Phonotactics in historical linguistics:
Quantitative interrogation of a novel data source
Abstract

Jayden Macklin-Cordes, The University of Queensland, 2020

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.