ICSTLL46
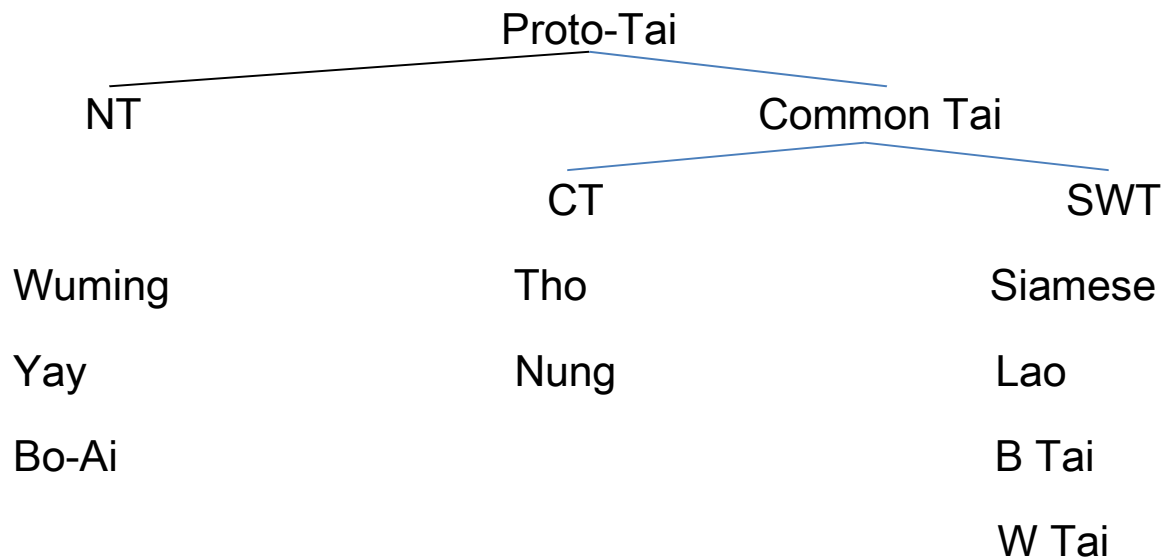
Tai subgrouping using phylogenetic estimation

Jerold A. Edmondson

Professor Emeritus of The University of Texas at Arlington
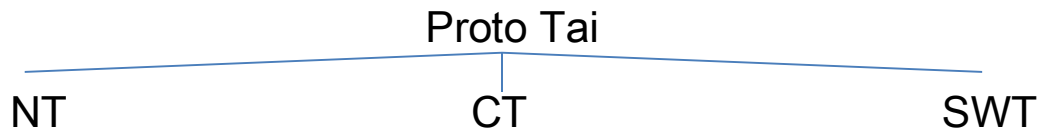
j.edmondson@sbcglobal.net or j.edmondson@gmail.com

1. The Tai stock of Tai-Kadai has subgroupings as follows

a. A. Haurdricourt 1956—examining ~20 lgs from Vietnam, Thailand, & China concluded the parent language divided

```
                        Proto-Tai
        NT                              Common Tai
                              CT                      SWT

    Wuming                   Tho                      Siamese

    Yay                      Nung                     Lao

    Bo-Ai                                             B Tai

                                                      W Tai
```

Based on H's study of dictionaries and reports from French clerics and civil servants.

2. Li Fang Kuei 1977 concluded

Proto Tai

| NT | CT | SWT |

Li began with the three representative languages

| Bo-Ai | Longzhou | Siamese |
| Wuming | | Ahom |

3. Phylogenetic analyses for settle controversies and providing new approaches to lg subgrouping
    a. Developed by molecular biologists for dealing with DNA and complex morphological data
    b. For lg data Requires linguistics encoding into machine readable form
    c. Can represent results as 'treelike' and 'non-treelike' development of linguistic phylogeny

4. We used
    a. 200 lexical and phonological data points
    b. 30 languages
    c. Data consists of taxa=the languages & characters=lexical-phonological points
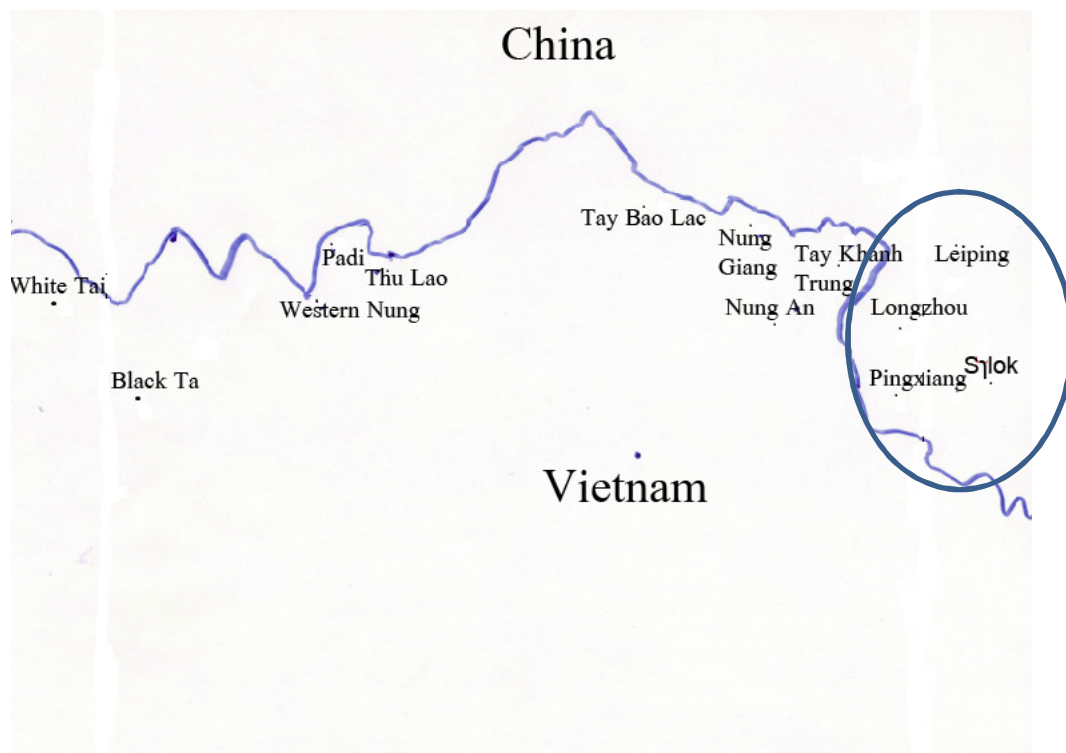
5. Example the characters, character states, taxa, & gloss

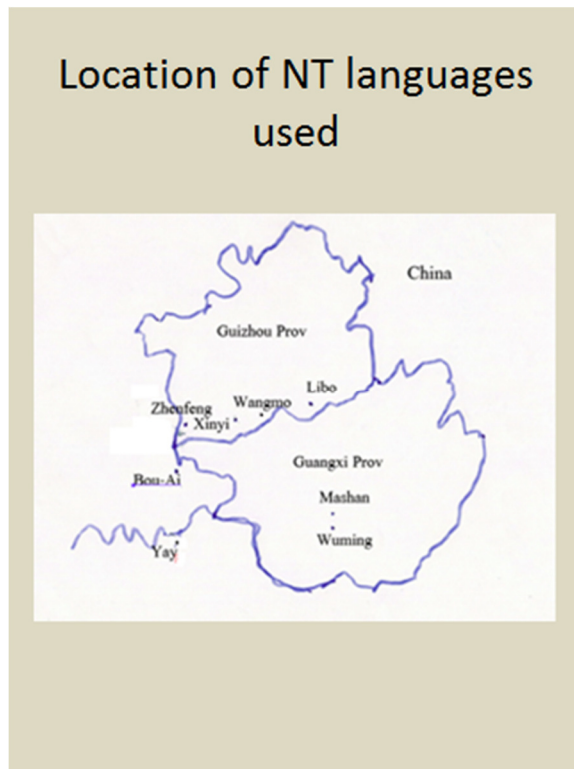| Gloss	die	taxon1	taxon2	taxon3……………….
Characters		thaai	thaai	
	die		haai	
	die			taai taai taai

6. The languages

Kam, Nung_Am, Nung_Giang, Western Nung, Thu_Lao, Tay BaoLac, Tay KhanhTrung, CaoLan, S_1Lok, PingXiang, LeiPing, NingMing, NungChao (Zhao)., Ahom, ThaiCentral, SouthernShan, WhiteTai, BlackTai, HPShan, Padi Mường Khương, Dehong, Lao, TaiTeeng, NiuChang Buyi, Wangmo Buyi, Libo Buyi, Xingxi Buyi, MaShan Zhuang, & Wuming Zhuang.

7. Some of the lesser studied languages

    a Caolan—endangered found in Tuyên Quang Prov, field Study

    a. ThuLao—found in Mường Khương area LàoCai Prov, also studied by Prof. Therapan L.Thongkum; field study

    b. Padi found in Mường Khương, SWT similar to WhiteTai.

    c. Nùng An and Nùng Giang (Yang); N. Cao Bằng Prov. Nùng Giang (Yang) is identical to the major lg of Jingxi Country 靖西县 in Guangxi

    d. Tày Khánh Trung and Tày Bảo Lạc of Cao Bằng Province are Tày languages (formerly called Thô)

    e. Tai Theeng—a SWT lg from Nghệ An Province, far to the south of the Red River Valley of Vietnam.

b. Map of the lgs along the Sino-Vietnamese border. Circled area used by Li Fangkuei 1977 for CT

China

Tay Bao Lac

White Tai
Padi
Thu Lao
Western Nung

Nung Giang
Tay Khanh
Trung
Nung An

Leiping
Longzhou

Black Ta

Pingxiang S'lok

Vietnam

9. Map of NT languages

Location of NT languages used

China

Guizhou Prov

Libo

Zhenfeng Wangmo
Xinyi

Bou-Ai
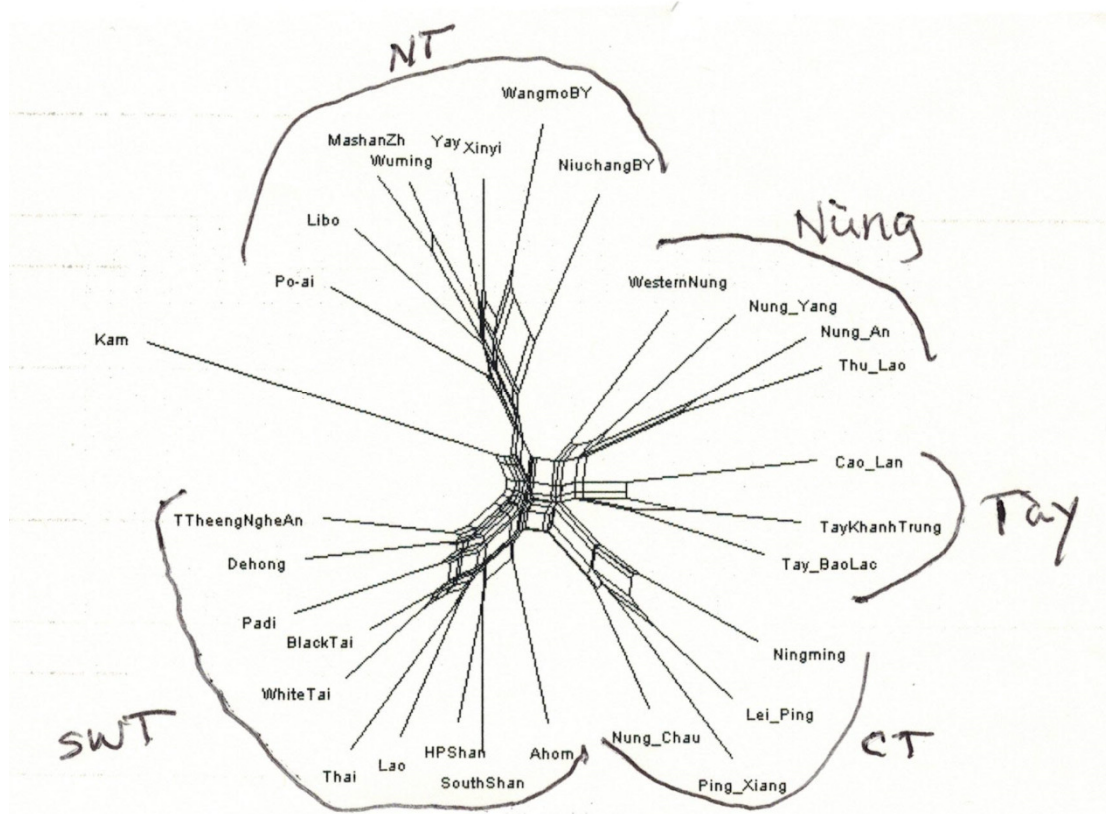
Guangxi Prov

Mashan

Wuming

Yay

10 Great interest in last 15 yrs in computation methods for estimating language histories—Bantu (Holden & Gray), Austronesian (Gray et al), IE (Gray & Atkinson 2003).

Computational solutions can help settle conflicting views or provide a new solution for an unstudied lg.

For 30 rooted languages there are 4.9518×10^{38} possible binary trees; for multifurcating trees, there are 7.0717×10^{41} tree solutions. We chose a binary solution.

No direct computational search can consider so many possible trees, so other search strategies must be used.

NeighborNet solutions are for lgs with non-treelike development, e.g. West African "3-days languages" with much interaction and/or convergence situations. We use the implementation Splits-Tree 4.0, D.H. Huson and D. Bryant 2006. Application of Phylogenetic Networks in Evolutionary Studies, Molecular Biology & Evolution 23(2), 254-267

This network show centers of clear division in the network, NT, SWT, CT, Tày, & Nùng. Surprisingly, there appears to be a CT', which includes Li (1977) CT as well as Nùng and Tày branches.

Other relationships are as expected, e.g. Thai & Lao, Black & White Tai share a branch, etc.

Since there are clear examples of branches, a tree phylogeny is a possible analysis, i.e. not just non-treelike evolution.

10. Bayesian Inference. Implemented here with MrBayes by JP Huelsenbeck.
    a. Starting from initial model tree-in Ling study we pick a tree at random
    b. Computer takes a 'Random walk' through 'model tree space'.

c. Computer Markov chain Monte Carlo* algorithm (by Metropolis) creates random trees

d. The process continues until new trees don't increase probability.

Result for Tai phylogeny with 1 million generations

Kam

Nung Chau

100

Ping Xiang

Lei Ping ←

Ningming

100

WesternNung

100

Nung Yang

100

Nung An

100

Thu Lao ←

92

Tay BaoLac

98

TayKhanhTrung

Cao Lan

Ahom

100

Thai

99

Lao

BlackTai

100

WhiteTai

75

Padi ←

100

99

Dehong

TTheengNgheAn

100

SouthShan

HPShan

100

NiuchangBY

WangmoBY

100

Libo

99

Xinyi

98

Yay ←

63

PoAi

100

Wuming

100

MashanZh

Conclusions

The two blue arrows show that, contra Li (1977), there is a CT' with two main branches, one a Nùng and Tày sister subgroup and one for the CT (of Li 1977).

The red arrow shows Ahom to be a SWT' supergroup, and a SWT branch with Thai and Lao, and Black Tai, Padi & White Tai sister groups, and the Shan languages forming the remaining branch.

The black arrow shows the NT branch with Wuming and Mashan Zhuang as sisters. The remaining branch has a Xinyi & Yay sister branch and the remaining languages forming a hierarchy.

Kam is not a member of the Tai stock and it show up outside the stock and forms the root.

Neither Haudricourt (1956) nor Li (1977) was confirmed.

The number Y on branches represents a consensus values; i.e. for the last X trees considered, a given branch occurred Y% of the time. (50% is 9 trees, 99% is 50 trees (approximately for 1 million generations). PSRF value = 1.000 indicating convergence.

One concern is that there are often polymorphs* separating NT from CT' and SWT, e.g. for *sky* fa ~ ban, *thunder*, fa dong ~ pja, etc. Since the polymorphs largely define a branch, including them with phonological changes is justified. *polymorphs = Eng stone vs. rock (Fr loan).

Sources

- Felsenstein, J (1985). *Inferring Phylogenies*. Sinauer: Sunderland, MA.

- Haudricourt, A. (1956). De la restitution des initiales dan les langues monosyllabiques: le problèm du Thai commun. *Bulletin de la Société de linguistique de Paris* 52.307-22

- Li Fang Kuei (1977). *A handbook or comparative Tai*.   The U Hawaii Press: Honolulu.

- W.Nung, Ningming, Leiping & Pingxiang Western Nung from T. Hudak (1995) Gedney's Central Tai languages. Univ of Michigan Press:Ann Arbor .

- White Tai, Black Tai, & HP Shan from T.Hudak 1994 Gedney's SW Tai languages U Michigan Press: Ann Arbor.

- Nungchau, Dehong, Wuming, & Wangmo from 壮侗语族语言间志 1983. Beijing:Minzu Press

- Po-ai 剥隘. From Li (1977).

- *Bouyei Chinese English Thai Dictionary* 2001. Mahidol U Press

- *Northern Zhuang Chinese Thai English* 2006 Mahidol U Press

- *Libo Buyi-Chinese English Glossary*. 1995. Yu Jiongbiao-Wil Snyder SIL

- Xingyi Buyi from 布依语调查报告. 1959. 科学出版社. 北京

- Tày—KhanhTrung & Tày Bảo Lạc Field notes, Cao Lan-Tuyên Quang Field notes, Tai Teeng Nghệ An Field notes, Padi –field notes & Thu Lao – field notes

- Holden, Clare J., and Russell Gray.(2006) Rapid radiation, borrowing, and dialect continua in the Bantu languages. In *Phylogenetic methods and the prehistory of languages*, ed. by Peter Forster and Colin Renfrew, 19-31. Cambridge, UK: MacDonald Institute Press, University of Cambridge. For cases of diffusion

- Gray, Russell D. and Quentin D. Atkinson. (2003). Language-tree Divergence Times Support the Anatolian Theory of Indo-European Origin. *Nature* 426: 435–39.

- Nichols, Johanna and Tandy Warnow (2008) Tutorial on Computational Linguistic Phylogeny. *Linguistics and Language Compass*, Vol. 2, No. 5: 760-820

- Edmondson, Jerold A., Ken Gregerson, & Paul Sidwell (2011). The North Bahnaric Clade: a numerical approach in ICAAL4-2 EG&S 250911