WICL-6 Abstracts
Title: Analyzing Cantonese Corpus Data

A short abstract for the corpus workshop

In this workshop, we introduce how digital technologies can be leveraged to process and study Cantonese corpus data. Several different use cases will be discussed to showcase the convenience and effectiveness of the digital tools in addressing a wide range of research questions. While the workshop is intended for complete beginners, we encourage the participants to take a quick look at the pre-workshop Python tutorial [https://github.com/jacksonllee/pycantonese/blob/main/docs/tutorials/lee-python-2021-april.ipynb].

One-paragraph abstracts for the three demonstrations

- Professor Charles T. K. Lam, The Hang Seng University of Hong Kong
    - Title: Analyzing Cantopop with corpus methods
    - Abstract:
      Cantopop is not only a popular form of entertainment among Cantonese speakers, but also seen as a vehicle of expression of emotion and identity. Lyrics in Cantopop are also interesting in that it mixes Cantonese and Mandarin in the text, despite being performed in Cantonese. In this part, I demonstrate (a) how a small corpus can be built using Python (specifically the PyCantonese package), (b) how one can obtain basic statistics from the lyrics corpus, and (c) how to extract information about the themes in the songs. The challenges in analyzing and interpreting Cantopop through corpus methods will also be discussed.

- Professor Chaak-ming Lau, The Education University of Hong Kong
    - Title: Extracting Cantonese data from Hong Kong Chinese corpora
    - Abstract:
      Hong Kong Chinese data is often a mixture of Modern Standard Chinese (MSC) and Written Cantonese, which is a feature of written data in a diglossic setting, which leads to linguistic inquiries about code choice and code-mixing strategies. I will use crawled data from two newspapers to demonstrate how to programmatically (a) determine whether a piece of writing is MSC, Cantonese, or a mix of the two, (b) classify the sub-genre of a piece of text with Written Cantonese elements, and (c) calculate the distribution of these codes over time and across genres or individual writers.

- Dr. Jackson Lee, Author of PyCantonese
    - Title: Working with Cantonese CHILDES data
    - Abstract:
      Thanks to recent research on language acquisition related to Cantonese, the CHILDES ecosystem has become the largest repository of Cantonese corpora with publicly available data files. The large amount of data necessitates methods that can efficiently handle it. In this demonstration, I show how some of the latest computational tools can be leveraged by replicating published research involving Cantonese CHILDES data with such tools.