 1(1). 1–6. https://doi.org/10.1093/jole/lzv001.

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. 1970. Languages of the Cairns rain forest region. In Stephen A. Wurm & Donald C. Laycock (eds.), *Pacific linguistic studies in honour of arthur capell*, vol. 13 (Pacific Linguistics: Series C), 651–687. Canberra: Research School of Pacific & Asian Studies, Australian National University.

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

Dixon, R. M. W. 2002. *Australian languages: Their nature and development* (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-2017)*. New York: Linguistic Society of America. https://doi.org/10.3765/amp.v5i0.4238.

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.

Donohue, Mark, Tim Denham & Stephen Oppenheimer. 2012. New methodologies for historical linguistics?: Calibrating a lexicon-based methodology for diffusion vs. subgrouping. *Diachronica* 29(4). 505–522. https://doi.org/10.1075/dia.29.4.04don.

Downey, A. B. 2001. The structural cause of file size distributions. In *Modeling, analysis and simulation of computer and telecommunication systems (MASCOTS)*, 361–370. Cincinnati, OH. https://doi.org/10.1109/MASCOT.2001.948888.

Dresher, B. Elan. 2009. *The contrastive hierarchy in phonology*. Cambridge: Cambridge University Press.

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.

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

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

Dunn, Michael. 2015. Language phylogenies. In Claire Bowern & Bethwyn Evans (eds.), *The routledge handbook of historical linguistics* (Routledge handbooks in linguistics), 190–211. New York: Routledge.

Dunn, Michael, Niclas Burenhult, Nicole Kruspe, Sylvia Tufvesson & Neele Becker. 2011. Aslian linguistic prehistory: A case study in computational phylogenetics. *Diachronica* 28(3). 291–323. https://doi.org/10.1075/dia.28.3.01dun.

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, Stephen C. Levinson, Eva Lindström, Ger Reesink & Angela Terrill. 2008. Structural phylogeny in historical linguistics: Methodological explorations applied in island Melanesia. *Language* 84(4). 710–759. https://doi.org/10.1353/lan.0.0069.

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. 330 pp.

Eddington, David. 2004. *Spanish phonology and morphology: Experimental and quantitative perspectives*. Amsterdam: John Benjamins.

Eeckhout, Jan. 2004. Gibrat's law for (all) cities. *American Economic Review* 94(5). 1429–1451. https://doi.org/10.1257/0002828043052303.

Ernestus, Mirjam Theresia Constantia & 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.

Eska, Joseph F. & Donald A. Ringe. 2004. Recent work in computational linguistic phylogeny. *Language* 80(3). 569–582.

Estoup, J. B. 1916. *Gammes sténographiques: Méthode et exercices pour l'acquisition de la vitesse*. Paris: Institut Sténographique. 151 pp.

Evans, Nicholas. 1990. The Minkin language of the Bourketown region. In G. N. O'Grady & D. T. Tryon (eds.), *Studies in comparative Pama-Nyungan*, 173–207. Canberra: Pacific Linguistics.

Evans, Nicholas. 1995. *A grammar of Kayardild: With historical-comparative notes on Tangkic* (Mouton grammar library 15). Berlin: Mouton de Gruyter.

Evans, Nicholas. 1997. The cradle of the Pama-Nyungans: Archaeological and linguistic speculations. In Patrick McConvell & Nicholas Evans (eds.), *Archaeology and linguistics: Aboriginal Australia in global perspective*, 385–417. Melbourne: Oxford University Press.

Evans, Nicholas. 2003. *The Non-Pama-Nyungan languages of northern Australia: Comparative studies of the continent's most linguistically complex region* (Pacific Linguistics 552). Canberra: Pacific Linguistics.

Evans, Nicholas & Patrick McConvell. 1998. The enigma of Pama-Nyungan expansion in Australia. In Roger Blench & Matthew Spriggs (eds.), *Correlating archaeological and linguistic hypotheses* (Archaeology and Language II), 174–191. London: Routledge.

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

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

Everett, Caleb D. 2018b. The similar rates of occurrence of consonants across the world's languages: A quantitative analysis of phonetically transcribed word lists. *Language Sciences* 69. 125–135. https://doi.org/10.1016/j.langsci.2018.07.003.

Everett, Caleb D., Damián E. Blasi & Seán G. Roberts. 2015. Climate, vocal folds, and tonal languages: Connecting the physiological and geographic dots. *Proceedings of the National Academy of Sciences* 112(5). 1322–1327. https://doi.org/10.1073/pnas.1417413112.

Faloutsos, Michalis, Petros Faloutsos & Christos Faloutsos. 1999. On power-law relationships of the Internet topology. In *Applications, technologies, architectures, and protocols for computer communication (SIGCOMM)*, 251–262. New York: Association for Computing Machinery (ACM). https://doi.org/10.1145/316188.316229.

Farrell, Patrick J. & Katrina Rogers-Stewart. 2006. Comprehensive study of tests for normality and symmetry: Extending the Spiegelhalter test. *Journal of Statistical Computation and Simulation* 76(9). 803–816. https://doi.org/10.1080/10629360500109023.

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

Felsenstein, Joseph. 1988. Phylogenies and quantitative characters. *Annual Review of Ecology and Systematics* 19. 445–471.

Flege, James Emil. 1987. The production of "new" and "similar" phones in a foreign language: Evidence for the effect of equivalence classification. *Journal of phonetics* 15(1). 47–65.

Fletcher, Janet & Andrew Butcher. 2014. Sound patterns of Australian languages. In Harold Koch & Rachel Nordlinger (eds.), *The languages and linguistics of Australia: A comprehensive guide* (The World of Linguistics 3), 91–138. Boston: De Gruyter Mouton.

Florin, S. Anna, Andrew S. Fairbairn, May Nango, Djaykuk Djandjomerr, Ben Marwick, Richard Fullagar, Mike Smith, Lynley A. Wallis & Chris Clarkson. 2020. The first Australian plant foods at Madjedbebe, 65,000–53,000 years ago. *Nature Communications* 11(1). 924. https://doi.org/10.1038/s41467-020-14723-0.

Foley, William A. 1986. *The Papuan languages of New Guinea*. Cambridge: Cambridge University Press.

Frank, Steven A. 2014. How to read probability distributions as statements about process. *Entropy* 16(11). 6059–6098. https://doi.org/10.3390/e16116059.

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.

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.

Gabaix, Xavier. 1999. Zipf's law for cities: An explanation. *The Quarterly Journal of Economics* 114(3). 739–767. https://doi.org/10.1162/003355399556133.

Gaby, Alice. 2006. *A grammar of kuuk thaayorre*. Melbourne, Australia: University of Melbourne dissertation.

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.

Garland, Theodore, Ramón Díaz-Uriarte & M. Westneat. 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.

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

Gasser, Emily & Claire Bowern. 2014. Revisiting phonotactic generalizations in Australian languages. In *Annual meeting on phonology (AMP-2013)*. University of Massachusetts, Amherst: Linguistic Society of America. https://doi.org/10.3765/amp.v1i1.17.

Gillespie, Colin S. 2014. Fitting heavy tailed distributions: The poweRlaw package. *arXiv e-prints*, arXiv:1407.3492.

Gittleman, John L. 1981. The phylogeny of parental care in fishes. *Animal Behaviour* 29(3). 936–941. https://doi.org/10.1016/S0003-3472(81)80031-0.

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

Goddard, Cliff & Cliff Goddard. 1985. *A grammar of yankunytjatjara*. Alice Springs, N.T.: Australian National University Thesis (Ph. D.)

Goldsmith, John A. 1990. *Autosegmental and metrical phonology*. Oxford: B. Blackwell.

Good, I. J. 1969. Statistics of language. In A. R. Meetham & Richard A. Hudson (eds.), *Encyclopaedia of linguistics, information, and control*, 1st edn., 567–580. Oxford: Pergamon Press.

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*. New York: Oxford University Press. 378 pp.

Gould, Stephen Jay. 1987. *An urchin in the storm: Essays about books and ideas*. 1st. New York: W.W. Norton.

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

Gray, R. D., A. J. Drummond & S. J. Greenhill. 2009. Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* 323(5913). 479–483. https://doi.org/10.1126/science.1166858.

Gray, Russell D. & Quentin D. Atkinson. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* 426(6965). 435–439. https://doi.org/10.1038/nature02029.

Gray, Russell D. & Fiona M. Jordan. 2000. Language trees support the express-train sequence of Austronesian expansion. *Nature* 405(6790). 1052–1055. https://doi.org/10.1038/35016575.

Greenhill, Simon J. 2015. TransNewGuinea.org: An online database of New Guinea languages. *PloS One* 10(10). e0141563. https://doi.org/10.1371/journal.pone.0141563.

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*. https://doi.org/10.1073/pnas.1700388114.

Grey, Sir George. 1841. *Journals of two expeditions of discovery in north-west and western Australia: During the years 1837, 38, and 39, under the authority of Her Majesty's Government. describing many newly discovered, important, and fertile districts, with observations on the moral and physical condition of the Aboriginal inhabitants, &c., &c*. T. & W. Boone.

Gusein-Zade, S. M. 1988. On the distribution of letters of the Russian language by frequencies. *Problems of the Transmission of Information* 23. 103–107.

Hale, Kenneth. 1964. Classification of Northen Paman languages, Cape York Peninsula, Australia: A research report. *Oceanic Linguistics* 3(2). 248–265. https://doi.org/10.2307/3622881.

Hamilton, Philip J. 1996. *Phonetic constraints and markedness in the phonotactics of Australian Aboriginal languages*. Toronto: University of Toronto thesis.

Hammarström, Harald, Robert Forkel, Martin Haspelmath & Sebastian Bank. 2020. *Glottolog 4.2.1*. Jena: Max Planck Institute for the Science of Human History. https://doi.org/10.5281/zenodo.3754591. https://glottolog.org/%20accessed%202020-09-17.

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

Harvey, Mark. 2001. *A grammar of Limilngan: A language of the Mary River region, Northern Territory, Australia* (Pacific Linguistics 516). Canberra: Pacific Linguistics. 209 pp. https://doi.org/10.15144/PL-516.

Harvey, Mark & Robert Mailhammer. 2017. Reconstructing remote relationships. *Diachronica* 34(4). 470–515. https://doi.org/10.1075/dia.15032.har.

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

Haspelmath, Martin. 2020. Human linguisticality and the building blocks of languages. *Frontiers in Psychology* 10. https://doi.org/10.3389/fpsyg.2019.03056.

Hassler, Gabriel, Max R. Tolkoff, William L. Allen, Lam Si Tung Ho, Philippe Lemey & Marc A. Suchard. 2020. Inferring phenotypic trait evolution on large trees with many incomplete measurements. *Journal of the American Statistical Association*. 1–15. https://doi.org/10.1080/01621459.2020.1799812.

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.

Heath, Jeffrey. 1978a. *Linguistic diffusion in Arnhem Land*. Australian Institute of Aboriginal Studies.

Heggarty, Paul. 2006. Interdisciplinary indiscipline? Can phylogenetic methods meaningfully be applied to language data—and to dating language. In Peter Forster & Colin Renfrew (eds.), *Phylogenetic methods and the prehistory of languages* (McDonald Institute monographs), 183–194. Cambridge: McDonald Institute for Archaeological Research.

Hockett, Charles F. 1963. The problem of universals in language. In Joseph Greenberg (ed.), *Universals of language*, 1–29. Cambridge, MA: MIT Press.

Hoenigswald, Henry M. 1965. Language change and linguistic reconstruction.

Holbrook, Andrew J., Philippe Lemey, Guy Baele, Simon Dellicour, Dirk Brockmann, Andrew Rambaut & Marc A. Suchard. 2020. Massive parallelization boosts big Bayesian multidimensional

scaling. *Journal of Computational and Graphical Statistics*. 1–14. https://doi.org/10.1080/10618600.2020.1754226.

Holden, Clare J & Russell D Gray. 2006. Rapid radiation, borrowing and dialect continua in the bantu languages. In Peter Forster & Colin Renfrew (eds.), *Phylogenetic methods and the prehistory of languages*, 9–31. Cambridge: McDonald Institute for Archaeological Research.

Holden, Clare Janaki. 2002. Bantu language trees reflect the spread of farming across sub-Saharan Africa: A maximum-parsimony analysis. *Proceedings of the Royal Society of London. Series B: Biological Sciences* 269(1493). 793–799. https://doi.org/10.1098/rspb.2002.1955.

Hollis, Jordan, Genevieve C. Richards, Jayden L. Macklin-Cordes & Erich R. Round. 2016. The Cape York lexical records of Bruce Sommer. In *Australian Linguistics Society (ALS) annual conference*. Monash University, Caulfield, Australia. https://doi.org/10.6084/m9.figshare.4299377.

Holm, Hans J. 2007. The new arboretum of Indo-European "trees": Can new algorithms reveal the phylogeny and even prehistory of indo-european? *Journal of Quantitative Linguistics* 14(2). 167–214. https://doi.org/10.1080/09296170701378916.

Hruschka, Daniel J., Simon Branford, Eric D. Smith, Jon Wilkins, Andrew Meade, Mark Pagel & Tanmoy Bhattacharya. 2015. Detecting regular sound changes in linguistics as events of concerted evolution. *Current Biology* 25(1). 1–9. https://doi.org/10.1016/j.cub.2014.10.064.

Hudjashov, Georgi, Toomas Kivisild, Peter A. Underhill, Phillip Endicott, Juan J. Sanchez, Alice A. Lin, Peidong Shen, Peter Oefner, Colin Renfrew, Richard Villems & Peter Forster. 2007. Revealing the prehistoric settlement of Australia by Y chromosome and mtDNA analysis. *Proceedings of the National Academy of Sciences* 104(21). 8726–8730. https://doi.org/10.1073/pnas.0702928104.

Hudson, Joyce & Eirlys Richards. 1993. Walmajarri dictionary. Australian Institute of Aboriginal and Torres Strait Islander Studies, Australian Indigenous Languages Collection. ASEDA 0167. Canberra.

Hunley, Keith, Michael Dunn, Eva Lindström, Ger Reesink, Angela Terrill, Meghan E. Healy, George Koki, Françoise R. Friedlaender & Jonathan S. Friedlaender. 2008. Genetic and linguistic coevolution in Northern Island Melanesia. *PLOS Genetics* 4(10). e1000239. https://doi.org/10.1371/journal.pgen.1000239.

Hutchinson, Matthew C., Marília P. Gaiarsa & Daniel B. Stouffer. 2018. Contemporary ecological interactions improve models of past trait evolution. *Systematic Biology* 0(0). 1–13. https://doi.org/10.1093/sysbio/syy012.

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

Hyman, Larry M. 2008. Universals in phonology. *The Linguistic Review* 25(1). 83–137. https://doi.org/10.1515/TLIR.2008.003.

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

Itō, Junko. 1988. *Syllable theory in prosodic phonology* (Outstanding dissertations in linguistics). New York: Garland.

Ives, Anthony R., Peter E. Midford, Theodore Garland & Todd Oakley. 2007. Within-species variation and measurement error in phylogenetic comparative methods. *Systematic Biology* 56(2). 252–270. https://doi.org/10.1080/10635150701313830.

Jackson, Joshua Conrad, Joseph Watts, Teague R. Henry, Johann-Mattis List, Robert Forkel, Peter J. Mucha, Simon J. Greenhill, Russell D. Gray & Kristen A. Lindquist. 2019. Emotion semantics show both cultural variation and universal structure. *Science* 366(6472). 1517–1522. https://doi.org/10.1126/science.aaw8160.

Jäger, Gerhard. 2019. Computational historical linguistics. *Theoretical Linguistics* 45(3). 151–182. https://doi.org/10.1515/tl-2019-0011.

Jäger, Gerhard & Søren Wichmann. 2016. Inferring the world tree of languages from word lists. In Seán G. Roberts, Christine Cuskley, Luke McCrohon, Lluís Barceló-Coblijn, Feher Olga & Tessa Verhoef (eds.), *The evolution of language: Proceedings of the 11th international conference on the evolution of language (EvoLang XI)*, 8. New Orleans. http://www.sfs.uni-tuebingen.de/~gjaeger/publications/evolangWorldTreeProceedings.pdf.

Johnson, Keith, Peter Ladefoged & Mona Lindau. 1993. Individual differences in vowel production. *The Journal of the Acoustical Society of America* 94(2). 701–714.

Kang, Yoonjung. 2003. Perceptual similarity in loanword adaptation: English postvocalic word-final stops in Korean. *Phonology*. 219–273. https://doi.org/10.1017/S0952675703004524.

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.

Kass, Robert E. & Adrian E. Raftery. 1995. Bayes factors. *Journal of the American Statistical Association* 90(430). 773–795. https://doi.org/10.1080/01621459.1995.10476572.

Kealy, Shimona, Julien Louys & Sue O'Connor. 2018. Least-cost pathway models indicate northern human dispersal from Sunda to Sahul. *Journal of Human Evolution* 125. 59–70. https://doi.org/10.1016/j.jhevol.2018.10.003.

Kembel, Steven W., Peter D. Cowan, Mattew 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.

Kenny, James Andrew. 1975. *A numerical taxonomy of ethnic units using Murdock's 1967 world sample.* Bloomington: Indiana University Ph.D. thesis. https://elibrary.ru/item.asp?id=7109468.

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.

Klokeid, Terry J. 1969. *Thargari phonology and morphology* (Pacific Linguistics Series B 12). Canberra: Australian National University.

Koch, Harold. 2004. A methodological history of Australian linguistic classification. In Claire Bowern & Harold Koch (eds.), *Australian languages: Classification and the comparative method* (Current Issues in Linguistic Theory 249), 17–59. Amsterdam: John Benjamins Publishing Company.

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.

Koch, Harold & Rachel Nordlinger. 2014. The languages of Australia in linguistic research: Context and issues. In Harold Koch & Rachel Nordlinger (eds.), *The languages and linguistics of Australia: A comprehensive guide* (The world of linguistics 3), 3–21. Berlin: Mouton de Gruyter.

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.

Kroeber, Alfred Louis. 1923. Relationships of the Australian languages. *Journal and Proceedings of the Royal Society of New South Wales* 57. 101–117.

Kuba, Markus & Alois Panholzer. 2012. Limiting distributions for a class of diminishing urn models. *Advances in Applied Probability* 44(1). 87–116.

Kucera, Henry, W. Nelson Francis, John B. Carroll & W. F. Twaddell. 1967. *Computational analysis of present day American English*. 1st edn. Providence, RI: Brown University Press. 424 pp.

Kullback, S. & R. A. Leibler. 1951. On information and sufficiency. *The Annals of Mathematical Statistics* 22(1). 79–86.

Labov, William. 2020. The regularity of regular sound change. *Language* 96(1). 42–59. https://doi.org/10.1353/lan.2020.0001.

Landis, Michael J., Joshua G. Schraiber & Mason Liang. 2012. Phylogenetic analysis using Lévy processes: Finding jumps in the evolution of continuous traits. *Systematic biology* 62(2). 193–204.

Landis, Michael J & Joshua G Schraiber. 2017. Pulsed evolution shaped modern vertebrate body sizes. *Proceedings of the National Academy of Sciences* 114(50). 13224–13229.

Lass, Roger. 1984. Vowel system universals and typology: Prologue to theory. *Phonology Yearbook* 1. 75–111. https://doi.org/10.1017/S0952675700000300.

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

Leaché, Adam D., Barbara L. Banbury, Joseph Felsenstein, Adrián Nieto-Montes de Oca & Alexandros Stamatakis. 2015. Short tree, long tree, right tree, wrong tree: New acquisition bias corrections for

inferring SNP phylogenies. *Systematic Biology* 64(6). 1032–1047. https://doi.org/10.1093/sysbio/syv053.

Lee, Jennifer R. 1987. *Tiwi today: A study of language change in a contact situation* (Pacific linguistics Series C 96). Canberra: Australian National University.

Lee, Michael S. Y. & Alessandro Palci. 2015. Morphological phylogenetics in the genomic age. *Current Biology* 25(19). R922–R929. https://doi.org/10.1016/j.cub.2015.07.009.

Lee, Yuan & Insin Kim. 2018. Change and stability in shopping tourist destination networks: The case of Seoul in Korea. *Journal of Destination Marketing & Management* 9. 267–278. https://doi.org/10.1016/j.jdmm.2018.02.004.

Leff, Jonathan W., Richard D. Bardgett, Anna Wilkinson, Benjamin G. Jackson, William J. Pritchard, Jonathan R. Long, Simon Oakley, Kelly E. Mason, Nicholas J. Ostle, David Johnson, Elizabeth M. Baggs & Noah Fierer. 2018. Predicting the structure of soil communities from plant community taxonomy, phylogeny, and traits. *The ISME Journal*. 1. https://doi.org/10.1038/s41396-018-0089-x.

Lev-Ari, Shiri, Marcela San Giacomo & Sharon Peperkamp. 2014. The effect of domain prestige and interlocutors' bilingualism on loanword adaptations. *Journal of Sociolinguistics* 18(5). 658–684.

Levy, Moshe. 2009. Gibrat's law for (all) cities: Comment. *American Economic Review* 99(4). 1672–1675. https://doi.org/10.1257/aer.99.4.1672.

Li, W. 1992. Random texts exhibit Zipf's-law-like word frequency distribution. *IEEE Transactions on Information Theory* 38(6). 1842–1845. https://doi.org/10.1109/18.165464.

Liljencrants, Johan & Björn Lindblom. 1972. Numerical simulation of vowel quality systems: The role of perceptual contrast. *Language* 48. 839–862.

List, Johann-Mattis. 2018. Automatic inference of sound correspondence patterns across multiple languages. *Computational Linguistics* 45(1). 137–161. https://doi.org/10.1162/coli_a_00344.

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/lzy006.

Loeb, Daniel E & Gian-Carlo Rota. 1989. Formal power series of logarithmic type. *Advances in Mathematics* 75(1). 1–118.

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.

Lyell, Sir Charles. 1863. *The geological evidences of the antiquity of man: With remarks on theories of the origin of species by variation*. John Murray.

Macklin-Cordes, Jayden L. 2017. High-definition phonotactic typology in Sahul: Choosing the data-rich approach. In *Association of linguistic typology (ALT-12)*. Australian National University, Canberra, Australia. https://doi.org/10.6084/m9.figshare.5861067.

Macklin-Cordes, Jayden L., Nathaniel L. Blackbourne, Thomas J. Bott, Jacqueline Cook, T. Mark Ellison, Jordan Hollis, Edith E. Kirlew, Genevieve C. Richards, Sanle Zhao & Erich R. Round. 2017. Robots who read grammars [poster]. In *CoEDL fest*. Alexandra Park Conference Centre, Alexandra Headlands, QLD, Australia. https://doi.org/10.6084/m9.figshare.4625248.

Macklin-Cordes, Jayden L., Claire Bowern & Erich R. Round. 2021. Phylogenetic signal in phonotactics. *Diachronica* 38. https://doi.org/10.1075/dia.20004.mac.

Macklin-Cordes, Jayden L., Steven P. Moran & Erich R. Round. 2016. Evaluating information loss from phonological dimensionality reduction. In *Speech science and technology (SST) satellite symposium: The role of predictability in shaping human language sound patterns*. Western Sydney University, Australia. https://doi.org/10.6084/m9.figshare.4465976.

Macklin-Cordes, Jayden L. & Erich R. Round. Submitted. Challenges of sampling and how phylogenetic comparative methods help: With a case study of the Pama-Nyungan laminal contrast. Preprint. https://doi.org/10.5281/zenodo.4796981.

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/10.15496/publikation-8609.

Macklin-Cordes, Jayden L. & Erich R. Round. 2016a. High-definition phonotactic data contain phylogenetic signal [poster]. In *90th annual meeting of the Linguistic Society of America (LSA)*. Washington, DC. https://doi.org/10.6084/m9.figshare.4534238.

Macklin-Cordes, Jayden L. & Erich R. Round. 2016b. High-definition phonotactics reflect linguistic pasts. Yale Linguistics Friday Lunch Talk. Yale University, New Haven, CT.

Macklin-Cordes, Jayden L. & Erich R. Round. 2016c. Reflections of linguistic history in quantitative phonotactics. In *Australian Linguistics Society (ALS) annual conference*. Monash University, Caulfield, Australia. https://doi.org/10.6084/m9.figshare.4299365.

Macklin-Cordes, Jayden L. & Erich R. Round. 2017. Fine-grained phonotactics in Sahul: Data-rich comparison for deep-time analysis [poster]. LSA Linguistic Institute Poster Session. University of Kentucky, Lexington.

Macklin-Cordes, Jayden L. & Erich R. Round. 2018. Phylogeny in phonotactics: A multivariate approach for stronger historical signal. In *Australian Linguistics Society (ALS) annual conference*. University of South Australia, Adelaide.

Macklin-Cordes, Jayden L. & Erich R. Round. 2019. Historical signal beyond the cognate: Phonotactics in linguistic phylogenetics. In *International conference on historical linguistics (ICHL24)*. Australian National University, Canberra, Australia. https://cloudstor.aarnet.edu.au/plus/s/zN5AbJ4fXoVR0ax.

Macklin-Cordes, Jayden L. & Erich R. Round. 2020a. Re-evaluating phoneme frequencies. *Frontiers in Psychology* 11. https://doi.org/10.3389/fpsyg.2020.570895.

Macklin-Cordes, Jayden L. & Erich R. Round. 2020b. Zipf's law? Re-evaluating phoneme frequencies. In *Australian linguistics society (ALS) annual conference*. [online conference].

Maddieson, Ian. 1984. *Patterns of sounds* (Cambridge studies in speech science and communication). Cambridge: Cambridge University Press.

Malaspinas, Anna-Sapfo, Michael Westaway, Craig Muller, Vitor Sousa, Oscar Lao, Isabel Alves, Anders Bergström, Georgios Athanasiadis, Jade Cheng, Jacob Crawford, Tim Heupink, Enrico Macholdt, Stephan Peischl, Simon Rasmussen, Stephan Schiffels, Sankar Subramanian, Joanne Wright, Anders Albrechtsen, Chiara Barbieri, Isabelle Dupanloup, Anders Eriksson, Ashot Margaryan, Ida Moltke, Irina Pugach, Thorfinn Korneliussen, Ivan Levkivskyi, J. Moreno-Mayar, Shengyu Ni, Fernando Racimo, Martin Sikora, Yali Xue, Farhang Aghakhanian, Nicolas Brucato, Søren Brunak, Paula Campos, Warren Clark, Sturla Ellingvåg, Gudjugudju Fourmile, Pascale Gerbault, Darren Injie, George Koki, Matthew Leavesley, Betty Logan, Aubrey Lynch, Elizabeth Matisoo-Smith, Peter Mcallister, Alexander Mentzer, Mait Metspalu, Andrea Migliano, Les Murgha, Maude Phipps, William Pomat, Doc Reynolds, Francois-Xavier Ricaut, Peter Siba, Mark Thomas, Thomas Wales, Colleen Wall, Stephen Oppenheimer, Chris Tyler-Smith, Richard Durbin, Joe Dortch, Andrea Manica, Mikkel Schierup, Robert Foley, Marta Lahr, Claire Bowern, Jeffrey Wall, Thomas Mailund, Mark Stoneking, Rasmus Nielsen, Manjinder Sandhu, Laurent Excoffier, David Lambert & Eske Willerslev. 2016. A genomic history of Aboriginal Australia. *Nature* 538(7624). 207–214L. https://doi.org/10.1038/nature18299.

Malevergne, Yannick, Vladilen Pisarenko & Didier Sornette. 2011. Testing the Pareto against the lognormal distributions with the uniformly most powerful unbiased test applied to the distribution of cities. *Physical Review E* 83(3). 036111. https://doi.org/10.1103/PhysRevE.83.036111.

Mandelbrot, Benoit. 1954. Structure formelle des textes et communication. *WORD* 10(1). 1–27. https://doi.org/10.1080/00437956.1954.11659509.

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.

Martindale, Colin, S. M. Gusein-Zade, Dean McKenzie & Mark Yu Borodovsky. 1996. Comparison of equations describing the ranked frequency distributions of graphemes and phonemes. *Journal of Quantitative Linguistics* 3(2). 106–112. https://doi.org/10.1080/09296179608599620.

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.

McEntee, John & Pearl McKenzie. 1992. *Adna-mat-na English dictionary*. Adelaide: the authors. 125 pp.

McMahon, April & Robert McMahon. 2003. Finding families: Quantitative methods in language classification. *Transactions of the Philological Society* 101(1). 7–55.

Meillet, Antoine. 1925. *La méthode comparative en linguistique historique*. Paris: Honoré Champion.

Memmott, Paul, Erich R. Round, Daniel Rosendahl & Sean Ulm. 2016. Fission, fusion and syncretism: Linguistic and environmental changes amongst the Tangkic people of the southern Gulf of Carpentaria, northern Australia. In Jean-Christophe Verstraete & Diane Hafner (eds.), *Land and Language in Cape York Peninsula and the Gulf Country. Culture and Language Use*, 105–136. Amsterdam: John Benjamins. https://doi.org/10.1075/clu.18.06mem.

Mesoudi, Alex. 2017. Pursuing Darwin's curious parallel: Prospects for a science of cultural evolution. *Proceedings of the National Academy of Sciences* 114(30). 7853–7860. https://doi.org/10.1073/pnas.1620741114.

Miceli, Luisa & Erich R. Round. N.d. Where have all the sound changes gone? on the scarcity of regular sound change in Australia [manuscript]. unpublished manuscript.

Michael, Lev, Natalia Chousou-Polydouri, Keith Bartolomei, Erin Donnelly, Sérgio Meira, Vivian Wauters & Zachary O'Hagan. 2015. A Bayesian phylogenetic classification of Tupí-Guaraní. *LIAMES: Línguas Indígenas Americanas* 15(2). 193–221. https://doi.org/10.20396/liames.v15i2.8642301.

Mielke, Jeff. 2008. *The emergence of distinctive features* (Oxford linguistics). Oxford: Oxford University Press.

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

Miller, John E., Tiago Tresoldi, Roberto Zariquiey, César A. Beltrán Castañón, Natalia Morozova & Johann-Mattis List. 2020. Using lexical language models to detect borrowings in monolingual wordlists. *PLOS ONE* 15(12). e0242709. https://doi.org/10.1371/journal.pone.0242709.

Mitzenmacher, Michael. 2004. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics* 1(2). 226–251. https://doi.org/10.1080/15427951.2004.10129088.

Montemurro, Marcelo A. 2001. Beyond the Zipf–Mandelbrot law in quantitative linguistics. *Physica A: Statistical Mechanics and its Applications* 300(3). 567–578. https://doi.org/10.1016/S0378-4371(01)00355-7.

Moorhouse, Matthew. 1846. *A vocabulary, and outline of the grammatical structure of the Murray River language: Spoken by the natives of South Australia, from Wellington on the Murray, as far as the Rufus*. Andrew Murray. http://hdl.handle.net/2440/24872.

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

Moran, Steven. 2012. *Phonetics information base and lexicon*. Seattle: University of Washington Ph.D. Dissertation.

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. 2018a. 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)*. Torun, Poland: NCU Press. https://doi.org/10.12775/3991-1.077. http://evolang.org/torun/proceedings/papertemplate.html?p=98.

Moran, Steven & Annemarie Verkerk. 2018b. 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.

Mühlhäusler, Peter, Darrell T Tryon & Stephen A. Wurm. 1996. *Atlas of languages of intercultural communication in the Pacific, Asia, and the Americas: vol i: maps. vol II: texts*. Berlin: De Gruyter.

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

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

Nagle, Nano, Mannis van Oven, Stephen Wilcox, Sheila van Holst Pellekaan, Chris Tyler-Smith, Yali Xue, Kaye N. Ballantyne, Leah Wilcox, Luka Papac, Karen Cooke, Roland A. H. van Oorschot, Peter McAllister, Lesley Williams, Manfred Kayser & R. John Mitchell. 2017. Aboriginal australian mitochondrial genome variation – an increased understanding of population antiquity and diversity. *Scientific Reports* 7(1). 43041. https://doi.org/10.1038/srep43041.

Naranan, S. & V. K. Balasubrahmanyan. 1992. Information theoretic models in statistical linguistics—part i: A model for word frequencies. *Current Science* 63(5). 261–269.

Naranan, S. & V. K. Balasubrahmanyan. 1998. Models for power law relations in linguistics and information science. *Journal of Quantitative Linguistics* 5(1). 35–61. https://doi.org/10.1080/09296179808590110.

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

Nash, David, Patrick McConvell, Arthur Capell, Kenneth Hale, Peter Sutton, Deborah Bird Rose & Jim Wafer. 1988. Mudburra wordlist. Australian Institute of Aboriginal and Torres Strait Islander Studies, Australian Indigenous Languages Collection. ASEDA 0031. Canberra. http://aiatsis.gov.au/sites/default/files/catalogue_resources/0031_access.zip.

Newman, M. E. J. 2005. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics* 46(5). 323–351. https://doi.org/10.1080/00107510500052444.

Nicholls, Geoff K. & Russell D. Gray. 2008. Dated ancestral trees from binary trait data and their application to the diversification of languages. *Journal of the Royal Statistical Society: Series B*

(Statistical Methodology) 70(3). 545–566. https://doi.org/10.1111/j.1467-9868.2007.00648.x.

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

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.

Nölle, Jonas, Stefan Hartmann & Peeter Tinits. 2020. Language evolution research in the year 2020: A survey of new directions. *Language Dynamics and Change* 10(1). 3–26. https://doi.org/10.1163/22105832-bja10005.

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

O'Connell, J. F. & J. Allen. 2015. The process, biotic impact, and global implications of the human colonization of Sahul about 47,000 years ago. *Journal of Archaeological Science*. Scoping the Future of Archaeological Science: Papers in Honour of Richard Klein 56. 73–84. https://doi.org/10.1016/j.jas.2015.02.020.

O'Grady, Geoffrey N. 1979. Preliminaries to a Proto Nuclear Pama-Nyungan stem list. In Stephen A. Wurm (ed.), *Australian linguistic studies*, 107–139. Canberra, Australia: Pacific Linguistics.

O'Grady, Geoffrey N. & Kenneth L. Hale. 2004. The coherence and distinctiveness of the Pama-Nyungan language family within the Australian linguistic phylum. In Claire Bowern & Harold Koch (eds.), *Australian languages: Classification and the comparative method* (Current Issues in Linguistic Theory 249), 69–92. Amsterdam/Philadelphia: John Benjamins.

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.

Osthoff, Hermann & Karl Brugmann. 1878. *Morphologische Untersuchungen auf dem Gebiete der indogermanischen Sprachen*. Vol. 1. Leipzig: Hirzel.

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

Pagel, Mark. 2017. Q&a: What is human language, when did it evolve and why should we care? *BMC Biology* 15. https://doi.org/10.1186/s12915-017-0405-3.

Paradis, Carole & Darlene LaCharité. 1997. Preservation and minimality in loanword adaptation. *Journal of linguistics*. 379–430. https://doi.org/10.1017/S0022226797006786.

Pareto, Vilfredo. 1897. *Cours d'economie politique*. Vol. 2. Lausanne, France: F. Rouge.

Parins-Fukuchi, Caroline. 2018. Use of continuous traits can improve morphological phylogenetics. *Systematic Biology* 67(2). 328–339. https://doi.org/10.1093/sysbio/syx072.

Pedro, Nicole, Nicolas Brucato, Veronica Fernandes, Mathilde André, Lauri Saag, William Pomat, Céline Besse, Anne Boland, Jean-François Deleuze, Chris Clarkson, Herawati Sudoyo, Mait Metspalu, Mark Stoneking, Murray P. Cox, Matthew Leavesley, Luisa Pereira & François-Xavier Ricaut. 2020. Papuan mitochondrial genomes and the settlement of Sahul. *Journal of Human Genetics* 65(10). 875–887. https://doi.org/10.1038/s10038-020-0781-3.

Perkins, Revere D. 1980. *The Evolution of Culture and Grammar*. Buffalo, New York: State University of New York Ph.D. thesis.

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

Perkins, Revere D. 1989. Statistical techniques for determining language sample size. *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language"* 13(2). 293–315.

Piantadosi, Steven T. 2014. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review* 21(5). 1112–1130.

Proctor, Michael Ian. 2009. *Gestural characterization of a phonological class: The liquids*. New Haven: Yale University Ph.D. Dissertation.

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

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

de Queiroz, Kevin. 1996. Including the characters of interest during tree reconstruction and the problems of circularity and bias in studies of character evolution. *The American Naturalist* 148(4). 700–708.

R Core Team. 2017a. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/.

R Core Team. 2017b. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. https://www.R-project.org/.

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.

Rambaut, Andrew, Alexei J. Drummond, Dong Xie, Guy Baele & Marc A. Suchard. 2018. Posterior summarization in Bayesian phylogenetics using tracer 1.7. *Systematic Biology* 67(5). 901–904. https://doi.org/10.1093/sysbio/syy032.

Reesink, Ger & Michael Dunn. 2012. Systematic typological comparison as a tool for investigating language history. In Nicholas Evans & Marian Klamer (eds.), *Melanesian languages on the*

edge of asia: challenges for the 21st century (Language Documentation & Conservation Special Publication 5), 34–71. Honolulu: University of Hawai'i Press.

Reesink, Ger, Ruth Singer & Michael Dunn. 2009. Explaining the linguistic diversity of Sahul using population models. *PLOS Biol* 7(11). e1000241. https://doi.org/10.1371/journal.pbio.1000241.

Reeves, Jessica M., Allan R. Chivas, Adriana García, Sabine Holt, Martine J. J. Couapel, Brian G. Jones, Dioni I. Cendón & David Fink. 2008. The sedimentary record of palaeoenvironments and sea-level change in the Gulf of Carpentaria, Australia, through the last glacial cycle. *Quaternary International*. Subaerially exposed continental shelves: Contributions from INQUA Project 0419 183(1). 3–22. https://doi.org/10.1016/j.quaint.2007.11.019.

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.

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

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

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

Round, Erich R. 2017a. Continent-wide, full-lexicon typology overturns our view of the Australian laminal contrast. In *Association of linguistic typology (ALT-12)*. Australian National University, Canberra, Australia.

Round, Erich R. 2017b. Matthew K. Gordon: Phonological typology [book review]. *Folia Linguistica* 51(3). 745–755. https://doi.org/10.1515/flin-2017-0027.

Round, Erich R. 2017c. The AusPhon-Lexicon project: 2 million normalized segments across 300 Australian languages. In *47th poznań linguistic meeting*. Poznań, Poland. http://wa.amu.edu.pl/plm_old/2017/files/abstracts/PLM2017_Abstract_Round.pdf.

Round, Erich R. 2019a. *Australian phonemic inventories contributed to PHOIBLE 2.0: Essential explanatory notes*. https://doi.org/10.5281/zenodo.3464333.

Round, Erich R. 2019b. 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. https://doi.org/10.5281/zenodo.3464333. https://phoible.org/contributors/ER.

Round, Erich R. 2021a. Phonotactics in Australian languages. In Claire Bowern (ed.), *Oxford handbook of Australian languages*. Oxford: Oxford University Press.

Round, Erich R. 2021b. Segment inventories in Australian languages. In Claire Bowern (ed.), *Oxford handbook of Australian languages*. Oxford: Oxford University Press.

Round, Erich R., Jacqueline Cook, T. Mark Ellison, Jordan Hollis, Jayden L. Macklin-Cordes & Genevieve C. Richards. 2016. The grammar harvester. In *Australian Linguistics Society (ALS) annual conference*. Monash University, Caulfield, Australia.

Round, Erich R. & Greville G. Corbett. 2020. Comparability and measurement in typological science: The bright future for linguistics. *Linguistic Typology* 24(3). 489–525. https://doi.org/10.1515/lingty-2020-2060.

Round, Erich R., T. Mark Ellison, Jayden L. Macklin-Cordes & Sacha Beniamine. 2020. Automated parsing of interlinear glossed text from page images of grammatical descriptions. In *12th language resources and evaluation conference (LREC-2020)*, 2871–2876. Marseille, France: European Language Resources Association. https://www.aclweb.org/anthology/2020.lrec-1.350.

Round, Erich R. & Jayden L. Macklin-Cordes. 2015. On the design, in practice, of typological microvariables. In *New developments in the quantitative study of languages*. Helsinki, Finland.

Round, Erich R. & Jayden L. Macklin-Cordes. 2019. Clouded vision: Why insights improve when uncertainty about data is made explicit. In *International conference on historical linguistics (ICHL24)*. Australian National University, Canberra, Australia. https://cloudstor.aarnet.edu.au/plus/s/WOmUM73NVJQFqf6.

Ryder, Robin J. & Geoff K. Nicholls. 2011. Missing data in a stochastic Dollo model for binary trait data, and its application to the dating of Proto-Indo-European. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 60(1). 71–92. https://doi.org/10.1111/j.1467-9876.2010.00743.x.

Sagart, Laurent, Guillaume Jacques, Yunfan Lai, Robin J. Ryder, Valentin Thouzeau, Simon J. Greenhill & Johann-Mattis List. 2019. Dated language phylogenies shed light on the ancestry of Sino-Tibetan. *Proceedings of the National Academy of Sciences* 116(21). 10317–10322. https://doi.org/10.1073/pnas.1817972116.

Salesky, Elizabeth, Eleanor Chodroff, Tiago Pimentel, Matthew Wiesner, Ryan Cotterell, Alan W Black & Jason Eisner. 2020. A corpus for large-scale phonetic typology. In *Association for computational linguistics (ACL)*, 4526–4546. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.415.

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.

Savelyev, Alexander & Martine Robbeets. 2020. Bayesian phylolinguistics infers the internal structure and the time-depth of the Turkic language family. *Journal of Language Evolution*. https://doi.org/10.1093/jole/lzz010.

Schleicher, August. 1853. Die ersten Spaltungen des indogermanischen Urvolkes. *Allgemeine Monatsschrift für Wissenschaft und Literatur* 1853. 876–787.

Schleicher, August. 1863. *Die Darwinsche Theorie und die sprachwissenschaft: Offenes Sendschreiben an Ernst Häckel*. Weimar: H. Böhlau.

Schmidt, Wilhelm. 1919. *Die Gliederung australischen Sprachen: Geographische, bibliographische, linguistische Grundzüge der Erforschung der australischen Sprachen*. Vienna: Druck und Verlag der Mechitharisten-Buchdruckerei.

Segovia-Martín, José & Sergio Balari. 2020. Eco-evo-devo and iterated learning: Towards an integrated approach in the light of niche construction. *Biology & Philosophy* 35(4). 42. https://doi.org/10.1007/s10539-020-09761-3.

Sharpe, Margaret C. 1972. *Alawa phonology and grammar*. Vol. 37 (Australian Aboriginal Studies 37, Linguistic Series 15). Canberra: Australian Institute of Aboriginal Studies.

Sicoli, Mark A. & Gary Holton. 2014. Linguistic phylogenies support back-migration from Beringia to Asia. *PLOS ONE* 9(3). e91722. https://doi.org/10.1371/journal.pone.0091722.

Sigurd, Bengt. 1968. Rank-frequency distributions for phonemes. *Phonetica* 18(1). 1–15. https://doi.org/10.1159/000258595.

Silverman, Daniel. 1992. Multiple scansions in loanword phonology: Evidence from Cantonese. *Phonology* 9(2). 289–328.

Simon, Herbert A. 1955. On a class of skew distribution functions. *Biometrika* 42(3). 425–440. https://doi.org/10.1093/biomet/42.3-4.425.

Sommer, Bruce A. N.d.(b). Koko Narr. Fryer Library Bruce Sommer Collection. UQFL476_b10f03_64, UQFL476_b10f03_65. Brisbane.

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.

Spiegelhalter, D. J. 1980. An omnibus test for normality for small samples. *Biometrika* 67(2). 493–496. https://doi.org/10.1093/biomet/67.2.493.

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.

Stevens, Kenneth N. 1989. On the quantal nature of speech. *Journal of phonetics* 17(1-2). 3–45.

Stumpf, Michael P. H. & Mason A. Porter. 2012. Critical truths about power laws. *Science* 335(6069). 665–666. https://doi.org/10.1126/science.1216142.

Suchard, Marc A., Philippe Lemey, Guy Baele, Daniel L. Ayres, Alexei J. Drummond & Andrew Rambaut. 2018. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution* 4(1). https://doi.org/10.1093/ve/vey016.

Swadesh, Morris. 1952. Lexico-statistic dating of prehistoric ethnic contacts: With special reference to North American Indians and Eskimos. *Proceedings of the American philosophical society*. 452–463.

Swadesh, Morris. 1955. Towards greater accuracy in lexicostatistic dating. *International journal of American linguistics* 21(2). 121–137. https://doi.org/10.1086/464321.

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

Tambovtsev, Yuri & Colin Martindale. 2007. Phoneme frequencies follow a Yule distribution. *SKASE Journal of Theoretical Linguistics* 4(2). 1–11. http://www.skase.sk/Volumes/JTL09/pdf_doc/1.pdf.

Thomason, Sarah Grey & Terrence Kaufman. 1988. *Language contact, creolization, and genetic linguistics*. Berkeley: University of California Press.

Tobler, Ray, Adam Rohrlach, Julien Soubrier, Pere Bover, Bastien Llamas, Jonathan Tuke, Nigel Bean, Ali Abdullah-Highfold, Shane Agius, Amy O'Donoghue, Isabel O'Loughlin, Peter Sutton, Fran Zilio, Keryn Walshe, Alan N. Williams, Chris S. M. Turney, Matthew Williams, Stephen M. Richards, Robert J. Mitchell, Emma Kowal, John R. Stephen, Lesley Williams, Wolfgang Haak & Alan Cooper. 2017. Aboriginal mitogenomes reveal 50,000 years of regionalism in Australia. *Nature* 544(7649). 180–184. https://doi.org/10.1038/nature21416.

Tolkoff, Max R., Michael E. Alfaro, Guy Baele, Philippe Lemey & Marc A. Suchard. 2018. Phylogenetic factor analysis. *Systematic Biology* 67(3). 384–399. https://doi.org/10.1093/sysbio/syx066.

Touboul, Jonathan & Alain Destexhe. 2010. Can power-law scaling and neuronal avalanches arise from stochastic dynamics? *PLOS ONE* 5(2). e8982. https://doi.org/10.1371/journal.pone.0008982.

Tryon, Darrell T. 1974. *Daly family languages, Australia*. Department of Linguistics, Research School of Pacific Studies, The Australian National University.

Urzúa, Carlos M. 2011. Testing for Zipf's law: A common pitfall. *Economics Letters* 112(3). 254–255. https://doi.org/10.1016/j.econlet.2011.05.049.

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.

Van Egmond, Marie-Elaine. 2012. *Enindhilyakwa phonology, morphosyntax and genetic position*. University of Sydney dissertation.

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. 2015. Where do all the motion verbs come from?: The speed of development of manner verbs and path verbs in Indo-European. *Diachronica* 32(1). 69–104. https://doi.org/10.1075/dia.32.1.03ver.

Verkerk, Annemarie. 2017. Phylogenetic comparative methods for typologists (Focusing on families and regions: A plea for using phylogenetic comparative methods in linguistic typology). In *Quanti-*

tative Analysis in Typology: The logic of choice among methods (workshop at the 12th Conference of the Association for Linguistic Typology. Australian National University, Canberra, Australia.

Verstraete, Jean-Christophe. 2019. Mbarrumbathama (lamalama). *International Phonetic Association. Journal of the International Phonetic Association* 49(2). 265–288. https://doi.org/http://dx.doi.org/10.1017/S0025100318000105.

de Villemereuil, Pierre & 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, C. F. & F. M. Voegelin. 1966. Index of languages of the world. *Anthropological Linguistics* 8(6). 1–222.

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.

Vuong, Quang H. 1989. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* 57(2). 307–333. https://doi.org/10.2307/1912557.

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.

Wang, William S.-Y. 1977. *The lexicon in phonological change*. The Hague: Mouton.

Wangka Maya Pilbara Aboriginal Language Centre & Australian Institute of Aboriginal and Torres Strait Islander Studies. 2004. *Putijarra-English wordlist, English-Putijarra finder topical wordlist &sketch morphology*. South Hedland, WA, Australia: Wangka Maya Pilbara Aboriginal Language Centre. 118 pp.

Webb, Campbell O., David D. Ackerly, Mark A. McPeek & 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.

Whiteley, Peter M., Ming Xue & Ward C. Wheeler. 2019. Revising the Bantu tree. *Cladistics* 35(3). 329–348. https://doi.org/10.1111/cla.12353.

Whitworth, William Allen. 1901. *Choice and chance: With 1,000 exercises*. Cambridge: Deighton Bell & Company.

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.

Wilkins, David P. 1989. *Mparntwe Arrernte (Aranda): Studies in the structure and semantics of grammar*. Canberra: Australian National University dissertation. https://openresearch-repository.anu.edu.au/handle/1885/9908.

Williams, Alan N., Sean Ulm, Tom Sapienza, Stephen Lewis & Chris S. M. Turney. 2018. Sea-level change and demography during the last glacial termination and early Holocene across the Australian continent. *Quaternary Science Reviews* 182. 144–154. https://doi.org/10.1016/j.quascirev.2017.11.030.

Witten, Ian H. & Timothy C. Bell. 1990. Source models for natural language text. *International Journal of Man-Machine Studies* 32(5). 545–579. https://doi.org/10.1016/S0020-7373(05)80033-1.

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.

Wright, April M. 2019. A systematist's guide to estimating bayesian phylogenies from morphological data. *Insect Systematics and Diversity* 3(3). https://doi.org/10.1093/isd/ixz006.

Wu, Mei-Shin, Nathanael E. Schweikhard, Timotheus A. Bodt, Nathan W. Hill & Johann-Mattis List. 2020. Computer-assisted language comparison: State of the art. *Journal of Open Humanities Data* 6(2). 1–14. https://doi.org/10.5334/johd.12.

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.

Wurm, Stephen A. 1975. *Papuan languages and the New Guinea linguistic scene* (Pacific linguistics Series C 38). Canberra: Australian National University.

Yanovich, Igor. 2020. Phylogenetic linguistic evidence and the Dene-Yeniseian homeland. https://doi.org/10.1075/dia.17038.yan.

Yule, G. Udny. 1925. A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F. R. S. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 213(402). 21–87. https://doi.org/10.1098/rstb.1925.0002.

Z'Graggen, John A. 1980. *A comparative word list of the Northern Adelbert Range languages, Mandang Province, Papua New Guinea*. Canberra: Pacific Linguistics. https://clics.clld.org/contributions/zgraggenmadang.

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

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

Zipf, George K. 1932. *Selective studies and the principle of relative frequency in language*. Cambridge, MA: Harvard University Press.

Zipf, George K. 1949. *Human behavior and the principle of least effort*. Reading, MA: Addison-Wesley.

Zuraw, Kie R. 2000. *Patterned exceptions in phonology*. Los Angeles: University of California dissertation.

Appendix A

Supplementary figure for Chapter 5

Quantifying phylogenetic signal requires an independently-derived reference phylogeny as a yardstick. Our reference phylogeny is a 285-tip Pama-Nyungan phylogeny inferred by the second author. Figure S1 gives a 112-tip subset of this phylogeny, corresponding to the 112 doculects used in this study.

As discussed in the main paper, the reference phylogeny was constructed using Bayesian phylogenetic methods in the software BEAST2 (Bouckaert et al. 2014). Bayesian phylogenetic methods use a Markov Chain Monte Carlo (MCMC) procedure to efficiently search the hypothesis space of possible trees and return a large posterior sample of similarly credible alternatives (capturing phylogenetic uncertainty). The tree in Figure S1 above is a maximum clade credibility tree, which is a summation of the posterior sample where the likelihood of all nodes in the tree (in terms of how frequently a given node reappears across the posterior sample) is maximized.

For further details on the Pama-Nyungan phylogeny used as a reference phylogeny in this study, see Bowern (2015). See also Bowern & Atkinson (2012), which infers a Pama-Nyungan phylogeny in the same way, using exactly the same evolutionary model parameters, but with an earlier iteration of the dataset containing fewer doculects. Additional discussion of the general process of constructing language phylogenies in BEAST2 can be found in Bouckaert, Bowern & Atkinson (2018), although a different evolutionary model is used.

The reference phylogeny was inferred using lexical cognate data, coded according to the principles of the Comparative Method. The cognate data used in the reference phylogeny is publicly available on Zenodo (Bowern 2018b) and also as a subset of the 305-language dataset in Bouckaert, Bowern & Atkinson (2018). This latter source also includes a Perl script for converting multistate cognate judgements into a binary matrix for use with BEAST2 phylogenetic software and will include information on underlying sources.

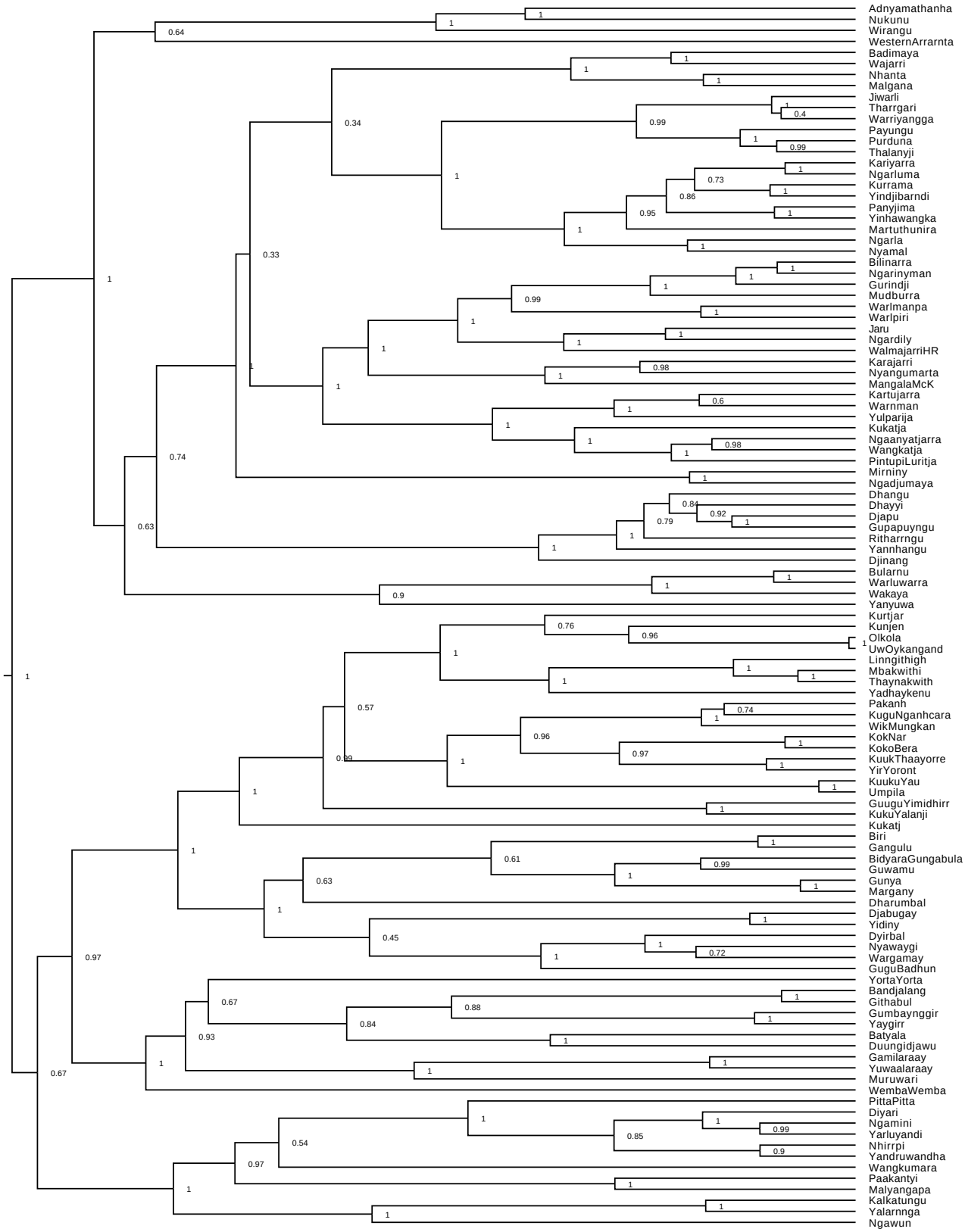

Figure A.1: Pama-Nyungan reference phylogeny. Node labels are posterior probabilities, giving an indication of support for each node. Although there is no strict, conventional cut-off, clades with posterior values above 0.5 are considered supported and values above 0.8 are considered strongly supported (Bowern & Atkinson 2012, p. 829).

Appendix B

Wordlist sources

The following list contains the details of original wordlist sources, CHIRILA source information where applicable, and details of any phonemic normalisation procedures used to maintain comparability between phonemicisations in Ausphon. The list covers all languages used throughout the thesis.

Adnyamathanha

CHIRILA source: CHIRILA/v2/McEnteeMcKenzie

John McEntee & Pearl McKenzie. 1992. *Adna-mat-na English dictionary*. Adelaide: the authors. 125 pp.

Phonemic normalisation: Coda tap normalized as vibrant. Otherwise, voiced stops, taps and fricatives normalized to lenis obstruents.

Alawa

Margaret C. Sharpe. 2001. *Alawa nanggaya nindanya yalanu junggulu = Alawa-Kriol-English dictionary*. Prospect, South Australia: Caitlin Press. 245 pp.

Amurdak

Robert Handelsmann. 1991. *Towards a description of Amurdak: A language of northern Australia*. Melbourne: University of Melbourne Honours Thesis

Phonemic normalisation: Prelateralized stops normalized as flapped laterals.

Angkamuthi

Terry Crowley. 1983. Uradhi. In R. M. W. Dixon & Barry J. Blake (eds.), *Handbook of Australian languages*, vol. 3, 5 vols., 307–428. Amsterdam: John Benjamins

Anguthimri

CHIRILA source: CHIRILA/v1/ASEDA0240

Terry Crowley. 1989. Mbakwithi vocabulary. Australian Institute of Aboriginal and Torres Strait Islander Studies, Australian Indigenous Languages Collection. ASEDA 0240. Canberra

Atampaya

Terry Crowley. 1983. Uradhi. In R. M. W. Dixon & Barry J. Blake (eds.), *Handbook of Australian languages*, vol. 3, 5 vols., 307–428. Amsterdam: John Benjamins

Badimaya

Doug Marmion. 1995. Badimaya dictionary. Australian Institute of Aboriginal and Torres Strait Islander Studies, Australian Indigenous Languages Collection. ASEDA 0615. Canberra

Phonemic normalisation: Double a normalized to long vowel.

Bakanh

Philip J. Hamilton. 1997a. *Pakanh Alphabetical Search Index*. Oykangand and Olkola Dictionary. `http://www.oocities.org/athens/delphi/2970/pakalpha.htm`

Bardi

CHIRILA source: CHIRILA/v1/akl99

Gedda Aklif & Kimberley Language Resource Centre. 1999. *Ardiyooloon Bardi Ngaanka: One Arm Point Bardi dictionary*. Halls Creek, WA, Australia: Kimberley Language Resource Centre. 222 pp.

Bidyara

Gavan Breen. 1973. *Bidyara and Gungabula grammar and vocabulary*. Vol. 8 (Linguistic Communications). Melbourne: Monash University. 227 pp.

Bilinarra

Felicity Meakins, Lauren Campbell, et al. 2013. *Bilinarra to English dictionary*. Batchelor, NT, Australia: Batchelor Press. 264 pp.

Biri

CHIRILA source: CHIRILA/v1/Terrell

Angela Terrill. 1999. Biri lexicons. Australian Institute of Aboriginal and Torres Strait Islander Studies, Australian Indigenous Languages Collection. ASEDA 0700. Canberra. `http://aiatsis.gov.au/sites/default/files/catalogue_resources/0700_access.zip`

Bularnu

Gavan Breen. 1988. Bularnu grammar and vocabulary machine-readable files. Australian Institute of Aboriginal and Torres Strait Islander Studies, Australian Indigenous Languages Collection. ASEDA 0007. Canberra. http://aiatsis.gov.au/sites/default/files/catalogue_resources/0007_access.zip

Bunuba

Kimberley Language Resource Centre. 2010. *Bunuba draft dictionary*. Halls Creek, WA, Australia: Kimberley Language Resource Centre. https://www.klrc.org.au/dictionary/bunuba/lexicon/01.htm (26 July, 2018)

Burarra

CHIRILA source: CHIRILA/v1/Glasgow

Kathleen Glasgow. 1994. *Burarra-Gun-nartpa dictionary with English finder list*. SIL

Butchulla

Jeanie Bell. 2003. *A sketch grammar of the Badjala language of Gari (Fraser Island)*. Melbourne: University of Melbourne M.A. Thesis

Central Arrernte

David P. Wilkins. N.d. Mparntwe Arrernte (Aranda): Studies in the structure and semantics of grammar. Australian Institute of Aboriginal and Torres Strait Islander Studies, Australian Indigenous Languages Collection. ASEDA 0476. Canberra

Phonemic normalisation: Labialized consonants normalized to C + w. Prestopped nasals normalized to stop + nasal sequence. Prepalatalized consonants normalized to j + C.

Dalabon

Nicholas Evans, Francesca Merlan & Maggie Tukumba. 2004. *A first dictionary of Dalabon (Ngalkbon)*. Maningrida, NT, Australia: Maningrida Arts & Culture. 489 pp.

Dhangu

R. David Zorc. 2004. Yolngu Matha dictionary. Australian Institute of Aboriginal and Torres Strait Islander Studies, Australian Indigenous Languages Collection. ASEDA 0778. Canberra. http://aiatsis.gov.au/sites/default/files/catalogue_resources/0778_Access.zip

Phonemic normalisation: Lenis retroflex stop normalized to retroflex flap.

Dharumbal

CHIRILA source: CHIRILA/v2/ter02

Angela Terrill. 2002. *Dharumbal: The language of Rockhampton, Australia* (Pacific Linguistics 525). Canberra: Pacific Linguistics. 108 pp. https://doi.org/10.15144/PL-525

Dhay'yi

Djarrayang Wunungmurra. 1993. Dhalwangu dictionary. Australian Institute of Aboriginal and Torres Strait Islander Studies, Australian Indigenous Languages Collection. ASEDA 0502. Canberra

Phonemic normalisation: Lenis retroflex stop normalized to retroflex flap; all other voicing is allophonic.

Diyari

Peter K. Austin. 1981a. *A grammar of Diyari, South Australia* (Cambridge Studies in Linguistics 32). Cambridge; New York: Cambridge University Press. 269 pp.

Phonemic normalisation: Phonetic trill-released stop normalized as stop + trill. Otherwise, voiced stops normalized as taps.

Djabugay

Sue Robertson & Bruce A. Sommer. 1997. Jaabugay dictionary. Australian Institute of Aboriginal and Torres Strait Islander Studies, Australian Indigenous Languages Collection. ASEDA 0013. Canberra

Djapu

CHIRILA source: CHIRILA/v1/mor83

Frances Morphy. 1983. Djapu, a Yolngu dialect. In R. M. W. Dixon & Barry Blake (eds.), *Handbook of Australian languages*, vol. 3, 5 vols., 1–188. Amsterdam: John Benjamins

Phonemic normalisation: Lenis retroflex stop normalized to retroflex flap.

Djinang

CHIRILA source: CHIRILA/v1/ASEDA0009

Bruce E. Waters. 1988. Djinang dictionary. Australian Institute of Aboriginal and Torres Strait Islander Studies, Australian Indigenous Languages Collection. ASEDA 0009. Canberra

Phonemic normalisation: Glottal closure normalized to a segment phoneme.

Duungidjawu

CHIRILA source: CHIRILA/v2/K&W 04

Suzanne Kite & Stephen A. Wurm. 2004. *The Duungidjawu language of southeast Queensland: Grammar, texts and vocabulary* (Pacific Linguistics 553). Canberra: Pacific Linguistics. 298 pp. https://doi.org/10.15144/PL-553

Dyirbal

CHIRILA source: CHIRILA/v1/dix72

R. M. W. Dixon. 1972. *The Dyirbal language of North Queensland*. Cambridge: Cambridge University Press

Emmi

Lysbeth J. Ford. 1998. *A description of the Emmi language of the Northern Territory of Australia*. Canberra: The Australian National University dissertation. 446 pp. http://hdl.handle.net/1885/10796 (12 July, 2018)

Erre

Bruce Birch. 2006. *A first dictionary of Erre, Mengerrdji and Urningangk: Three languages from the Alligator Rivers region of the north western Arnhem Land, Northern Territory, Australia*. Jabiru, NT, Australia: Gundjeihmi Aboriginal Corporation. 125 pp.

Phonemic normalisation: Double stops normalized as fortis.

Gamilaraay

CHIRILA source: CHIRILA/v1/ash03

Anna Ash, John Giacon & Amanda Lissarrague. 2003. *Gamilaraay, Yuwaalaraay & Yuwaalayaay dictionary*. Alice Springs, NT, Australia: IAD Press. 344 pp.

Gangulu

CHIRILA source: CHIRILA/v1/Terrell

Angela Terrill. 1999. Biri lexicons. Australian Institute of Aboriginal and Torres Strait Islander Studies, Australian Indigenous Languages Collection. ASEDA 0700. Canberra. http://aiatsis.gov.au/sites/default/files/catalogue_resources/0700_access.zip

Gidabal

CHIRILA source: CHIRILA/v1/cro78

Terry Crowley. 1978. *The Middle Clarence dialects of Bandjalang*. Vol. 12 (Research and regional studies). Canberra: Australian Institute of Aboriginal Studies

Gooniyandi

Kimberley Language Resource Centre. 1993. *Gooniyandi wordbook*. Halls Creek, Western Australia: Kimberley Language Resource Centre. 84 pp.

Gugu Badhun

CHIRILA source: CHIRILA/v1/sut73

Peter John Sutton. 1973. Gugu-Badhun and its neighbours. In *Gugu-Badhun and its neighbours: A linguistic salvage study*, 24–67. Sydney: Macquarie University

Gumbaynggir

Murrbay Aboriginal and Culture Cooperative. 2001. *A Gumbaynggir language dictionary = Gumbayngirr bijarr jandaygam*. Canberra: Aboriginal Studies Press. 160 pp.

Gunya

CHIRILA source: CHIRILA/v1/dixbla81

Gavan Breen. 1981a. Margany and Gunya. In R. M. W. Dixon & Barry Blake (eds.), *Handbook of Australian languages*, vol. 2, 275–394. Amsterdam: John Benjamins

Gupapuyngu

CHIRILA source: CHIRILA/v1/BL

Beulah Lowe & Beulah Lowe. 1976. Temporary Gupapuyngu dictionary. Milingimbi, NT, Australia

Gurindji

Felicity Meakins, Patrick McConvell, et al. 2013. *Gurindji to English dictionary*. Batchelor, NT, Australia: Batchelor Press. 596 pp.

Gurr-Goni

Rebecca Green & Leila Nimbadja (eds.). 2015. *Gurr-Goni to English dictionary*. Google-Books-ID: gIGpDAEACAAJ. Batchelor, NT, Australia: Batchelor Press. 357 pp.

Phonemic normalisation: Geminate stops normalized to fortis.

Guugu Yimidhirr

John B. Haviland. 1979. Guugu Yimidhirr. In R. M. W. Dixon & Barry Blake (eds.), *Handbook of Australian languages*, vol. 1, 5 vols., 26–180. Amsterdam: John Benjamins

Guwamu

CHIRILA source: CHIRILA/v1/Austin 1980

Peter K. Austin. 1980. Guwamu vocabulary and English-Guwamu finder list. Cambridge, MA

Iwaidja

Noreen Pym & Bonnie Larrimore. 2011. *Iwaidja-English Interactive Dictionary*. AuSIL Interactive Dictionary Series A-2. In collab. with Charles E. Grimes & Maarten Lecompte. `http://ausil.org/Dictionary/Iwaidja/lexicon/mainintro.htm` (23 July, 2018)

Jaru

Tasaku Tsunoda. 1981. Jaru wordlist. In David Nash (ed.). In collab. with Kathleen Menning, Joyce Hudson & G Cooling, *Sourcebook for Central Australian languages*. ASEDA 0119. Alice Springs, NT, Australia: Institute for Aboriginal Development

Phonemic normalisation: iji and uwu normalized as long high vowels.

Jawoyn

Francesca Merlan & Pascale Jacq. 2005. *Jawoyn-English dictionary and English finder-list*. In collab. with Jawoyn elders. Katherine, NT, Australia: Diwurruwurru-jaru Aboriginal Corporation. 342 pp.

Jiwarli

CHIRILA source: CHIRILA/v2/ASEDA0435

Peter K. Austin. N.d.(a). A dictionary of Jiwarli. Australian Institute of Aboriginal and Torres Strait Islander Studies, Australian Indigenous Languages Collection. ASEDA 0435. Canberra

Kalkatungu

CHIRILA source: CHIRILA/v2/ASEDA0205

Barry J. Blake. 1990a. Kalkatungu vocabulary. Australian Institute of Aboriginal and Torres Strait Islander Studies, Australian Indigenous Languages Collection. ASEDA 0205. Canberra

Phonemic normalisation: Double short vowels normalized as long.

Karajarri

Kevin R. McKelson. 1989. Studies in Karajarri

Kariyarra

Sue Smythe & Manny Lockyer. N.d. Kariyarra wordlist. Australian Institute of Aboriginal and Torres Strait Islander Studies, Australian Indigenous Languages Collection. ASEDA 0582. Canberra. `http://aiatsis.gov.au/sites/default/files/catalogue_resources/0582_access.zip`

Kartujarra

Geoffrey N. O'Grady. 1988a. Gardudjarra wordlist. Australian Institute of Aboriginal and Torres Strait Islander Studies, Australian Indigenous Languages Collection. ASEDA 0067. Canberra. `http://aiatsis.gov.au/sites/default/files/catalogue_resources/0067_access.zip`

Kija

CHIRILA source: CHIRILA/v1/bly01

Noel Blyth. 2001. Wangka dictionary and grammar. Australian Institute of Aboriginal and Torres Strait Islander Studies, Australian Indigenous Languages Collection. ASEDA 0709. Canberra. `http://aiatsis.gov.au/sites/default/files/catalogue_resources/0709_access.zip`

Kok Nar

Bruce A. Sommer. N.d.(b). Koko Narr. Fryer Library Bruce Sommer Collection. UQFL476_b10f03_64, UQFL476_b10f03_65. Brisbane

Koko Bera

Paul D. Black & Kokoberrin Tribal Aboriginal Corporation. 2007. The Kokoberrin and their languages. Australian Institute of Aboriginal and Torres Strait Islander Studies, Australian Indigenous Languages Collection. MS 4584

Kugu Nganhcara

CHIRILA source: CHIRILA/v1/ASEDA0021

Ian Smith & Steve Johnson. 1989. Kugu Nganchara. Australian Institute of Aboriginal and Torres Strait Islander Studies, Australian Indigenous Languages Collection. ASEDA 0021. Canberra

Kukatj

Gavan Breen. 1991. Kukatj grammar machine-readable files. Australian Institute of Aboriginal and Torres Strait Islander Studies, Australian Indigenous Languages Collection. ASEDA 0022. Canberra. `http://aiatsis.gov.au/sites/default/files/catalogue_resources/0022_access.zip`

Phonemic normalisation: Featureless vowel normalized as schwa.

Kukatja

CHIRILA source: CHIRILA/v1/ASEDA0504

Anthony Rex Peile & Hilaire Valiquette. N.d. A basic Kukatja to English dictionary. Australian Institute of Aboriginal and Torres Strait Islander Studies, Australian Indigenous Languages Collection. ASEDA 0504. Canberra

Kuku Yalanji

Henry D. Hershberger & Ruth Hershberger. 1986. *Kuku-Yalanji dictionary*. In collab. with Australian Aborigines Branch Summer Institute of Linguistics. Vol. 7 (Work Papers of SIL - AAIB. Series B). Darwin: Summer Institute of Linguistics, Australian Aborigines Branch. 294 pp.

Kurrama

Alan C. Dench. N.d. Kurrama. Australian Institute of Aboriginal and Torres Strait Islander Studies, Australian Indigenous Languages Collection. ASEDA 0481. Canberra. http://aiatsis.gov.au/sites/default/files/catalogue_resources/0481_access.zip

Kurtjar

CHIRILA source: CHIRILA/v1/ASEDA0026

Paul D. Black & Rolly Gilbert. 1988. Kurtjar dictionary. Australian Institute of Aboriginal and Torres Strait Islander Studies, Australian Indigenous Languages Collection. ASEDA 0026. Canberra

Phonemic normalisation: Retroflex glide˜tap normalized as glide.

Kuugu Ya'u

David A. Thompson. 1988. "Sand Beach" language: An outline of Kuuku Ya'u and Umpila. Australian Institute of Aboriginal and Torres Strait Islander Studies, Australian Indigenous Languages Collection. ASEDA 0027. Canberra

Lardil

Kenneth Hale & Ngakulmungan Kangka Leman. 1997. *Lardil dictionary: A vocabulary of the language of the Lardil people, Mornington Island, Gulf of Carpentaria, Queensland; with English-Lardil finder list*. Gununa, QLD, Australia: Mornington Shire Council. 347 pp.

Larrakia

Mark Harvey. 2004. *Larrakia dictionary*. Darwin: ATSIC, Yirra Bandoo Aboriginal Corporation. 92 pp.

Limilngan

Mark Harvey. 2001. *A grammar of Limilngan: A language of the Mary River region, Northern Territory, Australia* (Pacific Linguistics 516). Canberra: Pacific Linguistics. 209 pp. https://doi.org/10.15144/PL-516

Linngithigh

Kenneth Hale. 1999. A Linngithigh vocabulary. Australian Institute of Aboriginal and Torres Strait Islander Studies, Australian Indigenous Languages Collection. ASEDA 0687. Canberra

Phonemic normalisation: Trill-released stop normalized as stop + trill. Prenasalized stops normalized to nasal + lenis stop.

Malkana

Andrew Gargett. 2011. *A salvage grammar of Malgana, the language of Shark Bay, Western Australia* (Pacific Linguistics 624). Canberra: Pacific Linguistics. 102 pp. https://doi.org/10.15144/PL-624

Malyangapa

Luise A. Hercus. 1989. Maljangapa-Wadigali vocabulary. Australian Institute of Aboriginal and Torres Strait Islander Studies, Australian Indigenous Languages Collection. ASEDA 0246. Canberra. http://aiatsis.gov.au/sites/default/files/catalogue_resources/0246_access.zip

Mangala

Kevin McKelson. 1989a. Mangala wordlist. Australian Institute of Aboriginal and Torres Strait Islander Studies, Australian Indigenous Languages Collection. ASEDA 0220. Canberra. http://aiatsis.gov.au/sites/default/files/catalogue_resources/0220_access.zip

Margany

CHIRILA source: CHIRILA/v1/bre81

Gavan Breen. 1981a. Margany and Gunya. In R. M. W. Dixon & Barry Blake (eds.), *Handbook of Australian languages*, vol. 2, 275–394. Amsterdam: John Benjamins

Marra

Jeffrey Heath. 1981. *Basic materials in Mara: Grammar, texts and dictionary* (Pacific Linguistics Series C 60). Canberra: Pacific Linguistics. 534 pp. https://doi.org/10.15144/PL-C60

Martuthunira

Alan C. Dench. 1995. *Martuthunira, a language of the Pilbara region of Western Australia* (Pacific Linguistics Series C 125). Canberra: Pacific Linguistics. 406 pp. https://doi.org/10.15144/PL-C125

Matngele

Franklin D. Zandvoort. 1999. *A grammar of Matngele*. Armidale, NSW, Australia: University of New England B.A. (Hons)

Mawng

Ruth Singer et al. 2015. *Mawng dictionary v1.0*. mawngngaralk.org.au/main/dictionary.php (23 July, 2018)

Mbabaram

R. M. W. Dixon. 1991a. Mbabaram. In R. M. W. Dixon & Barry J. Blake (eds.), *Handbook of Australian languages*, vol. 4, 5 vols., 348–402. Melbourne: Oxford University Press

Phonemic normalisation: Labialized consonants normalized to C + w. Final schwa treated as a phoneme.

Mengerrdji

Bruce Birch. 2006. *A first dictionary of Erre, Mengerrdji and Urningangk: Three languages from the Alligator Rivers region of the north western Arnhem Land, Northern Territory, Australia*. Jabiru, NT, Australia: Gundjeihmi Aboriginal Corporation. 125 pp.

Phonemic normalisation: Double stops normalized as fortis.

Miriwoong

F. M. Kofod. 1976. Miriwung - English. Australian Institute of Aboriginal and Torres Strait Islander Studies, Australian Indigenous Languages Collection. MS 1896. Canberra

Mirniny

Geoffrey N. O'Grady & Edward M. Curr. 1988. Mirniny wordlist. Australian Institute of Aboriginal and Torres Strait Islander Studies, Australian Indigenous Languages Collection. ASEDA 0070. Canberra. `http://aiatsis.gov.au/sites/default/files/catalogue_resources/0070_access.zip`

Mudburra

David Nash et al. 1988. Mudburra wordlist. Australian Institute of Aboriginal and Torres Strait Islander Studies, Australian Indigenous Languages Collection. ASEDA 0031. Canberra. `http://aiatsis.gov.au/sites/default/files/catalogue_resources/0031_access.zip`

Murrinh-patha

Chester S. Street. 1987. *An introduction to the language and culture of the Murrinh-Patha*. Darwin: Summer Institute of Linguistics. Australian Aborigines Branch. 117, plus audio cassette

Muruwari

CHIRILA source: CHIRILA/v1/ASEDA0252

Lynette Frances Oates. 1992. *Muruwari (Moo-roo-warri) dictionary: Words of an Aboriginal language of north-western New South Wales*. Albury, NSW, Australia: Graeme van Brummelen, produced with the assistance of the Australian Institute of Aboriginal & Torres Strait Islander Studies. 97 pp.

Nakara

Bronwyn Eather, Yurrbukka Community & Bawinanga Aboriginal Corporation. 2005. *A first dictionary of Na-Kara*. In collab. with Jimmy Kalamirnda. Winnellie, NT, Australia: Maningrida Arts & Culture. 199 pp.

Phonemic normalisation: Geminate stops normalized to fortis.

Ngaanyatjarra

Amee Glass. 1988. Ngaanyatjarra wordlist. Australian Institute of Aboriginal and Torres Strait Islander Studies, Australian Indigenous Languages Collection. ASEDA 0033. Canberra. http://aiatsis.gov.au/sites/default/files/catalogue_resources/0033_access.zip

Ngadjunmaya

Wangka Maya Pilbara Aboriginal Language Centre. 2008a. *Ngajumaya dictionary 2008*. South Hedland, WA, Australia: Wangka Maya Pilbara Aboriginal Language Centre. 16 pp.

Ngalakgan

CHIRILA source: CHIRILA/v1/mor83

Francesca C. Merlan. 1983. *Ngalakan grammar, texts and vocabulary* (Pacific Linguistics Series B 89). Canberra: Pacific Linguistics. 229 pp. https://doi.org/10.15144/PL-B89

Phonemic normalisation: Geminate stops normalized to fortis.

Ngandi

Jeffrey Heath. 1978b. *Ngandi grammar, texts and dictionary*. Canberra: Australian Institute of Aboriginal Studies

Ngardily

Thomas M. Green. 1988. Ngardily wordlist. Australian Institute of Aboriginal and Torres Strait Islander Studies, Australian Indigenous Languages Collection. ASEDA 0034. Canberra. http://aiatsis.gov.au/sites/default/files/catalogue_resources/0034_access.zip

Ngarinyin

Howard H. J. Coate & A. P. Elkin. 1974. *Ngarinjin-English dictionary* (Oceania linguistic monographs 16). Sydney: University of Sydney. 534 pp.

Ngarinyman

Caroline Jones. 2005. Ngarinman vocabulary. Australian Institute of Aboriginal and Torres Strait Islander Studies, Australian Indigenous Languages Collection. ASEDA 0796. Canberra

Ngarla

Alexander Brown & Brian Geytenbeek. N.d. Ngarla-English dictionary (interim), English-Ngarla wordlist. Australian Institute of Aboriginal and Torres Strait Islander Studies, Australian Indigenous Languages Collection. ASEDA 0060. Canberra

Ngarluma

Kenneth Hale. 1989. Ngarluma wordlist. Australian Institute of Aboriginal and Torres Strait Islander Studies, Australian Indigenous Languages Collection. ASEDA 0037. Canberra. `http://aiatsis.gov.au/sites/default/files/catalogue_resources/0037_access.zip`

Ngawun

CHIRILA source: CHIRILA/v2/BreenMayi

Gavan Breen. 1981b. *The Mayi languages of the Queensland Gulf Country* (A.I.A.S. New Series 29). Canberra: Australian Institute of Aboriginal Studies. 238 pp.

Ngiyambaa

CHIRILA source: CHIRILA/v1/don-lex

Tamsin Donaldson. 1997. *Ngiyambaa wordworld*. Canberra: The Author, Australian Institute of Aboriginal & Torres Strait Islander Studies

Nhanda

CHIRILA source: CHIRILA/v1/ble01

Juliette Blevins. 2001. *Nhanda: An Aboriginal language of Western Australia* (Oceanic linguistics special publication 30). Honolulu: University of Hawai'i Press. 170 pp.

Nhangu

CHIRILA source: CHIRILA/v1/CB-fieldnotes

Bentley James. 2003. *Yan-nhangu dictionary*. In collab. with Laurie Baymarrwanga et al. Milingimbi, NT, Australia: B. James. 34 pp.

Nhirrpi

CHIRILA source: CHIRILA/v1/bow-nhi

Claire Bowern. 1999. Nhirrpi vocabulary, based on fieldnotes of S. A. Wurm

Nukunu

CHIRILA source: CHIRILA/v2/her92

Luise A. Hercus. 1992a. *A Nukunu dictionary*. Canberra: Department of Linguistics, Australian National University. 51 pp.

Phonemic normalisation: Voiced retroflex stop normalized to retroflex tap.

Nungali

Janet E. Bolt, W. G. Hoddinott & F. M. Kofod. 1971. *An elementary grammar of the Ngaliwuru language of the Northern Territory*. MS 211, mimeographed. Canberra: Australian Institute of Aboriginal & Torres Strait Islander Studies, Australian Indigenous Languages Collection

Nyamal

Albert Burgman. 2007b. *Nyamal dictionary: English-Nyamal finderlist and topical wordlist*. In collab. with Wangka Maya Pilbara Aboriginal Language Centre. South Hedland, WA, Australia: Wangka Maya Pilbara Aboriginal Language Centre. 59 pp.

Nyangumarta

Brian Geytenbeek, Helen Geytenbeek & Wangka Maya Pilbara Aboriginal Language Centre. 1991. *Nyangumarta-English dictionary (interim), with an English-Nyangumarta finder list*. Port Hedland, WA, Australia: Wangka Maya Pilbara Aboriginal Language Centre. 119 pp.

Nyawaygi

R. M. W. Dixon. 1983. Nyawaygi. In R. M. W. Dixon & Barry J. Blake (eds.), *Handbook of australian languages*, vol. 3, 5 vols., 431–531. Amsterdam: John Benjamins

Nyikina

CHIRILA source: CHIRILA/v1/Stokes

Bronwyn Stokes. N.d. Nyikina-English: A first lexicon. Australian Institute of Aboriginal and Torres Strait Islander Studies, Australian Indigenous Languages Collection. ASEDA0472. Canberra. `http://aiatsis.gov.au/sites/default/files/catalogue_resources/0472_access.zip`

Nyiyaparli

Geoffrey N. O'Grady. 1988b. Nyiyabali wordlist. Australian Institute of Aboriginal and Torres Strait Islander Studies, Australian Indigenous Languages Collection. ASEDA 0074. Canberra. `http://aiatsis.gov.au/sites/default/files/catalogue_resources/0074_access.zip`

Ogh Angkula

Bruce A. Sommer. N.d.(a). ANkula. Fryer Library Bruce Sommer Collection. UQFL476_b10f04. Brisbane

Ogh Unyjan

Bruce A. Sommer. N.d.(c). Ogh Unydjan. Fryer Library Bruce Sommer Collection. UQFL476_b09f03_s05. Brisbane

Olkol

Philip J. Hamilton. 1997b. *Uw Olkola and Uw Oykangand Alphabetical Search Index*. Oykangand and Olkola Multimedia Dictionary. http://www.oocities.org/athens/delphi/2970/olkola.htm

Oykangand

Philip J. Hamilton. 1997b. *Uw Olkola and Uw Oykangand Alphabetical Search Index*. Oykangand and Olkola Multimedia Dictionary. http://www.oocities.org/athens/delphi/2970/olkola.htm

Panyjima

Alan C. Dench. 1991. Panyjima. Australian Institute of Aboriginal and Torres Strait Islander Studies, Australian Indigenous Languages Collection. ASEDA 0375. Canberra. http://aiatsis.gov.au/sites/default/files/catalogue_resources/0375_access.zip

Patjtjamalh

Lysbeth J. Ford. 1997. *Batjamalh: Dictionary and texts*. Bungendore, NSW, Australia: L.J. Ford. 108 pp.

Payungu

CHIRILA source: CHIRILA/v1/ASEDA0394

Peter K. Austin. N.d.(d). Payungu - English dictionary. Australian Institute of Aboriginal and Torres Strait Islander Studies, Australian Indigenous Languages Collection. ASEDA 0394. Canberra

Pintupi

Kenneth Hansen & Lesley Hansen. 1992. *Pintupi/Luritja dictionary*. 3rd edn. Alice Springs, NT, Australia: Institute for Aboriginal Development. 267 pp.

Pitta Pitta

CHIRILA source: CHIRILA/v1/bla0275

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

Purduna

Albert Burgman. 2007a. *Burduna dictionary: English-Burduna wordlist and thematic wordlist*. In collab. with Wangka Maya Pilbara Aboriginal Language Centre. South Hedland, WA, Australia: Wangka Maya Pilbara Aboriginal Language Centre. 86 pp.

Putijarra

Wangka Maya Pilbara Aboriginal Language Centre & Australian Institute of Aboriginal and Torres Strait Islander Studies. 2004. *Putijarra-English wordlist, English-Putijarra finder topical wordlist &sketch morphology*. South Hedland, WA, Australia: Wangka Maya Pilbara Aboriginal Language Centre. 118 pp.

Rembarrnga

Graham McKay. 2011. Rembarrnga dictionary and grammar. Australian Institute of Aboriginal and Torres Strait Islander Studies, Australian Indigenous Languages Collection. ASEDA 0607. Canberra. `https://iats.ent.sirsidynix.net.au/client/en_AU/external/search/detailnonmodal/ent:$002f$002fSD_ILS$002f0$002fSD_ILS:402512/one?qu=AILEC+0607`

Phonemic normalisation: Geminate stops normalized to fortis.

Ritharrngu

CHIRILA source: CHIRILA/v1/Heath

Jeffrey Heath. 1976b. Ritharngu. In Robert M. W. Dixon (ed.), *Grammatical categories in australian languages* (Linguistic series 22), 285–287. Canberra: Australian Institute of Aboriginal Studies

Southern Paakintyi

Luise A. Hercus. N.d.(a). Paakantyi dictionary. Australian Institute of Aboriginal and Torres Strait Islander Studies, Australian Indigenous Languages Collection. ASEDA 0525. Canberra. `http://aiatsis.gov.au/sites/default/files/catalogue_resources/0525_access.zip`

Thaayorre

Tom Foote & Allen Hall. 1993. *Kuuk Thaayorre dictionary: Thaayorre/English; September, 1966-92*. Brisbane: Jolien Press. 239 pp.

Thalanyji

CHIRILA source: CHIRILA/v2/ASEDA0437

Peter K. Austin. N.d.(b). A dictionary of Thalanyji. Australian Institute of Aboriginal and Torres Strait Islander Studies, Australian Indigenous Languages Collection. ASEDA 0437. Canberra

Tharrkari

Peter K. Austin. 1992. *A dictionary of Tharrgari, Western Australia*. Bundoora, Victoria, Australia: La Trobe University. 60 pp.

Thaynakwithi

Gloria Thancoupie Fletcher. 2007. *Thanakupi's guide to language and culture: A Thaynakwith dictionary*. North Sydney, NSW, Australia: Jennifer Isaacs Arts & Publishing. 144 pp.

Thirarri

Peter K. Austin. 1981a. *A grammar of Diyari, South Australia* (Cambridge Studies in Linguistics 32). Cambridge; New York: Cambridge University Press. 269 pp.

Phonemic normalisation: Phonetic trill-released stop normalized as stop + trill. Otherwise, voiced stops normalized as taps.

Tiwi

Jennifer R. Lee. 2013. *Tiwi-English Interactive Dictionary*. AuSIL Interactive Dictionary Series A-4. In collab. with Charles E. Grimes & Maarten Lecompte. `http://ausil.org/Dictionary/Tiwi/lexicon/main.htm` (23 July, 2018)

Phonemic normalisation: r + alveolar normalized as retroflex. Following Breen 1979, (w)o is normalized as wa.

Umpila

Geoffrey N. O'Grady. 1988c. Umpila wordlist. Australian Institute of Aboriginal and Torres Strait Islander Studies, Australian Indigenous Languages Collection. ASEDA 0094. Canberra. `http://aiatsis.gov.au/sites/default/files/catalogue_resources/0094_access.zip`

Unggumi

William McGregor. 1985. *Handbook of Kimberley languages: Word lists*. Vol. 2. Broome: Kimberley Resource Centre

Urningangg

Bruce Birch. 2006. *A first dictionary of Erre, Mengerrdji and Urningangk: Three languages from the Alligator Rivers region of the north western Arnhem Land, Northern Territory, Australia*. Jabiru, NT, Australia: Gundjeihmi Aboriginal Corporation. 125 pp.

Waalubal

CHIRILA source: CHIRILA/v1/cro78

Terry Crowley. 1978. *The Middle Clarence dialects of Bandjalang*. Vol. 12 (Research and regional studies). Canberra: Australian Institute of Aboriginal Studies

Waanyi

CHIRILA source: CHIRILA/v1/WaanyiDict

Mary Laughren. 2016. Waanyi dictionary database. Brisbane

Wagiman

Stephen Wilson & Mark Harvey. 2001. *The Wagiman online dictionary*. The University of Syndey. In collab. with Lulu Martin Dalpbalngali. http://sydney.edu.au/arts/linguistics/research/wagiman/dict/dict.html (9 July, 2018)

Walmajarri

Joyce Hudson & Eirlys Richards. 1993. Walmajarri dictionary. Australian Institute of Aboriginal and Torres Strait Islander Studies, Australian Indigenous Languages Collection. ASEDA 0167. Canberra

Wambaya

CHIRILA source: CHIRILA/v1/RN Wambaya 29,269,292; NE NA

Wangkatja

Noel Blyth. 2001. Wangka dictionary and grammar. Australian Institute of Aboriginal and Torres Strait Islander Studies, Australian Indigenous Languages Collection. ASEDA 0709. Canberra. http://aiatsis.gov.au/sites/default/files/catalogue_resources/0709_access.zip

Wangkumara

CHIRILA source: CHIRILA/v1/robnd

Carol Robertson. 1985. *Wangkumara grammar and dictionary*. Sydney: Department of Technical & Further Education, Aboriginal Education Unit. 90 pp.

Phonemic normalisation: Double a normalized to long vowel.

Wanyjirra

Chikako Senge. 2015. *A grammar of Wanyjirra, a language of northern Australia*. Canberra: The Australian National University dissertation. https://openresearch-repository.anu.edu.au/bitstream/1885/109341/3/Senge%20Thesis%202016.pdf (23 July, 2018)

Wardaman

CHIRILA source: CHIRILA/v1/Merlan NA

Warlmanpa

David Nash, Kenneth Hale & Gavan Breen. 1984. Preliminary vocabulary of Warlmanpa. Australian Institute of Aboriginal and Torres Strait Islander Studies, Australian Indigenous Languages Collection. ASEDA 0049. Canberra

Warlpiri

CHIRILA source: CHIRILA/v2/WarlpiriDict

Steve Swartz. 1996. Warlpiri draft dictionary

Warluwarra

Gavan Breen. 1990. Warluwara grammar and wordlist. Australian Institute of Aboriginal and Torres Strait Islander Studies, Australian Indigenous Languages Collection. ASEDA 0253. Canberra

Phonemic normalisation: Prenasalized stops normalized to nasal + lenis stop. Tense glides normalized to fricatives. Tense lateral normalized to double lateral.

Warndarrang

CHIRILA source: CHIRILA/v1/hea80

Jeffrey Heath. 1980. *Basic materials in Warndarang: Grammar, texts and dictionary* (Pacific Linguistics Series B 72). Canberra: Pacific Linguistics. 186 pp. https://doi.org/10.15144/PL-B72

Warnman

CHIRILA source: CHIRILA/v2/ASEDA0334

Wangka Maya Pilbara Aboriginal Language Centre. N.d. Warnman wordlist. Australian Institute of Aboriginal and Torres Strait Islander Studies, Australian Indigenous Languages Collection. ASEDA 0334. Canberra. http://aiatsis.gov.au/sites/default/files/catalogue_resources/0334_Access.zip

Warrgamay

CHIRILA source: CHIRILA/v1/dixbla81

R. M. W. Dixon. 1981. Wargamay. In R. M. W. Dixon & Barry Blake (eds.), *Handbook of Australian languages*, vol. 2, 5 vols., 1–145. Amsterdam: John Benjamins

Warriyangga

Peter K. Austin. N.d.(c). A dictionary of Warriyangka. Australian Institute of Aboriginal and Torres Strait Islander Studies, Australian Indigenous Languages Collection. ASEDA 0439. Canberra

Watjarri

Doreen Mackman (ed.). 2012. *Wajarri dictionary: The language of the Murchison Region of Western Australia*. In collab. with Irra Wangga Language Centre & Yamaji Language Aboriginal Corporation. Geraldton, WA, Australia: Irra Wangga Language Centre. 249 pp. http://www.bundiyarra.com.au/wajarriApp/ (23 July, 2018)

Wemba Wemba

Luise A. Hercus. 1992b. *Wembawemba dictionary*. Canberra: L.A. Hercus. 116 pp.

Western Arrernte

Gavan Breen. 2000. *Introductory dictionary of Western Arrernte*. In collab. with John Pfitzner. Alice Springs, NT, Australia: IAD Press. 120 pp.

Phonemic normalisation: Labialized consonants normalized to C + w. Prestopped nasals normalized to stop + nasal sequence. Prepalatalized consonants normalized to j + C.

Western Wakaya

Gavan Breen. 2006. Wakaya. Australian Institute of Aboriginal and Torres Strait Islander Studies, Australian Indigenous Languages Collection. ASEDA 0047. Canberra. `http://aiatsis.gov.au/sites/default/files/catalogue_resources/0047_access.zip` (30 July, 2018)

Wik Mungkan

Christine Kilham et al. 2011. *Wik Mungkan-English Interactive Dictionary*. AuSIL Interactive Dictionary Series A-6. In collab. with Charles E. Grimes & Maarten Lecompte. `http://ausil.org/Dictionary/Wik-Mungkan/lexicon/mainintro.htm` (26 July, 2018)

Wik-Ngathan

CHIRILA source: CHIRILA/v2/PS

Peter Sutton. 1995. *Wik-Ngathan dictionary*. Prospect, South Australia: Caitlin Press. 182 pp.

Wirangu

Luise A. Hercus. 1999. *A grammar of the Wirangu language from the West Coast of South Australia* (Pacific Linguistics Series C 150). Canberra: Pacific Linguistics. 239 pp. `https://doi.org/10.15144/PL-C150`

Phonemic normalisation: Double a normalized to long vowel.

Wiri

Angela Terrill. 1999. Biri lexicons. Australian Institute of Aboriginal and Torres Strait Islander Studies, Australian Indigenous Languages Collection. ASEDA 0700. Canberra. `http://aiatsis.gov.au/sites/default/files/catalogue_resources/0700_access.zip`

Worrorra

William McGregor. 1985. *Handbook of Kimberley languages: Word lists*. Vol. 2. Broome: Kimberley Resource Centre

Wotjobaluk

Luise A. Hercus. N.d.(b). Wergaia vocabulary 1. Australian Institute of Aboriginal and Torres Strait Islander Studies, Australian Indigenous Languages Collection. ASEDA 0273. Canberra. `http://aiatsis.gov.au/sites/default/files/catalogue_resources/0273_access.zip`

Wubuy

CHIRILA source: CHIRILA/v1/Nunggubuyu Flora Fauna

Jeffrey Heath. 1976a. Nunggubuyu flora and fauna terminology. Australian Institute of Aboriginal and Torres Strait Islander Studies, Australian Indigenous Languages Collection. Canberra

Phonemic normalisation: Morphophonemic w1, w2 represented as their phonemic realizations w,b,k.

Yadhaykenu

Terry Crowley. 1983. Uradhi. In R. M. W. Dixon & Barry J. Blake (eds.), *Handbook of Australian languages*, vol. 3, 5 vols., 307–428. Amsterdam: John Benjamins

Yalarnnga

CHIRILA source: CHIRILA/v1/ASEDA0204

Gavan Breen & Barry J Blake. N.d. Yalarnnga vocab. Australian Institute of Aboriginal and Torres Strait Islander Studies, Australian Indigenous Languages Collection. ASEDA 0204. Canberra

Yandruwandha

CHIRILA source: CHIRILA/v1/breyandr

Gavan Breen. 2004. *Innamincka talk: A grammar of the Innamincka dialect of Yandruwandha with notes on other dialects* (Pacific Linguistics 558). Canberra: Pacific Linguistics. 245 pp. `https://doi.org/10.15144/PL-558`

Phonemic normalisation: Trill-released stop normalized as stop + trill. Prestopped laterals normalized to stop + lateral sequence.

Yanyuwa

John Bradley. N.d. Yanyuwa dictionary. Australian Institute of Aboriginal and Torres Strait Islander Studies, Australian Indigenous Languages Collection. ASEDA 0382. Canberra

Phonemic normalisation: Prenasalized stops normalized to nasal + stop sequence.

Yawijibaya

William McGregor. 1985. *Handbook of Kimberley languages: Word lists*. Vol. 2. Broome: Kimberley Resource Centre

Yaygir

CHIRILA source: CHIRILA/v1/morelli2011

Steve Morelli. 2012. *Yaygirr dictionary and grammar*. In collab. with Many Rivers Aboriginal Language Centre. Nambucca Heads, NSW, Australia: Muurrbay Aboriginal Language & Culture Co-operative. 254 pp.

Yidiny

CHIRILA source: CHIRILA/v2/dix91

R. M. W. Dixon. 1991b. *Words of our country: Stories, place names, and vocabulary in Yidiny, the Aboriginal language of the Cairns-Yarrabah region*. In collab. with Tony Irvine. St Lucia, Qld: University of Queensland Press. 312 pp.

Yindjibarndi

Bruce Anderson, E. Richards & Summer Institute of Linguistics. N.d. Yindjibarndi dictionary. Australian Institute of Aboriginal and Torres Strait Islander Studies, Australian Indigenous Languages Collection. ASEDA 0297. Canberra. `http://aiatsis.gov.au/sites/default/files/catalogue_resources/0297_access.zip`

Yinhawangka

Wangka Maya Pilbara Aboriginal Language Centre. 2008b. *Yinhawangka dictionary: English-Yinhawangka wordlist and topical wordlists 2008: draft 1*. South Hedland, WA, Australia: Wangka Maya Pilbara Aboriginal Language Centre. 92 pp.

Yintyingka

Jean-Christophe Verstraete & Bruce Rigsby. 2015. *A grammar and lexicon of Yintyingka* (Pacific Linguistics 648). Boston, Berlin: De Gruyter Mouton. 414 pp. `https://doi.org/10.1515/9781614519003`

Yir Yoront

Barry J. Alpher. 1991. *Yir-Yoront lexicon: Sketch and dictionary of an Australian language*. Vol. 6 (Trends in Linguistics: Documentation). Berlin, New York: Mouton de Gruyter

Yorta Yorta

CHIRILA source: CHIRILA/v1/bowmor99

Heather Bowe & Stephen Morey. 1999. *The Yorta Yorta (Bangerang) language of the Murray Goulburn including Yabula Yabula* (Pacific Linguistics Series C 154). Canberra: Pacific Linguistics. 286 pp.

Yulparija

Kevin McKelson. 1989b. Yulparija. Australian Institute of Aboriginal and Torres Strait Islander Studies, Australian Indigenous Languages Collection. ASEDA 0032. Canberra. http://aiatsis.gov.au/sites/default/files/catalogue_resources/0032_access.zip

Yuwaalaraay

CHIRILA source: CHIRILA/v1/ash03

Anna Ash, John Giacon & Amanda Lissarrague. 2003. *Gamilaraay, Yuwaalaraay & Yuwaalayaay dictionary*. Alice Springs, NT, Australia: IAD Press. 344 pp.

Yuwaliyaay

CHIRILA source: CHIRILA/v1/ash03

Anna Ash, John Giacon & Amanda Lissarrague. 2003. *Gamilaraay, Yuwaalaraay & Yuwaalayaay dictionary*. Alice Springs, NT, Australia: IAD Press. 344 pp.