原住民族語言發展論叢：
理論與實務

行政院原住民族委員會　出版
台東大學　華語文學系　編輯

台灣南島語數位典藏*

齊莉莎
中央研究院語言學研究所
hsez@gate.sinica.edu.tw

余清華†
中央研究院語言學研究所
harryyu@gate.sinica.edu.tw

1. 前言

台灣的南島語言類型極爲豐富，而且各種語言和方言之間的差異也頗大。然而，台灣南島語不同於漢語，主要在於沒有文字的記載，因此一旦語言或方言的使用者不復存在，則這個語言或方言也就跟著消失，毫無紀錄可循。目前我們除了以各種方式延續語言的生命外，更實際的作法即是蒐集和保存現有的語料。「台灣南島語數位典藏」不僅能使蒐集的語料彙整、流通，並且在語言、民族、文化、教育等各方面都將具有相當重要的參考價值。

「台灣南島語數位典藏」是國科會數位典藏國家型機構計畫下分項主題「語言典藏」的子計畫之一，整體建置作業之規劃暨主持人爲齊莉莎女士。此計畫之建置目的在於建立一數位圖書館（digital library），期能蒐集、保存、編輯及透過網路散播語言資源，供使用者存取已錄製及記音的南島語。第一期計畫爲：2002-2006，現在已經進入第二期計畫：2007-2011。

「台灣南島語數位典藏」（http://formosan.sinica.edu.tw）的建置已經接近六年了，包括語料庫、語言地理系統及書目資料庫等，最終目標爲建立所有台灣南島語的語音、詞彙、單句和長篇語料，並加以中、英文翻譯。在應用上，所設計的語料庫查詢介面可讓使用者依語言及方言別和類別等參數自訂語料庫的檢索範圍，做統計及比較研究。語言地理資訊系統的應用則是希望讓使用者了解台灣南島語的分布圖，並觀察同源詞與非同源詞的分布情形。此外，所建立的台灣南島語書目資料庫可以檢索下列不同的書目資訊：語言學、語言教學、文學及音樂等。

本文除了描述典藏的建立過程及內容外，還提出語言分析之層面所遇到的困難，讓我們重新思考台灣南島語之間的差異問題。另外，我們也將說明如何運用南島語典藏進行研究和教學。

* 本篇文章之校稿，感謝朱黛華、林志憲及莊雲翔之協助。
† 在此要特別感謝本篇文章的合撰作者—余清華先生（1963-2007）。余清華先生於民國 90 年 4 月 1 日到職，期間不斷協助第一作者建構語言典藏資料庫，是台灣地區第一位成功開發台灣南島語語料庫之系統工程師，成效卓越、功不可沒，且貢獻良多。

2. 台灣南島語數位典藏之內容

台灣南島語數位典藏之描述內容從以下兩方面進行：

（1） 說明計畫目的、工作進度及出版現況（請參考§2.1）。
（2） 介紹網路介面（請參考§2.2）。

2.1 計畫目標、工作進度、出版現況

2.1.1 計畫目標

除了蒐集、整理及保存台灣南島語語料之外，我們的目標可以進一步的分成兩部分來說明：

（1） 建立研究工具，讓國內外專家學者或從事台灣南島語研究有興趣之原住民人士、社會大眾或學生從不同角度（語音、詞彙、構詞、句法、對話等）進行研究及比較語言之間的差異。
（2） 發展教學工具。

2.1.2 數位典藏工作進度

我們在 2001 年從實驗角度進行了南島語的數位化，建立子資訊架構，並完成單一語言（魯凱語萬山方言）的語料庫。過去六年（2002-2007），我們在「台灣南島語數位典藏」工作小組之努力之下（ 請參考表 1）已經建立魯凱語（萬山、茂林、多納及大南等方言）、賽夏語（東河方言）、泰雅語（賽考利克方言）、鄒語、阿美語、布農語（南部方言）、排灣語（南排）、卑南語（南王）、巴宰語、卡那卡那富語及西拉雅語的語言典藏，包括數位化及部分分析（請參考表 2）。

表 1：「台灣南島語數位典藏」工作小組之成員及工作分配

・語料分析	魯凱語		齊莉莎、林惠娟、*邢天馨[1]
	賽夏語		朱黛華、齊莉莎
	泰雅語		葉郁婷、齊莉莎
	布農語		齊莉莎、*劉秋雲、林秀蓮、林聖賢、曾思奇、李文甦
	排灣語		*華加婧、齊莉莎
	卑南語		Josiane Cauquelin（戈格林）、齊莉莎、鄧芳青
・語料輸入	布農語 鄒語 巴宰語 卡那卡那富語 西拉雅語 魯凱語（大南方言） 排灣語		林志憲
	阿美語		*林翠鐘
・語料翻譯	鄒語	德文→英文	*Christopher Smidt
		英文→中文	*吳貞慧、潘家榮
	阿美語	法文→中文	林秀蓮
		法文→英文	齊莉莎
	布農語	中文→英文	*劉秋雲、齊莉莎
	排灣語	英文→中文	莊雲翔（北排方言）
		中文→英文	齊莉莎（南排方言）
	魯凱語	英文→中文	齊莉莎、*王秀梅（多納方言）
			*劉秋雲、齊莉莎（大南方言）
・地理系統			*華加婧、*白璧玲
・教學網站			齊莉莎、林志憲
・系統工程師			余清華[†]
・後設資料小組			*翁翠霞、*沈漢聰
・書目資料庫			齊莉莎、*劉秋雲、*華加婧、林志憲（持續更新語言學相關書目資料庫）

[1] * 表示已經不在此計畫下服務。

表 2：「台灣南島語數位典藏」語料蒐集、翻譯、分析校對及展出內容

語言	方言	田野調查者	原始記錄使用之語言	語言分析者	進行蒐集、翻譯、分析及校對時間	詞料	詞數	句子	音檔（mp3）	網路展示
魯凱語	萬山	1）齊莉莎、林惠娟	中英文	齊莉莎	1992-2003	14	6598	764	60MB	✓
		2）齊莉莎、林惠娟	中英文	齊莉莎	2002-2003	21	7000	1200	65MB	
	茂林	邢天馨	中英文	邢天馨、齊莉莎	2001-2002	24	3945	419	50MB	✓
	多納	1）齊莉莎	中英文	齊莉莎	2001-2003	12	11281	899	60MB	✓
		2）齊莉莎	中英文	齊莉莎	2003~	8	3400	500	35MB	
	大武	齊莉莎	中英文	齊莉莎	1992-1994, 2003	9	650	200	14MB	
	大南	李壬癸 [1975]	英文	齊莉莎	2002-2003	26	10656	1237	尚未統計	
	霧台	齊莉莎	中英文	齊莉莎	1992-1994, 2001-2003	尚未統計	尚未統計	尚未統計	尚未統計	
賽夏語	東河	1）朱黛華	中英文	朱黛華、齊莉莎	2001-2003	14	4479	374	30MB	✓
		2）朱黛華	中英文	朱黛華	2005~	3	800	250	15MB	
泰雅語	賽考利克	葉郁婷	中英文	葉郁婷、齊莉莎	2002-2003	20	10439	1476	80MB	✓
鄒語	特富野	董同龢等[1964]	英文	待分析	--	48	9088	1362	70MB	✓
	達邦		英文	待分析	--	57	8334	1003	66MB	✓
	久美		英文	待分析	--	29	5589	661	43MB	✓
阿美語	中部	Fey *et al.* [1993]	中英文	待分析	--	25	50000	1780	200MB	✓
卡那卡那富語		Tsuchida [1964]	英文	待分析	--	13	5412	605	尚未聲音檔	✓
巴宰語		Li and Tsuchida [2001]	英文、中英文	待分析	--	10	5961	781	尚未聲音檔	✓
布農語	南部	曾思奇等[1998]	中文	齊莉莎、劉秋雲曾思奇、李文甦林聖賢	2004, 2006~	49	35089	1265	尚未聲音檔	✓
排灣語	南部	華加婧	中文	華加婧齊莉莎	2005-2006	20	12000	800	55MB	✓
卑南語	南王	Cauquelin [To appear]	英文	Cauquelin齊莉莎	1983-2006	31	尚未統計	尚未統計	尚未統計	

2.1.3 出版現況

這幾年我們不斷發表論文及專書，可以分爲兩類：

(1) 與語料及語言分析相關（表 3）；

(2) 與語料庫架構相關（表 4）。

表 3：語料分析及語言分析之相關著作

語料分析		
1.	Cauquelin, Josiane（To appear）	卑南語
2.	Zeitoun, Elizabeth and Lin Hui-chuan（2003）	魯凱語萬山方言
語言分析		
3.	Zeitoun, Elizabeth（2002, 2007）	魯凱語萬山方言
4.	Zeitoun, Elizabeth and Hsin Tien-hsin（2002）	魯凱語茂林方言
5.	Zeitoun, Elizabeth and Wu Chen-huei（2006）	類型比較
碩士論文		
6.	Wang, May Hsiu-mei（2003）	魯凱語多納方言

表 4：典藏架構之相關著作

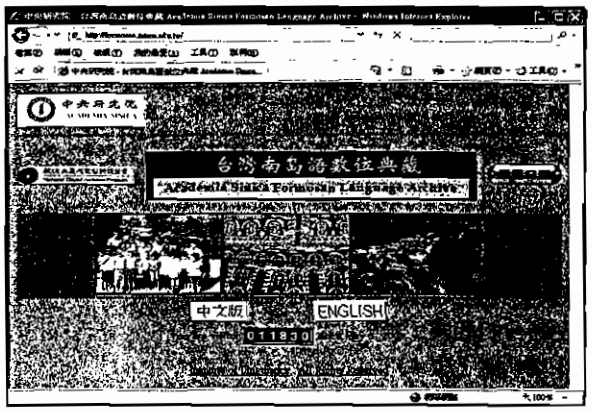
1.	Zeitoun, Elizabeth, Yu Ching-hua and Weng Cui-xia（2003）
2.	Zeitoun, Elizabeth and Yu Ching-hua（2005）

2.2 網路介面

「台灣南島語數位典藏」包括三種主要資料庫：語料庫、語言地理資訊系統及書目資料庫等。目前，這三種資料庫是獨立建構，不過爲了能建立各種台灣南島語的語音、詞彙、單句和長篇語料，最終目標會將它們結合起來。

所有的資料分爲兩種不同介面：中文及英文。

圖 1：首頁



2.2.1 語料庫

2.2.1.1 檢索系統

所建構的語料庫查詢介面可供使用者依語言、語料類別等參數設定語料庫範圍，做統計及比較研究。

瀏覽語料有不同的呈現方式：使用者可選取一種方言，以句子或段落存取所錄製的長篇語料，同時也可以比較原始語料（raw data）和重新分析的語料（edited data）。此外所提供的詞彙用字索引可讓使用者檢視每個詞彙的分布。再者，爲進一步分析及比較語料，每個詞也加以適當的斷詞處理，並附上中文與英文單詞對譯，然而語料可從三方面進行檢索及搜尋：關鍵詞、詞綴及詞類，可讓使用者從語料庫中找到任何出現的詞素，並且了解其分布與功能。

綜合上述之架構，以下說明台灣南島語語料庫（與其他能夠在網路瀏覽的典藏內容比較起來）之優點與特色：

(1) 我們儘可能依據錄音帶或 MD 記下原始的語料內容，因此使用者可以一邊聆聽聲音，一邊邊循句子／段落之記音。在其他的語言典藏，聲音檔未必能對應於文本上的記音，或者未能對應到語料。

(2) 可以從不同的角度來進行研究及語料的比較（句子、段落、詞彙分布、詞綴、詞類等）。在其他的典藏網頁裡，語料通常以 Acrobat Reader 格式（PDF）建立，如此一來難以查詢或應用；此外也無法對語料中出現的詞綴或詞類進行交互參考。

2.2.1.2 後設資料

台灣南島語的後設資料（metadata），2002 年在中研院計算中心後設資料工作組（MAAT）的協助下，亦已完成設計，希望在第二年的計畫執行期間加強資料擷取與資源分享，目前還在持續測試中。

所謂「後設資料」指有關資料的結構性資料，它一方面能夠幫典藏者利用結構化的典藏方式、標記及文本編目等方式來加以管理所收集的資料；另一方面，使用者也能夠更容易從不同的典藏品存取所需的資料。

後設資料的描述主要爲超連結的資訊。換句話說，在南島語典藏裡，它包含如下的資訊：（1）語料的側寫（如：題目、語言及方言、文體）；（2）田野調查（如：發音人、田野工作者／謄寫者／編輯者／翻譯者、資料收集／編輯／翻譯的日期）；以及（3）管理策略（版權宣告）。爲了統一術語起見，以及降低每個文本編目的負擔，分別建立語言資料庫與個人資訊的資料庫。語言資料庫包含的資訊爲地理分布、方言及方言變體、人口以及語言使用狀況，而個人資訊的資料庫則提供了發音人姓名、所屬的語言族群、年齡、語言能力、出生地等等，以及其他關於田野工作者／謄寫者／編輯者／翻譯者的基本資訊。「超連結」進一步讓使用者能夠從文本後設資料內的「語言」、「發音人」及「參與者」等元素取得更多資訊。

2.2.2 語言地理資訊系統

語言地理資訊系統的應用則是希望讓使用者了解台灣南島語的分布圖，並觀察同源詞與非同源詞的分布情形。我們持續新增各種語言及方言的詞彙，以擴大現有的資料庫。

2.2.3 書目資料庫

台灣南島語相關的書目資料庫可以檢索下列不同的書目資訊：語言學、語言教學、文學及音樂等。新的書目資訊不斷加入，希望達到更完整的台灣南島語書目資料庫，供使用者檢索所需。

3. 語言分析及工程技術的問題及解決

我們在建立台灣南島語語料庫時面臨的困難不但讓我們重新思考台灣南島語言之間的類型差異（請參考§3.1），更因而讓我們發展語言處理之軟體（請參考§3.2），以下將作進一步說明。

3.1 語言上之主要原則

所遇到的困難包括：書寫系統之選擇、記錄（包括斷詞及斷句）之正確性、縮寫及標記之原則、母語—中文—英文翻譯之一致性、構詞及句法之相似及差異。

3.1.1 書寫系統之選擇：以國際音標為標準

雖然台灣南島語之語音系統較簡單，但是過去不同的記錄方式（包括羅馬拼音、漢字、音節文字（片假名）及注音符號）使得台灣南島語尚未有「標準化」的書寫系統。這個問題，可以從兩個觀點來說明：

（1） 無論從不同語言或方言來看，至今台灣原住民尚未建立可以互用的書寫系統，例如，阿美語的 *d* 代表擦邊音[ɬ]；但是在其他語言，則用於齒音或齒齦音[d]。

（2） 學者之間所使用的書寫符號也不完全一致：例如，Blust（2003）將 c 用來記音齒間擦音 [θ]，可是其他學者卻[θ]將成寫 th。

爲了避免混淆並使國際音標能符合國際標準，我們對於所有重新分析及編輯或近來蒐集之語料皆以國際音標（IPA）爲標準[2]。

3.1.2 記錄之一致性及正確性

如何確認語言記錄的正確性及一致性，對每一個田野調查者爲一大考驗。這種問題

[2] 將來，將考慮和教育部書寫系統作連結，提供轉換系統。

可以從幾方面來探討：（1）記錄之一致性、（2）斷句或段落之標準，以下分別討論。

3.1.2.1 記錄之一致性

記錄之一致性可以從三方面來說明：（1）記音之正確性、（2）詞彙語意之判斷方式及（3）斷詞方式，以下分別討論之。

3.1.2.1.1 記音之正確性

有些問題只要仔細校對就可以解決，比如在李壬癸教授（1975）所蒐集的大南語料中 *aramor* 及 *ʔaramor*「很」同時出現，調查之後才發現 *ʔaramor* 爲正確發音；此外，齊莉莎（2002）在所蒐集的多納語料也找到不一致的詞彙對立，如：*ʔikay*「在」～*ta-ikay-anə*「放在」，在校對之後修改爲 *ta-ʔikay-anə*。

有些詞彙的不同記音不容易解釋，比如在萬山方言 *ivoko* 「朋友」不含喉塞音而 *la-ʔivoko*「朋友們」卻有，便無法解釋。雖然如此記錄方式會產生「不一致性」的問題，但唯有如此記錄方式才能正確表達出喉塞音在萬山魯凱語爲一個音位的事實。同樣的，我們必須注意語音轉換所帶來的變化，比如魯凱語萬山方言的喉塞音會因不同動詞之形式而出現的位置有所不同，比較：*o-vaʔai*「給（動態.限定）」，*vaaʔi*「給（動態.非限定）」及 *voaʔi*「給（動態.虛擬式）」。

3.1.2.1.2 詞彙語意之判斷：避免從上下文去決定詞彙之語意或句法功能

在很多文獻中，作者提供一個單詞之翻譯，但是這種作法往往產生不一致性的情況，以下進一步說明：

李壬癸教授（1999:145）有時將賽夏語的 *ʔi*（*k*）分析爲不獨立的成分，如：*ʔokaʔ ʔilaʔi* ...「沒 了 ...」，有時卻處理成一種獨立成分而且翻譯成「不」，如：*ʔokaʔ ʔila ʔik* ...「沒 了 不 ...」。事實上，*ʔi*／*ʔik* 可分析爲連繫詞，前者出現在動態動詞前，而後者則出現在狀態動詞前，其功能爲連結一個否定詞比如 *ʔokaʔ*「沒有、不」或 *ʔiðiʔi*「別」及一個動詞（請參考 Yeh 1991, 2000 and Zeitoun 2001），此外，*ʔi*／*ʔik* 並沒有特定的語意（lexical meaning）。

相對來說，同一詞可包含不同語意，比如：魯凱語萬山方言的 *ʔoponoho* 可指的是「萬山人」或「萬山村」，或不區分其他語言所區分的語意，比如：魯凱語萬山方言使用一個單詞 *taka* 表示「兄姊」，未區分「哥哥」或「姊姊」。

再者，許多詞彙或詞綴爲同一形式而不同義，比較萬山方言的：*ʔini-tolitoliʔi*「盪鞦韆」（< *ʔini-*「移動」）～ *ʔini-to-taləkə*「請客」（< *ʔini-*「飲」）～ *ʔini-taloðo*「過橋」（< *ʔini-*「過」）～*ʔini-ca-colo*「自己殺豬」（< *ʔini-*「反身」）～*ʔini-kana-kan-aə*「假裝吃」（< *ʔini-...-aə*「假裝」）。

這幾個例子證明無論一個詞有沒有句法功能（grammatical function）或詞義（lexical content），不能根據其出現的上下文而判斷其功能或語意，必須從其分佈及語意核心而決定。

3.1.2.1.3 斷詞方式：以詞素為標準

爲了掌握詞彙之語意並且瞭解其構詞方式，我們記錄的每一個詞彙以「詞素」爲標準，分爲兩大類：「自由形式」及「附著形式」。「自由形式」詞素通常指的是「詞根」；「附著形式」可以進一步區分爲「詞綴」及「依附詞」。

雖然理論基礎相當固定，但是在進行分析的過程中往往會遇到如何斷詞等問題，以下簡單的舉幾個例子來說明這一點：魯凱族萬山人自稱爲 *ʔoponoho*；仔細比較不同方言，可以得知與霧台（或大武、大南）的 *swaponogo* 是同源（古魯凱語*s >萬山 *ʔ*；古魯凱語 *oa >萬山 *o*；古魯凱語 *g >萬山 *h*）；再進一步觀察，就會發現 *swaponogo* 可以分爲：*sw-* + *a-* + *ponogo*「居住 + 事實 + 地名」而在霧台或其他魯凱語之東南部方言可以使用 *swaponogo* 或 *ponogo* 來指萬山人。萬山人已經不承認或不認得 *ʔo* 及 *ponoho* 這兩種詞素，因此在語料庫中尚未區分 *ʔoponoho*。

3.1.2 斷句或段落之標準

我們通常根據兩種因素來斷句：（1）子句或句子結構（structure of the clause or sentence）；（2）停頓（pause），在此並沒有限制必須從子句或句子斷語料，原因在於有時候子句的翻譯不一定完整。

此外，我們通常以句子內容之相關性來區分段落；有時候停頓也會幫助我們決定段落之界線。

3.1.3 母語—中文—英文翻譯之一致性及相對性

在翻譯的一致性上，我們所面臨的困難可以從（1）文化產生的多種語意概念及（2）同源及語意變化兩方面說明。

3.1.3.1 文化產生的多種語意概念

詞彙之翻譯，無論在同一個語言或不同語言常常造成困擾，原因在於翻譯詞彙的過程中，必須考慮最典型的語意。魯凱語六個方言中，有很多詞彙涵蓋的語意概念多，難以用一個中文或英文單詞來呈現。

3.1.3.2 同源及語意變化（semantic shift）

所蒐集的語料中，常可以找出（方言或語言之間）的同源詞，不但要掌握其翻譯之一致性，另外也必須注意其語意變化。魯凱語中的萬山、多納、霧台等方言的 *ləpəŋə* 表示「完成」，但是在大南方言，除了表示「完成」 *ləʔəŋə* 也可以表示「全部」。另外，魯凱語中之斜格代名詞由 *-a*（*nə*） 來表示，如：霧台方言的 *nakuanə*。但是，霧台方言的 *nakuanə* 也可以表示「擁有、領屬」（比如：「是誰的？」「是我的！」）。

3.1.4 縮寫及標記之原則

詞素之縮寫包括詞彙及語法翻譯（lexical and grammatical translations）。上述已經說明詞根之分析及翻譯標準。主要的句法詞素與德國 Max Plank Institute 所提出的原則及縮寫相當接近，在語言學上使用性很普遍。

3.1.5 構詞及句法之相似性及差異性

無論是分析一個語言或比較不同方言／語言，我們必須注意構詞及句法之相似性及差異性。

3.2 轉檔技術及語言處理軟體之發展

我們在技術上所面臨的困難有兩種：（1）如何在網路上呈現語料之轉檔；（2）如何協助分析者提供更正確及一致性高的語料。爲了解決這兩種問題，我們開發不同資料庫及語言處理之軟體，以下從三方面進一步說明：（1）語音轉換；（2）檔案轉換及（3）軟體之開發，以下分別討論之。

3.2.1 語音轉換

雖然在網路上呈現國際音標已經不是一件難事，但我們在 2001 年剛建立台灣南島語料庫時，這種技術尚未發展得很普遍，最後，我們決定使用 Unicode 代碼來轉換及呈現國際音標。

3.2.2 檔案轉換

我們建立了兩種資料庫，一種爲語言及方言代碼，語言代碼完全遵循 SIL 所提供的標準代碼。不過 SIL 卻未提供方言的代碼，因此我們自己列出相關的簡碼（參見表 5）。

表 5：語言及方言代碼

語言	SIL 代碼	方言	代碼
Rukai	DRU	Mantauran Maga Tona Budai Tanan Labuan	Mn Mg To Bu Ta La
Saisiyat	SAI	Taai Tungho	Ta Tu
Atayal	TAY	Squliq C'uli'	Sq Cu
Tsou	TSY	Duhtu Tfuya Tapangu	Du Tf Ta

Amis	ALV	Sakizaya Northern Tavalong-Vata'an Central Southern	Sa No Ta Ce So
Bunun	BNN	Takituduh Takibakha Takbanuaz Takivatan Isbukun	Td Th Tb Tn Is
Kanakanavu	QNB	Kanakanavu	Ka
Puyuma	PYU	Nanwang Kapitul	Na Ka
Paiwan	PWN	t-dialect tj-dialect	Td Tj

此外，我們將同一方言之長篇語料合併起來，產生一個機讀格式的 flatfile，這個檔案分成許多區塊（block），由六行文字所組成。第一行表示語言及方言別、文本別、段落別及句子別，其中句子別 a 表示第一句，b 表示第二句，其餘類推；第二行表示原文，也是本文主要處理的對象；第三、四行分別表示中、英文標記；最後兩行則分別表示中、英文翻譯，然後這種橫向的 flatfile 利用程式將原始資料匯入至關聯式資料庫（SQL Server），此資料庫不但可以呈現單詞、句子、段落三種語言之間的對齊，也可連結到語音檔（請參考圖 2 及圖 3）。

圖 2： 句子層面資料庫

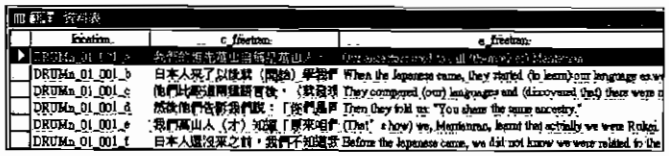
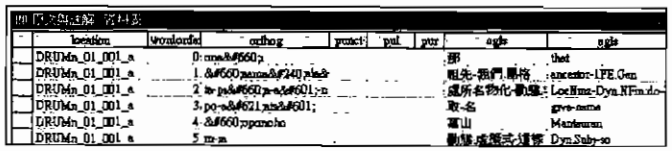


圖 3： 詞彙層面資料庫



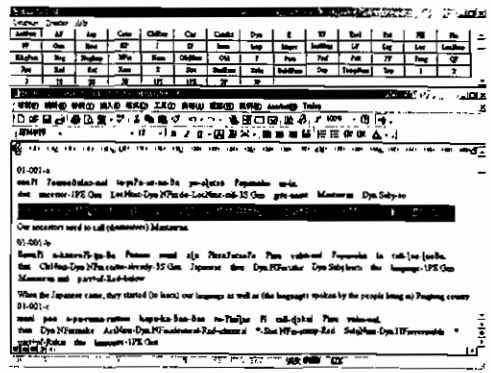
3.2.3 軟體之開發

爲了協助分析者提供更正確及一致性高的語料，我們開發出 AnnoTool 及 Chkgloss 兩種軟體。

3.2.3.1 AnnoTool

AnnoTool 具有兩種功能：（1）提供在語料庫中所使用的中英文縮寫（gloss）；（2）將中文或英文縮寫翻譯成英文或中文，不但可以達到更高的一致性，也可以節省時間。

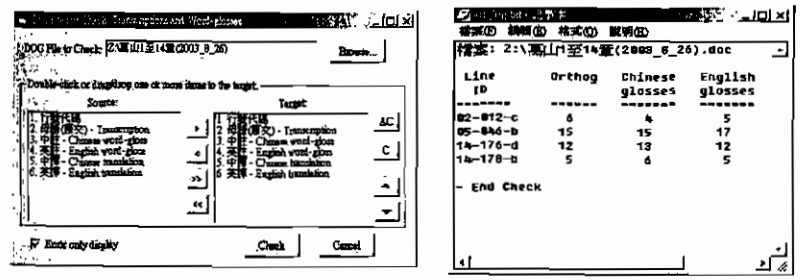
圖 4： AnnoTool 程式



3.2.3.2 Chkgloss

此外，也建立了 Chkgloss 程式檢查不同語言之間詞彙的數量，若發現不一致（如圖 5 所示），分析者必須回頭仔細檢查所分析過的語料。

圖 5： Chkgloss 程式



4. 如何運用南島語典藏進行教學

除了蒐集、整理及保存台灣南島語語料之外，台灣南島數位典藏提供一個完整的教學平台，教學內容從以下三方面進行:

（1）分布圖；

（2）語音；

（3）詞彙及簡單句。

所建構的分布圖介面可供使用者依語言及方言類別，選取一種方言後，可以聽此方言的語音，點進去可以同時觀察一種語音在某個字以及帶有相同語音的詞彙在句子裡的分佈。

此外，自從 96 年度開始發展網路辭典及參考語法，讓使用者更能夠進一步瞭解台灣南島語言的複雜性與差別。。

5. 未來展望

語言是表達思想、感情及傳遞知識的主要媒介，更是人們溝通不可或缺的工具。因此，語言的典藏，不但可成爲各種語言學研究的素材，更能提供當代語言、文化、教學、生活、社會的一扇窗口。

台灣南島語言典藏之任務主要是爲了保存並利用語言流傳的當代知識。然而在台灣南島語言瀕臨滅絕的危機下，「台灣南島語數位典藏」可確保語言消失後，語言的面貌仍得以保存下來。Cauquelin（To appear）從 1983 年開始蒐集、錄音並且記錄許多卑南族男女祭師所唸的祭文，同時將祭文翻譯成英文。此外，其分析語料的方式亦符合台灣南島語數位典藏之分析標準，此實爲卑南族後代留下寶貴的文化遺產，因爲至今已經沒有幾位男女祭師可以習讀此類祭文。我們相信此種呈現方式不僅可以促進原住民文化的復甦，同時也可維持多元的民族文化。

在文化的傳承與振興上，台灣南島語數位典藏具有深遠的影響，傳說故事能反映各族群（參見 Cauquelin 2004）對其歷史及移民的看法，其次口述語料往往或多或少也能反應當時的生活及社會的一面（參見 Zeitoun and Lin 2003）。

在學術上，台灣南島語言典藏的語料可作爲語言學（包括語音、詞彙、語法、語意等）、文學以及各種社會科學等的研究素材並提升此領域的研究及發展。

在教育上，台灣南島語言典藏的語料對當代語言的了解、詞典或文法專書的編纂可以作爲極佳的輔助工具。

以上報告只能很簡略地描述台灣南島語言數位典藏的主要架構，因此項工作非常繁重而且語言消失的速度非常快，希望更多專家學者能共襄盛舉，積極參與台灣南島語數位典藏的任務。

參考文獻

李壬癸（編輯）。2007。《高雄南島語言》。高雄縣文獻叢書系列。高雄：高雄縣政府。

曾思奇等。1988。《走過時空的月亮》。台中：晨星出版社。

Blust, Robert. 2003. *Thao Dictionary*. Language and Linguistics Monograph Series, No. A5. Taipei: Academia Sinica.

Cauquelin, Josiane. 2004. *The Aborigines of Taiwan. The Puyuma: From Headhunting to the Modern World*. London: RoutledgeCurzon.

Cauquelin, Josiane. To appear. *The Ritual Texts of the Last Religious Practioners of Nanwang Puyuma*. Language and Linguistics Monograph Series. Taipei: Academia Sinica.

Chu, Tai-hwa. 2003. Saisiyat texts. ms

Fey, Virginia et al. *The culture of the Amis*. Taipei: The Bible Society.

Hsin, Tien-hsin. 2002. Maga （Rukai） texts. ms

Li, Paul Jen-kuei. 1973. *Rukai Structure*. Institute of History and Philology, Speeial Publications No. 64. Taipei: Academia Sinica.

Li, Paul Jen-kuei. 1975. *Rukai Texts*. Institute of History and Philology, Special Publications No. 64-2. Taipei: Academia Sinica.

Li, Paul Jen-kuei. 1999. *The History of Formosan Aborigines -- Linguistic Perspectives*. Nantou: The Historical Commission of Taiwan Province. [In Chincse]

Li, Jen-kuei and Shigeru Tsuchida. 2002. *Pazih texts and songs*. Language and Linguistics Monograph Series, No. A2-2. Taipei: Academia Sinica.

Tsuchida, Shigeru（ed.）. 2003. *Kanakanavu Texts（Austronesian Formosan）*. Endangered Languages of the Pacific Rim. ELPR Publications Series A3-014.

Tung, T'ung ho et al. 1964. *A descriptive study of the Tsou language, Formosa*. Institute of History and Philology, Special publications No. 48. Taipei: Academia Sinica.

Wang, May Hsiu-mei. 2003. *Morphosyntactic manifestation of participants in Tona（Rukai）*. MA thesis. Taipei: National Taiwan Normal University.

Ye, Maya Yu-ting. 2003. Atayal texts. ms.

Yeh, Mei-li. 1991. *Saisiyat Structure*. MA thesis. Hsinchu: National Tsing Hua Univcrsity.

Yeh, Mei-li. 2000. Syntax and Semantics of the Saisiyat Negators. In V. De Guzman and B. Bender （eds） *Grammatical analysis: morphology, syntax and semantics: studies in honor of Stanley Starosta*. Occanic Linguistic Spccial Publication, No. 29, 258-273. Honolulu: University of Hawai'i at Mānoa.

Zeitoun, Elizabeth. 2001. Negation in Saisiyat: another perspective. *Oceanic Linguistics* 40.1:125-134.

Zeitoun, Elizabeth. 2002. Nominalization in Mantauran （Rukai）. *Language and Linguistics* 3.2: 241-282.

Zeitoun, Elizabeth. 2007. *A Grammar of Mantauran Rukai*. Language and Linguistics Monograph Series, No. A4-2. Taipei: Academia Sinica.

Zeitoun, Elizabeth and Tien-hsin Hsin. 2002. Glottal hopping in Mantauran （Rukai）. Paper read at IsCCL8, Taipei, November 8-10.

Zeitoun, Elizabeth and Lin, Hui-chuan. 2002. We should not forget the stories of thc Mantauran, Vol.2: Traditional folktales. ms.

Zeitoun, Elizabeth and Lin, Hui-chuan. 2003. *We should not forget the stories of the Mantauran, Vol.1: Memories of the past*. Language and Linguistics Monograph Series, No. A4. Taipei: Academia Sinica.

Zeitoun, Elizabeth, Yu Ching-hua and Weng Cui-xia. 2003. The Formosan Language Archive: development of a multimedia tool to salvage the languages and oral traditions of the indigenous tribes of Taiwan. *Oceanic Linguistics* 42.1:218-232.

Zeitoun, Elizabeth and Yu Chin-hua. 2005. Language analysis and language processing, *Computational Linguistics and Chinese Language Processing* 10.2: 167-200.

Zeitoun, Elizabeth and Chen-huei Wu. 2006. Reduplication in Formosan languages. In Chang, Henry Yung-li, Lillian M. Huang and Dah-an Ho （eds.） *Streams converging into an Ocean: Festschrift in Honor of Prof. Paul Jen-kuei Li on His 70th Birthday*, 97-142. Language and Linguistics Monograph Series W-5. Taipei: Academia Sinica.

中央研究院之台灣南島語數位典藏：http://formosan.sinica.edu.tw/
中央研究院之現代漢語平衡語料庫：http://www.sinica.edu.tw/SinicaCorpus
法國 LACITO 典藏網站：http://lacito.vjf.cnrs.fr/archivage/index.htm
德國 Max Plank Institute 縮寫原則網站：
http://www.eva.mpg.de/lingua/files/morpheme.html

台灣南島語語料庫詮釋資料發展之研究*

陳雪華
台灣大學圖書資訊學系教授

1. 語言與數位典藏

語言是所有知識保存的媒介、人類思想的工具，也是人際溝通與學習的重要手段。人類的智慧與文化，因語言而得到傳承；語言消失，文化就會失根、漂泊，語言的重要性不言而喻。台灣因地處亞太地區的樞紐，是亞洲到太平洋的門戶，地理位置的重要性非常明顯，而本島上的南島民族，其語言文化保存許多其他地區所沒有的特徵，台灣南島語言在南島語族中的地位舉足輕重，有愈來愈多的南島語言學者都承認台灣南島語的重要性。但由於種種歷史因果的影響，台灣島上各原住民族的語言生態，遭到嚴重的破壞，瘖啞而逐漸失語，若不及時搶救研究，將會越來越困難，因此研究與保存的工作刻不容緩（李壬癸 1999）。

台灣南島語因爲沒有文字的記載，一旦語言或方言的使用者不復存在，則該語言或方言便會面臨消失的命運。因此，如何在語者凋零、弱勢語言瀕臨滅亡的危機中保留珍貴文化遺產，是必須要關心的議題。除了以各種方式延續語言的生命外，更實際的作法即是進行語言的數位化典藏（高湘如等 2003）。即蒐集和保存現有的語料，將所蒐集的語料彙整、流通後，透過數位化形式，置於資料庫中，藉由良好的組織方式，將典藏資料呈現在網際網路上，提供使用者方便的檢索、分享及再利用。除可保存重要的文化資產外，亦滿足學者專家對語言之研究需求，並重現語言資料之活力與價值。因此，本研究旨在爲台大語言學研究所之台灣南島語語料庫擬定詮釋資料著錄格式及著錄範例，爲有效建置與檢索語言典藏資料提供一完善機制，以保存及典藏珍貴的台灣南島語語料。

2. 詮釋資料定義與功能

2.1 詮釋資料的定義

詮釋資料（Mctadata）最普遍的解釋是「data about data」可直譯爲描述資料的資料，即資料的描述性資訊。Locan Dempsey 和 Rachel Heery（1997）定義詮釋資料爲用來描述資料屬性的資料，用來支持如指示儲存位置、資源尋找、文件紀錄、評價、過濾等功

* 感謝國立臺灣大學資訊電子科技整合研究中心「多媒體整合實驗室」計畫的經費支援，並感謝臺大語言研究所蘇以文與宋麗梅教授研究團隊，以及圖書資訊學研究所陳怡妏同學的協助，使本研究得以完成。