Advancing explanatory and tonal dialectometry

Sung, H.W.M.

CHAPTER 9

Conclusion

The two main goals of this thesis are 1) to develop a methodology for automatic feature detection in order to increase the explainability of a dialect classification and 2) to develop and refine a tone distance calculation method for performing dialectometry on tonal languages (*dialect tonometry*). These methods have been applied to a dataset of Yue and Pinghua dialects in order to address a number of general and language-specific questions in dialectology.

The dataset which the thesis is based on comes from several dialect surveys and individual studies of the dialects in the Yue-Pinghua region, in Southern China. This dataset consists of around 130 monosyllabic words in 113 dialects.

After giving the background on the issues in Yue dialect classifications and an introduction to dialectometric methods, in Chapter 5, Levenshtein distance was applied to the segmental part of the dataset, and it was compared with the traditional classification, namely the scheme from the *Language Atlas of China*. Through the dialectometric analysis, it has been found that not all traditional Pinghua varieties are part of a continuum with Yue, as some of the traditional dialectologists argued, nor are they completely separate from Yue. Under the aggregate view, Northern (Guibei) Pinghua dialects appear to be quite different, and isolated from the rest of the Yue, as well as Southern

(Guinan) Pinghua dialects. The pattern we see suggests that it is likely that Northern Pinghua does not belong the Yue continuum. Focusing on the Yue continuum only, the segmental variation has shown many parallels with the patterns found in previous dialectometric studies in other languages. For instance, a geographical continuum pattern is observed, rather than finding distinct dialect areas with abrupt boundaries. This is illustrated by the presence of transition dialects between the Siyi and Guangfu dialects, for example. In addition, variation of Yue can be explained by several correlates. For instance, rivers have shown to be an important medium for dialect transplantation. This is reflected by the high degree of dialect similarities (low distances) between the dialects connected by rivers (e.g. the traditional Yongxun dialects with Guangfu dialects). Changes in political borders also seem to be reflected in the coastal dialects. It has been proposed that dialectal variation is closely linked to 'density of communication' (Bloomfield 1933). This can also be observed in the Yue dialects near the Guangdong-Guangxi border.

Whilst Levenshtein distance can convert qualitative data into dialect distances and these distances can be visualised in multiple ways, the conversion process causes information loss, meaning we do not know which features are characteristic for the dialect groups we see on the classification maps. In Chapter 6, I have made used of normalised pointwise mutual information (nPMI) to extract features which are closely associated with each of the five dialect groups of Yue. This method has proven useful for identifying important features of different Yue dialect groups, as it revealed exclusive features that were not considered in the traditional classification in the LAC. An additional finding is that not all dialect groups have equally exclusive and representative features. Perhaps one further usage of the nPMI scores is being an indicator for whether we should further divide the dialect group into smaller groups, which seems to apply for the case of Goulou dialects. This is an area for future studies.

Tonal languages have not been analysed a lot with dialectometric methods. Although there have been a handful of studies, the focuses of these studies are not dialect classification, and the datasets used tend to be relatively small, around 30 dialects or fewer. Additionally, there are a number of different methods in calculating tone distances, though no systematic comparison has been done in evaluating these methods, other than Tang (2009) (it should be noted that Tang's data consists of Sinitic languages instead of varieties within one Sinitic branch, which

is arguably closer to comparing related languages rather than dialects within a continuum). Chapter 7 started by reviewing the notations and representations of tones, and concluded that Chao's (1930) tone letters are the most suitable representation for cross-dialectal comparison of tones. Next, four tone distance calculation methods have been compared based on five criteria (tone overlap, perceptual dimensions, local incoherence, ARI score with the traditional classification and the ARI score with the segmental classification). Onset-Contour-Offset (OCO) and the Gandour-Harshman-Tang tone distance measurement (GH-T) turn out to be the more linguistic coherent methods for measuring tone distances. OCO is slightly preferred to GH-T because it can be easily combined with segments before calculating the Levenshtein distance. However, OCO suffers from not being able to distinguish all the tones in the Yue dataset. This implies that OCO still requires further refinement to be used for dialectometry. One additional observation being made during the comparison is that no matter which tone distance calculation method is used, the local incoherence is still quite high. This implies that tonal variation might behave differently from segmental variation, resulting a much higher local incoherence value.

Chapter 8 begins with the introduction of the mOCO representation, a modified version of OCO, and mOCO is able to distinguish 72 out of 73 tones in the dataset. Additionally, the first two MDS dimensions extracted from the tone distances calculated with mOCO correlate strongly to the perceptual dimensions identified by Gandour and Harshman (1978). mOCO is then applied to the Yue data for dialect classification, and there are a number of discoveries. First of all, as shown in the MDS map (Figure 8.4), we can find abrupt boundaries between dialect areas, very much similar to a cluster map (Figure 8.5, from cluster analysis). This pattern is different from what we have seen in the segmental variation. Furthermore, a comparison is made between tonal and segmental variation. The distance matrices from the segmental analysis (Levenshtein distance) and tonal analysis (mOCO) are correlated using the Mantel test. The results have shown that both matrices show only mediocre correlations. By further correlating the first three dimensions extracted through multidimensional scaling from the two matrices, it was found that not all dialect groups are correlated with one another; only Siyi and Western dialects (dimensions) show strong correlations. The reason for the lack of correlation between other dialect groups is unclear at this moment. One major difference between the segmental

MDS map and the tonal MDS map is that the westward diffusion of the Guangfu/Inland dialect group is not as clear in the tonal map as in the segmental map. It is currently unclear whether this is due to the differences between the rate of change for tones versus segments during the transplantation and formation of the western Guangfu dialects. Unfortunately, the current dialectal data cannot offer any answers to this question, especially when dialect tonometry is still at its infancy.

This dissertation has provided a case study for two new methods in dialectometry, one on feature extraction and one on dialect tonometry. Yue has served as a laboratory for testing dialectological theories and tools that are developed to help our understanding for language-specific topics. There are further potentials and further refinements needed for these methodologies.

Multiple sequence alignment has played an important role in feature extraction, as it allows us to retrieve precise linguistic variants that are highly associated with a particular dialect group. In addition to feature extraction, MSA also allows other kinds of analysis beyond the scope of the current thesis, such as the identification of dialect changes (Sung and Prokić submitted). One further usage of dialectometry is to determine the direction of dialect change in relation to the Standard variety (Heeringa and Hinskens 2015; Buurke et al. 2022). In order to observe the changes across two generations of speakers, one can compare the linguistic forms from older and younger speakers against the Standard variety by means of three-dimensional distance calculations. These studies, however, also suffer from the loss of qualitative information during the calculations of linguistic distances (Sung and Prokić submitted). By multi-aligning the linguistic forms of the Standard variety, the dialect spoken by the older generation and the dialect (of the same location as the older speakers) spoken by the younger generation, it is also possible to identify the direction of change (e.g. convergence to the Standard, divergence from the Standard, no change). Furthermore, the multi-aligned data also allows us to extract the precise sound changes of a certain type automatically, which potentially allows us to investigate the social and attitudinal factors of these changes.

For feature extraction using nPMI, the method has proven useful in German (Sung and Prokić 2024a) and Dutch (Sung and Prokić 2025). Furthermore, nPMI is not limited to the extraction of phonetic features only, it is also possible for morpho-syntactic data (e.g. Sung and Prokić 2024c) and dialect dictionary entries (e.g. Sung and De Tier submit-

ted), given that the input data is categorical. There can be further uses to the features extracted using nPMI. For instance, as we have seen in Chapter 6, shared innovations can be found through this method. Sung and Prokić (forthcoming) have made use of nPMI to find shared innovations in Yue and based on these results, they reconstructed the relative chronology of several sound change of an under-studied dialect group of Yue. On the other hand, Sung and Prokić (submitted) adapted Séguy's (1973a) idea from the gradient map of Gasconity (see Section 3.2.2), and used the top characteristic features to measure 'dialect typicality' and identify the core areas in different German dialect groups.

Although feature extraction has been successfully applied to dialect survey data, it should be noted that the items collected in these surveys are highly selective, elicited items. Some scholars criticised that dialect survey (or linguistic atlas) data are not naturalistic enough since the data were collected through elicitation (Szmrecsanyi 2013:3-4). This led to dialectologists turning to using corpora for dialectometry as an alternative. Corpus data are more naturalistic, and they reveal more about the "context and magnitude in which linguistic features are used" (Kuparinen and Scherrer 2024). Future studies on automatic feature extraction should aim for methods which can also process other types of dialect data, including dialect corpora or even recordings of dialect speech.

In terms of dialect tonometry, mOCO is only a stepping stone towards having a more adequate tone distance metric for looking at tonal variation. It is by no means a perfect method and it still requires some improvements. For instance, there are still a number of concave and convex tones which need to be differentiated in the mOCO representation. mOCO can serve as an intermediate method towards a better understanding of tonal variation, starting from Southeast Asia. However, this representation might pose challenges in measuring tone distances in tone languages spoken in other parts of the world, such as African and South American tone languages, as well as languages with pitch accent variation, like Japanese. This is because in many of these languages, they use an opaque notation system (see Section 7.3) to represent the tones. This means that there are no contours available to be converted into mOCO representation from the sources. One then needs to first prepare the transcriptions of the tone contours in the data in these languages before applying the mOCO distance calculation method. However, not all tonal phenomena are suitable to be represented using Chao's (1930) tone

letters. An alternative method, which perhaps can be applied to more languages, is to develop a method to measure tone distances directly from the F0 contours. However, methods for F0 measurements will not be applicable for dialects which are already extinct (which is the case for many traditional or base dialects in the world as a consequence of increase mobility, technological advances, increase access to education and negative attitudes towards dialects). mOCO, then, is still a valuable addition to the dialectometric toolkit, as it can easily be implemented to the dialect survey transcription data which were collected in the 20th century (before recordings could be made).

Yue has provided a massive amount of insights to dialectometry, both in terms of linguistic patterns and for methodological developments. Only with a typologically diverse set of languages can we discover the methodological gaps and new insights in language variation. Future studies can aim to explore dialects spoken in more unexplored territories of the world.