參與中研院「口語語料庫研究討論工作坊」心得分享

撰寫者／研究員 Akiw、研究員 Dawa、研究助理黃堂維

3 月 17 日，我們有幸報名參加中研院語言所曾淑娟研究員主講的「口語語料庫研究討論工作坊」。議程如下：

8:40-9:00	報到
9:00-10:20	口語語料庫建置與分析（一）曾淑娟（中央研究院語言學研究所）
10:20-10:30	休息
10:30-12:00	口語語料庫建置與分析（二）曾淑娟（中央研究院語言學研究所）
12:00-13:00	午餐
13:00-14:00	口語語料庫計畫分享（一） 1990-2005年金門口頭敘事田野普查錄音資料整理保存方法芻議 唐蕙韻（金門大學華語文學系） Taiwanese Min Spontaneous Speech Database 潘荷仙（交通大學外國語文學系） 華語學前幼兒整合性語言能力評估：華語KIDEVAL的開發與應用 張顯達（香港教育大學語言學及現代語言系）
14:00-14:50	全體參與人員討論交流
14:50-15:00	休息
15:00-15:40	口語語料庫計畫分享（二） 大型台灣廣播語料庫建置 廖元甫（台北科技大學電子工程系） 閩南語語音辨識與合成系統客製之口語語料庫 薛丞宏（itaigi工程師/中央研究院資訊科學研究所）
15:40-16:10	全體參與人員討論交流
16:10~	賦歸

而根據議程我們可以了解，本次工作坊分成上下午，上午場的主講是曾研究員，主要分享他們在做口語語料庫的經驗及應用，下午場則是邀請在建置口語語料庫的研究人員，一起討論在實際建置口語語料庫的過程中所面臨的困難或是相關經驗分享，互相交流、彼此提供建議。以下將以簡要的方式介紹各場次討論的精華：

〈口語語料庫建置與分析〉

從曾淑娟研究員的演講中，我們很清楚的得知口語語料庫建置的脈絡，然而，目前語料庫建置及公開的最大困境，即是在研究倫理的部分。因過去的語料收集，往往沒有填寫知情同意

書，就算有知情同意書，可能也沒有進一步的約定可以將語料作為語料庫公開使用，因此曾研究員提醒，在錄製口語語料時，一定要很注意這個部分。另外，口語語料轉寫為文字需要有輔助工具，以及特殊的標記，相關人員的訊練可能都至少要三個月。在訪談時，因為口語語料通常是以自然對話為主，因此建議用不同的麥克風收音，分開音軌，以利後續的研究及分析。在應用方面，像是建立兒童的口語語料庫，就有助於發展遲緩的兒童在語言學習方面有所對照，對於臨床治療分析很有幫助。除此之外，口語語料庫的證據也可以和社會語言學上的分析做對照，例如男性、女性、年紀、社經背景、學歷所可能影響的語言使用情形，往往也是相符的。

〈1990-2005 年金門口頭敘事田野普查錄音資料整理保存方法芻議〉

唐薰韻教授因早期參與計畫，協助金榮華教授整理 1990 年於金門採錄的口傳民間故事，並轉錄為文字，於 1996 年由金門社教館出版，目前錄音帶由金教授授權給唐教授保存。後 94 年唐教授獲金門縣政府補助，再進行口傳民間敘事普查，於 2007 年由金門縣文化局出版。

因此目前唐教授所持有 50-120 小時錄音檔案，1990 年多為錄音帶，而後則有錄音帶及電子音檔，限書籍篇幅及人力限制，調查內容並未全面整理出版。目前有試著用用錄音機轉錄出來（數位化），但音質不佳，因此想尋求解決的方法。在場的研究人員是建議錄音帶的內容應盡早數位化，也可以透過專業的影音修復機構協助，如臺南藝術大學音像紀錄與影像維護研究所就有這方面的技術。

“Taiwanese Min Spontaneous Speech Database”

潘荷仙老師分享他們所建置的閩南語口語語料庫，口語語料的收集的來源主要來自臺灣六個區域的發音人，分析所使用的參照資料有教育部臺灣閩南語常用詞辭典、國立中正大學台灣閩南語口語語料庫及建立在 CHILDES 系統上的資料(註：CHILDES 系統有含有世界各國語言的兒童口語語料資料，持續有研究人員以相同作業標準流程建置口語資料標記及分享這些口語語料，目前系統中有國立中正大學蔡素娟老師在系統上分享臺灣閩南語兒童語料庫(Taiwanese Child Language Corpus)的資料。)潘老師也分享在建置臺灣閩南語分析口語語料的工程，標記聲調等 9 項轉錄，使用自動化系統協助標記工作，目前研究還在進行中及思考如何在系統中建立搜尋以供後續研究分析更便利。

〈華語學前幼兒整合性語言能力評估：華語 KIDEVAL 的開發與應用〉

張顯達老師則是分享他們正在開發的 KIDEVAL 華語版。每一種語言各有各自的特色，張老師所帶領的團隊是跨國合作的研究，由於國外發展的 CHILDES 在開發時是以英語為出發，後續有愈來愈多語言採用此系統進行分析兒童語言，CHILDES 系統裡有協助分析兒童語料的軟體，KIDEVAL 是分析軟體中的一項，軟體可以馬上知道兒童的語言詞彙、平均語句長度、句法類型、流暢度等分析指標，但因為國外開發的軟體若要適用在華語，就需要調整，而張老師所帶領的團隊也發現所分析的指標中有不適用於華語上，因而著手發展調整華語版的 KIDEVAL 平台，只要此平台建置完成，對於兒童的華語能力發展及變化可以在很短的時間內分析出來，為了達到研究計畫的目的，有許多的基礎研究工作需要進行，如中文斷詞系統的標

記和自動斷詞等標準化的工作流程都需要進行分析及討論，完成後一定可以對華語兒童語言發展的研究有很大的幫助。

〈大型台灣廣播語料庫建置〉

廖元甫老師則是參加 AI 科技大擂台比賽，而在廣搜閩、客、華、原各語的口語語料，另外也開發一套系統可以辨識音檔，將華語分音節轉成文字，也會同時輸出時間軌，再結合影片後可成為字幕檔，仍在努力調整中，希望能夠搜集更多口語語料。

〈閩南語語音辨識與合成系統客製之口語語料庫〉

最後薛丞宏工程師目前主要參與中研院資訊所高明達老師的語音辨識計畫，專長是台灣語言相關的機器翻譯、語音合成及語音辨識和語料庫設計。他主要的報告是在分享語音辨識計畫整理語料時使用的聽打標記軟體，比較 praat 及 transcriber 以及客製化語料庫的優缺點，並分享對應的管理方法。

另外薛工程師也提到目前做語音合成實務上遇到的困境，如果定義及邏輯不夠明確，就會有一些問題，例如應用語音合成的相關技術在阿美語萌典上，僅靠一些語法概論的描述，可能無法擴及全部的詞彙、句子都適用，這是工程師端無法克服的問題，還是要有語言學者或族人提供更明確的語音結構定義，合成的語音會更準確。

分享：

參與本次工作坊，也提醒我們原住民族的口語語料及相關的研究仍不足，如兒童語言的調查與分析，目前尚未有長期觀察的研究，讓我們可以了解在不同語言環境的孩子真實的語言能力發展，以作為族語發展對照與評估。

而綜觀台灣原住民族的口語語料庫，仍有公開且經分析的如台大台灣南島語多媒體語料庫（噶瑪蘭語、賽夏語、鄒語、阿美語、撒奇萊雅語、賽德克語、布農語）、蘭嶼達悟語口語資料典藏網。而過去中研院也有台灣南島語數位典藏（魯凱語、賽夏語、泰雅語、布農語、排灣語、卑南語），也因此培養很多學術人才，但往往可能因為計畫結束就沒有足夠的人力及經費繼續支持營運，實為可惜。上述三個口語語料庫，都是由計畫團隊訪談、錄音後再進行轉寫及分析，再加上語料庫程式的工作，需要一個團隊才能執行。因此目前我們希望用「先典藏，後應用」的方式，首先要持續搜集早期的口語及書面語料，並盡快的將之數位化，建置數位族語資料庫存放這些珍貴的資料，以利未來可以讓更多人看見這些資料，進一步應用於族語的傳承與發展。

目前原住民族語言已有書寫符號可供記錄，會族語書寫的大多為族語老師等，口語的記錄仍是非常重要的一部份，也就是口語與書面都需要做記錄。唯有記錄才能做為後續研究人員進行轉寫分析記錄，並以科學的方式記錄聲學的特徵，有了這些資訊，未來的應用性才有可能做的到。族語口語研究也是很少有人去進行研究，小孩子的口語、成人的口語、長者的口語等，整個口語語言發展的研究及記錄是與閩南語及客語研究是相對落後的，若與華語口語研究來比較，更是落後更多。沒有留下口語記錄資料，會說族語的長輩若離開，沒有人學會的話，族語消失斷層會更加嚴重，對於搶救族語是非常不利的。

口語記錄由於是記錄語音，從個人資料的角度出發，語音是屬於可識別性之個人資料，有關資料之取得及後續口語資料保存及研究授權使，會是未來研究機構單位需要特別注意的一個環節，也就是在記錄前必須需簽同意書及資料之使用。唯有建立完善的保護機制，才可以永續進行族語研究及族語發展。