A New Dataset for Tonal and Segmental Dialectometry from the Yue- and Pinghua-Speaking Area

Ho Wang Matthew Sung[1], Jelena Prokić[2] and **Yiya Chen[3]**
Leiden University / Leiden, Netherlands
{h.w.m.sung[1], j.prokic[2], yiya.chen[3]}@hum.leidenuniv.nl

Abstract

Traditional dialectology or dialect geography is the study of geographical variation of language. Originated in Europe and pioneered in Germany and France, this field has predominantly been focusing on sounds, more specifically, on segments. Similarly, quantitative approaches to language variation concerned with the phonetic level are in most cases focusing on segments as well. However, more than half of the world's languages include lexical tones (Yip 2002). Despite this, tones are still underexplored in quantitative language comparison, partly due to the low accessibility of the suitable data. This paper aims to introduce a newly digitised dataset which comes from the Yue- and Pinghua-speaking areas in Southern China, with over 100 dialects. This dataset consists of two parts: tones and segments. In this paper, we illustrate how we can computationaly model tones in order to explore linguistic variation. We have applied a tone distance metric on our data, and we have found that 1) dialects also form a continuum on the tonal level and 2) other than tonemic (inventory) and tonetic differences, dialects can also differ in the lexical distribution of tones. The availability of this dataset will hopefully enable further exploration of the role of tones in quantitative typology and NLP research.

1 Introduction

Traditional dialectology or dialect geography (Chambers and Trudgill 1998: 14), is the study of geographical variation of language. This field originated in Europe, pioneered in Germany and France, and has always been focusing on sounds, more specifically, on segments. In the second half of the 20th century, the quantitative turn in dialectology, known as *dialectometry*, is no exception to this. While these methodologies have been widely used in Europe and America, there are only limited regions in the rest of the world which employ these methodologies, although there is a sign of growth in recent years. For instance, Yucatec Mayan (Pfeiler and Skopeteas 2022), Bantu languages (Nerbonne 2010), Japanese (Jeszenszky et al. 2019).

Tonal languages are defined as "[languages] in which an indication of pitch enters into the lexical realisation of at least some morphemes" (Hyman 2006: 229), and more than half of the world' s languages include lexical tones (Yip 2002). Despite this, tones are still gaining little attention in quantitative models of language variation. This lack of attention on tones is not surprising, however, since most European languages mostly do not use pitch to differentiate lexical meaning. One other reason could be the fact that digital data is not accessible and freely available. These factors cause barriers to the development of computational methods for tonal languages. For example, it is unclear whether the existing methods (e.g. Yang and Castro 2008) are suitable and adequate to deal with tones in tonal languages (Sung et al. forthcoming).

Take 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 tone categories (from the Middle Chinese period, based on the ancient rhyme dictionary descriptions). Although there is a huge amount of dialect data available for Chinese (in the form of IPA transcriptions, including tones), they are mostly printed on paper and are not digitised, ready to be used for quantitative analyses.

Until today, there is generally a very limited number of digital datasets which allows us to quantitatively model variation of tones, which is problematic given that the majority of the world's languages are tonal. Furthermore, although there are tools which allow us to align Southeast Asian tone

Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (SIGTYP 2024), pages 25–36

languages (Wu et al. 2020), and then visualize 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 paper aims to introduce a newly digitised dataset which comes from the Yue- and Pinghua-speaking areas in Southern China, with over 100 dialects.[1] This dataset consists of two parts: segments (Section 3) and tones (Section 4). The availability of this dataset will hopefully be an invitation to researchers around the world to initiate an exploration of tonal variation, which has long been neglected. In section 5 we present out preliminary research on tonal variation, followed by a conclusion.

2 Data Sources

The data presented in this paper consists of segments and tones. Segments contain impressionistic transcriptions of consonants and vowels of the words. Impressionistic tone transcriptions of pitch contours are represented 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.

There are two main sources for the dataset, namely word lists and homonymic syllabaries, which came from various dialect surveys and individual studies. 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. 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).

2.1 Sources

Our dataset consists of IPA transcriptions of over 130 words in 104 dialects. These dialects include traditional Yue and (Southern) Pinghua dialects (Chinese Academy of Social Sciences (CASS) 2012), 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). Other (individual) studies include Liu (2015), Zhong (2015), Huang (2006), Chen (2009), Yang (2013), Tan (2017), Shi (2009) and Chen and Weng (2010).

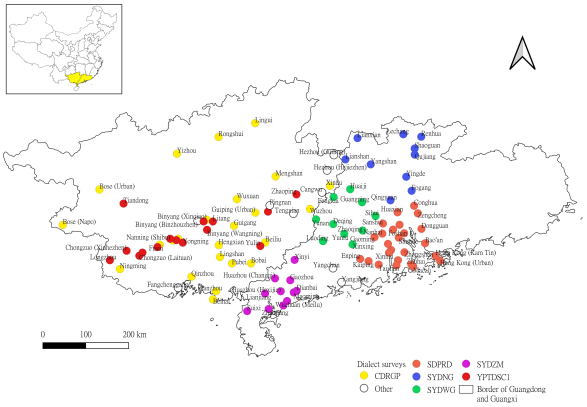

Figure 1: Localities and their respective sources.[2]

2.2 Selection of Words

Out of the 130 words in our 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 are, however, suitable for the data extraction process. One

[1] The datasets can be found under **Supplementary Material**.

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

group of such items are polysyllabic words. The data collected in the Yue and Pinghua dialect surveys are mainly monosyllabic words (or cognate morphemes), because records of polysyllabic words are not available (collected) for a big portion of the dialects in this dataset. Therefore, 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 items from the Swadesh list was excluded because they were not included in the dialect survey. An example is 'tongue'. The pronunciation of the written form of this word (舌 *sit3*) is included in the Swadesh list, but the actual spoken form used in Cantonese, 脷 *lei3*, is not included. One other group of words include items which can have two pronunciations, namely *literary* and *colloquial* pronunciations. Colloquial pronunciations of the characters usually reflect the pronunciations inherited by the dialects from their ancestors, while literary pronunciations are borrowings from the koine from different historical periods (Li 2007: 93). Although Yue has relatively fewer characters with literary pronunciations (Lau 2001: 134-135), it is still present in the Swadesh list, like 聽 'listen' (Lit. *ting3*, Col. *teng1*). Therefore, such items are discarded. Lastly, some items are doublets, meaning that both the spoken and written forms can be found. Both variants are included in the dataset.

In total, 54 words in our 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 10 (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.

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 due to the different uses of phonetic symbols to represent the same sound by different transcribers. 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, we have made some modifications to the original data.

3.1 Comparison with Existing Recordings

The data sources we have used 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. We have used 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) for such comparisons. For instance, this task allows us to identify subphonemic contrasts such as Cantonese [ɵ] (International Phonetic Association 2005) before -n and -t, which are often transcribed as <œ> in the transcriptions in varieties such as Guangzhou and Hong Kong (Urban) dialects.

3.2 Maintaining Contrasts

Another approach to reducing transcriber's 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 transcriptions is the high back vowel symbol before -ŋ, namely [ʊŋ]. The tendency across Yue dialects 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. We have chosen <oŋ> to be the default in representing [ʊŋ]. However, the tendency does not imply all instances of <uŋ> represent a [ʊŋ]. To make the more plausible judgement, we have 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 sound change).

In contrast, in the Nanning (Urban) dialect, <uŋ> does not form a minimal pair with <oŋ> (since 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 also the case in the recording from the Yubao database (under '南寧白話').

3.3 Removal of Redundant Characters

There are cases where symbols were added to the transcription in the original data, but they do not actually contribute to the actual phonetic realization 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 dialect comparison. Therefore, such redundant information (for dialect comparison) was removed.

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, we have removed diacritics for the vowel backness and height parameters. For example, Hong Kong (Kam Tin) dialect has a non-standard IPA symbol <A>, which stands for [a̠]. 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. We have decided to level these contrasts to full segments.

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 our knowledge, 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 further reports of the presence (or absence) of an initial glide for the dialects in our dataset.

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 [a̠] and [e̞]/[ɛ̝] (between [e] and [ɛ]) respectively, Handel 2015; Li 2017: 31) were converted into diacritic-less IPA symbols. 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 <ȵ> and laminal <ʃ>[4]are replaced by IPA <ɲ> and <s> respectively.

[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, we 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.

3.7 Phonetic Alignment

We have also modified the kv- and kʊ-, versus the ku- sequence. 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. 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.

3.8 Descriptive Statistics

In Table 1, we have illustrated how the data look before and after the cleaning process.

Dialect	Item	Raw	Cleaned
Guangzhou	'water'	sœy	sœy
Guangzhou	'skin1'	p'ei	p^hei

Table 1: Examples of Raw vs. Cleaned transcriptions

A reduction in the contrasts from the raw data can yield information loss. We have calculated the *Normalized Levenshtein distance* (Levenshtein 1966; Heeringa 2004) to see how much our cleaned transcription deviate from the raw transcription. The distribution of the deviation scores per dialect can be found in Figure 2 below.

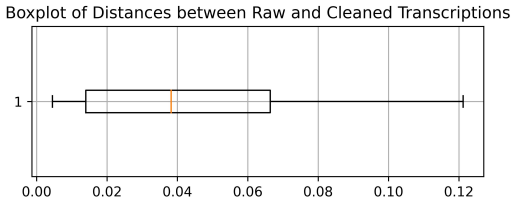

Figure 2: Boxplot of Distances between Raw and Cleaned Transcriptions.

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

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

4 Tone Data

The second half of the dataset consists of the tonal data from the same words and the same dialects. Up to date, there are no large-scale dialectometric studies on tones; the highest number of dialects involved are no more than 20 dialects (see Yang and Castro (2008); Tang (2009)). In some dialectometric studies, tones were neglected (e.g. Wichmann and Ran (2019)), 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 to 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.

Our tonal dataset is different from existing digital datasets, 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 together (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).

The tonal transcriptions cannot directly be used for the purpose of dialectometry, though. Sung et al. (forthcoming) have found that directly applying Levenshtein distance on the tone letters and comparing these tones categorically (the 'binary' method, Sung et al. (forthcoming)) do not yield satisfying tone distances for the purpose of dialectometry. Further conversion of these tones to another representation (e.g. Onset-Contour-Offset, Yang and Castro 2008) is required in order to get more meaningful tone distances. The availability of tones as tone letters allows users to apply any conversion of their choice. The question raised in Sung et al. (forthcoming) shows that currently there is no existing satisfying tone distance metric for dialectometry. In the subsections below, we briefly introduce three quantitative models of tone representation, tested in Sung et al. (forthcoming), as well as our modified version of the existing rep-

resentation proposed by Yang and Castro (2008)[5].

4.1 Chao's representation

The *tone-to-string* method applies the Levenshtein distance algorithm directly to Chao's (1930) tone letters. Levenshtein distance (Levenshtein 1966) is a string distance metric which seeks the least amount of operations, namely *insertions* (addition of element in string), *deletions* (removal of element in string) and *substitutions* (replacement of element in string) in order to transform one string into the other. The degrees of difference in the digits (pitch contours) are not accounted in this method, i.e. a substitution from 2 to 1 costs the same distance as from 4 to 1. This implementation of the tone-to-string method follows Tang (2009), where a two-digit tone aligns with a three-digit tone from the second digit of the three-digit tone, see the example below. Note that short tones are not distinguished from their longer counterpart, since it has not been proposed yet how tones like a short concave tone are represented under this method. Users should remove the '#' (length marker) in their data before applying this method. Take two tones, 15 vs. 325, as an example, we calculate the Levenshtein distance between the tones, which is demonstrated in Table 2. In this example, one substitution and deletion are required to convert 325 to 15, and that yields 2 / 3 = 0.67 difference between the two tones.

Slot 1	Slot 2	Slot 3	Operations	Distance
3	2	5	-	-
-	2	5	Deletion of 3	1
-	1	5	Substitution of 2 > 1	1
Sum				2

Table 2: Calculation of Levenshtein Distance between 325 and 15 with the tone-to-string method

4.2 Onset-Contour-Offset (OCO)

Onset-Contour-Offset (OCO hereafter) is a representation of tones proposed by Yang and Castro (2008). This representation gives a more phonetic representation of tones, instead of an abstract one, as its purpose is to approximate multiple cues of tones in the distance measure in order to generate a more accurate prediction for intelligibility between dialects (the purpose of Yang and Castro's study).

OCO involves a transformation of the tone letters/ 5-level transcription (Chao 1930) into a representation which consists of three components: *Onset*, *Contour* and *Offset*, each represented with one character, except for Contour, which can have up to two characters. Onset and Offset are the starting and ending contour levels of the tone, and the Contour is the shape of the tone. For the contour levels, the original 5-level transcription is converted into three categories, which are *H(igh)*, *M(id)* and *L(ow)*. H represents levels 4 and 5, M represents 3 and L represents 1 and 2. For contours, the basic shapes include *R(ising)*, *F(alling)*, *L(evel)*, and the complex tones are represented by the combination of the basic shapes, hence it has up to two characters. Examples of the Contour representations can be found in Table 3 below.

Representation	Contour	Example
L	Level	11, 33
R	Rising	12, 35
F	Falling	31, 52
RF	Convex	131, 253
FR	Concave	213, 424

Table 3: Contours in OCO representation with examples

As an example, the OCO representation of 221 would be LLFL, and for 24, it would be LRH. To calculate tone distances, Yang and Castro (2008) applied the Levenshtein distance algorithm on the OCO representation. This is illustrated in Table 4 below.

When two tones with different lengths are compared (length of three and four, like in Table 4), the Onset (Slot 1) and Offset (Slot 4) are always aligned together. In this example, we can find two substitutions and one deletion out of four alignment slots, which yields a Levenshtein distance of 0.75 between the tone pair.

Slot 1	Slot 2	Slot 3	Slot 4	Operations	Distance
L	L	F	L	-	-
L	R	F	L	Substitution of L > R	1
L	R	-	L	Deletion of F	1
L	R	-	H	Substitution of L > H	1
Sum					3

Table 4: Calculation of Levenshtein Distance between 221 and 24 with the OCO method

4.3 Modified Onset-Contour-Offset (mOCO)

In Sung et al. (forthcoming), it has been shown that the biggest drawback of the OCO represen-

[5]The scripts for the conversion of the original tone data to each of the representations introduced below are provided in the **supplementary material**. The converted tones can then be processed by existing dialectometric tools online, such as *Gabmap* (Nerbonne et al. 2011; Leinonen et al. 2016).

tation is that it is not able to distinguish enough tones in the Yue dataset (only 43.8%). Therefore, we made some adjustments in order to differentiate more tones present in our data. This representation is largely the same as OCO, and it still operates with a transformed representation of Chao's tone letters into onset-contour-offset.

Firstly, the pitch levels are expanded from originally differentiating three levels (merging 1 and 2 and merging 4 and 5) to distinguishing all five levels found in Chao's (1930) tone letters. Our modification creates a five-level contrast by having HH (5), H (4), M (3), L (2) and LL (1). The double representation 'HH' and 'LL' make their immediate neighbouring pitch levels, i.e. H and L respectively, costing a difference of 1, while all other pitch levels cost a difference of 2. Another modification we have made has to do with tone length. The differences in tone length are usually found between checked syllables and non-checked syllables. We have decided to represent tone length with a superscript ^{h}, which indicates a difference of 0.5.[6]

This representation of the highest and lowest pitch levels maintains the dimensions of 'direction' and 'average pitch', which has been identified in Gandour and Harshman (1978).

With the mOCO representation, when we apply Levenshtein distance between the tones in our dataset, we can differentiate 72 out of 73 tones (98.6%). The tone distances calcualted with the mOCO representation can be found in Figure 3. Since the mOCO representation can differentiate 72 tones, it should also be sufficient for other tonal languages in Southeast Asia, and perhaps in other parts of the world (given the same tone notation is used in the documentation, so that the conversion can be done).

5 Preliminary Results of Tonal Dialectometry of Yue Dialects

In this section, we will present our preliminary analysis of the tonal data, using the mOCO tone representation. The first question that we will try to answer is whether dialects form a dialect continuum, which has previously been observed in different languages on the segmental level. A follow-up question we investigated is in what ways do tones differ gradually from one dialect to another.

[6]Superscripted characters are counted as a difference of 0.5 in the Levenshtein algorithm implemented in *Gabmap*, if the last character of the *Offset* of both tones (but not the length) are identical. This implies that the tone length is only differentiated if the final character of the offset in the mOCO representation is identical. Please note that LED-A.org does not have the same implementation of the superscript ^{h}.

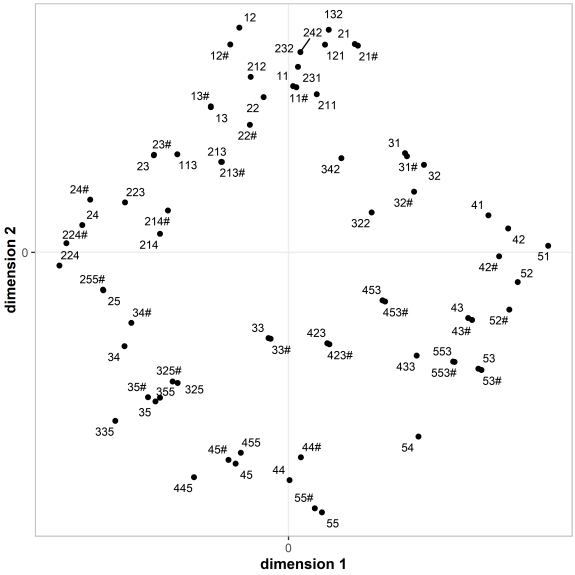

Figure 3: Multidimensional Scaling plot of tone distances calcualted with the mOCO representation.

5.1 Methodology

Firstly, the whole tonal dataset is converted to the mOCO representation. Next, for each pair of dialects, the tone distances for all the lexical items are calculated using normalized Levenshtein distance, summed and then divided by the total number of lexical items compared in the pairwise dialect comparison. These procedures yield a normalised aggregate tonal distance between all pairs of dialects in the data, which we store in a distance matrix. The distance calculation procedures described above were computed with *Gabmap* (Nerbonne et al. 2011; Leinonen et al. 2016).

5.2 Multidimensional Scaling Plot

Since a distance matrix is not interpretable for human eyes, we have employed *multidimensional scaling* to our distance matrix of the dialect tonal distances, in order to gain further insights into the tonal variation in our data. Multidimensional Scaling (MDS hereafter) is a dimensionality reduction method which represents "measurements of similarity (or dissimilarity) among pairs of objects as distances between points" (Borg and Groenen 2005: 3). In our case, an MDS plot would represent the dialects as points, and the further the points are from each other, the more different they

are. Unlike cluster analysis, the points on an MDS plot are not partitioned into discrete groups. In addition, no geographical information is added to the plot, so the distances projected on the plot is simply based on the distance matrix generated in the distance calculation. This technique is useful to visualize continuum-like dialect relations (existence of transitional dialects), as well as clusters. However, it requires one to interpret the plot themselves, including in what ways dialects differ from each other.

It is also important to check how much the distances represented in an MDS plot correlate to the original distance matrix. This is indicated by the explained variance (r^2) or by the Stress value (Heeringa 2004).

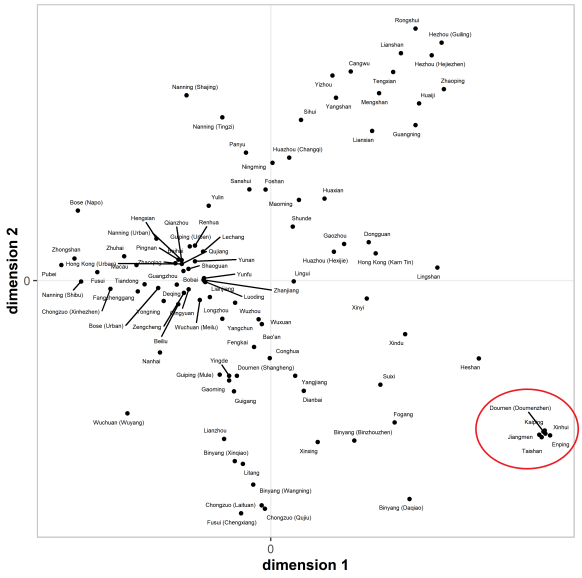

Figure 4: MDS plot of tone distances between Yue dialects (r^2= 0.70).

In Figure 4, we see a continuum-like distribution for the majority of the dialects in our data, with the possible exception of the Siyi dialects. They are marked with a red circle in Figure 4 and are clearly separate from the rest of the dialects. This corresponds to the analysis done on the segmental level (Sung 2023; Sung and Prokic 2023). This dialect group serves as our preliminary investigation into the ways in which tones vary in between dialects.

Figure 5 is a zoomed-in view of the Siyi cluster in Figure 4. We can see that although these dialects are relatively similar to each other in Figure 4, they do not completely overlap, meaning their tones are not completely identical. To gain more

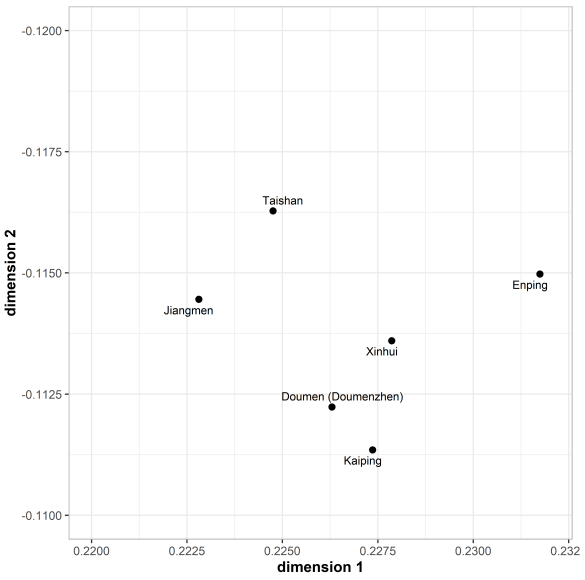

Figure 5: MDS plot of tone distances between Siyi dialects (Figure 4 zoomed in).

insights into how their tones differ from each other, we turn to the tonal inventories of these dialects.

In Table 5, the tonal inventories of the Siyi dialects are listed as the reflexes of the Middle Chinese (MC) tone categories. We can see that Taishan, Doumen and Kaiping dialects share the exact same inventory. Enping dialect has an almost identical inventory as these three dialects, except two MC tone categories share the same reflex, indicated by the merged cell in gray. Another group of dialects consists of Jiangmen and Xinhui. Their tonal inventories only differ from Taishan, Doumen and Kaiping dialects by one tone: the reflex of Yin Shang category is 45 instead of 55. Based on the inventories, we would expect the MDS plot to show overlaps of Taishan-type dialects (with Enping slightly further away), and the Jiangmen-type dialects to be even further away. This is however not the case in Figure 4. If we look at the tone correspondences between the Taishan and Kaiping dialects (see Appendix B), we can see that even though the tones in their inventories are identical, their correspondences are not perfect. This suggests that there are *lexical distribution* (Wells 1982) differences between these dialects occurring in the data.

Our preliminary results suggest that mOCO can detect both tonemic (inventory) and sub-tonemic (phonetic) differences between dialects. In addition, it can also detect lexical distributional differences between dialects with identical tone invento-

Tone Categories	Taishan	Kaiping	Doumen	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#

Table 5: Tone inventories of Siyi dialects (based on Middle Chinese tone categories)

ries.

6 Conclusion

Our Yue dataset has provided 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 our knowledge, this is one of the biggest dialectal dataset for tones within one language area. Our tonal dataset is digitised from dialects surveys which were transcribed in Chao's (1930), which means that it can be converted to any existing tone representations for further dialectometric analyses. In this paper, we have briefly demonstrated how we can use one of these representations to investigate how tones vary across different dialects. By using the mOCO representation, which can differentiate almost 99% of the tones in our data, we have identified a dialect continuum as well as a dialect island, namely the Siyi dialect group. Through a comparison of tone inventories and tone correspondences of Siyi dialects, we have further identified that dialects can differ on the tonal level tonemically, sub-tonemically and in terms of lexical distribution.

Tonal languages have been neglected in the study of linguistic variation for decades, partly due to the lack of available data. We hope this dataset will serve as the first step to remove the barrier for any scholars who are interested in variation of tones.

References

I. Borg and P. J. F. Groenen. 2005. *Modern multidimensional scaling: Theory and applications*. Springer Science and Business Media.

L. Campbell. 2013. *Historical linguistics*. Edinburgh University Press.

J. K. Chambers and P. Trudgill. 1998. *Dialectology*, 2nd edition. Cambridge University Press.

Y.-R. Chao. 1930. A system of tone letters. *Le maître phonétique*, 8(45):24–27.

H. Chen and Y. Lin. 2009. 粵語平話土話方音字彙第 *1* 編: 廣西粵語、桂南平話部分 *[Yue, Pinghua and Tuhua Dialect Survey Collection Part 1]*. Shanghai Educational Publishing House.

X. Chen. 2009. 廣西賀州八步 (桂岭) 本地话音系 [the phonology of the hezhou babu (guiling) dialect in guangxi]. 方言 *[Dialect]*, (1):53–71.

X. Chen and Z. Weng. 2010. 粵語西翼考察–廣西貴港粵語個案研究 *[Investigating Western Yue - A case study on Guigang Yue in Guangxi]*. Jinan University Press.

Chinese Academy of Social Sciences (CASS). 2012. 中國語言地圖集 *[Language Atlas of China]*, 2nd edition. Commercial Press.

W. N. Francis. 1983. *Dialectology: an introduction*. Longman Group Limited.

J. Gandour and R. A. Harshman. 1978. Crosslanguage differences in tone perception: A multidimensional scaling investigation. *Language and Speech*, 21(1):1–33.

Z. Handel. 2015. Non-ipa symbols in ipa transcriptions in china. In R. Sybesma, editor, *Encyclopedia of Chinese Language and Linguistics*. Brill Reference Online.

W. Heeringa. 2004. *Measuring dialect pronunciation using Levenshtein distance*. Ph.D. thesis, University of Groningen.

Q. Huang. 2006. 賀州市賀街本地話同音字匯 [homonymic syllabary of the hezhou hezhoujie local vernacular]. *Journal of Guilin Normal College*, 20(3):6–13.

L. M. Hyman. 2006. Word-prosodic typology. *Phonology*, 23(2):225–257.

International Phonetic Association. 2005. *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press.

P. Jeszenszky, Y. Hikosaka, S. Imamura, and K. Yano. 2019. Japanese lexical variation explained by spatial contact patterns. *ISPRS International Journal of Geo-Information*, 8(9):400.

C. Lau. 2001. 粵客方言文白異讀的比較 [the comparison between literary and colloquial readings in yue and hakka dialects]. In C. Lau, editor, 香港粵客方言比較研究 *[The Comparative Study of Hong Kong Yue and Hakka Dialects]*, pages 134–147. Jinan University Press.

T. Leinonen, Ç. Çöltekin, and J. Nerbonne. 2016. Using gabmap. *Lingua*, 178:71–83.

V. I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady*, 10(8):707–710.

R. Li. 2007. 漢語方言學 *[Chinese Dialectology]*, 2nd edition. Higher Education Press.

R. Li. 2017. 漢語方言調查 *[Surveying Chinese Dialects]*. Commercial Press.

Johann-Mattis List. 2019. Automatic inference of sound correspondence patterns across multiple languages. *Computational Linguistics*, 45(1):137–161.

C. Liu. 2015. 廣東兩陽粵語語音研究 *[Research in the Phonetics of Yue in the Guangdong Liangyang Area]*. Ph.D. thesis, Jinan University.

A Mathussek. 2016. On the problem of field worker isoglosses. In M.-H. Côté, R. Knooihuize, and J. Nerbonne, editors, *The future of dialects*, pages 99–116. Language Science Press.

J. Nerbonne. 2010. Measuring the diffusion of linguistic change. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1559):3821–3828.

J. Nerbonne, R. Colen, C. Gooskens, P. Kleiweg, and T. Leinonen. 2011. Gabmap-a web application for dialectology. *Dialectologia: revista electrònica*, pages 65–89.

B. Pfeiler and S. Skopeteas. 2022. Sources of convergence in indigenous languages: Lexical variation in yucatec maya. *PLoS ONE*, 17(5):e0268448.

QGIS Development Team. 2022. *QGIS Geographic Information System*. Open Source Geospatial Foundation Project.

H. Shao. 2016. 粵西湛茂地區粵語語音研究 *[The Phonological Study of the Yue Dialects spoken in the Zhan-Mao area in Western Guangdong]*. Sun Yat-Sen University Press.

R. Shi. 2009. 廣西防城區粵語音系 [the phonology of the fangcheng yue dialect in guangxi]. 百色學院學報 *[Journal of Baise University]*, 22(2):106–116.

J. N. Stanford. 2012. One size fits all? dialectometry in a small clan-based indigenous society. *Language Variation and Change*, 24(2):247–278.

H. W. M. Sung. 2023. Is a typologically, genetically different language similar to european languages? a dialectometrical analysis on yue and pinghua. In *73. Studentischen Tagung Sprachwissenschaft (StuTS), Oral Presentation*, Frankfurt, Germany.

H. W. M. Sung and J. Prokic. 2023. What are guangfu dialects? In *27th International Conference on Yue Dialects*, Ohio State University, Online Presentation.

H. W. M. Sung, J. Prokic, and Y. Chen. forthcoming. Applying the state–of–the–art tonal distance metrics to a large dialectal dataset. In U. Stange-Hundsdörfer S. Wagner, editor, *(Dia)lects in the 21st century: Selected papers from Methods in Dialectology XVII (Mainz, 2022)*. Language Science Press. Forthcoming.

M. Swadesh. 1955. Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics*, 21:121–137.

Y. Tan. 2017. 廣西賓陽縣 (賓州鎮) 本地話音系 [the phonology of binzhouzhen in the binyang county in guangxi]. 梧州學院學報 *[Journal of Wuzhou University]*, 27(5):58–71.

C. Tang. 2009. *Mutual intelligibility of Chinese dialects: an experimental approach*. Ph.D. thesis, Leiden University.

P. Trudgill. 1983. *On dialect: Social and geographical perspectives*. New York University Press.

F. Wang and W. S. Y. Wang. 2004. Basic words and language evolution. *Language and linguistics*, 5(3):643–662.

J.C. Wells. 1982. *Accents of English: Introduction*, volume 1. Cambridge University Press.

S. Wichmann and Q. Ran. 2019. Asjp 模式的漢語方言計算分析——以 65 個漢語方言語檔為例 [a phylogenetic study on 65 chinese doculects: with asjp tools]. 現代語文 *[Modern Chinese]*, (5):4–13.

M.-S. Wu, N.E. Schweikhard, T.A. Bodt, N.W. Hill, and J.-M. List. 2020. Computer-assisted language comparison: State of the art. *Journal of Open Humanities Data*, 6(1):2.

J. Xie. 2007. 廣西漢語方言研究 *[Studies on the Chinese dialects in Guangxi]*. People's Publishing House of Guangxi.

C. Yang and A. Castro. 2008. Representing tone in levenshtein distance. *International Journal of Humanities and Arts Computing*, 2(1-2):205–219.

S. Yang. 2013. 廣西藤縣濛江方言音系 [the phonology of the tengxian mengjiang dialect in guangxi]. 方言 *[Dialect]*, (1):71–85.

M. Yip. 2002. *Tone*. Cambridge University Press.

B. Zhan and Y. Cheung. 1987. *A Survey of Dialects in the Pearl River Delta, Vol. 1, Comparative Morpheme-Syllabary*. People's Publishing House of Guangdung.

B. Zhan and Y. Cheung. 1990. *A Survey of Dialects in the Pearl River Delta, Vol. 3, A Synthetic View*. People's Publishing House of Guangdung.

B. Zhan and Y. Cheung. 1994. *A Survey of Yue Dialects in North Guangdong*. Jinan University Press.

B. Zhan and Y. Cheung. 1998. *A Survey of Yue Dialects in West Guangdong*. Jinan University Press.

Z. Zhong. 2015. 廣西蒼梧本地話音系 [the phonology of cangwu local vernacular in guangxi]. 方言 *[Dialect]*, (2):177–192.

中國語言資源保護研究中心 [Research Centre of Linguistic Resource Reservation in China]. 2022. 中國語言資源保護工程採錄展示平台 [platform of linguistic resource reservation].

Supplementary Material

The datasets and the tone conversion scripts can be found in https://osf.io/m9g2a/.

Appendix A: List of Items in the Data

Chinese	English	Chinese	English
一	one	二	two
三	three	四	four
五	five	六	six
七	seven	八	eight
九	nine	十	ten
我	I	你	you
全	all	多	many
大	big	長	long
細	small_col	小	small_lit
男	man	女	woman
人	person	魚	fish
鳥	bird_lit	雀	bird_col
狗	dog	虱	lice
樹	tree	葉	leaf
根	root	皮	skin1
膚	skin2	肉	meat
血	blood	骨	bone
脂	fat	角	horn
尾	tail	羽	feather
髮	hair_head	毛	hair_body
頭	head	耳	ear
眼	eye	鼻	nose
口	mouth	牙	tooth1
齒	tooth2	爪	claws
腳	leg	膝	knee
手	hand	肚	abdomen
胸	breast	心	heart
肝	liver	飲	to drink
食	to eat	咬	to bite
看	to see_lit	知	to know
睡	to sleep	死	to die
殺	to kill	游	to swim
飛	to fly	走	to walk
來	to sit	企	to stand
講	to speak_col	日	sun
月	moon	水	water
雨	rain	石	stone
沙	sand	土	soil/earth
地	floor/ground	雲	cloud
煙	smoke	火	fire
灰	ash	燒	to burn
路	road	山	mountain
紅	red	綠	green
黃	yellow	藍	blue
白	white	黑	black
夜	night	熱	hot
凍	cold	滿	full
新	new	好	good
圓	round	乾	dry
史	history	蛇	snake
虎	tiger	鼠	mouse/rat
馬	horse	牛	cow
船	boat	春	Spring
夏	Summer	秋	Autumn
冬	Winter	西	West
北	North	出	out
入	enter	墳	tomb
想	to think	雙	double
見	to see_col	雞	chicken
豬	pig	湖	lake
合	together/to merge	村	village
愛	love	鴨	duck
奇	strange	具	tool
花	flower	光	light
師	teacher	去	to go

Appendix B: Tone Correspondences

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)