Undocumented labor:
How old fieldwork sheds new light on Tai tone system diversification

RIKKER DOCKUM, YALE UNIVERSITY

92nd Annual Meeting of the Linguistic Society of America

4-7 January 2018, Salt Lake City

BCS-1528386

Yale MACMILLAN CENTER

Grey Literature, Grey Data

What is 'grey literature'?

- In a nutshell: hard-to-find research output & data resources
- Exist outside of traditional publishing and distribution channels
  - Organization reports
  - Working papers
  - Government publications
  - White papers
  - Dissertations and theses

Grey Literature, Grey Data

Extended definition of grey literature for linguistics:

Only available in typically **disjoint** publishing and distribution channels

- Research output in languages other than English/European lgs
  - Engagement by English-speaking academia varies depending on the language
- Research output in countries that engage with English-speaking academia less frequently
  - If ever presented at an international conference, usually area focused e.g. SEALS
  - Students of faculty trained in Western academia who return home to teach
- Government-sponsored research in other countries
  - Language and dialect surveys
  - Often available only in physical libraries in distant places, or in obscure online repositories, or exclusively to local citizens (gov't ID number required to access)

Grey Literature, Grey Data

Challenges to discovery – physical access

- Single copy may exist
- Remote universities

Challenges to discovery – digital discovery

- Unnavigable library catalogs
- Inconsistent subject heading tagging by librarians
- Data siloing – theses in separate portion of the catalog

Challenges to discovery – digital access

- Restricted permissions for PDF downloads
- Low-resolution scans, extremely compressed/artifacty

Bringing Thailand's linguistic grey data to a wider audience

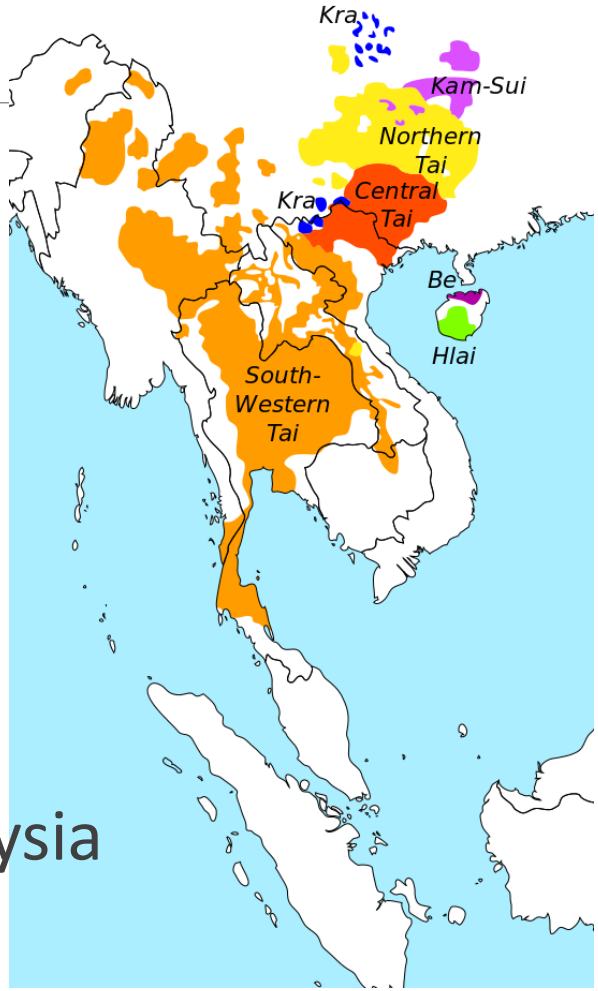
Avenues for discovery and access:

- Thai national research repository (indexing)
- Individual university efforts (thesis scanning projects)
- Open digital access on the web (some universities)
- On-site only digital access (other universities)
- Paper copies photocopied and scanned

While my efforts focused on Kra-Dai languages and dialects (the majority of Thailand's language documentation output), there are also many theses from other language families, particularly Austroasiatic, Austronesian, Hmong-Mien, and Sino-Tibetan

Kra-Dai

- ~100 languages
- ~100 million speakers (E&S 2008)
  - Thailand 60+ million
  - China 15+ million
  - Laos 3 million
  - Burma 3 million
  - India, Vietnam, Cambodia, Malaysia



A half century of Tai dialectology theses in Thailand (1969-2013)

Documenting tone systems:

Akharawatthanakun 1998, 2003; Anusurain 1998; Aruneeung 1990; Awirutthiyothin 2010; Banditkul 1993; Bunmee 2007; Canilao 2010; Chaimano 2009; Chinchest 1989; Debavalya 1983; Hanpanich 1992; Kantong 2007; Kewkasem 2003; Khamrueangsi 2002; Khemkhaeng 2002; Khotchanthuek 2002; Khumdee 2000; Kitivongprateep 2005; Kobsirikarn 1992; Komontha 1996; Kongthong 2006; Koowatthanasiri 1981; Kopprayun 1986; Krisnapan 1995; Lertthana 2005; Malaichalern 1988; Namwang 2001; Nasanee 2002; Nualjansaeng 1992; Panroj 1991; Pintasaard 2004; Pornsib 1994; Prapaipet 1989; Pratankiet 2001; Ratanadilok Na Phuket 1983; Saeng-ngam 2006; Sawangwan 1991; Sitthi 2006; Sittiprapaporn 1997; Soiyana 2009; Sritararat 1983; Taengko 1987; Tanlaput 1988; Tanprasert 2003; Thawarorit 2006; Tingsabadh 1990; Worawong 2000; Yooyen 2013

More general documentation:

Ache 1986; Ampornpan 1986; Angsuwiriya 2003; Arpakul 1995; Beadnok 1989; Bencha 2000; Boonabha 1969; Boonkao 1989; Boonsawasd 2012; Boonsner 1984; Buranasing 1988; Chai-arun 1998; Chanaingoon 1970; Chanavong 1980; Chaokhamin 1988; Charoenphol 1985; Charoenvalaya 1991; Chativong 1986; Chawsuan 1994; Choophan 2004; Chotecheun 1986; Chulkeeree 1991; Chummalee 2010; Dumruks 1970; Eam-eium 1986; Hasonnary 2000; Jantanakom 1983; Jidlang 2012; Jitbanjong 2002; Junlawan 2011; Jurjanad 1987; Kamwachirapitak 2005; Khamboonchoo 1985; Khwanritti 1987; Kitprasert 1985; Kongsuwan 1988; Kummun 1992; Laksanawong 2008; Lamchiagdase 1984; Lengtai 2009; Mahaphunthong 1996; Maneewong 1987; Manoosawet 1993; Mapawongse 1979; Maryprasith 1992; Massupong 1982; Matchikanang 1999; Nakorn 2000; Nakpuntawong 1987; Narkphong 1982; Ninjinda 1989; Osatananda 1997; Paiboonwangcharoen 1984; Panarat 1990; Panka 1980; Patpong 1997; Peamphermphoon 1986; Petsuk 1978; Phantachat 1983; Pimpa 1986; Plodkaew 2008; Ploykaew 1985; Plungsuwan 1981; Poo-Israkij 1985; Poonpholwattanaporn 2010; Praphin 1996; Pumma 2003; Pungpawpun 1984; Punthong 1979; Rakmoh 2007; Rakpaet 1998, 2010; Ratanapraseart 1985; Rinprom 1977; Rittiwong 1997; Saeneetontikul 1985; Sakdanuwatwong 1995; Seangsrichan 1998; Senisrisant 1983; Shen 2003; Sila 1975; Siriwisitkun 1986; Sombatmaungkan 1990; Somnuk 1982; Soongsumaln 2002; Sornjitti 2007; Subcharoen 1989; Sukpiti 1989; Sukpreedee 1988; Sumransook 1995; Sungkep 1983; Sungvanthrup 1991; Suppasin 2011; Sutadarat 1978; Suwanmusik 2004; Suwanratt 1991; Tanyong 1983; Tebpawan 2012; Teeranuwat 2002; Thavorn 2013; Thianthaworn 1998; Thongmark 1983; Thongphiew 1989; Thongrat 1988; Thumsaro 1993; Tippol 1988; Tisapong 1985; Udomphan 2000; Unakornsawat 1993; Vaitayavanich 1991; Weesakul 1983; Wetchasit 1987; Withayasakpan 1979; Worachin 2009; Wuttheerapon 2004; Yensamut 1981; Yoojaroensuk 1991

Using the data

Challenges to use – general issues

- Comparability!
- Transcription quality/reliability/consistency
- Laborious data entry required

Challenges to use – tonal research

- Varying tone notation standards
  - Chao numerals
  - Chao letters
  - Prose descriptions (usually in Thai)
- Inconsistent notation of creaky voice, glottal constriction, etc.

Language change & lexical tone

Language change and lexical tone

- Much big picture historical work focuses on tonogenesis
- Some posited implicational universals
  - Hyman & Schuh 1974, Maddieson 1978, Hyman 2007, Cahill 2008, e.g.
    - If a language has contour tones, it will have level tones
    - If a language has complex contours, it has simple contours
    - Raising more common than lowering
    - Low tone vowels longer than high tone vowels
    - Rising tone vowels longer than falling tone vowels
- A complete theory of sound change for tone systems is still far from complete

Language change and lexical tone

Some things that bear repeating/emphasizing:

- Tones are born from the collapse of other distinctions
  - voicing quality of consonant (onset or coda)
  - laryngeal configuration of consonants (voicing, aspiration)
  - vowel length
- Tone paradigms change **as a system**
- Number of tone contrasts are not acquired linearly
- **Complexity** exhibits itself in ways beyond pitch
  - Murmur, creak, glottal closure, duration, f0 range, tone sandhi

Tai tone system diversification

Gedney 1972

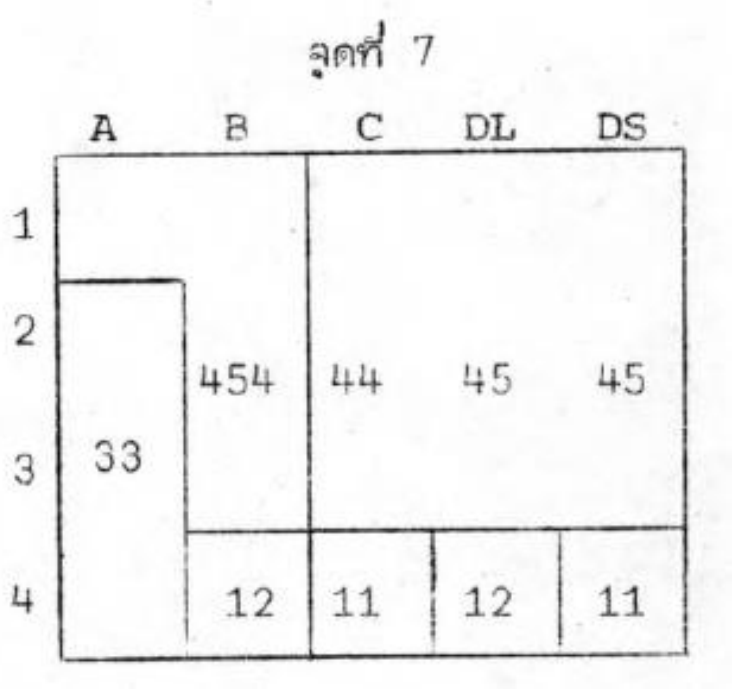
Historical tone category

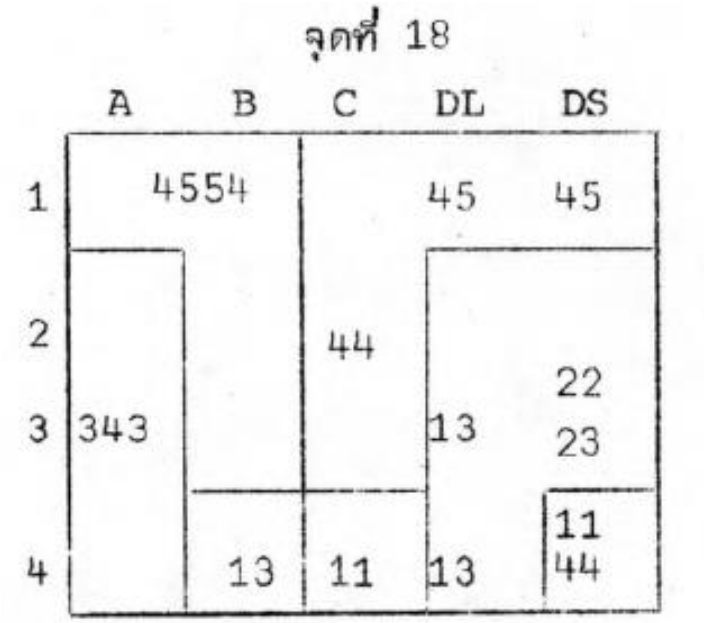
Proto-onsets	A	B	C	D-short	D-long
Voiceless w/ friction *pʰ, *tʰ, *kʰ, *s, *m̥, etc.	A1	B1	C1	DS1	DL1
Voiceless unaspirated *p, *t, *k, etc.	A2	B2	C2	DS2	DL2
Glottalized *ʔ, *ʔb, *ʔj, etc.	A3	B3	C3	DS3	DL3
Voiced *b, *m, *l, *z, etc.	A4	B4	C4	DS4	DL4
	Modal	Creaky	Glottal constriction		

(Pittayaporn 2009)

Tai tone system diversification

For each Tai doculect, we can use a diagnostic checklist (Gedney 1972) to determine how the tonogenetic scenario played out in that lect

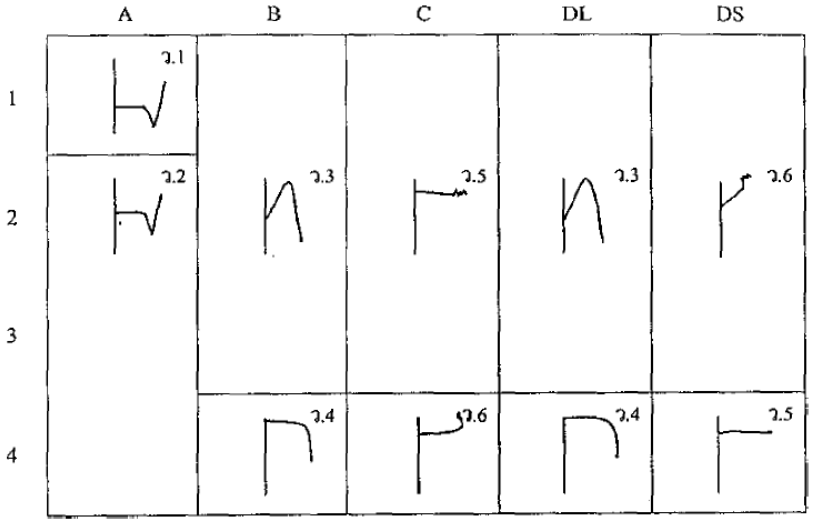




Southern Thai dialects, from Debavalya 1983

Tai tone system diversification

Variation in tone notation Gedney boxes abounds:



	A	B	C	DL	DS
1	˧.1	˧˥.3	˧˥.5	˧˥.3	˧.1
2					
3	˧.2				
4		˧.4	˧.6	˧.4	

Nyo, from Khamrueangsi 2002

Yong, from Soiyana 2009

Two avenues to explore today

And now, two brief case studies in how we can use large amounts of this liberated 'grey data' to generate new hypotheses and answer new research questions:

Question 1. Does more dialect data on Tai surface tones and their mappings to historical categories, give us evidence for historical surface tones?

Question 2. Is there statistically detectable phylogenetic signal in the series of historical tone splits and mergers? (For quantitative analysis)

Mapping surface tones to historical categories

Question 1

Does more dialect data on Tai surface tones and their mappings to historical categories, give us evidence for historical surface tones?

Note: Not every description includes both surface tone descriptions and the historical tone categories (or enough lexical data to identify them).

Data available for **279 doculects**

Mapping surface tones to historical categories

The Chao numerals are too noisy, coming from so many linguists

Reduced to the following notation:

L	Low level tone (Chao: 11, 22)
M	Mid level tone (Chao: 33)
H	High level tone (Chao: 44, 55)
F	Falling contour (Chao: 53, 51, 42, etc.)
R	Rising contour (Chao: 13, 15, 24, etc.)
ʔ, X	Leave marking of glottal constriction/creak in place

Mapping surface tones to historical categories

Complex contours (3-4 Chao numerals)

FL	Fall followed by low level (Chao: 422, 533, etc)
RH	Rise followed by high level (Chao: 455, 133, etc)
MR	Mid level tone with rise (Chao: 334, 335)
RM	Rise followed by mid level (Chao: 133, 233)
RF	Rise followed by fall (Chao: 253, 342, etc)

Next: Compare the proportion of each contour category with each cell of the Gedney box

Mapping surface tones to historical categories

Two result highlights (from 279 doculects):

- **88%** of modern surface tones for the A1 cell of the Gedney box have a rising component
  - A1 = modal voice + voiceless “friction” onset (fricatives, aspirated stops, devoiced liquids and nasals)
  - No other tone type compared with any cell of the paradigm exceeded 54%

- Of 122 doculects with glottal constriction notated on at least one modern tone, **64%** have it on reflexes of the proto C tone **and only** on reflexes of that tone.
  - More support for reconstruction by Pittayaporn 2009
  - That proportion could rise if creak and glottal constriction are more carefully teased apart through close reading of the source material

Phylogenetic signal in tone splits and mergers

Question 2

Is there statistically detectable phylogenetic signal in the series of tone splits and mergers (ignoring surface tones)?

Data available for **362 doculects**

Phylogenetic signal in tone splits and mergers

	A1	A2	A3	A4	B1	B2	B3	B4	C1	C2	C3	C4
A1												
A2	158											
A3	136	340										
A4	8	144	166									
B1	48	4	5	8								
B2	42	10	11	10	353							
B3	41	10	11	10	354	361						
B4	1	19	21	5	58	59	59					
C1	3	1	1	1	34	34	34	76				
C2	0	1	1	2	3	3	3	98	271			
C3	0	1	1	2	1	1	1	100	266	357		
C4	0	1	1	3	1	1	1	33	18	91	95	

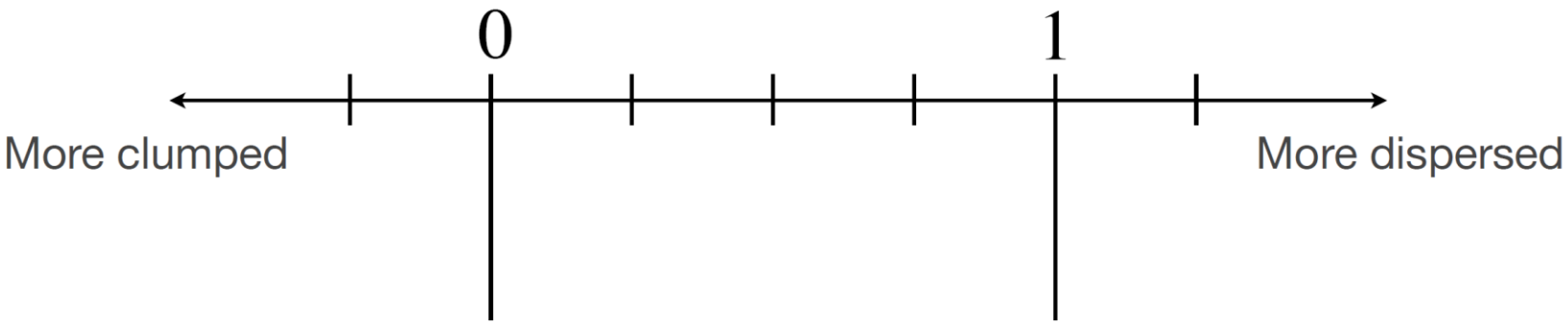
Pairwise comparison

Over 362 Tai doculects, how often were each pair of historical tone categories merged?

How often not?

(Note: these are simple counts, controlling for nothing)

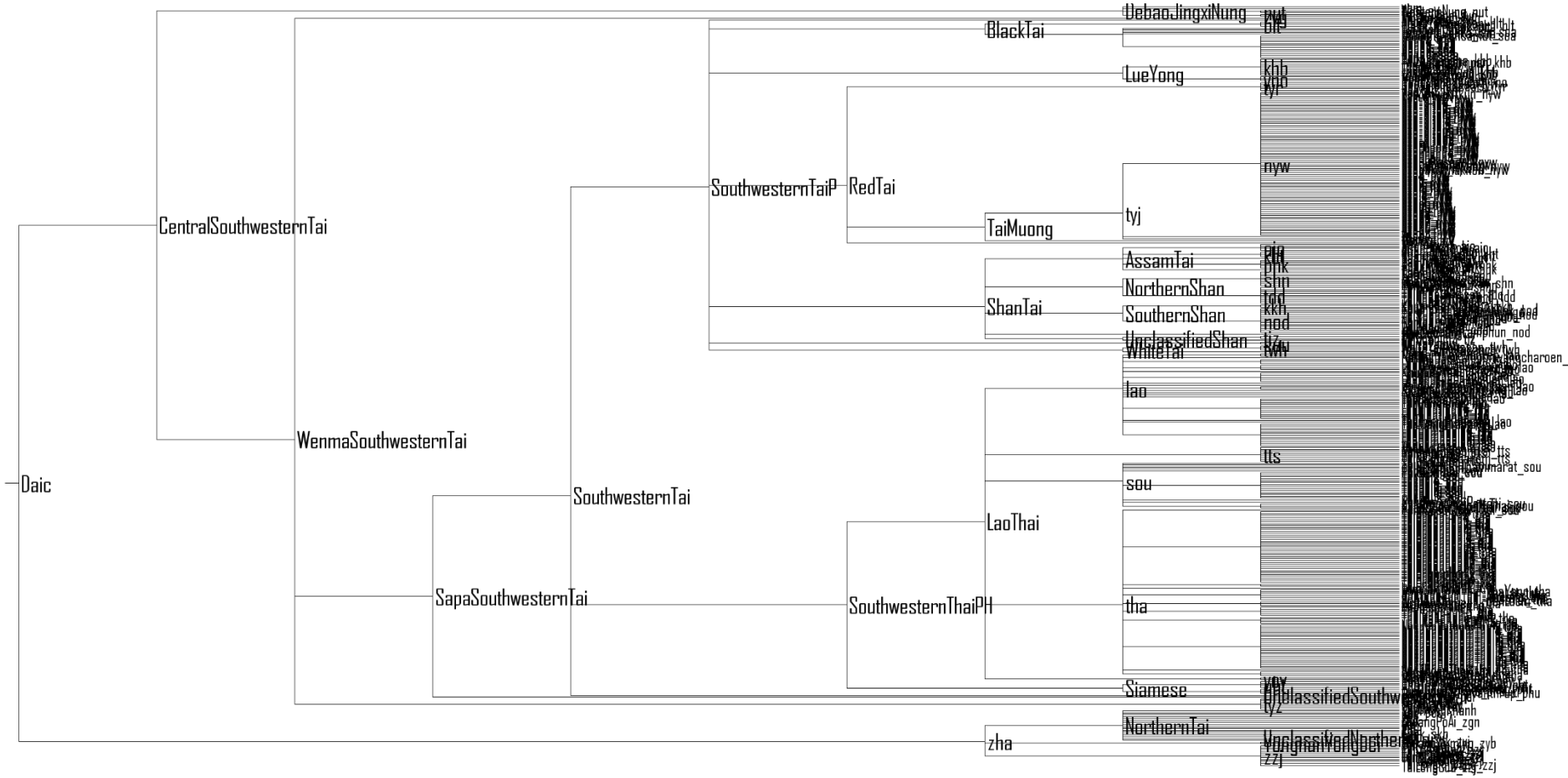
D test (Fritz & Purvis 2010)



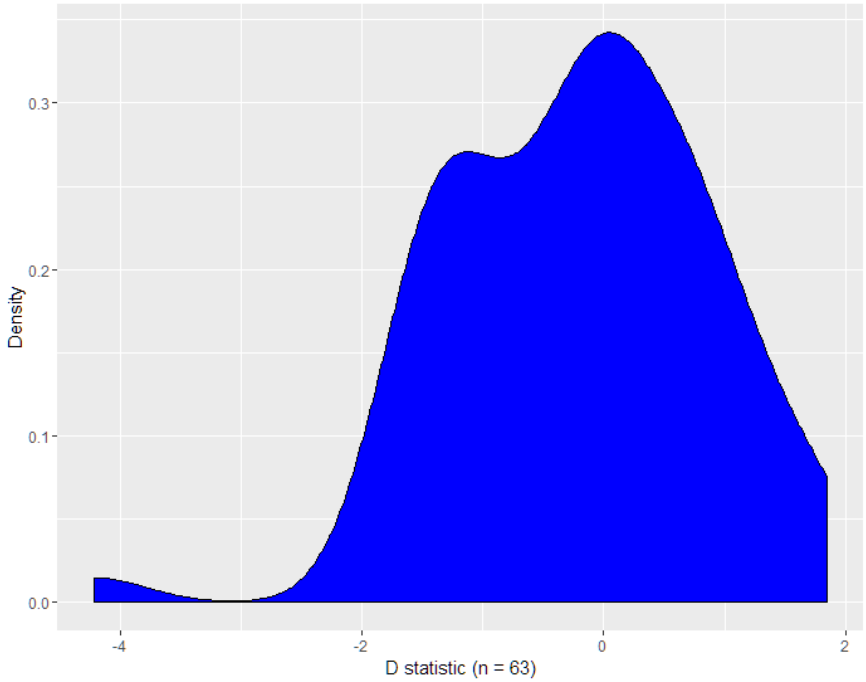
Brownian evolution

Un-tree-like randomness

Glottolog-based tree



Phylogenetic signal in tone splits and mergers



For tones occurring on open syllables & sonorant codas (A, B, C)

Phylogenetic signal in tone splits and mergers

Top ten traits with strongest signal:

B2=B3	**B1=B3**
B1=C4	**A2=C2**
B3=C4	**A3=C2**
B2=C4	**A2=C3**
A3=C4	**A3=C3**

Tones on open syllables & sonorant codas (A, B, C)

	A1	A2	A3	A4	B1	B2	B3	B4	C1	C2	C3
A1											
A2	-0.198										
A3	-0.143	-1.181									
A4	0.657	-0.086	0.026								
B1	-0.426	0.76	0.896	-1.295							
B2	-0.489	-1.284	-0.951	-0.707	-1.049						
B3	-0.547	-1.281	-0.95	-0.69	-1.481	-3.893					
B4	0.487	-1.12	-0.852	0.119	-0.243	-0.205	-0.204				
C1	1.044	1.705	1.866	1.837	0.329	0.33	0.324	-0.358			
C2	X	-1.442	-1.366	0.86	0.558	0.559	0.566	0.058	-0.145		
C3	X	-1.334	-1.308	0.853	1.322	1.277	1.317	0.074	-0.085	-0.036	
C4	X	-1.295	-1.515	0.993	-1.715	-1.662	-1.675	0.878	0.632	0.068	0.111

Phylogenetic signal in tone splits and mergers

Historical tone category

Proto-Tai onsets	A	B	C	D-short	D-long
Voiceless w/ friction *p^h, *t^h, *k^h, *s, *$m̥$, etc.	A1	B1	C1	DS1	DL1
Voiceless unaspirated *p, *t, *k, etc.	A2	B2	C2	DS2	DL2
Glottalized *ʔ, *ʔb, *ʔj, etc.	A3	B3	C3	DS3	DL3
Voiced *b, *m, *l, *z, etc.	A4	B4	C4	DS4	DL4
	Modal	Creaky	Glottal constriction		

Phylogenetic signal in tone splits and mergers

Historical tone category

Proto-Tai onsets	A	B	C	D-short	D-long
Voiceless w/ friction *p^h, *t^h, *k^h, *s, *m̥, etc.	A1	B1	C1	DS1	DL1
Voiceless unaspirated *p, *t, *k, etc.	A2	B2	C2	DS2	DL2
Glottalized *ʔ, *ʔb, *ʔj, etc.	A3	B3	C3	DS3	DL3
Voiced *b, *m, *l, *z, etc.	A4	B4	C4	DS4	DL4

B2=B3

Modal Creaky Glottal constriction

Phylogenetic signal in tone splits and mergers

Historical tone category

Proto-Tai onsets	A	B	C	D-short	D-long
Voiceless w/ friction *p^h, *t^h, *k^h, *s, *m̥, etc.	A1	B1	C1	DS1	DL1
Voiceless unaspirated *p, *t, *k, etc.	A2	B2	C2	DS2	DL2
Glottalized *ʔ, *ʔb, *ʔj, etc.	A3	B3	C3	DS3	DL3
Voiced *b, *m, *l, *z, etc.	A4	B4	C4	DS4	DL4

B2=B3
B1=C4

Modal Creaky Glottal constriction

Phylogenetic signal in tone splits and mergers

Historical tone category

Proto-Tai onsets	A	B	C	D-short	D-long
Voiceless w/ friction *p^h, *t^h, *k^h, *s, *m̥, etc.	A1	B1	C1	DS1	DL1
Voiceless unaspirated *p, *t, *k, etc.	A2	B2	C2	DS2	DL2
Glottalized *ʔ, *ʔb, *ʔj, etc.	A3	B3	C3	DS3	DL3
Voiced *b, *m, *l, *z, etc.	A4	B4	C4	DS4	DL4

B2=B3
B1=C4
B3=C4

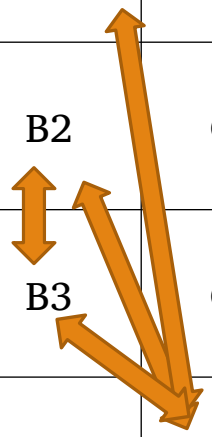
Modal Creaky Glottal constriction

Phylogenetic signal in tone splits and mergers

Historical tone category

Proto-Tai onsets

	A	B	C	D-short	D-long
Voiceless w/ friction *p^h, *t^h, *k^h, *s, *m̥, etc.	A1	B1	C1	DS1	DL1
Voiceless unaspirated *p, *t, *k, etc.	A2	B2	C2	DS2	DL2
Glottalized *ʔ, *ʔb, *ʔj, etc.	A3	B3	C3	DS3	DL3
Voiced *b, *m, *l, *z, etc.	A4	B4	C4	DS4	DL4



B2=B3
B1=C4
B3=C4
B2=C4

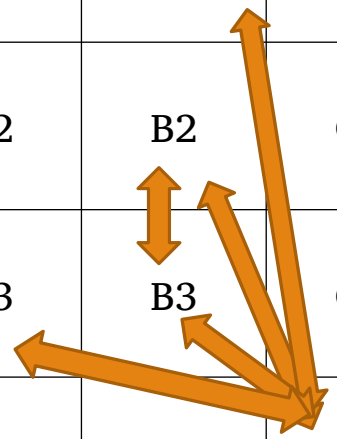
Modal Creaky Glottal constriction

Phylogenetic signal in tone splits and mergers

Historical tone category

Proto-Tai onsets

	A	B	C	D-short	D-long
Voiceless w/ friction *p^h, *t^h, *k^h, *s, *m̥, etc.	A1	B1	C1	DS1	DL1
Voiceless unaspirated *p, *t, *k, etc.	A2	B2	C2	DS2	DL2
Glottalized *ʔ, *ʔb, *ʔj, etc.	A3	B3	C3	DS3	DL3
Voiced *b, *m, *l, *z, etc.	A4	B4	C4	DS4	DL4



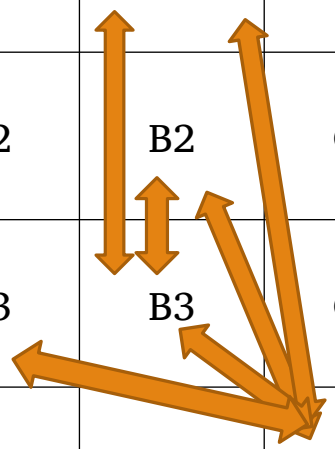
B2=B3
B1=C4
B3=C4
B2=C4
A3=C4

Modal Creaky Glottal constriction

Phylogenetic signal in tone splits and mergers

Historical tone category

Proto-Tai onsets	A	B	C	D-short	D-long
Voiceless w/ friction *p^h, *t^h, *k^h, *s, *m̥, etc.	A1	B1	C1	DS1	DL1
Voiceless unaspirated *p, *t, *k, etc.	A2	B2	C2	DS2	DL2
Glottalized *ʔ, *ʔb, *ʔj, etc.	A3	B3	C3	DS3	DL3
Voiced *b, *m, *l, *z, etc.	A4	B4	C4	DS4	DL4



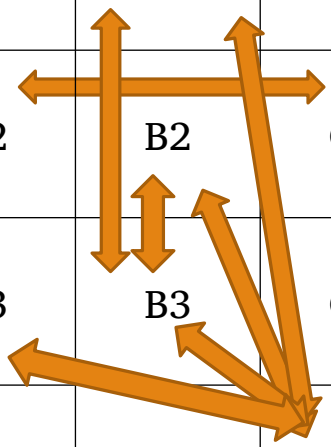
B2=B3
B1=C4
B3=C4
B2=C4
A3=C4
B1=B3

Modal Creaky Glottal constriction

Phylogenetic signal in tone splits and mergers

Historical tone category

Proto-Tai onsets	A	B	C	D-short	D-long
Voiceless w/ friction *p^h, *t^h, *k^h, *s, *m̥, etc.	A1	B1	C1	DS1	DL1
Voiceless unaspirated *p, *t, *k, etc.	A2	B2	C2	DS2	DL2
Glottalized *ʔ, *ʔb, *ʔj, etc.	A3	B3	C3	DS3	DL3
Voiced *b, *m, *l, *z, etc.	A4	B4	C4	DS4	DL4



B2=B3
B1=C4
B3=C4
B2=C4
A3=C4
B1=B3
A2=C2

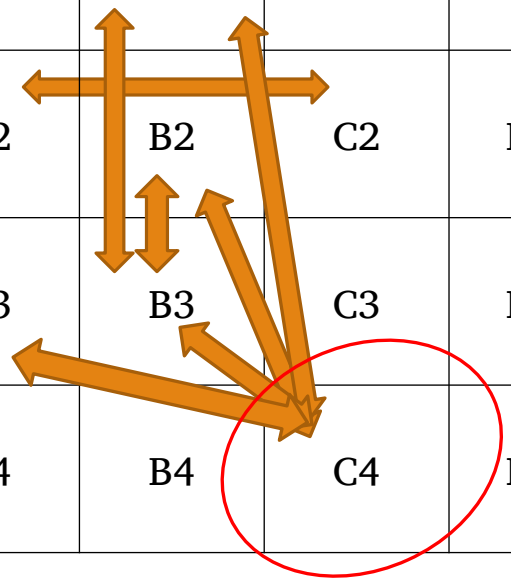
Modal Creaky Glottal constriction

Phylogenetic signal in tone splits and mergers

Historical tone category

Proto-Tai onsets

	A	B	C	D-short	D-long
Voiceless w/ friction *p^h, *t^h, *k^h, *s, *m̥, etc.	A1	B1	C1	DS1	DL1
Voiceless unaspirated *p, *t, *k, etc.	A2	B2	C2	DS2	DL2
Glottalized *ʔ, *ʔb, *ʔj, etc.	A3	B3	C3	DS3	DL3
Voiced *b, *m, *l, *z, etc.	A4	B4	C4	DS4	DL4



Modal Creaky Glottal constriction

B2=B3
B1=C4
B3=C4
B2=C4
A3=C4
B1=B3
A2=C2

Phylogenetic signal in tone splits and mergers

What makes a tonal trait informative for reconstructing phylogeny?

Trait	Count	D-score
B2=B3	361	-3.893
B1=C4	1	-1.715
B3=C4	1	-1.675
B2=C4	1	-1.662
A3=C4	1	-1.515
B1=B3	354	-1.481
A2=C2	1	-1.442
A3=C2	1	-1.366
A2=C3	1	-1.334
A3=C3	1	-1.308

N=362 (doculects)

When we compare D-score with trait frequency, two sides of same coin emerge:

1. The categories rarely merge
2. The categories rarely split

Uncommon splits/mergers contain the most phylogenetic signal. **But what makes a particular split/merger uncommon?** Data at this scale will let us answer these questions (soon).

It's just this type of insight that is fundamental to reconstruction with the comparative method in the segmental domain. No surprise it also extends to the tonal domain.

Conclusions

- Underdistributed data abounds, if we look for it
- It's another category of legacy data that deserves our attention

- **Question 1**
  - Good evidence to suggest that proto-A1 tone was rising
  - More support for glottal constriction in the proto-C tone (confirming Pittayaporn 2009)
  - Enough granular data can enable and strengthen claims about the surface tones at different historical stages
- **Question 2**
  - Strong phylogenetic signal in tone splits and mergers
  - Identifies specific tone changes particularly informative for granular language classification, giving us evidence beyond the segments

Thank you!