An Outline of Contemporary Vietnamese Cantonese

Dana Scott Bourgerie.
Brigham Young University

Abstract

There has been a Han Chinese presence in mainland Southeast Asia for centuries, including in what are now the countries of Cambodia, Laos, and Vietnam. Within this Chinese diaspora, five Chinese varieties prevail in varying proportions: Chaozhou Hainanese, Fujianese, Hakka, and Cantonese. Among these three countries,, the Vietnamese Chinese diaspora has the greatest number of Cantonese speakers. Even after major migrations post-Vietnam War, there are an estimated 500,000 Chinese speakers in Vietnam today, most of whom speak of variety of Cantonese.

Because long contact with the dominant Vietnamese language and separation from other varieties, local Cantonese has developed differently than in Hong Kong, mainland China. and elsewhere. Based on survey results, local collection, and existing literature, this paper outlines some salient features of Cantonese as it is spoken in Vietnam today; For example:

Archaic /less common forms

- 冇相干 mou5seung1gon1 vs 唔紧要 /冇问题
- 客棧 haak3zaan5 'hotel' (Vietnamese kháchsản)
- 算術 syun3seot6 'arithmetic' vs. 數學 sou3hok6
- 返諺 faan1ce2 vs. 返屋企 "'return home'

Phonology

- The historical 55/54 tone distinction is more widely maintained
- Conservative initial n- widely preserved over l-

Lexicon

- Ngộ đi dậm chẩu hầy 我 **đi** 飲酒啊'I'm going to go out drinking' (Vietnamese *đi* for 去 *hui)* (Hoang 2015)

Syntax

- [just] + Cantonese sentence... + thôi = enough/simply ... [sentence] instead of 只係 w/Vietnamese syntax e.g., (只係)食呢個 thôi 'just have one taste'
- 食飯冇 sik6faan6 mou5 for 食飯未 sik6faan6 mei6 'have you eaten yet?'

Reference:

Hoàng, Quốc. 2015. *Cảnh huống song ngữ Việt - Hoa tại đồng bằng sông Cửu Long*. Hanoi: Social Sciences

The CANGLISH Bilingual Corpus:
A Scalable Approach to Analyzing Cantonese-English Code-Switching

Ariel Chan,[1] Grace Wong[2], and Billy Gao[3]
Stanford University
[1]arielchan@stanford.edu, [2]grwong@stanford.edu, [3]billygao@stanford.edu

Code-switching, the alternation between languages, is a pervasive yet complex phenomenon among bilinguals. Analyzing large-scale code-switching data poses challenges due to difficulties in elicitation, segmentation, and annotation in naturalistic settings. While monolingual corpora exist for both Cantonese and English, code-switching databases for these languages remain scarce.

In this presentation, we introduce new methods to streamline the annotation, segmentation, and analysis of code-switching data from the CANGLISH Bilingual Corpus (Chan, 2023), using Voice Onset Time (VOT) analysis as an example. This corpus consists of two datasets: (1) a map task designed to elicit spontaneous code-switching in naturalistic conversations and (2) a sociolinguistic interview exploring language use and identity between the participant and experimenter. The data include conversational speech from 42 Cantonese-English bilinguals across three diasporic communities: 14 heritage bilinguals raised in Cantonese-speaking households in the U.S., 14 homeland bilinguals raised in Hong Kong in a Cantonese-dominant environment and educated in English, and 14 immersed bilinguals from Hong Kong who later moved to the U.S. Data were collected online via Zoom between September and December 2022. In total, the corpus contains 36 hours and 17 minutes of audio and video recordings.

Recordings were initially transcribed using PyTranscriber (Version 1.5, 2022) and manually reviewed by researchers in ELAN (Version 6.8; Sloetjes, & Wittenburg, 2008). Cantonese was transcribed in traditional Chinese characters. To facilitate analysis, we developed a Python script that processes each transcript by inserting spaces between Chinese characters and segmenting Cantonese and English speech into separate .eaf files. This segmentation enables the use of different acoustic models for force alignment in Cantonese and English, as well as the extraction of key phonetic information, including VOT values, stop consonants, preceding vowels, and the words containing the stop consonants in Praat (Version 6.4.12; Boersma & Weenink, 2025). Another Python script was designed to batch-extract speaker identity, whether the extracted token is a code-switched word, and the time elapsed since the last code-switching instance. We propose that this automated approach enhances the efficiency of code-switching research by reducing the labor-intensive process of manual annotation and facilitating large-scale linguistic analysis.

References:

Boersma, Paul & Weenink, David (2025). Praat: doing phonetics by computer [Computer program]. Version 6.4.12, retrieved 2 May 2024 from http://www.praat.org/.

Chan, A. (2023). The Diaspora of Bilinguals: Code-Switching in Three Groups of Cantonese-English Bilinguals. University of California, Los Angeles. University of California, Los Angeles [Doctoral dissertation]. Retrieved from https://escholarship.org/uc/item/8ns2t4fm

PyTranscriber. (2022). PyTranscriber: Open-source automatic transcription software (Version 1.5) Computer software. Retrieved from https://github.com/RafikGuerbi/PyTranscriber

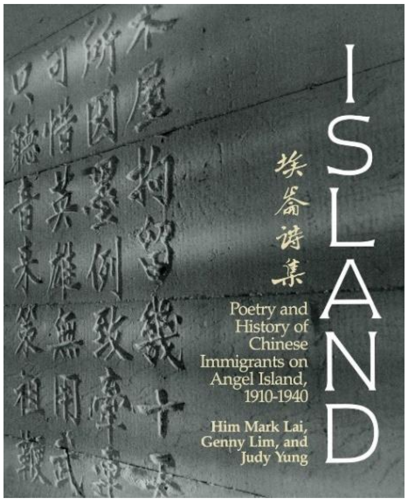
Sloetjes, H., & Wittenburg, P. (2008). Annotation by category - ELAN and ISO DCR. In: Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008).

Abstract for WICL-7. 1-2 November 2024, Brigham Young University, Provo, Utah

Poetry of the Chinese Immigrants from the Pearl River Delta on Angel Island (1910-1940): A Preliminary Linguistic Study

Marjorie K.M. Chan
chan.9@osu.edu
The Ohio State University

Angel Island, the largest natural island in the San Francisco Bay, served as a U.S. immigration station from 1910 to 1940. Although its creation was modelled after that on Ellis Island in New York, the Angel Island Immigration Center was, in fact, primarily an immigration detention center. Most detainees there were from China as a result of the Chinese Exclusion Act of 1882, an anti-Chinese law that prohibited Chinese laborers from entering the U.S. The vast majority of those processed at the Angel Island Immigration Station were Chinese, and they were mainly from the Pearl River Delta in Guangdong, and primarily from following counties: Siyi (四邑, Four Counties), Sanyi (三邑, Three Counties), and Zhongshan (中山, formerly Xiangshan 香山). The poetry left behind on the barrack walls reveal their frustrations, helplessness, anger and other emotions. A total of 135 poems that have been recorded are published with their English translation in Lai et al. (1980, 2014), which includes informative historical background.


ISLAND
埃崙詩集
Poetry and History of Chinese Immigrants on Angel Island, 1910-1940
Him Mark Lai, Genny Lim, and Judy Yung

These 135 poems—almost all anonymous with a small number providing information on their home district—serve as the corpus for this preliminary linguistic study. While a classical style was used for composing mainly heptasyllabic lines, one finds vernacular Cantonese words scattered here and there in many of the poems. For example, one sees 點知 *dim.zi* ‘how would one know,’ 為乜 *wai.mat* ‘for what reason,’ 咁 *gam* ‘so,’ etc. In addition, colloquial words and terms used by Cantonese speakers also appear in these poems, including: 唐山 *Tong.saan* ‘China (lit., Tang (dynasty) + mountain),’ 金山 *gam.saan* ‘California,’ 花旗 *Faa.kei* ‘U.S. (lit., Flowery Flag),’ 埃崙 *Oi.leon* ‘Angel Island (name given by the Chinese immigrants),’ and so forth.

This presentation provides a preliminary linguistic study, exploring in more detail these 135 poems, its vocabulary, choice of rhyming, etc.

Select References

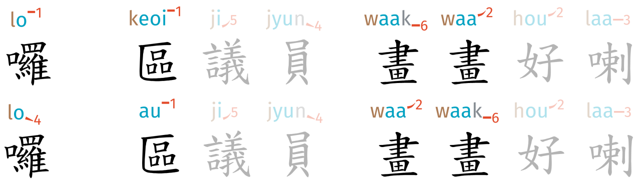
Lai, Him Mark, Genny Lim, and Judy Yung. 1980. *Island: Poetry and History of Chinese Immigrants on Angel Island, 1910-1940.* First edition. Seattle: University of Washington Press.

Lai, Him Mark, Genny Lim, and Judy Yung (eds.). 2014. *Island: Poetry and History of Chinese Immigrants on Angel Island, 1910-1940.* Second edition. Seattle: University of Washington Press. [This second edition more than doubled the length of the first edition. It added four poems from the immigration station on Ellis Island and eight poems from the immigration station in Victoria, Canada, and replaced short, translated excerpts of oral histories of Angel Island detainees recorded in the 1970s with a focus on 20 full-length translated interviews.]

A Markup for Cantonese

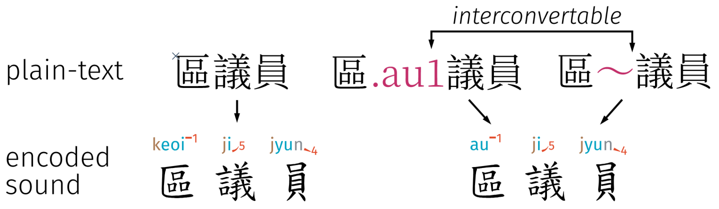
Jon Chui jon@visual-fonts.com Visual Fonts / non-student / applied linguistics

Written representation of Cantonese conventionally uses solely Chinese ideograms, assuming that the glyph drives both pronunciation and the meaning. However, often the combination of glyph and phone is needed to pinpoint a meaning, and omission of phonics introduces ambiguity at character, word, and phrase levels of granularity. This has important real-world implications: the character 囉, the word 區議員, and the phrase 晝畫好喇 simply cannot be uniquely encoded as input to a Text-to-Speech system.

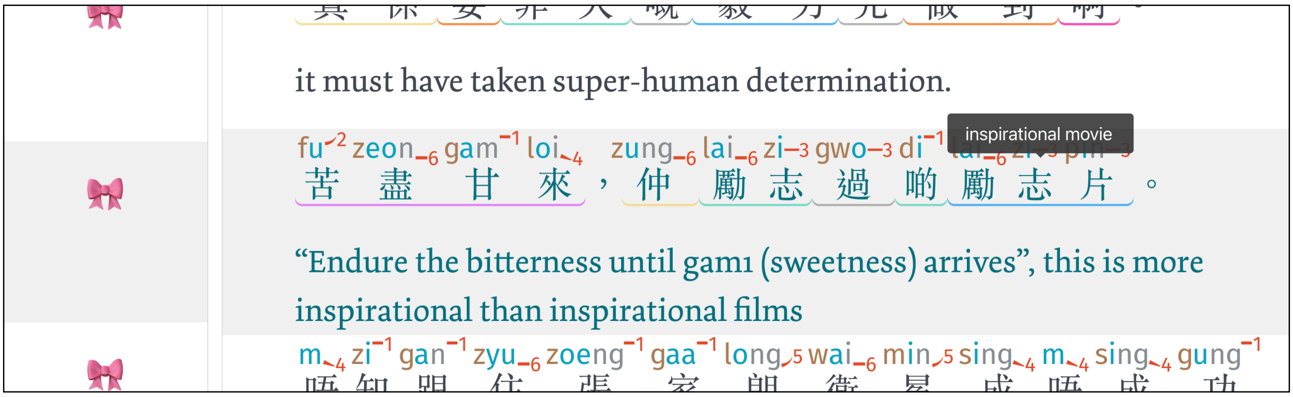


I present a markup specification that resolves this impasse. This specification was implemented in the Cantonese Font family of products. The Font integrates Chinese characters and Jyutping into font glyphs, and uses voluminous OpenType rules to deterministically control which Jyutping is displayed, achieving over 99.7% accuracy on mixtures of vernacular and standard written Chinese.

The markup acts on the different levels of granularity. On character level, mechanisms are provided to specify the exact pronunciation using the *.jyutping* or ~ notation after the character to be modified. This terse, plain-text notation is human-readable and easy to parse programmatically; the downstream applications can choose how this is handled. In the Cantonese Font, the *.jyutping* notation first changes the glyph that is displayed, and the string is then visually subsumed into the glyph.



On the word / phrase levels, the specification introduces mechanisms for indicating segmentation cut-points as well as composable inline tags. Together these allow text to be annotated with multi-layered yet unambiguous meaning.



it must have taken super-human determination.

fu2 zeon6 gam1 loi4 zung6 lai6 zi3 gwo3 di1 lai6 zi3 pin3

苦盡甘來，仲勵志過啲勵志片。

inspirational movie

“Endure the bitterness until gam1 (sweetness) arrives”, this is more inspirational than inspirational films

m4 zi1 gan1 zyu6 zoeng1 gaa1 long5 wai6 min5 sing4 m4 sing4 gung1

I conclude by presenting real-life applications of this markup system. These includes a 150 pages children’s book, a set of 文言文 for HKDSE Chinese, open and closed subtitled videos, and finally, a series of radio dramas with rich audio-synced subtitles. These subtitles, totalling over 25,000 characters, faithfully reflects what the voice actors spoke and is fully and accurately annotated with Jyutping; phrases were segmented and each segment coloured by its parts-of-speech, and translations are available on phrase and word levels.

The Perception of Sentence-Final Particles in Different Cantonese Speakers: A Pilot Study

Ka Fai Law
The Ohio State University
law.246@buckeyemail.osu.edu

Linguistic analyses of Cantonese sentence-final particles (SFPs) have been focused on aspects such as phonology (Law 1990, Wu 2009), syntax (Tang 2002), semantics and pragmatics (Kwok 1984, Bourgerie 1987), sociolinguistics (Chan 1996 and 2001), historical linguistics (Cheung 2009). However, research on the acquisition of SFPs in the Cantonese remains limited. Ko (2000) and Lim (2018) investigated the acquisition sequence of Cantonese SFPs in young children in Hong Kong. Lee and Law (2001) studied children's acquisition of five evidential SFPs. Although these few studies provide valuable insights into the acquisition of Cantonese SFPs in young children, research on adult Cantonese speakers have been much scarce.

To bridge this gap, this pilot study explores the perception of SFPs among different Cantonese speakers, evaluating their sensitivity to the use of SFPs and analyzing the factors that affect their judgement. This study aims to address the question: How sensitive are Cantonese speakers from diverse backgrounds to the use of SFPs?

This study consisted of two components: a sociolinguistic survey, which collected demographic information and language background, and a sentence judgement experiment designed to assess the participants' perception of SFP usage in different contexts. Four SFPs—*aa3*, *lo1*, *zek1* and *ge2*—were selected and formed three pairs of SFPs for the experiment: *aa3* vs *ge2*, *lo1* vs *ge2*, and *zek1* vs *ge2*. An example of stimuli is provided in (1) below. The experiment included 2 lists, each containing 36 context-sentence stimuli and 36 fillers, for a total of 72 contexts. Based on the context, participants were asked to evaluate how natural the sentences sounded on a Likert scale from 1 to 7.

Sixteen participants voluntarily took part in the study, with 7 males and 9 females. The participants contained two main groups: 1) Cantonese native speakers who immigrated to the U.S. and 2) Cantonese heritage speakers who were born in the U.S.

Initial observations reveal that participants in Group 1 (native speakers) experienced minimal difficulty identifying the infelicitous use of SFPs. In contrast, participants in Group 2 (heritage speakers) faced greater challenges in identifying improper uses of SFPs, particularly in sentences where *ge2* is used. The results indicate that this difficulty is associated with differences in language input, suggesting that Cantonese speakers with reduced exposure to Cantonese are less sensitive to the infelicitous use of SFPs.

This pilot study contributes to the field of Cantonese SFP acquisition, shedding light on perceptual differences among various Cantonese-speaking populations and highlighting the role of language input in developing sensitivity to SFP usage.

(1) Context: You and your friends are discussing what to eat for lunch. One friend suggests getting fried chicken, the other friend agrees. And then they ask you what you think. You tell them:

a. 我唔肚餓呀 (aa3)。
'I am not hungry.'

b. *我唔肚餓嘅 (ge2)。
'I am not hungry.'

Selected references

Cheung, H. N. S. (2009). Cantonese made easy: Sentence-final particles in early Cantonese. *Bulletin of Chinese Linguistics*, 3(2), 131-170.

Lee, T. H. T., & Law, A. (2001). Epistemic modality and the acquisition of Cantonese final particles. *Issues in East Asian language acquisition*, 67-128.

Tang, S. W. (2002). Asymmetric Distribution of Cantonese Sentence Final Particles. *Studies in Chinese Linguistics*, 2, 75-84.

Cantonese Poets on Culture and Cantonese Poets in Cantonese on Love

David B. Honey, david_honey@byu.edu
Brigham Young Univeristy

Abstract

This presentation will introduce the major exponents of a distinct Cantonese thematic poetic muse and its chief historical exponents, starting with the Tang poet Zhang Jiuling 張九齡, Song writers Li Maoying 李昴英 _and Yu Jing 余靖, early Qing poets The Three Great Masters of Lingnan 嶺南三大家, Qu Dajun 屈大均, Chen Gongyin 陳恭尹, and Liang Peilan 梁佩蘭. An alternate framework to analyze the development of a Cantonese muse is to trace the various iterations of the leading poetry society founded in the late Yuan and enduring through the early Ming, The Southern Garden Poetry Garden 南園詩社. It persisted on and off throughout the Ming and Qing periods until a last reconvention during the early Republican era.

In contrast to these mainstream poets on Cantonese culture and custom, we will next examine two different types of Cantonese poets who composed verses in Cantonese. First is the famous singer from Hong Kong during the seventies Sam Hui, the king of so-called Cantpop. Among my favorite songs of him celebrating the hard life of lower and middle class Hong Kong-ites are 鬼馬雙星, 天才與白痴, 半斤八兩, 賣身契, 制水歌, 摩登保鑣, and 念奴嬌. Second is Jiu Ji-yung (pinyin Zhao Ziyong] 招子庸, a mid-Qing poet who composed 92 poems on the theme of love.

“Cockroach” in Cantonese

Yingxin HUO & *Abraham Y.S. CHAN*
yhuo@uow.edu.au *abrahamc@uow.edu.au*
UOW College Hong Kong

Abstract

Publication of the *Hong Kong Lexical Lists for Primary Learning* by the HKSAR Education Bureau in 2007 caused quite a stir among local educators as it recommends writing the Cantonese word for cockroach [gaat6 zaat6] as 甴曱 instead of the more common 曱甴. Online discussions gradually accumulate, leading some to argue that the original 曱甴 was mistaken as 甴曱, perhaps first by Eitel in his *Chinese Dictionary* published in 1877.

We notice, however, that Williams’ *Tonic Dictionary* printed in 1856 already listed the word as 甴曱, but came with a vastly different pronunciation [꜀*ká tsát*꜅̲]. Citing myriad sources including dictionaries and dialectal surveys, we conclude that 甴曱 was indeed an earlier written representation, and the subsequent reversal 曱甴 a result of reinterpretation of a 19th-century sound change. We further propose that Williams’ pronunciation concurs with a common term for cockroach, shared among Min, Hakka and Yue dialects, that was probably a Hmong-Mien loanword.

References

Ernest John Eitel, *A Chinese Dictionary in the Cantonese Dialect.* London: Trübner & Co., 1877.

Samuel Wells Williams, *A Tonic Dictionary of the Chinese Language in the Canton Dialect.* Canton: Office of the Chinese Repository, 1856.

陳凱文：〈「曱甴」還是「甴曱」？〉，2017 年 1 月 10 日，讀取自 jonathanovsky.wordpress.com/2017/01/10/「曱甴」還是「甴曱」？/

香港特別行政區政府教育局課程發展處中國語文教育組編：《香港小學學習字詞表》。香港：香港教育局，2007 年。

Application of Corpus Linguistics and GenAI – Creating Teaching and Learning Experience of Teaching Cantonese to Non-Chinese speaking students

LAU, Cindy Wan Yee
College of Professional and Continuing Education
Cindy.lau@cpce-polyu.edu.hk

Abstract

Corpora have primarily been used in linguistic research, but they have not yet become a pedagogical mainstream of language teaching and assessment practices, albeit many positive comments in learners' language development in many countries. Therefore, this paper aims at giving some insights to frontline practitioners to develop teaching materials, lesson plan and synchronous and asynchronous class activities with the aid of language data from GenAI and public corpus. Learners can also benefit from acquiring various digital tools to facilitate their learning. This paper includes: (1) elucidating important notions, namely TPACK, DDL and CBLP and their potential to shape the teaching and learning experiences of educators and learners; (2) explaining how language data from GenAI and corpus can assist teachers' design of teaching curriculum and guide learners to be self-motivated, and (3) comparing pros and cons of language data drawn from GenAI and corpus.

WICL-7 The Zoeng Jyut Gaai Story-telling Speech Dataset
A multi-purposed Cantonese story-telling speech dataset for ASR, TTS linguistic analysis and more

Mingfei Lau (non-student)
Cantonese Computational Linguistics Infrastructure Development Workgroup
laubonghaudoi@icloud.com / support@jyutping.org

Topic: Language Resource, Language Technology.

Abstract

Zoeng Jyut Gaai 張悦楷 is a renowned Cantonese drama actor, stand-up comedian and story-telling artist (講古佬) in 20th century Canton. His story-telling performance represents the highest standard of the art of Cantonese story-telling. His story-telling works are high-quality sources that can be repurposed as speech datasets in various applications, such as Text-to-Speech (TTS) and Automatic Speech Recognition (ASR). In this paper, we present a speech dataset reconstructed from three of his most famous artistic pieces, story-telling *Romance of the Three Kingdoms* (三國演義), *Water Margins* (水滸傳) and *The Final Days of Mao Zedong* (毛澤東的黃昏歲月). We constructed the dataset by first dubbing the audio into subtitles, and then segmenting the audio into sentences based on the aligned timestamps. We ended up having more than 100 hours[1] of high-quality annotated speech recordings with high expressiveness, emotion, register and vocabulary coverage.

To demonstrate a possible usage / application of this dataset, we created a TTS demo https://huggingface.co/spaces/laubonghaudoi/zoengjyutgaai_tts with only ~30 hours of training data which produces fairly satisfying quality. Further applications include using the dataset as an ASR test set [2] and more.

The dataset is open free to the public under Creative Commons Zero v1.0 Universal license on CanCLID/zoengjyutgaai · Datasets at Hugging Face, The home page for the dataset is https://canclid.github.io/zoengjyutgaai/ . The dataset belongs to the world-wide public domain and is available to commercial or private use. We believe by opening this dataset to the public, we can boost the advancements of Cantonese TTS, ASR and other speech technologies, as well as the linguistic research towards the story-telling art performance, and the promotion of traditional Cantonese culture.

References

[1] GPT-SoVITS WebUI https://github.com/RVC-Boss/GPT-SoVITS
[2] Cantonese ASR Eval https://github.com/AlienKevin/cantonese_asr_eval

[1] The dataset is still Work-In-Progress and we have currently published 104 hours of data. We are still actively working on curating the dataset and by the time it is finished, it is expected to have more than 120 hours of data.

Identity negotiation of heritage Cantonese learners: Navigating learner investments and language ideologies in multicultural Canada

Raymond Pai, The University of British Columbia, raymond.pai@ubc.ca

Abstract:
Cantonese, historically the dominant Chinese language variety in the multicultural landscape of Canada, faces an increasing need for preservation and promotion. Heritage Cantonese learners try to preserve their linguistic identity amidst prevailing societal language ideologies (Xiao 1998; Yu & Chan, 2017), with English/French being the official languages of Canada while Mandarin gaining dominance in terms of educational and economic values. This presentation is part of a larger study that explores how adult heritage learners of Cantonese negotiate their cultural and linguistic heritage while integrating into a predominantly English-speaking society. The research focuses on the strategies and practices employed by heritage Cantonese learners in their (in)formal learning of Cantonese, and how these efforts interact with broader societal expectations and pressures.

Using narrative and thematic analysis approaches, this study combines survey and interview data from heritage Cantonese learners across Canada. The surveys provide an overview of language use patterns, while in-depth interviews of twelve selected focal participants offer insights into their personal experiences and perceptions regarding their linguistic identity and language learning investments. Preliminary findings indicate that heritage Cantonese learners employ various resources and strategies in their Cantonese learning trajectories, such as enrolling in heritage language schools and university courses, creating Cantonese-speaking environments at home and social domains, and utilizing popular media and platforms in Cantonese. However, these efforts often encounter challenges, including the dominance of English and Mandarin in educational institutions, peer pressure to conform to English norms, and the internalization of societal attitudes that may devalue heritage languages.

This study highlights the tension between heritage language investments and social assimilation, revealing how heritage Cantonese learners navigate these dynamics to maintain linguistic identity. It also underscores the role of community support and policy initiatives in promoting multilingualism and cultural diversity. The findings contribute to the broader discourse on language preservation, identity, and multiculturalism in Canada, offering implications for educators, policymakers, and community organizations working towards inclusive and equitable linguistic landscapes.

Xiao, H. (1998). Chinese language maintenance in Winnipeg. *Canadian Ethnic Studies Journal*, *30*(1), 86-98.
Yu, H., & Chan, S. (2017). The Cantonese Pacific: Migration Networks and Mobility Across Space and Time. *Trans-Pacific Mobilities: The Chinese in Canada*, 25-48.

WICL 7: April 2025

Sources on the North American Cantonese Diaspora

Kelly Summers*
FamilySearch.org

Abstract

Cantonese-speaking Chinese immigrants began arriving in North America in large numbers during the mid-1800s, primarily from Guangdong province. Despite their contributions, they faced intense discrimination and were targeted by exclusionary laws, such as the U.S. Chinese Exclusion Act of 1882 and Canada's Chinese Head Tax and Chinese Immigration Act of 1923. The records created because of these laws provide rich sources of information about these early Chinese immigrants. We will examine records, collections & tools to help identify the Chinese Diaspora.

Kelly Summers, Accredited Genealogist®

Kelly works for FamilySearch International as a Content Strategist for Asia. She identifies, recommends acquisition, and supervises the indexing of records and collections throughout many countries in Asia. She is especially interested in records pertaining to the Chinese Diaspora.

Abstract for WICL-7 Conference

Family Language Policies for Cantonese Heritage Language Maintenance

Wah Sun SZE
University of British Columbia
celiasws@gmail.com

Once a dominant language among Chinese diaspora in Canada, Cantonese is becoming increasingly vulnerable, as demonstrated by the decline of the intergenerational transmission of Cantonese as a home language and the relatively low full and partial retention rate of Cantonese compared with other immigrant languages commonly spoken at home, such as Mandarin and Punjabi (Statistics Canada, 2017; 2022). This decline underscores the pressing need to explore Family Language Policies, the foundation for language maintenance, conducive to the intergenerational transmission of Cantonese.

By adopting the Family Language Policy (FLP) framework (Spolsky, 2004), this conceptual paper draws on existing literature and case studies to examine common beliefs, practices and strategies that promote Cantonese language maintenance at home milieu. FLP, consisting of three components: language ideology and beliefs, language practices, language management and planning (Spolsky, 2004), refers to the implicit and explicit planning of language and literacy practices in home domains between family members (King et al., 2006). By exploring these three components of FLP, this paper aims to inform Cantonese-speaking families of various language beliefs, language practices, and language management strategies conducive to the preservation of Cantonese.

Findings suggest that positive parental attitudes towards Cantonese, bilingualism and bidialectalism, coupled with maximizing Cantonese home language use, and strategic planning of home language environment and schooling options are crucial in Cantonese language maintenance. Based on these findings, this paper offers practical recommendations for Cantonese-speaking parents, and proposes a parents' guide to facilitate knowledge mobilization for Cantonese transmission.

Key References:

King, K. A., & Fogle, L. W. (2006). Bilingual parenting as good parenting: Parents' perspectives on family language policy for additive bilingualism. International Journal of Bilingual Education and Bilingualism, 9(6), 695–712. https://doi.org/10.2167/beb362.0

Spolsky, B. (2004). Language Policy (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511615245

Statistics Canada. (2017, August 31). Linguistic diversity and multilingualism in Canadian homes. Statistics Canada. https://www12.statcan.gc.ca/census-recensement/2016/as-sa/98-200-x/2016010/98-200-x2016010-eng.cfm

Statistics Canada. (2022). Increasing diversity of languages, other than English or French, spoken at home [Infographic]. https://www150.statcan.gc.ca/n1/pub/11-627-m/11-627-m2022051-eng.htm

Harnessing Generative Artificial Intelligence to Study Cantonese Linguistic Variation

Paul Ueda[1], Ka Fai Law[1], Justin Leung[2], Marjorie K.M. Chan[1]
[1] The Ohio State University, [2] University of Toronto

Generative Artificial Intelligence (GenAI) programs are often based upon probabilistic methods, which can introduce bias into the output. For example, Gillespie (2024) examines how GenAI programs tend to produce socially-normative cultural works. Gautam et al. (2024) likewise note that GenAI systems are prone to distort and stereotype their representation of marginalized groups. However, little research has been done to see to what degree GenAI programs are able to reproduce sociolinguistic variation.

The present project utilizes multilingual translations of *Le Petit Prince* (such as Tsoi's (2021) translation of Saint-Exupéry 1943) to examine GenAI's ability to authentically reproduce sociolinguistic variation—in this case regionally-marked variants of Cantonese. This was achieved with two primary methodologies. The first was a declarative knowledge task in which ChatGPT was asked to explain differences between regional subvarieties of Cantonese. The second task consisted of a translation task for which ChatGPT was tasked with outputting regionally-marked translations of *Le Petit Prince.* Each response was regenerated three times.

For declarative knowledge, ChatGPT contradicted previous answers—such as calling both Hong Kong Cantonese (HKC) and Guangzhou Cantonese (GZC) the more phonologically conservative variant, for example, in lexical tones. Overall, the answers were exceedingly vague and represented stereotypes, echoing Gautam et al.'s (2024) findings. When asked to output translation, ChatGPT showed higher rates of formality or Mandarinization when compared to Tsoi's published translation, as shown in Table 1. This occurred regardless of the targeted regional variety, with the implication ChatGPT was unable to differentiate them, although there was a slight bias towards Hong Kong with the usage of Traditional characters. The results display that GenAI programs are often unable to reproduce authentic language variation—instead, relying upon stereotypes or whatever bias was present in the original training data. One implication of this is the importance of understanding sociolinguistic variation in training GenAI programs to produce authentic speech. Furthermore, this project presents implications and possible future directions on the degree of linguistic prejudice that Cantonese faces when participating in the technological arena.

Table 1: Comparison of lexical choice between Tsoi's Translation and GenAI output.

Vocabulary in Tsoi's Translation		GenAI Output	
上面	"on"	上	"on"
掗埞	"take up space"	佔(地)位	"take up space"
朝早	"morning"	早晨	"morning"

References

Gautam, Sanjana, Pranav Narayanan Venkit, Sourojit Ghosh. 2024. From Melting Pots to Misrepresentations: Exploring Harms in Generative AI. In *CHI'24: Generative AI and HCI workshop in CHI '24, May 11-16, 2024, Honululu Hawaii*. Not paginated, 7 pages. New York: ACM.

Gillespie, Tarleton. 2024. Generative AI and the politics of visibility. *Big Data & Society* 11(2): 1-14. www.doi.org/10.1177/20539517241252131.

Saint-Exupéry, Antoine de. 1943. *Le Petit Prince. The Little Prince in Cantonese with Jyutping.* 小王子香港粵拼版. Translated in 2021 by Wai-chuen Thomas Tsoi (蔡偉泉). Edited by Wai-man Zoe Lam and Suet-yee Suki You, and Jyutping romanization by Chaak-ming Lau, Chung-man Cui, Wing-yan Grace Chan, Hoi-lam Tiffany Pang, Wai-man Zoe Lam, and Wing-yan Vanessa Tsang. Hong Kong: Bleu Publications.

Revisiting Cantonese syllable-final consonant variation: an ultrasound study

Meihao Wan, Peggy Mok
The Chinese University of Hong Kong
wanmeihao@link.cuhk.edu.hk, peggymok@cuhk.edu.hk

Introduction: Cantonese has two sets of consonant codas: nasal codas (-m, -n, -ng) and plosive codas (-p, -t, -k). An ongoing sound change has been observed in Hong Kong Cantonese, involving the merging of [-t] and [-k], as well as [-n] and [-ng] [1, 2, 3, 4, 5]. Notably, the direction of this change is not unidirectional: there are the replacements of [-ng] with [-n], [-n] with [-ng], [-k] with [-t] and [-t] with [-k], and the direction may be influenced by the preceding vowels [5]. Similar instability of final consonants has also been observed in other Chinese dialects [6]. This study aims to provide an articulatory description of this sound change and explore the possible mechanisms that drive it.

Method: This study investigated the merging of consonant codas in Hong Kong Cantonese. Ultrasound imaging was used to capture tongue movements during the production of target words. Two native Cantonese speakers (1 male, 1 female, ages 21-23) participated in the study. Both were students at the Chinese University of Hong Kong. All target words were embedded in a carrier sentence: zoi3 gong2 ___ ko2 ko3 zi6 ('Say the word __ again'). Stimuli included different Cantonese vowels ([a, ɐ, ɛ, œ, ɵ, ɔ, i, ɪ, y, u, ʊ]) to examine the influence of vowel contexts. There are 336 tokens (28 rimes x 4 words x 3 repetitions) per participant. Audio signals and ultrasound imaging were recorded simultaneously, with target words presented in a random order.

Results: Ultrasound imaging revealed that the tongue tips for /-t/ and /-n/ are significantly higher than those for velar codas. While the sound change is occurring, it was only observed in a limited number of words. For the majority of words, no evidence of the sound change was found in the two speakers. We only observed the replacement of [-ng] with [-n] and [-k] with [-t] in certain words but not the reverse. Moreover, although the preceding vowels influenced tongue shape during the coda production, no significant vowel effect was observed. **Discussion**: The limited extent of the sound change, with the two speakers still distinguishing most words, suggests that the process may be in an early stage of lexical diffusion, primarily affecting less common words rather than high-frequency words. No significant vowel effect indicates that vowel context is unlikely to be a primary phonetic

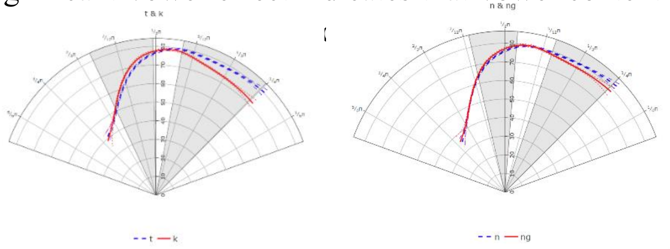


Figure 1: GAMM results of Cantonese consonant codas

Reference: [1] Bauer, R. S. (1979). Alveolarization in Cantonese: a case of lexical diffusion. *Journal of Chinese Linguistics*, *7*(1), 132-141. [2] Yeung, S. W. (1981). Some aspects of phonological variations in the Cantonese spoken in Hong Kong. *HKU Theses Online (HKUTO)*. [3] Zee, E. (1999). Change and variation in the syllable-initial and syllable-final consonants in Hong Kong Cantonese. *Journal of Chinese linguistics*, *27*(1), 120-167. [4] Chen, N. L. L. (1999). The velar coda variation in Hong Kong Cantonese. *Unpublished Masters Thesis. City University of Hong Kong, Hong Kong*. [5] To, C. K., McLeod, S., & Cheung, P. S. (2015). Phonetic variations and sound changes in Hong Kong Cantonese: Diachronic review, synchronic study and implications for speech sound assessment. *Clinical linguistics & phonetics*, *29*(5), 333-353. [6] Chen, M. Y., & SY, W. (1975). Sound change: actuation and implementation. *Language*, 255-281.

Sociative causative *zing* in Cantonese

Pui Yee Yuen, Humboldt of University of Berlin

Previous research on periphrastic causatives has explored two main areas: (1) the distinction between lexical and periphrastic causatives (Wolff, 2003; Song & Wolff, 2003), and (2) the variations between different forms of periphrastic causatives (Vichit-Vadakan, 1976; Nadathur & Lauer, 2020). Studies have shown that periphrastic causatives are associated with indirect and intentional causation and can express either causal sufficiency or necessity.

While multiple periphrastic causatives exist in Cantonese, they have remained largely unexplored. This study investigates the 整 *zing*-causative by comparing it with the 令 *ling*- and lexical causatives. Contrary to findings in other languages, the initial tests do not reveal significant differences between *zing* and *ling* or lexical ones in terms of directness or intentionality, nor do they associate with causal sufficiency and necessity, respectively.

Shibatani & Chung (2002) have discovered causatives in Japanese and Korean, which are neither direct nor indirect, and framed them as "sociative causatives". Building on Shibatani (1973)'s framework of manipulative and directive causation, sociative causatives require: shared time and space between causation participants, an agentive causer who is actively involved in causation without physical manipulation and agentive causee(s) who executes the action. In (1), the *mother* supervises rather than physically forces the action:

(1)	hahaoya-ga	kodomo-ni	hon-o	yoma-se-te	i-ru
	mother-NOM	child-ACC	book-ACC	read-CAUS-CONJ	be-PRS

'Mother is making the child read a book.' [Japanese]

Sociative causatives are situated on the continuum between one end of direct causation and another end of indirect causation. Depending on the level of active involvement the casuer has, they can be further subcategorised into three types: joint-action, assistive, and supervision (Shibatani & Chung (2002)).

To further investigate sociative causatives, the current study analyses corpora (Baroni & et al., 2009; Chin & Tweed, 2019) and conducts a forced-choice judgment test. In the test, native speakers are given scenarios where a caused event occurred and asked to choose between *zing*- and *ling*- or lexical causatives. The result discovers that the *zing*-causative aligns with the concept of sociative causatives. Furthermore, the study refines the concept of active involvement in sociative causatives by emphasising responsibility and controllability. The *zing*-causative is indeed preferred in the judgemental test:

(2) Your mum lit up an incense stick. When you got back home and opened the door, the wind blew in and the stick extinguished. In other words···

nei^{23}	*zing*^{35}*dou*^{33}/*ling*^{22}*dou*^{33}	*zi*^{55}	*hoeng*^{55}	*sik*^{55}	*zo*^{35}
you	make/cause	CL	incense stick	extinguish	PFV

'You made the stick extinguish.' [Cantonese]

In (2), while "you" do not directly extinguish the *stick*, the responsibility traces back to the causer ("you"), with the *wind* acting as the active causee, so participants tend to choose *zing* in this scenario. The experimental result reveals that *zing*-causative usage is a continuum, increasing with the causer's level of responsibility, distinguishing it from the periphrastic *ling*-causative's indirect causation.

Another novel finding of this study is that the association of responsibility with an unwanted outcome, contrasting with lexical causatives which typically describe welcomed results, compare the actions of *dad* in (3) below:

(3) a. Lexical causative

baa^{21}*baa*^{55}	*ting*^{21}	*zo*^{35}	*jam*^{55}*ngok*^{22}
Dad	stop	PFV	music

'Dad stopped the music'

b. Zing-causative

baa^{21}*baa*^{55}	*zing*^{35}*dou*^{33}	*jam*^{55}*ngok*^{22}	*ting*^{21}	*zo*^{35}
Dad	make/cause	music	stop	PFV

'Dad made the music stop' [Cantonese]

The findings solidify the concept of sociative causative and extend the understanding of it beyond certain languages, suggesting that it may be a universal linguistic property rather than a language-specific phenomenon. Yet, languages differ in how they express sociative causatives: shared form with periphrastic causative (e.g., Japanese), shared form with lexical causative (e.g., Korean), or even a distinctive (periphrastic) form (e.g. Cantonese). This also reveals the multifacetedness of the causation relationship in languages.

References [1] Baroni et al. 2009. In *Language resources and evaluation.* [2] Chin et al. 2019. In *Workshop on Cantonese (WOC): Cantonese study: An empirical approach.* [3] Nadathur et al. 2020. In *Glossa: a journal of general linguistics.* [4] Shibatani 1973. In *Journal of Linguistics.* [5] Shibatani et al. 2002. In *Japanese/Korean Linguistics.* [6] Song et al. 2003. In *Language, culture, and mind.* [7] Vichit-Vadakan 1976. In *The grammar of causative constructions.* [8] Wolf 2003. In *Cognition.*