

Overcoming Challenges in Cantonese Digital Humanities, and Why It Matters to the World

Cantonese, a major lingua franca in Hong Kong, Southern China, and Southeast Asia, serves as the medium for countless songs and movies over the past century and forms the core of Hongkongers' cultural heritage. Despite its widespread use and cultural significance, Cantonese remains underrepresented in digital humanities research. While the focus on larger, historically prominent languages and resource limitations are partial reasons, a closer examination of recent work in Cantonese NLP reveals that its linguistic properties pose significant challenges at multiple levels.

1. **Text Encoding:** Unicode assumes a stable set of characters. However, for Cantonese (and many Sinitic languages), the lack of standardisation requires users to create new characters on the fly or repurpose obsolete characters to represent speech sounds or words. Variations that often occur in the g2p layer in other languages are found at the morpheme-to-character level. This practice challenges the principles of CJK ideograph inclusions and normative character-based information in character databases.
2. **Text Normalisation:** For a language without an established standard, or when there are competing conventions, there is a need to constantly assess how words are used. Lines of development can change in a matter of months, and different preferences can coexist at any given time.
3. **Language Classification:** Commonly used language taxonomies (e.g., ISO 639-3) assume clear boundaries between languages. The diglossic practice, the “interlinguality” (or “trans-languaging practices”, innovative mixing of languages) and the availability of different register-based markers cannot be captured with a system of fixed language labels. Existing language classification taxonomies (Language Code - Locale - Orthography) do not support this complexity.
4. **The Role of DH in Decentralised Language Development:** Data modelling of the language, done by a small group of practitioners, profoundly impacts how Cantonese is developed. Decisions in transliteration or glyph standards, input methods, visualisation tools, dictionary structure, translation tools, etc. may lead to cascading effects on writing norms and conventions. In this sense, DH practitioners carry the role of language planners and bear significant responsibilities, similar to publishers of Early Modern English.

These issues are not unique to Cantonese; it merely reflects the situation of a better-developed peripheral language in DH. By addressing the obstacles faced in the pursuit of Cantonese DH, we aim to highlight our insights for the development of non-standardized or marginalised languages with Han-influenced orthographies. The solutions and experience from research and engineering teams should raise the awareness of future projects that promote digital humanities studies for other peripheral communities.

Speakers

Dr Chaak-Ming Lau

Founder of words.hk and TypeDuck, Writer and Cantonese linguist

Dr Jon Chui

Inventor of Cantonese Visual Fonts

Mr Henry Chan

Ideographic Research Group, Unicode Consortium

Chinese Language Interface Advisory Committee, HKSAR