CBRC, Department of Linguistics and Modern Languages, CUHK

Topic:	PyCan <mark>tonese</mark> : Cantonese linguistic research in the age of big data
Date:	Septe <mark>mber 1</mark> 5, 2015 (Tuesday)
Time:	2:30pm
Venue:	Childhood Bilingualism Research Centre, CUHK

Abstract

The age of big data encourages strongly empirical linguistic research and brings us more linguistic data than we could possibly work on. This is even more true for Cantonese than for other well-studied languages, because datasets and tools for Cantonese (if publicly available at all) are scattered and there has been little effort in systematizing the resources. In response to the lack of a general toolkit for Cantonese big data linguistic research, PyCantonese (http://pycantonese.github.io/) is under active development as a general-purpose, open-source tool for this purpose. PyCantonese is a library that runs in Python, the programming language that is the lingua franca in computational linguistics and natural language processing. In this talk, I introduce PyCantonese and demonstrate its basic usage such as search functions and romanization processing. I also describe ongoing work ranging from part-of-speech tagging to Cantonese monolingual/Cantonese-English bilingual child language development.

Jackson Lee is a PhD candidate at the Department of Linguistics at the University of Chicago. He obtained a BA in Linguistics and French at the University of Hong Kong and an MA in Linguistics at the University of Manchester. Currently, he works on computational linguistics. His research focuses on the modeling of how linguistic structure is learned from unstructured data. His current work revolves mostly around morphology and phonology, with his dissertation focusing on morphological paradigms. Being a Hong Kong native, he also finds time to contribute to Cantonese linguistics, and has worked and published on various topics on Cantonese. Marrying his interests in computational linguistics and Cantonese, he has recently initiated the PyCantonese project.