Phylogeny in Phonology: How Tai Sound Systems Encode Their Past

Rikker Dockum
Yale University

1 Introduction

The study of sound change is foundational to traditional historical linguistics, particularly the linguistic comparative method. As such, it is well established that the phonology of modern languages encodes useful data for studying the history of those languages, and their genetic relationships to one another. However, the utility of phonological systems for computational phylogenetic study has been more restricted. Phonology has typically been the means to the end, as traditional analysis enables coding for lexical cognacy in a dataset. Once coded for lexical cognacy through the comparative method, the particular sounds involved, indeed the phonology entirely, no longer factor into the analysis.

This study examines whether the phoneme inventories and phonotactic profiles of a set of languages themselves contain phylogenetic signal detectable using established statistical tests. This adds to the small but growing body of work on the use of phonological traits in computational phylogenetics for linguistics, including Macklin-Cordes (2015) and Macklin-Cordes and Round (2015). These two works explore phylogenetic signal contained in the phonotactics of the Ngumpin-Yapa languages, a 10-language subgroup of the Pama-Nyungan language family, spoken in the Pilbara region of Western Australia. The present study confirms and extends the findings of that work, using a set of data from 20 lects of the Tai branch of the Kra-Dai language family. The study confirms several findings from that previous work, including a strong degree of signal in the more high-resolution continuous traits drawn from phoneme frequency and biphone transition probabilities. New findings from this study include a relatively strong phylogenetic signal detected in even coarse-grained phoneme and biphone presence/absence, which previous work was unable to do.

2 Data

The data for this study come from Hudak 2008. Compiled from extensive mid-20th century fieldwork by William J. Gedney, this source consists of 1,159 cognate sets covering 19 languages from the Tai subgroup of the Kra-Dai family.[1] One language, Saek, is also subdivided in the dataset into younger generation Saek and older generation Saek, for a total of 20 lects.[2] The total Tai dataset from Hudak is 14,609 lexical items, giving an average lexicon size of about 750 items. The fact that the data consist entirely of posited cognate sets, as opposed to raw lexicons, could be argued to be sampling bias that will predispose the dataset to a positive result. However, given that these methods are still in need of validation using phonological traits to begin with, a noise-free dataset gives the best chance at detecting phylogenetic signal. This dataset essentially puts the tests used in this study on equal footing with a linguist using the traditional comparative method: building a historical analysis around a group of cognate sets. And indeed, if signal cannot be found in such a subset of the lexicon, then it is quite unlikely that the results would prove fruitful on full lexicons, either.

Data that can be extracted from the raw lexical material falls into two broad types: binary data on the presence or absence of phonological traits in the languages, and continuous data on the distribution or probability of those traits in their languages. Each of these two categories is further divided into phoneme

---

[1] The definition of 'cognate' used by Gedney and Hudak requires remark: they have included all forms believed to be modern reflexes of a particular form in Proto-Tai, and not just modern forms that coincide precisely in modern lexical meaning. This usage aligns with what Michael et al (2015) term 'quasi-cognates'. They also argue that for quantitative phylogenetic applications, the use of quasi-cognates better hews to the assumptions of the traditional comparative method.

[2] See Appendix for a full list of languages in the datasets.

data, which looks at individual segments, and biphone data, which looks at how those segments combine. These traits were extracted from the lexicons using Python scripts that I modified from ones developed for Gasser and Bowern (2014), in conjunction with my own Python and R scripts.[3]

2.1   *Binary data*   The first type of binary data records the presence or absence of phonemes in each language within its respective family dataset, encoded as a matrix of ones and zeroes. Of the 54 phonemes in the Tai data, 41 showed variation. When necessary for the nature of one of the phylogenetic tests, phonemes found in every language of a given family were pruned from the dataset, as there is no phylogenetic signal in a trait that exhibits no variation. The binary data are summarized in Table 1.

Lects	Phonemes	
	Total	w/Variation
20	54	41

Table 1: Variation in Tai phoneme data (binary).

The second type of binary data regards the presence/absence of segmental transitions between each phoneme found in a given language. This is thus a rudimentary representation of the phonotactics of the language. Binary data are only generated for attested transitions, as the probability of any unattested transition is always zero. These data are summarized in Table 2.

Lects	Biphones	
	Total	w/Variation
20	555	526

Table 2: Variation in biphone transition data (binary).

2.2   *Continuous data*   The first type of continuous data that is extracted from the lexicons for 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 genetically. 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.

Phoneme frequencies are calculated language by language. As discussed in Gasser and Bowern (2014), there are two methods to measure phoneme frequency in a lexicon: (a) the quotient of the occurrences of a phoneme in a language and the total number of segments in that language, and (b) the quotient of the number of lexical items that a phoneme occurs in and the total number of lexical items in that language. Given the variable length of items in the lexicon and the potential for multiple instances of a phoneme within a lexical item, the first method is used here, and is also the method used by Macklin-Cordes (2015).

The second type of continuous data is phoneme transition probabilities. These are computed as Markov chain transition probabilities, according to the formula in Ching & Ng (2006:3):

$$P_{ij} = P(x_{n+1} = j \,/\, x_n = i)$$

The probability of the transition *ij* is the probability that the system will next transition to state (i.e. segment) *j*, given the current state (i.e. segment) *i*.

Markov chain transition probabilities 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

---

[3] Thanks to Aidan Kaplan for assistance with scripts.

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 genetically closer those two languages are.

3   Methods

This study uses three tests to examine phylogenetic signal in different types of data: *D* statistic, *NeighborNet*, and Blomberg's *K*. These are described below.

3.1   *D statistic*   First developed by Fritz & Purvis (2010), this test measures phylogenetic signal in binary traits. The formula for the *D* statistic is as follows (Fritz & Purvis 2010:1044):

$$D = (\Sigma d_{obs} - \Sigma d_o) / (\Sigma d_b - \Sigma d_r)$$

Macklin-Cordes provides a succinct explanation of this formula (2015:37): "*D* is calculated by summing the differences between sister tips and sister nodes together across the phylogeny… This sum of differences is then deducted from its expectation if the trait were distributed according to a Brownian model of evolution (Nunn 2011:111), and then divided by the expected sum under Brownian evolution less the expected sum of the phylogenetically random distribution… The resulting *D* statistic is then tested for statistical significance against two null hypotheses—one where tip values are phylogenetically 'clumped' ($D = 0$), and one where they are dispersed randomly ($D = 1$)." And as also noted, while 0 and 1 are the null hypotheses, these are not the limits of the *D* statistics, which can exceed those two values if the nodes in the phylogeny are clumped even more conservatively or distributed even more evenly (Fritz & Purvis 2010: 1045).

3.2   *NeighborNet*   A NeighborNet analysis (Bryant & Moulton 2004) is a graph of the connectedness of data in the dataset, though it does not distinguish between horizontal (contact) and vertical (genetic) signal. It does allow us to derive two additional statistics, however: the delta-score, introduced by Holland et al (2002), and the mean *Q*-residual (Gray et al 2010). These tell us about how "treelike" a given dataset is, and are calculated by considering all of the taxa in a dataset in groups of four. Each combination of four taxa is considered by its possible pair-wise combinations (i.e. with four taxa *a*, *b*, *c* and *d*, we can combine them as *ab+cd*, *ac+bd*, and *ad+bc*). The distances between the taxa is then summed according to these combinations and ordered largest to smallest. If we label these summed combinations as Σ1, Σ2, and Σ3, where Σ1 represents the largest sum and Σ3 the smallest, then we can express the delta statistic and the *Q*-residual as follows (Gray et al 2010:3926):

$(\Sigma_1 - \Sigma_2) / (\Sigma_1 - \Sigma_3)$ delta-score
$(\Sigma_1 - \Sigma_2)^2$ *Q*-residual

For both statistics, they will equal zero if the data is perfectly tree-like. Gray et al argue for the *Q*-residual as obscuring less of the signal than the delta-score (2010:3926).

3.3   *Blomberg's K*   While both *D* statistic and NeighborNet test binary data, the *K* statistic is a test for phylogenetic signal in continuous data proposed by Blomberg et al (2003), based on variances of phylogenetically independent contrasts (PICs). The PICs for a given trait are calculated by considering all values of the phylogeny tips for that trait pairwise, and taking the contrast of two tips divided by the square root of the branch length distance that separates those tips (Felsenstein 1985:8). These are then compared against the expected distribution under Brownian evolution.

4   Results and discussion

4.1   *D statistic results*   This study calculated a *D* score for two types of binary Tai data (phonemes and biphone transitions) using the *phylo.d* function of the R package caper (Orme et al 2013). The density plot of *D* values for the first type, Tai binary phoneme data, is presented in Figure 1.
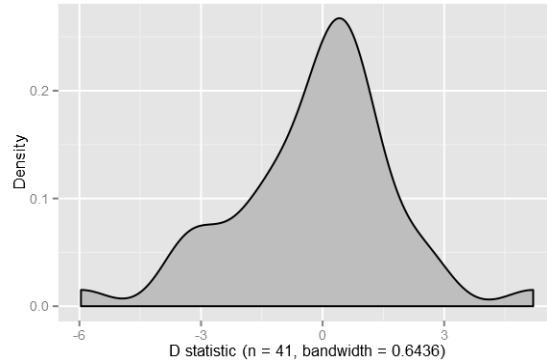
Figure 1: Density plot of *D* values for Tai phonemes (binary).

On binary phoneme data, Macklin-Cordes wrote that it is "unsurprising" that a list of present and absent segments is not phylogenetically informative, given the homogeneity of segmental inventories in Ngumpin-Yapa, which only had three testable segments to begin with (2015:69). His *D* results are given in Table 3.

	D	*p* (D=0)	*p* (D=1)
i:	2.659	0.113	0.735
u:	2.218	0.106	.0734
rR~	1.994	0.465	0.398
Mean **D**	2.29		
SD	0.33		

Table 3: Ngumpin-Yapa phoneme *D* scores and *p* values (Macklin-Cordes 2015:69)

The difference between the Tai data and the Ngumpin-Yapa data is striking. All three *D* scores from Ngumpin-Yapa are well above 1, the threshold indicating clustering characteristic of random distribution, yielding a mean *D* of 2.29. Contrast this with the *D* scores in Table 4.

	D	*p* (D=0)	*p* (D=1)		*D*	*p* (D=0)	*p* (D=1)
				š	0.372692	0.3994	0.1517
ɣ	-6.02706	0.9496	0	i:	0.392216	0.3822	0.1697
ʔy	-3.36769	0.6626	0.103	e:	0.397467	0.3835	0.166
ʔd	-3.3493	0.6708	0.0992	u:	0.402494	0.3739	0.1692
ʁ	-3.12975	0.9512	0	s	0.503545	0.4719	0.1407
ʔb	-3.01933	0.6591	0.097	ch	0.504791	0.3575	0.1413
θ	-2.38469	0.858	0.0104	ɔ:	0.549171	0.3264	0.1323
th	-2.31074	0.8556	0.0103	ɲ	0.636984	0.3037	0.1445
kh	-1.60053	0.9656	0	ɛ:	0.733751	0.2674	0.1902
ă	-1.54586	0.7805	0.0263	ɤ:	0.777366	0.2406	0.2204
r	-1.32693	0.8592	0.0039	o:	0.986454	0.151	0.4283
ɔ	-1.22613	0.8612	0.0019	ɯ:	0.997876	0.1492	0.4253
ph	-1.13276	0.861	0.0059	b	1.087499	0.1805	0.4069
ɬ	-0.73393	0.8379	0.0022	e	1.420117	0.2461	0.4974
ɰ	-0.72056	0.8483	0.002	d	1.495991	0.0934	0.6163
ɯ	-0.35617	0.6508	0.0406	c	1.517884	0.042	0.7487
u	-0.3087	0.6433	0.0375	k	2.460614	0.2287	0.3501
f	-0.24027	0.5792	0.0645	o	2.474653	0.229	0.3529
x	-0.23395	0.7031	0.0166	ð	2.506732	0.2301	0.352
v	-0.10507	0.618	0.045	sh	4.94437	0	0.7052
ʔw	0.184971	0.5429	0.1177	**Mean** *D*	-0.11911		
ɛ	0.18923	0.4991	0.046	**SD**	1.98		

Table 4: Tai phoneme *D* scores and *p* values, ordered by *D*.

As Table 4 shows, approximately half of the phonemes have *D* statistics below zero, and the mean *D* score for the Tai data is also negative. The dataset on average scores slightly better even than the null hypothesis that the trait distribution is the result of Brownian evolution. We can conclude from this that there is a strong phylogenetic signal in the Tai binary phoneme data alone. Now let us compare this with the *D* statistic for Tai biphone transitions, given as a density plot in Figure 2.
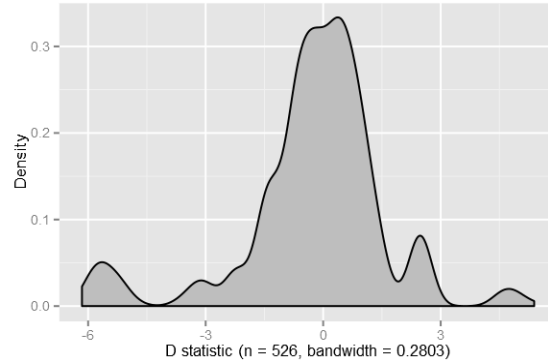

Figure 2: Density plot of *D* values for Tai biphone transitions (binary)

By comparison, Ngumpin-Yapa data showed a mean *D* for biphone transitions of 0.79 with standard deviation of 2.8 (Macklin-Cordes 2015:80), showing no significant detectable phylogenetic signal (2015:79). To provide a more granular look at the data from the Tai side, the twenty largest and smallest *D* scores for Tai biphones, from the total of more than 500 traits, are given in Table 5.

	D	*p* (D=0)	*p* (D=1)	…	…	…	…
rɣ:	-6.15479292	0.9522	0	ɣɯ	2.541741	0.2326	0.3496
hă	-5.977343211	0.953	0	ʔa:	2.55298	0.2327	0.356
ɣa:	-5.955821874	0.9483	0	iŋ	2.559538	0.2364	0.3498
re:	-5.925654008	0.9488	0	at	2.567858	0.2328	0.3538
ɛl	-5.919286765	0.8326	0	ɣa:	2.583812	0.2273	0.3543
wɣ:	-5.8423661	0.9517	0	#y	2.586476	0.235	0.349
rɯ:	-5.81630465	0.9468	0	#ð	2.623106	0.2288	0.3532
#ɣ	-5.812017916	0.9491	0	pa	2.682555	0.228	0.3621
u:l	-5.675415323	0.8366	0	#t	2.695217	0.231	0.3606
ă#	-5.672371502	0.9479	0	#k	2.742947	0.2306	0.3603
ɣɔ:	-5.654050941	0.9467	0	oʔ	4.466114	0	0.696
ɲi:	-5.636265478	0.947	0	aʔ	4.554462	0	0.7024
xe:	-5.594470499	0.9492	0.0068	ɣp	4.572078	0	0.6949
a:l	-5.579237163	0.8313	0	eʔ	4.637947	0	0.6944
xo:	-5.56562961	0.9524	0.0074	#sh	4.673463	0	0.7015
xi:	-5.546468993	0.9491	0.0062	shɯ	4.814715	0	0.6994
il	-5.494735837	0.8331	0	shɣ	4.876179	0	0.6996
ɔ:l	-5.39947864	0.8344	0	shi	4.899288	0	0.7051
al	-5.387467228	0.8361	0	ɣʔ	4.905214	0	0.7
sɣ:	-5.378904613	0.9472	0.0058	bɣ	4.933594	0	0.7016
…	…	…	…	**Mean *D***	-0.23937		
				SD	1.86		

Table 5: Twenty largest and smallest *D* scores for Tai biphone transitions, ordered by *D*.

From Table 5 we can again see that as with the binary phoneme data, the binary biphone transition data has a mean *D* value less than zero and (if we look at the ellipsed set of *D* scores) that fully 267 of the 526 observed biphone transitions are below zero, with many more above zero but very close to it. This test would

suggests strong phylogenetic signal in the binary biphone transition data, which raises the question of whether there was simply insufficient variation in the language sample used by Macklin-Cordes, rather than a general lack of useful signal in this type of data.

4.2 *NeighborNet results*   In the next test, three types of NeighborNet graphs were produced from nexus files of the Tai data. These three NeighborNet graphs are for Tai phonemes, Tai biphone transitions, and more traditional Tai cognate analysis, since the Hudak data is already organized into proposed cognate sets. The delta-scores and *Q*-residuals for each dataset are presented in Table 6.

Dataset	*Delta*	*Q-residual*
Tai phonemes	0.3115	0.03942
Tai biphone transitions	0.2988	0.02615
Tai cognates	0.2808	0.04088

Table 6: Variation in segmental transition data (binary).

The proper interpretation of delta-scores and *Q*-residuals is not an entirely straightforward matter, but for comparison, Gray et al describe a dataset with delta-score of 0.29 and *Q*-residual of 0.05 as "moderately tree-like" and delta-score of 0.41 and *Q*-residual of 0.02 as "strikingly non-tree like" (2010:3926-3927). Given these characterizations, all three of the NeighborNet graphs produced here for Tai data fall close to the "moderately tree-like" category. The first of the NeighborNet graphs is given in Figure 3.
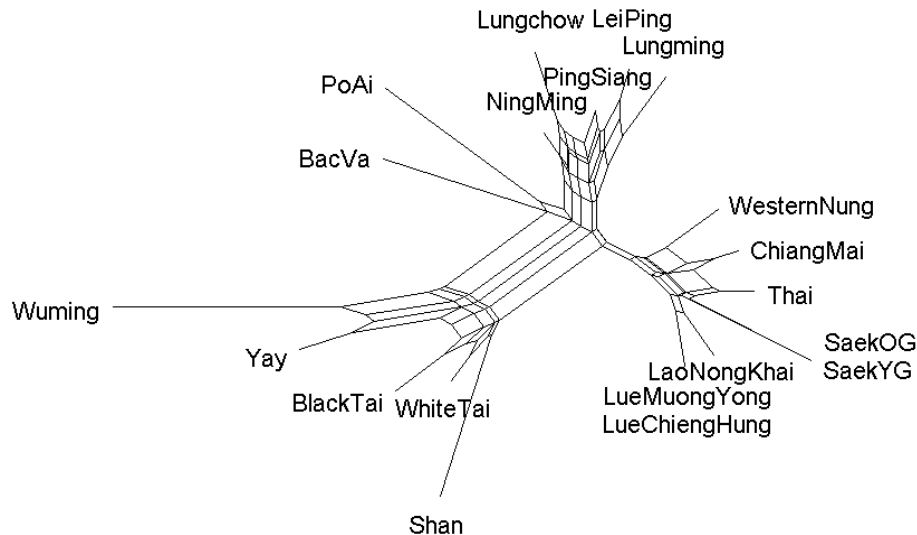

Figure 3: NeighborNet graph of Tai phoneme segments (binary).

For comparison, a traditional Tai family tree, modified from Hudak (2008), is given in Figure 4.[4] Figure 3 above picks out some clusters that closely match the tree, including five of the seven languages from the Central Tai branch in one cluster, five of the eight languages of Southwestern Tai mixed in another cluster with three Northern Tai languages. The other three Southwestern Tai languages, Black Tai, White Tai and Shan, are also close together, together with two other languages from Northern Tai. Assuming that the tree in Figure 4 is accurate, this would seem to indicate two subgroups each for Southwestern and Northern Tai, with horizontal transmission between the two groups.

---

[4] While this tree certainly does not represent the state of the art in Tai subgrouping, something still very much in flux, the three-branch Northern, Central, and Southwestern Tai tree, and variations on it, have been the most commonly cited classification for several decades. A new lexical phylogenetic subgrouping is currently in preparation by the author.
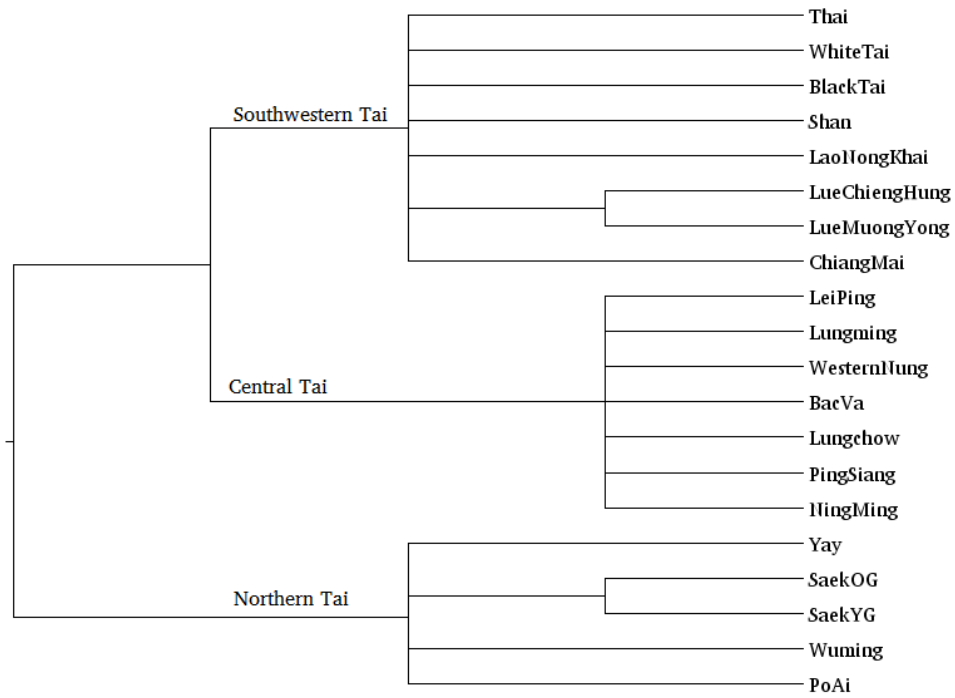
Figure 4: Traditional family tree of Tai languages (modified from Hudak 2008).

Next, compare the tree in Figure 4 against Figure 5, a NeighborNet graph of Tai segmental transitions.
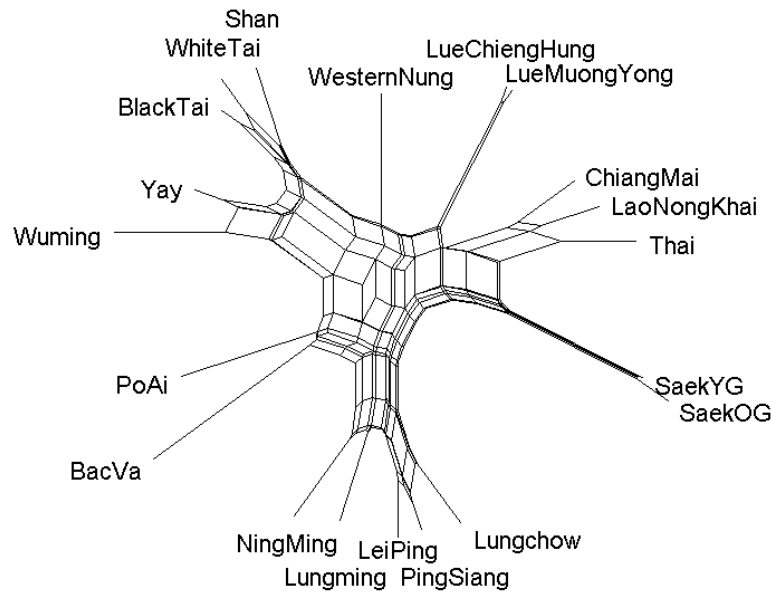

Figure 5: NeighborNet graph of Tai segmental transitions (binary).

In Figure 5 the NeighborNet has some clusters that more closely resemble the portions of the Tai tree, but in many ways picks out the same groupings as the binary phoneme data. For instance, Thai, Chiang Mai, and Lao Nong Khai cluster most closely together, which is expected given they are not only all from the

Southwestern branch, but are all in intense contact with one another due to all being spoken in Thailand and each having millions of speakers. The two Lue varieties, also Southwestern Tai, are not far from the Thai-Lao-Chiang Mai cluster, but are much more clearly grouped distinctly as a pair than in the previous graph. Once again five of the seven Central Tai languages are clearly clustered together at the bottom of the graph. And the same set of five languages spanning both Southwestern Tai (Shan, White Tai, Black Tai) and Northern Tai (Yay, Wuming) are also grouped closely together.

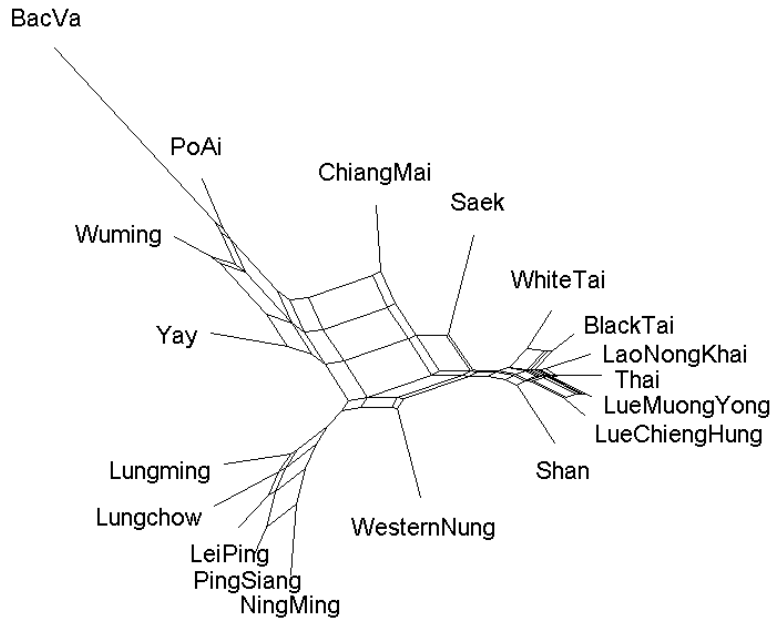Finally, the NeighborNet graph for shared lexical cognate data is given in Figure 6.

Figure 6: NeighborNet graph of Tai lexical cognates.

Here we see perhaps the best representation of at least the Southwestern Tai branch of our reference tree yet. Seven of the eight Southwestern Tai language form an obvious cluster, with Chiang Mai being the surprising outlier, though the fact that it does not cluster with anything else may indicate lexical innovation in that language, but at the very least it is not misgrouped with anything else. Yet again five of the seven Central Tai languages cluster together, but of the two missing, Western Nung and Bac Va, neither seems to cluster with any other language, either. The fact that we don't see Northern Tai languages grouping with Southwestern Tai languages in the way we saw before would seem to indicate that there has been horizontal phonological transfer between different Southwestern and Northern Tai languages, but often without lexical replacement.

4.3   *Blomberg's K results*   The final tests performed dealt not with binary data, but with continuous Tai data of both subtypes: phoneme segment frequency and Markov chain transition probabilities. K statistics were calculated using the *multiPhylosignal* function of the R package picante (Kembel et al 2010).

The K test requires at least some variation in each trait examined, so traits which showed no variation across their respective datasets were dropped for this test. The mean K for each language family and data type are given in Table 7.

	Probabilities	
	Phonemes	*Transitions*
Tai	0.71	0.68

Table 7: Mean *K* values for four sets of continuous data.

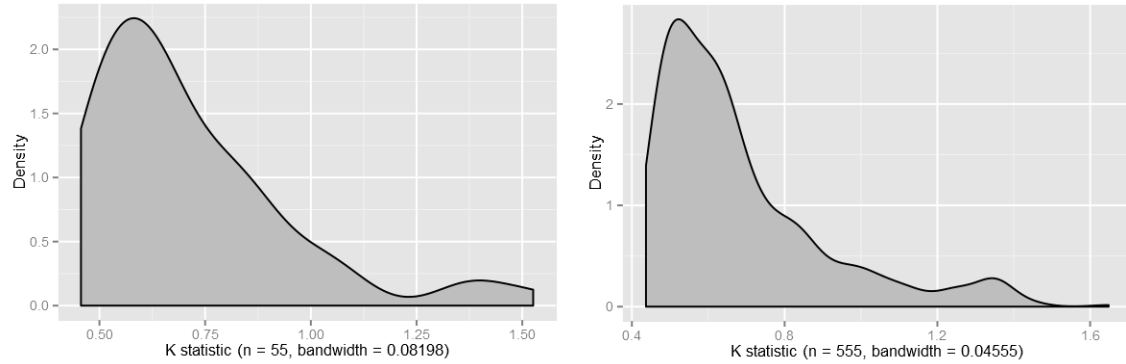Density plots for the *K* values of these two datasets are presented in Figure 7.

Figure 7: Density plot of *K* statistics for Tai phoneme frequency (left) and biphone transitions (right).

In testing with Blomberg's *K*, a score of zero indicates a fully random trait distribution, while a score of one indicates trait distribution as expected under Brownian evolution. For comparison, Ngumpin-Yapa phone probabilities had a mean *K* of 0.9, and biphone transitions 0.87 (Macklin-Cordes 2015:89). The tests on the Tai data return a lower *K* on average, but still a strong indication of phylogenetic signal over, and indeed many individual traits have very high *K* scores.

5 Conclusion

The preceding three tests for phylogenetic signal in Tai data produced a variety of positive results. In some areas, such as the *D* test on binary data, data from Ngumpin-Yapa proved too homogeneous to identify a phylogenetic signal, whereas with the Tai data the presence of signal was clear. In other areas, like the *K* test of continuous phoneme frequency and biphone transition probabilities, Ngumpin-Yapa produced *K* values indicating a stronger phylogenetic signal on average than the corresponding Tai data, but the set overal still showed signal, and many individual Tai traits had very strong signal.

For NeighborNet tests, the two types of Tai phonological data were compared against data coded for lexical cognacy. Better performance of the cognate data is to be expected, given that that kind of data is the core of traditional language classification generally. However, there were some clusters, notably Central Tai and parts of Southwestern Tai, that were found across all three NeighborNet analyses, indicating that contra Macklin-Cordes' dismissal of binary phoneme and transition data, given sufficient phonological variation in a set of related languages, some language clusters are recoverable from both coarse- and fine-grained binary phonological data.

Additional tests are available to be performed to test for phylogenetic signal in this data, including Abouheif's (1999) *Cmean* and Pagel's *lambda* (1999) among other possibilities. In some cases these tests are still being refined themselves, and both their statistical power and the correct interpretation of their output is still also under development. As such, comparing and contrasting results from multiple tests on novel datasets in the manner done in this study is an important part of the process of methodological refinement.

This study confirms findings of e.g. Macklin-Cordes (2015) of phylogenetic signal in the phonotactics of language, but also affirms the presence of detectable signal in some areas where other studies have been unable to do so. By applying these methods to a Tai dataset, the results of these tests are made more robust, as they are shown to be useful for data from additional language families. As such, 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 significant traction. The Tai clade 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.

6   Appendix: List of Tai lects used in the study

Black Tai   Tai of Lungchow
Lao dialect of Nong Khai   Tai of Lungming
Lue of Chieng Hung   Tai of Ning Ming
Lue of Muong Yong   Tai of Piang Siang
Saek (Old Generation)   Tai of Po-ai
Saek (Young Generation)   Tai of Western Nung
Shan   Tai of Wuming
Tai of Bac Va   Thai
Tai of Chiang Mai   White Tai
Tai of Lei Ping   Yay

References

Abouheif, E. (1999). A Method for Testing the Assumption of Phylogenetic Independence in Comparative Data. *Evolutionary Ecology Research* 1(8): 895–909

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

Bowern, Claire (2016). Chirila: Contemporary and Historical Resources for the Indigenous Languages of Australia. *Language Documentation and Conservation* 10:1–44.

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

Ching, W. K., and M. K. Ng (2006). *Markov Chains: Models, Algorithms and Applications*. Springer.

Felsenstein, J. (1985). Phylogenies and the comparative method. *American Naturalist*: 1–15.

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

Gasser, Emily, and Claire Bowern (2014). Revisiting Phonological Generalizations in Australian Languages. *Proceedings of the Annual Meetings on Phonology*.

Gray, R. D., D. Bryant, S. J. Greenhill (2010). On the shape and fabric of human history. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 365(1559): 3923–3933. doi:10.1098/rstb.2010.0162

Holland, B. R., K. T. Huber, A. Dress, and V. Moulton (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.

Kembel, S. W., P. D. Cowan, M. R. Helmus, W. K. Cornwell, H. Morlon, D. Ackerly, et al. (2010). Picante: R tools for integrating phylogenies and ecology. *Bioinformatics* 26(11): 1463–1464.

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

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

Michael, Lev, Natalia Chousou-Polydouri, Zachary O'Hagan, Keith Bartolomei, Diamantis Sellis, Emily Clem, and Erin Donnelly (2015). A Bayesian Phylogenetic Internal Classification of the Tupí-Guaraní Family. Talk presented at the 92nd Annual Meeting of the Linguistic Society of America. Portland, Oregon.

Nunn, C. L. (2011). *The Comparative Approach in Evolutionary Anthropology and Biology*. Chicago: The University of Chicago Press.

Orme, David, Rob Freckleton, Gavin Thomas, Thomas Petzoldt, Susanne Fritz, Nick Isaac and Will Pearse (2013). caper: Comparative Analyses of Phylogenetics and Evolution in R. R package version 0.5.2. https://CRAN.R-project.org/package=caper

Pagel, Mark (1999). Inferring the historical patterns of biological evolution. *Nature* 401:877–884.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.