Advancing explanatory and tonal dialectometry

Sung, H.W.M.

Summary

Dialectometry is a quantitative branch of dialectology, which makes use of computational and statistical methods on dialect data in order to understand language variation in space. The current dissertation presents how dialectometry can deepen our understanding of the variation of Yue dialects spoken in Southern China, as well as how Yue can help us broaden the scope of computational methods used in dialectometry, in order to account for tonal languages which are common in the world, but not so common as a subject of study within dialectometry.

A number of research questions are addressed in this dissertation, and they fall under the following themes: 1) segmental classification of Yue dialects, 2) identification of characteristic features of Yue dialects, 3) tonal classification of Yue dialects and 4) comparison between segmental and tonal variation of Yue dialects.

The dataset used in the current dissertation consists of the IPA transcription of around 130 words in 113 Yue-Pinghua dialects. For the segmental classification, Levenshtein distance was used to calculate the phonetic distances, followed by multidimensional scaling as a dimensionality reduction technique and cluster analysis to explore the internal structure of the Yue-Pinghua dialect landscape. Traditional Northern Pinghua has been found to be outside the Yue continuum, but that does not apply to traditional Southern Pinghua. A deeper analysis was then proceeded with the remaining 104 dialects (with Northern Pinghua removed as outliers). Yue dialects are found to lie in a big continuum on the segmental level, and the dialects can be divided into 2 to 5 big groups, depending on the level of detail one seeks for.

A dialectometric classification often receives criticisms for the lack of

details or explanations of the identified dialect groups. This is because classifications were based on distances, and it is difficult to retrieve the qualitative information (dialect features) after the quantification into distances. For this reason, multiple sequence alignment (MSA hereafter) was employed before calculating the dialect distances for the segmental classification of Yue. MSA breaks the phonetic transcriptions down to historically related segments, and this was done to all the dialects simultaneously. The transformation of the transcription data makes sound segments 1) historically more accurately aligned and 2) more suitable to be analysed with post-hoc analyses, such as automatic feature extraction. Multi-aligned data can be easily integrated to the usual dialectometric workflow: dialect distances can be calculated with the MSA data, followed by analyses like cluster analysis and multidimensional scaling. Using normalised Pointwise Mutual Information, an association measure commonly used in natural language processing, characteristic features closely associated to each dialect group (identified with the cluster analysis) can be identified. This technique has increased the explanatory component of dialectometry, which goes beyond a mere dialect classification.

On the other hand, tone languages, despite being quite common within the world's languages, are not studied a lot using dialectometric techniques. One problem is that it is not immediately clear how tone distances can be measured. There have been several approaches, from simply binary similarities to sophisticated perception-based distance calculation methods. The current dissertation applies Levenshtein distance on a representation of tones, modified Onset-Contour-Offset (mOCO hereafter), to obtain dialect distances and to explore how tones vary between different dialects in space. mOCO is a more suitable tone representation for dialectometry comparing to others since it can differentiate 72 of the 73 tones attested in the dataset, with gradual distances from one tone to another, which matches human perception. When applying mOCO to a dialectometric analysis of Yue dialects on the tonal level, the pattern of variation differs from segments. While segmental variation shows a continuum pattern, tonal variation shows a more categorical, dialect area pattern. A further analysis has shown that not all the dialect areas show the same pattern on the segmental and tonal levels. The discrepancies shown in other dialect groups are intriguing and pose further research questions which await for further research.