Advancing explanatory and tonal dialectometry

Sung, H.W.M.

CHAPTER 4

Data[1]

4.1 Introduction

In Chinese dialectology, a lot of the dialect survey data are still published in physical books, including recent published data (e.g. Zhuang and Bei 2022). Digital data is not always accessible and freely available. Despite the fact that there are platforms which allow you to view the dialect survey data digitally (e.g. 中國語言資源保護研究中心 [Research Centre of Linguistic Resource Reservation in China] 2022), they are still not downloadable for users. These factors cause barriers to the development and application of computational methods for Chinese dialects, as well as many non-Chinese tonal languages. One of these problems includes whether the existing dialectometric methods (e.g. Yang and Castro 2008) are suitable and adequate to deal with tones in tonal languages (Sung et al. 2025).

Taking Chinese dialectology as an example, there are numerous studies on dialects spoken in China, and it has a century-long tradition, but most studies on tonal variation are descriptive. Traditional studies usually report the tonal inventory of a dialect after a fieldwork investigation, and/or tones are analysed in terms of how they correspond to historical

[1]This chapter is based on Sung et al. (2024).

tone categories (from the Middle Chinese period, based on the ancient rhyme dictionary descriptions).

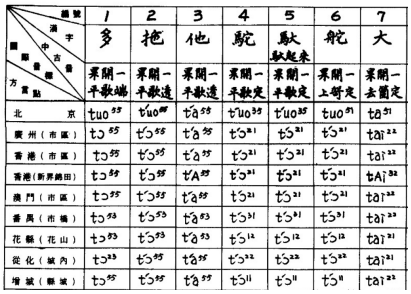
Until today, there is still a very limited number of digital datasets which allow us to quantitatively model variation of tones, which is problematic given that the majority of the world's languages are tonal (Yip 2002). Furthermore, even though there are tools which allow us to align Southeast Asian tone languages (Wu et al. 2020), and then extract and visualise the correspondences (both tones and segments) in table form (List 2019), these tools were developed for historical linguistics. In order to understand the synchronic dialect variation on the tonal level, alternative methods are needed in order to investigate how tones vary beyond correspondences.

This chapter will introduce the dataset which is used throughout the entire dissertation. The digital dataset that the dissertation is based on comes from Sung et al. (2024), with additional data from nine Northern Pinghua varieties, which are used to address the Yue-Pinghua dichotomy controversy in Chapter 5. This introduction starts with the sources of the dataset in Section 4.2. Next, the segmental data and tonal data are introduced in Section 4.3 and 4.4 respectively. The chapter ends with a conclusion in Section 4.5.

4.2 Data sources

The data presented in this chapter consists of phonetic transcriptions (in IPA) for both segments and tones. The dataset covers over 130 words in 113 dialects (104 dialects from Sung et al. (2024) and 9 additional locations in the current thesis) over the Yue-Pinghua-speaking area in Southern China.

The data were digitised from a number of dialect surveys and individual studies. There are two main sources for the dataset, namely word lists and homonymic syllabaries. Both sources are based on impressionistic transcriptions from word elicitation, but they are presented differently. Word lists are word-based, meaning words are organised in a tabular format (Francis 1983: 105-106), where the IPA transcriptions of each word are listed for each dialect all at once (see Figure 4.1a). On the other hand, homonymic syllabaries are pronunciation-based, meaning words with the same pronunciation are grouped together under one pronunciation (represented by the IPA transcriptions, see Figure 4.1b).



調號 / 字 / 中古音聲母 / 方言點	1 多 果開一平歌端	2 拖 果開一平歌透	3 他 果開一平歌透	4 駝 果開一平歌定	5 馱 馱起來 果開一平歌定	6 舵 果開一上哿定	7 大 果開一去箇定
北京	tuo^{55}	t'uo^{214}	t'a^{55}	t'uo^{35}	t'uo^{35}	tuo^{51}	ta^{51}
廣州（市區）	tɔ^{55}	tʻɔ^{35}	tʻa^{33}	tʻɔ^{21}	tʻɔ^{21}	tʻɔ^{21}	taːi^{22}
香港（市區）	tɔ^{55}	tʻɔ^{35}	tʻa^{33}	tʻɔ^{21}	tʻɔ^{21}	tʻɔ^{21}	taːi^{22}
香港（新界錦田）	tɔ^{55}	tʻɔ^{35}	tʻA^{33}	tʻɔ^{21}	tʻɔ^{21}	tʻɔ^{21}	tʌːi^{22}
澳門（市區）	tɔ^{55}	tʻɔ^{35}	tʻa^{33}	tʻɔ^{21}	tʻɔ^{21}	tʻɔ^{21}	taːi^{22}
番禺（市橋）	tɔ^{53}	tʻɔ^{35}	tʻa^{33}	tʻɔ^{21}	tʻɔ^{21}	tʻɔ^{21}	taːi^{22}
花縣（花山）	tɔ^{53}	tʻɔ^{35}	tʻa^{33}	tʻɔ^{12}	tʻɔ^{12}	tʻɔ^{13}	taːi^{21}
從化（城內）	tɔ^{33}	tʻɔ^{35}	tʻa^{35}	tʻɔ^{22}	tʻɔ^{22}	tʻɔ^{22}	taːi^{21}
增城（縣城）	tɔ^{55}	tʻɔ^{35}	tʻa^{33}	tʻɔ^{11}	tʻɔ^{11}	tʻɔ^{11}	taːi^{21}

(a) Sample of a dialect survey (from Zhan and Cheung 1987)

(b) Sample of a homonymic syllabary (from Zhong 2015)

Figure 4.1: Sources of Yue dialect transcriptions

Segments contain impressionistic transcriptions of consonants and vowels of the words. On the other hand, tones are represented by impressionistic transcriptions of pitch contours, using Chao's (1930) *tone letters*. The two sets of transcriptions are from the same sources; they were extracted from the same words (see below) and from the same dialects in the same sources.

4.2.1 Sources

The dataset used in the current dissertation consists of over 130 words in 113 dialects. These dialects include traditional Yue and Pinghua dialects (Chinese Academy of Social Sciences (CASS) 2012, see Chapter 2), which are Sinitic languages spoken in the Guangdong and Guangxi provinces in Southern China.

The dialect surveys include *Survey of Dialects in the Pearl River Delta* (*SDPRD*, Zhan and Cheung 1987), *Survey of Yue Dialects in Northern Guangdong* (*SYDNG*, Zhan and Cheung 1994), *Survey of Yue Dialects in Western Guangdong* (*SYDWG*, Zhan and Cheung 1998), *The Phonological Study of the Yue Dialects spoken in the Zhan-Mao area in Western Guangdong* (*SYDZM*, Shao 2016), *Chinese Dialect Research in the Guangxi Province* (*CDRGP*, Xie 2007), *Yue, Pinghua and Tuhua Dialect Survey Collection Part 1* (*YPTDSC1*, Chen and Lin 2009) and *Yue, Pinghua and Tuhua Dialect Survey Collection Part 2* (*YPTDSC2*, Chen and Liu 2009). Other (individual) studies include Liu (2015), Zhong (2015), Huang (2006), Chen (2009), Yang (2013), Tan (2017), Shi (2009)

and Chen and Weng (2010).

A map of the localities and the source of the dialect data can be found in Figure 4.2.

4.2.2 Selection of words

Out of the 130 words in the Yue dialect dataset, a portion of the items comes from the Swadesh 100-word list (Swadesh 1955), while some additional items come from outside this list. The Swadesh list is chosen because it is a standard word list for language comparison, with the assumption that words on this list represent the basic or core vocabulary – words that are universal, relatively culture-free and thus less likely to be replaced compared to other vocabulary (Campbell 2013: 448). In addition, Swadesh's 100 basic-word list has been tested by Wang and Wang (2004) to be the most suitable word list for sub-grouping Chinese dialects.

Not all items from the Swadesh list, however, are applicable for the dialectometric analysis in the current thesis. These include polysyllabic words, items excluded from the Yue dialect surveys and literary readings of characters.

The data collected in the Yue and Pinghua dialect surveys are mainly monosyllabic words (also known as cognate morphemes, character reading), because records of polysyllabic words are not available (collected) for a big portion of the dialects in the sources. In this dissertation, 'word' is used to refer to these monosyllabic words. Due to the specific type of data available, only a subset of the items in the Swadesh list is used, in order to ensure the commensurability of the dataset for all dialects.

Another group of (monosyllabic) items from the Swadesh list was excluded because they were not included in the dialect surveys used in this dissertation. For example, the Cantonese word for 'tongue' is '脷' *lei3* is not included in the Yue dialect surveys. Therefore, these items are not included in the dataset.

The third group of words which is not applicable for the dissertation includes items (characters) which can have two pronunciations, namely *literary* and *colloquial* pronunciations. Colloquial pronunciations of the characters usually reflect the pronunciation inherited by the dialects from their ancestors, while literary pronunciations are borrowings from

[2]Map created using *QGIS* (QGIS Development Team 2022).

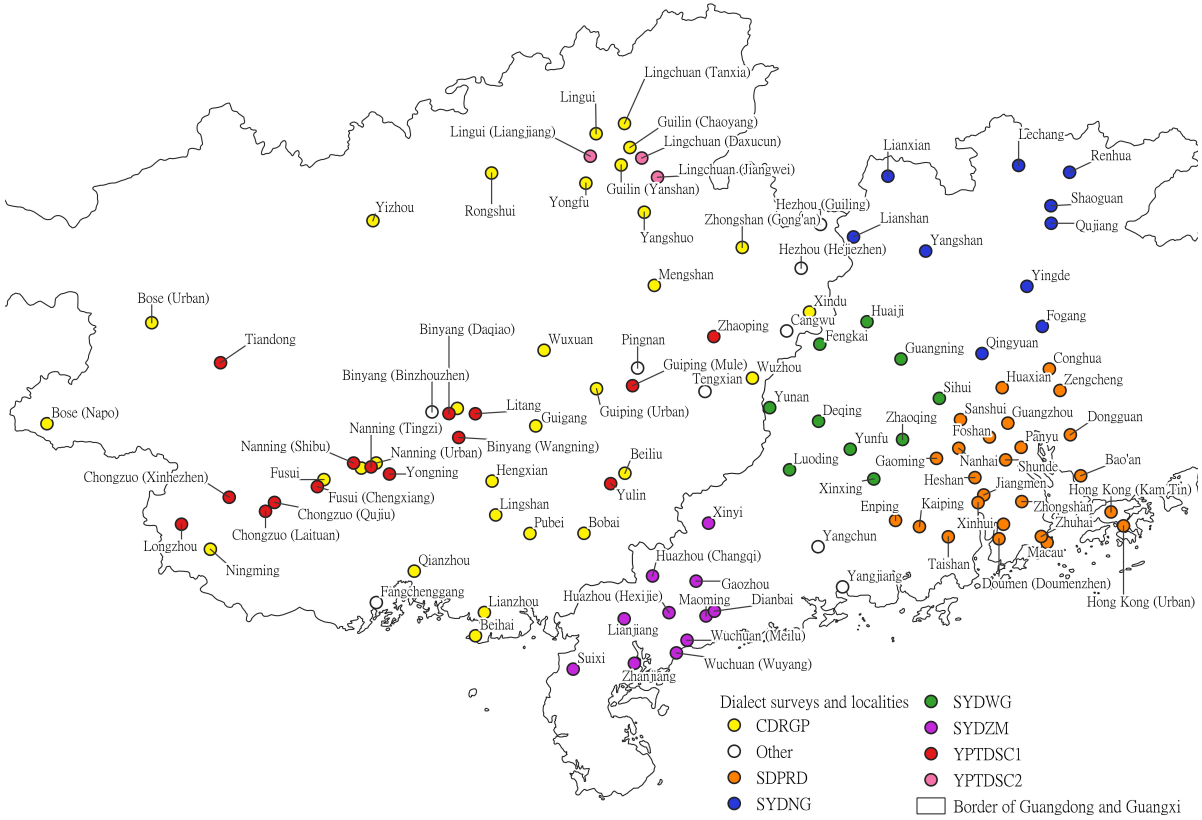


Figure 4.2: Localities and their respective sources[2]

the koine from different historical periods (Li 2007: 93). Although Yue has relatively fewer characters with literary pronunciations (Lau 2001: 134-135), they are still present in the Swadesh list, like 聽 'listen' (Lit. *ting3*, Col. *teng1*). These pronunciations are often recorded in the Yue dialect surveys, and therefore, such items are discarded.

It should be mentioned that some words consist of synonyms, which have the same meaning, but their characters (and pronunciations) are different. For instance, '鳥' and '雀' both mean 'bird'. '鳥' is used more in texts (although people do say terms like '觀鳥' *bird watching*), and '雀' is used predominantly in spoken Cantonese. Both words are included in the dataset for a number of reasons. Synchronically, even though '鳥' is not a common spoken word in some varieties of Yue, it has been found to yield interesting insights for dialectal comparison, since in a synchronic comparison, words more often found in texts also show interesting patterns. In the case for 'bird', diachronically speaking, '鳥' also demonstrates later sound changes that occurred after the word entered these varieties. For instance, this word shows an n-l alternation between dialects, which is a common merger or sound change in the Guangfu area. Furthermore, a number of Pinghua varieties have a t- onset for '鳥', which is quite unique for these varieties. Items like these should not be neglected simply because they are more common in textual language nowadays. They have their own merits in a synchronic (and to some extent, diachronic) analysis. In total, there are 3 pairs of synonyms above in the data. This would make 74 items (including synonyms) from the Swadesh list being included in the dataset.

In addition to the items in original Swadesh list, additional items were added. In total, 56 words in the dataset do not come from the Swadesh list. These words can be considered to be common, although not 'basic' or 'core' as such. The domains of these supplementary words include the rest of the numbers up to ten (Swadesh list only includes 'one' and 'two'), colour terms, direction, animals, and some words with known phonological variation, like 'flower', 'spring', and 'duck'. This addition is set out to enlarge the range of variation within the Yue dialects which are not present in the Swadesh list already.

The list of items can be found in Appendix A.

4.3 Modifications to the original segmental transcriptions

Dialect survey data are often found with transcribers' differences (or fieldworker isoglosses, Trudgill 1983; Mathussek 2016) when there are more than one fieldworker documenting dialects in the field. Transcribers'differences are inconsistencies of impressionistic transcriptions. These inconsistencies could arise due to the different uses of phonetic symbols to represent the same sound by different transcribers; transcribers perceive the same sounds differently or the transcriptions differ in the level of details. In other words, the differences we see in the data might be due to the habitual difference of the fieldworker instead of 'real' linguistic difference. To reduce the effects from the transcribers' differences, some modifications have been made to the original data.

4.3.1 Comparison with existing recordings

The data sources used in this dissertation do not have acoustic data accompanying the transcriptions. One of the ways to find out whether transcribers used different symbols to represent the same sound is to compare these transcriptions with the existing recordings from different projects on the same or nearby dialects. Recordings from the Yubao database (中國語言資源保護研究中心 [Research Centre of Linguistic Resource Reservation in China] 2022) were used for such comparisons. For instance, this task allows us to identify sub-phonemic contrasts such as Cantonese [ɵ] (International Phonetic Association 2005) before -n, -t and -y, which are often transcribed as <œ> in the transcriptions in varieties such as Guangzhou and Hong Kong (Urban) dialects.

4.3.2 Maintaining contrasts

Another approach to reducing transcribers' differences is to collapse contrasts between different notations, i.e. to merge symbols. However, this would potentially lead to a loss of information, with the risk of merging actual contrasts which are present in different dialects. To avoid collapsing unnecessary contrasts, when minimal pairs could be found in the rhyme inventory (provided in the dialect surveys for all localities), contrasts would be kept.

For example, one common difference in the transcriptions is the high back vowel symbol before -ŋ, namely [ʊŋ]. The tendency across Yue di-

alects is that there are two non-low back vowels which commonly pair with -ŋ, namely /ʊ/ and /ɔ/. Based on this tendency, we can derive the phonetic values of the vowels by inspecting the symbols used and the phonemic contrasts in the dialects. The main transcriptions of [ʊŋ] are <oŋ> and <uŋ> cross-dialectally. <oŋ> has been chosen to be the default in representing [ʊŋ]. However, the tendency does not imply all instances of <uŋ> represent a [ʊŋ]. To make the judgement more plausible, the following have been checked: 1) whether the inventory also has <oŋ>, and 2) whether <oŋ> could represent some other sounds, such as [ɔŋ]. This relies on the presence of minimal pairs. In the Hong Kong (Kam Tin) dialect, the original data have <uŋ> and <oŋ>. In addition, the Hong Kong (Kam Tin) dialect also has <ɔŋ>. Because [ɔ] already occupies the vowel in <ɔŋ>, <oŋ> that implies the pronunciation [ʊŋ]. At the same time, it implies that <uŋ> has the value [uŋ], a combination of a sound sequence uncommon across the Yue dialects (as a result of a sound change, < *-un).

In contrast, in the Nanning (Urban) dialect, <uŋ> does not form a minimal pair with <oŋ> (as it does not exist). Furthermore, the absent <oŋ> cannot be [ɔŋ] since <ɔŋ> already exists in the inventory. This implies that <uŋ> represents [ʊŋ]. This is indeed the case in the recording from the Yubao database (under '南寧白話' [Nanning Cantonese]).

4.3.3 Removal of redundant characters

There are cases where symbols were added to the transcription in the original data, but they do not contribute to the actual phonetic realisation of the word. The <ɲi-> sequence is an example. In words such as 人 'man/human/people' (which is typically transcribed as <ɲiɐn> in Western Yue dialects), the -i- medial is not really perceptible in the Yubao recordings. The addition of <-i-> is perhaps due to the fact that [ɲ] often appears before an -i- medial, and it is analysed as an allophone of /ŋ/ (Shao 2016: 42) or /n/ (e.g. Zhan and Cheung 1998). For <ɲiɐn>, since [ɐ] is not a high vowel, the medial -i- then could be a convention which indicates the presence of /i/ (but phonetically silent). While this information could be useful in the synchronic phonological analysis of the dialect, it creates inconsistencies for the computational dialect comparison. Therefore, such redundant information was removed.

4.3.4 Simplification of overly detailed transcriptions

Different transcribers would transcribe sounds in different broadness. Some (usually a minority) are narrower, with all the diacritics included, while some are broader, without diacritics.

The different degrees of transcription broadness cause additional inconsistencies to the data. In order to level the broadness, diacritics have been removed for the vowel backness and height parameters. For example, Hong Kong (Kam Tin) dialect has a non-standard IPA symbol <A>, which stands for [a̱]. This is further simplified to <a>. Superscripted segments, such as ^{u}V, NC (nasal+obstruent, could be ^{N}C or N^{C}), were all treated as full segments. This is because it is difficult to verify the status of the ^{u} in the ^{u}V sequence. For the nasal+obstruent sequence, some descriptions noted that these sequences have variation, like the Guangning dialect (Zhan and Cheung 1998: 14), which were not reflected in the data (the actual transcription of words). Hence a decision has been made to level these contrasts to full segments.

4.3.5 Consistency of onsets

Consistency of onsets mainly concerns word-initial high vowels. In Yue dialectology, it is common to see a zero-onset plus a medial (i.e. starting with i-, u- or y- instead of j- and w-) in the transcription, but not all transcribers do this. To my knowledge, only the Zhongshan dialect (Zhan and Cheung 1990: 72) and a few dialects in Guangxi (Xie 2007) do not start with a glide before a high vowel nucleus. For other dialects, it is unclear whether the choice between the vowel-initial vs. glide-initial reflects transcribers' differences. Therefore, the chosen normalised form is an onset with a glide for these syllables, until a further systematic report or survey of the presence (or absence) of an initial glide for the dialects in the dataset.

4.3.6 Converting Chinese IPA to Standard IPA

There are a few differences between the Chinese IPA and Standard IPA (International Phonetic Association 2005). These non-standard IPA symbols were converted to Standard IPA. For instance, the symbol for aspiration <'> was replaced with <^{h}>; capital vowel symbols <A> and <E> (roughly [a̱] and [e̝]/[ɛ̝] (between [e] and [ɛ]) respectively, Handel 2015; Li 2017: 31) were converted into diacritic-less IPA symbols (see

Section 4.3.4). One exception is the apical vowel <ɿ>, which remains in the dataset as a contrastive sound to the existing IPA symbols.[3] In terms of consonants, palatal nasals <ɲ> and laminal <ʃ>[4]are replaced by IPA <ɲ> and <s> respectively.

4.3.7 Phonetic alignment

A modification has been made for the ku- sequence, opposed to the kv- and kʊ- sequences. All the ku- sequences were converted to kw-, so that the medial -u- would be treated as a consonant. In quantitative language comparison, phonetic transcriptions are often aligned using pairwise or multiple sequence alignment algorithms (see Chapter 5 and 6). Introducing the above mentioned modification allows the medial -u- to be aligned with -v- and -ʊ-, instead of a nucleus vowel which does not belong to the onset.

4.3.8 Descriptive statistics

In Table 4.1, some examples have been listed for how the data would look before and after the modification.

Dialect	Character	Item	Raw	Modified
Guangzhou	水	'water'	sœy	sɵy
Guangzhou	秋	'Autumn'	ts'ɐu	ts^{h}ɐu

Table 4.1: Examples of Raw vs. Modified Transcriptions

A reduction in the contrasts from the raw data can yield information loss. I have calculated the *Normalised Levenshtein distance* (Levenshtein 1966; Heeringa 2004, see Section 3.3.1) to see how much the modified transcription deviate from the raw transcription. The distribution of the deviation scores per dialect can be found in Figure 4.3 below.

The mean Levenshtein distance is 0.043, and the standard deviation is 0.029. The minimum distance found between the raw and the modified transcriptions within a dialect is 0.004 (found in Zengcheng), while the maximum distance is 0.121 (found in Binyang (Binzhouzhen)).

[3]There could actually be more than one phonetic realisation for what is represented as an 'apical vowel'. However, since this information is not available in the dialect survey data, I treat this pool of possible sounds as one homogenous sound value by using the apical vowel symbol.

[4]Chinese IPA uses tongue positions instead of the palate as the places of articulation.

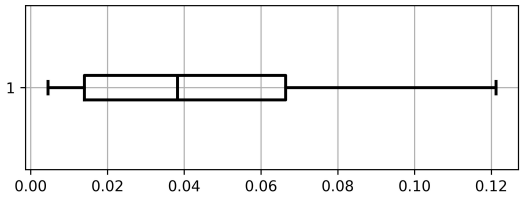


Figure 4.3: Boxplot of Distances between Raw and Modified Transcriptions.

The descriptive statistics of the raw vs. modified transcriptions do not suggest a huge deviation from the raw data after I have removed some potential transcribers' differences. On average, we might see 4 changes per 100 segments (mean value 0.04 multiplied by 100).

4.4 Tonal data

The second half of the dataset consists of the tonal data from the same words and the same dialects. To date, there are no large-scale dialectometric studies on tones; the highest number of dialects involved is no more than 30 dialects (see Yang and Castro 2008; Tang 2009). In some dialectometric studies, tones were neglected (e.g. Wichmann and Ran 2019), while others used a rather simplified method (e.g. Stanford 2012). In addition, there are studies on the correlation between phonetic distance and the perception of tones (e.g. Yang and Castro 2008), which do not focus on the application of these measures on dialect classification. Research questions regarding the variation of tones in larger dialect areas, or if there is a correlation between tonal and segmental variation, cannot be researched upon using these datasets.

The current tonal dataset is different from the existing ones, since it allows comparisons between tonal and segmental levels. The tones were transcribed in Chao's (1930) tone letters, which is a system for tone transcription consisting of 5 digits, 1, 2, 3, 4, 5, representing different (possible) contour levels in a tone. In this system, 1 represents the lowest contour level and 5 represents the highest. When combined (with two digits or three digits), they can indicate a change in the contour, which represents the shape of the tone. For example, 53 is a falling tone, whereas 213 is a dipping tone (a falling contour followed by a rising

contour). More about tonal notations can be found in Chapter 7.

4.5 Conclusion

Tonal languages have been neglected in the study of linguistic variation for decades, partly due to the lack of available data. The Yue dataset created by Sung et al. (2024) provide new possibilities in the study of language variation. It consists of both tonal and segmental data for the same lexical items for over 100 dialects. To my knowledge, this is one of the biggest dialectal dataset for tones within one language area.

This chapter introduces the dataset used in this dissertation. The dataset is split into segmental and tonal parts. The segmental data are used in Chapter 5 and 6; the tonal data are used in Chapter 7 and 8.

The dataset digitised by Sung et al. (2024) is publically available at: `https://osf.io/m9g2a/`.