A Bayesian phylogeny of Patkaian (Northern Naga)

statistical methods with large data on small languages

Kellen Parker van Dam

Universität Zürich Switzerland
Department of Comparative Language Science
Center for the Interdisciplinary Study of Language Evolution

La Trobe University Australia
Department of Languages & Cultures

Universität Passau Germany
Multicultural Computational Linguistics

Universität Zürich UZH

LA TROBE UNIVERSITY

FNSNF
Fonds national suisse
de la recherche scientifique

Sino-Tibetan Phylogenies

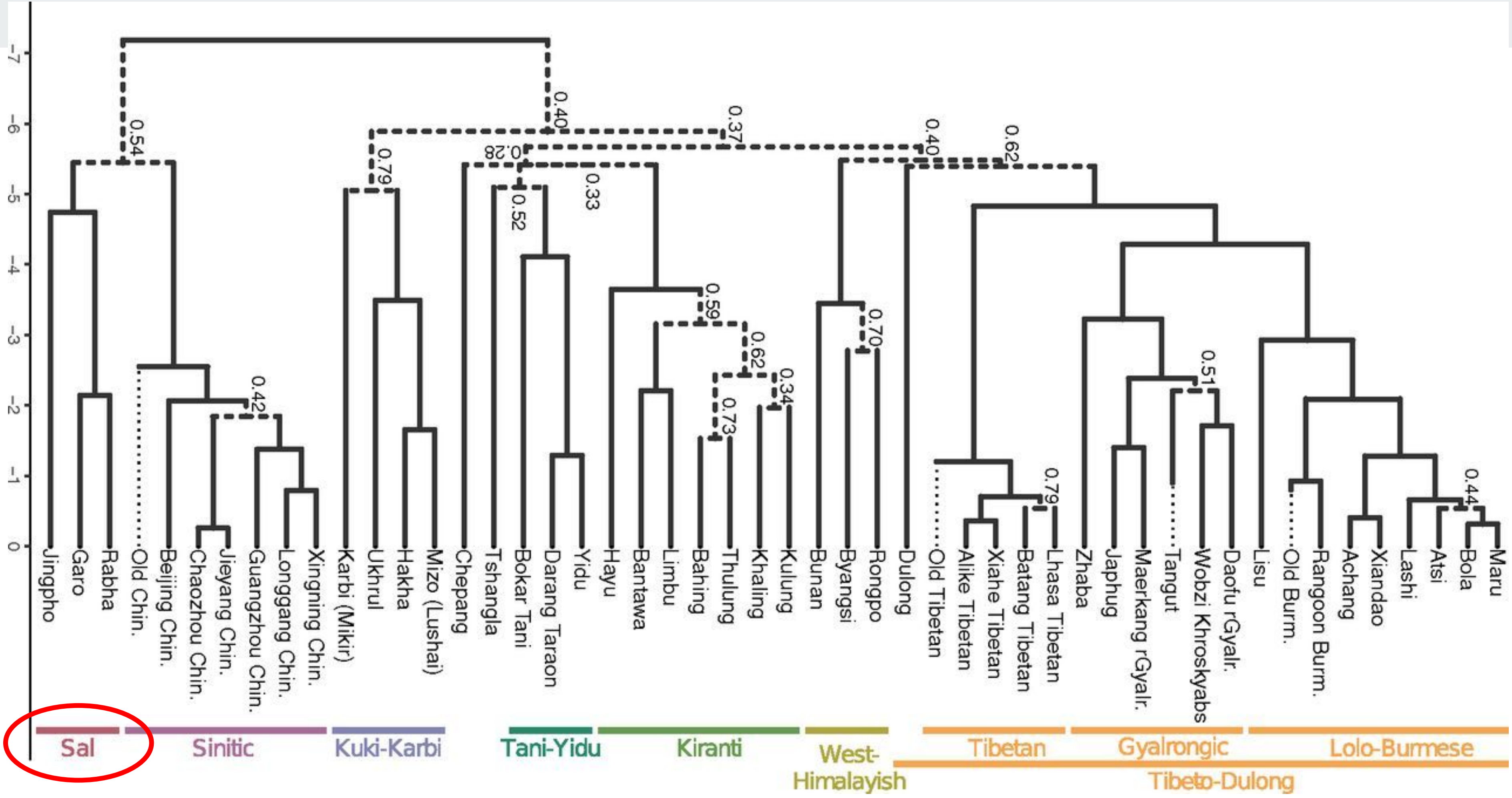
Recent Sino-Tibetan phylogenies offer interesting suggestions of large-scale relationships in the family.

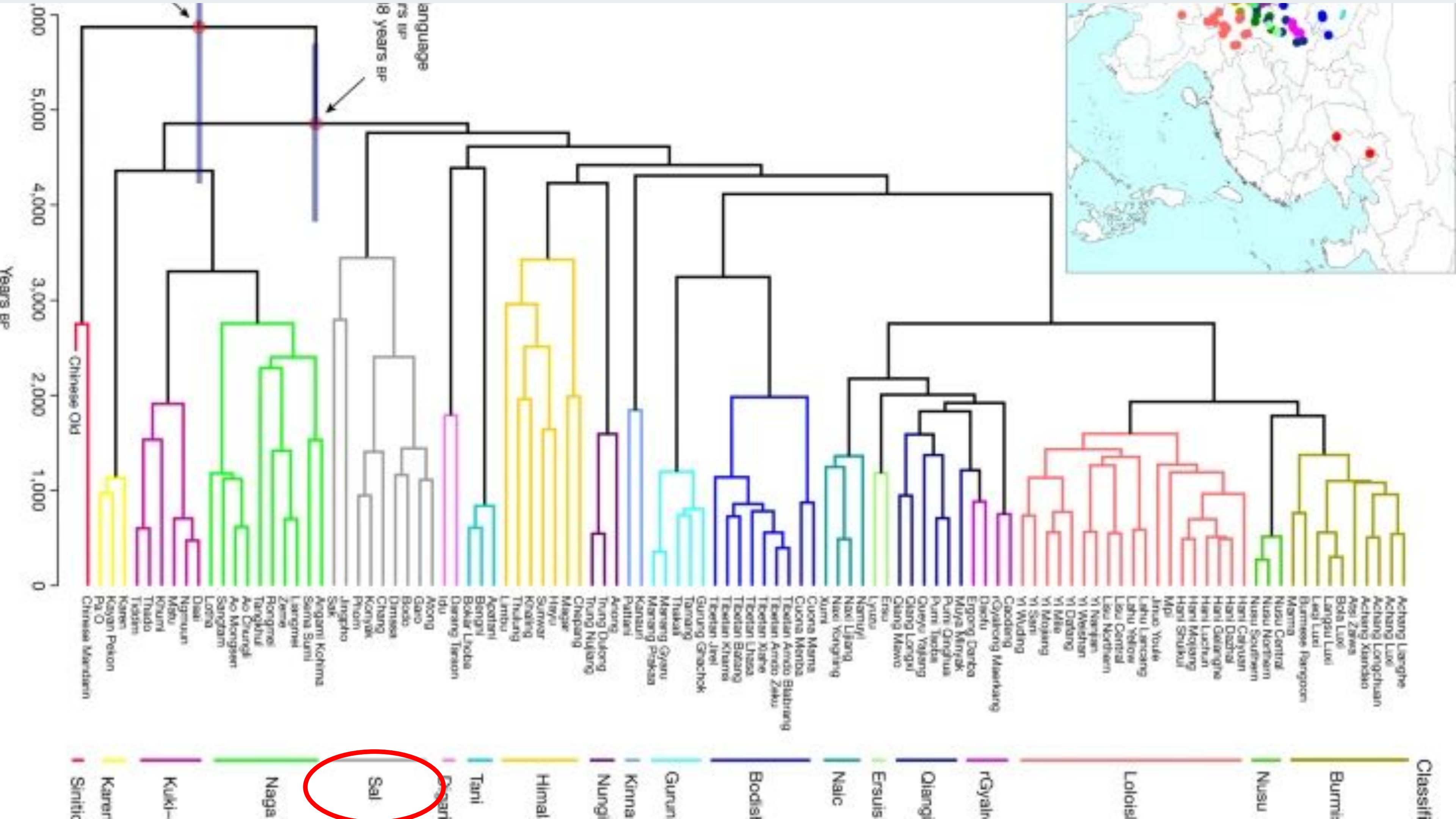
Methods and approaches vary, resulting in vastly different trees.

Posteriors are often quite low.

Sagart, L., Jacques, G., Lai, Y., Ryder, R.J., Thouzeau, V., Greenhill, S.J. and List, J.M., 2019. Dated language phylogenies shed light on the ancestry of Sino-Tibetan. Proceedings of the National Academy of Sciences, 116(21), pp.10317-10322.

Zhang, M., Yan, S., Pan, W. and Jin, L., 2019. Phylogenetic evidence for Sino-Tibetan origin in northern China in the Late Neolithic. Nature, 569(7754), pp.112-115.


-7
-6
-5
-4
-3
-2
-1
0
0.54
0.40
0.79
0.28
0.52
0.33
0.37
0.40
0.62
0.42
0.59
0.62
0.34
0.73
0.70
0.51
0.79
0.44
Jingpho
Garo
Rabha
Old Chin.
Beijing Chin.
Chaozhou Chin.
Jieyang Chin.
Guangzhou Chin.
Longgang Chin.
Xingning Chin.
Karbi (Mikir)
Ukhrul
Hakha
Mizo (Lushai)
Chepang
Tshangla
Bokar Tani
Darang Taraon
Yidu
Hayu
Bantawa
Limbu
Bahing
Thulung
Khaling
Kulung
Bunan
Byangsi
Rongpo
Dulong
Old Tibetan
Alike Tibetan
Xiahe Tibetan
Batang Tibetan
Lhasa Tibetan
Zhaba
Japhug
Maerkang rGyalr.
Tangut
Wobzi Khroskyabs
Daofu rGyalr.
Lisu
Old Burm.
Rangoon Burm.
Achang
Xiandao
Lashi
Atsi
Bola
Maru
Sal
Sinitic
Kuki-Karbi
Tani-Yidu
Kiranti
West-Himalayish
Tibetan
Gyalrongic
Lolo-Burmese
Tibeto-Dulong

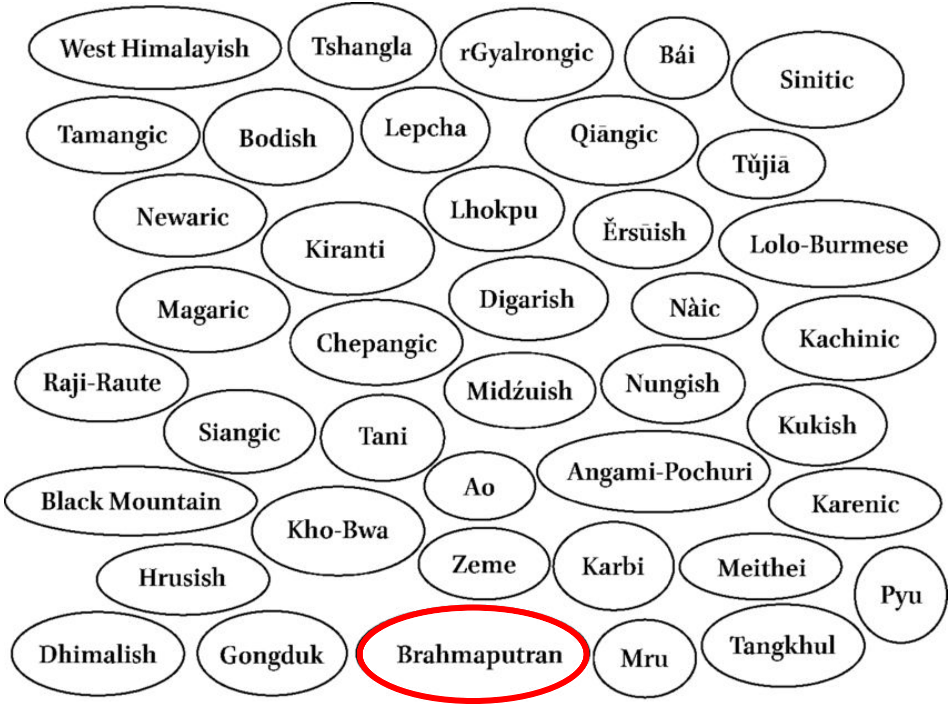


anguage
rs BP
8 years BP
Years BP
,000
5,000
4,000
3,000
2,000
1,000
0
Chinese Old
Chinese Mandarin
Pa O
Kayan Pekon
Karen
Tiddim
Thado
Khumi
Mizo
Ngmuun
Daai
Lotha
Sangtam
Ao Mongsen
Ao Chungli
Tangkhul
Rongmei
Zeme
Liangmei
Sema Sumi
Angami Kohima
Sal
Jingpho
Phom
Konyak
Chang
Dimasa
Bodo
Garo
Atong
Idu
Darang Taraon
Bokar Lhoba
Bengni
Apatani
Limbu
Thulung
Khaling
Sunwar
Hayu
Magar
Chepang
Trung Dulong
Trung Nujiang
Anong
Pattani
Kanauri
Manang Gyaru
Manang Prakaa
Thakali
Tamang
Gurung Ghachok
Tibetan Jirel
Tibetan Kham
Tibetan Batang
Tibetan Lhasa
Tibetan Xiahe
Tibetan Amdo Zeku
Tibetan Amdo Blabrang
Cuona Menba
Cuona Mama
Xumi
Naxi Yongning
Naxi Lijiang
Namuyi
Lyuzu
Ersu
Qiang Mawo
Qiang Longxi
Queyu Yajiang
Pumi Taoba
Pumi Qinghua
Muya Minyak
Ergong Danba
Daofu
rGyalrong Maerkang
Caodeng
Yi Wuding
Yi Sani
Yi Mojiang
Yi Mile
Yi Dafang
Yi Weishan
Yi Nanjian
Lisu Northern
Lisu Central
Lahu Yellow
Lahu Lancang
Jinuo Youle
Mpi
Hani Mojiang
Hani Luchun
Hani Shuikui
Hani Gelanghe
Hani Dazhai
Hani Caiyuan
Nusu Southern
Nusu Northern
Nusu Central
Marma
Burmese Rangoon
Leqi Luxi
Langsu Luxi
Bola Luxi
Atsi Zaiwa
Achang Xiandao
Achang Longchuan
Achang Lianghe
Sinitic
Karen
Kuki-
Naga
Sal
Deng
Tani
Himal
Nungi
Kinna
Gurun
Bodisl
Naic
Ersuis
Qiangi
rGyalr
Loloisl
Nusu
Burmis
Classifi

Sino-Tibetan Phylogenies

Recent Sino-Tibetan phylogenies offer interesting suggestions of large-scale relationships in the family.

However, significant differences in results and approach still leave us reliant on the “Fallen Leaves” model of van Driem (2012).



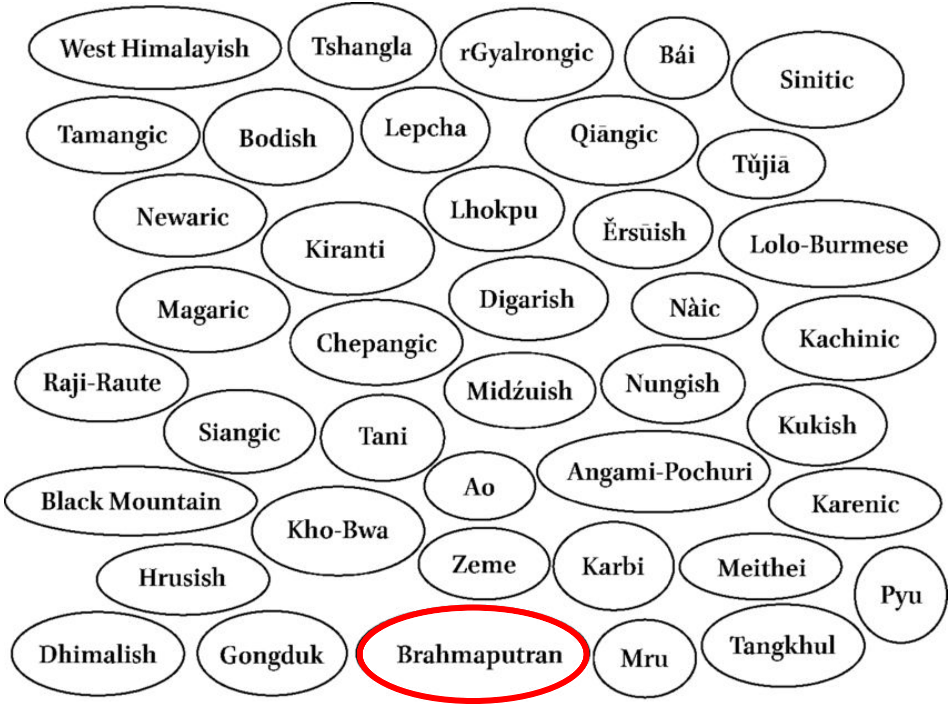
See also: Orlandi, G., 2021.

Sino-Tibetan Phylogenies

Recent Sino-Tibetan phylogenies offer interesting suggestions of large-scale relationships in the family.

However, significant differences in results and approach still leave us reliant on the “Fallen Leaves” model of van Driem (2012).

One solution: shifting to a intensive **bottom-up** approach to resolve some of these issues.



See also: Orlandi, G., 2021.

The approach:

For each branch of Sal:

- collect data for all attested varieties / all published doculects based on the ~750 concept “SALIST” word list
- curate the data to account for biases / mistakes in elicitation, morphological features, semantic splits &c, omitting external borrowings and identifying internal borrowings
- produce trees following as closely as possible the methods of Sagart et al (2019)

Stage 1 - Bodo-Garo

- *A Bayesian phylogeny of Bodo-Garo: Testing novel methods on established groupings.* North East Indian Linguistics Society conference. Guwahati, Assam, India. January 2023
- *Developing Bayesian language phylogenies from previously published data: A case study of Bodo-Garo.* Workshop on New Results and Methods in Reconstructing Population History. Universität Zürich, Zürich, Switzerland. 30 January – 1 February 2023

Conclusion:

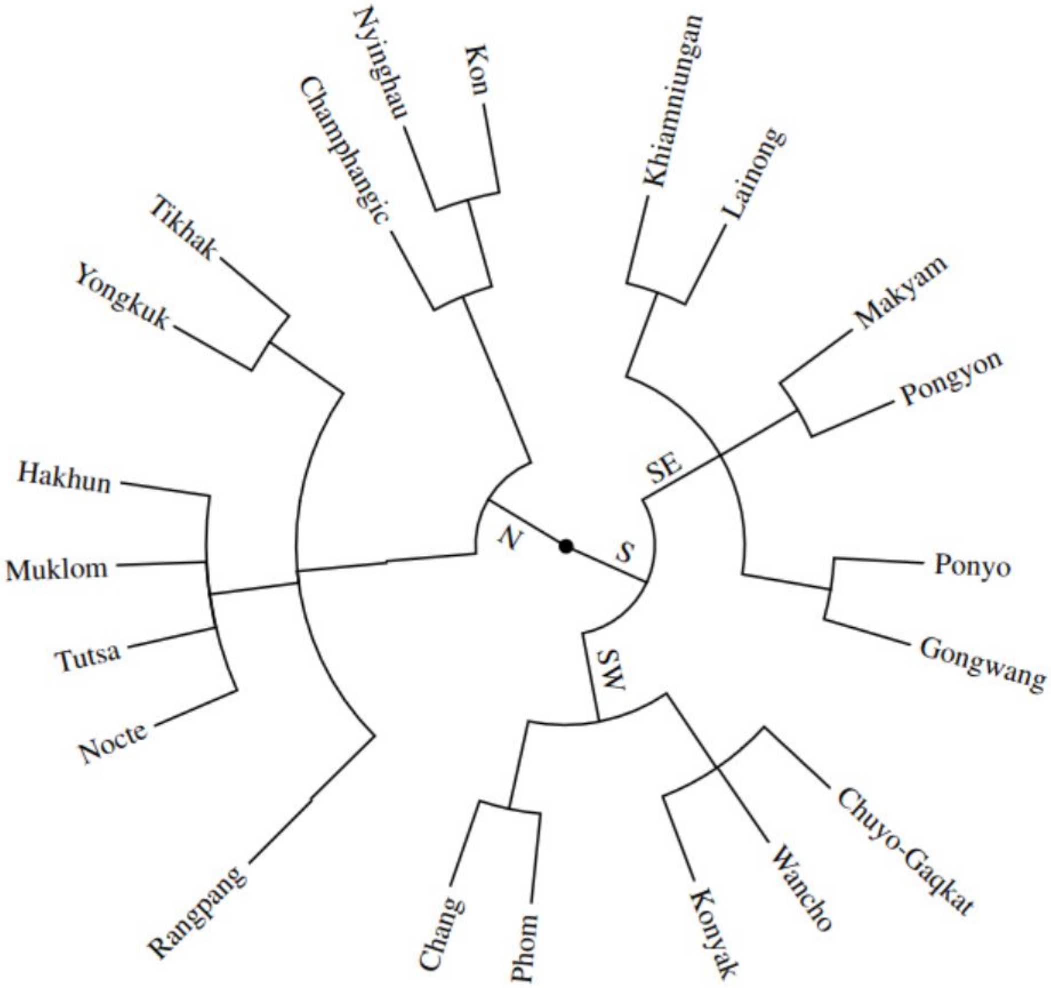
Uncontrolled data sources / elicitation is an issue, but the nature of Bodo-Garo as a heavily creolised former lingua franca is a bigger issues.

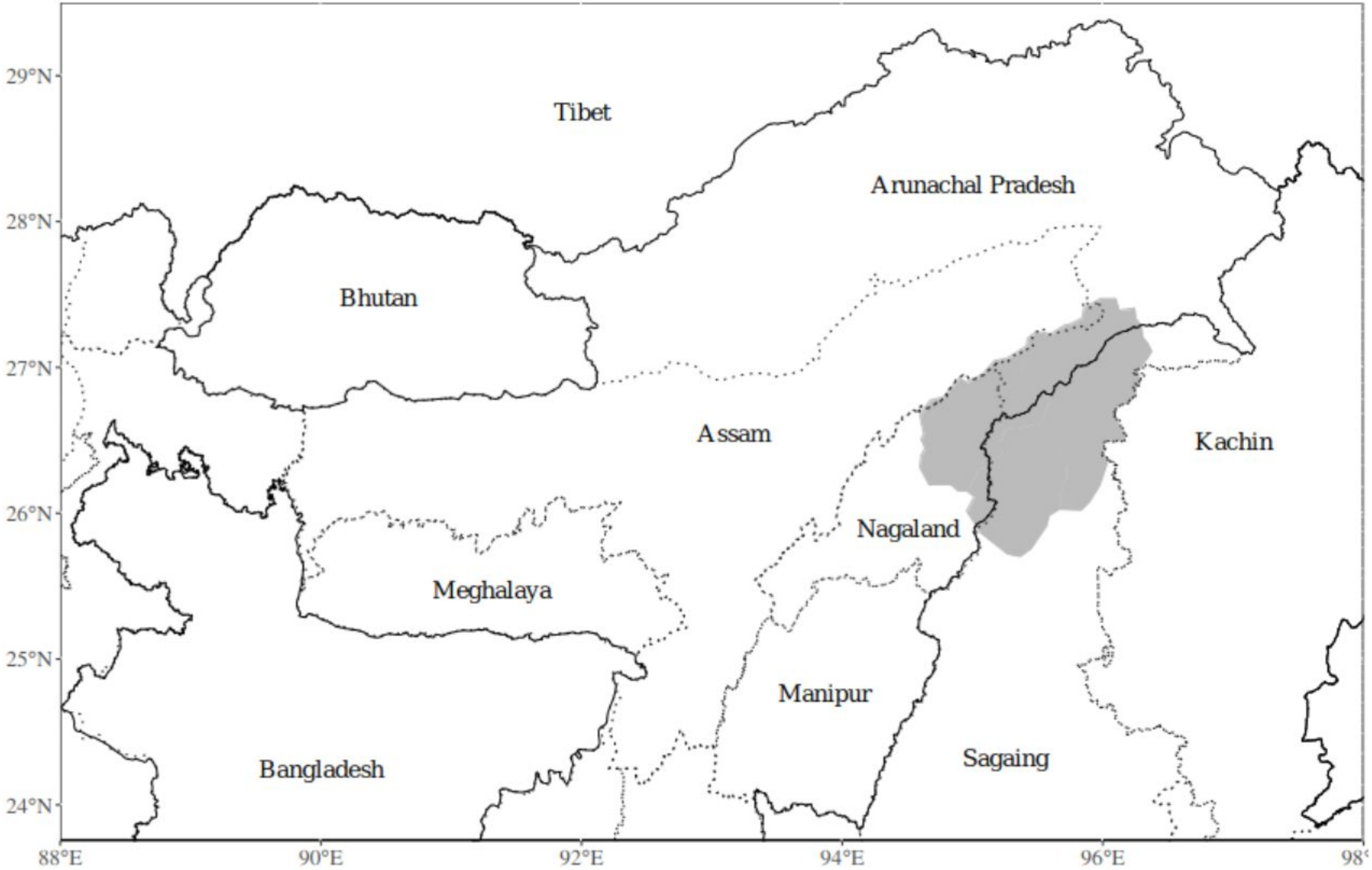
Stage 2 - Patkaian

Initial grouping based on speaker self-reporting & impressionistic descriptions in the literature.

Phonological reconstructions done for each major parent node, checked against neighbouring varieties or those for which some other connection may become apparent.

Through these reconstructions, regular correspondences have been established based on which cognacy can be judged.



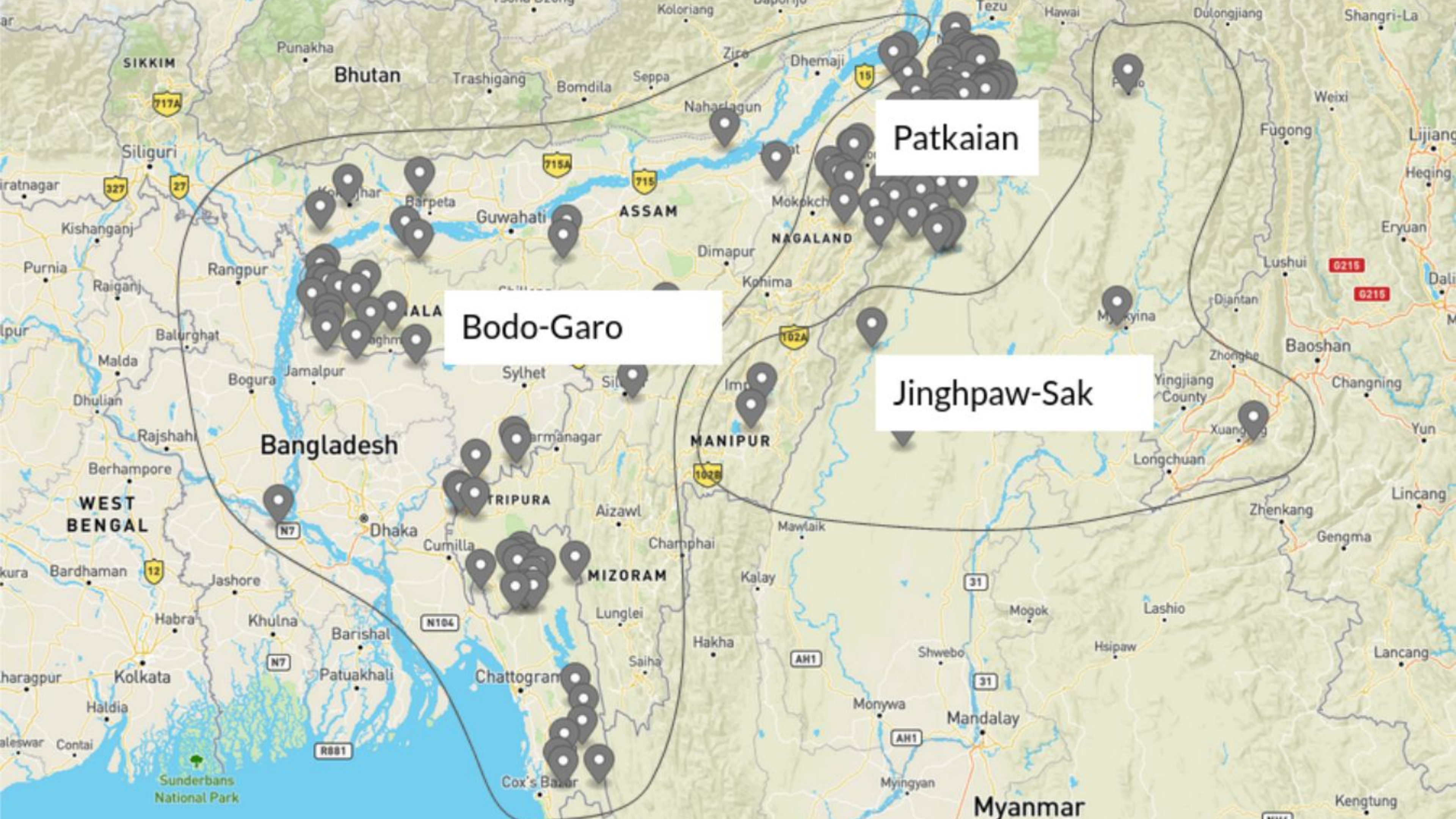

Tibet
Arunachal Pradesh
Bhutan
Assam
Kachin
Nagaland
Meghalaya
Manipur
Sagaing
Bangladesh
29°N
28°N
27°N
26°N
25°N
24°N
88°E
90°E
92°E
94°E
96°E
98°

Patkaian (Northern Naga)

Not closely related to the other “Naga” languages (Angami, Ao, Sumi etc)

Typical community size is around 2’000 speakers for most varieties, some are much larger: 60’000 each for Khiamniungan & Wancho

The most diverse branch within the proposed Sal family.

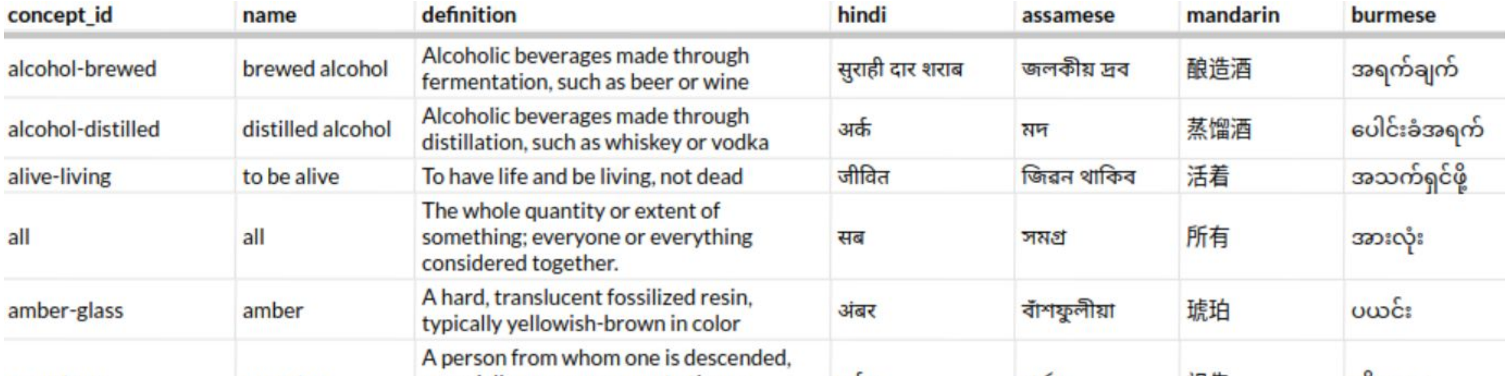


Patkaian
Bodo-Garo
Jinghpaw-Sak
SIKKIM
Bhutan
Punakha
Trashigang
Bomdila
Seppa
Ziro
Koloriang
Dhemaji
Tezu
Hawai
Dulongjiang
Shangri-La
Naharlagun
Siliguri
Biratnagar
Kishanganj
Purnia
Raiganj
Rangpur
Barpeta
Guwahati
ASSAM
Mokokchung
NAGALAND
Dimapur
Kohima
Fugong
Lijiang
Heqing
Weixi
Eryuan
Lushui
Diantan
Myitkyina
Baoshan
Zhonghe
Changning
Yingjiang County
Xuanggang
Longchuan
Lincang
Zhenkang
Gengma
Malda
Balurghat
Bogura
Jamalpur
Dhulian
Rajshahi
Bangladesh
Sylhet
Silchar
Imphal
MANIPUR
Dharmanagar
TRIPURA
Aizawl
Champhai
Mawlaik
Berhampore
WEST BENGAL
Dhaka
Cumilla
MIZORAM
Kalay
Mogok
Lashio
Bardhaman
Jashore
Khulna
Barishal
Lunglei
Hakha
Saiha
Shwebo
Hsipaw
Lancang
Habra
Kolkata
Patuakhali
Chattogram
Monywa
Mandalay
Kharagpur
Haldia
Contai
Sunderbans National Park
Cox's Bazar
Myingyan
Myanmar
Kengtung
717A
327
27
715A
715
15
102A
102B
12
N7
N104
R881
AH1
31
G215

data collection & methodology

Data collection

- **~750 concepts** covering **~175 doculects** of Patkaian
- Concepts derived from the CALMSEA word list (Matisoff 1978), the Intercontinental Dictionary Series word list (Key & Comrie 2023), and a large number of additional concepts which are widely attested in the doculects, forming the **SALIST** (Sal Area Lexical Inventory for Sino-Tibetan)



concept_id	name	definition	hindi	assamese	mandarin	burmese
alcohol-brewed	brewed alcohol	Alcoholic beverages made through fermentation, such as beer or wine	सुराही दार शराब	জলকীয় দ্ৰব	酿造酒	အရက်ချက်
alcohol-distilled	distilled alcohol	Alcoholic beverages made through distillation, such as whiskey or vodka	अर्क	মদ	蒸馏酒	ပေါင်းခံအရက်
alive-living	to be alive	To have life and be living, not dead	जीवित	জিৱন থাকিব	活着	အသက်ရှင်ဖို့
all	all	The whole quantity or extent of something; everyone or everything considered together.	सब	সমগ্ৰ	所有	အားလုံး
amber-glass	amber	A hard, translucent fossilized resin, typically yellowish-brown in color	अंबर	বাঁশফুলীয়া	琥珀	ပယင်း
		A person from whom one is descended,				

Data collection

- Only those terms which represent the most basic / typical word for a concept are included
- mini sketch grammars have been developed for each language regarding morphology, nominalisation, affixes in general
- Concepts with low coverage across branches were omitted in the end (< 8)

Cognacy & mesolanguages

- Cognacy is enforced through regular sound correspondences.
  - However, irregular sound changes are common

    *ɣap → shoot, by extension kick/propel

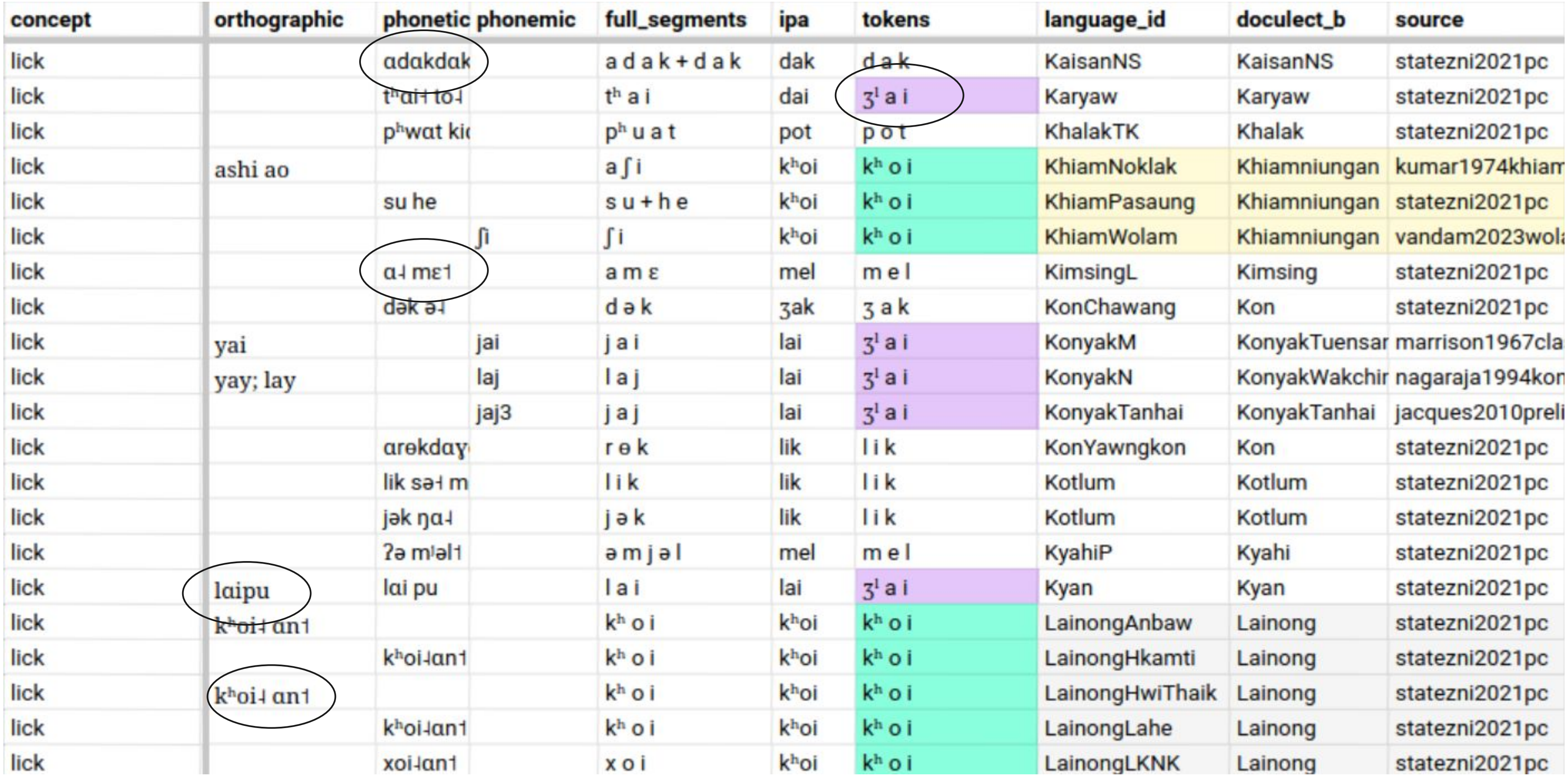
    ⁑k^{h}ɨŋ$_1$ → *k^{h}ɨŋ$_1$, of sky; *k^{h}ɨŋ$_3$ → of water
- Such irregularities require some variability be allowed in cognate judgements re strictness of sound changes
- all terms have been coded for cognacy at the **morpheme** level

Methodology

LingPy (List et al 2021) used for the creation of a MrBayes (Ronquist et al 2012) nexus file. Cognate assignment was still done manually.

Use of **MrBayes** Markov chain Monte Carlo method. No available means to calibrate a clock for Patkaian (or Sal more generally), so BEAST is not an option (as in Sagart et al 2019, Zhang et al 2019).

Amri Karbi (Konnerth, p.c; 2014) used as an outgroup, with additional Sal-internal outgroups for sub-branch trees (e.g. for determining three-way split within Patkaian)

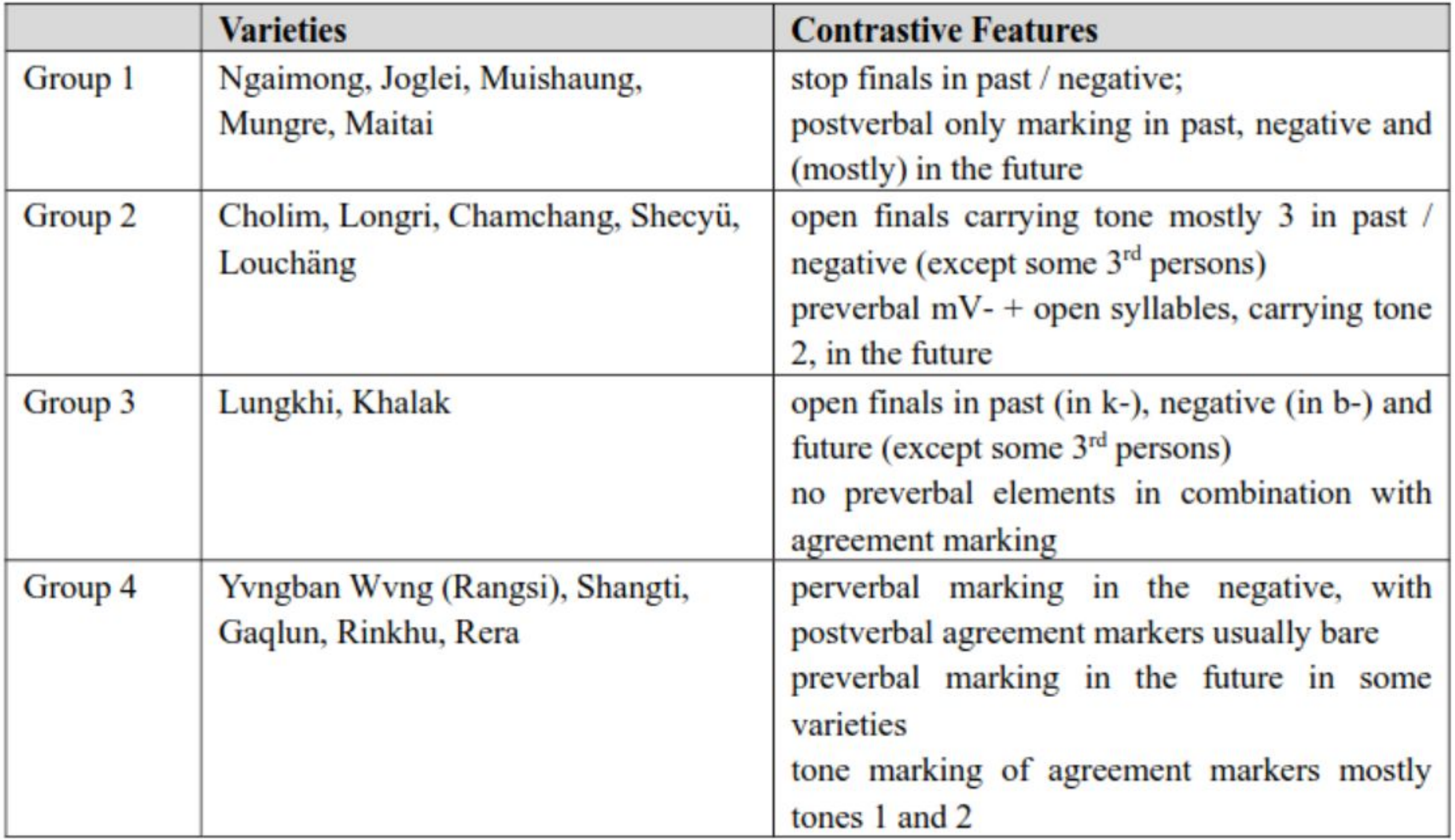


concept	orthographic	phonetic	phonemic	full_segments	ipa	tokens	language_id	doculect_b	source
lick		ɑdɑkdɑk		a d a k + d a k	dak	d a k	KaisanNS	KaisanNS	statezni2021pc
lick		tʰai˧ to˨		tʰ a i	dai	ʒˡ a i	Karyaw	Karyaw	statezni2021pc
lick		pʰwɑt ki		pʰ u a t	pot	p o t	KhalakTK	Khalak	statezni2021pc
lick	ashi ao			a ʃ i	kʰoi	kʰ o i	KhiamNoklak	Khiamniungan	kumar1974khiam
lick		su he		s u + h e	kʰoi	kʰ o i	KhiamPasaung	Khiamniungan	statezni2021pc
lick			ʃì	ʃ i	kʰoi	kʰ o i	KhiamWolam	Khiamniungan	vandam2023wola
lick		ɑ˨ mɛ˥		a m ɛ	mel	m e l	KimsingL	Kimsing	statezni2021pc
lick		dək ə˨		d ə k	ʒak	ʒ a k	KonChawang	Kon	statezni2021pc
lick	yai		jai	j a i	lai	ʒˡ a i	KonyakM	KonyakTuensar	marrison1967cla
lick	yay; lay		laj	l a j	lai	ʒˡ a i	KonyakN	KonyakWakchir	nagaraja1994kon
lick			jaj3	j a j	lai	ʒˡ a i	KonyakTanhai	KonyakTanhai	jacques2010preli
lick		ɑrɵkdɑy		r ɵ k	lik	l i k	KonYawngkon	Kon	statezni2021pc
lick		lik sə˧ m		l i k	lik	l i k	Kotlum	Kotlum	statezni2021pc
lick		jək ŋɑ˨		j ə k	lik	l i k	Kotlum	Kotlum	statezni2021pc
lick		ʔə mʲəl˥		ə m j ə l	mel	m e l	KyahiP	Kyahi	statezni2021pc
lick	lɑipu	lɑi pu		l a i	lai	ʒˡ a i	Kyan	Kyan	statezni2021pc
lick	kʰoi˨ ɑn˥			kʰ o i	kʰoi	kʰ o i	LainongAnbaw	Lainong	statezni2021pc
lick		kʰoi˨ɑn˥		kʰ o i	kʰoi	kʰ o i	LainongHkamti	Lainong	statezni2021pc
lick	kʰoi˨ ɑn˥			kʰ o i	kʰoi	kʰ o i	LainongHwiThaik	Lainong	statezni2021pc
lick		kʰoi˨ɑn˥		kʰ o i	kʰoi	kʰ o i	LainongLahe	Lainong	statezni2021pc
lick		xoi˨ɑn˥		x o i	kʰoi	kʰ o i	LainongLKNK	Lainong	statezni2021pc

a mini-sketch grammar has been worked out for each variety covering word formation & basic morphology

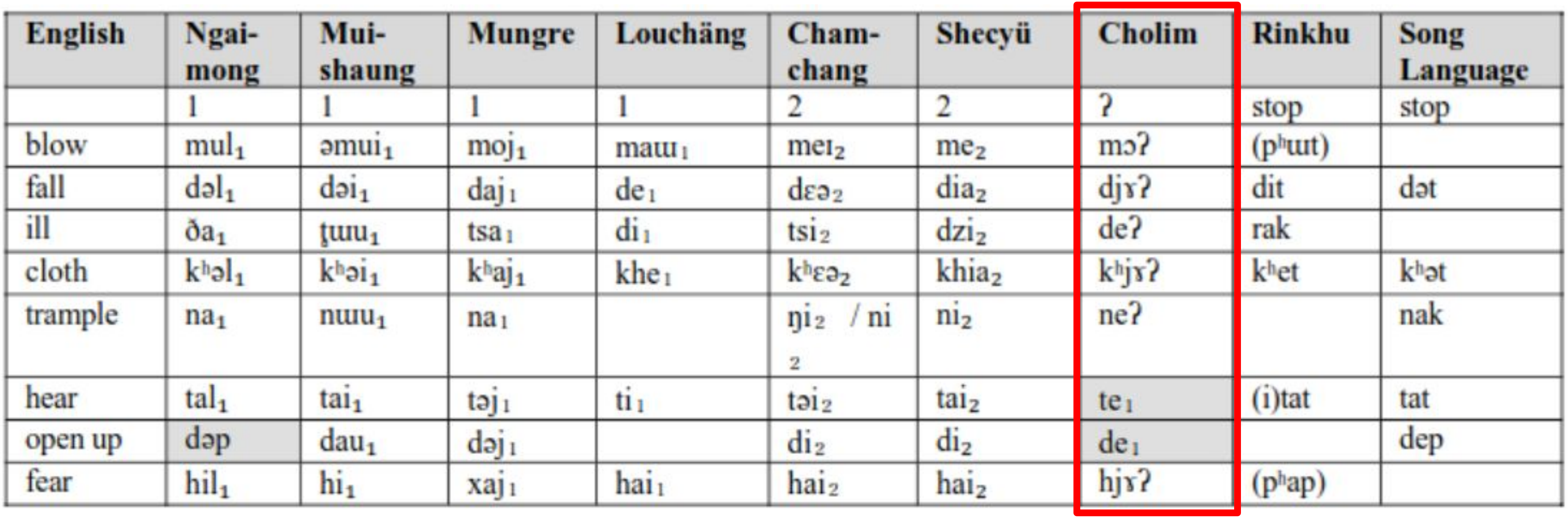
Results

Morey, S., 2019. Pangwa Tangsa agreement markers and verbal operators. *Himalayan Linguistics*, *18*(1).

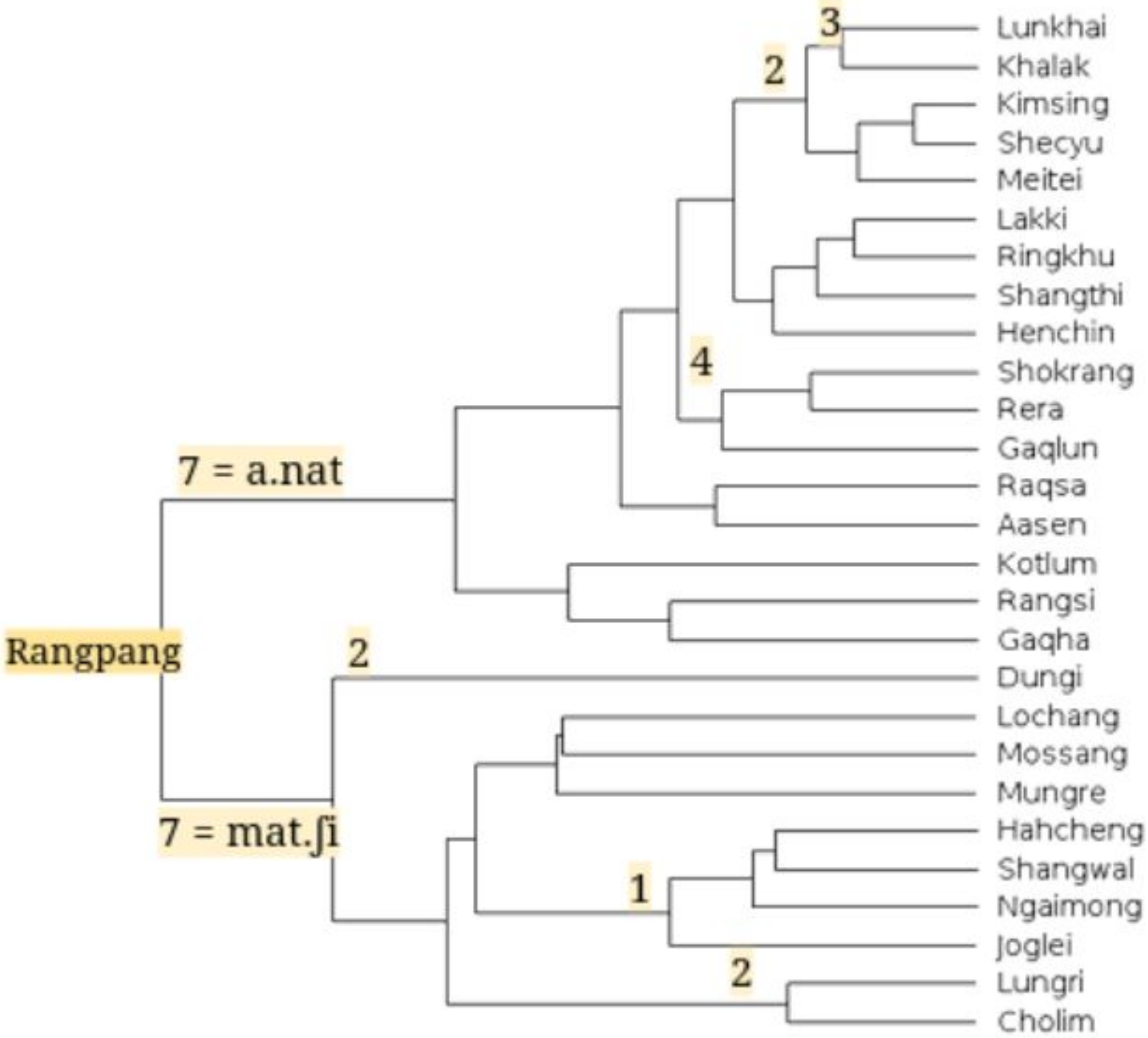


	Varieties	**Contrastive Features**
Group 1	Ngaimong, Joglei, Muishaung, Mungre, Maitai	stop finals in past / negative; postverbal only marking in past, negative and (mostly) in the future
Group 2	Cholim, Longri, Chamchang, Shecyü, Louchäng	open finals carrying tone mostly 3 in past / negative (except some 3rd persons) preverbal mV- + open syllables, carrying tone 2, in the future
Group 3	Lungkhi, Khalak	open finals in past (in k-), negative (in b-) and future (except some 3rd persons) no preverbal elements in combination with agreement marking
Group 4	Yvngban Wvng (Rangsi), Shangti, Gaqlun, Rinkhu, Rera	perverbal marking in the negative, with postverbal agreement markers usually bare preverbal marking in the future in some varieties tone marking of agreement markers mostly tones 1 and 2

Morey, S., 2019. Pangwa Tangsa agreement markers and verbal operators. *Himalayan Linguistics*, *18*(1).



English	Ngai-mong	Mui-shaung	Mungre	Louchäng	Cham-chang	Shecyü	Cholim	Rinkhu	Song Language
	1	1	1	1	2	2	ʔ	stop	stop
blow	mul$_1$	əmui$_1$	moj$_1$	maɯ$_1$	meɪ$_2$	me$_2$	mɔʔ	(p^{h}ɯt)	
fall	dəl$_1$	dəi$_1$	daj$_1$	de$_1$	dɛə$_2$	dia$_2$	djɤʔ	dit	dət
ill	ða$_1$	ʈɯu$_1$	tsa$_1$	di$_1$	tsi$_2$	dzi$_2$	deʔ	rak	
cloth	k^{h}əl$_1$	k^{h}əi$_1$	k^{h}aj$_1$	khe$_1$	k^{h}ɛə$_2$	khia$_2$	k^{h}jɤʔ	k^{h}et	k^{h}ət
trample	na$_1$	nɯu$_1$	na$_1$		ŋi$_2$ / ni$_2$	ni$_2$	neʔ		nak
hear	tal$_1$	tai$_1$	təj$_1$	ti$_1$	təi$_2$	tai$_2$	te$_1$	(i)tat	tat
open up	dəp	dau$_1$	dəj$_1$		di$_2$	di$_2$	de$_1$		dep
fear	hil$_1$	hi$_1$	xaj$_1$	hai$_1$	hai$_2$	hai$_2$	hjɤʔ	(p^{h}ap)	

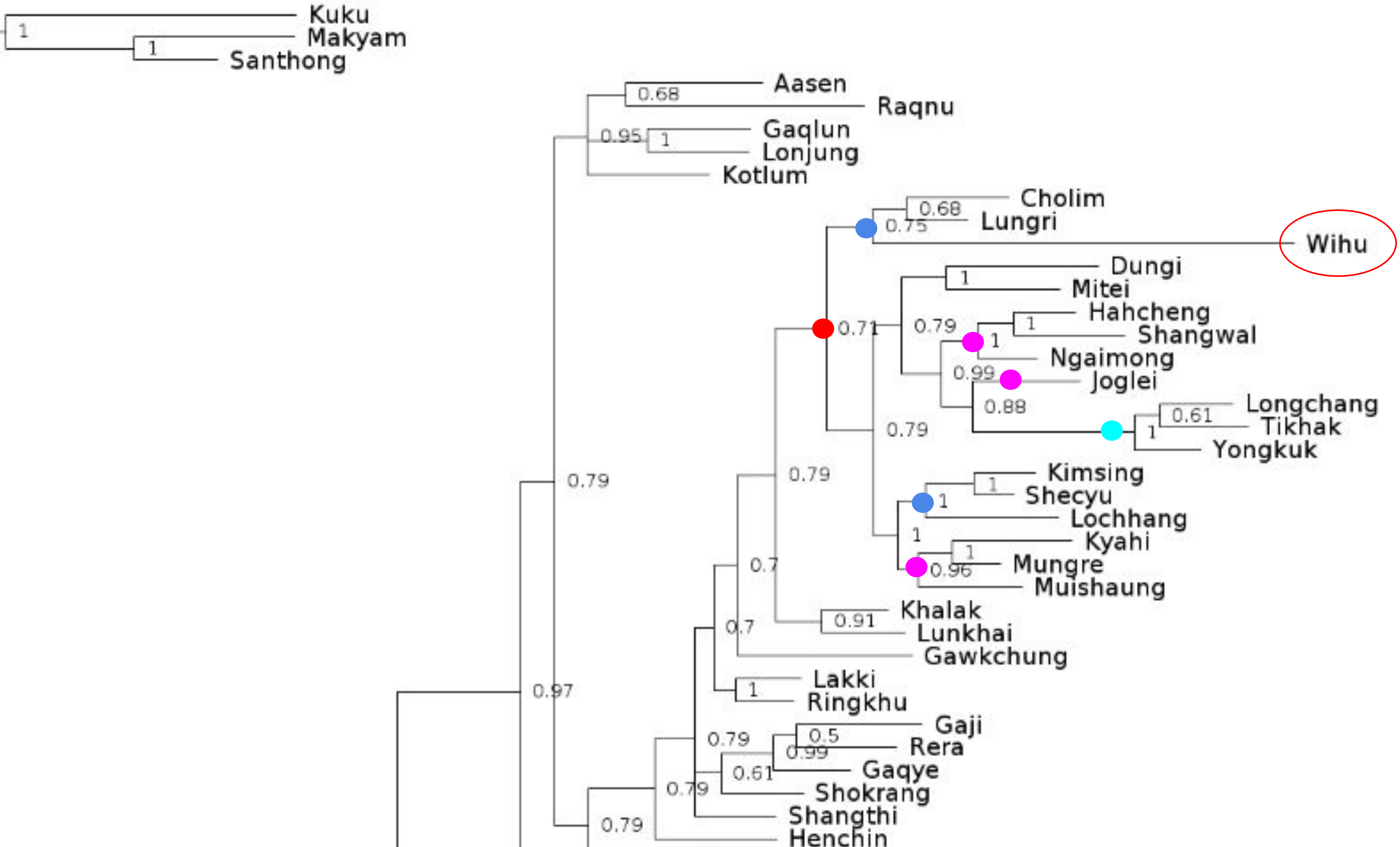


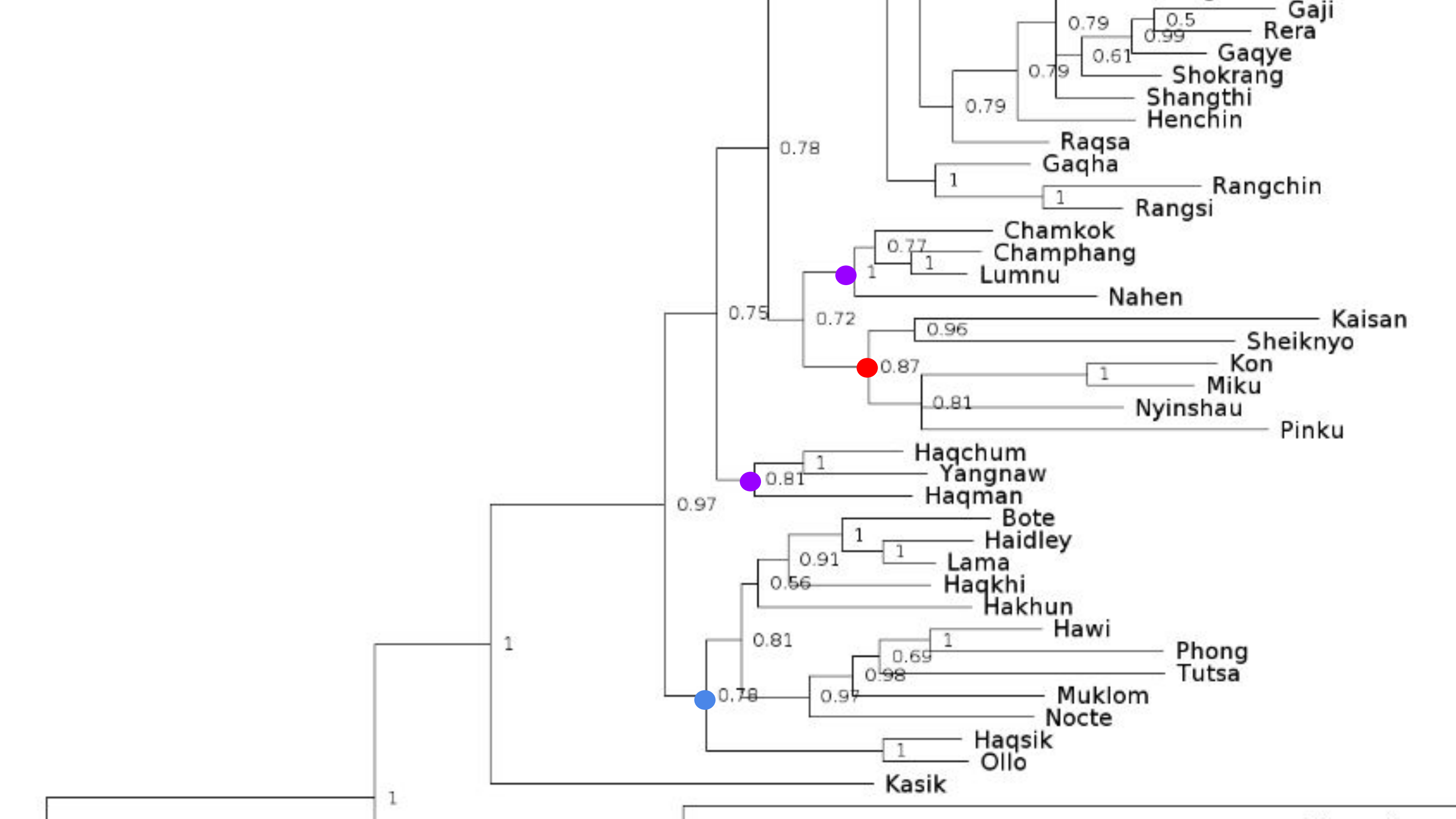
Morey’s groups correspond to phonological & morphological features.

Many of these may have other explanations, e.g.:

- common sound changes for phonological features
- loss of productive *-ʔ nominaliser for tonal differences
- speculatively: Jesperson’s cycle for negation placement, or IRR (vs FUT) < NEG, or just well attested innovation in TB negation marking (DeLancey 2015)
- esoterogeny!

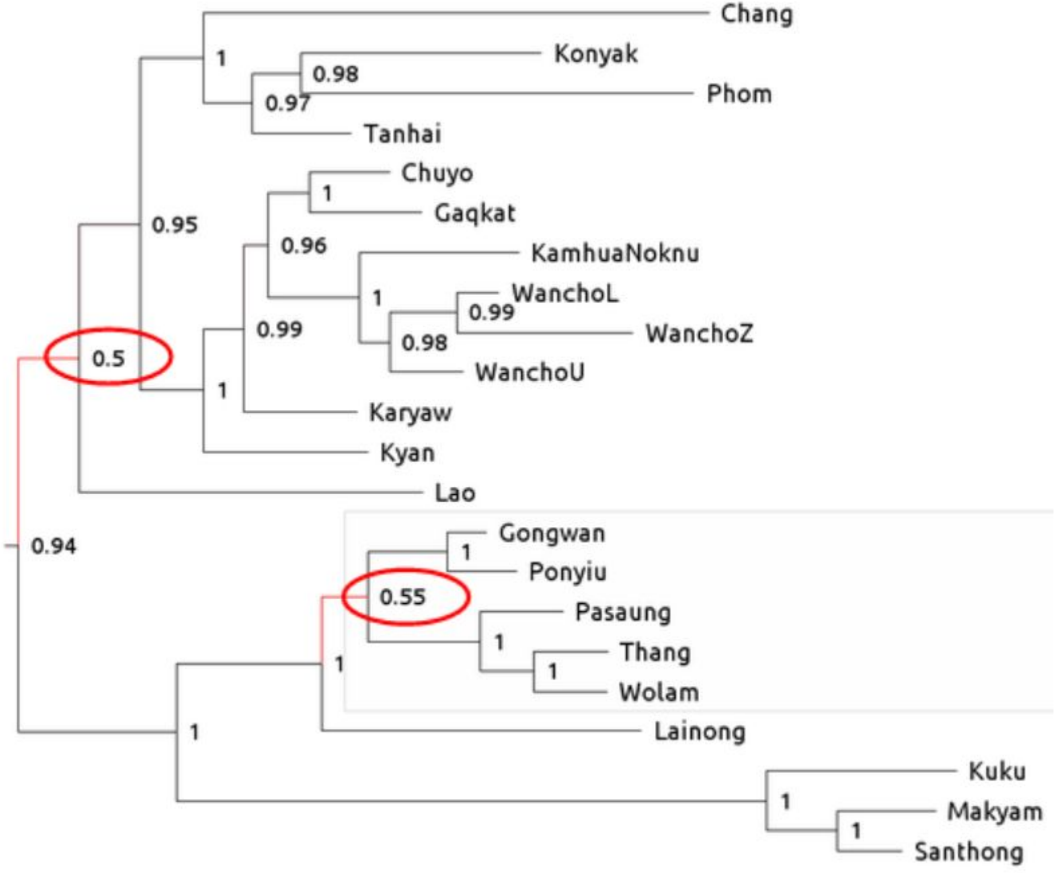
etc.


Kuku
Makyam
Santhong
1
1
Aasen
Raqnu
0.68
Gaqlun
Lonjung
0.95
1
Kotlum
Cholim
Lungri
0.68
0.75
Wihu
Dungi
Mitei
1
Hahcheng
Shangwal
1
0.71
0.79
1
Ngaimong
Joglei
0.99
0.88
Longchang
Tikhak
Yongkuk
0.61
1
0.79
Kimsing
Shecyu
1
1
Lochhang
1
Kyahi
Mungre
1
0.96
Muishaung
0.79
0.7
Khalak
Lunkhai
0.91
0.7
Gawkchung
0.79
0.97
Lakki
Ringkhu
1
Gaji
Rera
0.5
0.99
0.79
Gaqye
0.61
Shokrang
0.79
Shangthi
Henchin
0.79

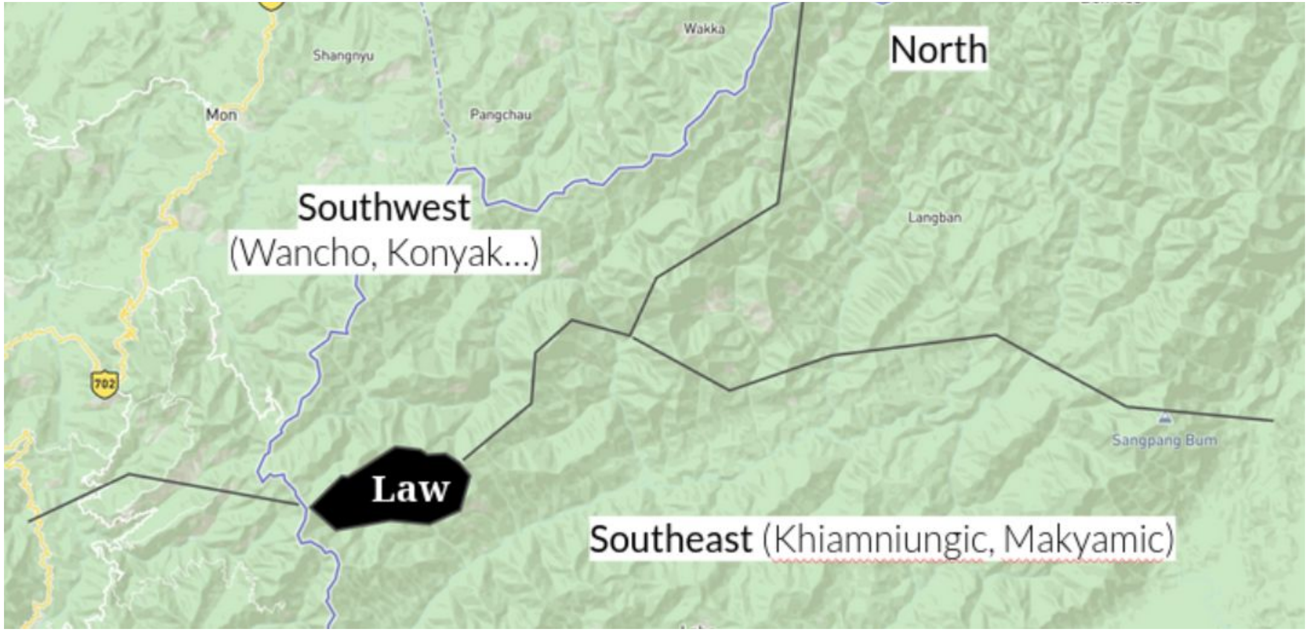

0.79
0.5
Gaji
Rera
0.99
0.61
Gaqye
0.79
Shokrang
Shangthi
0.79
Henchin
0.78
Raqsa
Gaqha
1
Rangchin
1
Rangsi
Chamkok
0.77
Champhang
1
1
Lumnu
Nahen
0.75
0.72
Kaisan
0.96
Sheiknyo
Kon
0.87
1
Miku
0.81
Nyinshau
Pinku
Haqchum
1
Yangnaw
0.81
Haqman
0.97
Bote
1
Haidley
1
0.91
Lama
0.56
Haqkhi
Hakhun
1
Hawi
0.81
1
Phong
0.69
0.98
Tutsa
0.78
0.97
Muklom
Nocte
Haqsik
1
Ollo
Kasik
1

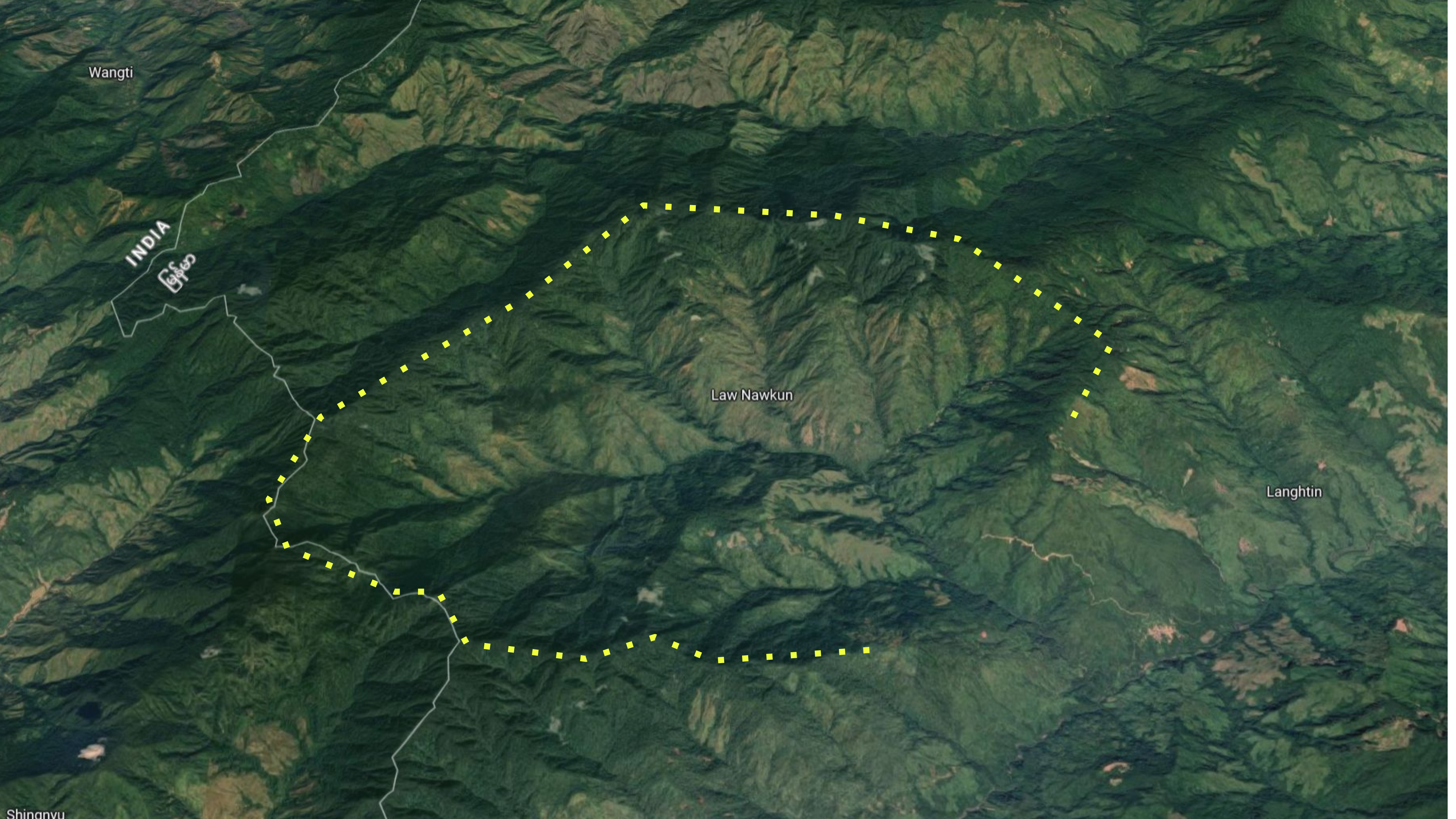
Southern branches

Law and Ponyiu-Gongwan placement has interesting complications



Southern branches - Law




Wangti
INDIA
မြန်မာ
Law Nawkun
Langhtin
Shingnyu

Conclusions

General conclusions

Many conventional groupings often do now hold up to scrutiny, rely more on administrative boundaries etc.

Often the result of modern (not historical) geographic proximity

Bayesian methods aren’t always well applied, and may not always give the answers, but can often point us in important and unconsidered directions.

By combination of phonological **reconstruction**, elicitation **consistency**, incorporation of traditional migration **narratives** to reconstruct potential historical **contact** with particular focus on **geography**, and an understanding of word formation to control for elicitation inconsistencies, such methods can be of great value.

Much more care is required than is often taken.

General conclusions

Bayesian methods aren't always well applied, and may not always give the answers even if done well.

The methods can still point us in directions we may otherwise miss, while also helping reduce **some** forms of researcher bias.

Ideally:

- phonological **reconstruction**, at all major branches
- consistent **elicitation** methods
- incorporation of migration **narratives** & historical **contact**
- an accounting of **geography** factors

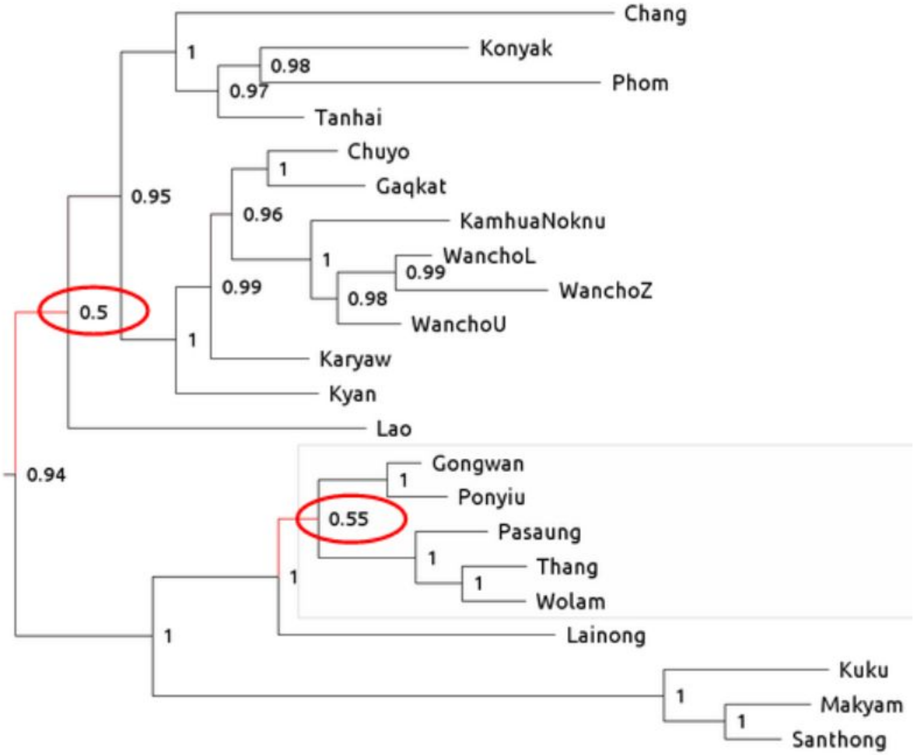
Much more care is required than is often taken.

Specific conclusions

“Noctean” as proposed by van Dam & Rahman (2019; 2021) is viable and includes Tutsa, Ollo, Hakhun & Muklom

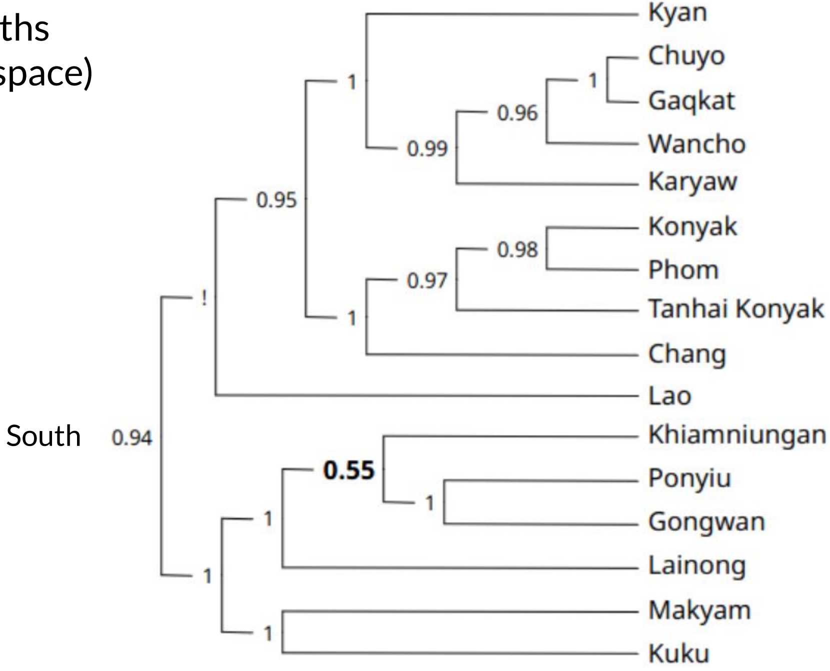
Champhangic as including Haqchum is not viable

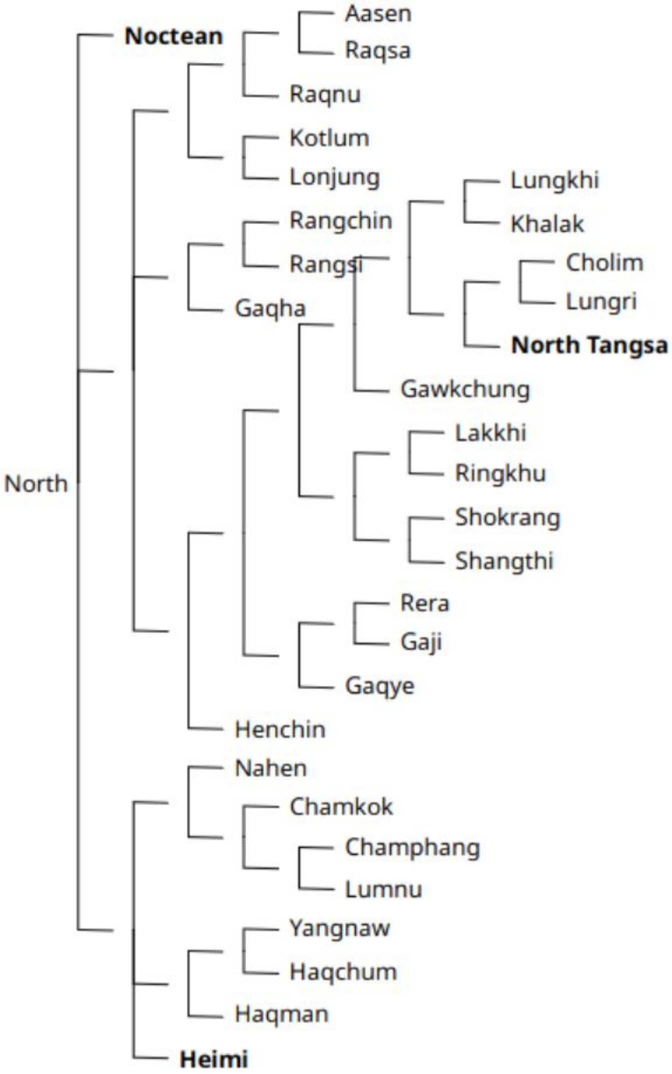
Khasik is properly placed in the northern branch, not Southwest, but as an earliest branch off Noctean

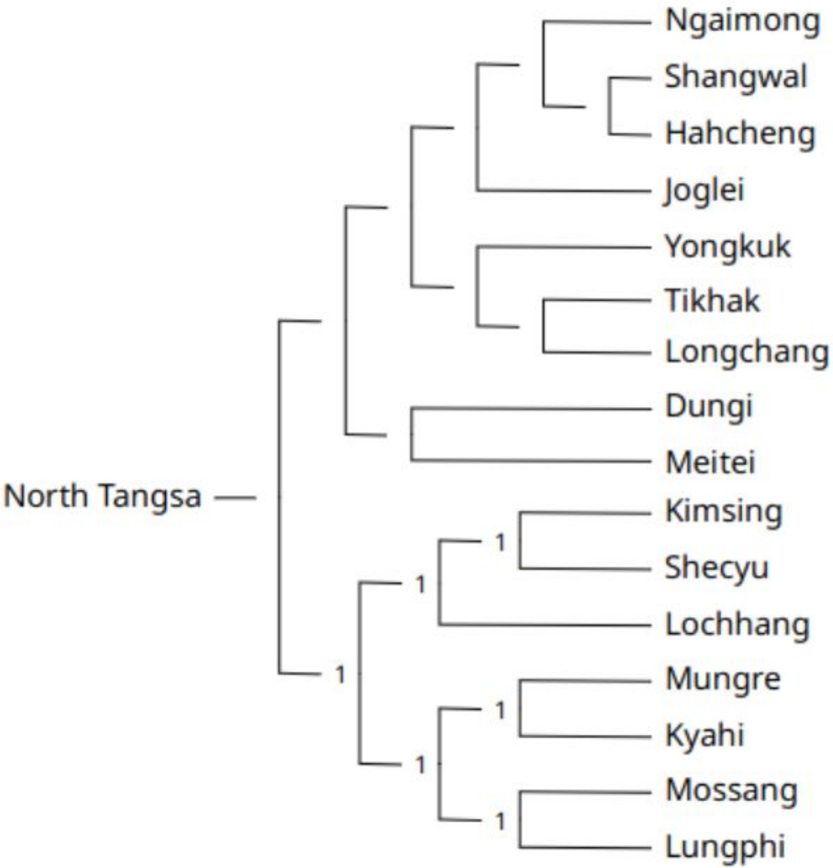


The tree(s)

(branch lengths omitted for space)







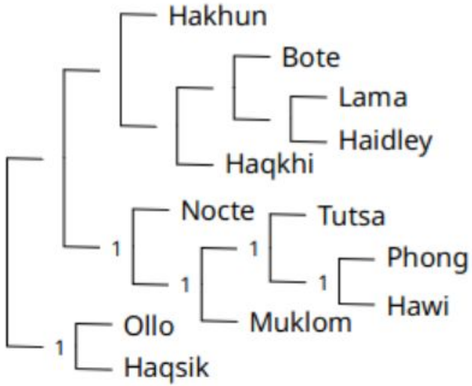


Figure 3: Noctean

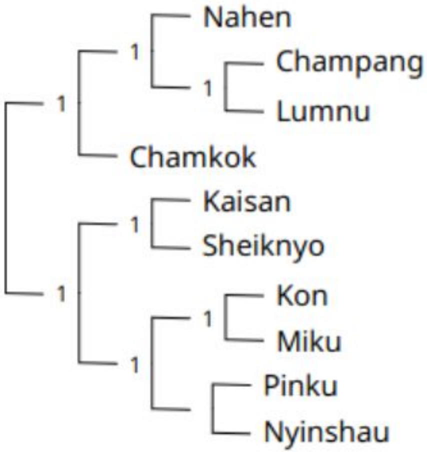


Figure 4: Heimi

Next steps

- Additional filtering by coverage

- Additional data collection for low-coverage varieties

- Collection of data for entirely missing varieties, including other ‘liturgical’ language doculects

tʰaimi hai (Wolam Khiamniungan)
ketʒu əʒuŋ (Muishaung Tangsa)
ɲem pʰoi (Kamhua Noknu Wancho)
thank you (West Michigan American English)

contact: **kellenparker@gmail.com**

Special thanks:
Natalia Chousou-Polydouri (UZH), **Dipjyoti Goswami** (La Trobe University), **Keen Thaam** (Nagaland University), **Anui Sainu** (Myanmar), **Linda Konnerth** (Bern), **Nathan Statezni** (SIL), **Stephen Morey** (La Trobe University)

References

- DeLancey, Scott. 2015. The origins of postverbal negation in Kuki-Chin. North East Indian Linguistics 7, 203-212. Canberra, Australian National University: Asia-Pacific Linguistics Open Access.
- Konnerth, L.A., 2014. A grammar of Karbi. University of Oregon. // — 2023, personal communication
- List, Johann-Mattis and Forkel, Robert (2021): LingPy. A Python library for historical linguistics. Version 2.6.9. URL: https://lingpy.org, DOI: https://zenodo.org/badge/latestdoi/5137/lingpy/lingpy. With contributions by Greenhill, Simon, Tresoldi, Tiago, Christoph Rzymski, Gereon Kaiping, Steven Moran, Peter Bouda, Johannes Dellert, Taraka Rama, Frank Nagel. Leizpig: Max Planck Institute for Evolutionary Anthropology.
- Morey, S., 2019. Pangwa Tangsa agreement markers and verbal operators. Himalayan Linguistics, 18(1).
- Orlandi, G., 2021. Once again on the history and validity of the Sino-Tibetan bifurcate model. Journal of Language Relationship, 19(3-4), pp.263-292.
- Ronquist, F., M. Teslenko, P. van der Mark, D.L. Ayres, A. Darling, S. Höhna, B. Larget, L. Liu, M.A. Suchard, and J.P. Huelsenbeck. 2012. MRBAYES 3.2: Efficient Bayesian phylogenetic inference and model selection across a large model space. Syst. Biol. 61:539-542.
- Sagart, L., Jacques, G., Lai, Y., Ryder, R.J., Thouzeau, V., Greenhill, S.J. and List, J.M., 2019. Dated language phylogenies shed light on the ancestry of Sino-Tibetan. Proceedings of the National Academy of Sciences, 116(21), pp.10317-10322.
- Zhang, M., Yan, S., Pan, W. and Jin, L., 2019. Phylogenetic evidence for Sino-Tibetan origin in northern China in the Late Neolithic. Nature, 569(7754), pp.112-115.