PHYLOGENY IN PHONOLOGY: TAI SOUND SYSTEMS ENCODE THEIR EVOLUTIONARY HISTORY

RIKKER DOCKUM [*1]

[*]Corresponding Author: rikker.dockum@yale.edu
[1]Department 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).

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.
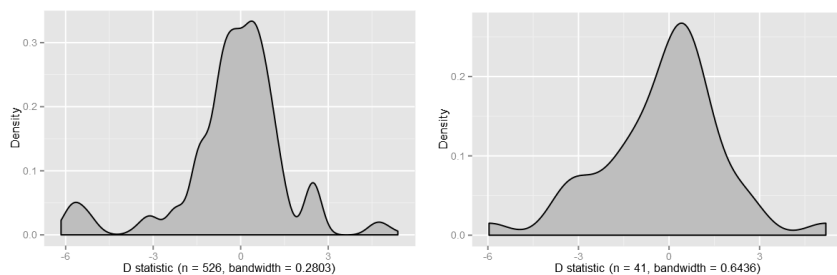
Figures 1 and 2. Density plots of *D* for phoneme and biphone presence/absence data.
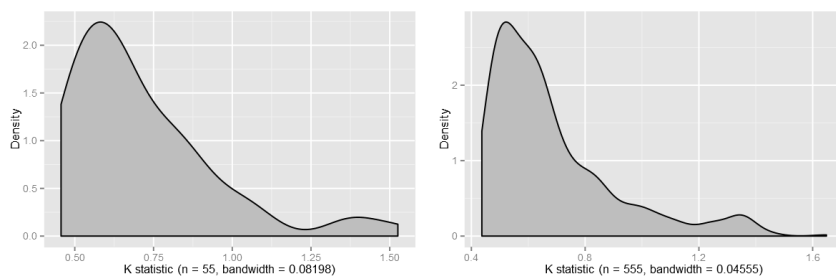
Figures 3 and 4. Density plots of *K* for phoneme frequency and biphone transition probability.

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

Blomberg, S. P., T. Garland, & A. R. Ives. (2003). Testing for Phylogenetic Signal in Comparative Data: Behavioral Traits Are More Labile. *Evolution* 57(4): 717–745.

Bryant, D., & Moulton, V. (2004). Neighbor-Net: An Agglomerative Method for the Construction of Phylogenetic Networks. *Molecular biology and evolution* 21(2): 255–265.

Fritz, S. A., & Purvis, A. (2010). Selectivity in Mammalian Extinction Risk and Threat Types: A New Measure of Phylogenetic Strength in Binary Traits. *Conservation Biology* 24(4): 1042–1051.

Gray, R. D., Bryant, D., & Greenhill, S. J. (2010). On the shape and fabric of human history. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 365(1559): 3923-3933.

Holland, B. R., Huber, K. T., Dress, A., & Moulton, V. (2002). Delta Plots: A Tool for Analyzing Phylogenetic Distance Data. *Molecular biology and evolution* 19(12): 2051–2059.

Hudak, Thomas John. (2008). *William J. Gedney's Comparative Tai Sourcebook.* Oceanic Linguistics Special Publication No. 34. Honolulu: University of Hawai'i Press.

Macklin-Cordes, Jayden. (2015). *Phylogeny and Phonotactics: Quantifying Historical Signal in Sequences of Sound*. B.A. thesis. University of Queensland.

Macklin-Cordes, Jayden & Round, Erich. (2015). High-definition phonotactics reflect linguistic pasts. Proceedings of the 6th Conference on Quantitative Investigations in Theoretical Linguistics, Tubingen, Germany.