香港教育大學
The Education University of Hong Kong

ONLINE

School of Cantonese Studies

粵語研究研習班

15-16 May 2021

School Handbook

https://www.eduhk.hk/lml/scs2021/

語言學及現代語言系
LINGUISTICS AND MODERN LANGUAGE STUDIES

Centre for Research on Linguistics and Language Studies

Contents

Organizing Committee.....1

Acknowledgements.....2

About School of Cantonese Studies 3

About the Speakers 4

Schedule.....12

Session 1: Frontiers in Cantonese corpus-based research.....13

Session 2: Linguistic issues in constructing Cantonese corpora 15

Session 3: Digital processing of Cantonese corpus data.....20

Session 4: Demonstration of Cantonese digital resources and tools and Cantonese learning apps.....21

(1) Demonstration of Cantonese digital resources and tools.....21

(2) Demonstration of Cantonese learning apps.....23

Poster.....25

Organizing Committee

Hintat Cheung 張顯達
Department of Linguistics & Modern Language Studies, Centre for Research on Linguistics and Language Studies, The Education University of Hong Kong

Andy Chin 錢志安
Department of Linguistics & Modern Language Studies, Centre for Research on Linguistics and Language Studies, The Education University of Hong Kong

Shin Kataoka 片岡新
Department of Linguistics & Modern Language Studies, Centre for Research on Linguistics and Language Studies, The Education University of Hong Kong

Yik Po Lai 黎奕葆
Department of Linguistics & Modern Language Studies, Centre for Research on Linguistics and Language Studies, The Education University of Hong Kong

Chaak Ming Lau 劉擇明
Department of Linguistics & Modern Language Studies, Centre for Research on Linguistics and Language Studies, The Education University of Hong Kong

Cherry Yeung 楊舜鈴
Department of Linguistics & Modern Language Studies, Centre for Research on Linguistics and Language Studies, The Education University of Hong Kong

Acknowledgements

The Organizing Committee is grateful for the sponsorship from The Linguistic Society of Hong Kong (香港語言學學會).

About School of Cantonese Studies

The Department of Linguistics and Modern Language Studies, and the Centre for Research on Linguistics and Language Studies at The Education University of Hong Kong (EdUHK) organized the first School of Cantonese Studies in May 2019. The five-day event covered nine lectures delivered by twelve scholars specializing in major topics pertinent to the Cantonese language. There were 60 participants coming from different parts of the world.

With the positive feedback on the first School, the two units at EdUHK organized the School of Cantonese Studies again in 2021 which carries the theme **Cantonese Studies in the Digital Age**. We are now living in the Information Age with different kinds of data. It is important for us to apply appropriate methods to collect, process, interpret, and represent these data. Digital technologies play a significant role in this regard. The second School of Cantonese Studies invited scholars to introduce their research work in Cantonese involving digital technologies, such as corpus-based research, online tools and resources for Cantonese studies, as well as construction and processing of Cantonese corpus data.

Aims of the School

- To introduce recent developments and knowledge on different domains in Cantonese Studies to the participants;
- To introduce systematic and rigorous methodologies for conducting research on Cantonese;
- To provide a venue for scholarly exchange and interaction between scholars and participants of different backgrounds who are interested in Cantonese Studies.

About the Speakers (In order of last names)

CHEUNG Hin Tat 張顯達

The Education University of Hong Kong

Cheung Hin Tat is Professor at the Department of Linguistics and Modern Language Studies and Director of the Centre for Research in Linguistics and Language Studies at The Education University of Hong Kong. He has been investigating a wide range of issues in developmental psycholinguistics, including grammatical acquisition, language impairment and narrative development. For promoting research in first and second language acquisition, he and his associates have constructed two corpora: Taiwan Corpus of Child Mandarin and the LTTC English Learner Corpus.

Andy CHIN 錢志安

The Education University of Hong Kong

Andy Chin is Head of the Department of Linguistics and Modern Language Studies at The Education University of Hong Kong. His research interests include Cantonese studies, sociolinguistics, discourse analysis and corpus linguistics. His Corpus of Mid-20th Century Hong Kong Cantonese won the gold medal and special award at the Silicon Valley International Invention Festival in 2019. In March 2021, Andy' **CanPro** Cantonese language mobile app won the silver medal in the virtual edition of Geneva's Inventions Expo. CanPro is an interactive listening exercise consisting of 400 common Cantonese expressions extracted from his Cantonese corpus.

Luis Morgado da COSTA

Nanyang Technological University

Luis Morgado Costa is a PhD student at the Interdisciplinary Graduate School, Nanyang Technological University (NTU), Singapore. Before that, he was a research associate in the Computational Linguistics Lab, Division of Linguistics and Multilingual Studies, also at NTU. He is a member of DELPH-IN, sharing the communal commitment of open-source development of NLP tools for high quality (linguistically motivated) syntactic and semantic parsing. And also a member of the Global Wordnet Association, contributing to open-source research on computational lexical semantics. He has a broad range of research interests, including Parsing and Generation, Computational Lexicography, Computer Assisted Language Learning, Word Sense Disambiguation, Sentiment Analysis, Machine Translation, as well as general Mandarin Chinese and Japanese Linguistics.

KATAOKA Shin 片岡新

The Education University of Hong Kong

Shin Kataoka is an Assistant Professor at the Department of Linguistics and Modern Language Studies, The Education University of Hong Kong. His research interests include diachronic study of Cantonese, missionary works in Cantonese, etc. He recently constructed The Early Cantonese Bible Database (早期粵語聖經資料庫). He also co-authored many Cantonese textbooks and constructed EdUHK Cantonese Self-Learning Website (香港教育大學粵語自學平台).

KI Mei Ying 祁美瑩

The Chinese University of Hong Kong

Ki Mei Ying is currently a Research Assistant at the Department of Chinese Language and Literature, The Chinese University of Hong Kong. Her research interests are Cantonese phonetics and its interface studies. She is also passionate about different areas of applied linguistics, including language teaching, language testing, and computational linguistics. She, with Lai Yik Po and Yip Ka Fai, developed CanTONEse [iOS] [Android], a mobile app for non-native speakers to learn Cantonese tones featuring application of sound-colour synaesthesia.

Charles LAM 林子鈞

The Hang Seng University of Hong Kong

Charles Lam is Assistant Professor at the Department of English, The Hang Seng University of Hong Kong. His primary research interest is syntax-semantics interface under the generative framework. In addition, he also conducts research in digital humanities and applications of semantics.

LAI Yik Po 黎奕葆

The Education University of Hong Kong

Lai Yik Po is a Post-doctoral Fellow at the Department of Linguistics and Modern Language Studies, The Education University of Hong Kong. His research interest is in Cantonese linguistics, Chinese comparative dialectology, and historical linguistics. He, with Ki Mei Ying and Yip Ka Fai, developed CanTONEse [iOS] [Android], a mobile app for non-native speakers to learn Cantonese tones featuring application of sound-colour synaesthesia.

LAU Chaak Ming 劉擇明

The Education University of Hong Kong

Dr Lau Chaak Ming is currently Assistant Professor at Department of Linguistics and Modern Language Studies of the Education University of Hong Kong. He is a linguist and a digital humanities practitioner, interested in both theoretical and computational linguistics. He founded 粵典 (words.hk) in 2014, a dictionary project with over 500 volunteers and contains over 50,000 word entries, and is the main developer of gamification platform CantoSounds and mobile input method CantoKey.

LAU Ming Fei 劉銘霏

Cantonese Computational Linguistics Infrastructure Development Group (CanCLID)

Lau Mingfei is a computational linguist and the co-founder of CanCLID. He is a main contributor of rime-cantonese input method lexicon. He is interested in corpus development and digitalization for low-resource languages.

Jackson LEE 李麟

Author of PyCantonese

Jackson Lee works on computational linguistics. He is the author of PyCantonese, a Python library for Cantonese linguistics and natural language processing.

LEE Hun-Tak Thomas 李行德

Tianjin Normal University and the Chinese University of Hong Kong

Lee Hun-Tak Thomas is Professor and Principal Researcher at the Institute of Theoretical Psycholinguistics of Tianjin Normal University and Emeritus Professor at The Department of Linguistics and Modern Languages, The Chinese University of Hong Kong. His research has centered on the acquisition of syntax and semantics, with special reference to the quantificational competence of Mandarin-speaking and Cantonese-speaking children. Along with other colleagues then based at Hong Kong Polytechnic University and University of Hong Kong (Colleen Wong and Sam Leung), Professor Lee developed The Hong Kong Cantonese Child Language Corpus (CANCORP) in the 1990s, which has an updated version made available in 2012.

LEUNG Wai Mun 梁慧敏

The Hong Kong Polytechnic University

Leung Wai-mun is Assistant Professor of Applied Chinese Linguistics at The Hong Kong Polytechnic University. Her research interests include Cantonese studies, sociolinguistics, Chinese language education and the teaching of Chinese to non-Chinese speaking students. She recently constructed "The 19th Century (1865-1894) Cantonese Christian Writings Database 十九世紀中後期（1865-1894）粵語基督教典籍資料庫". She is also the first author of the monograph *Biliteracy and Trilingualism: Language Education Policy Research in Hong Kong*（兩文三語：香港語文教育政策研究, CityU Press, 2020) in which a chapter is devoted to discussing Cantonese.

LUKE Kang Kwong 陸鏡光

Nanyang Technological University

Professor Luke is NTU President's Chair Professor of Linguistics and Chair of the School of Social Sciences at Nanyang Technological University, Singapore. He has done work on a number of areas of linguistics, including phonology, syntax, sociolinguistics and natural language understanding. The main focus of Prof. Luke's recent research is on talk and social interaction using an Ethnomethodological Conversation Analytic approach. In the 1990s, Prof. Luke developed The Hong Kong Cantonese Corpus with about 230,000 Chinese words.

Joanna Ut-Seong SIO 蕭月嫦

Palacký University Olomouc

Joanna Ut-Seong Sio is Assistant Professor at the Department of Asian Studies at Palacký University Olomouc at Czech Republic. Her research interests include Chinese languages, especially in the area of syntax and semantics, as well as the use of verbal arts in the training of communication skills. She has developed the Cantonese WordNet.

Raymond TSE 謝家尉

粵典 words.hk

Raymond Tse is currently the chief editor of words.hk (《粵典》), a crowdsourcing-based Cantonese-Cantonese dictionary project including over 50,000 word entries. Apart from practical lexicography, he is committed to cross-sector promotion of Cantonese. In 2020, he taught an introductory workshop on phonological and lexical changes, grammar topics, tone-melody match, and the development of written Cantonese to students in Shaw College, the Chinese University of Hong Kong. Recently he reviewed Cantonese subtitles in The Way We Keep Dancing (《狂舞派 3》) as invited.

Benjamin T'SOU 鄒嘉彥

City University of Hong Kong / The Hong Kong University of Science and Technology

Benjamin T'sou is Emeritus Chair Professor (CityUHK) and member of Académie Royale des Sciences d'Outre-Mer (Belgium). Since 1995, his team has been cultivating the Cross-language Linguistic Variations in Chinese Corpus, LIVAC.

His publications include *Chinese Language and Chinese Society*, *Anthology on Language Contact*, *A Textbook on Sociolinguistics*, *Quantitative and Computational Studies on the Chinese Language*, *Linguistic Corpus and Corpus Linguistics in the Chinese Context*, *Quadra-syllabic Idiomatic Expressions in Cantonese: Inheritance and Innovation*.

YIP Ka Fai 葉家煇

Yale University

Yip Ka Fai is a PhD student in the Department of Linguistics at Yale University. His research areas are syntax and semantics with a focus on Cantonese, Mandarin, and Vietnamese. He is also interested in linguistic variations under a corpus-based approach. He, with Ki Mei Ying and Lai Yik Po, developed CanTONEse [iOS] [Android], a mobile app for non-native speakers to learn Cantonese tones featuring application of sound-colour synaesthesia.

Carine YIU 姚玉敏

The Hong Kong University of Science and Technology

Carine Yiu is Associate Professor of Humanities at the Hong Kong University of Science and Technology. Her research interests include Chinese linguistics with a focus on Cantonese; history of Chinese dialects; syntax, semantics, typology. She has developed a number of corpora of Cantonese and other Chinese dialects: Early Cantonese Colloquial Texts: A Database, Early Cantonese Tagged Database, Database of Early Chinese Dialects.

ZHANG Ling 張凌

The Education University of Hong Kong

Zhang Ling is Assistant Professor at the Department of Chinese Language Studies at The Education University of Hong Kong. Her research interests include Cantonese phonetics and phonology, studies of sentence-final particles, and applied linguistics. She recently developed a mobile App 《古詩粵唱粵啱 Key》 to help students learn Chinese poems through singing.

Schedule*

15 May (0830 – 0900, HKT)	Registration (Password: 056407)
15 May (0900 – 0915, HKT)	Opening Remarks and Group Photos (Password: 056407)
15 May (0915 – 1130, HKT)	Frontiers in Cantonese Corpus-based Research (Password: 056407) Benjamin T'SOU
15 May (1400 – 1630, HKT)	Linguistic Issues in Constructing Cantonese Corpora (Password: 964884) Discussants: Andy CHIN, Thomas Hun-Tak LEE, LUKE Kang Kwong, Carine YIU Moderator: CHEUNG Hin Tat
16 May (0900 – 1200, HKT)	Digital Processing of Cantonese Corpus Data (Password: 175729) Charles LAM, LAU Chaak Ming, Jackson LEE
16 May (1400 – 1730, HKT)	Demonstration of Cantonese Digital Resources and Tools and Learning Apps (Password: 798685) 1. 《成語填字遊戲》: Benjamin T'SOU (14:00 – 14:20) 2. 《古詩粵唱粵啱 Key》: ZHANG Ling (14:20 – 14:40) 3. CanTONEse: KI Mei Ying, LAI Yik Po, YIP Ka Fai (14:40-15:00) 4. Q&A (15:00 – 15:10) 5. The 19th Century (1865-1894) Cantonese Christian Writings Database: LEUNG Wai Mun (15:10-15:30) 6. The Early Cantonese Bible Database: Shin KATAOKA (15:30-15:50) 7. Early Cantonese Colloquial Texts: A Database, Early Cantonese Tagged Database, Database of Early Chinese Dialects: Carine YIU (15:50-16:10) 8. Q&A (16:10-16:20) 9. words.hk《粵典》: Raymond TSE (16:20-16:40) 10. Cantonese Wordnet: Luis Morgado da COSTA, Joanna Ut-Seong SIO (16:40-17:00) 11. Rime-Cantonese and Inject-Jyutping: LAU Ming Fei (17:00-17:20) 12. Q&A (17:20-17:30)
16 May (1730 – 1745, HKT)	Concluding Remarks (Password: 798685)

* All talks are delivered in English supplemented with Chinese. Materials of the School can be accessed at the Github page.

Session 1

Some Frontiers in Cantonese Corpus-based Research

Benjamin K. T'sou

City University of Hong Kong

The Hong Kong University of Science and Technology

Much work is in evidence in corpus linguistics and in Cantonese linguistics. The cross-fertilization between the two will be beneficial to both and would promote new frontiers in Cantonese linguistics in the study of the Chinese language and its underlying social and cultural diversity. A major question may be raised on how the goals of Cantonese linguistics may differ from Chinese linguistics. It leads to some relevant issues on what constitute salient linguistic variations and their broader societal significance when Cantonese is compared with Mandarin and other dialects. The approach I would like to take is to draw attention to the overt, opaque and covert variations between Cantonese and other dialects and the deeper variations they reveal about the dialect communities.

There can be gradations of overt to covert variations. The clearly marked and overt differences between morphemes in Cantonese and Mandarin (e.g.咩 in Cantonese and 嗎 in Mandarin) could shed light on major issues in grammar, (e.g. interrogation). This could be compared with opaque variations, for example: 1) MSC 派發: colloquial Mandarin 發(獎金) vs Cantonese 派(獎金), 2) MSC 房屋: colloquial Mandarin(有)房 vs Cantonese(有)屋, and 3) MSC 溶化: colloquial Mandarin(冰)化 or (冰)溶化 vs Cantonese (冰)溶 only. There are other more subtle differences in spite of the identity of morphemes: 1) Cantonese 你車我去 vs Mandarin 你用車載我去; 2) Cantonese 你車倒個學生 vs Mandarin 你的車撞倒了學生; 3) Cantonese (a) 架車要賣了, (b) 架架車都要賣 vs Mandarin (a) 那部/輛車要賣了(b) 每一部/輛車都要賣了 and 4) 打擊 vs. 擊打, Cantonese 打爛/打破 vs Mandarin 打破.

Still many other kinds of variations are found, which would also benefit from the judicious use of well curated corpus and other means in digital humanities to uncover and monitor a wide range of linguistic and collateral social and cultural undercurrents of significance in the Cantonese speaking communities.

This presentation will explore the scope of corpus-based research and its usefulness and of open ended or closed corpus. Furthermore, we shall explore the limitations and challenges to be faced in purposeful cultivation and curation of corpus.

References

Chin, Andy Chi-on. (1998). A quantitative and qualitative analysis of words in Chinese news headlines. In B. Tsou, T. Lai, S. Chan, W. S-Y. Wang (Eds.), *Quantitative and Computational Studies on the Chinese Language*, pp.235-252. Hong Kong: Language Information Sciences Research Centre, City University of Hong Kong.

Chin, Andy Chi-on. (2019). Initiatives of digital humanities in Cantonese studies: A corpus of mid-twentieth-century Hong Kong Cantonese. In A. T. W. Bo (Ed.), *Digital Humanities and New Ways of Teaching*, pp. 71-88. Singapore: Springer.

Chin, Andy Chi-on. (2020). What can the corpus of mid-20th century Hong Kong Cantonese tell us about Hong Kong society of half a century ago? In B. Basciano, F. Gatti, & A. Morbiato (Eds.), *Corpus-Based Research on Chinese Language and Linguistics*, pp. 243-262. Venezia: Edizioni Ca' Foscari - Digital Publishing.

Tsou, Benjamin, Lai, Tom, Chan, Samuel and Wang, William S.-Y. (Eds.) (1998). *Quantitative and Computational Studies on the Chinese Language* 《漢語計量與計研究》. Hong Kong: Language Information Sciences Research Center, City University of Hong Kong.

鄒嘉彥, 黎邦洋. (2003). "漢語共時語料庫與信息開發" (Chinese synchronous corpus and data mining). In *Critical Issues in Chinese Information Processing* 《中文資訊處理干重要問題》"973 計劃國家語言自然語言理解與知識挖掘", 徐波 孫茂松 靳光主編, pp.147-165. 北京: 科學出版社.

Tsou, Benjamin K. & Kwong, Olivia. (2015). LIVAC as a monitoring corpus for tracking trends beyond linguistics. Linguistic Corpus and Corpus Linguistics in the Chinese Context. *Journal of Chinese Linguistics* Monograph Series 25, 447-471.

Tsou, Benjamin K. (2019). Sociolinguistic aspects of the Chinese language. (commissioned work) In *Oxford Bibliographies in Chinese Studies*, Oxford University Press.

Tsou, Benjamin K. and Yip, Ka Fai. (2020). "A corpus-based comparative study of light verbs in three Chinese speech communities" selected papers from 34th PACLIC Workshop on MWEA, 2020 Oct., Hanoi, Vietnam.

Tsou, Benjamin K. (2020). "From "Ding^{1}Dong^{1}" to "Ding^{4}Ling^{1}Dong^{4}Long^{4}": Reflections on Yue – Cantonese Quadra-syllabic Expressions". In *Current Research in Chinese Linguistics*: *Selected paper from 22nd International Conference on Yue Dialects Hong Kong* (December 2017), *99*(1), 11-20.

Tsou, Benjamin K. (2021). Some reflections on developments in linguistics in Hong Kong on the occasion the 35th Anniversary of LSHK. In commemorative volume on the 35th Anniversary of LSHK.

Yip, Ka Fai, Tsou, Benjamin and Ji, Yaxuan. (2020). "漢語動詞虛化初探：港澳京三地同中之異" Annual Research Forum, The Linguistic Society of Hong Kong.

Yiu, Carine. (2021). The origin and development of the question particle *me1* 咩 in Cantonese. *Lingua*, 254. https://doi.org/10.1016/j.lingua.2021.103049.

Session 2

Linguistic Issues in Constructing Cantonese Corpora

Discussants: Andy CHIN
Thomas Hun-Tak LEE
LUKE Kang Kwong
Carine YIU
Moderator: CHEUNG Hin Tat

Part 1 Each discussant has 15 -20 minutes to summarize their corpus construction with respect to: (a) part of speech tag sets, (b) colloquial morpheme representations and (c) other challenging issues related to corpus construction

Part 2 Open Discussion. Discussants will be invited to address some open-ended questions, as well as questions raised from the floor. Below are some of the open-ended questions:

1. Will it be possible to have a unified POS tag set for the future development in Cantonese corpora? What are the alternatives to a unified POS tag set?
2. For the variations in Cantonese orthography in the existing corpora, what can be done to provide effective text search? How can we measure/evaluate the reception of particular orthographic choices from users?
3. What strategies are found to be effective in coping with the linguistic issues when constructing your corpus? What should be avoided?
4. Any recommendation for specific Cantonese corpus to be developed in the future?
5. What research questions can be further explored by using Cantonese corpora?

The Corpus of Mid-20th Century Hong Kong Cantonese 二十世紀中期香港粵語語料庫

Andy CHIN (The Education University of Hong Kong)

The Corpus of Mid-20th Century Hong Kong Cantonese developed at The Education University of Hong Kong in 2012, with the support of RGC's Early Career Scheme and EdUHK's internal research grant, provides a snapshot of the Cantonese language spoken in Hong Kong half a century ago. The data of the corpus was collected by transcribing the speech dialogues in 80 black-and-white movies produced between 1940 and 1970. There are two phases of the corpus. The two phases of the corpus have accumulated around 800,000 character tokens. Besides diachronic studies of the Cantonese language, the corpus data is also useful for examining the socio-cultural issues of early Hong Kong.

URL: https://hkcc.eduhk.hk/

Publications:

錢志安. (2013). 粵語研究新資源：《香港二十世紀中期粵語語料庫》. 《中國語文通訊》, 92(1):7-16.

Chin, Andy Chi-on. (2019). Initiatives of digital humanities in Cantonese studies: A corpus of mid-twentieth-century Hong Kong Cantonese. In A. T. W. Bo (Ed.), *Digital humanities and new ways of teaching* (pp. 71-88). Singapore: Springer.

Chin, Andy Chi-on. (2020). What can the corpus of mid-20th century Hong Kong Cantonese tell us about Hong Kong society of half a century ago? In B. Basciano, F. Gatti, & A. Morbiato (Eds.), *Corpus-based research on Chinese language and linguistics* (pp. 243-262). Venezia: Edizioni Ca' Foscari - Digital Publishing.

Chin, Andy Chi-on. (forthcoming). A corpus-based approach to learning and teaching Cantonese. In C. S.-l. Lee (Ed.), *Cantonese as a second language: Learning needs, teaching methodology and curriculum design*. London: Routledge.

黎奕葆, 錢志安. (2018). 粵語的動詞後綴"着". 收錄於何大安, 姚玉敏, 孫景濤, 陳忠敏, 張洪年編, 《漢語與漢藏語前沿研究—— 丁邦新先生八秩壽慶論文集》, 頁 697-710. 北京: 社會科學文獻出版社.

Lai, Y. P. (2020). Multiple functions of the Cantonese 'wait' verb *dang2* and their historical development. *Studies in Language*, *44*(4): 917-963.

Hong Kong Cantonese Child Language Corpus (CANCORP)

Thomas Hun-Tak Lee (Tianjin Normal University & The Chinese University of Hong Kong)

CANCORP, jointly developed by The Chinese University of Hong Kong, The Hong Kong Polytechnic University, and The University of Hong Kong in 1996, with the support of RGC's earmarked grant, is the most important resources for studying the development of grammatical competence in Cantonese-speaking children. The data of the corpus was collected by transcribing a set of audio recordings of conversational exchanges between child subjects and adults. Eight Cantonese-speaking children were each observed for around 12 months. The beginning age of observation was between 1;07 and 2;08. The mean number of observation sessions for each child, with each session resulting in a one-hour audio recording, was 21.

URL: http://www.arts.cuhk.edu.hk/~lal/corpora.html#CANCORP

Publications:

Lee, H. T. (2000). Finiteness and null arguments in Child Cantonese. *The Tsinghua Journal of Chinese Studies*, New Series 30:365-393.

李行德. (2009). 粵語兒童對粵語結構分析的啟示。收錄於錢志安、郭必之、李寶倫、鄒嘉彥編《粵語跨學科研究：第十三屆國際粵方言研討會論文集》，1-21 頁。香港： 香港城市大學語言資訊科學研究中心。

Lee, H. T., & Law, A. (2001). Epistemic modality and the acquisition of Cantonese final particles. In Issues in East Asian language acquisition, ed. Mineharu Nakayama, 67-128. Tokyo: Kuroshio Publishers.

Lee, H. T., & Wong, C. (1998). CANCORP: The Hong Kong Cantonese Child Language Corpus. *Cahiers de Linguistique – Asie Orientale*, *27*(2): 211-228.

Hong Kong Cantonese Corpus (HKCanCor) 香港粵語語料庫

Luke Kang Kwong (Nanyang Technological University)

HKCanCor, developed at The University of Hong Kong, with the support of RGC Grant, is one of the earliest and most useful corpora for data-driven linguistic analysis of contemporary Hong Kong Cantonese. The data of HKCanCor was collected by transcribing audio recordings of spontaneous conversations (involving 2 to 4 speakers), radio programmes, and a monologue, all of which were recorded between 1997 and 1998. About 230,000 words were collected.

URL: http://compling.hss.ntu.edu.sg/hkcancor/

Publications:

陸鏡光. (2007). 粵語句末助詞的書寫方式. In Joanna Ut-Soeng Sio, & Sze-Wing Tang (Eds.), Studies in Cantonese Linguistics 2 (pp. 95-107). Hong Kong: Linguistics Society of Hong Kong.

Luke, K. K., & Wong, M. L.Y. (2015). The Hong Kong Cantonese Corpus: Design and uses. In B. K. Tsou & O. Y. Kwong (Eds.), *Linguistic Corpus and Corpus Linguistics in the Chinese Context* (pp. 312-333). Journal of Chinese Linguistics Monograph Series, no. 25. Hong Kong: The Chinese University Press.

Wong, P. W. (2006). The specification of POS tagging of the Hong Kong University Cantonese Corpus. *International Journal of Technology and Human Interaction* 2 (1): 21-38.

The Early Cantonese Tagged Database 早期粵語標註語料庫

Carine YIU (The Hong Kong University of Science and Technology)

The Early Cantonese Tagged Database developed at The Hong Kong University of Science and Technology in 2012, with the support of RGC's General Research Fund, provides a window into the Cantonese language spoken between 1870s and 1930s. The database includes ten Cantonese colloquial texts, from the Cantonese translation of the *St. Mark's Gospel* (1872) to A. Fulton's *Progressive and Idiomatic Sentences in Cantonese Colloquial* (1931), with around 160,000 character tokens. It was the first corpus that provides diachronic data of Cantonese with syntactic tagging.

URL: http://database.shss.ust.hk/Cantag/

Publications:

姚玉敏. (2010)〈早期粵語中的"畀"字句〉[Bei-sentences in early Cantonese].《歷時演變與語言接觸—中國東南方言》*Diachronic Change and Language Contact – Dialects in South East China*.《中國語言學報專刊 24》*Journal of Chinese Linguistics*, monograph series (The Chinese University of Hong Kong): no.24, pp. 162-185.

Yiu, Y. M. C. (2013). Directional verbs in Cantonese: A typological and historical study. *Language and Linguistics*, *14*(3): 511-569. (Academia Sinica, Taipei).

Yiu, Y. M. C. (2014). Typology of word order in Chinese dialects: Revisiting the classification of Min. *Language and Linguistics*, *15*(4): 539-573. (Academia Sinica, Taipei).

Yiu, Y. M. C. (2014). *The typology of motion events: An empirical study of Chinese dialects*. Berlin: De Gruyter Mouton.

Session 3

Digital Processing of Cantonese Corpus Data

Charles LAM, LAU Chaak Ming, Jackson LEE

There is an abundance of Cantonese language resources on the internet that can be tapped into with the help of appropriate programmatic tools. This session will explore how PyCantonese, a free Python library, can be used along with text corpora to perform both theoretical and applied linguistic research tasks. The three speakers will walk you through simple Python code on the Google Colab platform. Programmers and absolute beginners are welcome.

Session 4

Demonstration of Cantonese Digital Resources and Tools and Cantonese Learning Apps

(1) Demonstration of Cantonese Digital Resources and Tools

(a) Cantonese Wordnet: Luis Morgado da COSTA, Joanna Ut-Seong SIO

The Cantonese Wordnet is a lexical resource for Cantonese (based on Hong Kong Cantonese) that organizes words in a semantic network. Words are connected using relations such as synonymy, hypernymy and hyponymy. It is built pivoted on the Princeton WordNet (Fellbaum 1998), following many of the assumptions of the Chinese Open Wordnet (Wang and Bond, 2013). It is also linked to the Collaborative Interlingual Index (Bond et al., 2016) – an open, language agnostic, flat-structured index that links wordnets across languages. In this talk, we will discuss Cantonese-specific issues in the development of this wordnet and some of its potential applications.

(b) The Early Cantonese Bible Database: Shin KATAOKA

Cantonese Bible was first produced in the mid-nineteenth century as a by-product of Western missionaries' mission in Guangdong and Hong Kong. Launched in 2020, the Early Cantonese Bible Database provides Internet users with the following three functions: 1. The brief history of Cantonese Bible translation; 2. The information of missionaries who were involved in the production of Cantonese Bible; and 3. The corpus of the Gospel of Luke in six versions published between 1867 to 1927. Using some examples from the corpus, I will talk about how the information in the database can be utilized in the diachronic study of Cantonese or the study of Bible translation in China.

(c) Rime-Cantonese and Inject-Jyutping: LAU Ming Fei

Rime-Cantonese is the state-of-the-art Cantonese input method lexicon. It is a normalized lexicon. By "normalized", we mean both phonetically and orthographically, which is:

1. Written Cantonese has no standardized orthography, and written Cantonese in real world are highly irregular. To tackle this issue, we devise a recommended orthography for written Cantonese, including a set of character choices for Cantonese sentence final particles. This lexicon strictly follows this standard in its character.

2. Cantonese use Chinese characters in its writing system, while one character may have multiple variants. We normalize the character variants in the lexicon with a standard character set.

3. Most extant lexicons do not contrast the ng/0 initials and often collects both pronunciations for characters. We keep the contrast to reduce redundancy and ambiguities.

Inject-Jyutping is a browser extension (supports Chrome, Edge and Firefox) based on rime-cantonese. It automatically annotates Jyutping above Chinese characters in web pages, turning any web into a live Cantonese textbook. It is a powerful tool for any Cantonese learners.

(d) The 19th Century (1865-1894) Cantonese Christian Writings Database: LEUNG Wai Mun

"Database of the 19th Century (1865-1894) Cantonese Christian Writings"（十九世紀中後期（1865-1894）粵語基督教典籍資料庫）developed at The Hong Kong Polytechnic University in 2020, with the support of The Lord Wilson Heritage Trust and PolyU's departmental research grant, provides a snapshot of the Cantonese Christian writings in late Qing dynasty within a 30-year period. The database contains 15 books produced by the early western Christian missionaries between 1865 and 1894, and has accumulated around 466,000 character tokens. Features of the database include images of the book contents and text comparison between the Cantonese translation of the biblical books and the corresponding Mandarin (1919) and modern Cantonese (2010) translations. It is believed that the database will be useful to diachronic studies of Cantonese as well as religious historical and translation studies.

(e) words.hk《粵典》: Raymond TSE

https://words.hk/, founded in 2014, is a project to build a crowdsourced, descriptivist Cantonese dictionary with complete explanations and illustrative example sentences in Cantonese and English, in order to document the actual contemporary state of the Cantonese language in Hong Kong. In addition to usage that is accepted by mainstream Cantonese users (e.g. characters and pronunciations), we also document those that are used by a substantial minority.

Currently, at least 10,000 out of 50,000 word entries have been already published. In addition, some of our data that may be useful for developing input methods, natural language processing, etc. is released in the Public Domain.

(f) Early Cantonese Colloquial Texts: A Database, Early Cantonese Tagged Database, Database of Early Chinese Dialects: Carine YIU

The Database of Early Chinese Dialects was developed as a by-product of Carine Yiu's GRF project "Reconstructing the history of Chinese dialectal grammar: A study of word order". It includes twenty-one texts (more than 418,000 Chinese characters), covering Cantonese, Hakka, Mandarin, Min and Wu and including both Bible translation and textbooks. The texts, searchable according to specified Chinese characters or chapters and verses in the Gospel of Mark, render the database an indispensable tool for research into the grammar of early Chinese dialects. The Database of Early Chinese Dialects and two other databases on early Cantonese (Early Cantonese Colloquial Texts: A Database and Early Cantonese Tagged Database) are hosted by the Center for Chinese Linguistics, The Hong Kong University of Science and Technology http://ccl.ust.hk/ (under 'Useful Resources').

(2) Demonstration of Cantonese Learning Apps

(a) CanTONEse: KI Mei Ying, LAI Yik Po, YIP Ka Fai

Tone is arguably the hardest area for self-learners of Cantonese, especially if one's native language is not tonal. In order to help non-Cantonese native speaking students to master this important and significant aspect of Cantonese sound system, a mobile app, **CanTONEse**, was designed to provide them with a self-learning tool for Cantonese tones. Teaching materials with recordings, self-assessment exercises, an interactive game and daily conversations are available in the app. In order to turn abstract pitch difference of tones into a concrete and intuitive representation, tones are displayed via different means in addition to numerals (1-6). Throughout the app, tones are (1) colour-coded according to sound-colour synaesthesia (Ward, Huckstep & Tsakanikos 2006), (2) represented by movements of animals according to pitch heights and contours, and (3) drawn on a music score sheet.

(b)《成語填字遊戲》: Benjamin T'SOU

Quadra-syllabic Idiomatic Expressions (QIEs) are emblematic of the Chinese language and many other languages in and adjacent to China, including genetically unrelated ones such as Japanese and Korean. Members from all walks of life are fond of using them to showcase verbal flair in personal communications. As a popular form of speech art, it shares the foundations of poetic forms, which according to Jacobson, embody both the poetry of grammar and the grammar of poetry.

A major function of QIEs is to provide semantic argumentation through the projection of metaphorical and figurative meaning. They include devices such as the insertion and transposition of linguistic elements, onomatopoetic and semantic expansion, and dialectal variations. For example, Mandarin "pipa" 噼啪/ "pilipala" 噼里啪啦; Cantonese "king^{4}kang^{4}" / "king^{4}ling^{1}kang^{4}lang^{4}", "養妻活兒", and "早行晚拆".

Given their systematic and quadra-syllabic format, QIEs lend themselves readily to crossword puzzle formulations. A typical example is played with hints for the player to progressively fill the empty squares. Linguistic skills are needed which involve appreciation of structural parallelism, analogy, and metaphorical extrapolation. The relative ease to play the game could vary according to the QIE's structure and the language skills of the players. Evaluative feedback is given after each game as well as access to relevant etymology, synonymous and antonymous QIEs, and actual QIE usage examples.

A large set of such crossword puzzle games have been created from about 20000 such examples taken from LIVAC (https://en.wikipedia.org/wiki/LIVAC_Synchronous_Corpus). They have been already accessed by about 700,000 players in and outside Mainland China.

The demonstration will provide actual games for participants to have some hands-on experience.

(c)《古詩粵唱粵啱 Key》: ZHANG Ling

This mobile APP is a product of knowledge transfer from Chinese language and linguistics to educational applications. The knowledge basis involves two points. Firstly, classical Chinese poems were probably in a singing style as their oral form in ancient time, and this APP revives this tradition, which can enhance memorization. Secondly, Cantonese tones have a declination trend in speaking and non-declination requirement in singing. The reverse direction application of this linguistic theory can help to turn speaking utterances into the singing style. There are three modes in the APP: (1) Demo of poems in Cantonese singing; (2) Karaoke; (3) Creative composing. For Modes (2) and (3), users can audio-record their own singing. They can review, upload, and share their works with their peers and teachers, and they can appreciate each other's works and give "likes". The design of these modes can cater for needs at different levels and make the APP more interesting and interactive.

香港教育大學
The Education University of Hong Kong

ONLINE

School of Cantonese Studies 2021

粵語研究研習班

15-16 May 2021

CANTONESE STUDIES IN THE DIGITAL AGE

Topics

- Frontiers in Cantonese corpus-based research
- Linguistic issues in constructing Cantonese corpora
- Digital processing of Cantonese corpus data
- Cantonese digital resources and tools

SPEAKERS

- 張顯達 CHEUNG Hin Tat (The Education University of Hong Kong)
- 錢志安 Andy CHIN (The Education University of Hong Kong)
- Luis Morgado da COSTA (Nanyang Technological University)
- 片岡新 Shin KATAOKA (The Education University of Hong Kong)
- 祁美瑩 KI Mei Ying (The Chinese University of Hong Kong)
- 黎奕葆 LAI Yik Po (The Education University of Hong Kong)
- 林子鈞 Charles LAM (The Hang Seng University of Hong Kong)
- 劉擇明 LAU Chaak Ming (The Education University of Hong Kong)
- 劉銘霏 LAU Ming Fei (Cantonese Computational Linguistics Infrastructure Development Group)
- 李行德 LEE Hun-Tak Thomas (Tianjin Normal University and The Chinese University of Hong Kong)
- 李　麟 Jackson LEE (Author of PyCantonese)
- 梁慧敏 LEUNG Wai Mun (The Hong Kong Polytechnic University)
- 陸鏡光 LUKE Kang Kwong (Nanyang Technological University)
- 蕭月嫦 Joanna Ut-Seong SIO (Palacký University Olomouc)
- 謝家尉 Raymond TSE (粵典words.hk)
- 鄒嘉彥 Benjamin T'SOU (City University of Hong Kong)
- 葉家煇 YIP Ka Fai (Yale University)
- 姚玉敏 Carine YIU (The Hong Kong University of Science and Technology)
- 張　凌 ZHANG Ling (The Education University of Hong Kong)

scs2021@eduhk.hk　　www.eduhk.hk/lml/scs2021/

Co-organised by:
語言學及現代語言系 LINGUISTICS AND MODERN LANGUAGE STUDIES
Centre for Research on Linguistics and Language Studies

Sponsored by:
香港語言學學會 The Linguistic Society of Hong Kong

Department of Linguistics and Modern Language Studies

https://www.eduhk.hk/lml
lml@eduhk.hk

Centre for Research on Linguistics and Language Studies

https://www.eduhk.hk/crlls
crlls@eduhk.hk