The 6th Workshop on Innovations in Cantonese Linguistics (WICL-6)
The Ohio State University
May 27 – 28 2022

Learning Cantonese with a Big Data Approach

Andy Chin
The Education University of Hong Kong

Abstract

This presentation introduces a mobile app developed for Cantonese learning with corpus linguistic methods. There is a huge amount of linguistic information that one can access when learning a new language. How to select the most appropriate and useful linguistic data/features/structures is an important issue for both learners and instructors.

Traditional reference materials (such as dictionaries) provide basic information of words, such as number of strokes, pronunciations, word senses, etc. Word usage is mainly presented with a handful of example sentences. There are some questions however traditional reference materials cannot answer:

(1) How many **characters** and **words** does one need to learn?
(2) What are **the most commonly used verbs, nouns, and adjectives**?
(3) What are **the common object nouns** of the verb “eat 食”?

Answers to these questions require quantitative data and linguistic knowledge which can be obtained through corpus data. At the same time, we should bear in mind that the materials designed for the learners should be practical and meet the needs for daily communication.

In 2019, The Education University of Hong Kong received support and funding from the Standing Committee on Language Education and Research (SCOLAR) and the Language Fund to construct a mobile app of Chinese (including Cantonese) self-learning with a big data approach. The design of the app combines corpus technology and linguistic knowledge. The corpus has about two million characters and the data covers a number of genres.

Another feature of the app is the information on **collocation**. Words do not exist alone when used. They will usually combine with other words to form larger units: phrases, sentences and texts. It is thus meaningful to explore how words appear with other words in the language. In this app, the **"verb-object nouns"** and **"classifier-nouns"** collocation pairs were illustrated with corpus examples.