Tonal aspects of Cantonese-English code-switching in HLVC corpus

Katrina Kechun Li, Christopher Bryant, Li Nguyen
University of Cambridge
kl502@cam.ac.uk, cjb255@cam.ac.uk, nhbn2@cam.ac.uk

This paper investigates the possible phonological effects of lexical tone on predicting code-switching patterns between English and Cantonese in a large corpus using semi-automatic natural language processing (NLP) method. Data comes from the Heritage Language Variation and Change in Toronto Project (HLVC, Nagy 2011), which contains native speakers from Hong Kong as well as second-generation heritage speakers from Toronto, Canada. Previous studies have suggested that certain tones in a tonal language are more conducive to code-switching to a non-tonal language; e.g. code switching from Vietnamese to English is facilitated more by high tones than low tones (Tuc 1997), and switching from Mandarin to English occurs more frequently following falling and neutral tones (Zheng 1997). These studies, however, were small in scale and did not consider other confounding factors such as tone frequencies and syntactic structure.

In this work, we address this gap by focusing on Cantonese-English – a new language pair that has never been examined. We processed the data in line with Nguyen and Bryant's study (2020) on Vietnamese-English to facilitate cross-linguistic comparison. Specifically, we sent the largest contiguous sequences of Cantonese or English text in a sentence (based on character encodings) to the relevant tokenizer and part-of-speech (POS) tagger. We used PyCantonese (Lee 2015) and spaCy (Honnibal & Montani 2017) to process Cantonese and English respectively. Each token was assigned a language label: Cantonese, English or Language-neutral, where language neutral terms include fillers such as 'yeah' and 'umm', as well as sentence-final particles such as '啊/ah' and '啦/lah'. The tonal information at the switch point was then extracted.

After preprocessing, the preliminary results show that 7.09% (N = 1046/14757) of the sentences contain code-switching between Cantonese and English. In particular, the mid and low level tones (T3, T6) occur most often at Cantonese to English switch points (T3: 30%, N = 315/1957; T6: 23%, N = 443/1957), whereas the falling tone (T4) is least likely to occur (T4: 3%, N = 69/1957). The pattern also holds if we compare the switch point occurrence relative to overall occurrence for each tone. The facilitatory effect of level tones here resonates with the word prosody of Cantonese English words which assigns level tones to English syllables (Gussenhoven 2012). This differs from the facilitatory effect of high/mid tone in Vietnamese which is associated with English stressed/unstressed syllables (Tuc 1997), or that of the falling tone in Mandarin which is similar to the English intonational contour (Zheng 1997). Ultimately, our initial findings confirm tonal effects previously found in other language pairs, but more importantly suggest some patterns peculiar to code-switching between Cantonese and English. We further probe these patterns using regression analysis to account for part-of-speech information and word frequency, as well as the differences between native and heritage Cantonese speakers.

Key References

Nagy, N. (2011). A multilingual corpus to explore variation in language contact situations. *Rassegna Italiana Di Linguistica Applicata, 43*(1–2), 65–84.

Tuc, H.-D. (1997). Tonal facilitation of code-switching. *Australian Review of Applied Linguistics, 20*(2), 129–151.

Zheng, L. (1997). Tonal aspects of code-switching. *Monash University Linguistics Papers, 1*(1), 53–63.

Nguyen, L., & Bryant, C. (2020). CanVEC - the Canberra Vietnamese-English code-switching natural speech corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, 4121–4129.

Lee, J. L. (2015). *PyCantonese: Cantonese linguistic research in the age of big data* [Talk]. The Childhood Bilingualism Research Centre, Chinese University of Hong Kong.

Honnibal, M., & Montani, I. (2017). *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*.

Gussenhoven, C. (2012). Tone and intonation in Cantonese English. In *Proceedings of the Third International Symposium on Tonal Aspects of Languages*, Nanjing, China.