

## The sociolinguistic situation of Dejing Zhuang: A case study in what counts as a language

Eric M. Jackson<sup>1</sup>  
SIL International, East Asia Group  
Chinese University of Hong Kong, 25 January 2011

This talk has two goals—one general, and one specific. The general goal is that the audience would understand the issues surrounding the determination of the referents of ISO language identifiers, and would be exposed to some of the methods of dialectology which are used to establish them. The more specific goal is that the audience would understand the sociolinguistic situation of Taic in Guangxi—and in particular of the Zhuang varieties spoken in the area traditionally known as the Dejing Zhuang dialect area—and would understand the specific results and implication of our recent dialect survey work there.

### 1 Introduction: The context of “dialect survey” and ISO 639-3 codes

- (1) ISO 639-3: Codes for the representation of names of languages.

“ISO 639-3 attempts to provide as complete an enumeration of languages as possible, including living, extinct, ancient, and constructed languages, whether major or minor, written or unwritten.”

(from <http://www.sil.org/iso639-3/>, accessed 3 January 2011)

- The current standard for identifying languages of human communication

- (2) *The Ethnologue: Languages of the world* (Lewis 2009)

- Included the previous commonly used three-letter language identifiers
- First produced in 1951; sixteenth edition published in 2009

- (3) “What counts as a language? What counts as a dialect?”

What factors are relevant to answer these questions?

“A language is a dialect with an army and a navy.”

*quote often attributed to the Yiddish linguist Max Weinreich*

- Political factors
- Social factors
- Linguistic factors

- (4) From sociolinguists and linguistic anthropologists:

- A language is not a “particle”, but what is used by community members in community interactions → delineated by social networks (Mühlhäusler 1996, Hornberger 2002)

- (5) Example of a case with controversy: Serbo-Croatian vs. Serbian and Croatian (and Bosnian, and Montenegrin...)

- Solution: two levels of ISO 639-3 codes – three “languages” [bos], [hrv], [srp], one “macro-language” [hbs]

- (6) ISO criteria for assigning distinct 639-3 language codes (summary of criteria from <http://www.sil.org/iso639-3/scope.asp>, accessed 3 January 2010)

- Two speech varieties with high inherent mutual intelligibility  
→ same language code
- Spoken intelligibility is low, but common literature or ethnolinguistic identity  
→ same language code (or different codes, same macro-language)
- Spoken intelligibility is high, but well-established distinct ethno-linguistic identity  
→ different language codes (or different codes, same macro-language)

➤ And if you disagree with the current codes, submit a change request

- (7) Case in point: “Dejing Zhuang” [zyg]

- Described as a dialect of Zhuang [zha] (ISO macrolanguage) by Zhang et al (1999)
- Distinct ISO code in 2007; named “Yang” after the most populous autonym

- (8) Our question: Was this an accurate description of the language situation?

### 2 The language situation of Dejing Zhuang

- (9) Zhuang, politically: largest of the 55 minority nationalities in China [census 2003]

- roughly 16 million speakers as of 2000 census
- roughly 14 million speakers in Guangxi Zhuang Autonomous Region; remainder primarily in Yunnan, with smaller groups in Guangdong, Guizhou, Hunan

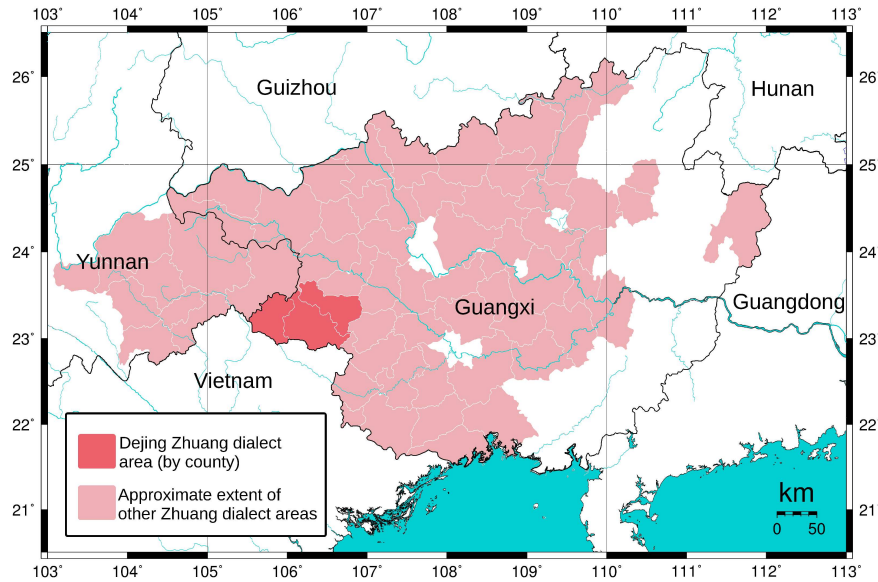
- (10) Zhuang, linguistically: not a phylogenetic linguistic group (excluding other languages)

- Part of the Taic branch of the Kra-dai (or Tai-Kadai) family
- Traditional reconstruction of Taic (Li 1977) posits Southwestern, Central, and Northern branches (though other reconstructions, eg Pittayaporn 2009, are more complex)
- Southern Zhuang part of Central Taic; Northern Zhuang part of Northern Taic
- Southern Zhuang varieties closer to Central Taic languages—eg Nùng [nut], Tày [tyz] in Vietnam—than to many Northern Zhuang varieties
- Northern Zhuang varieties closer to Northern Taic languages—eg Bouyei [pcc] in Guizhou and Vietnam, Yoy [yoy] in Thailand and Laos—than to Southern Zhuang

- (11) Dejing Zhuang: Jingxi, Napo & Debao Counties, about 1 million people, 97% Zhuang

<sup>1</sup> Even though only one person is giving this talk, the research described in this talk is by no means the work of just one person. The fieldwork was carried out by a team of SIL linguists consisting of the author, Lau Shuh Huey, and Emily Jackson, in cooperation with the Guangxi Minorities Language and Scripts Work Commission and its county-level affiliates. For a full account of this work, see Jackson et al (forthcoming).

(12) The Dejing Zhuang dialect area (county-level information from Luo et al 2005)



(13) Linguistic information on Zhuang in the Dejing area

- Chinese national census data: official ethnicity information only
- Local county almanacs: published 1998-2002
- Three wordlist datapoints in Zhang et al 1999
- A detailed grammatical description of the Jingxi County Seat variety: Zheng 1996

(14) Use of Zhuang in general (according to almanacs, Zheng 1996)

- Local Zhuang widely used across all social domains
- Jingxi County Seat variety (called “Yang”) well understood across the region

(15) Difficulties with county almanac linguistic information

- *Not uniform*: almanacs vary in their names, descriptions, detail of local varieties
- *Chinese ethnonyms are not standardized*: the same name can be applied to linguistically different groups, or linguistically similar groups can have different names
- *People movements in the past*: a variety that is now “local” may have come from somewhere else in the recent or distant past

(16) Is the published information accurate?

- Impressions of local Zhuang likely reflect *sociolinguistic* situation, not strictly linguistic
- We can look for evidence to support or refute it!

(17) With assumptions regarding linking varieties and average village size (in Napo), and normalizing populations to 2004, we can compare population of Zhuang varieties

(18) Dejing dialect area population by Zhuang variety, estimated (normalized to 2004)

Zhuang variety name(s)	Jingxi	Debao	Napo	Total	Percent total Zhuang
Jingxi Yang (仰)/Nongshun (农顺)	371892		16351	388243	39%
Fu (府)/Lang (狼)/Nongfu (农府)	1146	266250	8077	275473	28%
Min (敏)		17750	76633	94383	9.4%
Zong (宗)	75957			75957	7.6%
Long'an (隆安)/Nong'an (侬安)	26102		9062	35164	3.5%
Napo Yang (央)			32702	32702	3.3%
Zuozhou (佐州)/Zhazhou (炸州)	27011		3940	30951	3.1%
Debao Nong (侬)		17750		17750	1.8%
Sheng (省)	14718		2167	16885	1.7%
Rui (锐)	11304		3152	14456	1.4%
Napo Bunong (布侬)			8865	8865	0.9%
Ao (嗷)			4334	4334	0.4%
Jue (决)			3152	3152	0.3%
Yong (拥)			1379	1379	0.1%
Dong (峒)			≥ 591	≥ 591	≥ 0.1%

(19) This lets us prioritize language development efforts—or dialect survey—to address the largest population groups first

- Jingxi Yang (and related varieties) spoken by almost 40% of the area's Zhuang (But if it's that large, is it spoken uniformly? Or close enough to develop together?)
- Targeting the two largest groups together accounts for roughly two-thirds of the area's Zhuang (if they're spoken uniformly enough to be targeted together)
- The remaining one-third of the area's Zhuang seem to show a lot of variation—could they be associated with larger speech communities in other areas?
- **BUT** at least we know better what questions we need to answer through fieldwork!

### 3 Preparing for fieldwork: Research questions and survey design

#### 3.1 What questions do we want to answer?

- (20) What makes language development in a community successful?
- *High mutual intelligibility of spoken forms*: Can people naturally understand each other?
  - *Phonetic and phonological similarity*: Can we make a single common orthography?
    - These two factors both parallel the ISO criterion of intelligibility, linguistic similarity
  - *Social acceptance*: Will the entire community accept the developed form as representing their own language, their own social community?
    - This factor parallels the ISO criterion of common ethnolinguistic identity
- (21) Assume published reports are accurate that Jingxi County Seat Yang is widely understood
- **THUS** that variety of Yang likely best to use for language development in the community that accepts and understands it
  - **BUT** published reports also indicate that Yang isn't all that's spoken in the Dejing Zhuang dialect area. How do all the other varieties relate?
- (22) Practical considerations:
- We had only about 20 weeks for fieldwork (March – June 2008)
  - The government would like to have useful results from our fieldwork partnership
- (23) We will prioritize our work based on population—target the largest populations first
- **SO** we take as our goal to evaluate how widely the Jingxi County Seat variety of Yang is understood, as a de facto standard for development of Yang Zhuang, even among communities that don't call themselves “Yang”
  - **THUS** we will not test a wide range of candidate varieties to find which one is most widely understood
  - **BUT** we do know of some variation that could pose problems for wide intelligibility...

#### Variation within Jingxi Yang

- (24) Historical merger for some varieties of Yang:
- Yang in syllable onsets: Proto-Tai \*m, \*hm → Yang m; Proto-Tai \*n, \*hn → Yang n
  - Conservative Yang varieties: Proto-Tai \*ʔb → Yang ʔb, Proto-Tai \*ʔd → Yang ʔd  
/ʔb/~/m/, /ʔd/~/n/ contrast is preserved
  - Merging Yang varieties: Proto-Tai \*ʔb → Yang m, Proto-Tai \*ʔd → Yang n  
/ʔb/~/m/, /ʔd/~/n/ contrast is lost
  - An example of rhinoglottophilia (Matisoff 1975) leading to diachronic change?

- (25) Conservative varieties and merging varieties have different intelligibility?
- Liao (2008): upland Yang (north and west of Jingxi County Seat) is conservative, lowland Yang (south and east, including Jingxi County Seat) is merging
  - Milliken and Milliken (1996): this dialect relationship reduces intelligibility and learnability of the merging variety for listeners from the conservative variety
  - **SO** would a hypothetical *non-merging* variety of Jingxi County Seat Yang be more widely understood than the actual variety? Would it make a better standard?
- (26) Primary goal: determine the suitability of language development based on two varieties of Yang Zhuang:
- actual Jingxi County Seat Yang; we'll call this the *urban reference variety*
  - a conservative variety of Jingxi Yang, differing from Jingxi County Seat Yang as little as possible except for retaining ʔb, ʔd; we'll call it the *rural reference variety*
- (27) Phonetic and phonological similarity—requires information on not just two varieties
- Need to look concretely at phonetic and phonological variation among all varieties that will constitute part of the Yang language community being developed
- (28) What if we discover areas where *neither* variety of Yang is understood or accepted?
- What varieties are spoken in that area? What is the intelligibility of different varieties, social acceptance of different varieties? What phonetic and phonological variation?
  - **BUT** since our time for fieldwork is very short, make these areas a secondary priority to finishing the evaluation of Yang for development
- (29) Research questions so far:
- R1: Over what area do speakers of Dejing Zhuang adequately understand the urban reference variety?
  - R2: What are the attitudes of Dejing Zhuang speakers towards the urban reference variety?
  - R3: Over what area do speakers of Dejing Zhuang adequately understand the rural reference variety?
  - R4: What are the attitudes of Dejing Zhuang speakers towards the rural reference variety?
  - R5: What phonetic and phonemic differences are found within varieties of Dejing Zhuang within the Dejing language area?
  - R6 (Supplementary): Where the two reference varieties will not support language development, what other variety(ies) are adequately and widely understood within that community, and are socially accepted within that community?

### 3.2 What information will answer those questions?

- (30) Survey design: What are the concepts that are referenced in our research questions? What types of research instruments will measure these concepts?
- **Understanding or intelligibility of a language variety** (R1, R3, R6): functional assessment via Recorded Text Test (RTT)
  - **Attitudes and opinions toward a language variety and the individuals who speak it** (R2, R4, R6): a well-designed Sociolinguistic Questionnaire (SLQ)
  - **Phonetic and phonemic similarity** (R5): a standard wordlist
- (31) Other factors relevant to successful language development activities:
- Language vitality; language shift; language attitude

## 4 Methodology

### 4.1 Where to start: Sampling and site selection

- (32) Your sample of language always represents something—but does it represent what you think it does, or what you want it to? This takes careful thought!
- (33) The ideal village for our survey:
- within an administrative village composed entirely of the same variety of Zhuang
  - have as little as possible interaction with outside varieties of Zhuang
  - therefore, far from major roads, far from the township administrative center
  - (but, we also need to be able to get to it easily!)
- (34) The ideal set of villages for our survey:
- sample the named varieties (from published sources) with largest populations
  - for very populous or spread-out varieties, sample in more than one location BUT our time is short—our secondary samples within each speech variety just need to ensure that variation within that variety isn't too large
- (35) We also need to find a suitable variety of Yang to represent the rural reference variety!
- (36) The final state:
- Nine datapoints which are taken to be “primary”
  - Nine datapoints which duplicate one of the primary varieties, to check variation

### 4.2 A comparative linguist's old friend, with a new twist: Wordlists

- (37) What is it:
- a list of prompts in Chinese, for which we transcribe in IPA the pronunciation of the equivalent word in the local Zhuang (list taken from Johnson & Wang 2010)
  - 490 prompts in the full list (but only 144 were collected at secondary datapoints)
- (38) The primary reason for collecting a wordlist: to answer research question R5 by cataloging the phonetic variation among varieties, for orthography development
- Secondary use: provide an estimate of mutual intelligibility of varieties based on phonetic similarity (more on that later...)
  - Other possible use: establish phylogenetic relationships between speech varieties by the historical comparative method
- (39) How do you transcribe tone categories for a Taic language you've only heard for the first time this morning? Start with historical relationships (thanks to Gedney 1972)
- Proto-Tai is traditionally reconstructed with four tones (eg, Li 1977)
  - Many Taic languages show tone splits based on the initial reconstructed consonant: voiced consonants; preglottalized consonants and glottal plosives; plain voiceless plosives; or “voiceless friction” sounds
  - Start by eliciting a full set of cognates with the right properties, and if you're lucky, you'll have instantiated all of the modern tone categories
- (40) The new twist: using wordlists to estimate intelligibility, based on phonetic similarity
- Intelligibility has traditionally been estimated from wordlists based on a shared cognate percentage; this is *lexicostatistics*
  - A new tool to analyze wordlist similarity: *string edit distance* (or *Levenshtein distance*), which can be automated by computer, making it fast and reproducible
  - String edit distance shown to correlate just as well with intelligibility as lexicostatistics: Beijering et al (2008) for Scandinavian languages, Castro & Yang (2009) for East Asian languages with lexical tone
- (41) Why do we need this, if we primarily estimate intelligibility by other means?
- Wordlist similarity estimates inherent intelligibility, distinct from acquired intelligibility

(42) String edit distance: What is it?

- Words at two locations—datapoint 1: [p a ŋ g a]; datapoint 2: [b a j n n a]
- If insertions and deletions each have a cost of 1 (thus, substitutions have a cost of 2), what is the least costly way to convert one string to the other?

Start: datapoint 1	p a ŋ g a	Cost
Substitute [b] for [p]	b a ŋ g a	2
Insert [j]	b a j ŋ g a	1
Substitute [n] for [ŋ]	b a j n g a	2
Substitute [n] for [g]	b a j n n a	2
Done: datapoint 2	b a j n n a	Total cost: 7

- This calculation is done for each wordlist item, and then averaged for each pair of wordlist locations
  - At the end: a distance matrix showing the average phonetic distance between each pair of wordlist locations
- (43) We used the open-source package Rug/L04 (Kleiweg 2008) for this calculation
- Provides tools for clustering analysis, to represent the distance matrix as a tree
  - Provides tools for multi-dimensional scaling, to represent the distance matrix as a two-dimensional plot
- (44) How do we know when our functional intelligibility test reflects a degree of acquired intelligibility, rather than inherent intelligibility?
- When predicted intelligibility (from wordlists) and functional intelligibility (from RTTs) diverge, we look for other evidence of strong acquired intelligibility (from SLQs)
- (45) What about phonemic differences?
- Important for writing system development—to make sure everyone can write the same words the same way
  - Also useful for estimating intelligibility (Milliken & Milliken 1996)
  - The mapping from the phones of a spoken language to the phonemic categories of a hearer's language influences the *inherent learnability* of that spoken variety
- (46) Establishing the entire phonetic system of a language based on a wordlist can be a big task—but we will at least be establishing tone systems, easy to compare

### 4.3 Tricky to interpret, but useful: Recorded Text Tests (RTTs)

(47) RTT: tool to estimate functional intelligibility of one variety for speakers of another

- Record a personal story in the language spoken at one location (*reference variety*)
- Play the story at another location (*test variety*)
- The amount of the story that listeners understand reflects the intelligibility of the reference variety by speakers of the test variety
- In use since the 1960s (Casad 1974); formats have changed (Blair 1990, Nahhas 2006)

(48) We use the Quantitative Retelling method (Nahhas 2006)

- Record a personal story of about 3 minutes
- Transcribe the story in IPA; gloss it, and develop a free translation
- Divide the story into discourse-natural pieces, about 3 clauses or 15 seconds each
- Establish a baseline re-telling: five residents of the same village listen to the complete story once, then listen to each segment and re-tell it in their own words
- From these re-tellings, note the re-told elements that were common to all five
- At the test location: participants listen to the whole story once, then re-tell each segment
- Test responses are scored against the elements from the baseline re-tellings

(49) A given speech variety can be intelligible to a person because:

- that speech variety is sufficiently similar to that person's first language—*inherent intelligibility*
- that person has learned that speech variety (or one similar to it) as a second language—*acquired intelligibility*

(50) What are the possible confounding factors?

- *Listeners have already learned (or partially learned) the reference language*  
In this case, test scores would represent acquired intelligibility rather than inherent intelligibility; so, screen participants for exposure to other languages, and do not test those who have likely learned the reference language; measure the standard deviation of test scores (ideally < 10%), look critically at outliers
- *Any given listener may be linguistically gifted, and score better than others*  
So, collect scores from a set of participants, and take the average score; ten participants is taken to be a minimum
- *Listeners can't perform the task well, even if they understand the story perfectly*  
So, administer an unscored practice test to each participant first, in a language they are known to understand well (the local language, or a regional language); this introduces the task so that participants get used to it, and if the participant cannot score well on a language that they know, they are not given the real test

(51) How should RTT results be interpreted?

- Easy cases: high intelligibility (average > 90%), low intelligibility (average < 50%)
- Hard cases: partial intelligibility—anything else
- Our cutoff: 85% or higher for a reference variety to serve as a standard or basis for language development of the test variety

(52) How comparable are any two RTT results?

- Two stories can differ in many variables: voice quality and clarity of the storyteller, rate of speech, vocabulary choice, cultural content
- When choosing stories to use for RTTs, these variables need to be at least roughly equal in order for scores to be comparable
- Our score cutoff (85%) is for an “easy” story—a slow, clear storyteller, with cultural content that is roughly shared by the storyteller and test listeners

#### 4.4 Not just asking for answers: Sociolinguistic Questionnaires (SLQs)

(53) Lots of guidelines for making questionnaires—this is not just a linguistic tool! (eg, Wright & Marsden 2010, Statistics Canada 2009)

- Making a good questionnaire is not as simple as “think of the information you want to know, and ask for it”
- Important: Does the concept invoked by your question really instantiate the concept you want to measure? Does your question answer what you think it does?  
“A good question is one that produces answers that are reliable and valid measures of something we want to describe” (Fowler 1995:1-2)

(54) Some principles of good question-making (also from Fowler 1995):

- Ask about firsthand experiences—current situations, feelings and perceptions, not hypothetical situations
- Ask one question at a time
- Avoid questions with hidden assumptions
- Word the question so that everyone understands it the same way
- The question must provide enough information that the participant can answer it
- Orient all participants to the task in the same way—make it consistent

(55) Recall our purposes in using SLQs—we wanted to answer research questions on

- attitudes and opinions toward a language variety and those who speak it
- language vitality, language use, language shift
- attitudes toward language development

(56) Types of SLQ that we used:

- *Post-RTT individual questionnaire*: measured attitudes of RTT participants toward specific speech varieties after hearing that variety, not using that variety's name  
Indicates attitudes of same/different or positive/negative toward the speech varieties represented by the RTTs
- *Group sociolinguistic questionnaire*: measured attitudes of groups of residents at each location toward specific speech varieties by name  
Also collected information for language vitality, use, and change, and attitudes toward language development
- *Leader sociolinguistic questionnaire*: provided background information for interpretation of other SLQs, village demographics, and language use information

(57) Interpretation of SLQ responses

- Because our sample was so small (generally 10 people for post-RTT SLQs, groups of no more than 10 for group SLQs), we did not perform a statistical analysis
- We categorized answers, and assembled categorized answers according to the concepts that they indicated (“language attitude” questions, “language use”, etc)
- Then we looked for dominant trends (eg, “Eight of ten people at datapoint 8 said they thought RTT01 sounded most harsh”)

#### 4.5 Management of materials

(58) Informed consent

- We read participants an informational script before we collected any data: who we were, what data we intended to collect, how we would compensate them for their time (if at all), what we would do with the data afterwards, and asked for voluntary consent

(59) Information management

- Anonymized or collectivized versions of our information created for sharing with partners, other interested researchers

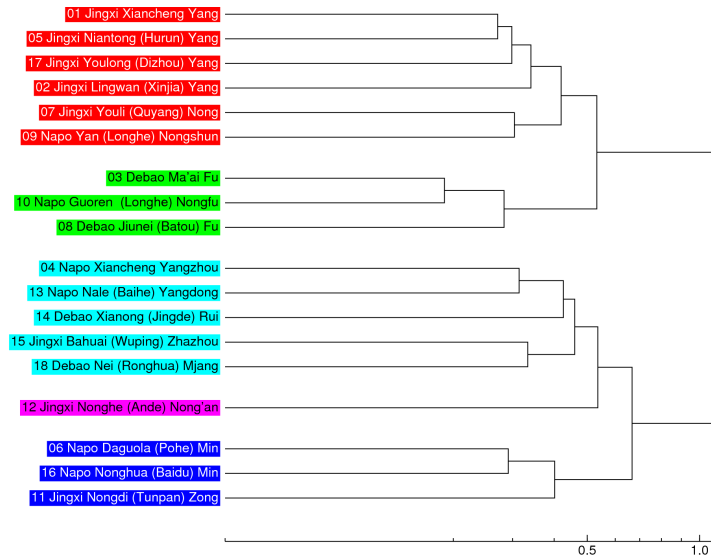
(60) Archiving

- All recordings, metadata, and appropriately anonymized data to be made available through SIL's soon-to-be-operational public archive

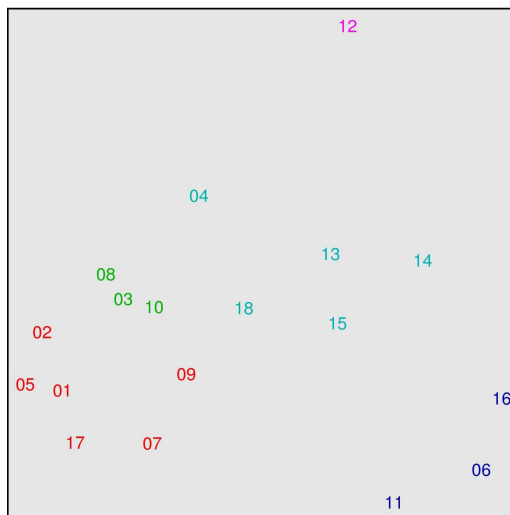
## 5 Survey results

### 5.1 Wordlist results: String edit distance and tone systems comparison

(61) Dendrogram and cluster analysis of tokenized wordlist data for all 18 datapoints (clustering using Ward's method)



(62) Multidimensional scale plot of tokenized wordlist data for all 18 datapoints



(63) Dendrogram showing clusters that were found regardless of clustering algorithm used



(64) We use the following labels for clusters:

- *Jingxi Yang cluster*: core Jingxi Yang varieties (datapoints 01, 02, 05, and 17), plus Jingxi Nong (datapoint 07) and Napo Nongshun (datapoint 09)
- *Fu cluster*: Debao Fu varieties (datapoints 03 and 08), plus Napo Nongfu (datapoint 10)
- *Yang-Nong cluster*: Fu cluster and Jingxi Yang cluster, taken together
- *Napo Yang cluster*: Napo Yangzhou and Yangdong (datapoints 04 and 13)
- *Zhazhou-Mjang cluster*: Jingxi Zhazhou (datapoint 15) and Debao Mjang (datapoint 18)
- *Min-Zong cluster*: Napo Min (datapoints 06 and 16) and Jingxi Zong (datapoint 11)

(65) A note: comparison of wordlists which include “Minz” data from Funing County (Johnson and Wang 2010) show Yunnan Minz clustering with this Min-Zong cluster

(66) Can we say what level of clustering corresponds to “a different language”? Or what degree of string edit distance?

- No! It's not that easy

(67) Tone systems of Dejing Zhuang varieties: color codes and tone numbers for the accompanying table in (70).<sup>2</sup>

red	= tone 1	(Proto-Tai tone A1)
orange	= tone 2	(Proto-Tai tone A2)
yellow	= tone 5	(Proto-Tai tone B1)
green	= tone 6	(Proto-Tai tone B2)
blue	= tone 3	(Proto-Tai tone C1)
purple	= tone 4	(Proto-Tai tone C2)
cyan	= tone 7 short	(Proto-Tai tone D1-short)
black	= tone 8 short	(Proto-Tai tone D2-short)
brown	= tone 7 long (or tone 9)	(Proto-Tai tone D1-long)
gray	= tone 8 long (or tone 10)	(Proto-Tai tone D2-long)

(68) Several points from comparing tone system properties shown in (70) to the clustering analysis based on phonetic similarity

- Fu cluster, Jingxi Yang cluster are internally similar in tone splits and pitch contours
- Some clusters with high internal similarity—Yang-Nong cluster and Min-Zong cluster—nevertheless show different development of tone systems, though often with similar pitch contours
- Jingxi Nong'an is truly an outlier!

(69) See Jackson et al (forthcoming) for further discussion and a system relations analysis (Milliken & Milliken 1996) of tone systems; conclusions are:

- Tone systems for the Jingxi Yang cluster, Napo Yang cluster, and Rui would easily allow development a common orthography, even though tone pitch values differ
- Tone systems for the Debao Fu cluster would be slightly difficult to reconcile into a common orthography with Jingxi Yang, but not impossible; a community member from Debao has in fact made one such proposal (Liao 2008)
- Other varieties—especially Zong and Nong'an—show tone systems that would pose significant problems for a common orthography with Jingxi Yang

(70) Pitch contour and historical development of tone systems of Dejing Zhuang varieties<sup>3</sup>

Gedney Box	Dejing Zhuang Varieties												Jingxi Nongdi (Tunpan) Zong		Jingxi Nonghe (Ande) Nong'an			
	Debao Ma'ai Fu	Debao Jiunei (Batou) Fu	Napo Gutoren (Longhe) Nongfu	Debao Nei (Ronghua) Mjiang	Jingxi Bahuai (Wuping) Zhazhou	Jingxi Xiancheng Yang	Jingxi Lingwan (Xinjia) Yang	Jingxi Niantong (Hurun) Yang	Jingxi Youlong (Dizhou) Yang	Napo Yan (Longhe) Nongshun	Jingxi Youli (Quyang) Nong	Napo Xiancheng Yangzhou	Napo Nale (Bahe) Yangdong	Debao Xianong (Jingde) Rui	Napo Daguola (Pohe) Min	Napo Nonghua (Baidu) Min	Jingxi Nongdi (Tunpan) Zong	Jingxi Nonghe (Ande) Nong'an
	03	08	10	18	15	01	02	05	17	09	07	04	13	14	06	16	11	12
1	53	53	53	45	35	52	53	53	45	55	45	35	35	24	24	35	35/24	45/43/21
2	53	53	53	45	35	52	53	53	45	55	45	35	35	24	24	35	35	45
3	31	31	31	32	32	52	53	53	45	55	45	35	35	24	24	35	35	21
4	31	31	31	32	32	31	21	31	31	31	31	53	33	33	55	55	44	33
5	45	44	45	44	44	45	354	45	34	24	34	24	213	22	22	22	32	34
6	45	44	45	44	44	45	354	45	34	24	34	24	213	22	22	22	32	34
7	33	22	33	33	21	45	354	45	34	24	34	24	213	22	22	22	21	43*
8	33	22	33	33	21	243	243	24	24	13	23	42	31	31	31	31	42	43*
9	35*	24*	34*	23*	24*	334*	44*	34*	32*	33*	55*	44*	55*	45*	33*	33*	32*	45
10	35*	24*	34*	23*	24*	334*	44*	34*	32*	33*	55*	44*	55*	45*	33*	33*	32*	45
11	35*	24*	34*	23*	24*	334*	44*	34*	32*	33*	55*	44*	55*	45*	33*	33*	32*	45
12	213	21	22*	21*	54*	24	214	13	23*	12*	22*	31	53*	54*	53	54*	54*	54*
13	55	44	44	44	44	44	44	33	44	55	55	44	55	44	55	55	55/11	55
14	55	44	44	44	44	44	44	33	44	55	55	44	55	44	55	55	55	55
15	55	44	44	44	44	44	44	33	44	55	55	44	55	44	55	55	55	55
16	21	21	21	22	44	21	21	23	33	22	21	31	44	33	55	55	55	44
17	55/33	44/22	44/33	33	44	35	55	55	45	35	34	13	213	22	22	22	33/11	24
18	55/33	44/22	44/33	33	44	35	55	55	45	35	34	13	213	22	22	22	33/11	24
19	55/33	44/22	44/33	33	22	35	55	55	45	35	34	13	213	22	22	22	11	44/43
20	33	22	33	23	22	24	24	24	23	12	13	33	32	31	31	31	43	44/43

<sup>2</sup> Colors in the datapoint label row in (70) correspond to status of that datapoint within the clustering analysis shown in (61).

<sup>3</sup> Pitch contours are transcribed using a five level system, where 1 represents the lowest relative tone level, and 5 the highest; thus, 55 represents a high level tone, 24 represents a mid rising tone, and so on.



### 5.2 RTT results

(71) Summary of intelligibility (inherent and acquired) results from RTTs

Test location and variety	Intelligibility of reference variety			
	Jingxi Yang	Debao Fu	Napo Yangzhou	Napo Min
05 Jingxi Hurun Yang	high			
17 Jingxi Dizhou Yang		high*	high*	
07 Jingxi Quyang Nong	high			
09 Napo Longhe Nongshun	high		high	
03 Debao Ma'ai Fu	high			
10 Napo Longhe Nongfu	high		high	
08 Debao Batou Fu	high		high*	
04 Napo County Seat Yangzhou	high			
13 Napo Baihe Yangdong	high		high	
14 Debao Jingde Rui	<i>high</i>	<i>high</i>		
15 Jingxi Wuping Zhazhou	<i>high</i>			
18 Debao Ronghua Mjang		low*		
12 Jingxi Ande Nong'an	<i>high</i>			
06 Napo Pohe Min	<i>high</i>		<i>high</i>	
16 Napo Baidu Min	low		<i>high</i>	
19 Napo Delong Min	low*	low*		
11 Jingxi Tunpan Zong	<i>high</i>			high*

- Italics indicate where the observed level of participants' understanding is judged to result from at least some amount of acquired intelligibility (judged by comparison of wordlist and RTT data, and participant screening information)
- Asterisks indicate a suggested level of intelligibility for locations with insufficient RTT information for confident conclusions (samples with only two to eight participants)
- Blank cells indicate untested combinations; gray cells indicate hometown locations

(72) Without bringing up ISO codes yet, what does this tell us about likely success of language development activities using these varieties?

- Using a variety of Jingxi Yang which closely resembles the varieties we tested would likely be successful—linguistically—at least for those varieties above the double line
- This includes language development activities such as early mother-tongue education programs and possibly orthography development and literacy activities
- Among the varieties below the double line, the Min and Zong varieties appear to have sufficient inherent intelligibility to be grouped for language development activities

### 5.3 SLQ results

(73) Can our SLQ results be taken as representative of the communities as a whole?

- Post-RTT SLQ sample ought to represent only that portion of the community with the lowest exposure to non-local Zhuang varieties, not the community as a whole
- Group SLQ sample represents the portion of the community present at the time
- Many community members ages 20-40 were absent from the community or excluded from RTT participation because of language exposure acquired through migrant labor!

(74) We assume:

- Our RTT sample did not (or could not) always sample those with *no* exposure to outside varieties, so it is not so far off from representative of the broader communities
- Significant exposure to other Zhuang varieties acquired through migrant labor experience will tend to make individuals more open to non-local Zhuang varieties (but we could be wrong!)

### Post-RTT SLQ

(75) Summary of attitude results from post-RTT SLQs

Location	Similarity of ...				Most positive non-local variety	Most negative non-local variety
	Jingxi County Seat (RTT01)	Rural Jingxi (RTT02)	Debao County Seat (RTT03)	Napo County Seat (RTT04)		
05	same	same			RTT02	RTT01
02	neutral					
07	mixed	mixed			RTT02	RTT01
09	mixed	mixed		different	RTT01/2	RTT04
03	different	different			unclear	
10	neutral	different	same	different	RTT03	RTT04
08	different	different	same		RTT03	RTT01
04	different	different			RTT01	RTT02
13	different	different		mixed	RTT04	RTT01
14	different	different	different		RTT03	unclear
15	different	different			unclear	unclear
12	different	different			RTT01	RTT11
06	different	different		different	RTT04	RTT01
16	different	different		different	RTT04	unclear
11	different	different			unclear	unclear

### Group SLQ

(76) Group SLQ responses concerning attitudes toward other Zhuang varieties

Location	JX County Seat Zhuang: soft or harsh?	Harshest Zhuang?	Softest Zhuang?	Which Zhuang is most well understood in this county?
05	softer	local or Zong	JX County Seat	JX County Seat
17	(it sounds good)	local	Debao	
02	softer, similar to local	local Yang	Dabang village Yang	Yang
07	(we understand it)	Nong'an or Rui	Debao Fu	JX and DB County Seats
09	softer	Nong'an		
03	softer		NP County Seat	the local variety
10	harsh	Jingxi	Debao	Yang or Min
08	harsh	Jingxi	Debao County Seat	the local variety
04	softer	Debao	JX or NP County Seat	Yangzhou
13	harsh	Debao	Napo	
14	soft		JX or DB County Seat	Fu
15	(it sounds nice)	Nong'an, Zhenzhou	DB County Seat	JX County Seat
12	(disagreement in group)	local Nong'an or JX County Seat	Debao	Yang and Zong
06	softer	Debao	Jingxi	
16	soft	Rui or [ŋau <sup>21</sup> ]		Min
11	harsh	local Zong	Nanpo Township	

(77) This information should be used to supplement the post-RTT attitude information, since it is based only on language names, not a concrete sample of the language in question

(78) We take softness to indicate loosely positive attitudes toward a variety—and Jingxi Yang is frequently viewed as somewhat soft

- Jingxi County Seat Yang generally viewed positively within Jingxi Yang cluster
- Jingxi County Seat Yang generally viewed negatively within Debao Fu cluster
- Other areas differ in their attitude toward Jingxi County Seat Yang, but more favorable than unfavorable

## 6 Interpretation and future research

### 6.1 Answers to research questions

(79) R1: *Over what area do speakers of Dejing Zhuang adequately understand the urban reference variety?*

R3: *Over what area do speakers of Dejing Zhuang adequately understand the rural reference variety?*

- Because of inequalities in the urban and rural Yang RTTs, we judged that these results were not directly numerically comparable
- At least some variety of Yang is highly inherently intelligible within the Yang-Nong and Napo Yang clusters
- Many other Zhuang language communities in Jingxi and Debao have high intelligibility in Jingxi Yang, but only as a result of extensive exposure to that variety

(80) R2: *What are the attitudes of Dejing Zhuang speakers towards the urban reference variety?*

R4: *What are the attitudes of Dejing Zhuang speakers towards the rural reference variety?*

- Most locations within the Jingxi Yang cluster expressed positive attitudes toward Jingxi County Seat Yang, and more positive attitudes toward the rural reference variety
- Most locations within the Debao Fu cluster expressed negative attitudes toward Jingxi Yang, with more positive attitudes toward the rural reference variety
- The two varieties of Napo Yang were divided in attitude toward both tested varieties of Jingxi Yang
- Attitudes among other low-population varieties are mixed

(81) R5: *What phonetic and phonemic differences are found within varieties of Dejing Zhuang in the Dejing language area?*

- For tone only: many varieties, including the Yang-Nong cluster, Napo Yang cluster, and some other varieties, are sufficiently similar to allow the use of a common orthography
- For tone only: other varieties, especially Zong and Nong'an, pose significant challenges for a common orthography

(82) R6 (Supplementary): *Where the two reference varieties (of Jingxi Yang) will not support language development, what other variety(ies) are adequately and widely understood within that community, and are socially accepted within that community?*

- Varieties of Min and Zong in Jingxi and Napo Counties appear to be highly inherently intelligible, with mildly positive attitudes
- Further testing will be needed within this cluster to determine the most widely understood and accepted variety to use as a focus for language development

## 6.2 Implications for language development and ISO language identifiers

(83) Language development within the Yang-Nong cluster

- Likely no linguistic challenges to using a variety of Yang as a basis for development
- Likely some attitude challenges to using a variety of Yang within the Fu subcluster—not deadly to a language development project, but you will want to address these attitudes!
- Our RTTs could not show this, but on linguistic grounds, a conservative Yang variety (maintaining preglottalized stops) may be a better basis than a merging variety

(84) Current [zyg] “Zhuang, Yang”

- Should be specified to include only varieties in the Yang-Nong cluster, and possibly the Napo Yang cluster—2004 population of 696,000 to 714,000 ( $\pm$  Debao Nong?)

(85) Language development within remaining varieties of the Dejing area

- Independent language development within the Min-Zong cluster would likely be successful, using a variety of Min-Zong as a focus
- Conclusions for other less-populous communities are not clear from this survey

(86) Current [zgm] “Zhuang, Minz”

- Currently only applied to the relatively small (~2600 people in 2007) Minz communities in Yunnan, on the basis of Johnson and Wang (2010)
- The much more populous Min and Zong communities in Guangxi (2004 population around 170,000) should likely also be included in this identifier

(87) In total, the Yang-Nong and Napo Yang clusters, plus the Min-Zong cluster, would account for nearly 90% of the Zhuang of the area—benefit for lots of people!

## 6.3 The future: How to best extend this work

(88) Doing more survey: Zhuang varieties with unclear associations

- *Mjang*: what kind of Zhuang does it resemble most? Is this what the Debao County Almanac refers to as “Debao Nong”?
- *Rui*: what kind of Zhuang does it resemble most? Is this a variety of what Johnson and Wang (2010) refer to as “Sha”—Northern Taic Qiubei Yei Zhuang [zqe] and Guibian Yei Zhuang [zgb]? (This is a guess based only on the name)
- *Nong'an*: is this a community of speakers of Zuojiang Zhuang [zzj] displaced from Long'an County? (This is a guess based only on the name)
- Many other Zhuang varieties—Ao, Jue, Yong, Dong, Sheng, Bunong—not even sampled

(89) “Intermediate” varieties: Napo Yang cluster, Mjang-Zhazhou cluster, Rui

- How much Jingxi Yang intelligibility is inherent, and how much is acquired?

(90) Min-Zong cluster: what's the ideal “central” variety, widely understood and accepted?

(91) Pittayaporn (2009) notes that surveys with a high degree of geographic resolution are still rare among Taic languages

- This concrete picture of variation can be helpful in figuring out historical change and historical people movements

- We take this as agreement that more surveys like this in other areas would be beneficial

## 7 Closing: what counts as a language

(92) “What counts as a language?” can be politically charged; we tried to avoid controversy

- Our criteria: high intelligibility, social acceptance (*a la* common ethnolinguistic identity)
- Others may use different criteria and make different classifications

(93) Benefits for other linguists

- We used some tools that other linguists might not be familiar with—RTTs, questionnaire-type SLQs
- We also used traditional wordlists, albeit with new methods of analysis—string-edit distance and systems relations
- Many linguists could easily adopt these tools and methods, adding them to existing fieldwork on other languages → to provide more complete language documentation
- This would benefit the broader linguistics community through improved accuracy of representation by ISO language identifiers

(94) Benefits for language communities

- Improved, targeted information for decision making in government or community-based language development efforts
- Makes language development efforts more successful by targeting right-sized language communities—not too large (too much linguistic variation), not too small (too many separate projects, not enough resources or critical mass to keep all of them going)

(95) Bringing benefit to language communities is not just for NGOs anymore!

- LSA ethics statement, under the heading “Responsibility to communities”:

“In general, linguists should strive to determine what will be constructive for all those involved in a research encounter, taking into account the community’s cultural norms and values.”

*Linguistic Society of America 2009:3*

## References

- Beijing, Karin, Charlotte Gooskens & Wilbert Heeringa. 2008. Predicting intelligibility and perceived linguistic distances by means of the Levenshtein algorithm. In Marjo van Koppen and Bert Botma (eds.), *Linguistics in the Netherlands 2008*, 13-24. Amsterdam: John Benjamins Publishing Company.
- Blair, Frank. 1990. *Survey on a shoestring*. Dallas: SIL International.
- Casad, Eugene. 1974. *Dialect intelligibility testing* (Summer Institute of Linguistics Publications in Linguistics and Related Fields 38). Norman: Summer Institute of Linguistics of the University of Oklahoma.
- Castro, Andy & Cathryn Yang. 2009. Representing tone in Levenshtein distance. *International Journal of Humanities and Arts Computing: Computing and Language Variation* 2. 205-219.
- Debao Xian Zhi. 1998. [Debao County Almanac.] Nanning, Guangxi: The People's Press.
- Fowler, Floyd, Jr. 1995. *Improving survey questions: Design and evaluation*. Thousand Oaks, CA: Sage Publications.
- Gedney, William. 1972. A checklist for determining tone in Tai dialects. In M. Estelle Smith (ed.), *Studies in honor of George L. Trager*, 423-438. The Hague: Mouton.
- Hornberger, Nancy. 2002. Multilingual language policies and the continue of biliteracy: An ecological approach. *Language policy*. 1:27-51.
- Jackson, Eric, Emily Jackson & Shuh Huey Lau. Forthcoming. A sociolinguistic survey of the Dejing Zhuang sub-dialect area (SIL Electronic Survey Reports). Draft available from [https://mail.link77.net/~eric\\_jackson@sil.org/research.html](https://mail.link77.net/~eric_jackson@sil.org/research.html).
- Jingxi Xian Zhi. 2000. [Jingxi County Almanac.] Nanning, Guangxi: The People's Press.
- Johnson, Eric & Wang Mingfu. 2010. A sociolinguistic introduction to the Central Taic languages of Wenshan Prefecture, China (SIL Electronic Survey Reports 2010-027). <http://www.sil.org/silesr/abstract.asp?ref=2010-027> (September 2010).
- Kleiweg, Peter. 2008. RuG/L04 (version of 2008/10/15) [computer program]. Retrieved from <http://www.let.rug.nl/~kleiweg/L04/> (accessed October 2008 through October 2010).
- Lewis, M. Paul (ed.). 2009. *Ethnologue: Languages of the World, Sixteenth edition*. Dallas: SIL International. Online version: [www.ethnologue.com](http://www.ethnologue.com).
- Li, Fang-Kuei. 1977. *A handbook of comparative Tai*. Honolulu: The University of Hawaii Press.
- Liao, Hanbo. 2008. Taiyuzhi zhongbuzu Yangnongyu (Dejingtuyu) yinxi gaikuang yu pinyin fang'an xiangjie [An overview of the sound system of the Central Taic language Yang-Nong (Dejing vernacular) and description of a phonetic spelling system]. Ms. Linguistics Society of America. 2009. Ethics statement (May 2009). Retrieved from [http://www.lsadc.org/info/pdf\\_files/Ethics\\_Statement.pdf](http://www.lsadc.org/info/pdf_files/Ethics_Statement.pdf) (checked January 2011). Ms.
- Luo, Liming, Ma Chaofa, Margaret Milliken, Nong Guangmin, Wu Jun, Du Zaijing, Lu Zhenyu, Lu Xiaoli, Yang Yijie, Chen Fulong, Zhang Haiying, Mo Jiayu, Liang Jinjie, Huang Dawu, Huang Peixing, Huang Rumeng, Huang Quanxi, Stuart Milliken, Qin Minggui, Qin Yaowu, Xie Lanyan, & Meng Yuanyao (eds.). 2005. *Zhuang-Chinese-English dictionary*. Beijing: Nationalities Press.
- Matisoff, James. 1975. Rhinoglottophilia: The mysterious connection between nasality and glottality. In Charles Ferguson, Larry Hyman and John Ohala (eds.), *Nasálfest: Papers from a Symposium on Nasals and Nasalization*, 265-287. Stanford, CA: Stanford University Press.
- Milliken, Margaret & Stuart Milliken. 1996. System relationships in assessing dialect intelligibility. *Notes on Linguistics* 72. 15-31.
- Mühlhäusler, Peter. 1996. *Linguistic ecology: Language change and linguistic imperialism in the Pacific region*. London: Routledge.
- Nahhas, Ramzi, compiler. 2006. *The steps of Recorded Text Testing: A practical guide*. Chiang Mai, Thailand: Payap University, ms.
- Napo Xian Zhi. 2002. [Napo County Almanac.] Nanning, Guangxi: The People's Press.
- National Bureau of Statistics, Office of Population and Social Science Statistics & National Ethnic Affairs Commission, Office of Economic Development (eds.). 2003. 2000 Nian Renkou Pucha Zhongguo Minzu Renkou Ziliao. [Year 2000 Population Census Data for China's Ethnic Groups]. Beijing: The Ethnic Publishing House.
- Pittayaporn, Pittawat. 2009. *The phonology of Proto-Tai*. Ithaca, NY: Cornell University dissertation.
- Statistics Canada. 2009. Statistics Canada quality guidelines (5<sup>th</sup> edition). Available online from <http://www.statcan.gc.ca/pub/12-539-x/12-539-x2009001-eng.htm>.
- Wright, James & Peter Marsden (eds.). 2010. *Handbook of survey research* (2<sup>nd</sup> edition). Bingley, UK: Emerald Publishing Group.
- Zhang, Junru, Liang Min, Ouyang Jueya, Zheng Yiqing, Li Xulian & Xie Jianyou (eds.). 1999. *Zhuangyu Fangyan Yanjiu* [Zhuang Dialect Research] (Zhongguo Shaoshu Minzu Yuyan Fangyan Yanjiu Congshu [Chinese Minority Language Dialect Research Series]). Chengdu: Sichuan Ethnic Publishing House.
- Zheng, Yiqing. 1996. *Jingxi Zhuangyu Yanjiu* [Jingxi Zhuang language research]. Beijing: Chinese Academy of Social Science, Nationalities Research Institute.

A PDF version of this handout is available for download from [https://mail.link77.net/~eric\\_jackson@sil.org/research.html](https://mail.link77.net/~eric_jackson@sil.org/research.html).