中文連音二字詞之語音合成

Coarticulation of Two-Syllable Words in Mandarin Speech Synthesis

Jun-Wen Hwang ， Ming-Shing Yu ， Shyh-Yang Hwang and Ming-Jer Wu

（黃 志 文）　（余 明 興）　（黃 世 陽）　（吳 明 哲）

Department of Applied Mathematics

National Chung-Hsing University

Taichung , 40227, Taiwan

E-mail ： MSYU@DRAGON.NCHU.EDU.TW

摘　要

本篇論文最主要在研究語音合成中的連音部份，我們從所錄製好的連音庫中，切取出連音二字詞，然後將之細分成三部份，針對每部份使用單音重新組合而成。在重新組合的過程，利用已知的連續音資訊，從單音中抓取最像連續音的部份來合成連音二字詞，並且模擬它的連音情形，例如考慮音量、基週走勢、音長等，使得日後在合成連續語音時，能達到類似連續語音的自然流暢。

1. 緒論

1.1 連音型態與特性

John R. Deller 等人在所著一書中[1]提到有關連音（ Coarticulation ）的解釋爲：『在語音的產生過程中，發音器官在產生一連串所須要的音素時，爲了達到語音的自然流暢性，發音器官的變化是平滑的，而連音正是從這平滑的變化過程中所產生的』。

在中文連音的研究方面，近來有陳志祥[7]在中文連音型態之初步研究一文中指出，連音型態分爲三個基本型態：

(1) 停頓連接

在兩音節間有一段靜音存在，通常發生在詞和詞之間，如（圖 1-a ）所示，本圖中爲『辛勞的播種』中的『的播』。

(2) 緊密連接

在兩音節間幾乎沒有停頓，但兩音節間並無重疊的波形存在。通常在詞內這種情形較常發生（圖 1-b ），本圖中爲『風吹草動』中的『風吹』。

(3) 重疊連接

在兩音節間，不僅不存在靜音，且其基週呈現出連續變化的情形，兩音節中間會有一段轉換過程中過渡的週期波存在（圖 1-c ），本圖中爲『政府官員』中的『官員』。

更詳細的說，停頓連接會發生於呼吸群[2]（李琳山教授於演講時稱之爲韻律段， Prosodic Segment ）結束和下個呼吸群開始之間。而緊密連接和重疊連接則會發生在一個呼吸群之中。至於是緊密連接還是重疊連接，則視此二字詞的結構而定，若後音節的子音是具週期性的子音，如 ㄇ、ㄋ、

ㄌ、ㄖ。或是沒有子音，單就介音（ㄧ、ㄨ、ㄩ）開始，或是只有母音，則屬於重疊連接，否則都屬於緊密連接。如表一所示：

	後音節之起始音
緊密連接	ㄅ、ㄆ、ㄈ、ㄉ、ㄊ、ㄍ、ㄎ、ㄐ、ㄑ、ㄒ、ㄓ、ㄔ、ㄕ、ㄗ、ㄘ、ㄙ。
重疊連接	ㄇ、ㄋ、ㄌ、ㄖ、ㄏ、ㄧ、ㄨ、ㄩ、ㄚ、ㄛ、ㄜ、ㄝ、ㄞ、ㄟ、ㄠ、ㄡ、ㄢ、ㄣ、ㄤ、ㄥ、ㄦ。

表一　緊密連接和重疊連接所相對應的後音起始音

此處較爲特殊的是以 ㄏ 爲起始的後音，會有重疊連接的情形，基本上是因爲連續語音中，ㄏ 常被省略之故。第 2.2.2 節會提到關於ㄏ的幾個例子。

1.2 研究方向及論文架構

在本篇論文中，因爲我們最終的目的是達到中文語音合成的自然流暢性，就時域合成而言，在停頓連接和緊密連接方面，假如不考慮音長所影響的因素，則並不會影響合成時的流暢性。所以在本篇文章的研究方向，是將整個焦點集中於重疊連接上，希望利用單音來合成連續語音時，在連接部份能模擬重疊連接，達到類似連續語音的自然流暢性。

本篇論文的整個架構敘述如下。在第二節中我們將討論研究進行的方式，介紹本論文所用到的語音分析技術和方法。在第三節則是本論文的重點所在，描述如何使用單音來合成連音二字詞。第四節則爲合成結果的度量實驗。第五節是本篇論文的總結，及未來可努力的方向。

2 研究進行方式

2.1 單音庫與連續音庫

爲了研究連續語音中的連音現象，我們請了一位女性錄音員，錄製好單音庫和連續語音庫，其取樣頻率（ Sampling Rate ）均爲 12 kHz 。在單音庫方面，因爲我們是利用單音來合成連續音，所以在單音的錄製方面，我們分成前音和後音，其中前音是所錄製的二字詞中的首字，而後音則是其末字。

在連續語音庫方面，我們依據某些報紙文章錄製而成，整個語音檔資料總共錄製有 55,786,883 Bytes ，共 13,999 個中文單字，共約一小時又十七餘分鐘，平均每個音長爲 330ms ，相當於每秒發 3.03 個音。連續語音庫提供我們在使用單音來合成連續語音時，如何模擬產生連續語音中重疊連接的連音段，進而提高語音合成的自然流暢性。

2.2 動態時間校準演算法之應用

由於我們所錄製的單音，用來合成連續音時，大部份都較真正的連續音爲長，所以我們必須從單音中抽取出真正用來合成連續音的部份，因此我們使用 DTW 來做單音和連續音的比對。首先，以中文單音爲參考樣本（ Reference ），再從連續音庫中，取出相同的音十個做爲測試樣本（ Test ）。又因爲各樣本的長度並不一致，所以我們以音框 （ Frame ）爲單位，音框的長度固定爲 20 ms ，測試樣本固定爲 50 個音框，參考樣本固定爲 80 個音框，音框和音框之間的重疊部份（ Overlap ），則視樣本的長度而定。也就是說，彈性調整音框重疊部份，使得測試樣本音框數固定成 50 個音框，參考樣本音框數固定成 80 個音框。

在此處的測試樣本音框數 50 及參考樣本音框數 80 的訂定，在測試樣本部份是依據在連續音庫中連續音的長度而定，根據對連續音庫中的個別音節做統計的結果其音長範圍約落在 100 ms 到 500 ms 之間，因音框的長度爲 20 ms ，所以我們選取音框數爲 50 ，使得對不同長度的測試音（在此處爲連續單音），只要調整音框重疊的長度，就可使得不同長度的單音具有相同的音框數。而固定的音框數是爲了使我們在做 DTW 比對時，不同長度的單音，卻仍擁有相同長度的比對路徑，方便觀察比較。

同理，因爲在單音庫中的單音長度大約落在 280 ms 到 800 ms 之間，大約是連續音的 8/5 倍，所以我們取音框數爲 80 。對於不同長度的單音利用重疊的長度來調整使其具有相同的參考音框數。

在 DTW 的演算法部份，全域路徑限制（ Global Path Constrains ）方面，定爲 1:4 大約爲連續音的音長比上單音的音長的最大值。ε 是比對範圍前端和尾端容許的鬆弛值（ Relax ），在本實驗中我們將ε 的值設成 6 ，使得前端和尾端的對應彈性較大。在音框的距離量測方面，對每個音框我們使用了 16 階的倒頻譜係數。

2.2.1 DTW 應用於母音部份之合成

首先我們從連續音庫中，切取出各類母音（韻母、複韻母、聲隨韻母）做爲測試樣本，從單音庫中切取相對之母音爲參考樣本，然後求取其 DTW 對應路徑。整個結果如圖 2 所示。

在圖 2 中，我們從所有母音類的 DTW 路徑對應圖中，發現絕大部份的路徑非常接近斜率 8/5 的直線。斜率 8/5 的直線表示在從單音中抽取我們所須要的連續音部份時，只要根據其音長的比例，等比例的從單音中抽取相對應的部份即可藉此合成連續音中之母音部份。

我們在真正利用單音來合成連續音時，是以基週（ Pitch ）爲合成的單

位，根據上面所發現的結果，我們以音長爲比例，等比例的從單音中抓取基週以合成連續音之母音部份。

2.2.2 DTW 應用於子音部份

在上節中我們提到利用音長的比率抓取對應的基週（ Pitch)合成連續音，但是子音部份並不存在穩定性的基週，所以我們在合成連續音中的子音時必須再特別處理。我們從連續音庫中切取各種子音，各約五十個，在切取這些子音時，必須在尾端保留開始和母音銜接約 2 至 3 個基週。取音框長度爲 7 ms ，測試樣本數爲 50 音框，參考樣本數爲 80 音框，進行 DTW 路徑對應。結果發現無氣塞音（ㄅ、ㄉ、ㄍ），送氣塞音（ㄆ、ㄊ、ㄎ）及無氣塞擦音（ㄗ、ㄓ、ㄐ）的路徑對應都落在同一類，非常合乎各類子音的特性，而且這些子音都屬於音長較短的子音，在連續音中的長度和在單音中的長度相差不大，所以我們決定在使用單音合成連續音時，這些子音並不特別處理，直接擷取單音中的子音作爲連續音中的子音部份。至於有聲子音（ㄇ、ㄋ、ㄌ、ㄖ），因其具有像母音週期性的特徵，所以將這部份當成是母音處理。

最後剩下的爲送氣塞擦音（ㄘ、ㄔ、ㄑ），及清音（ㄈ、ㄙ、ㄕ、ㄒ、ㄏ），因爲這些子音的音長較長，會影響連續音合成的結果。我們亦從連續音庫中，對每個子音利用中研院所提供的字轉音介面，各找到約五十個屬於詞內（含詞尾）和詞外（含詞首）的子音，求其平均長度。整個結果如圖 4 所示，其中單音（前）和單音（後）分別代表詞首的單音和詞尾的單音。我們從圖中可看出：

1.除了 ㄈ 和 ㄏ 之外，其餘的六個子音，在連續音中詞首和詞內的長度相差不大，所以在合成子音時，並不考慮是否在詞內的因素。

2.單音中子音的長度大約比連續音中的子音多出一倍。

針對這兩個現象，我們分別再對這些子音從連續音庫多取一些測試樣本，進行 DTW 路徑比對，結果如圖 3 所示。從圖 3 中我們發現ㄘ、ㄔ、ㄑ、ㄙ、ㄕ、ㄒ、ㄏ的 DTW 路徑非常靠近斜率 8/5 的直線，所以我們以單音合成連續音時，子音部份假設存有週期性，然後依其音長的比例抓取對應的基週合成連續音的子音部份。

至於 ㄈ 和 ㄏ 這兩個較特殊的子音，分成詞首和詞內處理。當ㄏ在詞首時，從圖 3 中可看出其路徑靠近斜率 8/5 的直線，所以使用等比例的方法抽取所須要的基週部份。至於在詞中時，基於ㄏ這個音會被省略的現象，例如『南韓、聯合』（圖 5），所以我們在合成時會省略子音ㄏ的部份。在ㄈ的部份，只考慮在詞首的部份，因爲ㄈ在詞中時並不產生重疊連接。從圖 3 中我們可看出ㄈ的 DTW 路徑的走勢偏重於後半部，所以我們採用直接取最靠近母音部份做爲合成連續音ㄈ的子音部份。

最後將子音的處理，歸納成表二：

	子　　音
直接取單音部份	ㄅ、ㄉ、ㄍ、ㄆ、ㄊ、ㄎ、ㄐ、ㄓ、ㄗ、ㄈ
依子音長度比	ㄘ、ㄔ、ㄑ、ㄙ、ㄕ、ㄒ、ㄏ（詞首）
忽　　略	ㄏ（詞中）

表二　子音合成時之分類總表

2.3 連音二字詞之合成架構

在本篇論文中，最主要的目的是利用單音模擬合成連續語音中的連音

二字詞，所以這個合成架構的必要條件是：『用來做爲模擬的連續語音二字詞必須存在』。一旦從連續音庫取得連音二字詞後，整個利用單音來模擬連音二字詞的步驟如下：

步驟一：找出整個連音二字詞的連音中點

步驟二：利用連音中點切出連音段

步驟三：利用單音合成前段音節

步驟四：合成連音段

步驟五：利用單音合成後段音節

整個流程如圖 6 所示。

3 連音二字詞之合成

3.1 連音中點

在連音中點的求取方面，我們利用兩相連單音和連音二字詞以倒頻譜係數差（ Delta － Cepstrum ）[3][5]爲參數，然後利用 DTW 路徑比對找出連音中點。在此，我們發現使用倒頻譜係數差所找到之連音中點準確性非常高。於是我們使用此方法協助我們決定連音中點。

3.2 決定連音段

連音段（ Coarticulation Segment ）是指重疊連接時，在兩音節中間所存在的一段轉換過程週期波，基本上它可能是一個存在兩音節中的音素。首先我們必須決定出連音段長度。因爲我們在 3.1 節中可以準確的決定連音中點，所以我們取連音中點附近 20ms 爲一音框，而且整個連音二字詞

亦以 20ms 爲一音框，每次重疊爲 10ms ，然後以連音中點的音框逐次和整個連音二字詞的音框進行比對，經三點平均平滑後，產生如圖 7 的音框距離對應曲線，從圖中可看出連音中點會是一個波谷產生點，於是我們對此曲線微分求連音中點兩端斜率最大的音框，則此兩端所切即爲連音段。

3.3 利用單音合成前後段音節

連音段求出之後，則整個二字連音詞被切成三部份，第一部份即前音節部份，第二部份即連音段，第三部份即後音節部份。因爲我們是利用單音來模擬二字詞連音，單音和前音節的部份音長一定不相同，我們分別使用兩種不同的方法來進行前音節部份的合成

我們將子音和母音分開處理，子音依表二所述，只有那些在單音中較長的子音必須特別處理。而母音部份，在連音二字詞部份，我們取到第二部份，即第一部份加上連音段作爲要利用單音來合成的目標樣本。我們用了兩種不同的方法，都是用來調整單音的長度使其和目標樣本的長度一致。

第一種是我們在 2.2.1 節所提，在母音方面以週期爲單位，以單音基週數和目標樣本的基週數爲比例，取所對應的單音基週來合成目標樣本。而那些子音太長的單音，在子音部份我們亦使用其長度的比例進行長度的調整。

而第二種方法則較爲複雜，將目標樣本和單音以基週爲音框的單位進行 DTW 路徑對應，然後依路徑取出所對應的單音基週合成目標樣本，見圖 8 。而子音太長的亦用 DTW 來進行長度上的調整。

3.4 合成連音段

我們在 3.3 節中，做單音節的合成時，前音節的部分有連音段所對應之單音部份，而後音節的部份亦有連音段所對應之單音，照合成的結構看來，多出了一段連音段。而所多出的連音段是爲了要利用圖 9 的方法合成連音段時用來重疊用。也就是將前單音節的連音段和後單音節的連音段重疊合成出連音段。

在合成連音段時，必須使得合成中的週期長度一樣，於是我們使用基週重建的方法來使得週期的長度一致。

3.5 音量及基週走勢

在連音二字詞的合成方面，我們的重點是放在整個合成二字詞是否聽起來順暢自然，而不會有第一個音節尚未結束，第二音節音就已搶先發出之感。

所以連音段即扮演從前一音節過渡到下一音節的角色。雖然音量及基週走勢亦非常重要，但這應該是跟整句話的韻律較相關，我們會在將合成的二字詞放回整句話時，調整其基週走勢及音量。圖 10 是一個連續音和單音的基週走勢比較圖，由圖中可看出，基週的走勢是非常重要的。

4 評估與度量

4.1 實驗評估

最後我們進行整個結果的評估，測試分成兩部份。第一部份是對二字詞的評估。第二部份則是對平衡句的評估。

在二字詞的評估方面，我們利用單音合成從連續音中取得的二字詞，並且分別使用本論文所提及的兩種方法：一種是以音長爲比例擷取相對應的基週，另一種則以 DTW 路徑來擷取相對應的基週。在平衡句的評估方面，我們從電信所提之平衡句中，切取一個連音二字詞，然後利用單音合成後，調整其音量及基週走勢，再放回平衡句中。也就是說，我們比較二種句子，一種是其中有某個二字詞是利用單音合成的平衡句，另一種則完全是原音。每一部份皆分別測試其自然度（ Naturalness ）和理解度（ Comprehensibility ）。

在自然度方面，我們從 60 個混著原音和合成音的二字詞中，隨機播放給聽者聽，請他們就所聽到的二字詞評分，分數從 1 到 100 分（ 100 分爲最高分）。理解度方面也是從混合著原音和合成音的二字詞中，任意挑選出一句二字詞，然後請他們將所聽到的二字詞寫出。在平衡句方面的測驗亦同於二字詞。受測對象分別爲 29 位國中、國小的老師，還有 13 位大學生，總共 42 位。

4.2 實驗結果

在二字詞的自然度方面，受測的二字詞中有原音及合成音，合成音部份又有兩種合成方式，第一種是根據音長的比例，均勻（ Uniform ）的取單音基週部份合成連音二字詞，而另一種則是在取基週時根據 DTW 的對應路徑來合成連音二字詞。在此測試部份，我們並不根據連音二字詞來調整基週走勢和音量大小。

結果顯示出，聽者對原音和合成音的喜好程度差異很小，表三是各種測試音的結果。絕對分數是聽者評分的平均分數，喜好度則是在某位聽者聽起來是三種中平均分數最高的，而厭惡度（ Dislike ），則是平均分數最低的。從表中可看出從連續音中抽取出之連音二字詞，單獨播放時可能比

合成音效果差，因爲它本身包含了連續音中的韻律訊息，可能只適合在連續音中。一旦從連續音中獨立出來，就可能產生如表四的結果。

	絕對分數	喜好度	厭惡度
原　　音	72.5	9.5 %	59.5 %
Uniform	74.3	42.8 %	28.6 %
DTW Path	74.9	47.6 %	12.0 %

表三：二字詞的自然度結果

	正確率
原　　音	86.3%
Uniform	79.1%

表四：二字詞的正確率

在二字詞的理解度方面，其結果如表四所示，我們的評量計算方式是先求出原音二字詞的正確率爲 86.3 %，再求出合成二字詞的正確率爲 79.1 %，最後假設原音二字詞的正確率爲 100 %，則相對的合成二字詞的正確率爲 91.7 %。

就表三的結果而言，使用 DTW 路徑對應比 Uniform 對應有較佳的效果，但是 DTW Path 比 Uniform 耗費較大的處理過程，基於語音合成即時的考量，我們在使用於整句平衡句的合成時，只使用 Uniform 的方式。

在平衡句的測試方面，自然度的結果如表五所示，合成音和原音相對的自然度爲 96.1 %，在這個測試中，從受測者的分數中可看出，百分之九十以上的聽者偏好原音。在理解度的測試方面，其結果如表六所示，原音

的理解度爲 94.8 %，而合成音方面其理解度爲 83.3 %，其相對之理解度可達到 87.8 %。

	絕 對 分 數
原　　音	86.7
合 成 音	83.3

表五：平衡句的自然度結果

	正 確 率
原　　音	94.8%
Uniform	83.3%

表六：平衡句的正確率

4.3 實驗結果討論

從上一節的實驗結果中可看出，在二字詞的測試方面，我們得到一個不錯的結果，但是放回平衡句時，效果不是很好，我們觀察的結果，可能有二個問題存在：

1.調整基週走勢的方法

2.所使用單音和連續音的差異

在調整基週走勢的方法方面，我們只有試過一般的線性和基週重建，可以試試其它的調整方法，如 Pitch Synchronous Overlap and Add （ PSOLA ）演算法[4][6]，許多文獻提及有不錯的效果。

在本篇論文中一直嘗試使用單音來模擬合成二字詞，但是單音和連續

音除了音長、音量的差異外，還有許多差異：第一，單音在錄製時期完全不含韻律訊息。第二，有些會產生變音、藕合效應的連續音（例如：『神機妙算』中的『機』受『妙』的影響而變成『ㄐㄧㄠ』），並不存在我們的單音庫中。因此，我們可嘗試從連續音庫來建單音庫，由經由適當的分類來建立單音庫，則這些單音可能已具備一些韻律的訊息，應該比我們現在用來合成的單音，更接近連續音。

5 結論

我們從事語音合成的目標在完成一套、能夠連續發音的中文文句翻語音系統，儘管有許多合成系統被提出，但是很少有文獻提出有關連音的探討。本篇論文從模擬連續語音中的連音二字詞開始，首先利用 DTW 對應找出單音中的子音和母音與連續音中的子音和母音是跟其音長呈等比例關係。在確定連音中點方面，我們使用倒頻譜係數差來改善原 DTW 的缺點，進而確定連音中點。再從已知之連音中點，找出連音段長度，到整個連音二字詞的完成。

整個實驗的結果說明，在二字詞的合成方面，我們從一個較長的句子中所切取之二字詞，反而不會比合成的二字詞好聽，根據此項訊息告訴我們，長度不同的句子應有不同的韻律訊息。可知韻律訊息在語音合成中之重要性。

我們希望能朝向自然的語音合成之路邁進，在做連音段的合成時可以在沒有連續音的資訊下，完成連音段的合成，這其實和我們的發音器官的變化性，還有聲道的頻譜特性有很大的相關性，我們一直很缺乏這方面的知識[1]，所以值得在此方面投入研究，以期獲得更佳的連音方法。

參考文獻

[1] John R. Deller, Jr.John G.Proakis, and John H. L. Hansen ，“Discrete-Time Processing of Speech Signals”, MACMILLAN 1993 。

[2]呂士南，周同春，初敏和陸亞民,“漢語合成系統中音高音長規則”，第三屆全國人機語音通訊學術會議論文集。中國， 1994 。

[3]丁培毅, 張保忠和黃英峰,“以分段量測的動態時間歸正法則改善混淆群集的辨認”，電信研究季刊，第 19 卷第三期, 1989 。

[4]Pei-yih Ting, Chun-Yu Tsai and Chi-Shi Liu, “The Post-Processing Stages of a Mandarin Waveform Synthesizer”, in Proc. of 1994 International Computer Symposium Taiwan , 1994 , p1262-p1266 。

[5]Carl D. Mitchell,Mary P.Harper and Leah H.Jamieson, “Using Explicit Segmentation to Improve HMM Phone Recognition”, in Proc. of ICASSP 1995 。

[6]Eric Moulines and Francis Charpentier, “Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones”, Speech Communication, September, 1990,p454-p467 。

[7]陳志祥，“國語連續語音連音型態之初步研究”，中興大學應數研究所碩士論文， 1995 。

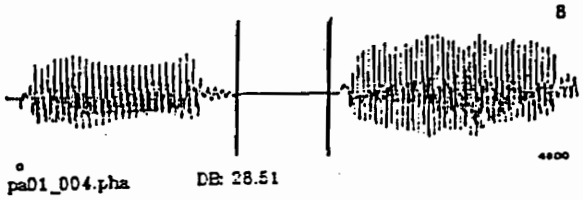


圖 1-a　停頓連接

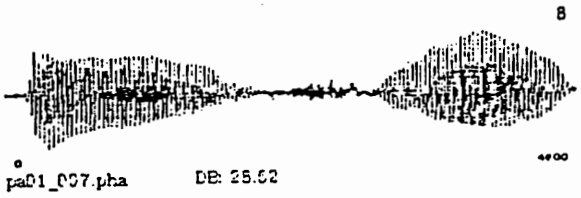


圖 1-b　緊密連接

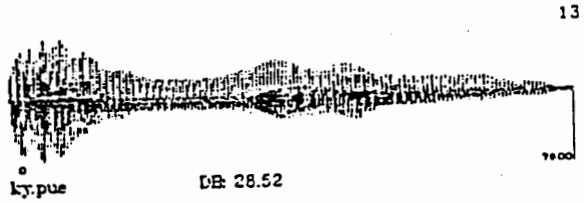


圖 1-c　重疊連接

圖 1　停頓連接、緊密連接和重疊連接之圖例

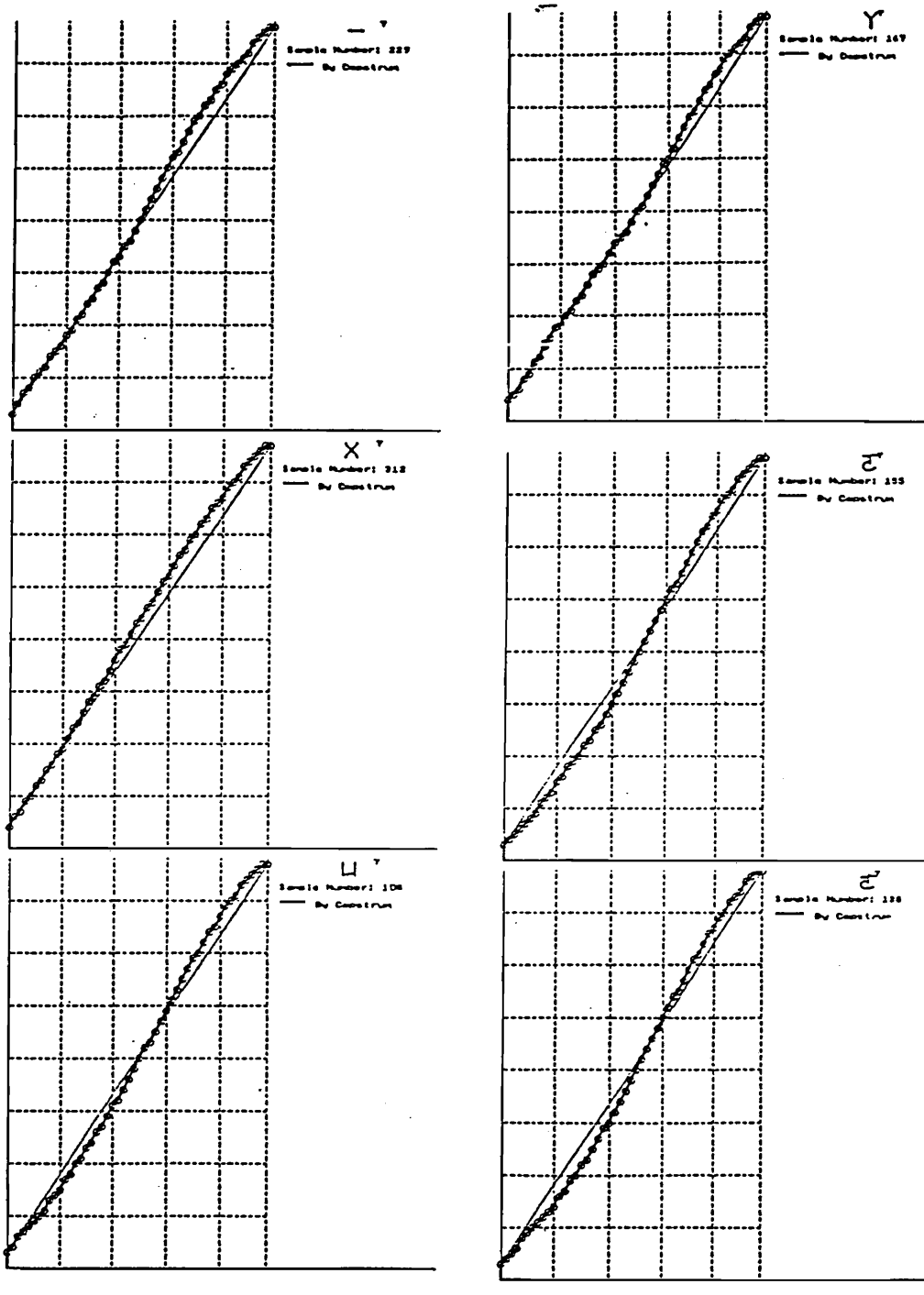
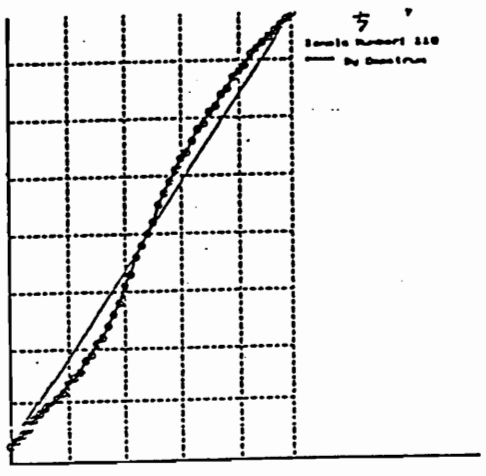
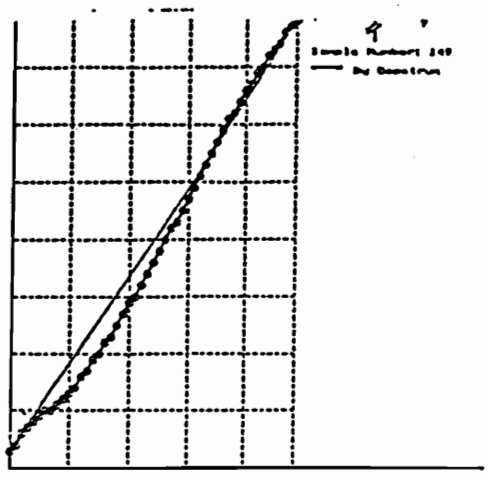
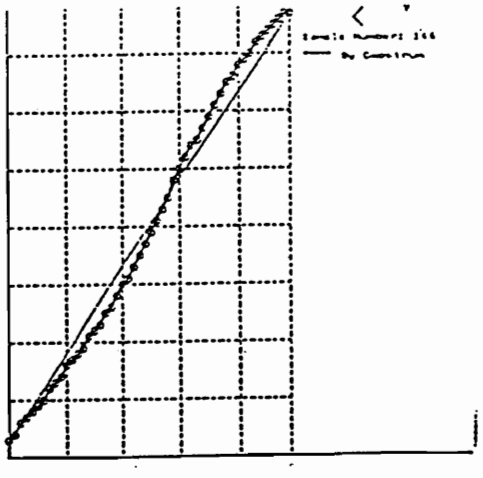
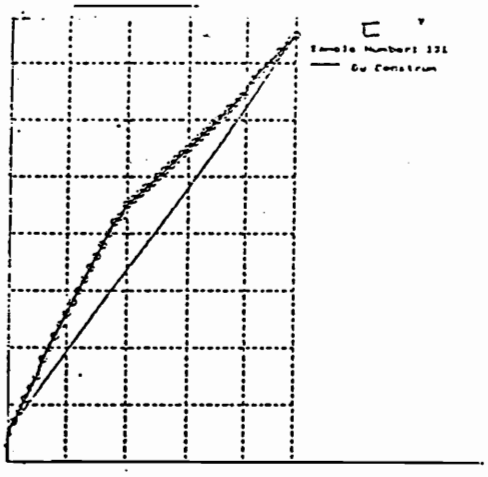


圖2 母音類 DTW 路徑對應圖









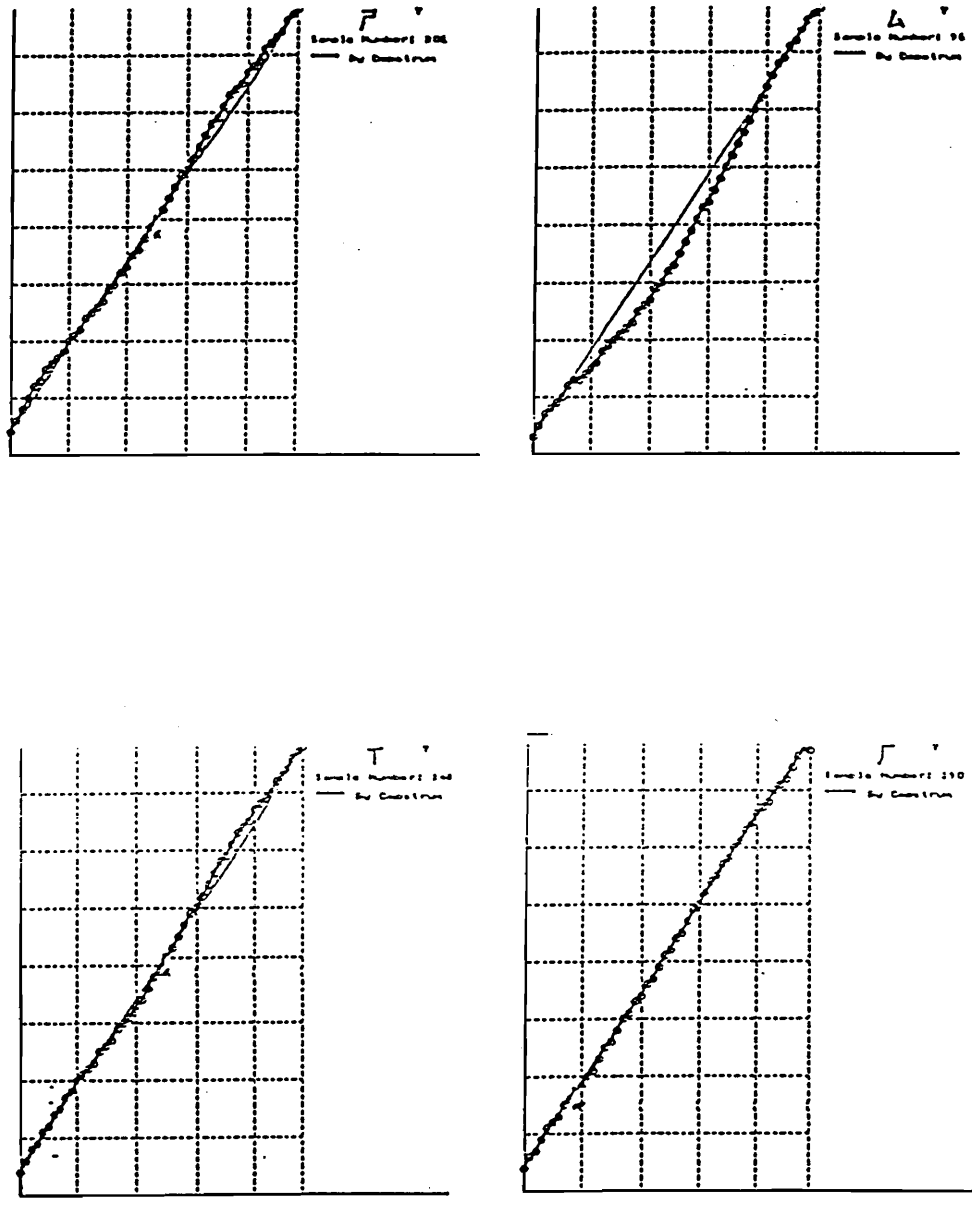


圖3　子音ㄘ、ㄔ、ㄑ、ㄈ、ㄕ、ㄙ、ㄒ、ㄏ 之DTW路徑對應圖

子　音	詞 首 (含落單)	詞 內 (含詞尾)	單　音(前)	單　音(後)
ㄘ	1185.15	1364.19	1991.08	2103.13
ㄔ	1238.82	1250.57	1920.19	1947.56
ㄑ	1327.41	1261.3	2126.43	2114.36
ㄈ	405.28	256.51	781.31	815.52
ㄙ	1451.23	1521.33	2675.54	2680.7
ㄕ	1637.35	1473.1	2778.71	2547.13
ㄒ	1472.75	1288.62	2839	2697
ㄏ	1036.97		1895.96	1700.47

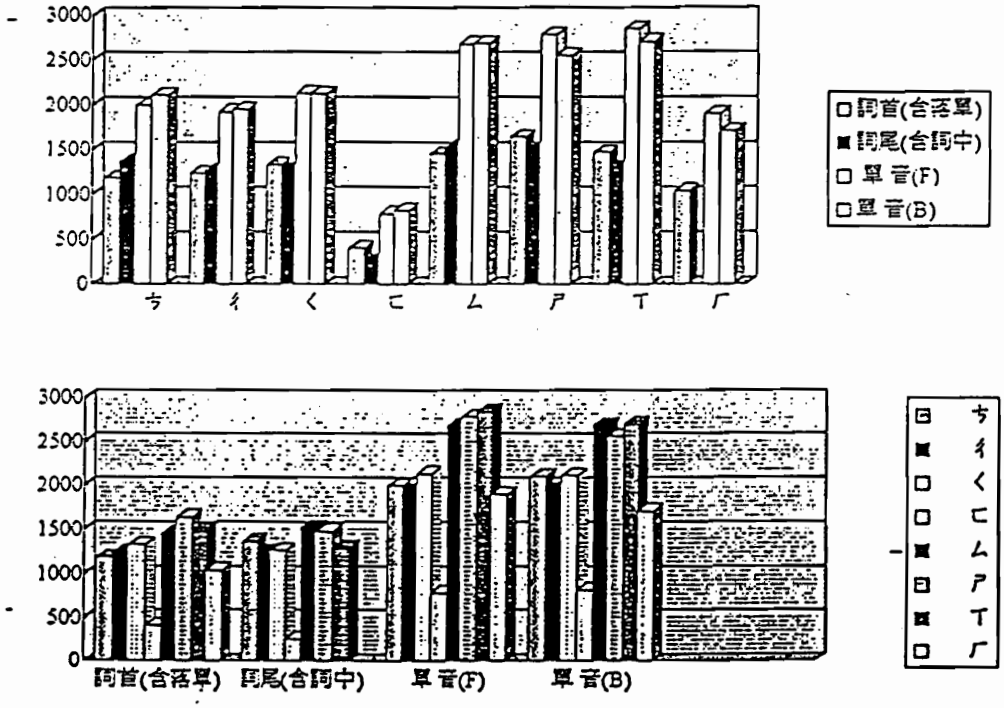


圖4　子音音長統計圖表

『南韓』的連音情形

『聯合』的連音情形

圖5　子音ㄏ在詞內產生連音的情形，"|"表示人工切音時所做的標記。

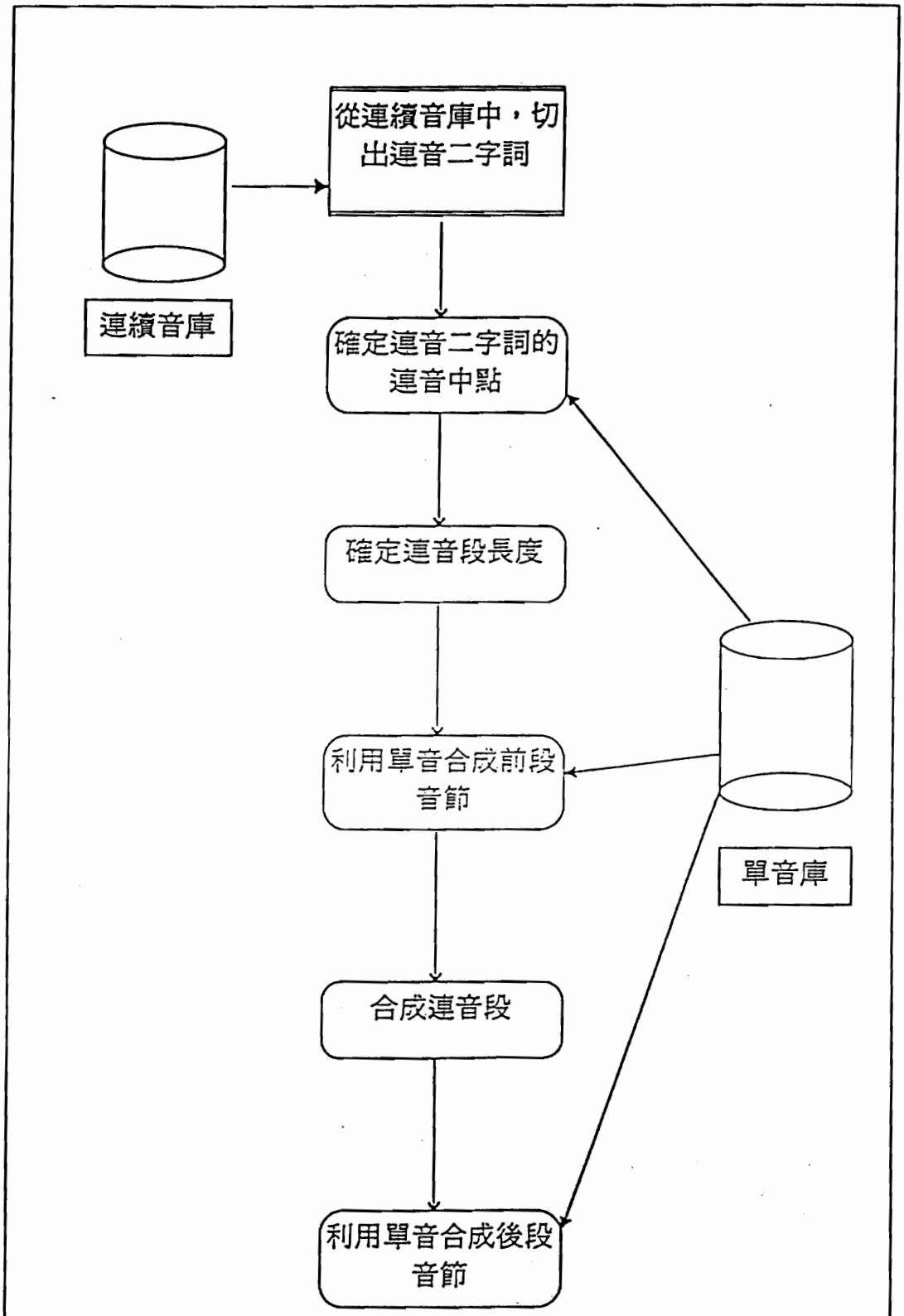


圖 6　連音二字詞合成架構流程圖

圖 7-a　連音二字詞『孤立』

圖 7-b　連音中點音框與整個連音二字詞之距離曲線

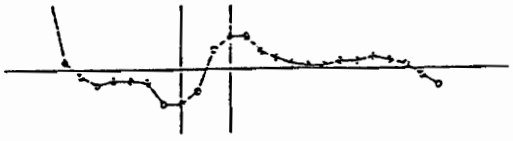
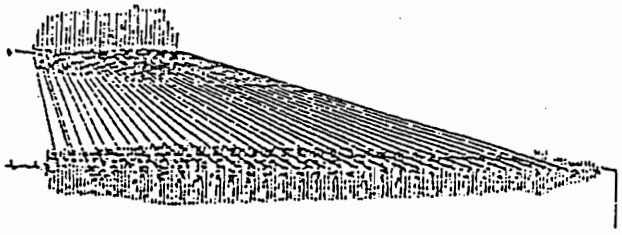
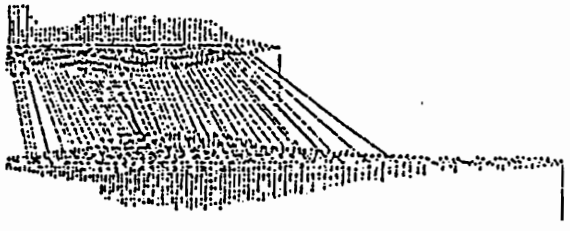


圖 7-c　圖 7-b 曲線之斜率變化走勢

圖 7　連音段之求取過程



前音節之 DTW 路徑對應圖示



後音節之 DTW 路徑對應圖示

圖 8　連音二字詞分別和前音及後音之以基週爲單位之 DTW 對應關係圖

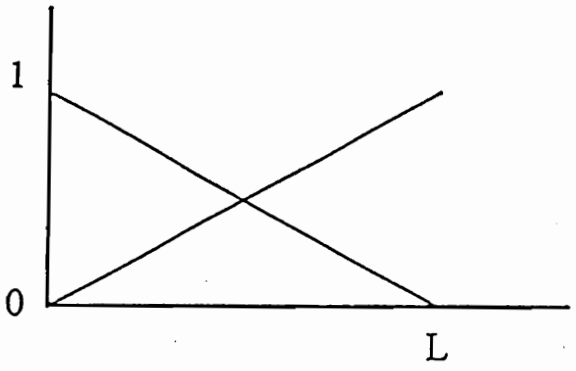


圖9　合成連音段之A、B係數走勢圖

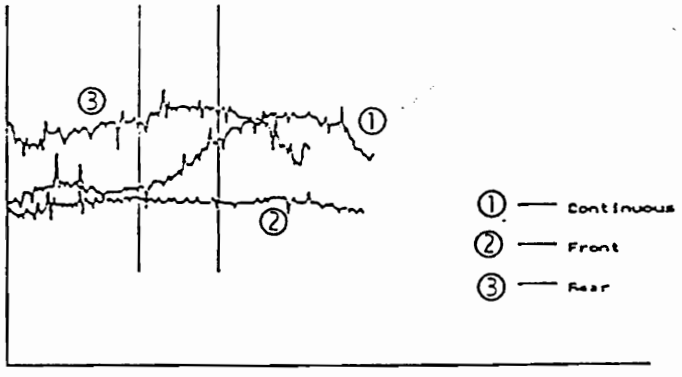


圖10　『官員』基週走勢比較圖，①是連音二字詞『官員』的基週走勢曲線，②是單音『官』的基週走勢曲線，③是單音『員』的基週走勢曲線

時間比例基週波形內差 -- 一個國語音節信號合成之新方法

Time-Proportionated Interpolation of Pitch Waveforms -- A New Method for Mandarin Syllable-Signal Synthesis

古鴻炎　　許文龍
Hung-yan Gu and Wen-lung Shiu

國立臺灣工業技術學院 電機系
台北市基隆路四段43號
Department of Electrical Engineering, National Taiwan Institute of Technology
Taipei, Taiwan, R.O.C
e-mail: root@guhy.ee.ntit.edu.tw

摘要

在許多文句翻國語語音的系統裡，都採用音節爲語音合成之單位，因此本論文針對文句翻國語語音系統研究了一個音節信號合成之新方法，稱爲時間比例基週波形內差法。其特色是，除了具有和其它時域合成方法一樣的清晰音質之外，還提供了較多的信號控制之自由度，包括音調(或基週軌跡)之控制，以便由第一聲音節去合成其它聲調的音節；音長之控制，以調整說話速度及反映其它因素對音節長度之影響；以及聲道(vocal track)長度之控制，以便使男生原音合成之女生聲音較爲自然，並可用以合成卡通人物的聲音，這是一個新的嚐試。雖然其它時域合成方法也有提供音調、音長之控制，但是我們的合成方法提供的自由度較高，且已讓這兩控制因素間的相互干擾降低很多。

國科會補助專題研究計畫，編號：NSC 85-2213-E-011-046

1、導言

一個中文文句翻國語語音的系統，可看成是由兩個主要的組件串接而成，其中一個我們稱爲注音與韻律(prosodic)處理單元，它負責將輸入文句所對應的注音查出，然後設定各個語音合成單位的韻律控制參數，包括音調(基頻軌跡)、音長(duration)、音強(intensity)、與音前停頓(pause)，對國語語音合成來說，音節是最常被採用的語音單位。另外一個組件我們稱爲語音信號(或波形)合成單元，它必須能夠依據前一個單元輸出的語音單位編號及相隨的韻律控制參數去合成出語音信號，除了要忠實地接受韻律參數的控制外，也要合成出具有清晰音質的信號，本論文所研究的主要就是在語音信號合成這個單元上，實際上則提出了一個國語音節合成的新方法，它可在具有一定清晰度的條件下，提供更多控制上的自由度。

過去，雖然已有一些中文文句翻語音的系統被提出〔1,2,3,4〕，但是許多系統的研究重點是在韻律處理單元，關於語音信號合成單元則採用既有的技術、或作部份的修改，因此，合成出來的語音信號常受所用技術的限制而有一些缺點，如 LPC 技術〔5,6〕合成的語音信號不清晰、音質不好，共振峰合成技術〔7,8〕雖可獲得較佳的音質，但是相鄰音素(phoneme)間信號轉移(transition)之模擬仍不理想，並且需以人工來調整、設定大量的音響(acoustic)控制參數(即未有一套理想的自動化的參數值估計方法)。最近，一種直接在時域波形上操作的技術(稱爲 PSOLA, pitch-synchronous overlap and add)被提出〔9,10,11〕，它不但可合成出清晰音質的語音信號，並且控制參數數值(即基週之時間位置)較容易決定，不過它只提供侷限的音長控制之自由度，即不接受音長比值(合成音長除以原始音長)被任意設定成合理範圍內(如 0.5~2.0)的一個數值，因此，後來一種變形的技術稱爲 LP-PSOLA 被提出〔11,12〕，以提高音長控制之自由度，可是它顧此失彼，反而導入雜訊使合成的語音信號變得不清晰。再者，PSOLA 技術對國語音節

之合成還存在有缺陷，一個例子如以第一聲/ai/音節去合成第四聲之/ai/時，原本第一聲/ai/之/a/與/i/若各佔一半的時間(這裡爲了說明才用二分法，實際上是逐漸轉移的)，則合成之第四聲/ai/裡，/i/部份會比/a/部份長，因爲第四聲在信號波形上的特徵是週期由小變大，而 PSOLA 技術爲了改變音調的一種作法是單純地把週期逐漸拉大，這樣就造成了/i/部份會比/a/部份長之副作用，實際上就是在時間軸上扭曲(time warping)頻譜變換之速度，另外一種作法是捨棄幾個信號週期後再把剩下的週期逐漸拉大，以使音長維持固定，但是直接丟棄信號週期必然會使頻譜走勢變成不連續(特別是在一個雙母音音節中)，這樣爲了控制一個因素而導致另一個因素失控(反觀人本身並無此種失控情形)，並不能算是一個好的解決方法。

爲了改進前述語音合成技術的缺點，本論文提出了一個國語音節信號合成的方法，稱爲時間比例基週波形內差法(Time-Proportionated Interpolation of Pitch Waveform, TPIP)，它也是在時域上直接對信號波形作處理，因此合成的語音信號具有一定的清晰度，不過它提供的音長控制之自由度卻是比 PSOLA 技術高很多，這也是當初研究此技術的一個重要動機。至於音調控制之自由度，對國語語音合成來說則是一項基本的、不可缺少的的控制因素，這樣才能接受韻律處理單元的控制去合成出國語裡的各個聲調或整句的句調，而所提出之方法不僅能配合各種音調去合成出語音信號，並且沒有 PSOLA 技術裡的因爲改變音調而破壞頻譜變換速度的副作用。除了音長、音調控制之自由度外，我們也增加了另外一個自由度，即可讓合成語音的共振峰頻率全體(F1,F2,F3,...)作相同倍率之升高或降低的控制，提供這種控制的動機是，當把音調(F0軌跡)提高以便由男生發的原始音節波形去合成女生聲音時，合成的語音信號聽起來總是有男生在假裝女生聲音的感覺，這說明男女生除了基頻的高低差異之外，還有其它先天上的差異(發音習慣可說是後天上學來的)，其中重要的一項是聲道長短的差異，一般說來男生聲道較女生的爲長，並且由聲道的音響模型可知

〔13,14〕，聲道長度和共振峰頻率值之間存在者反比的關係(聲道短則共振證頻率高)，因此我們相當於提供了聲道變長變短之控制。關於音長、音調、聲道長等三項因素之控制，我們提出的音節信號合成方法，在合理的參數值範圍內各個因素可說是獨立的、不相互干擾的，這可由頻譜分析圖及所建造的原型文句翻國語語音系統來得到驗證。

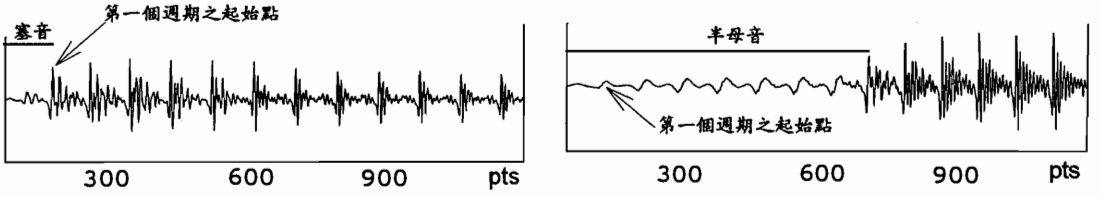
2、國語音節結構 與 無聲部份信號合成

一個國語音節的信號可看成是由無聲(voiceless)部份與有聲(voiced)部分兩部份串接組成，無聲部份的信號對應於波形無週期性之塞音(stop)、擦音(fricative)、或塞擦音(affricate)，而有聲部份的信號則對應於波形有週期性之鼻音(nasal)、滑音(glide)、流音(liquid)、或母音(vowel)，如果一個音節全由有聲之音素構成，則我們可把此音節當作無聲部分極爲短暫的音節，即把第一個週期前的信號看作是無聲的部分，如此，每一個國語音節都可說是具有 UV 之結構，U 與 V 分別表示無聲與有聲部份。

由於一個合成音節的時間長度是由韻律處理單元輸出的音長參數來決定的，因此，當要合成一個音節的信號時，需先依據音長參數去決定無聲、有聲兩部份各要佔據多少時間，然後才分別去合成無聲與有聲兩部份的信號。這裡將先介紹無聲部份在信號上的分類，然後對不同類的無聲信號去說明無、有聲兩部份的時間分配作法及無聲信號的合成方法，其重點在於如何配合韻律處理單元的音長要求但不破壞無聲音素的特性(即要清晰可辨)，至於有聲部份的合成方法則在下一節介紹。

對於音節之無聲部份的信號，我們歸類成兩種信號類別，以方便作時間分配與信號合成之處理，第一類稱爲短暫無聲，表示其無聲部份的時間

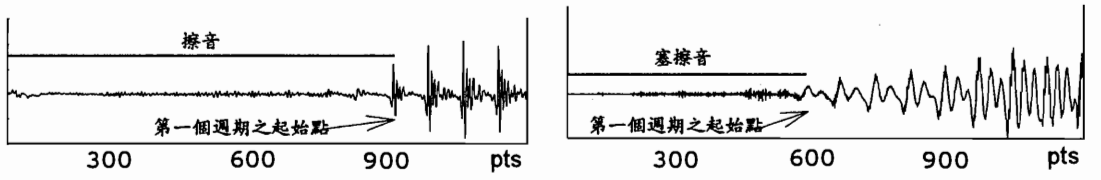
很短，第二類稱爲長帶無聲，表示其無聲部份的時間相對地長很多。我們希望能把非送氣塞音(如/ㄅ/)開頭之音節(波形例子如圖1(a)所示)、半母音(含鼻音、滑音、與流音)開頭之音節(波形例子如圖1(b)所示)、及母音



(a) 塞音開頭音節/ㄅㄚ/之波形　　(b) 半母音開頭音節/ㄇㄚ/之波形

圖1　具短暫無聲部份之音節波形

開頭之音節都歸屬爲具有短暫無聲部份之音節，而把擦音開頭(如/ㄏ/)之音節(波形例子如圖2(a)所示)、非送氣與送氣塞擦音(如/ㄐ,ㄑ/)開頭之音節(波形例子如圖2(b)所示)、及送氣塞音(如/ㄆ/)開頭之音節都歸屬爲具



(a) 擦音開頭音節/ㄏㄚ/之波形　　(b) 非送氣塞擦音音節/ㄐㄧㄚ/之波形

圖2　具長帶無聲部份之音節波形

有長帶無聲部份之音節，如此在實作上，就可以依據一個音節的第一個有聲信號週期之起始點的時間位置來對無聲部份作分類，我們使用的門檻是300個樣本點(取樣頻率爲11,025Hz)，即音節裡的第一個週期起始點在300點以內者就歸屬爲具有短暫無聲，否則就歸屬爲具有長帶無聲。

依據原始音節波形裡的第一個週期起始點的位置，對它的無聲部份作分類後，接著就要作無、有聲兩部份的時間分配及無聲部份之信號合成。如果原始音節具有短暫無聲部份，則將原始音節裡第一個有聲週期起始前

的樣本點直接拷貝到合成音節的起始部份，以作爲合成音節的無聲部份，然後將剩餘時間(從音長參數扣除無聲部份之時間)分配給有聲部份；如果原始音節具有長帶無聲部份，則要依據原始音節裡無、有聲兩部份的時間比例去分配音長參數所給定的時間，假設原始音節的長度爲300ms,且無、有聲兩部份的時間比例是4:6，再令音長參數給定的時間是 R 毫秒，此時，若 R*(4/10) > 300*(4/10)*1.5，則只分配 300*(4/10)*1.5 毫秒給合成音節之無聲部份，否則分配 R*(4/10) 毫秒給無聲部份，剩餘時間自然就分給有聲部份，如此之時間分配，相當於限定無聲部份最多只能佔據原始音節無聲部份的時間長度的1.5倍，這是考慮人自己將一個音節唸得很長時，主要是將有聲部份延長而無聲部份並非等比例延長，得知無聲部份的時間長度後，接著將原始音節裡開頭的300個樣本直接拷貝到合成音節的起始部份，以保存塞擦音的起始塞音特性，接著按照原始音節與合成音節之無聲部份的時間比例去內差出其它的無聲部份之樣本值，內差的方法以圖3來說明，

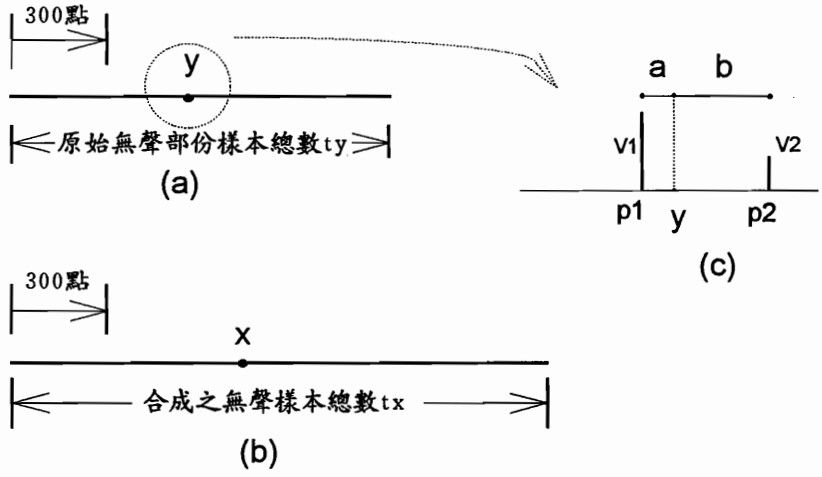


圖3　合成無聲樣本值之內差法示意圖

圖3(b)裡的 x 表示欲合成之無聲部份的一個樣本點，tx表示總共要合成的點數，圖3(a)裡的ty表示原始音節無聲部份的總點數，y是x依時間比例所對應之點，則

$$y = \frac{x-300}{tx-300} * (ty-300) + 300 \qquad (1)$$

很明顯的，由(1)式算出的 y 值並不會剛好爲整數，設介於圖3(c)中的p1和p2兩整數點之間，a爲y點到p1的距離，b爲y點到p2間的距離，v1、v2爲樣本值，我們就簡單地以線性內差來計算x點上的振幅值，如下式所示

$$x\_SampleValue = v1 * \frac{b}{a+b} + v2 * \frac{a}{a+b} \qquad (2)$$

雖然只應用簡單的線性內差來算時間位置及樣本值，但實際聽測時並未發現無聲部分有誤辨之情形。

3、有聲部份信號合成

前一節裡已說明如何將音長參數給定的時間分配給無、有聲兩部份，當知道有聲部份的時間長度後，接著就要依據韻律處理單元輸出的音調參數去計算出一序列的週期長度值 $L_1, L_2, L_3, \ldots$，即基週軌跡之計算，然後才去計算各個週期應具有的波形(或樣本值)。在觀念上，週期長度序列 $L_1, L_2, L_3, \ldots$ 的求取，和各個週期內的樣本值的求取是兩件獨立的工作，因此，在3.1節提出的週期長度值的計算方法，只代表我們的原型文句翻國語語音系統所使用的一種簡單、可行的作法，並不表示我們非常推薦它，至於3.2節提出的一個週期內各樣本值的求取方法，即本文題目所稱的時間比例基週波形內差法，則是我們非常推薦採用的。

3.1 基週軌跡計算

韻律處理單元爲了讓一個音節擁有特定的聲調(當然也可把句調考慮進來)，那麼它可透過音調參數之輸出去告知音節信號合成單元，在我們的原

型系統裡，音調參數是指用以逼近基週軌跡之6個線段的7個端點頻率值，這6個線段各分配音節有聲部份的 1/6 時間，當使用更多的線段時，基週軌跡自然會更平滑。

關於週期長度之計算，我們以圖4裡的一個線段爲例來說明，圖中Freq1和Frq2表示此線段的兩個端點頻率值，我們的目的是要求取在此線段

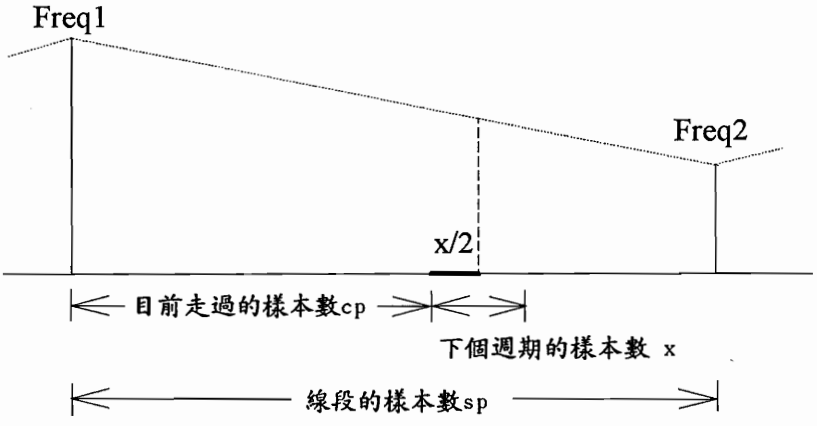


圖4　週期長度求取之示意圖

內各個週期以樣本點數來算的時間長度，設圖中目前要加入的週期，其長度爲 x 個樣本點，則 x 的數值可依據如下的線性時間比例關係來求取：

$$\left(\frac{cp+\frac{x}{2}}{sp}\right)\left(\frac{11025}{Freq2}-\frac{11025}{Freq1}\right)+\frac{11025}{Freq1}=x \qquad (3)$$

式子裡的 11,025 是取樣頻率，等式的觀念是以兩端點頻率對應的週期長度來內差出 x 的樣本點數，並且一個週期是以其中心點爲代表，所以cp要加上x/2。當一個週期跨越到下一個線段時，我們就看週期中心點是否已超出本次的線段，如未超出則保留此週期，否則就將剩餘時間轉給下一個線段。

3.2 週期內樣本值求取

在計算出一序列的週期長度值 $L_1, L_2, L_3, ...$之後，接著就要去合成各個週期的波形，即計算出週期內的各個樣本值，這裡我們提出一種新的求取方法，就是前面提到的時間比例基週波形內差法，它的詳細作法可以如下之四個處理步驟來說明：

(Step 1) 找尋兩個對應之原始信號週期及決定加權:

首先依據欲合成之信號週期的中心點c，如圖5的上方所示，去找出原始第一聲音節波形中兩個對應的信號週期，尋找方法是依據線性時間比例

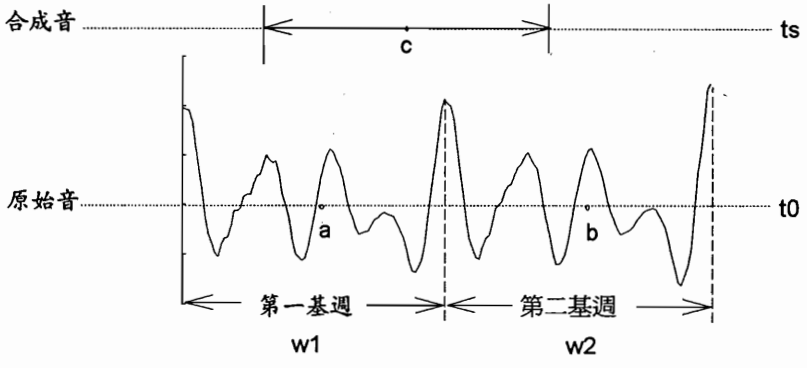


圖5　合成週期與對應之兩原始信號週期

之關係，從原始音節波形中找出兩個相鄰週期的中心點a與b，如圖5的下方所示，使得不等式

$$\frac{a}{to} \le \frac{c}{ts} \prec \frac{b}{to} \qquad (4)$$

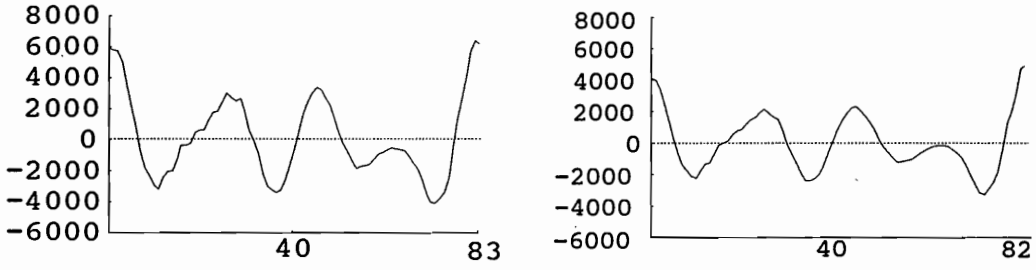
成立，其中ts表示合成音節之有聲部份的總點數，to表示原始音節裡有聲部分之總點數。找到兩個相對應的原始音基週之後，接著要計算二個原始週期波形當被內差組合以合成新的基週波形時，各自的加權值w1與w2要設爲多少，這裡我們仍是以線性時間比例來決定w1、w2的值，如下式:

$$令\ \alpha=\frac{a}{to}\ ,\ \beta=\frac{b}{to}\ ,\ \gamma=\frac{c}{ts}$$

$$設\ w1=\frac{\beta-\gamma}{\beta-\alpha}\ ,\ w2=\frac{\gamma-\alpha}{\beta-\alpha} \qquad (5)$$

(Step 2) 乘上加權值:

此步驟的動作是將找出之第一個原始週期的各個樣本點乘上w1，再將相鄰的第二個原始週期的各個樣本點乘上w2，圖6顯示乘上加權值後的結果，因爲欲合成週期的中心點較靠近第一個原始週期，所以圖6(a)裡的波形的振幅較圖6(a)裡的大。



第一週期乘w1後的波形　　第二週期乘上w2後的波形

圖6 乘上加權值後的兩原始週期波形

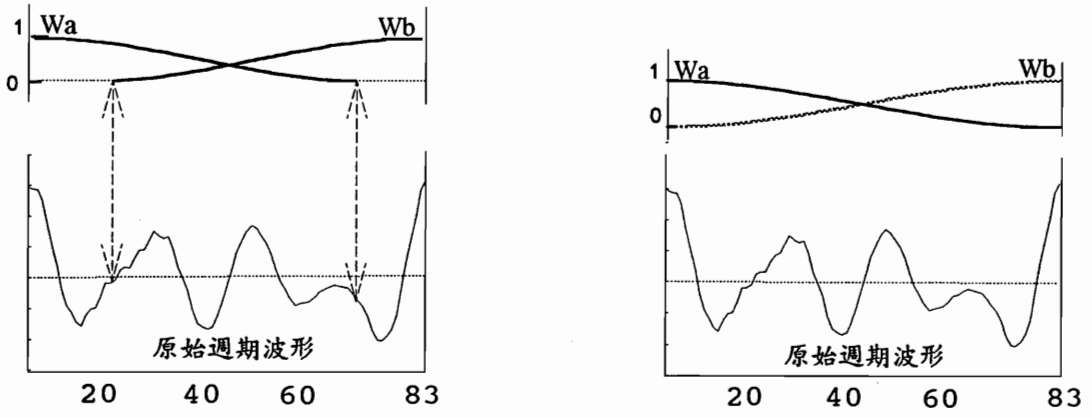
(Step 3) 乘上餘弦窗:

在把兩個原始週期乘上餘弦窗之前需先決定餘弦窗的長度，餘弦窗的長度由原始週期長度和新合成週期長度共同決定，如果原始週期的長度大於新合成週期的長度，則設定餘弦窗長度爲新合成週期長度的兩倍，如圖7(a)所示，圖中橫軸是樣本點數，否則以原始週期長度的兩倍作爲餘弦窗長度，如圖7(b)所示。餘弦窗之函數如(6)式所示：

$$w(n)=0.5+0.5*\cos\left(\frac{2\pi n}{N}\right),\quad 0\le n\le N-1 \qquad (6)$$

決定餘弦窗的長度後，接著就把兩個對應的原始週期個別乘上二個半邊的餘弦窗，如圖7裡的 Wa 和 Wb 表示二個半邊的餘弦窗，左邊的右半餘弦窗

(Wa所表示者)乘上後得到的信號波形就放在新合成週期的左邊，而右邊的



(a) 合成週期長度小於原始週期長時之餘弦窗　(b) 合成週期長度大於原始週期長時之餘弦窗

圖7　餘弦窗長度設定

左半餘弦窗(Wb所表示者)乘上後得到的信號波形就在放在右邊，此時就會得到如圖8(a)與8(c)所示的波形，圖8(a)表示對第一個原始週期之處理，而圖8(c)是第二個原始週期的，然後作疊加的動作，就會得到如圖8(b)與

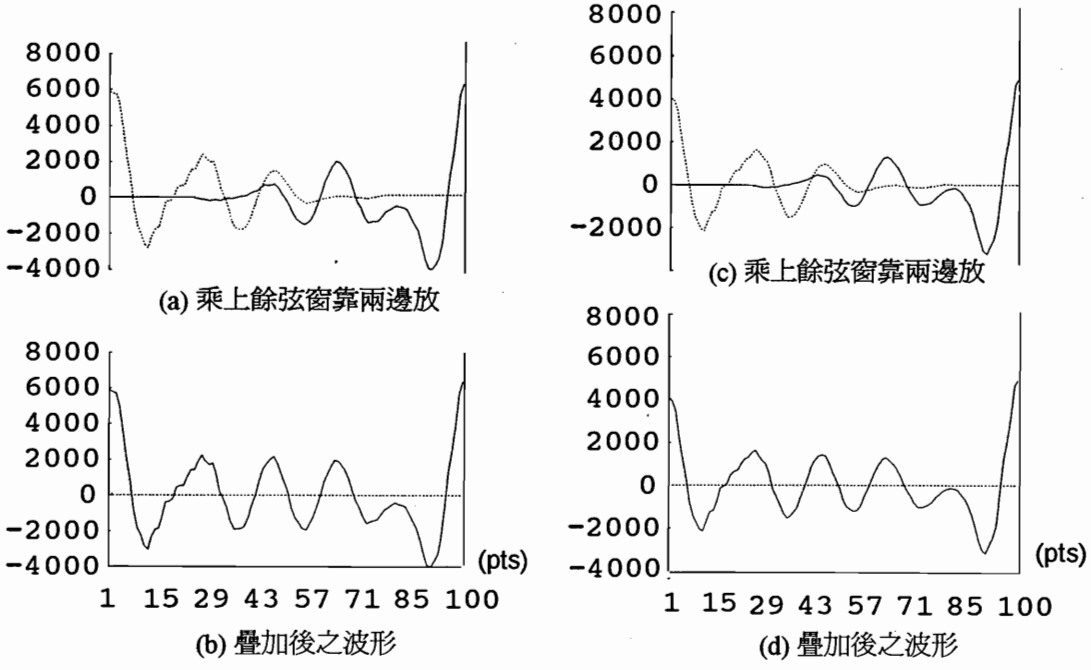


圖8 乘上餘弦窗後靠兩邊放、疊加之兩原始週期波形

8(d)所示的經過處理的原始週期波形(這裡假設新合成週期長度大於原始週期長)。

(Step 4) 相加兩處理過的原始週期波形:

把圖8(b)與圖8(d)所示的兩個經過處理的原始週期波形相加，最後就會得到如圖9所示的新合成之波形，比較圖9和圖5可發現，新合成之波形裡有5個波峰，而兩個原始週期之波形各只有4個波峰，這是因爲這裡的新合成週期長度被設爲100點，比兩個原始週期的長度都長，當週期長度變長，波峰數也要變多，這樣才能維持相同的共振頻率值。

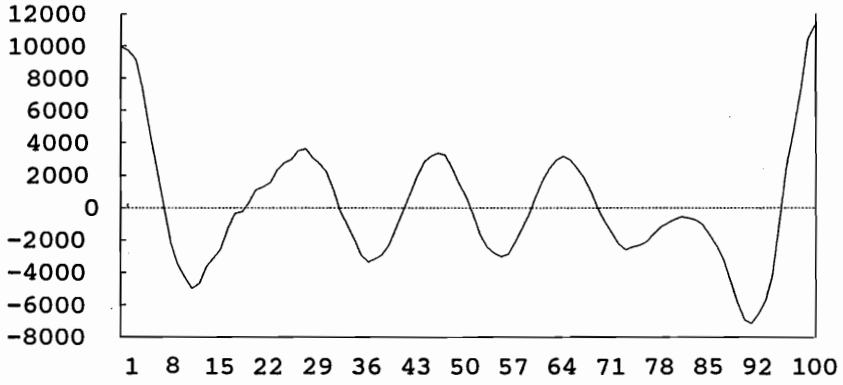


圖9 新合成的信號週期

如此，重覆(Step 1)至(Step 4)的步驟去處理基週軌跡裡的各個週期，就可將音節之有聲部份的信號合成出來，再加上無聲部分的合成處理，則整個音節的信號就可被合成出來了。

3.3 聲道長度調整

由於原始音節信號是由男生錄製的，因此當只把基頻軌跡整個提高來模仿女生聲音時，合成的語音聽起來會像一位男生在假裝發女生聲音，顯

得非常不自然。男、女聲除了發音習慣和基頻高度不同外，聲道的長度也不相同一般來說，男生的聲道較長，女生的聲道較短，因此，我們增加了聲道長度這個控制參數，希望在合成女生聲音時，聽起來較自然。

由聲道的音響(acoustic)模型〔13,14〕可知，聲道長度和共振頻率位置有密切關係存在，即聲道變短共振頻率會提高之反比關係，所以，調整聲道長短，就相當於把各個共振頻率按比例提高或降低。過去，在爲卡通人物配音時，常以錄音機快速放音之方式來將成人聲音轉成小孩聲音，錄音機快放，除了會把基頻提高外，也會把各個共振峰頻率按比例提高，不過，聲音的時間長度卻縮短了。我們希望設計一個具有相當彈性的國語音節合成器，能夠對音長、基頻(F0)軌跡、共振峰頻率(F1、F2...)整體等三項因素幾乎獨立地去控制，在3.2節我們已說明一個可獨立去控制音長與基頻軌跡的方法，這裡我們將說明一個可對共振峰頻率全體獨立去調整的作法，它可被簡單地嵌入3.2節的(Step 1)與(Step 2)之間，而不會破壞其它處理步驟，詳細說來就是應用錄音機快放的道理去作 resampling 的處理，不過是對3.2節(Step 1)找出的兩個原始基週波形去作 resampling，例如當要把共振峰頻率全體調高1.3倍時，就對應於在原始基週波形上，每次要走1.3個取樣點(若取樣頻率不變的話)，如此做，新的取樣點可能會落在原始週期裡的兩個樣本之間，此時，我們可以利用數位訊號處理中的取樣理論將其新樣本值求出，但此種做法太過於費時(計算量大)，在實作上是不可行的，所以我們使用一種以二次多項式去內插的作法，就是把新的取樣點 x 的左邊兩個舊取樣點 x0, x1 和右邊一個取樣點 x2 代入一個二次多項式，而建立如下的方程式：

$$\begin{cases} y_0 = f(x_0) = Ax_0^{\,2} + Bx_0^{\,1} + C \\ y_1 = f(x_1) = Ax_1^{\,2} + Bx_1^{\,1} + C \\ y_2 = f(x_2) = Ax_2^{\,2} + Bx_2^{\,1} + C \end{cases} \tag{7}$$

以圖10來說明，我們可以令x0、x1、x2之值爲0、1、2，而讓y0、y1、y2表示三個樣本值，如此式(7)便形成一個三元一次方程組，而其解爲：

$$\begin{cases} A = y_2 - 2y_1 + y_0 / 2 \\ B = -y_2 + 4y_1 - 3y_0 / 2 \\ C = y_0 \end{cases} \tag{8}$$

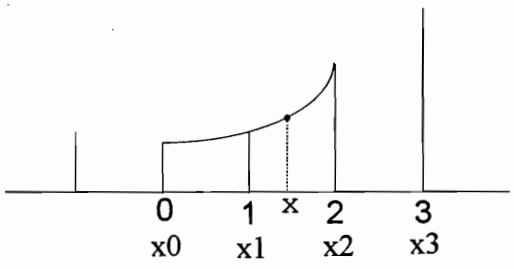


圖10 二次多項式內差之示意圖

求得A、B、C後，再將圖10的 x 座標值換成 x-x1+1 代入，如此繼續，最後便可得到如圖11所示的波形，明顯地週期長度已改變爲原先長度的 1/1.3 倍，接著再依據圖11的兩個經過 resampling 的原始週期去進行3.2節裡的

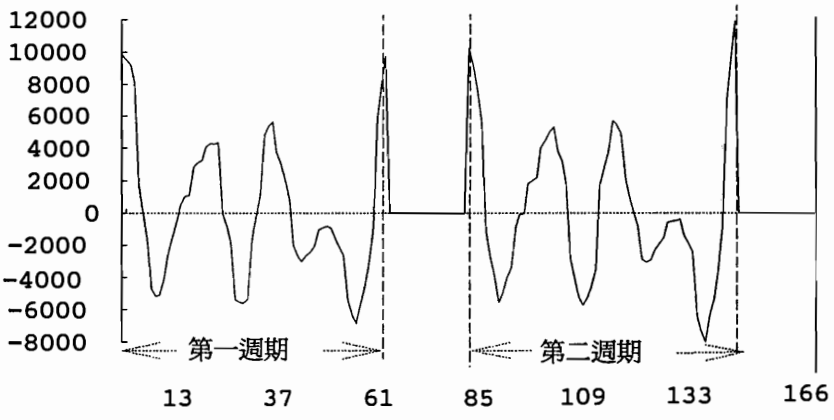


圖11 一次走1.3個取樣點之 resampling 後的兩個原始週期

(Step 2)至(Step 4)之處理，此時，餘弦窗的長度也必需依據 resampling 過的原始週期長度及合成週期長度來決定，如此，處理完後便會得到如圖12所示的波形，將此圖與圖9比較可發現，在相同週期長度的條件下，圖12

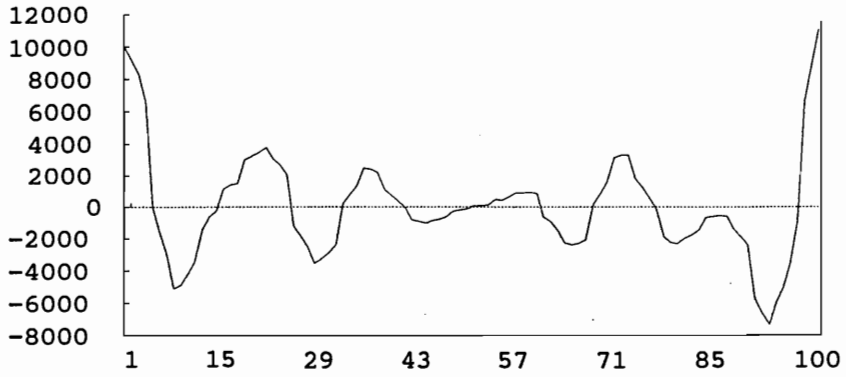


圖12 一次走1.3個取樣點之新合成週期

裡有6個波峰，而圖9只有5個波峰，所以，圖12裡的波形具有較高的共振頻率值，這正是我們我們原先所想要的。

一般來說女生的聲道比男生的短，所以共振頻率會較高，因此在調整基準音高時，我們也希望共振峰頻率(F1,F2...)整體一起被連帶地調整，如此以避免發生男生裝女生聲音的感覺，所以，我們便依據男女生的平均基頻值和平均第一共振峰F1的比值來建立如下的調整公式：

$$WalkPoints = 1 + 0.15 * \left(\frac{NewF0}{120} - \frac{OldF0}{120} \right) \qquad (9)$$

基中，NewF0 表示新設定的基準音高，OldF0 表示原錄製音節信號的基準音高，120是男生的平均基頻值，而1.15是發/ㄜ/音時，男女生平均的F1值之比值，如此，WalkPoint就表示在原始週期波形上每次要走的取樣點數。此外，我們也可設計成讓使用者單獨去設定 WalkPoint 的值，而不改變基準音高值。利用上述作法去調整聲道長度，就可以得到如卡通人物的聲音，或是老沈的聲音。

4、本合成方法之實驗驗證

對於一個音節信號合成單元來說，所合成信號的品質或清晰度(是否混濁、是否有雜訊)是一項重要的評估項目，可是在實際使用上，音節信號合成單元並不是單獨存在的，它通常需和韻律處理單元一起工作(即構成一個文句翻語音系統)，也就是說我們懷疑單獨工作時的表現是否能代表和韻律處理單元一起工作時的表現，因此，我們就實際去建造一個原型的中文文句翻國語語音系統，原始的409個第一聲音節波形是請一位男性播音員來錄製的，使用 11,025Hz 取樣頻率及 16bits/sample 之解析度，在切除 silence 信號後，信號波形共佔 2.24 Mbytes，這顯示時域上的音節信號合成作法，並不如想像中那麼佔記憶體，關於韻律處理單元的製作，我們基本上是參考前人的 rule-based 作法[1]，但也做了一些修改，目前整個原型系統已可即時地唸出國語語音，初步聽測試驗顯示，音質相當清晰，具有和其它時域合成方法一樣的清晰特性，至於可辯度和自然流利度，則和韻律處理單元的好壞有密切關係，因此不宜以此兩項目去評估信號合成單元。

除了清晰度之外，一個好的信號合成單元尚需具備充分的信號控制上之自由度，自由度愈大，則韻律處理單元愈有發揮的空間，例如唸快或唸慢，高昂或低沈，角色扮演(男生、女生或小孩聲音)等。所提出之音節信號合成方法，提供了音調、音長、聲道長之信號控制的自由度，值得注意的是，這三項控制因素在合理的參數值範圍內，都幾乎可獨立去控制(改變數值)，下面就以聲譜(spectrogram)分析來檢查，當一項控制因素被改變數值時，是否會發生副作用。

4.1 音調變換

由於原始音節信號的聲調都是第一聲，因此，當合成出第四聲(或其它

聲調)的音節信號時，共振頻率(F1,F2,...)結構(頻率值及走勢)是否會受到影想，就成爲一個很令人關切的問題。這裡，以音節/ㄚ/爲例，去分析原始之第一聲/ㄚ/音節，其波形如圖13所示，和合成之第四聲/ㄚ/音節，其波形如圖14所示，使用了 Hyperception 公司出品之 Hypersignal 信號分析軟體〔15〕，結果得到如圖15和圖16之聲譜圖，由圖14可看出信號的週期長度一直在增加中，使得對應的圖16之聲譜圖上，基頻及其諧波隨著時間在下降，但是共振頻率則維持水平方向前進，並且具有和圖15之共振頻率一樣的垂直高度。

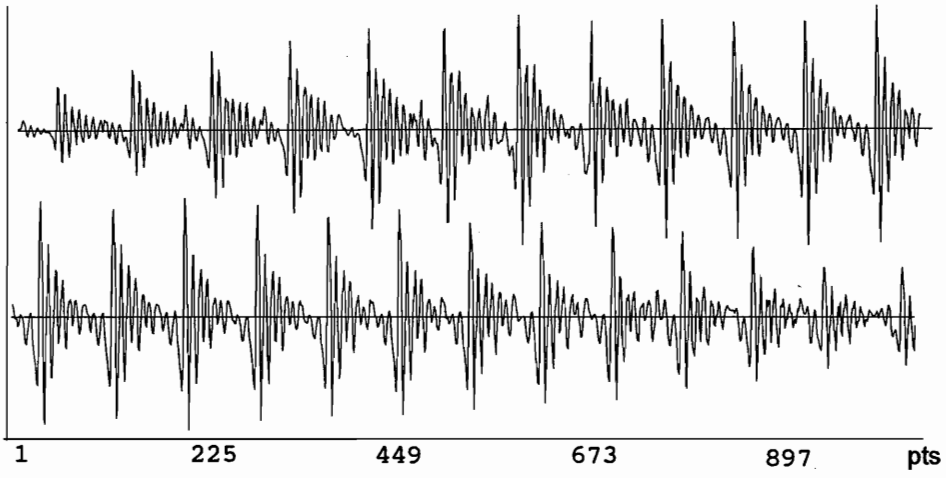


圖13 原始第一聲/ㄚ/之波形

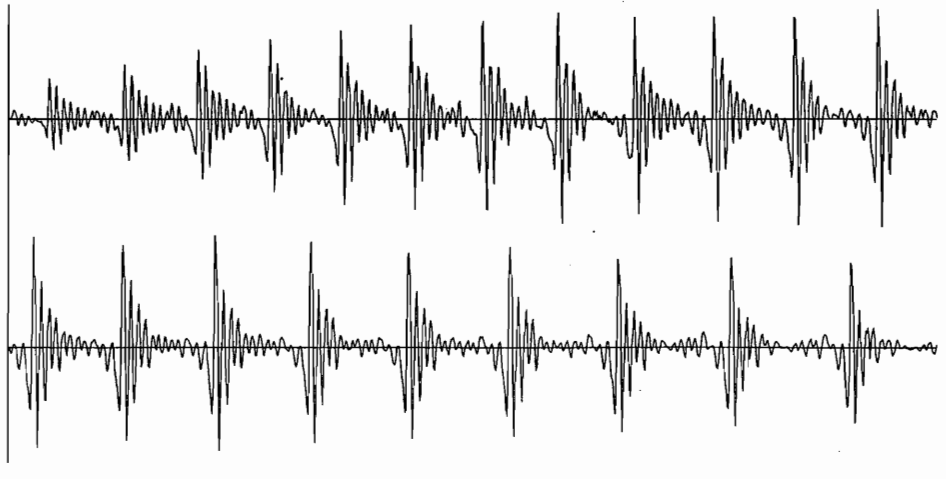


圖14 合成之第四聲/ㄚ/之波形

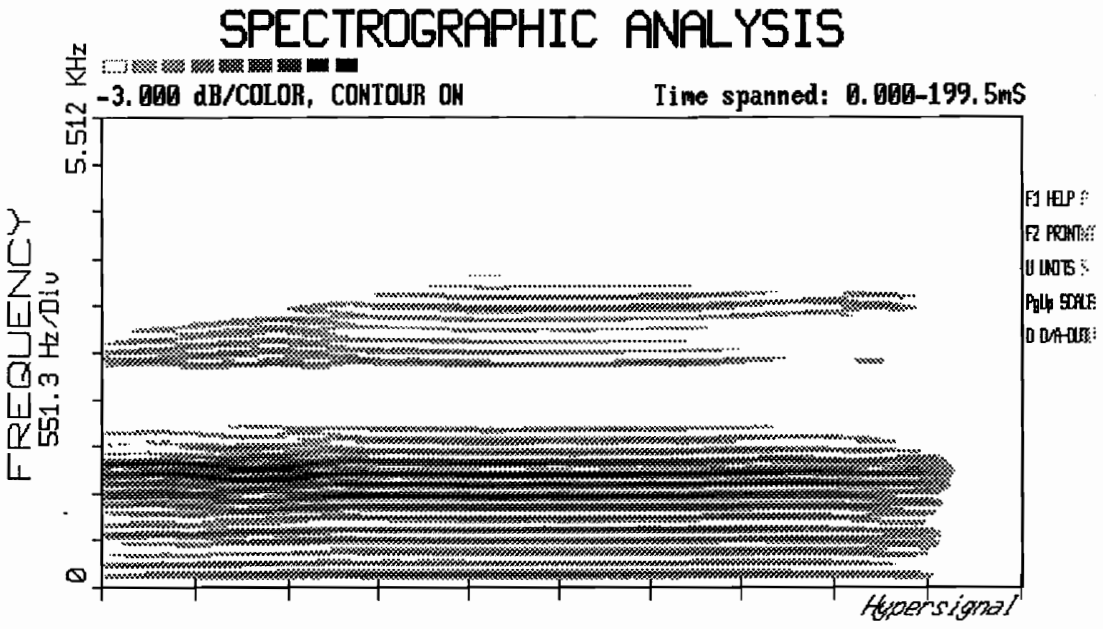


圖15 原始第一聲/Y/音之聲譜圖

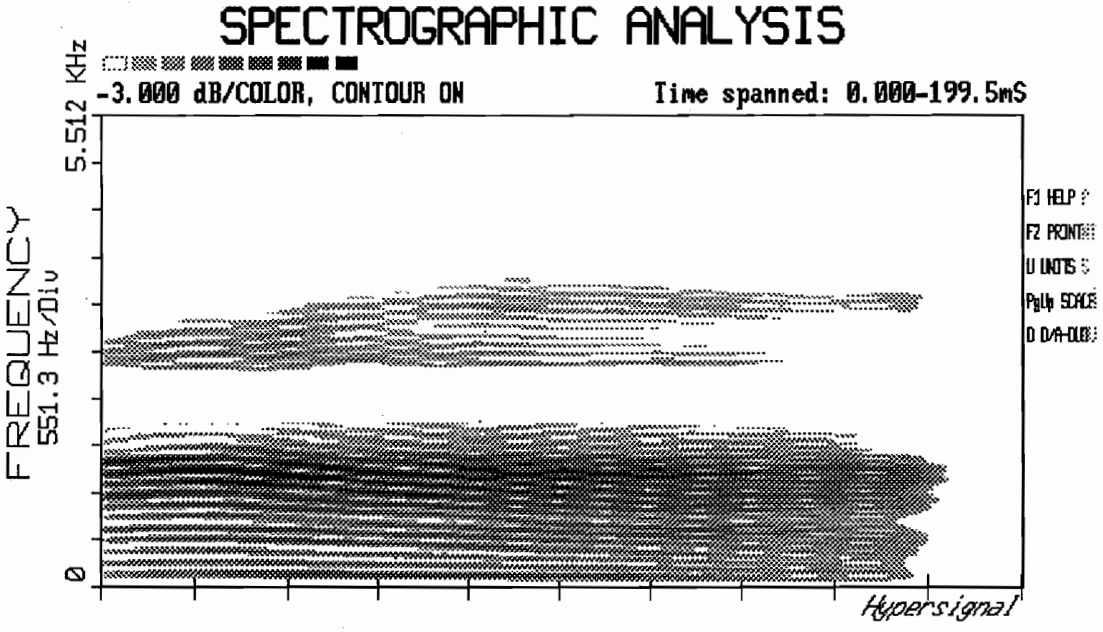


圖16 合成之第四聲/Y/音之聲譜圖

4.2 音長變換

當合成的音節信號，其音長比原始音長大(或小)許多時，共振頻率的走勢(軌跡)是否要成比例地延長(或縮短)，或是延長特定的時間部份(如頻

譜穩定或呈水平時），即要採取線性的或非線性的 time-warping，是一個值得探討的問題，這可以從觀察人自己發雙母音時唸長唸短的差異開始。我們的音節信號合成方法，理論上會採取線性方式來作時間之延長或縮短，不過，這裡仍以實驗分析方式來驗證，以音節/ㄞ/爲例，去分析原始之/ㄞ/音節，其波形如圖17所示，和合成的兩倍長之/ㄞ/音節，其波形如圖18所示，結果得到如圖19和圖20之聲譜圖，比較圖19與圖20，我們看到共振頻率F1、F2的高度及走勢並無不同之地方，不同的是時間軸的尺度，所以我們的合成方法的確是以線性方式來延長時間。

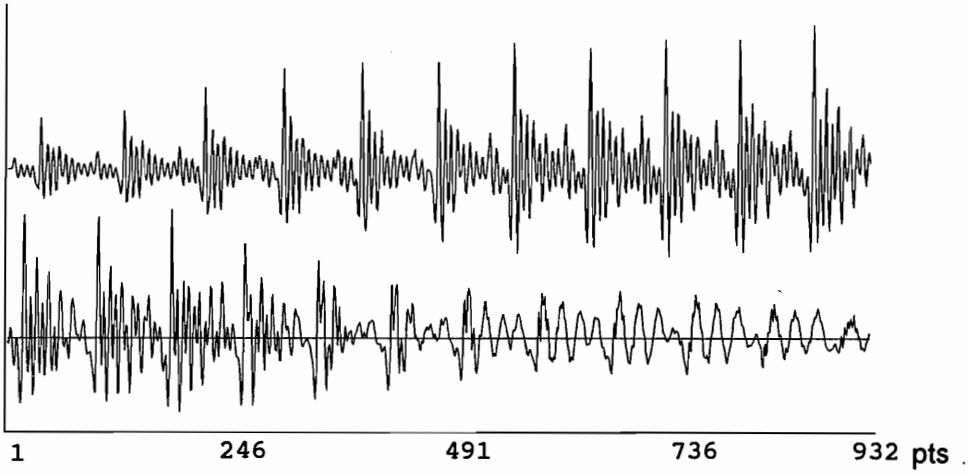


圖17 原始/ㄞ/音之波形

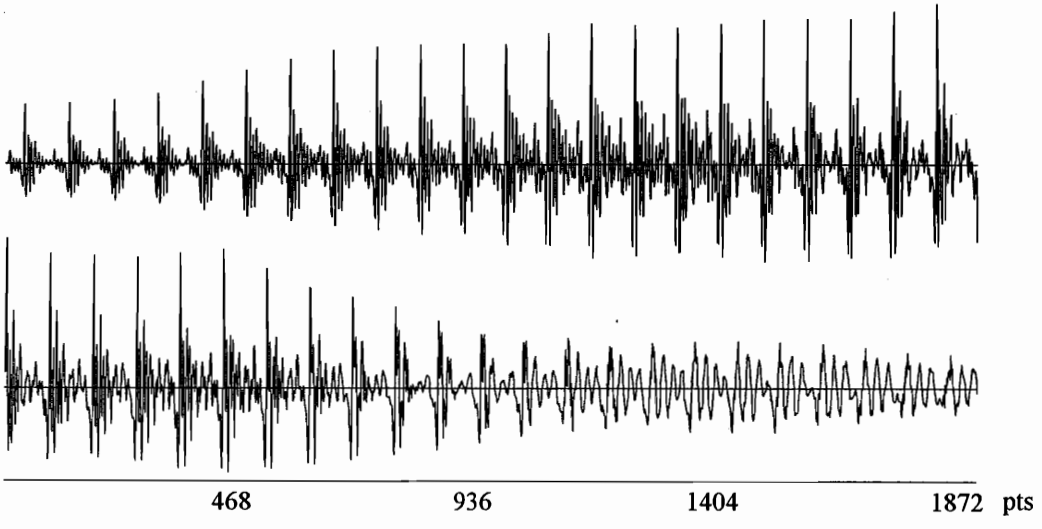


圖18 合成之二倍音長/ㄞ/音之波形

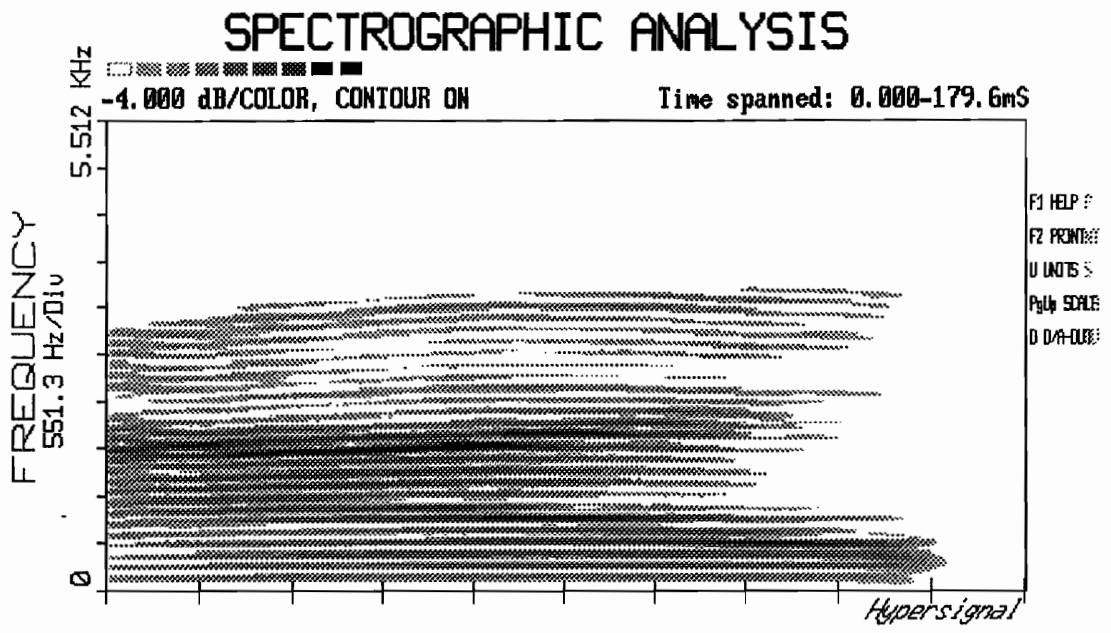


圖19 原始/ㄌ/音之聲譜圖

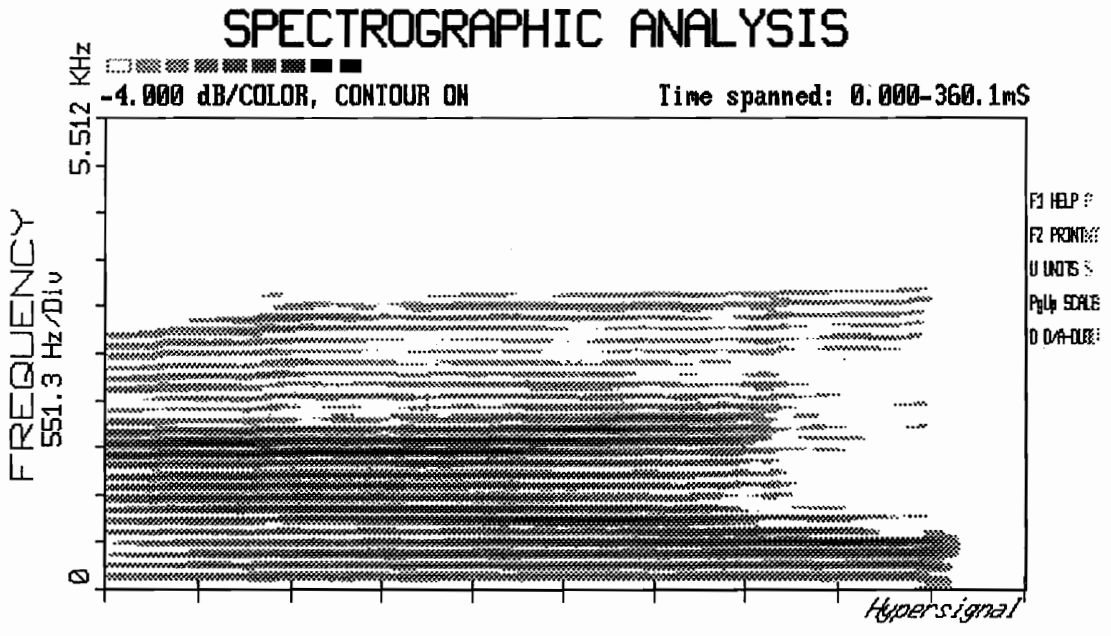


圖20 合成之二倍長/ㄌ/音之聲譜圖

4.3 聲道長變換

我們的音節合成方法以 resampling 來達到提高(或降低)共振頻率的目的，這就相當於縮短(或加長)聲道長度，不過，作 resampling 和音調控制

是兩件獨立的事情，也就是說可在相同音調的條件下去調整共振頻率高度，如圖21和圖22的波形，它們具有相同的音調、相同的週期長度，但是圖21是作一次走1.3點之 resampling 來合成的/ㄚ/音波形，而圖22則是作一次走0.7點之 resampling 來合成的/ㄚ/音波形。圖21和圖22波形所對應的聲譜圖分別如圖23和圖24所示，圖21波形每個週期裡的波峰較多，意味有較高的共振頻率，所以圖23裡我們看到的共振頻率高度要比圖24裡的高許多，事實上，圖23裡的共振頻率會是圖15的共振頻率的1.3倍高，而圖24裡的共振頻率會是圖15的共振頻率的0.7倍高。

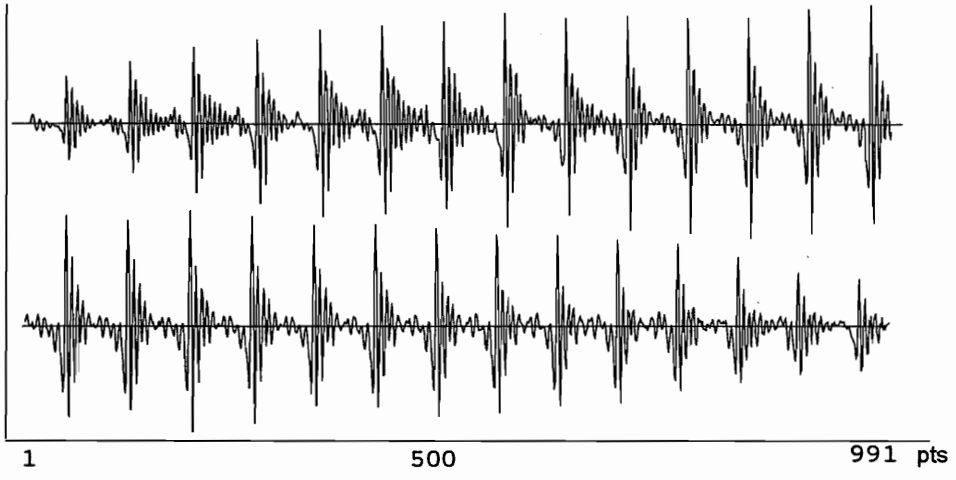


圖21 一次走1.3點之合成/ㄚ/音的波形

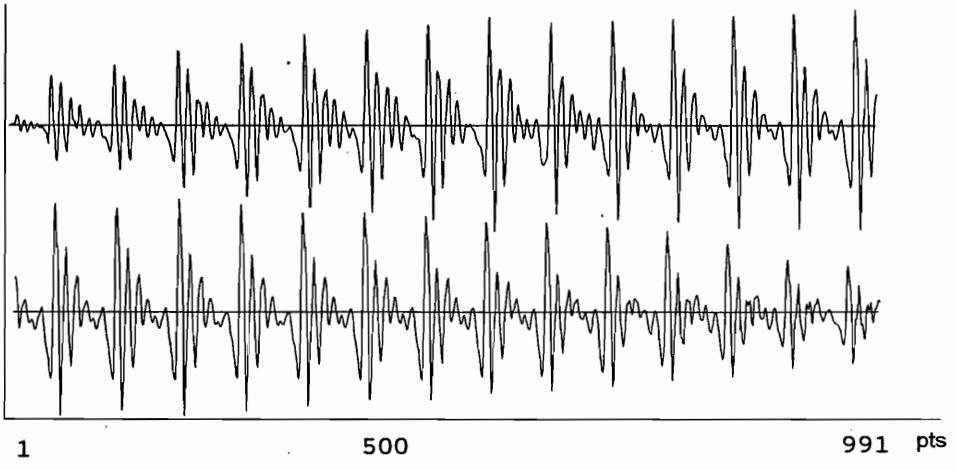


圖22 一次走0.7點之合成/ㄚ/音的波形

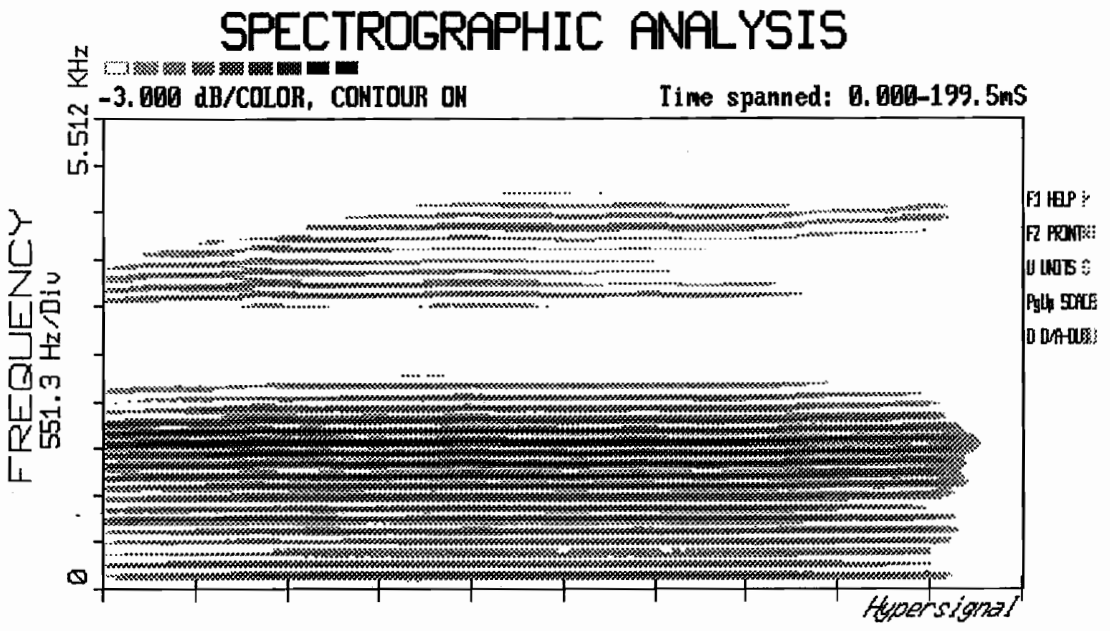


圖23 一次走1.3點之合成/ㄚ/音的聲譜圖

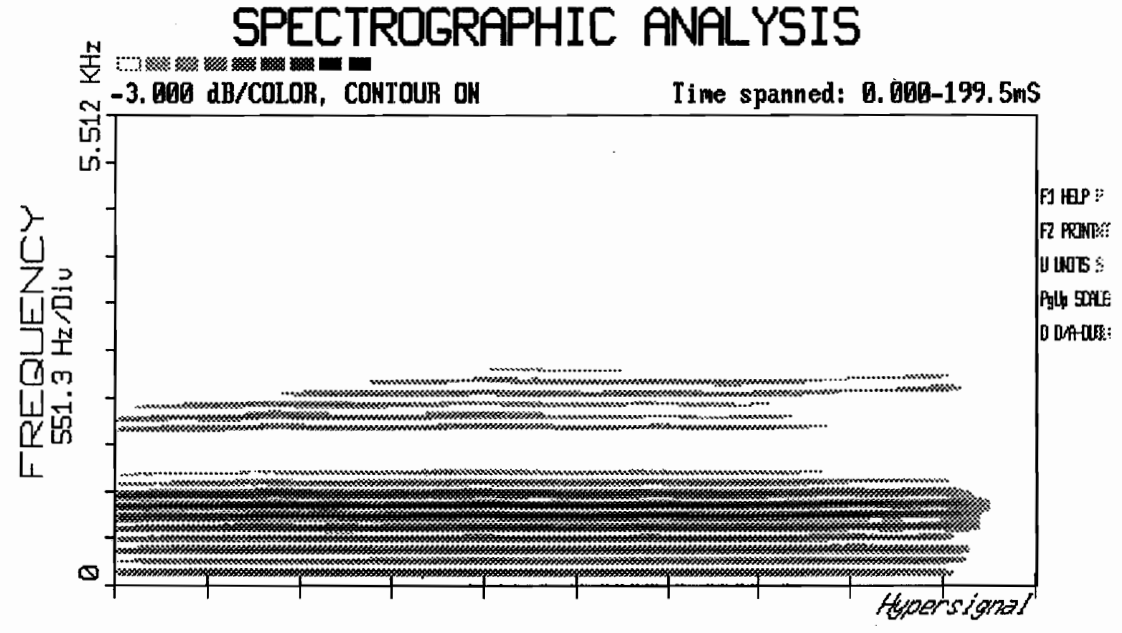


圖24 一次走0.7點之合成/ㄚ/音的聲譜圖

5、結語

本論文提出一個國語音節信號合成的新方法，它的特色在於保留時域

合成方法的清晰音質之條件下，加強了信號控制之自由度，如音調之控制，不會再導致頻譜或共振峰走勢在時間軸上被扭曲；音長之控制，較 PSOLA 技術更具有彈性；聲道長之控制，是一項新的嘗試，以前的文句翻語音系統並未提供，它使得較自然的、及更多的音色能被合成出來，而豐富的音色是拓展文句翻語音系統之應用範圍的基礎，新的應用範圍如雙(或多)主播之新聞播報、小說與故事之講述、甚至於戲劇裡的對話的合成。

除了提出音節信號合成之方法，我們也以此方法去建造了一個原型的中文文句翻國語語音系統，初步聽測合成之語音信號，顯示所提出之音節信號合成方法的確能合成出清晰的語音，並且能夠依照預先的想法，讓前述的三項控制因素獨立地去改變數值。

參考文獻

[1] Lee, L. S., C. Y. Tseng and C. J. Hsieh, "Improved Tone Concatenation Rules in a Formant-Based Chinese Text-to-Speech System", IEEE trans. Speech and Audio Processing, Vol. 1, No. 3, pp. 287-294, 1993.

[2] Chiou, H. B., H. C. Wang and Y. C. Chang, "Synthesis of Mandarin Speech Based on Hybrid Concatenation", Computer Processing of Chinese and Oriental Languages, Vol. 5, pp. 217-231, 1991.

[3] Chen, S. H., S. H. Hwang and C. Y. Tsai, "A First Study on Neural Net Based Generation of Prosodic and Spectral information for Mandarin text-to-speech", Int. Conf. ASSP, pp. 45-48, 1992.

[4] 吳宗憲、陳昭宏、莊欣中，「以 CELP 爲基處之文句翻語音中韻律訊息之產生與調整」，中華民國第八屆計算語言學研討會論文集，第233-251頁，1995.

[5] Atal, B. S. and S. L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave", J. Acoust. Soc. Am., pp. 637-655, 1971.

[6] Markel, J. D. and A. H. Gray, Jr., Linear Prediction of Speech, Springer-Verlag, 1976.

[7] Klatt, D. H., "Software for a Cascade/Parallel Formant Synthesizer", J. Acoust. Soc. Am., pp. 971-995, 1980.

[8] Holmes, J., "Formant Synthesizers - Cascade or Parallel ?", Speech Communication, pp. 251-273, 1983.

[9] Charpentier, F. and M. Stella, "Diphone Synthesis Using an Overlap-Add Technique for Speech Waveform Concatenation", Proc. Int. Conf. ASSP, pp. 2015-2018, 1986.

[10] Hamon, C., E. Moulines and F. Charpentier, "A Diphone Synthesis System Based on Time-Domain Prosodic Modification of Speech", Proc. Int. Conf. ASSP, pp. 238-241, 1986.

[11] Modulines, E. and F. Charpentier, "Pitch-Synchronous Waveform Processing Techniques for Text-to-speech Synthesis Using Diphones", Speech Communication, pp. 453-467, 1990.

[12] Galanes, F. M., M. H. Savoji and J. M. Pardo, "New Algorithm for Spectral Smoothing and Envelop Modification for LP-PSOLA Synthesis", Proc. Int. Conf. ASSP, pp. I-573-576, 1994.

[13] O'Shaughnessy, D., Speech Communication: Human and Machine, Addison-Wesley, 1987.

[14] Rabiner, L. and B. H. Juang, Foundmentals of Speech Recognition, Prentice-Hall, 1993.

[15] Hyperception, Hypersignal Users Manual, 1991.

中英文文句翻語音系統中連音處理之研究

吳宗憲，陳昭宏，林超群

國立成功大學 資訊工程研究所

e-mail address: chwu@server2.iie.ncku.edu.tw

摘要

在本論文中，我們建構了一套高音質的中英文文句翻語音系統，本系統可以將中英文參雜的輸入文句經由經系統處理後，轉為合成的語音訊號輸出，同時我們亦提出了一些在語音合成處理上的實用技術，這些技術能同時滿足在中文及英文合成處理的需求。為了能使系統合成出的語音更自然流利，除了在英文語音合成中採用包含連音資訊的雙音(Diphone)合成單元外，我們並且嘗試在中文連音(Coarticulation)處理方面能有些突破，我們可以將中文連音型態分成三大類來解決連音處理的問題，分別針對中文單音節(Monosyllable)間及英文雙音間的連接，就能量及基週軌跡進行平滑化，以期達到連音處理的目的。根據實驗評估結果顯示，在自然度方面，以 MOS 法評估出來的結果為 3.6 分，若以等級區分則約在良好及尚可之間。而在可辨度評估方面，正確率達 84.2%。

一、緒論

中文文句翻語音系統的研發，主要分佈在華人地區：中國大陸、香港、台灣。至今仍未見一套成熟、可商品化的系統，主要是自然度、流利度及破音字的處理上遇到瓶頸，此外，由於中文文章中也常含有英文文句，因此吾人希望藉由本論文的研究，開發一套發音清晰且具有高度自然語調的中英文語音合成系統。

就中文語音合成模組而言，首先我們錄製了 1432 個由女性發聲的國語單音(含四聲及輕聲)，同時對每一單音除了在錄音過程中要求音長標準化(Normalization)之外，事後又對音高、音量依聲調做標準化的處理。發音模組中我們採用由 CNET 實驗室提出的時域基週同步疊加法(Time Domain Pitch Synchronous Overlap and Add, TD-PSOLA)，在時域上對單音做音高、音長、音量的調整。音韻處理方面，我們均知中

文文法鬆散任意組合之語句甚多，而具有意義的最小代表單位為『詞』，因此我們將中研院八萬詞庫擴充，將日常口語化、習慣化、通俗化之詞加入，成為系統適用的詞庫。並且將詞庫中之詞以四聲相接的情形及字數做分類，統計各個分類之音高、音長、音量的變化情形，並對需要例外處理之詞額外記錄。因此雖然我們採用規則條列法則(Rule Base)來決定如何調整音韻的變化，但我們細分音韻調整為『詞內音韻處理』及『整句音韻處理』兩個模組，先後分別調整音韻，當語音合成時根據輸入的文字，做整合性的音韻調整。在破音字處理上採用破音字分類法，並逐類用有限的規則解決。

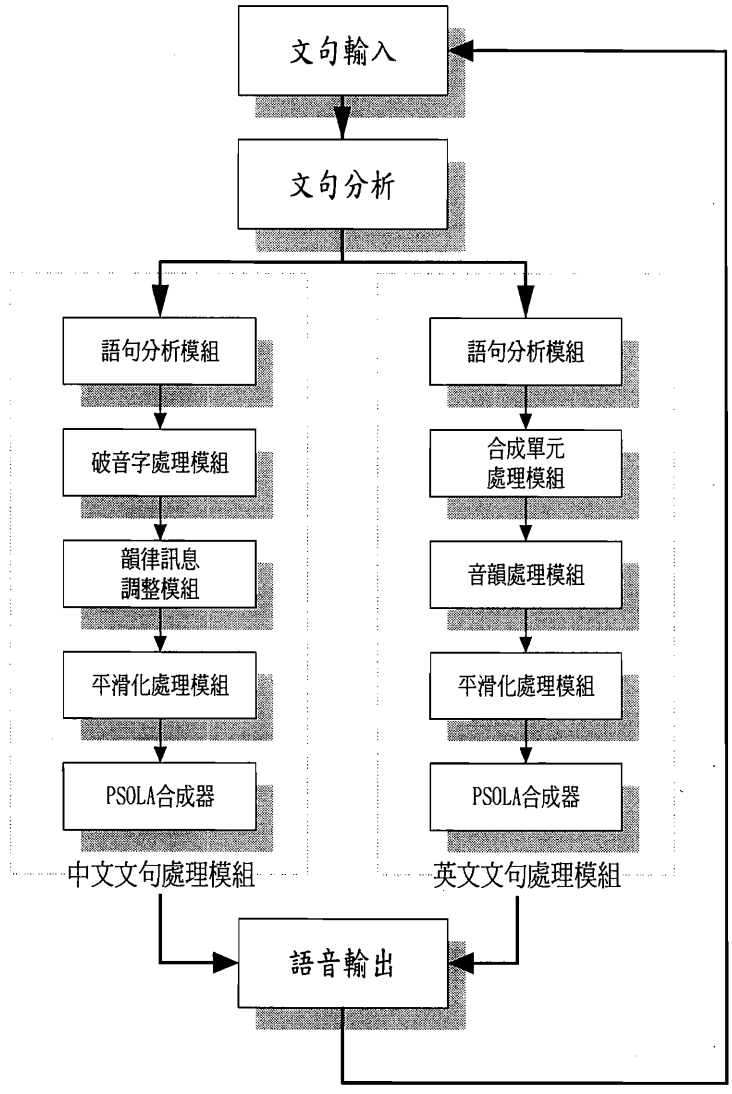
就英文語音合成模組而言，本系統採雙音(Diphone)為基本合成單元。首先我們製作出包含約十三萬英文字之電子字典，再根據電子字典找出 691 個包含所有雙音組合之平衡字(Balanced Word)，我們根據此預先錄製所有平衡字的音檔，再由音檔中分離出所要的雙音使成為單一合成單元。錄音至儲存的過程不做任何失真的壓縮處理；同時對每一平衡字要求在錄音過程中，音長、音高、音量均限制在一範圍內。發音模組我們仍採用時域基週同步疊加法，在時域上對各個組合之雙音作音高、音長、音量的調整。另外我們亦製作出自然拼音法的『字轉音規則庫』，處理創造字或不存在電子字典中之單字。詞綴庫的建立也是必要的，協助字轉音規則庫處理，如此就能完善的建立出英文文句翻語音系統。

二、系統架構

此套中英文文句翻語音系統主要可分為文句分析模組、中文文句處理模組、英文文句處理模組等幾個部份，其中，中文處理模組又可細分為語句分析模組、破音字處理模組、韻律訊息調整模組、平滑化處理模組及 PSOLA 合成器等五個基本模組。而英文處理模組又可細分為語句分析模組、合成單元處理模組、音韻處理模組、平滑化處理模組及 PSOLA 合成器等五個基本模組。透過這些模組的運作，可將任意輸入的中英文夾雜之文句，轉成語音輸出。基本架構圖如圖(一)所示，各模組功能簡介如下：

◆ 文句分析模組：

此模組之功能乃是分辨輸入之文句，何處是中文句(含標點符號及阿拉伯數字)的起始點及結束點；何處是英文文句及空白的起始點及結束點；如此分離出不同文句之屬性，即可將文句送入適當的處理模組加以處理。



圖(一) 系統基本架構圖

◆ 中文文句處理模組：

中文處理模組專責處理中文語句，包含中文字串、標點符號及阿拉伯數字的語音合成，其中又細分為：

1. 語句分析模組：

利用系統中文詞庫對輸入的中文語句進行構詞、斷詞的工作，找出詞邊界、詞屬性及斷句邊界，並配合發音字典取得各個字相對應的注音符號

與音檔編號。

2. 破音字處理模組：

使用破音字分類法及配合系統中文詞庫、讀音字典等來處理破音字，且修正相對應的注音符號與音檔編號。

3. 韻律訊息調整模組：

主要細分為詞內音韻處理及整句的音韻處理，因為使用我們特別處理過的詞庫包含了口語化之詞，況且詞庫是已知且有限的，所以詞內音韻處理我們事先可以做好，執行時只要針對句型判斷，做整句的音韻處理即可。

4. 平滑化處理模組：

針對可連音之中文字做連音處理，並儘量使之前後基週頻譜平滑化。

5. PSOLA 合成器：

採用時域基週同步疊加法，參考單音的基週資料，在時域上作音長、音高、音量的調整。

◆ 英文文句處理模組：

英文處理模組專責處裡英文語句，包含英文字串、空白(Space)的語音合成，其中又細分為

1. 語句分析模組：

此模組主要是處理輸入之英文字串，配合電子字典、詞綴庫及字轉音規則庫，將英文字串轉換成相對應的音標及重音組合。

2. 合成單元處理模組：

此模組的功能乃是接收由『字轉音標及重音處理模組』產生之音標及重音字串，由此字串至音檔庫中找出相對應之音檔。

3. 音韻處理模組：

根據音標標定之重音及次重音調整整個字的音韻，在調整此音韻變化之前必須對各個雙音(Diphone)的基週軌跡(Pitch Contour)做標準化，再由音

韻規則做調整。

4. 平滑化處理模組：

由於雙音是由其他平衡詞之音檔中切出來的，所以前後雙音的基週可能並不一樣(或相差極大)，即使經過音韻處理模組處理之後仍有基週不平滑之情形，此模組將前後雙音正規化(Nomorlization)，如此便能輸出自然、平順的合成語音。

5. PSOLA 合成器：

跟中文文句處理模組一樣，我們同樣也採用時域基週同步疊加法，參考合成單元的基週資料，在時域上作英文字內各個雙音的音長、音高、音量的調整。

三、系統資料庫建立

◆ 中文資料庫建立

1.語料庫：

在考慮實用性及高品質的原則下，本系統係採用成大資訊所中文實驗室錄製之音檔，此音檔採用 16 bits 音效卡及單向收音之單聲道麥克風錄製合成單元，且使錄音過程標準化(音長大約 0.27ms 、音高為錄音者之正常發音音高、音量差異在 20%以下)。

2.音韻規則庫：

詞內音韻規則庫：我們先將詞庫中的詞以四聲相接的情形及字數做分類，統計各個分類之音高、音長、音量變化情形並對需要例外處裡的詞額外記錄。

整句音韻規則庫：整句音韻調整我們採用規則條列式(Rule Base)的作法，在執行時，根據輸入的文句做整體的調整。其規則條列如下：

- 換氣循環：模擬人的換氣動作，約每六字為一個循環，遇到詞則提前或延後換氣。換氣前音量、音高逐漸降低，換氣時略為停頓，換氣後音量、音高較為提高。

- 句首處理：每個句子的句首音高略為提高。
- 句尾處理：句尾要拉長音長、降低音高、能量遞減。
- 二聲接二聲：提高第一字的音高。
- 介詞處理：遇到常用的字，如：的、著、是、到、和、得等，縮短音長、降低音量，並拉長前一字的音長。
- 標點符號：在文章中遇有各類標點符號時加入長短不等的停頓，如表(一)。
- 停頓：詞之前、句尾亦加入適當的停頓。
- 句型：將文句分為直述句、疑問句、驚嘆句、命令句，分別調整其音高、音長、音量的整體走勢。

表(一) 標點符號與停頓時間關係表

標點符號	，	；	。	？	！
停頓時間	480ms	500ms	520ms	540ms	560ms

3.基週資料庫：

基週標記位置的取得首先使用程式整批(Batch)作業，再由人工逐一檢查、校正，並連同每一個合成單元的音高、音量平均值(mean)、及音長、穩定區位置等特徵參數一起儲存，以增加系統效率。

◆ 英文資料庫建立

- 電子字典

一套包含大約十三萬字、約佔 2.8 Mega Bytes 儲存空間的英文電子字典。

- 雙音語料庫

在評估過各種合成單元的合成效果之後，我們採用了合成效果既不錯，所佔儲存空間又適當的『雙音』(Diphone)為本系統的基本合成單元。首先由電子字典中只考慮常用字，再根據音素(Phoneme)為組合雙音的基礎，找出

且錄製平衡字共 691 個，再由我們發展的切音工具程式在這些平衡字中標定出 1187 個雙音邊界。同樣的，我們仍然採用語錄製中文合成單元相同的語者及錄音規格，採用相同規格及語者是為了不使合成中英文句時，會有兩個不同人之聲音，而且如此音高、音量及唸法習慣皆相同，將有助於合成出高品質、高自然度之語音。

- 詞綴庫

我們另外從電子字典中分離出 383 個詞綴，其中包含前置詞綴 205 個後置詞綴 178 個，為處理英文時態字及複數字、組合字之用。

- 字轉音規則庫

當我們遇到創造字及組合字或不包含於此電子字典之中的單字時，若系統跳過不唸則會使使用者對本系統的效能大打折扣，[15] 對此種問題提出解決的方法。

- 基週資料庫

與中文模組相同採用時域基週同步疊加法(PSOLA)，不同的是用來調整英文字之字調，因此我們需要知道每個被合成單元聲音波型的基週標記位置(Pitch Mark Position)。所以我們利用程式標定再經過人工檢查，預先儲存每個被合成單元聲音波型的基週標記位置及音高、音量、音長及穩定區位置等特徵參數。

四、中文文句處理模組

◆ 中文文句處理流程

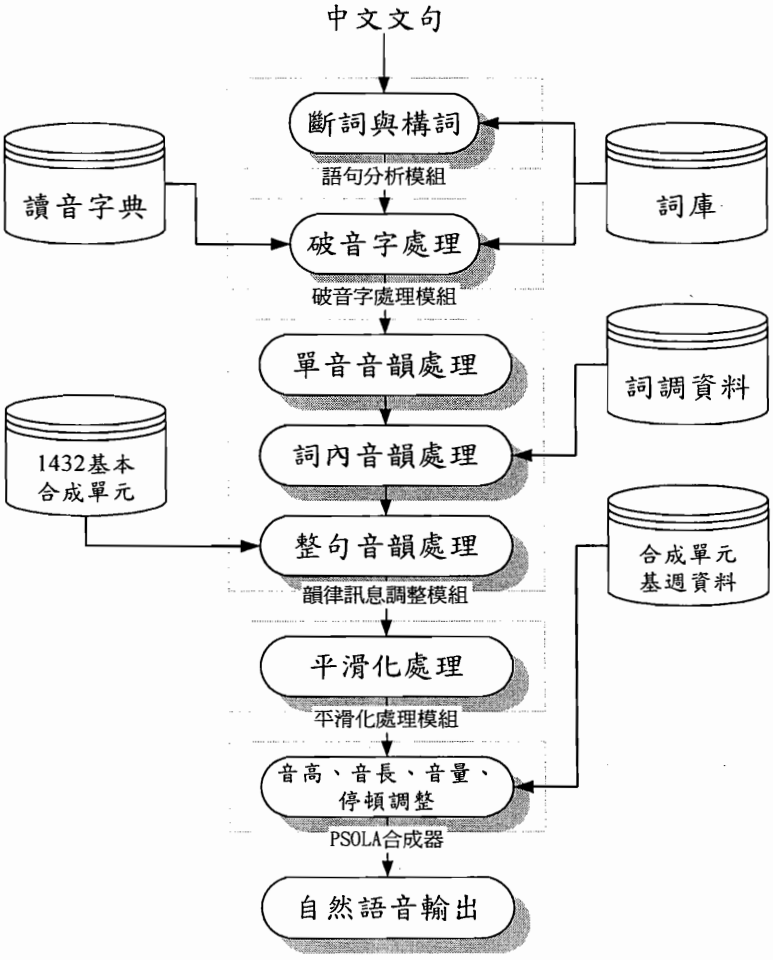
中文文句處理模組(如圖二)又可細分為語句分析模組、破音字處理模組、韻律訊息調整模組、平滑化處理模組及 PSOLA 合成器等五個基本模組。其中除平滑化處理模組將留在連音處理討論外，其餘模組將列在以下各節逐一說明。

◆ 語句分析模組

- 斷詞

基本上語句分析模組所處理的第一個步驟便是斷詞(word identification)，斷詞的目的在標示出輸入的中文文句中所包含之所有的詞，由於中文的文法極為鬆散且字與字之間的組合極多，所以同一中文字串可能會有許多不同的斷詞方式，也就是語意上及斷詞上的混淆(ambiguity)，因此中文的斷詞事很困難的事。

本論文採用的斷詞方式是利用統計的方式，將我們常用的詞事先建立一套詞庫，經過我們建立好我們適用的資料結構之後，如此就可正確且快速地找出所有包含於輸入中文句中的詞組合，再輸入構詞程序找出最佳之詞組合。



圖(二) 中文文句處理流程

- 構詞

雖然我們已經將原來引用的中研院詞庫修正注音成為口語化詞並且加入新詞，但想要包含所有的詞是不可能的，所以我們只能將非規律性組合的詞與出現頻率較高的詞蒐入詞庫中，具有規律性的部份則由構詞規則來處理。

本系統採用的規則如下：

- 長詞優先：以最少詞數涵蓋最多字數者優先。
- 左詞優先：兩個相鄰詞發生搶中間詞之時，以左邊的詞為優先。
- 詞長平均：例如『台南市政』可以斷成(台南、市政)或(台南市、政)兩種組合，我們採用兩詞長較相近的組合。
- 前綴詞處理：包括小、可、老、好等處理。
- 數量詞處理：利用詞性將數詞與量詞合併輸出。
- 重疊詞處理：如『快快樂樂』在斷詞後會斷成(快、快樂、樂)，處理之後將(快快樂樂)當成一個詞輸出。

◆ 破音字處理模組

為研究破音字的問題，我們總共統計出中文字共有 960 個字擁有兩個以上的讀音，而處理方面，我們採用本實驗室研發出之『破音字分類處理法』[13]來處理合成破音字語音時所遇到的問題。

◆ 音韻訊息調整模組

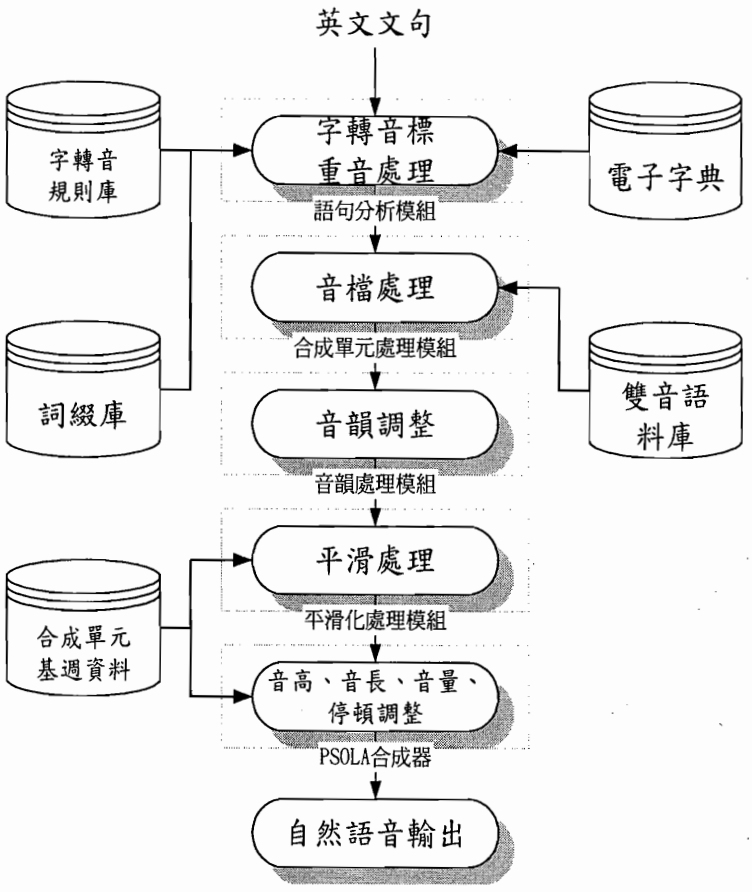
韻律訊息的調整本論文主要是採用音韻規則庫來作為調整韻律訊息的依據，並且使用 PSOLA 合成器調整音高、音長、音量，詞內音韻利用統計法建立規則庫，整句音韻規則庫採用條列法

五、英文文句處理模組

◆ 英文文句處理流程

英文處理模組(如圖三)可細分為語句分析模組、合成單元處理模組、音韻處理模組、平滑化處理模組及 PSOLA 合成器等五個基本模組。其中平滑化處理模

組將在下一節報告。

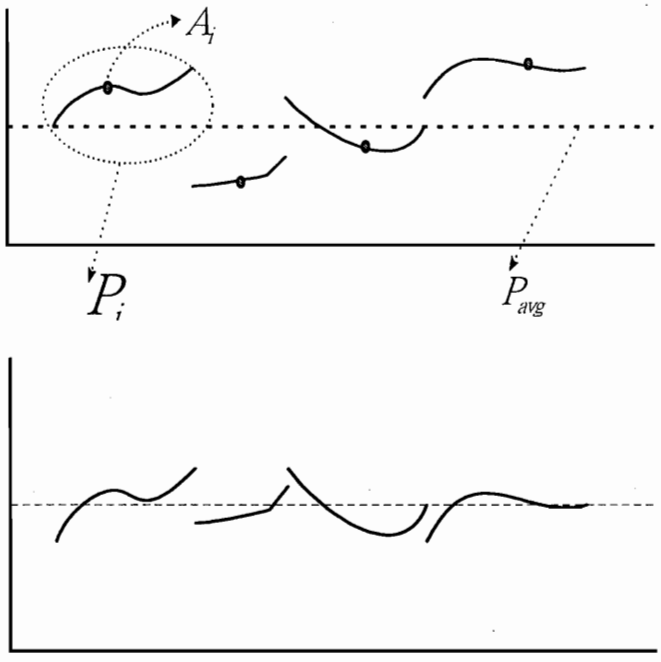


圖(三) 英文文句處理流程

◆ 語句分析模組

英文語句分析模組的處理較中文語句分析為簡單，因為英文的每一單字之間均有空白間隔存在。此模組的功能在處理輸入之英文字串，並且由字串中分離出英文單字逐一參考電子字典，找出此單字的音標、音節及重音標示，另外在此步驟中亦找出對應雙音(Diphone)檔案的索引(Index)及標示出那一雙音需做重音或輕音處理。

此外，為處理不存在於電子字典中之英文單字(包括組合字、創造字及時態字等)，我們解決的方法是參考系統中建立的字轉音規則庫及詞綴庫，若經由詞綴刪除之後可以在電子字典中找到單字，則以字典中之音標為主，若此詞綴處理



圖(四) 基週正規化

之後仍不在電子字典之中，則採用參考字轉音規則庫的自然拼音法將單字的音標、音節數、重音及輕音標示等創造出來。並仍需做雙音及音標的對應工作。

◆ 合成單元處理模組

此模組接收由語句分析模組處理過且輸出的音標及重音、輕音節標記，配合雙音索引檔由語料庫中找到雙音音檔並載入記憶體中。

◆ 音韻處理模組

● 正規化(Normalization)

由於各個雙音是由不同的平衡字語料中切割下來的，所以組合為另一單字發音時，會有前後音大小聲不同，基週變化太快的現象，雖然在錄音的過程中我們已對音量及基週大小做了一些限制，但仍有必要使被合成單元的能量、基週正規化(Normalization)，所以我們在音韻處理模組中首先處理的動作就是正規化，如圖(四)，過程簡述如下：

一、求各個 Diphone 之 Pitch Contour P_i 及平均 Pitch A_i

二、求 Pitch 之總平均 $P_{avg} = \frac{\sum_{i=1}^{n} A_i}{n}$ (5.1)

三、Normalize $\forall P_i,\ P_i = P_i \times \frac{P_{avg}}{A_i}$ (5.2)

- 音韻處理

本論文解決韻律訊息的問題是採用音韻規則的建立，音韻規則簡介如下：

一、音量規則：

1. 母音音量較子音為大。
2. 重音節的音量比非重音節大。
3. 有聲音節的音量通常隨著音高上升而增加。
4. 句子末了音節的音量通常較小，尤其是非重音節。

二、音長規則：

1. 重音節的音長比非重音節長。
2. 在停頓之前的音節中，增加其母音或子音的音長。
3. 在片語末端的音節中，增加其母音或子音的音長。
4. 在單字音節的音節段中，稍微縮短其中非最後音節段的音長。
5. 略為縮短多音節單字中的音節段的音長。
6. 縮短單字所有的子音中非字首的子音的音長。
7. 明顯增加被強調的單字中的母音的音長。

三、音高週期規則

1. 句首較句末有較高的音高。
2. 在無聲子音附近的音高比有聲子音附近的要高一點。
3. 重音節的地方有較高的音高。

六、連音處理

在中文的連音處理方面，由於本系統的中文音韻訊息調整模組已經將詞內音韻及整句音韻先後處理，且已有相當不錯的效果，所以我們只要針對單音(syllable)

與單音連接時做必要的處理即可。在比較過一些方法及試聽實作後的音質之後，我們決定應用中興大學應數所發展的連音處理模型[14]，加入我們發展的基週平滑法，改良成為新的連音處理模組來處理本系統合成語音的連音問題，連音型態分類及處理表如表(二)。

茲就重疊相加法分為能量平滑及基週平滑簡述如下：

表(二) 連音型態分類及處理表

前音節之最末音素	後音節之最前音素	連音型態	處理方式
ㄓㄔㄕㄖㄗㄘㄙㄚㄛㄜㄝㄞㄟㄠㄡㄢㄣㄤㄥㄦㄧㄨㄩ	清子音： ㄅㄆㄈㄉㄊㄍㄎㄏㄐㄑㄒㄓㄔㄕㄗㄘㄙ	停頓處理	加入 10ms 的靜音停頓
	濁子音： ㄇㄋㄌㄖ	緊密連接	無停頓
	母音： ㄚㄛㄜㄝㄞㄟㄠㄡㄢㄣㄤㄥㄦㄧㄨㄩ	重疊連接	重疊相加法

- 能量平滑：

$$s[i] = s_1[i] \times w_1[i] + s_2[i] \times w_2[i] \quad \text{for } i=0 \text{ to } M \tag{6.1}$$

$$\text{其中 } w_1[i] = \begin{cases} 1 & 0 \le i \le \alpha L \\ \dfrac{1}{(\alpha-1)L} i + \dfrac{1}{1-\alpha} & \alpha L \le i \le L \\ 0 & \alpha L \le i \le M \end{cases} \quad 0 < \alpha < 1 \tag{6.2}$$

$$w_2[i] = 1 - w_1[i] \tag{6.3}$$

S_1為前一 syllable 之數位訊號，S_2為後一 syllable 之數位訊號，且根據實驗，α取 10/12 至 11/12 為佳。

- 基週平滑：

我們觀察到如果屬於重疊連接連音型態，則前後語音訊號的基週軌跡(pitch contour)大部分都是平滑的，而我們錄製的語料是以單音錄製，所以當前後音連接時基週軌跡絕大多數並不平滑，導致音調上的不自然，為此我們也提出一基週軌跡平滑化的方法，簡述如下：

為了要保持原基週軌跡的走勢，我們採用旋轉的觀念

$$\begin{bmatrix} \sin\phi & \cos\phi \\ \cos\phi & -\sin\phi \end{bmatrix} \begin{bmatrix} \mathrm{x} \\ \mathrm{y} \end{bmatrix} = \begin{bmatrix} x' \\ y' \end{bmatrix} \tag{6.4}$$

其中

$$\phi = \arccos\left(\frac{a^2 + b^2 - c^2}{2ab}\right) \tag{6.5}$$

$$a = \sqrt{((1-\alpha)L)^2 + p(L)^2} \tag{6.6}$$

$$b = \sqrt{((1-\alpha)L)^2 + A} \tag{6.7}$$

$$c = p(L) - A \tag{6.8}$$

L 為基週軌跡長度，p(L)代表軌跡上第 L 點之值，A 為前一音之基週軌跡的最後一點值，與後一音之基週軌跡的第一點值之平均，αL 為旋轉起始點。另外由於旋轉過後之基週軌跡，長度上可能會比我們需求的要長或短，因此我們再使用 up-sampling 或 down-sampling 的方法來達到我們的要求。

以上處理是以中文為例，在英文平滑化處理上亦可採用相同的方法，唯因為英文的合成單元-雙音比中文的合成單元要短的多，而且相連接的雙音之間的自相關性更高，若經由一般的方法處理很難會有效果，經由以上的平滑化處理流程之後，我們可以相當輕易地得到自然度較高的合成語音。

七、實驗結果與系統效能評估

一般評估文句翻語音系統的效能大多採用『自然度』(naturalness)與『可辨度』(intelligibility)對系統進行測試，我們相同也朝這兩方面對系統效能進行評估，但此種評估方式較為主觀，因此我們也以圖表示出語音波型、基週軌跡等數據作為客觀評估的依據。實驗對象除實驗室同仁(專業領域)八人之外，另也對非實驗室(一般人士)五人進行測試，評估項目說明如下：

◆ 主觀評估：

- 自然度評估(MOS 法)：

在自然度評估的方法中，測試樣本分別包括中文(詞、句子、短文)及英文單字的內部測試(inside test)及外部測試(outside test)，其評估方法採用平均鑑定分數(Mean Opinion Scores, MOS)法，將結果分為優良(excellent)、良好(good)、尚可(fair)、差(poor)及極差(unsatisfactory)五個等級，分別給予 5 至 1 不等的分數，測試結果如表(三)所示。

表(三) 自然度評估表

測試種類		數量	自然度(MOS)
中文	二字詞	200	4.0
	三字詞	200	4.0
	四字詞	200	3.8
	句子	100	3.6
	短文	10	3.6
英文	inside test(word)	100	3.1
	outside test(word)	100	2.9
平均			3.6

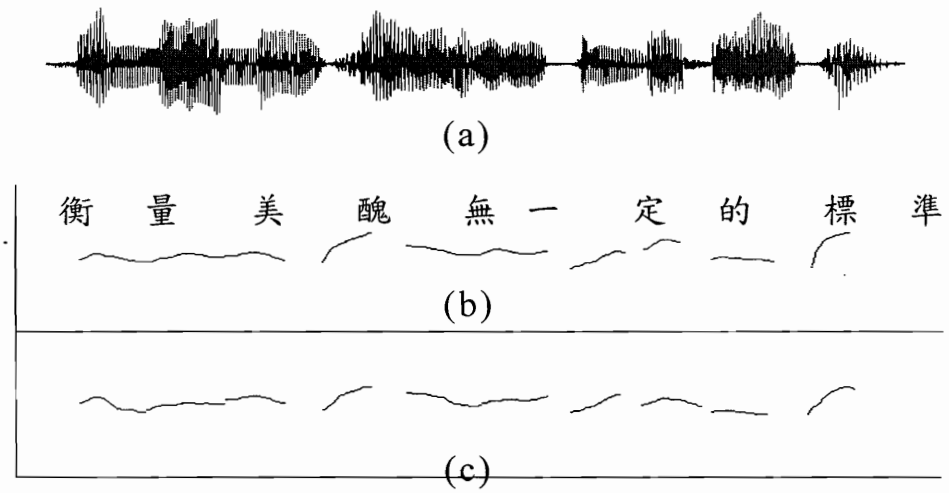
表(四) 可辨度評估表

測試種類		數量	可辨度
中文	單字詞	1432	88.9%
	二字詞	200	92.7%
	三字詞	200	90.4%
	四字詞	200	93.0%
	句子	100	93.5%
英文	inside test(word)	100	72.8%
	outside test(word)	100	57.9%
平均			84.2%

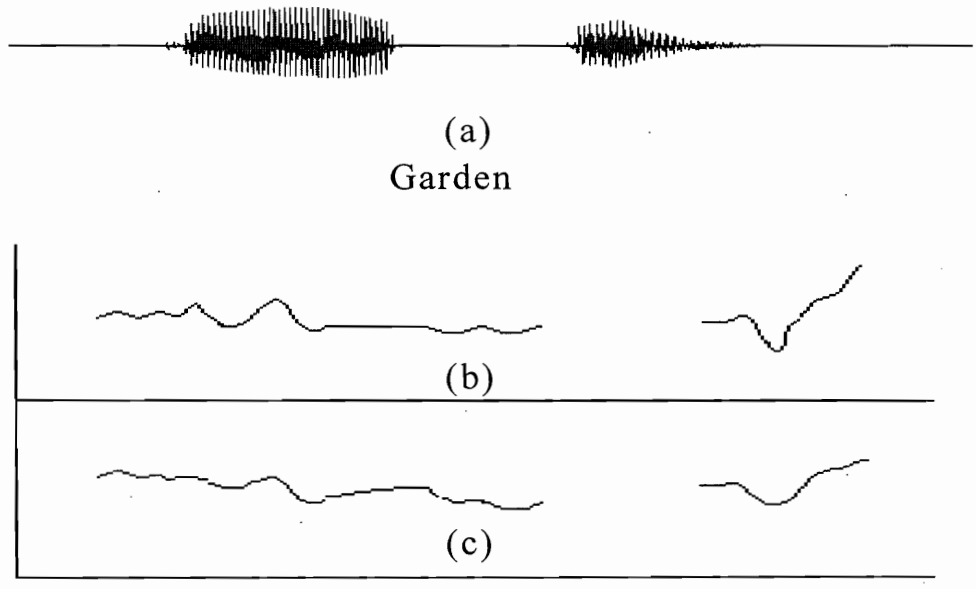
- 可辨度評估：

在可辨度的評估方面，測試樣本以系統合成的語音為主，分別包括中文(詞、句子)及英文單字的內部測試(inside test)及外部測試(outside test)，評估方法是我們由系統輸出合成的語音由受測人員將聽到的中文語音以聽寫的方式寫出標示的注音符號，聽到的英文語音默寫出單字拼法，若拼出的單字其唸法與原語音類似則以正確計算，最後統計正確的文句與受測者默寫結果之間的差異，可辨度評估結果如表(四)所示。

根據實驗評估結果顯示，在自然度方面，以 MOS 法評估出來的結果為 3.6 分，若以等級區分則約在良好及尚可之間。而在可辨度評估方面，平均有 84.2% 的正確率。綜合以上兩項的評估，雖然未能達到極高分數的評估結果，但已經能達到可以接受的程度了。



圖(五) (a)原始的語音波型(b)原始基週軌跡(c)合成語音之基週軌跡圖



圖(六) (a)原始的語音波型(b)原始基週軌跡(c)合成語音之基週軌跡圖

◆ 客觀評估：

我們以一句中文句及一個英文單字的語音波型圖為例，分別展示出原始的語音波型、基週軌跡及合成之後的基週軌跡，如圖(五)、(六)所示。

由基週軌跡比較圖中我們可一發現平滑化的處理已發生效果，他將原本基週軌跡相差過多的雙音合成單元都做了平滑化處理，使合成出的語音能有較自然的輸出。

八、結論與討論

在本論文中，我們建構了一套中英文文句翻語音系統，本系統可以將中英文參雜的文句(包括檔案)經由演算法的運算，配合語料庫及其他資料庫，轉為語音合成的聲音輸出，一般而言合成的語音還算自然流利，在時效上也可以達到即時處理的要求，值得一提的是，本系統提供了一些在語音訊號處理上的實用技術，這些技術均能在中文及英文處理上共同使用。經由實驗，本系統各方面的綜合表現均能令人滿意。

經由數次實驗之後，我們發現中文文句翻語音模組的整體表現較英文文句翻語音模組為佳，原因之一是因為本實驗室針對中文文句翻語音系統發展已行有多年，而英文的語音合成技術除了從一些國外的論文、期刊中可以得到一些有價值的資料外，有關雙音語料庫建立、英文語法規則等基本知識我們均可以說是從頭開始建立，另外，非母語的限制更是從錄製語料庫開始便困擾著我們，以下列出三大方向提供未來參考改進之：

◆ 雙音語料庫：

就中文語音合成而言，我們的單音語料是相當標準的，而且以此語料合成的效果也有口皆碑，但相較中文合成而言我們的英文合成就不成熟許多，我們認為最主要的原因是出在英文非我們的母語，首先在製作平衡字時，雖然我們已經儘量從常用字著手，但仍有許多字的讀音對語料錄製者而言是非常難錄製的，再者，我們在實驗中發現，有些字無論如何處理，合成的效果仍有瑕疵，我們判斷可能是有一些平衡字含有『破音』成份(如：sk、kr、scr 等)，而這些『破音』成份若被切成雙音且被合成於另一不含『破音』成

份的單字時，合成語音的品質就會大打折扣。最簡單的解決方法是的不採用這些含有『破音』成份的雙音，如果儲存空間允許，甚至可以將固定成形字首及字尾連同『破音』成份的雙音均由三音(Triphone)或多音(Polyphone)取代，如此將可大大提昇合成語音的品質。

◆ 音韻資料庫：

雖然我們已根據我們提出的音韻處理規則分別對重音及次重音加以處理，但效果仍有待改善，原因乃是除了具有重音、或次重音標記的音標符號我們要對其特別處理之外，對於重音或次重音標記附近的雙音合成單元要如何處理，我們至今仍無法提出一種有效的規則對其處理，對此我們構想未來應朝向類似連續語音資料庫中詞音韻歸納的方法，對應出合成單元在平衡字中的相對位置，並擷取其音長、音高、音量、停頓等韻律訊息特徵，加以參考之，另外，整句的韻律訊息調整亦可由此方法實現。

◆ 整句連音處理：

本論文的研究之中，為降低系統複雜度，因此我們並未對整句的連音做適當的處理，簡單的說，我們合成出的英文語音中單字與單字間應有的連音資訊都被我們忽略了(例如： make up on....、 would you mind...等)，未來應統計出那些音標的組合會產生連音，會產生『變調連音』或『不變調連音』還是連音之後省略後面母音或變化為輕音的各個類型，分別加以處理之。

九、誌謝

感謝國科會經費補助(計畫編號 NSC85-2622-E-006-003) ，使得本論文能順利完成。

參考文獻

[1] F.J. Charpentier and M.G. Stella, "Diphone Synthesis Using an Overlap-Add Technique For Speech Waveforms Conca-tenation", Intern. conf. on ASSP, ICASSP-86, pp. 2015-2019, 1986.

[2] J. Allen, "Synthesis of speech from unrestriced text", Proc, IEEE, vol. 64, pp. 422-

433, 1976.

[3] L. R. Rabiner and R. W. Schafer, "Digital Processing of Speech Signals", p398-p403, 1987.

[4] L. S. Lee, C. Y. Tseng and M. Ouh-Young, "The Synthesis Rules in a Chinese Text-to-Speech System", IEEE Trans. On Acoust Speech, and Signal Processing, vol. 37, No. 9, pp. 1309-1319, September 1989.

[5] E.Moulines, F. Charpentier, "Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis using Diphones", Speech Comm, Vol. 9, pp.453-467, 1990.

[6] 熊明德,"中文語音規則合成之研究", 淡江資工所碩士論文，民國 80 年

[7] 陳克建、陳正佳、林隆基, "中文語句分析的研究—斷詞與構詞", 技術報告, TR-86-004, 中央研究院, 1986

[8] H. Klatt, "Review of Text-to-Speech Conversion for English", J. Acoust. Soc. Amer., vol. 82, no. 3, pp. 737-793, Sep. 1987.

[9] Lin-Shan. Lee, Chiu-Yu. Tseng, and Ching-jiang. Hsieh, "Improved Tone Concatenation Rules in a Formant-Based Chinese Text-to-Speech System" IEEE Trans., Speech, Audio Processing, vol. 1, no. 3, pp. 287-294, Jul. 1993.

[10] 盧承周、余秀敏、劉繼謚,"中文發音辭典的建立和破音字的發音辨別系統", 電信研究雙季刊, pp. 287-295, 19 卷 13 期, 1989.

[11] Bigorgne, et al, "Multilingual PSOLA Text-to-Speech System", in Proc. ICASSP93, pp. II-187-II-190.

[12] Ngor-Chi Chan, Chorkin Chan, "Prosodic Rules for Connected Mandarin Synthesis", Journal of Information Science and Engineering, pp. 261-281, 1992.

[13] 蔡祈岩, "文句翻語音系統中破音字及音韻處理之研究", 成大資工所碩士論文，民國 84 年

[14] 陳志祥,"國語連續語音連音型態之初步研究", 中興大學應數所碩士論文，民國

83 年

[15] 林超群,“中英文文句翻語音系統中連音處理之研究”, 成大資工所碩士論文,民國 85 年

尋易(Csmart-II)：智慧型網路中文資訊檢索系統

(An Intelligent Chinese Information Retrieval System for the Internet)

簡立峰，李明哲，陳明權，陳宏明，陳丁旗，
董惟鳳，李宏俊，黃敦怡，張元貞

中央研究院資訊科學研究所

E-mail: lfchien@iis.sinica.edu.tw

摘要

考慮網路中文電子資源的有效利用，我們將原本針對中文全文檢索設計的尋易(Csmart)系統發展成新一代 Csmart-II 智慧型網路中文資訊檢索系統，包括整合資源發掘與過濾、資訊檢索與語音介面技術。使得 Csmart 系統可以以語音或鍵盤輸入近似自然語言查詢檢索網路即時新聞、BBS 論壇、中文 Web Pages 等網路資源。本文是有關 Csmart 系統的整體設計以及部份新發展技術的簡介。

一. 前言

隨著網際網路(Internet)的快速成長，網路上的電子資源，舉凡電子郵件、網路新聞、Web Pages、電子期刊、電子書等成長的相當迅速。為了使這些資源充分利用，網路資訊檢索系統 (Network Information Retrieval System)需求大為增加，包括Lycos, Infoseek, Alta Vista, Excite等在短短兩三年間陸續發展出來[1,2]，對網際網路的使用者而言這些系統無疑地是資訊瀚海的領航員，藉由這些系統的協助，網際網路資源的運用更加發揮。遺憾地，這些系統都是針對英文世界使用者設計的。

考慮中文電子資源的有效利用，適合網際網路資訊服務的高效率中文資訊檢索技術研究是相當迫切的。可惜國內已知的僅有中正大學的GAIS系統[3]、中央大學的CHARVEST系統等少數研究團體從事這方面的努力。這和英語世界動輒成千上萬研究人力物力的投入差距相當大。為此在過去一年中我們將原本針對中文全文檢索設計的尋易(Csmart)系統，由一般檢索核心程式(Searching Engine) [4]擴充具有網路資源發掘與簡易過濾能力，並且增加許多針對中文特性設計的檢索功能以及語音查詢技術，因此發展成新一代智慧型網路中文資訊檢索系統 [5]。

目前Csmart系統除了可以檢索包括電子辭典、建築文獻、佛學書目與摘要、產業技術報告等一般文件資料庫，也可以開始檢索網路即時新聞、BBS論壇、中文Web Pages等網路資源。

為了發展網路中文檢索，我們研究出許多技術包括 Fast Full-text Search [6]、Approximate Text Search、Qasi-Natural Language Query [7]、Speech Retrieval [8]、Word-based Text Search、Relevant Sentence Extraction、Relevance Feedback、Filtering and Subject Dissemination等，且在系統功能與結構設計上也仔細考量中文特性。這中間多數都是全新的嘗試。我們覺得這之中一些經驗可以提出來供大家參考，然而由於Csmart系統整合許多技術，限於篇幅本文嘗試僅就Csmart系統的整體設計及部份新發展技術作一簡介。進一步瞭解Csmart的技術內容可參考相關技術文件或利用Web Browser 連結至 http: // csmart.iis.sinica.edu.tw/。

二. 系統架構

Csmart 的系統架構如圖 1 所示包括 3 個子系統：資源發掘與過濾 (Resource Discovery and Filtering)、資訊檢索(Information Retrieval)與語音介面(Speech Interface)。為了發展網路檢索， Csmart 的研究朝兩方向發展。一個方向是網路資源的充分利用。我們初步設計了 Robot 程式可以在網路上自動發現有收藏價值的資源，藉此我們開始收錄網路中文 Web Pages ， BBS 論壇以及即時新聞並提供檢索，藉由 Robot 技術的發展我們得以開啓網路資源的眞正利用。以計算語言學的角度，我們可以取得源源不斷語料庫，可以統計不同領域的語言差異；以資訊服務的角度，我們有機會整合各個圖書館書目資料庫發展虛擬圖書書目檢索，整合網路新聞，發展虛擬新聞檢索；以語音辨認角度，我們可以建立豐富語言模型，發展語言模型調適技術， Client/Server 方式的語音辨認；最後以資訊檢索的角度，傳統文件分類、資訊抽取、資訊摘要、使用者行為分析等研究，由於有豐富的的資源與使用者，也因此可以較深入發展。

另一方面，我們開始研究語音與自然語言人機介面。由於我們發現許多智慧型檢索功能，如自然語言檢索，由於中文輸入的困難而無法發揮；另外我們發現智慧型的檢索技術多數必須藉由良好的人機互動才能展現，因而我們嘗試發展語音檢索技術與設計人機互動式查詢功能，目前我們以金聲三號為基礎已完成允許使用者以說話的方式詢問 Csmart 系統的語音檢索技術(Speech Retrieval)[8]。我們發現中文單音節特性使語音檢索技術相當接近實用程度。這對發展中文口語交談系統將是很好的開始。

為配合網路資源利用與語音自然語言人機介面的需求，在資訊檢索方面我們發展很多新的技術。舉例，由於網路傳輸不便，檢索結果必須有更佳的精確率以

減少不必要的網路傳輸。為此，我們除了持續加強近似自然語言檢索的檢索精確率與 Ranking 能力，也發展相關文句擷取功能，可使得檢索出的文件有更好的提示以提高檢索結果的可讀性，還有新增 Relevance Feedback 技術，允許從選取的查詢結果自動產生更精確查詢，以及新增主題自動選粹技術，讓使用者自訂新聞主題，而系統隨時主動提供相關新聞。這些技術的開發使得 Csmart 檢索功能的豐富如表 1 的比較說明與國際著名系統相較並不遜色[1,2]。

綜合上述整個Csmart系統運作如圖 1 說明。首先以左下使用者為中心，使用者可以透過網路連接Csmart，以線上(On-line)檢索方式選擇資料庫並輸入查詢以檢索Csmart所收錄的資源，另外也可以預先輸入查詢以建立個人資料檔 (User Profile)並以離線 (Off-line)方式檢索，當Csmart在網路上發現相關資源會以E-Mail方式主動通知使用者讀取。使用者在檢索時可以選擇以打字或者語音輸入。使用者可選擇的檢索功能如表1舉例說明包括邏輯查詢(Boolean Query),近似字串查詢(Approximate Query)，自然語言聯想查詢(Qasi-Natural Language Query)等，另外對檢索出的結果也可要求標示相關程度、顯示相關文句、標記相關字串直接檢索、以Relevance Feedback方式進一步檢索、以語音合成方式唸出檢索文件，以及建立個人資料檔案 (User Profile)。

另外，以資源為中心，如果是授權的特定資源如電子辭典、電子書等可以將全文儲存在 Csmart 主機，加以收錄整理提供線上檢索；若是網路資源如 BBS 、網路新聞、 Web Pages 則須透過 Robot 技術加以收錄、過濾、抽取提供資訊檢索模組建立索引。不論特定資源或網路資源如果須利用語音檢索，則須建立領域語言模型。所以資訊檢索子系統一方面須收錄資源發掘與過濾系統傳送的資源，另一方面又須提供使用者包括線上與離線以及打字與語音輸入等不同檢索。上述功能在目前 Csmart 系統都已提供。圖 2, 3, 4, 5 是有關使用者使用各種檢索方式舉例。其中打字輸入與語音輸入分屬不同介面。

三. 資源發掘與過濾

資源發掘與過濾系統是 Csmart 系統新的努力方向。如圖 6 所示資源發掘與過濾系統事實上包括發掘 (Discovery)、過濾 (Filtering)、抽取 (Extraction)技術。資源發掘主要是利用所謂 Information Spider 或 Robot 技術遊走網路發現值得收藏的資源。基本上網路資源種類很多有 Web Pages 、FTP 文件、News Groups 、BBS 等等，不同類型資源其收錄方式不一樣。以 Web Pages 收錄為例，必須利用 Hyper-link 有效遊走網路，避免收錄重複或品質差的資源、另外對有興趣的資源還必須有效加註(Annotation)。目前 Csmart 在 Web Pages 檢索方面還在實驗階段，因為中文資源相對英文還很少，因此所發展的 Robot 並不須時常上網收集資料以免造成網路擁塞。我們對有興趣的資源的摘要內容如圖 7 所示。

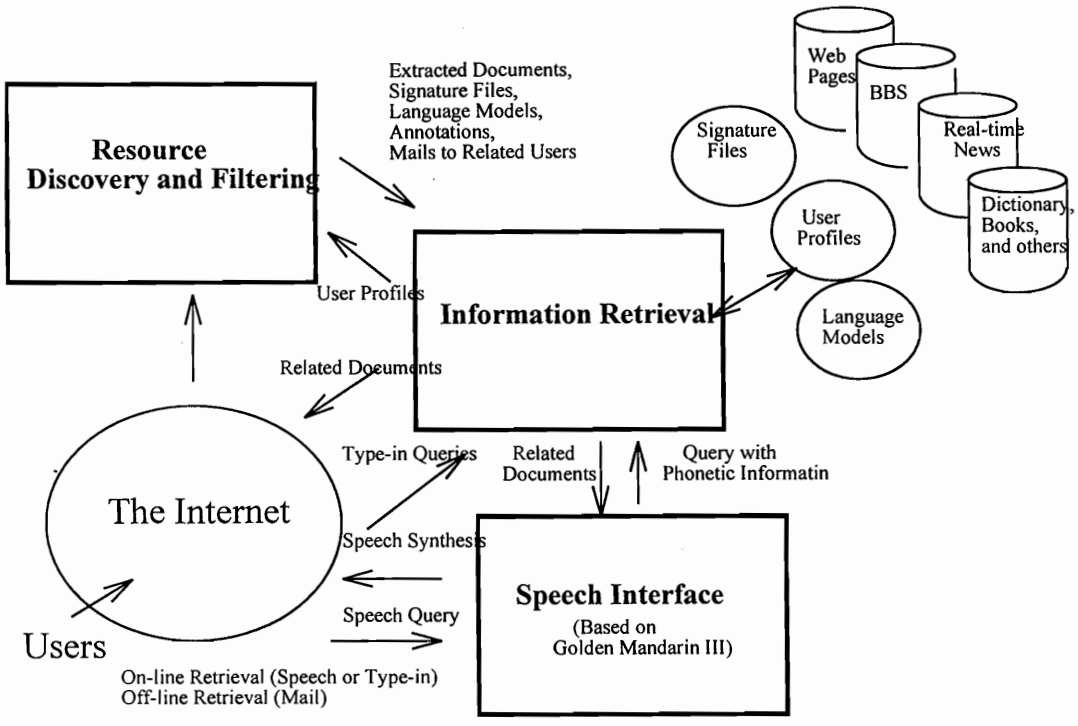


圖 1 Csmart 系統架構圖

至於資訊抽取部份，Csmart 會對收錄文件建立文件特徵(Signature File) [7]與語言模型[8]。前者提供資訊檢索子系統使用，後者是一種改良式馬可夫語言模型(Markov Language Model)提供語音介面提高辨認率以及資資訊系統作進一步特徵抽取分析之用。還有資訊過濾部份，和多數網路檢索系統一樣我們的研究還在起步階段[9]。我們以建立 User Profiles 方式進行資訊過濾，User Profiles 內主要是 User Query ，E-mail 帳號，以及從使用者提供或選定的文件(進一步說明可參見第六節) 抽取文件特徵。以即時新聞為例，使用者可輸入"國泰人壽"的查詢，並選取若干篇文章如有關國泰人壽除權、國泰人壽人事異動等，Csmart 系統嘗試從中抽取特徵，當有相關新聞即可提供使用者讀取。我們很關心資訊過濾技術的研究，包括資源自動分類(Classification) [10]、關鍵詞抽取 (Keyword Extraction)、個人化資訊服務(Personalized Service) [11]等。因為網路資源過多，如無有效分類使用者檢索負擔大；對收錄文件未能抽取出關鍵詞，則以全文檢索查詢成千上萬文件，檢索精確率會很低；還有每個使用者所需資源不同，長遠看檢索系統必須對不同使用者有不同檢索策略。目前 Csmart 在這方面的研究還須不斷加強。

	Functions/Systems	Lycos	Alta Vista	Excite	Csmart	Note
Indexing	Inverted File/Signature	Inverted File	Inverted File	Inverted File	Signature	
Language	English/Chinese	English	English&2- byte	English	Chinese&English*	
Searching Functions	Boolean Query	Min/Max	AND/OR/NOT	AND/OR	AND/OR	
	Approximate Query	Y	Y	Y	Y	工業技術研究院/工研院
	NLQ and Ranking	Y	Y	Y	Y	最新內閣名單/禁止英國牛肉國家
	Field Searching	Y	Y	Y	Y	標題/作者/關鍵詞/全文
	Relevance Feedback			Y	Y	
	Speech-Input Query				Y	
	Word String Match				Y	腦科/電腦科學
	User Profile and Mailing				Y	
Searching Results	Scoring	Y	Y	Y	Y	
	Relevant Sen. Extraction			Y	Y	張德培的排名/張德培擊敗..排名世界..
	Speech Synthesis				Y	
Information Extraction	Spider	Y	Y	Y	Y	
	Annotation of Web Page	Y	Y	Y	Y	
	Auto Keyword Ext.			Y	Under Developing	
	Information Filtering			Y	Under Developing	

表 1 Csmart 技術與功能和國際著名系統比較

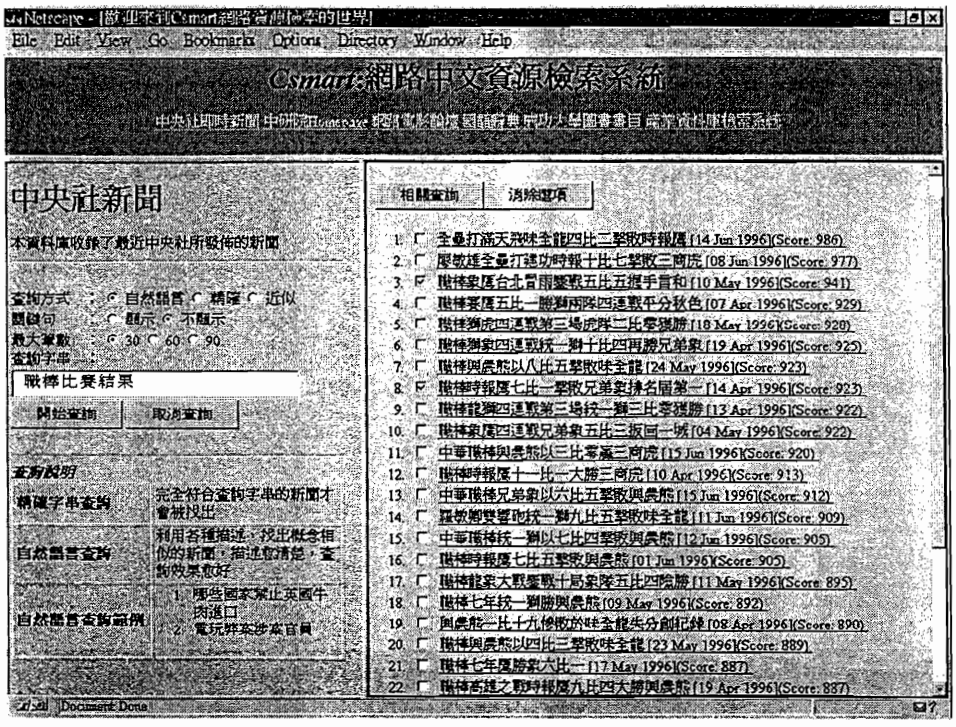


圖 2 Csmart 系統畫面 （以近似自然語言與 Relevance Feedback 檢索即時新聞）

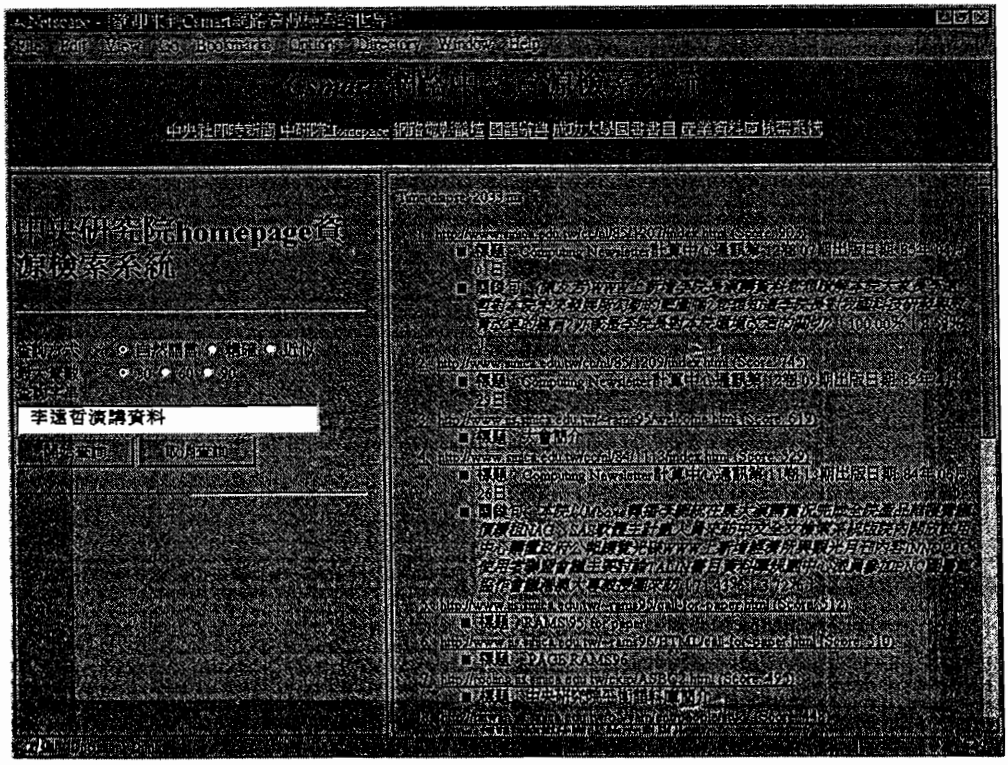


圖 3 Csmart 系統畫面（以近似自然語言檢索 Web Pages 資料庫）

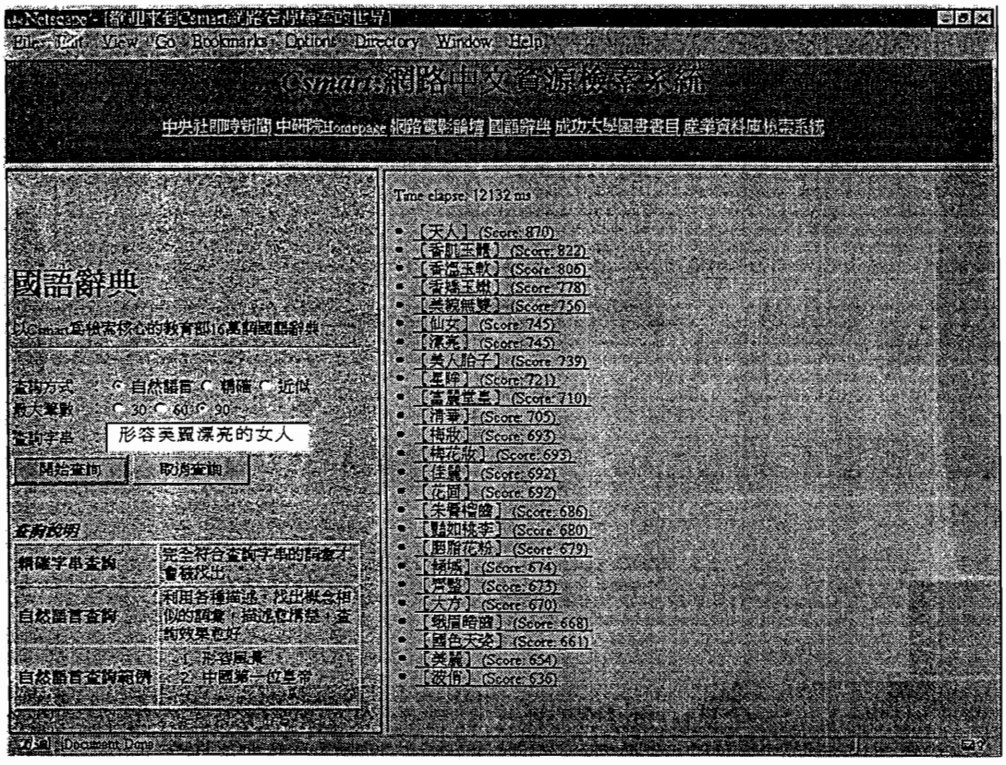


圖 4 Csmart 系統畫面 (以近似自然語言檢索國語辭典)

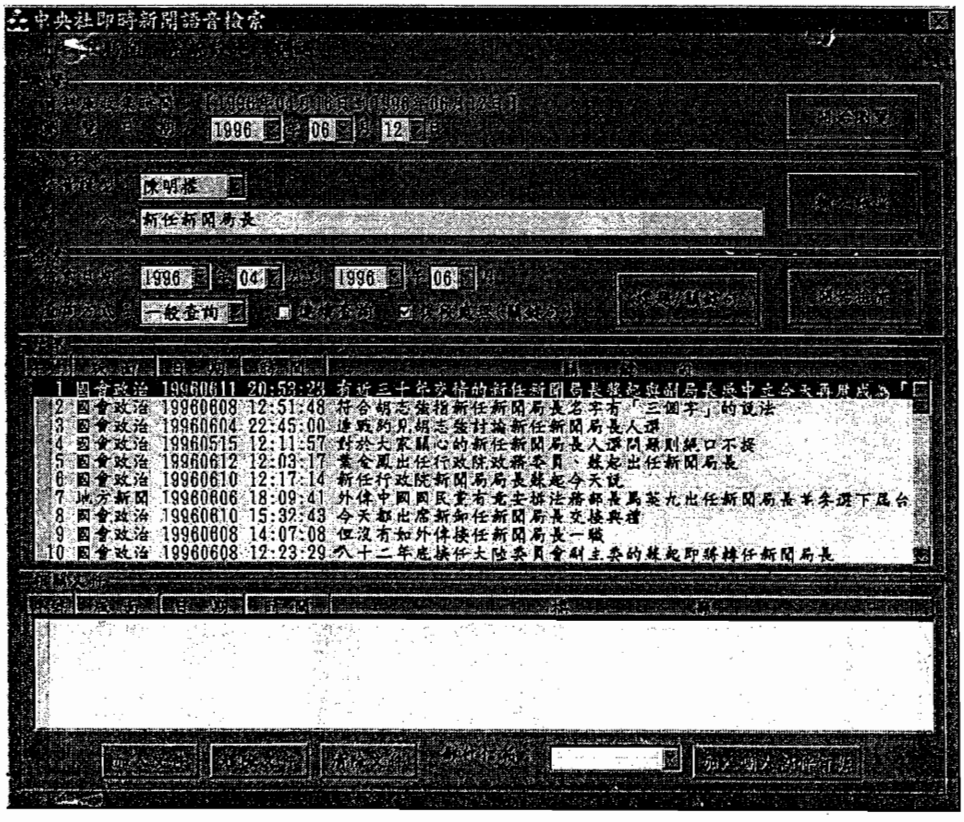


圖 5 Csmart 系統畫面 (以語音輸入查詢檢索即時新聞)

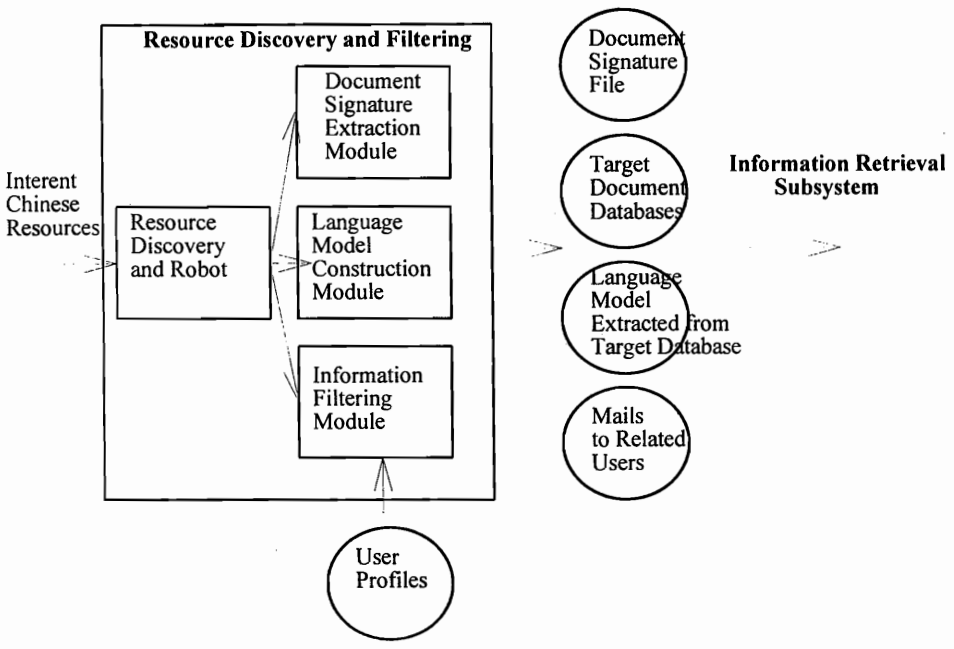


圖 6 資源發掘與過濾子系統

1. the URL
2. the first 200 Characters in the header fields
3. 20 lines or 20 percent of the document
4. number of other pages with links to this document
5. at most 20 lines of hyperlink text from other pages with links to this document
6. extractable keywords (under research)
7. the date the last downloaded
8. the date the last modified
9. document size in bytes

圖 7 Csmart 對收錄的 Web Pages 抽取的資料

四. 資訊檢索 -- 二段式搜尋

資訊檢索子系統是 Csmart 中比較有基礎的部份。Csmart 所使用的檢索技術是以特徵檔(Signature File)為核心的二段式搜尋。根據我們的經驗中文資訊檢索所須克服至少包括檢索詞的收錄與認定，專有名詞與新詞的抽取與斷詞歧異的解決，以及在近似檢索功能的提供 [12]。為此我們發展出以特徵檔(Signature File)為核心的二段式搜尋[7]。

這兩段式搜尋機制主要是將索引比對(Index Matching)與文件比對(Text Matching)分開以克服中文不易使用詞索引以及前述困難。這個方法如圖 8 所示包括第一階段以特徵檔為主的 Fast Search 以及第二階段以文件比對為主的

Detailed Search。由於中文斷詞困難，詞索引建立不易。因此我們覺得索引比對主要作用只是加速過濾多數無關文件，只要索引在比對時有很高的召回率且比對效率高，索引記錄的訊息可以較為模糊。因此我們發展出字層次的特徵檔技術。我們所發展的特徵檔搜尋方法與英文方法接近，不過在設計特徵擷取方式(Signature Extraction Method)時[13]，除了考慮資訊過濾程度，也要考慮給每個Signature Bit有較高語意蘊含以便施行近似檢索。至於文件比對部份，我們發展具備斷詞能力的比對技術，由於第一階段已把多數無關文件過濾掉，因此可利用豐富辭典內容與語言知識，施行精確斷詞以及近似分析，最後將正確文件選出。

在這種搜尋架構下，不論使用者選擇邏輯查詢、近似字串查詢或近似自然語言查詢都必需先產生查詢特徵(Query Signature)。若是邏輯或近似查詢即交由精確或近似比對搜尋程式處理，這包括第一階段先將查詢特徵和所有文件特徵比對，未滿足該查詢特徵的文件將被濾掉。未被濾掉的文件內容在第二階段將會被讀出及與查詢仔細比對，眞正滿足該查詢文件才會檢索出。若使用者是以近似自然語言方式查詢文件，則將交由最佳比對搜尋程序處理。在第一階段該程序會將查詢特徵與所有文件特徵一一比對估算出其查詢與文件之初步相似度(Relevance Value)，相似度足夠高的文件才會在第二階段繼續處理。第二階段基本上是將這些文件內文讀出，並對查詢句子進行關鍵語抽取，以及仔細比對這些關鍵語在相關文件中出現的頻率、位置與加權，以進一步判斷其相似度，最後相當相似的文件才會檢索出。事實上目前愈來愈多的東方語言檢索機制採用特徵檔技術，我們相信兩段式搜尋技術對東方語言檢索非常合適[14-16]。

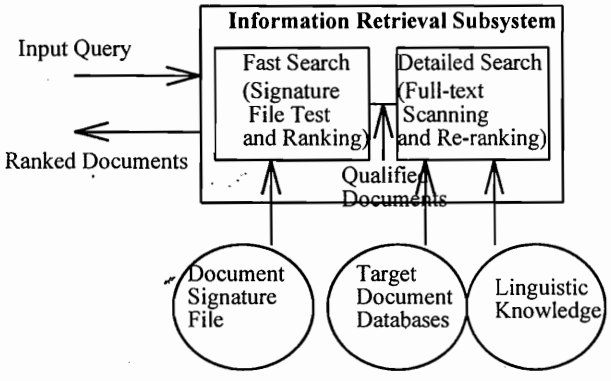


圖 8 Csmart 的二段式搜尋架構

五. 資訊檢索 -- 近似自然語言查詢與應用

Csmart 檢索技術最特殊之處應該是近似自然語言查詢技術的設計。早先我們觀察一般檢索系統成效發現多數使用者並不會使用邏輯查詢而且很多時候中

文查詢無法精確表達，如以圖書書目查詢為例當使用者不確定作者名稱，如"索忍尼新"或"索忍尼津"，書名記憶不精確，如"中國文化基本教材"或"中華文化基本教材"，出版社不能肯定，如"台視文化"或者"臺視文化"，這中間可能有簡稱，外來語言，也可能讀者記憶不清等，查詢這些資料如無高效率近似檢索功能則會形成相當不便。因此我們希望發展出的近似自然語言查詢滿足以下條件：

1. 能允許無限制(Non-constrained)的輸入查詢字串
   也就是查詢字串允許出現非控制字彙(Noncontrolled Vocabulary)
2. 檢索機制能有相當容錯能力
   這包括能檢索近似字串、容忍些許資料登錄錯誤、詞類變化等
3. 檢索出的結果能夠依據與查詢相關程度依序排列，且檢索結果相當合理

我們所發展的近似自然語言查詢方法基本概念包括以下幾個步驟：

1. 以 IDF 為主要參考依據決定查詢字串中每個可能單字與相連雙字的重要性，如"張德培網球排名"中"張"、"德"、"培"、"網"、"球"、"排"、"名"、"張德"、"德培"、"培網"、"網球"、"球排"、"排名"等每個單字與雙字。

2. 利用特徵檔快速檢測所有文件包含前述查詢單字與雙字出現的情形，據此計算出每個文件的初步相似度，相似度過低的文件則加以過濾。

3. 比對辭典，進一步抽取查詢中的可能詞彙，如"張德培"、"網球"、"排名"，將可能有關文件內容讀出，考慮相關文件包含這些詞彙的情形與可能的領域知識如詞彙出現在標題與內文會有不同加權，重新決定文件相似度。

這樣的近似自然語言查詢方法雖然簡單，卻也考慮中文斷詞困難、中文字意豐富、中文雙字鑑別率，中文檢索詞不易精確表達、大量文件檢索效率、領域知識的運用等因素。在實際觀察發現檢索精確率相當的高，目前我們還未作大規模檢測，但是在十多種資料庫超過上千次查詢，我們覺得中文近似自然語言檢索可能比英文有更好的效率，不過這只是臆測仍缺乏完整實驗證實。

自然語言查詢提供全新的查詢概念，只要使用者將可能的檢索概念盡量表達則查詢成效極高。如表 2 所示，形形色色的查詢透過以自然語言聯想查詢可以查出。

六. 資訊檢索 -- 進階檢索技術

1. 近似字串查詢

除了自然語言查詢，前述 Csmart 還發展許多特殊資訊檢索技術以克服中文檢索困難。在近似字串查詢方面，如表 3 所示很多時候中文如無近似字串檢索許多資訊無法加以檢索出來。 Csmart 的近似字串檢索主要是先比對字串頭尾含相同字(少數情形與同音字有關如巴塞隆納、巴塞隆那例外)，如中研院與中央研究院，資策會與資訊工業策進會，這些字串有一定長度關係且短字串內的很高比例的字出現在長字串中且字序一致，最重要的是短字除最後一個字外幾乎都出現在詞的左邊界上。

2. 能解決斷詞歧異的文件比對技術

中文檢索一般都是字串檢索，很少考慮斷詞歧異解決。然而一些查詢如表 4 如無精確斷詞，以詞的觀點則會出現 False Drops 。 Csmart 利用兩段式搜尋，在第二段文件比對時加上斷詞技術，所以多數斷詞歧異能夠排除。

查詢舉例	資料庫
李院長演講資料	中研院 Web Page 資料庫
資訊所圖書館	中研院 Web Page 資料庫
簡立峰電話	中研院 Web Page 資料庫
尋易 Csmart 系統	中研院 Web Page 資料庫
新內閣名單	網路即時新聞資料庫
張德培網球排名	網路即時新聞資料庫
芝加哥公牛對西雅圖超音速	網路即時新聞資料庫
禁止英國牛肉進口的國家	網路即時新聞資料庫
奧斯卡最佳影片	網路即時新聞資料庫
最好看的電影	網路電影論壇資料庫
形容美麗漂亮的女人	電子辭典
形容風景	電子辭典
諾貝爾獎得主	電子辭典
發明電燈的人	電子辭典
最新高溫超導體	產業技術報告資料庫
平行編譯器可行性	產業技術報告資料庫
中國文化	圖書書目資料庫
王姓電腦概論	圖書書目資料庫
世界最高的建築	建築文獻資料庫

表2　Csmart自然語言聯想查詢舉例

種類	舉例
頭銜	李登輝、李總統登輝、李主席登輝
單位簡稱	中研院、中央研究院
單位簡稱	台大、台灣大學
人名拼字	郭李建夫、郭李健夫
相似詞	中國文化、中華文化
譯名	巴塞隆納、巴塞隆那
打字錯誤	電腦概論、電腦概論
相似片語	台北市大安分局、台北市中山分局、台北市分局、
相似片語	高溫超導技術、高溫超導體技術、高溫超導材料技術

表 3　需求近似字串查詢舉例

檢索詞	須精確斷詞的字串
語言學	組合語言學、程式語言學
腦科	電腦科學
中共	其中共有、美中共同參與
陳健康	指陳健康的重要
化學	國際化學術會議、電腦化學理、動物演化學

表 4　須精確斷詞的檢索詞舉例

3. 相關文句擷取技術

網路檢索不只檢索精確率要高，檢索結果的提示須相當可讀性，以協助判斷檢索出的文件是否相關，特別是從檢索出標題不易理解與查詢的關係，或者檢索過長的全文文件時。所以檢索結果除了顯示文件標題、可能出處外，如表 5 所示相關文句擷取與顯示可以提高檢索結果的可讀性。相關文句擷取與近似字串比對觀念上相似，可是要注意查詢中關鍵詞的抽取與分佈，如"張德培"與"排名"。高水準相關文句擷取技術須有簡單文法剖析，目前 Csmart 並未運用這類技術。

查詢	相關文句擷取
張德培的網球排名	文件標題：大滿貫杯網球賽決賽 相關文句：張德培在大滿貫杯比賽擊敗貝克排名晉升第七
公牛隊第四場	文件標題：NBA 比賽公牛初嚐敗績 相關文句：NBA 總冠軍決賽第四場西雅圖超音速隊獲勝
禁止英國牛肉進口的國家	文件標題：世界主要報紙標題 相關文句：巴基斯坦今天宣佈禁止從英國進口牛肉
語音辨認	文件標題：聲控系統技術應用結案報告 相關文句：以期在語音辨認系統實現上仍能達到實驗室水準
虎象比賽最有價值球員	文件標題：職棒虎象纏戰十一局象隊二比一獲勝 相關文句：路易士同時獲選最有價值球員

表 5 關鍵文句擷取技術提高檢索結果的可讀性

4. Relevance Feedback

當使用者輸入的查詢過於簡單或者滿足查詢條件文件過多，舉例使用者原本想查詢馬英九先生新任職務，若以"馬英九"為檢索詞可能在一個月時新聞中可檢索到超過 500 則，其中多數是關於電玩弊案、反毒、工程弊案的報導。通常使用者不知道如何再過濾出相關新聞。這時透過 Relevance Feedback ，使用者只須選擇幾篇眞正相關報導，由系統重新產生更精確查詢即可。 Csmart 的 Relevance Feedback 方法是參考查詢與相關文件特徵(Signature)，利用 Signature Bit 交集與 IDF 加權重新產生查詢特徵(Query Signature)。我們發現 Relevance Feedback 特別適合網路檢索，因為網路資源量多質差，使用者不易一次決定適當的查詢，利用 Relevance Feedback 技術，不論系統與使用者負擔都可減輕。這一點從 Excite 的系統成效也可以觀察出 [2]。

七. 語音介面與語音檢索

為了有效克服中文輸入的困難，以及嘗試設計人機互動式查詢技術以模擬人與電腦對話的效果，我們以金聲 3 號為基礎完成初步具備語音檢索效果的語音介面與語音檢索，如圖 4 所示我們允許使用者以說話的方式詢問系統(Unconstrained Speech Query)。語音檢索並不是將語音辨認與資訊檢索合併即可。語音檢索要考慮語音辨認強健性特別是專有名詞，此外語音檢索更要考慮檢索強健性及容錯能力與檢索速度。在這方面 Csmart 發展許多技術[17, 18]。首先在語音辨認強健方面，我們將語言辨認的語音解碼模組(Linguistic Decoder)獨立出。以欲查詢的資料庫為語料訓練語言模型，這個語言模型是以字為基礎特別加

重專有名詞辨認率，而且具備動態訓練能力以因應資料庫的隨時異動。另外在資訊檢索的強健方面，我們允許語音辨認系統送出最可能字串與候選音節(Syllable)。

在資訊檢索子系統方面，如須以語音檢索的系統我們會把音節訊息加入文件特徵檔中，以提高檢索容錯率。在實際使用發現，由於語音輸入便利，使用者的查詢具有較多的訊息。一般可以說出 8~10 個字。只要語音辨認率維持 70%以上，對檢索系統而言即會有很高正確率。因為 Csmart 的自然語言查詢檢索成效主要是受查詢的資訊豐富與否的影響。以資訊豐富程度來看，70%的語音正確率所輸入的查詢與打字輸入查詢往往是接近的，因為打字輸入的查詢較短。我們實際觀察語音查詢的成效發現語音檢索較為便利、也比較容易表達出眞正自然的查詢、輸入也快得多，其可行性很高。這很可能是中文檢索的一大特色，對發展口語交談系統很有助益 [19]。

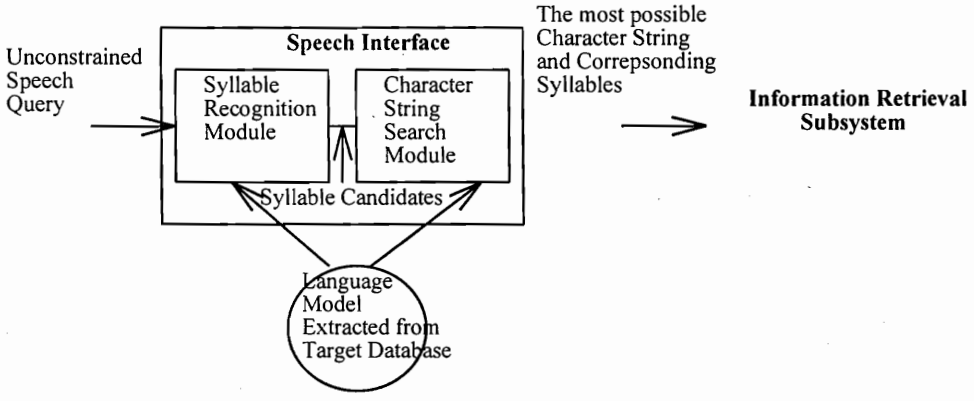


圖 9 語音介面子系統

八. 系統成效與未來研究

前述所有 Csmart 技術都已經開發完成，並且多數經過很長期測試。目前 Csmart 系統除了可以檢索包括電子辭典、建築文獻、佛學書目與摘要、產業技術報告等一般文件資料庫，也可以開始檢索網路即時新聞、BBS 論壇、中文 Web Pages 等網路資源。在搜尋速度方面，快速查詢在五千萬字(100MB)在 PC/486 環境檢索一般字串不到 1 秒。五億字(1GB)在 SPARC10 環境檢索一般字串約 2~5 秒。自然語言檢索則較花時間，在 SPARC10 檢索 2 萬則論文摘要約 1 秒，16 萬則約 4~5 秒。以台灣現有 URL 數目估計短時間不會超過 10 萬而且成長也不大因此檢索速度應無問題。而在索引成效方面，文件索引為可調式(Scalable)，索引大小視需要調整，一般文件所需之索引空間約只佔文件大小的 15~30%左右，另外索引建置時間極短，100MB 文件在 PC/486 環境約只需 4 分鐘，在 SPARC 10 只要 2 分鐘。至於檢索功能與語音檢索效率前述各節也已說明。大致上 Csmart

的資訊檢索技術已符合實用，語音介面與檢索在實驗室成效良好未來可行性高，而資源發現與擷取技術則須持續發展。為此我們已經開始研究關鍵詞抽取(Keyword Extraction)技術以因應大量網路資源所需，藉此希望發展出資訊分類與過濾技術，使我們有能力判斷出有興趣收集的資源，並且進而擷取出重要訊息加以建立索引，與發展個人化資訊服務。

參考資料

1. G. Venditto, Searching Engine Showdown, Internet World, May 1996.

2. M. Courtois, et al., Cool Tools for Searching the Web: A Performance Evaluation, Online, Nov. 1995.

3. S. Wu, Gais Home Page, http://gais.cs.ccu.edu.tw/

4. Lee-Feng Chien (95b), 尋易 (Csmart) -- A High-Performance Chinese Document Retrieval System, The 1995 International Conference of Computer Processing for Oriental Languages, *ICCPOL'95*.

5. Lee-Feng Chien , Hsiao-Tiech Pu, Ming-Chan Chen, Hung-Ming Chen and Ming-Jer Lee, Natural Language Information Retrieval with Speech Recognition Techniques for Network Chinese Resources Discovery, May 1996 International Workshop on Information Retrieval with Oriental Languages

6. Lee-Feng Chien , A Model-Based Signature File Approach for Full-text Retrieval of Chinese Document Databases. To appear on Computer Processing of Chinese and Oriental Languages, 1996.

7. Lee-Feng Chien , Fast and Quasi-Natural Language Search for Gigabytes of Chinese Texts, ACM SIGIR'95, 1995.

8. Sung-Chien lin, Lee-Feng Chien, Keh-Jiann Chen, Lin-Shan Lee, An Efficient Voice Retrieval System for Very-Large-Vocabulary Chinese Textual Databases with a Clustered Language Model, May 1996 (ICASSP'96).

9. Belkin, Nicholas J., et al, "Information Filtering and Information Retrieval: Two Sides of the Same Coin?", Communications of the ACM, Vol. 35, No 12, Dec. 1992).

10. D. Lewis, Evaluating and Optimizing Autonumous Text Classification Systems, SIGIR'95.

11. Foltz, Peter W., et al, "Personalized Information Delivery: An Analysis of

Information Filtering Methods", Communication of the ACM, Vol. 35, No. 12, Dec., 1992.

12. Lee-Feng Chien and Hsiao-Tiech Pu, Important Issues on Chinese Full-text Information Retrieval, Invited and to be submmited for Computational Linguisitcs and Chinese Language Processing.

13. Faloutsos, C., "Access Methods for Text", ACM Computing Surveys, March 1985, 49-74.

14. Tyne Liang, Suh-yin Lee and Wei-Pang Yang, Optimal Weight Assignment for a Chinese Signature File, Information Processing and Management, Vol 32, No. 2, pp. 227-237, 1996.

15. Lee, Ahn and Shin, An Effective Indexing Method for Korean Text Retrieval, International Workshop on Information Retrieval with Oriental Languages, Korea, 1996.

16. Y. Ogawa, A New Character-based Indexing Organization Using Frequency Data for Japanese Documents, SIGIR'95.

17. Sung-Chien Lin, Lee-Feng Chien and Lin-shan Lee, A Syllable-based very-Large-Vocabulary Voice Retrieval System for Chinese Databases with Textual Attributes, Proceedings of the 4th European Conference on Speech Communication and Technology, Sept. 1995.

18. Sung-chien Lin, Lee-Feng Chien and Lin-shan Lee, Unconstrained Speech Retrieval for Chinese Docment Databases with Very Large Vocabulary, Proceedings of the 4th European Conference on Speech Communication and Technology, Sept. 1995.

19. Yen-Ju Yang, Lee-Feng Chien and Lin-Shan Lee, An Efficient linguistic Decoding System with Adaptive Learning for Mandarin Speech Recogntion, accepted by CPCOL.

CORRECTING CHINESE REPETITION REPAIRS IN SPONTANEOUS SPEECH

Yue-Shi Lee and Hsin-Hsi Chen

Department of Computer Science and Information Engineering

National Taiwan University

Taipei, Taiwan, R.O.C.

e-mail. hh_chen@csie.ntu.edu.tw

Abstract

Disfluencies involving speech repairs pose serious problems for spoken language processing systems. However, which cues in speech signals may facilitate Chinese repair processing is not known. This paper concerns the acoustic and prosodic analysis for correcting Chinese repetition repairs in spontaneous speech. A large spoken corpus is examined in this study. The experimental results show that our method can achieve the precision rate of 93.87% and the recall rate of 90.65%, without using the lexical information.

1 Introduction

Most of the previous acoustic analyses of speech examined data from speakers who carefully pronounce their speech. Natural spontaneous or conversational speech differs from careful or reading speech in several ways. The most obvious difference is the use of speech repairs. In spontaneous speech, people often start talking and then think along the way. This causes spontaneous speech to have a variable speaking rate, and such speech often exhibits speech repairs, which are interrupts in the flow of speech,

where the speaker reiterates a portion of the immediately preceding speech, with or without a change.

Heeman and Allen [1] describe that 25% of turns contain at least one speech repair in their corpus. In our study, 17% of turns contain at least one speech repair. Thus, the speech repairs cannot be negligible and have influences to a certain extent. On the one hand, correctly recognizing speech repairs can help automatic speech recognizers to avoid textual errors. In most of the current speech recognition systems, words repeated in a speech repair are simply fed as word hypotheses to the language model of the recognizer. This may cause difficulties in having a proper recognition since the language model is usually trained on fluent text only.

On the other hand, even if all the words in a disfluent segment are correctly recognized, failure to detect a disfluency may lead to interpretation errors during subsequent processing. (1) is an example[1].

(1) 952 ..她應該%-- {A2,2-1,1}
953 ...不應該升工程--- {R1,2-1,2}
954 ..工程師那麼快的.\

There are an addition repair (A) and a repetition repair (R) in (1)[2]. For the first repair, the speaker changes his intention from "應該" to "不應該". If this kind of repair cannot be detected, the system will misunderstand the intention of the speaker.

Recently, text-first approach [1, 4, 5] and speech-first approach [6, 7] have been proposed to touch on repairs in English. The text-first approach assumes the speech

---

[1] The transcription system proposed by Bois, *et al.* [2] is used to transcribe the spoken data. The two symbols ... and .. denote a unfilled pause (silence) is medium and short, respectively. The symbol % denotes the glottal stop. The detailed transcribing conventions are shown in Appendix A. Relevant characters in examples are in **boldface** and underline.

[2] The annotating scheme of repair can refer to Chou [3].

recognizer could provide a correct transcription. That is, it tries to detect and correct speech repairs automatically using text alone. Hindle [4] adds rules to a deterministic parser to tackle the problem of correcting speech repairs. His parsing strategy depends on the successful disambiguation of the syntactic categories. Although syntactic categories can be determined well by local context, Hindle admits that speech repair disrupts the local context. Bear, *et al.* [5] firstly try to parse the input sentence and then invoke a repair processing when the parsing fails. For repair processing, a simple pattern matcher finds the candidates based on the lexical cues at the first stage. Then the syntactic and semantic processing filters out the impossible candidates. Heeman and Allen [1] present an algorithm that detects and corrects modification and abridged repairs. The algorithm uses some repair patterns to capture potential repairs. These patterns are built based on the identification of word fragments, editing terms[3], and word correspondences between the repaired segment and the repairing segment[4]. The resulting potential repairs are then passed to a statistical filter that judges the proposal as either fluent speech or an actual repair.

The speech-first approach tries to identify speech repairs using acoustic and prosodic cues. Nakatani and Hirschberg [6, 7] investigate the detection of the interruption point of speech repairs based on this line. The cues that they found are the occurrence of a filled pause, the presence of a word fragment, the energy peaks in each word and other features such as accent. However, they did not address the problem of correcting the speech repairs. In other words, they do not determine

---

[3] The editing terms can either be filled pauses (e.g., um, un, er) or cue phrases (e.g., I mean, I guess, well).

[4] A repair is composed of a repaired segment and a repairing segment which immediately follows the repaired segment. A repaired segment denotes the portion of the utterance which is being repaired, and a repairing segment denotes the portion which is accomplishing the repair [8]. That is, the repaired segment is replaced by the repairing segment.

which words are undesired.

These approaches cannot be adopted to deal with Chinese speech repairs for the following reasons. First, a Chinese sentence is composed of a string of characters without any word boundaries. In other words, it is necessary to segment Chinese sentence before tagging and parsing [9, 10]. Repairs make segmentation and text-first approach more difficult. Second, Chinese repairs may not always have an editing terms between a repaired segment and a repairing segment. In other words, editing terms do not have much effect in Chinese repair processing. Third, duplicate constructions (e.g., 幫幫忙, 陸陸續續) in Chinese utterances are used very often, but they do not always initiate a repair. That is, a simple pattern matching mechanism cannot be workable. Forth, the Chinese speech repairs may be initiated at various syntactic environment [11], e.g., before the subject, during the subject, after the subject and before the verb, during the verb, during a direct object, during a prepositional phrase, during subordination, and so on. The variety makes the identification of Chinese speech repairs more troublesome.

Because the identification of repairs in Chinese may not be deferred to the latter modules of the spoken language processing systems, this paper identifies several cues based on acoustic and prosodic analysis of repairs in a spoken corpus and proposes methods for exploiting these cues to correct the repetition repairs. Section 2 defines four major types of speech repairs. Section 3 introduces the spoken corpus. Sections 4 and 5 describe the acoustic and prosodic analysis of repairs. Section 6 is the concluding remarks.

2 Types of Chinese Speech Repairs

Heeman and Allen [1] divide English speech repairs into three types: fresh starts, modifications and abridged. For Chinese speech repairs, Chui [11] classify them into eleven patterns. In this section, we map these eleven patterns into four major types according to their surface forms.

Let $c_1 \ldots c_n\ c_{n+1} \ldots c_{n+m}$ be a sequence of Chinese characters. They may form an utterance or two consecutive utterances. The four major types of speech repairs are described as follows:

(a) Repetition Repair

There exists an i-character string, such that $c_{n-i+1} \ldots c_n$ (the repaired segment) is equal to $c_{n+1} \ldots c_{n+i}$ (the repairing segment). The repetition can range from a portion of a word up to several words. After being repaired, the utterances become $c_1 \ldots c_{n-i}\ c_{n+1} \ldots c_{n+m}$. (2) and (3) show two examples.

(2) 384 ..我--- {R1,1-1,1}
385 ..我是^知道我有這個毛病啊=,-

(3) 667 ..全國的]一起申-- {R1,1-1,1}
668 ..申請的=.\

The repetition repair occurs between utterances 384 (667) and 385 (668). The word "我" and the character "申" are repeated in (2) and (3), respectively. The character "申" is a portion of the word "申請". After being repaired, the utterances become "我是知道我有這個毛病啊" and "全國的一起申請的", respectively.

(b) Addition Repair

There are two types of addition repairs.

(i) There exist a j-character string and a k-character string, such that $c_{n-j+1} \ldots c_n$ (the repaired segment) is equal to $c_{n+k+1} \ldots c_{n+k+j}$. The character string, $c_{n+1} \ldots c_{n+k}\ c_{n+k+1} \ldots c_{n+k+j}$, forms the repairing segment in this case. After being repaired, the utterances become $c_1 \ldots c_{n-j}\ c_{n+1} \ldots c_{n+m}$. That is, $c_{n+1} \ldots c_{n+k}$ are added. (4) shows an example. The speaker's intention is "你們自已應該要".

(4) 313 ...你們自已要,- {A1,1-1,2}
314 ^應該要,-

(ii) There exist an i-character string, a j-character string and a k-character string, such that $c_{n-j-i+1} \ldots c_{n-j}$ is equal to $c_{n+1} \ldots c_{n+i}$ and $c_{n-j+1} \ldots c_n$ is equal to $c_{n+i+k+1} \ldots c_{n+i+k+j}$. The character string, $c_{n-j-i+1} \ldots c_{n-j}\ c_{n-j+1} \ldots c_n$, forms the repaired segment and the character string, $c_{n+1} \ldots c_{n+i}\ c_{n+i+1} \ldots c_{n+i+k}\ c_{n+i+k+1} \ldots c_{n+i+k+j}$, forms the repairing segment in this case. After being repaired, the utterances become $c_1 \ldots c_{n-j-i}\ c_{n+1} \ldots c_{n+m}$. (5) shows an example. The desired utterance is "他今天才說".

(5) 1953 Z:[<F^他說,- {A1,1-2,4}
1954 ..他今天]才說,-

(c) Replacement Repair

There are five types of replacement repairs.

(i) There exist an i-character string, an h-character string and a k-character string, such that $c_{n-h-i+1} \ldots c_{n-h}$ is equal to $c_{n+1} \ldots c_{n+i}$. The character string, $c_{n-h-i+1} \ldots c_{n-h}\ c_{n-h+1} \ldots c_n$, forms the repaired segment and the

character string, $c_{n+1} \ldots c_{n+i}\ c_{n+i+1} \ldots c_{n+i+k}$, forms the repairing segment in this case. After being repaired, the utterances become $c_1 \ldots c_{n-h-i}\ c_{n+1} \ldots c_{n+m}$. (6) shows an example. The final utterance is "你一定沒有講出來吧".

(6) 733 Z:[你一定沒有]^唸%-- {P1,1-3,3}
734 ...沒有講出來吧=,-

(ii) There exist a j-character string, an h-character string and a k-character string, such that $c_{n-j+1} \ldots c_n$ is equal to $c_{n+k+1} \ldots c_{n+k+j}$. The character string, $c_{n-j-h+1} \ldots c_{n-j}\ c_{n-j+1} \ldots c_n$, forms the repaired segment and the character string, $c_{n+1} \ldots c_{n+k}\ c_{n+k+1} \ldots c_{n+k+j}$, forms the repairing segment in this case. After being repaired, the utterances become $c_1 \ldots c_{n-j-h}\ c_{n+1} \ldots c_{n+m}$. (7) shows an example. The word "很多" is replaced by "非常多" and the utterances become "非常多人過來我們這邊買東西".

(7) 2612 Y:..很多,- {P2,2-1,2}
2613 ..^非常[多人過來我們這邊]買東西,-

(iii) There exist an i-character string, a j-character string, an h-character string and a k-character string, such that $c_{n-j-h-i+1} \ldots c_{n-j-h}$ is equal to $c_{n+1} \ldots c_{n+i}$ and $c_{n-j+1} \ldots c_n$ is equal to $c_{n+i+k+1} \ldots c_{n+i+k+j}$. The character string, $c_{n-j-h-i+1} \ldots c_{n-j-h}\ c_{n-j-h+1} \ldots c_{n-j}\ c_{n-j+1} \ldots c_n$, forms the repaired segment and the character string, $c_{n+1} \ldots c_{n+i}\ c_{n+i+1} \ldots c_{n+i+k}\ c_{n+i+k+1} \ldots c_{n+i+k+j}$, forms the repairing segment in this case. After being repaired, the utterances become $c_1 \ldots c_{n-j-h-i}\ c_{n+1} \ldots c_{n+m}$. (8) shows an example. The word "差一滴" is substituted by "差幾滴".

(8) 528 ...(1.1)eh 我這還差一滴-- {P1,3-1,3}
529 ...eh 差幾滴 ei=.\

(iv) There exist an i-character string, a j-character string and an h-character string, such that $c_{n-j-h-i+1} \ldots c_{n-j-h}$ is equal to $c_{n+1} \ldots c_{n+i}$ and $c_{n-j+1} \ldots c_n$ is equal to $c_{n+i+1} \ldots c_{n+i+j}$. The character string, $c_{n-j-h-i+1} \ldots c_{n-j-h}\ c_{n-j-h+1} \ldots c_{n-j}\ c_{n-j+1} \ldots c_n$, forms the repaired segment and the character string, $c_{n+1} \ldots c_{n+i}\ c_{n+i+1} \ldots c_{n+i+j}$, forms the repairing segment in this case. After being repaired, the utterances become $c_1 \ldots c_{n-j-h-i}\ c_{n+1} \ldots c_{n+m}$. (9) shows an example.

(9) 406 [我就是--- {P1,3-1,2}
407 ..我是-]--

(v) Different from the above replacement repairs, the repaired segment and the repairing segment in this type do not match any characters. (10) shows an example.

(10) 659 ..他們-- {P1,2-1,1}
660 ...她唸的,-

(d) Abandon Repair

The original utterance is discarded and a new utterance is initiated. (11) shows an example.

(11) 54 H:...(.9)那就不要一堆-- {B1,6-1,1}
55 ..大家坐在一起幹嘛啊?/

After being repaired, the utterances become "大家坐在一起幹嘛啊".

3 Spoken Corpus

The spoken corpus analyzed in this paper consists of two commonplace, everyday conversations among friends. Each is about forty-minute long. There are four and five speakers in these two conversations, respectively. It is originated from Professor Kaiwai Chui at National Chengchi University [11]. In total, this corpus contains 5395 utterances, 22409 words and 2602 turns. There are totally 440 self-repairs[5]. On the average, 17% of turns contain at least one repair. Tables 1 and 2 list the frequency distribution of each type of repairs in the two conversations.

Table 1. Frequency Distribution of Repairs in Conversation 1

Speaker	Repetition	Addition	Replacement	Abandon
L	23	8	3	1
H	54	12	9	4
Z	35	3	3	2
O	10	0	1	0
Total	122	23	16	7

Table 2. Frequency Distribution of Repairs in Conversation 2

Speaker	Repetition	Addition	Replacement	Abandon
L	39	5	4	4
W	53	9	6	8
Y	61	9	8	12
Z	44	2	4	1
J	2	1	0	0
Total	199	26	22	25

[5] The speech repairs discussed in this paper are all self-repairs. That is, only the repairs accomplished by the same speaker are considered. This is because this kind of repairs is the most common form of repairs. Nevertheless, the present study includes repairs placed across different turns.

In the above statistics, we find that the repetition repairs form the majority (72.62% in conversation 1 and 73.16% in conversation 2) of the repairs. Addition (Replacement) repairs have 13.69% (9.52%) and 9.56% (8.09%) in conversations 1 and 2, respectively. The rest (4.17% in conversation 1 and 9.19% in conversation 2) are the most complex type of repairs, i.e., Abandon. Because this paper corrects repairs based on acoustic and prosodic cues, the Chinese characters in the spoken corpus are converted into the corresponding syllables manually.

4 Basic Analysis Method

4.1 Simple Pattern Matching Mechanism

Because the repetition repairs form the majority, we focus on the repetition repairs in this paper. Although the repetition repairs have the simple surface form, correcting such a kind of speech repairs is not trivial. That is, a simple pattern matching mechanism cannot work perfectly. Table 3 explains this phenomenon. A repair is proposed when a string of syllables repeats within an utterance or between two consecutive utterances.

Table 3. The Experimental Results Using Simple Pattern Matching Mechanism

Conversation	Total Repairs	Proposed	Correct
1	122	243	118
2	199	412	196
Total	321	655	314

Columns 2, 3 and 4 denote the total repetition repairs, the number of repairs proposed by the simple pattern matcher and the number of correct proposed repairs, respectively.

For example, 243 repairs are proposed by the simple pattern matcher in conversation 1, but only 118 of them are correct. That is, there are 125 false alarms. Since there are 122 repetition repairs in conversation 1, 4 repetition repairs are not captured. Some of them are listed below.

(12) 794 L:...那假如<L2 ^**Mac** L2>-- {R1,1-1,1}
795 ...<L2 **Mac** L2>的<L2 set up L2>不起來的.\

(13) 1541 ...連<L2 **Genni**--- {R1,1-1,1}
1542 ...**Genni**fer L2>三個=.\

(14) 1835 ...它<L2 **supermar**--- {R1,1-1,1}
1836 ...**super]mar**ket L2>也是很多那種,-

Because only Chinese speech repairs are considered, English repairs are lost. Although this technique can achieve recall rate of 97.82%, it has a relatively low precision rate, i.e., 47.94%.

Since the simple pattern matching mechanism cannot solve this problem properly, two basic analyses are firstly considered in the next two subsections: the length of the repeated syllable string and the number of inter-utterances.

4.2 The Length of the Repeated Syllable String

One of the most important things that we want to know is "how many syllables are repeated in the repetition repairs?"

Table 4. The Distribution of the Length of the Repeated Syllable String

Conv.\Length	1	2	3	4
1	71	40	6	1
2	107	72	15	2
Total	178	112	21	3

Table 4 lists the distribution of the length of the repeated syllable string in the repair. The length ranges from 1 to 4. That is, when a string of syllables repeats and the length of this string is greater than 4, we do not regard it as a repetition repair.

4.3 The Number of Inter-Utterances

Basically, most of the repetition repairs occur within an utterance or between two consecutive utterances of one speaker without interrupting by other speakers. However, if enough utterances pronounced by other speakers are inserted between two utterances of one speaker, the repetition repairs usually do not occur between them. (15) is an example.

(15) 2445 L:...哦=,-
2446 .那不[是%]-\
2447 Y:[三]個多<@ 月 @>.\
2448 ...[[三個半月]].\
2449 J:[[好過份啊]],-
2450 Z:(0)對啊=?/
2451 ...我覺得?/
2452 Y:...我^這次我十月=?/
2453 ...十五號啊,-
2454 ..<P 十五號十六[號 P>.\
2455 Z:[huh huh],-
2456 J:...你那個時候訂了.\
2457 就沒有貨了=是不是.\
2458 ...還是說%,-
2459 Y:...對啊.\
2460 他的意思[是%]--
2461 L:[不是].\

Although there is a matched string "不是" between utterances 2446 and 2461 for speaker L, it is neither a repaired segment nor a repairing segment. This is because 14 utterances interrupt the flow of thought of the speaker L. After analyzing the spoken

corpus, some statistics are shown below.

(1) Total 13.69% of repetition repairs occur in the same utterance.

(2) Total 71.66% of repetition repairs occur between two consecutive utterances without interrupting by other speakers.

(3) Only 0.32% of repetition repairs occur across more than 3 utterances issued by other speakers.

According to these analyses, when more than 3 utterances pronounced by other speakers interrupt the speech of a speaker, we do not check whether there is a repetition repair or not.

5 Advanced Analysis Method

5.1 Unfilled Pause (...)

Observing the spoken corpus, we find that there is a significant unfilled pause (silence) between a repaired segment and a repairing segment for repetition repairs[6], whereas actual or intended repeated characters (syllables) usually do not have any unfilled pauses between them. Some typical examples are shown below.

(16) 505 ...(.8)不是說[你%-- {R1,1-1,1}
506 ..你..感覺到已經],-

(17) 606 Z:[那我--- {R1,2-1,2}
607 Z:..那我還=謝謝]]你們<F 啊 F>]?/
608 H:...(.7)^當然你要謝謝我們啊=,-

In the above examples, actual repeated characters (syllables) "謝謝" do not have any unfilled pause, whereas there is a unfilled pause between utterances 505 (606) and 506

[6] Because the filled pauses such as um, un and er do not occur frequently in the spoken corpus, the effects of filled pauses are not demonstrated in this paper.

(607). Based on the basic analysis and the unfilled pause information, the experimental results for two conversations are listed below.

Table 5. The Experimental Results Using Unfilled Pause

Conversation	Total Repairs	Proposed	Correct
1	122	99	86
2	199	191	158
Total	321	290	244

The experimental results show that the precision rate is increased to 84.14%, and the recall rate is decreased to 76.01%.

5.2 Glottal Stop (%)

Glottal stop has the similar functions to unfilled pause. That is, a glottal stop may occur between the repaired segment and the repairing segment for the repetition repairs, whereas actual repeated characters usually do not have such a marker between them. (16) is an example. Based on the basic analysis and the glottal stop information, the experimental results for two conversations are listed below.

Table 6. The Experimental Results Using Glottal Stop

Conversation	Total Repairs	Proposed	Correct
1	122	31	31
2	199	85	82
Total	321	116	113

From Table 6, we find that glottal stop is a more reliable cue than unfilled pause, but it does not occur as frequently as unfilled pause. These points are verified by the high

precision rate (97.41%) and the low recall rate (35.20%). When the basic analysis, the unfilled pause and glottal stop information are applied together, the experimental results for two conversations are listed in Table 7.

Table 7. The Experimental Results Using Unfilled Pause and Glottal Stop

Conversation	Total Repairs	Proposed	Correct
1	122	110	97
2	199	204	169
Total	321	314	266

Both the precision rate (84.71%) and the recall rate (82.87%) are all better than those in the former models.

5.3 Two Consecutive Equal Utterances

If two consecutive utterances are equal, repetition repairs usually do not occur within and between them when the length of the utterance is long enough. (18) is an example. The matched string is "是有這種" which denotes an emphasis. Thus, utterances 894 and 895 do not form a repair.

(18) 892 L:...(1.4)啊=,-
893 ..[對對對.\
894 ...是有這種,-
895 ..是有這種,-
896 對對 .\

Based on our spoken corpus, when the equal utterance length is greater than 2, no repetition repairs occur. This cue can eliminate some implausible repairs, so that the precision rate can be increased.

5.4 Cue Patterns

To increase the precision rate, another method is proposed. We collect the wrong proposed repairs that satisfy the criteria of basic analysis, unfilled pause and glottal stop. For the generalization, only the first syllable of each wrong proposed repair is concerned. The syllables whose frequency is larger than 1 is regarded as the type I cue pattern. By this way, six type I cue patterns, i.e., ㄧㄡˇ, ㄕˋ, ㄏㄠˇ, ㄛˊ, ㄚ• and ㄉㄨㄟˋ, are selected. (19) and (20) are two examples. Although the matched strings in these two examples satisfy the criteria of basic analysis, unfilled pause and glottal stop, they are not repairs.

(19) 69 J:[這什麼意思啊]?/
70 Y:...他是說他喜歡^當那隻,-
71 他說^錯了.\
72 J:...(1.3)^啊?/((WHAT?))

(20) 3362 Z:...(1.6)哦,\
3363 W:...其他的事情你們要...處理.\
3364 Y:(0)@@@=
3365 Z:...(1.3)哦真的啊,-

Because the first two patterns, ㄧㄡˇ and ㄕˋ, do not have the actual benefits in experiments, they are discarded. This is because only the negative examples (wrong proposed repairs) are used to generate this kind of patterns. Thus, a repair is proposed when a string of syllables repeats, satisfies criteria of basic analysis, unfilled pause and glottal stop, and the first syllable of the string does not belong to one of the four type I cue patterns.

Similarly, another kind of patterns is considered to increase the recall rate. Those repetition repairs that do not satisfy the criteria of unfilled pause and glottal stop are collected. The similar procedures for type I cue patterns are adopted to generate

type II cue patterns. Finally, four such patterns are selected, i.e., ㄕㄜˊ, ㄋㄧˇ, ㄨㄛˇ and ㄋㄚˋ. (21) and (22) are two examples. Although the matched strings in these two examples do not satisfy the criteria of unfilled pause and glottal stop, they are all repairs.

(21) 464 O:[[你那-]]-- {R1,2-1,2}
465 [[你那-]]-- {R1,2-1,2}
466 L:[[那^吐]]出來=.\
467 O:(0)你那時--- {R1,3-1,3}
468 你那時候已經,-

(22) 1251 Z:[那就]-- {R1,2-1,2}
1252 (0)那就表示他不想你[[^買=]].\

Based on type II patterns, some additional repairs can be proposed when a string of syllables repeats, it does not satisfy the criteria of unfilled pause and glottal stop, but the first syllable of the string belongs to one of the four type II cue patterns.

Based on the method described in Section 5.2, the equal utterances information and the cue patterns, the experimental results are listed below.

Table 8. The Experimental Results

Conversation	Total Repairs	Proposed	Correct
1	122	120	111
2	199	190	180
Total	321	310	291

The experimental results show that the precision rate of 93.87% and the recall rate of 90.65% can be achieved. Besides, we also test another spoken corpus. The corpus has 504 utterances. It is about ten-minute long. It is originated from Professor Shuanfan Huang at National Taiwan University [12]. There are four speakers and

totally 19 repetition repairs in this corpus. Table 9 lists the experimental results of the simple pattern matching and the method used in this section.

Table 9. The Experimental Results

Method	Total Repairs	Proposed	Correct
Pattern Matching	19	45	19
Our Method	19	21	18

Because the glottal stop is not annotated in this corpus, this cue is not used in this experiment. Apparently, our method (precision: 85.71%, recall: 94.74%) is better than simple pattern matching (precision: 42.22%, recall: 100%).

6 Concluding Remarks

Any spoken language systems will not perform well without treating speech repairs. Correcting speech repairs make more reliable environments for the subsequent processing. This paper identifies several cues based on acoustic and prosodic analysis of repairs in a large spoken corpus and proposes methods for exploiting these cues to correct the repetition repairs. The experimental results show that our method can achieve the precision rate of 93.87% and the recall rate of 90.65%, without using lexical information. O'Shaughnessy [13] claims that most speech repairs do not have lengthening prior to the hesitation pause. If this cue is used in our model, it can slightly increase the precision rate (95.37%), but the recall rate (76.95%) is greatly decreased.

Although our method can perform well in repetition repairs, other kinds of repairs such as addition, replacement and abandon repairs are not addressed in this paper. They have more complex surface forms and should be investigated further.

Acknowledgments

We are grateful to Professor Shuanfan Huang and Professor Kawai Chui for their kindly providing their spoken corpora to us.

References

[1] P. Heeman and J. Allen (1994) "Detecting and Correcting Speech Repairs," *Proceedings of ACL*, 1994, pp. 295-302.

[2] D. Bois, *et al.* (1992) "Discourse Transcription," *Santa Barbara Papers in Linguistics*, Vol. 4, 1992.

[3] M.L. Chou (1996) *Detecting and Correcting Chinese Speech Repairs*, Master Thesis, Department of Computer Science and Information Engineering, National Taiwan University, 1996.

[4] D. Hindle (1983) "Deterministic Parsing of Syntactic Nonfluencies," *Proceedings of ACL*, 1983, pp. 123-128.

[5] J. Bear, J. Dowding and E. Shriberg (1992) "Integrating Multiple Knowledge Sources for Detection and Correction of Repairs in Human-Computer Dialog," *Proceedings of ACL*, 1992, pp. 56-63.

[6] C. Nakatani and J. Hirschberg (1993a) "A Speech-First Model for Repair Detection and Correction," *Proceedings of EUROSPEECH*, 1993a, pp. 1173-1176.

[7] C. Nakatani and J. Hirschberg (1993b) "A Speech-First Model for Repair Detection and Correction," *Proceedings of ACL*, 1993b, pp. 46-53.

[8] B.A. Fox and R. Jasperson (forthcoming) "A Syntactic Exploration of Repair in English Conversation," *Descriptive and Theoretical Models in the Alternative*

Linguistics, Davis P. (Ed.), forthcoming.

[9] K.J. Chen and S.H. Liu (1992) "Word Identification for Mandarin Chinese Sentences," *Proceedings of COLING*, 1992, pp. 101-107.

[10] R. Sproat, *et al.* (1994) "A Stochastic Finite-State Word-Segmentation Algorithm for Chinese," *Proceedings of ACL*, 1994, pp. 66-73.

[11] K. Chui (1995) "Repair in Chinese Conversation," *Proceedings of the Second International Symposium on Language in Taiwan*, 1995, pp. 75-96.

[12] 黃宣範 (1996) "漢語口語語料庫的建立," 語言學門專題計畫研究成果發表會, 台北, 南港, 1996.

[13] D. O'Shaughnessy (1992) "Recognition of hesitation in Spontaneous Speech," *Proceedings of ICASSP*, 1992, pp. 521-524.

Appendix A The Transcribing Conventions of the Corpus

Units

{Carriage Return}	Intonation Unit
--	Truncated Intonation Unit
{Space}	Word
-	Truncated Word

Speakers

:	Speaker Identify / Turn Start
[ ]	Speech Overlap

Transitional Continuity

.	Final
,	Continuing
?	Appeal

Terminal Pitch Direction

\	Fall
/	Rise
_	Level

Accent and Lengthening

^	Primary Accent
=	Lengthening

Pause

...(N)	Long
...	Medium
..	Short
(0)	Latching

Vocal Noises

(H)	Inhalation
%	Glottal Stop
@	Laughter

Quality

< @ @ >	Laugh Quality
< Q Q >	Quotation Quality
< F F >	Fast Tempo
< PP PP >	Very Soft
< MRC MRC >	Each Word Distinct and Emphasized

Specialized Notations

< L2 L2 >	Code Switching from Mandarin to English
(( ))	Transcriber's Comment

國語語音辨認中多領域語言模型之訓練、偵測與調適

Training, Detection and Adaptation of Multi-Domain Language Models for Mandarin Speech Recognition

林頌堅[1]、蔡吉龍[1]、簡立峰[2]、陳克健[2]、李琳山[1,2]
[1]國立台灣大學資訊工程學研究所
[2]中央研究院資訊科學研究所
e-mail: lsc@speech.ee.ntu.edu.tw

摘要

在本論文中，我們提出適用於不同領域間中文語言模型的自動訓練、偵測與調適方法。應用這些方法在極大詞彙國語語音辨認的語言解碼方法將可以訓練出各種不同應用領域的語言模型，為輸入的語音選擇合適應用領域的中文語言模型，對訓練語料不足的特殊領域語音辨認可以進一步提昇辨認正確率。在初步的實驗中，我們利用多領域語言模型進行語言解碼的語音辨認正確率可以比利用一般語言模型高2~8%，從這樣的結果可以驗證語言模型調適確有其效果，並值得做進一步的研究。

一、緒論

本論文提出一系列適用於不同領域間中文語言模型(Chinese Language Models)的自動訓練(Training)、偵測(Detection)與調適(Adaptation)方法。應用這些方法在極大詞彙國語語音辨認(Mandarin Speech Recognition with Very Large Vocabulary)的語言解碼(Linguistic Decoding)方法將可以為輸入的語音選擇合適應用領域(Application Domain)的中文語言模型，進一步提昇在特定領域下的語音辨認正確率。這項結果並有助於我們瞭解在不同領域下語言的統計特性，拓展國語語音辨認在不同應用領域的適用性。

在極大詞彙國語語音辨認中，常用的語言模型是統計式的馬可夫模型(Markov Model)[1,2,3]。訓練一個可靠的統計式馬可夫語言模型需要蒐集極大量的訓練語料(Training Corpus)，但是在若干應用環境中我們無法提供大量的訓練語料[4]。一個解決的方法是利用語言模型調適的技術將目前針對極大詞彙語音辨認系統所訓練出來的語言模型加以轉換成這個領域的特定語言模型，這樣的技術便成為目前語言模型中極富挑戰性的研究[5,6]。

在本論文我們提出利用多個領域特定語言模型(Domain-Specific Language Models)來進行語言模型調適的方法，語音辨認系統會根據使用者所輸入的語音自動選擇合適的應用領域語言模型進行語言解碼。使用多領域語言模型的語音辨認系統，一來可以提昇語音辨認的正確性；二來有助於瞭解如多種資料庫存取等多領域語音辨認應用的研究。在我們的方法中，首先將蒐集到所有的訓練語料訓練一個一般語言模型(General Language Model)，希望能抽取出不分領域所共有的

語言統計特性，然後對每一個應用領域，便可以少量的領域特定語料所訓練出的語言模型組合一般語言模型，產生一個新的領域特定語言模型，以適用於新領域的語言解碼。於是，對每一個應用領域，我們都有一個最合適的領域特定語言模型來描述該領域的語言特性。在使用者輸入語音時，我們就可以利用目前輸入語音在不同領域語言模型的辨認結果，自動挑選最合適的領域特定語言模型。在以多領域語言模型進行語音辨認的實驗中，對不同領域的測試語料使用該領域特定的語言模型比只用一般語言模型，均可得到 2~8%正確率提昇，因此可以證實語言模型調適對特定領域下的語音辨認有所幫助。

在本論文所提出的多領域語言模型方法中有兩個問題亟待解決。一是對不同領域如何訓練出該領域特定的語言模型，另一個問題則是進行語音辨認時，如何偵測輸入語音的應用領域來選擇領域特定語言模型。對於多領域語言模型的訓練，我們以內插法組合一般與應用領域語言模型，也就是以不同的加權值(Weighting Value)組合兩者的詞雙連機率值(Word Bigram Probabilities)，作為新的詞雙連機率值。每當新收到一筆語料需要進行領域特定語言模型訓練時，我們就以各個領域特定語言模型來判讀這筆語料與哪一類應用領域的訓練語料較接近，判讀之後將這筆語料加入最接近的應用領域訓練語料中，訓練出新的應用領域語言模型；如果新的訓練語料與所有訓練語料都相差相當大，我們便以這筆語料新成立一類應用領域，並對該應用領域進行語言模型訓練。在本論文中，我們實驗以文字複雜度(Perplexity)[7,8]與詞雙連涵蓋率(Word Bigram Coverage Rate)作為判讀應用領域的資訊。在結合這兩種資訊後，從實驗中，我們比較這種自動判讀領域的方法，可以發現訓練語料的領域分類與由人工判讀的結果相同。

對於選擇用來進行語言解碼的語言模型，我們以領域特定語言模型對輸入語音的辨認結果來偵測這些語音最接近哪些應用領域。在使用者輸入語音到多領域的國語語音辨認系統時，語言解碼單元先用所有經調適後的領域特定語言模型對前面輸入數句語音的候選音節序列進行語言解碼，將所得到的分數作為應用領域偵測的資訊。蒐集足夠預測接下來輸入的語音可能屬於哪一個應用領域的資訊之後，便可以該應用領域的語言模型提供接下來語言解碼所需的文法限制之用。

本論文的其餘章節結構如下；第二節對語言模型調適的問題再作一個清楚的描述，並且回顧一些前人在這個研究方向的努力以及在本論文中我們所提出來的方法。第三節報告我們的多領域語言模型訓練法以及使用文字複雜度和詞雙連涵蓋率來進行訓練語料領域判讀的一些實驗結果。第四節我們介紹利用多領域語言模型進行極大詞彙語音辨認的方法和一些初步的結果。最後，我們以第五節對本論文作一個總結，並展望未來在語言模型調適的研究方向。

二、語言模型調適

統計式馬可夫語言模型需要大量的語言模型作為決定模型參數的依據，但大量訓練語料的蒐集需要花費相當長久的時間與大量的人力，所以是件非常困難的工作。以往的極大詞彙的國語語音辨認的研究便是個很好的例子，在以往發展出來的實驗性國語語音辨認系統中，由於蒐集訓練語料的困難，目前能提供穩定而大量訓練語料的來源，只有中文報紙，所以在這些實驗或系統中，都以所蒐集到的新聞作為訓練語料來訓練中文語言模型，所能發展的應用也都偏向新聞聽寫的應用。但是，隨著極大詞彙語音辨認技術發展的成功，發展這項技術到其他應用

領域已是件水到渠成的工作了，比方說著名的飛行旅遊資料庫存取(Air Travel Information Service, ATIS Project)，便是美國尖端研究計畫署(ARPA)所極力推動的利用語音來存取資料庫內容應用的群體計畫[9]。在這些應用領域下的語言模型技術，自然成為一項重要的研究方向。

對於這些特殊的應用領域而言，因為輸入語音的用字、句型等等語言特性與新聞語料大不相同，顯然不能以原先利用新聞語料所訓練出來的語言模型作為語言解碼的資訊。但針對這些應用領域蒐集大量訓練語料並不是件容易的事，尤其是在一些新的應用領域，系統的雛形階段所能蒐集的語料非常的少。因此，在這種限制之下，特殊領域的語音辨認自然很難得到理想的結果。以往對於特殊應用領域的語言解碼，一個可行的方法是利用詞群語言模型(Word-Class-Based Language Models)[10,11]，這個方法針對應用領域內的用語，根據它們在這領域內的語法(Syntax)、語意(Semantics)等訊息來分群，譬如在 ATIS Project 中，我們可以將這個領域中所有可能用到的城市名稱歸為同一群，因為它們在這個領域都是指出發地和目的地，扮演同樣語法和語意的角色。對詞分群的方法可以降低語言模型的參數量，提供訊息分享(Information Sharing)的能力，因此可以增加語言解碼時的強健性(Robustness)。但是由於所能取得的訓練語料，實在相當有限，所以無法以自動分群的方法來進行語言模型訓練，需要不少以專家知識介入的地方。

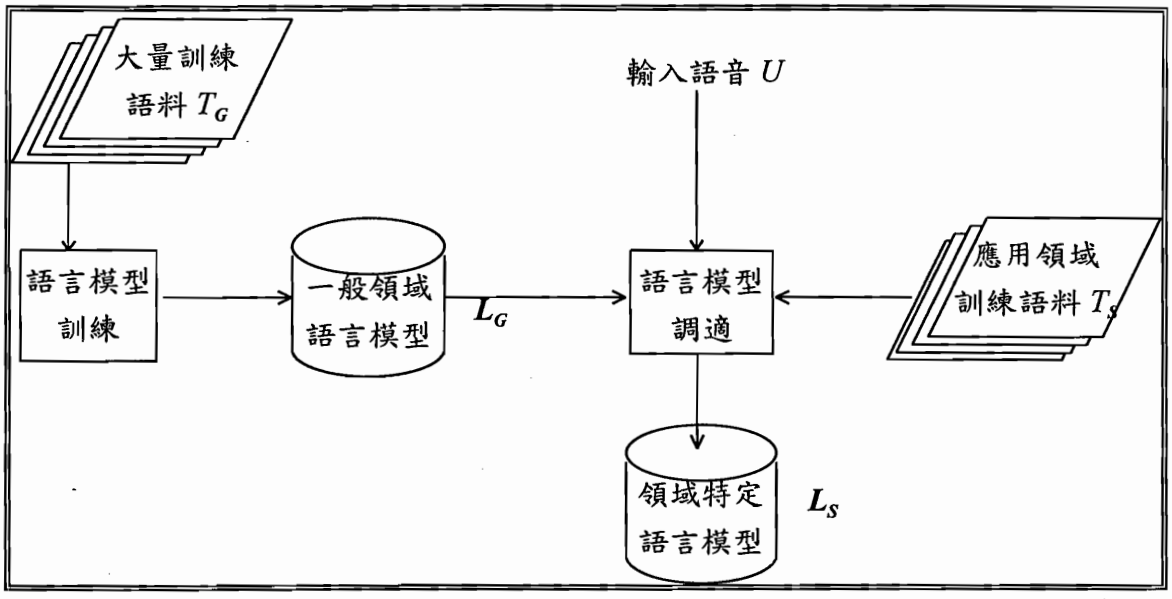


圖 1 語言模型調適示意圖

除了詞群語言模型之外，另外一種方式是以語言模型調適的方法來增加在特定領域下的語言解碼能力[12]。語言模型調適的概念是先以大量的訓練語料，訓練出一個一般語言模型 L_G，希望能從一個非常大量的訓練語料庫抽取出不分領域所共有的語言統計特性。然後當進行某應用領域 S 下的語音辨認時，對輸入的語音 U，抽取其中的語言特性資訊來對一般語言模型 L_G 進行調適，產生一個新的語言模型 L_S'，加強這個領域內特殊的用詞和語法等語言特性，以適用於新領域的語言解碼。用來對 L_G 調適的資訊可能只有 U 中的語言特性資訊，或是從少量應用領域 S 內蒐集的語料 T_S 中獲得。如圖 1 所示便是一個這樣的語言模型調適的過程。

過去在語言模型調適上的研究

在介紹我們的研究之前，我們首先回顧一些前人在這方面所做的努力。在特定領域的語料中，我們可以觀察到有某些詞經常重複出現的出現，比方說，有關「公共安全」領域內的語料中常常出現出現像“檢查、火災、安全、樓梯、消防”這一類的詞。在參考文獻[13]中就利用這個語言現象，以一個短期記憶體(Cache Memory, Short-Term Memory)來儲存辨認過的詞，當下次預測輸入語音中是否包含有這樣的詞時，便可以得到較高的機率。

再進一步觀察，可以發現領域內的高頻詞間往往存在某些關連性(Association)，於是參考文獻[14]便提出觸發對(Trigger Pair)的觀念，其觀念就是希望從特定領域語料內抽取兩兩具有具有關連性的詞，把這樣的一對詞稱為觸發對。當輸入的語音中有觸發對的其中一個詞出現時，另外一個詞在語言模型中的機率值也跟著動態調整，來提昇語言解碼的正確性。從實驗數據顯示，經過人工選取的觸發對的確可以降低約 12%的文字複雜度。

為了使語言模型具有隨著應用領域而調整的能力，以符合特殊應用領域之特性，參考文獻[15]和[16]都以最小區辨資訊(Minimum Discrimination Information, MDI)的方法改變最小的分佈差異來調整不同領域間不同的 n 連機率。實驗結果顯示這種方法可以降低語音辨認中詞的錯誤率約 10~14%，表示這樣的方法的確有它的發展空間，而這樣的一個實驗結果也引發我們進行分析不同應用領域中語言特性的動機，於是我們便著手研究多應用領域中文語言模型調適的方法。

本論文所提出的方法

本論文所針對語言模型調適的問題，是一個藉由多領域語言模型[12]來進行調適的問題，也就是在使用者輸入語音後，語音辨認系統根據所輸入的語音選擇合適的應用領域語言模型，作為調適的語言模型，進行語言解碼。使用多領域語言模型的語音辨認系統，一來因為所使用的語言模型較一般語言模型精確，可以對各領域內不同的語言特性做較精細的描述，有助於提昇語音辨認的正確性；二來目前如多種資料庫存取等多領域語音辨認的應用已逐漸興起，以多領域語言模型整合極大詞彙語音辨認系統，有助於瞭解這方面應用的研究。

在進行語音辨認時，當使用者開始輸入語音，我們首先以少量輸入語音來預測目前輸入語音是屬於哪一種領域，比方說 S 領域，然後便以 S 的領域特定語言模型 L_S'對接下來輸入的語音進行語言解碼。領域特定語言模型 L_S'包含了不分領域共同的語言特性與這個領域內獨特的語言特性，其訓練方法如下：首先我們利用蒐集到的大量訓練語料，訓練一個包含所有不分領域的一般語言模型 L_G。L_G 具有每一個應用領域所共有的語言統計特性，以 L_G 進行語言解碼，基本上對各種領域的語音，已經可以達到還不錯的辨認效果。然後對每一個應用領域 S，以所蒐集到少量的領域特定訓練語料 T_S 來訓練領域 S 的語言模型 L_S，加強領域 S 內所特有的語言特性，再組合 L_G 與 L_S，產生新的領域特定語言模型 L_S'。於是，對每一個應用領域，我們都有一個最合適的領域特定語言模型來描述該領域的語言特性。

在本論文中，所採用的語言模型是詞雙連語言模型。以內插法(Interpolation)[17]來組合一般詞雙連語言模型 L_G 和領域 S 詞雙連語言模型 L_S 中

的詞雙連機率值，作為領域特定語言模型 L_S'內的詞雙連機率值。我們可以用式(1)的數學形式來更清楚的表達這個方法。

$$\Pr_S'(w_i|w_j) \overset{def}{=} (1-d)\times \Pr_G(w_i|w_j) + d \times \Pr_S(w_i|w_j) \qquad \ldots\ldots(1)$$

在這個式子中，w_i 和 w_j 代表詞典中的任意兩個詞，$\Pr_G(w_i \mid w_j)$和 $\Pr_S(w_i \mid w_j)$分別代表(w_j, w_i)這對詞出現在 L_G和 L_S中的詞雙連機率值。d 是一個介於 0 和 1 之間的數值，用來調整調適詞雙連機率值 $\Pr_S'(w_i \mid w_j)$中 $\Pr_G(w_i \mid w_j)$和 $\Pr_S(w_i \mid w_j)$間的比重。一般而言，若是我們蒐集到較多的應用領域訓練語料，我們可以較信賴領域詞雙連語言模型 L_S，d 的值可以設一個較大的值；否則，我們需要仰賴一般詞雙連語言模型 L_G提供語言解碼時的資訊，d 就必須設定一個較小的值。下面，我們以一般以及領域特定詞雙連語言模型進行語言解碼的實驗，以驗證語言模型調適的效果。

實驗環境

在報告語言解碼的實驗及其結果之前，我們首先對實驗的環境參數作一介紹，爾後在本論文的實驗中，若非特別提及，則表示相同的參數。

(1)詞典：

由中研院詞庫小組提供的詞典，共 84464 目詞，取高頻之一字到四字詞共 43591 目，其詞目數分佈如表 1 所示：

詞長	一字詞	二字詞	三字詞	四字詞
詞目數	8,154 目	23,529 目	5,494 目	6,414 目

表 1 本論文所使用詞典的詞目數分佈情形

(2)語言模型訓練語料：

在語言模型訓練語料中，包括了一般與特定領域兩類。對於一般語料，我們所能蒐集到穩定與大量的中文語料是新聞語料，同時中文新聞語料包羅萬象，幾乎涵蓋了各個領域，因此我們以三家報社共九個月的新聞語料做為我們實驗中一般語言模型的訓練語料。至於特定領域語料的選取，我們從不同語料來源來蒐集，包括從現代漢語平衡語料庫所選取的哲學類(phi)和文學類(lit)相關文章、從電子佈告欄(BBS)中蒐集的棒球區(ball)和微軟視窗軟體區(win)內的討論文章、以及一些關於科技報告的文章(cs)。表 2 是這些訓練語料的大小。

領域別	一般	phi	lit	ball	win	cs
詞數	12,094,234	67,127	61,676	170,214	16,647	3,445
字數	18,384,664	100,054	87,987	226,864	23,463	5,612

表 2 本論文實驗所用各個領域訓練語料的大小

(3)測試語料：

對每一特定領域，我們從對應的領域中選取若干篇文章進行測試，但所選取的文章並不與訓練語料所選取的重複。換言之，這些語料全部都是外部測試(Outside Test)，所有的應用領域共計 20 篇測試語料。在語音輸入方面，我們以一位語者對每篇測試語料唸一遍，以金聲三號的聲學處理單元[1]將輸入的語音轉換成一序列的候選音節，將這些候選音節儲存起來，便於比較。我們

以候選音節中是否有出現正確音節的比率來衡量聲學處理單元對每一測試語料的正確性，這樣的比率我們稱為音節包含率(Inclusion Rate)。每個領域的測試語料大小及音節包含率如表 3 。

領域別	phi	lit	ball	win	cs
詞數	5,799	4,621	4,470	17,535	7,038
字數	8,407	6,269	6,011	24,271	11,223
音節包含率	99.82%	99.81%	99.75%	99.70%	99.88%

表 3 各測試文章所含有之字詞數及平均正確音節涵蓋率

實驗結果

表 4 是我們以一般及領域特定語言模型對各個領域的測試語料進行語言解碼所得到的字正確率。表中的各欄代表每一個不同領域測試語料對不同語言模型的結果，這些結果均是經過仔細調整式(1)中一般與領域語言模型的比重 d 而得到最佳辨認正確率的結果，我們也將每一應用領域的調整比重值列於最後一列中。

	phi	lit	ball	win	cs
一般 語言模型	82.52%	82.04%	81.43%	82.28%	87.04%
領域特定 語言模型	85.24%	87.03%	90.13%	87.54%	90.72%
調適比重 d	0.75	0.95	0.95	0.75	0.25

表 4 以一般及領域特定語言模型

對不同應用領域測試語料進行語言解碼所得到的結果

從上面的表中，我們可以觀察到對各個領域的測試語料，即使以一般語言模型進行語言解碼已可獲得不錯的效果，但以領域特定語言模型進行語言解碼，確實可以得到更好的效果，辨認正確率上昇的幅度從 2 到 8%。從上表中，我們也

觀察到一些有趣的現象：在這個實驗中，辨認正確率上昇幅度最高的是電子佈告欄的棒球類討論文章(ball)這個領域，經調適後，正確率可以從 81.43%上昇到 90.13%。經過仔細閱讀調適和測試語料，發現裡面內容與一般調適語料最大的不同是這類語料的文章主題(Subject)較狹隘，裡面包含了許多這領域內的術語，這些術語在調適和測試語料一再被提及，所以在經過調適之後，可以得到大幅度上昇的辨認正確性。

另外一個值得注意的現象是，科技報告文章領域(cs)的調整比重 d 遠比其他領域來的小，我們可以對照表 4 和表 2，發現這領域的語料量遠比其他領域小，因此可驗證了我們在前面的推論者，調整比重 d 與領域的語料大小有關。我們對這個現象進一步進行領域語料量與調整比重間關係的實驗，我們逐漸增加 cs 領域的語料量來觀察最佳辨認正確率與對應的比重，實驗結果列於表 5。從表 5 中，我們可以觀察到在語料量增加之下，語音辨認的正確率的確有增加的趨勢，同時，在訓練語料增加兩倍之後，調整比重已從傾向一般變成為傾向領域語言模型，可見得語料量的大小確實影響語言模型的調適與解碼效果。

語料量	5,612 字	10,768 字	17,373 字
辨認正確率	90.72%	91.31%	91.45%
調整比重 d	0.25	0.25	0.75

表 5 不同語料量大小與調整比重及辨認正確率間的關係

由上面這些實驗結果中，可以觀察對應用領域內的測試語料，以領域特定語言模型可以獲得更好的結果，同時增加語料量對語音正確率有顯著的幫助。因此如何自動判讀訓練語料的應用領域來訓練領域特定語言模型與語音辨認系統如何自動而迅速偵測使用者輸入語音的應用領域，便成為多領域語言模型的兩大問

題，在下面兩節中，我們將探討這兩個問題並試圖找出可能的解決方法。

三、多領域語言模型訓練

以多領域語言模型來進行語言解碼，我們首先要對每一個領域訓練一個領域的語言模型以及一個一般語言模型，然後以內插法組合一般與領域特定語言模型來訓練領域特定語言模型。從前一節的實驗中，我們可以瞭解到語料的增加可以進一步提昇這個領域內測試語料的辨認正確率，所以即使已經產生領域特定語言模型，若我們能蒐集到新的語料，最好能將它們歸類到最接近的應用領域，加入訓練語料中。在本節中，我們將描述多應用領域語言模型的訓練方法，尤其著重在判讀新語料的領域判讀上。

我們可以用圖 2 中的演算法來表示我們的多領域語言模型訓練方法。在初始的時候，在步驟 I-1 到 I-3，我們首先訓練每一個應用領域的詞雙連語言模型 L_S 與一般的詞雙連語言模型 L_G，並以內插法組合一般與領域特定的詞雙連機率值，作為新的領域特定詞雙連機率值，所以在步驟 I-3 完成後我們對每一個應用領域有一個領域特定語言模型 L_S'，已經可以進行多領域的語音辨認應用。而在系統已經使用之後，每當新收到一筆語料 T，這筆語料可以加入目前已有的領域，或另外成立一類新的應用領域。步驟 II-1 中首先以目前的領域語言模型來判讀 T 與哪一類應用領域的訓練語料較接近，判讀之後便可將將這筆語料加入最接近的應用領域 S 訓練新的領域特定語言模型 L_S'(步驟 II-2)；如果新的訓練語料 T 與所有訓練語料都相差相當大，我們便以這筆語料新成立一類應用領域 S，再訓練屬於該應用領域的詞雙連語言模型 L_S'(步驟 II-3)。

I.初始:

I-1. L_G = TrainingModel(T_G)

// 以極大量語料 T_G 訓練一般領域語言模型

I-2. Λ = {}

I-3. for any domain S

L_S = TrainingModel(T_S), Λ = Λ ∪ {L_S}

// 對應用領域 S，訓練領域語言模型 L_S

L_S' = Interpolation(L_G,L_S)

// 組合產生領域特定語言模型

II.當新語料 T 進入:

II-1. S = DomainDetermination(T,Λ)

// 判讀新進入的語料 T，屬於哪一個領域

II-2. if (S != NULL)

// 找到合適的應用領域

T_S += T, L_S = TrainingModel(T_S), L_S' = Interpolation(L_G,L_S)

// 將 T 加入應用領域 S 中，產生領域特定語言模型

II-3. else

// 沒找到合適的應用領域

T_S = T, L_S = TrainingModel(T_S), Λ = Λ ∪ {L_S}

L_S' = Interpolation(L_G,L_S)

// 以 T 成立新應用領域 S，產生領域特定語言模型

圖 3 多領域語言模型訓練與調適的演算法

訓練語料所屬領域特定語言模型的判讀

判讀新語料屬於哪一個應用領域，事實上是一種文件分類(Document Classification)的問題。在過去的資訊檢索(Information Retrieval)研究中，已經提出相當多自動化的文件分類方法[18,19,20]，這些方法可以歸納成三個步驟。

一、選定文件的集合與分類的類別，並選定文件的全部或者具有較多資訊的部份(比方說，題目或摘要等等)作為訓練或判讀的資料，稱為文件的簡述(Profile)，

簡述可以是為是該文件的代用品(Representation)。

二、根據訂定的篩檢規則(如：詞類和頻率等等)，從簡述中挑選出關鍵詞(Keywords)來。

三、可以用向量模式(Vector Space Model)或機率模式(Probability Model)來估測文件與類別間的相關程度，以進行類別的判讀。

由此可知，這些方法多以關鍵詞為分類的主要資訊，換言之，這樣的方法只考慮關鍵詞本身的局部性資訊。在本論文中，我們嘗試利用詞雙連語言模型中包含的詞與詞間的關連性(Association)來判讀語料所屬的應用領域，希望能夠捕捉到更進一步的資訊，得到更好的判讀結果。下面簡介本論文所用的兩類詞關連性資訊，文字複雜度(Perplexity)和詞雙連涵蓋率(Word Bigram Coverage)，同時報告以各種領域特定語言模型量測在第二節實驗中所用各種領域測試語料的文字版本(Text Transcription)做為測試語料所得到的文字複雜度和詞雙連涵蓋率來探討這兩類資訊做為領域判讀的可能性。

1. 文字複雜度

在本論文中，我們以詞雙連語言模型對測試語料的文字複雜度(Perplexity)[7,8]的大小來衡量領域特定語言模型與新的訓練語料間的關連性。在定義文字複雜度之前，我們首先定義詞群語言模型在長度為 N 測試語料 W 上的熵值(Entropy)H_p 為

$$H_p = -\frac{1}{\sum_{i=1}^{N}|w_i|}\sum_{i=1}^{N}\log \Pr(w_i|w_{i-1}) \qquad .....(2)$$

這裡，w_i 代表測試語料 W 中出現的詞，$|w_i|$則是代表詞 w_i 的字數。H_p 可以作為評估不同語言模型對測試語料預測能力的強弱。H_p 值愈大，代表這樣的語言模型對此篇測試語料的預測能力愈強，反之亦然。為了方便表示出每個字平均後接字的數目，接著我們定義文字複雜度為

$$PP = 2^{Hp} \qquad .....(3)$$

平均文字複雜度的物理意義可認為是用語言模型來預測測試文章平均每個字後面可以接字的數目。所以觀察語言模型的文字複雜度可以判斷用來訓練語言模型的語料與測試語料字詞間的關連性。舉例而言，如果一個語言模型 L 對某一測試語料 W 有較小的文字複雜度，也就是說，利用 L 來預測 W 中任意給定的字，平均每個字的後接字數都很少，這樣意謂著 L 的訓練語料與 W 的字詞間有很大的關連性。反之則 L 的訓練語料與 W 可能沒有關連性。所以我們可以用訓練好的領域特定語言模型來判讀新的訓練語料是否與這個領域的其他訓練語料有關連。我們可以對每一個領域特定語言模型來量測新的訓練語料的文字複雜度，將新的訓練語料加入那些得到較小文字複雜度的語言模型所屬的領域中。

表 6 是我們以各種領域特定語言模型量測不同領域測試語料所得到的文字複雜度。表 6 中，由左而右的各欄分別是 phi 、 lit 、 ball 、 win 和 cs 各領域的測試語料在各種領域特定語言模型下所得到的文字複雜度，而每一列分別代表各種領域特定語言模型對不同應用領域測試語料的文字複雜度。

	phi	lit	ball	win	cs
L_{phi}	**352**	765	1067	765	521
L_{lit}	554	**301**	1174	781	587
L_{ball}	601	838	**37**	604	534
L_{win}	876	1239	1349	**402**	473
L_{cs}	1249	2221	2376	1812	**362**

表 6 以各種領域特定語言模型對不同應用領域測試語料的文字複雜度

2. 詞雙連涵蓋率

我們定義詞雙連涵蓋率如下：對一篇具有 m 對詞雙連對的語料 W,$B = \{b_1, b_2, \ldots b_i, \ldots b_m\}$，代表語料中所有詞雙連對所成的集合，其中 b_i 代表測試文章中第 i 對詞雙連對。 M 則表示這個領域中所有訓練語料內不同詞雙連對所成的集合。則詞雙連涵蓋率可定義如下:

$$\text{詞雙連涵蓋率} = 100\% \times \frac{\sum_{i=1}^{m} \chi_i}{|B|} \qquad \chi_i = \begin{cases} 1 & \text{if } b_i \in M \\ 0 & \text{otherwise} \end{cases} \qquad \text{.....(4)}$$

由上面的定義，我們可以知道對一新語料 W 而言，詞雙連涵蓋率代表的是這個語料中的詞雙連對出現在應用領域訓練語料的比率。詞雙連對所代表的則是詞的前後相連關係，因此詞雙連涵蓋比率愈高，顯然新語料 W 中詞的前後關連性與這個應用領域愈像。所以，詞雙連涵蓋率可以用來作為新語料屬於哪一個應用領域的判讀標準。

表 7 是我們以各種領域特定語言模型量測不同領域測試語料所得到的詞雙連涵蓋率。表 7 中，由左而右的各欄分別是 phi 、 lit 、 ball 、 win 和 cs 各領域的測試語料在各種領域特定語言模型下所得到的詞雙連涵蓋率，而每一列分別代表各種領域特定語言模型對不同應用領域測試語料的詞雙連涵蓋率。

	phi	lit	ball	win	cs
L_{phi}	**37.28%**	31.39%	23.45%	28.62%	12.42%
L_{lit}	25.95%	**52.24%**	20.77%	27.89%	8.97%
L_{ball}	24.68%	29.28%	**99.70%**	35.55%	11.96%
L_{win}	13.73%	17.48%	15.93%	**40.28%**	12.94%
L_{cs}	5.02%	3.30%	3.30%	4.71%	**17.88%**

表 7 以各種領域特定語言模型對不同應用領域測試語料的詞雙連涵蓋率

應用領域判讀的討論

從表 6 和表 7 之中，我們可以觀察到每一應用領域測試語料在對應的領域特定語言模型下所得到的文字複雜度都比其他領域特定語言模型所得到者來得低；詞雙連涵蓋率中也有類似的現象發生，每一應用領域測試語料在對應的領域特定語言模型下的詞雙連涵蓋率都比其他領域特定語言模型來得高。我們可以驗證前面所推論者，新進入的語料在領域特定語言模型下得到較低的文字複雜度和較高的詞雙連涵蓋率，語料與這個應用領域間應存有某種關連性。這樣的結果證實我們可以用文字複雜度和詞雙連涵蓋率來作為應用領域判讀的資訊。

從以上兩表中，我們也可以觀察到一些有趣的現象。首先我們可以看到 ball 領域的測試語料對 ball 的領域特定語言模型 L_{ball} 有非常低的文字複雜度與非常高的詞雙連涵蓋率，這可以進一步驗證我們在前一節中所觀察的現象，我們所收集到的電子佈告欄的棒球討論區語料裡存在許多的術語，這些術語一再出現的結果造成 L_{ball} 對這領域的測試語料有非常低的文字複雜度與非常高的詞雙連涵蓋率。

另一個觀察是 win 和 cs 兩個領域都是有關電腦技術方面的領域，在這些領域中是主題最接近的兩個領域。這樣的關係對文字複雜度和詞雙連涵蓋率的量測

有什麼影響呢？以 win 的領域特定語言模型 L_{win} 來測量 cs 領域的測試語料得到在所有領域特定語言模型僅次於本身領域特定語言模型 L_{cs} 的效果，這再次的證實了我們以文字複雜度和詞雙連涵蓋率作為應用領域判讀資訊的可行性。反過來，雖然以 cs 的領域特定語言模型 L_{cs} 來測量 win 領域的測試語料得不到這樣好的效果，但造成這種現象的主要原因是由於 cs 領域的訓練語料量並不充足，所以我們無法得到非常可靠的語言模型估計。

在組合兩種資訊之後，我們以測試語料來進行應用領域判讀，在總共 20 篇的測試語料中，所有由自動判讀的結果都與原先人工分類者相同。雖然實驗的測試語料太少了，所以可以得到這麼好的結果。但是以文字複雜度與詞雙連涵蓋率作為應用領域判讀的優點也可略見其端倪。

四、使用多領域語言模型之語言解碼

在我們的多領域語言模型之語言解碼方法，一開始我們以所有應用領域的領域特定語言模型進行語言解碼，逐漸排除不可能的應用領域，一直到最後只剩少數一、二個可能的領域。選取應用領域所依據的資訊是以對應的領域特定語言模型對輸入語音進行語言解碼所得到的分數。圖 3 是我們的多領域語言模型語言解碼方法。在使用者輸入語音到多領域的國語語音辨認系統時，前面幾句我們先以所有領域特定語言模型來偵測這些語音最接近於哪些應用領域(步驟 I-2)。語言解碼單元在接受聲學處理單元比對出的候選音節序列 Σ 後，在步驟 II-2 時，以每一個領域特定語言模型對 C 進行語言解碼，從中搜尋出最有可能的字串。在進行完所有領域的語言解碼之後，步驟 II-3 裡，選取其中解碼分數最高的字串 C_{S^*}，

將這條字串輸出(步驟 II-4)，同時記錄下這條字串所屬的領域 S^*(步驟 II-5)。在蒐集足夠預測接下來輸入語音的可能應用領域資訊後，便可以目前可能應用領域的領域特定語言模型提供接下來語言解碼所需的文法限制之用(步驟 III-1)，完成語言模型調適。

I. 初始：
  I-1. $\Lambda = \{\}, \Lambda' = \{\}$
    // 將解碼集合 Λ 和修正領域集合 Λ'設為空集合
  I-2. for all L_S'
    $\Lambda = \Lambda \cup \{L_S'\}$
    // 將系統內所有領域的調適後領域特定語言模型 L_S'加入 Λ 中
II. 使用者輸入語音 U
  II-1. Σ = AcousticMatching(U)
    // 聲學處理單元將 U 辨認為候選音節序列 Σ
  II-2. for all L_S' in Λ
    (C_S, P_S)= LinguisticDecoding(Σ, L_S')
    // 以調適後領域特定語言模型 L_S'對 Σ 進行語言解碼，得到字串 C_S與解碼分數 P_S
  II-3. select S^* , such that $P_{S^*} \geqq P_S$ for all L_S' in Λ
    // 找出分數最高的領域 S^*
  II-4. output C_{S^*}
    // 輸出由 L_{S^*}'解碼出的字串 C_{S^*}
  II-5. $\Lambda' = \Lambda' \cup \{L_{S^*}'\}$
    // 更新修正領域集合 Λ'
III. 經過數句之後：
  III-1. $\Lambda = \Lambda'$
    // 以修正領域集合 Λ'更新解碼集合 Λ，調適語言模型

圖 3 多領域語言模型語言解碼的演算法

多領域語言模型語言解碼的實驗

這個實驗中，我們以前面實驗所用的五個應用領域來進行多領域語言模型的

語言解碼，討論語音所屬應用領域的偵測與比較一般語言模型與應用多語言模型語言解碼得到的辨認正確性。首先來看看輸入語音的句數與應用領域偵測正確性的關係，我們以本節中所提出的利用多領域語言模型的語言解碼方法，對 20 篇測試語料進行語言解碼，記錄下每一次輸入語句後所得到的前三個偵測到的應用領域，加以統計，圖 4 所示是以輸入句數為橫座標，應用領域偵測的正確率為縱座標所得到的關係圖。從圖 4 可以看出前三個偵測的結果就已經相當不錯。在本圖中，我們可以同時觀察到輸入句數與應用領域偵測的正確率有正面的影響，也就是說，輸入的句數愈多，可以偵測到愈正確的應用領域。這個結果證實了本節所提出的利用多領域語言模型的語言解碼方法具有語言模型調適的能力。

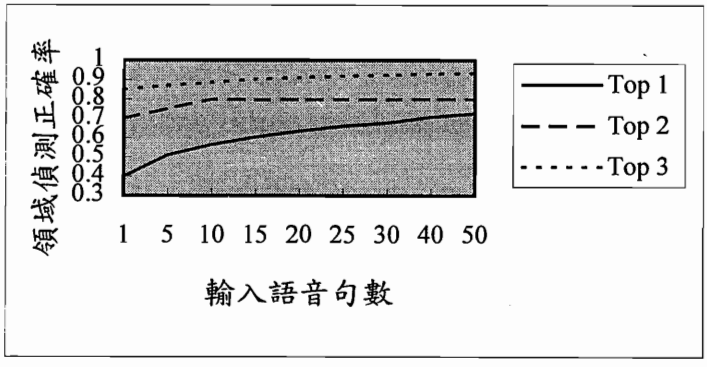


圖 4 輸入語音句數與應用領域偵測正確性的關係圖

接下來我們來看看實際應用這個多領域語言模型的語言解碼方法所得到的語音辨認效果。表 8 是對各種應用領域的測試語料所得到的字正確率，對照於在第二節中的表 4 中以一般語言模型進行語言解碼所得到的字正確率，可以看出這個多領域語言模型的語言解碼方法確實有助於辨認正確率的提昇。此外，我們可以發現在某些訓練語料相當缺乏的領域中，使用多領域語言模型甚至可以得到比已知應用領域時直接採用領域特定語言模型還要好的辨認結果，因為我們可以利

用其他主題較為接近的應用領域所訓練出來的語言模型來彌補某些沒有出現在這個領域訓練語料中的語言特性。

測試語料的領域	phi	lit	ball	win	cs
語音辨認字正確率	84.43%	85.68%	88.70%	87.13%	91.13%

表 8 利用多領域語言模型的語言模型調適法進行語音辨認得到的字正確率

五、結論

在本論文中，我們提出適用於不同領域間中文語言模型的自動訓練、偵測與調適方法。在多領域語言模型的訓練上，我們根據文字複雜度與詞雙連涵蓋率來判讀新訓練語料的應用領域，然後對每一個應用領域利用領域內的訓練語料訓練一個語言模型，再以內插法組合這個語言模型與由不分領域的大量語料訓練的一般語言模型，得到這個應用領用的領域特定語言模型。在進行語音辨認時，系統自動偵測輸入語音選擇最適合的領域特定語言模型，作為調適的語言模型，來提供語言解碼所需的文法限制。我們以利用領域特定語言模型進行語言解碼得到的分數作為偵測應用領域所需的資訊。應用這些方法在極大詞彙國語語音辨認的語言解碼方法將可以為輸入的語音選擇合適應用領域的中文語言模型，對訓練語料不足的特殊領域語音辨認可以進一步提昇辨認正確率。

目前我們利用多領域語言模型作為語言模型已經得到一些初步的結果，值得我們繼續發展這類技術的研究。在未來的方向上，我們希望能夠繼續提昇應用領域偵測的準確性，同時在多領域的應用環境中，需要更精簡的語言模型，結合前人在語言模型調適中所發展出來的技術，諸如：短期記憶體、觸發對和最小區辨資訊等等，希望能夠提供更有效的語言模型調適技術，找出更符合在應用領域中

的語言特性。另外，利用文字複雜度和詞雙連涵蓋率等詞關連性資訊可以得到相當令人滿意的應用領域判讀結果，我們希望擴展這類的研究到文件分類的技術上。

參考文獻

[1] H-W. Wang, et. al., "Complete Recognition of Continuous Mandarin Speech for Chinese Language with Very Large Vocabulary but Limited Training Data," in Proceedings of ICASSP'95, pp. 61-64, Detroit, USA, 1995.

[2] Y-J. Yang, S-C. Lin, L-F. Chien, K-J. Chen, and L-S. Lee, "An Intelligent and Efficient Word-Class-Based Chinese Language Model for Mandarin Speech Recognition with Very Large Vocabulary," Proc. ICSLP'94, pp. 1371-1374, Yokohama, Japan, 1994.

[3] Y-C. Chang, S-C. Lin, L-F. Chien, K-J. Chen, and L-S. Lee, "Methodology, Implementation and Application of Word-Class Based Language Model in Mandarin Speech Recognition," Proc. ROCLING VII, pp. 17-34, Hsinchu, ROC, 1994. (in Chinese)

[4] S-C. Lin, L-F. Chien, K-J. Chen, and L-S. Lee, "Unconstrained Speech Retrieval for Chinese Document Databases with Very Large Vocabulary and Unlimited Domains," Proc. EUROSPEECH'95, pp. 1203-1206, Madrid, Spain, 1995.

[5] C-H. Lee, "Stochastic Modeling in Spoken Dialogue System Design," Speech Communication, Vol. 15, pp. 311-322, 1994.

[6] R. Cole, et. al., "The Challenge of Spoken Language Systems: Research Directions for the Nineties," IEEE Trans. Speech and Audio Processing, Vol. 3, No. 1, pp.1-20, 1995.

[7] F. Jelinek, "The Development of an Experimental Discrete Dictation Recognizer," in Proceedings of IEEE, Vol. 73, No. 11, pp. 1616-1624, Nov. 1985.

[8] K.F. Lee, *Large-Vocabulary Speaker-Independent Continuous Speech Recognition: The SPHINX System*, Ph.D. Dissertation, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA, Apr. 1988.

[9] Speech and Natural Language: Proceedings of the ARPA Workshop, Morgan Kaufmann Publishers, CA, USA, 1994.

[10] B.Suhm and A. Waibel, "Towards Better Language Models for Spontaneous Speech," in Proceedings of ICSLP'94, Vol. II, pp. 831-834, Yokohama, Japan,

1994.

[11] M. McCandless and J. Glass, "Empirical Acquisition of Language Models for Speech Recognition," in Proceedings of ICSLP'94, Vol. II, pp. 835-838, Yokohama, Japan, 1994.

[12] S. Matsunaga, T. Yamada, and K. Shikano, "Task Adaptation in Stochastic Language Models for Continuous Speech Recognition," in Proceedings of ICASSP'92, Vol. I, pp. 165-168, San Francisco, California, USA, 1992.

[13] R. Kuhn and R. D. Mori, "A Cache-Based Natural Language Model for Speech Recognition," IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. PAMI-12, No. 6, pp. 570-583, Jun. 1990.

[14] R. Lau, R. Rosenfeld, and S. Roukos, "Trigger-Based Language Models: A Maximum Entropy Approach," in Proceedings of ICSSP'93, Vol. II, pp. 45-48, Adelaide, South Australia, 1993.

[15] S. D. Pietra, V. D. Pietra, R. L. Mercer, S. Roukos, "Adaptive Language Modeling Using Minimum Discriminant Estimation," in Proceedings of ICASSP'92, Vol. I, pp. 633-636, San Francisco, California, USA, 1992.

[16] R. Rosenfeld, "A Hybrid Approach to Adaptive Statistical Language Modeling," in Proceedings of Human Language Technology Workshop, pp. 76-81, 1994.

[17] P.S. Rao, M. D. Monkowski, and S. Roukos, "Language Model Adaptation via Minimum Discrimination Information," in Proceedings of ICASSP'95, Vol. I, pp. 161-165, Detroit, Michigan, USA, 1995.

[18] R. R. Larson, "Experiments in Automatic Library of Congress Classification," JASIS, Vol. 43, No. 2, pp. 130-149, 1992.

[19] D. D. Lewis, "An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task," in Proceedings of SIGIR'92, pp. 37-50, Copenhagen, Denmark, 1992.

[20] P. S. Jacobs, "Using Statistical Methods to Improve Knowledge-Based News Categorization," IEEE Expert, Vol. 8, No. 2, pp.13-23, Apr. 1993.

中英文字檔案區域調整資料壓縮方法

A Locally Adaptive Data Compression Scheme for Chinese-English Text Files

張克章* 徐熊健** 朱賢武+

*國防管理學院資訊管理學系

E-mail:chang@rs360.ndmc.edu.tw

**銘傳管理學院資訊管理學系

+國防管理學院資源管理研究所

摘要

隨著資訊科技快速發展與普及，人類仰賴資訊網路進行國際間資料交換的需求與日俱增，因此適合應用於網路傳輸的資料壓縮方法成爲一項紓解資訊網路擁塞問題必需的方法。由於中英文混合檔案資料在國內外一般的應用相當普遍。本文提出雙區域調整串列對照編碼法(Double Adaptive List Correspondence,DALC)方法，屬於單一回合過程且適合應用於網路傳輸的文字檔案壓縮方法，基本的構想係建立在二維串列與參考區域原則，改進以往二維串列區域調整編碼法無法處理中英文混合資料壓縮的缺失。經由本文運用move-to-front原則及提出的group-move-to-front原則，在資料壓縮過程中建置兩個串列對照輔助結構並配合前導位置整數表示法，有效結合字母導向調整方式，完成了本文之資料壓縮方法。爲測試本文中提出壓縮方法的優點與可行性，本文針對同時出現中英文混合資料與較可能使用的常用字爲著眼考量，乃從時報資訊立即新聞稿中擷取測試資料，包括三種全英文、全中文及中英文混合等三種檔案資料。經過測試並比較壓縮率發現本文中所提出編碼方法之壓縮效益優於其他方法；同時本文方法亦具有容易製作的優點，將可實際應用於資訊網路。

1 緒論

隨著資訊技術發展一日千里，人類仰賴國際間資訊網路通訊的需求與日俱增，無可避免地面臨了各類龐大資訊充斥於網路上所造成擁塞現象。資料壓縮(data compression)的重要性不喻可知，尤其快速且適合應用於網路傳輸的資料壓縮工具更是不可或缺。如同其他東方語系國家一樣，普遍存在著中文與英文資料混合的一般文章及報導資料，因此就實用價值觀點，資料壓縮的研究必

須實際地探討中英文字混合檔案壓縮技術，不應再侷限探討缺少實用程度的完全英文或中文檔案壓縮方法。現行許多中文檔案壓縮方法是由英文檔案壓縮方法改良而來。設計文字壓縮技術在基本策略上有整體最佳(global optimum)與區域最佳(local optimum)兩種策略。所謂整體最佳策略係針對檔案整體全部作爲編碼依據，而區域最佳策略則依據壓縮過程中按目前所讀取的檔案資料狀況作爲編碼依據。因此若採行整體最佳策略，進行壓縮作業時，至少須經過二個回合(two passes)以上的處理過程，而區域最佳策略則僅需一個回合(one pass)處理過程。許多文字檔案資料壓縮研究透過上述兩種策略，運用各種例如樹狀結構(tree)、表格(table)、串列(list)以及統計運算(statistics)等資料結構來探討文字檔案資料壓縮方法。有關文字檔案壓縮技術分類如圖1.1表示。就實用觀點，壓縮過程讀取檔案兩個回合，且壓縮(compress)及解壓縮(decompress)皆須負擔前置成本(overhead)，相當不適合應用於網路傳輸。

爲設計一個適合應用於網路傳輸的文字檔案資料壓縮方法，必須符合單一回合處理過程、壓縮及解壓縮過程簡單且快速等需求。因此本文選擇區域調整串列結構(locally adaptive list structure)作爲中英文字檔案壓縮之探討方向。本文所提出的雙區域調整串列對照編碼法(Double Adaptive List Correspondence,DALC)，係針對原二維串列(two dimensional linked list)區域調整編碼法[14]無法處理英文檔案資料壓縮的瓶頸。本文嘗試結合字母導向(alphabet-oriented)調整方式，採用前導位置整數(prefix positional integer)表示法並配合BIG-5中文碼特性，一併解決了中英文混合檔案資料壓縮、前導位置整數編碼長度及控制串列長度等成本問題，並驗證本文所提方法的實用價值及壓縮效益。爲驗證本文所提方法的可行性與優點，本文以三種檔案，包括全英文、全中文及中英文混合等形式，進行實證分析與比較，測試結果證實本文所提方法的可行性。

2 相關文獻探討

以下介紹與本文研究相關之文獻，並指出各方法之特性與缺點。

2.1 霍夫曼多群編碼法之中英文混合檔案壓縮方法

張眞誠與蔡文輝[11]於1991年依據霍夫曼多群編碼法(Huffman multigroup coding)完成了中英文混合檔案壓縮方法。中英文混合檔案所指的英文是泛指英

文字母、阿拉伯數字、英文標點符號等（暨非中文字）。壓縮過程中須分成兩回合來進行壓縮。霍夫曼多群編碼法所具有的特性如下：

(1)霍夫曼多群編碼法屬於統計式(account)編碼法。

(2)壓縮過程是需經過二回合處理，無論編碼或解碼都必須先行建構相同的霍夫曼樹，且建立霍夫曼樹時必須計算字元(byte)出現的次數。

(3)出現頻率最高的位元組，其霍夫曼碼最短；而出現頻率愈低者則霍夫曼碼愈長。

而霍夫曼多群編碼法所具有的限制及缺點如下：

(1)加入一筆新資料於原始檔案中，則所有霍夫曼樹都必須重新建構。

(2)若應用霍夫曼多群編碼法於網路傳輸時，首先必須先傳輸至接收端所需解壓縮時建構與傳送端相同三個唯一H、L及E霍夫曼樹所需的中序順序(inorder sequence)及後序順序(postorder sequence)資料所耗費的前置成本，使得壓縮效益大打折扣。顯而易見此方法不適合應用於網路傳輸。

2.2 藍波-立夫-衛曲編碼法之完全中文檔案壓縮方法

本節將介紹張眞誠與蔡文輝[12]陸續於1991年所提出另一個依據藍波-立夫-衛曲(Lempel-Ziv-Welch ,LZW)編碼法之完全中文檔案資料壓縮方法。所謂完全中文資料檔案乃指檔案中的資料完全皆爲中文字碼，暨中文字乃至於所有的阿拉伯數字、標點符號、英文字母及螢光幕上看不見的換行跳列符號（Carriage Return and Line Feed，CRLF）等皆視同中文字來處理。資料區分爲高位元組群、低位元組群和換行跳列群等三個子群(subgroup)。藍波-立夫-衛曲編碼法之完全中文檔案壓縮方法所具有的特性如下：

(1)屬於代換式字典(dictionary)編碼法。

(2)壓縮與還原過程同樣需要額外建置成本(表格)，但藍波-立夫-衛曲編碼法較之霍夫曼多群編碼法有相當小的壓縮還原所需表格的前置成本，因此較之霍夫曼多群編碼法有相當不錯的壓縮效益。

(3)依據藍波-立夫-衛曲編碼法之完全中文檔案資料壓縮方法所處理的檔案性質必須限制所有的阿拉伯數字、標點符號、英文符號及空白符號等均需完全佔用兩個位元組。

2.3 單一串列區域調整英文檔案壓縮方法

Bentley等人[6]於1986年所提出英文檔案資料壓縮的方法係屬於將固定長度(fixed length)轉換成不固定長度(variable length)輸出的檔案資料壓縮方法[3]。所處理的檔案資料被視爲空白(space)隔開的諸多文字串所組成的組合。這些被空白隔開的文字串(string)，在本文中皆視爲節點(node)。主要是利用單一串列(single list)來儲存壓縮處理過的文字串，並且對於曾經出現過的文字串，當其再次出現時，便以一個整數來取代，藉以縮短輸出碼長度。Bentley所提出的區域調整資料壓縮方法有下列限制及缺點：

(1)串列長度不宜無限制地增長。

(2)舊字串出現的頻率與出現的時間間距(interval)機會影響壓縮效果。

(3)在實際應用上可能文字串間相隔好幾個空白，超過一個以上的空白部份也當做一個空白文字串來處理，造成還原失眞(distortion)。

(4)將整數及實數等數字形態資料亦視同文字串來處理，由於這一類資料的重複出現次數低，影響壓縮效益。

3 雙區域調整串列對照編碼法(Double Adaptive List Correspondence，DALC)

本節提出改進陳信宏[14]於1995年所提出二維串列(two dimensional linked list)區域調整完全中文壓縮方法。所謂二維串列基本上可視爲座標分別有直、橫兩個基準線分別經由指標(pointer)自啓始節點由上而下、由左至右地單一方向串連(link)。在本節定位爲高位元組串列。主要將中文BIG-5碼之高位元組(high byte)與低位元組(low byte)予以分別開來存放，因此所有高位元組所存放的串列就稱之爲高位元組串列(high byte list)。而以高位元組串列上每個高位元組節點下各自領頭(lead)串連一組串列，其每個節點則用來存放文字中具有相同高位元組之低位元組，其二維串列結構詳如圖3.1。高、低位元組串列之新舊節點係依(move-to-front,MTF)原則排列。其實move-to-front原則相當於作業系統(operating system,OS)中記憶體管理策略(memory management strategy)中最近時間內置換最少被使用(least recently used replacement,LRU)的頁置換(page replacement)策略運用模式。可使出現頻率高的文字大部份可落在串列的前半部，而出現頻率低的文字便逐漸被推移至後半部，以期使高出現頻率的文字能獲得較小的位置整數。但是move-to-front原則較著重

於舊字元重複出現的時間間距(interval)，暨重複出現的時間間距愈長，則輸出前導位置整數碼長度就愈長。

因此本文爲期明白展示本文所提出彌補move-to-front原則缺乏對於最常出現的字元有最小編碼長度的掌握的限制，以及無法處理中英文混合檔案壓縮的不足與限制。另行建置一組與實施move-to-front原則串列一樣擁有相同內容且同步調整的串列。但此新建置的串列的字元排序調整係依照在壓縮過程中所有文字所累計的次數，由大至小，而同次數的群集合中元素的排列，仍依照move-to-front原則，稱之爲(group-move-to-front,GMTF)原則區域調整串列。藉以保證高次數出現率的文字能獲得最小的前導位置整數，二組調整串列在壓縮過程中相互各自同步調整，並自其中選擇最小位置整數碼的串列作爲輸出。如此一直在壓縮過程中充份相互發揮〝截長補短〞作用以有效縮減位置整數碼的長度，這就是本文在這一節中所提出的雙區域調整對照串列(Double Adaptive List Correspondence, DALC)架構原則。

因此，本文嘗試在中文字二維串列及英數字串列同時各建置另一組擁有相同內容的調整串列。而此串列裡各個字元的排列係分別依照壓縮過程中各中文字之高位元組及英數字ASCⅡ位元組被使用過累計的次數由大到小，並隨著資料內容的不斷地處理中而調整改變。換言之，存在著二組中文字二維串列及二組英數字串列。其中所不同的是，新建置的一組是以GMTF原則，係按照各字元出現次數總計由大到小，但同次數的字元元素群中則仍按move-to-front原則排列的串列。二組串列的內容相同，同步調整，目的是藉由二組串列進行比較哪一組串列被使用字元所得前導位置整數碼較小，便以此串列的整數碼輸出，藉此"截長補短"的作用，有效控制及縮短整數碼長度的成本。

至於如何縮短及控制串列長度成本問題。由於現行最普遍使用的BIG-5中文碼共包含13053個中文字，其中有5401個用字，而次常用則有7652個字，而5401個常用字的高位元組範圍落在A4至C6之間[2]。爲避免常用中文字之高位元組串列長度不受其他次常用或罕用中文字混合其間而增長，進而分散串列長度成本。嘗試另建置一組新中文字二維串列，用來儲放其他高位元組不屬於A4至C6範圍之間的中文字。

綜合上述，本文提出DALC架構及BIG-5中文編碼原則，原英數字及中文字二維MTF串列各增加了一組GMTF輔助串列，於是基本上產生了下列八個串列調整條件的狀況，其分類如下：

(1)載入新字元於英數字MTF及GMTF串列。

(2)字元已存在於英數字MTF串列中。

(3)字元已存在於英數字GMTF串列中。

(4)載入字元於中文字二維MTF及GMTF串列。

(5)已存在於常用中文字MTF串列中。

(6)已存在於常用中文字GMTF串列中。

(7)已存在於不常用中文字MTF串列中。

(8)已存在於不常用中文字GMTF串列中。

本文使用三個位元自"000"至"111"來分別表示上述八個串列調整條件狀況作爲固定長度辨識碼。在壓縮過程中除了辨識(pattern)前面佔用三個位元外，由二組MTF及GMTF串列相互比較而取其最小的值輸出。因此最後串列調整條件的表示及其輸出結果如下：

H表示高位元組；
L表示ASCⅡ值或低位元組；
prefix()表示前導位置整數編碼函數；
d_H表示高位元組的位置整數；
d_L表示低位元組的位置整數；

(1)載入新字元於英數字MTF及GMTF串列。

000 L

(2)字元已存在於英數字MTF串列中。

001 prefix(d_L)

(3)字元已存在於英數字GMTF串列中。

010 prefix(d_L)

(4)載入於常用或不常用中文字二維MTF及GMTF串列。

011 H L 或
011 prefix(d_H+1) L

(5)已存在於常用中文字MTF串列中。

100 prefix(d_H) prefix(d_L)

(6)已存在於常用中文字GMTF串列中。

101 prefix(d_H) prefix(d_L)

(7)已存在於不常用中文字MTF串列中。

110 prefix(d_H) prefix(d_L)

(8)已存在於不常用中文字GMTF串列中。

111 prefix(d_H) prefix(d_L)

本方法之流程圖如圖3.2所示。詳細說明本流程執行之過程及各串列內容的變化，請參考文獻[4]。

4 實證測試與分析比較

爲證實本文提出的雙區域調整串列對照編碼法的壓縮效益，並就霍夫曼中英文多群編碼法與二維串列區域調整壓縮方法兩種方法作比較，分別就全英文、全中文以及最後的中英文混合檔案資料，以彰顯本文所提出的壓縮方法的優點與可行性。

4.1 全英文檔案資料壓縮效益比較

在本節處理英文資料部份，所執行模擬比較的方法，除了二維串列區域調整編碼法及雙區域調整串列對照編碼法外，另包括在文獻探討中提到的霍夫曼中英文多群編碼法。選擇其作爲比較的原因，由於霍夫曼中英文多群編碼法具備可處理英文資料壓縮能力，可與之比較以表現本文提出的壓縮方法在英文壓縮的效益。所使用的測試資料爲本文檔案（text file），測試資料取自微軟視窗(Microsoft Windows)軟體中各項英文註解檔案，資料檔案大小範圍由10K 位元組至74K 位元組分成五組，效益分析與比較是以位元組(byte)爲單位。以其編碼後所需記憶空間大小作爲依據，模擬結果數據如表4.1。而圖4.1爲測試資料各方法所獲得的壓縮效益。圖4.2則爲各方法所展現的壓縮率(compression ratio ,CR)。所謂壓縮率定義係壓縮後所減少之位元數除以原始檔之位元數；壓縮率 =（ 原始檔案位元數 - 壓縮後檔案位元數）/ 原始檔案位元數。壓縮率愈高，表示省略的資料量愈多，因此壓縮效益越佳。

由表4.1之數據可以說明二維串列區域調整壓縮方法在處理英文資料壓縮效益顯然不佳，證實單純使用字母導向調整方式之英數字串列，僅應用MTF原則在英文資料壓縮的做法並不是一個可行的構想。本文所提出的雙區域調整串列對照編碼法在英文壓縮效益表現平均有50%的壓縮率。主要原因在於壓縮過程中，由於MTF及GMTF 兩組串列相互發揮輔助作用，確保高出現率的字元隨著壓縮過

程中一直控制在英數字GMTF串列的前端位置，經由掌握高出現率的字母及數字有較短小的前導位置整數編碼長度。證實本文提出DALC編碼方法也可經由字母導向調整方式而應用在英文檔案資料壓縮。

4.2 全中文檔案資料壓縮效益比較

在本文中所處理的中文碼係採用爲最爲普遍應用的 BIG-5碼，參與執行模擬比較的方法如同在4.1節所使用的三種方法。所測試的性質係採完全中文檔案資料，資料來源取自民國八十二年二月十五日「時報資訊」立即新聞稿，並將其中舉凡英文及數字等不屬BIG-5內碼資料皆予以剔除，再從中選擇五組作爲測試。檔案範圍由130K至920K位元組，模擬結果如表4.2。各方法之間壓縮結果比較圖（histograms）如圖4.3。壓縮率比較圖如圖4.4。由壓縮率比較圖所顯示壓縮率的表現，最佳爲雙區域調整串列對照編碼法，次爲霍夫曼多群編碼法，再次爲二維串列區域調整編碼法。由圖4.4中顯示，雙區域調整串列對照編碼法僅將二維串列區域調整邊碼法提升壓縮率約略10%，不若在英文資料壓縮率的差異程度。足可說明中文資料檔案之造字結構及分佈形態與英文資料檔案不同。中文是由固定雙位元組所組成，而英文則由固定的26個英文字母所組成。因此設計文字資料壓縮方法，必須考量其性質差異之處。

4.3 中英文混合檔案資料壓縮效益比較

最後在本節將探討中英文混合檔案資料壓縮效益的表現，在4.1 及 4.2節係個別針對全英文及全中文資料檔案在各種壓縮方法之比較。現就以中英文混合檔案作爲測試資料，資料來源除4.2節之時報新聞稿外，另外加上4.1節英文資料，同樣選擇五組資料作爲測試。檔案範圍自140K 至990K 位元組，同樣參考4.2 節所使用壓縮效益比較標準。各方法在中英混合檔案壓縮模擬結果如表4.3。壓縮結果檔案大小比較圖如圖4.5。壓縮率比較圖如圖4.6。由壓縮率比較圖顯示，雙區域調整串列對照與霍夫曼多群編碼法同時提昇2%壓縮率，相對地二維串列區域調整編碼法則減少3%壓縮率，這說明二維串列區域調整編碼法在英文資料方面壓縮的表現影響到中英文混合檔案壓縮效益，壓縮率比較圖仍顯示雙區域調整串列對照編碼法的壓縮率最佳，其餘依次爲霍夫曼多群編碼法、二維串列區域調整編碼法。

綜合以上分別就英文、中文及中英文混合檔案部份測試結果，證實雙區域調整串列對照編碼法由於增加GMTF原則串列分別掌握英文字元與中文BIG-5碼高位元組出現次數由高至低的排序，正與霍夫曼多群編碼法之中英文混合檔案壓縮方法中最高出現率字元擁有最短(小)碼長度原則相同，並且又搭配了最近時間內最久未使用的被置換之MTF原則，在壓縮過程中充份發揮相互「截長補短」作用。不僅提昇陳信宏所提出二維串列區域調整編碼法之壓縮效益外，並解決了無法處理中英文混合檔案壓縮的問題，增進實際應用的價值。

5 結論與未來發展方向

文字檔案資料壓縮目的在於減少文字資料中重複資訊，不論在儲存空間或網路傳輸成本皆有其探討的重要性。網際網路與人類生活緊密結合，更應突破純粹英文或中文資料檔案壓縮方法所欠缺實用價值或不利網路傳輸等瓶頸。本文選擇區域調整串列結構作爲中英文字混合資料檔案壓縮方面探討對象，原因在於區域調整串列結構有結構簡單、單一回合處理、無其它壓縮方法中壓縮還原所需額外前置成本、壓縮還原過程簡單快速，且應用網路傳輸時，傳送與接收端可立即同步進行壓縮及還原等諸多優點。不同於以往諸多壓縮方法乃先將傳輸資料予以壓縮後，再將壓縮還原所需前置成本先送至接收端後，再將壓縮檔案予以解壓還原等多回合過程所耗費的成本，而大大降低其優越性與實用價值。

本文針對上述目標，修正陳信宏所提出完全中文資料檔案壓縮方法中二維串列區域調整結構，在第四章分別就英文、中文及中英文混合部份實證分析其壓縮效益。由測試數據可以清楚了解本文提出的DALC編碼方法經由同時掌握壓縮過程中「最常被使用」及「最近時間內最常被使用」兩項原則，分別在英文、中文及中英文混合資料壓縮測試結果都獲得相當程度的效果。尤其在英文資料壓縮部份，更提供了另一種不同於Bentley原先所使用單字導向調整做法，證實字母導向調整做法經由本文提出DALC編碼方法，同樣可以獲致壓縮效益的另一種探討方向。

未來因應中文標準碼之頒行，屆時在中英文混合檔案資料壓縮領域同時也提供了一個值得深入研究探討的課題，也是本文繼續探討的方向與目標。

參考文獻

〔1〕施威銘（民七七),「 C語言實務 」, 旗標出版有限公司,台北。

〔2〕倚天資訊（民七九),「倚天中文系統使用手冊」,倚天資訊股份有限公司,台北。

〔3〕張眞誠（民八三),「資料壓縮原理與實務」,松崗圖書公司,台北。

〔4〕朱賢武（民八四),「中英文字檔案區域調整資料壓縮方法之設計與製作」,國防管理學院資源管理研究所碩士論文。

〔5〕Abrahamson, David M.:"An adaptive dependency source model for data compression," Communications of the ACM, Vol. 32, No. 1, pp. 77-83.(1989).

〔6〕Bentley, J. L., Sleator, D. D., Tarjan, R. E. and Wei, V. K.:"A Locally Adaptive Data Compression Scheme," Communications of the ACM, Vol. 29, No. 4, pp. 320-330.(1986).

〔7〕Bell, Timothy C.:"Better OPM/L text compression," IEEE Trans. Communications, Vol. COM-34, No. 12, pp. 1176-1182.(1986).

〔8〕Bell, T., Witten, I. H., and Cleary, J. G.:"Modeling for text compression," ACM Computing Surveys, Vol. 21, No. 4, pp.557-591.(1989).

〔9〕Bailey, R. L. and Mukkamala, R.:"Pipelining data compression algorithms," The Computer Journal, Vol. 33, No. 4, pp. 308-313.(1990).

〔10〕Chang, H. K. C. and Chen, S. H.:"A new locally adaptive data compression scheme using multilist structure," The Computer Journal, Vol. 36, No. 6, pp. 570-578.(1993).

〔11〕Chang, C. C. and Tsai, W. H.:" A data compression scheme for Chinese-English characters," Computer Processing of Chinese & Oriental Languages, Vol. 5, No. 2, pp. 154-182.(1991).

〔12〕Chang, C. C. and Tsai, W. H.:" A compression scheme based upon Lempel-Ziv method for chinese texts," Journal of Computers, Vol. 3, No. 2, pp. 1-10.(1991).

[13] Chang, H. K. C. and Chen, S. H.:"Extended predictive data coding scheme for Chinese text Files," Computer Processing of Chinese & Oriental Languages, Vol. 7, No. 2, pp. 154-182.(1993).

[14] Chang, H. K. C. and Chen, S. H.:" A locally adaptive coding scheme for Chinese text files," Journal of Information Science and Engineering, Vol. 11, No. 1, pp. 51-71.(1995).

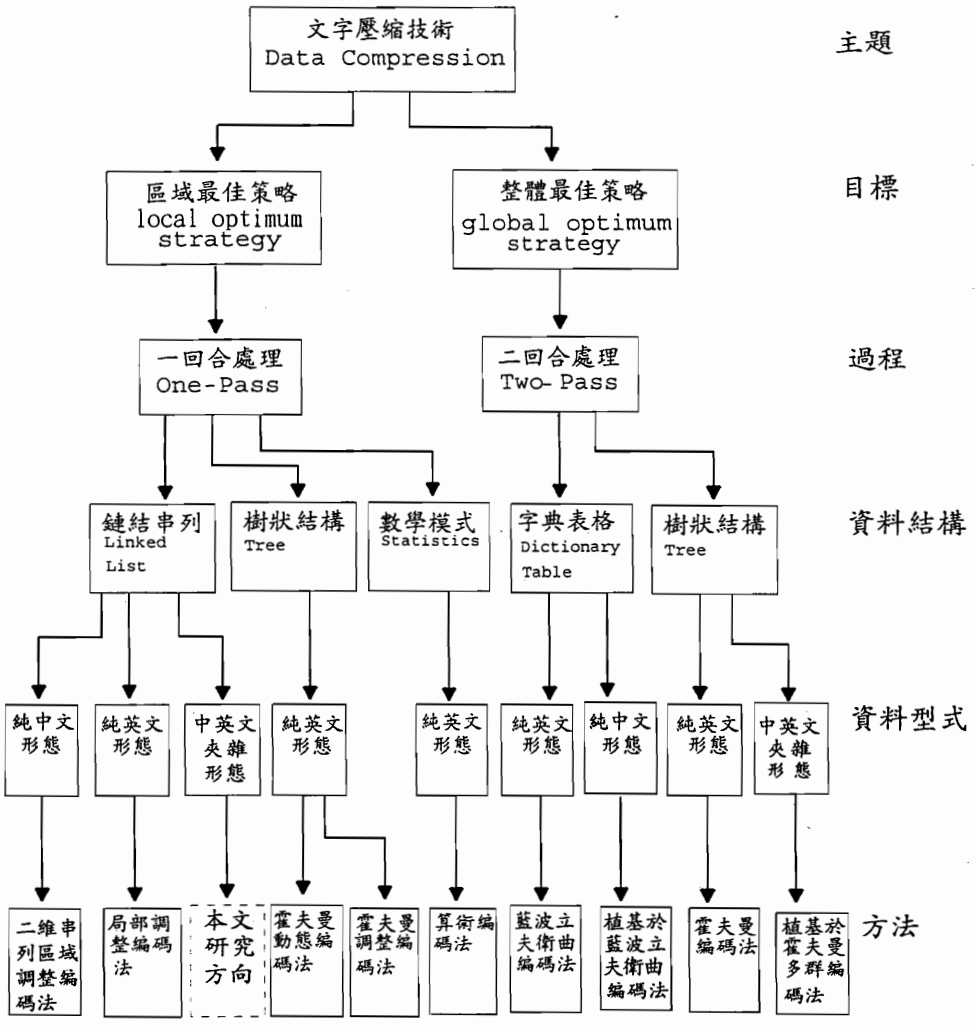


圖1.1　文字資料壓縮分類圖

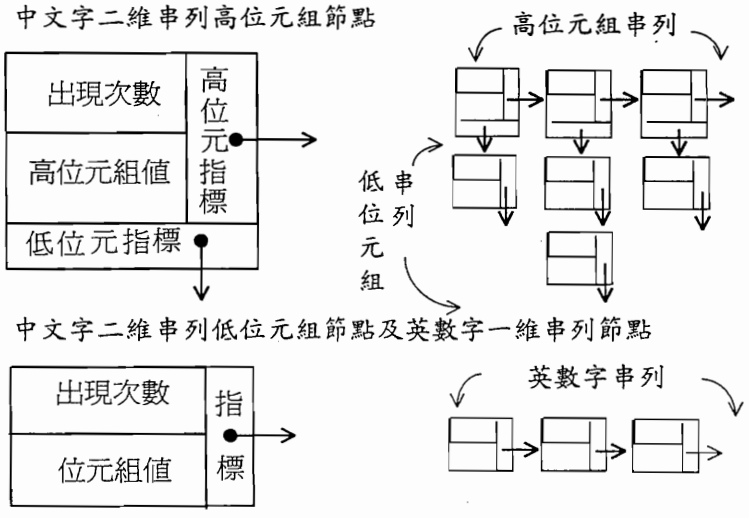
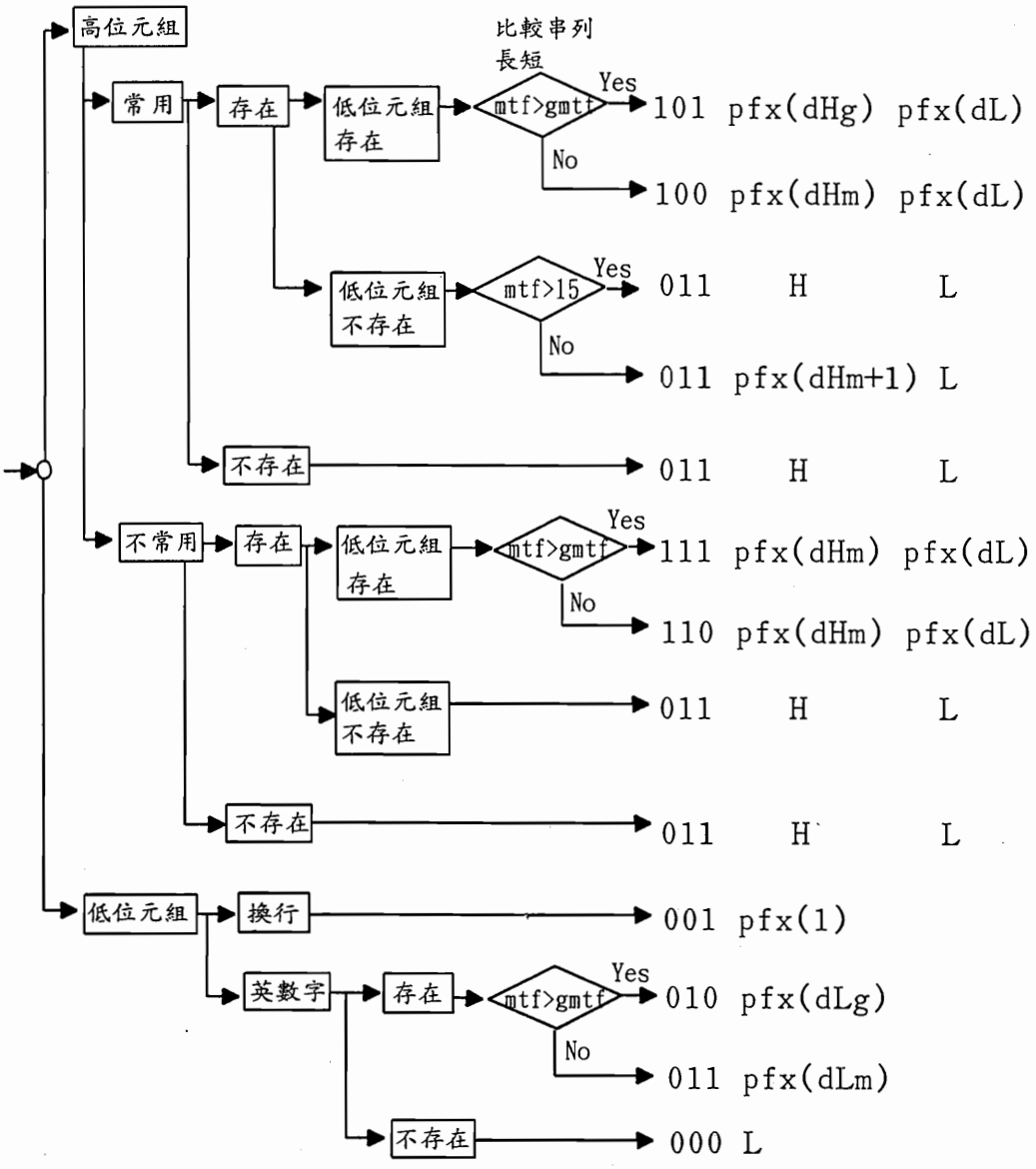


圖3.1 中英文字串列及節點結構



Note:

pfx():prefix code
dHg:integer position of high byte for gmtf list.
dLg:integer position of low byte for gmtf list.
dHm:integer position of high byte for mtf list.
dLm:integer position of low byte for mtf list.
H: high byte
L: low byte

圖3.2 本方法之流程圖

表4.1 全英文檔案各方法壓縮結果比較表

測試檔案長度（bytes）	11,614	20,950	32,947	43,431	76,379
二維串列區域調整編碼法	14,066	25,364	39,901	52,592	92,490
霍夫曼中英文多群編碼法	6,866	12,131	19,110	25,626	43,536
雙區域調整串列	5,796	10,412	16,347	21,531	37,828

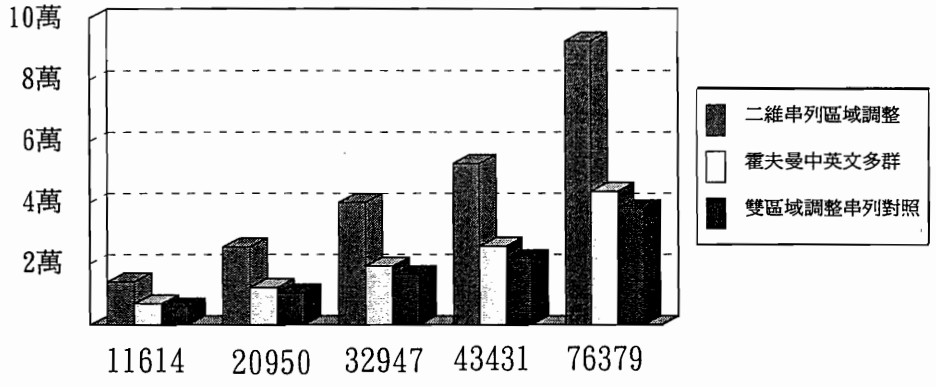


圖4.1 全英文檔案各方法壓縮結果比較圖

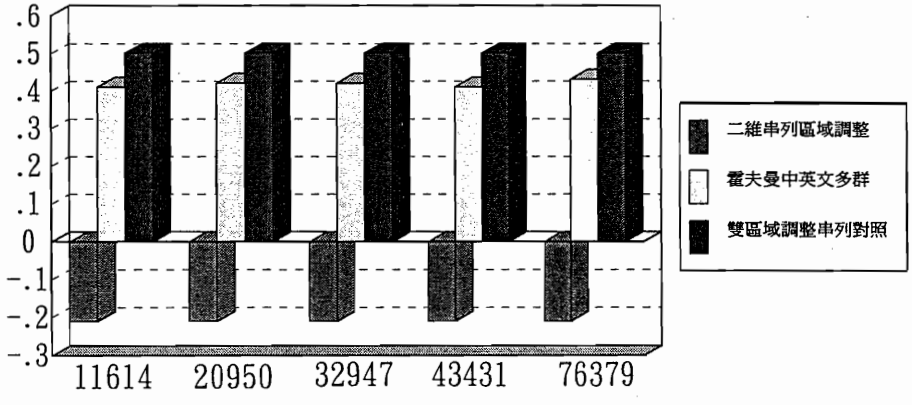


圖4.2 全英文檔案各方法壓縮率比較圖

表4.2 全中文檔案各方法壓縮結果比較表

測試檔案長度（bytes）	134,885	269,771	404,656	539,541	944,196
二維串列區域調整編碼法	108,309	216,805	325,300	433,759	759,279
霍夫曼中英文多群編碼法	101,165	199,641	295,344	393,871	689,262
雙區域調整串列對照編碼法	95,773	191,548	287,319	388,466	679,781

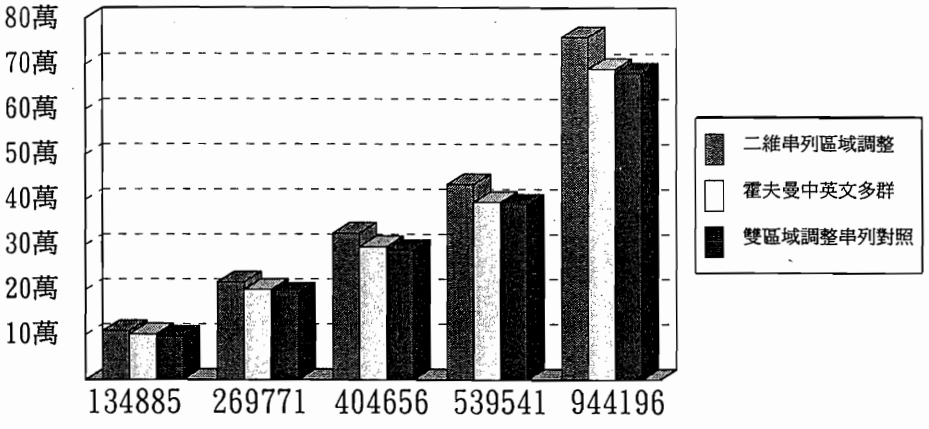


圖4.3 全中文檔案各方法壓縮結果比較圖

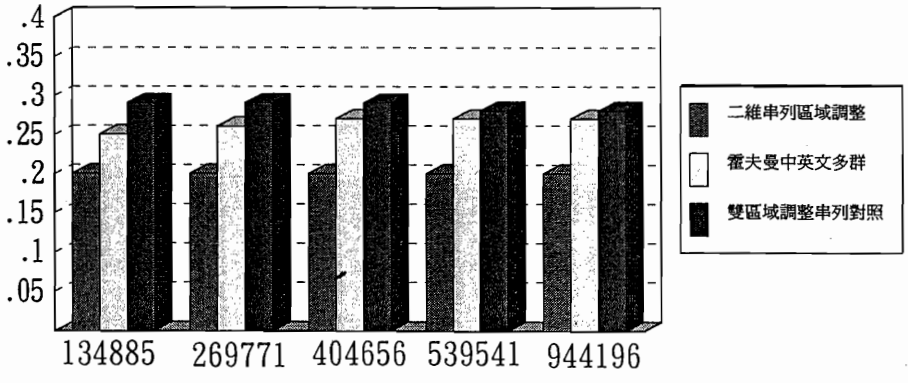


圖4.4 全中文檔案各方法壓縮率比較圖

表4.3 中英文混合檔案各方法壓縮結果比較表

測試檔案長度 (bytes)	146,499	290,721	437,603	582,972	1,020,575
二維串列區域調整編碼法	121,596	241,323	367,599	489,687	857,322
霍夫曼中英文多群編碼法	106,950	209,323	315,161	419,808	724,616
雙區域調整串	101,088	200,599	301,940	402,251	714,404

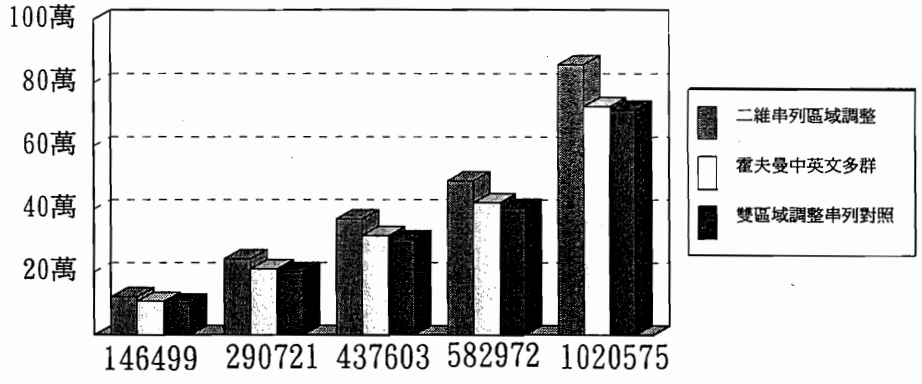


圖4.5 中英文混合檔案各方法壓縮結果比較圖

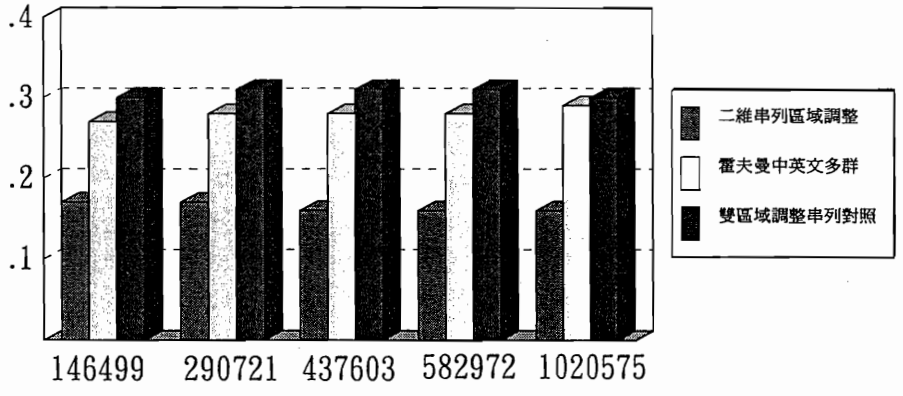


圖4.6 中英文混合檔案各方法壓縮率比較圖

A Preliminary Study of Disambiguating VO- and VN-Constructions Using Selection Preferences

Kok-Wee Gan

Department of Computer Science

Hong Kong University of Science and Technology

Clear Water Bay, Kowloon, Hong Kong

E-mail: gankw@cs.ust.hk

Abstract

In Chinese, a verb followed by a noun can be analyzed as either a verb-object (VO) construction or a verb-noun (VN) construction. In the latter, the verb acts as a modifier of the noun. This paper describes how selection preferences can be used to determine whether a Verb+Noun pair (V+N) is a VO-construction or a VN-construction. The approach also takes syntactic factors into consideration. These factors are expresssed in terms of likelihood measures of the tendency of verbs and nouns functioning as VN- and VO-constructions. The preliminary result based on 17 bi-syllabic, transitive verbs with a total of 880 V+N pairs is 88.4%.

1 Introduction

In Chinese, a verb followed by a noun can be analyzed as either a VO-construction or a VN-construction. For example, 訓練口才 *xun4lian4 kou3cai2* 'train oratorical skills' is a VO-construction, where 口才 *kou3cai2* 'oratorical skills' is the object of the verb 訓練

xun4lian4 'train'. However, 訓練方法 *xun4lian4 fang1fa3* 'training methods' is a VN-construction, with the verb 訓練 *xun4lian4* 'training' acting as the modifier of the noun 方法 *fang1fa3* 'methods'. There is no inflections in Chinese to distinguish between these two usages of verbs. This ambiguity poses a problem for a Chinese parser. In this paper, we describe an approach to automatically determine whether a V+N pair is a VO- or VN-construction using selection preferences.

Selection preferences cast selection restrictions in probabilistic terms. Selection restrictions of a predicate are specifications of the necessary and sufficient condition for a semantically acceptable argument. Such conditions are represented as boolean functions of semantic markers. Selection preferences, in contrast, represent such conditions as real-value functions. Such conditions are usually derived from corpora. For example, semantically acceptable arguments which can be the object of the predicate 吃 *chi1* 'eat' tend to be *physical*, *animate*, *edible*, and so forth. Measures of how likely the object of 吃 *chi1* 'eat' is *physical*, *animate*, etc., constitute the selection preferences of 吃 *chi1* 'eat'. In Section 2, we describe an information-theoretic approach of determining the selection preferences of a predicate [7]. We will explain how we make use of selection preferences to disambiguate VN- and VO-constructions in Section 3. The experimental results will be reported in Section 4. A comparison with related work will be covered in Section 5.

2 Determination of Selection Preferences

We define the selection preferences of a predicate over a taxonomy of 116 conceptual classes [1]. The taxonomy is primarily organized into a hyponymy (IS-A) hierarchy as shown in the appendix. Some of the conceptual classes, for example, *edible, flowers, fruits, holes, human, literature,* and *locative,* are features that serve to link together concepts which are otherwise not related in the hierarchy. These concepts are listed in the

appendix with a plus operator in front.

The information-theoretic approach of determining selection preferences as proposed in [7] is adopted and summarized as follows.

Let P be a random variable ranging over the set $\{p_1, p_2, ..., p_m\}$ of predicates. Let C be another random variable ranging over the set $\{c_1, c_2, ..., c_k\}$ of conceptual classes in a taxonomy. C is related to P by a particular predicate-argument relationship, such as verb-object, or adjective-noun. The preference of a particular predicate p_i is defined as the effect it has on the distribution of C. Let the distribution of C regardless of any particular predicate be the prior distribution, $p(c)$, and let the posterior distribution $p(c|p_i)$ be the distribution of C given the predicate p_i. The change between the prior distribution $p(c)$ and the posterior distribution $p(c|p_i)$ constitutes the selection preference strength of the predicate p_i, which can be measured by relative entropy. In information theory, the relative entropy of two probability distributions p and q is defined as:

$$D(p||q) = \sum_x p(x)\log\frac{p(x)}{q(x)} \tag{1}$$

Replacing p with $p(c|p_i)$, q with $p(c)$, the **selection preference strength** of the predicate p_i is:

$$D(p(c|p_i)||p(c)) = \sum_c p(c|p_i)\log\frac{p(c|p_i)}{p(c)} \tag{2}$$

The **selectional association** of a predicate, $A(p_i, c_j)$, for a particular argument class c_j is defined as:

$$A(p_i, c_j) = \frac{1}{\eta_i} p(c_j|p_i) \log \frac{p(c_j|p_i)}{p(c_j)} \quad (3)$$

where η_i is the selection preference strength of the predicate p_i as shown in equation (2).

The **selection preference** of a predicate p_i is a vector of selectional associations between p_i and a list of conceptual classes in a taxonomy. The statistical technique of maximum likelihood estimation (MLE) is used in deriving the prior distribution $p(c)$ and the posterior distribution $p(c|p_i)$. For a particular conceptual class c_j, $p(c_j)$ is derived by:

$$\hat{p}_{MLE}(c_j) = \frac{freq(c_j)}{N} \quad (4)$$

where

$$N = \sum_{j=1}^{116} freq(c_j) \quad (5)$$

and $freq(c_j)$ is the frequency of the conceptual class c_j, which is defined as

$$freq(c_j) = \sum_{w \in words(c_j)} \frac{freq(w)}{|classes(w)|} \quad (6)$$

$words(c_j)$ is the set of words that belong to the conceptual class c_j, $|classes(w)|$ is the number of conceptual classes of a word w, and $freq(w)$ is the frequency of w.

The conditional probability of a particular conceptual class c_j given a predicate p_i is estimated from:

$$\hat{p}_{MLE}(c_j|p_i)=\frac{\text{freq}(c_j,p_i)}{N} \tag{7}$$

where

$$N=\sum_{j=1}^{116}\text{freq}(c_j,p_i) \tag{8}$$

and

$$\text{freq}(c_j,p_i)=\sum_{w\in\text{words}(c_j)}\frac{\text{freq}(w,p_i)}{|\text{classes}(w)|} \tag{9}$$

words(c_j) is the set of words that belong to the conceptual class c_j, freq(w, p_i) is the co-occurrence frequency of the word w and the predicate p_i,[1] and |classes(w)| is the number of conceptual classes of w.

3 Disambiguation of VO- and VN-constructions

According to [2], ambiguities in V+N pairs are most difficult in transitive verbs. We therefore focus on disambiguating $V_{transitive}$+N pairs; in particular, we focus on bi-syllabic transitive verbs. We extracted a total of 880 $V_{transitive}$+N pairs from the Sinica corpus Version 1.0 [4]. 708 word pairs were used for training while the remaining 172 pairs were used for testing. The list of verbs covered are: 訓練 *xun3lian4* 'train', 表演 *biao2yan3* 'perform', 治療 *zhi4liao2* 'cure', 表達 *biao3da2* 'express', 學習 *xue2xi2* 'learn', 選擇 *xuan3ze2* 'choose', 生產 *sheng1can3* 'produce', 解決 *jie3jue2* 'solve', 教育

[1] In our experiment, the window size is set to 5. That is, a word w is regarded as co-occurring with a predicate p_i if it is not more than 5 words away from the predicate.

jiao4yu4 'educate', 發展 *fa1zhan3* 'develop', 處理 *chu2li3* 'handle', 參加 *can1jia1* 'participate', 管理 *guan2li3* 'manage', 建設 *jian4she4* 'build', 進行 *jin4xing2* 'go on', 使用 *shi3yong4* 'utilize', and 影響 *ying2xiang3* 'influence'. We manually separated the training set into VO-pairs and VN-pairs, from which we derived the selection preferences of each verb in a VO-relation and a VN-relation. The formulae used in the derivation have been covered in Section 2. These vectors of selection preferences ($Pref_{VO}$ and $Pref_{VN}$,) are later used to determine whether a V+N pair is a VO-construction or a VN-construction. We will illustrate this step with an example.

Figure 2 and 3 show the selection profiles of the verb 影響 *ying2xiang3* 'influence' in a VO- and VN-relations respectively.

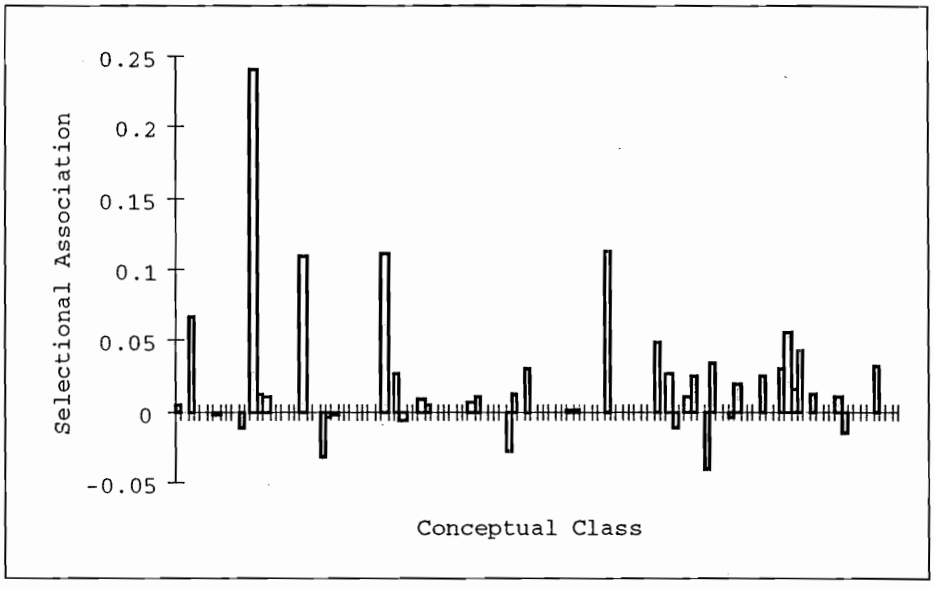


Figure 2. Selection profile of 影響 *ying2xiang3* 'influence' in a VO-relation

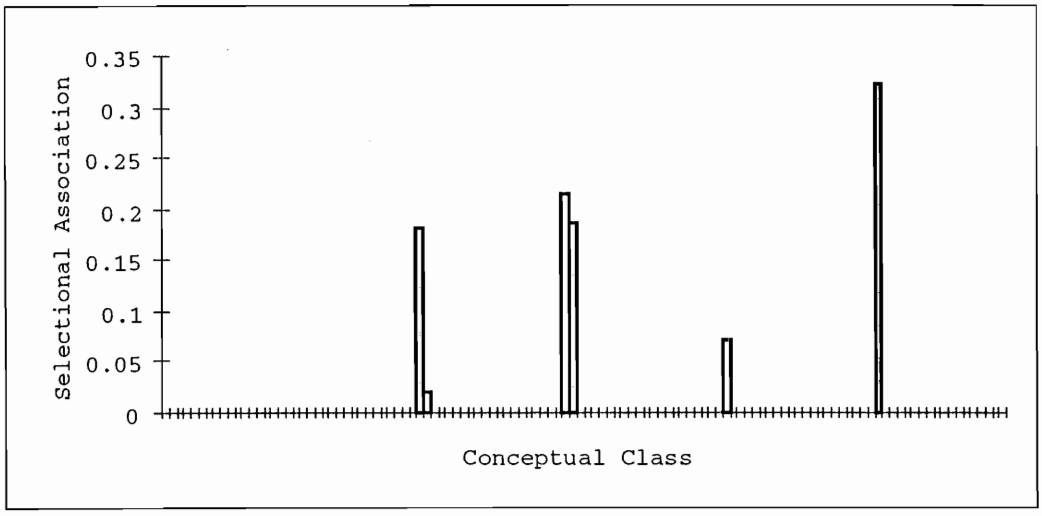


Figure 3. Selection profile of 影響 *ying2xiang3* 'influence' in a VN-relation

The selection profile of the verb 影響 *ying2xiang3* 'influence' in Figure 2 indicates that arguments of a wide range of conceptual classes could be the object of 影響 *ying2xiang3* 'influence'. On the other hand, the verb 影響 *ying2xiang3* 'influence' is modifiers of nouns of a much restricted set of conceptual classes (see Figure 3). During testing, say when 影響十成績 is presented, we first derive a binary vector Q that represents the conceptual classes of 成績 *cheng2ji1* 'results'. The vector is defined over the entire set of conceptual classes in a taxonomy (a total of 116 classes in our case). All conceptual classes that 成績 *cheng2ji1* 'results' belongs to will be assigned the value 1, with all the others 0. A **similarity measure** between Q and a preference vector *Pref* is defined as follows.

$$S = \frac{|Pref|}{M} Pref \cdot Q + T_V \cdot T_N \tag{10}$$

where $|Pref|$ is the total number of conceptual classes in *Pref* that have non-zero selectional association, and M is the total number of conceptual classes in a taxonomy. For the

similarity measure of a verb V in a VO-construction (S_{VO}), T_V is a probabilistic measure of the tendency that the verb appears in VO-construction, and T_N is a probabilistic measure of the tendency that a noun appears in VO-construction. The former is called the **VO-tendency of a verb** (T_V^{VO}) while the latter is called the **VO-tendency of a noun** (T_N^{VO}). The VO-tendency of a verb V is estimated from the percentage of $V+N_i$ pairs in a corpus that are VO-construction. The VO-tendency of a noun N is estimated from the percentage of V_i+N pairs where N is the object of V_i. The **VN-tendency of a verb** (T_V^{VN}) and the **VN-tendency of a noun** (T_N^{VN}) in determining the similarity measure of a verb in a VN-construction (S_{VN}) can be derived from the following equations.

$$T_V^{VN} = 1 - T_V^{VO} \tag{11}$$

$$T_N^{VN} = 1 - T_N^{VO} \tag{12}$$

Our experimental data of the similarity scores of 成績 *cheng2ji1* 'results' with respect to the verb 影響 *ying2xiang3* 'influence' under VO-construction (S_{VO}) and VN-construction (S_{VN}) are 0.32 and 0.063 respectively. Since S_{VO} is greater than S_{VN}, we conclude that 影響十成績 is a VO-construction.

The ratio |Pref|/M in equation (10) is a weight used to implement the preference for a construction which has a selection profile that covers a wide range of conceptual classes. In the example of the verb 影響 *ying2xiang3* 'influence', the selection profile of the VO-construction is more spread out than that of the VN-construction. When a new pair of 影響 +N is encountered, assigning it as a VO-construction would have a higher chance of

being correct. This heuristics is incorporated into the ratio. The additive term $T_v \cdot T_N$ incorporates the heuristics as observed by [2]: (i) When both the verb and noun in a V+N pair have a high VO-tendency, it is more likely that this is a VO-construction. Conversely, when both the verb and noun have a high VN-tendency, they are more likely to form a VN-construction. When the tendency of the noun and verb contradict each other, as well as when neither the noun nor the verb has a clear VN- or VO-tendency, selection preferences play a decisive role.

4 Experimental Results and Discussion

The experimental procedure is summarized as follows:

- extract from the Sinica corpus all sentences[2] which contain one of the 17 bi-syllabic, transitive verbs;
- extract semi-automatically all V+N pairs from the sentences found in step 1;
- manually split the V+N pairs into two groups: VO-pairs and VN-pairs;
- derive T_V^{VO}, T_N^{VO}, T_V^{VN} and T_N^{VN} in the manner as described in Section 3;
- use 80% of the VO-pairs and VN-pairs as training data to derive the selection preferences of each verb (see Section 2);
- use the remainding 20% to evaluate the performance of the proposed approach.

The similarity measure in equation (10) is used to determine whether a given V+N pair is a VO- or VN-construction. The decision is as follows:

[2] A sentence is defined as a sequence of characters delimited by punctuation marks.

if $S_{VO} >= S_{VN}$

then V+N is a VO-construction

else V+N is a VN-construction

An average recognition rate of 88.4% is obtained in our experiment. A detailed break down is shown in Table 1.

Table 1. Recognition Rate of Each Verb

Verbs	Recognition Rate(%)
訓練 *xun3lian4* 'train'	72.7
表演 *biao2yan3* 'perform'	100
治療 *zhi4liao2* 'cure'	100
表達 *biao3da2* 'express'	80
學習 *xue2xi2* 'learn'	71.4
選擇 *xuan3ze2* 'choose'	100
生產 *sheng1can3* 'produce'	66.7
解決 *jie3jue2* 'solve'	66.7
教育 *jiao4yu4* 'educate'	85.7
發展 *fa1zhan3* 'develop'	91.7
處理 *chu2li3* 'handle'	100
參加 *can1jia1* 'participate'	94.1
管理 *guan2li3* 'manage'	100
建設 *jian4she4* 'build'	90
進行 *jin4xing2* 'go on'	90
使用 *shi3yong4* 'utilize',	82.4
影響 *ying2xiang3* 'influence'	95.5
Average	88.4

The VN-tendency of the 17 verbs are shown in increasing order in Table 2. The same table can also be interpreted as displaying the VO-tendency of these verbs in decreasing order.

Table 2. VN-tendency (in %) of the 17 Verbs

影響	8
選擇	18
處理	28
進行	29
使用	31
學習	31
表達	32
生產	35
發展	41
管理	43
參加	44
解決	46
建設	67
訓練	68
治療	68
表演	79
教育	84

Three factors that have impacts on the derivation of selection preferences are:

- **Word Boundary Accuracy** The accuracy of word boundaries in the Sinica corpus will directly influence the derivation of selection preferences. First, a sentence that contains a predicate p_i will be missed if the predicate is incorrectly segmented. Second, any error in the word boundaries of words that co-occur with

the predicate will affect the estimate $\hat{p}_{MLE}(c_j|p_i)$ in equation (7). The Sinica corpus uses human labor to post-edit on the output of an automatic parts-of-speech tagger [6]. The post-editing work includes correcting errors in word boundaries and parts-of-speech [5]. In terms of word boundary accuracy, it is one of the best resources available currently.

- **V+N Pairs Extraction** Statistical techniques in general face the problem of insufficient data. Hence,the larger a test set is, the better are the statistical estimates. In this work, we used a semi-automatic approach to extract V+N pairs from the Sinica corpus. From all sentences that contain a particular verb, say 影響 *ying2xiang3* 'influence', we extract only those V+N pairs that are of these two patterns: V N+ and V N+ 的 N+.[3] In sentences (1) to (3), the followings: 影響十別人, 影響十世界, and 影響十形式 were extracted. We then manually went through all the extracted pairs to remove the erroneous ones and to decide whether they are VN- or VO-constructions.

(1) 即 卡耐基 的 那 本 《 如何
ji4 *ka3nai4ji1* *de* *nai4* *ben3* *ru2he2*
that is Carnegie DE[4] that CL[5] how
影響 別人 》。
ying2xiang3 *bie2ren2*
influence others
That is, the book "How to influence others" written by Carnegie.

[3] N+ refers to one or more nouns.

[4] DE refers to the structure word 的 de.

[5] CL stands for a classifier.

(2)

這些	都	不	足以	構成
zhei4xie1	*dou1*	*bu4*	*zu2yi3*	*gou4cheng2*
these	all	not	sufficient	constitute

衡量	一 位	影響	全	世界，
heng2liang4	*yi1 wei4*	*ying2xiang3*	*quan2*	*shi4jie4*
measure	one CL	influence	whole	world

These are not sufficient to measure a person who has influenced the whole world.

(3)

目的	在	了解	社會	因素
mu4di4	*zai4*	*liao2jie3*	*she4hui4*	*ying1shu4*
goal	at	understand	society	factor

如何	影響	語言	的	形式。
ru2he2	*ying2xiang3*	*yu3yan2*	*de*	*xing2shi4*
how	influence	language	DE	form

The goal is to understand how social factors influence language form.

Our simplistic approach inevitably leaves out many V+N pairs. Sentences (4) to (6) are some examples where this has happened. The object of 影響 *ying2xiang3* 'influence' in (4) is 日本 *ri4ben3* 'Japan', which appears in a passive sentence structure. Our approach missed this. In (5), the object has been wrongly identified as 部分 *bu4fen4* 'part' instead of 張力 *zhang1li4* 'tension'. In (6), 品質 *pin3zhi4* 'quality' instead of 生活 *sheng1huo2* 'living' should be the object of 影響 *ying2xiang3* 'influence'.

(4) 日本 受 儒家 影響。

ri4ben3 *shou4* *ru2jia1* *ying2xiang3*

Japan receive confucius influence

Japan is influenced by confucius thinking.

(5) 減少 不 影響 張力 的

ian3shao3 *bu4* *ying2xiang3* *zhang1li4* *de*

reduce not influence tension DE

部分。

bu4fen4

part

To reduce those parts which do not influence tension.

(6) 甚至 影響 生活 與 工作 的

shen3zhi4 *ying2xiang3* *sheng1huo2* *yu3* *gong1zuo4* *de*

even influence life and work DE

品質。

pin3zhi4

quality

even influences the quality of life and work

- **Polysemy Issue** The Sinica corpus is not sense-disambiguated. Therefore, the selectional behavior of multiple senses of a verb is conflated. This is not necessarily a problem, as the resulting selection profile of the verb has distinct groupings. In determining the similarity measure using equation (10), only groupings that match the conceptual classes of a noun are considered.

The issue of polysemy also occurs in nouns. Our conceptual classes of nouns were based on the CKIP dictionary [3]. This dictionary has altogether 78,410 lexical entries, out of which 34,984 are nouns. The average number of senses per noun is 1.0115. Thus, most of the nouns in the CKIP dictionary have only one sense.

5 Comparison With Related Work

The approach described in [2] uses the following algorithm to decide whether a V+N pair is a VN- or VO-construction.

```
if    V is intransitive/pseudotransitive
then  V+N is a VN-construction
else  if V can be nominalized
        then  if V has a strong VN-tendency
                then  if  N is not an individuated noun[6]
                        then  V+N is a VN-construction
                        else  V+N is a VO-construction
                else  V+N is a VO-construction
        else  V+N is a VO-construction
```

Our work replaced the following steps of the algorithm by an information-theoretic approach as described in Sections 2 and 3.

[6]The followings are considered as individuated nouns: proper nouns, count nouns, location nouns, and pronouns.

if V has a strong VN-tendency

then if N is not an individuated noun

then V+N is a VN-construction

else V+N is a VO-construction

else V+N is a VO-construction

It is not clear in [2] what threshold is used to decide whether a verb has a strong VN-tendency. The paper also did not explicitly state the performance of the algorithm. Thus, a quantitative comparison is not possible. Qualitatively, the approach in [2] uses the part-of-speech of nouns (i.e., whether a noun is an individuated noun) to decide whether a V+N pair is a VN- or VO-construction. Selection preferences in our approach in equation (10) is essentially a measure of the semantic compatibility between a verb and a noun. Our approach, in addition, has also incorparated syntactic factors. In [2], it is observed that individuated nouns usually have a strong VO-tendency while non-individuated nouns are more likely to have a strong VN-tendency. This insight on the syntactic behavior of nouns in V+N pairs is implicitly incorporated in the term T_N in equation (10). Individuated nouns will have a high VO-tendency value (T_N^{VO}) while non-individuated nouns will have a high VN-tendency value (T_N^{VN}). An advantage of our approach in comparison with the hard-and-fast rules in [2] is that exceptions to the rules can be handled better. For example, 影響十蹟象 will be identified as a VO-construction in [2] since the verb 影響 *ying2xiang3* 'influence' has an extremely weak VN-tendency (0.08 as shown in Table 2). The correct relation in this example should be a VN-construction, which is correctly identified in our approach because we consider not only the tendency of the verb, but also the tendency of the noun involved, as well as the selection preferences of the verb.

6 Conclusions

We have described in this paper a new approach to disambiguate VN- and VO-constructions using selection preferences. In addition to this semantic factor, our approach has also incorporated likelihood measures of the tendency of verbs and nouns functioning in VN- and VO-constructions. These measures are implicit syntactic factors. Our preliminary results based on 17 bi-syllabic, transitive verbs with a total of 880 V+N pairs is 88.4%. Our next goal is to evaluate this approach with a larger set of data.

Acknowledgments

We would like to express our thanks to the Academia Sinica, Taiwan, for giving us the permission to use the Sinica corpus and the CKIP dictionary.

References

[1] 莫若萍 (1992) 一個適用於剖析漢語的概念結構，中央研究院資訊科學研究所技術報告 92-04。

[2] 陳克健，洪偉美 (1995) 中文裡「動一名」述賓結構及「動一名」偏正結構的分析，第八屆計算語言學研討會論文集，1-13。

[3] 詞庫小組 (1993) 中文詞類分析（三版），中央研究院資訊科學研究所技術報告 93-05。

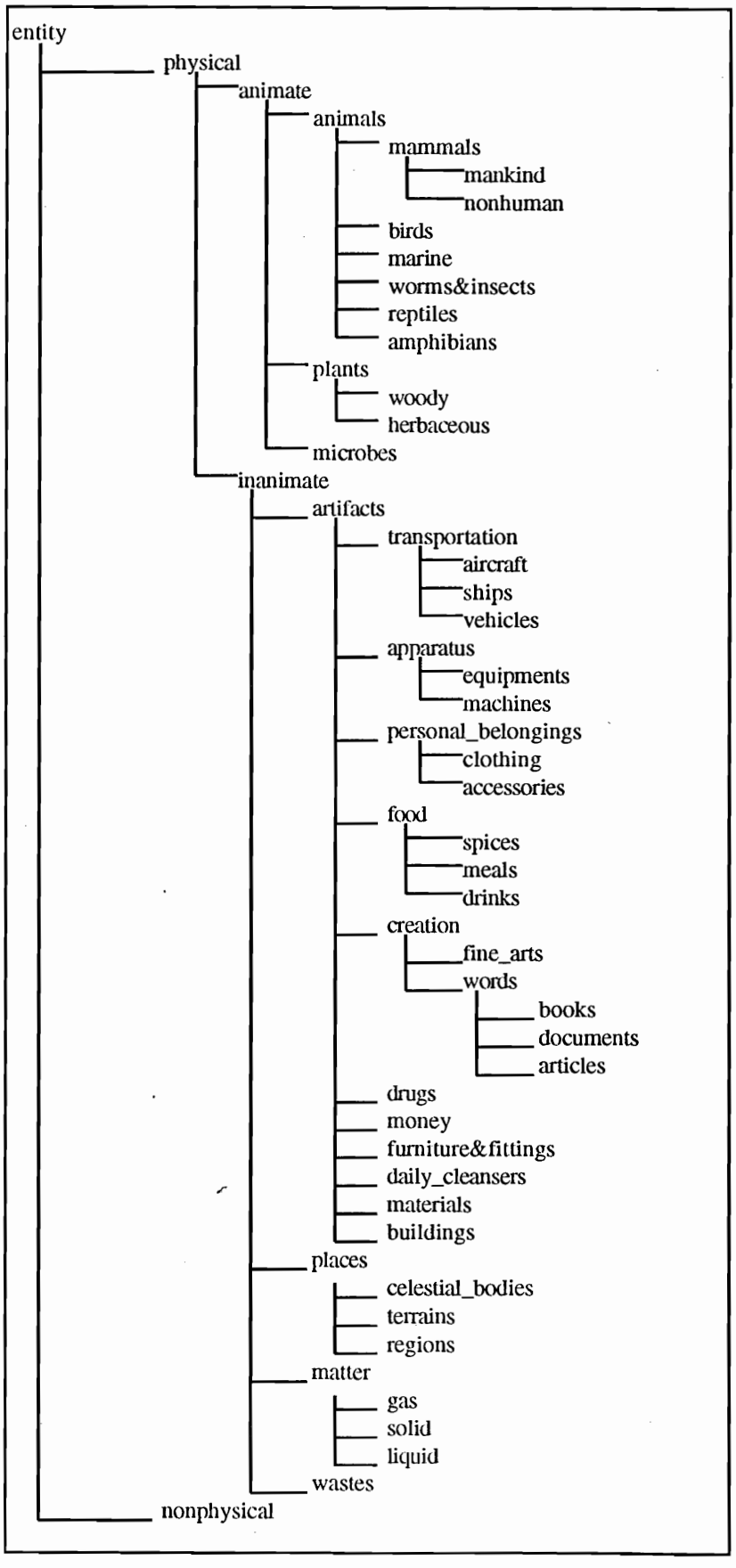
[4] 詞庫小組 (1995) 中央研究院平衡語料庫的內容與說明，中央研究院資訊科學研究所技術報告 95-02。

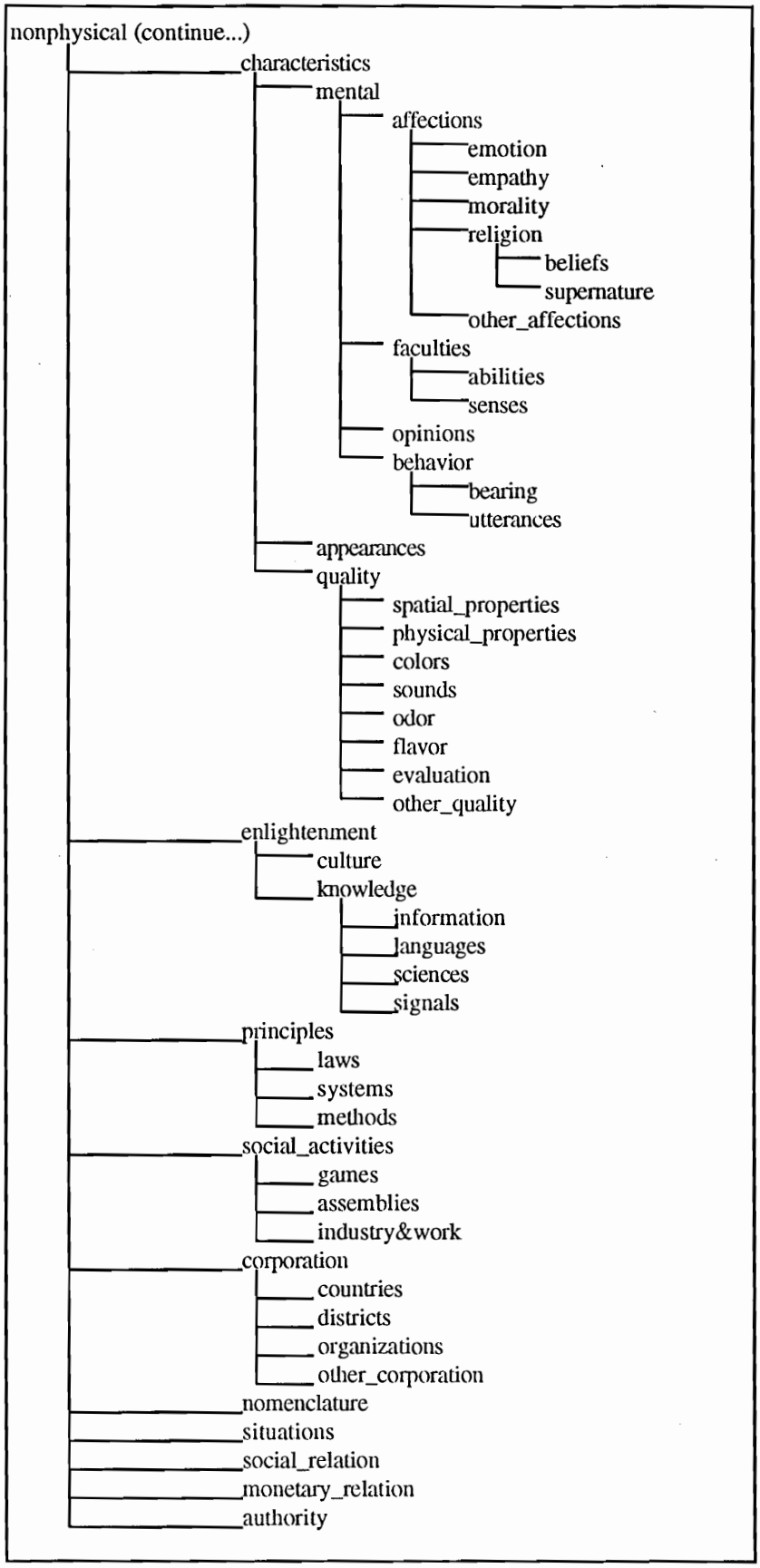
[5] Chang, Li-Ping, Chen Keh-Jiann (1995) "The CKIP part-of-speech tagging system for modern Chinese texts". *Proceedings of ICCPOL 95.*

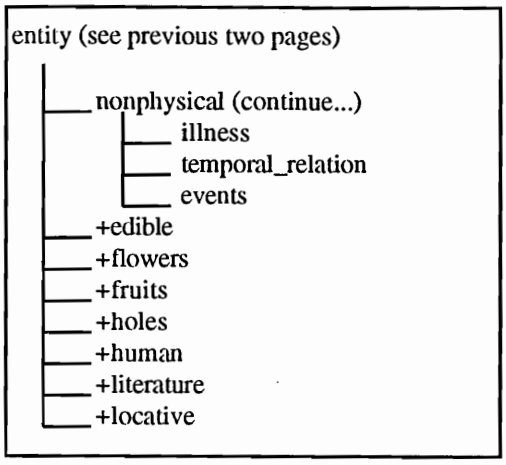
[6] Chen, Keh-Jiann, Liu Shing-Huan, Chang Li-Ping, Chin Yeh-Hao (1994) "A practical tagger for Chinese corpora". *Proceedings of ROCLING VII*, 111-126.

[7] Resnik, P. S. (1993) "Selection and information: a class-based approach to lexical relationships". Ph.D dissertation, University of Pennsylvania.

Appendix: A Taxonomy of Concept




nonphysical (continue...)
characteristics
mental
affections
emotion
empathy
morality
religion
beliefs
supernature
other_affections
faculties
abilities
senses
opinions
behavior
bearing
utterances
appearances
quality
spatial_properties
physical_properties
colors
sounds
odor
flavor
evaluation
other_quality
enlightenment
culture
knowledge
information
languages
sciences
signals
principles
laws
systems
methods
social_activities
games
assemblies
industry&work
corporation
countries
districts
organizations
other_corporation
nomenclature
situations
social_relation
monetary_relation
authority


entity (see previous two pages)
nonphysical (continue...)
illness
temporal_relation
events
+edible
+flowers
+fruits
+holes
+human
+literature
+locative

語料庫在辭典編輯上的運用

張麗麗*、陳克健*、黃居仁**

*中央研究院資訊科學研究所　**中央研究院歷史語言研究所

張麗麗：lili@iis.sinica.edu.tw
陳克健：kchen@iis.sinica.edu.tw
黃居仁：hschuren@ccvax.sinica.edu.tw

摘要

這篇文章介紹了詞庫小組利用中研院語料庫編輯辭典的方法和步驟，同時也探討語料庫對於辭典編輯的貢獻。我們所編輯的是量詞辭典和名量搭配辭典，兩本辭典是一體兩面，內容彼此相關。利用兩個資料庫和一個編輯介面系統，我們可以在線上直接抽取、統計、篩選相關語料，並且可以在線上進行編輯的工作，包含鍵入、轉換、挑選、排序、統計、篩選、傳送、增刪、複製、搬移、輸入等工作。根據我們的經驗，語料庫對於辭典編輯最大的貢獻在於它能提供各式各樣的語言事實，使得辭典能夠更廣泛、深入地記載每個詞彙的用法。豐富例句幫助編輯人員掌握每個詞彙全面的、當代的用法。各種詞頻資料是許多道編輯步驟的重要依據，像是決定詞項、選取例詞、排列例詞、調整用法分析的順序等。辭典中的例詞例句也都是從語料庫中擷取出來的，比起編輯人員自己造的例子，更能反應語言實際的使用情形。此外，由於我們使用的是標記語料庫，詞和詞之間是一個一個分開的，並標有詞類，搜尋起來更加方便有效。

1. 簡介

由於電腦發展一再突破了容量和速度的極限，因此利用電腦來處理大量語料才變成可能。近年來，語料庫為本（corpus-based）的研究成為語言學及計算語言學的一個重要發展，甚至也拓展到應用語言學的領域上。1987年英國辭典公司 Collins 出版了世界上第一本利用語料庫編輯的辭典（Sinclair 1987b），開啓了新的辭典編輯方式，更開創了新的辭典風格。[1]

Collins 公司委託英國伯明罕大學英文系成立了一個 COBUILD專案計畫。[2] 首先他們利用該系發展多年的語料庫針對每個詞項作索引檔。其次，編輯人員利用索引查詢語料庫的資料以及其他相關資料，並依照規畫好的格式撰寫每個詞項的基本資料，包含詞項、發音、各種詞形、意義、詞類、例句…等等。然後再人工鍵入這些資料，建立了基本詞彙資料庫。這個基本詞彙資料庫是該專

案計畫的核心，儲存了每一個詞彙的相關訊息，可以發展成各式各樣的辭典、文法書、語言教學用書……等等。[3] 下一步就是由電腦從基本詞彙資料庫中抽取適當資料組合成辭典或書本的形式，放入排版系統。編輯人員直接進入排版系統進行潤飾、修改的工作，完成編輯程序。該專案計畫的第一個成果就是這本極受歡迎的 Collins Cobuild English Language Dictionary。

基本說來，這是工具的突破帶動了辭典編輯方式和風格的突破。從整個編輯流程來看，電腦、機讀語料庫和機讀資料庫是新引入的編輯工具。究竟這些新的工具起了什麼樣的作用呢？該辭典的主編 John Sinclair （Sinclair 1987a pvii）認為機讀語料庫使得編輯人員「第一次能夠廣泛且深入地觀察語言」，並且「對於語言中最常用的表達方式有了成千上萬的新發現。」[4] 機讀資料庫成為另一種形式的語料來源，電腦可以依照資料庫中的詞類或語法屬性分類查詢，「使得編輯人員可以檢查分類上的一致性，或是經由比對修正某個詞的分析。」（Sinclair 1987a p42）電腦除了在機讀語料庫和機讀資料庫中建索引資料、作搜尋、排序、分類、統計…等工作外，還具有各式各樣的編輯功能，像是挑出拼寫錯誤、核對交錯索引、排版…等，甚至還可以成立電子郵件中心，流通編輯人員的提示。

詞庫小組已於一九九五年發表推出「中研院平衡語料庫」(Sinica Corpus)，是世界上第一個有完整詞類標記的漢語平衡語料庫（黃居仁等 1995、詞庫小組 1995）。該小組並於該年七月受國語日報社委託展開利用語料庫編輯辭典的計畫，預計一九九六年八月完成。早期詞庫小組已經採用語料庫進行詞彙分析，但是利用語料庫編輯辭典還是第一次嘗試。由於人力、時間及經費上的限制，我們必須從規模較小的辭典開始。我們使用的是標記語料庫，可以挑出特定的詞類來編輯，因此我們和國語日報社商定編輯兩本相關的辭典：量詞辭典和名量搭配辭典。就我們所知，在漢語領域中，這是首度利用語料庫來編輯辭典。我們依照語料庫的特性、漢語的特性以及所編辭典的特性，為這兩本辭典設計了新的編輯方法和流程，在這篇文章中將為各位一一介紹。除此之外，我們也會就這次編輯經驗討論語料庫對辭典編輯的貢獻。

2. 量詞辭典與名量搭配辭典的編輯

2.1 構想

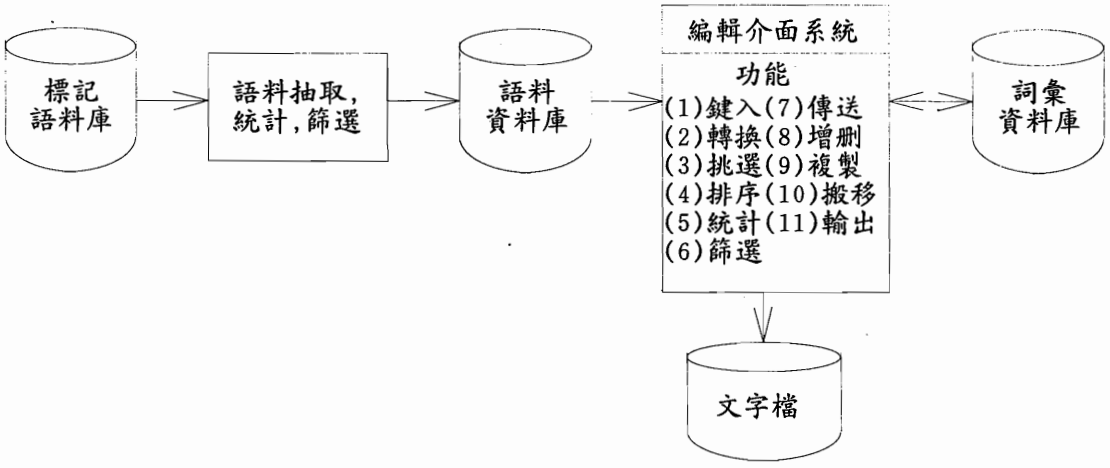
我們所要編輯的量詞辭典以及名量搭配辭典是一體兩面。量詞辭典是從量詞出發，探討每個量詞的用法以及可以搭配的名詞類型。名量搭配辭典是從名詞出發，探討每個名詞能夠搭配的量詞。在名量搭配辭典中，我們有一個創新的作法，就是依名詞尾將名詞分類，[5] 具有相同詞尾的名詞再分成次類，[6] 只需針對同一組的名詞標上適用量詞就可以了。

根據編輯體例（請參見附錄一），在量詞辭典中每個詞項下包含：量詞項、注音、分類、量詞次類標記、語意說明、例詞、例句、辨析及總辨析（可省略）；在名量搭配辭典中每個名詞尾下包含：名詞尾項、注音、語意分類、語意說明、次分類、例詞、所搭配量詞、辨析（可省略）。依照上述這兩本辭典的架構，語料庫可以提供下列資料：一、詞項清單：量詞、名詞、名詞尾；二、詞頻資料；三、名詞量詞搭配資料；四、例詞和例句。

2.2 方法

COBUILD 小組的編輯流程並非完全電子化，從語料庫到詞彙資料庫這一段是離線作業的。他們的作法是由編輯人員觀察語料庫中的資料，用紙和筆撰寫每個詞項的基本資料，再鍵入電腦中，建立基本詞彙資料庫。他們採取這樣的方式是可以理解的，因為每一個詞項都可能有成千上萬個例子有待觀察，而且除了例句外，語料庫並不能直接提供所需資料，都需要編輯人員融會貫通之後才能整理出基本資料。但是這樣的編輯流程將會增加鍵入及校正的工作。我們決定將這個步驟也電子化，也就是說我們可以直接將語料庫中的資料轉到詞彙資料庫中。因為這個編輯計畫所面臨的情況不同：我們所使用的是標記語料庫，可就需要挑選出最適當的資料；尤其是就名量搭配辭典這一部份來看，編輯只要參考語料庫中搜尋出的相關資料即可，而且只要加以適當的分類、編排，可以直接用於辭典中；再加上量詞辭典和名量搭配辭典的規模不大，就記憶空間和速率來看都不成問題。

我們的作法是在語料庫和詞彙資料庫之間多建立了一個語料資料庫，並加上一個編輯介面系統，以便傳送語料庫中的資料到詞彙資料庫中。因此針對每個詞項必須建立兩個資料庫：語料資料庫和詞彙資料庫。語料資料庫儲存從語料庫中抽出的相關資料，詞彙資料庫存放編輯過的資料。透過編輯介面系統可以直接擷取語料資料庫中的資料放置到詞彙資料庫中適當的欄位。詞彙資料庫的格式是依據辭典所需內容規畫好的，其中資料可以轉成辭典形式的文字檔，交給排版系統處理。語料庫和這兩個資料庫的對應關係如圖一所示：



圖一：三個資料庫間的關係以及編輯流程

2.3. 編輯流程

從圖一可以清楚看到建立詞彙資料庫的流程包含兩個步驟：第一個步驟是從標記語料庫中抽取語料，建立語料資料庫；第二個步驟是透過一個編輯介面系統建立詞彙資料庫。第一個步驟大部份是由電腦完成的，包括從大量語料中抽取所需資料，並進行排序、統計等工作。只有在少數情況下才需要人工進行篩選。第二個步驟則需要人工編輯詞彙的每項訊息。有一部份訊息，像是解釋、用法說明必須完全靠人工鍵入；但是有一部份訊息，像是例詞例句，卻可以到語料資料庫中挑選，再傳送到詞彙資料庫。

詞彙資料庫建立之後還有一個步驟，就是從詞彙資料庫中抽取資料轉成純文字檔，再交給排版系統排版。這個步驟也是在編輯介面系統中進行。

2.3.1 步驟一：建立語料資料庫

語料資料庫中存放的資料都是直接從語料庫中得出的資料，包含：一、詞項清單：量詞、名詞、名詞尾；二、詞頻資料；三、名詞量詞搭配資料；四、例詞和例句。其中只有只有例詞例句另外需要人工篩選，因為例詞例句的數量太大。例如最常用的量詞「個」在兩百萬詞的語料庫中出現15495次，次常用量詞「種」也出現4666次。因此我們特別設計了一個小程式，由編輯人員在語料中挑選出涵蓋各種用法的適量的例詞例句，加上規定的符號，啓動傳送功能，就會將這些例子傳送到語料資料庫中。建立語料資料庫包含了下列幾個步驟：

a. 詞項搜尋：

找出中研院語料庫中所有的量詞：537個量詞，共58,615例。並計算每個量詞的出現頻率。（請參見附錄二）

找出中研院語料庫中所有的常用普通名詞：30,888個名詞，共378,768例。另外補充2,167個名詞，[7] 共計33,055個名詞。並計算每個名詞的出現頻率。

將33055個名詞依名詞尾排序比對，共得2,787個詞尾。（請參見附錄三）

b. 搜尋名詞、量詞搭配關係

找出量詞右方5個詞內出現過的名詞的例子：502個量詞，共46,569例。

刪除名詞出現在某些標點符號（如，。：；？！）之後的例子，及名詞或量詞為外文的例子。

依上述資料建量詞和名詞的搭配資料：共計53,872對

找出的53,872對名量搭配資料將分別放入各個量詞和名詞的語料資料庫。

c. 挑選例詞和例句

從語料庫中挑選涵蓋各種用法的例詞和例句，標上規定的符號。

將資料傳送至語料資料庫。

d. 建立量詞和名詞尾的語料資料庫

量詞的語料資料庫包含：量詞項、詞頻、例詞、例句（由於有例詞例句，所以不將所搭配的名詞列入[8]）（請參見附錄四）

名詞尾的語料資料庫包含：名詞尾項、詞頻、組成名詞、詞頻、每個名詞所搭配量詞（請參見附錄五）

2.3.2 步驟二：利用編輯介面系統建立詞彙資料庫

為了編輯流程全面電子化，我們特別設計了一套編輯介面系統，可以讓編輯人員直接編輯詞彙資料庫。該編輯介面系統包含以下幾項功能：(1) 鍵入、(2) 轉換（進入語料資料庫）、(3) 挑選（挑選例詞和例句）、(4) 排序（就選取的例詞排序）、(5) 統計（統計同一組名詞所搭配的量詞及出現次數）、(6) 篩選（就統計出的量詞資料進行篩選）、(7) 傳送（將語料資料庫中挑選出且排序篩選過的語料傳到詞彙資料庫中適當的欄位）、(8) 增刪（增加或刪除整項分析）、(9) 複製（複製整項分析）、(10) 搬移（搬移整項分析）、(11)輸出（將資料轉成辭典形式的文字檔）。

在該系統中編輯人員可以隨時進入語料資料庫觀察語料，構想整個詞彙的分析架構，然後在編輯介面系統直接鍵入各項資料，或到語料資料庫挑選適當資料傳送到詞彙資料庫中。傳送資料的過程如下：(1)進入語料資料庫、(2)挑選例詞或例句、(3)編輯系統自動依照頻率將例詞例句排序、(4)編輯系統主動提出重新排序功能、[9] (5)編輯人員可以重新排序也可以跳開、(6)最後將挑出的且排好序的資料傳送到詞彙資料庫。在名量搭配辭典中，排序後還要進行兩個步驟才能傳送。在名詞尾的例詞排序後，編輯系統會主動替放在同一組的名詞累計所搭配的量詞及其頻率，並將統計出的量詞資料依頻率排列。這份資料的錯誤率比較高，所以統計之後，也提供篩選的功能。[10] 編輯人員篩選完畢才會將例詞和量詞一起傳送到詞彙資料庫（整個傳送過程請參見附錄六）。除此之外，針對整項分析，編輯系統還提供增加、刪除、複製和搬移的功能。

詞彙資料庫各項資料的建立大致可以利用鍵入和傳送的方式完成，如表一和表二所示。

表一：量詞辭典詞彙資料庫的資料建立方式

量詞辭典	電腦自動存入詞彙資料庫中	編輯人員構想鍵入詞彙資料庫中	編輯人員從語料資料庫中挑選，再由電腦傳送
量詞項	*（傳送）		
注音	*（傳送）		
分類		*（鍵入）	
量詞次類標記		*（鍵入）	
語意說明		*（鍵入）	
例詞			*（傳送）
例句			*（傳送）
辨析及總辨析（可省略）		*（鍵入）	

表二：名量搭配辭典詞彙資料庫的資料建立方式

名量搭配辭典	電腦自動存入詞彙資料庫中	編輯人員構想鍵入詞彙資料庫中	編輯人員從語料資料庫中挑選，再由電腦傳送
名詞尾項	*（傳送）		
注音	*（傳送）		
語意分類		*（鍵入）	
語意說明		*（鍵入）	
次分類		*（鍵入）	
例詞			*（傳送）
所搭配量詞		*（鍵入）	*（傳送）
辨析(可省略)		*（鍵入）	

雖然編輯人員需要鍵入的資料不少，但是這些都是需要靠觀察語料資料庫中的語料才能做出的分析。也就是說語料資料庫中的語料不是只有提供資料傳送到詞彙資料庫而已，還能供編輯人員觀察語言的使用情形。語料資料庫中的例詞例句及搭配資料都是根據需要精挑出來的，除去了冗贅的訊息，使得編輯人員在觀察和歸納語言現象時更加便利。例如在替名詞尾的語意作分類時，只要觀察包含該名詞尾的所有名詞即可，根本不需要觀察例句。又例如在分析量詞的用法時，由於語料資料庫中的例詞和例句是人工篩選出來的，能夠充分反應該量詞的各種用法，所以只要觀察這些挑出的例子就可以了。

2.3.3 步驟三：從詞彙資料庫中抽取資料轉成純文字檔，放入排版系統

編輯系統還提供了一個功能，就是從詞彙資料庫抽取資料，並依照編輯體例轉成類似辭典形式的純文字檔（請參見附錄七）。編輯人員可以利用文字檔進行集體校閱的工作。許多問題和現象在分析單個量詞時是看不出來的，但是有了整體的分析資料，要進行各種觀察和比較就十分容易。更正的工作仍然回編輯系統中進行，修改完畢再轉成新的文字檔。最後，內容正確無誤的文字檔才交給排版公司，放入排版系統排版。

3. 語料庫對辭典編輯的貢獻

近年來，語料庫為本(corpus-based)的研究成為語言學及計算語言學的一個重要發展（Svartvik 1992， Church and Mercer 1993， 陳克健 1994， 黃居仁 1995）。語料庫對於各種研究的最大貢獻在於它提供了豐富的語言事實，而且由於是機讀形式，可以迅速有效執行各式各樣的搜尋、排序及統計的工作。對於辭典編輯，它最大的貢獻也在此。在傳統編輯辭典的工作中，蒐集資料是件重要但是十分不容易的工作，要從浩瀚的語言中抓取一個詞彙的各種使用實例幾乎是不可能做到的。因此辭典內容往往只是反應編輯人員本身的語文素養、對語言的觀察能力、以及考證其他辭典或相關書籍的功夫，究竟一本辭典反映了多少語言實況根本無從得知。語料庫卻完全突破了上述的限制，編輯人員首度能夠充分掌握實際語料，做出準確、客觀、有根據的詞彙分析，使得辭典在反應語言實況上有了重大的進步。同時由於是機讀語料庫，在搜尋資料的效率和速度上也大大超越傳統方式的極限。

在實際編輯過程中，我們發現在編輯的每一個細節中都可能需要參考語料庫所提供的各種資料，因為語料庫所能提供辭典編輯的語言事實是多方面的：

a. 全面反應每個詞彙的各種用法。

有些詞彙的用法是非常複雜的，沒有掌握充分的語料，就沒辦法將該詞彙準確分析。以「片」為例，它可以用來計量許多的事物，有具體的，像是「幾片落葉、兩片玻璃、一片江山、一片茶園」，也有抽象的，像是「一片歡樂，

一片熱心、一片朝氣」，甚至於還有介於抽象和具體之間，像是「一片不景氣、一片陰影、一片笑聲」。沒有將各種用法蒐集齊全並經過仔細的觀察，我們很難將「片」的用法作一番詳盡且深入的描繪。語料庫可以很快的將出現量詞「片」的所有例句馬上找出，編輯人員才能免除以偏蓋全的錯誤。

試比較我們所編出的量詞辭典和其他的量詞辭典，可以看到一個明顯的差異：我們所分析的用法總是比較周全。以「道」為例，大陸編的「現代漢語量詞手冊」（郭先珍，1986）分析「道」用來計量三種物品：

1 計量江河或某些長條狀的東西。
2 計量門、墻或類似門墻的東西。
3 計量命令、題目等。

而我們從語料庫中所觀察到的「道」卻可以計量六種物品（編號前加上*的符號表示是本辭典才有的用法分析）：

1 用來計量長條形物品。
*2 用來計量成線狀的光線。
3 用來計量門或牆。
4 用來計量題目、命令。
*5 用來計量菜餚。
*6 用來計量程序。

我們從語料庫中所觀察到的豐富語言事實除了反映在用法分項上，也反映在「辨析」欄中。例如「片」和「聲」都可以用來計量聲音，但是除了意義不同外，用法也不同：

片：辨析：…… 計量聲音最常用的量詞是「聲」，但是它是個別計量每一次發出的聲音，如「三聲槍響、幾聲狗吠」，不像「片」是用來描述同時發出的大量聲音。因此，「聲」的前面可以接各種數詞，而「片」只能接「一」。請參見「聲」。

b. 忠實反應每個詞彙的當時用法。

試比較大陸出的「現代漢語量詞手冊」（郭先珍，1986）以及教育部1995年10月出的「常用量詞手冊」（教育部國語推行委員會，1995）和我們所編的量詞辭典，就以「支」為例。郭先珍（1986）分析「支」有五個用法：

1 計量隊伍。
2 計量歌曲或樂曲。
3 計量電燈的光度，相當於「瓦」。
4 計量某些杆狀的東西。
5 紗線粗細的計量單位。

國語會（1995）的分析十分類似，有四個用法：

1 計算隊伍的單位。
2 計算歌曲、樂曲的單位。
3 計算棉紗細度的單位。
4 計算燈光強度的單位。

再看看我們利用語料庫所觀察到的「支」的用法，共有十一種之多（編號前加上*的符號表示是本辭典才有的用法分析）：

1 用來計量長杆狀的物品。
*2 用來計量電話。
3 用來計量隊伍或團體。
*4 用來計量整體中分支出來的人、物。
*5 用來計量錄影帶、影碟。
6 用來計量歌曲、舞蹈或影片。
*7 用來計量棒球賽中擊出的打擊數。
*8 用來計量賞罰記錄。
*9 用來計量數字、號碼。
*10 用來計量股票。
11 用來計量紡紗。

我們所分析出來「支」的用法，有七種現代用法是大陸「現代漢語量詞手冊」以及教育部「常用量詞手冊」中所沒有的。這七種用法所計量的物品都是現代社會中的新產物，而且成為現代生活中重要物品，像是：電話、錄影帶、影碟、棒球、大過、股票、…等。若不是語料庫中豐富多樣的語料我們也是沒有辦法掌握到現代漢語的最新變化。同時我們在語料庫中也觀察到現在已經不用「支」來計量電燈的光度，而用「燭光」來取代。因此，和另外兩本辭典不同，我們並不放這個用法。

c. 詞頻資料可以幫助決定詞項。

以量詞「爿」為例，雖然我們可以在其他辭典中查到這個詞，知道它是用來計量商店，但是在語料庫中，它的出現次數是零。有了這項訊息，編輯人員就需要考慮是否要收「爿」在量詞辭典中。在分析名量搭配辭典中的「商店、雜貨店、旅店」等名詞項時，也要考慮是否要放「爿」這項搭配資料。

d. 頻率資料可以幫助挑選例詞、調整例詞順序。

在編輯中，我們總是從眾多例子中挑選出最具代表性的例子，而且排序上也以此為準則。什麼樣的例子算是具有代表性呢？出現頻率就是一項很重要的依據。以名詞尾「本」為例，它表示三種意思：根本、本子、本錢。表示「根本」之意有兩個例詞：「基本」（361次）、「根本」（53次）。很明顯的，「基本」的次數遠超過「根本」，依照這個訊息編輯人員應該將「基本」排在「根本」之前。另外表示「本子」的例子計有：劇本（69次）、版本（51次）、書本（47次）、課本（22次）、筆記本（11次）、……、中譯本（1次）、手抄本（1次）、標點本（一次）、印本（1次），共四十二個例子。編輯人員不需要將所有四十二個例子都放到辭典中，而是有所取捨。取捨的標準之一就是它們的出現次數，編輯人員會優先挑出常用詞，並且也會參考次數多寡排列例詞的順序。

e. 頻率資料和例詞數量可以幫助調整用法分析的順序。

以「張」為例，它可以用來計量臉、嘴、紙、桌子等物品。這些都是實體名詞，也都是日常生活中的常用名詞。究竟哪一個用法該放在前面？哪一個用法該放在後面？「張」的頻率資料和例詞數量可以作為參考（請參見附錄八）。根據這兩項準則，我們對「張」的分析順序如下：

1 用來計量平面的物品。
例詞：十張紙、四千多張罰單、一千張磁碟片、一張會員卡、九張照片、幾張鈔票、幾張郵票、一張稅單、兩張畫、幾張報紙、好多張地圖、八十張幻燈片、三張蔥油餅。

2 用來計量主要功能在平面上的物品，如桌子、床、椅子。
例詞：這張雙人床、那張床、十幾張圓桌、那張椅子、一張高腳小方桌、幾張書桌、那張長沙發。

3 用來計量臉面。

例詞：一張臉、那張驚慌的小臉、一張張猶顯稚氣的臉龐、一張張黧黑的面孔。

4用來計量可以張開的物品。

例詞：兩張嘴巴、一張伶牙利嘴、幾張弓、兩張魚網。

f. 標記語料庫可以提供各種語法行為的統計資料。

就編輯辭典的層面來看，標記語料庫的運用層面比起未標記語料庫來得廣。因為標記語料庫提供了更多的訊息，有利編輯人員利用這些訊息作各種搜尋的工作。由於中研院語料庫標有詞類，因此我們才能抓取語料中的名詞。也才能將名詞依詞尾排序，得出名詞尾資料。也因為同樣的理由，我們才能搜尋名詞和量詞的搭配關係供編輯人員參考，否則就得完全靠編輯人員的功力了。

就漢語書寫的特性而言，詞和詞是串連在一起的，中間沒有間隔標記。標記語料庫中的詞是一個一個分開的，無論是搜尋、排序、統計…等工作，其準確度和速率都會大為提高。以「本」為例，在兩百萬語料庫中，「本」字一共出現了5784次，但是其中只有313次是量詞的用法。利用標記語料庫可以直接將這313個例子準確的挑出，省去從5784個例子中挑選和核對的麻煩。

g. 提供實用的例詞和例句

從語料庫擷取的例詞和例句和編輯人員自行造出的例詞例句是有很大的區別的，這一點 Sinclair（1987b pxv）有很精闢的說明：

> ……語料庫中的例子反映了實際使用的情形。它們佐證解釋、反映用法、並提供了現代英語在說和寫上可以信賴的指引。
>
> 相反的，造出的句子實際上只是解釋的一部份。它們被用來闡明（refine）解釋，在多半情況下只是被用來闡述（clarify）解釋。

語料庫除了提供上述多方面的語言事實外，它對語言分析也起了間接的作用。就是因為語料庫能夠提供多種語言面貌，因此幫助編輯人員做出更準確的分析。有些影響是反映在個別詞彙的分析上，有些則是反映在整體的語法架構上。就以這次編輯的經驗來看，在量詞分析的過程中，由於對量詞的用法一直有新發現，因此我們對量詞的分類也一再修訂改善。一開始是依照 Chao(1968)

及詞庫小組（1993b）的架構將量詞分成十類，[11] 最後修訂出來的分類為：一般量詞、事件量詞、種類量詞、容器量詞、約量量詞、標準量詞、動量詞，一共七類。我們認為新的分類更能掌握量詞間的差異，更能將性質不同的量詞區分清楚（請參考詞庫小組（1996）以及Huang（in prep.））。

4. 討論

最後，我們也想提一下機讀資料庫對辭典編輯的助益。在這次辭典編輯的經驗中，無論是語料資料庫或是詞彙資料庫都具有以下的功能，增加了編輯流程的效率和便利。

一、可以隨時更新。
二、可以進行比對和比較的工作。
三、可以根據需要抽取所需資料。
四、可以變化各種輸出或排版格式，也可以根據需要將詞項作各種排序。

這一套編輯流程的設計粗具規模，還有許多待加強和改進之處。例如，詞彙資料庫和文字檔若能雙向傳輸，修改資料則更加便利。因為校稿工作多半是在印出的文字檔進行的，如果修改的工作也可以直接在文字檔中利用編校系統（editor）進行，再傳回詞彙資料庫就方便多了。但是由於目前詞彙資料庫和文字檔只能單向傳輸，所以只能進入編輯系統逐詞逐欄地修改了。其次，和COBUILD 作法不同的是，我們的整項電子化流程並沒有包含排版系統，我們是將純文字檔的最後版本交給打字公司排版的。由於這兩本辭典的適用對象是小學生，必須加上注音和插圖，這些都增加了排版的困難。所以我們認為就這兩本辭典編輯而言，並不適合加入排版系統。再者，我們還可以建立交互索引功能。在分析一個詞彙會提到一些相關詞彙，像是同義詞、似義詞、反義詞、用法相近詞、…等，也應該在這些相關詞彙下標示。也就是電腦應該提供一個備忘庫，提醒編輯分析該詞彙時可以參考的詞項。這項功能在編輯大部頭的辭典時格外重要。此外，編輯介面系統也可以增加比對資料的功能。由於這兩本辭典是相呼應的，我們必須確定兩本辭典的內容不會互相矛盾，或是有所遺漏。比對工作交給電腦將會比較周全，效率也較好。但是因為兩部辭典是到同樣的語料庫中搜尋資料，差異不至於太大，因此決定交由人工進行比對。

註：

1. 隨後牛津及朗文這兩個重要辭典出版中心也立即跟進，構建了以辭典編纂為目標的語料庫。

2. COBUILD是Collins Birmingham University International Language Database 的缩寫。

3. COBUILD 根據這個基本資料庫已經陸陸續續出版了一系列的辭典、文法書及語言教學用書，像是：Collins COBUILD English Language Dictionary、 Collins COBUILD Essential English Dictionary、Collins COBUILD Student's Dictionary、 Collins COBUILD Dictionary of Phrasal Verbs、Collins COBUILD English Grammar、Collins COBUILD Student's Grammar、Collins COBUILD English Guides: 1 Prepositions and 2 Word Formation、Collins COBUILD English Course: levelsl, 2, and 3.、Collins COBUILD English Course: Tests、 Collins COBUILD English Course: First Lessons。

4. Sinclair(1987a pvii)舉了幾個例子：see、give、keep都有基本的語意，一般人都會認為基本的意義也是最常用的，但是事實並不是如此。see最常用在 "I see, you see"，give最常用在 "give a talk"， keep 最常用在 "keep warm"。

5. 由於名詞數量十分龐大，分別一一標記，不但耗工費時，讀者在查詢時也不方便。語意相近或是分類上屬於同一類的名詞所搭配的量詞往往都一樣，所以針對同一類的名詞來給量詞是可行的。依照Sproat and Shih(1996)的說法，名詞尾通常都是表示「種類」的概念，因此我們依照名詞尾來分類，往往也能把屬於同一類的名詞放在一起。依照名詞尾來分類名詞還有許多好處，不但檢索容易，相當於逆查辭典，而且還提供了名詞尾的語意分類。

6. 依照編輯原則，帶有同一個詞尾的名詞依下列原則再分成次類：一、名詞尾本身的語意不同；二、名詞所搭配的量詞不同。

7. 根據詞庫的詞類來看，名詞屬於Na和Ncb兩類。但是中研院語料庫只標上簡單的詞類，有Na，但是Ncb歸入Nc。Nc包含了Nca、Ncb兩類，必須一一核對詞庫電子辭典才能確定其細類。因此在中研院語料庫只找Na，另外利用詞庫另一套一千萬詞語料庫核對詞庫電子辭典，找出出現兩次以上的Ncb，共1277個，以及出現兩次以上卻不在中研院語料庫中出現過的的Na，共890個。這就是為什麼補充了2167個名詞（1277個Ncb加上890個Na）的緣故。

8. 由於名量搭配資料是根據詞的間距來計算而非根據實際結構來判斷，因此誤差大，不比人工挑選出的例詞例句。所以在量詞的語料資料庫中不放電腦找出的名詞搭配資料，只放編輯人員挑選出的例詞例句。在名詞尾的語料資料庫中，雖然放了每個名詞所搭配的量詞，並且在編輯系統中提供統計每一組名詞所有的量詞和頻率的功能，但是仍然需要人工檢查、篩選和補充。

9. 根據挑選出的例詞排序的方法是：電腦先依照出現頻率將例詞排序並給予例詞010、020、…120等編號。編輯人員可以更改編號，然後電腦再依照編號大小重新排序，並放入詞彙資料庫。

10. 請參見註8。

11. 根據詞庫小組（1993b）的分類，量詞有十類：個體量詞、跟述賓式合用的量詞、群體量詞、部份量詞、容器量詞、暫時量詞、標準量詞、準量詞、動量詞、零量詞。

參考書目

郭先珍，1986，現代漢語量詞手冊，中國和平出版社。

教育部國語推行委員會，1995，常用量詞手冊，國語文教育叢書17，教育部。

陳克健，1994，素材語言學與本文處理，發表於ICCL-3會議，一九九四年七月，香港。

黃居仁，1995，科際整合與整合科技—談計算語言學與語料庫語言學之角色與發展。「語言學研究之現狀與發展」研討會，七月十五日，國立台灣師範大學。

黃居仁、陳克建、張莉萍、許蕙麗，1995，中央研究院平衡語料庫簡介，中華民國八十四年第八屆計算語言學研討會論文集。

詞庫小組，1993a，新聞常用名詞詞頻與分類－語料庫為本研究系列之四，技術報告93-04，中央研究院資訊科學研究所。

詞庫小組，1993b，中文詞類分析，技術報告93-05，中央研究院資訊科學研究所。

詞庫小組，1995，中央研究院平衡語料庫的內容與說明，技術報告95-02，中央研究院資訊科學研究所。

詞庫小組，1996，量詞的語意分類，搜文解字，計算語言學學會通訊第七卷第四期。

Allen, Keith, 1977. "Classifiers". Langauge. 53:285-311.

Chao, Yuen Ren, 1968. A Grammar of Spoken Chinese. Berkeley: University of California Press.

Church, K. W. and R. L. Mercer, 1993. "Introduction to the Special Issue on Computational Linguistics Using Large Corpora." Computational Linguistics, Vol.19, No.1, pp.1-24.

Huang, Chu-Ren, 1989. Citicization and Type-Lifting: A Unified Account of Mandarin NP de. Bloomington:IULC.

Huang, Chu-Ren, 1994. "Corpus-based Studies of Mandarin Chinese: Foundational Issues and Preliminary Results." In Mathew Chen and Ovid Tzeng Eds. In Honor of William S.-Y. Wang: Interdisciplinary Studies on Language and Language Change. pp165-186. Taipei:Pyramid.

Huang, Chu-Ren, in preparation. "Classifiers and Semantic Type Coercion: Motivating a New Classification of Classifiers."

Sinclair, John, 1987a. Ed. Looking up: An Account of Cobuild Project. London: William Collins Sons & Co Ltd.

Sinclair, John, 1987b. Ed. Collins Cobuild English Language Dictionary. London: William Collins Sons & Co Ltd.

Sproat and Shih, 1996. "A Corpus-based Analysis of Mandarin Nominal Root Compound." Journal of East Asian Linguistics 5, 46-71.

Svartvik, Jan, 1992. Ed. "Directions in Corpus Linguistics." Proceedings of Nobel Symposium 82, 4-8 August 1991. Trends in Linguistics Studies and Monographs 65. Berlin: Mouton.

Tai, James H-Y, 1994. "Chinese Classifier Systems and Human Categorization." In Honor of William S-Y. Wang: Interdisciplinary Studies on Language and Language Change. (ed.) by Mattew Y. Chen and Ovid J.L. Tzeng.

體例一：量詞辭典

◎ **條** ㄊㄧㄠˊ　個體量詞

①後接條狀物品，像是領帶、項鍊、毛巾、裙子、褲子、傷痕、吐司、腿、船。例：他的手腕上有一條傷痕。

②後接體形呈長條狀之動物，像是蛇、牛、魚、狗。例：前門屋簷上有條大蛇盤踞。

③後接條狀地面物，像是道路、公路、馬路、通路、財路、活路、生路、巷子、捷徑、河、排水溝、支流。例：水庫由三十條大小支流匯合而成。

④後接「線」的家族詞，像是路線、幹線、航線、陣線、防線、生產線、電線、電話線、地平線、視線、專線、線索。例：可以隱隱看見視線盡頭有一條墨綠的地平線。

⑤後接法律規則等詞，像是條文、條款、準則、規則、規定、指標、指示。例：根據清理法第廿三條規定，罰款四千五百元。

⑥後接「命」的家族詞，像是命、人命、生命。例：一場大火奪走了六十四條人命。

體例二：名詞量詞搭配辭典

法 ㄈㄚˇ ①方法、方式：辦法、方法、說法、作法、做法、手法、玩法、想法、標示法、走法、…等。個量：個、項。

群量：套、種、類、樣、系列、串。

部量：些、部份、堆。

②法律條文：憲法、選罷法、民法、建築法、稅法、刑法、訴訟法、…等。個量：條、項、道。群量：種、系列。部量：些、部份、堆。

附錄二 量詞依照頻率排列範例（第一頁）

記錄#	nf	nf_voice	nf_count
1	個（兒）	ㄍㄜ・	15495
2	種	ㄓㄨㄥˇ	4666
3	位	ㄨㄟˋ	3225
4	年	ㄋㄧㄢˊ	2397
5	次	ㄘˋ	2386
6	項	ㄒㄧㄤˋ	2011
7	名	ㄇㄧㄥˊ	1558
8	場	ㄔㄤˊ（俗讀ㄔㄤˇ）	919
9	天	ㄊㄧㄢ	879
10	些（兒）	ㄒㄧㄝ	872
11	條	ㄊㄧㄠˊ	823
12	家	ㄐㄧㄚ	822
13	隻	ㄓ	821
14	元	ㄩㄢˊ	771
15	件	ㄐㄧㄢˋ	718
16	座	ㄗㄨㄛˋ	608
17	歲	ㄙㄨㄟˋ	551
18	段（兒）	ㄉㄨㄢˋ	523
19	屆	ㄐㄧㄝˋ	515
20	張	ㄓㄤ	479
21	塊	ㄎㄨㄞˋ	458
22	支	ㄓ	453
23	句	ㄐㄩˋ	447
24	片（兒）	ㄆㄧㄢˋ	444
25	套	ㄊㄠˋ	441
26	本（兒）	ㄅㄣˇ	435
27	點（兒）	ㄉㄧㄢˇ	422
28	類	ㄌㄟˋ	407
29	部	ㄅㄨˋ	385
30	群	ㄑㄩㄣˊ	315
31	顆	ㄎㄜ	306
32	份	ㄈㄣˋ	305
33	組	ㄗㄨˇ	299
34	首	ㄕㄡˇ	289
35	筆	ㄅㄧˇ	274
36	篇（兒）	ㄆㄧㄢ	273
37	把	ㄅㄚˇ	250
38	期	ㄑㄧˊ	245
39	號（兒）	ㄏㄠˋ	242
40	日	ㄖˋ	240
41	輛	ㄌㄧㄤˋ	220
42	樣（兒）	ㄧㄤˋ	213
43	根（兒）	ㄍㄣ	212
44	度	ㄉㄨˋ	208
45	對	ㄉㄨㄟˋ	204
46	分（兒）	ㄈㄣ	200
47	股	ㄍㄨˇ	194
48	道	ㄉㄠˋ	191
49	批	ㄆㄧ	188
50	週	ㄓㄡ	179
51	口（兒）	ㄎㄡˇ	175
52	間	ㄐㄧㄢ	169
53	隊	ㄉㄨㄟˋ	163
54	層	ㄘㄥˊ	159
55	波	ㄅㄛ（又讀ㄆㄛ）	153

附錄三 名詞尾依照ASCII碼排列範例（第一頁）

NO	詞尾	筆數	包含此詞尾之NA			
1	一	34	國一(18)	高一(10)	大一(3)☆	研一(1)
			唯一(1)	等一(1)		
2	丁	10	園丁(3)	庖丁(2)	兵丁(1)	壯丁(1)
			男丁(1)	補丁(1)	辣子雞丁(1)	人丁(0)☆
			布丁丁(0)☆	柳丁(0)☆		
3	七	2	老七(1)	銫一三七(1)		
4	了	14	知了了(11)	了(3)		
5	二	26	老二二(10)	國二(9)	高二(4)	二(2)
			大二(1)☆			
6	人	16200	人(8921)★◎	女人(300)★	男人(255)★◎	投資人(247)
			中國人(245)◎	家人(170)★◎	國人(166)★◎	人人(150)★◎
			工人(139)★◎	老人(135)★◎	台灣人(130)	夫人(129)★◎
			主人(128)★◎	負責人(124)◎	被害人(121)★	日本人(114)◎
			客人(114)★◎	現代人(112)★◎	年輕人(105)★	軍人(92)★◎
			病人(90)★	大人(85)★◎	候選人(82)	商人(66)★◎
			友人(63)★	藝人(63)★	美國人(62)◎	發言人(62)★
			法人(57)★	敵人(57)★◎	世人(56)★	詩人(56)★◎
			華人(55)★	外省人(52)	古人(51)☆◎	漢人(51)☆
			巨人(50)★◎	主持人(49)★	衆人(47)★◎	強人(46)★
			好人(44)☆◎	老年人(44)	巴西人(42)	外人(42)★
			新人(41)★	窮人(38)	召集人(37)	一家人(36)◎
			婦人(36)★◎	當事人(36)★	領導人(35)	外國人(34)
			農人(33)◎	成人(32)★	製作人(31)	德國人(29)◎
			親人(29)★◎	白人(28)	行人(28)★	後人(28)☆◎
			陌生人(28)◎	旁人(28)☆	碼頭工人(28)	壞人(28)☆◎
			財團法人(27)★	學人(26)☆	小人(25)☆◎	發起人(24)★
			黑人(24)★	證駕駛人(23)	中年人(23)	文人(23)☆
			日人(23)	當地人(22)	法國人(22)	盲人(22)
			情人(22)★	佳人(20)☆	專人(21)★	經理人(21)
			獵人(21)☆◎	提名人(19)	客家人(20)	有心人(19)
			所有人(19)★	美人(17)	山人(18)☆	名人(18)
			泰人(18)	創辦人(17)	英國人(17)	偉人(17)☆◎
			族人(17)☆	浪人(16)	先人(16)☆	印度人(16)
			前人(16)☆◎	被保險人(15)☆	路人(16)★	僧人(16)
			承銷人(15)	某人(14)☆	戀人(15)☆	申請人(14)
			東方人(14)	本省人(13)	爲人(14)★	經紀人(14)
			女人(13)	阿拉伯人(12)	生意人(13)☆	企業人(13)
			西方人(12)	文化人(11)	雪亞人(12)☆	愛爾蘭人(12)
			聖誕老人(12)☆	自己人(10)☆	洲人(11)☆	活人(11)☆
			聖人(11)☆◎	管理人(10)◎	青年人(10)	納稅人(10)
			猶太人(10)☆	村人(9)☆	凡人(9)☆	代言人(9)
			有錢人(9)	行人(9)	所有權人(9)	香港人(9)
			倆人(9)	矮人(9)◎	荷蘭人(9)◎	愛人(9)☆
			愛人樂人(9)	尼泊爾人(8)	葡萄牙人(9)	漁人(9)◎
			代理人(8)	社會人(8)	各人(8)☆	局外人(8)☆
			受刑人(8)	澳洲人(8)	參選人(8)	發行人(8)☆
			瑞士人(8)	犯人(7)☆	土人(7)◎	中縣人(7)
			台北人(7)	承攬人(7)	同居人(7)	完人(7)◎
			宜蘭人(7)	常人(7)☆◎	俄羅斯人(7)	個人(7)
			原始人(7)	罪人(7)	植物人(7)☆	發明人(7)◎
			當選人(7)☆	丈人(6)	義大利人(7)◎	僕人(7)◎
			熟人(7)☆	受害人(6)	上海人(6)	今人(6)◎
			老人(6)	歐洲人(6)☆	俗人(6)	現耕人(6)
			超人(6)☆	主講人(5)	讀書人(6)	內人(5)☆
			少年人(5)	地球人(5)	平常人(5)◎	印地安人(5)
			吉普賽人(5)	牧羊人(5)	死方人(5)	佣人(5)
			波蘭人(5)		南方人(5)	南投人(5)

附錄四 量詞語料資料庫存放資料範例（量詞「條」）

例詞

這條河流
這條線
一條雙曲線
每一條大街
一條門徑
一條葉脈
好幾條魚
一兩條白絹
一兩條水道
一兩條鐵路
這條愛的連線
一條命
一條輸尿管
整條豬
四條路線
整條走廊
一條金鍊子
一條產業道路
一條獨木舟
一四條多條人命
這條海線
一條命
一條游泳褲
一長條弧線
憲法第一八條
這條金

幾條線索
四條弦
一條路徑
那條小溪
一條小魚
這一條山脈
一條繩子
一條金帶子
一條手帕
這條鐵路
這條電話熱線
一條馬路
一條財路
第二條市場路線
一條鐵鍊
一條長長的影子
這條線索
草案第二條
2條航線
這條主線
一條瀑布
一條規則
兩條笨重的犀牛
三十條罪狀
勞基法第八十四條
這條規定

整條街
一條蛇
一條船
一條銀河
一條巨龍
這條公路
這這條偉大的公路
兩條腿
兩條金魚
一條隧道
幾條鯽魚
一條巷子
該一條捷徑
一條狗
這條歌
一條墨綠的地平線
這條金屬線
每一條新聞
二條橋
整條香蕉
一條尾巴
數條項鍊
每條語法規律
二十五條條文
公務員法第二十四條

例句

這兩條線將時空分成兩類上下部份
春天穿過了每一條大街，也穿過了每一條小巷。
借書辦法很快的訂出來了，一共三條。
要不是大白鵝趕來救他，可能送掉一條命。
東邊的天空，出現了一條虹。
來往的汽車像流星，街燈連接成一條金龍。
大黃狗嚇得四條腿發軟。
客廳的玻璃第四條裡，養著兩條金魚。
依憲法第四十條規定：總統依法行使大赦。
她知道那條線只是幻影
整條沙灘潔白無比，沙細如粉
七色彩虹眞美麗，彎著就像是一條香蕉。
我相信這條水溝是全校最乾淨的水溝。
相信總有一天，一條魔毯將會出現在眼前。
每條拉麵已經細如髮絲
凡經過登山關卡者，每人贈送紀念品毛巾一條與摸彩券一張，送完爲止。
泰國人對國家未來要走那條路，意見分岐。

附錄五 名詞尾語料資料庫存放資料範例（「法」第一頁）

方法★◎	505	種(38) 個(33) 套(6) 項(6) 年(1) 次(1) 兩(1) 招(1) 株(1) 組
辦法★◎	337	個(15) 套(11) 項(10) 種(9) 戶(1) 年(1) 次(1) 折(1) 屆(1)
法★	272	本(1) 份(1) 次(1) 套(1) 條(1) 種(1) 類(1)
看法★	191	個(7) 種(7) 位(5) 名(1) 派(1) 家(1) 種種(1) 樣(1) 點(1)
說法★◎	141	種(21) 名(4) 項(3) 個(2) 元(1) 件(1) 些(1)
作法★	129	種(12) 個(4) 位(1) 步(1) 套(1) 週(1) 項(1) 種種(1) 點(1)
手法★	109	種(7) 起(2) 個(1) 家(1) 項(1) 點(1)
想法★◎	104	個(10) 種(9) 次(1) 類(1)
司法★	92	種(3) 個(2) 件(1) 門(1) 路(1)
書法★◎	89	手(2) 件(2) 幅(2) 位(1) 家(1) 張(1) 張張(1) 歲(1)
做法	84	種(11) 個(7) 項(2) 些(1)
語法☆	82	個(3) 類(3) 句(1) 年(1) 系列(1) 套(1) 條(1) 層(1)
憲法★	59	個(1) 群(1)
輸入法	34	種(1)
勞基法★	29	名(1) 次(1) 個(1)
演奏法	28	種(9) 個(1) 項(1)
演算法	24	種(2) 套(1)
鬆弛法	20	個(2)
技法☆	19	種(3)
指法	19	種(2) 件(1)
療法☆	19	種(2)
刑法★	18	
文法☆	15	
民法☆	15	
身法☆	15	套(4) 個(1) 盤(1)
出版法	13	場(1)
用法☆	13	種(1)
寫法☆	12	種(3)
公司法☆	11	
交易法	11	
選罷法★	11	
工會法	10	
玩法☆	10	種(4)
魔法	10	本(1) 個(1) 種(1)
打法☆	9	把(1) 套(1) 場(1) 種(1)
佛法☆	9	個(1)
治療法	9	種(1)
畫法	9	筆(1)
國安法★	8	
著作權法	8	
解決辦法	8	項(1)
比對法	7	種(1)
保育法	7	
建築法	7	
講法	7	個(2) 種(1)
人團法	6	
教學法☆	6	種(1)
賭法☆	6	種(3)
反列法	5	項(3)
吃法	5	個(1)
戒嚴法	5	
修持法	5	種(3)
消防法	5	
國際法☆	5	間(1)
教師法	5	
組織法	5	
解法	5	個(1) 家(1)
劃分法	5	
廣電法	5	

附錄六 傳送資料流程（名詞「法」）

a. 進入「法」的語料資料庫

b. 編輯人員挑選例詞

```
*方法★◎        505  種(38) 個(33) 套(6) 項(6) 年(1) 次(1) 兩(1) 招(1) 株(1) 組
*辦法★◎        337  個(15) 套(11) 項(10) 種(9) 戶(1) 年(1) 次(1) 折(1) 屆(1)
 法★           272  本(1) 份(1) 次(1) 套(1) 條(1) 種(1) 類(1)
 看法★          191  個(7) 種(7) 位(5) 名(1) 派(1) 家(1) 種種(1) 樣(1) 點(1)
 說法★◎         141  種(21) 名(4) 項(3) 個(2) 元(1) 件(1) 些(1)
*作法★          129  種(12) 個(4) 位(1) 步(1) 套(1) 週(1) 項(1) 種種(1) 點(1)
*手法★          109  種(7) 起(2) 個(1) 家(1) 項(1) 點(1)
 想法★◎         104  個(10) 種(9) 次(1) 類(1)
 司法★          92   種(3) 個(2) 件(1) 門(1) 路(1)
 書法★◎         89   手(2) 件(2) 幅(2) 位(1) 家(1) 張(1) 張張(1) 歲(1)
*做法            84   種(11) 個(7) 項(2) 些(1)
 語法☆          82   個(3) 類(3) 句(1) 年(1) 系列(1) 套(1) 條(1) 層(1)
 憲法★          59   個(1) 群(1)
 輸入法         34   種(1)
 勞基法★        29   名(1) 次(1) 個(1)
 演奏法         28   種(9) 個(1) 項(1)
*演算法         24   種(2) 套(1)
 鬆弛法         20   個(2)

                    [總共246項，第1/14頁]          [已選 10 項]
 PgUp上一頁 PgDw下一頁 ↑上 ↓下 Enter選擇 Ctrl-End選量詞 Esc結束(不選量詞)
MAIN    F:\user\dic\NA_C        記錄 9422/33055  檔案              Num    Ins
```

c. 編輯系統根據頻率自動排序

d. 編輯人員修改順序

```
 技法☆
 指法
*療法☆          010    方法
 刑法★          020    辦法
 文法☆          030    作法
 民法☆          040    手法
 身法☆          031    做法
 出版法         110    演算法
*用法☆          070    療法
*寫法☆          050    用法
 公司法☆        060    寫法
 交易法         100    玩法
 選罷法★
 工會法
*玩法☆
 魔法
 打法☆
 佛法☆

                PgUp/PgDw上下頁 Ctrl-End結束排序 Esc重新選擇
 PgUp上一頁 PgDw(不選量詞)
MAIN    F:\user\dic\NA_C        記錄 9668/33055  檔案              Num    Ins
```

e. 編輯系統自動統計量詞資料並依頻率排序

f. 編輯人員篩選量詞

法：方法、辦法、作法、做法、手法、用法、寫法、療法、玩法、演算法

種 (91)	個 (60)	項 (20) [a]	套 (19)	×年 (2)	×次 (2)
×點 (2) [ae]	×起 (2)	×兩 (1)	招 (1)	×株 (1)	組 (1)
×塊 (1) [ae]	×種種 (1)	×戶 (1)	×折 (1)	×屆 (1)	×條 (1)
×位 (1)	×步 (1)	×週 (1)	×家 (1)	些 (1)	

g. 傳送到詞彙資料庫

◣ 名 詞 量 詞 搭 配 詞 典 分 析 ◢ F1:儲存／列印

詞項	法	編號	na0562_1	注音	ㄈㄚˇ	分析	T	獨立詞	F
▲語意	1 方法、方式								

▲凡例

說明 1：
例子：方法、辦法、作法、做法、手法、用法、寫法、療法、玩法
●一般：個、項、套、招、組　●事件：
●種類：種　●容器：
●約量：些
●辨析：

▲特例

例子：
●一般：　●事件：
●種類：　●容器：
●約量：
●辨析：

PgUp/PgDw上下 Ctrl-PgDw增刪插 Ctrl-PgUp拷貝/重排 Ctrl-Home選擇 Ctrl-End結束
MAIN F:\user\dic\NA_C 記錄 1/33055 檔案 Num Ins

附錄七 文字檔範例（量詞：「條」，名量搭配：「法」）

條ㄊㄧㄠˊ

①〔事物〕用來計量長條狀物品。例詞：四條船、一條葉脈、一條繩子、一條金帶子、一條手帕、一條鐵鍊、一條長長的影子、一條游泳褲。例句：天空的東邊出現了一條虹。

辨析：「根」和「條」都可以用來計量長條狀物品，但是有一些差別。「根」只能計量橫切面爲圓的物品，因此形狀不規則的長條狀物品只能用「條」來計量，如「一條船、一條圍巾、一條馬路」。此外，「條」傾向計量軟的或可以彎曲的長條狀物品，因此硬的長條狀物品，如「樹枝、鋼管」，就只能用「根」來計量。請參見「根」。

②〔事物〕用來計量體形或尾巴呈長條狀的動物。例詞：一條小魚、一條巨龍、整條豬、一條狗、兩條笨重的犀牛、兩條蛇、一條毛毛蟲。例句：來往的汽車像流星，街燈連接成一條金龍。

辨析：不過有些動物有尾巴卻不用「條」來計量，例如我們不說「一條貓、一條老虎」。用「條」計量的動物通常都可以用「隻」來計量。請參見「隻」。

③〔事物〕用來計量條狀地面物。例詞：這條河流、整條街、這條公路、兩條水道、一條鐵路、一條隧道、一條馬路、一條巷子、一條產業道路。例句：春天穿過了每一條大街和每一條小巷。

④〔事物〕用來計量「線」，也包括抽象意涵的線。例詞：這條線、一長條弧線、一條雙曲線、四條路線、一條墨綠的地平線、這條線索、兩條航線、這條海線。例句：這兩條線將時空分成上下兩部份。

⑤〔事物〕用來計量法律、規則、訊息。例詞：這條規定、一條規則、每一條新聞、三十條罪狀、二十五條條文、憲法第十八條、草案第二條、兩條消息。例句：依憲法第四十條規定：總統依法行使大赦。

辨析：「條」和「項」、「款」都可以用來計量條文，不過就公文的陳述次序來看，「項」是「條」下面的分項，「款」是項下面的分項，如「地方自治綱要第五十一條第一項第三款」。請參見「項、款」。

⑥〔事物〕用來計量「生命」的相關詞。例詞：一條命、一條生命、四十多條人命。例句：要不是救生員及時趕到，他可能送掉一條命。

法ㄈㄚˇ

①方法或方式。

◎方法、辦法、作法、做法、手法、用法、寫法、療法、玩法、演算法、…。指方法或方式。〔一般〕：個、項、套、道、招、組。〔種類〕：樣、式。

◎看法、說法、想法、講法、…。指意見。〔一般〕：個、項、點。〔種類〕：派、樣、式。

②法律或規律。

◎憲法、勞基法、刑法、民法、交易法、選罷法、國安法、著作權法、保育法、國際法、軍法、稅法、…。指各種法律。通常不搭配量詞。

辨析：「憲法」還可以說「一部憲法」。

◎語法、文法、句法、…。指語文的規律。〔一般〕：套、條、個。

◎佛法、魔法、…。通常不搭配量詞。

③技藝或法力。

◎技法、槍法、劍法、箭法、刀法、指法、…。〔一般〕：套。

◎書法。〔一般〕：幅、張、篇、件。

辨析：「書法」除了和上述量詞搭配之外，還有「他寫得一手好書法」這樣的說法。

20 張 290

照片 (18)
牌 (7)
門 (4)
股票 (3)
錢 (3)
卡片 (2)
成紅報圖靠雙小天方主外甘床拘者長宣耶原海病迷國報稅陽新義雷監影歷牆證
照片 (18)
牌 (7)
門 (4)
股價 (3)
票 (3)
錢 (3)
卡片 (2)
成績單 (2)
紅色 (2)
報紙 (2)
圖片 (2)
靠背 (2)
雙人床 (2)
小鈔 (1)
天量 (1)
方桌 (1)
主義 (1)
外匯券 (1)
甘談 (1)
床柱 (1)
卦 (1)
拘捕令 (1)
者 (1)
長笛 (1)
宣傳 (1)
耶穌頭 (1)
原圖 (1)
海 (1)
病歷表 (1)
迷宮 (1)
國樂 (1)
報稅名單 (1)
陽光 (1)
新聞 (1)
義賣品 (1)
雷射 (1)
監測網 (1)
影子 (1)
歷史 (1)
牆明 (1)
專輯 (12)
紙條 (6)
量 (4)
便條 (3)
單 (3)
臉孔 (3)
白紙 (2)
車票 (2)
風景 (2)
椅子 (2)
歌 (2)
遺失啓事 (2)
三角形 (1)
五彩 (1)
幻燈片 (1)
日期 (1)
古蹟 (1)
布告 (1)
安非他命 (1)
沙發 (1)
姊姊 (1)
明信片 (1)
花容 (1)
門神 (1)
宦臣 (1)
面額 (1)
唇 (1)
海灣 (1)
笑臉 (1)
陣線 (1)
措施 (1)
畫卷 (1)
視線 (1)
集 (1)
會員證 (1)
腳 (1)
圖 (1)
網 (1)
獎金 (1)
螢幕 (1)
聯盟 (1)
攝影 (1)
臉 (12)
賀年片 (6)
人 (3)
海報 (3)
鈔票 (3)
支票 (2)
光碟 (2)
直徑 (2)
書桌 (2)
畫 (2)
磁碟片 (2)
錫箔紙 (2)
大花臉 (1)
內閣 (1)
手 (1)
水準 (1)
司機 (1)
正片 (1)
收藏版 (1)
身分證 (1)
底片 (1)
治安 (1)
表 (1)
勉勵卡 (1)
拷貝 (1)
音樂 (1)
校友 (1)
留言 (1)
紙屑 (1)
剪報 (1)
啓事 (1)
畫像 (1)
賀卡 (1)
塑膠紙 (1)
經典 (1)
號碼 (1)
演唱會 (1)
網子 (1)
稿紙 (1)
頭髮 (1)
舊事 (1)
尪仔標 (1)
床 (8)
名片 (4)
毛毯 (3)
紙 (3)
慰問卡 (3)
出地圖片 (2)
信用卡 (2)
設計圖 (2)
圓桌 (2)
影本 (2)
錦緞 (2)
大頭照 (1)
切結書 (1)
文憑 (1)
牛皮卡 (1)
外親事 (1)
母行事曆 (1)
事所得 (1)
玩具 (1)
表姊們 (1)
娃娃臉 (1)
美金 (1)
倍 (1)
桌子 (1)
病床 (1)
紙頭 (1)
唱片 (1)
場次 (1)
發票 (1)
郵政 (1)
新台幣 (1)
罩蓋 (1)
雷公臉 (1)
漫畫稿 (1)
嘴 (1)
墨色 (1)
嬰兒 (1)
藥膏 (1)
妹 (1)

語料庫為本的語義訊息抽取與辨析
以近義詞研究為例

蔡美智* 黃居仁* 陳克健**
*中研院史語所
**中研院資訊所

電子郵件：tsmei@hp.iis.sinica.edu.tw

摘要

本文以近義詞研究為例，說明如何利用語料庫進行語意抽取與辨析。傳統的研究方法在舉證過程中，因為缺少豐富語料作參考，難免遺漏一些有趣的語言現象，同時也無法反應語言事實。反觀語料庫研究，雖然能巨細靡遺地記錄語言現象，提供各種數據，然而往往會為繁複的語料所困，掌握不到現象背後的眞正導因。這裡我們配合語料庫進行語意方面的研究，一方面詳細觀察詞項之間的句法功能，計算其使用頻率；另一方面從各種差異中找出基本原由，證明詞彙的句法表現取決於自身的語意特性。

1. 緒論：「熱心」與「熱情」

一般而言，近義詞的定義並不明確，而被認定為近義詞的詞項之間彼此到底有什麼差別，也沒有一個嚴格的界定方法。

以 Teng (1994)主編的中文近義詞詞典為例，書中針對漢英翻譯需要，收錄一些英文譯文相同的中文詞，視為近義詞，然後利用各自不同的用法對這些詞加以區分。例如「熱心」和「熱情」這對詞雖然都表示對事物的全心投入，可是彼此的用法並不完全相同，在名詞、形容詞、副詞、動詞四種用法[1]當中，「熱心」沒有名詞功能，無法接受定語修飾(例1)；反之，「熱情」沒有動詞功能，不能後接賓語(例4)。至於兩者都適合的用法，只有形容詞和副詞兩種(例2-3)。

(1) 懷著滿腔(*熱心+熱情)去前方勞軍。

(2) 他十分(熱心+熱情)。

(3) 他總是(熱心+熱情)地幫助別人。

(4) 她一向(熱心+*熱情)慈善事業。

由下表，我們可以清楚看出這兩個詞項在用法上的異同。

(5)

用法＼詞項	熱心	熱情
名詞	-	+
形容詞	+	+
副詞	+	+
動詞	+	-

這種比較詞彙功能的方法，確實提供了一套辨識近義詞的具體標準，可是限於例句的局部性，觀察結果難免有所偏失，例如Teng (1994)根據例(6)，判定「熱情」比「熱心」適合接上定語標誌「的」修飾後面的名詞組，事實上我們發現，只要將句中內容稍作修改，「熱心」一樣可以擔任定語功能（例7）。

(6) (*熱心+熱情)的觀眾為精彩的表演熱烈鼓掌。

(7) 熱心的觀眾為這次的表演四處奔走。

其次，Teng (1994)作了一個有趣的觀察，那就是當介詞「對」引出的對象是人的時候，搭配的謂語以「熱情」為宜(例8a)；反之，如果引出的對象是物的話，像「朋友的事」，那麼「熱心」就較適合(例8b)。

(8) a. 他對人很(*熱心+熱情)。

b. 他一向對朋友的事都很(熱心+*熱情)。

不過，只要我們觀察更多的語料便可以發現，像「人」與「物」這樣的對立關係，不僅只存在於事件所涉及的對象上面，從句子的主語成份，也看得到類似的現象。下面三組例句顯示，「熱心」和「熱情」都可以自由搭配像「民眾、男人、各位貴賓」一類的屬人主語，可是只有「熱情」能搭配「氣氛、陽光、太陽」等非屬人主語。

(9) a. (熱心+熱情)民眾送兩箱草莓慰問。

b. (*熱心+熱情)的氣氛備受世界各地的遊客青睞。

(10) a. 男人不可太(熱心+熱情)。

b. 陽光不再像以前那麼(*熱心+熱情)。

(11) a. 感謝各位貴賓遠道而來，(熱心+熱情)參與。

b. 太陽還是衝破烏雲，(*熱心+熱情)的放出光和熱。

另一個問題是，這種針對幾個例句進行觀察比較的研究方法，可以得到像表(5)的結果，知道一個詞項具備哪些功能，不具備哪些功能，然而該詞項最重要的功能是什麼，以及各項功能間的使用比例為何，卻無從知曉。

諸如此類的問題，唯有配合語料庫的應用，才可以獲致圓滿的解答(參見黃居仁1995)。我們利用中央研究院詞庫小組所建立的兩百萬語料庫(黃居仁、陳克健等1995)進行搜尋，總計取得用例「熱心」71條，「熱情」77條，下表記載兩者各項功能的使用情形。

(12)

用法＼詞項	熱心 (71)	熱情 (77)
名物化	5 7.0%	39 50.6%
定語 ∅	7 9.9%	4 5.2%
的	7 9.9%	14 18.2%
狀語 ∅	20 28.2%	4 5.2%
地	10 14.1%	2 2.6%
謂語 ∅	14 19.7%	14 18.2%
賓	7 9.9%	
介	1 1.3%	

由表中提供的數據可以看出，這兩個詞項除了謂語及物用法[2]及名物化用法[3]有明顯差異之外，其他功能的分佈情形也頗有距離。儘管兩者都具備定語和狀語功能，但事實上「熱心」用作狀語的次數是定語用法的兩倍，而「熱情」恰恰相反，狀語用法僅及定語用法的三分之一。兩者在用法上唯一差不多的地方，只有謂語不及物功能一種。另外值得注意的是，「熱心」仍然有少數幾個名物化用例，並不像預期中的絕對禁止。

由此可見，語料庫應用能幫助我們進一步了解近義詞之間的關係。這些詞事實上只有在某些用法底下表現類似，整體而言仍是很不相同。接下來我們將利用語料庫提供的訊息，觀察另一對近義詞「高興」和「快樂」句法行為的異同，並探究其原因。

2. 語料庫應用實例

在兩百萬詞中央研究院平衡語料庫中，「高興」和「快樂」出現的次數分別為280與365，都是使用頻率相當高的動詞，由中文詞知識小組(1994)編制的新聞語料詞頻來看，兩者都是收錄的28326目詞中，前四千個最常用的詞彙。這對詞無論按 Vendler (1967)、Teng (1975)或 Smith (1991)提出的分類原則，都是狀態動詞，具有可以接受程度副詞「很」修飾的語法特徵(例13)。

(13) a. 她很快樂。

b. 她很高興。

2.1 「高興」不等於「快樂」

儘管語意和用法頗類似，兩者的句法表現並不盡相同，如例(14)所示，「高興」可以後接賓語子句，「快樂」就不行。

(14) 她很(高興+*快樂)張三來了。

所以依照中文詞知識庫小組(1993)採用的論元結構標準，這對詞分別列入兩個不同的類：「高興」屬狀態句賓述詞，而「快樂」屬狀態不及物述詞。

可是我們發現除了後接賓語子句這項功能以外，這對詞之間還有其它不同的用法，譬如例(15)的名物化功能，例(16)的定語功能，例(17)的結果補語功能，以及例(18)與動貌標誌「了」搭配，兩者的表現一概相反。

(15) 人有追求(*高興+快樂)的本能。

(16) 如何做個(*高興+快樂)的上班族。

(17) 他看得很(高興+*快樂)。

(18) 客人(高興+*快樂)了會賞你錢。

這些對比並非「狀態句賓」和「狀態不及物」兩個次類劃分所能預測的，因此我們決定先從語料庫中觀察這對詞的實際使用情形，再比較兩者各自特有的句法功能，從中抽離出導致差異的語意因素。

2.2 句法功能與範例

我們將使用平衡語料庫的多項配備，包括詞類標記、關鍵詞檢索、排序、過濾、詞類統計、列印等多種功能，比較這對詞的句法功能，包括謂語、補語、狀語、定語及名物化等用法。

首先，我們利用關鍵詞搜尋(KWIC)功能找出所有包含關鍵詞的例句，接著利用詞類標記統計各種功能的出現次數。例如標有<+NOM>的詞項具名物化功能，出現於<DE>後面的則具補語功能。其他僅由關鍵詞標記無法辨識的功能，如狀語和定語，則利用語料庫可預設視窗範圍的功能，以及詞類檢視、濾除等功能來判斷。各種功能的分佈情形簡要敘述如下：

2.2.1 謂語功能

在語料庫中，「高興」的標記為VK，擔任謂語的情形計有224次，是該詞數項功能中最重要的一項，使用率為百分之八十；標記為VH的「快樂」則有119次充當謂語，使用率約百分之三十。兩者用作謂語的時候，絕大部分都呈現不及物用法(例19)。可是「高興」卻獨自擁有及物用法，後面可以接上子句(例20)。

(19)　出爐了，大家吃得津津有味，高興又快樂。

(20)　我們很高興創刊號終於發行了。

再者，兩個詞後方都可以直接加上補語成份，如例(21)中的「萬分、無比」，或是形成由「得」字帶出的補語結構(例22)。

(21) a.　他眞是高興萬分。

b.　全家一定快樂無比。

(22) a.　我和妹妹都高興得大聲叫好。

b.　一聽到披薩，眞是快樂得不得了。

2.2.2 補語功能

「高興」和「快樂」也都可以出現在「得」字後面，發揮補語功能(例23)。不過，兩者的使用頻率都不高，一個是百分之三，另一個也只有百分之五。

(23) a. 他看得很高興。

b. 爸爸忙得很快樂。

2.2.3 狀語功能

這兩個動詞扮演狀語的次數都比扮演補語的情形來得多，總計「高興」有47條用例，「快樂」有30條，兩者的差別在於「高興」作狀語時一定帶有狀語標誌「的」或「地」(例24)，因為不帶標誌的話,就會被理解成謂語的句賓用法。至於「快樂」帶標誌也可以(例25a)，不帶標誌，直接出現在謂語前方進行修飾也可以(例25b)，帶標誌的情形大約是不帶標誌的兩倍。

(24) 小明高興的跳起來。

(25) a. 她快樂地叫出來。

b. 快樂過新年！

2.2.4 定語功能

在定語功能方面，「高興」的用例為零，「快樂」則有116次之多，使用率達百分之三十，和謂語一樣重要。「快樂」扮演定語功能的時候，如下則例句所示，定語標誌「的」可有可無，但是和「的」一起出現的頻率較高，約為不帶「的」的兩倍。

(26) a. 露出了快樂的笑容

b. 在一起的種種快樂情景

2.2.5 名物化功能

與上一項定語功能的情況類似，兩個動詞名物化的頻率相差十分懸殊，「高興」全部只有一個名物化的用例(例27)，出現率還不到百分之一，至於「快樂」名物化的情形則相當普遍，既可以擔任主語(例28a)，也可以擔任賓語(例28b-c)，出現率超過百分之二十。

(27) 不過卻帶有一些高興，因為終於可以回到自己的家了。

(28) a. 快樂在哪裡?

b. 人有追求快樂，逃避痛苦的本能。

c. 祕密組織以快樂和暴力為尚。

由以上的觀察，我們可以得到一個初步的認識，那就是「高興」和「快樂」這兩個詞雖然意義相近，而且都具有謂語、補語、狀語等用法，但是分佈比例卻大不相同。以下將語料庫研究所得的數據製成圖表：

(29)

功能＼詞項	高興 (280)	快樂 (365)
謂語	224 (80%)	119 (32%)
補語	8 (3%)	17 (5%)
狀語 地	47 (17%)	19 (5%)
狀語 ∅		11 (3%)
定語 的		77 (21%)
定語 ∅		39 (11%)
名物化	1 (0.3%)	83 (23%)

從表中可以清楚看出，除了能否後接子句之外，「高興」的諸多功能當中，以謂語的用法最具代表性，而定語和名物化的用例則微乎其微。相反地，「快樂」除了謂語以外，也經常有擔任定語或名物化的情形。這些句法上的差異，不是「狀態句賓」和「狀態不及物」兩個次類劃分所能預見的，因此我們將在下一節進一步探索，是什麼因素致使「高興」和「快樂」句法結構互異。

3. 詞彙語意特徵研究

我們再次將收集到的語料依照語言現象分門別類，諸如詞項扮演的句法功能，表述的事件動貌，搭配的主語屬性，及建構的句型，以便從中抽離出導致上述許多語法差異的語意特性。

3.1 狀態有無變化 <±change of state>

首先，我們利用句法功能分佈與事件動貌，證明「高興」和「快樂」表述不同的事件類型，前者隱含某種狀態變化，後者屬於沒有變化、均質的狀態。

3.1.1 句法功能

綜觀第二節的功能分佈介紹，「高興」和「快樂」在用法上有以下三大不同點：第一，「快樂」名物化的情形相當普遍，「高興」則有困難(例30)。第二，「快樂」可以扮演定語功能修飾名詞組，「高興」則不行(例31)。第三，「高興」後面可以接句子，「快樂」則無此用法(例32)。

(30) 人有追求(*高興+快樂)，逃避痛苦的本能。

(31) 如何做個(*高興+快樂)的上班族。

(32) 我們很(高興+*快樂)創刊號終於出來了。

這些對比顯示「高興」和「快樂」屬於不同的事件型態，「高興」的動作性較強，後面接的句子雖然不是眞正的受事賓語，但也是促使某種狀態發生的導因(例32)，「快樂」則名詞性較強，可以指涉認知世界中的某種特性(例30)，也可以次類劃分一個集群(例31)。因此就語意觀點而言，我們可以說「高興」含有狀態變化(change- of-state)的語意特性，「快樂」則是穩定、均質的狀態(homogeneous state)。

3.1.2 事件動貌

一旦弄清楚這兩個近義詞間不同的語意特性，下面的語法現象都可以得到合理的解釋。例(33)中「高興」因為詞彙語意本身即涉及狀態改變，所以能夠帶上

完成貌標誌「了」，並搭配表瞬間動作的時間子句「聽了」。反之，同樣的搭配完全不適合「快樂」這種穩定均質的狀態。

(33) a. 客人(高興+*快樂)了會賞你錢。

b. 父親聽了很(高興+*快樂)。

有關時間副詞的搭配方面，也出現類似的對比情形，表瞬間狀態的副詞「正」適合修飾「高興」，而表恆久狀態的副詞「永遠」則適合修飾「快樂」。

(34) a. 我們談得正(高興+*快樂)，突然…

b. 永遠(快樂+*高興)

在補語的搭配上，兩者都可以接表程度的補語，如例(35a)中的「不得了」，但只有「高興」容許後面接上由句子組成的結果補語，如例(35b)中的「大聲叫好」，表示某事件發生之後所產生的新事件。

(35) a. 一聽到披薩，眞是(快樂+高興)得不得了。

b. 我和妹妹都(高興+*快樂)得大聲叫好。

擔任狀語的時候，兩者和被修飾的謂語呈現不同的選擇限制，「高興」很適合搭配由動作動詞組成的謂語，如例(36a)中的「叫出來」，但遇到像例(36b)中動作性低的事件「成長」就不太自然。至於「快樂」則沒有這層限制，修飾的事件無論是動態或靜態都可以。

(36) a. 她(快樂+高興)地叫出來。

b. 全家人(快樂+*高興)地一起成長。

擔任補語的時候，兩者和前面的謂語同樣表現出不同的選擇限制，「高興」搭配狀態動詞或動作動詞都可以，分別表示狀態的程度和動作的結果，如(37a)中的「忙」和(37b)中的「看」。「快樂」則只適合搭配狀態動詞表程度，搭配動作動詞的話，詮釋起來會很奇怪。

(37) a. 爸爸忙得(很快樂+很高興)。

b. 他看得(很高興+*很快樂)。

(38) a. 玩/2、跳/1、吃/1、看/1、說/1、談/1、欣賞/1 + 得 + 高興

b. 過/8、活/6、玩/2、忙/1 + 得 + 快樂

換句話說，只有「高興」可以表示一個事件過後產生的新狀態。事實上，根據我們對實際語料的統計，如例(38)所示，「高興」當補語時搭配的謂語對象多是「跳、吃、看、說」一類的動作動詞，相反地，「快樂」搭配的對象則絕大部分是像「過、活」這種描寫持續狀態的動詞。

上述各種現象從不同的角度證明，「高興」和「快樂」之間的許多句法差異，都與兩者本身的事件型態有關，「高興」意味著狀態有所變化，而「快樂」則屬穩定均質的狀態，無涉任何變化。

3.2 意志能否控制 <±volition>

接下來我們從語料中透過主語屬性以及句型類別，歸納出另一項可以區別「高興」和「快樂」的語意特性，即其本身所表述的狀態能否由主觀意志控制。

3.2.1 有生主語

「高興」和「快樂」與句中主語也表現出不同的選擇限制，如例(39)所示，「高興」可以接受屬人代名詞「她們」擔任主語，但無法接受無生名詞「生活」當主語。至於「快樂」搭配兩種屬性的主語都沒有問題。

(39) a. 她們(快樂+高興)嗎？

b. 心胸開闊，生活(快樂+*高興)最重要。

3.2.2 句型

在搭配句型方面，兩者也呈現出以下兩點差異：第一，只有「高興」可以形成命令句(例40)；第二，只有「快樂」才適合祝願的句型(例41)。

(40) a. （高興+*快樂）一點！

b. 別(高興+*快樂)！

(41) a. 祝你(快樂+*高興)！

b. 媽媽過節(快樂+*高興)！

由於命令句是要求聽話者作出相關的反應，聽話者顯然具有執行能力，而祝願語

係純屬說話者的願望，無關聽話者的執行能力。因此，既然「高興」能與命令句相容，本身應該是個主觀意識可以控制的狀態。反之，「快樂」適合祝願語而排斥命令句，所以應為主觀意識所無法操控。

這項推論可以解釋前一小節有關主語屬性的觀察，「高興」之所以必須搭配屬人主語，便是因為那才具有控制的能力。「快樂」因為屬自然生成的狀態，不受外力左右，所以搭配的對象是有生主語也好，是無生主語也行。

下面幾個例句分別從不同的角度來支持這項論點。首先，兩個近義詞之間只有含意志控制特性的詞項「高興」才能搭配情態動詞「應該、要」(例42)。

(42) a. 你應該(高興+*快樂)。

b. 要我(高興+*快樂)就陪我打麻將！

其次，也只有「高興」這種心理狀態，當事人能夠清楚認知，進而以言行來表達(例43)。

(43) a. 打從心裡（高興+*快樂）。

b. 研究人員表示(高興+*快樂)。

再者，可以憑常理判斷，作出適當情緒反應的，同樣也只有「高興」（例44）。

(44) a. 陳菊出獄不值得(高興+*快樂)。

b. 為陳老師(高興+*快樂)。

最後，只有「快樂」這種心理的自然反應，意志所無法控制的情緒，才可能連當事人都沒有辦法掌握，進而發生自問或者弄錯的情形（例45）。

(45) a. 可是，我(快樂+*高興)嗎？

b. 他以為自己非常(快樂+*高興)。

根據上述的考量和比較，我們得以從語料中複雜的語法現象，歸納出兩項基本的詞彙語意特性，一為狀態有無發生變化<±change-of-state>，二為意志能否自由控制<±volition>，而兩者都與動詞本身所表達的事件類型有關。這一點印證了Pustejovsky 1991，Levin 1993 等所採用的詞彙-句法互動原則，並肯定了詞彙語意在句法中的主導地位。

4. 結論

以上我們利用語料庫的多項功能配備，成功地剖析出致使近義詞之間彼此句法行為迥異的詞彙語意特性。這個示範說明語料庫應用是當前語言學研究的利器，無論是語法現象的考察，語意因素的探索，或是語用效應的評估，我們都能夠賴以進行最詳盡的比對和觀察，並且由統計數據了解語言現象的全貌，這一點是傳統方法中檢驗有限幾條例句所無法探究的。

附註

[1] Teng (1994)顯然依照英文文法的觀點，將「熱心、熱情」當作形容詞，它們除了本身的用法以外，還兼具其他詞類的功能。我們則主張將詞類和句法功能的界限劃分清楚：一種詞類可以扮演多種句法功能，如狀態動詞「熱情」可以同時具有名物化、狀語、謂語等多種功能；同理，一種句法功能可以由不同的詞類來扮演，如「昨天觀眾一直熱情地鼓掌」一句中時間名詞「昨天」、副詞「一直」、動詞「熱情」等都可以勝任狀語功能。

[2] 即Teng (1994)所謂的動詞用法。

[3] 「熱情」名詞性用法相當突出，佔總出現率一半以上，在詞庫制定的詞類分析裡面便以多重詞類看待，詞類標記可能是Na名詞或是VH狀態不及物述詞。另外也有人主張將之視為名詞，因為就構詞率而言，「熱情」屬偏正結構複合詞，理應保有中心語「情」名詞的特性。可是反觀「熱心」，雖然中心語「心」也是名詞，但屬性卻毫無疑問是動詞。因此我們採用單一詞類觀點，將這個詞視為狀態不及物述詞，至於這三種作法哪一種比較妥當，這裡暫時不予討論。

參考書目

中文詞知識庫小組. 1993. *詞庫小組技術報告93-05 中文詞類分析.* 台北南港: 中央研究院.

中文詞知識庫小組. 1994. *詞庫小組技術報告94-01 中文書面語頻率詞點(新聞語料詞頻統計).* 台北南港: 中央研究院.

黃居仁 1995. 科技整合與整合科技—談計算語言學與語料庫語言學之角色與發展. 語言學門現況與發展研討會. 台北: 國立師範大學.

黃居仁, 陳克健, 張莉萍, 許蕙麗 1995. 中央研究院平衡語料庫簡介. *第八屆計算語言學研討會論文集.* pp. 81-99. 內壢: 元智工學院.

LEVIN, B. 1993. *English Verb Classes and Alternations: A Preliminary Investigation.* Chicago: University of Chicago Press.

PUSTEJOVSKY, J. 1991. The Generative Lexicon. *Computational Linguistics* 17.4.
SMITH, C. 1991. *The Parameter of Aspect.* Dordrecht: Kluwer.
TENG, S.-H. 1975. *A semantic study in transitivity relations in Chinese.* Ph.D dissertation. University of California at Berkeley.
TENG, S.-H. 1994. *Chinese Synonyms Usage Dictionary.* Taipei: Crane Publishing.
VENDLER, Z. 1967. *Linguistics in Philosophy.* Ithaca: Cornell University Press.

介詞翻譯法則的自動擷取

張俊盛、徐瑞宏、陳惠群
國立清華大學資訊科學學系
新竹市 30043 光復路二段 101 號
Email: jschang@cs.nthu.edu.tw

摘要

介詞在自然語言的處理上一直扮演著重要的角色；而介詞所表示的不同語意關係，更直接影響了介詞的翻譯。我們將設法由機讀語料中擷取出介詞的翻譯法則。

我們將以介詞 with 為例，先依其不同的翻譯特性加以分析，共分成片語、直接受詞、功能、原因、情狀、屬性、共事、對象、關係等九類。其中對直接受詞類，我們利用了句法的次分類資訊；功能及屬性兩類，我們則提出先分類再利用辭彙語意結構來選擇翻譯時所選用的辭彙。我們引入屬性和關係兩類是用以描述 V-O-P-N 結構中的 O 和 N 兩者的關係，以彌補格位分析的不足。

實驗以朗文當代英漢雙語詞典中的雙語例句為訓練語料，抽取句子的 V-O-P-N 結構，觀察其在同義詞詞林的類語意碼，再利用決策串列(decision list)來產生法則。外部測試的結果獲得超過 75%的高正確率。

一、前言

介詞是虛詞(function word)，在句中表示詞與詞或詞與句子間的關係，因此介詞本身與其所構成的短語往往有重要作用。其種類不多，但使用量很大。編譯館國中三年級英文課本有 36.4%的句子至少包含一個介詞(Chen,1991) 。介詞的用法也相當複雜，多數的介詞有多個詞義：

(1-1) He bound it on with rope. (他用繩子把它捆綁起來)

(1-2) Will you walk up to the shop with me? (你和我一塊到那店舖好嗎?)

朗文當代英漢雙解詞典(Longman English-Chinese dictionary of contemporary English)中， with 就有二十個詞義。此外，介詞形成的結構分歧，例如介詞組連繫歧異的解決(PP attachment disambiguation)也仍是研究重點。

學習外國語時，介詞用法往往令人困惑。因為母語中介詞的用法及現象，常跟外國語的介詞有很大的差異。以中文和英文為例，主要有以下幾點(Tang,1979)：

1. 英文介詞是虛詞，而中文裡，其語意卻常帶有動詞功能，甚至根本就是動詞。如 at (在)、 to (到)、 into (進)。一旦轉化為不同詞性，更增翻譯的困難。
2. 英文介詞組由介詞及介詞賓語的名詞片語組成。而中文常要再加上一個像「上面」 、「裡面」的定位語 (localizer) 。例如 on the desk (在桌子上面)。
3. 英文介詞組的語意差異由介詞本身即可反應出。中文卻需要定位語的幫忙。例如： beside the desk (在桌子旁邊)、 in the desk (在桌子裡面)。
4. 英文介詞的選擇有時是由其介詞賓語的名詞片語決定。例如： at 6 o'clock 、 on Friday 。有時則由動詞或形容詞決定。例如： arrive at 、 angry

with 。此種搭配 (collocation)現象在中文中不存在。

5. 英文介詞常有多個意思，但不同的介詞有時卻表示相同的意義。例如： 10 minutes before / to 6 o'clock 。中文少有類似的情形。

由此可見，不解決介詞的問題，翻譯品質難以提升。

二、有關介詞翻譯的研究

(一)介詞的分析

(A.)格位語法(case grammar)

格位語法結合句法及語意解釋，期能了解深層及表面結構(Cook, 1989)(Rich and Knight, 1991)。 Ravin (1990)用字典中「 to VERB with NP 」形式的解釋，解決介詞定義歧異。他依介詞組不同的事態(states of affairs)，對 with 的意義予以分類：

1. USE ：相當於「 by means of 」「 using 」。如「 to obscure with a cloud 」「 to surround with an army 」。又分為 USE-of-Instrument 等五小類。
2. MANNER ：此類用法相當於副詞。如「 to anticipate with anxiety 」即相當於「 to anticipate anxiously 」。又分為 Intention-as-MANNER 等五小類。
3. ALTERATION：相當於「 make 」「 put into/onto 」。如「 to mark with a bar 」「 to impregnate with alcohol 」。又分 ALTERATION-by-Marking 等四小類。
4. CO-AGENCY 或 PARTICIPATION：相當於「 and 」。例如「 to combine with other parts 」。
5. PROVISION ：相當於「 give 」。例如「 to fit with clothes 」。
6. PHRASAL ：即片語。

Ravin 分析的沒有考慮連繫到名詞的可能性。此外，分類也有許多模糊不清之處。如 USE-of-Substance 跟 ALTERATION 的分界就不明顯。

Chen(1991)大致遵循格位語法概念。介詞組做修飾語時，有多種論旨角色

(thematic role)。依其角色，分為 Agent 、 Source 、 Goal 、 Location 、 Instrument 、 Path 、 Benefactive 、 End 、 Extent Time 、 Point Time 十類。 Durand(1993)用修飾語的語意關係(semantic relation for modifier)對介詞組分類；其依據不脫格位語法範疇。

雖然格位分析對介詞組做了區分，卻無法和實際翻譯有較直接的關聯：

(2-1) The cars were all bedecked with flowers. (車子都裝飾上花)

(2-2) She covered her ears with her hands. (她用手捂住耳朵)

這兩句都歸在 Object 類，但一譯「上」，一譯「用」。又如：

(2-3) Her hair became grey with the passing of the year.

(2-4) a child with a dirty face

格位分析並沒有分類可反應出(2-3)become 和 pass 間的關係。(2-4)child 跟 face 間的關係，和名詞間的語意網路有關，同樣也無法依格位分析加以分類。

(B.)朗文詞典的分析

在朗文當代英漢雙解詞典中，對 with 一字的分析極細微，有 20 個不同的意義：

1. 跟、同、和...在一起、帶著、連、加上： staying with a friend (待在朋友那兒) 、 living with one's children (跟自己的孩子住在一起)
2. 有、顯出： a book with a green cover (綠色封面的書)， a child with a dirty face (臉髒的小孩)， a factory with its chimney smoking (煙囪在冒煙的工廠)
3. 用： fight with a sword (用劍打鬥)， to hear with one's ears (用耳朵聽)
4. 用；拿(指材料、內容)： a cake made with eggs (雞蛋做的糕點)
5. 支持；贊成；向著： to vote with the government (投政府的票)
6. 跟；與；和(指對抗)： to compete with foreign company (跟外商競爭)
7. 順著；跟著： to sail with the wind(順著風航行)、 carried along with the crowd (隨著人群移動)
8. 隨著...；與...同時： Her hair became grey with the passing of the year (隨著歲月的消逝，她的

頭髮變白了)

9. 跟；與；和(對同等的比較)： to compare chalk with cheese (拿乾酪來和粉筆比較)、 to match a coat with a skirt (裙子配外套)、 level with the street (與街成一水平)
10. 跟；與；和(指分離)： to part with money (掏錢)、 to break with the past (和過去斷絕關係)
11. 雖然...可是；儘管...可是： With the best will in the world, I can't make her like me (我好意去接近她，但總是得不到她的歡心)
12. 因為；...得...；因為有： singing with joy (高興地歌唱)， grass wet with rain (被雨淋濕的草地)、 eyes bright with excitement (眼睛閃著興奮)、 With 3 children we cannot afford new furniture (因為有三個孩子，我們買不起新傢俱)
13. 交給...看： to trust someone with a secret (把秘密告訴某人)
14. 有關；至於；對於： Be careful with that glass (那個玻璃杯要小心使用)
15. 跟；與；和(指連接)： connect with
16. (用於命令)： Down with the school (打倒學校)
17. 由...選出： The decision rests with you (由你決定)
18. in with (與朋友鬼混)
19. with it (指穿著、思想、行為等方面很入時)
20. with me/you (聽懂我的/你的話)： Are wou still with me? (你還明白我的話嗎)

針對介詞的機器翻譯，我們認為上述分類需要做調整：

1. 相同翻譯的解釋合併：以 6.及 9.為例，都譯成「跟」，似乎沒有必要區分。
2. 不同翻譯的解釋應區分：以 12.為例，當副詞的「得、地」和說明原因的「因」，翻譯上有明顯差異。加上「得」可對應格位語法的「 Object 」及 Ravin 所提的「 Manner 」，而「因」卻找不到對應的分類，證實有細分的必要。

(二)介詞翻譯知識的來源

(A.)手工撰寫的法則

Chen (1991)的作法中，動詞組、介詞組、及名詞組須先給予語意特徵 (semantic feature)，例如 the man who he met in the park 中的 met 、 in 、 park 的語意特徵分別為

$$\text{met}:\begin{bmatrix}\text{cat}:\text{tv}\\ \text{subj}:\begin{bmatrix}\text{feature}:\text{animate}\\ \text{case}:\text{agent}\end{bmatrix}\\ \text{obj}:\begin{bmatrix}\text{feature}:\text{animate}\\ \text{case}:\text{patient}\end{bmatrix}\\ \text{chinese}:\text{看見}\end{bmatrix},\ \text{park}:\begin{bmatrix}\text{cat}:\text{noun}\\ \text{num}:\text{singular}\\ \text{feature}:\text{location}\\ \text{chinese}:\begin{bmatrix}\text{cl}:\text{座}\\ \text{n}:\text{公園}\end{bmatrix}\end{bmatrix},\ \text{in}:\begin{bmatrix}\text{cat}:\text{prep}\\ \text{chinese}:\begin{bmatrix}\text{cl}:\text{在}\\ \text{c2}:\text{X}\\ \text{c3}:\text{裡}\end{bmatrix}\\ \text{n}:\text{X}\end{bmatrix}$$

動詞組除語意特徵外，又分 23 次分類(subcategory)，應用在 $\overline{X}$ 理論(Sells,1985)，決定介詞組的用法。依此撰寫句法規則(syntactic rule)和轉換規則(transfer rule)。

(B.)例句資料庫(example database)

例句為本的作法，知識來源是例句資料庫及同義詞典(thesaurus)。 Sumita 和 Iida(1992)用英日文 ATR 語料庫， 17,000 句(270,000 字)，限定在研討會註冊的對話。

(C.)目標語語料

許(1994)先用雙語語料窮舉目標語的可能翻譯，再用一千萬字中文目標語語料，經斷詞及詞性標注，統計中文詞頻，計算考慮距離的二元接續表(distance bigram)，輔以同義詞詞林的語意碼，彌補取樣不足，最後以機率模型解決辭彙選擇。

(D.)語意網路(semantic network)

語意網路如 WordNet ，儲存名詞，動詞，形容詞及副詞同義群(synonym group)間的反義詞(antonym)、下義詞(hyponym)、同義詞(synonym) …等關係。朗文英漢雙解多功能分類詞典(Longman Lexicon of Contemporary English)及中文的同義詞詞林，將每個字分別給予語意碼，建構成階層式結構，也視為語意網路。

(E.)中英文語意碼

中英文語料在訓練時，不易涵蓋所有辭彙，導致翻譯時不易找到對應辭彙。因此在語料為本的機器翻譯中，適當的把辭彙分類，以分類來處理，有助於解決一詞多義問題，降低參數空間，也降低儲存空間而提升搜尋速度(柯,1993)。 Sekine 和 Tsujii(1995)曾提到，辭彙根據語意分類，特別是根據主題的分類，有助於解決句法歧異。類似作法如 Guthrie(1991)、 Yarowsky(1992)、 Chen 和 Chang(1995)。

本實驗先透過辭彙語意分析前處理，獲得實詞(content word)的英文語意碼及中文語意碼，這些碼分別取自朗文英漢雙解多功能分類辭典和同義詞詞林。朗文以 14 個主題(subject)來編排，每個主題分 7 至 12 個標題(title)，每個標題含 10 至 50 個相關字的集合(set)。總計在朗文中有 2504 個集合。詞林的編排，則以詞義為分類原則。將語意碼分成大、中、小三級，包括 12 大類， 94 中類， 1428 小類。以「狗」為例，有兩個語意碼： Bi07(B 類屬「物」類， Bi 屬「動物」類， Bi07 則是「豬狗兔」)和 Dd15(D 類屬「抽象事物」類， Dd 屬「性能」類， Dd15 屬「名稱姓名生肖性別」)。

(F.)衍生辭彙理論 (Theory of Generative Lexicon)

Pustejovsky(1991)提出辭彙語意結構(qualia structure)，每個字的語意就其組成(Constitutive role)、形態(Formal role)、功能(Telic role)、生成(Agentive role)四方面加以討論；以「書」為例，其辭彙語意結構為 (Pustejovsky,1993)：

$$\begin{bmatrix} \text{book}(x,y) \\ \text{CONST} = \text{information}(y) \\ \text{FORMAL} = \text{physobj}(x) \\ \text{TELIC} = \text{read}(T,W,y) \\ \text{AGENTIVE} = \text{write}(T,z,y) \end{bmatrix}$$

物體又分容器(container)、工具(instrument)、空間物體(figure-ground object)等型

態。然後又分別定義詞項概念範例(Lexical Conceptual Paradigms, LCP)，使名詞的辭彙語意結構可預設常用的功能及生成動詞。如 door 是空間物體，因此可優選 paint 及 scrub 等作用於具象物體的動詞，也可優選 walk through 和 fill 之類的空間動作。

(G.)辭彙句法資訊

前面提到 Chen(1991)將動詞分成 23 個子分類，用以決定介詞組該連繫到名詞還是動詞。由其實驗中可發現，辭彙在句法上的次分類資訊，對翻譯有一定程度的幫助。在分析介詞 with 時，我們就利用了朗文詞典中的句法分析代碼。

(三)介詞翻譯的計算模式

(A.)法則式的作法

Chen (1991)的作法是用字的語意特徵等資訊以及人工撰寫的法則，而在翻譯時應用符合條件的法則。這要花費很多時間在撰寫法則以及語意特徵分析上，法則涵蓋範圍也有限，而且未解決一詞多義的問題。 Durand(1993)的作法跟 Chen 很接近。

(B.)例句式的作法

Sumita 和 Iida(1992)的例句為本翻譯系統，英譯日的名詞性(adnominal)介詞正確率 87%，副詞性(adverbial)介詞更達 90%。但用英日文 ATR 語料庫範圍較小，因此旦沒有對應例句時，就容易錯誤。且其對介詞連繫等問題，採前處理方式事先解決。

(C.)語料式的作法

許(1994)由目標語出發，從所有可能的翻譯中，找出最可能的組合。例如：

(2-5) I ate a fish with a fork.

從中擷取出的 S-V-O-P-N 單元結構為

E = {I,eat,fish,with,fork}

t("I")={我，余，吾，本人}, a("I")=4
t("with")={用，和，以，跟，支持，隨著，.....}, a("with")=12
t("eat")={吃，吃飯，腐蝕，侵蝕}, a("eat")=4

t(e)是英文詞經辭彙對應的中文集合， a(e)是指英文詞 e 譯成中文時，所有可能的個數。另有一組規則來變化翻譯辭彙的位置。處理後，從解答空間 T(E)中找出最佳解：

T(E)={(我，用，刀叉，吃，魚)，(我，和，刀叉，吃，魚)，(我，以，刀叉，吃，魚).....}

此作法的問題在於介詞對應的中文很難列舉, 參數計算空間太大, 需要大量語料。

(四)目標語的辭彙選擇

我們將辭彙選擇問題分三層次：最簡單的層次是由互等的翻譯中，選取較佳者：

(2-6) I played a game with her.

譯成「我跟她一起玩遊戲」或「我和她一起玩遊戲」。我們也常省略中文虛字：

(2-7) Glass; handle with care.

其譯文「玻璃，小心地搬運」中的「地」往往省略。辭彙選擇中最簡單的層次，只要從可能的辭彙中選擇一個。選擇時，用許(1994)的作法，應可得到流暢的翻譯。

第二層次，是意思正確的轉換。例如：

(2-8) Every night she finished her prayers with a chaplet.

with 在目標語中會轉為動詞。雖可譯成「她每晚用唸珠唸完禱告」，但常會選擇和「唸珠」搭配的功能動詞 (telic verb) ，而譯成「她每晚數著唸珠唸完禱告」。功能動詞可由前述辭彙語意結構中得到。這層次的辭彙翻譯，無法簡單直接地找到。

由於語言風格(style)不同，常為求詞句順暢，而在目標語中配合上下文來進行詞彙修飾，這是最複雜的層次。因來源語及目標語常會發生語意相同而用詞不盡相同、來源語衍生度(fertility) 大於一、零代詞 (zero ananphor)(陳, 1994)等等的現象：

(2-9) He acted with a Shakespeare company.

可直譯為「他跟一個莎士比亞劇團一起演出」，但此句常轉為「他是莎士比亞劇團的演員」，把動詞「act」轉為名詞「演員」。如此難由目標語中找到介詞的翻譯。

(2-10) The ladies were ablaze with jewels.

(2-11) He was ablaze with anger.

不譯為「女士們因珠寶而發光」和「他因憤怒而激動」；而是「女士們滿身珠光寶氣」和「他怒容滿面」。同時也發現在句型、用字一致下，翻譯也非一成不變；因介詞賓語不同，會有不同的翻譯。另外還有一種可能就是諺語及慣用法：

(2-12) She was born with a silver spoon in her mouth.

此句的含意非字面的「她出生時嘴裡有一支銀湯匙」，而是「她生而富貴」。這類句子，都需處理，才能得到較好的翻譯品質。這個層次並不在本文的討論範圍內。

(五)歸納及整理

根據前面幾節所述，我們得知，若依不同事態或語意關係來分類，可區分介詞的不同語意。但機器翻譯應考慮實際選字的問題。這使得我們必須做適當調整。而且知識來源應減少人工部份。若翻譯每個句子都需人參與決策過程，那只算是交談式(interactive)，或機器輔助翻譯。再者須考量參數的數量，以免訓練語料不足，即使是超大型語料。我們提出語意碼的模式來降低參數數量，同時解決一詞多義問題。

三、介詞的分析—以 with 為例

(一)with 的九大類翻譯

(A.)片語 (Phrasal)

出現於片語、慣用法、及祈使句或命令句中。由機讀詞典可取得相關資訊。

(3-1) I cannot put up with your behaviour any longer. (我再也不能容忍你的行為了)

(3-2) Nothing is the matter with me. (我沒什麼)

(B.)直接受詞 (Direct Object)

V-O-with-N 結構中的動詞 V 在朗文詞典中的語法次分類碼(syntactic subcategorization code)是以下其中之一時， with 即為此類：

[A] ：雙賓動詞(double-object verb)，需一直接受詞和一間接受詞。如 fill 、 emboss 。

[B] [Wv5]：常帶-ed 而作形容詞用之動詞。如 bedeck 、 ornament 、 pinch 。

[C] [T1 ，通常被動]： T1 指及物動詞，且緊接在動詞後的受詞或補語位置要置一名詞。此類動詞常以被動式出現。如 beset 、 encompass 。

格位語法分析時，這些動詞幾乎都在 Object 類。而介詞組中心語往往是直接受詞，故為「直接受詞」類。翻譯有兩種情形，一是譯成「被」，特別是用於被動句：

(3-3) The river was contaminated with waste from the factory. (河川被工廠排出的廢物污染了)

另一種因動詞性質，使 with 譯成一定位語(如「上」)或修飾語(例如「滿」)：

(3-4) A machine is clogged with dirt. (機器塞滿了髒物)

(3-5) to spray a wall with paint (將牆噴上油漆)

此種動詞，如 emboss 代碼 D1+*on/with* ，使 with 分在「直接受詞」類。其翻譯從 *on* 中可得到線索，表 emboss 動作作用在物體表面上；故 with 譯「上」。又 embed 代碼 D1+*in/with* 表動作作用到物體內部； with 譯為「入」。 emboss 的中文是「加浮雕花紋於...上」， fill 的解釋是「填滿」，也指出 with 的翻譯。

(C.)功能 (Telic)

第三類為「功能」類，表示「用...工具」「以...方法(手段)」等概念。

(3-6) Polish your shoes with a brush. (用刷子擦亮你的鞋)
Fa Bq Bp

上肢動作　　衣物　　用品

(3-7) He fobbed me off with a story. (他編了一個故事來騙我)
Hn　Aa　Dk
惡行　人的泛稱　文教

(3-8) He refreshed himeself with a glass of beer. (他喝了杯啤酒提提神)
Je　Aa　Br
影響　人的泛稱　食品

明顯看出英文介詞在中文會轉成動詞。句中 with 都譯成「用」或「以」，意思雖不變，卻未必合習慣。因此處理時將介詞正確歸類，再依中文辭彙語意結構中的動詞加以翻譯；「故事」的生成動詞是「編」，「啤酒」的功能動詞是「喝」。

過去的分析，都有「工具」(instrument)類，跟「功能」類意義相近。不同的是，「功能」類範圍大。(3-6)的「刷子」是工具，但(3-7)的「故事」視為工具就很勉強。但兩者均可以相同方式選取正確目標語，故應屬同類。

(D.)原因 (Cause)

第四類為「因」類，表示「原因」「條件」：

(3-9) The child's eyes rounded with excitement. (孩子因興奮眼睛睜得圓圓的)
Ih　Ga
變化　心理狀態

(3-10) His back was bent with age. (他因年老而背部彎曲)
Fd　Ca
全身動作　時間

(3-11) Your hands are　blue with cold. (你的雙手凍得發青了)
XX　Ec　Eb
be 動詞　顏色　表象

Durand(1993)的分析也有「原因」類。若不潤飾，幾乎都譯成「因」。然可為原因的東西很多；像是時間(Ca)、物體(Bo)、溫度(Eb)，或是情緒(Ga)。稍後實驗證明「原因」這種語意關係，不易從少量資料中擷取。

(E.)情狀 (Manner)

第五類為「情狀」類，當介詞片語用以修飾動詞，用法相當於副詞時即為此類：

(3-12) He assailed the difficulty with eagerness. (他興致勃勃地克服困難)
Jd 存在　Da 事情　Df 意識

(3-13) The boy's heart fluttered with excitement. (這男孩興奮得心砰砰地跳)
Ib 生理現象　Ga 心理狀態

不經潤飾，幾乎都譯成「地」。介詞賓語的中心語都是些與情緒或人的特質有關的辭彙，涵蓋範圍包括意識 (Df)、德才 (Ee)、心理狀態 (Ga)、心理活動 (Gb)等，分類時較明確簡單。Ravin 及 Durand 的分類，也有 Manner 類。

(F.)屬性 (Attribute)

第六類為「屬性」類。當 N 為具體的東西時我們常說「S(O)帶著 N」，而當 N 是抽象概念時，則常說是「S(O)有 N」。換言之，N 可視為 S 或 O 的一個屬性：

(3-14) some fresh blood with new ideas (有新觀念的新人)
Al 才識　Dk 文教

(3-15) people with certain diseases (患某種病的人)
Aa 人的泛稱　Dl 疾病

(3-16) The servant came into the bedroom with a cup of tea. (僕人端著一杯茶進入臥室)
Hj 生活　Bn 建築物　Br 食品

(3-17) He acted with great nobility of purpose. (他懷著崇高的目的行事)
Hi 社交　Db 事理

這類也有譯成中文動詞的現象。因此用中文辭彙語意結構中的動詞來選字。和「功能」類不同的是，「屬性」類用形態(formal)方面的資訊，而非功能方面。不同

型態和不同動作搭配。例如目的、理想，用「懷著」來描述。若無適當動詞，則用預設的「帶」和「有」。未必流暢，但能正確傳達介詞代表的語意關係。

此外，「屬性」類描述的是介詞組連繫到名詞的關係，Durand (1993) 的 Quality 類中有類似的概念，而在 Ravin(1990) 的研究及格位語法中則找不到對應的分析。

(G.)共事 (Co-agency)

第七類為「共事」類，通常出現於「S 跟(與/和/隨)N 一起 VO」情境。基本上這類就是出現在 S 與 N 一起做 V 這件事的情形下，相當於格位語法中的 Agent 類：

(3-18) I am quits with him. (我和他互不相欠)
XX Ed Aa
be 動詞 性質 人的泛稱

(3-19) Sweeden has frontiers with Norway. (瑞典與挪威接壤)
Jd Cb Di
存在 空間 社會政法

(3-20) She is always fun to be with. (和她在一起總是很有趣)
XX
be 動詞

大致是在「跟」「與」「和」三者中選擇，輔以「在一起」等字眼。可用許(1994)目標語語料為本的作法，選擇一個最常用，最流暢的辭彙。

(H.)對象 (Concern)

第八類為「對象」類，表「對於」「關於」等概念。Durand (1993)也有此類：

(3-21) I am losing my patience with him. (我對他失去耐心了)
Jd Ee Aa
存在 德才 人的泛稱

(3-22) I have some authority with the young boy. (我對那小男孩有些影響力)
Jd Je Ab
存在 影響 男女老幼

翻譯時，如不考慮潤飾，此類的直譯為「對」或是「對於」。

(I.)關係 (Relation)

第九類為「關係」類，表示O和N間的關係，例如所謂的Part-Whole Relationship：

(3-23) He has got a good position with an oil company. (他已在油公司謀得一份好工作)
Jd 存在 Di 社會政法 Dm 機構

(3-24) I am back to square one with the work. (這項工作得從頭再來)
XX be動詞 Da 事情 Di 社會政法

或是O帶有動作意味，而N為其賓語，例如：

(3-25) John is a great favourite with his grandmother. (約翰是他祖母的最愛)
XX be動詞 Ag 人的狀況 Ah 親人

(3-26) I need some guidance with my studies. (我的功課需人指導)
Jc 配合 Hg 教衛科研 Dk 文教

這類描述O與N的關係，僅Durand(1993)有類似分析。O與N的關係，用其他方法訓練，涵蓋範圍也許較廣。如語意網路。「關係」類的翻譯，似乎沒有規則。

	朗文	格位語法	Ravin	本實驗	備註
1	跟	Agent	Co-Agency	共事	
2	有、顯出			屬性/關係	Part-Whole
3	用	Instrument	Use	功能	
4	用、拿(材料)	Object	Alteration/Provision	功能/直接受詞	
5	支持	Object	Co-Agency	共事	
6	跟(對抗)	Agent	Co-Agency	共事	
7	順著	Agent	Co-Agency	共事	
8	隨著		Co-Agency	共事	While
9	跟(比較)	Object	Co-Agency	共事	
10	跟(分離)			共事	Phrasal
11	雖然			原因	Constraint
12	因為、得	Object	Manner	原因/情狀	Cause
13	交給...看	Object	Co-Agency	共事	
14	有關、對於	Benefactive		對象	
15	跟(連接)		Phrasal	共事	Phrasal
16	(命令)		Phrasal	片語	Phrasal
17	由...選出		Phrasal	共事	Phrasal
18	in with		Phrasal	片語	Phrasal
19	with it		Phrasal	片語	Phrasal

20	with me/you		Phrasal	共事	Phrasal

表 3-1　Longman 、Case Grammar 、Ravin 、及本實驗四種分析的對照表

跟過去比較起來，我們的分析有以下的優點及特性：

1. 最終目標是正確翻譯，所以不需分析細微的語意差異，因為即使分析正確，對實際翻譯可能沒有幫助。所以分析時應以目標語的翻譯作為依據。
2. 在「功能」跟「屬性」兩類中，提出先分類後再選字的觀念。若配合如辭彙語意結構的處理，只要分類正確，即可得到正確的介詞翻譯。
3. 引入「直接受詞」類。句法次分類可決定介詞的分析與翻譯。
4. 引入「屬性」及「關係」類，表 O 與 N 的關係。格位語法無類似分類。

(二)其他介詞的分類

本實驗的介詞分類似乎免不了人力介入。觀察那些已有對應到中文翻譯的 with ，其實分類相當明顯。再者，介詞數目不多且不會增加，人力介入也不困難。

譯文經常是經過潤飾的。本實驗訓練語料， 48.8%的 with 無對應中文。除去「片語」類，仍有 34.1%因目標語潤飾而無對應。目前採人工輔助的方式解決這個問題。

四、介詞分析的計算模式

(一)雙語資料的處理

我們以朗文當代英漢雙解字典含 with 的中英對照例句為訓練語料。英文句先做詞性標註(Church,1988)及原形化處理，然後由朗文英漢雙解多功能分類詞典，得到英文辭彙的英文語意碼集合。而對應的中文句斷詞後查閱同義詞詞林，得到中文辭彙及其中文語意碼集合和詞性。再做辭彙語意分析，給予每個英文字正確語意碼(Chen and Chang,1994; Chang and Chen,1996a; Chang and Chen,1996b; Chang et al.,

1996)，並完成中英文辭彙對應(柯,1993; Ker and Chang,1995; Ker and Chang, 1996)。

有關介詞的問題，多以 V-O-P-N 單元結構[1]表示句子 (Hindle and Rooth, 1993) (許, 1994) (Chen and Chang, 1995)。取得所需的 V-O-with-N 結構(許,1994)方法如下：

1. with 之前第一個出現的動詞為 V-O-with-N 結構中的 V 。

2. 動詞與 with 間的名詞組為 O ，其最末字為中心語。若無名詞，取形容詞。

3. with 之後為介詞賓語，取名詞組的中心語為 V-O-with-N 結構中的 N 。

取出 1,424 個 V-O-with-N 結構後，將句中 with 中文翻譯歸類至九大類中。

(4-1) I open the door with a key. (我用鎖匙 開 門)
Mb065 Db024 Hd131 Bo03 Fa31 Bn04
打開與關鎖 門及其組成 鎖與鑰匙 機件 開關 門窗

假設語意碼取第二級，動詞(open)屬手的動作，受詞(door)是建築物的一部份，介詞賓語(key)是工具。因介詞表達「關係」，手的動作跟工具間的關係，就是「手用工具作用於建築物上」。所以從句子組成的語意，可推出組成成份間可能的關係。

(4-2) I went out with a doctor. (我跟一個醫生一起出去)
Bj166 Ae15

(4-3) I went out with a lawyer. (我跟一個律師一起出去)
Ck201 Ae12

本實驗以同義詞詞林 94 個中類語意碼[2]，來研究 V-O-with-N 結構和 with 的翻譯。

(二)法則的產生

採用決策串列(decision list)(Rivest,1987; Yarowsky,1994a; Yarowsky,1994b)來產

---

1. [1] V 表動詞組的主要動詞，O 表受詞，P 表介詞，N 表介詞賓語的中心語。若進一步推廣至 S-V-O-P-N 結構的話，其中的 S 表主詞
2. [2]若 V 、O 、N 其中一欄沒有字的話，則以—表之。例如：I go with you.句中的 O 的中文語意碼即標為—。

生法則。基本上法則就是說明搭配辭和 with 翻譯的關係，以及其關係的可靠性分數。

但產生法則前，要先加以處理兩種特殊情形：

1. 特殊用法：當 with 以固定的搭配形式和某些特定字共現時，可直譯成常用中文，而不需看其語意碼或是其他。例如出現 mix with ， with 直接歸到「跟」類(...跟...混合)。 au fait with 常直譯成「熟悉」， with 就歸到「片語」類。
2. 「直接受詞」類：當句子中的動詞在朗文詞典中的文法分析代號為「 D 」、「 Wv5 」、「 T1 通常被動」這三類時，我們將 with 分到「直接受詞」類。

實驗中的 V 、 O 、 N ，所可能出現的搭配形式有以下幾種：

1. VON ：是要三個都符合才算。亦即應用了所有包含在 V 、 O 、 N 這三個變數中的資訊。因為要符合所有的條件，所以接近於例句式的作法。
2. VO 、 VN 、 ON ：只要二個符合。其中 VN 類代表動詞及其論元的關係， ON 類代表 Part-Whole 等字與字間的語意關係。三類間，不預設優先權，完全由分數值來決定。
3. 只要一個符合。 V 類由動詞來看介詞翻譯。 N 類則由修飾語(adjunct)為出發點。 O 類雖無理論基礎，但 O 空白或為代名詞時，介詞組此時無法連繫到 O ，將間接解決結構歧異問題，所以也有助於介詞的翻譯。而這三類間也不預設優先權，完全由分數值來決定。

表 4-1 根據可能出現的搭配形式，分別舉了幾個例子：

搭配形式	with 分類	V	O	N
VON	功能	Fa(上肢動作)	Bp(用品)	Bk(全身)
VO	對象	Hi(社交)	Aa(人的泛稱)	
VN	共事	Hj(生活)		Aa(人的泛稱)
ON	對象		Ee(德才)	Aa(人的泛稱)
V	功能	Ig(始末)		
O	功能		Bm(材料)	
N	共事			Aj(人的關係)

表 4-1　不同形式搭配的法則

以表 4-1 第一列 Fa-Bp-Bk 為例，在訓練語料中共有下表 4-2 中的四個例句：

with 分類	原句
功能	She untied the knots with dexterous fingers.

功能	He drummed on the table with his fingers.
功能	Hit the ball with a long free swing of the arm.
功能	If you pick up the ball with your hand in golf, you suffer a penalty.

表 4-2　訓練語料中 Fa-Bp-Bk 的例句

我們假設 VON 都符合者所含資訊較精確，優先權最高；只符合一個的優先權最低。而對每個可能的搭配 C，對應的 with 分類 P，其分數值為

$$Score(C,P)=\frac{Count(C,P)}{Count(C,\overline{P})}$$

$Count(C,P)$ 表示搭配 C 且 with 的分類為 P 的次數，而 $\overline{P}$ 表示其他可能的 with 分類。分數高則正確率高，會優先使用。分數值大於一時，表示對的次數比較多；因此可為門檻值。若不設門檻值(或為 0 時)，表示只要是訓練語料中取出的訊息，就儘量使用。再以表 4-2 為例，此時搭配 C 為 Fa-Bk-Bp，P 為「功能」，規則的分數值為

$$Score(Fa\text{-}Bk\text{-}Bp, 功能)=\frac{4}{0}$$

即 Fa-Bk-Bp 出現四次，with 皆屬「功能」類，且找不到非功能類的例句。這表示此法則可靠度極高。

但為解決資料稀疏(sparse)的問題，還要做修勻(smoothing)。假設所有的零值都是由於抽樣不足，用簡單方式來修勻：在 I × J 表格中，每個方格加上 1/J (Agresti,1990)。如表 4-2，VON 類法則可視為在一 J = 7 的表格中[3]，因此搭配為 Fa-Bp-Bk 且介詞 with 屬「功能」類的這條法則，分數值為

$$Score(Fa\text{-}Bk\text{-}Bp, 功能)=\frac{4+\frac{1}{7}}{0+\frac{7-1}{7}}=4.8333$$

---

3. [3] 全部可能的介詞分類共有九類，可是由於「片語」類跟「直接受詞」類的資訊明確，在之前就先處理掉了，所以在產生法則時只考慮剩下的七類。

五、實驗結果與討論

(一)實驗結果

表 5-1 到 5-3 是產生出來的法則。在此我們只列舉各種形式分數最高的前十名。

分類	V	O	N	分數	計數值	例句
共事	Hj	--	Aa	28.714	25	You must come with us, I insist.
共事	XX	--	Aa	11.571	10	The poor are always with us.
情狀	Fc	--	Ga	7.0000	6	The children squealed with delight.
功能	Fb	--	Bk	7.0000	6	One old lady walked with heavy foot.
功能	Ig	--	Dk	5.8571	5	The party finished with a song.
對象	XX	Ee	Dj	4.7143	4	He is carefree with his money.
對象	XX	Ee	Ab	4.7143	4	They are very strict with their children.
對象	XX	Ee	Aa	4.7143	4	Are you being straight with me?
共事	Hj	--	Aj	4.7143	4	She is shacking up with her boyfriend.
共事	Hi	--	Aa	4.7143	4	You must not joke with him.

表 5-1　VON 類法則的前十名

1,424 個 V-O-with-N 結構中，「共事」類最多，佔 361 句(25.3%)，其次為「功能」類，324 句(22.8%)；可見 with 的分析相當困難。只取最高頻對應，準確率極低。

我們以應用率(applicability)及準確率(precision)作為評估實驗結果的依據：

$$\text{應用率} = \frac{\text{完成猜測的VOPN數}}{\text{所有測試的VOPN數}} \qquad \text{準確率} = \frac{\text{正確猜測的VOPN數}}{\text{完成測試的VOPN數}}$$

分類	V	O	N	分數	計數值	例句
共事	Hj		Aa	44.573	39	It is a pleasure to do business with you.
功能	Fa		Bo	21.716	19	Cut it with the scissors.
共事	XX	--		16.002	14	She was here with her betrothed.
功能	Ig	--		13.716	12	The story opens with a snowstorm.
功能	Fa		Bp	13.716	12	Bind the prisoner with rope.
功能	Fa		Bk	13.716	12	Hit him with your right.
屬性	--	Aa		10.287	9	Someone with creativity is needed.
屬性	--	Ab		9.1444	8	The man with the big dog came in.
功能	Fa	Aa		9.1444	8	I cut myself free with an axe.
功能	Ig		Dk	9.1444	8	The party finished with a song.

表 5-2　VO/VN/ON 類法則的前十名

表 5-4 是 1,424 句內部測試。外部測試用牛津高級英英英漢雙解辭典以及 Collins Cobuild English Langauge Dictionary 的 109 個 with 的解釋例句。實驗結果如表 5-5 。

分類	V	O	N	分數	計數值	例句
功能	Fa			10.043	59	He splashed his face with cold water.
功能			Fa	5.7143	5	He set the machine going with a push.
功能	Ig			4.9032	19	Let's top off the evening with a drink.
功能	Hd			4.8696	14	They feed the pig with the parings.
功能		Fa		4.5714	4	He gave me a rap with her pencil.
情狀			Gb	4.5217	13	Cross the road with care.
共事			Af	4.2667	8	He arrived with several attendant helper.
共事	Ja			3.7333	7	The decision lies with you.
共事			Aj	3.4872	17	I'm storing my television with a friend.
功能		Bm		3.4286	3	He probed the mud with a stick.

表 5-3　V/O/N 類法則的前十名

門檻值	0	1	2	3
總數	1424	1424	1424	1424
正確	1397	1396	1174	1005
錯誤	27	26	106	48
未知	0	2	144	371
應用率	100%	99.9%	89.9%	74.0%
準確率	98.1%	98.2%	91.7%	95.4%

表 5-4 內部測試的結果

門檻值	1	2	3
總數	109	109	109
正確	82	72	54
錯誤	27	21	9
未知	0	16	46
應用率	100%	85%	58%
準確率	75%	77%	86%

表 5-5 外部測試的結果

外部測試可看出，門檻值越高，應用率越低，準確率越高。若以應用的法則種類來觀察(門檻值為 1)，可獲表 5-6 的結果：

種類	總數	正確	錯誤
直接受詞	7	7(100%)	0(0%)
片語	21	19(90%)	2(10%)
VON	27	23(85%)	4(15%)
VO/VN/ON	49	33(67%)	16(33%)
V/O/N	5	0(0%)	5(100%)

表 5-6　外部測試結果依法則種類加以分類

這和我們之前「符合的欄位數越多，可靠度越高」的假設完全符合。只符合一個欄位的五句全部都錯，雖說是資訊太少，所以可靠度較低，但不代表這些法則沒有利用價值。以只看 N 的欄位為例，部份的修飾語其實已可左右介詞 with 的翻譯：

(5-1) Do your work with care. (小心做你的工作)
Hj　Dk　Gb
生活　文教　心理活動

(5-2) Cross the road with care. (小心過馬路)
Hf　Bn　Gb

交通運輸 建築物 心理活動

(5-3) Wash the cups with care. (小心洗杯子)
Fa Bp Gb
上肢動作 用具 心理活動

由這三例可發現，雖然動詞受詞不同，但大部份"with care"都屬「情狀」類 (訓練語料中 7 句皆屬「情狀」類)。只是由於字典偏差，導致法則有過份特殊(over-specialization)的現象，使得在應用時，較傾向使用符合兩個以上欄位的法則。

(二)錯誤結果分析

觀察外部測試(門檻值 1)，若以 with 分類來看，可得到如下表 5-7 的數據：

分類	總數	正確	錯誤
片語	14	14(100%)	0(0%)
直接受詞	7	7(100%)	0(0%)
情狀	4	4(100%)	0(0%)
功能	16	15(94%)	1(6%)
共事	36	27(75%)	9(25%)
對象	8	6(75%)	2(25%)
原因	7	3(43%)	4(57%)
關係	5	2(40%)	3(60%)
屬性	12	4(33%)	8(66%)

表 5-7 依 with 翻譯分類結果分類

前六類正確率在 75%以上，後三類卻不到 50%。我們的解釋是：

1. 「原因」： Yarowsky(1992)在解決字義的歧異時提到，若一個字可出現在多種上下文中(如 interest)，就難由鄰近字的標題(topic)推測其含意。本實驗中也有類似現象。一件事的原因可能是一個人，一個東西，甚至只是一個人的心理狀態。光從 V 、 O 、 N 似乎無法看出介詞賓語是用來表示原因的。也許要設法從句子的語意中觀察出介詞賓語是用來說明前面的動詞或全句，才可能將其正確譯成「因為」。例如"Metal expands with heat."，需

要現實世界的知識—東西因熱會膨脹；否則「原因」和其他類並不易區別。

2. 「關係」：「關係」這類出現於描述 O 與 N 的關係。然而 with 連繫到動詞的機率原本就較高(Sumita and Iida,1992)，使得從訓練語料中得到的資訊有限。Iris(1988)曾將 Part-Whole 關係分四個模型來解決，其中，有些要有現實世界知識(如 Valentine's Day-February)，有些在用字上有特殊的搭配(如 sheep-flock)。這些資訊龐大，從實驗中的 1,400 句中，是取不到的。

3. 「屬性」：在錯誤的八句中，有六句的介詞片語是用來主詞的述語，而主詞在此並未加以考慮。將實驗擴大至 S-V-O-with-N ，也許可以解決。

而其他的錯誤原因包括：

1. 語料不足：只有一千四百多個訓練句，部份的現象無法涵蓋。

2. 同義詞詞林的分類：在同一個同義詞詞林的中類裏，仍有很多語意分歧。這也就是 Yarowsky(1992)的「同一分類中的些微差異」。以"Cut the cloth with the knife." 以及 "Put the cloth with the knife."為例，我們用刀子來剪(Fa28)，但也會放(Fa13)一樣東西跟刀子在一起。可以改用小類來做，或搭配其他相關的資訊(如英文的語意碼)，但先決條件是要有更大量的語料。

3. 前處理的錯誤：前處理的錯誤會間接影響結果。詞性標注，以布朗語料庫測試，有 97.495%的準確率。解決字的歧義，有 94%的準確率(Chang and Chen, 1996)。辭彙對應，是 93.3%(Ker and Chang,1995)。中英文語意碼對應，則為 93.4%(Ker and Chang, 1996)。改進前處理，也是未來研究的重點。

4. 機讀字典的缺點：字典列舉所有可能，所以用於處理片語等特殊用法時，

不易因語料不足而取樣不到。但字典通常偏好特殊現象，加上編纂者的個人觀點(Zernik, 1991)，因此也不易反應正常的語言現象。 Hearst (1991)提到，機讀字典對某些語意結構的偏好，會導致不平衡的共現資訊。

六、結論與未來展望

根據實驗的結果與分析，我們提出一些有待加強之處及可能的發展方向：

1. 推廣至其他的介詞，甚至是其他的虛詞，以期得到更好的機器翻譯品質。
2. 解決需人工分類訓練語料的問題，使整個流程能做到完全自動。
3. 收集目標語中的辭彙相關資訊，用以輔助辭彙選擇的問題。對中文做較深入的語法分析，應該可能對介詞在中文中扮演的角色有更進一步的認識。
4. 利用其他機讀字典中可以擷取的資訊，例如領域碼 (domain code)，定義，語意碼間的階層關係，以得到更多有用的訊息。
5. 利用其他雙語語料，配合現有機讀字典，以得到較一般性、平衡性的法則。

致謝

本實驗獲得行政院國科會計畫編號 NSC85-2213-E-007-042 贊助，特此致謝。

參考文獻

1. Agresti, Alen, *Categorical Data Analysis*, New York: John Wiley & Sons, 1990.
2. Chang, Jason J.S., and J.N. Chen, "Acquisition of Computaional-Semantic Lexicons from Machine Readable Lexical Resources", in *Proceeding of ACL SIGLEX Workshop*, 1996.
3. Chang, Jason J.S., and J.N. Chen, "Word Sense Division Based on Dictionary and Thesauras", unpublished manuscript, 1996.
4. Chang, Jason J.S., H.H. Sheng, J.N. Chen, S.J. Ker, "Combining Machine Readable Lexical Resources and Bilingual Copora for Board Word Sense Disambiguaion", unpublished manuscript, 1996.
5. Chen, Hsin-Hsi, "The Transfer of Prepositional Pharse in English-Chinese Machine Translation System", *Conference Book, ACH/ALLC*, 1991.
6. Chen, Jen-Nan, and Jyun-Sheng Chang, "Towards Generality and Modularity in

Statistical Word Sense Disambiguation", in *Proceedings of the Asian Conference on Language, Information and Computation*, 1994.
7. Chen, Mathis H.C., and Jason J.S. Chang, "Structural Ambiguity and Conceptural Information Retrieval", in *Proceedings of the 10th Pacific Asia Conference*, HongKong, 1995.
8. Church, Kenneth Ward, "A Stocastic Parts Program and Noun Phrase Parser for Unrestricted Text", in *Proceeding of 2nd Conference on Applied Natural Language Proceesing*, Austin, Texas, 1988.
9. Cook, Walter A., *Case Grammar Theory*, Washington, D.C.: Georgetown University Press, 1989.
10. Dagan, Ido, and Alon Itai, "Word Sense Disampbiguation Using a Second Language Monolingual Corpus", *Computational Linguisics 20(4)*, 1994.
11. Durand, Jacques, "On the Translation of Prepositions in Multilingual MT", In Frank Van Eynde, editor, *Linguistic Issues in Machine Translation*, London: Pinter Publishers, 1993.
12. Gale, William, and Kenneth Church, "What is wrong with adding one?", *AT&T Bell Laboratories Statistical Research Report No.90*, 1989.
13. Hearst, Marti A., "Noun Homograph Disambiguation Using Local Context in Large Text Corpora", in *Proceedings of the 7th Annual Conference of the UW Centre for the New OED and the Text Research: Using Corpora*, 1991.
14. Hindle, Donald, and Mats Rooth, "Structural Ambiguity and Lexical Relation", *Computational Linguistic 19(1)*, 1993.
15. Hutchins, W. John, and Harold L. Somers, *An Introduction to Machine Translation*, London: Academic Press, 1992.
16. Iris, Madelyn Anne, Bonnie E. Litowitz, and Martha Evens, "Problems of the Part-Whole Relation", in Martha Walton Evens, editor, *Relational Models of the Lexicon: Representing Knowledge in Semantic Networks*, Cambridge: Cambridge University Press, 1988.
17. Kennedy, Graeme, "Between and Through: The Company They Keep and Fuctions They Serve", in Karin Aijmer and Bengt Altenberg, editors, *English Corpus Lingusitics*, New York: Longman Inc., 1991.
18. Ker, Sur-Jin, and Jason J.S. Chang, "Automatic Acquisition of Class-based Rules for Word Alighment", in *Proceedings of the 10th Pacific Asia Conference*, HongKong, 1995.
19. Ker, Sur-Jin, and Jason J.S. Chang, "Aligning More Words with High Precision for Small Bilingual Corpora", in *Proceedings of 31th Annual Meeting of the Association for Computer Linguistics*, 1996.
20. Longman, *Longman English-Chinese Dictionary of Contemporary English*, Hong Kong: Longman Group(Far East) Ltd., 1992.
21. McArthur, Tom, *Longman Lexicon of Contemporary English (English-Chinese Edition)*, Hong Kong: Longman Group(Far East) Ltd., 1993.
22. McRoy, Susan W., "Using Multiple Knowledge Source for Word Sense Disambiguation", *Computational Linguistics 18(1)*, 1992.
23. Pustejovsky, James, "The Generative Lexicon", *Computational Linguistics 17(4)*, 1991.

24. Pustejovsky, James, "Lexical Semantic Techniques for Corpus Analysis", *Computational Linguistics 19(2)*, 1993.
25. Ravin, Uael, "Disambiguating and Interpreting Verb Definitions", in *Proceedings of 28th Annual Meeting of the Association for Computer Linguistics*, Pittsburgh, PA, 1990.
26. Rich, Elaine, and Kevin Knight, *Artificial Intelligence*, Singarpore: McGraw-Hill, Inc., 2nd ed., 1991.
27. Rivest, Ronald L., "Learning Decision List", in *Machine Translation*, 1987.
28. Sekine, Satoshi, and Jun-Ichi Tsujii, "Automatic Acquisition of Semantic Collocation from Corpora", in *Machine Translation*, Netherland: Kluwer Academic Publishers, 1995.
29. Sells, Peter, *Lectures on Contemporary Syntactic Theories*, United States: Center for the Study of Language and Information, 2nd Ed., 1985.
30. Sinclair, John, *Collins Cobuild English Language Dictionary*, Great British: William Collins Sons & Co Ltd., 1988.
31. Sumita, Eiichiro, and Hitoshi Iida, "Example-Based NLP Techniques - A Case Study of Machine Translation", in *Statistically-Based Natural Language Programming Techniques -- Papers from the 1992 AAAI Workshop Technical Report W-92-01*, Menlo Park, California: AAAI Press, 1992.
32. Tang, Ting-Chi, "Contrastive Approach in Biligual Education", *Studies in Chinese Syntax: Monograms on Modern Linguistics*, Taipei: Students Book Co., 1979.
33. Yarowsky, David, "Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora", in *Proceedings of 30th Annual Meeting of of the Association for Computational Linguistics*, Nantes, 1992.
34. Yarowsky, David, "Decision List for Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French", in *Proceeding of 32nd Annual Meeting of the Association for Computational Linguistics*, Las Cruces, NM, 1994.
35. Yarowsky, David, "A Comparison od Corpus-based Techniques for Restoring Accents in Spanish and French Text", in *Proceeding of 2nd Annual Workshop on Very Large Corpora*, Kyoto, Japan, 1994.
36. Zernik, Uri, "Introduction", in Uri Zernik, editor, *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, Hillsdale, NJ: Lawrence Erlbaum Associates, 1991.
37. 柯淑津,"英中辭彙對應與語意分析之研究", 行政院國家科學委員會專題研究計劃成果報告 NSC82-0408-E-007-195, 1993.
38. 張芳杰, 牛津高級英英英漢雙解辭典, 台北: 東華書局, 6 版, 1989.
39. 梅家駒, 竺一鳴, 高蘊琦, 殷鴻翔, 同義詞詞林, 台北: 東華書局, 1993.
40. 許曙峰,"英中機器翻譯－介詞組分析、轉換、生成", 清華大學資訊科學研究所碩士論文, 1994.
41. 陳正誼,"英中機器翻譯－中文補詞增刪之研究", 清華大學資訊科學研究所碩士論文, 1994.

Context-Centered Template Matching for Chinese Lexicon Construction

GUO JIN
Institute of Systems Science
National University of Singapore
Kent Ridge, Singapore 119597
email: guojin@iss.nus.sg

ABSTRACT

Our objective is to develop productive approaches for constructing million-entry Chinese lexicon. The emphasis in this paper is on the development of the fresh context-centered lexicon construction methodology. Beginning with an intuitive explanation, we gradually worked out the formal mathematical model and the productive lexicon construction algorithm. The effectiveness of the proposed scheme is demonstrated in a detailed case study.

KEYWORDS: Lexicon Construction, Term Identification, Chinese Computing, Language Modeling, Statistical and Corpus-based NLP.

1 INTRODUCTION

The objective in this paper is to build a lexicon with huge number of entries. The task is strongly application motivated. For example, the Chinese Dictation Kit (CDK) developed at Apple-ISS (Yuan, etc., 1996) is based on a lexicon of over 350,000 entries. The Chinese-English Machine Translation System from SYSTRAN (Yang and Gerber, 1996) employs a lexicon of over 600,000 words. From my personal experiences in Chinese text spell checking and proofreading (Guo, 1994) and Pinyin-to-Hanzi transcription (Guo, 1993), huge lexicon on the size of million entries would be of great help, especially for alternative suggestion and error correction. Large lexicon can also be found in many other applications such as text-to-speech (Sproat, 1994) and information retrieval (Harman, 1995, Harman, 1996).

Moreover, such lexicon is largely for general purpose rather than domain specific. For example, I was told that most obvious domain specific and/or technical terms are purposely excluded from the Apple's CDK lexicon. This is to ensure the general suitability of the system. Domain specific entries are supported with a function limited supplementary user dictionary management mechanism.

Inevitably, many entries in such huge lexicon are not simple words but compounds, idioms, rigid phrases, or even complete (short) sentences. The question here is not the justification of whether or not entries in such lexicon are words, or such huge collection is still a lexicon proper. Rather, what is challenging us is: how do we build up such a huge lexicon and make it useful?

Past efforts in unknown word identification and dictionary construction can be classified into the following categories. Manual construction is of course the first way. The 600,000 words SYSTRAN lexicon (Yang and Gerber, 1996) is from a government department. Automatic extraction from large (partially) parsed corpora such as Brown corpus (Kucera and Francis, 1967) and the Penn TreeBank (Markus, et al., 1993) also falls into this category. This is definitely a quality approach, provided that enough resources are available.

Grammar-based automatic generation is the second way. Basically, the 350,000 words Apple CDK lexicon (Yuan, etc., 1996) is from the enumeration of some well-designed ***Prefix-Body-Suffix*** patterns on a core dictionary of about 10,000 multi-character words. For example, as an entry, the Chinese phrase ***"很美丽的 (very beautiful)"*** follows the pattern ***"adv + adj + de"*** where "很 (very)" is the adverbial prefix, "美丽 (beauty)" the body and "的 (de mark)" the suffix sometimes functioned

similar to the English adjective suffix "-ful". One of the obvious problems is the lexicon's (lack of) coverage. Too many words and terms are not compositional, at least not composed in such a simple way.

Sub-language modeling is the third way. For example, Sun and Huang (1996) systematically presented several intelligent ***agents*** each for identifying a type of constructs such as Chinese names (CName Agent), Transcribed Foreign Names (TFName Agent) and Chinese Place Names (CPName Agent). Song, etc., (1996) presented sets of rules for company name and person's name identification. Mo, etc., (1991) derived the grammar for Chinese-specific determination-measure compounds and implemented an identification parser. Chen and Liu (1992) implemented rules for reduplication and A-not-A constructions and some other kinds of derivable constructions. Sub-language approach is the mainstream in unknown word processing in Chinese text processing (Chang, 1994, Nie, Jin and Hannan, 1994, Zheng and Liu, 1993, Lee, Lee and Chen, 1994, among many others). All these works are limited to their predefined type of constructs.

Entity-centered statistical modeling is yet another way. The assumption behind this school is that "words are tightly-bounding and frequently-using entities". Fung and Wu (1994) applied ***CXtract***, a localized version of ***Xtract*** by Smadja (1993), to extract statistically significant Chinese character ngrams as possible entries for dictionary augmentation. This approach is not applicable for the identification of the majority of low frequency words, a drawback deeply rooted in its underlying assumption (Smadja, 1993)*. From a corpus of about 2 million Chinese characters, Fung and Wu (1994) only extracted about 5,000 new entities. Other examples of works in this category are Chiang, etc., (1992), Lin, etc., (1993) and Sproat and Shih (1990). ***Mutual information*** and ***t-test*** are the two representative means in this school (Church and Hanks, 1990).

The only work I am aware of which is more or less away from the above mentioned categories is by Luk (1995). He proposed to use his heuristically defined "lexicographic index" as an indicator to highlight potential "words". The unique characteristics of his approach is its capability of extracting low frequency words: a character string is highlighted as long as it occurred in at least two different contexts. With this approach, Luk (1996) extracted 75,535 new entities from the 4-million-character PH corpus (Guo and Lui, 1992), clearly higher recall than what Fung and Wu (1994) achieved with their entity-centered ngram statistical approach, if the two are comparable.

To reach the million-entry target, however, the recall rate must be significantly higher. As Luk (1995) has already loosed the requirement to two different occurs, the question here basically becomes: ***how do we extract words that occur only ONCE in corpus***?

Note, even for huge corpus, the majority are still those occurring few times. For instance, in a collection of about 60 million characters news articles from China's Xinhua News Agency, there are 1,129,313 unique word bigram types, of which 339,839 bigram types occur once, and 172,467 twice. That is, even for a corpus of that huge size, there are still about half bigram types occurring merely once or twice. That is yet a result based on close-dictionary tokenization (that is, there is no effort taken for unknown word identification in the process of text tokenization and all tokenized words are in the given tokenization dictionary). Situations for trigrams and general ngrams under open-dictionary tokenization will be even more apparent.

Some readers might argue that those low frequency entities are not important, simply because each of them, as an individual entity, is next to ***never been used***. That is true in one sense. Collectively, however, the "silent majority" forms a large mass. If all bigram types occurring less than 11 times are filtered out as Fung and Wu (1994) did, on average, there will be at least one unknown bigram type in each sentence†. As bigrams are the core of unknown words or compounds, discarding low frequency

---

* Smadja (1993, page 165) himself pointed out that: "***Xtract*** has only been effective at retrieving collocations for words appearing at least several dozen times in the corpus". However, "For the 10 million-word stock market corpus, there are some 60,000 different word forms" and "Out of the 60,000 words in the corpus, only 8,000 were repeated more than 50 times."

† This could also be validated on data from many other publications. For example, Church and Gale (1991) listed in their Table 1 that, in a corpus of 22 million words, bigram types with frequency up to 9 totally count for 5,210,157 occurrences. That is, on average, for every four bigrams encountered, there is a low-frequency bigram type. And normally sentence length is far more than four words.

bigram types effectively ruled out the possibility of collecting and utilizing that significant portion of knowledge.

Entity-centered statistical approaches are good for highlighting those few "top stars". In contrast, our attention is on the largely ignored "silent majority". The key idea here is to let context play the center role.

We will first in Section 2 give an intuitive explanation of our basic idea, and then, in Section 3, present a detailed case study. Section 4 is reserved for formal presentation of our mathematical modeling. Then, the baseline lexicon construction algorithm is given in Section 5. Then a short conclusion is in Section 6. Complete data list for the case study is in Appendix.

2 INTUITION

In this section, we will first give an intuitive start of our context-centered thinking, then contrast it with traditional entity-centered thinking, and sketch our context-centered lexicon augmentation algorithm.

2.1 "This is a word."

Suppose I have no idea of English but was told that the sentences below are all legitimate in English:

(1.a) This is a ***pear***.
(1.b) This is a ***table***.
(1.c) This is a ***man***.
(1.d) This is a ***dog***.

Moreover, I was told that ***pear***, ***table***, ***man*** and ***dog*** are all legitimate English ***words***. Then, if I know the sentence

(1.e) This is a ***xxx***.

is also correct, by analogy, I will feel comfortable to agree that the entity ***xxx*** here ought to be an English ***word*** also. Moreover, I am happy to accept the ***assertion*** that: ***any single entity taking the position of "xxx" in the sentence above is a*** **word**. In other words, any entity which could be filled up into the empty slot in

(2) This is a _____.

is a word. In this paper, the empty slot is named a ***word holding slot*** or simply ***(word) slot*** and (part of) the sentence containing such word holding slot a ***word holding template*** or simply ***(word) template***.

To go a step further, let us agree that "***a pear***", "***a table***", "***a man***" and "***a dog***" are all a type of entity called ***chunk*** (Abney, 1991, Abney, 1996). Notice that all these chunks take the empty slot in sentence:

(3) This is _____.

Then, it should sound reasonable to accept that: ***anything taking the empty slot above is a chunk***. Or, the empty slot is a ***chunk holding slot*** and (part of) the sentence a ***chunk holding template***.

2.2 Context-centered Definition

Note, we do define what are words or chunks above. Explicitly, ***a word is an entity that could take up a word holding slot in a word holding template.***

This definition makes itself apart from the tradition of entity-centered thinking. Linguistically, as evidenced in grammar books and dictionaries, "word" is normally defined as an entity with "phonological, lexicographic, syntactic and/or semantic independence (meaning)", and "free usage in text" (e.g., Hu, 1987). In NLP community, the widely followed criterion for justifying word is the notorious eight Chinese characters: ***"结合紧密, 使用稳定 (tight-in-combination and stable-in-use)"*** (GB-13715, 1993). We call them ***entity-centered definitions*** since the judgment is mostly based on the characteristics of the entity to be judged, but with no explicit reference to context.

Such entity-centered definitions worked fine for those high frequency "top stars". If an entity occurred thousand or even million times in text, it is easy to judge whether or not it is "tight-in-combination

and stable-in-use" or "with phonological, lexicographic, syntactic and/or semantic independence (meaning)". However, such high frequency entities, if they *are* words, are by and large already recorded in out-of-the-shelf dictionaries, a resource already in use. What becomes problematic are exactly those with low usage frequencies. Yet, without high profile in text, we are not able to reliably determine their (statistical) characteristics.

To the "silent majority", context becomes the primary source. Yes, as long as you believe in the correctness of the sentence "This is a xxx", you will not hesitate to agree with the acceptability of "xxx" as a word, no matter how rare, odd, exotic, strange or mysterious it is. We only need to see *one* occurrence of the word and confidently make our judgment. That is the power of context, an information source largely neglected in literature.

Note, we are not rejecting the entity-centered thinking. What we are emphasizing here is that, with the compensation of the context-centered thinking, we may even have a more complete understanding. The entity-centered model and the context-centered model are somehow in opposition. Both have their pros and cons. Yet they could co-exist in one system, as they are the complement of each other in nature. In this paper I will purposely ignore the good part of the entity-centered model.

2.3 Algorithm Sketch

The intuitive illustration above suggests itself the following two-phase dictionary augmentation *algorithm*:

Phase 1: Template Preparation

This is to collect a large set of high quality word holding templates such as "This is a ___.". Later, we will discuss how this could be done in an automatic manner on corpus and dictionary.

Phase 2: Template Matching

This is to identify words from text by matching sentences against above prepared templates. Those entities in text with surrounding context matching some templates and themselves taking word holding slots will be collected as candidates for dictionary augmentation.

The core of our dictionary augmentation algorithm is that simple. To actually put the algorithm into practical use, of course, some preprocessing such as dictionary-based tokenization and part-of-speech tagging, and some postprocessing such as linguistic filtering and statistical selection, are also required. Our focus, however, is only on the above-mentioned two phases. Before going into details of the two phases, let us have a real world case study first.

3 CASE STUDY

The task in this section is to have a detailed case study on context-centered template matching word identification. We will first introduce our template pattern, and discuss various properties of a particular template used in this case study. Then, word candidates extracted from corpus with that particular template are presented, categorized and analyzed. We will also make a short conclusion in the last subsection.

3.1 Template Pattern

Practically, to make the collection of word holding templates manageable, most word templates could not be full sentences but small sentence fragments. Although various template types are possible, we only tried the simplest one as given below:

```
<LeftContext> <WordHoldingSlot> <RightContext>
```

Depending on the level of preprocessing, the `<LeftContext>` and `<RightContext>` could be strings of characters (for un-tokenized text), words (for tokenized text), part-of-speeches (for part-of-speech tagged text), or their combinations. Similarly, the `<WordHoldingSlot>` could take up a fixed or variable number of characters, words and/or part-of-speeches.

3.2 <发展> <character bigram> <comma>

To be specific, let us examine in detail a relatively simple real world example given below.

<发展> <character bigram> <comma>

This template is exactly five characters long, with the left two characters fixed to the two Chinese characters "发展 (develop or development, if used as a word)", and the rightmost character bounded to the specific punctuation mark comma. The middle two character positions are left unconstrained to form the word holding slot. For this template, there is no text preprocessing required.

This template is not as trivial as the template "This is a ___." we illustrated in the section above. It is purposely selected for highlighting some potential inherent difficulties.

First, we assume no text pre-tokenization. At least theoretically, this will bring in all kinds of traps associated with the notorious Chinese text tokenization problem widely believed to be caused by the lack of English blank space equivalent word delimiter in Chinese text. In Chinese, "发展 (develop, development)" itself is a legitimate word, but "发展中 (developing)" is also an entry in some dictionaries. Moreover, in sentence "张发展翅飞翔 (Zhang Fa is flying high)", the character "发 (Fa)", the first character in "发展", is in fact the given name of the person "张发 (Zhang Fa)" and thus a part of the proper noun, and "展", the second character in "发展", is the initial character of the predicative idiom "展翅飞翔 (fly high)". In short, depending on the context it appeared, the continuous character bigram "发展" could be used as a word, or part of a word, or parts of two adjacent words.

Even if the text is properly tokenized and "发展" is confirmed to be used as a word in text, its part-of-speech is nevertheless ambiguous. Chinese is a language lack of inflections. Depending on the context it is used, the two-character word "发展" could be used as either a noun ("development") or a verb ("develop"), yet written exactly the same. We have to count on the context for part-of-speech determination and/or disambiguation. But in the template, in terms of the part-of-speech ambiguous Chinese word "发展", its right context is by design a "word holding slot" meaning an unknown word, yet its left context is not mentioned at all.

In addition, the template is not a full sentence. We do not even know whether or not the five-character-long template is a syntactically or semantically self-contained unit like a phrase or a term. Rather, in terms of the word holding slot, only two left characters and one right character are given. That is the only indicator or constraint.

In a word, the information provided in this template is rather poor. Many ambiguities and unknowns could be expected to arise even for human expert explorers simply because of the lack of information. Realistically, we could not expect high predictability of the template in word recognition.

Nevertheless, not to make the thing too hopeless, two positive constraints are also built into the template. One, the right context of the word holding slot is chosen to be a punctuation mark (the comma). This effectively removes the right boundary ambiguity of the word in question. Definitely, if the comma is replaced with a Chinese character string with an ambiguity level comparable with the left context "发展", the effectiveness of thus formed template will be even less predictable.

Second, the word holding slot is restricted to exactly two Chinese characters long. This effectively reduces the variety of words in question. We will report elsewhere case studies for other word holding slot lengths.

3.3 Data

We extracted all partial sentences matching the above template from the 4-million-character PH corpus (Guo and Lui, 1992). The complete list of all the 257 matches is in Appendix A. As the number of matches is not large, readers are recommended to have their own detailed examination. All the character pairs taking the template's word holding slot are listed in the table below. They are in the same order as in Appendix A. Duplications are not removed, since they may correspond to different context.

Table 1: The 257 Chinese character bigrams taking the word holding slot in template "<发展><bigram><comma>". Extracted from the PH corpus. Duplication preserved.

迅速	合作	较快	着想	较快	下去	速度	势头	经济	方向
经济	战略	规划	方向	起来	后劲	起来	经济	基金	养鱼
计划	迅速	指标	很快	腰果	道路	纲要	速度	目标	较快
潜力	目标	方向	畜牧	顺利	规划	资金	品种	生产	方针
壮大	较快	战略	同时	经济	情况	方面	更快	规律	生产
战略	战略	下去	很快	下去	经济	起来	时期	计划	最快
计划	计划	阶段	计划	水平	战略	项目	道路	不快	生产
趋势	后劲	动态	缓慢	关系	速度	服务	基金	基金	需要
政策	生产	生产	趋势	规划	较快	情况	生产	关系	农业
过多	过快	很快	基金	资金	规划	计划	阶段	迅速	历史
来说	太快	阶段	战略	水平	过程	实际	较快	规划	进程
迅速	纲要	很快	体育	服务	规划	方向	很快	生产	党员
较快	迅速	生产	中国	规划	计划	很快	情况	援助	生产
林业	顺利	任务	过程	壮大	阶段	能力	经济	道路	很快
较快	需要	道路	着眼	规划	很快	方向	经济	经费	模式
方向	旅游	变化	迅速	历史	农业	进步	战略	品种	缓慢
经济	规划	之中	迅速	多了	水平	重点	目标	旅游	情况
计划	公司	时期	速度	大计	规划	阶段	之路	之路	养羊
迅速	较快	迅速	来看	的县	上去	关系	迅速	战略	能力
任务	方向	过程	规划	壮大	情况	迅速	缓慢	起来	加工
需要	迅速	现状	方向	以后	很快	战略	经济	顺利	农业
经济	工作	减慢	迅速	态势	迅速	途径	重点	壮大	前景
计划	战略	需要	规划	经济	水平	速度	前景	情况	计划
农业	势头	探索	进程	阶段	绵羊	生产	蓝图	战略	条件
目标	中国	生产	战略	前途	关系	机遇	繁荣	中国	来说
趋势	战略	速度	战略	经济	基金	蓝图	*	*	*

3.4 Analysis

The majority listed above are legitimate dictionary words. There are only 39 bigrams not found in XianHan (1983), a famous medium size authentic Chinese dictionary with about 50,000 entries.

Table 2: analysis of the Chinese character bigrams not in the XianHan dictionary.

structure	num	Unknowns	words in XianHan
adv + adj	24	很快/9,较快/9,更快/1 最快/1,不快/1,过快/1 太快/1, 过多/1	大红
prop noun	3	中国/3	中国人民解放军
verb + verb	3	来说/2, 来看/1	(而言; 说来,看来)
的+ noun	3	之路/2, 的县/1	的话
verb + noun	2	养鱼/1, 养羊/1	养兵, 养地
之 + loco	1	之中/1	之前, 之后
adj + 了	1	多了/1	
verb + adj	1	减慢/1	减低,减轻,减弱,减少

Among the 39 non-dictionary bigrams, the dominant portion are those following the compounding pattern: ***"adv. + adj."***, where both the adverb and the adjective are a single Chinese character. There are totally 7 distinct types (很快(very fast)/9, 较快(relatively fast)/9, 更快(faster)/1, 最快(fastest)/1, 不快(not fast)/1, 过快(too fast)/1, 太快(overly fast)/1, 过多(too many)/1) accumulatively occurred 24 times. Note, ***"adv. + adj."*** is a valid Chinese word formation pattern (the so-called ***状中结构, modifier-predicate compounding***). There do exist in XianHan words like "大红 (bright red)" following this pattern. Moreover, all these bigrams are with high usage frequency and expressive mutual information score. That is, they are all "tight-in-combination and stable-in-use". Because

Chinese words, phrases, and sentences are formed under the same principle (Zhu, 1985), there is essentially no clear boundary for words, phrases and even sentences. Whether or not they are words might largely depend on the taste of individual lexicographers. Nevertheless, they are all legitimate and self-contained syntactic and/or semantic entities.

The rest 14 occurrences represents seven different types. Proper nouns such as country names like "中国 (China)" above are normally not collected in the main body of a dictionary. Rather, they are conventionally compiled as supplementary dictionary appendices. What we want to emphasize here is that such sub-language entities emerge themselves naturally in context.

"来说" and "来看" are not recorded in XianHan, but they are nevertheless tightly bounded and frequently used. In addition, they have quite unique characteristics in usage and peculiar (syntactic) meaning.

"的县" and "之路" are not in XianHan. But "的话" does. Similarly, XianHan has "之后" and "之前" but no "之中". Both "养羊" and "养鱼" are not in XianHan, but "养兵" and "养地" are in it. XianHan has "减低", "减轻", "减弱" and "减少", but no "减慢".

3.5 Conclusion

Frankly, I would hesitate to quantify the precision and/or recall achieved from the particular template above, as there exist significant inter-judge variances on whether or not what listed above are words. Instead, I make relevant data available in full and suggest readers to do their own calculation. What I hope to accomplish in this section is to convince readers the following two observations.

(1) Word/chunk identification/extraction by template matching is effective. At least this is true for the particular non-trivial template of "<发展><bigram><comma>". This observation is important as it provides us with an empirical support of our context-centered template matching word identification scheme.

(2) The effectiveness of a template in word identification is quantitatively measurable with respect to given corpus and dictionary. This observation is important as it implies the learnability or trainability of word identification templates. That is, we may have automatic ways for template preparation.

4 MATHEMATICAL MODELING

In this section, we will put the context-centered thinking into a broad world. Through formal mathematical modeling, we will achieve deep understanding on both the word identification problem and its different problem solving strategies. Entity modeling and context modeling are to be introduced and contrasted.

4.1 The World

Given context as a word holding template and entity taking up the word holding slot of the template, the question here is: whether or not the entity is a word. Assume there are only two clear-cut answers: "***yes***" or "***no***". Then, in the language of ***probability***, we have created a tiny world with three citizens or ***random variables***: the ***word holding template T*** taking set of mutually independent templates as the universe, the ***word holding slot S*** which could be filled up with certain type of entities, and the ***answer A*** to the question taking either "***yes***" or "***no***". Moreover, the joint probability

(1) $Pr(A, S, T)$

gives us the precise and complete description of the three-random-variable world. That is, having the joint probability (1) known, any question about the world could be answered.

4.2 Word Identification Modeling

In particular, suppose word holding template $T=t$ match a natural language sentence extracted from a corpus, and $S=s$ be the sentence fragment taking the word holding slot. Then, it has been well established (Duda and Hart, 1973) that, on average, the ***minimum error decision*** is, $S=s$ is a word in sentence if and only if there holds

(2) $Pr(A=yes, S=s, T=t) > Pr(A=no, S=s, T=t).$

Assume

(3) $Pr(T,S|A) = Pr(T|A)Pr(S|A),$

and denote

(4) $Lt = ln\ Pr(A=yes|T=t) / Pr(A=no|T=t),$

(5) $Ls = ln\ Pr(A=yes|S=s) / Pr(A=no|S=s),$

(6) $La = ln\ Pr(A=yes) / Pr(A=no),$

the decision rule (2) could be rewritten as

(7) $Lt + Ls > La.$

That is, under the assumption given in formula (3), to have minimum error solution to the word identification problem is equivalent to calculate the ***context likelihood Lt***, the ***entity likelihood Ls*** and the ***solution likelihood La***, and to make decision with rule (7). This is in turn equivalent to estimate the ***context probability Pr(A\|T)***, the ***entity probability Pr(A\|S)*** and the ***solution probability Pr(A).***

4.3 Entity Modeling

The task of entity modeling is to estimate for any possible entity $S=s$ the entity likelihood defined in formula (5):

(5) $Lt = ln\ Pr(A=yes|S=s) / Pr(A=no|S=s).$

As we elaborated before, depending on system configuration, an entity could be a plain character string, a string of simple words, or some type of character, word, and part-of-speech combinations. We use the term ***entity*** for generality.

Note that, since we have assumed only two possible answers, there holds

(8) $Pr(A=yes|S=s) + Pr(A=no|S=s) = 1.$

Then, the entity likelihood could be equivalently written as

(9) $Lt = ln\ Pr(A=yes|S=s) / (1 - Pr(A=yes|S=s)).$

Hence, the heart of entity modeling is in modeling the probability $Pr(A=yes|S=s)$ for any possible entity s. In essence, the probability $Pr(A=yes|S=s)$ is expressing the fitness of entity $S=s$ as a word out of context.

Entity modeling is nothing but the theme in both sub-language oriented and entity-centered word identification research. Numerous modeling approaches have been proposed in literature. For example, in Lee, Lee, and Chen (1994), the probability of character trigram $C_1C_2C_3$ used as a Chinese name is estimated as the product of $Pr(C_1|surname)$, the probability of the first character C_1 as a surname, $Pr(C_2|middlename)$, the probability of the second character C_2 as a middle name, and $Pr(C_3|givenname)$, the probability of the third character C_3 as a given name:

(10) $Pr(A=yes|S=C_1C_2C_3)$
$= Pr(C_1|surname)\ Pr(C_2|middlename)\ Pr(C_3|givenname).$

The ngram-based approach by Fung and Wu (1994) is in fact implicitly modeling $Pr(A=yes|S=ngram)$ through a statistical decision procedure, while their linguistic filters (Wu and Fung, 1994) are implicitly modeling $Pr(A=no|S=ngram)$ through a set of linguistic selection rules.

4.4 Solution Modeling

The task of solution modeling is to estimate the solution likelihood defined in formula (6):

(6) $La = ln\ Pr(A=yes) / Pr(A=no).$

Similar to what we did for entity modeling, only the probability $Pr(A=yes)$ is to be estimated. Theoretically, this could be done directly from a corpus by counting the cases where positive and negative decisions are made.

In practice, however, the likelihood of $La=lnPr(A=yes)/Pr(A=no)$ is better obtained from user assignment than from corpus training. Similar to what we did in Chinese spell checking and proofreading (Guo, 1994), the solution likelihood could be used as a control variable for recall/precision balancing. Different solution likelihood settings will result in different recall and precision rates. You may be able to get high precision at the cost of low recall, or to go the other way around. We found this is a practical mechanism for fulfilling different preferences within the single system.

4.5 Context Modeling

The task of context modeling is to estimate from any possible word holding template $T=t$ the context likelihood defined in formula (4):

$$(4) \qquad Lt = ln\ Pr(A=yes|T=t) / Pr(A=no|T=t).$$

Similar to the discussion above, this likelihood could also be written as

$$(11) \qquad Lt = ln\ Pr(A=yes|T=t) / (1 - Pr(A=yes|T=t)).$$

The core here is to estimate $Pr(A=yes|T=t)$, the probability of entity taking the word holding slot of the word holding template $T=t$ a valid word. Given a tokenized corpus, its ***Maximum Likelihood Estimation*** (*MLE*) is:

$$(12) \qquad Pr(A=yes|T=t) = n/N$$

where N = "the number of times template $T=t$ matches a corpus sentence" and n = "the number of times the entity in the matched template slot is a valid word".

For example, in the case study above, there is $N=257$. If we treat only those in XianHan (1983) as valid word, there is $n=219$. Thus, there are:

$$Pr(A=yes|T="<发展><bigram><comma>")=219/257=0.85,$$

$$Lt = ln\ 219/38=1.75.$$

If only the three cases under the pattern ***"的 + noun"*** are considered non-words, there are:

$$Pr(A=yes|T="<发展><bigram><comma>")=254/257=0.99,$$

$$Lt = ln\ 254/3=4.44.$$

4.6 Template-Slot Independence Assumption

It must be pointed out that, to reach the word identification model (7), we explicitly made the following assumption in formula (3):

$$(3) \qquad Pr(T,S|A) = Pr(T|A)Pr(S|A).$$

This assumption could be read as that, with respect to any specific answer $A=a$, the word holding template T and the word holding slot S are probabilistically independent. In this paper, this is referred to as ***Template-Slot Independence Assumption***.

This assumption does not hold for templates such as parts of collocational patterns. This assumption is adopted for two reasons. First, we are not aware of any practically computable modeling approach which could take into account all the three random variables simultaneously. We have to decompose the three variable world into several one or two variable subworlds. That is, we have to accept some compromise.

Second, tight-bounding collocations are relatively rare in real text. For an English corpus of 10 million words, according to Smadja (1993), only about 8,000 collocations are reliable, a neglectable portion among millions of other bigrams and ngrams. Moreover, as words which could take up slots in such collocational templates are highly constrained, such collocations, even if chosen as word holding templates, are not productive.

In short, the assumption we adopt is necessary for efficient computational modeling and feasible for productive practical application.

4.7 Short Summary

According to the word identification model given in formula (7), to solve the problem unbiasly, both entity modeling and context modeling are necessary. In literature, however, the emphasis has almost exclusively been put on entity modeling and it is hardly possible to find a proposal with the context model explicitly incorporated.

In contrast, by exploring the word identification problem in such a systematic way, we make it explicit the importance of context modeling, and proposed a concrete context-modeling approach.

The use of context information in such an explicit and systematic way is the fundamental difference between the approach proposed in this paper and the rest in literature.

5 ALGORITHM

In this section, we will first go to the extreme by trying to construct lexicon and perform tokenization ***without dictionary.*** Such an algorithm is proposed. Some notes, variations and improvements are then given to contrast the key idea and to make the thing realistic.

5.1 Start from Empty

In Section 3, we presented in detail a case study which gives us strong confidence on context-centered word identification and lexicon construction. Note, in that case study, we did not do any preprocessing. In particular, we did not do text word tokenization. Moreover, we did not use any dictionary in word identification. The dictionary XianHan (1983) is employed only for human evaluation of the effectiveness of the scheme. Explicitly, we are building up lexicon from scratch and doing text tokenization without lexicon. This section is on the algorithm proper.

5.2 Baseline Algorithm

The baseline algorithm has four phases:

5.2.1 Initialization (n=1)

- Put all none Chinese character symbols in the working character set such as GB, Big5 or UNICODE into the lexicon. Punctuation marks, numerical digits and Roman alphabets are part of these symbols.
- Put all function words, such as prepositions, pronouns, particles and classifiers, into the lexicon.
- Put a few special Chinese characters such as "是" and "有" into the lexicon.
- Put a special sentence begin symbol and a special sentence end symbol into the lexicon.
- Prepare a large text corpus by appending the sentence begin and sentence end symbols at the two ends of every sentence in corpus.
- Set a maximum word length.

5.2.2 Bootstrapping (n=2)

- For each word W_1 and W_2 in the above initialized lexicon, form word holding template $T(W_1,W_2)$ = ***<W1><Chinese character bigram><W2>,*** where the word ***W1*** and ***W2*** are the left and right context respectively, and the ***Chinese character bigram*** takes the word holding slot.
- For each thus formed word holding template and for each sentence in corpus, check to see if there is a match.
- If yes, put the Chinese character bigram found in the matched Chinese sentence and taking the word holding slot of the word holding template into lexicon.

5.2.3 Iteration (n>2)

Suppose lexicon up to word length n-1, n>2, has been constructed. Now, to augment the lexicon with words of length n,

- For each word W_1 and W_2 in the previously constructed lexicon, form word holding template $T(W_1,W_2)$ = ***<W1><Chinese character ngram><W2>***, where the word *W1* and *W2* are the left and right context respectively, and the ***Chinese character ngram*** takes the word holding slot.
- For each thus formed word holding template, evaluate its context likelihood *Lt* for word holding slot of length $m<n$ Chinese characters. This is done by (1) matching the modified (that is, the requirement for the word holding slot has been relaxed to allow any Chinese $m<n$ grams) word holding template against the whole corpus, (2) counting the number of m-grams which are taking the word holding slot and are or are not words in the latest lexicon, and (3) calculating the log ratio according to the formula $Lt = ln\ Pr(A=yes|T=t) / Pr(A=no|T=t)$ introduced in the last section.
- If the context likelihood is better than a preset threshold (the solution likelihood), the template is confirmed.
- For each thus confirmed word holding template and for each sentence in corpus, check to see if there is a match.
- If yes, put the Chinese character ngram, which is found in the matched Chinese sentence and takes the word holding slot of the confirmed word holding template, into lexicon.

Repeat the above procedure successively for each n, until it reaches the preset maximum word length limit.

5.2.4 Completion (n=1)

- Add all single Chinese characters into the lexicon.

5.3 Notes

First, in Phase 2, only bigrams are collected, and bigrams involving none Chinese character symbols are excluded.

Second, compared with Phase 2, template evaluation and confirmation are added to Phase 3. Operations in this phase will be executed consecutively for n=3 up to the preset maximum word length.

Third, the last phase, Phase 4, is to make the lexicon complete. As defined in (Guo, 1996), a lexicon is complete if all words in open text are in the lexicon. Operationally, this is proven to be equivalent to have all characters in the working character set be included in the lexicon (Guo, 1996).

5.4 Improvements and Variations

Countless improvements and variations of the baseline algorithm are possible. To keep the description brief, only a few are discussed below.

5.4.1 "Top Star" Templates

In the baseline algorithm above, both the left and right template context are single word long. This could be updated to allow multi-word context for higher reliability. Most rare templates are better to be discarded in production, since their word predicting power could not be faithfully estimated. A practical way is to generate various context length templates but only keep those appearing at least several dozen times in the corpus. This trick will both boost the lexicon's quality and the algorithm's computational efficiency. The efficiency comes from the fact that, only high frequency entries in the lexicon need to be considered in template preparation, evaluation and matching, yet these high frequency entries are always a tiny portion of the whole lexicon. Moreover, the frequent appearance of a template in corpus makes the estimation of its context likelihood operationally reliable, and thus reduce the noise level.

Methodologically, this is essentially ***to let a few "top star" context templates call back the "silent majority" slot words.***

5.4.2 Resource

In what we did above, there is neither lexicon nor lexicon-based tokenizer employed. However, the two resources are not too difficult to obtain. Suppose we already have a machine readable high quality dictionary and a lexicon-based tokenizer such as the simplest longest (maximum) matching segmentor. Then, we could start by using that dictionary as initial lexicon. Moreover, before each iteration begin, we could tokenize the corpus according to the current available lexicon. This will help us both in reducing errors caused by cross word boundary template matching and in enhancing computational efficiency through precompiling corpus ngram statistics. Furthermore, part-of-speech tagger could also be incorporated.

5.4.3 Linear Pattern Matching

It is also worth noting here the ***Aho-Corasick*** pattern matching algorithm (Aho and Corasick, 1975, Aho, 1990) and its derivation (Pinter, 1985) which allows the inclusion of don't-care symbols in patterns. By organizing lexicon and/or templates in ***trie*** (Knuth, 1973) structure, the pattern matching can be implemented in time linearly proportional to the corpus size and independent of the size of lexicon and/or template. However, detailed discussions on efficiently applying these algorithms to lexicon-based tokenizations are out of the scope of this paper. Interested readers are referred to (Guo, 1996).

6 CONCLUSIONS

In this paper, we have established the fresh context-centered Chinese lexicon construction methodology which is first introduced with an intuitive explanation and an informative case study, and then generalized with formal mathematical modeling and algorithm development. Due to space limit, experimental evaluations will be reported elsewhere.

The primary advantage of the context-centered lexicon construction approach is its unmatchable productivity of recalling the "silent majority". The algorithm we proposed could extract almost all words and phrases in text, even if they only appear once or twice. Moreover, as our case study depicted, the precision could also be very good.

Under the guidance of our mathematical word identification model, several possible improvements are also highlighted in this paper. It is understood that we have restricted ourselves to be within the framework of traditional syntactic analysis. What we believe to be worth further pursuing are the following two aspects:

- ***Chunks***: In English, a chunk is essentially a "non-recursive core of an intra-clausal constituent" (Abney, 1991, Abney, 1996). To make it applicable to Chinese, some adaptations are required. The principle, however, is nevertheless there. Only after introducing the concept of chunks, we can have a better understanding of our task. In Chinese, lexicon construction is nothing but chunk collection. It is practically not critical to argue the definition of words. Rather, if an ngram is a chunk, we collect it.
- ***Semantics***: The more effective way of enhancing the quality of a lexicon is by introducing semantic databases such as the multilingual ***SenseWeb*** (Dong and Guo, 1996) or the English ***WordNet*** (Miller, 1990). Normally we do not work from scratch but augment an already validated lexicon. Then, with SenseWeb or WordNet, we could check the similarity between what we are extracting from corpus and what are already in the lexicon.

ACKNOWLEDGEMENTS

Great thanks to Prof. Dong Zhen Dong and Dr. Robert Luk for insightful discussion, genius help and paper critiquing. Nurdini helped me a lot in improving the English writing.

REFERENCES

Abney, Steven (1991), ***Parsing by Chunks***, In Robert Berwick, Steven Abney and Carol Tenny (eds.), Principle-Based Parsing, Kluwer Academic Publishers, Dordrecht.

Abney, Steven (1996), ***Chunk Stylebook***, On-line document, http://www.sfs.nphil.uni-tuebingen.de/~abney/96i.ps.gz.

Aho, A. V., (1990), ***Algorithms for Finding Patterns in Strings***, in: (ven Leeuwen, J., eds.) Handbook of Theoretical Computer Science, Volume A, Algorithms and Complexity, Chapter 5, pp. 273-278, The MIT Press.

Aho, A. V., and Corasick, M. J., (1975), ***Efficient String Matching: An Aid to Bibliographic Search***, Communications of ACM, 18 (6), pp. 333-340.

Black, Ezra, Roger Garside, and Geoffery Leech, (1993), ***Statistically-Driven Computer Grammars of English: The IBM/Lancaster Approach***, Amsterdam: Rodopi Publishers.

Chang, J. S., Chen, S. D., Ker, S. J., Chan, Y. and Liu J. S., (1994), ***A Multi-Corpus Approach to Recognition of Proper Names in Chinese Texts***, Computer Processing of Chinese and Oriental Languages, Vol. 8, No. 1, page 73-86.

Chen, Keh-Jiann, and Shing-Huan Liu, (1992), ***Word Identification for Mandarin Chinese Sentences***, In Proceedings COLING-92, Nantes, pp. 101-107.

Chiang, Tung-Hui, Jing-Shin Chang, Ming-Yu Lin, and Keh-Yih Su, (1992), ***Statistical Models for Word Segmentation and Unknown Word Resolution***, In Proceedings of ROCLING-92, pp. 121-146.

Church, Kenneth. W., and William A. Gale, (1991), ***A Comparison of the Enhanced Good-Turing and Deleted Estimation Methods for Estimating Probabilities of English Bigrams***, Computer Speech and Language, Vol. 5, No. 1, page 19-54.

Church, Kenneth. W., and P. Hanks, (1990), ***Word Association Norms, Mutual Information and Lexicography***, Computational Linguistics, Vol. 16, No. 1, pp. 22-29.

Dong, Zhen Dong and Jin Guo, (1996), ***Construction and Application of SenseWeb -- a Multilingual Semantic Lexicoal Database***, A Tutorial at the 1996 International Conference on Chinese Computing (ICCC96), Singapore.

Duda, R. O., and P. E. Hart, (1973), ***Pattern Classification and Scene Analysis***, New York, John Wiley & Sons.

GB-13715, (1993), ***Contemporary Chinese Language Word Segmentation Specification for Information Processing***, PRC National Standard, China National Standard Bureau.

Guo, Jin, and Ho Chung Lui, (1992), ***PH: a Chinese Corpus for Pinyin-Hanzi Transcription***, Technical Report TR93-112-0, Institute of Systems Science, National University of Singapore.

Guo, Jin, (1993), ***Statistical Language Modeling and Some Experimental Results on Chinese Syllables to Words Transcription***, Journal of Chinese Information Processing, Vol. 7, No. 1, pp. 18-27.

Guo, Jin, (1994), ***Automatic Chinese Spelling Checking***, A Tutorial at the 1994 International Conference on Chinese Computing (ICCC94), Singapore, Part I, 31 pages, Part II, 44 pages.

Guo, Jin, (1996), ***An Efficient and Complete Algorithm for Unambiguous Word Boundary Identification***, (to be submitted), available on-line: http://sunzi.iss.nus.sg:1996/guojin/papers/acbci.ps.gz.

Fung, Pascale, and Dekai Wu, (1994), ***Statistical Augmentation of a Chinese Machine-Readable Dictionary***, In Proceedings of the Second Annual Workshop on Very Large Corpora (WVLC-2), Kyoto.

Harman, Donna K. (eds.) (1995), ***Overview of the Third Text REtrieval Conference (TREC-3),*** NIST Special Publication, Washington, DC: US Government Printing Office.

Harman, Donna K. (eds.) (1996), ***Overview of the Fourth Text REtrieval Conference (TREC-4),*** NIST Special Publication, Washington, DC: US Government Printing Office.

Hu, Yu Shu, (1987), ***Xiandai Hanyu (Modern Chinese),*** Shanghai Education Press.

Knuth, D. E., (1973), ***Fundamental Algorithms***, second edition, The Art of Programming, Vol. 1. Addison-Wesley, Reading, Mass., pp. 487-499.

Kucera, H., and Francis, W. N., (1967), ***Computational Analysis of Present-Day American English***, Providence, Brown University Press.

Lee, C, Y. Lee, and H. Chen, (1994), ***Research of the Identification of Names in Chinese Text***, In Proceedings of ROCLING VII, pp. 203-222.

Lin, Ming-Yu, Tung-Hui Chiang, and Keh-Yih Su, (1993), ***A Preliminary Study on Unknown Word Problem in Chinese Word Segmentation***, In Proceedings of ROCLING VI, pp. 119-141.

Lua, K. T., (1995), ***Experiments on the Use of Bigram Mutual Information in Chinese Natural Language Processing***, In Proceedings of the 1995 International Conference on Computer Processing of Oriental Languages (ICCPOL-95), Hawaii, pp. 306-313.

Luk, Robert W. P., (1995), ***Automatic Tokenization of Chinese Text Driven by a Lexicographic Index with Linguistic Pattern Filtering***, In Proceedings of the 1995 International Conference on Computer Processing of Oriental Languages (ICCPOL-95), Hawaii, pp. 217-224.

Luk, Robert, W. P., (1996), Personal Communications.

Markus, M., M. A. Marcinkiewicz, and B. Santorini, (1993), ***Building a Large Annotated Corpus of English: The Penn Treebank***, Computational Linguistics, Vol. 19, No. 2, page 313-329.

Miller, George A., (1990), ***WordNet: An On-line Lexical Database***, International Journal of Lexicography, Vol. 3, No. 4, page 235-312.

Mo, Ruo-ping, Yao-Jung Yang, Keh-Jiann Chen, and Chu-Ren Huang, (1991), ***Determinative-Measure Compounds in Mandarin Chinese: Formation Rules and Parser Implementation***, In Proceedings of ROCLING-IV, pp. 111-134.

Nie, J-Y., Jin W-Y., and Hannan, M-L., (1994), ***A Hybrid Approach to Unknown Word Detection and Segmentation of Chinese***, In Proceedings of International Conference on Chinese Computing 1994 (ICCC-94), Singapore, page 326-335.

Pinter, R. Y., (1985), ***Efficient String Matching with Don't-Care Patterns***, in: A. Apostolico and Z. Galil, (eds.), Computational Algorithms on Words, Springer Berlin.

Smadja, Frank, (1993), ***Retrieving Collocations from text: Xtract***, Computational Linguistics, Vol. 19, No. 1, pp. 143-177.

Song, Rou, Chaojie Qiu, Longgeng Ouyang, Lubing Xu, and Xin Wang, (1996), ***Bi-Orderly-Neighborship and Its Application to Chinese Word Segmentation and Proofreading***, In Proceedings of the 1996 International Conference on Chinese Computing (ICCC96), Singapore, pp. 428-433.

Sproat, Richard, and Shih, Chilin, (1990), ***A Statistical Method for Finding Word Boundaries in Chinese Text***, Computer Processing of Chinese and Oriental Languages, Vol. 4, No. 4, page 336-349.

Sproat, Richard, (1994), ***English Noun-Phrase Accent Prediction for Text-to-Speech***, Computer Speech and Language, 1994, No. 8, pp. 79-94.

Sun, Maosong, and Changning Huang, (1996), ***Word Segmentation and Part-of-Speech Tagging for Unrestricted Chinese Texts***, A Tutorial at the 1996 International Conference on Chinese Computing (ICCC96), Singapore.

Wu, Dekai, and Pascale Fung, (1994), ***Improving Chinese Tokenization with Linguistic Filters on Statistical Acquisition***, In Proceedings of the 4th Conference on Applied Natural Language Processing (ANLP-94), Stuttgart, pp. 180-181.

Wu, H-J., Jin Guo, Ho Chung Lui., and Hwee Boon Low., (1994), ***Corpus-Based Speech and Language Research in the Institute of Systems Science***, In Proceedings of International Symposium on Speech, Image Processing and Neural Networks (ISSIPNN-94), Hong Kong, pp. 142-145.

XianHan (1983), ***现代汉语词典 (Modern Chinese Dictionary)***, 2nd edition, Commercial Press, Beijing.

Yang, Jin, and Laurie Gerber, (1996), ***SYSTRAN Chinese-English Machine Translation System***, In Proceedings of the 1996 International Conference on Chinese Computing (ICCC96), Singapore, pp. 205-210.

Yuan, Baosheng, Yuqin Gao, Hisao-Wuen Hon, Jean-Luc Lebrun, Zhiwei Lin, Gareth Loudon, and Xi Han, (1996), ***Chinese Dictation Kit: A Very Large Vocabulary Mandarin Speech Input System***, In Proceedings of the 1996 International Conference on Chinese Computing (ICCC96), Singapore, pp. 1-4.

Zheng, J. and K. Liu, (1993), ***Approach of Processing Tactics on the Names and Surnames in Chinese Automatic Segmenting System***, In Chen and Yuan (eds.) Computational Linguistics: Research and Applications, Beijing Linguistics Institute Publisher, Beijing, pp. 139-143.

Zhu, Dexi, (1985), ***语法答问 (Questions and Answers on Chinese Grammar)***, Commercial Press, Beijing.

APPENDIX

Listed below are the 257 partial sentences in the PH corpus matching template "<发展><bigram><comma>". See section 3 for analysis.

1.制盐业和有色金属产业发展迅速,
2.苏联外长谢瓦尔德纳泽和１９日抵苏访问的民主德国外长菲舍尔在２０日会谈中主张在华约范围内进一步发展合作,
3.全区勤工俭学近几年发展较快,
4.鞍钢党政领导从企业的长期发展着想,
5.技术出口发展较快,
6.稳定地发展下去,
7.在困难的条件下保持了正常的发展速度,
8.我国地方铁路建设呈现持续发展势头,
9.努力发展经济,
10.关于乡镇企业今后的发展方向,
11.它对发展经济,
12.实现分三步走的经济发展战略,
13.二是制定了发展规划,
14.稳定面积攻单产的发展方向,
15.把生产力发展起来,
16.增强农业发展后劲,
17.羽绒厂等加工企业迅速发展起来,
18.向全县人民发出了“发展经济,
19.并尽快建立竹类科技发展基金,
20.现在全区已有２０多万劳力从事专业发展养鱼,
21.在研究制定经济和社会发展计划,
22.乡镇企业起步晚但发展迅速,
23.强调以农业特别是粮食发展指标,
24.农村经济发展很快,
25.在沿海发展腰果,
26.找到并坚持了符合本国国情的发展道路,
27.省政府制订了农田水利发展纲要,
28.这一惊人的发展速度,
29.生态发展目标,
30.卷烟工业发展较快,
31.又有发展潜力,
32.对围绕产业发展目标,
33.低残留的发展方向,
34.发展畜牧,
35.两国在各个领域的友好合作关系发展顺利,
36.孟加拉国政府制订了长期发展规划,
37.由政府向全国足球协会提供发展资金,
38.发展品种,
39.帮助群众发展生产,
40.抓落实”的发展方针,
41.指出中国的振兴最终要靠高科技和新技术产业的发展壮大,
42.去年捷同非社会主义国家的易货贸易发展较快,
43.江西省各级党政部门自觉围绕这一发展战略,
44.在畜牧业稳步发展同时,
45.会议呼吁发达国家帮助最不发达国家更快发展经济,

46.他这次访泰的目的是研究柬埔寨局势发展情况，
47.找到最佳发展方面，
48.发展更快，
49.不断揭示各种灾害的成因和发展规律，
50.不少地方应用这种方式发展生产，
51.各国政府有责任根据本国国情制定发展战略，
52.制定出切合本国实际的发展战略，
53.如此发展下去，
54.近年来亚太地区的经济发展很快，
55.如果这种局势发展下去，
56.南亚各国才能发展经济，
57.我国海洋石油工业是在改革开放中发展起来，
58.今年我国旅游业进人了恢复和发展时期，
59.批准１９８９年国民经济和社会发展计划执行情况和１９９０年国民经济和社会发展计划，
60.步人了历史上又一个发展最快，
61.批准１９８９年国民经济和社会发展计划执行情况和１９９０年国民经济和社会发展计划，
62.批准１９８９年国民经济和社会发展计划执行情况和１９９０年国民经济和社会发展计划，
63.进人了一个新的发展阶段，
64.我们制定方针政策和经济社会发展计划，
65.多数民族自治地区将会逐步接近或赶上当时全国的中等发展水平，
66.根据国家的产业政策和发展战略，
67.农林牧方面的发展项目，
68.使匈牙利迈向现代化的发展道路，
69.目前铁路发展不快，
70.发展生产，
71.对我国横向经济联合的发展趋势，
72.增强企业发展后劲，
73.密切注视国际造船业的发展动态，
74.欧洲企业减少在非洲的投资主要是因为非洲市场发展缓慢，
75.各地区及有关国际组织保持和发展关系，
76.工业生产保持一定的发展速度，
77.新课程的设置与发展服务，
78.建立农业发展基金，
79.海安县利用筹集的约２０００万元农业发展基金，
80.又结合香港的实际情况和发展需要，
81.他们希望政府从速制订中长期发展政策，
82.牲畜和农用机具等物资以发展生产，
83.企业承包经营责任制无论是对发展生产，
84.充分了解和科学地预测世界高技术发展趋势，
85.大多数省市都已制定了特殊教育的发展规划，
86.西藏自治区社会福利事业近年来发展较快，
87.双方讨论了巴勒斯坦问题的最近发展情况，
88.双方都要求通过紧密的协作发展生产，
89.特别是伊朗发展关系，
90.进一步发展农业，
91.宾馆饭店发展过多，
92.近几年棉纺能力发展过快，
93.蔡塘村发展很快，
94.山东省各级财政部门还将征集的４亿多元的农业发展基金，
95.这个省还注意积累渔业发展资金，
96.成立了这个旨在研讨产业政策布局和发展规划，
97.根据中国的发展计划，
98.新疆维吾尔自治区民族团结进人一个新的发展阶段，
99.保险业务发展迅速，
100.回顾我们党的新闻事业的发展历史，
101.就新闻战线自身的发展来说，
102.由于近年来新闻队伍发展太快，
103.中国出版业跨人了全新的发展阶段，
104.必须服从国家的总体发展战略，
105.不能脱离我国的经济文化整体发展水平，
106.回顾总结党的十一届三中全会以来的历史发展过程，
107.紧密围绕天津教育发展实际，
108.港口等单位运输１０吨集装箱的业务发展较快，
109.制定出一个发展规划，
110.发达国家限制纺织品进口阻碍了发展中国家的发展进程，
111.我国戏曲电视剧近年发展迅速，
112.科学技术长期合作发展纲要，
113.“近年来首钢的事业发展很快，
114.“发展体育，
115.为各项事业的发展服务，
116.市政府就制定了“菜篮子”工程５年发展规划，
117.麦棉套种是发展方向，
118.国外某个领域的技术发展很快，
119.尽快恢复和发展生产，
120.大冶钢厂党委重视在生产一线发展党员，
121.除了江南地区近几年经济发展较快，
122.由于发展迅速，
123.努力发展生产，
124.才能发展中国，
125.我国高技术研究发展规划，

126.为了今后更好地实施“８６３”高技术研究发展计划，
127.河南省小火电机组近年来发展很快，
128.双方回顾了１０年来两国教育学术交流的发展情况，
129.即把它们国民生产总值的百分之零点七提供为官方发展援助，
130.继续努力发展生产，
131.龙泉市重视依靠科技发展林业，
132.“如果事态发展顺利，
133.按照全会提出的国民经济和社会发展任务，
134.从单项服务到系列化服务的发展过程，
135.集体经济越是发展壮大，
136.这标志着我国社会主义现代化建设进入了一个新的发展阶段，
137.对增强贫困地区经济发展能力，
138.波兰现在正进行改革和发展经济，
139.走什么样的发展道路，
140.蒙中两国关系在各个领域里发展很快，
141.这个地区东部发展较快，
142.产业结构调整及外向型企业发展需要，
143.选择什么发展道路，
144.二是从贫困地区的长期发展着眼，
145.必须把扶贫开发列入本地区的国民经济和社会发展规划，
146.也不会发展很快，
147.要在对社会实际情况进行深入分析和研究的基础上确定发展方向，
148.不仅要求重视发展经济，
149.国务院有关部门还拨给西藏一些专项教育发展经费，
150.科研机构要积极探索多种形式的管理和发展模式，
151.选择什么发展方向，
152.发展旅游，
153.随着国际形势和匈中两国情况的发展变化，
154.我国橡胶工业发展迅速，
155.他在谈话中回顾了匈中关系的发展历史，
156.就是要充分运用一切科技成果发展农业，
157.共同为中华妇女事业的发展进步，
158.有必要重新审视上海的发展战略，
159.发展品种，
160.因而奶业的发展缓慢，
161.发展经济，
162.十年改变贫困面貌的发展规划，
163.竭力把科学技术物化活化在经济与社会发展之中，
164.西藏自治区卫生事业发展迅速，
165.个体和私营经济不是发展多了，
166.达到小康发展水平，
167.钢铁等基础工业是中国“十年规划”和“八五计划”中的发展重点，
168.既要考虑有利于解决经济发展中的深层次问题和实现中长期的发展目标，
169.加之具有长期贸易和旅游历史的一些其他国家仍在大力发展旅游，
170.阿里维勃沃一行应国务院特区办邀请前来了解中国经济特区的发展情况，
171.农业部将尽快修订颁发“八五”农机化发展计划，
172.由内地大学毕业生为主组建的新疆唐布拉贫困与发展公司，
173.检察技术工作目前正处于一个重要的发展时期，
174.之发展速度，
175.一些著名科技专家最近在京聚会研讨我国航天事业发展大计，
176.这个省各级税务部门围绕本地经济发展规划，
177.财贸工会进入了新的发展阶段，
178.贵州虹山轴承厂走外向型企业发展之路，
179.贵州虹山轴承厂走外向型企业发展之路，
180.村里发展养羊，
181.发展迅速，
182.体育事业发展较快，
183.以乡镇工业为主体的农村非农产业发展迅速，
184.因为从香港长远的经济发展来看，
185.就是涌现了一批经济持续稳定发展的县，
186.一个村的经济能不能发展上去，
187.同日本自民党和社会党以及日本其他政党扩大与发展关系，
188.但女子足球在朝鲜却发展迅速，
189.李铁锤制定了以技术进步为主导的发展战略，
190.初步形成了自我调节和自我发展能力，
191.按照《建议》提出的国民经济和社会发展任务，
192.它进一步明确了农村的基本经济政策和发展方向，
193.从单项服务到系列化服务的发展过程，
194.农村工作开创新局面的发展规划，
195.集体经济越是发展壮大，
196.双方相互介绍了各自国家当前科技和经济发展情况，
197.广西对外保险事业发展迅速，
198.多年来发展缓慢，
199.使服务体系尽快地发展起来，

200.有些产品可以采取城乡联合或协作的办法发展加工，
201.适应旅游业发展需要，
202.近年来我国农业机械化事业发展迅速，
203.国家科委根据世界科学技术发展现状，
204.这一技术已被公认为通信网的发展方向，
205.村办工业发展以后，
206.珠海十年来发展很快，
207.共同探讨电信事业的发展战略，
208.为进一步帮助边疆地区发展经济，
209.台资在这里发展顺利，
210.重视发展农业，
211.为调动各个方面的积极因素发展经济，
212.以确保妇女更加踊跃地参与国家的社会和经济发展工作，
213.生产发展减慢，
214.改革开放以来通讯事业发展迅速，
215.为各级政府了解社会经济发展态势，
216.在短期内发展迅速，
217.发展途径，
218.确定发展重点，
219.保护现有企业的生产力并使之不断发展壮大，
220.两国牢固的关系有着广阔的发展前景，
221.被列人市国民经济发展计划，
222.从北往南推进的发展战略，
223.中国将根据经济发展需要，
224.他准备到国家民委汇报全州发展规划，
225.发展经济，
226.是适应我国生产力发展水平，
227.今后１０年保持百分之六的年均发展速度，
228.这次会关系大陆发展前景，
229.双方相互介绍了各自国家当前科技和经济发展情况，
230.同时港府正在执行一项研究和发展计划，
231.依靠科技进步发展农业，
232.显示出蒸蒸日上的发展势头，
233.十年的发展探索，
234.符合中国的国情和历史发展进程，
235.李鹏的报告清楚地表现出中国第三代领导把９０年代视为标志着中国进人一个新的发展阶段，
236.如果在南方建立人工草地１亿多亩发展绵羊，
237.贵阳市劳动局还想方设法帮助集体企业发展生产，
238.“李鹏总理的报告勾画出了今后十年的发展蓝图，
239.制订国家长期科技发展战略，
240.也是改善妇女参与社会发展条件，
241.实现教育发展目标，
242.只有社会主义才能发展中国，
243.我们党和国家的中心任务是发展生产，
244.我们必须确立“科技兴国”的经济发展战略，
245.只要有发展前途，
246.特别是中日两国应该积极发展关系，
247.民族地区面临着极好的发展机遇，
248.“李鹏总理的报告为未来十年新疆的发展繁荣，
249.在稳定中发展中国，
250.这对于一个新的产业体系的建立和发展来说，
251.预测发展趋势，
252.大力调整了指导思想和发展战略，
253.山东的有关部门就全省的经济发展速度，
254.山东在今后十年已进一步明确了以效益为中心的发展战略，
255.联合发展经济，
256.深圳市已设立经济合作发展基金，
257.今年的“两会”描绘今后十年的发展蓝图，