國　立　交　通　大　學

電信工程學系碩士班

碩士論文

台語斷詞器之改進

An improvement in Taiwanese Parser

研究生：蕭希群

指導教授：王逸如　博士

中華民國九十四年七月

台語斷詞器之改進

An improvement in Taiwanese Parser

研究生：蕭希群　　Student: Hsi-Chun Hsiao

指導教授：王逸如　博士　　Advisor: Dr. Yih-Ru Wang

國立交通大學

電信工程學系碩士班

碩士論文

A Thesis
Submitted to Department of Communication Engineering
College of Electric Engineering and Computer Science
National Chiao Tung University
In Partial Fulfillment of the Requirements
For the Degree of
Master of Science
In
Electrical Engineering
July, 2005
Hsinchu, Taiwan, Republic of China

中華民國九十四年七月

台語斷詞器之改進

研究生　蕭希群　　指導教授：王逸如

國立交通大學電信工程學系碩士班

中文摘要

在本論文裡建立了一套完整的台語斷詞器以期提升台語文字轉語音器的效能。首先將中文 TTS 系統裡文字分析部份所使用到的構詞規則、文字正規化規則引入台語文字分析單元中，並依據台語文字的特性，增加了文/白讀處理機制、變調機制，並針對台語構詞的特性，增加了部份構詞規則以構出特殊的台語結構詞。

針對台語詞庫資料方面，除了詞數的擴充外，並增加了詞性(Part of Speech)資訊。詞性資訊可以作爲部分構詞規則的依據，並可因此決定變調邊界，除此之外也可以依此產生更準確的韻律訊息。對台語構詞資料方面，除了增加台語特有的構詞形式外，並針對數字做了文/白讀法的修正。針對變調規則方面，則考慮構詞/文字正規化後變調規則的套用，以期達到更接近真實台語變調的現象。

中華民國九十四年七月

An improvement in Taiwanese Parser

Student : Hsi-Chun Hsiao Advisor : Dr. Yih-Ru Wang

Department of Communication Engineering

National Chiao Tung University

Abstract

In this thesis, the capability of Taiwanese parser is improved. We use the structure of Chinese parser developed in our laboratory, with modifying the word combination module to satisfy the special property of Taiwanese, and add some module such as tone sandhi rules module、pronunciation selection module for digit number pronunciation, and so on.

We also improve our Taiwanese dictionary database, including extending dictionary capacity and adding POS (part of speech) information to word.

Finally, we use this modified parser to parse Taiwanese texts to analysis the advantage of our modification, and discussing what effort can we do in the future.

誌謝

人的際遇是非常奇妙的事情。在某些時刻所做的某些選擇，遇到了某些人，做了某些決定，在在影響自己接下來該走什麼路、該做些什麼事情。際遇就像上天安排好的棋局一樣，每件事情的發生與否，就像早已被安排好了一樣不停地運作著。

很慶幸能夠進入這個溫暖的實驗室，有兩位老師細心地指導著，還有一群可愛的同學和學長、學弟們，在專業問題上、生活經驗上所做的討論和分享，豐富了我兩年的碩士生涯，也留下了很多美好的回憶。這些回憶是最珍貴且無價的資產，也是碩士生涯中最值得我珍惜的部分。

我要特別感謝指導我的王逸如老師和陳信宏老師。王老師在關鍵問題上總會適時提供有效的做法，使我不至於以錯誤的方法埋首下去，徒然浪費寶貴時間；陳老師在問題的大方向上總是能描述得很清楚，使我能依著這個方向走下去而不會對研究的方向感到困惑和茫然。真的很謝謝兩位老師細心的引導。

感謝振宇學長，你的程式能力超強，寫出很多很經典的程式片段，我從閱讀這些片段的過程中學到了不少程式的技巧。感謝佩穎，能和你討論生活上的種種困惑和心得是一件幸福的事情，雖然從今以後見面的機會變得更少了，但是你所帶給我的快樂卻是永恆的。隆勳、金翰是我們之中最聰明的兩個人，工作時沉穩的表情讓我不禁羨慕起你們的聰明腦袋。應順總是我們之中最早來且工作態度最穩定的一個，非常有工程師該有的架式，我相信將來在工作上會有很不錯的成績的。lubo 的個性非常豪爽，抽煙的樣子超痞超帥，希望你當兵順利啦！

我要感謝我媽媽支持我繼續唸博士班，讓我沒有經濟上的憂慮，能專心做我想做的事情，真的非常感謝。還有 NDL 的 Iris 小姐，認識你真的讓我眼界大開，讓我從足不出戶變成老愛往外跑，心胸開闊了許多。

感謝的話總是說不完的，我只想說，謝謝大家豐富了我的人生，謝謝！

目錄

表目錄

圖目錄

第一章 緒論

在交通大學語音實驗室所發展的國語 Text to Speech(TTS)系統[1]中，其合成出的語音品質已達到一定的水準，不僅能產生自然的抑揚頓挫，在發音的品質上也做了一些努力。國語 TTS 在這些年來所做的努力和所獲得的經驗，有很多是台語 TTS 系統可以拿來利用的。但目前的台語 TTS 系統受到台語書寫法以及語言習慣的限制而無法得到很好的效能，但仔細分析後仍然可發現有許多可以努力的方向。本論文將針對台語文字和發音的特性，以及利用國語文字分析上得到的經驗和結果，期望能製作出更接近實際語言現象的台語 TTS 系統。

1.1 研究動機

本系統嘗試將文句由機器以自然的方式唸出，希望能達到與人類口語對話相近的自然度與流利度。由於人類在與其他人發生對話行爲時，會以一般聽者能夠適應且理解的語流速度發音，所以當我們想將文字轉爲語音時，若僅是將輸入的文字逐字發音，而不針對自然語言現象(語流速度、停頓、語句的抑揚頓挫等)做分析和使用，則合成出來的語音將會變得非常不自然。本系統嘗試經由正確的台語文字分析，以擷取正確的文字結構資訊提供文字轉語音合成系統使用，以期獲得更自然的合成語音。

1.2 研究方向

本論文針對台語文字以及語言的現象，將明顯可見的構詞規則加入台語文字分析單元中。所使用的規則包含文字和語言現象兩方面：

(1) 針對台語特有的構詞結構『附加』『重疊』等現象，搭配本實驗室中文文字分析器中所使用的構詞規則，結合成可以處理大部分台語構詞現象的構詞單元。另外並加入了詞類標記的動作，使得輸出的詞含有詞類的資訊。

(2) 加入了文字正規化單元，將數字串、符號等轉換成文字以利語音合成，並且針對台語數字部份的文/白讀音做處理，並套用變調規則於斷詞、構詞和文字正規化所形成的詞裡，期望能得到較接近實際台語變調現象的讀音。

1.3 章節概要

第一章　緒論：簡介研究動機、研究方向

第二章　台語 TTS 系統流程：簡述目前台語 TTS 整體系統運作所做的考量

第三章　台語文字分析器之設計：依據台語文字以及發音特性，細分數種情況設計出來的文字分析流程

第四章　台語文字分析器的效能分析：以定量和定性方式，討論目前的文字分析器所能達成的系統功能與系統改善的程度

第五章　結論與未來展望

第二章 台語文字轉語音系統流程

以自然語言現象爲考量的語音合成系統，應該包含自然語言現象的分析和使用，以增加發音的自然度。自然語言現象的分析可以分爲兩種層面研究，一種是藉由語者所講出的語料做字意上的分析，分析的內容包含語意、語境、詞類等。另一個層面是人類發聲器官特性對語音所造成的影響，包含了發聲器官所造成的限制，如換氣時機、語調的音高限制、語調轉折速度的限制等。

目前交大電信語音實驗室的語音合成系統是以前者做爲分析自然語言現象的依據，因爲在文字分析的部分目前已有大量的研究結果，包含中研院詞庫分析小組針對語言現象所做的語料收集和統計分析，還有各個語言學相關的學術單位所整理出來的語言規則和詞庫等，資料量非常龐大且有深度，適合以工程的方式做分析。

至於第二個層面對語音所造成的影響，所牽涉的層面包含了更複雜的語意、語境等考量，譬如語者所選擇的換氣時機，常會選在一句話的某個意思已經充分表達完整時。目前的語音合成系統並沒有針對此方面做研究。

2.1 台語 TTS 系統架構

一個語音合成系統主要分爲三個部分，包含文字分析(Text Analysis)、韻律訊息(Prosody information)的產生，以及聲音波形(Waveform)的輸出三部分，如圖 2.1 所示，這也是目前交大語音實驗室的台語 TTS 系統所使用的系統架構。

以下將針對目前台語 TTS 系統架構的三個部分簡單地描述系統運作流程、各流程所使用到的資訊，以及各流程所做的考量。

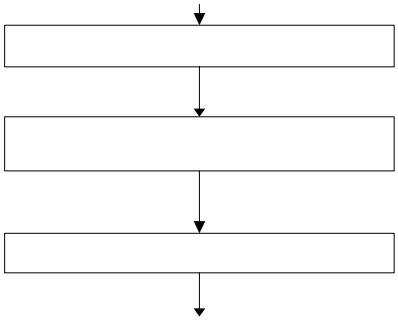

圖 2.1 語音合成系統架構

(1) 文字分析部分：將文章中的文字做斷詞分析、詞性分析、語言現象分析等，以獲得足夠的語言訊息參數。目前的語言訊息參數包含拼音、斷詞資訊(詞長、詞性、一個字在詞裡的詞首/詞中/詞尾等)、句法邊界(標點符號)等。

(2) 韻律訊息產生部份：韻律訊息包括聲調、音節長短、停頓時間長短等。

由文字分析所產生的語言訊息參數，可以獲得適合的韻律資訊，以決定聲音該如何發出。

韻律訊息的產生一般有兩種方法，一種是使用規則法(Rule based)，將某些詞法組合可能的唸法列出韻律規則，另一種就是使用較複雜的統計與自我學習的方式學習韻律規則。前者譬如句首起始音調提高、遇到問號時音調上揚、遇到句號時語調下降等。以此種方式列可能列出上百種規則，並以簡單的統計方式對規則加以修正，以產生較爲正確的韻律訊息參數。此種方式需要大量的語言學知識，適合擁有深厚語言學基礎的人定義出適當的規則來使用。

本實驗室採用的方式爲後者，嘗試以程式自我學習的方式學習到正確的韻律訊息。所使用的方式爲 Recurrent Neural Network(RNN)，以語言參數對照由聲音波形提取的韻律訊息參數做訓練，得到一個韻律訊息產生器的 training model。得到此 model 以後，便可依據語言參數自動產生韻律訊息參數。此種方法的好處是，不需要具備深厚的語言學知識，便可利用工程方法達到同樣的效果，且效果的好壞隨著訓練語料的增加和語言參數種類的擴充而變得更好，日後可再搭配少數的規則使系統更接近自然語言現象。

(3) 聲音輸出單元：目前在聲音波形部分的做法有兩種，一種是以單音節波形串接成句子(Syllable Template based)，另一種是以長句的聲音波形爲單位串接(Corpus Based)。前者是以文字分析後的結果，從大量的聲音資料庫中挑選合適的句子整段串接成句子，缺點是需要非常大量的聲音波形資料，因爲詞的組合種類非常多，很難以少量資料達到很好的效果。後者是以基本音節(不考慮音調時，中文爲 411 個基本音節，台語爲 877 個基本音節)做爲合成聲音波形的樣板波形(Template)，依據上層所得到的韻律訊息參數，以基本音節做音調、音量、音長及音節間靜音長度的改變等，串接成一個完整的句子。本實驗室目前使用後者，因爲只要少許的波形資料就可以做出基本的系統。

第三章 台語文字分析器之設計

在台語 Text to Speech(TTS)系統中，文字分析的精確與否決定系統的整體表現甚劇。主要原因是台語的一字詞通常擁有多重發音，而這種現象在中文字裡是不會發生的。如果將一篇中文文章全部斷爲一字詞，大部分的發音仍舊會正確，只是在韻律方面較不自然而已。然而若將台語文章全部斷爲一字詞，則因爲台語文字一字多音的現象，使得發音結果與實際上的正確唸法相去甚遠。

本章將針對台語文字特性，以及基於目前中文文字分析器架構所延伸出來的系統架構做分項的說明。

3.1 台語文字轉語音所面臨的問題

台語文字轉發音主要有以下幾種問題：(1) 一字多音 (2) 文句中混雜著許多口語化的詞綴及疊詞 (3) 台語變調邊界的決定，以上三種特性將影響台語 TTS 系統決定的發音結果。

接著我們將詳述這幾種問題：

(1) 一字多音：

這是台語文字中最主要的問題之一。由於台語並沒有屬於自己的文字，所以常常借用漢字中發音相近的字或者意義相近的字使用。然而這些字的選用並沒有一個標準的法則，所以常常是依照作者的習慣而選用，如此則造成一個字可能出現在很多的場合，且在不同場合擁有不同的發音。譬如『打』就有 keN5、koaN7、phah、poah、taN2 等五種可能發音。如果沒有辦法正確斷詞而使『打』以一字詞的形式出現，則將會有很大的機會將『打』唸成不適當的音。

(2) 文句中夾雜口語中的詞綴及疊詞：

台語文句中，詞綴及疊詞的使用遠較國語頻繁，以下將針對台語詞綴和

台語疊詞的部分分別做討論。

台語詞綴包含了前接詞綴與後接詞綴，前者譬如『阿明』中使用『阿』等字修飾人名或親屬稱謂，後者譬如『雞仔』、『狗仔』等爲了描述小物品或小動物所加的修飾。此類文字組合嚴格來說並不能稱得上是一個『詞』，然而在口語上及文章書寫上常常出現此類的組合。

台語的疊詞主要是爲了做口語上的強調。譬如『燒燙燙』、『冷支支』等，針對所使用的形容詞『燒』及『冷』做強調。此類文字組合本身不算是一個具有特別詞類特性的詞，但是在口語上經常連在一起使用，且這種現象也常出現在台語的文章之中。

(3) 台語變調邊界的決定

『變調』是指在語句中，某一個單字的聲調會因其位在文中位置的不同而產生變化。在國語中最常出現的變調規則就是『三三調』，也就是當一個詞中有兩個連續的三聲字接在一起，第一個三聲字的聲調會轉爲二聲。

台語的變調現象較中文複雜許多，且在語言上的意義也比國語的變調來得重要。通常台語中所有字都會變調，但在某些語意或詞意分界點的前一個字卻不會變調，我們將此分接點稱爲變調邊界。因爲台語的變調邊界時常伴隨有文法上的意義及講話時強調的重點所在，若不遵循變調規則，立刻會讓人無法理解講話的內容，或者會令聽者感到困惑。

台語變調的現象在台語口語上非常常見，但是變調的邊界卻是很難用單純的方式決定。變調邊界可決定整個句子的聲調變化，若將變調邊界放置在句子裡不適當的地方，則整個句子聽起來將很不自然。

由上述三點可以發現，在製作一個台語 TTS 的斷詞器時，如果我們僅是以字典查詢的方式作爲斷詞的依據時，將很容易將文章內容斷成大量的一字詞，造成發音選取上的困擾。然而這些一字詞是有機會根據特定的構詞結構組合成多字詞的。

當我們將數個一字詞以構詞方法構成多字詞時，除了可以解決一字詞發音的問題外，還可以根據構詞決定變調邊界。變調是台語發音中普遍存在的現象，而變調邊界通常位於詞的最後一個字。此處所提到的詞包含口語化的詞組，並不只侷限於含有特定意義、特定詞性的詞。構詞能夠將部份口語化的詞組合起來，並可依此界定構詞邊界，使得變調邊界的決定更接近實際的變調邊界。另外，我們也可以依據構詞規則決定數字發音。台語的數字發音有很明顯的文白讀之分，以往並沒有針對此種現象作處理，加入構詞單元之後，可以針對不同的構詞情形對數字發音部分作分類，以決定數字發音。

針對上述考量，我們所設計的台語文字分析流程將會較國語文字分析複雜。詳細的系統流程將在下一個小節分項探討。

3.2 台語文字分析系統流程

針對台語文字及發音的特性所設計出來的台語文字分析流程如圖 3.1 所示。

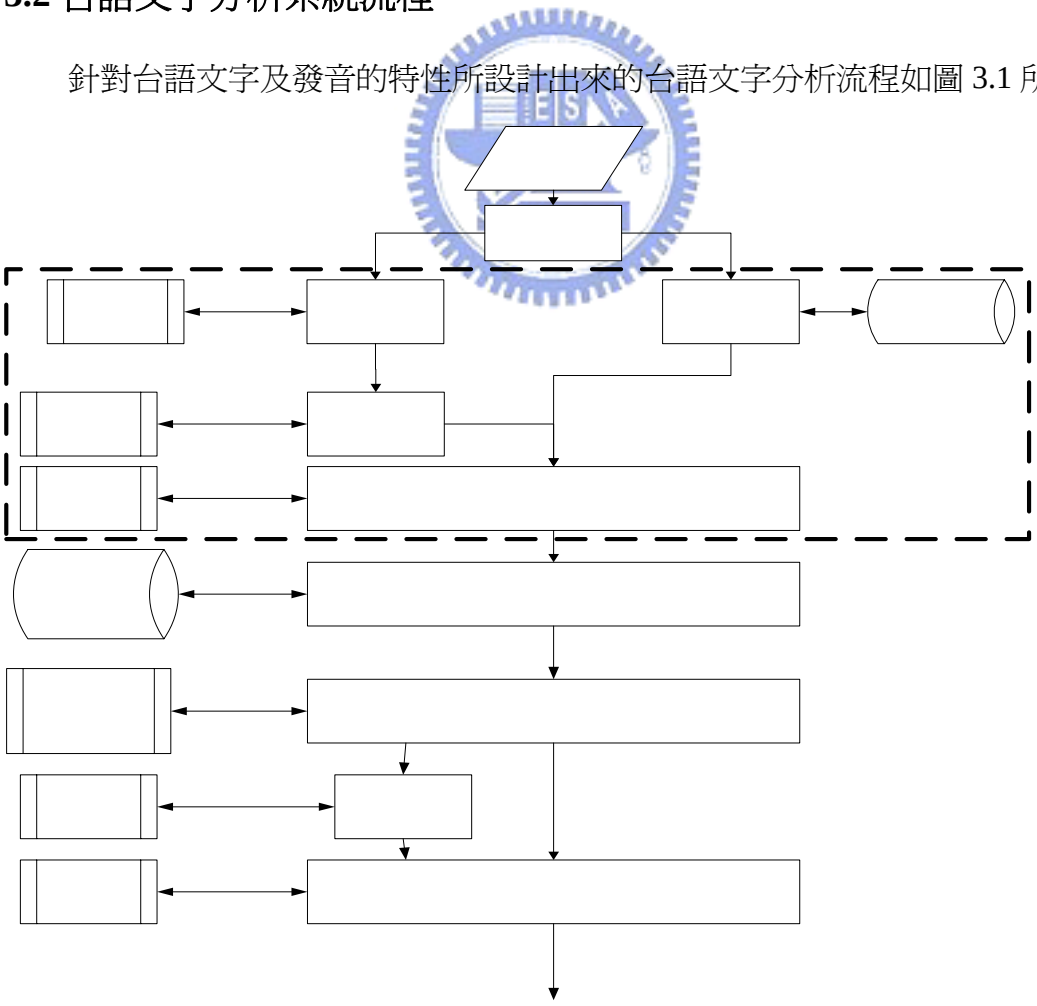

圖 3.1 台語文字分析流程

針對前一小節所做的考量而設計出的台語文字分析流程共包含五個部分，分別爲前處理單元、斷詞單元、詞類標記單元、文字正規化單元、變調處理單元。以下分別簡述五個單元各自的功能：

(1) 前處理單元

台語文字通常爲漢字/拼音字夾雜，此單元將拼音字部份轉換成台語斷詞器內部的 Extended code 格式，以利辭典樹的建立與搜尋所謂 Extended code 就是將台語拼音文字轉成其發音，再借用 BIG5 碼中保留區域來表示這些發音，如此就可以用以解決漢字、拼音夾雜的問題了。另外，此單元也可以處理簡單的拼音連寫現象。

(2) 斷詞單元

此單元主要做辭典查詢以及構詞的動作，依據斷詞規則選取最適當的斷詞結果，並針對數字部份做少量的文白讀音修正。文白讀音不同也是台語中一個獨特的現象，通常發生在姓氏、數字的讀音上。譬如必須對數字串單獨發音的狀況，所使用的讀音爲文讀音，譬如電話號碼、西元的年份等狀況。而在唸一般的數值時使用白讀音，譬如價錢的數字部分等。

構詞單元爲延續中文構詞規則部份，加入部份台語構詞規則，斷詞規則則延用中文斷詞器所使用的六條斷詞規則[11]。此六條規則由中研院詞庫小組所提出，依序爲長詞優先、詞長標準差小者優先、附著語素最小者優先、定量複合詞字數合最小者優先、一字詞詞頻最高者優先，及總詞頻最高者優先等。

(3) 詞類標記單元

使用中文斷詞器中，針對中研院平衡語料庫 3.0 版所做的詞類雙連文統計資料，做爲標示台語斷詞結果中 POS 資訊的依據。

(4) 文字正規化單元

使用中文文字正規化規則，針對發音部份做修改，其中加入了部份的文白讀發音規則。

(5) 變調處理單元

以斷詞結果的詞尾作爲變調邊界，針對詞做變調規則的套用

接著我們將詳細介紹此系統流程裡各單元中的各元件運作原理與設計時所做的考量。

3.3 台語斷詞系統之製作

3.3.1 字典查詢單元

在斷詞單元中，一個完整的辭典是最重要的。在論文中我們做了台語詞庫詞數的擴充、一字詞發音的修正及詞庫資訊的增加等三部分的改進，將分別詳述如下：

(1) 詞庫詞數擴充部分

交大語音實驗室中原有一套台語辭典，主要是由鄭良偉老師整理的台語辭典及中研院的國語辭典合併而成，並以人工方式加入『一分鐘台語單字速成』等書籍內含的詞[12]而成。

針對詞數擴充部分，主要來源爲樟樹出版社出版的『新編華台語對照典』所篩選出來的詞條。此部辭典的特色爲，收錄了很多台語極爲口語化的詞條和一些名詞(植物名等)、中文詞等。口語化詞條譬如『一人一業』、『人看人』等，植物名譬如"七葉蓮"，中文成語譬如"三顧茅廬"等。因爲這些詞條接近台語文章的真實現象，所以此次所擴充的詞條將更能爲台語文句標示出正確的辭彙。交大語音實驗室原始辭典及新擴充辭典的辭彙數比較如表 3.1 所示，我們可以看到新增加的辭彙數約有兩萬字。

表 3.1 擴充後的台語辭典詞數統計

	一字詞	二字詞	三字詞	四字詞	五字詞	六字詞	總數
原始辭典詞數	13914	59233	16671	5074	318	17	95227
新增詞數	0	6703	8601	4040	1260	400	21004
擴充後辭典總詞數	13914	65936	25272	9114	1578	417	116231

(2) 一字詞發音修正部份：

台語斷詞經常斷出一字詞，然而一字詞通常有數種發音，造成選取發音上的困難，所以希望針對一字詞的某個發音在辭典中出現的頻率做統計並分類，依此決定此發音是否爲主要發音。目前台語詞庫中的一字詞部份，大部份發音是由中文一字詞發音直接轉成台語相近音，如此做的結果是，部份的一字詞所擁有的發音並不一定是正確的。以往對於辭典中的一字詞部份並沒有特別針對發音的正確性做校正，使得斷詞結果的一字詞拼音錯誤情況相當嚴重。在此希望能針對一字詞發音部份做簡單的分析和修正，希望能讓一字詞發音正確度更高。

以下列步驟將辭典中的一字詞分類：

(a) 僅一種拼音可能，只出現在一字詞

(b) 僅一種拼音可能，曾出現在多字詞

(c) 多種拼音可能，但此單字在辭典中出現頻率極低

(d) 多種拼音可能，其中一種拼音頻率遠高過其他拼音

(e) 多種拼音可能，有兩種以上的拼音出現頻率很高

以此方式將一字詞分類後，各類字數如表 3.2 所示：

表 3.2 針對一字詞發音情況所做的分類統計

單音 Freq=1	單音 Freq>1	多音，單字出現頻率極低 (Freq<15)	多音，其中一種發音頻率較高	多音，各發音頻率均高	總字數
7791	2295	1392	1400	1036	13028

由表 3.2 可見，台語具有一字多音特性的單字詞約有 3800 個，約佔了總單字詞的 30%，比起國語 510 個破音字[13]還要高出許多，所以一字詞發音選擇的正確與否對台語 TTS 的效能影響十分之大。在論文中我們會對台語辭典中的一字詞發音現象作進一步探討。

由表 3.2 的分類結果可看出，(a)(c)兩類屬於少用字，(b)類屬於僅含單一種發音的字，均不會造成一字詞發音選擇上的問題，只有(d)、(e)兩類單字會造成大部分的一字詞發音選取問題。

針對(d)、(e)兩類一字詞，我們以人工觀察後可大致歸納出以下兩種特性：

(1) 多種念法，但各念法相近，原因大部份爲南北腔調不同。

(2) 單字本身具有文白讀特性。

此兩類單字詞的發音不確定性高，因爲必須用人工確認各發音是否真的適合這個單字詞，且必須考慮文白讀音出現的場合和時機，是個複雜且還無法解決的問題。由之後的斷詞結果分析可知，影響一字詞發音正確性的關鍵字大致發生在此兩類單字上。但腔調及文白讀場合選擇的問題並不在本論文探討的範圍。

目前的做法是，選取在詞庫中出現頻率最高的發音做爲一字詞發音的選取原則。在後面的斷詞結果分析中，會再針對一字詞發音部份做修正的動作。

(3) 詞庫資訊擴充部份：

以往台語詞庫僅含讀音資訊，不若中文詞庫有著詞頻、詞類等資訊。由於台語已做標示斷詞、詞類資訊的文章數量極少，無法爲台語詞統計詞頻，然而卻有機會爲部份台語詞增加詞類標記(POS,Part of Speech)的資訊，所以我們將台語辭典擴充一個欄位用以標示詞目的 POS。

在本論文中，於台語辭典增加詞類標記資訊的方式爲

(a) 利用鄭良偉老師提供的五萬三千詞辭典及『新編華台語對照典』中具有的台語/中文詞對照，由中文辭典中將所對應到的中文詞目之 POS 填入台語辭典中。在這兩部辭典中的台語/中文詞對照資料常有一個台語詞對應多個中文詞的現象。我們將所有對應到的中文詞所擁有的 POS 資訊聯集起來，作爲此台語詞的詞類標記資訊來源。

(b) 直接以台語詞彙查詢中文辭典，若找出相符的詞，則將其 POS 填入台語辭典中。

以此兩步驟大約可得六萬五千筆台語詞的詞類標記資訊，大約占了辭典總詞數的 60%左右。雖然未能將所有詞目加上 POS 資訊，但有 POS 資訊的詞目已可在後面將提到的台語特有夠詞規則加以應用。詞類資訊在未來的台語 TTS 系統上將成爲重要的資訊，有了詞性的資訊，將可以依據詞與詞間詞性的關係連接成更長的詞，不只可以改善台語變調邊界的判定，更能提供韻律訊息產生器更多資訊，以產生更好的韻律。

3.3.2 台語的構詞現象

台語文字常出現特殊的複合詞(compound word)，包含『重疊』、『附加』和『複合』([9],p.157)等構詞規則所產生的詞。此類複合詞具有緊密結合的關係，可獨立形成一個變調組([9]p.153)，所以在斷詞器中不只可依此兩類規則將詞構出，還可對此類構出詞套用變調規則，以更符合台語詞的實際念法。

台語的**『重疊』**構詞規則指的是重複同一個語位兩次或三次，以構造成

詞的辦法([9],p.157)。此處所提及的語位，爲語言裡最小的有意義的單位，且不能再分割成更小的有意義的單位([9],p.150)。台語所提及的『重疊』與中文所提到的『疊詞』意義不同。中文的疊詞所描述的，是此疊詞在文字重疊之前仍然有自己的意義。例如『快快樂樂』由『快樂』衍生而得，而『快樂』本身就具有自己的意義，重疊是爲了加強此形容詞的程度。台語的疊詞在語位重疊前可能有，也可能沒有本身的意義。譬如『冷支支』在第二個語位重疊前爲『冷支』，並沒有自己的意義，而『無理無由』中的『理由』具有本身的意義。

台語的**『附加』**構詞規則指的是詞根和詞頭或詞尾黏合的構詞方式。譬如『阿明』、『厝仔』等本身不屬於一個詞，但是在口語上經常會將『阿』、『仔』等元素併入名詞的前或後，且此類元素在合併之後所成的詞將符合變調規律。

台語的**『複合』**構詞規則指的是兩個或者兩個以上的詞根成分組成合成詞的構詞方式。大部分的複合詞已經成爲普通常見的詞，譬如『心肝』、『兄弟』、『風雨』等，在台語的斷詞系統裡面並不需要特別處理這類詞，僅需經由擴充辭典就可做到正確的斷詞結果。另外一類的複合詞爲『定詞／量詞』組合而成的詞，這類複合詞將使用國語斷詞器中的構詞規則構出。

(1) 台語構詞中的『重疊』、『附加』與『複合』現象

一、重疊：

重疊可分爲九類：AA、AAA、ABB、AAB、AABB、ABAB、ABCC、ABAC、ABCB 等。

(1) AA：可發生在名詞、時間詞、數詞、量詞、形容詞、動詞、副詞上。

名詞：儂儂(人人)、

量詞：日日(每天)、步步(每一步驟)

數詞：一一(一項一項)、萬萬(狀語)

副詞：足足(狀語)、白白(狀語)

形容詞：lo5 lo5(渾濁)

動詞：收收(含有催促之涵義)

(2) AAA：只發生在形容詞。

粉粉粉(誇飾)。譬如用在形容臉蛋很粉嫩的時候。

長長長(誇飾) ，形容物品很長。

(3) ABB：發生在 A 爲形容詞時。

譬如『燒燙燙』、『冷支支』等，將形容詞後面的譬況詞重複。

多數譬況詞看不出和前面的形容詞有密切的語意關聯[9]。有些形容詞可加不同的譬況詞，譬如『白』可後接『蒼蒼』、『sut4 sut4』、『siak4 siak4』、『chin chiN』等。

(4) AAB：發生在 B 爲動詞時。

譬如 『ko ko 纏』(糾纏不清)、『溜溜去』等。重疊的部分主要用來增加生動感，有些和後面的動詞有語意關連，有些看不出。

(5) AABB：

這類『重疊』有些是 AB 本身就是一個詞，有些是 AB 不爲詞。前者譬如『清清楚楚』、『零零星星』，後者譬如『挨挨陣陣』(人群擁擠的樣子)、『七七八八』(顛三倒四的樣子)。

(6) ABAB：AB 通常是形容詞。

譬如『神經神經』、『古意古意』、『老實老實』等。雙音節形容詞幾乎都可以重疊成這一類疊詞([9],p.161)

(7) ABCC：AB 爲雙音節形容詞，性質和 ABB 相同。

例如『清氣溜溜』、『熱鬧滾滾』等。

(8) ABAC：BC 是並列詞，修飾語 A 重疊後，分別修飾 B 和 C。

譬如『無煩無惱』、『無明無朗』、『無冥無日』等。

(9) ABCB：通常 AC 是並列的名詞，且都具有 B 的特性。

譬如『皮癢骨癢』、『喙笑目笑』等。

(10) AABC：譬如『包包起來』。

二、附加：

附加是指詞根和詞頭或詞尾黏合的構詞方式，該合成詞的語意主要由詞根表達，詞頭詞尾只有抽象的語法意義。以下對常出現的詞頭／詞尾做分析。

(1) 詞頭：典型的詞頭爲『阿』字。通常加在單一個字的人名或親屬稱謂前面，譬如『阿英』、『阿輝』等。

(2) 詞尾：典型的詞尾有『仔/a2』、『的／e5』、『le0』、『了／ka／著』，例如：

(a) 『仔/a2』：使用時機有二，一種是加在形體不大的普通名詞後面，有細小的意思，譬如『雞 a2』、『狗 a2』。另一種情況是加在職業後面表示輕賤之意，如『牽交子』、『剃頭仔』等。

(b) 『的／e5』：相當於中文的『的』在詞語中的地位。譬如『狗 e5 耳 a2 真利』、『伊 e5 杉破去矣』等。

(c) 『le0』：加在動詞後面表示動作的持續性。譬如『坐 le0』、『倒 le0』、『跪 le0』、『貼 le0』等。

(d) 『了／ka／著』：接在動作動詞後面，譬如『阿英走了傷緊』(阿英跑得太快了)，『阿英杉收了真整齊』。『了／ka／著』三者的差別僅在語意上的不同。譬如『hit 隻馬走了真緊／hit 隻馬走 ka 真緊／hit 隻馬走著真緊』三句話，使用『了』是在評斷馬跑的結果，使用『ka』是在描述馬跑的狀況，使用『著』是在肯定馬跑的能力。

在經由觀察實際文章的詞尾用法後，可找出常出現於詞尾做爲附加字的集合有：{啊、咧、啦、喔、啥、個、e5、了、仔}等。在實際將此集合使用於詞尾附加性的構詞後，可明顯增加構詞率，且以人工檢查時，並無看到明顯搶詞的情形。

三、複合：

此處所提及的複合，只針對『數量複合詞』的部分作討論。台語的數詞和量詞大部分都可沿用中文的數詞及量詞，但是有部分的詞組是屬於台語特有的，譬如『半冥』、『逐 e5』等。

上述三項台語構詞方式的描述，是以台語實際上的構詞現象做構詞的考量。由於目前所使用的台語辭典僅包含詞類的資訊，所以在實做台語構詞單元時，將以詞類資訊作爲實做上的依據。實作台語構詞單元的方法將在下一小節中詳述。

3.3.3 台語構詞單元的設計

由上一小節所描述的台語構詞方法，我們可以以列出規則的方式套用到實際的系統中。在設計台語構詞單元時可以考慮兩方面的資訊，一方面是針對台語詞的特性所列出的構詞規則，另一方面是繼承中文構詞單元所使用的規則結構。

接下來將描述構詞單元設計時所做的系統架構考量，分爲台語構詞規則部分、中文構詞規則部分、決定變調邊界、決定數字發音四部分討論。

(1) 針對台語詞的特性部分所設計之構詞規則

以上一小節所述的台語構詞現象，針對『重疊』、『附加』和『複合』等現象列出系統實做上所使用的構詞規則。

(a) 重疊詞部分：

依據十條重疊規則出現的時機做爲構詞的限制。目前的做法是以詞類做爲構詞的限制。以下列出十條重疊規則所使用的構詞限制。

(1) AA：所有的詞性皆可(名詞、時間詞、數詞、量詞、形容詞、動詞、副詞等)。

(2) AAA：限 A 爲形容詞。

(3) ABB：限 A 爲形容詞。

(4) AAB：限 B 爲動詞。

(5) AABB：沒有明顯的詞類限制。

(6) ABAB：限 AB 爲雙音節形容詞。

(7) ABCC：限 AB 爲雙音節形容詞。

(8) ABAC：沒有明顯的詞類限制。

(9) ABCB：A、C 各自皆爲名詞。

(10) AABC：A 爲動詞。

(b) 附加詞部分：

針對詞頭和詞尾部分作構詞上的限制。以下分爲詞頭、詞尾附加兩部分做討論。

(1) 詞頭附加：限制後接詞爲單音節。譬如『阿明』、『阿花』等。

(2) 詞尾附加：前接名詞、代名詞、動詞和形容詞。

譬如『狗 a2』、『伊 e5』、『倒 le0』、『收了』、『狡猾 e5』等。此類構成詞的變調邊界不一定在這些詞尾處，譬如『狡猾 e5 儂無歡迎』的變調邊界應該在『儂』而不是在 e5。

(c) 複合詞部分：

針對台語特有的數詞和量詞做收集，譬如 tak8(逐)、e5(個)等，併入中文構詞中數量複合詞的集合裡。

(2) 中文構詞規則部分

由附錄一的中文構詞規則及相關例子，標注合用的規則直接套用。一種規則代表一種構詞結構，可發現大部份的結構都可以直接讓台語使用。

由中文構詞規則拿來使用的部份，可以分爲定詞、量詞集合與構詞歸則兩部份討論。

定詞、量詞集合及部份範例如表 3.3 所示，構詞規則部分則於表 3.4：

表 3.3 定詞量詞種類與集合例

定詞、量詞種類	**定詞、量詞集合例**
數詞定詞	一,二,兩,三,四,五,甲,乙,丙,丁,戊
特殊數量定詞	半,多,許,整,正
全體數量定詞	全,滿,整,成,一切,所有
疑問數量定詞	多少,若干,幾多
前程度副詞	很,挺,怪,真,好,極,滿,更,再
程度數量定詞	一些,多,許多,部份
部分數量定詞	半,若干,有的
指示定詞	這,那,哪,這些,那些,哪些
特指定詞	上,下,前,後,不到, ,左右,不等
時間名詞	早上,晚上,上午,中午,下午,週一,週二
個體量詞	行,戶,件,家,架,卷,具,節,款,客,輛,粒,片
群體量詞	對,雙,宗,番,畦,餐,行,身,列,系列,排
部分量詞	團,堆,泡,綹,撮,把,股,灘, 部分,坨,匹,疋
容器量詞	匣,箱,櫃,櫥,籃,簍,包,袋
準量詞	弄,段,號,地,街,樓,學期,下子,版,冊
動量詞	度,輪,回,次,遍,趟,下,口,刀,仞,槌
暫時量詞	身,頭,臉,鼻子,嘴,肚子,手,腳,桌子,院子
標準量詞	公寸,公分,公尺,公克,公斤角,毛,元

表 3.4 構詞規則與部分範例

基本的複合定詞	一百多萬,一百多,一點三
基本的定量複合詞	三大個,四小半個,五個,一公升
全體數量定詞與定量複合詞再結合	整個,整整一百個,
疑問數量定詞與量詞結合	若干個之多,若干個
程度數量定詞與定量複合詞再結合	好幾十個,好幾個,那麼多個,好幾百
部分數量定詞與定量複合詞再結合	數百個之多,數個,數十餘萬
特指定詞與定量複合詞再結合	本項,貴班,
表示日期或時間	一個多月,一分十五秒三
表座標	一度一分一秒
表地址	一段一號
表溫度	攝氏五度
其他	一百個左右,不到一百個

由定詞、量詞集合及構詞歸則中可以發現，部份定詞、量詞集合所構出的詞並不適用於台語的情況，譬如台語並沒有『若干個之多』這類的講法。即使如此，目前仍然以直接使用此類集合並針對定詞、量詞做少量的擴充爲主，並不將中文特性強的定詞、量詞集合及對應的構詞規則做刪減。理由是，由觀察部份台語文章的書寫法可知，有些台語文章傾向於使用中文文章的筆法書寫，部份使用中文詞以取代真正符合台語念法的台語詞。面對此種狀況，目前的做法爲直接爲此類中文詞以逐字發音的方式標上台語發音，日後可以考慮針對此種情況，設計出較好的中文詞轉換成台語發音的機制。

(3)變調邊界之決定

台語變調規律[7]事實上是十分複雜的，在我們的系統中，主要還是以一個詞作爲一個變調單元，但是對於使用構詞規則所建立的詞之變調邊界的決定則有經過特別的處理程序。

由構詞規則建構出的複合詞，它們的變調規則會有以下三種可能性：

(a) 完全不套用變調規則，譬如數字串。

(b) 直接以構詞後的詞尾做爲變調邊界，對詞直接套用構詞規則。

(c) 變調邊界不僅存在於所構出詞的詞尾，且存在於構出詞的詞中間。此情況發生於當構詞規則爲較小的子規則結合而成，而在子規則部份已經決定變調邊界時。以下舉出兩個構詞中決定變調邊界的例子

(1) 規則 268：IN1 {月} (IN1 ({日,號}))，代表的是『數字+月+數字+日(號)』，如五月三號。此規則的變調邊界標記爲 0，表示詞的中間存在變調邊界。而此規則又是由子規則 264：{IN1,元} {月,月份}，和 265：{IN1} {日,號}組合而成，此兩規則的變調邊界標記爲 1，表示由此兩個規則已經可以決定變調邊界，規則 268 不需要再考慮變調邊界的問題。

(2) 『一百多萬』符合規則 200，其變調邊界在『萬』，然而又有可能和『個』形成規則 214 的例子『一百多萬個』，此時變調邊界將從『萬』移到『個』。此類邊界會因爲後接字而改變的情形可分爲兩種狀況討論，一種是本身就是自己的子規則(只有規則 199 一個例子)需要另外處理，另一種情況就是變調邊界可以在確定爲某個規則以後決定。譬如若構詞結果符合規則 200，則直接定義詞尾爲變調邊界。如果能再構出較上層規則 214 的詞，則可重新定義新構出詞的詞尾爲變調邊界。如果繼續構詞過程發現還可以再往上層規則，則因爲在規則 214 時已經決定變調邊界，繼續構詞的結果並不會影響變調邊界的界定。以此調整變調邊界的好處在於，利用構詞的過程就可以決定較正確的變調邊界，且一旦決定了就不會受到更高層規則的影響，如此在遇到構詞層次非常高的情況時，也不用擔心變調邊界如何決定的問題。

(4) 由構詞規則決定數字部份的文/白讀發音

數字發音最麻煩的部分就是文/白讀音的選用。以往並沒有較好的機制選擇正確的發音，當構詞單元加入台語斷詞系統後，有機會將數字發音部分做處理。

數字發音有以下三種可能的情況：

(a) 所有數字都念成文讀音。譬如規則 263 例『西元一九八一年』的數字部份。

(b) 所有數字都念成白讀音。譬如規則 237 例『一年一班』，不管使用哪個數字都用白讀音。

(c) 可能白讀也可能文讀，但是有一定的規律。

譬如規則 264 例『一月』使用文讀，其他月份使用白讀。規則 199 例『一百二十一』中，只有接在『十』後面的『一』念文讀，接其他數字均念白讀音，其餘數字也都是白讀音。可以發現例外都發生在『一』，可針對此情況特別處理。

文/白讀機制跟著構詞規則的好處是，不需要一開始就將各種可能情況處理，只要將情況歸類，針對某些類別使用適合的發音，剩下少數無法簡單歸類的情況則可再單獨處理，簡化了處理的手續，且更能精確解決特殊情況。

3.3.4 由候選詞組決定詞單元

在圖 3.1 斷詞器架構中可以看到，當我們在做辭典查詢和構詞之後，可以獲得許多候選詞組。後選詞組的建立方式爲，由輸入字串的第一個字開始，連續組成三個詞，成爲一個候選詞組。以此方式將所有可能候選詞組找出來。

譬如當我們輸入『有一張桌子』時，會產生以下三個候選詞組：

(a) 有 一 張

(b) 有 一張 桌

(c) 有 一張 桌子

接著我們要由這三組候選詞組中找出最佳的一組詞組組合。中研院詞庫處理小組所發展的國語斷詞器中已定義出一套十分有效的選詞規則，在我們的台語斷詞器中也是沿用這些規則，但在去除與詞頻相關部份。

下面將中研院詞庫處理小組所定義的六條斷詞規則簡述如下：

Rule 1: 長詞優先：以候選詞組中詞長總合最大的詞組做爲選詞的結果。

(一個候選詞組最多由 3 個詞組成)

Rule 2: 詞長標準差少優先：選擇候選詞組中各詞詞長的標準差最小的詞組。

Rule 3: 附著語素少者優先：附著語素的定義，在中文斷詞器裡是以『很少單獨出現在語句中的一字詞』爲標準。

Rule 4: 定量複合詞字數合最少者優先

Rule 5: 一字詞詞頻最高者優先

Rule 6: 總詞頻最高者優先

由上述六條規則可知，斷詞規則需要用到的資訊：詞長、詞頻、構詞資訊。其中 Rule 1、2 使用詞長資訊、Rule 3、5、6 使用詞頻資訊，Rule 4 使用構詞資訊。

由於台語辭典並沒有自己的詞頻，現階段先暫時不引用 Rule 3、5、6 ，而僅以 Rule1、2、4 做爲選擇候選詞組的依據。

3.3.5 詞類標記單元

由江振宇學長[1]所做的『中文詞類雙連文統計資料』做爲詞組中詞性的決定依據。直接使用中文的統計是因爲台語詞性結構與中文相近，且台語目前無法自行做這項統計(沒有已標注詞類的語料)。

目前台語詞庫約有 60%的詞擁有詞類資訊，可做爲部份構詞規則的依據。

3.3.6 文字正規化單元

在中文的文字正規化部份，針對了以下幾種情形做文字正規化動作：

(a) 英文字母部分：ＴＥＬ、ＦＡＸ、AM、PM、12cm、10kg 等。

(b) 符號部分：90%、2005/4/15、12.5 等。

(c) 數字加單位部分：100 公斤、1 月 1 日、2005 年等。

針對以上情形定出數字、符號的發音方式及發音的順序。

在台語文字正規化部分，可直接套用中文的文字正規化規則，大部分的台語發音也可以直接填入，只有在數字的發音部分較爲麻煩，牽涉到數字的文讀/白讀發音及變調。

文白讀發音部份，可利用構詞規則中的定義處理。

變調部份，可分爲全部不變調和全部變調兩種。前者譬如電話號碼，後者譬如『100 公斤』等。將文字正規化後的情形分成這兩類處理。

3.3.7 變調單元

依據斷詞/構詞結果套用變調規則。

變調規則如黃競億所整理描述的原則，略述如下

(a) 詞尾不變調，詞裡其餘字均需變調。

(b) 非入聲音變調規則：1→7→3→2→1，5→3(北部)或者 5→7(南部)

(c) 入聲音(p,t,k,h 結尾音)變調規則：

p,t,k 結尾：4→8→4

h 結尾：4→2，8→3

(d) 『仔』的前一個音節變調後的再變調：若變調後爲 2 或 3，則再變調成 1 或 7。

每個台語單字都有屬於自己的本調音和變調音。通常變調會發生在某單字和別的單字接成詞以後。譬如『電』『風』本調音分別爲"tian7"、"hong"，但是在組成『電風』一詞時，"tian7"將變調爲"tian**3**"，而"hong"位於詞尾不需變調，使得讀音變成"tian**3**-hong"。

台語變調規則必須套入所有的詞的發音。此處所指的詞包含詞庫的詞、構詞形成的詞、文字正規化後所成的詞等。構詞前後變調的情形會不同，譬如『三個』可和『人』構成『三個人』一詞，前者的本調音和變調音(讀音)分別爲" saN-e5"、" saN**7**-e5"，後者則分別爲 " saN-e5-lang5"和" saN**7**-e**3**-lang5"。造成此變調結果不同的原因，在於變調規則會套用到一個詞除了詞尾的所有音上面，所以變調規則必須在構詞結束之後再套入。

目前詞庫內詞的拼音皆爲本調，在斷詞/構詞/文字正規化之後的詞拼音將維持本調。我們可以在前面的步驟做完之後加入變調規則做音調的修正。

第四章 台語文字分析器的效能分析

在上一章中，我們已詳細考慮台語文字的各種現象，但受限於台語辭典資料難以取得，所以我們已經現階段可以實現的各項文字規則加入我們所製作的台語斷詞器中。在這一章中，我們要評估改進後之台語斷詞器的效能。也受限於無法大量獲得有正確斷詞標示的台於文章，所以我們只能以舊有的文字分析單元和目前的文字分析單元所得到的斷詞結果做比對，以定性和定量兩種方式做效能的分析，以說明此系統所達到的改善效果。其中定性的方式，就是以舉例的方式，針對台語文章所出現的各種現象，以舊有處理方式和目前處理方式所得到的結果，說明目前系統在面對各種現象所能達到的改進。而定量方式則是針對大量台語文章做斷詞分析，以評估斷詞結果、發音正確度等。

4.1 以定性方式做效能分析

本節將特別針對台語構詞的部份做斷詞結果分析，以實際例子判斷是否保留此條規則。保留與否的原則爲，此規則是否會造成無法彌補的搶詞錯誤。

如上一章所述，構詞分爲兩部份，一部份爲沿用中文構詞規則(擁有固定數量的定詞、量詞集合)，另一部份爲配合台語文章特性所做的構詞(沒有特定的集合，單以詞性等做爲構詞的限制)。前者由於有固定的集合，造成錯誤構詞的機會較小，而後者因爲不受到有限集合數的限制，僅以詞性或者字數做爲限制，搶詞的情況較前者嚴重。

接著將對各個構詞規則做斷詞結果分析：

(1) 詞頭附加

例句：『乖乖及阿德是兩隻流浪狗』

不使用構詞：『乖 乖 及 阿 德 是 兩隻 流浪 狗』

使用構詞： 『**乖乖及 阿德** 是 兩隻 流浪 狗』(搶詞發生在『乖乖及』)

使用構詞(不使用 AAB)：『乖乖 及 阿德 是 兩隻 流浪 狗』(較正確)

可以發現此句話有兩個地方使用到構詞規則，一個是重疊詞規則(AAB)，另一個是詞頭附加規則(阿+一字詞)。

重疊詞部份，可發現使用 AAB 規則所造成的搶詞，是非常嚴重的搶詞情況。發生此類搶詞的原因爲，AAB 構詞規則只使用 B 的詞類(限制 B 爲動詞)做爲限制的依據，然而在『及』所含有的詞性裡就包含了動詞，造成了搶詞。同樣的情況也發生在 ABB 構詞規則，將在下一小節以例子做討論。

詞頭附加部份，觀察大量斷詞結果可發現，在此類詞頭附加部份擁有很高的構詞正確率，其中一個原因是在實做中，詞頭附加是以斷詞結果出來後，判斷『阿』的後接詞是否爲單字詞，若是，則將之構出。以此方式可較符合詞頭附加的原則，且以人工確認斷詞結果時也發現，並沒有因爲加入了此構詞規則而造成任何搶詞的情況。

(2) 詞尾附加

詞尾附加的原則爲『詞尾附加集合與前接詞構成詞』，做法如詞頭附加，將斷詞結果符合詞尾附加條件的兩個詞做構詞的動作。

詞尾集合如下：{啊、咧、啦、喔、唅、阿、個、e5、了、仔}，以下以實際例子說明詞尾附加造成的影響。

例句：『早暗啊斟茶、燒香啊』

斷詞結果(不使用構詞)：『早暗 啊 斟茶 、 燒香 啊』

斷詞結果(使用構詞) ：『**早暗啊** 斟茶 、 **燒香啊**』

可以發現採取詞尾構詞的動作，可以成功地構出正確的詞。

例句：『厝邊仔的腳步聲嘛愛認啊！』

斷詞結果(不使用構詞)：『厝邊 仔 的 腳步 聲 嘛 愛認 啊 ！』

斷詞結果(使用構詞)： 『**厝邊仔** 的 腳步 聲 嘛 **愛認啊** ！』

此例句如同上一句，只使用了詞尾附加，可成功構成詞。

例句：『一粒種子，會當發做青菜、稻仔、麥仔、果子抑是樹仔。』

斷詞結果(不使用構詞)：『一 粒 種子，會當 發做 青菜、稻 仔、麥仔、果子 抑是 樹 仔。』

斷詞結果(使用構詞)： 『一粒 種子，會當 發做 青菜、**稻仔**、麥仔、果子抑是 **樹仔**。』

此例句可以將『仔』這個常用於描寫微小物品的情況展示出來。辭典無法將所有可能的此類情況列出，譬如辭典內沒有『樹仔』這個詞，必須用構詞方法構出。

例句：『我看著灶腳有一隻鳥鼠仔』

斷詞結果(不使用構詞)：『我 看著 灶腳 有 一 隻 鳥鼠 仔』

斷詞結果(使用構詞) ：『我 看著 灶腳 有 一隻 鳥鼠仔』

此例句使用了兩個構詞規則，一是數量複合詞(一隻)，另一個是詞尾附加(鳥鼠仔)。由大量斷詞結果的人工確認可知，詞尾附加可達到很好的效果，原因是此類詞尾附加字在台語文章裡出現的頻率極高，幾乎每兩句話就會出現一次，且通常此類附加字不容易與後接字形成一個詞，不易造成搶詞的問題，所以通常都可順利與前接詞構成一個詞。以下列出數例，可突顯此類附加字出現的頻率的確是非常高的：『自從啊』、『安呢啦』、『就著啦』、『囝仔的』、『平常時啊』、『兄弟仔』、『水珠仔』…等。符合此類構詞原則的詞非常多，且在不加入構詞規則時均無法構成一個詞。

(3) 疊詞

(a) AAA 規則：

例句：『將阿德管甲乖乖乖』

斷詞結果(未使用構詞)：『將 阿 德 管甲 乖 乖 乖』

斷詞結果(使用構詞)： 『將 阿德 管甲 乖乖乖』

此句話使用到兩個構詞規則，一個是詞頭附加，另一個是疊詞(AAA)。由人工觀察此類構詞方式所看到的斷詞結果，發現此構詞規則效果非常好，因爲在連續三個相同字出現時，的確就是屬於此類疊詞。

例句：『嘴仔開開開，就共阿德踢一下』

斷詞結果(未使用構詞)：『嘴 仔 開 開 開，就 共 阿 德 踢 一下』

斷詞結果(使用構詞)：『**嘴仔 開開開**，就 共 **阿德** 踢 一下』

此句話使用詞頭、詞尾附加、疊詞等構詞規則。可發現 AAA 在很多文章中幾乎是可獨立成詞，而不會與前後字造成搶詞的現象。

例句：『四箍輦轉攏白白白、平平平、綿綿綿，清氣閣安祥』

不加構詞：『四 箍 輦 轉 攏 白白 白、平平 平、綿 綿 綿，清氣 閣 安祥』

加入構詞：『四 箍 輦 轉 攏 **白白白**、**平平平**、**綿綿綿**，清氣 閣 安祥』

此例僅使用了構詞規則 AAA，但可達到極佳的效果。

例句：『空空空啊就好好好卜拼逐 e 來拼安呢啦。』

不加構詞：『空空 空 啊 就 好好 好 卜 拼 逐 e 來 拼 安呢 啦。』

加入構詞：『**空空空啊** 就 **好好好** 卜 拼 **逐 e** 來 拼 **安呢啦**。』

此例使用了詞尾附加、AAA 兩個規則。『空空』和『好好』是辭典內擁有的詞，經由構詞後可形成更長的詞。

(b) AAB、ABB 規則：

例句：『恬恬仔將這個世界變做白 siak4 siak4』

斷詞結果(不加構詞)：『恬 恬 仔 將 這個 世界 變做 白 siak4 siak4』

斷詞結果(加入構詞)：『**恬恬仔** 將 這個 世界 變做 **白 siak4 siak4**』

此句話使用了 AAB 與 ABB 兩個構詞規則，可成功構成詞。

例句：**『忽然感覺厝直直搖』**

斷詞結果(不加構詞)：『忽然 感覺 厝 直直 搖』

斷詞結果(加入構詞)：『忽然 感覺 厝 直直搖』

『直直』爲原本辭典內擁有的詞，可依構詞規則構成『直直搖』。

然而，除了在成功構成詞的例子，這兩類構詞規則也造成嚴重的搶詞。譬如：

例句：『乖乖及阿德是兩隻流浪狗』

此處會將『乖乖及』構成一個詞，原因是『及』本身就有詞性爲動詞的可能，而 AAB 規則僅限制 B 符合動詞即可，造成了搶詞。

此類搶詞出現次數頻繁，例如『食乎空空』被斷成『食 乎空空』而非正確斷詞『食 乎 空空』，造成搶詞。另外，『媽媽牽我落車』會被斷成『媽媽牽 我 落車』、『我共爸爸講』被斷成『我 共 爸爸講』，均造成搶詞。

由於我們不希望經過構詞後造成無法挽回的搶詞，且此類疊詞可以經過日後的收集等方式得到，所以將這兩類構詞規則去除(雖然構詞效果很好)。

(c) AABC 規則：

例句：『樹枝當作愛人仝款共 in7 包包起來』

不加構詞：『樹枝 當作 愛人 仝款 共 in7 包 包 起來』

加入構詞：『樹枝 當作 愛人 仝款 共 in7 包包起來』

例句：『我嘛愛看雪將大地當作寶貝仝款共 In7 蓋蓋起來』

不加構詞：『我 嘛 愛看 雪 將 大地 當作 寶貝 仝款 共 In7 蓋 蓋 起來』

加入構詞：『我 嘛 愛看 雪 將 大地 當作 寶貝 仝款 共 In7 蓋蓋起來』

此類構詞規則可套用的例句，數量較少，但可看出仍然有構成詞的可能性。所做的構詞規則限制僅爲 A 爲動詞。

(d) ABAB 規則：

例句：『死去的家庭足濟足濟，有夠可憐。』

不加構詞：『死去 的 家庭 足 濟 足 濟，有夠 可憐。』

加入構詞：『死去的 家庭 足 濟 足 濟，有夠 可憐。』

加入構詞(辭典加入『足濟』一詞後)：『死去的 家庭 足濟足濟，有夠 可憐。』

此類構詞規則所做的構詞限制爲，AB 必須是一個詞。所以當詞典內沒有 AB 這個詞，便無法構出 ABAB 這個詞。

在這一小節中，我們展示了部分構詞規則所造成的斷詞結果，並簡短評估斷詞結果的優缺點。在下一小節中，我們將針對這些優缺點選出適合的台語構詞規則。

4.2 斷詞效果分析與構詞方法的修正

由上一小節的分析，可看出僅以詞性做爲構詞的依據是較危險的事情。中文構詞規則部份使用了有限的定詞和量詞集合，將構成詞的可能性縮小到有限度的組合數以內，以此方式所做的構詞，將有較少的搶詞現象。

在台語構詞規則裡，雖然將規則全數套用可以達到很好的構詞效果，卻也造成大量的搶詞，原因在於所依循的規則可套用到所有的字，並沒有一個方法將可套用字的集合縮小。

基於以上的理由，在台語構詞規則的選用上，以人工確認斷詞結果的經驗與構詞方法的可取代性，將適用的構詞方式列下：

詞頭附加語：{阿}+一字詞

詞尾附加語：集合{啊、咧、啦、喔、啥、個、e5、了、仔}。

重疊詞：套用 AA、AAA、ABAB、AABB、AABC 等五類。此五類經過斷詞結果評估後，有很高的構詞率。

4.3 修改構詞規則後的斷詞結果之定量分析

由上一小節的分析，我們將構詞規則可靠的部份加入台語斷詞器中，做大量文章的斷詞處理，以定量的方式分析斷詞結果。由於沒有大量的已標示斷詞結果之文字資料，所以在此我們希望針對下列三點做分析：

(1) 降低一字詞率：希望能藉由構詞等方式，降低斷詞後一字詞出現的比率。台語斷詞結果若出現過多一字詞，可能造成的影響爲發音無法確定、無法正確判斷變調邊界、韻律錯誤等問題。

(2) 降低音節錯誤率：針對一字詞、多字詞的音節正確率做分析，以人工檢視的方式，觀察自動標記拼音的正確性，分析錯誤來源以及找出可能改善的方法。

(3) 增加平均斷詞詞長：觀察加入構詞規則後，平均詞長是否有改善。

由於台語並沒有大量正確斷詞結果的文章，使得很多分析只能以人工做判斷，無法統計斷詞正確率等數據。不過經由人工分析後，可得到很多寶貴的改善方法和經驗。

4.3.1 斷詞結果分析的資料來源：

在台語語音資料庫中有標音資料，因其標音方式是以人工完成，所以可以認爲是正確對應台語文字的讀音，所以我們使用台語語音資料庫所使用的文章對應拼音做爲斷詞結果分析的對象。

此文字對應拼音資料庫有以下缺點：

(1) 文字和拼音字數不一致：由於當初在做音檔的 transcription 時，只有針對音檔部份做聽寫，而文字檔部份，只有部份有針對文字和拼音的對應做校正，導致文章中常出現有此字卻沒有對應拼音，或者有拼音卻沒有字，或者漏掉幾個字或幾個拼音的情況。

(2) 文章內容並不是原始型式：由於部份文章有經過人工修改(以對應音檔內容)，所以在部份內容上並沒有保留原始的型式，而是修改成可以逐字念出的型式。這代表著這些文章並不需要做文字正規化的動作，無法評估斷詞器中文字正規化單元之效能對。

由於此文字資料庫並沒有經過人工仔細確認過，在使用這些資料時常會造成分析的困擾。因爲此資料庫數量龐大，很難短時間內用人工確認音節拼音是否和文字部份對齊，所以用以下步驟將有問題的文字檔案挑除：

(1) 挑掉字數與拼音數不穩合者：拼音數和文字個數不穩合，必需經過人工檢視拼音在哪個地方多了(或者少了)音節，檢查的過程非常耗時，暫時不使用這類檔案做分析。

(2) 檢視斷詞後拼音錯誤率過高的檔案：在針對剩餘文章做斷詞分析時，發現有部份文章的拼音錯誤率極高(50%以上)，以初步斷詞結果分析的平均錯誤率30%而言，此類文章的錯誤率過高了些。以人工檢視此類檔案，可發現大部份檔案雖然拼音數和字數相符，但是拼音和文字的對應卻在文章中的某些句子以後卻完全沒有對齊到，表示其中有缺字或者缺拼音的現象。以人工修正或者人工挑掉此類文章，剩下 396 篇文字/拼音對照檔，共有 28225 個音節個數可供使用。

4.3.2 斷詞結果分析：分爲三部份討論，第一部份爲一字詞拼音錯誤情況分析，第二部份爲多字詞拼音錯誤情況分析。

[一] 一字詞拼音錯誤情況分析

分析步驟：分爲拼音正確率和斷詞平均詞長兩部份做分析

(1) 拼音正確率部份：在做初步斷詞分析時，發現拼音錯誤率高達 29%，將近三分之一的拼音錯誤。以下爲各數據的統計量：

表 4.1 斷詞結果錯誤率統計(修正一字詞發音前)

音節數	音節錯誤個數	音節錯誤率	多字詞個數	一字詞個數	一字詞錯誤數	一字詞錯誤率
28225	8208	29.08%	8792	8757	3832	43.76%

我們以人工分析一字詞發音錯誤，可將錯誤情形分爲以下幾類：

(a) 辭典本身標音錯誤：譬如『雪』在辭典內的標音爲”sap4”，但是經過一字詞錯誤分析後，發現所有的『雪』都使用拼音”seh4”，可以依此確定此類錯誤原因爲辭典本身的錯誤造成的。

(b) 未正確斷詞：台語單字發音會因爲是否正確斷詞造成發音上的不同。譬如『m 但』的發音爲”m7 na7”，『但是』的發音爲”tan7 si7“，兩個『但』發音不同。如果沒有正確斷詞，使得『但』單獨出現時，就會面臨不知道該選則哪個發音的現象。例如『但是啊』的斷詞結果爲『但』『是啊』，原因爲辭典裡擁有『但是』和『是啊』兩個台語詞。然而原先辭典內『但』的發音爲”na7”，使得斷詞結果的『但』發音錯誤。

(c) 使用場合不同：有些字屬於較活躍的字，在作爲一字詞使用時擁有多種發音。譬如『食』擁有”sit8”和”chiah7”兩種常用發音，在口語上傾向於使用後者的發音，辭典內使用前者。另外，『住』有兩種發音”chu7”和”toa2”，前者發生在多字詞的發音上，譬如『住戶』”chu7 hou7”、『住址』“chu7 chi2”等，辭典所使用的拼音爲此。後者發生在當『住』爲單音節動詞時。此類一

字詞有自己的固定念法，必須針對此拼音做修正。

(d) 文/白讀音的問題：譬如『人』有”lin5 和 lang5”兩種發音，前者爲文讀音，後者爲白讀音。目前沒有很好的機制處理文白讀音的問題。

(e) 文章標音上的錯誤：此類錯誤不屬於斷詞器所造成的錯誤。在檢視文章拼音的正確性時可發現部份人工標音並不正確，或者當輸入文字爲拼音文字時，標上的拼音卻與拼音文字本身不同(譬如將文字”kap4”標成”kah8”)。此類錯誤不該列入斷詞拼音錯誤的統計裡。

由表 4.1 可以發現，以詞數而言，一字詞幾乎佔了所有斷詞結果的一半左右，造成平均斷詞後詞長爲 1.6083 音節/詞，相對於國語斷詞器斷詞後之平均詞長 1.715 音節/詞低了些[1]。由統計又可發現，一字詞錯誤率(一字詞錯誤數/一字詞個數)約達 44%，可以發現只要斷詞結果爲一字詞，大約將近一半的發音都會錯誤，比率非常高。我們希望從修正一字詞發音下手，以改善一字詞發音錯誤率過高的問題。

由觀察得知，大部份的一字詞都能獨立成詞，無法經過構詞構成長詞。且除了文白讀類型的單字外，大部份一字詞發音都傾向於一種或者兩種發音，少部份字擁有多種發音。以此假設下，統計一字詞在文章中的發音情況，修正辭典一字詞(包含構詞的一字詞部份)發音，並觀察音節錯誤率如表 4.2：

表 4.2 斷詞結果錯誤率統計(修正一字詞發音後)

音節數	音節錯誤個數	音節錯誤率	一字詞個數	一字詞錯誤數	一字詞錯誤率	一字詞錯誤佔總錯誤比例
28225	5744	19.02%	8757	1274	14.55%	22.18%

可以發現，經過修正一字詞發音後，整體錯誤率已大幅下降，表示在大量文章中，一字詞發音並沒有如以往所預期的那樣混亂，大部份的一字詞發音是可以被正確決定的。

(2) 平均詞長提昇部份：分爲加入/不加入構詞規則的統計數據如表 4.3：

表 4.3 構詞所造成的平均詞長提升現象

	總音節數	總詞數	平均詞長	一字詞數	多字詞數	一字詞率
不加構詞	28225	19250	1.4662	11230	8020	58.34%
加入構詞	28225	17549	1.6084	8757	8792	49.90%

可以發現加入構詞規則後，的確有增加構成詞的機會，大幅降低一字詞出現的次數。

[二] 多字詞拼音錯誤情況分析：

以下將針對拼音錯誤情況做討論。

由表 4.2 的數據來看，多字詞拼音錯誤佔了總錯誤音節數的 77.82%，以整體來看，大約造成 14%的錯誤率，非常高。然而仔細分析後可發現，這些錯誤並不是真正的發音錯誤，絕大多數都是相近音，只有少數錯誤是真正的發音錯誤。

譬如『無像』自動標音結果爲”bo3 siang5”，人工標音結果爲”bou7 seng5”，兩者發音接近，卻被判斷成兩個發音都錯誤。台語發音較中文發音難以掌握，相近音非常普遍，且音與音之間的差別有時近到難以區分。譬如”bo”和”bou”一般人很難將之區分開來，造成人工標音時有些微的誤差，而這些誤差造成錯誤率大幅上升。另外一種可能是，詞本身就有幾種可行的念法。譬如『人人』就有”lin3 lin5”和”lang3 lang5”兩種發音，前者爲自動標音結果，後者爲人工標示結果，兩者發音都是正確的。

除去拼音相近部份，還有些屬於多字詞拼音錯誤的情況列舉如下：

[1] 字義上的錯誤：譬如『一樣』這個詞，自動斷詞標音爲”chit4 iuN7”，爲

一個定量複合詞。人工標音爲”kang3 khoan2”，爲『相同』之意。兩者由於語意上的不同，使得發音差異很大。

[2] 斷詞錯誤：例如短句『三十三到三十四年』的斷詞結果爲『三十』、『三到』、『三十四年』，自動標音結果爲“saN7 chap8”、” sam7-to3”、” saN7-chap4-si2-ni5”，其中錯誤的爲『到』的拼音，正確拼音應該爲”kau2”，但是因爲辭典有『三到』這個詞，所以在斷詞錯誤的情況下，造成拼音的錯誤。

[3] 文白讀錯誤：譬如部份姓氏使用文讀音而非白讀音，此類情況目前並沒有方法可以有效處理。

4.3.3 斷詞結果分析後之構詞規則的改進：

經改進過之台語斷詞系統之標音錯誤率還是遠高於國語斷詞器，雖然這是台語一字多音特性之必然結果，但我們仍認爲未來還可以在構詞規則上面做改進，以達到更高的構詞率和更正確的拼音。

由統計結果來看，加入構詞機制後，一字詞出現的頻率仍然過高(占所有斷詞結果的 50%左右)，可見以傳統的構詞方式並沒有辦法達到我們希望的效果。以語音合成的角度而言，過多的一字詞，即使拼音皆正確，也會面臨韻律不正確的問題，所以在檢討拼音正確率的同時，還是得思考是否能用其他方式將零碎的一字詞盡量組合成長詞。

觀察斷詞結果，可看出連續一字詞可能出現在以下的情況

(1) 俚語：譬如『食拔仔放統子』等俚語，斷詞結果爲『食 拔仔 放 統 子』，可觀察到一字詞現象非常明顯。台語的俚語講法通常很固定，且以韻律的角度而言，俚語通常需要有韻律連貫的感覺，如果拼音皆正確，卻沒有做到韻律上的連貫，則聽來會很不自然。

(2) 修飾語：譬如形容詞、副詞等修飾語，常被斷成一字詞。譬如『真好食』被斷成『真』、『好食』，其中的『真』屬於修飾『好食』的副詞；或者像『這』、『那』等修飾語，以韻律角度而言，應該與後面的詞做構詞的動作，譬如『這老師』被斷成『這 老師』，看起來將『這』與後接詞『老師』構成詞會有比較好的韻律效果。

(3) 中文詞：有些中文詞被直接拿來當做台語詞用，譬如『這句話』被斷成三個一字詞，『毛色』被斷成兩個一字詞等。

(4) 未收錄詞：可以以文章斷詞結果，配合人工檢視，找出目前辭典內沒有但可能可以構成詞的詞。譬如『豆腐湯』被斷成『豆腐 湯』，但是『豆腐湯』應該能夠成爲一個詞。

未來可以針對這四點原則，做構詞規則與詞庫上的擴充，相信可以再增加斷詞的正確性，以提升斷詞後平均詞長，進而提升發音正確性、韻律流暢度以及音調的正確度。

4.3.4 變調邊界的評估

由斷詞後平均詞長 1.6 來看，做變調邊界的正確率統計是沒有意義的事情。難以做分析的原因有二：

(1) 文章本身的音調標記並非全然正確。

台語的音調很難掌握，文章內的拼音雖然由台語基礎良好的工讀生做標記，但是因爲調值難以掌握，常常出現斷詞結果中，理論上正確的調值與文章所標記的調值不同的情況。

(2) 斷詞結果中，一字詞佔全體斷詞結果 50%左右。

由於變調邊界完全由斷詞結果決定，在一字詞率爲 50%的情況下，無法標示上正確的調值。

變調邊界的決定其實是很複雜的問題，並不能僅靠簡單的規則決定出來。只有在日後增加斷詞平均詞長後，所做的分析才有意義。

五、結論與未來展望

本論文嘗試以以下方式改善台語斷詞器的效能，包含

(1) 增加構詞單元，包含使用中文構詞機制部份與加入台語構詞規則部份。

(2) 增加數字部份文白讀音的選取機制，增加數字發音的正確度。

(3) 修正變調規則套用的範圍，譬如針對構詞後的詞調整變調邊界點。

(4) 修正並擴充台語辭典所含資訊，包括詞數的擴充、詞類資訊的加入、一字詞拼音的修正等。

加入構詞機制後，針對測試語料所統計出的平均詞長，由 1.46 增加至 1.6。增幅雖然不多，但是卻可從中得到有用的結論。

(a) 刪除錯誤率高的構詞規則，的確可以增加構詞成功率，顯示構詞的確對台語文章的斷詞起了作用。

(b) 依附在構詞動作下的數字文白讀音選擇機制，可以將原先認爲難以處理的數字發音部份做有效的初步分類，日後可再針對部分數字發音的特例做修正，可以更確定數字部分的發音。

(c) 依附在構詞規則下的變調邊界決定機制，雖然目前看來成效不彰(極少長詞被構出)，但是可做爲日後變調邊界決定方式的依據之一。待日後有能力擴充構詞的效能時，可以考慮使用此種簡單的標記方式來達成變調邊界的決定。

(d) 在數字文白讀部份，以構詞規則分類的方式決定數字的文白讀，可以將複雜的數字文白讀情況分類，較傳統的窮舉法更能準確決定數字的文白讀發音。

在未來展望部份，我們由目前斷詞器針對實際文章斷詞後結果的分析可知，還有許多可改善的部份，包含

(a) 詞庫的擴充。以往詞庫擴充的來源均爲標準辭典，傾向於具有詞的特性(詞性、意義等)，然而實際觀察斷詞結果後發現，有很多常寫在一起的字組，經常同時出現在文章裡，但是字組本身又沒有一個特定的詞性。此類字組可經由人工大量觀察文章而逐漸擴充。另外可能擴充的詞包含一些中文詞，譬如『大叔』、『接連』等中文特性較重的詞，也常在台語文章中出現。另外俚語的收集也可以增加斷出長詞的機會。

(b) 構詞規則的擴充。有一些詞綴，譬如『副總統』的『副』等，或者一些修飾語，譬如『真樸實』、『人真土直』的『真』等，有機會經過仔細評估後加入構詞規則中(評估原則爲，是否會造成嚴重搶詞，因爲此類規則並不是以有限的集合互相構成詞，而是以規則法構出)。

(c) 變調邊界的決定。在提高斷詞詞長後，必須檢視變調邊界的決定方式。目前仍然是以詞本身做爲變調邊界的決定，但是將來可以考慮使用詞類等資訊，搭配構詞方式的分類，以決定較正確的變調邊界。

Reference

[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] 鐘祥睿，”台語 TTS 系統之改進”，國立交通大學電信工程學系碩士論文，民國九十一年六月

[3] 黃競億，”台語 TTS 變調規則與斷詞器之製作”，國立交通大學電信工程學系碩士論文，民國九十年六月

[4] 楊鈺清，”台語文句翻語音系統之製作”， 國立交通大學電信工程學系碩士論文，民國八十八年六月

[5] 黃紹華，”A Study on Prosodic Information Generator for Mandarin Text to \ Speech”，國立交通大學電子研究所博士論文，民國八十五年六月

[6] 鐘榮富，”The Segment Phonology of Southern Min in Taiwan”，文鶴出版社，民國九十一年十一月(中文版書名爲”台語的語音基礎”)

[7] 鄭良偉，”台語的語音與詞法”，遠流出版社，民國八十六年

[8] 吳秀麗，“實用漢字台語讀音”，自立晚報文化出版部，民國八十一年五月

[9] 楊秀芳，”台灣閩南語語法稿”，大安出版社，民國八十年四月

[10] 張振興，”台灣閩南方言記略”，文史哲出版社，民國七十年九月

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

[12] 王文德，"台語語音辨識與文字處理之研究"，民國九十三年七月

[13] 吳季芳，"表列國語一字多音"，文化出版社，民國九十二年三月

附錄一：變調邊界標記

以國語構詞規則定義變調邊界種類，以做為構詞時決定邊調邊界的參考。

1：詞尾為變調邊界　2：變調邊界受後接詞影響

類別	regular expression	邊界	範例
NOP_	IN1 -> NO1*	2	一百
	IN2 -> NO2*	2	壹佰
	IN3 -> {IN1,IN2} {多,餘,來} ({萬,億,兆})	2	一百多萬 一百多
	DN -> (IN1) {點} IN1 ;	2	一點三
	FN2 -> (IN1 {又}) IN1 {分之} {IN1,DN} ({強,弱})	2	一又三分之一
	DN_1 -> IN1 {成} (IN1)	2	三成五
	NOP_1 -> IN1 (DESC) {半}	2	四小半 四半
	NOP_1 -> IN1 DESC	2	四大
	NOP_2 -> DESC {半}	2	大半
	NOP_3 -> IN1 PNM	2	一百整 一百多
NOP	NOP1 -> IN1 (DESC) ({半}) M	1	三大個 四小半個 五個
	NOP2 -> DESC ({半}) M	1	大半個
	NOP3 -> IN1 M PNM	1	一百個之多
	NOP4 -> M PNM	1	個之多
	NOP5 -> {IN3,DN,FN2,雙} M	1	一百多萬個
	NOP_6 -> {IN1, IN3,NOP_3,DN,FN2} {平方,立方} Nfg_1	1	一百平方公分
	NOP_6 -> {IN1, IN3,NOP_3,DN,FN2} Nfg_2	2	三畝
	NOP_7 -> {IN1,IN3,NOP_3,DN,FN2} Nfg_1	2	一百公分
	NOP_8 -> {IN1,IN3,NOP_3,DN,FN2} Nfg_3	2	一百多公斤
	NOP_9 -> {IN1, IN3,NOP_3,DN,FN2} Nfg_4	2	一公升
	NOP_10 -> {IN1, IN3,NOP_3,DN,FN2} Nfg_5	2	一小時
	NOP_11 -> {IN1,IN3,NOP_3,DN,FN2} Nfg_6	2	一元
	ONP -> ON M	1	甲方 乙方
RD	RNOP2 -> {半} M	2	半個
	RNOP3 -> {DESC,成} M	2	大個
	RD10 -> ({NOP1,NOP_1 }({又}) {NOP1,NOP_1} (({又}) {NOP1,NOP_1})*	2	一個又一個
	RD11 -> RNOP2 RNOP2	2	半個半個
	RD12 -> RNOP3 RNOP3	2	大個大個
	RD14 -> {這,那} {一} M M	2	這一支支
	RD13 -> {一} M M	2	一個個
WQP	WQP -> WQ M	2	整個
	WQP -> WQ Nff	1	全身
	WQP -> {整整,滿滿} {NOP1, NOP2,NOP3,NOP5,NOP_6, NOP_7,NOP_8,NOP_9,NOP_10,NOP_11}	2	整整一百個
WQP_	WQP_ -> {整整,滿滿} {IN1,IN3,NOP_3,DN,FN2}	2	整整一百
QQP	QQP -> QQ {NOP4,M}	2	若干個之多 若干個
DQP	DQ2 -> DFa {多,少}	2	很多
	DQP1 -> {好幾} {M,NOP1, NOP2,NOP3,NOP5,NOP_6, NOP_7,NOP4,NOP_8,NOP_9,NOP_10,NOP_11}	2	好幾十個 好幾大半個 好幾個之多 好幾個 數個之多
	DQP2 -> {DQ1,DQ2} M	2	那麼多個
DQP_	DQP_1 -> {好幾} {IN1,IN3,NOP_2,NOP_3}	2	好幾百
PQP	PQP1 -> {數} {M,NOP1,NOP2,NOP3,NOP4,NOP_6,NOP_7,NOP_8,NOP_9,NOP_10 ,NOP_11}	2	數百個 數大半個 數百個之多 數個
	PQP2 -> PQ {NOP4,M}	2	若干個之多 半個之多
PQP_	PQP_1 -> {數} {IN1,IN3}	2	數百 數十餘萬
CNP	CNP -> IN1 {年} {IN1,ON,N} {班}	1	一年一班 一年甲班 一年忠班
DSP	DSP1 -> DS M	1	本項 貴班
	DSP1 -> {他} {國,省,州,縣,鄉,村,鎮,鄰,里,郡,區,站,巷,弄,段,號, 地,樓,衖,市,洲} ;	2	他國
	DSP2 -> {該} {M,NOP1,NOP2,NOP3,NOP_6,NOP_7,NOP_8,NOP_9,NOP_10,NOP_ 11,PQP1}	1	該一支 該數十支
DSP_	DSP_2 ->{該} {IN1,IN3,DN,FN2,NOP_2,NOP_3,PQP_1}	2	該數十

OSP	OSP2 -> {各} {DESC,M,雙,XQP,NOP}	1	各項
	OSP2 -> {逐} M ;	1	逐項
	OSP3 -> {另外,近,將近} {PQP,NOP1,NOP2,NOP3,NOP5,NOP_6,NOP_7,NOP_8,NOP_9,NOP_10,NOP_11} ;	2	將近數百個 將近一百個
	OSP4 -> OS {PQP,NOP1,NOP2,NOP3,NOP5,NOP_6,NOP_7,NOP_8,NOP_9,NOP_10,NOP_11} ;	2	前一百個
OSP_	OSP_1 -> {第} {IN1,IN3,DN,FN2,NOP_3}	2	第一百
	OSP_2 -> {各} {IN1,IN3,DN,FN2,NOP_3}	2	各一百
	OSP_2 -> {每} {IN1,IN3,DN,FN2,NOP_3}	2	每一百
	OSP_3 -> {另外,近,將近} {IN1,IN3,DN,FN2,NOP_3} ;	2	將近數百 將近一百
	OSP_4 -> OS {IN1,IN3,DN,FN2,NOP_3} ;	2	前一百
DDP_	DDP_1 -> DD	2	這一百項
	DDP_2 -> {此} {WQP_,DQP_1,DQP_2,PQP_1,IN1,IN3,DN,FN2,NOP_3} ;	2	此一百個
OHSP	OHSP -> ({其它,其他,其餘}) {任何} {PQP1,NOP1,NOP2,NOP3,NOP5,NOP_6,NOP_7,NOP_8,NOP_9,NOP_10,NOP_11}	2	任何其餘一百個
OHSP_	OHSP_ -> ({其它,其他,其餘}) {任何} {IN1,IN3,DN,FN2,NOP_3} ;	2	其餘任何一百個
TIME	TDM -> {IN1} {個} ({多,餘,來,半}) Nfg_5	2	一個多月
	STDM -> IN1 {分} {IN1 {秒} (IN1)}	1	一分十五秒三
	TDM1 -> IN1 {小時} (STDM) (TPNM)	1	一小時一分秒
	TDM2_ -> IN1 {時,點} IN1 {刻} (TPNM)	1	一點一刻整
	TDM2 -> IN1 {小時} IN1 {刻} (TPNM)	1	一小時一刻整
	TDM2 -> ({Ndaac,Ndaad}) {元} {年} ({元} {月} (IN1({日,號})))	1	民國元年元月十號
	TDM2 -> ({Ndaac,Ndaad}) IN1 {年} ({元} {月} (IN1({日,號})))	1	民國三十年元月十號
	TDM2 -> ({Ndaac,Ndaad}) {元} {年} (IN1 {月} (IN1({日,號})))	1	民國元年四月十號
	TDM2 -> ({Ndaac,Ndaad}) IN1 {年} (IN1 {月} (IN1({日,號})))	1	民國三十年四月十號
	TDM3 -> ({Ndaac,Ndaad}) {元} {年} {元} {月份} ;	1	民國元年元月份
	TDM3 -> ({Ndaac,Ndaad}) IN1 {年} {元} {月份} ;	1	民國七十年元月份
	TDM3 -> ({Ndaac,Ndaad}) {元} {年} IN1 {月份} ;	1	民國元年三月份
	TDM3 -> ({Ndaac,Ndaad}) IN1 {年} IN1 {月份} ;	1	民國五年三月份
	TDM4 -> IN1 {月} (IN1 ({日,號})) ;	1	一月一日
	TDM4 -> {元,正,上,下,每,本} {月} (IN1 ({日,號})) ;	1	本月四號
	TDM5 -> IN1 {日,號};	1	一號
	TDM7 -> Ndabd1 (Ndabe) TDM1_;	1	星期一傍晚五點
	TDM8 -> Ndadb1;	1	星期一傍晚
	TDM9 -> Ndabe TDM1;	1	傍晚五點
	TDM10 -> {每,上,下,本} ({個}) {TDM7,TDM8};	1	每個禮拜五
LLP	LLP -> IN1 {度} (IN1 {分} (IN1 {秒}))	1	一度一分一秒
ADP	ADP -> (IN1 {段}) (IN1 {巷}) (IN1 {弄}) IN1 ({之} IN1) {號} (IN1 {樓}) ({之} IN1)	1	一段一號
TDP	TDP -> {攝氏,華氏} ({零下}) {IN1,DN} {度}	1	攝氏五度
BD	BD2 -> {NOP1,NOP2,NOP3,NOP4,NOP_6,NOP_7,NOP_8,NOP_9,NOP_10,NOP_11} BD	1	一百個左右
	BD2 -> {不到}{NOP1,NOP2,NOP3,NOP4,NOP_6,NOP_7,NOP_8,NOP_9,NOP_10,NOP_11}	2	不到一百個
BD_	BD_2 -> {IN1,IN3,NOP_2,NOP_3,STDM,LLP,TDP,STDM,TDM1~10…..etc} BD	1	一百左右
	BD_2 -> {不到} {IN1,IN3,NOP_2,NOP_3,STDM,LLP,TDP,STDM,TDM1~10…..etc}	2	不到一百
MONEY	MON -> $ {INI,IN3,DN}	1	$100