國 立 交 通 大 學

電信工程學系

碩 士 論 文

中文字轉音系統之文句分析
的進一步研究

A Further Study on Text Analysis
For Mandarin TTS

研 究 生：傅明榮

指導教授：王逸如　博士

中　華　民　國　九十六　年　七　月

中文字轉音系統之文句分析的進一步研究

A Further Study on Text Analysis for Mandarin TTS

研 究 生：傅明榮　　Student：Ming-Zong Fu

指導教授：王逸如 博士　　Advisor：Dr. Yih-Ru Wang

國 立 交 通 大 學

電 信 工 程 學 系

碩 士 論 文

A Thesis
Submitted to Department of Communication Engineering
College of Electrical Engineering and Computer Science
National Chiao Tung University
in partial Fulfillment of the Requirements
for the Degree of Master
in
Electrical Engineering

July 2007
Hsinchu, Taiwan, Republic of China

中華民國九十六年七月

中文字轉音系統之文句分析的進一步研究

研究生：傅明榮　　指導教授：王逸如 博士

國立交通大學電信工程學系碩士班

中文摘要

在本論文中，我們建立詞綴構詞單元模組至中文斷詞器內，以改善某些衍生詞無法窮舉於詞典的問題，使中文斷詞器的架構更加完善，並將中文斷詞器製成一便於使用的視窗工具。整個最核心的構詞單元是採用中研院中文分詞規範所提供的“詞綴，接頭／接尾詞”列表，經由統計整理，並以詞類作為規則法構詞的依據。另外，從詞綴著手，並加上介詞與連結詞，觀察三者對於口語語音停頓類型的特殊現象，從中挑選特別字詞，提供除詞類、詞長等參數外，作為未來從文字預估停頓的研究上另一新參數。本文也針對中文語音合成系統當中，破音字的問題作前處理，提供正確的語料可供未來研究使用。為了評量詞綴構詞單元之效能，我們以《中研院平衡語料庫 3.0 版》作為測試語料，測試結果顯示構詞正確率達八成左右。最後我們分析各個構詞規則的錯誤率，探討構詞錯誤的更正方法。

A Further Study on Text Analysis for Mandarin TTS

Student : Ming-Zong Fu Advisor : Dr. Yih-Ru Wang

Department of Communication Engineering

National Chiao Tung University

Abstract

In this thesis, the further research about text on Mandarin Text-to-Speech(TTS) System. First, we hand on the multiphone characters and affix characters which proposed by Mandarin Promotion Committee, the Ministry of Education and Chinese Knowledge Information Processing group(CKIP), Academia Sinica. We design the new word combination module after the word identification to dispose of unknown word by using 74 rules. And, we observed some special words that can affect prosodic pause. This observation may improve on predicting break indices form Chinese text. At last, the Sinica Corpus published by CKIP is used to evaluate the performance of the new combination module. We achieve a precision rate of 0.829 in word combination. We also analyze word combination results to give advices in furure word

誌謝

首先，我要感謝指導教授王逸如老師兩年來諄諄的教導，不厭其煩的聆聽我的報告，糾正邏輯上的錯誤，帶領我一步一步地學會做事的方法。也謝謝陳信宏老師，在碩二的這一年間，陳老師不僅適時的給予如何做研究的建議外，對於所作的每一步其前因後果以及未來能延伸的問題都鉅細靡遺的講解，讓我深刻感受到老師們那清楚的思緒和思考問題的思考邏輯。讓我這兩年來獲益匪淺，也讓我對於讀書的態度有了不同以往的大轉變。

再來，我還要感謝我的室友們阿弟、小民、輝昇、大雕、大師、問賢、神龍，雖然認識你們短短兩年，但卻有種相見恨晚的感受。跟你們一起打球，一起聊天真的很快樂，也能讓我從實驗室回來而煩悶疲憊的心得以紓解，沒有你們，我一定會撐不下去的，如今要離開不能天天見到你們真有點捨不得。

實驗室的大家也是不可忘卻的，小鄧的牛頭不對馬嘴個性，小迷彩說話一針見血，跟獻文大大討論海賊王，很愛吃麥當勞的宏宇，超愛家的啟風，超愛紅襪的胤賢，很愛兇的友駿，很愛熱血誰呀的銘彥以及很愛打飛機的性獸、很會嗆聲的 Barking、會陪我跑步的智合、會陪我看 NBA 的阿德以及靜悄悄的希群，有了你們的幫忙，讓我在做事上更如虎添翼，尤其是性獸學長，沒有你就沒有現在的我啊！

最後不免俗的要感謝我的家人，爸爸、媽媽、哥哥、姐姐以及大嫂，你們一直以來的支持讓我沒有經濟上的問題，對我如此的好吃的用的住的玩的都會想到我，讓我每次回家都能一再地充電以備作研究的能量，我很高興你們以我為榮，我也盡力的做好我能做的。

經過這麼多年來的學習，在交大這兩年感觸特別深，因為讓我看見許多強者，讓我了解「沒看見好的，不知自己能多好。」。謝謝大家，有你們才有現在的我，未來的我會繼續努力的。

目錄

表目錄

圖目錄

第一章 緒論

1.1 研究動機

目前，交通大學語音處理實驗室所研究發展的“中文文字轉語音系統”(Text-to-Speech System 簡稱 TTS System)，已有相當不錯的合成品質。而此系統的最前級必然是將輸入文句加以解析並斷詞的斷詞器。江振宇[1]的中文斷詞器，採用中研院提出的六條斷詞規則[2]，並以詞典樹(Lexicon Tree)的資料結構儲存詞典，建立基本的斷詞單元。並將具備某種規則的詞，例如：定量複合詞、重疊詞等，參考中研院提出的構詞規則[3,4]設計構詞單元。隨後還有給予斷詞詞類的詞類標記單元，及將某些阿拉伯數字、詞或符號由寫法傳為語音讀法的文字正規化單元，整體架構如圖 1-1 所示

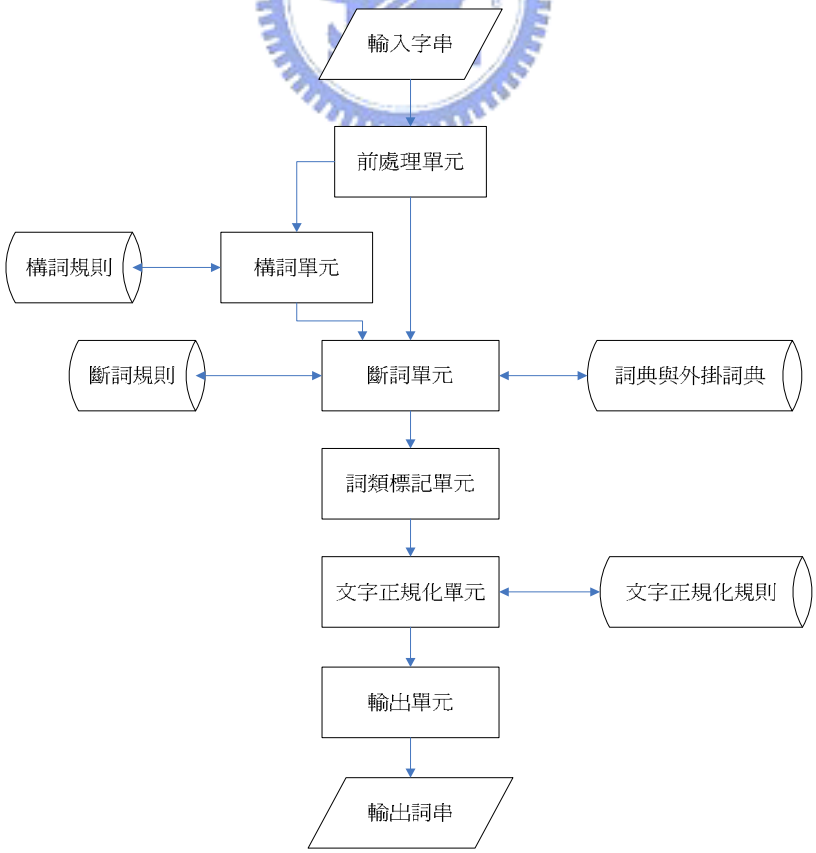

圖 1-1：中文斷詞器架構

中文斷詞器輸出的為語言參數，這些語言參數包括詞、詞長、詞類以及基本音節等資訊，其後將這些語言參數送往韻律訊息產生器產生所需的韻律參數，如基頻軌跡(pitch contour)、音長(duration)、停頓(pause)和音量(energy)，最後語音合成器利用 PSOLA 的技術，依據韻律參數將基本音節調合成語音。

為了合成出來的語音聽起來自然，符合人說話的方式，這些韻律參數簡而言之就像是人類說話的“抑揚頓挫”。而最為影響一個人了解另一個人說話的內容的參數莫過於停頓了，一個人說話，在不該停頓的地方停頓或是該停頓的地方不停頓都會讓人覺得不自然，進而無法了解一個人說話的真正涵義。但現在斷詞器輸出的詞串是詞典中的詞或是構詞規則單元產生的詞，詞典大量的詞彙會造成搶詞而造成不該出現的停頓；構詞規則可能構出太長的詞而造成該停頓的點卻沒出現停頓，這兩種問題皆會造成人聽覺上的不舒服。所以我們希望能從文字方面得到更多關於停頓方面的資訊，如此一來，能夠給予斷詞器更多韻律方面的資訊減少斷出過多的短詞或是搶詞與構詞規則構出過長的詞，之後更進一步的冀望這類停頓的資訊能幫助未來的斷詞器斷出韻律詞邊界(Prosodic Word Boundary)。

1.2 研究方向

本文研究的重點是希望能從斷完的短詞去預估每個短詞之間的停頓以及長詞應斷為若干短語的停頓，將停頓的資訊送往後級的單元使用。基於這個方向，針對短語與短語之間停頓的現象和長詞斷為若干短語的現象加以分析。首先我們先對中研院提出的“詞綴、接頭 / 接尾詞參考表”[5]著手，此類的短詞因可與其他短詞相接組合成另一新的短詞，所以詞典無法完全收錄，而導致斷詞時易被斷為兩個短詞，而介詞與連接詞兩者，因為其詞類(Part of Speech)功用特別，其停頓點應也有其特別之處。針對這兩方面，加以詞類等資訊，在未來的研究中套用數學模型，設計一自動預估停頓標記的方法，給予中文斷詞器更多韻律方面

的資訊。除此之外，我們進一步地將詞綴整理、歸納出詞綴構詞規則，並將規則寫入中文斷詞器當中。為了之後對於破音字的研究，於本文也事先對破音字作了前處理的工作以供未來研究之使用。我們以《中研院平衡語料庫 3.0 版》部分語料經人工斷詞以及江振宇自動標記停頓標記的語料做為統計來源，給予詞綴、連接詞及介詞三者，停頓分佈的詳細分析。提出特別字詞，以供未來從文字預估停頓的研究上另一新的參數。最後，我們採用定量的測試方式來檢測詞綴構詞規則的準確性。

1.3 章節概要

第一章　**緒論**：介紹本論文的研究動機與方向。

第二章　**中文斷詞器之概述**：簡述目前中文斷詞器的整體架構。

第三章　**破音字前處理及詞綴構詞單元之設計**：說明對於未來必然要面對之破音字的問題，敘述如何處理以及選定破音字等前處理工作，以及建立詞綴構詞規則並加入中文斷詞器中的演算法。

第四章　**詞綴構詞單元之效能分析**：以《中研院平衡語料庫 3.0 版》進行加入詞綴構詞單元的中文斷詞器之定性與定量實驗分析。

第五章　**文字中的特別字詞與停頓標記關係之統計**：說明本論文針對“詞綴、接頭 / 接尾詞參考表”以及介詞與連接詞三者進行停頓分佈的統計結果並給予詳細的分析。

第六章　**結論與未來展望**。

第二章 中文斷詞器概述

2.1 分詞標準

交通大學所發展的中文斷詞器主要是以語音合成的角度來設計，因此希望斷出的詞能夠適於語音合成的單位，但是，中文對於詞的定義較模糊，所以我們的首要工作便是定義分詞單位，也就是訂定「分詞標準」。對於資訊處理而言，根據中研院：「詞為一個具有獨立意義，且扮演特定語法功能的字串應視為一個詞」，然而對於語音合成而言，分詞單位與資訊處理的標準不盡相同，例如「八位學生」，依資訊處理的標準會斷成「八位」、「學生」，「八位」含有學生的數量資訊，「學生」含有某個事物的名稱資訊，但對於語音合成的角度而言，「八位學生」應為一個分詞單位，所以可以以資訊處理標準斷詞的結果「八位」、「學生」在構詞成一個語音合成的分詞單位。

由以上的敘述得知，如果要達到以語音合成考量的斷詞結果，可以利用「資訊處理標準」的斷詞結果，再經由一些規則，將斷詞結果提升到適用於語音合成，所以這裡提出的中文斷詞器，將斷詞處理依據不同的「分詞標準」可分為不同的前後級處理，分級如下：

第一級：由「構詞單元」及「斷詞單元」構成，詞集合為「詞典」以及「構詞單元產生的詞」，由此級斷出的詞希望能達到「資訊處理」的標準，使得斷詞結果含有充分的語法與語意資訊，以供後級利用。

第二級：由「後置構詞單元」及「未知詞構詞單元」組成。由於第一級產生的斷詞結果含有充分的資訊，因此第二級可以利用這些資訊再輔以一些規則或統計資訊，將衍生詞或未知詞(如:人名、專有名詞等)斷出。

第三級：以「語音合成考量的構詞單元」。這一級的「分詞標準」是以語音合成為考量，利用前兩級所給予的資訊，將分詞單位屬於語音合成考量的詞構出。

2.2 中文斷詞器系統架構

目前完成的系統模組已達到部份「第二級分詞標準」，系統概述如下：

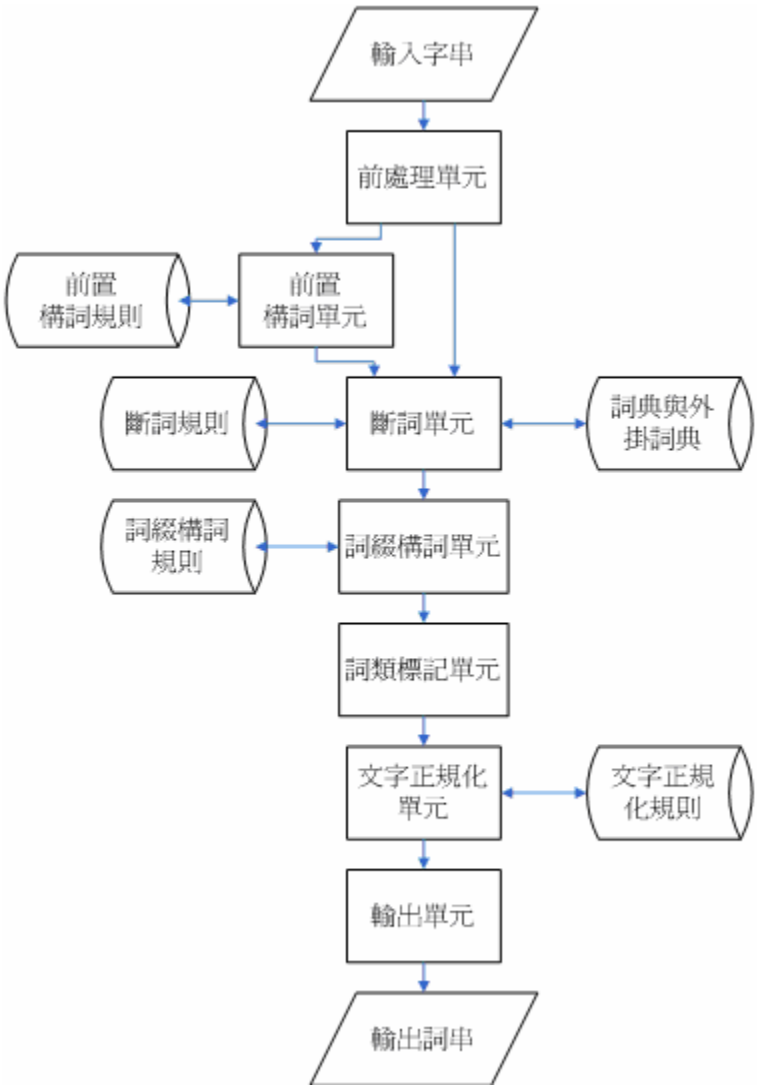

圖 2-1：中文斷詞器整體架構圖

(1)前處理單元：

由於是要使用於 TTS 的斷詞器，因此輸入的字串有可能含有 ASCII code 或是 Big-5 code，因此為了使系統處理的字串格式統一，在進入斷詞器以前，首先我們將所有的ASCII code轉為Big-5 code,例如字串「比去年同期減少了8.6%」會先轉為 Big-5 字串「比去年同期減少了８・６％」。

(2)前置構詞單元：

由於中文斷詞的方法是將輸入的文句與詞典做搜尋比對的工作，但是中文詞無窮無盡，要將所有的詞收錄至詞典當中是不可能的，而有些具有規則而無法詳列於詞典的詞，如定量複合詞以及重疊詞等，這些詞的組成是有規律的，可藉由輸入的文字串中，經由構詞規則將這些詞結合出來，這般構出的詞如同比對詞典的動作。由此模組構出的詞，會留下其構詞的結構，以便後級的模組使用。在此級分別分為(1)定量複合詞構詞單元(包含有「定量複合詞」、「數詞定詞」、「數量定詞」、「時間詞」、「地方詞」)、(2)重疊詞構詞單元，示意圖如下：

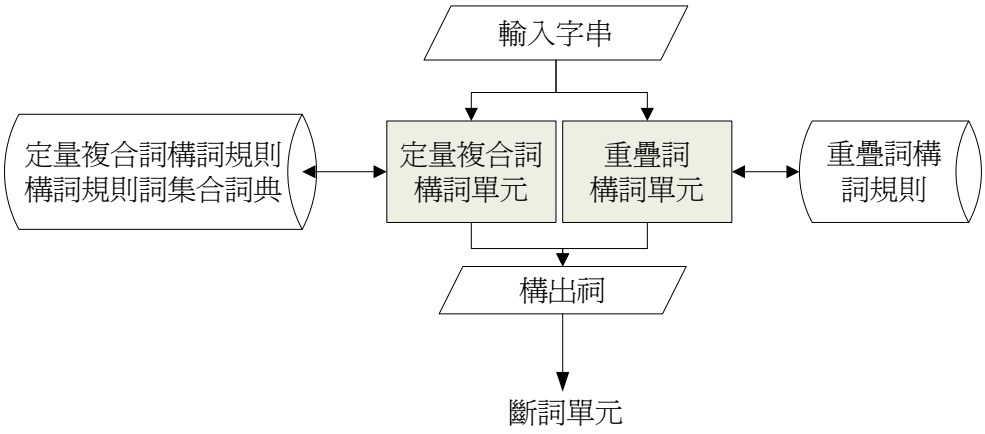

圖 2-2：前置構詞單元示意圖

(3)詞典及外掛詞典：

由於中文斷詞的方式是將輸入的文字和詞典進行搜尋比對，詞典的好壞會直接影響到「斷詞單元」的斷詞結果。我們先對詞典作了增減修正動作，將過於合詞的詞自詞典中刪除，目前中文斷詞器詞典的詞數統計如下表 2-1

表 2-1：詞典詞數統計表

	詞數
一字詞	13,110
二字詞	64,886
三字詞	26,042
四字詞	16,066
五字詞	999
六字詞	155
七字詞	65
八字詞	9
總計	121,332

除此之外，如有遇到特定使用環境的時候，會遭遇到輸入文句有許多的特殊詞，我們可以藉著新增這些詞於外掛詞典，解決特殊詞的斷詞。

(4)斷詞單元：

此單元是中文斷詞器的核心部份，目的是將輸入的字串做適當的斷詞。此單元分為兩大步驟(1)建立候選詞組(2)挑選候選詞組。建立候選詞組是將輸入的文字串利用詞典、外掛詞典以及前置構詞單元構出的詞，以 Matching Algorithm 找出所有可能的詞串組合。然而，這些所有可能的詞串組合只有一組是適當的斷詞結果，所以必須經由一些規則或是統計的方法，挑選候選詞組，選定最適合的詞串。在此我們沿用中研院提出的六條斷詞規則，這些規則分別是(1)長詞優先(2)標準差小的優先(3)附著語素最少者優先(4)候選詞組中定量複合詞字數和最

少者優先(5)一字詞詞頻最高者優先(6)總詞頻最高者優先，最後再加上第七條規則，以防經過六條斷詞規則仍不能選出詞組，規則為將任意選擇候選詞組中的任一個詞組的第一個詞作為斷詞結果輸出。「斷詞單元」是最前級的分詞單元，之後的「詞綴構詞單元」、「詞類標記單元」和「文字正規化單元」皆是利用此單元的斷詞結果再做處理。

(5)詞綴構詞單元：

經由斷詞單元斷出的詞串當中，仍然含有前置構詞單元尚未構出的衍生詞，這些衍生詞也是無法窮舉於詞典，所以經過斷詞單元，這類衍生詞會被斷為短詞。但是我們所需要的是以語音合成為考量的斷詞器，因此這類衍生詞不應該被斷開，我們希望能藉由「後置構詞單元」將這類衍生詞構出，如詞綴、人名等未知詞。因此我們於斷詞單元後設計「詞綴構詞單元」，藉由「斷詞單元」輸出的資訊，來將詞綴衍生詞構出。而至於人名等未知詞留待未來在做處理。詳細的「詞綴構詞單元」設計於第三章有詳細說明。

(6)詞類標記單元：

有了斷完詞的詞串，而中文的詞有時具有多種詞類(Part of speech, POS)，因此還需對這些詞串找到其最適當的對應詞類，目前採用的「詞類標記單元」，為沿用江振宇用《中研院平衡語料庫 3.0 版》所建立的「詞類雙連文模型」，應用 Viterbi-search 給予斷出的詞串最佳的詞類標記串。

(7)文字正規化單元：

以語音合成為考量的中文斷詞器，在輸入的文句之中，有時會有阿拉伯數字、詞或符號必須由書寫法轉為語音讀法，此過程稱為文字正規化，此模組利用

斷詞單元及前後構詞單元留下的詞結構進行文字正規化，舉例來說「９０・５％」應該將寫法轉為語音讀法「百分之九十點五」。目前的文字正規化單元，針對了以下幾種情形做正規化動作：

(a) 英文字母部份：TEL、FAX、AM、PM、30cm、100kg 等

(b) 符號部份；90.5%、2007/12/25、19.6 等

(c) 數字加單位部份：100 公斤、1 月 1 日、2007 年等

針對以上情形定出數字、符號的發音方式及發音順序。

(8)輸出單元：

此單元為輸出韻律產生器所需要的語言參數：詞、詞類和音碼。

第三章　破音字前處理及詞綴構詞規則之設計

目前的中文斷詞器多是以利用構詞規則將含有某種特性的詞構出以及查詢大量詞彙構成的詞典這兩部份將一句文句斷為若干個分詞。但是在這當中，給予每個字基本音節的資訊已被詞典以及規則表兩者限制住。然而中文有許許多多的字是具有一字多音的特性，因此會導致斷詞器斷出的詞，其中字的基本音節不符合文句所要闡述的意義，這樣合成出來的語音聽起來會使人不舒服。所以正視破音字的問題是必然的。

交通大學語音處理實驗室現階段的中文斷詞器[1]已將定量複合詞和重疊詞兩類具有明顯規則的詞，設計一前置構詞單元，輔助這類無法完全收錄至詞典的詞，避免這類詞被切分為一個一個的字。這部份的工作可在「分詞標準」三級中的第一級完成。但是還有些可以依規則構成的詞，例如：副總統、副會長、副班長、副廠長；大人們、家長們、小孩們、老師們等這類由中研院提出的“詞綴、接頭 / 接尾詞參考表”中由某個雙音節的詞緊接一個單音節的詞綴所衍生出的詞，這類型的詞可由後置構詞單元來構成，可於「分詞標準」的第二級完成。底下分兩小節來個別詳細說明破音字前處理以及後置構詞單元之詞綴構詞單元。

3.1 破音字前處理之說明

國語語音部份，常會有一字多音的情形，這類文字我們統稱為“破音字”。破音字在國語語音合成系統中，易造成混淆使得系統合成出的發音與口語發音不符，造成人聽覺的怪異感，因此針對破音字的問題做研究，對國語語音合成系統是有其必要性的。

首先第一步需了解國語中有哪些字是破音字，我們參照教育部國語推行委員

會所制定的國語辭典當中，所審定的《國語一字多音審定表》[6]建立初步的破音字表，此表總收錄的破音字共有 4253 個。但這 4253 個破音字當中，有許多的古字，或是有許多的破音字只在古文當中才被使用到，這對於現今的口語習慣當中甚少使用，因此我們並不收錄這些文字於我們所需處理的破音字當中，我們僅只針對現今口語及書寫經常會用到的文字做處理。

依此方向作“破音字表”的收錄，我們從 4253 個字刪減至 896 個字，但仍然過多，因此我們再針對某些破音字只有在特定的詞當中才會發生，我們將此類破音字的詞，收錄至我們國語詞典當中，以利事後斷詞時，能使斷詞的標音正確無誤且正規。例如：《單》通常音【ㄉㄢ】，但當作地名或姓氏時常音【ㄕㄢˋ】。而表匈奴首領一詞《單于》音【ㄔㄢˊ　ㄩˊ】。因此我們再針對此類情形，將大部分僅止某些詞的破音字，收錄至詞典並對我們現有的國語詞典做修正。至於姓氏方面的單字破音字如：《曾》通常音【ㄘㄥˊ】但當做姓氏的時候音【ㄗㄥ】，此類姓氏暫時未做處理，留待未來斷詞器收錄姓氏的外掛詞典來處理。依此方向下來，我們更進一步從 896 個破音字當中節錄 126 個，修正紀錄如下：

表 3-1-1：依破音字修正詞典紀錄節錄

編號	國字	注音	註解
1	上	ㄕㄤˇ	字典一字詞收錄為（ㄕㄤˋ） “上聲”一詞(詞典已有) 將”平上去入”加入四字詞中
2	乘	ㄕㄥˋ	字典一字詞收錄（ㄔㄥˊ） 其餘標記為（ㄕㄥˋ)的詞皆已納入辭典，如：萬乘之國、大乘、小乘等
3	仇	ㄑㄧㄡˊ	辭典一字詞收錄為（ㄔㄡˊ）、姓氏
4	任	ㄖㄣˋ	姓氏，納入詞典中 合成：ㄖㄣˋ 辨識：ㄖㄣˊ and ㄖㄣˋ 皆標註
5	佔	ㄍㄨㄚˋ	二字詞詞典已有”佔(ㄍㄨㄚˋ)衣”一詞 三字詞詞典新增”佔(ㄍㄨㄚˋ)衣舖”一詞

表 3-1-2：常用破音字表節錄

編號	國字	讀音	說明
1	中	ㄓㄨㄥ	中央、中國、中學、中立
		ㄓㄨㄥˋ	中的、中毒、中意
2	乾	ㄍㄧㄢˊ	乾卦、乾坤、乾乾
		ㄍㄢ	餅乾、乾杯、乾淨、乾脆
3	了	ㄌㄧㄠˇ	了結、了解、了不起、受不了
		ㄌㄜ•	「做完了！」、「這就難怪了！」
4	供	ㄍㄨㄥ	ㄖ供、供給
		ㄍㄨㄥˋ	供品、供奉
5	便	ㄅㄧㄢˋ	方便、便利
		ㄆㄧㄢˊ	便宜、便辟、大腹便便
6	倒	ㄉㄠˇ	倒閉、跌倒、倒塌、顛倒
		ㄉㄠˋ	倒影、倒轉
7	假	ㄐㄧㄚˇ	假借、假裝
		ㄐㄧㄚˋ	假期、放假
8	傍	ㄅㄤˋ	依山傍水、依傍
		ㄅㄤ	傍晚、傍午
9	傳	ㄔㄨㄢˊ	傳單、傳神、傳染
		ㄓㄨㄢˋ	左傳、傳記
…	…	…	…
58	暈	ㄩㄣ	頭暈眼花、暈車、暈倒
		ㄩㄣˋ	月暈、燈暈、酒暈、血暈
…	…	…	…
83	種	ㄓㄨㄥˇ	種子、種類
		ㄓㄨㄥˋ	種田、接種
…	…	…	…

表 3-1-1 是說明我們依照上述刪減某特定的破音字，對詞典做修正的紀錄，表 3-1-2 是常用破音字表的節錄，此兩表完整的內容請詳閱附錄 1、2。

完成破音字表，對於未來研究所需使用的 Treebank 文字資料庫也要做處理，此語料庫並未經過斷詞，所以我們利用江振宇的中文斷詞器[1]做斷詞之後再加以修正，此語料庫約十二萬字，修正記錄在此不贅述，詳閱附錄 3。

3.2 詞綴構詞單元之設計

中研院提出的中文資訊處理分詞規範所提及的層次劃分當中，依電腦自動化處理分詞的難易程度及實際使用情況，分信、達、雅三級，(1)信級；凡是收錄在標準詞典的詞一律斷開，(2)達級：能以構詞律組合出來的詞在達級合併，(3)雅級：無法完全收錄至詞典中的詞在雅級合併。但是在語音合成的角度而言，對於分詞的標準不盡與資訊處理的標準相同，這於第二章中文斷詞器概述的起頭就有提到，於是我們也採用了我們以語音合成角度所提出的分詞標準。在江振宇的中文斷詞器[1]中加入了「定量複合詞」、「數量定詞」、「數詞定詞」、「時間詞」、「地方辭」、「重疊詞」等希望能夠經由構詞單元，將這些詞由輸入文句詞組中合併出來，相同於查詞典的地位，因此構詞單元構出的詞與詞典中的詞皆為建立候選詞組的詞集合。

但經斷詞單元斷出的詞串中，仍然有些字能夠彼此結合成長詞，因此我們希望能經由後置構詞單元將能結合的詞結合成長詞，例如：【大人們】會被斷為【大人】和【們】、【老前輩】會被斷為【老】以及【前輩】。這類如【們】和【老】與【大人】和【前輩】合併為另一個詞的字，我們稱為“詞綴、接頭/接尾詞”，但在此論文當中皆統稱為“詞綴”。3.2.1 將說明詞綴構詞規則的建立，3.2.2 將說明「詞綴構詞單元」在程式語言中的構詞演算法。

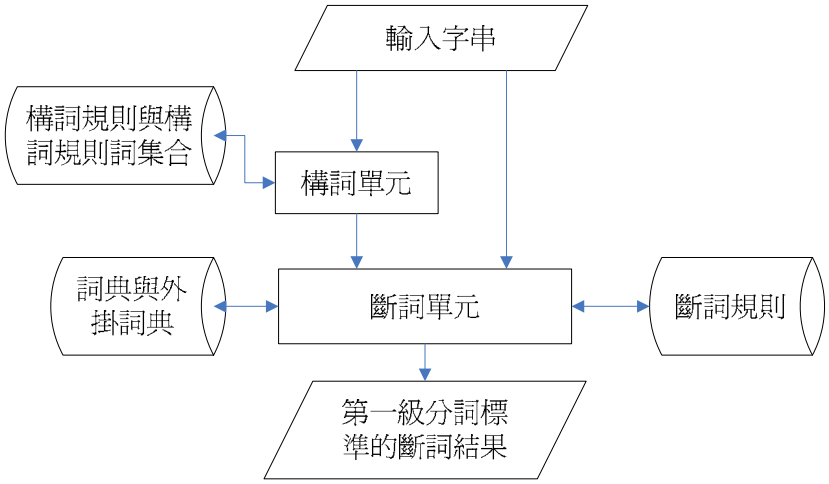

圖 3-2-1：構詞單元在斷詞器中的位置

3.2.1 詞綴構詞規則的建立

中研院提出的中文資訊處理分詞規範當中有詳細的附錄提供“詞綴、接頭／接尾詞參考表”，除了這項資訊之外，我們再輔以清華大學張俊盛教授整理出的前、後詞綴表務求完整。但這許許多多的詞綴當中，有某些詞綴也符合定量複合詞的特性，因此在建構詞綴表之前，我們對江振宇的定量複合詞構詞單元瀏覽一遍，將這些詞綴加進定量複合詞構詞規則當中，加強定量複合詞構詞規則的完整性。而清華大學張俊盛教授整理出的詞綴表，其中含有許多數詞的詞綴，類似這種情形的詞綴因已加入定量複合詞構詞規則當中，因此並不收錄至詞綴表。

經過上述初步的處理，我們刪減了一些詞綴，但詞綴的數量依然稍微過多。因此我們針對五萬字的 Treebank 文字資料庫做統計，統計詞綴表當中的詞綴是否有於語料庫當中出現，未出現的詞綴視為鮮少出現的詞綴，暫時從詞綴表當中剔除。最後我們收錄了前詞綴 121 個、後詞綴 195 個。

由[1、3]我們得知樹狀結構的構詞規則可以用 Regular Expression 表示，而且以 Regular Expression 表示的構詞規則使人容易看懂規則，但為了要將規則應用在程式語言中，必須把 Regular Expression 轉為以「規則標記」表示的

Chomsky Normal Form 表示方法，應用在程式語言中的好處是，如果對於某一項規則要修改或增加新規則，只需針對這些「規則標記」進行簡單的修改即可。

我們以表 3-2-1 來說明，「規則標記」301～330 所對應的，為前詞綴的集合，401～443 和 458 為後詞綴的集合。這些詞綴就是詞綴構詞規則的基本元素，而同一規則標記的詞綴是具有相同詞類或是句法一致的特性，詳細的「規則標記」請詳閱附錄 4。這些詞綴「規則標記」再搭配前後文的詞類，就可建構出「詞綴構詞規則」，見下表 3-2-2，詳細的「詞綴構詞規則」請參閱附錄 5。

表 3-2-1：詞綴規則標記

規則標記	Fixword	詞類
301	非	1,11,36
302	多	11,20,37
303	老,大,小,高,低,粗,細,淡,淺,深,古 冷,長,易,乾,軟,硬,短,新,輕,薄,舊	37
…	…	…
307	曾,又,剛,仍,已,正,早 即,尚,便,常,現,就,遂	11
…	…	…
312	太,更,很,略,最,稍,微,極,較,頗,遠	7,11
…	…	…
318	好	7,37
…	…	…
330	變	37
401	們	12
…	…	…
418	國,省,村,鄉,鎮,縣,市,課,社	12,14
…	…	…
442	上,中,下,內,時,前,後,來,底,起,裡,頭	9,15,23
…	…	…

表 3-2-2：詞綴構詞規則表節錄

規則標記	類別(國)	regular expression	範例
301	非	非(A) + {Na}	非博士,非親骨肉
…	…	…	…
303	老,大,小,高,低,…	{老,大,小,…}(VH) +{Na,DE,Di}	老賊,老士官店
…	…	…	…
330	變	變(VH) + {VH}	變酸,變黑
401	們	{Nab,Naea} + 們{Neqb,Naea}	大人們,情侶們
…	…	…	…
437	賽,會,式,制	{Na,VA,…} + {賽,會,…}{Na}	邀請賽,季級賽
…	…	…	…

至於新構出的詞其詞類，大部分的後詞綴多為名詞所以給予其詞類為「Na 普通名詞」，剩餘的如「Nc 地方詞」給予新詞類仍為「Nc 地方詞」,「Nb 專有名詞」給予「Nb 專有名詞」的詞類，其餘的新詞，暫定給予原本相同的詞類。前詞綴的部份，如詞綴本身是「Na 普通名詞」,新詞也保留同樣的詞類。而前詞綴含有許多的副詞以及狀態不及物動詞，這兩類的詞綴，我們給予的詞類暫設定為前詞綴的後接詞其詞類，例如：【老】的詞類為「VH 狀態不及物動詞」，後接詞【士官】為「Na 普通名詞」，兩個詞構成一新詞【老士官】我們給予「Na 普通名詞」的詞類；【很】的詞類為「Dfa 前程度副詞」，後接詞【高興】為「VH 狀態不及物動詞」,兩個詞構成一新詞【很高興】我們給予「VH 狀態不及物動詞」的詞類。

3.2.2 詞綴構詞單元之演算法

由圖 3-2-1 可以看出，江振宇的斷詞器[1]已經達到我們分詞標準的第一級，之後我們加入後置構詞單元之詞綴構詞單元。對於已經被斷詞的詞串，我們應將那些詞綴找出，並利用詞綴構詞規則將短詞合併為長詞。整個詞綴構詞單元

的工作流程，表示在下圖 3-2-2：

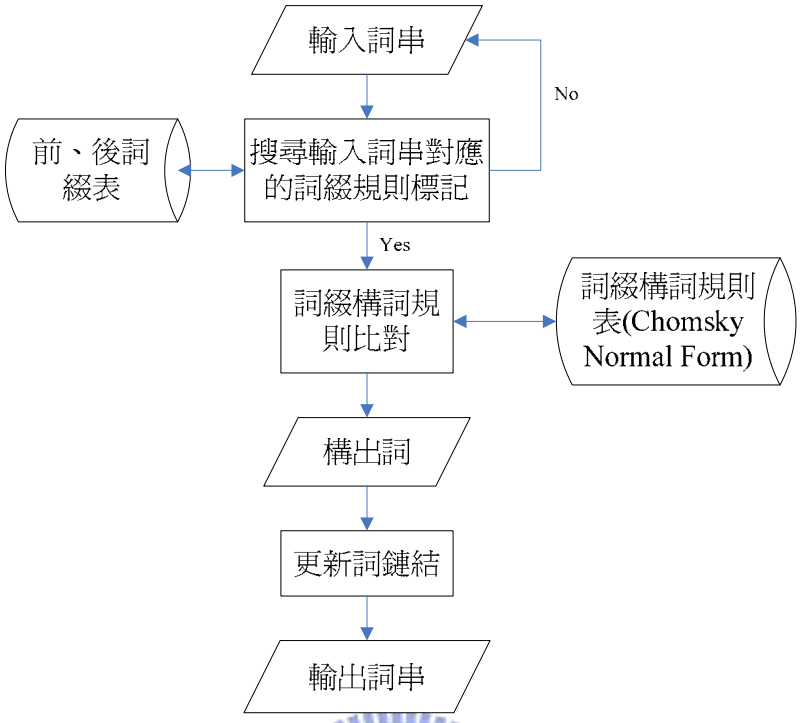

圖 3-2-2：詞綴構詞單元流程圖

以下我們皆以文句「老先生與大人們在討論事情」作為例子說明，首先，此文句經過前端的斷詞器會被斷為如下圖的詞串

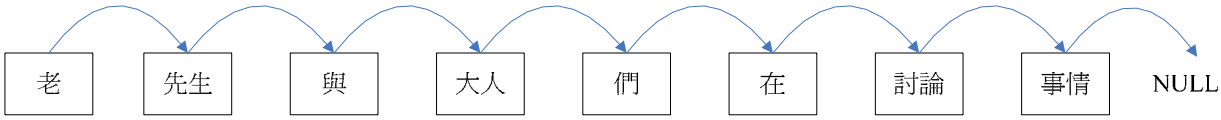

圖 3-2-3：Linked-list Output Words

這些詞串會送往後級的詞綴構詞單元，經過以下幾個詞綴處理步驟：

(1)搜尋輸入詞串對應的詞綴規則標記：

這一步驟的目的就是要將詞串當中，符合詞綴的詞找出來，並給予詞串對應的詞綴規則標記。

類似於查詞典的方式，我們對已建立的前、後詞綴列表進行搜尋。因為詞綴為單音節的一字詞，所以我們於搜尋之前建立一簡單的判斷機制，就是判定輸入的詞串其詞長是否為一字詞，如並非一字詞，則進行下一個詞的比對工作。如果符合一字詞的條件則進行比對前、後詞綴表的工作。如此一來，詞串當中符合詞綴的一字詞將會被搜尋出來並給予詞綴規則標記，由例句說明可得到下表 3-2-3

表 3-2-3：「老先生與大人們在討論事情」之詞綴規則標記集合表

詞串	規則標記	詞類	
老	303	37(VH)	
們	401	12(Na)	21(Neqb)

找到所有的詞綴以及給予規則標記之後，便可以進行比對詞綴構詞規則的動作。

(2)詞綴構詞規則比對：

此步驟的目的，是將已找出符合的詞綴文字，經由與詞綴構詞規則的比對，進而合併成為新的衍生詞。而構詞規則是以樹狀結構表示，一個樹代表一個構詞規則，且規則標記為 301～330 的屬於前詞綴，401～443 和 458 的屬於後詞綴，因此我們只需針對規則標記尋找構詞規則，前詞綴則依據構詞規則比對下一個詞的詞類；後詞綴則依據規則比對上一個詞的詞類。如符合規則，則將兩個詞合併為一新詞輸出。詞綴構詞規則比對步驟如下：

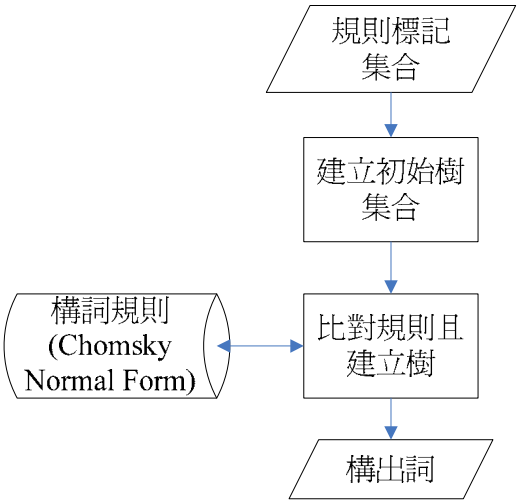

圖 3-2-4：詞綴構詞規則比對流程圖

(2.a)建立初始樹集合：

由於是承襲著斷詞單元的結果，每一個詞皆視為一棵樹，而且詞綴構詞規則是參考定量複合詞構詞規則，所以我們必須將輸入的規則標記集合，轉為以樹狀結構表示的資料結構，如下圖 3-2-5：

詞串	規則標記	詞類	
老	303	37(VH)	
們	401	12(Na)	21(Neqb)

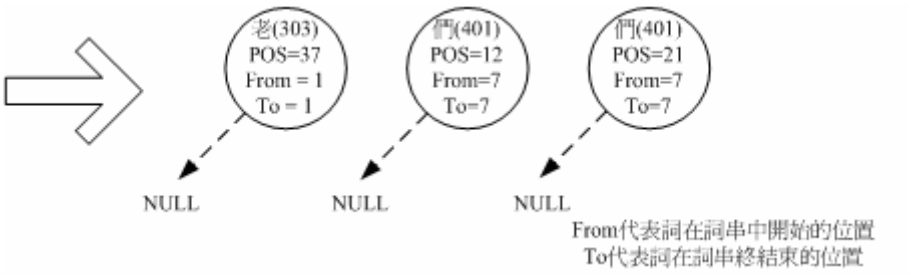

圖 3-2-5：建立初始樹集合

(2.b)比對規則且建立樹：

此步驟是詞綴構詞規則比對的核心，目的是將所有可能的樹狀結構組合找出來(可能的構詞組合)，若樹集合中的前詞綴與其在詞串中的下一棵樹或是後詞綴與其在詞串中的前一顆樹可以某構詞規則結合，我們便建立一棵新的樹(新的詞)，並給予新的樹「規則標記」(代表符合哪項規則)及詞類(一條規則對應到一個特定的詞類)。

如果符合詞綴的特性而存在於樹集合當中，但卻不存在符合的規則，則新建之樹必為一個空集合。如果此情形發生，我們則將新建的樹刪除，避免浪費過多不必要的記憶體空間。整個比對流程如下圖 3-2-6

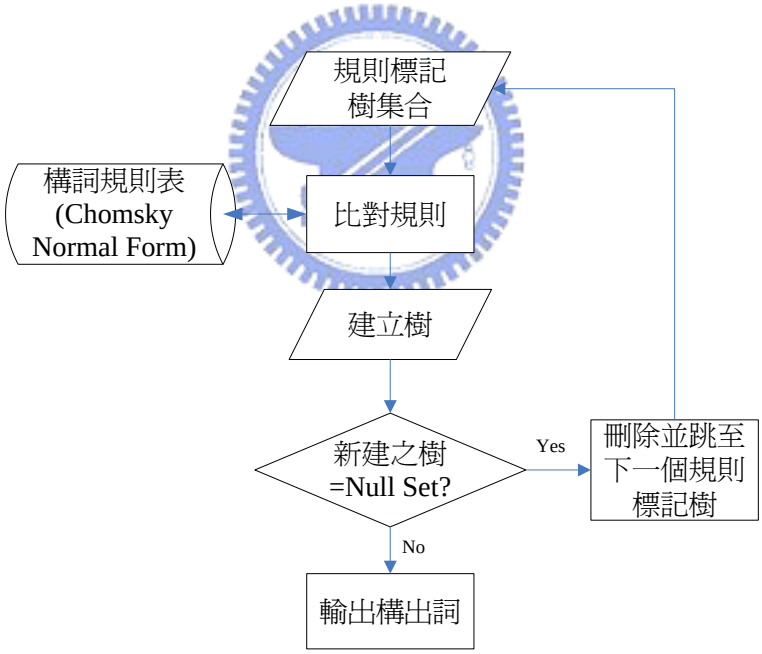

圖 3-2-6：詞綴構詞規則比對流程圖

為了增加規則比對的速度，我們將詞綴構詞規則依照詞典樹的作法，將規則存成樹狀結構，整個規則樹為一個 general tree，同一層的節點依照規則標記的數字大小排序好，是一個記憶體動態改變大小的陣列，往後如有新增規則，便會將新規則插入這樹狀結構當中。如下圖 3-2-7

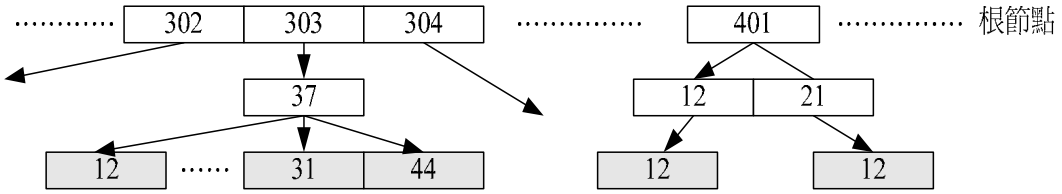

圖 3-2-7：詞綴構詞規則樹

深色底的代表規則的終點，代表從根節點走到終點經過的節點為一個構詞規則，在終點記錄這個規則所對應的規則資訊(規則標記、詞類)，若無法走到終點則表示沒有對應的規則。以「老(303)　先生」為例，由根節點 303 出發，「老(VH)」符合下一個節點詞類代碼 37，「先生(Na)」符合終端點詞類代碼 12，如此便符合一項規則，建立一個新的樹「老先生」並給予規則標記 303 及詞類「普通名詞 Na(12)」。

(3)更新詞鏈結：

經由上面兩步驟之後，所有能夠更進一步結合成長詞的皆已構出。但為了能夠使後級的單元使用，我們必須更新舊的詞串鏈結。此步驟不僅更新了舊的詞串鏈結，也將舊詞的樹狀結構保留，經由圖 3-2-8 便可一目了然。

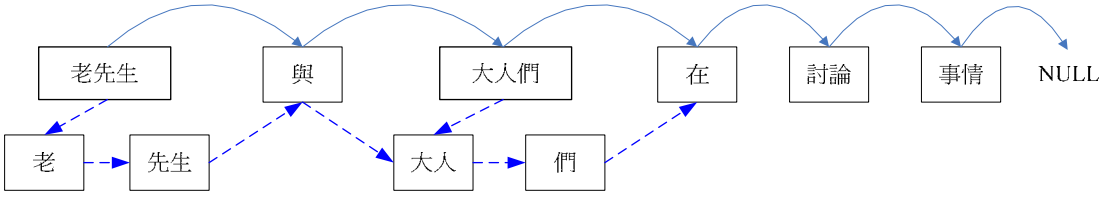

圖 3-2-8：更新詞串鏈結圖

由上圖，上層為新的詞串鏈結，下層為舊的詞串鏈結。

(4)輸出構出詞：

這一步是將更新詞串鏈結完畢之後的詞串鏈結輸出，最後輸出的詞串將會是「老先生　與　大人們　在　討論　事情」。

經過以上四個步驟，詞綴構詞單元便算完成，而所有的規則標記與詞類資訊皆完整保留，且規則標記與之前的定量複合詞並無重複，因此輸出的詞串送往後級的文字正規化並不會造成混淆。

3.3 漢語文句斷詞器

我們將完整的中文斷詞器製作成一個獨立的工具，如此一來，能在未來幫助中文斷詞上的研究，也能將斷詞結果輸出給予其他單位使用。首先先看下圖

圖 3-3-1：漢語文句斷詞器介面

以下我們用代號 1～7 來說明各個功能。

1. 輸入文句：

這是一個文字編輯方塊，可以自行輸入文字，也能藉由按鈕 4 開啟已存在的文字檔案，開啟的文字檔案內容會顯示在「輸入文句」此文字編輯方塊。

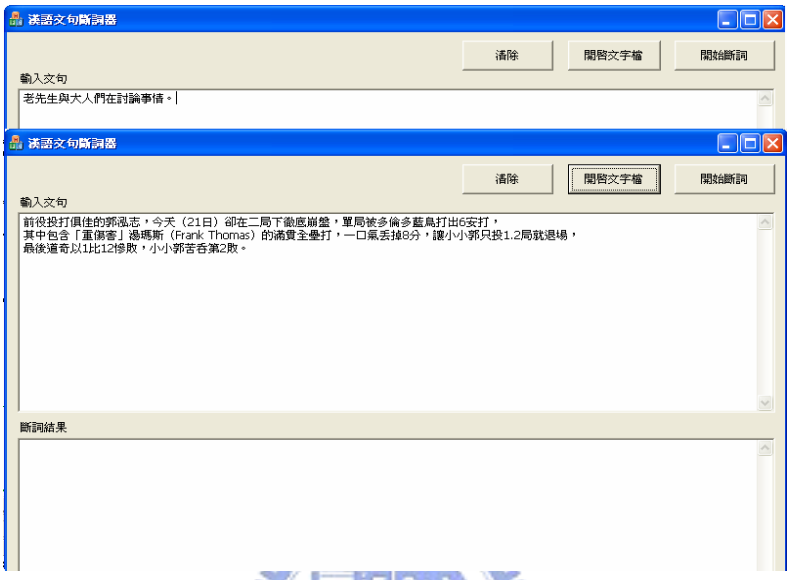

圖 3-3-2：自行輸入與開啟文字檔

2. 斷詞結果：

此為一文字方塊，無法自行輸入文字，也無法編輯，此方塊所顯示的，為啟動按鈕 5 之後，將「輸入文句」的文字內容斷詞輸出至此文字方塊

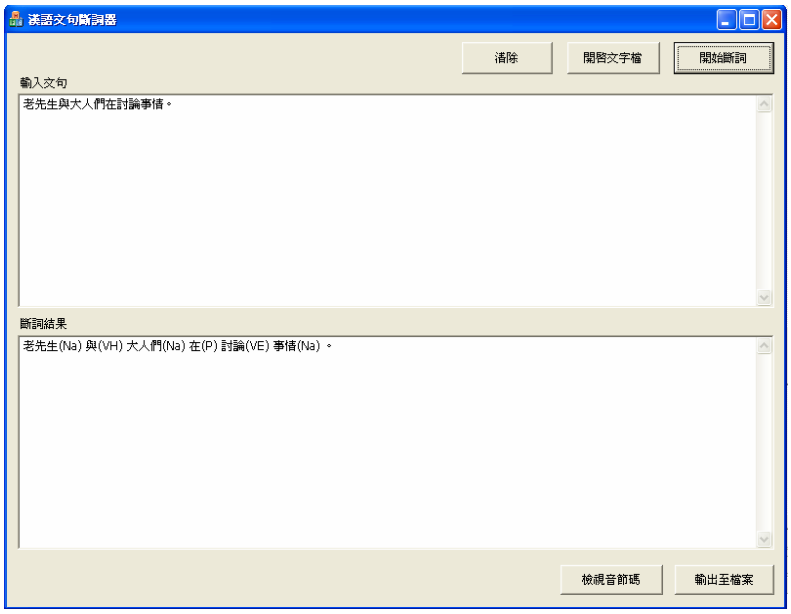

圖 3-3-3：顯示斷詞結果

3. 清除：

此按鈕的功能為將「輸入文句」及「斷詞結果」兩文字方塊之內容清除。

4. 開啟文字檔：

此按鈕的功能為將已存在的文字檔案內容輸入至「輸入文句」文字編輯方塊。

5. 斷詞：

此按鈕為整體中文斷詞器的核心，按此按鈕能將「輸入文句」文字編輯方塊中的內容，進入我們已設計的中文斷詞器，並將斷詞結果輸出至「斷詞結果」文字方塊內。

6. 輸出至檔案：

此按鈕為將「輸入文句」文字編輯方塊內的內容，其斷詞結果另存至一新的檔案中，方便未來研究上的使用。儲存格式如下：

表 3-3-1：斷詞結果輸入至檔案格式

字	411 音節碼	詞長及詞內位置	詞類
老	3092	301	12
先	1261	302	12
生	1170	303	12
與	3216	101	37
大	4018	301	12
人	2137	302	12
們	5147	303	12
在	4051	101	26
討	3090	201	34
論	4365	202	34
事	4003	201	12
情	2286	202	12
。	6003	101	50

7. 檢視音節碼：

此按鈕為將「輸入文句」文字編輯方塊內的內容，經過斷詞器得到各文字的音碼，為求便於觀察，將輸出格式由 411 個音節碼轉換為注音形式。

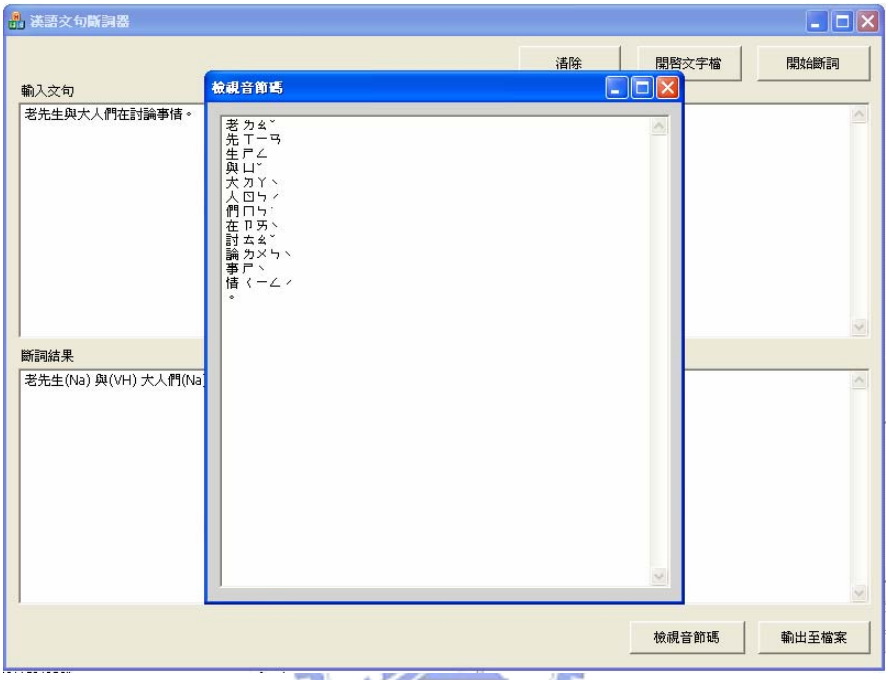

圖 3-3-4：檢視音節碼(注音形式)

第四章 詞綴構詞單元之效能分析

在本章中，我們將驗證之前所建立的詞綴構詞規則加入中文斷詞器[1]之後，構出的詞在整體斷詞結果中的效能

4.1 測試語料

我們採用《中研院平衡語料庫 3.0 版》部份語料作為測試語料，語料庫已經過正確之斷詞與詞類標記，以下為此測試語料的統計資訊：

表 4-1：部份《中研院平衡語料庫 3.0 版》語料庫統計

文章篇數	1,263
總詞數	880,861
總中文詞數	748,616
中文專有名詞數	12,812
外文詞數	2,718
標點符號數	129,527

表 4-2：《中研院平衡語料庫 3.0 版》語料庫統計

文章篇數	9,286
總詞數	5,841,942
總中文詞數	4,883,661
中文專有名詞數	94,121
外文詞數	27,502
標點符號數	930,779

4.2 斷詞結果之評量

因為設計的詞綴構詞單元是利用斷完詞的結果，因此在此先評量斷詞器的斷詞結果。由於專有名詞在整個語料庫裡佔了將近 2%，因此在評量斷詞器效能時，同時觀察含與不含專有名詞的斷詞結果，在這裡我們定義召回率(recall)及精確率(precision)作為斷詞結果的評量標準，定義如下：

N1 =（平衡語料庫的中文詞數）

N2 =（經斷詞器輸出之中文詞數）

N3 =（經斷詞器斷詞且與平衡語料庫一致的中文詞數）

斷詞召回率 = N3 / N1

斷詞精確率 = N3 / N2

斷詞結果如下表：

表 4-3：斷詞結果

	僅以詞典斷詞		加入前置構詞單元	
	含專名	不含專名	含專名	不含專名
N1	4,883,661	4,789,540	4,883,661	4,789,540
N2	4,773,152	4,679,443	4,434,692	4,341,015
N3	4,008,905	3,964,242	3,844,914	3,799,378
召回率	0.821	0.828	0.787	0.793
精確率	0.84	0.847	0.867	0.875

4.3 詞綴構詞單元效能之分析

在這節我們要討論詞綴構詞單元的效能，以下以定量的方式來分析。我們依照構詞規則編號，列出某構詞規則針對部分語料庫構出的詞數目，如下兩表

表 4-4：前詞綴構詞規則使用之分佈

規則標記	構出詞數 N1	正確構詞數 N2	rate = N2/N1	規則標記	構出詞數 N1	正確構詞數 N2	rate = N2/N1
301	48	43	0.896	316	81	73	0.901
302	131	117	0.891	317	10	10	1.000
303	2669	2387	0.894	318	262	237	0.905
304	518	473	0.913	319	0	0	0.000
305	171	144	0.842	320	1253	1094	0.873
306	53	37	0.698	321	177	168	0.949
307	4287	3905	0.911	322	127	118	0.929
308	4475	3894	0.870	323	8	6	0.750
309	86	86	1.000	324	2742	2229	0.813
310	2273	1983	0.872	325	67	54	0.806
311	122	106	0.869	326	76	71	0.934
312	2921	2384	0.816	327	239	206	0.862
313	2151	1648	0.766	328	66	61	0.924
314	303	243	0.802	329	867	750	0.865
315	521	482	0.925	330	32	26	0.813
Total	26736	22450	0.840				

表 4-5：後詞綴構詞規則使用分佈

規則標記	構出詞數 N1	正確構詞數 N2	rate = N2/N1	規則標記	構出詞數 N1	正確構詞數 N2	rate = N2/N1
401	162	162	1.000	412	19	16	0.842
402	51	51	1.000	413	889	814	0.916
403	356	299	0.840	414	784	744	0.949
404	0	0	0.000	415	178	144	0.809
405	570	431	0.756	416	79	65	0.823
406	71	57	0.803	417	17	11	0.647
407	417	342	0.820	418	66	50	0.758
408	58	51	0.879	419	19	19	1.000
409	109	90	0.826	420	1525	1378	0.904
410	24	19	0.792	421	53	48	0.906
411	25	21	0.840	422	32	26	0.813

規則標記	構出詞數 N1	正確構詞數 N2	rate = N2/N1	規則標記	構出詞數 N1	正確構詞數 N2	rate = N2/N1
423	13	8	0.615	434	691	429	0.621
424	72	63	0.875	435	57	47	0.825
425	10	10	1.000	436	21	19	0.905
426	39	30	0.769	437	90	86	0.956
427	12	12	1.000	438	0	0	0.000
428	725	604	0.833	439	59	51	0.864
429	0	0	0.000	440	0	0	0.000
430	87	87	1.000	441	9	6	0.667
431	111	98	0.883	442	9285	7235	0.779
432	50	42	0.840	443	11	8	0.727
433	72	61	0.847	458	743	654	0.880
Total	17661	14388	0.815				

上述兩表中的 N1 表示經由詞綴構詞單元構出的總詞數，N2 表示經由人工檢察構出詞是否合適，如會造成些微的不通順既屬於構詞錯誤。整體的正確率有 0.829，其中被引用最多的規則為規則標記 442，佔構詞單元構出詞的 20.5%，此類詞綴的詞類有“Ng”、“Ncd”等方位詞，構出的詞例如有：大會中、地底下、山上等，但其正確率只有 0.779，是因為此規則的後詞綴集合當中含有“時”、“來”等，直接與前詞相接，會造成語意上的不通順，所以我們將這些例子不算在正確的構詞數中。

第二多的是規則 307 與 308，這兩類前詞綴的詞集合皆屬於副詞的特性。依照中研院的分詞規範，這類副詞並不屬於詞綴。是依據清華大學張俊盛統計 OOV(Out Of Vocabulary)列表，才將此類副詞收錄至詞綴，判斷依據也是依照 OOV 列表所列之詞。所以雖然這兩類合出的正確率將近九成，但對於斷詞後的語意並不通順，建議將這兩類的詞從構詞規則當中刪去。

至於正確率較低的規則標記 306、323、417、441，是因為數量太少，導致刪除部分構詞會使正確率變化較大，未來使用完整語料庫統計會更正確，而規則標記 434，當中有個詞為“文”，在測試的語料庫當中有許多人名有使用“文”字，且因為詞綴構詞規則是利用詞類來做為判斷構詞的依據，不如定量複合詞有固定的詞集合，易造成構詞上許多的錯誤。

其餘正確率較高的部份，其詞綴的特性較明顯，且符合構詞的詞類也較單一，因此數量少、正確率高，除了專有名詞造成的錯誤和前級斷詞器搶詞的錯誤使得構詞錯誤外，並沒有過於嚴重的問題。

我們在此不統計平均詞長，因為經過詞綴構詞單元之後，平均詞長必然會比較長，但其中含有許多錯誤，所以平均詞長變長並沒有意義，因有可能造成過於合詞使得語意更模糊。

第五章 文字中的特別字詞與停頓標記關係之統計

以自然語言現象為考量的文字翻語音系統，除了採用恰當的語音合成技術外，最重要的問題即為如何從文字抽取出能夠體現人類語音特徵的韻律參數給予後級產生合適的韻律，如：基頻軌跡、音長、停頓、能量等。在自然語音中，人們不會將一句長句一口氣讀出，而是分成若干個短語；也不會將一句話皆一一斷開為獨立的短語，而是根據短語之間的結合緊密程度插入不同長度的停頓間隔。這不僅是人類生理上的限制，也含有人類說話韻律節奏上的需要以及語意上的涵義，以利於語者和聽者間的溝通。

但是，從文字自動產生韻律參數到現今為止仍然是一個相當困難的問題。因為，韻律其一層包覆一層的架構，從文字上並不是那麼容易就能夠辨別出來，而且很多研究[7、8、9、10]也指出，韻律的組成跟語法上並不是那麼完全的一致，因此這兩者之間的關係尚未被完全的了解。

經過上面的闡述發現，要去預估停頓這項韻律參數，本身能得到的信息就受限了，再加上本實驗的結果是要整合至 TTS 系統中，使用的信息只能從文字方面獲得，所以又更加無法使用重音或是聲學特徵，更何況重音本身就是一項需要被預測的信息。

未來的研究是希望使文字分析器，能僅由文字去預估停頓。除了會採用前人所使用的詞類、詞長、離句首及句尾的距離等參數，還加入了特別字詞的參數。基本想法是以前述中，已建立的詞綴表為出發點，再加上連接詞、介詞，這三種特別的字詞著手。採用的語料庫為五萬字的 Treebank 文字資料庫，以下表 5-1 為資料庫的統計結果。

表 5-1：Treebank 五萬字文字資料庫統計

文章篇數	379
總詞數	57,266
中文詞數	53,697
標點符號數	3,569

因為需要觀察的特徵參數有詞類、詞長等，為求正確，因此本語料庫是經由人工斷詞並標記詞類後，利用江振宇自動標記停頓類型的模型標註停頓標記。而標記的停頓類型，採用中研院所提出的六種停頓標記[12]，包括：B0、B1、B2-1、B2-2、B3 及 B4。其中 B0 代表 highly-coupled 音節邊界，B1 代表 normal 音節邊界，B2-1 代表人可分辨的 prosodic word(PW)邊界，具有 F0 movement，B2-2 代表人可分辨的 prosodic word(PW)邊界，具有停頓，B3 代表 prosodic phrase(PPh)邊界，B4 代表 breath group/prosodic phrase group(BG/PG)邊界。

5.1 從詞綴觀察停頓分佈之統計

我們目前建立的詞綴表中，前接詞綴共有 121 個，後接詞綴有 195 個。在本節我們將直觀文字資料中，前、後詞綴與其接頭、接尾詞連接。因此第一步便著手統計前、後接詞綴與其接頭、接尾詞個別停頓標記的數量，程度請見下表 5-1-1 與 5-1-2

表 5-1-1：前詞綴與接頭詞之 Break type 統計表節錄

前詞綴	字詞	“B0”	“B1”	“B2-1”	“B2-2”	“B3”	“B4”
新	one	2	13	0	0	0	0
	more	2	3	0	0	0	0
…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…

大	one	3	31	2	0	0	0
	more	7	33	0	0	0	0
…	…	…	…	…	…	…	…

表 5-1-2：後詞綴與接尾詞之 Break type 統計表節錄

後詞綴	字詞	“B0”	“B1”	“B2-1”	“B2-2”	“B3”	“B4”
地	one	1	27	2	1	0	0
	more	0	4	0	0	0	0
…	…	…	…	…	…	…	…
到	one	0	16	0	0	0	0
	more	4	141	0	0	0	0
…	…	…	…	…	…	…	…

上面兩表的第一列中，標記著〝B0〞、〝B1〞、〝B2-1〞、〝B2-2〞、〝B3〞、〝B4〞，在表 5-1-1 中表示前詞綴與其後接頭詞兩者間的相接點的停頓標記，在表 5-1-2 中表示後詞綴與其前接尾詞兩者間的相接點的停頓標記。

由上述兩表可觀察到，前詞綴與其接頭詞、後詞綴與其接尾詞兩者的停頓標記多為〝B0〞與〝B1〞，說明了，前詞綴與其接頭詞之間沒有很長的停頓，後詞綴與其接尾詞之間同樣的也沒有很長的停頓，偶爾只有短暫的停頓或是強調重音的形式，其餘幾乎皆是緊密相接的，這與我們的直覺也相當一致。

詞綴因點綴其他詞而成一新詞，所以上面的統計只是更進一步使我們確信這些詞的確屬於詞綴。但詞綴不僅能與其他詞相接出現，也能單獨出現，那詞綴在什麼情況下會是詞綴呢？詞綴這麼特殊的詞，其停頓的特徵真的只與其接頭、尾詞緊密相接而已嗎？

為了探討詞綴與前後文的關係，我們對語料庫進行一次人工觀察，發現某些詞綴不單單只跟其接頭、接尾詞緊密相接，某些詞綴還會有很高的機率與上下文斷開，因此我們接著統計詞綴與前後文的停頓標記，請看下表 5-1-3 及 5-1-4

表 5-1-3：前詞綴與前後文之停頓標記統計節錄

不	前＼後	Null	non	minor	major
	Null	0	98	4	0
	non	0	141	11	0
	minor	0	210	7	0
	major	0	60	6	0

表 5-1-4：後詞綴與前後文之停頓標記統計節錄

人	前＼後	Null	non	minor	major
	Null	0	0	0	0
	non	0	166	102	136
	minor	0	4	1	0
	major	0	0	0	0

如按照原本分為六類停頓標記做統計，因為各別的數量過少，對於未來的研究，實質上並無太大的幫助，因此將六類合併為三類做統計，〝B0〞與〝B1〞併為一類成〝non〞代表沒有停頓，〝B2-1〞與〝B2-2〞併為一類成〝minor〞代表有短暫的停頓，〝B3〞與〝B4〞併為一類成〝major〞代表有較長的停頓，〝Null〞代表前後文為標點符號或同為句子啟始。

經由上面的統計發現，前接詞綴當中的“不”、“可”、“無”、“全”、“共”不僅有很高的比例與後詞相接，也有較高的比例與前接詞斷開。而後詞綴當中的“面”、“地”、“黨”、“隊”、“力”、“感”、“家”、“權”、“者”、“場”、“人”、“額”、“員”、“物”、“心”、“出”、“到”等不僅會與前詞相接，還有較高的比例會與後詞斷開。

在統計的結果當中，後詞綴有一類其詞類很特別“Ng”，我們在下一節討論此詞類的停頓標記。

5.2　詞類“Ng”之停頓分佈統計

在 5.1 節後半的統計當中，如下表 5-2-1，我們發現了一類特別的詞綴，這一類的詞綴其所擁有的詞類為“Ng 後置詞”以及“Ncda 位置詞”。

表 5-2-1：後詞綴詞類“Ng”、“Ncda”與前後文停頓標記之統計節錄

上	後 / 前	Null	non	minor	major
front	Null	0	0	0	0
	non	0	22	7	31
	minor	0	0	0	0
	major	0	0	0	0
後	後 / 前	Null	non	minor	major
front	Null	0	0	0	0
	non	0	10	13	66
	minor	0	1	0	1
	major	0	0	0	0

在之前的統計當中，因為考量到數量的問題，並無將這兩類詞類分開統計，且在目前的統計當中，皆只是單字詞且為詞綴的部份。但是語料庫當中，有些二字詞或是其他詞的詞類也是“Ng”或“Ncda”。所以我們對語料庫再作了一次“Ng”及“Ncda”兩類詞類的完整統計，在這次的統計，我們分開計算“Ng”與“Ncda”兩類詞類其前後接詞的停頓分布，結果如下表 5-2-2

表 5-2-2：“Ng”及“Ncda”與前後文停頓標記之統計

Ng	No break	Minor break	Major break	Total
前接詞	535	15	0	550
後接詞	107	75	368	550
Ncda	No break	Minor break	Major break	Total
前接詞	112	9	0	121
後接詞	49	25	47	121

由表中可觀察出，“Ng”與後接詞之間有停頓的的比例高達 0.805，而在“Ng”與後接詞之間沒有停頓的 107 例中，後接詞類為 DE 的有 38 例、為 VE 的有 10 例，為 Ng 的有 1 例。而“Ncda”經過統計，它和後接詞間的停頓關係雖然並沒有 Ng 來的明顯，但與後接詞之間有停頓的比例也有 0.595，而沒有停頓的 49 例當中，有 17 例是後接詞類 DE，3 例是後接詞類 VE。

由觀察結果發現兩個很特別的詞類 VE 與 DE，Ng 後接此兩個詞類時，會使得詞類為 Ng 的詞與後接詞不僅不會斷開，反而與後接詞緊密相接，而將停頓點由 Ng 轉至 VE，例如：

表 5-2-3：“Ng”因“VE”而停頓後移範例

文章	文句	詞長及詞內位置	詞類	Prosody State	Break Type	Pause Duration
treebank_274	合	301	Ncb	13	0	0.001625
treebank_274	議	302	Ncb	12	1	0.001688
treebank_274	庭	303	Ncb	10	1	0.001500
treebank_274	審	201	VC2	10	0	0.000500
treebank_274	理	202	VC2	12	1	0.001813
treebank_274	後	101	Ng	6	0	0.001625
treebank_274	認	201	VE2	6	0	0.001625
treebank_274	為	202	VE2	5	4	0.500870

而“DE 的”本身就形如附著語素，獨立存在並無意義，所以會與前詞緊密連接是相當合理的。

觀察到如此特別的詞類 VE，在下一節，我們做進一步統計。

5.3 詞類“VE”之停頓分佈統計

經由上一節的統計發現,詞類為“VE”的詞在語料庫當中會將詞類“Ng”本該停頓的點沒有出現停頓，反而將停頓往後移。參閱中研院平衡語料庫的技術報告，當中對“VE”的定義為動作句賓述詞，後接句賓語的動作及物述詞。

由定義[11]而言，此類的詞後面必須接一子句，由直觀的猜測，為了闡述後面承接的子句，於此動詞之後應會出現停頓，將整體說話的狀態重新還原。因此我們統計“VE”與其後接詞的停頓分佈，分佈結果如下表 5-3-1：

表 5-3-1：“VE”與其後接詞之停頓分佈統計

VE	No break	Minor break	Major break	Total
後接詞	241	210	395	846

經由統計結果發現“VE”與其後接詞之間出現停頓的比例高達 0.715。且在所有的統計量當中，“VE”與後接詞之間出現長停頓的次數是最高的，這映證了我們的猜測，“VE”之後為了闡述一子句，不僅會停頓，還有較高的比例是產生長停頓將整體說話的狀態重新還原。

雖然產生停頓的比例高達 0.715，但在 846 次的統計量當中還是出現 241 次“VE”與其後接詞之間是沒有停頓產生。

因此我們更進一步的仔細探究“VE”其結構發現，“VE”依其論元個數不同共分為兩大類。一類為三元述詞“VE1”，另一類為二元述詞“VE2”，而其中“VE1”又可細分為問類“VE11”及說類“VE12”。

- 三元述詞表示：以主事者(agent)為主語，以終點(goal)為間接賓語，客體(theme)為直接賓語(句賓語)。
- 二元述詞表示：以主事者(agent)為主語，終點(goal)為句賓語。

由上述的定義可以得知，“VE1”後面會承接某個間接賓語之後再承接一個句賓語，而“VE2”後面是直接承接一個句賓語。或許“VE”之後沒產生停頓的原因，是因為“VE1”在其後承接一間接賓語造成，因此我們做了以下的統計，請見下表 5-3-2：

表 5-3-2：“VE11”及“VE12”之停頓分佈統計

“VE11”	No break	Minor break	Major break	Total
後接詞	12	2	19	33

“VE12”	No break	Minor break	Major break	Total
後接詞	27	9	11	47

上表當中，“VE11”之後出現長停頓的數量最多，造成此現象的原因是因為，“VE11”表問類，在語料庫當中後面常接標點符號，所以大多都是長停頓，至於後面不是接標點符號但卻出現長停頓的次數只有 3 次。而“VE12”表說類，其後較少直接連接標點符號，所以出現長停頓的次數不像“VE11”是所有統計量當中最多的，統計結果反而如我們所預測的，沒有停頓的次數有 27 次，是所有統計量中最高的。且“VE12”其後出現長停頓的 11 次統計量當中，有 6 次其後

是連接標點符號而造成長停頓。

經由上表的統計，確實再次映證了我們的猜測，因“VE1”會承接一個間接賓語而導致“VE1”與間接賓語之間沒有停頓產生。但是扣除掉“VE1”的影響還是有 202 次的統計量是沒有停頓的。因此用人工觀察這些沒有出現停頓的例子，概要的將原因分為三類，見下表 5-3-3：

表 5-3-3：“VE2”後無停頓之分類表

“VE2”後無停頓之分類	承接的子句開頭為一賓語或修飾詞	承接的詞類為 DE、Di、Ng、T	其他	Total
Number of “No break”	90	78	34	202

由上表可得知，第一類：“VE2”之後承接的子句開頭為一賓語或修飾詞，會造成詞類上的混淆，使得“VE2”將承接子句的主體或修飾語作為間接賓語，導致“VE2”與後接詞之間沒有停頓產生，舉下表 5-3-4 說明：

表 5-3-4：“VE2”後無停頓之範例

文章	文句	詞長及詞內位置	詞類	Prosody State	Break Type	Pause Duration
treebank_007	人	201	Naeb	11	0	0.001063
treebank_007	們	202	Naeb	12	1	0.013813
treebank_007	常	101	Dd	12	1	0.000188
treebank_007	見	101	VE2	12	0	0.001063
treebank_007	一	201	DM	11	1	0.032938
treebank_007	種	202	DM	10	2-2	0.124190
treebank_007	大	201	Nad	13	1	0.001375
treebank_007	型	202	Nad	9	1	0.000875
treebank_007	卷	201	Nab	8	0	0.001063
treebank_007	毛	202	Nab	8	2-1	0.001063
treebank_007	黑	201	Nab	11	1	0.018562
treebank_007	犬	202	Nab	7	3	0.369880

另外一類是“VE2”後面承接“DE、Di、Ng、T”此四類詞類，這四類詞類的詞本身就是後置或是語助詞的特徵，在講話時會輕輕的帶過，而不會在此四類詞類之前產生停頓，所以使“VE2”與此四類詞類相接時，“VE2”的停頓會往後移，等待一更大的子句，舉下表 5-3-5 說明：

表 5-3-5：“VE2”停頓後移之範例

文章	文句	詞長及詞內位置	詞類	Prosody State	Break Type	Pause Duration
treebank_044	以	101	P11	13	2-1	0.001625
treebank_044	行	201	Nad	14	1	0.001625
treebank_044	動	202	Nad	15	1	0.001688
treebank_044	說	201	VE2	6	0	0.001688
treebank_044	明	202	VE2	7	0	0.001688
treebank_044	了	101	Di	7	3	0.248310
treebank_044	他	101	Nhaa	5	1	0.001625
treebank_044	在	201	VK1	8	1	0.002500
treebank_044	乎	202	VK1	9	2-1	0.001688
treebank_044	妳	101	Nhaa	12	1	0.001688
treebank_044	的	101	DE	11	2-2	0.045000
treebank_044	感	201	Nac	9	1	0.060000
treebank_044	覺	202	Nac	9	2-2	0.118120
treebank_044	與	101	Caa	7	1	0.005375
treebank_044	期	201	Nac	9	0	0.002063
treebank_044	望	202	Nac	9	4	0.663310

其餘的 34 例，尚無法歸納出是何原因造成“VE2”沒有產生停頓。

5.4 連接詞與介詞之停頓分佈統計

前面三節，從詞綴作為特別字詞的初始著眼點開始，緊接著討論兩個特別的詞類“Ng”與“VE”。這節將要討論的是連接詞與介詞。為什麼要討論這兩類詞類，主要是因為這兩類詞類的功能單一，具備較鮮明的特徵，且這兩類詞類是使

句子結構更加豐富的兩個重要詞類，以下分兩小節討論這兩類的停頓分佈統計，並依據統計結果，粗劣的選定特殊詞。

5.4.1 連接詞之停頓分佈統計

連接詞，顧名思義就是連接兩個或兩個以上的語言單位，組成較大的語言單位。連接詞所連接的範圍可能是在一簡單句內，組合兩個對等的詞組成分。不過大部分連接詞的作用範圍超過一個簡單句，藉以標明兩分句間的承接關係。

根據連接詞作用的範圍，及其在句中扮演的角色地位，可將連結詞分為表詞組並列關係的並列連接詞，以及表分句關係的關聯連接詞。

5.4.1.1 （Ca）並列連接詞

在簡單句的範圍內，連接兩個概念相似的成份，組成成份的作用與其所連接的成份相同。由於並列連接詞具有以上功能，故將並列連接詞作為中心，觀察其前後兩個相接成份的停頓狀態，而並列連接詞又可細分為兩類（1）Caa 對等連接詞（如：和、跟、或者）與（2）Cab 列舉連接詞（僅只：等、等等、之類）。

因為列舉連接詞的詞有限，較好觀察。但對等連接詞不然，因此我先將語料庫當中的並列連接詞抽出並統計數量。如某連接詞出現次數低於 20 次，因為其本身的樣本數過少，於本文當中將不再討論。

（1） Caa 對等連接詞：

經過統計之後，出現超過 20 次的詞共有：及、與、或、和、以及五個詞。

而對等連接詞與前、後成份的停頓共有四類：(1) 前後皆不停 (2) 前停後不停 (3) 前不停後停 (4) 前後皆停，我們利用下表 5-4-1-1 來說明：

表 5-4-1-1：對等連接詞“及”與前後成份之停頓分佈統計

及	前＼後	NULL	B0	B1	B2-1	B2-2	B3	B4
	NULL	0	0	0	0	2	0	0
	B0	0	0	0	0	0	0	0
	B1	0	0	0	0	0	0	0
	B2-1	0	0	0	0	0	0	0
	B2-2	0	3	5	9	2	0	0
	B3	0	8	21	13	12	1	0
	B4	0	0	0	0	0	0	0

由上表發現，大多的對等連接詞“及”都與前接成份斷開，不然就是與前後成份皆斷開，這樣的現象不僅“及”如此而已，經過統計結果發現，另外的對等連接詞“與”和“或”，也有如此的現象。而“和”的結果如下表 5-4-1-2：

表 5-4-1-2：對等連接詞“和”與前後成份之停頓分佈統計

和	前＼後	NULL	B0	B1	B2-1	B2-2	B3	B4
	NULL	0	0	1	1	1	0	0
	B0	0	0	0	0	0	0	0
	B1	0	4	5	8	3	0	0
	B2-1	0	0	0	0	0	0	0
	B2-2	0	1	0	18	3	0	0
	B3	0	0	3	24	8	0	0
	B4	0	0	0	0	0	0	0

由上表得知，對等連接詞“和”大多也是與前後成份斷開，但有些微的部份是與前後皆不停或是前不停後停，而前不停後停的例子當中，與後成分之間的停頓多在 0.03 秒以下，聽覺上不易察覺。

至於“以及”的停頓統計分佈請見下表 5-4-1-3：

表 5-4-1-3：對等連接詞“以及”與前後成份之停頓分佈統計

以及	前＼後	NULL	B0	B1	B2-1	B2-2	B3	B4
	NULL	0	2	6	2	2	5	0
	B0	0	0	0	0	0	0	0
	B1	0	0	0	0	0	0	0
	B2-1	0	0	0	0	0	0	0
	B2-2	0	0	0	0	0	0	0
	B3	0	0	1	3	2	0	0
	B4	0	0	0	0	1	0	0

由上表，只可做初步的猜測“以及”常置於句首或是與前接成分斷開。而置於句首也表示“以及”之前常有標點符號，也可表示成與前接成分充份斷開。

（2） Cab 列舉連接詞：

這一類的詞只有「等、等等、之類」三個。而在語料庫當中“等等”與“之類”出現次數太少，所以以下只有列舉連接詞“等”的討論，首先請先看下表 5-4-1-4：

表 5-4-1-4：列舉連接詞“等”與前後成份之停頓分佈統計

等	前＼後	NULL	B0	B1	B2-1	B2-2	B3	B4
	NULL	0	0	0	0	0	0	0
	B0	0	0	0	0	0	0	0
	B1	0	5	10	0	7	4	10
	B2-1	0	0	1	0	1	0	0
	B2-2	0	1	0	1	0	1	0
	B3	0	0	0	0	0	0	0
	B4	0	0	0	0	0	0	0

由上表得知，列舉連接詞“等”多與前詞相接，與前詞之間並沒有停頓產生，而跟後接詞的停頓關係，初步觀察，可能是語意上的影響，如“等”之後的詞與之前的詞有關聯(如：後成分的詞需是比前成分較高層次的集合類稱)，則“等”與後詞相接不停頓。下舉一表說明：

表 5-4-1-5：列舉連接詞後未停頓之範例

文章	文句	詞長及詞內位置	詞類	Prosody State	Break type	Pause Duration
treebank_248	其	201	Neqa	7	1	0.020000
treebank_248	他	202	Neqa	4	0	0.001688
treebank_248	內	301	Nad	10	1	0.001688
treebank_248	在	302	Nad	6	1	0.001625
treebank_248	性	303	Nad	3	0	0.001688
treebank_248	美	101	Nv4	4	4	0.556750
treebank_248	、	101		-1	-1	0.000000
treebank_248	人	101	Nab	13	0	0.001688
treebank_248	之	101	DE	14	0	0.001688
treebank_248	美	101	Nv4	12	3	0.278310
treebank_248	。	101		-1	-1	0.000000
treebank_248	興	201	Nac	8	1	0.001625
treebank_248	趣	202	Nac	7	1	0.020813
treebank_248	投	201	VH11	6	1	0.001625
treebank_248	合	202	VH11	6	1	0.026437
treebank_248	等	101	Cab	8	1	0.001625
treebank_248	方	201	Nac	5	0	0.001688
treebank_248	面	202	Nac	5	4	0.545000

反之，如果沒有關聯(後成分非前成分之詞集合)，則“等”與後詞之間產生停頓，而重新還原說話的狀態，闡述之後的語意，因此，統計的結果顯示，如有停頓的發生，多為較長的停頓。下舉一表說明：

表 5-4-1-6：列舉連接詞後有停頓之範例

文章	文句	詞長及詞內位置	詞類	Prosody State	Break Type	Pause Duration
treebank_298	加	201	Cbcb	5	1	0.001688
treebank_298	上	202	Cbcb	4	3	0.265000
treebank_298	政	201	Nac	13	1	0.000125
treebank_298	府	202	Nac	9	1	0.001688
treebank_298	所	201	VK2	5	0	0.001688
treebank_298	謂	202	VK2	5	2-2	0.086625
treebank_298	統	201	Nab	4	1	0.046625
treebank_298	派	202	Nab	8	4	0.405000
treebank_298	。	101		-1	-1	0.000000
treebank_298	反	301	Nab	12	1	0.001688
treebank_298	動	302	Nab	15	1	0.003250
treebank_298	派	303	Nab	9	1	0.020000
treebank_298	等	101	Cab	6	2-2	0.180310
treebank_298	保	201	VH11	8	1	0.001688
treebank_298	守	202	VH11	5	0	0.001688
treebank_298	人	201	Nab	6	1	0.001625
treebank_298	士	202	Nab	6	1	0.001688
treebank_298	之	101	DE	5	1	0.040000
treebank_298	打	201	Nv1	5	0	0.001625
treebank_298	壓	202	Nv1	6	3	0.266690
treebank_298	、	101		-1	-1	0.000000

5.4.1.2 (Cb)關聯連接詞

關聯連接詞的功能是把分句連成複句的形式，是句子層次的修飾語。關聯連接詞可以出現再前一分句或是後一分句，且多半位於一分句的動詞之前，只有少數的例外

根據分句之間地位關係的不同，複句可分為聯合複句和偏正複句。前者分句地位平等；後者分句有主從之分，偏句對主句有說明限制的作用。例如：

(1)只要太陽一出來，雪人馬上不見。(偏正句)

(2)他不但圖文並茂，而且唱作俱佳。(聯合句)

偏正句的語意可分為轉折、假設、因果、條件、取捨、目的。聯合句的語意有選擇、遞進、並列。部分偏證據的偏句可倒置，成為後一分句。例如上例(1)也可說「雪人馬上不見，只要太陽一出來。」

根據上述的接續關係，共可將關聯連接詞分為以下各類。在說明分類的同時，並將在語料庫當中出現一定數量的關聯連接詞列舉出來，其餘未列舉出來的關聯連接詞，於本節將不被討論。

1. Cba　移動性前繫連接詞：語意上具起頭作用，後面常需接一個分句，其所在分句可能移位至複句的後半段。(下分兩類)

   - Cbaa　偏正句移動性連接詞。例：因、因為、如果、由於
   - Cbab　偏正句句尾連接詞。這一類只有“的話”和“起見”。

2. Cbb　非移動性前繫連接詞：語意上具起頭作用，後面常需接一個分句，位置固定在前一分句。(下分兩類)

   - Cbba　偏正句非移動性前繫連接詞。例：就是
   - Cbbb　聯合句前繫連接詞。例：首先、一來(此兩例語料庫中數量過少)

3. Cbc　後繫連接詞：能將一個分句聯繫於前一個句子的連接詞。(下分兩類)

   - Cbca　偏正句後繫連接詞。例：以、而、但、然而、因此、所以、不過、但是
   - Cbcb　聯合句後繫連接詞。例：並

以下將上述的關聯連接詞，依其分類再加上語意劃分得下表 5-4-1-7 及 5-4-1-8

表 5-4-1-7：聯合複句連結詞語意及位置分類表

選擇	要麼、要不	要麼、要不
遞進	非但、不獨、不但、不僅	而且、並且、且、反而
並列	首先、一來、一方面	其次、二來、二方面

表 5-4-1-8：偏正複句連接詞語意及位置分類表

	前繫		後繫
	移動性	非移動性	
轉折			而、但、然而、不過、但是
			+contrast
因果	因、因為、由於		所以，因此
	+reason		+result
假設	如果	就是	
	+hypothesis	+uncondition	
目的			以+purpose

以下從最單一特性的詞“的話”討論起，請見下表

表 5-4-1-9：句尾連接詞“的話”與前後文之停頓分佈統計

的話	後 / 前	NULL	B0	B1	B2-1	B2-2	B3	B4
	NULL	0	0	0	0	0	0	0
	B0	0	0	0	0	0	0	0
	B1	0	0	0	0	0	0	5
	B2-1	0	0	0	0	0	0	1
	B2-2	0	0	0	0	0	0	0
	B3	0	0	0	0	0	0	0
	B4	0	0	0	0	0	0	0

由上表觀察，的確“的話”符合句尾連接詞的特性且需連結在前的分句，所以所有的統計量皆為與前詞連接，並在最後結束。至於有一個例外的統計量，根據人

工觀察的結果，應為標記錯誤導致。

接著討論同一語意當中，個數較少的開始討論，首先先從表“目的”的“以”開始，且其位置屬於後繫連接詞，因此通常所在的出現位置會是後句的句首，我們看下表的統計結果

表 5-4-1-10：偏正後繫連接詞“以”與前後成分之停頓分佈統計

以	前＼後	NULL	B0	B1	B2-1	B2-2	B3	B4
	NULL	0	1	11	3	0	0	0
	B0	0	0	0	1	1	0	0
	B1	0	0	0	0	0	0	0
	B2-1	0	0	0	0	0	0	0
	B2-2	0	0	0	0	0	0	0
	B3	0	1	1	0	0	0	0
	B4	0	0	0	0	0	0	0

由上表的統計結果，確實“以”符合置於句首的特性，不然就是在其之前會產生較長的停頓。而“以”的語意是表“目的”，所以其後會承接一分句闡述目的，因此通常其後較不易出現停頓。

其次我們要討論語意表“假設”的詞“如果”，“如果”有主動的限制假設條件，看下表的統計結果

表 5-4-1-11：偏正移動性前繫連接詞“如果”與前後成分之停頓分佈統計

如果	前＼後	NULL	B0	B1	B2-1	B2-2	B3	B4
	NULL	0	9	14	8	6	2	0
	B0	0	0	1	0	0	0	0
	B1	0	0	0	0	0	0	0
	B2-1	0	0	1	0	0	0	0
	B2-2	0	0	0	0	1	0	0
	B3	0	1	3	2	0	0	0
	B4	0	0	0	0	0	0	0

由上表發現，依舊符合其置於句首的特性，且“如果”的語意為含有主動條件的假設，所以會與後詞連接來闡述假設句，因此也符合統計量大多與後詞為無停頓或是短暫的停頓。

再來我們討論表“因果”關係的連接詞，因特性類同，所以只在語意表“因”及“果”的詞當中各取其中一個詞來討論，就拿“由於”跟“因此”說明先見下表

表 5-4-1-12：偏正移動性前繫連接詞“由於”與前後成分之停頓分佈統計

由於	前＼後	NULL	B0	B1	B2-1	B2-2	B3	B4
	NULL	0	3	14	3	0	2	0
	B0	0	1	1	1	1	1	0
	B1	0	0	0	0	0	0	0
	B2-1	0	0	0	1	0	0	0
	B2-2	0	0	0	0	0	0	0
	B3	0	1	1	1	0	0	0
	B4	0	0	0	0	0	0	0

由上表可見，“由於”此類偏正移動性前繫連接詞不僅只置於句首，偶有移動至複句的後半段，造成與前詞成分之間沒有停頓。且因“由於”表原因，後面需承接原因子句，因此統計量結果顯示，大多為連接其後的原因句而沒有出現停頓。

表 5-4-1-13：偏正後繫連接詞“因此”與前後成分之停頓分佈統計

因此	前＼後	NULL	B0	B1	B2-1	B2-2	B3	B4
	NULL	0	0	0	1	4	30	4
	B0	0	0	0	0	0	0	0
	B1	0	0	0	0	0	0	0
	B2-1	0	0	0	0	0	0	0
	B2-2	0	0	0	0	0	0	0
	B3	0	0	0	0	1	0	0
	B4	0	0	0	0	0	0	0

而相對於表“因”的“因此”本身為一後繫連接詞，必位於後半句的句首，所以觀察上表的統計結果，確實符合此特性，其前面多為標點符號，不然就是在其前端有長的停頓出現。就語意而言，“因此”為表結果，其後要連接一表結果的子句，但是“因此”、“所以”等詞，不同於“因”、“因為”等本身就含有表原因的詞意，“因此”本身並無含有表“結果”的詞意。

統計結果也顯示出，“因此”之後必然會出現停頓，因為本身特性跟其後的子句格格不入，如相接在一起，會導致語意不清。

最後一組，含有“轉折”語意的連接詞，在此我們以數量較多的“而”作為討論，至於“但”、“然而”、“但是”皆很明顯的出現在句首，這都很符合其詞類特性，“而”的統計結果請見下表

表 5-4-1-14：偏正後繫連接詞“而”與前後成分之停頓分佈統計

而	前＼後	NULL	B0	B1	B2-1	B2-2	B3	B4
	NULL	0	16	38	18	4	0	0
	B0	0	2	3	1	0	0	0
	B1	0	0	1	0	0	0	0
	B2-1	0	0	0	0	0	0	0
	B2-2	0	3	8	4	3	0	0
	B3	0	2	7	5	2	0	0
	B4	0	0	0	0	0	0	0

上表明顯地顯示“而”出現在句首的特性，但有時文句的書寫上，轉折分句與前分句並非一定會經由標點符號隔開。如有此情形，上表中也顯示出如連接詞之前並無標點符號，亦通常會出現停頓。

且“轉折”連接詞其後需承接一子句，闡述與前分句不同語氣或是不同語意的分句，所以其後不應會出現長停頓。由統計結果也顯示，與後接詞之間的停頓的多為 minor break，major break 完全沒有

5.4.2 介詞之停頓分佈統計

根據詞庫小組的定義，介詞在漢語中屬於前置詞(preposition)，同時也是功能詞的一種。依據典型語法級型態特徵以下列三點作為判斷介詞的標準：

1. 介詞必須引介一論元，且此論元成份不可省略。
2. 介詞不作謂語中心。
3. 介詞沒有時態，沒有嘗試貌。

且介詞是個封閉的集合，在中研院詞庫小組的《中文詞類分析》報告中有窮舉漢語中的介詞，並依介詞的語法表現和扮演語意角色歸類細分為 65 類。

以下的討論，我們針對這 65 類來做討論，不依各個介詞來作為討論的最小單元，因為介詞本身是可窮舉的，亦表示介詞本身就為一特殊詞，所以目前假設同一類的介詞會表現相同的特性。因此我們依據類別，對語料庫作統計，數量不足 30 次的將不在討論的範圍之內。

因此，在本節將被討論的類別有

表 5-4-2-1：介詞分類集合及其集合文字

P02	被、受、叫、備受
P03	為、為了
P06	由
P07	把、將
P11	以
P19	從
P21	在
P23	於、于
P31	針對、對
P35	與、同、和
P61	至、到
P62	向
P63	跟

底下列舉數量較多的部份統計結果：

表 5-4-2-2：P07 與前後詞之停頓分佈統計

P07	後 / 前	NULL	B0	B1	B2-1	B2-2	B3	B4
	NULL	0	2	8	10	0	0	0
	B0	0	0	0	0	0	0	0
	B1	0	2	19	15	4	0	0
	B2-1	0	0	0	0	0	0	0
	B2-2	0	1	6	1	0	0	0
	B3	0	2	5	10	3	0	0
	B4	0	0	0	0	0	0	0

表 5-4-2-3：P21 與前後詞之停頓分佈統計

P21	後 / 前	NULL	B0	B1	B2-1	B2-2	B3	B4
	NULL	0	10	25	55	2	0	0
	B0	0	0	0	0	0	0	0
	B1	0	13	48	86	20	2	0
	B2-1	0	1	1	0	0	0	0
	B2-2	0	1	7	13	3	1	0
	B3	0	3	6	47	2	0	0
	B4	0	0	0	0	0	0	0

表 5-4-2-4：P31 與前後詞之停頓分佈統計

P31	後 / 前	NULL	B0	B1	B2-1	B2-2	B3	B4
	NULL	0	8	13	14	0	0	0
	B0	0	0	0	0	0	0	0
	B1	0	9	17	26	6	1	0
	B2-1	0	1	3	2	1	0	0
	B2-2	0	6	10	6	4	1	0
	B3	0	1	4	18	2	0	0
	B4	0	0	0	0	0	0	0

介詞雖然屬於前置詞，但經由上列三表的統計結果顯示，介詞並沒有很顯然的與後詞成分相接，反而有點偏向與前詞成分相接的跡象。而且與前後詞的各個統計量分佈都相當的平均，表示介詞對於是否造成停頓與否的影響並不高，相當的隨機。

會造成此現象的原因或許是因為介詞的語法特性使然。由於現在語言中的介詞大部分是從述詞演變而來，所以有部份的介詞仍然保留述詞的用法。且介詞同一個詞，大部分皆擁有多功能的語意特徵，必須由引介的論元或是更後續的成分才可決定介詞的語意角色。導致雖然介詞的詞類單一，但介詞本身的功能很複雜，不如前面數節當中討論的特殊詞功能單一。

5.5 本章結論

本章共分四小節來討論詞綴、“Ng”詞、“VE”詞以及連接詞和介詞關於停頓的關係，也從中挑選出特別的字詞來作為影響停頓的依據。

對於未來要從文字預估停頓的研究上，除了詞類、詞長等參數，或是僅考慮功能詞與實詞的粗糙分類外，提供了經過統計觀察而抽取出的特殊字、詞與詞類等，新的額外考量，或許對於未來的研究上有更佳的幫助。

第六章 結論與未來展望

經過前面各章的討論與實驗結果，本章提出下列結論，以及未來研究的目標

1. 在詞綴構詞單元方面，雖然整體的正確率有 82.9%，但因為是利用斷詞之後的結果來加以構詞，因此斷詞器本身的錯誤，構詞單元要來加以承受，會造成構詞不正確或是本該可構出詞的卻未構出。未來可將詞綴構詞單元改置於斷詞器之前，增加斷詞器可挑選的候選詞組，減少斷詞錯誤。目前的詞綴構詞規則當中有許多的副詞，這類的詞要如何給定正確的詞類也是需要再改善的。而專有名詞也造成了構詞上不少的錯誤，主要原因是斷詞器並未具有處理專有名詞的能力，這也是未來斷詞器須考量的問題。

2. 我們於第三章有說明對於破音字表的編定以及語料庫的修正。破音字的處理對於語音合成來說是很重要的工作，經過我們人工方式修正語料庫，未來有正確的語料庫能提供我們建立處理破音字的模組加入斷詞器當中。

3. 我們於第四章從詞綴著手，討論了許多與停頓有關聯的特別字詞。未來在從文字預估停頓標記的研究上，提供除考量前後詞類相接、前後詞長等影響外，可以著眼這些特別字詞，嘗試套用數學模型，評量對於此研究課題是否有進一步的改善。最終設計模組並加入斷詞器中，給予斷詞結果除音節碼、詞長和詞類等額外的韻律參數，使斷詞結果富含更多韻律方面的參數，使得韻律產生器能產生更好的結果。

參考文獻

[1] 江振宇，“中文斷詞器之改進”，國立交通大學電信工程學系碩士論文，民國九十三年七月

[2] Chen,Keh-jiann,Shing-Huang Liu, "Word Identification for Mandarin Chinese Sentences", Proceedings COLING `92,pp.101-105,Nantes,France,1992

[3] Mo,Ruo-ping Jean,Yao-Jung Yang,K.J.Chen,Chu-Ren Huang, "Determinative-Measure Compounds in Mandarin Chinese: Their Formation Rules and Parser Implementation",Proceedings of ROCLING V(R.O.C. Computational Linguistics Conference) pp.215-223,1992

[4] Chen,Feng-Yi,Ruo-Ping Jean,Mo,Chu-Ren Huang and Keh-Jiann Chen, "Reduplication in Mandarin Chinese: Their Formation Rules,Syntactic Behavior and ICG Representation", Proceeding of ROCLING V(R.O.C. Computational Linguistics Conference) pp.215-223,1992

[5] 黃居仁，陳克健，陳鳳儀，魏文真，張麗麗，“「資訊處理用中文分詞規範」設計理念及規範內容”，中央研究院歷史語言研究所，中央研究院資訊科學研究所

[6] 李鍌，張孝裕等，“國語一字多音審訂表”，教育部國語推行委員會，國語文教育叢書 25，中華民國八十八年三月

[7] Liberman, M. Y. and Prince, A. S., “On Stress and Linguistic Rhythm”, Linguistic Inquiry, Vol.8, 1977, pp. 249-336.

[8] Gee, J. P. and Grosjean, F., “Performance Structure: A Psycholinguistic and Linguistic Appraisal”, Cognitive Psychology, Vol. 15, 1983, pp. 411-458.

[9] Selkirk, E., Phonology and syntax: The relationship between sound and structure, MIT press,1984.

[10] Ladd, D. R. and Campbell, N., “Theories of Prosodic Structure: Evidence from Syllable Duration”, Proceeding of the 12nd International Congress of Phonetic Sciences, 1991.

[11] 中央研究院中文詞知識庫，“中文詞類分析”技術報告 CKIP-93-05

[12] Chiu-yu Tseng and Fu-chiang Chou, " Machine readable phonetic transcription system for Chinese dialects spoken in Taiwan", The Journal of the Acoustical Society of Japan (E), Vol.20, No.3, (May 1999), 215-223

附錄 1：針對某特定破音字對詞典修正之紀錄

編號	國字	注音	註解
1	上	ㄕㄤˇ	字典一字詞收錄爲（ㄕㄤˋ） “上聲”一詞(詞典已有) 將”平上去入”加入四字詞中
2	乘	ㄕㄥˋ	字典一字詞收錄（ㄔㄥˊ） 其餘標記爲（ㄕㄥˋ)的詞皆已納入辭典 如：萬乘之國、大乘、小乘等
3	仇	ㄑㄧㄡˊ	辭典一字詞收錄爲（ㄔㄡˊ） 姓氏，討論是否納入詞典中
4	任	ㄖㄣˋ	姓氏，納入詞典中 合成：ㄖㄣˋ 辨識：ㄖㄣˊ and ㄖㄣˋ 皆標註
5	估	ㄍㄨˋ	二字詞詞典已有”估(ㄍㄨˋ)衣”一詞 三字詞詞典新增”估(ㄍㄨˋ)衣舖”一詞
6	便	ㄅㄧㄢˋ ㄆㄧㄢˊ	新增至破音字表
7	倔	ㄐㄩㄝˊ	辭典皆標註 ㄐㄩㄝˋ 將辭典”倔(ㄐㄩㄝˋ)強”改爲(ㄐㄩㄝˊ) 三字詞辭典納入”倔脾氣”一詞
8	刨	ㄅㄠˋ	一字詞僅收錄(ㄆㄠˊ) 標記爲（ㄅㄠˋ)的詞皆已納入辭典
9	伺	ㄘˋ	辭典已有”伺(ㄘˋ)候”一詞 合成標伺(ㄘˋ)候 辨識標伺(ㄘˋ；ㄙˋ)候 其餘皆標(ㄙˋ)
10	剌	ㄌㄚˊ ㄌㄚˋ ㄌㄚ	字典一字詞收錄爲（ㄌㄚˊ） 辭典僅收錄”潑剌(ㄌㄚˋ)”一詞 辭典僅收錄”大剌剌(ㄌㄚ)”一詞
11	削	ㄒㄩㄝˋ ㄒㄧㄠ	字典一字詞收錄爲（ㄒㄩㄝˋ） 削足適履、瘦削、削職等詞典已收錄 其餘詞典收錄之詞皆音(ㄒㄧㄠ)
12	創	ㄔㄨㄤ ㄔㄨㄤˋ	字典一字詞收錄爲（ㄔㄨㄤˋ） 納入破音字表 並對字典做修正
13	區	ㄡ	姓氏，納入詞典中

14	午	ㄏㄨㄛ•	“晌午”收錄至二字詞辭典 ㄕㄤˇ　ㄏㄨㄛ•
15	卒	ㄗㄨˊ ㄘㄨˋ	一字詞詞典已收錄 將音(ㄘㄨˋ)的詞收錄至辭典內 如：倉卒、卒然
16	南無	ㄋㄚˊ ㄇㄛˊ	“南無”(ㄋㄚˊ　ㄇㄛˊ)一詞收錄至合成辭典 辨識辭典收錄”南無”(ㄋㄢˊ　ㄨˊ)及”南無”(ㄋㄚˊ　ㄇㄛˊ)
17	卷	ㄑㄩㄢˊ	字典一字詞收錄爲 (ㄐㄩㄢˇ) 詞典內的卷(ㄐㄩㄢˇ)曲改爲(ㄑㄩㄢˊ)
18	厭	ㄧㄢ	字典一字詞收錄爲 (ㄧㄢˋ) 詞典已正確收錄”厭厭”一詞
19	吃	ㄐㄧˊ	字典一字詞收錄爲 (ㄔ) “口吃” 詞典已正確收錄
20	合	ㄍㄜˇ	字典一字詞收錄爲 (ㄏㄜˊ) 將詞典“公合(ㄏㄜˊ)”改爲(ㄍㄜˇ)
21	呱	ㄍㄨ	字典一字詞收錄爲 (ㄍㄨㄚ) 非”呱呱墜地(ㄍㄨ)”，正確音爲(ㄨㄚ) 辭典已正確收錄此詞
22	咯	ㄍㄜ	字典一字詞收錄爲 (ㄍㄜˊ) 將詞典”咯咯”(ㄍㄜˊ)改爲(ㄍㄜ)
23	哈	ㄏㄚˇ	“哈巴(ㄏㄚˇ　ㄅㄚ•)狗”詞典收錄爲 (ㄏㄚ　ㄅㄚ) 合成標示：ㄏㄚˇ　ㄅㄚ• 辨識摽示：兩者皆標
24	哪	ㄋㄨㄛˊ	字典一字詞收錄爲 (ㄋㄚˇ) 將”哪吒(ㄋㄨㄛˊ　ㄓㄚˋ)”收錄至詞典
25	噹	ㄉㄧㄤ	“噹噹兒”查無此詞及音
26	大	ㄉㄞˋ	字典一字詞收錄爲 (ㄉㄚˋ) “大夫”一詞標記爲(ㄉㄞˋ　ㄈㄨ)
27	尉	ㄩˋ	字典一字詞收錄爲 (ㄨㄟˋ) “尉遲”複姓，收錄至辭典
28	屬	ㄓㄨˇ	同”囑”，是否要將詞收錄至辭典???

29	強	ㄑㄧㄤˇ	字典一字詞收錄爲（ㄑㄧㄤˊ） 將辭典內”勉強(ㄑㄧㄤˊ)”修改爲(ㄑㄧㄤˇ) “倔強”一詞已修正
30	彷	ㄆㄤˊ	字典一字詞收錄爲（ㄈㄤˇ） 彷徨辭典已正確收錄 將辭典內”彷(ㄈㄤˇ)徉”修改爲(ㄆㄤˊ)
31	惡	ㄜˇ	字典一字詞收錄爲（ㄜˋ） 辭典已有“惡(ㄜˋ)心”一詞，修正爲(ㄜˇ)
32	抹	ㄇㄛˋ	字典一字詞收錄爲（ㄇㄛˇ） 辭典已有”拐彎抹角”一詞 “抹頭”、”抹胸”等詞，音(ㄇㄛˋ)(ㄇㄛˇ)皆有 代表不同的意思，要取哪一個音????
33	查	ㄓㄚ	字典一字詞收錄爲（ㄔㄚˊ） 姓氏，新增至辭典
34	滑	ㄍㄨˇ	字典一字詞收錄爲（ㄏㄨㄚˊ） 辭典已有”滑(ㄏㄨㄚˊ)稽”，修改爲滑(ㄍㄨˇ)稽。 合成標示爲”滑(ㄍㄨˇ)稽” 辨識兩者皆標示
35	爪	ㄓㄠˇ	字典一字詞收錄爲（ㄓㄨㄚˇ） 已對辭典內的詞做修正
36	瞿	ㄑㄩˊ	字典一字詞收錄爲（ㄐㄩˋ） 姓氏，新增至辭典
37	磨	ㄇㄛˋ	辭典已正確收錄”石磨”一詞，其餘音(ㄇㄛˋ)的如何處理，再討論
38	秤	ㄆㄧㄥˊ	字典一字詞收錄爲（ㄔㄥˋ） 辭典收錄爲“天秤(ㄔㄥˋ)” 合成標示：天秤(ㄔㄥˋ) 辨識摽示：天秤(ㄔㄥˋ)(ㄆㄧㄥˊ)皆標
39	翟	ㄓㄞˊ	姓氏，新增至辭典
40	舍	ㄕㄜˇ	通”捨”，先紀錄之後再討論如何處理
41	酢	ㄘㄨˋ	是否將一字詞標音(ㄕㄨㄛˋ)修改爲(ㄘㄨˋ)??
42	陸	ㄌㄧㄡˋ	記錄起來，在構詞規則中做處理

43	叨	ㄊㄠ	字典一字詞收錄爲（ㄉㄠ） 二字詞辭典皆以正確收錄 將四字詞”叨陪末座”新增至詞典
44	娜	ㄋㄨㄛˊ	字典一字詞收錄爲（ㄋㄚˋ） “婀娜”一詞辭典已正確收錄
45	屏	ㄅㄧㄥˇ	字典一字詞收錄爲（ㄆㄧㄥˊ） “屏息”、”屏棄”等詞，辭典皆已正確收錄
46	差	ㄘ ㄔㄚ	字典一字詞收錄爲（ㄔㄚ） 其餘的破音字詞大多都已正確收錄至詞典中
47	扒	ㄅㄚ	字典一字詞收錄爲（ㄆㄚˊ） 將音(ㄅㄚ)的常用詞新增至詞典
48	拓	ㄊㄚˋ	字典一字詞收錄爲（ㄊㄨㄛˋ） 詞典新增"拓碑"一詞、"拓本"辭典已正確收錄
49	撇	ㄆㄧㄝ	字典一字詞收錄爲（ㄆㄧㄝˇ） 已對詞典做修正，新增”八字還沒一撇”、”撇嘴” 兩詞至詞典
50	闢	ㄑㄩㄝ	字典一字詞收錄爲（ㄑㄩㄝˋ） 音(ㄑㄩㄝ)的詞，辭典皆已正確收錄

附錄 2：破音字表總表及常用破音字表

編號	國字	讀音	說明
1	中	ㄓㄨㄥ	中央、中國、中學、中立
		ㄓㄨㄥˋ	中的、中毒、中意
2	乾	ㄑㄧㄢˊ	乾卦、乾坤、乾乾
		ㄍㄢ	餅乾、乾杯、乾淨、乾脆
3	了	ㄌㄧㄠˇ	了結、了解、了不起、受不了
		ㄌㄜ•	「做完了！」、「這就難怪了！」
4	供	ㄍㄨㄥ	口供、供給
		ㄍㄨㄥˋ	供品、供奉
5	便	ㄅㄧㄢˋ	方便、便利
		ㄆㄧㄢˊ	便宜、便辟、大腹便便
6	倒	ㄉㄠˇ	倒閉、跌倒、倒塌、顛倒
		ㄉㄠˋ	倒影、倒轉
7	假	ㄐㄧㄚˇ	假借、假裝
		ㄐㄧㄚˋ	假期、放假
8	傍	ㄅㄤˋ	依山傍水、依傍
		ㄅㄤ	傍晚、傍午
9	傳	ㄔㄨㄢˊ	傳單、傳神、傳染
		ㄓㄨㄢˋ	左傳、傳記
10	冠	ㄍㄨㄢ	皇冠、雞冠、冠冕堂皇
		ㄍㄨㄢˋ	冠禮、冠軍、豔冠群芳、冠絕一時、冠詞
11	切	ㄑㄧㄝˋ	密切、一切、反切
		ㄑㄧㄝ	切麵、切斷、切磋
12	分	ㄈㄣ	分數、分析
		ㄈㄣˋ	部分、本分
13	創	ㄔㄨㄤˋ	開創、創舉
		ㄔㄨㄤ	重創、創傷
14	勒	ㄌㄜˋ	韁勒、勒碑、勒令
		ㄌㄟˋ	勒緊、勒死
15	勞	ㄌㄠˊ	功勞、勞工、勞駕、疲勞、勞什子
		ㄌㄠˋ	慰勞、勞軍
16	勝	ㄕㄥˋ	戰勝、尋幽覽勝、勝會
		ㄕㄥ	勝任、不勝枚舉
17	匹	ㄆㄧ	單槍匹馬、馬匹
		ㄆㄧˇ	布匹、匹配、匹夫之勇、匹敵

18	占	ㄓㄢ	占卜、占夢、占星
		ㄓㄢˋ	占有、占據、占領
19	參	ㄘㄢ	參議員、參觀、參加、曹參
		ㄕㄣ	人參、動如參商、曾參（即曾子）
20	吐	ㄊㄨˇ	吐痰、吐露
		ㄊㄨˋ	吐血、嘔吐
21	否	ㄈㄡˇ	否定、是否、不置可否
		ㄆㄧˇ	否卦、否極泰來、臧否
22	和	ㄏㄜˊ	總和、大和民族、和尚、和平、和風、和談
		ㄏㄜˋ	唱和、附和
		ㄏㄢˋ	我和你
23	哄	ㄏㄨㄥ	哄傳、哄動、哄堂大笑
		ㄏㄨㄥˇ	連哄帶騙、哄小孩
24	咽	ㄧㄢ	咽喉
		ㄧㄝˋ	哽咽
25	哪	ㄋㄚˇ	哪知、哪個
		ㄋㄚ•	「你又走哪？」（兒女英雄傳・第七回）
26	喪	ㄙㄤ	居喪、喪亡、喪服
		ㄙㄤˋ	喪失、沮喪、喪心病狂
27	喝	ㄏㄜ	喝水、喝酒
		ㄏㄜˋ	喝責、喝采、喝令、呼喝
28	嘔	ㄡˇ	嘔吐、嘔心瀝血
		ㄡ	嘔歌、嘔啞
		ㄡˋ	嘔氣、存心嘔我
29	嚇	ㄏㄜˋ	恐嚇、嚇嚇
		ㄒㄧㄚˋ	嚇唬、嚇了一跳
30	嚼	ㄐㄧㄠˊ	嚼舌、細嚼慢嚥
		ㄐㄩㄝˊ	咀嚼
31	圈	ㄑㄩㄢ	圓圈、圈套、墳圈子、圈選、商業圈
		ㄐㄩㄢˋ	豬圈、羊圈
32	地	ㄉㄧˋ	大地
		ㄉㄜ•	慢慢地
33	塞	ㄙㄜˋ	阻塞、充塞、閉塞、推諉塞責
		ㄙㄞˋ	邊塞、要塞
		ㄙㄞ	塞車、活塞、瓶塞

34	奇	ㄑㄧˊ	奇怪、奇門遁甲、傳奇、奇裝異服
		ㄐㄧ	奇數、奇拜、奇利、奇贏
35	好	ㄏㄠˇ	好朋友、好人好事
		ㄏㄠˋ	投其所好、好學不倦
36	宿	ㄙㄨˋ	宿舍、宿命、宿願、住宿
		ㄒㄧㄡˇ	一宿、整宿
37	將	ㄐㄧㄤ	將軍、將就、打將下去、將近
		ㄐㄧㄤˋ	勇將、上將、將兵
38	少	ㄕㄠˇ	缺少、少頃
		ㄕㄠˋ	少傅、少尉、少年
39	屬	ㄕㄨˇ	親屬、金屬、部屬、屬於
		ㄓㄨˇ	屬託、屬文
40	差	ㄔㄚ	差數、差勁、差不多
		ㄔㄞ	郵差、出差、夫差
41	幾	ㄐㄧˇ	幾何學、「他才十幾歲。」
		ㄐㄧ	庶幾、幾乎、「天之降罔，維其幾矣！」
42	度	ㄉㄨˋ	制度、態度、梅開二度、置之度外、度日如年
		ㄉㄨㄛˋ	忖度、量度、度長絜大
43	強	ㄑㄧㄤˊ	強壯、強身、弱肉強食、強盜、強權政治
		ㄑㄧㄤˇ	勉強
		ㄐㄧㄤˋ	倔強
44	彈	ㄉㄢˋ	彈弓、炸彈、彈道飛彈、彈子房、彈盡援絕
		ㄊㄢˊ	彈性、彈劾、彈腿、彈簧床
45	待	ㄉㄞˋ	對待、待遇、坐以待斃、待人接物
		ㄉㄞ	待不住、待會兒
46	得	ㄉㄜˊ	得到、得意
		ㄉㄟˇ	總得、「這事得虧他幫忙。」
		ㄉㄜ˙	飛得高、跳得遠
47	從	ㄘㄨㄥˊ	跟從、順從、力不從心、言聽計從、從前
		ㄗㄨㄥˋ	侍從、從犯、從兄弟
48	悶	ㄇㄣˋ	煩悶、悶得慌
		ㄇㄣ	悶熱、悶飯
49	惡	ㄜˋ	罪惡、惡化
		ㄨˋ	羞惡、交惡、可惡、厭惡
50	應	ㄧㄥ	應該、應非難事
		ㄧㄥˋ	應驗、應對、報應、姓

51	扇	ㄕㄢˋ	門扇、扇子
		ㄕㄢ	扇風、扇惑
52	挑	ㄊㄧㄠ	挑選、挑夫、挑剔、挑食、挑水
		ㄊㄧㄠˇ	挑撥、挑燈、挑戰、挑釁、挑逗
53	掃	ㄙㄠˇ	掃地、掃興
		ㄙㄠˋ	掃帚、掃把
54	擔	ㄉㄢ	負擔、擔擱
		ㄉㄢˋ	重擔、扁擔
55	教	ㄐㄧㄠˋ	佛教、教唆、教育、教材教法、諄諄教誨、
		ㄐㄧㄠ	教學生、教書匠
56	散	ㄙㄢˋ	分散、散播、散熱、散步
		ㄙㄢˇ	丸散、廣陵散（琴曲名）、散文、一盤散沙
57	數	ㄕㄨˋ	數目、數學
		ㄕㄨˇ	數來寶、數落
58	暈	ㄩㄣ	頭暈眼花、暈車、暈倒
		ㄩㄣˋ	月暈、燈暈、酒暈、血暈
59	暴	ㄅㄠˋ	暴虐、暴躁、暴斃
		ㄆㄨˋ	暴露、一暴十寒
60	曲	ㄑㄩ	彎曲、曲膝、歪曲、委曲求全、曲棍球、曲線、曲突徙薪
		ㄑㄩˇ	歌曲、元曲、曲高和寡
61	曾	ㄗㄥ	曾孫、姓（如清代有曾國藩）
		ㄘㄥˊ	曾經
62	會	ㄏㄨㄟˋ	農會、都會、機會、領會
		ㄏㄨㄟˇ	限於「一會兒」、「多會兒」等詞音
63	朝	ㄓㄠ	朝露、朝氣蓬勃、朝會、有朝一日
		ㄔㄠˊ	朝代、朝廷、朝覲、朝聖
64	校	ㄒㄧㄠˋ	學校、上校、校慶、校長、校友
		ㄐㄧㄠˋ	校量、校稿、校閱
65	樂	ㄩㄝˋ	音樂、樂經、樂器、姓（如戰國時燕國名將樂毅）
		ㄌㄜˋ	快樂、樂天知命、樂觀其成、樂此不疲
66	橫	ㄏㄥˊ	縱橫、「雲橫秦嶺家何在？」（唐・韓愈・自詠詩）
		ㄏㄥˋ	橫政、橫死、橫財
67	沒	ㄇㄛˋ	沉沒、沒收
		ㄇㄟˊ	沒有、沒用
68	泥	ㄋㄧˊ	泥土、棗泥、拖泥帶水、爛醉如泥、泥娃娃
		ㄋㄧˋ	拘泥、泥古、泥牆

69	漲	ㄓㄤˋ	熱漲冷縮
		ㄓㄤˇ	漲潮、漲價、水漲船高
70	漂	ㄆㄧㄠ	漂浮、漂泊
		ㄆㄧㄠˇ	漂母、漂白
		ㄆㄧㄠˋ	漂亮、漂賬、漂了
71	爲	ㄨㄟˊ	行爲、爲政、天下爲公、爲非作歹
		ㄨㄟˋ	爲什麼、爲民服務
72	率	ㄕㄨㄞˋ	表率、率領、率由舊章、草率、坦率、率同
		ㄌㄩˋ	速率、或然率
73	甚	ㄕㄣˋ	甚好、甚至、欺人太甚
		ㄕㄣˊ	甚麼、作甚、甚處
74	畜	ㄔㄨˋ	畜生、家畜
		ㄒㄩˋ	畜產、畜養、畜牧
75	當	ㄉㄤ	當權、當選、當時、安步當車、當機立斷、當作（應當作）、當年、當日
		ㄉㄤˋ	當鋪、勾當、當作（視爲、認爲是）
76	的	ㄉㄧˋ	目的、標的
		ㄉㄧˊ	的確
		ㄉㄜ•	美麗的、慢慢的
77	盛	ㄕㄥˋ	盛情難卻、盛氣凌人、興盛、盛會、姓（如清代有盛宣懷）
		ㄔㄥˊ	盛飯、粢盛、盛湯
78	省	ㄕㄥˇ	省分、節省、中書省
		ㄒㄧㄥˇ	反省、省親、省悟、省視
79	看	ㄎㄢˋ	看見、看病、試試看、看板
		ㄎㄢ	看門、看守、看護、看押
80	相	ㄒㄧㄤ	相像、相親相愛、互相
		ㄒㄧㄤˋ	福相、吃相、丞相、相親
81	禁	ㄐㄧㄣˋ	宵禁、禁止、囚禁
		ㄐㄧㄣ	弱不禁風、禁受、禁得起
82	禪	ㄔㄢˊ	禪坐、禪語
		ㄕㄢˋ	禪讓、封禪
83	種	ㄓㄨㄥˇ	種子、種類
		ㄓㄨㄥˋ	種田、接種
84	稱	ㄔㄥ	稱號、稱讚、稱一稱
		ㄔㄥˋ	稱職、對稱、桿稱

85	空	ㄎㄨㄥ	天空、空間、空手、空歡喜
		ㄎㄨㄥˋ	空閒、空白
86	答	ㄉㄚˊ	答數、回答
		ㄉㄚ	答應、羞答答
87	累	ㄌㄟˇ	累犯、累積、累贅、「係累其子弟。」（孟子・梁惠王下）
		ㄌㄟˋ	家累、連累、勞累、「累你多走一趟。」
88	給	ㄐㄧˇ	年給、給事中、供給、配給、給假、給與、職務加給
		ㄍㄟˇ	「我給他一本書。」、「給我拿來！」、「快給他道歉！」
89	縫	ㄈㄥˊ	裁縫、縫紉
		ㄈㄥˋ	門縫、衣縫
90	署	ㄕㄨˇ	官署、署長、環保署
		ㄕㄨˋ	部署、簽署、署理
91	翹	ㄑㄧㄠˊ	翹楚、翹首、翹舌
		ㄑㄧㄠˋ	翹翹板、翹辮子
92	聽	ㄊㄧㄥ	聽講、聽戲、垂簾聽政
		ㄊㄧㄥˋ	聽其自然、聽天由命
93	背	ㄅㄟˋ	背後、背景、違背、背書、離鄉背井
		ㄅㄟ	背書包
94	脈	ㄇㄞˋ	動脈、脈搏
		ㄇㄛˋ	脈脈含情
95	興	ㄒㄧㄥ	興建、興旺、興奮
		ㄒㄧㄥˋ	興趣、高興、興匆匆
96	荷	ㄏㄜˊ	荷花、薄荷
		ㄏㄜˋ	荷鋤、負荷
97	著	ㄓㄨˋ	名著、著作
		ㄓㄨㄛˊ	棋高一著、著手、著實
		ㄓㄠˊ	著火、「睡著了！」、「打著了！」
		ㄓㄠ	著涼、著急、「著哇！這正合我意。」
		ㄓㄜ•	坐著、「你且慢著！」
98	藏	ㄘㄤˊ	藏匿
		ㄗㄤˋ	西藏、三藏、寶藏、藏青
99	藉	ㄐㄧㄝˋ	憑藉、藉口、慰藉、醞藉
		ㄐㄧˊ	藉藉、聲名狼藉
100	處	ㄔㄨˋ	住處、益處
		ㄔㄨˇ	處理、相處、處士、處女、姓（如漢代有處興）
101	號	ㄏㄠˋ	記號、坐號、號外、號召、號碼
		ㄏㄠˊ	號哭、呼號

102	行	ㄒㄧㄥˊ	人行道、行書、五行、流行、行程
		ㄏㄤˊ	行列、同行、內行、行家、太行山
103	衝	ㄔㄨㄥ	要衝、衝突
		ㄔㄨㄥˋ	衝南走、太衝、「衝你的面子。」
104	要	ㄧㄠˋ	摘要、需要、要職
		ㄧㄠ	要求、要功、要挾、姓（如春秋時吳國有要離）
105	覺	ㄐㄩㄝˊ	知覺、發覺、覺醒、覺悟、覺得
		ㄐㄧㄠˋ	睡覺
106	說	ㄕㄨㄛ	邪說、小說、說話、說明、光說不練
		ㄕㄨㄟˋ	說客、游說
107	調	ㄊㄧㄠˊ	調羹、調色、調笑、調養、調皮
		ㄉㄧㄠˋ	租庸調法、聲調、格調、調換
108	車	ㄐㄩ	車馬炮、學富五車
		ㄔㄜ	汽車、試車、車衣服、姓
109	載	ㄗㄞˋ	刊載、文以載道、載重量、姓
		ㄗㄞˇ	一年半載、千年萬載
110	轉	ㄓㄨㄢˇ	轉學、轉彎、旋轉、颱風轉向
		ㄓㄨㄢˋ	轉圈兒、暈頭轉向、地球自轉、公轉
111	還	ㄏㄨㄢˊ	還原、償還、還債
		ㄏㄞˊ	「時間還早。」、「還是老樣子。」
112	那	ㄋㄚˋ	那個、那麼著、剎那
		ㄋㄚˇ	「那有這種事？」、「這是那門子的規矩？」
113	都	ㄉㄨ	首都、都市
		ㄉㄡ	大都如此、都是、都好
114	釘	ㄉㄧㄥ	鐵釘、補釘
		ㄉㄧㄥˋ	釘書機、釘門牌、釘扣子
115	重	ㄓㄨㄥˋ	體重、慎重
		ㄔㄨㄥˊ	重複、重來
116	量	ㄌㄧㄤˋ	容量、重量、量力而爲、較量
		ㄌㄧㄤˊ	量體重、思量、測量、度量衡
117	鑽	ㄗㄨㄢ	鑽研、鑽洞、鑽木取火
		ㄗㄨㄢˋ	鑽子、鑽戒
118	長	ㄔㄤˊ	專長、長短、長久、冗長、長生果
		ㄓㄤˇ	尊長、首長、生長
119	間	ㄐㄧㄢ	隔間、房間、時間
		ㄐㄧㄢˋ	間隙、間諜、離間、間或

120	阿	ㄜ	山阿、阿諛、阿房宮、姓（如明代有阿其麟）
		ㄚ	阿拉伯、阿伯、阿斗
121	降	ㄐㄧㄤˋ	降落傘、霜降、降雨
		ㄒㄧㄤˊ	降龍伏虎、投降、降服
122	難	ㄋㄢˊ	難堪、進退兩難、難題、爲難
		ㄋㄢˋ	災難、問難
123	露	ㄌㄨˋ	玫瑰露、暴露、顯露
		ㄌㄡˋ	露出馬腳、露了口風、衣角外露
124	養	ㄧㄤˇ	養育、撫養小孩、養蘭、養雞鴨
		ㄧㄤˋ	奉養、供養父母
125	鮮	ㄒㄧㄢ	新鮮、海鮮
		ㄒㄧㄢˇ	鮮有、鮮少
126	更	ㄍㄥ	變更、更夫、三更半夜、少不更事
		ㄍㄥˋ	自力更生、更好、更生人

附錄 3：依破音字表針對 Treebank 十二萬字修正紀錄

字	數目	紀錄
中	97	當”中”做爲辭綴的時候被斷爲一字詞，例如：生活中、生命中，以及另有”中蘇”一詞等等皆從(ㄓㄨㄥˋ)改爲(ㄓㄨㄥ) 較嚴重的錯誤爲 treebank_953.trans 檔案中其一子句： 由於今年仍循往例，集中(ㄅㄨˊ→ ㄓㄨㄥ)於同一天除權 本來猜測是一字詞詞典標音爲(ㄓㄨㄥˋ)，但查看的結果發現，詞典一字詞的標音爲(ㄓㄨㄥ)
乾	1	錯誤出現在人名的部份
了	5	有一個嚴重錯誤在 treebank_026_1.trans 比我高半個頭就可以了(ㄑㄧㄢˇ → ㄌㄜ˙)，而我喜歡靠在這樣男生的肩膀上 其餘的都是(ㄌㄜ˙)與(ㄌㄧㄠˇ)的錯誤
供	2	treebank_882.trans 但由於股票求過於供（ㄍㄨㄥˋ→ㄍㄨㄥ），基本上只買不賣又 treebank_946.trans 如何操作呢？下列之策略供（ㄍㄨㄥˋ→ㄍㄨㄥ）爲參考
勒	3	錯誤皆爲勒(ㄌㄜ→ㄌㄜˋ)索，錯誤原因爲詞典本身二字詞標音爲(ㄌㄜ)，以改正爲(ㄌㄜˋ)
冠	1	Treebank_463.trans 的獎助金給球賽中獲得冠（ㄐㄩㄢ→ㄍㄨㄢˋ）。亞軍的俱樂部，作爲
倒	1	treebank_719.trans 個兒女和外國人結婚。倒（ㄉㄠˇ→ㄉㄠˋ）不是有什麼種族成見
和	11	ㄏㄜˊ與ㄏㄢˋ的問題
地	57	(ㄉㄧˋ)與(ㄉㄜ˙)的問題，例如：輕輕地，徐緩地等等
塞	3	錯誤皆出現在專有名詞，活塞隊、拉塞克
好	2	treebank_248_2.trans 力是很神質的，一個好（ㄏㄠˇˋ→ㄏㄠˋ）說話的人會欣賞一位 treebank_351_3.trans 公司，眾志瓷器公司。好（ㄏㄠˇˋ→ㄏㄠˋ）堡公司。理新工業。
得	52	很多錯誤來自辭典的”使得(ㄉㄜˊ)”，修正爲”使得(ㄉㄜ˙)” 此種錯誤有 20 個 其餘的尚有”記得”、”曉得”、”值得”
從	7	錯誤的標音皆標成(ㄐㄝˊ)，以作修正，爲什麼會標音錯誤如此嚴重，未知
惡	1	treebank_214_2.trans 因其離開祖國，而發生惡(ㄜˋ→ ㄨˋ)感或歧視。而我們各民

應	27+2	斷爲一字詞的”應”大多都唸成(ㄧㄥ)，而字典一字詞也是收錄(ㄧㄥ)，但仍有 27 個應該唸成(ㄧㄥ)的標記成(ㄧㄥˋ) 其中有兩個錯誤爲:”應(ㄧㄥˋ)邀”標記成”應(ㄧㄥ)邀”，已將辭典做修正
教	2	treebank_250_2.trans 後來下嫁一位教授，專心理家，共教(ㄐㄧㄠ→ㄐㄧㄠˋ)導 treebank_648.trans 回曆七月戰鬥對伊斯蘭教(ㄐㄧㄠ→ㄐㄧㄠˋ)民而言是很不道德的
散	10	全部的錯誤皆爲”散戶”一詞，應唸爲”散(ㄙㄢˇ)戶”，但字典當中標爲”散(ㄙㄢˋ)戶”，已對詞典做修正
暈	1	treebank_090_1.trans 澀味的酒，以微漾，輕暈(ㄩㄣ → ㄩㄣˋ)的夜的風華。不是什麼
曾	4	treebank_376_1.trans 括陽字號驅逐艦，從未曾(ㄏㄨㄟˋ→ ㄘㄥˊ)駛達日本海域的公開 其餘三個皆出錯在人名的部份
樂	2	treebank_574.trans 一百多位景美女中的樂(ㄌㄜˋ→ ㄩㄝˋ)儀隊揭開比賽的序幕 treebank_1111.trans 在原始的人類部落裡，樂(ㄌㄜˋ→ ㄩㄝˋ)，舞，劇三項其實是不
漲	17	將正規讀音應唸(ㄓㄤˇ)卻標音爲(ㄓㄤˋ)的修正，並對詞典收錄的詞加以檢查並修正
爲	127	錯誤皆是將(ㄨㄟˋ→ㄨㄟˊ)
當	23+2	當中的 23 個是非常嚴重的錯誤，"當”都標成(ㄇㄤ) treebank_612.trans 不過當(ㄉㄤˋ→ ㄉㄤ)蘇聯總統戈巴契夫本周 treebank_661.trans 不過當(ㄉㄤˋ→ ㄉㄤ)蘇聯總統戈巴契夫本周
率	2	treebank_659.trans 百分之九的經濟年成長率(ㄕㄨㄞˋ→ ㄌㄩˋ) 。其國民平均所得目 treebank_984.trans 億一千五百萬元，核列率(ㄕㄨㄞˋ→ ㄌㄩˋ)由去年百分之八十四
相	1	treebank_176_2.trans 品質提昇到與日本產品相(ㄒㄧㄤˋ→ ㄒㄧㄤ)抗，猶言抵制日貨
的	2	treebank_332_0.trans 媒體達成其施放消息目的(ㄉㄜ•→ ㄉㄧˋ)，決策單位頗感憂慮。 treebank_896.trans 路逮捕上次掃黑行動中的(ㄉㄧˋ→ ㄉㄜ•) 漏網之魚官浩君
種	42	錯誤皆是將(ㄓㄨㄥˇ→ㄓㄨㄥˋ)

稱	3	treebank_061_2.trans 史達林極盡謾罵能事，稱(ㄔㄥˋ→ ㄔㄥ)史達林爲史賊，絕口不 treebank_363_1.trans 辦公室副主任馬曉文則稱(ㄔㄥˋ→ ㄔㄥ)，由澳門飛廣州的航機 treebank_390.trans 代理中樞，結果強森很稱(ㄔㄣˋ→ ㄔㄥˋ)職的得了四十二分，以
給	7	全部是標音標成(ㄍㄟˋˋ)修改成(ㄍㄟˇ)
署	9	文字修正爲正規唸法”部署(ㄕㄨˋ)”、”簽署(ㄕㄨˋ)”、 ”署(ㄕㄨˋ)名”
著	3	treebank_052_1.trans 這期間，我曾矛盾的試著(ㄓㄠˊ→ㄓㄜ•)想分手，但卻受不了對 treebank_141_0.trans 獨立。事見，鄒魯先生著(ㄓㄜ• →ㄓㄨˋ)中國國民黨史稿頁九七 treebank_1142.trans 前不著(ㄓㄜ•→ㄓㄠˊ)村後不著(ㄓㄜ•→ㄓㄠˊ)店，
藏	1	treebank_080_1.trans 之類的小蟲，蠹婦把它藏(ㄔㄤˊ→ㄘㄤˊ)在不引人注目的牆石縫
行	2	treebank_283_1.trans 長聯手配合的結果，在行(ㄏㄤˊ→ㄒㄧㄥˊ)政院治安會報召 treebank_617.trans 預料布希夫人將與他同行(ㄏㄤˊ→ㄒㄧㄥˊ)。
那	46+2	treebank_046_2.trans 婚中的委曲與憤怒在剎那(ㄋㄨㄛˊ→ㄋㄚˋ)間爆發了，我把婆婆 treebank_179_0.trans 家當作只是個台灣人，那(ㄉㄡ→ㄋㄚˋ)未免太被小看了 其餘 46 個都是應該唸(ㄋㄚˋ)卻標音成(ㄋㄚˇ)，原因爲例如：那些、那裡、那個等詞，如語待疑問的意思時唸(ㄋㄚˇ)，而辭典收錄這些詞的標音皆爲(ㄋㄚˇ)
都	7	修正某些詞”都(ㄉㄨ)會”成”都(ㄉㄡ)會”
重	3	treebank_165_2.trans 也能恢復性趣和性慾，重(ㄓㄨㄥˋ→ㄔㄨㄥˊ)度美滿的性生活。 treebank_588.trans 盛王朝。如今，比利雖重(ㄓㄨㄥˋ→ㄔㄨㄥˊ)披戰袍，但他已是 treebank_1064.trans 仍須做大幅改變，發回重(ㄓㄨㄥˋ→ㄔㄨㄥˊ)議
量	1	treebank_438.trans 何者會造成過大行政裁量(ㄌㄧㄤˋ→ㄌㄧㄤˊ)權之問題。蓋行政裁量(ㄌㄧㄤˋ→ㄌㄧㄤˊ)

長	3	treebank_092_1.trans 翠湖賓館，打了一個很長(ㄓㄤˇ→ㄔㄤˊ)的電話給不願回家的女 treebank_460.trans 第一場長(ㄓㄤˇ→ㄔㄤˊ)榮中學對翔宇旅行社， treebank_1130.trans 供其居住，可是女孩越長(ㄔㄤˊ→ㄓㄤˇ)越大，新蓋屋宇已容
露	3	treebank_879.trans 要求所有與會者不得洩露(ㄌㄡˋ→ㄌㄨˋ)內容，但據與會經濟部 treebank_901.trans 資產股尾盤急拉似也透露(ㄌㄡˋ→ㄌㄨˋ)出多頭預作反彈的 treebank_1173.trans 紋章，深鎖的眉宇，顯露(ㄌㄡˋ→ㄌㄨˋ)出痛苦掙扎的表情 已對詞典中的二字詞作修正
分	6	部分(ㄈㄣ)→(ㄈㄣˋ)
鮮	1	treebank_249_1.trans 形成負相關的卻屢見不鮮(ㄒㄧㄢˇ→ㄒㄧㄢ)，一位邱姓友人出身

附錄 4：前、後詞綴列表

規則標記	prefix	詞類
301	非	1,11,36
302	多	11,20,37
303	老,大,小,高,低,粗,細,淡,淺,深,古 冷,長,易,乾,軟,硬,短,新,輕,薄,舊	37
304	前,後,本,各,諸,該	18
305	總,原	11
306	代,特,專	11
307	曾,又,剛,仍,已,正,早 即,尚,便,常,現,就,遂	11
308	則,卻,才,只,也,亦,光,再,先,快,還	11
309	男,女	12
310	不,未	11
311	反	31
312	太,更,很,略,最,稍,微,極,較,頗,遠	7,11
313	來,去	11
314	可,別,該	11
315	清,甜,紅,白,黃,黑,綠,青,藍,活,遠	37
316	全	6,11
317	共,僅	6,11,37
318	好	7,37
319	有	43
320	此,這	19
321	自	11,26
322	含,帶,具	31,40
323	供,請,擬	35
324	所,皆,都	11
325	造,建,設	31
326	相,互	11
327	耐,抗,重,無,獲	40
328	駐,赴	32
329	會	11
330	變	37

規則標記	postfix	詞類
401	們	12
402	某,氏	12,13,18
403	論,點,線,篇,網,罪 庫,展,面,派,度,型	12
404	方,島,面,角,地	15
405	樹,槍,團,組,產,紋,草,根,書,集, 科,花,河,角,池,石,卡,片,山,費, 攤,販,景,魚,湖,販	12
406	欄,號,表,函,單,版	12
407	國,部,處,室,科,所 課,區,班,社,系,局	14
408	檔,餐,量,商,班,枝,率	12
409	熱,業,貨,術,風,股,狂,迷	12
410	黨,會,社	12
411	隊	12
412	罐,桶,瓶,巢,屋	12
413	心,感,性,力,慾	12
414	地	44
415	觀,學,圈,金,額,品	12
416	化	37
417	孔,戶	12
418	國,省,村,鄉,鎮,縣,市,課,社	12,14
419	箱,袋,盒	12
420	人	12
421	籍,腔,商,圈,案,制,貨,餐,裔	12
422	姓,君	12
423	價	12
424	機,器	12
425	期,曲	12
426	藥,劑,油	12
427	曲,劇,節,險	12
428	者,師,員,家,手,軍	12
429	界,門,綱,目,科,屬,種	12

430	廠,場,站,界,舍	14
431	類,群,案,板,形	12
432	權	12
433	記,物	12
434	歌,舞,史,文,宴,圖,語,話	12
435	聲	12
436	數	12
437	賽,會,式,制	12
438	桌,椅	12
439	獎,章	12
440	層	12
441	園	12,14
442	上,中,下,內, 時,前,後,來, 底,起,裡,頭	9,15,23
443	票	12
458	課,樹,線,孔,檔,餐,箱,獎,語,歌,舞,網,團,戶,路,話,號,類,節,集,隊,腔,層,期,景,魚,貨,處,組,票,球,桿,巢,假,級,病,班,圈,根,案,書,師,家,面,軍,科,派,花,股,河,板,姓,制,角,局,君,行,色,片,爪,文,手,力,箱,袋,盒,險	22
458	類,餐,檔,篇,欄,層,槍,團,道,頭,集,隊,尊,場,部,袋,處,組,粒,票,條,員,級,班,根,席,面,科,度,枝,角,局,行,色,曲,式,名,石,台,文,手,塊,	22

附錄 5：系統採用之詞綴構詞規則

GROUP	regular expression	範例
301	非(A) + {Na}	非博士,非親骨肉
301	非{VG,Dc} + {VL,Nc}	非讓,非受
302	多(D) + {D,P,VJ,VE,VG,VK}	多以,多與,多發生
302	多{VH,Neqa} + {VE,Na,Nc}	多聽,多子,多處
303	{老,大,小,高,低,粗,細,淡,淺,深,古,冷,長,易,乾,軟,硬,短,新,輕,薄,舊}(VH) + {Na,DE,Di}	老賊,老士官,大花心 小事,小裁縫店
304	{前,後,本,各,諸,該}(Nes) + {Na,Nb,Nc,Nd,Neu}	前理事,前美國,前二
305	{總,原}{(D)} + {D,VA,VC,VD,VE,VF,VJ,VK,VH}	總認爲,總該,原少不了
306	{代,特,專}(D) + {VC,VF}	代清掃,特請,專供
307	{又,已,仍,正,先,早,即,快,尚,便,剛,常,稍}(Dd) +{VA,VC,VCL,VE,VG,VH,VJ,VK1,VL}	又交過,已造成,早知道
308	{則,卻,才,只,也,亦,光,再,先,快,還}(D) + {VA,VC,VCL,VE,VF,VH,VJ,VK,VL}	也沒有,也填補,只願,只拿
309	{男,女}(Na) + {Na}	女尼,女教師,男醫生,男病人
310	不(Dc) + {A,VA,VC,VCL,VE,VH,VJ,VK,VL+C13}	不包括,不正確的
311	反(VC) + {Na,Nb,Nc}	反杜邦,反五輕,反日,反民主
312	{太,更,很,略,最,稍,微,極,較,頗,遠}(Dfa) + {VH,VJ,VK}	更成熟,最高興
313	{去,來}(D) + {VA,VC,VE,VF,VH,VJ,VK,VL)}	去愛,去參觀,來接獲
314	{可,別,該}(D) + {VC,VE,VF,P}	可用,可接獲,別以爲
315	{清,甜,紅,白,黃,黑,綠,青,藍,活,遠}(VH) + {Na,DE}	黑武士,青紅燈,甜菜
316	全(Neqa) + {Na,Nca,Nf}	全案,全國,全台灣
317	{共,僅}(D) + {VA,VC,VJ,Neu}	共載有,共教導
318	好(VH) +{VC,Na,VH,VA}	好的,好消化,好丈夫

319	有(V_2) + {Neqa,Na,DM,VH,VA}	有多少,有三個
320	{此,這}(Nep) + {Na,Nc,Nf,VC,Neu}	此事,此案,
321	自(P59) + {(VA),(VC),(Nca),(Ndaab),(Ndaad),(Ndabc)}	自民國三十八年
321	自(Dh) + {(VE2)}	自以爲
322	{含,帶,具}(VC) + {Na}	含水
323	{供,請,擬}(VC) + {Na,VA,VC,VF}	供食用,供觀賞,請參閱
324	{所,皆,都}(D) +{VK,VJ,VD,VC,VE,VF,D,VL,VH,V_2}	所造成,所期望,皆認爲
325	{建,設,造}(VC) + {P,Na}	建寺,建燈塔,設在
326	{相,互}(D) + {VJ,VC,VA}	相抗,相謀議
327	{耐,抗,重,無,獲}(VJ) + {Na,VJ}	重意境,重韻味
328	{駐,赴}(VC) + {Nb,Nc}	駐中共,駐美,駐英
329	會(D) + {VK,VC,P,VA,VH,VL,VF,VE,VD,V_2}	會造成,會有,會發生
330	變(VH) + {VH}	變酸,變黑

GROUP	regular expression	範例
401	{Nab,Naea} + 們{Neqb,Naea}	大人們,情侶們
402	{Nb} + {某,氏}{Na,Nb,Nes}	陳氏,許某
403	{A,VA,VC,VH,VJ,Na}+{論,點,線,篇,網,罪, 庫,展,面,派,度,型}{Na}	特定點,彙整點 亡國論,文學論
404	{Nes,Neu,Ncd} + {方,島,面,角,地}(Ncd)	一方,雙方
405	{Na,VC} + {樹,槍,團,組,產,紋,草,根,書,集,科 ,花,河,角,池,石,卡,片,山}{Na}	種樹,奪槍
406	{Na}+{欄,號,表,函}{Na}	進出表,
407	{Na,Nb,Nc,Nes,Neu,Ncd,Neqa,VC,VA} + {國,部,處,室,科,所,課,區,班,社,系}{Nc}	病灶處
408	{A,Na,VC} + {檔,餐,量,商,班,枝,率}{Na}	進口商,發生率
409	{Na,VC,VH} + {熱,業,貨,術,風,股,狂,迷}{Na}	哈雷熱,電玩迷
410	{Na,Neu,Nes,Nep,Neqa} + {黨,會,社}{Na}	本黨,學友會
411	{A,Na,Nb,Nc,VC} + 隊{Na}	代表隊,衝鋒隊
412	{Na} + {罐,桶,瓶,巢,屋}{Na}	飲料罐,塑膠瓶
413	{A,Na,VC,VH,VI,VJ,VK} + {心,感,性,力,慾}{Na}	同理心,自信心
414	{VK1,VK2,VH11,Dd,VE2,Dh} + 地{DE}	明白地,充分地
415	{Na} + {觀,學,圈,金}{Na}	價值觀,人際觀 心理學,行爲學
416	{A,Na,VH} + {VH}	軍事化,多媒體化
417	{Na,VA,VH} + {Na}	實戶,往來戶
418	{VC,VCL,Neu,Nes,Nep,Neqa,Na} + {國,省,村,鄉,鎮,縣,市}{Na}	兩黨,本黨,全國
419	{Na,VC,VD,VI} + {箱,袋,盒}{Na}	水果盒,塑膠袋
420	{Na,VB,VA,VC,Neu,VE,VH,VG} + 人{Na}	多數人
421	{Nc} + {籍,腔,商,圈,案,制,貨,餐,裔}{Na}	馬來西亞籍
422	{Nb,VC,FW} + {姓,君}{Na}	陳君,許姓
423	{A,Na,VA,VC,VE,VH,VHC,VJ} + 價{Na}	公定價
424	{A,Na,Nc,VA,VC,VD,VF,VH}+{機,器}{Na}	耕作機
425	{Na,VA,VC,VH} + {期,曲}{Na}	蛻變期,成長期
426	{A,Na,VA,VAC,VC,VH} + {藥,劑,油}{Na}	國產藥,
427	{A,Na} + {劇,節,曲,險}{Na}	愛情劇
428	{Na,VB,VA,VC,Neu,VE,VH,VG} + {者,師,員,家,手}{Na}	領導者,示威者,理容師

429	{Nes,Neu,Na} + {界,門,綱,目,科,屬,種}{Na}	蝸牛屬
430	{A,Na,Nb,Nc,VA,VC,VHC,Neu,DM} + {廠,場,站,界,舍}{Nc}	針織廠
431	{Na,VC,VH,FW} + {類,群,案,板,形}{Na}	經濟類,藝術類
432	{Na,VA,VC,VE,VH,VHC,VI} + 權{Na}	自由權,專利權
433	{Na,VA,VB,VC,VE,VH,VK} + {記,物}{Na}	復仇記,漂流記
434	{Na,Nb,Nc,VC} + {歌,舞,史,文,宴,圖,語,話}{Na}	劍舞,文學史
435	{A,Na,VA,VB,VE,VH} + 聲{Na}	木魚聲,討論聲
436	{A,Na,Nc,VH} + 數{Na}	大概數,特別數
437	{Na,VA,VC,VE,VG} + {賽,會,式,制}{Na}	邀請賽,季級賽
438	{VA} + {桌,椅}{Na}	辦公桌,躺椅
439	{Na,Nb,Nc} + {獎,章}{Na}	傑出獎,許可章
440	{Nc,Na} + 層{Na}	地質層,年齡層
441	{Na} + 園{Na,Nc}	動物園
442	{Nc,VC,Na,VF,VA,Nb,VH,VK,DM,P} + {上,中,下,內,時,前,後,來,底,起,裡,頭}{Ng,Ncd,Di}	地表下,襯托下
443	{Na,VC,VD,VJ,VK} + 票{Na}	電影票
458	{Neu,DM,Neqa,Nes,Nep} + {課,樹,線,檔,餐,孔,箱,獎,語,歌,舞,網,槍,圖,團,戶,路,話,號,類,節,集,隊,腔,層,期,景,魚,章,處,組,票,球,桿,巢,假,病,班,圈,根,案,書,師,家,面,軍,科,派,花,股,河,板,姓,制,角,局,君,行,片,爪,文,手,力,,袋,盒}{Na}	三科,五箱,九層
458	{Neu,DM,Neqa,Nes,Nep} + {類,餐,檔,篇,欄,層,槍,團,道,頭,集,隊,尊,場,部,袋,處,組,粒,票,條,員,級,班,根,席,面,科,度,枝,角,局,行,式,名,石,台,文,手,塊,回}{Nf}	兩類,三餐,五檔