知識天地

以語料庫語言學研究台灣國語口語使用

曾淑娟（語言學研究所副研究員）

語言學研究的主體是語言。語言形式大體上分書面語，口語以及時下流行的網路通信產生的口語化書面語。書面語包含報章雜誌，小說，學術文章等不同文體。口語也有朗讀，準備過的談話與沒有準備過的談話等不同形式。就語言學研究方法而論，語料庫研究算是一門較新的領域。它透過採集系統性的語料、標註語言學現象後，再以統計方法分析現象的細節與現象之間的相關性。本文主要介紹的是沒有準備過的自發性口語，與其中蘊含的語言學現象。

要收集自發性口語語料，首重情境設計。情境設計可以是完全無限制的自由談話，也可以是針對特定主題的會話，或是執行預先設計好的任務對話、訪問，或是請發音人口頭陳述事件。發音人的選取，影響極大。發音人的年齡、教育程度、性別及職業，都與詞彙和口音有很大的關係。錄音程序包括發音人所獲得的指示，閱讀或簽署的文件，以及錄音程序的說明。錄音地點可以是在戶外、錄音間、或普通房間。錄音設備與錄音器械、麥克風、錄音取樣品率、立體聲與否等有關。聲音處理相關的，有檔案轉錄的過程、聲音檔處理所用的語音軟體、以及語料是否經過切音處理。文字處理，則是語言內容的轉寫與標記。標記依語料庫建立的目的不同，可以是幾個特定語言現象，也可以是完整的言談或語法標記。語料庫語言學主要的研究內容，也包括標記系統的分析，與語料建置的工具開發。所謂的標記，像是口語對話的言談結構，與口語語音的變體。依語言材料的不同，言談結構的呈現亦會跟著改變。因此，言談標記必須依照語料庫建立的目的與形態而設計。口語語音的變化，若非實際審視語料，無法想像其中的多樣性。因此如果開發工具不佳，不僅收集速度慢，而且也無法確保所收集與標記的內容前後一致。

在中央研究院語言學研究所錄製的四個不同情境設計的台灣國語口語語料庫，分別是「現代漢語連續口語對話語音語料庫」、「現代漢語地圖導引口語語音語料庫」、「現代漢語主題對話語音語料庫」與「現代漢語新聞朗讀語音語料庫」，細節可以參見下表。相關的技術與內容也可參看多媒體語言的呈現與典藏網址 http://mmc.sinica.edu.tw。網路檢索系統正在建置當中，完成後將可開放提供語言學研究高效率的口語語料。

語料庫語言學能提供語言學家個別專業觀察以外的實證數據。但是所選取的標記系統，必須能確保標記員有明確的操作型定義，以及所設定的語言現象能反映語料的特性。圖上顯示「現代漢語連續口語對話語音語料庫」的標記系統，對所有說話者表現出的結果分佈，有很高的一致性。這表示所定義的現象，是合理而且能反映實際語言的使用。標記只是第一步，後續的分析研究才是語料庫語言學有用而且有趣的地方。有關標記現象詳細的分析與解讀，請參看筆者其他論著 http://corpus.ling.sinica.edu.tw/member/tsengsc/tsengsc_publications.html。

表一：台灣國語口語語料庫

	現代漢語連續口語 對話語音語料庫	現代漢語地圖導引 口語語音語料庫	現代漢語主題對話 語音語料庫	現代漢語新聞朗讀 語音語料庫
情境	陌生人之間的自然對話。談話內容除了限定路徑問題及自我介紹外，不限定特定主題，隨發音人自由交談。	地理導向對話。發音人雙方熟識，分持詳細地圖與刪減資訊後地圖，由持詳圖者依序引導持簡圖者至指定目的地。	熟識者之間的自然對話。兩位發音人選定 2001 年發生的特定新聞主題或事件進行對談。	新聞稿的朗讀。新聞內容是從 2001 年各類十大熱門新聞中選出八篇報導，再統一刪減至適當長度。
時間	1999.03 – 1999.07	2002.01 - 2002.03	2002.01 - 2002.03	2002.01 - 2002.03
語料	30 對話 （25.6 小時，平均長度約 50 分鐘）	30 對話 （5 小時，平均長度約 10 分鐘）	30 對話 （11 小時，平均長度約 22 分鐘）	60 朗讀 （2 小時，每篇新聞約 2 分鐘）
發音人	男 23 位，女 37 位 年齡：16-45	男 27 位，女 33 位 年齡：14-63	男 27 位，女 33 位 年齡：14-63	男 27 位，女 33 位 年齡：14-63
標記	詳細自發性口語現象 共標記 8 個對話	特殊語音現象 共標記 26 個對話	言談標記 共標記 29 個對話	無
音檔	18 GB	2.8 GB	6.78 GB	1.3 GB

以下僅舉三項語料庫研究，說明語料庫語言學能反映語言使用不同面向的觀察。

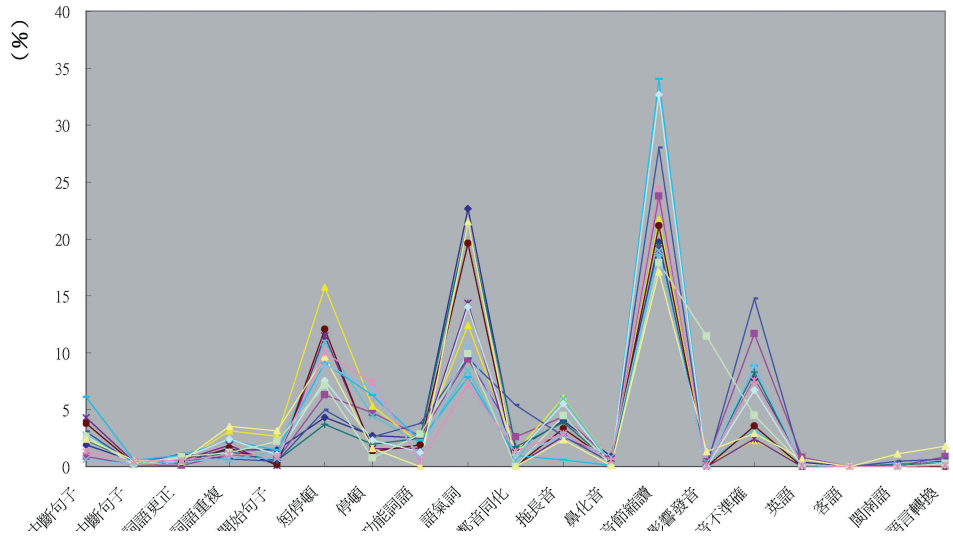
（一）高頻詞的使用：在現代漢語口語對話語料庫中，總共使用了 117,215 個音節，也就是方塊字。以「葡萄」爲例。「葡萄」是兩個字，兩個音節，但卻是一個詞。如果以「詞」爲單位，語料庫共含 81,053 個詞。實際語料統計中，「的」是最常被使用的詞，出現了 3,307 次。其次是「我」，「是」，「對」，「你」分別出現 3225,2648,1597,1447 次。圖中中的四十個高頻詞使用的次數佔所有語料的 47%。也就是說，我們在交談時不管是討論什麼議題，說的話有一半的機率是這四十個詞其中之一。這些詞語多半是虛詞、連接詞或人稱代名詞。

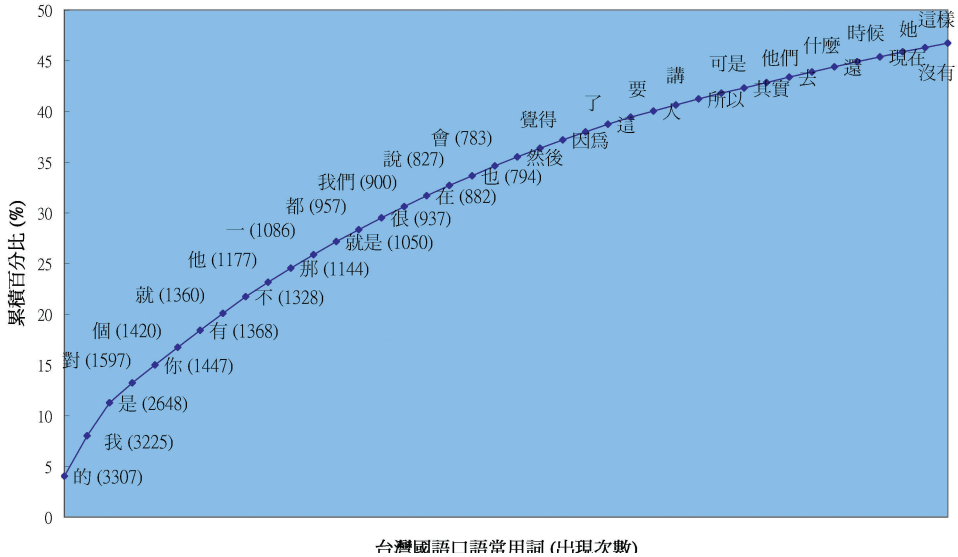
對說話者來說，即使這些詞發音上不完整，甚至根本不存在，我們都能依照語境推測而理解。但是對自動語音辨識系統來說，高度弱化的語音卻是最具挑戰性的任務。因此只要系統先能認出這四十個高頻詞，接下來相鄰的詞語可以依語音形式，語法環境以及語意相關性逐步解析。自然口語的發音，不像我們朗讀時那麼清楚、抑揚頓挫也沒那麼明顯，而且停頓也不像朗讀時那樣可以依標點符號預先安排。因此整理出高頻詞不同的語音變體，是語言學家對自然語言處理可以有具體貢獻的地方。這些語音變體又與合音詞有關。合音詞通常都是高頻的虛詞。由於經常使用，語意又不顯著，所以大家在發音時就比較偷懶，講得快一點。結果就是語音弱化，或與其他音節連併。

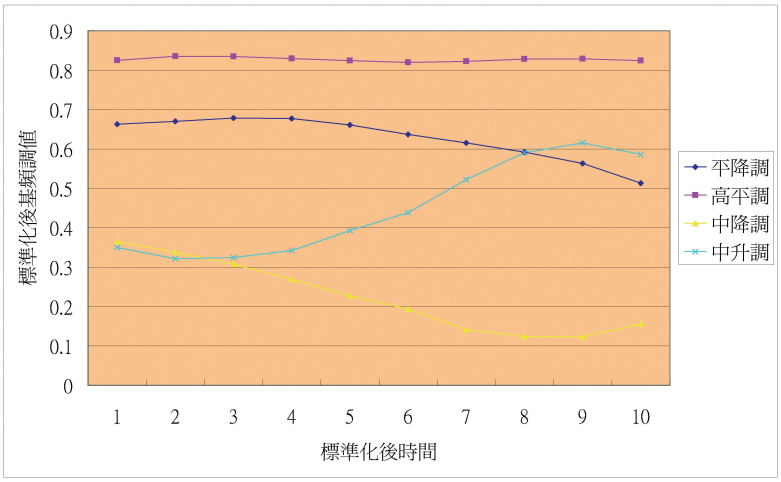
（二）合音詞的產生：龍宇純在其 1979 年論文中提到「徐言爲二，疾言爲一的雙音節詞合音爲單音節詞的現象 … 按反切之語，自漢以上即已有之。宋沈括謂古語已有二聲合爲一字者，如不可爲叵，何不爲盍，如是爲爾，而已爲爾（…），之乎爲諸。」可見合音詞在中國很早就存在了。梅祖麟院士也在 2002 年的文章中提到，閩南語中人稱代詞的複數形式，可能是由作爲複數詞綴的「儂」字聲母/n/連併弱化而來。漢語文獻提供的，只是書面的資料，所以合音詞的產生過程到底如何，沒辦法研究。但是現在因爲科技發達，我們可以建立語音資料庫，從語音現象一窺究竟。以目前網路通信流行的「ㄅㄧㄤˋ」代替「不一樣」、或「醬子」代替「這樣子」爲例，由於口語中出現頻繁，因此在口語化的書面交談中自然就承襲了口語的習慣而經常出現。一般合音的過程，是前一字的聲母加上後一字的韻母。所以「不」取其聲母「ㄅ」，「一」的韻母與「樣」的介音同爲「ㄧ」，加上「樣」的「ㄤ」，就連併成「ㄅㄧㄤˋ」。「醬」取代「這樣」，也是同樣的道理，只是現代漢語裡沒有「ㄓㄧㄤˋ」，就以「ㄐㄧㄤˋ」代替。聲波頻譜圖的分析，也顯示這兩組連併，在口語中依情境和語流速度的快慢，其弱化程度會有所不同。由完整的兩或三個音節，到連併成一個音節，弱化程度由輕到重的過程是連續的。這樣的研究爲合音詞在實際口語發音上找到佐證。

（三）語氣詞的使用：中文慣用語氣詞，表現語意之外的語氣。一個句子常因爲搭配了語氣詞，而使得原本的語意逆轉。以最常使用的語氣詞「啊」爲例。在語料庫中，「啊」佔了所有語氣詞的四分之一，一共出現了 2008 次。「啊」的語氣表現是最爲複雜的，也是最多樣的。我們將所有 2008 個「啊」聲音檔切出後，再把不同語者，不同基頻標準化後整體比較，可將「啊」的調性分爲四類：平降調、高平調、中降調與中升調（見圖下）。這四種調性都與一種以上的語氣並用。其中較常使用的，是平降調表贊同語氣，高平調表說話輪延續的意願，中降調表示回答問題，與中升調表總結一段談話。在台灣國語的口語裡，也使用相當多由閩南語承襲來的語氣詞，例如 HON，HEIN 等。但是卻很少使用由客語承襲來的語氣詞。有時語氣詞的使用也有社會語言學的意義。爲了表現對某個群體的親和力，會特地使用特別的語氣詞以拉近與對方的距離。群體可以是以語言、年齡、職業或性別爲導向。這也說明了語料庫能針對不同特徵子集進行檢索的重要性。

以上簡單介紹語料庫語言學的研究內容，若讀者有相關問題歡迎來信指教 tsengsc@gate.sinica.edu.tw。







圖上：語料庫標註結果
圖中：台灣國語口語高頻詞分佈
圖下：「啊」的四種調性