談南島語語料庫典藏之數位化－從資料庫架構建立的觀點

余清華
中研院語言學研究所
台北市南港區 115
研究院路二段 128 號
886-2-27863300 ext. 362
harryyu@gate.sinica.edu.tw

周鳳瑛
龍華科技大學
桃園縣龜山鄉 333
萬壽路一段 300 號
886-2-82093211 ext. 647
fyc@mail.lhu.edu.tw

1. 前言

拜電腦科技之賜，語料庫(corpus)近幾年來不斷蓬勃發展，尤其用在語言典藏上助益甚多，但要進一步利用這些語料庫，實屬不易。理由是光有典藏，而沒有檢索機制的建立，將失去典藏之精神。所以如何將語料庫轉換成電腦的資料庫(database)，以供進一步檢索及應用，一直是語料工作者努力的目標。

傳統上，語言學家比較關心如何將語料呈現於平面媒體上，所以發明了很多符號及簡碼，以解決紙上(paper-and-pencil)作業的問題。語言學家在田野調查所收集到的語料(linguistic data)，通常會利用文書處理軟體予以建檔，最後付梓成書，成為書面格式(book format)之資料呈現，亦即，整篇文章(text)被視為一個物件，其所含的詞彙由左至右水平呈現，相對應的註解及翻譯則緊隨於後。這種文字上的安排，都只考慮到印刷上的問題，對於資料的深層結構仍未加以探究。例如，同樣的資料項可能有許多註解或翻譯，但為了印刷上的要求，祇得遷就於書面格式的篇幅與編排。因此，如何將電腦資料庫技術應用於語料的表達，乃語言學研究的新範疇。

本文將介紹國家數位典藏計畫之一的南島語語料庫，其乃利用現代資料儲存與擷取技術，將原始語料庫的檔案轉換成資料庫。但是，在應用電腦科技之前，我們必須正視並重新思考我們要表達的資料之本質為何？其資料項之間的關係為何？何種資料結構可以處理這些關係？以及資料間的關係要如何呈現給使用者？

因此，本文節次安排如后：1. 前言，2. 南島語語料庫與資料庫，3. 點-線-面 vs 單字-句子-文章， 4. 關聯式資料庫，5. 詞彙庫，與 6. 結論。

2. 南島語語料庫與資料庫

南島語語料庫為國家數位典藏的子計劃之一。基本上，這是一個書寫形式的語料庫(written corpus)，以國際音標(IPA)謄寫，並加上文字的註解與翻譯等資訊。其目標乃要收集南島民族語言，包括魯凱語、雅美語、鄒語、賽夏語、泰雅語、排灣語、布農語、阿美語以及卑南語等語言及所屬方言。原始語料均經過嚴謹的剖析及加註、翻譯，這些語料的加工部份均為中研院語言所齊莉莎女士親自審閱及不斷修訂，是一個頗具語言學內涵的語料庫。本文將以魯凱語(Rukai)的萬山方言(Mantauran)為例，表一所示為該語料庫之樣例。

001a. 1.

Onaʔi	ʔaamaðalaə-nai	ta-piʔa-aə-na-ða	po-aʟacə	ʔoponoho	m-ia
那	祖先-我們.屬格	處所名物化-動態.非限定:做-處所名物化-還-他.屬格	取-名	萬山	動態.虛擬式-這樣
that	ancestor-1PE.Gen	LocNmz-Dyn.NFin:do-LocNmz-still-3S.Gen	give-name	Mantauran	Dyn.Subj-so
我們的祖先萬山（自己是）萬山人。					
Our ancestors used to call (themselves) Mantauran.					

表一 萬山方言語料庫之樣例

這個語料庫建立於每個句子(sentence)的分析上，每一句的資訊均形成一個文字區塊(block)，由六行文字所組成。第一行表示行號代碼及文章代碼，其中行號代碼可再區分為兩部分，前面部份為前三碼的數字，表示段落代碼，後面部份為尾碼，以英文字母識別之，表示某一段落內的行數，a 表示第一行，b 表示第二行，其餘類推；第二行表示母語本身(以 IPA 拼寫)；第三行及第四行分別表示中、英文註解(glosses)；最後兩行則分別表示中、英文翻譯(translation)。

該計劃目前使用微軟資料庫產品 Access 作為關聯式資料庫管理系統。在將語料庫轉換為資料庫之前，我們必須檢視原來書面格式的檔案結構與文字規劃(lay out)，然後撰寫一個轉換工具程式，以將其轉成機讀格式(computer-readable format)的資料檔，並把所有剖析過的資料項分別放到相關的資料表上。日後資料一旦擴增，則可考慮較大型的資料庫軟體(如 SQL Server)，以增進查詢效益。

3. 點-線-面 vs 單字-句子-文章

書寫語言(written language)的基本組成元素為單字，然後單字再組成句子，句子進一步組成文章，最後便構成了語言面本身。將「單字-句子-文章」

(word-sentence-text)想像成「點-線-面」的類比，從最小的單位，即單字(word)著手，我們以此作為紀錄(record)單位，再細分為若干資料項(欄位)，依此產生單字層級(word-level)的資料表，如表二所示。接著，我們也以同樣方法產生句子層級(sentence-level)的資料表，如表三所示。這兩個表格均衍生於表一的句子資訊。乍見之下，似乎令人覺得多此一舉，其實不然。誠然，我們已從書面格式的水平呈現轉變為垂直格式的資料表達了。

更重要的是，資料的定義更為嚴謹，而且資料的組織結構(organization)與呈現(presentation)彼此獨立。一旦我們對於資料正確地加以結構化，那麼將來要用任何格式來呈現語料，將是一件輕易的事。以表二為例，母語拼寫(orthog)的形式可透過 wordorder 欄位來還原成原來的句子；而整個段落也可利用 location 欄位及 wordorder 欄位而完成，關於進一步的說明，詳見於後續章節。從這些資料表可看到，資料與資料之間的連結更為明確，不必讓研究者去費心建立。

location	wordorder	orthog	cgls	egls	textid
001a	0	onaʔi	那	That	1
001a	1	ʔaamaðalaə-nai	祖先-我們.屬格	ancestor-1PE.Gen	1
001a	2	ta-piʔa-aə-na-ða	處所名物化-動態.非限定:做-處所名物化-還-他.屬格	LocNmz-Dyn.NFin:do-LocNmz-still-3S.Gen	1
001a	3	po-aɭacə	取-名	give-name	1
001a	4	ʔoponoho	萬山	Mantauran	1
001a	5	m-ia	動態.虛擬式-這樣	Dyn.Subj-so	1

表二 單字層級(word-level)的資料表 –「單字與註解」

location	c_freetran	e_freetran	soundpath	textid
001a	我們的祖先萬山（自己是）萬山人。	Our ancestors used to call (themselves) Mantauran.	Mantauran/001a.mp3	1

表三 句子層級(sentence-level)的資料表 –「翻譯」

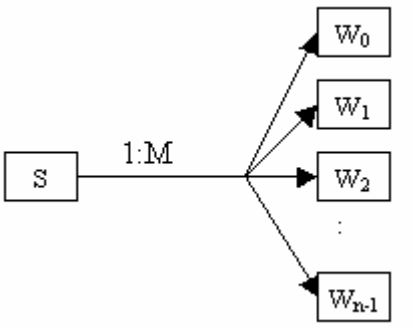
截至目前為止，我們考慮到資料的組織結構，相同句子的資料被分散到兩個資料表，此外，我們儘量對資料庫進行正規化，使其資料項不會重複(redundant)。即使如此，為了了解資料本身，我們必須知道以上資料表的 location, orthog, cgls, textid, e_freetran 等各自代表什麼意義，一旦我們明瞭這些，那麼也就自然明瞭欄

位和紀錄為何如此建構，以及彼等資訊為何置於不同資料表上。我們相信，資料庫的結構在某種程度上反映了語言學上的理論。

4. 關聯式資料庫

關聯式資料庫(relational database)的精神在於每一筆資料至少有一識別代碼(identifier)，以指出該筆資料(例如，字或句)出自於何處(例如，哪一篇或者哪一段落)，如此，利用這些識別代碼，我們就可以產生任意組合的句子形式。

從表二與表三的資料表來看，我們不難發現，這些資料表含有相同欄位可以串聯起來，形成了關聯式資料庫。再者，由於一個句子包含多個單字，而一個單字只屬於一個句子，此種資料架構便構成了典型的主要/明細(Master/Detail)關係，如同客戶訂單系統，一位客戶可有許多訂單，但一個訂單只屬於一位客戶所有；又如同學生選課系統，一位學生可以選修多門課程，但是一門課程只屬於一位學生所選修。換言之，兩者之間存在著一對多(One-to-Many)的關係，如圖一所示。當我們把這種關係實作於資料庫管理系統(如 Access 或 SQL Server 等)時，利用 SQL 查詢語法即能方便且迅速地檢索出不同的結果。當然，資料庫查詢的好處絕對不僅於此，我們也可還原成原始語料樣式(raw data)，或者從現有資料庫內容再萃取詞彙庫(lexicon)或標記集(tagset)等。



圖一 One-to-Many 關係圖

首先，讓我們來看看如何從資料庫還原為原始語料的樣子。從前述可知，句子層級的資料表(名稱為*翻譯*)與單字層級的資料表(名稱為*單字與註解*)兩者之間存在著一對多(One-to-Many)的關係，換言之，兩者之間有著親子關係(parent/child relationship)的存在。例如，一個句子對映到一些不同的單字，這種關係最適合使

用 Data-shaped 的 SQL 子句來表達，使用這個子句意謂著不用 join 以及複雜的過濾(filter)動作，更不需要拼湊型程式碼(spaghetti-code)來製作呈現邏輯。此指令簡單、易於瞭解，降低了網路流量(network traffic)，且提供了與 XML 工具整合的介面。我們使用 Shape 語法如下：

SHAPE {SELECT * FROM 翻譯} APPEND ({SELECT * FROM 原文與註解 RELATE Location TO Location}) As rsDetail

以上 Shape 命令傳回一個紀錄集(物件)，其內又包含另一紀錄集(物件)，亦即建立了紀錄集中的紀錄集(recordsets within recordsets)，如果我們將此一傳回的物件進一步檢視，將得到如表四所示的欄位結構：

欄位名稱	資料類型
location	文字
c_freetran	文字
e_freetran	文字
soundpath	文字
textid	數字
rsDetail	物件

表四 主要物件的欄位結構

其中最後一個欄位為 rsDetail，是一個物件，其所包含的欄位結構，如表五所示。

欄位名稱	資料類型
location	文字
wordorder	文字
orthog	文字
cgls	文字
egls	文字
textid	數字

表五 rsDetail 物件的欄位結構

當我們要進行資料呈現時，就可以針對這兩個物件予以操作一番。首先，為了呈現母語的原文時，我們可以使用 rsDetail 欄位(物件)，也就是說，根據其

wordorder 欄位將母語的拼音形態、中英文註解等欄位值依序列出，由於我們的資料庫是針對單字為基礎所設計的，所以在語料輸出上可以適當地對齊(alignment)，利用表格的框線可以很容易的框住每字及其註解資料，其效果等同於原始的語料檔格式，如表六所示。

0	1	2	3	4	5
onaʔi	ʔaamaðalaə-nai	ta-piʔa-aə-na-ða	po-aɭacə	ʔoponoho	m-ia
那	祖先-我們.屬格	處所名物化-動態.非限定:做-處所名物化-還-他.屬格	取-名	萬山	動態.虛擬式-這樣
that	ancestor-1PE.Gen	LocNmz-Dyn.NFin:do-LocNmz-still-3S.Gen	give-name	Mantauran	Dyn.Subj-so

表六 rsDetail 物件所展現的語料輸出

同理，該句的中英文翻譯及其他資訊等，亦很容易地予以取出，我們可將其內容附加在表六文字的後面：

我們的祖先萬山（自己是）萬山人。

Our ancestors used to call (themselves) Mantauran.

因為此句是屬於文章代碼(textid)為 1，且行號(location)為 001a 的資料錄，所以可以很快地還原成原始語料(見表一)。當然，我們的語料不祇一句或一段而已，在龐大的語料中也可進行部分篩選，其法是使用 WHERE 子句來調整查詢。

以上是屬於逐句的語料輸出，如果要做整個段落(全文)的母語輸出的話，仍然可依據上述語法，再加上 WHERE LEFT(location, 3) = ‘001’ AND textid = 1，就可以把第 001 段抽離出來，再逐句加以列出。

由上可知，一旦資料庫的結構建立完整之後，要做出多樣化的查詢其實不難。而南島語語料庫數位化的結構即依循語言學的原理予以設計，故，語言學家即可利用該資料庫查詢的功能，進一步理解及處理該語料的輸出。

5. 詞彙庫

母語的全文顯示是由一系列的拼音形態所形成的。利用表二的單字層級資料表中之 orthog 欄位，我們也可建立某一方言的詞彙庫(lexicon)，至少就目前所收集的語料而言。下列 SQL 語法可產生按字母排列而且沒有重覆的詞彙集：

SELECT DISTINCT orthog FROM 原文與註解 ORDER BY orthog

甚至，我們也可檢索出某單字在文章中的分佈情形等。這些都是語料庫轉換為資料庫的優點。資料庫是由若干資料表所組成，有些欄位名稱使用代碼或簡碼之類的識別字，可讀性較低，但是經過資料庫的關聯特性及操作，就可以產生易於閱讀的文字輸出。

6. 結論

對於語料庫的應用，語言學家關心的是如何呈現該語言原來的面貌，而電腦科學家則希望能將語料加以組織及結構化，再導入資料庫技術，以應付使用者不同的檢索需求。因此，語言典藏數位化一方面將克服傳統紙筆技術的問題，另方面也可摒棄書面格式的語料輸出，而這些理想都必須有賴電腦關聯式資料庫的技術予以達成。

從書面格式的語料庫進展到關聯式資料庫，代表著複雜度的增加，但是也在資料的有效運用及操控性上獲得相對的回報。細究之下，複雜度的增加並不是真實的，那些不同但相連結的資料表都可被認為是與語言學家的專業知識更加密切關聯。資料庫理論無疑是如何設計欄位、紀錄及資料表，正如同語言學要如何呈現單字、句子及文章一樣，在彼此之間建立一個緊密而有效的連結。

本文所介紹的數位化南島語語料庫即利用現代資料儲存與擷取技術，以電腦的資料結構將原始語料庫的檔案轉換成資料庫。其中，對於語料庫的結構化、與正規化，乃利用關聯式資料庫的精神，一方面將語料資料定義的更為嚴謹，另方面對於資料與資料之間的連結也更為明確。雖然本文所述之議題可能已超過實際的技術問題，但透過南島語語料庫數位化計畫的嘗試，相信對於未來語料庫的研究將有著極深刻的影響。

參考文獻

中文部分

林惠娟，1999，我們來說萬山話(冊 1-6)，文鶴，台北

葉榮木，1999，資料結構使用 Visual Basic，松崗，台北

溫賢發譯，2001，程式設計實務：SQL 資料庫技術，碁峰，台北

詹嘉文、王秀卿譯，2001，SQL 程式設計入門手冊，麥格羅．希爾國際出版公司，台北

黎逸凡、黃淑敏、張家齊譯，2000，Hitchhiker's Guide to Visual Basic and SQL Server 應用與設計(第六版)，Microsoft，台北

英文部分

Leech, G., Myers, G., and Thomas, J. (eds), 1995 Spoken English on Computer. Longman Publishing, New York

Jacobson, M., Michailovsky, B., and Lowe, B., 2000. *Linguistic documents synchronizing sound and text*, Speech Communication 33 (2001) 79-96.

Zeitoun and Lin. To appear. *We should not forget the stories of the Mantauran, Vol.1: Memories of the past*. Submitted to the Institute of Linguistics, Academia Sinica.