標點符號在語言資料庫中的設計與實作
Design and Implementation of Punctuation in a Linguistic Database

余清華
中研院語言所
台北市南港區 115
研究院路二段 128 號
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
fyc@mail.lhu.edu.tw

摘要

在現代的書寫語言中，標點符號與語言文字不可分割。在語言保存的過程裡，標點符號應被視為一個典藏物件。然而要如何客觀地與文字保存在一起，並且發揮語言典藏的功能，則是本文探討的重點。本文嘗試用關聯式資料庫的概念與方法，將每個文字視為記錄，而把標點符號看作為相關的欄位屬性，希望藉此能有效地儲存完整的語料，這樣不但可成功地回復原始的語言面貌，還具有資料庫本身提供的額外優點。

關鍵詞：標點符號，語料庫，資料庫，數位化，自然語言處理

1. 前言

在理論或計算語言學的研究範疇，標點符號的處理往往被忽略。一方面，它缺乏一套明確而詳細的符號功能理論，以適用於計算領域，另一方面，也因為寫作習慣的不同而有個別差異。所以這方面的研究不易探討。[1] 就現代的書寫語言而言，文字與標點符號是不可分開的。換言之，標點符號對於文字的處理及語言理解自有其重要性，甚至被認為是一種禮貌、智慧的表徵。[7]

事實上，沒有考慮標點符號的存在，要進一步了解及處理書寫語言將是不可能。雖然有人會看待標點符號不過是反映朗讀語氣的工具而已，但是從語言學的角度來看，其本身可能更具有語法或語意結構上的意涵。[3]

從形態上而言，標點符號乃書寫語言上的一個明顯且具有獨立特質的記號(token)—可視為一個單獨的拼寫物件(orthographical component)。一旦我們認清此一事實，那麼本文就有充分理由研究它。

本文在這裡並不強調標點符號在書寫上的重要性，這類說明在坊間的文法書籍或寫作指南早已俯拾皆是，而是要從「台灣南島語數位典藏」的實務經驗中，說明標點符號應如何與文字一起保存，卻又分開處理的原則與方法。

「台灣南島語數位典藏」為「中央研究院國家典藏數位化計畫」下分項主題「語言典藏」的子計畫之一。基本上，這是一個書面形式的語料庫(written corpus)，使用國際音標(IPA)記錄台灣瀕危語言的原住民族語，輔以文字的語法註解(gloss)與翻譯(free translation)等資訊。其目標乃要收集大多數的台灣南島語，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，建立一個語言資料庫。原始語料均經過嚴謹的分析、註解及翻譯，這些加工的部份均為中研院語言所齊莉莎女士親自審閱及不斷修訂，是一個極具語言學內涵的典型多語語料庫。[8] [9] [10]

語料庫的原文完全依據羅馬拼音的方式，若有語音無法用字母表達，則使用 IPA 符號。句中的字與字之間以空格區隔，每句至少有一個標點符號，即 full stop，但其間使用逗號、分號或引號的情形，亦所在皆有。本文嘗試將標點符號視為一個獨立的處理元素，即一個典藏的記號物件，希望從客觀描述的角度找出有效的儲存方式並將其保存於資料庫。

另一方面，由於南島語句子的記音轉寫部份採取與歐美語系同一套的標點符號集，所以，為了說明起見，在某些情況下，我們會使用英文例句。

本文的結構說明安排如下：第一節說明本文處理標點符號的基本背景，第二節說明詞語的定義，第三節說明句子的界定方式，第四節則說明語料庫的內容架構，第五節提出標點符號的資料庫解決方案。最後，第六節則總結本文的發現並提出新的議題，以提供未來可以研究的方向。

2. 何謂詞語？

Kucera 與 Francis (1967)對於詞語(word)的定義如下：一個詞語是由連續的字母與數字符號所構成的字串，兩邊用空格予以隔開；且可能包含連字符(hyphen)及撇號(apostrophe)，但不包含其他的標點符號。[6]

基於此定義，只要在句中找到有空格隔開的字串，便可認定其為詞語。起初，這種判定方法似乎很合理，也很便利，但事實卻不然。標點符號有時會尾隨所謂詞語的後面。要移除這些標點

符號雖然很容易，但是以英文為例，有些標點符號可能代表縮寫，如 etc.或 Calif.等。

所以，在這種情況下，我們便要辨別 Washington 或大寫的動詞 Wash。

…in Seattle, Wash.
Wash your hands.

幸運的是，我們處理的對象並沒有這種縮寫形式，故可以避開這個問題。然而，跨越詞組或子句的引號、引號內夾用其他符號、甚至省略號(ellipsis)的使用，時常見於語料庫。當語料庫要轉換成資料庫或其他形式(如 XML)時，這些符號都不能省略，因為標點符號也是語料的一部分。請見下列例句：

(1) "mit qilaˊ " so-n-nahaᵐ .
People said, "The ass is lazy."
(2) "iaʔə, oəla-ka-nomi lo ʔomaləŋa."
"Sleep well too."
(3) "ʔo..." la ya.
"Oh...," he replied.

另一個要注意的是連字符。在語料庫所記錄的詞彙中，連字符(-)是用來連接最小且有意義的構詞單位，即詞綴(affix)，最後則構成一個完整的詞語(或稱單字，word)，如上例所示。

吾人皆知，語言是由句子(sentence)所組成，句子則是由詞語(word)所組成。然而細察之下，句子其實不光包含詞語，還可能包含數字(number)或標點符號(punctuation)。在亞洲語系(中文、日文、韓文)，字與字之間並不以空格隔開，且標點符號也不同於歐美語系。因此斷詞(segmentation)是一件頗具挑戰性的工作。相反地，這個語料庫的書寫採用羅馬拼音系統，故可從有利的資訊角度來處理語言中的文字及符號物件。

3. 何謂句子？

明確地說，句子是一個長長的字串，由以空格隔開的若干記號(token)所組成，而且每個記號可以是詞語、數字或標點符號。通常，句尾處會有一個代表句子結尾的 full stop，而句中可能穿插其他標點符號。[5]

一個完整句在句尾處會以句號(.)、問號(?)或驚歎號(!)表示結束，這些符號稱為 full stop。但是問號也可能出現在引述句裡，如下：

(4) "oai-ŋa-ta olopo?"
"Where shall we go hunting?"

在上例中，句尾處有兩個符號，一為問號，一為引號，但它們屬於不同類的符號，不能刻意放在同一欄位。關於此，容後說明。

有了以上對於詞語與句子的認識作為前提，語料庫的文字處理，便可以考慮下列標點符號、詞語與句子的關係：

1. 任何推定的句子後面會出現.?!
2. 引號或括號可能跨越詞語、詞組或句子
3. 取消常見縮寫的句點，避免與句尾符號衝突
4. 符號.?!視為句子邊界(sentence boundary)
5. 空格表示詞語邊界(word boundary)

於是，我們可以根據這些概括的關係和規則，得到句子和詞語的邊界判定演算法，如下所示：

句子邊界 =
句點 + 空格 + 另一個句子
或者 句點 + 引號 + 空格 + 另一個句子
或者 句點 + 空格 + 引號 + 另一個句子

詞語邊界 =
空格 + 另一個詞語
或者 引號 + 空格 + 另一個詞語
或者 句點 + 空格 + 另一個詞語
或者 句點 + 引號 + 空格 + 另一個詞語

4. 語料庫的正規化

原始的語料庫將每篇故事(或稱文本)放在不同的檔案上，而且每篇故事以句子為基礎加以分析及翻譯。研究者必須從不同檔案尋找所要處理的對象，然後將這些對象合併起來，以產生一個機讀格式的 flatfile。

請檢視下列 flatfile 的例子，注意，每行開頭的數字在此僅作說明，實際上不存在於語料庫：

1: DRUMn-01-001-a
2: onaʔi ʔaamaðalaə-nai ta-piʔa-aə-na-ða po-aɭatsə ʔoponoho m-ia.
3: 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
4: that ancestor-1PE.Gen LocNmz-Dyn.NFin:do-LocNmz-still-3S.Gen give-name Mantauran Dyn.Subj-so
5: 我們的祖先萬山自稱是萬山人。
6: Our ancestors used to call (themselves) Mantauran.
7: DRUMn-01-001-b
8: ðonaʔi a-kaavaʔi-ŋa-ða ʔaomo mani aɭa ʔitsaʔotsaʔo ʔina vaha-nai ʔoponoho la tali-ɭao-ɭaoðo.
9: 那 分句名物化-動態.非限定:來-已經-他.屬格 日本人 就 動態.非限定:拿 動態.虛擬式:學 這 話-我們.屬格 萬山 和 屬於-重疊-下面
10: that ClsNmz-Dyn.NFin:come-already-3S.Gen Japanese then Dyn.NFin:take Dyn.Subj:learn this language-1PE.Gen Mantauran and part+of-Red-below

11: 日本人來了以後就（開始）學我們萬山和屏東縣（排灣及魯凱族）的話。
12: When the Japanese came, they started (to learn) our language as well as (the languages spoken by the people living in) Pingtung county.
...

由上可知，語料庫乃建立在每句分析的基礎上，每句的相關資訊形成一個區塊(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(glosses)；最後兩行則分別表示中、英文翻譯。

從這種橫向的 flatfile 轉換到縱向格式的資料庫，須撰寫一個資料庫轉換程式。[10] 惟本文擬針對詞語層級的資料庫，深入探討詞語與標點符號的關係與處理。

資料庫設計的首要之務就是解決標點符號的問題。[4] 標點符號雖非詞彙的一部份，但卻是句子不可或缺的記號物件。台灣南島語語料庫對於標點符號並未特別獨立出來進行標記，不過在原始句子仍使用標點符號，此意謂著記錄人對於該語言的理解與詮釋。

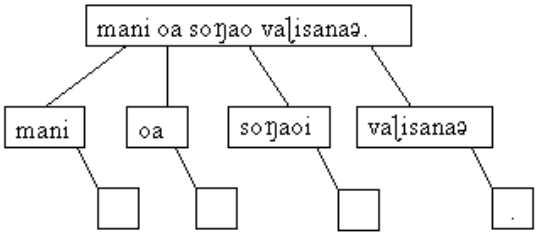
從語言典藏的觀點，我們不能忽略標點符號的存在，反而要將其視為獨立的物件，與其他文字並列，共存於資料庫的欄位，以便日後擷取。

5. 標點符號的解決方案

首先，讓我們檢視一個典型的句子：

(5) mani oa soŋao vaḻisanaə.
Then, it was the Bunun village Valisanae.

轉成階層結構後，得圖一：



圖一 例句(5)的樹結構

當句子進行剖析時，標點符號會被視為單獨記號，別於詞語本身。但是，如果記錄以詞語或標點符號為基礎，那麼就會產生不同類別的記錄以及不同組的欄位。如果把詞語間的空格也視為一種標點符號的話，那麼每個詞語可以說具有一個相關聯的標點符號。雖然這些符號代表著詞語邊界，但在實際上，它則鄰接於前一個詞語。據此，我們可以將標點符號視為前一個詞語的欄位。所以相對應的資料表，如表一所示：

location	wordorder	orthog	punct
DRUMn-01-001-a	0	mani	
DRUMn-01-001-a	1	oa	
DRUMn-01-001-a	2	soŋao	
DRUMn-01-001-a	3	vaḻisanaə	.

表一 對應於圖一的資料欄位

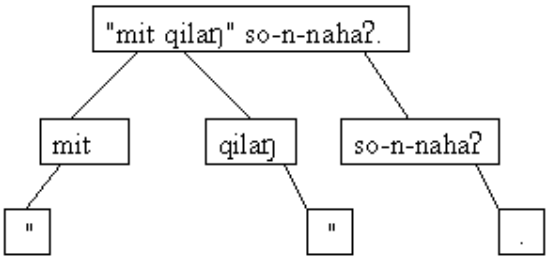
每一筆記錄都建立在一個明顯的語言單位上，即詞語。每個詞語(*orthog*)的對應資訊則放在其他欄位上。*location* 表示所屬句子的識別字，*wordorder* 表示詞序，而 *punct* 則存放標點符號(含空格)。

這個表格以縱向格式呈現了句子結構，取代了我們過去所熟悉的橫向格式。更重要的是，資料的定義更加嚴謹，資料組織與資料呈現的問題予以分開。如果想要組成任何格式的資料，將隨心所欲。例如，使用上面資料表的詞語與標點符號欄位，就可以還原一個標準的文本句。

然而，語料庫的內容並不如此單純，許多句子都含有引號：

(6) "mit qilaɻ " so-n-nahaᵐ .
People said, "The ass is lazy."

轉成階層結構後，得圖二：



圖二 例句(6)的樹結構

顯然地，我們不能依據表一將標點符號放到欄位上。引號是成對的，分成左右兩個，其含括的範圍可能是一個字、一個詞組或一個子句。括號使用也是同樣情形。因此，有必要考慮兩組的標點符號：

1. 一般符號 (punct)：. , ; : ! ?
2. 引號或括號 (pul & pur)：' " [ ] ( )

根據這樣的劃分，上句放到資料表的情形，將如表二所示：

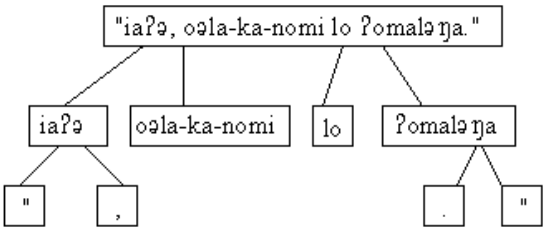
location	wordorder	orthog	punct	pul	pur
TAYSq-01-001-c	0	mit		"	
TAYSq-01-001-c	1	qilaɻ			"
TAYSq-01-001-c	2	so-n-nahaᵐ	.		

表二 對應於圖二的資料欄位

同樣地，這個處理模式也可套用至引號跨越整句的例子：

(7) "iaʔə, oəla-ka-nomi lo ʔomaləŋa."
"Sleep well too."

首先，我們將其視為一個整體，然後利用空格作為區隔符號，找出候選詞語，進而取出標點符號。該句剖析樹如圖三所示：



圖三 例句(7)的樹結構

此句雖複雜，但轉成階層結構後，標點符號位於詞語節點之下，成為直屬的子節點(immediate daughter)。放到資料表後，可得到表三：

location	Word order	orthog	punct	pul	pur
DRUMn-11-156-b	0	iaʔə	,	"	
DRUMn-11-156-b	1	oəla-ka-nomi			
DRUMn-11-156-b	2	lo			
DRUMn-11-156-b	3	ʔomaləŋa	.		"

表三 對應於圖三的資料欄位

在這樣的設計理念下，詞語本身與其後面(或前面)緊鄰的標點符號便隔開了，但卻屬於同一筆記錄。語句的還原工作便是連接詞語(orthog)與標點符號(punct)欄位，然後檢查該詞語左右是否有相關聯的引號或括號，若有的話，便加上對應的符號。

所以，我們可利用 SQL 指令寫出連接欄位的運算式：[2]

TRIM(pul) + CONCAT(orthog, punct) + TRIM(pur)

上式使用下列規則：

1. 若 pul 欄位有非空白的值，則使用該值，否則使用空字串。
2. 若 pur 欄位有非空白的值，則使用該值，否則使用空字串。
3. 若 punct 欄位有非空白的值，則直接使用該值，否則使用一個空格。

對於規則 3，我們之前已將空格(word boundary)視為詞語所關聯的標點符號，所以大多數情形符合前述的公式。

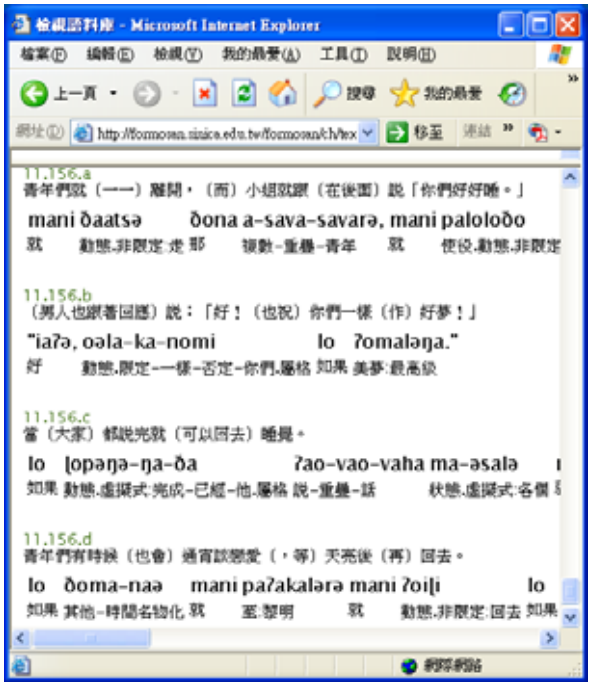
例如，當引號夾住肯定句結尾時，最後一字的組成序列可能為：

左引號	文字	.	右引號

根據上述公式，因 punct 欄位已有非空白的值，故一切正常。但是，當文字位於句首或句中時，若 punct 欄位是空白且右邊有引號時，依上述處理原則，將有一個空格遺留在文字與右引號之間。所以我們必須試著將該空格移至字串後面，修改的式子將如下：

TRIM(pul) + CONCAT(orthog, TRIM(punct)) + TRIM(pur) + SPACE(1)

最後，在相同句子編號的索引下，經由上式所產生的字串輸出加以串接，即可還原為原始的語句了，如圖四所示。



圖四 資料庫輸出範例

在此，我們要強調，本文純粹是從語言客觀的角度來描述文字及標點符號的保存。至於符號用法的正確性與其語法或語意上的探討，並不在本文討論範圍內。

6. 結論

標點符號在自然語言處理以及語言理解上扮演重要的角色。它不再是朗讀語氣上的一個暫停標記，而更應視為語法或語意上的一個有效節點。沒有它，要了解及剖析自然語言則不可能。

本文從實證經驗中探討標點符號在資料庫上的設計和實作，然後又可從資料庫還原為原來的語料庫，甚至還有其他更多查詢的可用性。然而，這僅限於事先有良好結構的語料庫格式，面臨 E 世代的網路資源，充滿著無數的語言素材(linguistic material)，我們將難以處理五花八門的語言表達形式，如 c@p、Micro$oft、h*ll 等字母取代的情形或者 :-)、:-D、Q_Q 等表情符號。

此外，在計算語言學的符號處理上，我們發

現可能需要一套適合的標點符號理論，用以發揮其在語言文字處理上的積極角色，以有效處理人類語言豐富而多樣的表現形式。

參考文獻

[1] Bayraktar, Murat and Say, Bilge and Akman, Varol. 1998. *An Analysis of English Punctuation: The Special Case of Comma*. International Journal of Corpus Linguistics 3 (1): 33-57.

[2] Houlette, Forrest. *SQL: A Beginner's Guide*. New York: McGraw-Hill, 2001.

[3] Jones, Bernard. 1994. “Exploring the Role of Puncutuation in Parsing Natural Languages”. In *Proceedings of the 15th International Conference on Computational Linguistics* (COLING-94). Kyoto, Japan, 421-425.

[4] Knowles, Gerry. Converting a corpus into a relational database: SEC becomes MARSEC. In Leech, G., Myers, G., and Thomas, J. (eds), *Spoken English on Computer*. New York: Longman Publishing. pp.208-219, 1995.

[5] Kucera H. and W.N. Francis. 1967. *Computational analysis of present-day American English*. Providence: Brown University Press.

[6] Kucera H. and W.N. Francis. 1982. *Frequency Analysis of English Usage: Lexicon and Grammar*. Boston, Massachusetts: Houghton Mifflin.

[7] Nunberg, Geoffrey. 1990. *The Linguistics of Punctuation*. Standford, California: CSLI Publications.

[8] Zeitoun et al. 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. Volume 42, no. 1. Hawaii: University of Hawai‘i Press.

[9] Zeitoun, Elizabeth and Yu, Ching-hua. 2004. “The Formosan Language Archive: Language Processing and Linguistc Analysis”. The 1st International Joint Conference on Language Language Processing (IJCNLP-04) Asian Language Resources Workshop. Sanya City. Hainan Island, China. 25 March, 2004 .

[10] Zeitoun, Elizabeth and Yu, Ching-hua. “The Formosan Language Archive: Language Processing and Linguistc Analysis.” International Journal of Computational Linguistics & Chinese Language Processing. 10.2. June 2005.