Research Assessment Exercise 2020
Impact Case Study

University: The Education University of Hong Kong
Unit of Assessment (UoA): 33 Linguistics & language studies

Title of case study: Corpus-based Studies of Cantonese

(1) Summary of the impact

The UoA identified *Corpus-based Studies of Cantonese* as the impact case for this submission. The research project was led by Dr Andy Chin who constructed a Cantonese corpus (in two phases) with about 1,000,000 Chinese characters. The corpus offers an alternative perspective for studying Cantonese with both quantitative and qualitative data which traditional dialectological studies of Cantonese cannot provide. The corpus offers objective and authentic data for the design and compilation of teaching and learning materials for Cantonese. Positive feedback on the use of the corpus data for teaching and learning Cantonese was received from some Cantonese learners in Hong Kong and instructors in Canada and Mainland China.

(2) Underpinning research

Hong Kong is a multilingual society (specifically 兩文三語) where Cantonese is an important language on a par with Chinese/Putonghua and English. Nearly 90% of the population in Hong Kong speak Cantonese as their home language or mother tongue, according to the latest census data. Many non-local people living and working in Hong Kong are eager to learn and speak Cantonese in order to integrate themselves into the community. Furthermore, Cantonese is also used in many formal situations, such as the Government, Legislative Council, classrooms as the medium of instruction, media, etc. It has close contact and extensive interaction with other languages and thus gives rise to many interesting linguistic phenomena. In spite of its important status, Cantonese has never been formalised and implemented in the language curriculum (primary and secondary schools). The language has not been standardised in terms of its grammar, pronunciations, and writing system. The learning and teaching materials of Cantonese available in the market thus vary in terms of content, and the teaching methods are always subject to the preference and choice of the compilers and instructors. The UoA thus argues that a more scientific and objective approach is needed for promoting Cantonese studies, including learning and teaching of Cantonese.

Studies of Cantonese flourished in the 1970s and over the past four decades, many works such as fieldwork reports and reference grammars on Cantonese have been published. However, most of these studies were undertaken through fieldworks by interviewing senior informants. These works undoubtedly have enriched our understanding of Cantonese such as its lexicon, phonology and grammar. Some deeper issues such as pragmatics, semantics and discourse however remain to be explored and this kind of research requires a significant amount of authentic and natural language data. In this regard, a Cantonese corpus was proposed to expand the scope of Cantonese linguistic research. Another major advantage of a Cantonese corpus is the provision of both quantitative and qualitative data which are objective and unbiased for research and also other applications such as compilation of language materials, and natural language processing such speech-to-text/text-to-speech algorithms. In this regard, the UoA proposed a research programme entitled *Corpus-based Studies of Cantonese* through the construction of a Cantonese corpus.

The research programme started in 2011 and was supported by a number of internal research grants from the UoA's university plus a prestigious competitive research grant (The Early Career Scheme)

from the Research Grants Council (Project title*: Linguistic Analysis of Mid-20th Century Hong Kong Cantonese by Constructing an Annotated Spoken Corpus* (Project No.: 859713). [1]
Cantonese is a spoken language, and the corpus data are thus in the oral form. The data of the corpus was collected by transcribing the dialogues of 60 black-and-white movies produced between the 1950s and 1970s. The corpus has a size of about 1,000,000 Chinese characters (around 200,000 in the first phase and 800,000 in the second phase). The construction of the corpus also touches upon some challenging tasks in natural language processing in Chinese (as well as Cantonese) such as word segmentation and POS tagging. The tagging provides more precise information on words, especially those that have multiple meanings (such as 過 as an aspect marker, a verb and a preposition in the comparative construction). The transcription work was done on the ELAN platform developed by the Max Planck Institute for Psycholinguistics so that the search results can be linked with the corresponding video segments. The purpose of this linking is to obtain extra-linguistic information such as physical settings, facial expressions, and gestures of speakers when making utterances. In other words, users of the corpus do not only get "what was said in the utterance", but also "how the utterance was made". This feature is not found in other existing Cantonese corpora, and is particularly important for learners of Cantonese. In addition, the corpus is beneficial for the future development of Cantonese studies under the theme of Digital Humanities which draws upon digital technology and big data.

(3) References to the research

1. 錢志安. (2011). 粵語語法的多角度研究. 《中國語文研究》, 31/32:33-43. In this paper, the author argues for a multi-dimensional perspective to study Cantonese grammar in addition to the traditional dialectological approach.
2. Chin, Andy Chi-on. (2011). Grammaticalization of the Cantonese Double Object Verb [pei^{35}] 畀 in Typological and Areal Perspectives. Language and Linguistics (語言暨語言學), 12(3):529-563. This paper surveyed *a corpus of early textual materials* of Cantonese (including some 19th century missionary texts) and discussed the grammaticalization of the Cantonese double-object verb.
3. 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. This paper introduces the Cantonese corpus described in this impact case statement.
4. 錢志安. (2017). 粵語(四字)歇後語的修辭功能. 《粵語研究》(增刊) - 中國南方語言四音節慣用語研究, 125-134. This paper studies the *xie-hou-yu* found in the corpus, and its rhetorical function.
5. 黎奕葆, 錢志安. (2018). 粵語的動詞後綴"着". 輯錄於何大安、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. This paper studies the verbal suffix 着 found in the Cantonese corpus. This suffix was seldom discussed in Cantonese grammar due to its infrequent usage in contemporary Cantonese.
6. Chin, Andy Chi-on. (2019). Initiatives of digital humanities in Cantonese studies: A corpus of mid-twentieth-century Hong Kong Cantonese. Edited by Anna Tso Wing-Bo, *Digital humanities and new ways of teaching* (71-88). Singapore: Springer. This paper argues how the Cantonese corpus can benefit research in Digital Humanities.

(4) Details of the impact

There are a number of ways the UoA used to document the impact of the case. The corpus was available online for free. According to our record, there are nearly 3000 registered users of the first

[1] URL of the corpus: http://hkcc.eduhk.hk. Username: demo. Password: hkccdemo19.

phase of the corpus. For the second phase of the corpus, two workshops were organised to engage the potential users and share with them how the corpus can enhance teaching and learning of Cantonese. Follow-up surveys were conducted with the participants of the workshops. 26 participants provided feedback on the survey. 22 of them found the corpus having ***significant or some changes*** on their previous views or ways of teaching/learning/studying Cantonese. Some of these changes are listed below:

(1) Learning a new language with movies or real language data is more beneficial.
(2) A new resource connected to media that is of interest to me, targeted and allowing me to already access movies and actors to learn more from, that users can play at their own pace.
(3) The corpus provides data such as word association, frequencies which are useful in learning Cantonese
(4) The inclusion of video clips in the corpus can allow users to study more effectively
(5) I'm already tired of traditional learning resources. This database changes my previous views on what is POSSIBLE in teaching / learning / studying Cantonese.
(6) The quantitative data (such as frequencies) can help learners and teachers to select appropriate materials.
(7) The inclusion of video clips opens up opportunities to analyze gestures and speakers' tones, which is not available in existing text-based corpora.
(8) It inspires me to incorporate corpus in language teaching and curriculum design

We also invited some overseas instructors of Cantonese to give us feedback on the corpus and how it benefits to their teaching of Cantonese, and to what extent it enriches their students' experience in learning Cantonese. Please see the last item in Section 5 below.

The UoA organized a professional workshop - School of Cantonese Studies - in May 2019 with an aim to introduce the public the knowledge of Cantonese. The topics were not limited to linguistics, but also culture, history and ethnography. The School attracted about 60 participants from different parts of the world, including Hong Kong, Mainland China, Taiwan, France, the United Kingdom, the USA, Singapore, Malaysia and Russia. Besides undergraduate and postgraduate students, there were general public who took this opportunity to deepen their knowledge about Cantonese. The Cantonese corpus was introduced under the theme "Online resources for Cantonese studies".

Other impact generated from this case include media interviews with the PI, Dr Andy Chin. Over the past few years, Dr Chin was invited by the media to share his research experience on Cantonese and in particular to rationales of constructing the corpus. In his interviews in April 2019, the focus was on the potential application value of the corpus, i.e. facilitating the learning and teaching of Cantonese. He is currently working with the Communications Office of the University to launch an online platform (Instagram, Facebook, YouTube) to share with the public some interesting linguistic issues on Cantonese (observed from the corpus) and promote how Cantonese can be studied in a systematic and rigorous way.

In the summer of 2019, Dr Andy Chin (as the PI) secured a project from the Standing Committee on Language Education and Research (SCOLAR) to develop a self-learning online platform for non-Chinese speakers to learn Cantonese (as well as Chinese). The project draws on the PI and his team's prior experience and expertise in Cantonese studies and corpus linguistic research. According to the by-census data of the Hong Kong Government (2016), "64.3% and 52.9% of ethnic minorities aged 5-14 and 15-24 were able to read Chinese respectively".[2] It is anticipated that the output of the project (i.e. the online platform) can help improve the Chinese proficiency of the ethnic minorities living in Hong Kong.

[2] https://www.bycensus2016.gov.hk/data/snapshotPDF/Snapshot10.pdf, page 5.

(5) Sources to corroborate the impact

The impact of the case submitted for this exercise can be found in the following. Please refer to the webpage http://hkcc.eduhk.hk/impactcase for the details:

(a) Media interviews on the corpus and its applications:
1. 星島日報 (Singtao Daily), 2013.6.13, 《教院語料庫 集粵語長片對白》(A corpus of early Cantonese movie dialogues) [LINK]
2. 灼見名家 (Master Insights), 2017.5.12 《教大學者醉心粵語研究　以懷舊電影創建網上語料庫》(An interview by Master Insights on the development of an online corpus with old Cantonese movies) [LINK]
3.《粵講粵爛鬼3》, 2017.8.19. [a TV programme on early Hong Kong] [LINK]
4.《粵講粵爛鬼3》, 2017.8.26. [a TV programme on early Hong Kong] [LINK]
The following is a series of interviews on the application of the Cantonese corpus data in teaching and learning Cantonese:
5. 教大建語料庫助外國人趣學廣東話, 2019.4.11 [LINK]
6. 對唔住呀 對唔住啡意思大不同 教大新版語料庫助學粵語, 2019.4.11 [LINK]
7. 老片教粵語 老外都讚好, 2019.4.11 [LINK]
8. 教大語庫輯70粵語片 學廣東話神髓, 2019.4.11 [LINK]
9. 輸入70粵語長片百萬字對白 教大粵語庫增播片學例句 [LINK]
10. 教大學者夥「粵語迷」澳洲籍助理創語料庫　助外籍人士學廣東話, 2019.4.11 [LINK]
11. 教大推新版粵語語料庫 助用家「趣學廣東話」, 2019.4.29 [LINK]

(b) 12. The Cantonese Corpus won a gold medal and a special award for its *educational technology innovation* in the Silicon Valley International Festival, June 2019

(c) 13. Testimonials from instructors of Cantonese. Instructors of Cantonese based in Mainland China and Canada provide their comments on the use of the Cantonese corpus in their teaching.