Phylogeny in phonology: how Tai sound systems encode their past
Rikker Dockum, Yale University

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 a means to the end: traditional comparative analysis enables coding a dataset for lexical cognacy. But once coded in this manner, the specifics of the phonology no longer factor into the analysis. This paper uses Tai languages to examine how the facts of phonological systems themselves encode phylogenetic signal, and to evaluate how phonologically-focused methods compare with lexically-focused analyses.

This study builds on work by Macklin-Cordes (2015) and Macklin-Cordes and Round (2015), which drew their data from the CHIRILA database of Australian languages (Bowern 2016). Data used in this study are of two broad types: binary data, encoding the presence or absence of phones and biphones in the lexicon of each language under examination, and continuous data, which encodes the frequency with which each phone and biphone appears in the lexicons. Tai lexical data are from Hudak (2008). For the binary data, only those phones and biphones that exhibit variation in the dataset were used for study, as summarized in Table 1.

<table>
<thead>
<tr>
<th>Datatype</th>
<th>Lects</th>
<th>Total</th>
<th>w/Variation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phonemes</td>
<td>20</td>
<td>54</td>
<td>41</td>
</tr>
<tr>
<td>Biphones</td>
<td>20</td>
<td>555</td>
<td>526</td>
</tr>
</tbody>
</table>

Table 1. Variation in phoneme and biphone data (binary)

The first type of continuous data extracted from the lexicons in this study is frequency data for phonemes found in each language. The hypothesis underlying this type of data is that two identical languages would share both the same phonology and the same lexicon, and thus the closer two languages are in both phonemes present and the distribution of those phonemes across their lexicons, the closer those languages are to each other genealogically. Distribution of a phoneme across the dataset would be expected to be more phylogenetically informative than simple binary data, as two languages may share a phoneme, but it may be a core phoneme with high functional load in one language, but a marginal phoneme found only in a few loanwords in the other language.

The second type of continuous data is phoneme transition probabilities. Following Macklin-Cordes (2015), these are computed as Markov chain transition probabilities, which provide a more robust representation of the phonotactics of a language than either phoneme probability, which does not consider the environment in which phonemes appear, or simple binary presence/absence information about the transitions, which does not take into account their likelihood of appearing in a given lexicon. Thus, similar to the discussion of phoneme probability data above, it is hypothesized that the closer two languages are in sharing a profile of Markov chain transition probabilities, the genealogically closer those two languages are.

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). Sample density plots of $D$ and $K$ tests for biphones are given in Figures 1 and 2.
Figures 1 and 2. Density plots of D (binary) and K (continuous) scores for Tai biphones.

The results of these statistical tests confirm findings in Macklin-Cordes (2015) of phylogenetic signal in the phonotactics of language, but also affirms the presence of some detectable signal in areas 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