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。