PyCantonese: Developing computational tools for Cantonese linguistics

Jackson L. Lee, Litong Chen, & Tsz-Him Tsui
University of Chicago & The Ohio State University

Introduction: In this talk, we introducce PyCantonese, an open-source Python library for computational research in Cantonese linguistics. There are two primary motivations for this project. First, while an increasing number of Cantonese corpora are available (e.g., the Hong Kong Cantonese Corpus (Luke & Wong 2015), HKCAC (Leung & Law 2001, Fung & Law 2013), the Cantonese Radio Corpus (Francis & Matthews 2005)), these resources are in incompatible formats and there are no general toolkits for handling Cantonese corpus data. Second, computational linguistics is a largely undeveloped sub-field for Cantonese. In response to these gaps, PyCantonese is designed to provide general tools for the manipulation, annotation, and analysis of Cantonese corpus data. We demonstrate the implemented tools including the handling of Jyutping romanization and corpus search functions, and show how PyCantonese can facilitate Cantonese linguistic research.

Handling Jyutping: A common scenario in Cantonese corpus work is that a corpus is available and transcribed in Jyutping, but no tools are readily available to parse Jyutping in order to identify onsets, nuclei, codas, and tones. We demonstrate the relevant functionalities of PyCantonese, and how they facilitate research areas such as phonotactics and phonological development using child-directed speech data.

Search functions: Another frequent task is to search for particular items in corpus data. Depending on the exact nature of the dataset being used, PyCantonese provides search functions for some given Jyutping elements, part-of-speech tags, and Chinese characters. We show how simple searches are performed using PyCantonese, and how to combine these functions and programming techniques to achieve what would be of great interest to linguists (e.g., find verbal and prepositional phrases).

Ongoing work: Because high-quality data in a consistent format are essential in corpus- based research, part of the ongoing work involves the reformatting of corpus data. The datasets thus made available act as training data for various annotation and analysis tools being developed, including word segmentation, conversion between Jyutping and Chinese characters, and part-of-speech tagging.

Conclusions: Computational linguistic research for Cantonese is very much at its infancy. In the current age of big data research, computationally heavy work coupled with large corpus datasets necessitates the availability of suitable tools. As an expanding and evolving project, PyCantonese is envisioned to play the facilitating role that interfaces between the researchers and the large volume of Cantonese data.

References:

Francis, Elaine J. and Stephen Matthews. 2005. A multi-dimensional approach to the category ‘verb’ in Cantonese. Journal of Linguistics 41: 269-305.

Fung, Suk-Yee and Sam-Po Law. 2013. A phonetically annotated corpus of spoken Cantonese: The Hong Kong Cantonese Adult Language Corpus. Newsletter of Chinese Language 92(1): 15.

Leung,Man-Tak and Sam-Po Law. 2001. HKCAC: The Hong Kong Cantonese adult language corpus. International Journal of Corpus Linguistics 6: 305-326

Luke, Kang-Kwong & May Lai-Yin Wong. 2015. The Hong Kong Cantonese Corpus: Design and uses. Journal of Chinese Linguistics.