Universiteit
Leiden
The Netherlands

Advancing explanatory and tonal dialectometry

Sung, H.W.M.

Citation
Sung, H. W. M. (2026, February 13). *Advancing explanatory and tonal dialectometry*. *LOT dissertation series*. LOT, Amsterdam. Retrieved from https://hdl.handle.net/1887/4291801

Version:	Publisher's Version
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/4291801

Note: To cite this publication please use the final published version (if applicable).

CHAPTER 1

Introduction

1.1 Definition of 'Dialect'

The subjects of study in Chinese dialectology are the 'Chinese dialects' (You 2016:27). The Chinese term, '方言' (fong1yin4 in Cantonese, or fāngyán in Mandarin), is often translated as 'dialect' in English, although literally the term means 'regional speech' (Tang 2018). The translation of '方言' as 'dialect' is controversial (Mair 1991). Mair argues that the criterion often used in Western dialectology, where 'dialects' are mutually intelligible (opposed to 'language', a "collection of mutually intelligible dialects", Chambers and Trudgill (1998:3)[1]), is not satisfied for the so-called Chinese 'dialects'. Another common criteria, that different Chinese 'dialects' share the same writing language, also does not hold according to Mair (1991). Possible alternatives to address Chinese 'dialects' include Sinitic languages, *regionalect*[2] (DeFrancis 1986) or *topolect* (Mair 1991). In addition, the term 'dialect' in Chinese di-

[1]It should be noted that it is not always the case in Western dialectology that 'dialects' are mutually intelligible. Chambers and Trudgill (1998:4) used German as an example.

[2]Not to be confused with *regiolect* (Hoppenbrouwers 1983; Kehrein 2020), which is an intermediate variety (or rather a continuum of varieties) between the base dialect and the standard language.

alectology can also be used to refer to an entire Sinitic branch (e.g. Yue dialect), the varieties within a region (e.g. Teochew) or the speech variety of one location (e.g. Guangzhou dialect) (Kurpaska 2010).

In this dissertation, I do not use the term 'dialect' in the same way as in Chinese dialectology, i.e. referring to a branch of Sinitic (which includes a lot of sub-varieties). This is because in the following chapters, I focus on the (micro-)variation within a branch of Sinitic, and not the differences between major branches of Sinitic (with only a handful of representative dialects). A *dialect*, in this thesis, refers to the variety of a locality in which we have data for, i.e. a locality in dialect survey. This is based on the view of the linguistic geographers that each location has its own dialect (Wiesinger 1983).[3] A group of dialects which consists of more than one dialect is referred to as a *dialect group*. The usage of 'dialect' here is somewhat similar to concepts like *patois* in the French tradition (Martinet 1954), *traditional dialect* in the British English tradition (Wells 1982; Trudgill 1999; Hughes et al. 2013) and *base dialect*, *Dialekt* in the German tradition (Bellmann 1998; Wiesinger 1983). The term 'dialect' is also interchangeable with 'variety' throughout the dissertation.

1.2 Background

Yue and Pinghua are two branches of the Sinitic language family spoken in Southern China, namely in the Guangdong and Guangxi provinces (Chinese Academy of Social Sciences (CASS) 2012). In Chinese dialectology, dialectologists have shown interests in classifying the sub-dialects of Yue since the 1960s. However, many classifications often focus on the Yue dialects in Guangdong only, while others only in Guangxi. In addition, the classification criteria are often unknown. One of the exceptions, which is also the more authoritative classification, is from the *Language Atlas of China (2nd edition)* (hereafter LAC) (Chinese Academy of Social Sciences (CASS) 2012). This classification covers the entire Yue-speaking region, and it provides the criteria for the classification. However, the motivations for the choices of features are unknown, and whether these features are really characteristic for each dialect group is also in question. Despite having a classification for the Yue-speaking re-

[3]Original text: "Nach sprachgeographischer Anschauung besitzt daher jeder Ort seinen eigenen Dialekt." (Wiesinger 1983:807)

gion, the unanswered questions regarding the LAC classification beg for further studies to enhance our understanding of the internal structure of Yue. On the other hand, Pinghua is an interesting case because since its first proposal as a separate branch of Sinitic in the 1980s, its status has been controversial. Scholars have been debating for decades whether Pinghua belongs to Yue or not, and there seems to be no common ground up to this day. In the LAC classification, Pinghua is discussed in a separate chapter, with its own classification, without much consideration of Yue. In Chapter 2, an overview of the existing classifications of Yue and Pinghua dialects are provided, followed by a discussion of the problems in these classifications. This chapter identifies several research gaps in Yue dialectology, and these gaps invite the use of computational methods to bring new perspectives to the field.

Dialectometry is a branch of quantitative dialectology which makes use of computational and statistical methods. A central goal in dialectometry is to classify dialects through considering a substantial amount of data, in order to reduce the selection biases from picking a handful of features as classification criteria. One of the most important techniques in dialectometry is the measurement of dialect distances (the term 'dialectometry' means the measurement of dialect [similarities or distances], Séguy 1973b). By obtaining dialect distances (through a range of methods, depending on the data type), one can then use techniques from machine learning and numerous visualisation techniques to explore the underlying structure and relationships between dialects. Chapter 3 provides a brief history of the field, followed by an introduction to the methods commonly used in dialectometry, including those that are used throughout the thesis. These methods, as shown in Chapter 5, can bring new insights to Yue and Pinghua dialectology. First, there is a comparison between the LAC classification and the dialectometric classification, followed by a new classification based on the quantitative analysis of the dialect survey data. Lastly, some extralinguistic correlates are explored based on the aggregate analysis.

In terms of the data, Yue and Pinghua dialects have a relatively good coverage of dialect localities compared to some other branches of Sinitic, such as Hakka. Towards the end of the 20th century and the beginning of the 21st century, numerous dialect surveys were conducted, e.g. the *Survey of Dialects of the Pearl River Delta* (Zhan and Cheung 1987) and *Chinese Dialect Research in the Guangxi Province* (Xie 2007). A more detailed description of the data used in the current thesis can be found

in Chapter 4.

1.3 Motivations of the dissertation

In the last decade, the field of dialectometry has rapidly grown in terms of its methods, research directions and popularity. However, most studies are still largely focused on European languages, and not so much on languages in other parts of the world. Dialectometric methods allow us to extract more abstract (non-language-specific) patterns of linguistic variation from one area, which then allows cross-linguistic comparisons. It is of great interest for dialectologists to see whether non-European languages with different historical, typological and sociolinguistic backgrounds share similar variation patterns in Europe.

Tonal languages[4], on the contrary, have not been extensively explored under the view of an aggregate analysis. Furthermore, it is unclear how tone distances should be measured for the purpose of dialectometry (and how adequate the existing methods are). This reflected in the lack of consensus of a unified tone distance calculation approach (e.g. Yang and Castro (2008) and Stanford (2012) would use different tone calculation methods for the same type of tone notations). Without (finding) an adequate method, it is not possible to investigate patterns in tonal variation using dialectometric methods that have been developed in the past few decades. We have reached a bottleneck for quantitative tonal dialectology.

On a different note, mainstream dialectometric approaches (see Section 3.2.2) convert qualitative dialect data to linguistic distances, which has made a huge advance in the understanding of dialectal variation. However, these approaches suffer from a lack of explainability, due to a loss of information during the conversion process. Although there are several attempts to address the problem (e.g. Wieling and Nerbonne 2011; Prokić et al. 2012; Pickl 2016), these methods have not been evaluated systematically. Feature extraction is an important and unmissable step in quantitative dialectology, as it provides indications for what exactly makes one dialect group distinctive from another, which completes the procedures of what dialect classification is really about.

[4] 'Tonal languages' in this thesis do not include accentual or pitch accent languages (Yip 2002:4). See Chapter 7.

The issues raised above provide the motivations for the current dissertation. There are two main directions in the current thesis: 1) I aim to extend the scope of dialectometry to tonal languages by assessing and proposing a new way to measure tone distances and 2) I aim to enhance the explainability of dialectometric methods by proposing a novel approach which can extract features exclusive to a particular dialect cluster. Enhancing and further developing these two areas will allow me to discover both general and language-specific dialectological patterns based on a case study of Yue dialects. The findings in this thesis will hopefully fill in knowledge gaps which lie within and outside European dialectological traditions.

1.4 Research questions

The main goals of the current thesis are to apply and expand current dialectometric techniques to Yue and Pinghua dialects in order to address a number of general and language-specific dialectological questions. A number of issues are addressed throughout. First of all, the segmental classification of Yue and Pinghua is explored, which sets the stage for the following chapters. Moreover, to detect the segmental characteristics of the Yue dialects, the segmental analysis is indispensable. In terms of the dialectometric analysis of tonal variation, Yue provides a laboratory for such an exploration, and segmental variation becomes relevant again in the comparison of the two linguistic levels. Perhaps one would expect tonal variation to show similar patterns as segmental variation, since both segments and tones are part of the phonetics components in word production. The comparison between the two linguistic levels will reveal to what extent this assumption is correct.

To explore the segmental variation of Yue-Pinghua, existing methods in dialectometry (Levenshtein distance) are utilised. Chapter 5 begins with the Yue-Pinghua dichotomy, which is highly relevant for the background in Yue-Pinghua studies. Pinghua is classified as a sister branch of Sinitic to Yue, but its existence has caused a huge debate since its first proposal. It is important to assess the status of Pinghua, as the inclusion of highly different varieties (which are not part of a continuum) can skew the data, damping some major patterns for the dialects we are interested in (in this case, dialects in the Yue continuum).

When Yue dialects are defined using data-driven, quantitative ap-

proaches from dialectometry, a classification on the segmental level can be performed. These approaches involve calculating dialect distances with Levenshtein distance, followed by analysing the distances using cluster analysis and multidimensional scaling. First of all, the results of the segmental classification can tell us what the dialect landscape of Yue looks like on the aggregate level. In addition, extralinguistic correlates can be identified and they often reflect historical events (such as migration) and other factors which shaped the dialect landscape we see in the classification, as well as the quantitative maps.

The following questions regarding segmental variation of Yue are raised in Chapter 5:

(1) Is Pinghua part of a continuum with Yue?

(2) How does Yue vary on the segmental level?

(3) Can geographical variation in Yue be explained by certain extralinguistic correlates?

The methodology in Chapter 5 is based on the use of Levenshtein distance. However, this distance metric does not return the characteristic (segmental) features that are present in each dialect cluster. To understand the characteristics of each dialect group identified in a segmental dialectometric analysis, we need alternative methods. One of these methods include *normalised pointwise mutual information* (nPMI). In Chapter 6, nPMI is used as a feature extraction technique for the Yue dialects. This chapter involves two language-specific questions:

(4) What are the characteristic features associated with each dialect group of Yue?

(5) To what extent does the traditional classification capture these characteristic features?

Variation of tonal languages has not been well-studied through dialectometric methods. It is also unclear how adequate existing methods can be applied to a bigger dialect dataset for the purpose of dialect

classification. In Chapter 7, the main goal is to assess the state-of-the-art methods in calculating tone distances. The chapter starts with an overview of the tone representations and the existing methods in quantifying tone distances. Next, the existing tone distance calculation methods are compared using a number of criteria, including assessing the cohesion between the calculated tone distances and the perceptual tonal distance. The comparison serves as the foundation for Chapter 8, which seeks a tone distance calculation method for refinement, in order to perform a dialectometric analysis on the tonal variation in Yue.

The research questions relevant to the comparison of existing tone distance calculation methods are:

(6) Which tone representation should be used for the computation of tone distances?

(7) Which existing tone distance calculation method(s) is/are suitable for the purpose of dialect classification?

The final content chapter of this dissertation focuses on the application of *dialect tonometry* (calculating the aggregate tone distances for the purpose of dialectometry) on the Yue dialects. This chapter aims to uncover the patterns in tonal variation in space, for the first time, under the aggregate perspective on over 100 dialects. Using a new tone representation (modified Onset-Contour-Offset), aggregate distances are calculated, and the tonal variation of Yue dialects is explored using common techniques in dialectometry. In addition, a comparison with the segmental analysis in Chapter 5 is performed in order to observe the similarities and differences between these two linguistic levels, which are often put together under phonetic variation.

In Chapter 8, I will explore the tonal variation on the aggregate level by answering the following questions:

(8) Do dialects form a geographical continuum on the tonal level?

(9) To what extent does tonal variation correlate with segmental variation?

Addressing the research questions above can enhance our understanding of general and language-specific (Yue) aspects of language variation. In addition, the current dissertation also proposes new methodologies in dialectometry, which can be applied to other languages in the world.

The structure of the dissertation is as follows: Chapter 2 provides an overview of Yue and Pinghua classifications and a discussion of their problems; Chapter 3 gives a brief history of the dialectometry, followed by an introduction to the methods commonly used in the field. Chapter 4 gives a description of the data used in this thesis, and Chapter 5 provides the segmental analysis of Yue-Pinghua. Chapter 6 extends from the previous chapter, by exploring the characteristic features of different Yue dialect groups using automatic feature extraction. Chapter 7 and 8 concerns tonal dialectometry. The former chapter introduces what a tonal language is; provides an overview of existing tone representations and lastly, compares the existing tone distance calculation methods in order to assess to what extent existing methods are fit for a dialectometric analysis on tones. The latter chapter builds up on one of the tone distance calculation methods introduced in the previous chapter, and the tonal variation of Yue is explored using the modified OCO representation of tones. Moreover, segmental (from Chapter 5) and tonal distances are compared using various statistical techniques. Lastly, the thesis ends with a conclusion.