Project Discussion: The Austro-Tai Hypothesis

Matthias Gerner

City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong

Abstract

By using phylogenetic computational methods, this project aims at testing the Austro-Tai Hypothesis which proposes a genetic relation between Tai-Kadai and Austronesian languages (Benedict 1942, 1975). We test the Austro-Tai Hypothesis at two levels. We first measure the degree of genetic relatedness between Tai-Kadai and Taiwanese (i.e. Formosan) languages. Austronesian languages are believed to have originated from Taiwan. In a short lapse of time, between 4,000 and 2,000 BC, Austronesian speakers rapidly moved throughout the Pacific. This migration wave is called by Austronesian specialists the “Express Train to Polynesia” (Gray & Jordan, 2000). At a second level, we test a hypothesis of population expansion, the *Northeast-to-Southwest Hypothesis*, which is the idea that Tai-Kadai groups migrated from the Northeast (Taiwan, Guangdong, Guangxi) to the Southwest (Malay Peninsula). This migration movement would be the sole model compatible with the Austronesian Hypothesis.

Keywords: Tai-Kadai, Austronesian, The Austro-Tai Hypothesis;

1. Previous phylogenetic work on Tai-Kadai languages

‘Tai-Kadai’ is used as provisional term. The family name is the subject of ongoing discussion over the past twenty years.

The Tai-Kadai (or Kadai) languages are spoken in a large area of Southeast Asia extending from Guizhou Province (China) in the North to half way down the Malay Peninsula. Westernmost Tai-Kadai languages are Shan dialects of Myanmar. In the East, we find Zhuang dialects spoken in Guangdong Province (see Map 2 below).

The term ‘Kadai’ was coined by Benedict in the 1940s from the Gelao prefix *ka-* for *man* and from *dai*, one of the selfnames of the Hlai living on Hainan island (China). It originally accounted for non-Tai groups outside of Thailand. Since then the label Kadai has undergone several transformations. Four types of internal classification of Tai-Kadai languages were proposed (see Table 1 below).

Benedict (1975) and Edmondson & Solnit (1988) arranged the Kadai family into three groups: Kam-Tai, Hlai and Geyang (a residual group of lesser known languages). They later revised this classification for Kam-Tai by dissociating Tai and Kam-Sui (Edmondson & Solnit, 1997; Chamberlain, 1997; Diller, Edmondson & Luo, 2008).

Subsequent classifications detailed the internal structure of the other subgroups of the Kadai family such as the Kra group, a new name given to the Geyang group (Ostapirat, 2000) or the Hlai group (Norquest, 2007).

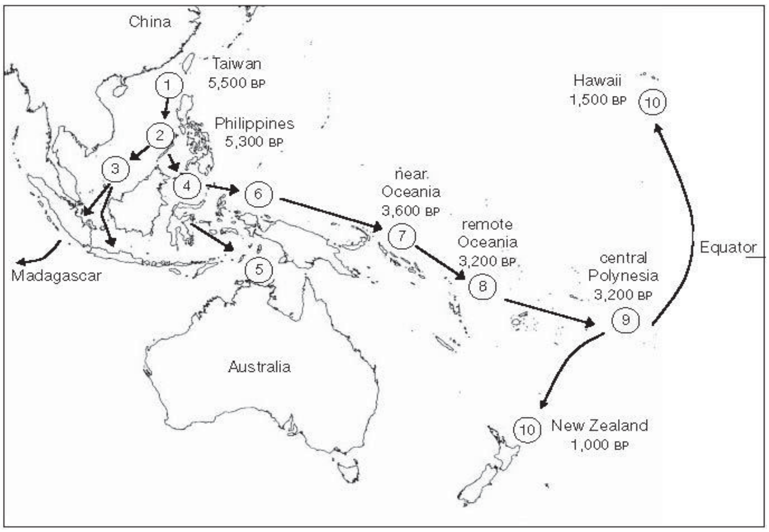
Ostapirat’s “Krai-Dai” grouping provides an elaborate classification of 17 Gelao dialects (previously called Geyang or Kadai). Ospirat’s classification is based on three shared innovations: partition of implosives, loss of labial endings and lexical innovations. Ostapirat (2000:15) adopts the term Kra as the reconstructed selfname of the Gelao ancestor. He discards the term Kadai and Geyang. Norquest’s dissertation is a phonological reconstruction of Proto-Hlai on the basis of 12 Modern Hlai varieties all spoken on the island of Hainan (Norquest, 2007).

All previous phylogenetic work on Kadai languages was undertaken by linguists using the comparative method to a greater or lesser degree (Fox 1995:17-36; Trask 1996: 208-216). No computational phylogenetic work was undertaken so far. An integrative approach is necessary as the reconstruction work previously done emphasized smaller segments of the Tai-Kadai languages (especially for Ostapirat and Norquest).

Kadai	Tai-Kadai	Kra-Dai	Kra-Dai
Benedict (1975) Edmondson & Solnit (1988)	Chamberlain (1997) Edmondson & Solnit (1997) Diller, Edmondson & Luo (2008)	Ostapirat (2000)	Norquest (2007)
Kam-Tai	Tai	Kra	Hlai
Lakkja, Biao	Northern	Southwestern	Bouhin (Heitu)
Kam-Sui	Northern Zhuang, Saek,	Western	Greater Hlai
Kam, Sui, Maonan,	E, Bouyei, Yay, Mène	Gelao, Lachi	Ha Em (Zhongsha)
Mulam, Then, Mak	Central	Southern	Central Hlai
Be	Nung, Tày,	Laha	East-Central Hlai
Tai	Southern Zhuang	Centraleast	Lauhut (Baoding)
Northern	Southwest	Central	Qi
Northern Zhuang, Saek	Thai, Lao, Shan,	Paha	Tongzha
E, Bouyei, Yay, Mène	B, R, W, Tai, Ahom	Eastern	Zandui
Central	Be	Buyang, Pubiao	Baoting
Nung, Tày,	Lakkja, Biao	Hlai (Li)	North-Central Hlai
Southern Zhuang	Kam-Sui	Kam-Tai	NWC Hlai
Southwest	Kam, Sui, Maonan,	Be	Cun
Thai, Lao, Shan,	Mulam, Then, Mak	Kam-Sui	Nadou
B, R, W, Tai, Ahom	Hlai (Li)	Kam, Sui, Maonan,	NEC Hlai
Hlai (Li)	Kadai	Mulam, Then, Mak	Meifu
Geyang	Gelao, Lachi, Buyang,	Tai	Changjiang
Gelao, Lachi, Buyang,	Pubiao, Yerong, Laha	Northern	Moyfaw
Pubiao, Yerong, Laha		North Zhuang, Saek,	Run
		E, Bouyei, Yay, Mène	Baisha
		Central	Yuanmen
		Nung, Tày,	Tai
		Southern Zhuang	Lakkja, Biao
		Southwest	Kam-Sui
		Thai, Lao, Shan,	Kra
		B, R, W, Tai, Ahom	

Table 1: The Tai-Kadai languages

The comparative method produces phylogenetic trees resulting from a chain of fuzzy decisions made by the researcher. In some cases, the results are difficult to verify and seem arbitrary.[1] Computational phylogenetic approaches calculate through several possible reconstructions and suggest one as the most likely one.

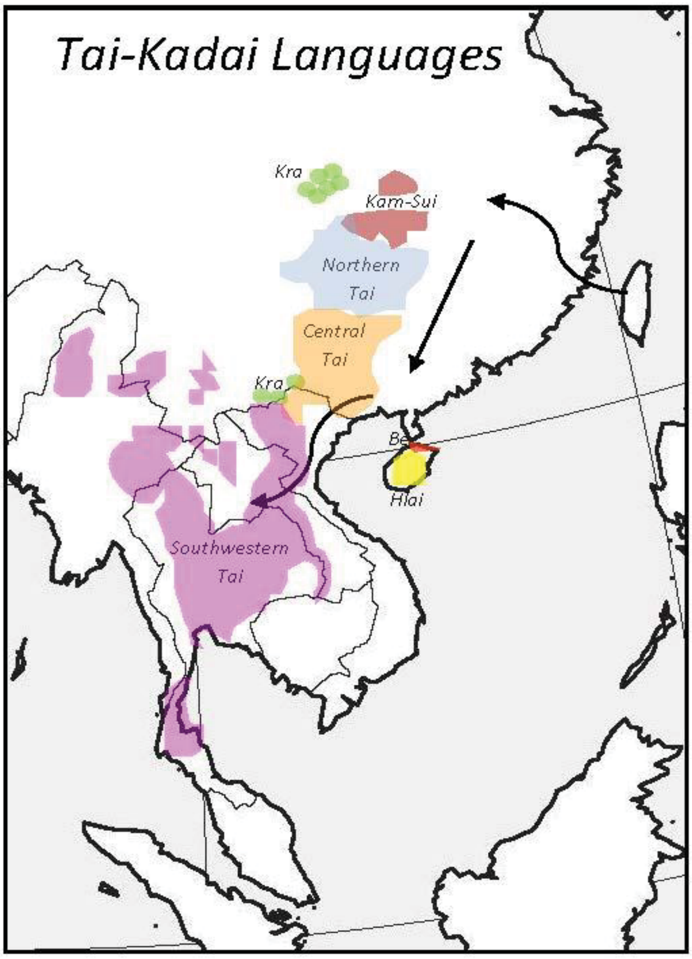


Map 1: The Austronesian colonization of the Pacific (Map from Gray & Jordan 2000:1053)

[1] See Sagart's reaction to Matisoff's reconstruction of Tibeto-Burman (Sagart, 2008).

The relation of the Tai-Kadai family to other language families in East Asia is controversial. Benedict (1942, 1975) related Tai-Kadai to Austronesian languages (*Austro-Tai Hypothesis*), but Austronesian linguists were skeptical about the reconstructions made and have characterized them as "too loose" (Ross 1994: 96). Austronesian languages are believed to have originated from Southeast China and Taiwan. In a short lapse of time (between 4,000 and 2,000 BC) Austronesian speakers rapidly moved throughout the Pacific. This migration wave is called by Austronesian specialists the "express train to Polynesia" (see Map 1). It was validated by computational phylogenetic methods (Gray & Jordan, 2000).

If Tai-Kadai languages are genetically related to the Austronesian phylum, the dispersion of Tai-Kadai languages must have occurred from the Northeast (China) to the Southwest (Malay Peninsula), see Map 2. The internal genetic relations of Tai-Kadai languages would reflect the direction of this migration. We call this idea the *Northeast to Southwest Hypothesis*.



Map 2: Geographic expansion of Tai-Kadai peoples
(Blank map is Courtesy Arizona Geographic Alliance, Terry Dorschied, cartographer)

In this project, we test the Austro-Tai Hypothesis by measuring the degree of genetic relatedness between Tai-Kadai and Formosan languages and by refuting or showing evidence for the Northeast-to-Southwest Hypothesis.

2. Underlying Datasets

About 90% of the following datasets are available and input in a Microsoft Access database:

- 54 Tai-Kadai languages: Wordlists of more than 2500 items of the basic vocabulary
- 7 Formosan languages: Wordlists of 2500 items of the basic vocabulary

3. Mathematical and computational methods

Phylogenetics was applied to linguistics about 10 years ago with two seminal papers: Gray & Jordan (2000) on a phylogenetic subgrouping of the Austronesian stock, and Gray & Atkinson (2003) on a phylogenetic inference method applied to the Indo-European family.

In linguistics, language features (sound shape, word order) represent the input data for phylogenetic methods in the same way DNA sequences supply the raw data in biology. Phylogenetics can be applied to linguistics in two ways. First, it allows establishing the evolutionary family tree for a set of contemporary languages. Second, phylogenetics can gauge the statistical probability, and thereby test hypotheses, of how language properties distribute over a given family tree.

In linguistics, the notion of *shared innovation* (Trask 1996: 182) is the single most important principle of language classification. The idea is that languages which did not undergo a particular linguistic change split off early from languages which do share the innovation. This concept is implemented in phylogenetics by the notion of *parsimony score of a character*.

We involve the following mathematical notions which are laid out in textbooks of statistics, genetics and computer science (e.g. Semple & Steel, 2003; Huson et al, 2010; Foulds & Graham 1982): *graph*, *X-tree* and *random variable*, *characters*, *parsimony score of character*, *most-parsimonious trees*, *Markov process*. Markov processes are applied to matrices of cognates and allow the reconstruction of phylogenetic trees that model ancestry trees of a group of languages. The algorithms that produce these trees use Bayesian inference methods such as the *Markov Chain Monte Carlo Sampling*. Available Software packages such as MrBayes implement these algorithms.

3.1 Establishment of cognate matrix

For each/most of the 2500 senses, we will establish cognate forms. Forms that are only differentiated by regular and attributable sound changes are cognate. They potentially descend from the same form in an ancestral language (Trask 1996: 205). The establishment of cognates for a set of related languages, if done manually, represents a time-consuming task. We program a half-automated procedure that allows human control over the decision which forms are cognate and which forms are probably not.

The basic syllable structure of Tai-Kadai languages is C(C)V(V)(C) discarding suprasegmental effects such as tones, aspiration and voicing. As consonants and vowels occupy coordinates in a 2-dimensional space whose two axes are *degree of openness* and *point of articulation*, we can encode each phoneme with a pair of numbers. To show how it works, we encode the degree of openness on a scale 1 (stop) – 8 (glide) – 9 (close vowel) – 12 (open vowel), and the point of articulation on a scale 1 (bilabial, apical vowel) – 9 (back vowel) – 11 (glottal). The following example illustrates the translation algorithm from CVC syllables into vectors of the $\mathbf{R}^6$ real vector space.

Language	*Province*	*County*	*‘wear’*	*Vector*
pu^{42}ʔjai^{42}	贵州省	贞丰县	tan^{35}	(1,4; 12,4; 0,4)
pu^{22}jəi^{11}	云南省	广南县	tən^{11}	(1,4; 10.5,5; 0,4)
vun^{213}tɕuːŋ^{22}	广西壮族自治区	南宁市	møːk^{44}	(0,1; 10,3; 1,8)

For $\mathbf{R}^6$, we employ the usual real metric which calculates the distance of two syllables in the sound space (understood as the multi-dimensional vector space $\mathbf{R}^6$).

‘wear’	tan^{35}	tən^{11}	møːk^{44}
tan^{35}	0	3.25	32
tən^{11}	3.25	0	31.25
møːk^{44}	32	31.25	0

This example shows that tan^{35} and tən^{11} are cognates, whereas mø:k^{44} is not cognate with the other two forms.

3.2 Family tree of the Tai-Kadai languages and relatedness to Formosan languages

Let *n* be the number of languages from which wordlists are fed into the database. Let *m* be the number of cognate forms found in the procedure mentioned in §3.1. We can establish the cognate matrix, in which the place (i,j) is 1 if the ith language exhibits the jth cognate and 0 otherwise. The resulting $n \times m$ matrix is the input matrix for the most-parsimonious tree algorithm. It is also the input of Bayesian Markov chain Monte Carlo methods in which it approximates the Bayesian posterior probability distribution of the trees. Bayesian methods allow incorporating the subgroupings of Tai-Kadai languages previously proposed (e.g.) as prior probability distribution. In a second run, the matrix can be expanded by the Formosan languages to calculate genetic relatedness.

3.3 Testing the Northeast-to-Southwest Hypothesis of population expansion

The Northeast-to-Southwest hypothesis claims that the Tai-Kadai populations expanded along an axis from the Northeast (Guizhou, Guanxi, Guangdong) to the Southwest (Malay Peninsula). For this geographical axis, we number the different stations that Tai-Kadai populations might have walked through. This will provide a sequence of characters which we submit to Bayesian methods of phylogeny.

References

[1]Benedict, Paul K. 1942. Thai, Kadai, and Indonesian: A new alignment in South-Eastern Asia. *American Anthropologist* 44:576-601.

[2]Benedict, Paul K. 1975. *Austro-Tai: language and culture*. New Haven: HRAF Press.

[3]Chamberlain, James R. 1998. The Origin of the Sek: Implications for Tai and Vietnamese History. *Journal of the Siam Society* 86.1 & 86.2: 27-48.

[4]Diller, Anthony, Jerold Edmondson and Yongxian Luo (eds.) (2008). *The Tai–Kadai Languages*. London: Routledge.

[5]Diller, Anthony, Jerold Edmondson and Luo Yongxian (eds.) (2008). “Chadong, a newly discovered Kam-Sui language in Northern Guangxi.” *The Tai-Kadai languages,* 596-620. London: Routledge.

[6]Foulds, L. R. and R. L. Graham. 1982. The Steiner Problem in phylogeny is NP-complete. *Advances in Applied Mathematics* 3:43-49.

[7]Fox, A. 1995. *Linguistic Reconstruction*. Oxford: Oxford University Press.

[8]Gray, Russell and Fiona Jordan. 2000. Language trees support the express-train sequence of Austronesian expansion. *Nature* 405:1052-1055.

[9]Gray, Russell & Quentin Atkinson. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* 426:435-439.

[10]Huson, Daniel, Regula Rupp and Celine Scornavacca. 2010. *Phylogenetic Networks*. Cambridge: Cambridge University Press.

[11]Li, Jenkuei Paul. 2004. *Selected Papers on Formosan Languages,* vol. 2. Taipei, Taiwan: Institute of Linguistics, Academia Sinica.

[12]Norquest, Peter 2007. *A Phonological Reconstruction of Proto-Hlai*. Ph.D Dissertation. Department of Anthropology, University of Arizona.

[13]Ostapirat, Weera. 2000. Proto-Kra. *Linguistics of the Tibeto-Burman Area* 23.1:1-251.

[14]Ross, Malcom D. 1995. Some current issues in Austronesian linguistics. In *Comparative Austronesian dictionary: An introduction to Austronesian studies*, Darrell T. Tryon (ed.), 45-120. Trends in Linguistics, Volume 1. Berlin and New York: Mouton de Gruyter.

[15]Sagart, Laurent. 2008. Reply to Matisoff on the Handbook of Proto-Tibeto-Burman: System and philosophy of Sino-Tibetan reconstruction. *Diachronica* 25:153-155.

[16]Semple, C. and M. Steel. 2003. *Phylogenetics*. Oxford: Oxford University Press.

[17]Trask, R. L. 1996. *Historical Linguistics*. London: Edward Arnolds Publishers.

[18]Tryon, Darrell T. (ed.) 1995. Trends in Linguistics, Volume 1-5. Berlin and New York: Mouton de Gruyter.