A New Dataset for Tonal Segmental Dialectometry

From the Yue- and Pinghua-Speaking Area

Matthew Sung, Jelena Prokic & Yiya Chen| 6th SIGTYP Workshop (Malta)

Universiteit
Leiden
The Netherlands

Motivation

Traditional Dialectology

- Originated in Europe
- Predominantly been focusing on sounds (segments)

Motivation

Traditional Dialectology

- Originated in Europe
- Predominantly been focusing on sounds (segments)

Dialectometry/ Quantitative Dialectology

- Similar to traditional dialectology
- Mostly on European languages, but have been expanding beyond Europe recently

Motivation

Traditional Dialectology

- Originated in Europe
- Predominantly been focusing on sounds (segments)

Dialectometry/ Quantitative Dialectology

- Similar to traditional dialectology
- Mostly on European languages, but have been expanding beyond Europe recently

Tone languages

- More than half of the world's languages are tonal (Yip 2002)
- Not many datasets with more than 20 dialects (of the same language) are publically available which contains both tones and segments

New Dataset!

Sinitic languages

- Yue and Pinghua
- Spoken in Southern China
- Tone languages

New Dataset!

Sinitic languages

- Yue and Pinghua
- Spoken in Southern China
- Tone languages

It consists of:

- 104 varieties
- 130 monosyllabic words
  - E.g. Body parts, numbers, animals, geographical features etc.
- Both segments and tones

Data sources

Dialect Surveys and Homonymic Syllabaries

編號 / 地點	1 多	2 拖	3 他	4 駝	5 馱 駄延来	6 舵	7 大	8 馱 馱子	9 椰	10 哪 哪個？
	果開一平歌端	果開一平歌透	果開一平歌透	果開一平歌定	果開一平歌定	果開一上哿定	果開一去箇定	果開一去箇定	果開一平歌定(俗)	果開一上哿泥(俗)
北京	tuo55	tʻuo214	tʻa55	tʻuo35	tʻuo35	tuo51	ta51	tuo51	huo35	na214
廣州（市區）	tɔ55	tʻɔ35	tʻa55	tɔ21	tɔ21	tɔ21	tai22	tɔ22	nɔ21	nai13
香港（市區）	tɔ55	tʻɔ35	tʻa55	tɔ21	tɔ21	tɔ21	tai13	tɔ13	nɔ21	nai13
香港（新界錦田）	tɔ55	tʻɔ35	tʻA55	tɔ21	tɔ21	tɔ21	tAi32		nɔ21	nA35
澳門（市區）	tɔ55	tʻɔ35	tʻa55	tɔ21	tɔ21	tɔ21	tai22	tɔ21	lɔ21	lai13
番禺（市橋）	tɔ53	tʻɔ35	tʻa53	tɔ21	tɔ21	tɔ21	tai22		lɔ21	lai13
花都（花山）	tɔ53	tʻɔ35	tʻa53	tɔ12	tɔ12	tɔ12	tai21	tɔ12	nai13	nai35
從化（城內）	tɔ23	tʻɔ35	tʻa35	tɔ22	tɔ22	tɔ22	tai21		nɔ22	nai35
增城（縣城）	tɔ55	tʻɔ35	tʻa35	tɔ11	tɔ11	tɔ11	tai21	tɔ22	nɔ11	nai13
佛山（市區）	tɔ55	tʻɔ35	tʻa55	tɔ44	tɔ44	tɔ44	tai22		nɔ44	nai13
南海（沙頭）	tɔ55	tʻɔ35	tʻa55	tɔ44	tɔ44	tɔ44	ta22		nɔ44	nai13
順德（大良）	tɔ55	tʻɔ35	tʻa55	tɔ42	tɔ44	tɔ42	tai21	tɔ42	lɔ42	lai13
三水（西南）	tɔ53	tʻɔ35	tʻa53	tɔ21	tɔ21	tɔ21	tai21	tɔ21	nɔ21	nai13
高明（明城）	tɔ445	tʻɔ445	tʻa445	tɔ21	tɔ21	tɔ21	tai21		nɔ21	nai21
中山（石岐）	tɔ55	tʻɔ35	tʻa55	tɔ51	tɔ51	tɔ51	tai35	tɔ51	nɔ51	nɔ51
珠海（前山）	tɔ55	tʻɔ35	tʻa55	tɔ21	tɔ21	tɔ21	tai35	tɔ21	nai21	nai13
斗門（上橫水上話）	tɔ45	tʻɔ45	tʻa45	tɔ43	tɔ43	tɔ43	tai21	tɔ43	nɔ43	nai13
斗門（斗門城）	tʻɔ33	hʷɔ33	ha35	hʷɔ33	hʷɔ22	hʷɔ21	tai33	hʷɔ33	ⁿdʷɔ21	ⁿdʷɔ22
江門（白沙）	tɔ35	hɔ33	ha35	hɔ21	hɔ21	hɔ21	tai21	hɔ22	lɔ21	lɔ22
新會（會城）	tʻɔ33	hʷɔ33	ha33	hʷɔ21	hʷɔ21	hʷɔ21	tai21	hʷɔ22	ⁿdʷɔ25	ⁿdʷɔ22
台山（台城）	ʉ35	hʷɔ35	h ɐ55	hʷɔ21	hʷɔ21	hʷɔ21	ɑi21		ⁿdʷɔ22	ⁿdɑ21
開平（赤坎）	u55	hu35	ha35	hu21	hu21	hu21	ɑi21		ⁿdɔ22	ⁿdɑ21
恩平（牛江）	tua33	hua35	ha33	hua21	hua21	hua21	taj21		ⁿdua21	ⁿdua22
鶴山（雅瑤）	ɔu33	hɔu33	tʻA33	hɔu13	hɔu13	hɔ21	ɔ21		nɔu12	
東莞（莞城）	tɔ213	tʻɔ413	tʻa213	tɔ21	tɔ21	tɔ21	tai21	tɔ21	nɔ21	na21
寶安（沙井）	tɔ213/tɔ55	tʻɔ413/tɔ55	tʻa213/tʻa35	tɔ21	tɔ21	tɔ21	tai21	tɔ21	nɔ21	na11
惠州（市區）	tɔ33	tʻɔ33	tʻa33	tɔ11	tɔ11	tɔ11	tʻai31	tɔ11	nɔ11	na35
東莞（清溪）	tɔ35	tʻɔ35	tʻa35	tɔ11	tɔ11	tɔ11	tʻai44		la11	la21
深圳（沙頭角）	tɔ33	tʻɔ35	tʻa35	tɔ11	tɔ11	tɔ11	tʻai44		la33	la21
從化（呂田）	tɔ44	tʻɔ44	tʻa44	tɔ35	tɔ35	tʻɔ53	tʻai21		la44	la44
中山（南蓢合水）	tɔ55	tʻɔ35	tʻa33	tɔ21	tɔ21	tɔ21	tʻai55	tai55	nɔ21	nai35
中山（龜都）	tɔ213/tɔ51	tʻɔ213/tʻɔ35	tʻa213	tɔ21	tɔ21	tɔ21	tai35/tʻai35	tɔ21	hɔk11	hai24/tɔ51

Dialect Survey (*Survey of Dialects in the Pearl River Delta*)

ɛ

ŋ [ㄱ]惹□额□ŋjɐpㄱ~：蜻蜓 [ㄴ]嘢东西

o

p [ㄅ]□~老：大舅母 [ㄆ]婆袍大棉~：棉袄 [ㄇ]抱□泡沫 [ㄈ]菢孵暴~露□一~：一串

b [ㄅ]煲 [ㄆ]□像蝙蝠的一种鸟 [ㄉ]保宝堡碉~ [ㄌ]报

m [ㄅ]□脸肿 [ㄆ]毛 [ㄇ]冒帽

f [ㄅ]搔骚臊 [ㄆ]曹漕槽 [ㄉ]嫂 [ㄋ]造 [ㄌ]扫~地｜~把 [ㄈ]□早~晚~：早稻晚稻

t [ㄅ]糟遭 [ㄆ]掏滔涛桃陶淘逃萄葡~绸捆扎 [ㄉ]早枣蚤 [ㄋ]导祷祈~ [ㄌ]灶~头：灶 [ㄈ]道稻盗

tʻ [ㄅ]操 [ㄉ]讨草 [ㄌ]套澡糙燥

dˀ [ㄅ]刀叨 [ㄉ]岛捣倒打~｜~茶 [ㄌ]到

n [ㄋ]恼脑 [ㄈ]□瞪眼睛

l [ㄆ]劳牢痨 [ㄉ]佬 [ㄋ]老 [ㄈ]涝

ʃ [ㄅ]□形容密而多 [ㄉ]□做~：做什么 [ㄈ]傻

k [ㄅ]高糕篙膏羔蒿商~ [ㄉ]稿 [ㄌ]个告箇~ɓokㄱ：这里

ŋ [ㄋ]我 [ㄌ]傲 [ㄈ]熬

h [ㄉ]好~人 [ㄌ]耗好喜~

ø [ㄋ]豪毫壕 [ㄌ]澳~门 [ㄈ]浩号蚝~油

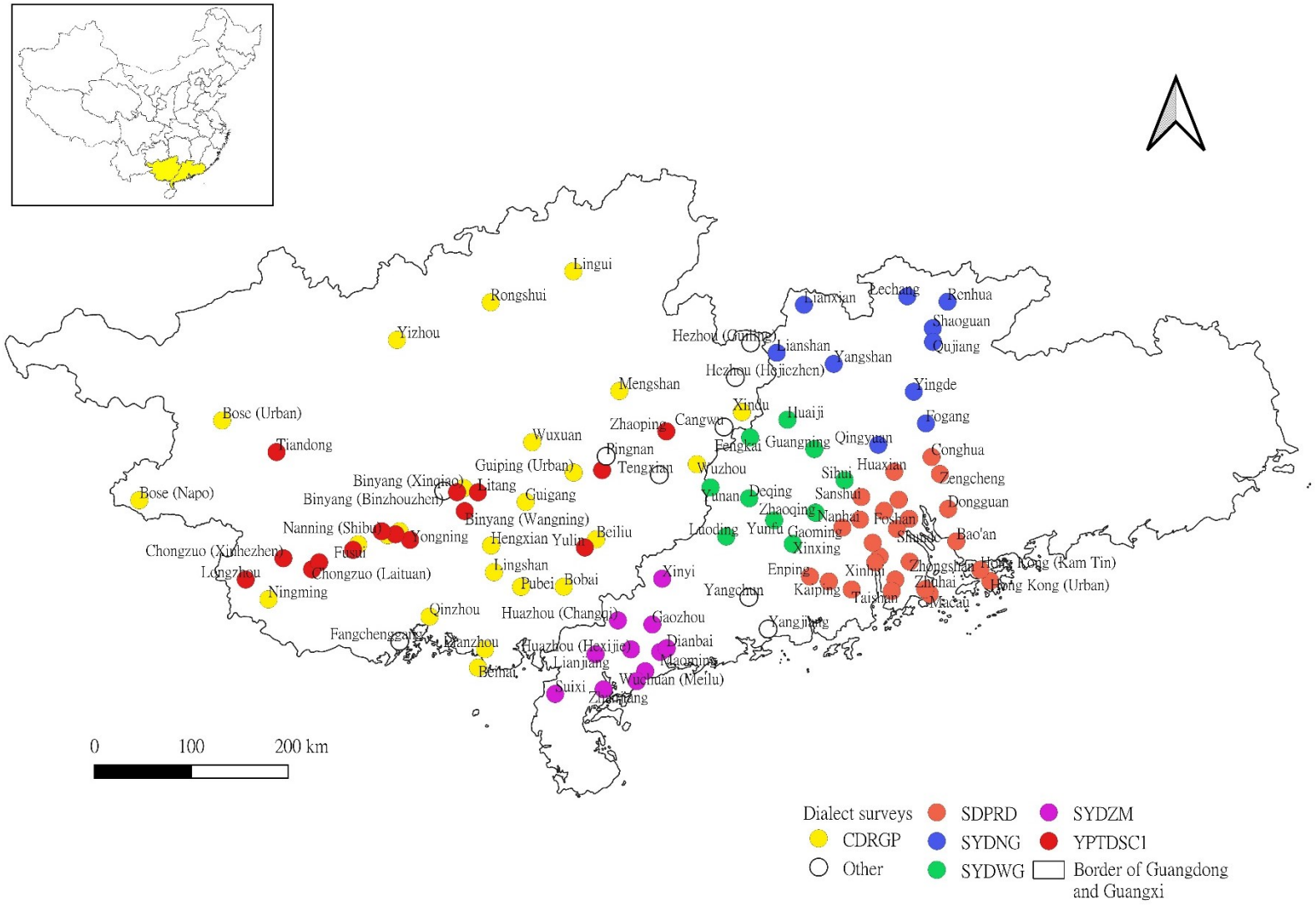
w [ㄌ]浼~lɐnㄚ：骗

i

p [ㄋ]皮肥脾琵~琶枇~杷 [ㄋ]婢被棉~｜~迫 [ㄈ]备鼻避

Homonymic Syllabary (*The Phonology of Cangwu Local Vernacular in Guangxi*)

Localities and their respective sources



Segmental Data

Universiteit
Leiden
The Netherlands

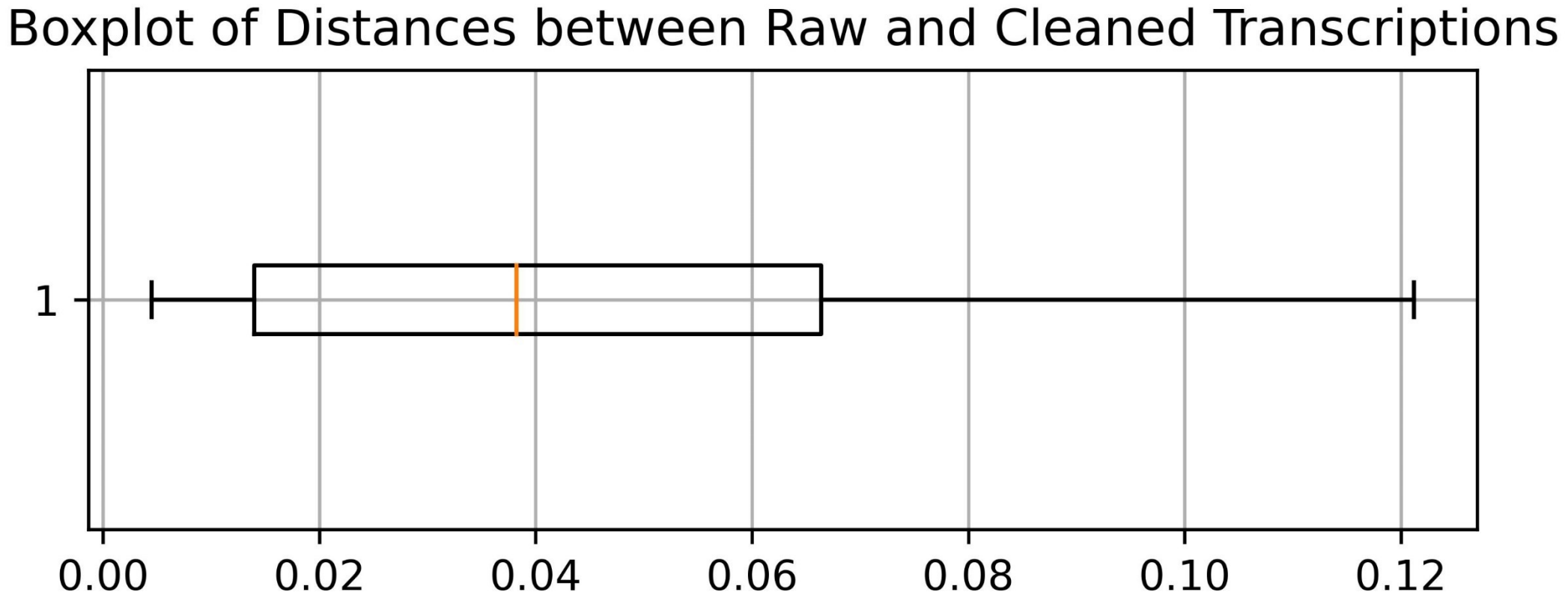
Modifications to the original transcriptions

- Transcribers' differences
  - More than 1 transcriber
  - Different conventions/ habits of transcription for the same sound
  - Can influence the dialectometric analysis
- Normalisation/ cleaning required

Modifications

1. Comparison with existing recordings
   - Guangzhou ‘water’ sœy → sɵy
2. Maintaining contrasts
3. Removal of redundant characters
   - Gaoyang dialects ‘person’ ɲiɐn → ɲɐn
4. Simplification of overly detailed transcriptions
   - HK (Kam Tin) a̲ → a
5. Consistency of onsets
6. Conversion from Chinese IPA to Standard IPA
   - Guangzhou ‘skin’ p’ei → pʰei
7. Phonetic alignment

Levenshtein Distance between Raw vs. Cleaned Transcriptions



Tonal Data

Universiteit
Leiden
The Netherlands

Tonal data

- The existing digital tonal datasets generally consist of about 20-30 dialects
- Our dataset
  - Same number of words as the segmental data
  - Same items (130)
  - Same dialects (104)
  - Allows us to compare the segmental and tonal variation of the same area

Tone notations

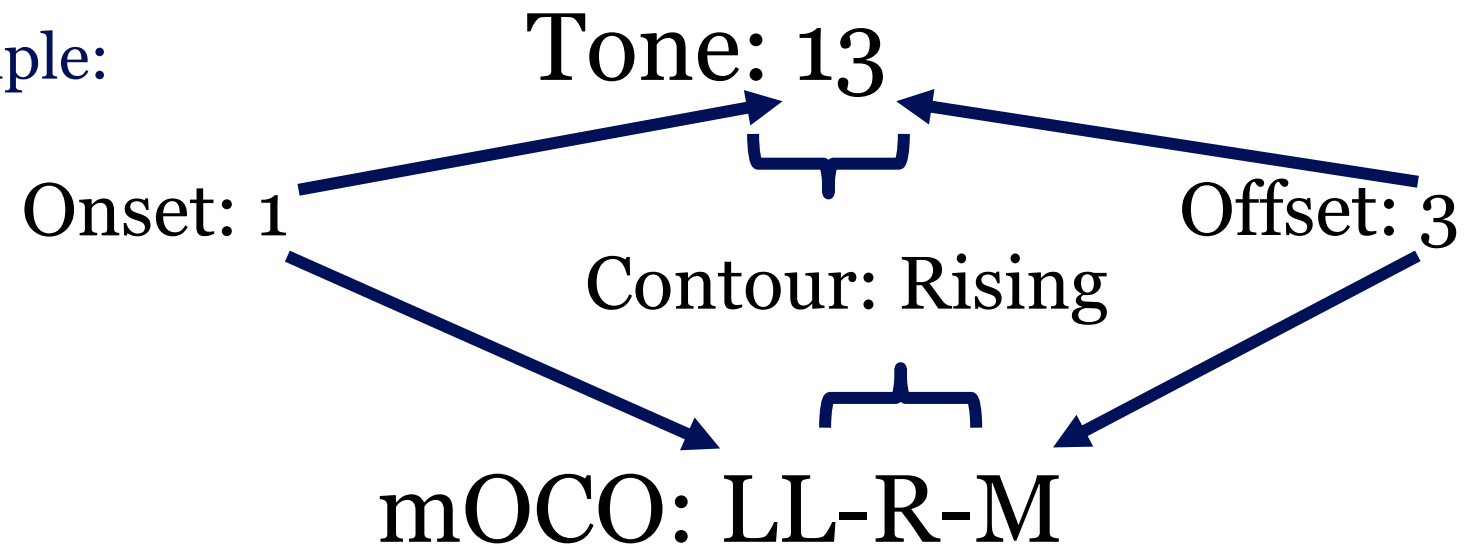
- Chao's (1930) tone letters
- Contour levels: 1, 2, 3, 4, 5
  - 1 is the lower pitch level, 5 is the highest
- Combination of pitch levels to represent a tone contour
  - Level tones: 11, 22, 33, 44, 55
  - Rising tones: e.g. 12, 24, 25, 45
  - Falling tones: e.g. 51, 42, 31, 43
  - Concave tones: e.g. 413, 512
  - Convex tones: e.g. 232, 253

Tone representation in dialectometry

- Chao's (1930) transcription system cannot directly be used for dialectometry (Sung et al. Forthcoming)
- Requires further conversion to yield meaningful tone distances (when using tools like *Gabmap* or *LED-A.org*)
  - Tone-to-string (Tang 2009)
  - Onset-Contour-Offset (Yang and Castro 2008)
  - **Modified Onset-Contour-Offset**

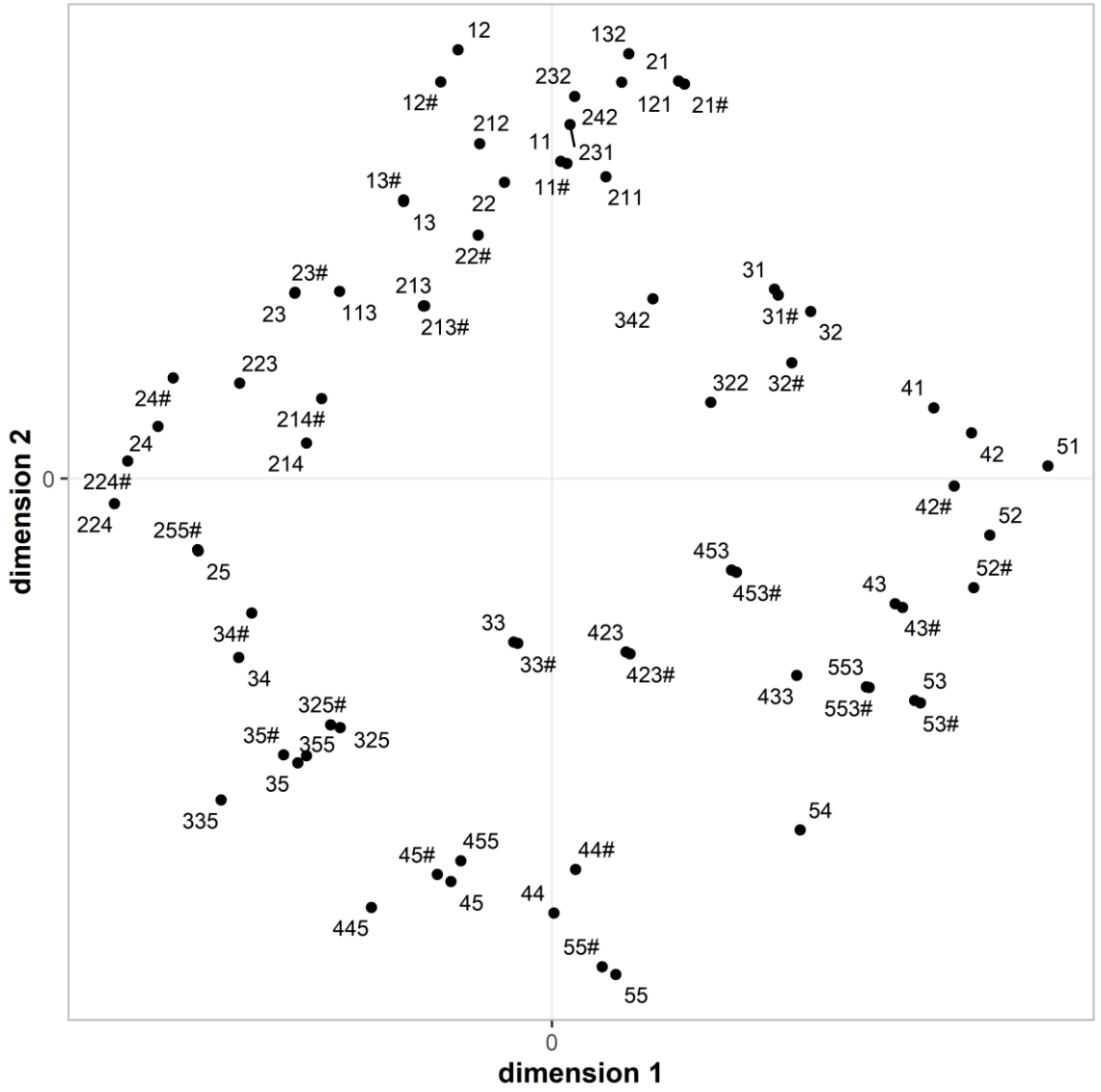
Modified Onset-Contour-Offset (mOCO)

- Modified from Yang and Castro's (2008) representation
- Converts Chao's (1930) transcription system into 3 parts:
  - Onset (starting pitch of the tone)
  - Contour (shape of the tone)
  - Offset (ending pitch of the tone)
- Here is an example:



Tone distances calculated with mOCO

- Tone distances: applying Levenshtein Distance (Levenshtein 1966, Heeringa 2004) on the mOCO representation
- This representation corresponds to the perceptual dimensions (Gandour and Harshman 1978)
- It can differentiate 72 out of 73 of the tones in the data



Preliminary analysis of tonal variation

Universiteit
Leiden
The Netherlands

Tonal variation

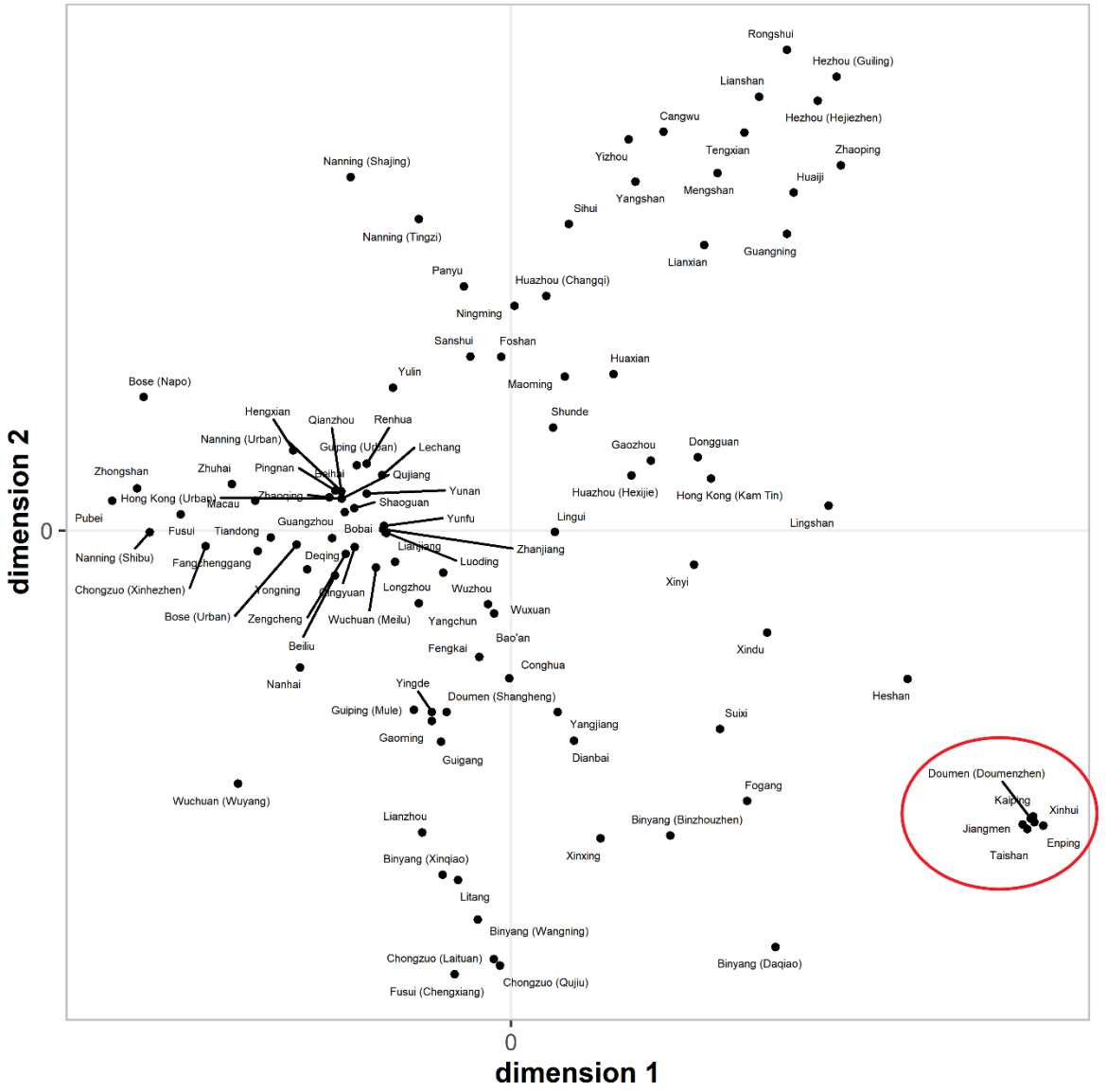
- In Chinese dialectology, very often studies on tones are descriptive
  - Correspondences between Middle Chinese tone categories and present-day pronunciation
- Existing dialectometric studies of tonal languages
  - Sometimes tones were neglected
  - Typically with a small dataset
  - Often combined with segments
  - The tone distance metrics cannot differentiate a lot of the tones in the data (Sung et al. Forthcoming)

Tone distances between dialects

- Between each pair of dialects in the data
  - -Calculate the Levenshtein distance between the tones of each word
  - -Sum the distances and divide it by the number of words compared (normalisation)

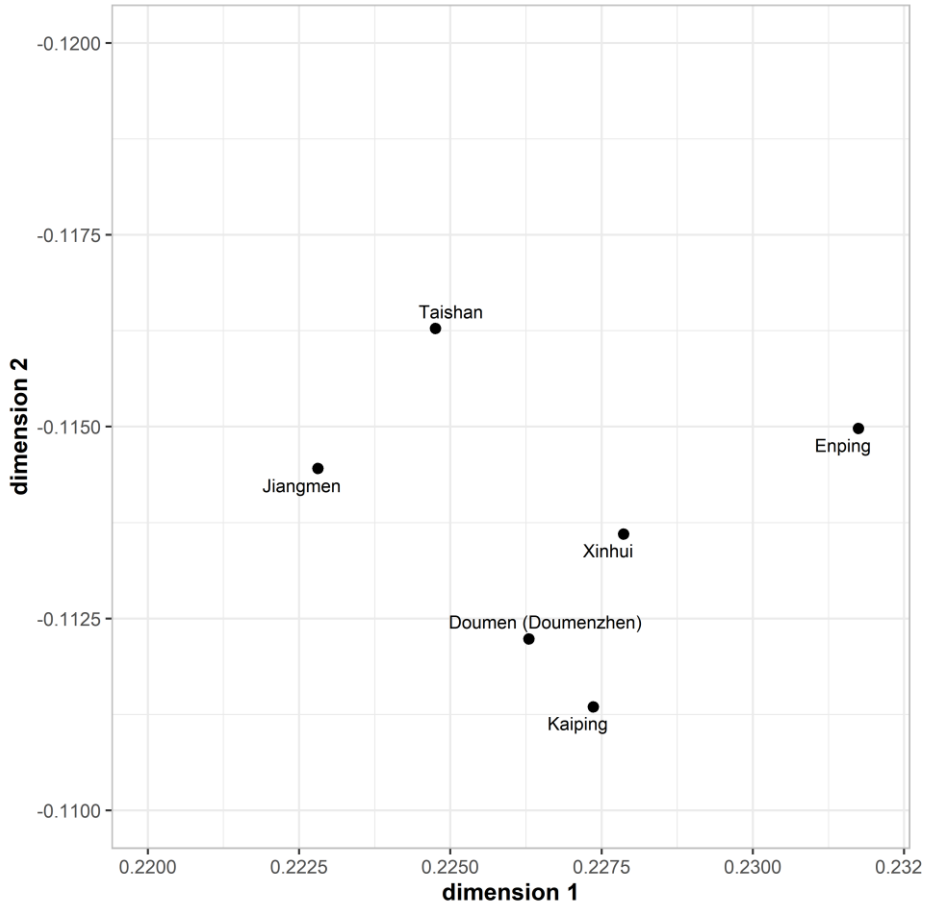
Dialect distances of Yue (tones only)

- MDS plot of 104 Yue dialects
- General pattern: continuum-like
- There is a clear cluster which seem to be isolated from the broader continuum
  - Siyi dialects (red circle)



Dialect distances of Siyi dialects (tones only)

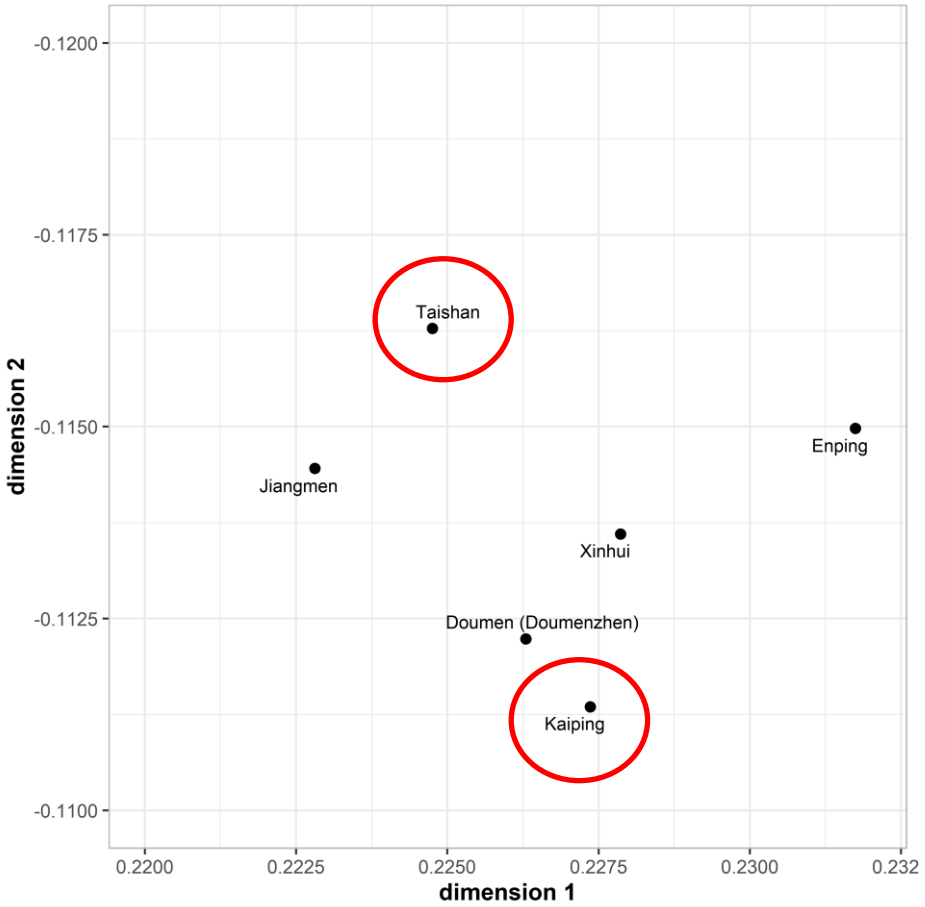
- Siyi dialects
- We can see 3 patterns here:
  - Group 1: Doumen, Taishan, Kaiping
  - Group 2: Enping
  - Group 3: Jiangman, Xinhui



Tone Categories	Doumen	Taishan	Kaiping	Enping	Jiangmen	Xinhui
Yin Ping	33	33	33	33	23	23
Yang Ping	22	22	22	22	22	22
Yin Shang	55	55	55	55	**45**	**45**
Yang Shang	21	21	21	**31**	21	21
Qu	31	31	31		31	31
Yin Ru1	55#	55#	55#	55#	55#	55#
Yin Ru2	33#	33#	33#	33#	33#	33#
Yang Ru	21#	21#	21#	21#	21#	21#

Dialect distances of Siyi dialects (tones only)

- Group 1 dialects
- Same tonemic inventory
- Why do they not overlap on the MDS plot?
- How do we explain this pattern?



Tone Categories	Doumen	Taishan	Kaiping	Enping	Jiangmen	Xinhui
Yin Ping	33	33	33	33	23	23
Yang Ping	22	22	22	22	22	22
Yin Shang	55	55	55	55	**45**	**45**
Yang Shang	21	21	21	**31**	21	21
Qu	31	31	31		31	31
Yin Ru1	55#	55#	55#	55#	55#	55#
Yin Ru2	33#	33#	33#	33#	33#	33#
Yang Ru	21#	21#	21#	21#	21#	21#

Tone correspondences

- Tone distances has to be accompanied with a tone correspondence table to get further insights
- We see *Lexical Distribution* differences (Wells 1982), or 'exceptions' in grey
- These are detected and reflected in the aggregate tone distances

Correspondences	No. of Items
11# : 21#	1
21:21	4
21:31	2
21# : 21#	12
21# : 33#	1
22:22	21
22:55	1
31:31	10
33:21	2
33:33	33
33:55	1
33# : 21#	1
33# : 33#	3
35:21	1
55:55	26
55# : 55#	11

Correspondence Table of Tones between Taishan (left) and Kaiping (right) Dialects (irregular correspondences in gray)

Conclusion

New Dataset

- 104 Yue-Pinghua dialects
- 130 words
- Segments and tones

New tone distance metric

- mOCO is able to differentiate 72/73 tones in the dataset

New discovery on tonal variation

- There seems to be a dialect continuum on the tonal level in the Yue-speaking area
- Tones can vary on the tonetic, tonemic as well as lexical distribution

Data availability

OSF repository:

https://osf.io/j5gxz/

- Contains the data and

Tone conversion scripts (Python)

Full datasets are under embargo at the moment

Contact info:
Matthew Sung:
h.w.m.sung@hum.leidenuniv.nl

Thank you for your attention!

Universiteit Leiden
The Netherlands

Contact info:
Matthew Sung:
h.w.m.sung@hum.leidenuniv.nl
Personal website:
https://sites.google.com/view/matthew-sung-dialectologist/home

Discover the world at Leiden University