Applying the state-of-the-art tonal distance metrics to large dialectal data

Matthew Sung, Jelena Prokic & Yiya Chen
Leiden University

From Seguy's (1971, 1973) early dialectometric studies to the application of Levenshtein distance in dialectometry (e.g. Heeringa 2004) nowadays, the calculation of phonetic distances between dialects has largely been focused on segments. Despite the fact that tonal languages make up to 70% of the languages in the world (Yip 2002: 1), tones are still largely neglected or simplified in comparative dialectological studies. For instance, Stanford (2012) treats any two tones as either the same or different, there is no in between distances.

In the current literature, there are two studies that propose new tone distance measures, namely Yang & Castro (2008) and Tang (2009). Yang & Castro (2008) has found that their Onset-Contour(-Offset)/OC(O) representation correlates best with mutual intelligibility based on applying Levenshtein distance on a range of tonal representations, including Chao's (1930) tone letters, tones as autosegments (e.g. Duanmu 1994), tones as approximated pitch targets (Xu & Wang 2001). Tang (2009) on the other hand tested several approaches on 15 Chinese dialects, including an inventory-based comparison, Levenshtein distance on Chao's (1930) tone letters and Yang & Castro's (2008) OCO tone representation. In addition, she compared Cheng's (1991) published distances based on 17 dialects (see Cheng 1991, Tang 2009 for more). By counting the misclassifications in the split between Mandarin and non-Mandarin dialects, she found that Cheng's distances work the best, with only 1 misclassification. However, she also noted that this method "fails to reflect any of the internal taxonomy" after distinguishing the Mandarin and non-Mandarin dialects (Tang 2009: 137). Thus far, no existing study measures tone distances for the purpose of dialect classification.

In this presentation we examine the existing tone distance measures and apply them on a large, newly compiled, dialect dataset. The data that we will be using comes from Zhan & Cheung (1987), Zhan & Cheung (1994), Zhan & Cheung (1998), Shao (2016), Chen & Lin (2009a), Chen & Lin (2009b) and Xie (2007), and it consists of 123 Yue and Pinghua dialects represented with over 120 words each. Our results show that the current state-of-the-art method proposed by Yang & Castro can distinguish less than 50% of the tones in our data, which makes it unsuitable for classifying dialects at a lower level. In addition, we will also show how other representations (binary comparison and Levenshtein distance on Chao's (1930) tone letters) perform. Lastly, we will compare for the first time segmental dialect classification and tonal dialect classification and see how much similarity the two linguistic levels share.

References:

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

Chen, H. & Lin, Y. (2009a). *Yue yu ping hua tu hua fang yin zi hui di 1 pian: Guangxi yue yu, Guinan ping hua bu fen 粵語平話土話方音字彙第 1 編: 廣西粵語、桂南平話部分 [The Lexicon of Yue, Pinghua and Tuhua Volume 1: Guangxi Yue and Guinan Pinghua]*. Shanghai: Shanghai Education Publishing.

Chen, H. & Lin, Y. (2009b). *Yue yu ping hua tu hua fang yin zi hui di 2 pian: Guangxi yue yu, Guinan ping hua bu fen 粵語平話土話方音字彙第 2 編: 桂北、桂東及周邊平話、土話部分 [The Lexicon of Yue, Pinghua and Tuhua Volume 1: Guangxi Yue and Guinan Pinghua]*. Shanghai: Shanghai Education Publishing.

Cheng, C. C. (1991). “Quantifying affinity among Chinese dialects”. *Journal of Chinese Linguistics Monograph Series*, (3), 76-110.

Duanmu, S. (1994). “Against contour tone units”. *Linguistic Inquiry*, 25(4), 555–608.

Heeringa, W. (2004). *Measuring dialect pronunciation using Levenshtein distance*. Groningen: University Library Groningen. [PhD dissertation, University of Groningen.]

Séguy J. 1971. La relation entre la distance spatiale et la distance lexicale. *Revue de linguistique romane* 35, 335–57.

Séguy J. 1973. La dialectométrie dans l’atlas linguistique de la Gascogne. *Revue de linguistique romane* 37, 1–24.

Shao, H. (2016). *Yue xi zhan mao di qu yue yu yu yin yan jiu 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]*. Guangzhou: Sun Yat-Sen University Press.

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

Tang, C. (2009). *Mutual intelligibility of Chinese dialects: an experimental approach*. Netherlands Graduate School of Linguistics. [PhD dissertation, Leiden University.]

Xie, J. (2007). *Guangxi han yu fang yan yan jiu 廣西漢語方言研究 [Studies on the Chinese dialects in Guangxi]*. Guangxi People’s Publishing.

Xu, Y. and Wang, Q. E. (2001). “Pitch targets and their realization: Evidence from Chinese”. *Speech Communication*, 33, 319–337.

Yang, C., & Castro, A. (2008). “Representing tone in Levenshtein distance”. *International Journal of Humanities and Arts Computing*, 2(1-2), 205-219.

Yip, M. (2002). Tone. Cambridge University Press.

Zhan, B. & Cheung, Y. (1987). *A Survey of Dialects in the Pearl River Delta, Vol. 1, Comparative Morpheme-Syllabary*. People’s Publishing House of Guangdong.

Zhan, B. & Cheung, Y. (1994). *A Survey of Yue Dialects in North Guangdong*. Jinan University Press.

Zhan, B. & Cheung, Y. (1998). *A Survey of Yue Dialects in West Guangdong*. Jinan University Press.