PHYLOGENY IN PHONOLOGY: TAI SOUND SYSTEMS ENCODE THEIR EVOLUTIONARY HISTORY

RIKKER DOCKUM *1

*Corresponding Author: rikker.dockum@yale.edu
1Department of Linguistics, Yale University, New Haven, CT, USA

It is well established that the phonology of modern languages encodes useful data for studying the history of languages and their genealogical relationships to one another. The study of sound change is a foundation of traditional historical linguistics, particularly the linguistic comparative method. However, the utility of phonological systems for computational phylogenetic study has been more restricted. Phonology has typically been the means to an end: traditional comparative analysis enables coding a lexical dataset for cognacy. But once coded in this manner, the specifics of the phonology no longer factor into the quantitative analysis. This paper uses Tai languages to demonstrate that the facts of phonological systems themselves encode phylogenetic signal directly, suitable for use in quantitative historical analysis.

This study extends findings in previous by e.g. Macklin-Cordes (2015) and Macklin-Cordes and Round (2015). In this study, lexical data from Hudak (2008) comprising 1,159 cognate sets covering 20 Tai lects were used to generate datasets of two broad types: (1) binary data, encoding the presence or absence of phones and biphones in the lexicon of each language under examination, and (2) continuous data, which encodes phone frequency and biphone Markov chain transition probability (Ching & Ng 2006) in each language. Together these form a phonotactic profile for each language. Only those phones and biphones that exhibit variation in the overall dataset were included in the statistical analyses. The hypothesis underlying this type of data is that the more closely related two languages are to one another, the more similar their phonotactic profile will be. The phylogenetic tree required by these statistical tests was adapted from Chamberlain (1975).

This paper is distributed under a Creative Commons CC-BY-ND license.
DOI:10.12775/3991-1.021
Three tests are used to examine phylogenetic signal in the datasets described: 
*D test* (Fritz & Purvis 2010), Blomberg’s *K* (Blomberg et al 2003), and 
*NeighborNet* (Bryant & Moulton 2004), which also involves the calculation of a 
delta-score (Holland et al 2002), and the mean Q-residual (Gray et al 2010). The 
Density plots of *D* and *K* values for phones and biphones are in Figures 1 to 4.

**Figures 1 and 2.** Density plots of *D* for phoneme and biphone presence/absence data.

**Figures 3 and 4.** Density plots of *K* for phoneme frequency and biphone transition probability.

The results of these statistical tests confirm findings in Macklin-Cordes (2015) of phylogenetic signal in the “high-resolution” phonotactic profiles of 
languages. They also extend those findings by detecting strong signal in binary 
presence/absence data where that study was unable to do so. The results of the 
present study are of interest to linguists generally in the ongoing work of 
developing and testing phylogenetic methods of linguistic analysis. While the 
relative difficulty of using the traditional linguistic comparative method with 
Australian languages makes phylogenetic tools especially attractive and useful, 
the demonstrated results with the Tai data also shows the potential utility of these 
methods in other language families where traditional methods already have 
traction. This thus serves as a model for the application of these tests to language 
families and geographical regions in need of improved language classification 
throughout the world.
References


