香港成人粵語口語語料庫

馮淑儀[1]、羅心寶[2]

香港理工大學[1]、香港大學[2]

1. 構建緣起

《香港成人粵語口語語料庫（Hong Kong Cantonese Adult Language Corpus）》[1]是一個在上世紀九十年末構建的自然語言語料庫（見 Leung and Law 2001 及 Law, Fung and Leung 2004）。當時的粵語語料庫只有 CANCORP 及 Fletcher, Leung, Stokes and Weizman（2000）這兩個以收錄香港兒童粵語為主的語料庫，粵語語料庫的建設工作相對薄弱。這是由於當代粵語並沒有大量的現成文本可供機器自動閱讀和處理，構建粵語口語語料庫就只得全人手製作，是一項耗時耗力的工作。可是，語言研究工作必須建基於大量的語言事實，而語料庫的建設則大大節省了語言學者重複蒐集語料的時間和人力。有見及此，我們構建了這個一共收錄了八個多小時，約十七萬字的成人粵語自然語言語料庫。我們着力為所蒐集到的語料提供較細緻的文字和語音轉寫，期望它可以成為研究語法、語音和話語的學者的可靠參考資料。十年過去了，我們樂於看見各種粵語語料庫相繼建成。不過，我們相信《香港成人粵語口語語料庫》的獨特性使它仍然有相當的使用價值。

2. 語料性質

為了比較有效地呈現當代香港粵語口語的真實使用面貌，本語料庫徵得香港電台同意，採用它們在一九九八年十一月至二零零零年二月期間製作的七個廣播節目作為語料庫的主要內容。這些節目都大多是無預設文本，並以即時對話的形式進行，頗能忠實地呈現當代粵語的語音、語法、詞匯和語用面貌。這七個廣播節分屬論壇和峰煙（phone-in）節目兩大類型。論壇收錄了“政黨論壇”；“特區年代財經學人”兩個節目。內容主要是由節目主持人邀請社會上的知名人士就某一時事或財經議題作出即場的對談和辯論。峰煙節目則收錄了“平息你的風波”；“有冇心情顏聯武”；“星空奇遇鐵達尼”；“海琪的天空”；“恐怖熱線”等五個節目。每個節目的話題和內容

[1] 語料庫獲研資局撥款予羅心寶、馮淑儀和梁文德共同開發（#HKU5190/98H）。羅心寶、馮淑儀負責構建語料庫和核實所有文字及語音轉寫，梁文德則負責設計和編寫檢索系統。

各異，其中包括了政治時事、經濟民生、家庭生活、人際關係、兩性相處，以致靈界故事。聽眾致電電台就某一話題自由地發表意見，或向主持人抒發個人生活感受。主持人因應聽眾的內容作互動交流。這些節目因而涵蓋了多種不同的語體風格，既包括了較正規的高層粵語，也包括了較口語的低層粵語。說話人包括了不同性別、年齡、職業和文化階層人士。除了節目主持人以外，語料庫一共收錄了六十九位不同聽眾的語料。 由於論壇的參與者都是社會上的知名人士，他們的年齡和語言背境比較容易翻查。至於峰煙節目部分，雖然我們沒法取得這些聽眾的個人資料，但我們根據談話內容，估計他們分別來自中、青、幼三代。致電“有冇心情顏聯武”、“星空奇遇鐵達尼”以及“海琪的天空”三個節目的多數是在學青少年，談話內容往往涉及他們的感情煩惱和考試壓力。致電“平息你的風波”和“恐怖熱線”的，大多是青年至中年人。

3. 組成架構

這個語料庫由四個部分組成：語音庫、文本庫、標音庫和檢索系統。

3.1. 語音庫

語音庫收錄了這七個節目的 WAV 格式音檔。每個節目的時長由最短的四十七分鐘到最長的一百分鐘不等。八個節目合共約提供八個多小時的錄音。這些廣播錄音真實地記錄了不同人士的不同粵語發音，包括較正規的標準發音、較新派的發音（包括被貶稱為懶音的發音）、語誤，和各種共時音變。用家可以直接使用這些音檔作進一步的聲學分析。

3.2. 文字庫

由於這些廣播節目都沒有文本，我們為這些錄音提供了約共十七萬字的文字轉寫。我們參考了 Du Bois, Schuetze-Cumming, and Paolino（1993）的轉寫標記法來進行轉寫。例如：

表一：話語轉寫符號舉例

符號	意義
...	停頓
[重疊發言]	兩個說話人之間的重疊發言
[[重疊發言]]	三個或以上說話人之間的重疊發言
X	無法分辨的發言
＜不肯定＞	無法確定的發言
@	非語言動作，如笑聲、咳嗽、呵欠聲等等

以下是“特區年代財經學人”的部分文字庫和語音庫的轉寫（第一欄的 M 代表男性；第二欄的 G 代表嘉賓，H 代表主持人。）：

M	G	仲	有	香	港	係	一	個	好	細	嘅	地	方
M	G	tsoŋ˩	jɐu˨	hœŋ˥	kɔŋ˧	hɐi˩	ɐk˥	kɔ˩	hou˧	sɐi˩	kɛ˩	tei˩	fɔŋ˥
M	G	之	但	係	麻	雀	雖	小	五	臟	俱	全	
M	G	tsʰi˥	tan˩	ɐi˩	ma˨	tsœk˩	søy˥	siu˧	m˨	tsɔŋ˩	kʰøy˥	tsʰyn˨	
M	G	各	式	各	樣	金	融	業	[	有	]		
M	G	kɔk˩	sik˥	kɔk˩	jœŋ˩	kɐm˥	joŋ˨	jip˩	[	jɐu˨	]		
M	H	[	唔	]									
M	H	[	m̩˩	]									

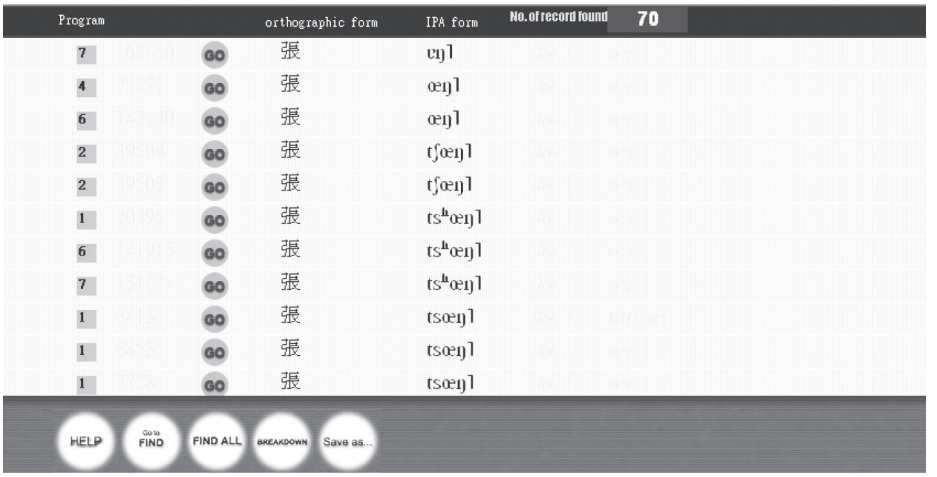
3.3. 標音庫

標音庫是構建整個語料庫最費力的部分。我們不採用較省力的音位標注法，而是使用國際音標，頗仔細地描寫了每個說話人的實際發音。 標音庫記錄了每個音節的各個語音變體、語誤，以及各種語流音變，如減音、合音、弱化、同化作用、異化作用等現象。例如：“張”一詞在語料庫內就有以下五種變體：[tsœŋ˥] [tʃœŋ˥] 、[tsʰœŋ˥] 、[œŋ˥] 和 [ɐŋ˥] 。[tsʰœŋ˥] 可以被判斷為一個語誤；[œŋ˥] 是出現在“收到呢張”的句子中；而 [ɐŋ˥] 則出現在“咁我哋而家因其中一張……”的句子中。另外，減音和合音現象在語料庫中也很常見，如“係”[hɐi˩] 一詞的聲母在實際話語中往往被刪去，發為 [ɐi˩]；“會唔會”[wui˨ m̩˩ wui˨] 在某些語境會合音為 [wui˨ mui˨]；“即係”[tsɛ˥ hɐi˩] 會合音為 [tsɛ˩] ，“但係”一詞會很規範地發成 [tan˩ hɐi˩] 或合音成 [tɐi˩]，或進一步弱化為 [tə˩]。

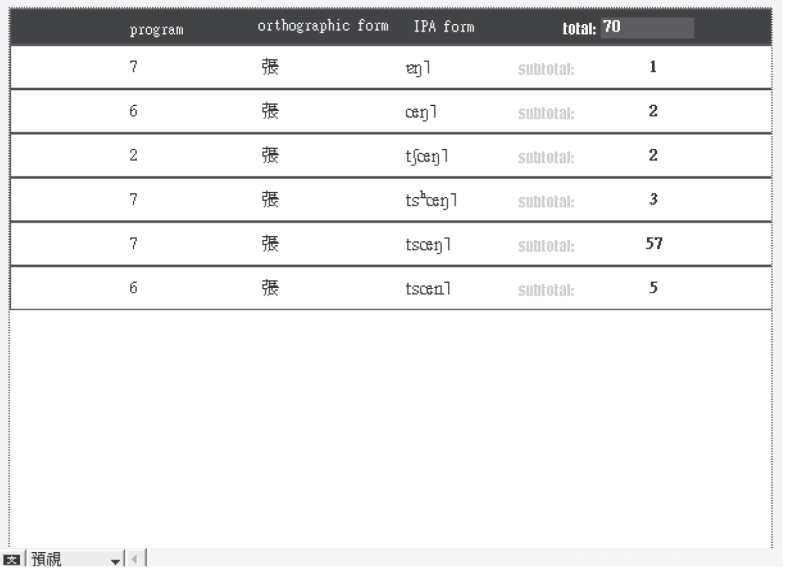
3.4. 檢索系統

為了使用的方便，語料庫設置了一個綜合檢索系統。該系統分為單字和句子檢索兩大部分，而每個部分又可以選擇以漢字輸入或音標輸入來進行檢索。例如：當用家輸入單字“張”時，檢索系統會馬上回饋該漢字在語料庫中的各種實際發音變體以及各變體的出處和語境（見圖一）。如果輸入音標，檢索系統會馬上回饋該跟音節相對應的漢字 / 句子以及它們的出處和語境。用家並且可以限定檢索某一特定節目；或某特定話語角色，如主持或聽眾、男性或女性等。除了做搜尋器，檢索系統還可以提供某個發音變體的出現總次數和頻率統計（見圖二）。這個設置為語音研究提供了重要的數據資料：詞頻對語言教學人員編寫粵語教程有很大幫助；而音節頻率則在認知語言學中佔很重要地位，因為它涉及自然言語處理，語言習得，腦神經病變引致的音韻失調，言語治療設計和心理語言學實驗設計等等範疇。

圖一：單字檢索結果



圖二：頻次統計



4. 現狀與未來

當語料庫構建完成後，我們曾經把它公開上載到香港大學網站。不少同行也使用了這個語料庫進行粵語語音、語法和言語治療的研究，如 Barry, Blamey and Fletcher（2006），Wong（2009），Kirby and Yu（2007），以及不少學位論文，如 Wong（2006）等等。可是，後來網站遭到黑客入侵而被迫暫時下架。我們一直考慮用其他方式跟同行分享這些資源。但是，我們當年編寫電腦程式時所採用的電腦配置系

統已經過時，不利用家獨立掛載使用。我們正計劃申請撥款，把語料庫進行技術更新，並加添詞類自動標注功能，希望在不久將來，重新上載上網，供各界人士參考和使用。

參考文獻

Barry, Johanna, Peter Balmey, and Janet Fletcher. 2006. Factors affecting the acquisition of vowel phonemes by pre-linguistically deafened cochlear implant users learning Cantonese. *Clinical Linguistics and Phonetics* 20(10): 761-780.

Du Bois, John W., Stephan Schuetze-Coburn, Susanna Cumming, and Danae Paolino. 1993. Outline of Discourse Transcription. *Talking Data: Transcription and Coding Methods for Discourse Research*, ed. Jane A. Edwards and Martin D. Lampert, 45-89. Hillsdale, NJ: Lawrence Erlbaum Associates.

Fletcher, Paul, Cheung-Shing Samuel Leung, Stephanie Stokes, and Zehava Weizman. 2000. *Cantonese Pre-School Language Development: A Guide* (Report of the project "Milestones in the learning of spoken Cantonese by pre-school children"). Hong Kong: Language Fund.

Kirby, James P. and Alan C. L. Yu. 2007. Lexical and phonotactic effects on wordlikeness judgments in Cantonese. In *Proceedings of the 16th International Congress of Phonetic Sciences*, ed. Jürgen Trouvain and W. J. Barry, 1389-1392. Saarbrücken : Univ. des Saarlandes.

Leung, Man-Tak, and Sam-Po Law. 2001. HKCAC: The Hong Kong Cantonese adult language corpus. *International Journal of Corpus Linguistics* 6: 305-326.

Leung, Man-Tak, Sam-Po Law, and Suk-Yee Fung. 2004. Type and token frequencies of phonological units in Hong Kong Cantonese. *Behavior Research Methods, Instruments, and Computer* 36(3): 500-505.

Wong, Lai-Yin May. 2009. Gei constructions in Mandarin Chinese and Bei constructions in Cantonese: A corpus-driven contrastive study. *International Journal of Corpus Linguistics* 14: 60-80.

Wong, Wai-Yi Peggy. 2006. Syllable fusion in Hong Kong Cantonese connected speech. Doctoral dissertation, The Ohio State University.

通訊地址：香港 九龍 紅磡 香港理工大學 中文及雙語學系（馮淑儀）
香港 薄扶林道 香港大學 言語及聽覺科學部（羅心寶）
電郵地址：roxana.fung@polyu.edu.hk（馮淑儀）、splaw@hku.hk（羅心寶）
收稿日期：2012年10月15日
接受日期：2012年11月18日