Notes on the Subdivisions in Kra

Jerold A. Edmondson
The University of Texas at Arlington, Arlington, TX 76019-0559 USA
Jerold.A. Edmondson@gmail.com/j.edmondson@sbcglobal.net

1. **Introduction**. This paper will focus on a few details of the Kra languages that have not yet been treated in the elegant and exhaustive account by Ostapirat (2000).[1] In particular, I will be singling out several members of this group: Laha Noong Lay and Laha Ta Mit, the Nung Ven (Anh), and the Gelao languages of Northern Vietnam and China. At the time of his writing, Ostapirat did not yet have access to all materials on these languages, because audio files and analysis of Laha were not yet available, and because Red Gelao (in Vietnam and in China) and Nung Ven had only very recently been discovered and a definitive study of Sanchong Gelao was still yet to be published. In situ study of these groups took place only in 1997-2000 as a part of an NSF-NEH grant to the Dr. Kenneth J. Gregerson and author, which included a number of expeditions to language sites in Lao Cai, Sơn La, Cao Bằng, Hà Giang, et al. to collect data on the minority languages of the border provinces of Vietnam. Li Jinfang found a small group of people in Malipo County of Yunnan Province speaking a nearly identical language to the Nung Ven. And Shen Yu May had not yet started work on Sanchong Gelao of Longlin County in Guangxi Province until after this time. Thus, the rationale for this paper is to consider Proto-Kra (Ostapirat 2000) in the light of new information about languages of this stock. In almost all cases the results serves only to strengthen its claims, as the following exposition will show.

2. **The Laha Language**. The Laha language occupies a special position among the Kra languages. It survives today in two separated communities 35 km apart on opposite sides of the Black River. According to speakers of Laha Noong Lay (LNL) on the south side, speakers of Laha Ta Mit (LTM) have not visited their village in living memory.

The time of division among this ethnicity is unknown, but the palpable differences in the sound and lexical structure today suggest a separation of some time. Laha Noong Lay (LNL) possesses as

[1] The research reported on here was supported by grants NEH RT-21754-95 from the National Endowment for the Humanities and by the grants SBR 9511285 and SBR 9729043 from the National Science Foundation to the author and Dr Kenneth J. Gregerson all entitled 'Languages of the Vietnam-China Borderlands'. I wish also to acknowledge the assistance of Profs. Kenneth J. Gregerson for various help and advice, as well as Nguyễn Văn Lợi, Hoàng Văn Ma, and To Văn Thang, who arranged and accompanied me on the field trips that led to the data and analysis here. I am also grateful to Yao Shih-ping for helping me compile the data, Dr. Nancy Rowe for help with statistics, and Ms Shen Yu-May MA for data on the Sanchong Gelao language. All reconstructions are from Ostapirat, Weera. (2000). Proto-Kra. LTBA 23.1.

one of its consonant codas a final *–l*, a rare feature in this language family and has a number of unique and archaic lexical forms; Laha Ta Mit (LTM) often shows disyllabic forms where LNL has monosyllables. Also, especially the tones LNL and LTM are in need of more attention. In Solntseva & Hoàng (1986) approximate tones values based on auditory impressions were given and in Edmondson & Gregerson (1997) there are errors and inconsistencies in some examples. This section is dedicated to clarifying those issues.

Our study of LNL with two native-speaker consultants (one of whom was also studied by Solnetseva & Hoàng) shows extensive contouring of the tone trajectories in this language. We can confirm the treatment of Ostapirat with the evidence for four basic tonal divisions A, B, C, and D in the parent language of Laha with a split conditioned by original voiceless and voiced initials, which resulted in Series 1 and Series 2 tones, as in Tai and Kam-Sui languages. The tone values I found are represented the scale-of-five system of Y. R. Chao (1931), as in (1).

(1) Tone values of Laha Noong Lay

	A	B	C	D
Series 1	534	34	14ʔ	14
Series 2	44ʔ	441	41ʔ	44

Note also the sharp glottal constriction syllable-finally in A2, C1, and C2, which leads to a shortened vowel nucleus and a rapid decline in pitch. D tones end in voiceless applosives /-p –t –k/.

The tone values for LTM show great similarity in shape and height to LNL.

(2) Tone values of Laha Ta Mit

	A	B	C	D
Series 1	343	53	53ʔ	33
Series 2	33	24	31ʔ	23

Comparing 1 and 2, the two Laha varieties have common features such as the shortening of syllable length in the C class. The value of the B tone in Ta Mit was difficult at first to determine, because, when the syllable begins with a voiced sound, then the difference between A1 and B1 categories was masked by the pitch depressing effects of the such voiced consonants.

Ostapirat (2000), reacting to differing values in Solntseva & Hoàng (1986) for the A category Series 1 tone, speculated that there might be a Tone A1', a slightly higher series of Tone 1 for voiceless aspirates/continuants initials, as is found in some Tai and Kam-Sui languages. I compared three repetitions each of the lexical items: (a) 'dog', 'pig', 'unhusked rice', 'ginger',and 'road' and (b) 'head louse', 'eye', 'grass', 'fire', and 'sunlight' from the Laha Noong Lay speakers. The first group had initials showing voiceless aspirates/continuants and the second group started with unaspirated voiceless stops. Tone 1 is realized with a fall-rise contour, so three measurements were taken; at onset, at the medial and offset points in the trajectory. Values are seen in 3.

(3)

Voiceless aspirated/continuant	Voiceless stops	(values in Semitones)
Onset 55.126 ± 1.432813	55.08667 ± 2.753874	
Medial 47.686 ± 1.83154	47.35267 ± 3.662089	
Offset 49.342 ± 1.872436	49.73467 ± 4.107142	

A Type 3 Fixed Effects Test was applied to these data and the least square means of difference were so small that they could not be detected. That indicates that for the speakers I studied there does not appear to be a significant height difference of pitch for these two categories.

3. **The Gelao Language**. One of the most important contributions of Ostapirat (2000) is the division of the subgroups within Gelao into: (a) Central, represented by Wanzizhai (W) near Anshun in Guizhou Province, China, (b) Southwestern, represented by Laozhai (LZ), and (c) Northern, represented by Qiaoshang (Q). Since I have obtained data from several other types of Gelao, I have added those to the analysis in Ostapirat (2000); the general features are that Southwestern Gelao has fewer tones and better retention of voicing distinctions, and Northern has reflexes of the original retroflex series. The SW Gelao and W Gelao have preserved fewer nasal finals than Central Gelao types.

Ostapirat (p. 33) suggested that those Gelao language groups with autonym Haki (Hagei) "…most likely…) belong to the SW branch and added that this branch seems to be the only one that lies outside Guizhou Province in China. Shen (2003) investigated the Sanchong Gelao (SC), a Hagei group from Longlin County in Guangxi Province, China and was able to determine than Sanchong Gelao, in fact, does not belong to the SW Branch. Her results showed conclusively that Sanchong was a member of the Central group, having patterns of shared innovation with other Central Gelao languages. So, for example, the pattern of changes that jointly swept over both Wanzizhai and the Sanchong as a part of their history are numerous, so I repeat only briefly some of Shen's work on this topic. Sanchong and Wanzizhai have common mutations with regard to the devoicing of some original voiced consonant initials, cf. 4.

(4)	SC	W	LZ	Q	Proto-Kra
to do A2	^{n}taw^{31}	t^{h}a^{44},	di^{35}	tɤu^{31}	*du A
bone D2	taŋ^{33}	taŋ^{33}	dæ^{31}	tɒ^{21}	*dək D
y. brother B2	nt¢u^{33}	tsəɯ^{13}	ʑu^{31}	so^{21}	*ʒaɯ B
chopsticks C2	t¢u^{33}	tsəɯ^{13}	dzau^{33}	tso^{33}	*dzaɯ B/C

One of the most important shared innovations between Sanchong and Wanzizhai is the reforming of the coda /*-ak/, which is realized as the homo-organic nasal coda /-aŋ/, cf. the example given in 5.

(5)	SC	W	LZ	Q	
deep	l^{j}aŋ53	laŋ	ʐɹ	lɒ	*(h)ləkD
to hear	^{n}tʂaŋ	tsaŋ	---	---	*dʒəkD
bone	taŋ33	taŋ	dæ	tɒ	*dəkD

In this case final -k was transformed into a final nasal coda.

Shen describes further innovations demonstrating *affrication* and *devoicing*, etc., which I will not chronicle here further, since the strength of the case for Sanchong belonging to the Central Gelao subgroup seems compelling. But the force of the argument of shared innovation as a feature of subgroups can now be used to determine the heritage of Red Gelao, a highly threatened, newly discovered language of Hà Giang Province, Vietnam..

3.1. Red Gelao. On May 27, 1997 Nguyễn Văn Lợi, Hoàng Văn Ma, and I were able to collect data on the highly endangered Red Gelao language, which in Vietnam seems to be preserved by only 50 speakers at Na Khê and Bìch Đich of Yên Minh District. These people call themselves *vɑ^{35} ntə^{31}* 'people group-red', from the root *krɑ C *dət D, cf. Laha Noong Lay *ʔdɯt* D1 'red'. Li Jinfang found a Gelao group in Fanpo Village in Malipo County of Yunnan Province in China, whose language corresponds in sound structure and lexicon to a very large degree with the Red Gelao speakers of Vietnam; they call themselves *ɯ^{3} weĭ^{55}*. In the following I will be using Na Khê data as representative, as I possess more data from this variety.

The first notable feature tying Na Khê to Laozhai regards tone, since both have but four.[2] The A tone is the only tone that has split.

(6)	A	B	C	D
Series 1	44	31	35	31
Series 2	33	31	35	31

But the segmental affinity of Na Khê and Fanpo to Laozhai is even more persuasive. Consider first the development of *kʒ (Ostapirat 2000:119). Wanzi and Na Khê have stop reflexes, whereas Wanzi and Qiaoshang have continuant reflexes, as 7 shows.

(7)	NK	LZ	W	Q	Proto-Kra
heavy A1	koŋ^{44}	qɤɯ	xau	---	*kʒəl A
lightweight C	kɯ^{31}	qɤɯ	xau	χe	*kʒa C
dry B	kɯ^{35}	qo	xen	χø	* kʒa B

Laozhai and Na Khê have fronted retroflex sonorant to alveolar consonants,

(8)	NK	LZ	W	Q	Proto-Gelao

[2] It must be noted that the contrast between A1 and A2 in Na Khê Village is not easy to distinguish. I have set up four tones, but it is possible that the A tone in this place has not split. In Fanpo Village in Malipo County of Yunnan Province, there seems no doubt about the division.

fat A2	lo33	no	nan	zø	*ɳ-
near C2	le35	lɤɯ	lau	za	*l̩-
hawk/eagle C2	lua35	lu	li	zø	*l̩-

The vowel /a/ becomes /i/ after acute, and becomes /ɯ/after grave initials, almost perfectly matching developments in LZ, cf. Ostapirat (p. 124).

(9)	NK	LZ	W	Q	Proto-Kra
eye A2	te33	ti	tau	ze	*m-ʈa A
horse C2	ɲi35	ɲi	ɳ̊tɕau	ɳ̊dʑe	*ŋja C
hand A2	mən33	mi	mpau	mbe	*mja A
snake A2	ŋɯ33	ŋɤɯ	ŋkau	ŋge	*ŋa A
lightweight C1	kɯ35	qɤɯ	xau	χe	*kʒa C
dry B1	kɯ31	qɤɯ	xau	---	*kʒa B

There is similar identity of the reflexes of /aɯ/, where are manifested /u/ or /e/ after labials, while they surface as /u/ or /au/ in LZ.

(10)	NK	LZ	W	Q	
y. brother B2	zu̯ɯ31	ʑu31	tsəɯ[13]	so21	*ʒaɯ B
duck A2	^{m}be33	blu	---	plo	*blaɯ A
chopsticks C2	^{n}dʑu35	dʑau	tsəɯ	tso	*dzaɯ B/C
male/ C2	pu35	pau	---	po	*C-paɯ C

The rhyme /an/ has become /o/ or /i/ in LZ and /o/ in NK, forms quite different from /an/ in Wanzi and /ø/ in Qiaoshang.

(11)	NK	LZ	W	Q	
egg A1	ndo44	to	tan	zø	*ʈam A
bitter A1	qoŋ44	qo	qan	---	*kəm A
teeth A1	poŋ44	pi	pan	py	*l-pən A

Finally, there is the initial fricatives *v-, *(ɣ)w-, and *x-. Ostapirat (p. 115) considers the ɣ- before w- an innovative onglide. It is found LZ and also, in NK, while in Qiaoshang it has become /ai/ and in Wanzi /u/.

(12)	NK		LZ	W	Q	
go C2	(ŋ)ve[44]	A1	---	vu	fo	* ɣwa C
sun A2	(ŋ)va[33]		ɣwə	---	---	*l-wən A
hat A1	xo31	B1	hau	hu	---	

While there is good evidence that NK and LZ are in the same subgroup, there are also several examples of independent developments. The correspondences between *vj/vr between these two SW languages is quite irregular, as is the set with initial *pwl/pwr reconstructed by Ostapirat (p. 117).

(13)	NK	LZ	W	Q	Proto-Gelao
ten D1	kuei^{31}	---	pe	vlo	*pwl- D
year A1	kuei^{44}	prə	plei	vlen A2	*pwr- A
die A1	^{n}tua^{44}	plen	pen	vlen A2	*pwr- A

One could speculate that a dissimilation or possibly a perceptual confusion might have led *pwl first to become *kwl* and then *kw*; the development of ku- occurs in *p-vj- initials.

One of the very unique features of Red Gelao are the reflexes of 'dog' and 'pig'.

(14)	NK	MLP	W	Proto-Kra
dog	xɑŋ^{44}	xɑŋ^{44} - xɑŋ^{44}	mpau	*x-ma A1
pig	foŋ^{44}	foŋ^{44}	mpa	*x-mu A1

One is reminded of the behavior of these lexical items in some forms of Sui of the Kam-Sui group, where m̥aa A1 'dog' can become hãã A1 (from Suiyu Diaocha Baogao).

4. Nung Ven (Anh). The Nung Ven ethnicity lives in Cao Bằng Province near the China border at a location 12 km east of Nội Thôn City. When we studied them in 1998, they reported a population of 200 persons. Nung Ven (NV), as they are known to the local Nung (Tai) ethnicity, or Anh [aiŋ], as they call themselves, is a member of the Buyang complex of small languages, which includes Paha, Yalhong, Ecun and Longjia. Data on these has been reported in Li Jinfang (2000).

The Nung Ven seems lexically quite similar to the Buyang complex of languages including Ecun, Langjia, and Yalhong. Many of the lexical item found in these languages are not found in the other Kra languages, including lexical items for 'belly', 'blood', 'chest', 'excrement', 'nose', 'skin', 'tendon', 'rice', 'ladder', 'short' (not long), 'wet', 'close the eyes', and 'right' (not left). So, there appears to be strong evidence for the subgroup that Ostapirat calls Central-East Kra, which is made up of Paha, Buyang (including Nung Ven) and Pubiao (Qabiao).

Another confounding influence in regard to the affiliation of the Nung Ven people stems from their cohabitation in the same village with a Tai group (Nùng). For that reason there is much influence from language contact in areas of shared cultural items. But before we consider the position of Nung Ven within Central-East-Kra, we need to report on some features of the Nung Ven language.

One the special features of this language are its tone system, cf. 15 below. Tone values were determined from multiple repetitions of: *saa* A1 'two', *tuu* A1 'three' *taa* A1 'eye'; *ŋaa* A2 'snake', *maan* A2 'new', *hau* B1 'dry'; *kuu* B1 'old', *taa* B1 'short'; *jau* B2 'younger brother', *muu* B2 'smelly', *θii* C1 'intestines', *ʔau* C1 'meat', *luum* C2 'steal', *waa* C2 'go', *ʔuk* D1 'white', *ʂaa* A1 'two' ([ʂ] represents a narrow groove retroflex sibilant fricative, similar to that used by Southern States speakers of the highlands, (p.c. Jimmy G. Harris) or the s of many speakers of Iberian Spanish), *nok* D2 'bird' and *ŋaat* D2 'needle'. The tone categories are given in terms of the system developed Gedney (1972) and Li Fang-kuei (1977)

The Nung Ven has a tone values as follows:

(15)	A	B	C	D
Series 1	34	342	454ʔ	44
Series 2	11	32	12ʔ	33

Ostapirat (177-205) gives an account of the history of Central-Eastern Kra subgroup, which includes Paha, Ecun (being the representative of Buyang complex), and Pubiao (Qabiao). Aside from Ecun, I have also taken data from Li Jinfang (2000) for Langjia and Yalhong to help in the comparison with Nung Ven.

There is a general tendency in NV not to be tolerant of consonant stop initials that lead to more marked forms of linguistic categories, such as retroflexed, postvelar, clustered consonant or complex onsets. So, for example, the /t- ʈ- nt- C-ʈ-/ as well as /-ʈ-/ have all merged to /t-/, while /k- ʔ-/ have remained unchanged. These mutations are illustrated in 16:

(16)	source	Nung Ven	Buyang (Ecun)	Paha	Qabiao	
A1	*t-	tuu	tuu	tuu	tuu	three
D1	*t-	tap	tap	tap	tjap	liver
A1	*ʈ-	tam	tam	ðam		egg
B1	*ʈ-	tai		ðai		bite
B1	*nt-	tuu	tuu	duu	tau	ashes
D1	*nt-	tik	tiak	tek	dɛɛk	full
A1	*C-ʈ-	tuu	tuu	ðuu		head louse
A1	*k-	kam	ʔam	qam		bitter
A1	*k-	kaa	ʔaa	qaa	qaa	cogon grass
B1	*k-	kuu	ʔuu	qaa	qau	old
C1	*C-k-	ʔau		ɣaa		alcohol
A1	*ʔ-	ʔai		ʔaai	ʔai	good
C1	*ʔ-	ʔuŋ	ʔɔɔŋ	ʔɔɔŋ	ʔɔŋ	water

The two classes of sibilant fricatives are much less dramatic in regard to change. Just as in the case of the stops, sibilants can be initial or medial with a preceding sequi-syllabic consonant. Thus there are *s- and *C-s as well as *ʃ- and *C-ʃ- and these have different reflexes in Nung Ven as is seen in 17 (in my data set I have no examples derived from *ʃ- or *ʧ-:

(17)	source	Nung Ven	Buyang (Ecun)	Paha	Qabiao	
A1	*s-	ṣaa	θaa	θaa	ɕee	two
A1	*s-	ṣam	θam		θam	hair of head
A1	*ts-	tṣɯ		tɕɛn		buy

C1	*ʔ-s- θii		θai	ðhii	intestine
B1	*ʔ-s- θuən	θui	θei	ðhɛɛ	garlic
A1	*ʔ-ʧ- ṣuŋ	θɔɔŋ	jɔɔŋ	θuaŋ	teeth
D1	*ɦ-ʃ- ṣuk	ɕaak D2	jhuu	θaak	rope
D1	*ɦ-ʧ-tṣət	ɕut D2	jɛt	θɑt	tail

Beyond these important changes, the voiced preglottalized stops /ʔb ʔd/ and the nasals in Nung Ven have the same reflexes as Buyang, so *ɗ- is realized as *ʔdap*^{D1} (B)/*ʔdap*^{D1} (NV) 'remember'and *nɗ- is realized as *ʔdii*^{A1} (B)/ *ʔdii*^{A1} 'gallbladder'(NV) and also *m- becomes *maan*^{A2} (B)/*maan*^{A2} (NV) and from *ʔ-m- becomes *mui*^{A2} (NV) 'bear' and from *ɦ-m- the forms *maa*^{A2} (B) / *maa*^{A2} (NV) 'five'. One difference between Buyang (Ecun) and Nung Ven is that Buyang initial *(ɣ)w- become w- and not v-, e.g. 'wind' is *wən*^{A2} in NV, corresponding to *vən*^{A2} in Buyang.

5. **A phylogentic analysis**. In this section I will describe an attempt to estimate the evolutionary history of the Kra language family using computation methods. Data will be from Ostapirat (2000) augmented by additional materials above and from notes by Li Jinfang. The idea of using an algorithmic approach for language history among linguists is probably associated with the now discredited proposal of *lexicostatistics* and *glottochronology* by Morris Swadesh, who was influenced by the idea of radiocarbon dating. He sensed that languages like radioactive substances lose some of their lexical substance over time, i.e. they have a *lexical half-life*. While glottochronology and lexicostatistics share a similar kind of algorithm, the agglomerative clustering technique, which can show errors from the assumption of a constant clocklike evolution, (Nichols & Warnow 2008:776). So, aside from the general idea, modern approaches to the phylogeny of linguistic system use algorithms quite different from those used by Swadesh.

Most of the work for computational estimation of language history is using software packages that were developed for determining the phylogenetic descent of biological species. Two of these are regarded as especially useful for linguistic analysis, Bayesian analysis and Network analysis. The software employed in this paper is identical to that used in a recent study of the evolution of language structures in Papuan languages on the island of New Guinea by Dunn et al (2008). We will be using *Splits Tree 4.0* developed by Huson &. Bryant (2006).

Linguistic phylogeny to determine language subgroups is not a replacement for classical comparative reconstruction, but rather focuses on a somewhat different but related task, approximating the evolutionary history of a language family. It will not produce a reconstruction of the states ancestral to the cognate sets found in contemporary daughter languages (Nichols & Warnow, 2008). True to its heritage in systematic biology, the contemporary languages are the "leaves" of a genetic tree and are called the *Taxa*. The input for analysis is a matrix of data, typically first assembled with a spreadsheet, for example, MS-Excel, in which the rows represent Taxa and the columns *Characters*.

Characters are conceived of as lexical items, rules, presence or absence of syntactic structures, and many other linguistic elements that maps a *gloss* onto the language in question. So stating it formally, *cat* is the Character that turns 猫 into *English.* One often begins assembling a lexical of inherited forms based upon comparative study. Thus, a line from my lexical database may appear as in (18):

(18) Lexical Database

	Gel-SC	Gel-W	Gel-Q	Gel-LZ	Gel-NK	Gel-MLP	Lac-BP	Lac-MLP	Lah-NL	Lah-MT	Paha	Ecun	NV	Langj	Yal	Qab
'骨头'	*taŋD2*	*taŋD2*	*tɒD2*	*daeD2*	*nduaD2*	*dua31*	*thɔD2*	*thjõD2S*	*dakD2*	*thakD2*	--	--	*ʔdakD1*	--	--	*ʔdakD1*

Note that in 18, there is only one state for the gloss 骨头 in all languages save for Paha, Ecun, Langjia, and Yalhong, which don't have this character. These four must use another character for the 骨头 gloss.

The next step takes us from the lexical database to a Character State database, which is an elaboration of the lexical database. An example line from the Character State Database is seen in 19.

(19) Character State Database

	Gelao-SC	Gelao-W	Gelao-Q	Gelao-LZ	Gelao-NK	Gelao-MLP	Lachi-BP	Lachi-ML	Laha-NL	Laha-T	Paha	Ecun	NungV	Langjia	Yalhong	Qabiao
bone	taŋD2	taŋD2														
bone				daeD2	nduaD2	dua31										
bone			tɒD2													
bone							thɔD2									
bone								thjõD2S								
bone									dakD2							
bone										thakD2						
bone													ʔdakD1			ʔdakD1

This screen shot from MS-Excel shows the characters for the gloss 骨头 have been assigned to eight different *character states* on successive rows, as the characters are seen to be close cognates or further cognates. Some situations are different from the case of the gloss 骨头, for example for the gloss 女孩子 English has *girl,* German has *Mädchen,* and Swedish has *flicka*, all not cognate. So the gloss 女孩子 is *polymorphic*, i.e. has three character states. Kra also demonstrates polymorphisms. So, for example, the gloss 血 can be realized as a descendent of *plat D for most Kra languages and realized as a descendent of *kɣa C for Buyang and Pubiao (Qabiao). Such polymorphism can be helpful in establishing subgroups, in this case the subgroup Buyang + Pubiao.

There is another consideration. Linguistic subgroups should be decided on the basis of *homology* 'shared origins' and not on the basis of *homoplasy* 'similar form' or 'parallel developments'. To give an example from biology, there are several kinds of organisms that have developed toxin to protect themselves from predators, e.g. reptiles, insects, platypuses, capibaras, birds, etc., but these are not descended from a common biological source but rather developed in parallel from common utility. That means in linguistic research we should be distrustful of subgroups that could have developed

independently because of typological, cross-linguistic principles or natural rules, as these may occur in unrelated species or at two or more places in a tree of descent of one species without them being because of common origins. As Nichols & Warnow (2008) view on this issue is that the quirkier the innovation, the better for determining subgroups.

(20) Transposed Character State Database—continuing to the right for all character states

1	Gloss	alive	alive	alive	armpit	armpit	armpit	beard	beard	beard	beard	belly
2	Gelao-SC	plu35			ɖiŋC2				miB2			
3	Gelao-W	pleuu			tsenB2		?		menC2		?	
4	Gelao-Q			pu	?		?					
5	Gelao-LZ		pu			tɕiC1						
6	Gelao-NK		pai		?		?		miC2		?	fo31
7	Gelao-MLP											mo31
8	Lachi-BP				?					miC2		maD1
9	Lachi-ML				tjąC1		?		?		?	?
10	Laha-NL				taaiC1		?		?		ɲuutD2	?
11	Laha-T				?		?		?		ɲutD2	?
12	Paha				taaiB1		?		?		?	mkookD1
13	Ecun				?		liɛA2	mɯ̃mC2			?	?
14	NungV				?		?	mẽmC2			?	
15	Langjia											
16	Yalhong											
17	Qabiao				?		lhiiA2	mɯmC2			?	m̥okD1

The character state database must now be translated into the Nexus file structure. The first step in this process is to transpose the taxa and characters. The result is seen in 20. Then, the filled in cells in 20 must be replaced with 1, the empty cells with 0, and the questions marks cells remain and signify no data available. One must also add in a few command lines to the input file to inform the process about the number of taxa and characters, and when to stop. A portion of the Nexus file for the Kra computation is seen in 21.

(21) Nexus File for the Kra computation (in part only)

```
#NEXUS
begin data;
dimensions ntax=18 nchar=748;
FORMAT
MISSING=? [GAP=?] Datatype=STANDARD [SYMBOLS = "0 1"];
MATRIX
Kam 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0
Zhuang 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0
Gelao-SC 1 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0
Gelao-W 1 0 0 1 0 ? 0 0 0 1 0 ? 0 0 0 0 0 1 0 0 0 0 0 0 1
Gelao-Q 0 0 1 ? 0 ? 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0
Gelao-LZ 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0
Gelao-NK 0 1 0 ? 0 ? 0 0 0 1 0 ? 1 0 0 0 0 1 0 0 0 0 0 0
Gelao-MLP 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0
Lachi-BP 0 0 0 ? 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 1 0 0 0 0 0
Lachi-ML 0 0 0 1 0 ? 0 0 0 ? 0 ? ? 0 0 0 0 0 0 1 0 0 0 0 0
Laha-NL 0 0 0 1 0 ? 0 0 0 ? 0 1 ? 0 0 0 0 0 0 0 1 0 0 0 0 0
Laha-T 0 0 0 ? 0 ? 0 0 0 ? 0 1 ? 0 0 0 0 0 0 0 1 0 0 0 0 0
Paha 0 0 0 1 0 ? 0 0 0 ? 0 ? 1 0 0 0 0 0 1 0 0 0 0 0 0 0
Ecun 0 0 0 ? 0 1 0 0 1 0 0 ? ? 0 0 0 0 0 0 0 0 1 0 0 0 0 0
NungV 0 0 0 ? 0 ? 0 0 1 0 0 ? 0 1 0 0 0 0 0 0 0 1 0 0 0 0
Langjia 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0
Yalhong 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 ? 0 0 0
Qabiao 0 0 0 ? 0 1 0 0 1 0 0 ? 1 0 0 0 0 0 0 0 0 0 1 0 0 0
;
END;
```

The file must be saved in *.txt format for input to SplitsTree 4.0. Tutorials and examples of how to run a file are included with the software instructions.

6. **Results of the phylogenetic analysis and discussion.** SplitsTree 4.0 results are given in Figures 1 & 2 below. The Nexus file for this calculation contained 756 character states and 16 taxa.

Some of the main features of the resulting network are:

a. The Kra network is divided into two sections Gelao-Lachi in the west and Laha-Paha-Pubiao (Qabiao)-the Buyang group (Nung-Ven, Ecun, Langjia, and Yalhong) in the east.

b. The two Lachi taxa show a long and a very thin branch radiating from the center of the network, indicating that these taxa share a common heritage and were created by "rapid radiation". This result appears plausible since the Lachi live in communities just opposite each other on the Vietnam-China border. The only difference in the two varieties I noticed is that Lachi-BP has changed s- -> f- for some items.

c. Central Gelao represented by the taxa, Gelao Wanzizhai and Gelao Sanchong, are separated by about 100 km of mountainous territory, though they seem close.

d. Gelao-NK, Gelao-Malipo, and Gelao Laozhai are also linked together as are Laha Noong Lay and Laha Tamit, but the branching suggests greater distance and interaction among the varieties.

e. The Buyang languages (Nung-Ven, Ecun, Langjia and somewhat more distantly Yalhong) represent an area of linguistic interaction and appear to be more distantly connected to Pubiao

(Qabiao). Ostapirat recognizes this feature as well. Paha appears to be closer to the Laha branch.

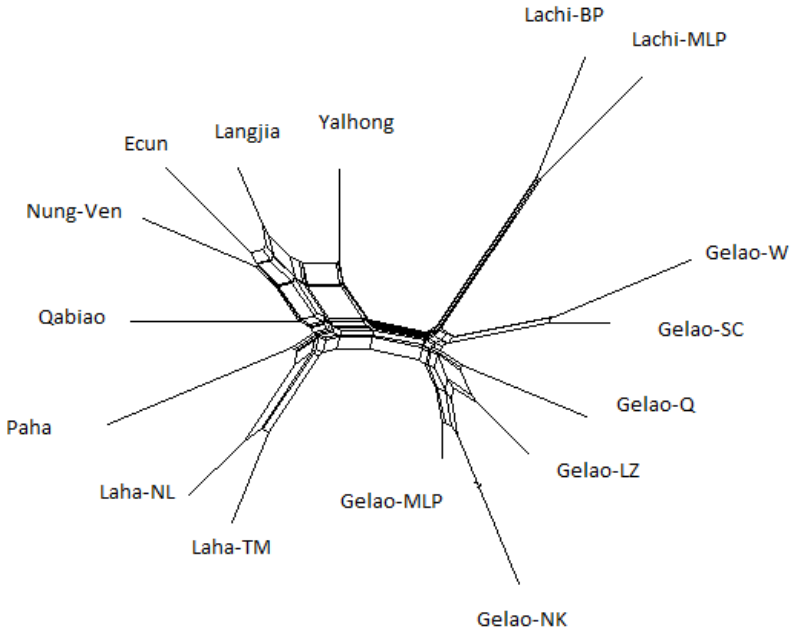

Figure 1: A SplitsTree 4.0 Network of the relationship among the Kra languages

Figure 2 was drawn from the same data that produced Figure 1. It looks specifically for language mixing. One can see the connections between two or more branches in the tree, specifically the Buyang complex. It also shows that the Gelao languages are close to the root of the tree. That could have been predicted from the observation of Sapir in 1916 that language diversity (as well as biological diversity) is greatest near the origin of a phylogeny. Gelao fills that description well, having a large population (though not too many speakers), settled over a large amount of territory, with the vernacular forms showing great diversity.

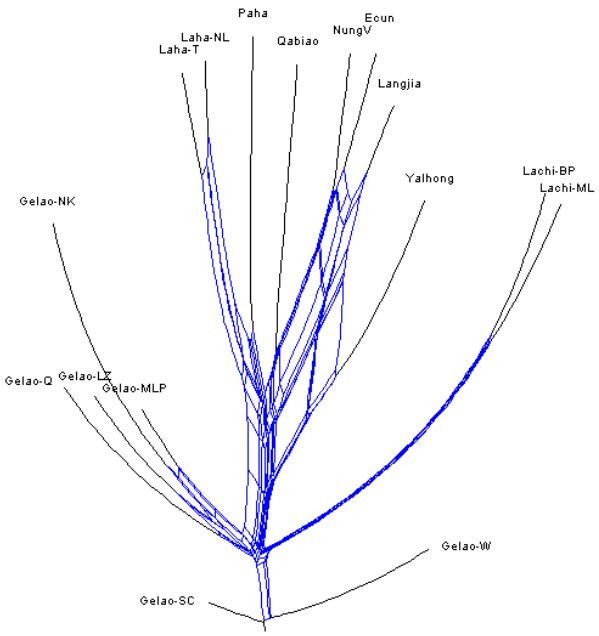

Figure 2: With a rooted hybrid network showing the branches of the tree with significant interaction among Nung-Ven, Ecun, and Langjia with Yalhong more distant from the area of mixing

7. Discussion and conclusions. The estimation of the distance among the branches corresponded with the benchmark, the diagram Ostapirat (p. 41), except that Laha relentlessly preferred to adjoin to Paha, Buyang, and Pubiao and not to Gelao and Lachi. In the input data I kept those etyma with manifest CVC structure in different character states from those that lost final oral stops and that might be responsible. I attempted to model the commonality of the voiced stops becoming voiced breathy and then voiceless aspirated as in Lachi to help represent that innovation. But it was not sufficient to change the allegiance of Laha. So, my estimation must count as a failure for the moment. Otherwise the estimated structure matches the benchmark quite well and adds a bit more information about distance between the taxa as captured in the branch length.

References

Chao, Yuen-Re. (1930). A system of tone-letters Le Maître Phonétique 45. 24–27

Edmondson, Jerold A. & Kenneth J. Gregerson. (1997). Outlying Kam-Tai; notes on Ta Mit Laha. Mon-Khmer Studies 27. 257-69.

Gedney, William. J. (1972) A checklist for determining tones in Tai dialects. In Studies in Linguistics in honor of George L. Trager. The Hague: Mouton.

Huelsenbeck, John and Fredrik Ronquist, 2001. MrBayes: Bayesian inference of phylogeny, Bioinformatics 17.754–5.

Huson, D., and D. Bryant 2006. Application of phylogenetic networks in evolutionary studies. Molecular Biology and Evolution 32.254–67.

Li Fang-Kuei. (1977). A handbook of comparative Tai. Honolulu: The University Press of Hawaii.

Michael Dunn, Stephen C. Levinson, Eva Lindström, Ger Reesink & Angela Terrill. (2008) "Structural phylogeny in historical linguistics: methodological explorations applied in Island Melanesia". *Language* 84:4.710–759.

Nichols, Johanna & Tandy Warnow. (2008). Tutorial on Computational Linguistic Phylogeny. *Language and Linguistics Compass* 2/5 (2008): 760–820, 10.1111/j.1749-818x.2008.00082.x

Ostapirat, Weera. 2000. Proto Kra *LTBA* 23.1-251.

Sapir, Edward. 1968 (1916). Time perspective in aboriginal American culture: a study in method. In *Selected writings of Edward Sapir in language, culture and personality* (D.G. Mandelbaum ed.), 389- 467. Berkeley: University of California Press

Solntseva, Nina V. & Hoàng Văn Ma. (1986). Materials from the Soviet-Vietnamese linguistic expedition for the year 1970: Laha Language. Moscow: Nauka.

李锦芳. 2000. 怖央语言研究. 北京. 中央民族大学出版社