台語文字與語音語料庫之建置

Development of a Taiwanese Speech and Text Corpus

廖子宇　Tzu-Yu Liao
中央研究院資訊科學所
Academia Sinica
Institute of Information Science
ziiyu@iis.sinica.edu.tw

呂仁園　Ren-yuan Lyu
長庚大學資訊工程學系
Department of Computer Science and Information Engineering
Chang Gung University
renyuan.lyu@gmail.com

高明達　Ming-Tat Ko
中央研究院資訊科學所
Academia Sinica
Institute of Information Science
mtko@iis.sinica.edu.tw

江永進　Yuang-chin Chiang
國立清華大學統計學系
Institute of Statistics
National Tsing Hua University
chiang@stat.nthu.edu.tw

張智星　Jyh-Shing Roger Jang
國立清華大學資訊工程學系
Department of Computer Science
National Tsing Hua University
jang@cs.nthu.edu.tw

摘要

台語在台灣是三大主要語言之一，台語的使用人口約為 70%的人口，可是，其台語方面的相關研究卻是很少、研究論文主要還是以華語為主。優質的計算語言學研究需要大規模的語料來支持，本計畫目的是建立大規模的台語語料庫，建立台語計算語言學研究發展的厚實基礎。同時希望以此經驗嘗試建立台灣弱勢語言的計算語言學研究發展模式。本計畫中，將建立一個台語語料，語料來源類型為台語朗讀、新聞、戲劇及談話。建立 200 個小時的台語文字與語音語料。

Abstract

The main goal of this paper is to develop a large scale Taiwanese corpus. In the mean time, we try to establish a successful model for the computational linguistic research on other minority Taiwanese languages such as Haka.In this paper, we will build a Taiwanese speech corpus. The source of speech corpus is Taiwanese dramas and news from TV stations. The goal of the corpus is 200 hours speech material with annotation.

關鍵詞：corpus, speech recognition, Taiwanese, transcription

一、緒論

根據 2005 年的統計[1]，台語的總人口數為四千五百萬人，台灣是其中台語使用人口較多的地區，約有一千五百萬人。由使用人口數而言，台語是個不可忽視的語言。台語在台灣是三大（華語、台語以及客語）主要語言之一。在台灣，大部份的人有能力說這兩種或三種語言，根據統計[2]，華語的使用人口約為 82%，而台語的使用人口約為 70%的人口，且台灣是目前唯一以台語為重要語言的國家。

然而，目前台灣對於台語方面的相關研究，如語言學、應用科學（語音辨識、語言辨識、網路資訊檢索，等等），其論文數量卻是很少、研究論文主要還是以華語為主。在台灣，台語是重要語言，使用人口眾多，使用環境相當好，不論在學術或社會應用而言，台語研究都應該受到重視。在台灣台語和華語的使用人口並沒相差多少，但研究台語的論文量和研究華語的論文量是相差非常多。其中一個原因是因為台語目前沒有統一的文字書寫系統，甚至大部分使用台語的人口並不會書寫台語。這一方面造成台語文字資訊來源貧乏，另一方面也造成自動化處理台語文字的困難。這兩者使得台語的研究發展產生了阻礙，而間接的影響到台語研究的數量。

現今優質的計算語言學研究都需要大規模的語料來支持。使用的語料規模越大通常代表其訓練出來的應用系統的準確率越高。本計畫的目的即是建立大規模的台語語料庫，以建立台語計算語言學研究發展的厚實基礎。同時希望以此經驗嘗試建立台灣弱勢語言的計算語言學研究的發展模式。

台語文字語料的收集因書寫系統的分歧與書寫人口的缺乏，有其困難。但台語的口語人口相當多，台語的語音資料非常豐富，加上大多數的台語人口都能使用華語，提供建置大量台語語料的可能性。

二、語料庫處理流程

以下，我們描述整個語料庫處理的流程，分成下列幾個步驟來敘述：

（一） 語音蒐集

我們先設法取得高品質之語料作為來源，首先對語音資料做些格式的正規化，指定取樣頻率、聲道數、以及取樣位元數分別為 16000Hz, 單聲道以及 16 bit/sample。

（二） 切音

依據語音串流中較長的靜音段，做為切音成段落的依據，再進一步根據所設定文句長度來切到句的層級。

（三） 聽打

將每一段經過切音的聲音段做聲音內容的聽打，使用的拼音系統為台語通用拼音。目前聽打工作只做聲音內容的音節聽打，並未做聲音內容的漢字聽打，若是語音蒐集時同時蒐集語音和文本內容時，會將文本內容保存下來並將音節聽打增加在文本內容之後。

聽打的儲存格式為 XML 檔案格式，副檔名為.trs，下面為經過聽打所儲存的 trs file 的儲存檔案內容範例。

<?xml version="1.0" encoding="UTF-8"?>

<Sync time="2.746"/>

001 牛墟 （hi）//de3-it1-pinn1-,qu3-hi1-

<Sync time="5.982"/>

紀傳洲（18/04/07thk 改寫）//zok1-zia4-,gi4-tuan2-ziu1-

<Sync time="8.821"/>

古早，//go1-za4-

….

此範例中，聽打的內容是用 UTF-8 編碼，同時記錄了每一句的時間、拼音以及蒐集取得的文字。

1. 聽打工具

使用 LDC(Linguistic Data Consortium)提供的 Transcriber[3]系統來聽打與標記節目錄音資料。

2. 聽打說明

（1） 聽打人員

以各地方教育單位之鄉土語言老師為主，每位老師從事鄉土教育工作皆有一定之年資。

（2） 拼音系統

本語料蒐集計畫使用台語通用拼音，並以自然調型的方式做聽打。

（3） 事件標記

聽打過程中所出現的非台語聲音會使用標記 Event 的方式作註記。紀錄聽打過程中出現的 Event 需紀錄三種資料分別是 Type、Description、Extent。

Type 是用來區分聲音的種類，Description 為說明，Extent 則是標記起始與終始點或該段時間。

3. 成果儲存格式

本計畫決定運用 xml 格式來做為資料聽打成果的儲存方式，其內記錄以下四個項目：

（1） 檔頭：記錄檔案建立時間及對應聲音檔

（2） 語者(Speakers)：記錄內容出現所有語者的名字以及 id 編號，id 編號於後續內容中使用。

（3） 主題(Topics)：記錄內容所有出現的主題段落名稱與 id 編號，id 編號於後續內容中使用。

（4） 內容(Episode)：記錄所有段落、語者、標音的時間位置。

（5） 段落標記(Section)

此標記包含於內容(Episode)之中，子項目分別為：Topic 標記主題的 id 編號，startTime 段落內容開始時間，endTime 段落內容結束時間。

（6） 語者標記(Turn)

此標記包含於內容(Episode)之中，Speaker 標記說話人的 id 編號，startTime 段落內容開始時間，endTime 段落內容結束時間。

（7） 標音標記(Sync)

此標記包含於內容(Episode)之中，主要記錄聽打內容與開始時間，子項目分別為：Time 標音開始時間。

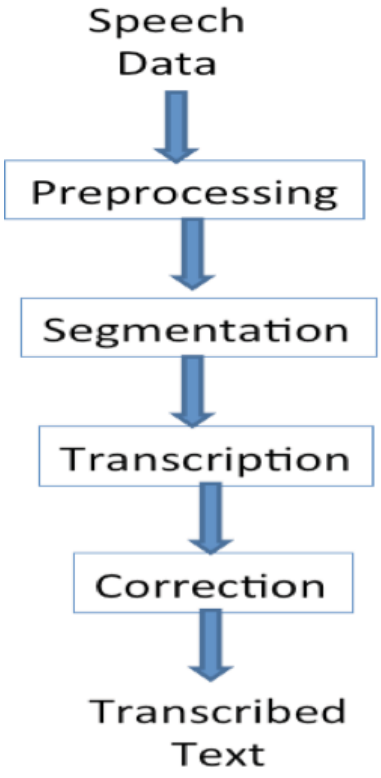
（8） 事件標記(Event)

此標記包含於內容(Episode)之中，所有語音中出現的非台語聲音都會以此做標記，子項目分別為：type 事件種類，desc 事件內容，exten 事件位置。

（四） 校正

校正工作以聽打人員兩人為一組互相校正，兩人於聽打完成時交換檔案做校正，校正者於聽打內容中標記其認為聽打錯誤之內容並加註其認為正確之聽打內容，保留聽打者之聽打標音與校正者之校正標音，並於後續之二校中做選擇。二校由聽打者執行，搜尋校正者所做之校正標記並從原始聽打標音與校正標音中選擇其認為正確之標音。

以上流程我們以圖一綜合呈現之。



圖一、聽打工作流程

三、已建立之語料內容簡介及數量統計

本計畫準備以三年的時間蒐集及處理 200 小時的台語語音資料，語料類型分為朗讀、新聞、戲劇、以及談話。以下分別就各類型語料做些描述：

（一） 朗讀語料

所謂朗讀語料，在我們的研究中，指的是先有朗讀文稿，說話人完全照本宣科，完全不加入說話人的思考或情緒，是一種文字和語音完全相符的語音資資料。目前，我們有 2 套朗讀語料，分別如下：

1. TW03-GS

這是長庚大學多媒體實驗室於 2003 年所錄製的錄音語料，由呂仁園教授主持的語料蒐集計畫，錄製文本內容以台語所有音節(約一千五百個)平均分布在各個句子當中，並邀請四位台語教學老師來錄製，所得到的語料庫。

2. TwEdu

由教育部邀請學者專家組成小組分組進行閩南語文章之選錄，並聘請專人將文章內容改寫成適合朗讀之文稿，名為「閩南語朗讀文章選輯」，也請國立教育廣播電台協助錄製聲音檔。「閩南語朗讀文章選輯」收錄文章 133 篇(含重複 7 篇共 140 篇)，其用字除依教育部公布之台語閩南語推薦用字外，部分不屬於前述用字者，為便利閱讀，依原著用字呈現。[4]

（二） 新聞語料

所謂新聞語料，在我們的研究中，指的是電視新聞記者在新聞節目中所報導之新聞內容，經過我們的蒐集以及人工聽打所製成的語料，原則上新聞主播有預先寫好的新聞稿來念，但是有些時候，新聞主播會自由發揮，加入個人臨場的反應，而且免不了有個人情緒參雜其中。典型的新聞語料，包括主播、記者、受訪者以及插播廣告，在我們蒐集的語料中，主播一定是用台語來播報，外場記者及受訪者不必然使用台語，我們的人工聽打，對於非台語的部分，不做音標轉寫，只標註該段聲音的語言屬性如英語、或華語。

目前，我們有 5 套新聞語料，其中 4 套來自民視台語新聞[5]簡稱 FTVN-1, FTVN-2, FTVN-3, FTVN-4，FTVN 為民視台語新聞的縮寫，後續編號則為不同時間所做的聽打工作時間順序，1 套來自公視台語新聞[6]，名稱為 PTSN-1，其內容詳情分別如下：

1. FTVN-1, FTVN-2

經過剪接後的新聞節目，只含主播以及台語記者，所以整個節目，幾乎都是以台語發音為主。

2. FTVN-3, FTVN-4

經過剪接後的新聞節目，只含主播以及台語記者，所以整個節目，幾乎都是以台語發音為主。

3. PTSN-1

同為完整的新聞節目，與前項 FTVN-1, FTVN-2 雷同。

（三） 戲劇語料

所謂戲劇語料，在我們的研究中，指的是電視台之台語戲劇節目，如民視的浪淘沙[7]或娘家[8]，我們取得該戲劇節目的現場錄音檔案，僅含演員的純語音，不含製成節目之後的混音，使得我們不必費神處理那些非語音的聲音。原則上戲劇演員應有劇本可供念誦，演員對劇本內容有一定的熟悉度，配合當下的情緒，說出含豐富情緒內容的對白。目前，我們有 2 套戲劇語料，都是來自民視，簡介

如下：

1. 浪淘沙

戲劇浪淘沙總集數為 30 集，從中挑選 10 集，挑選之集數為 3、6、9、12、15、18、21、24、27、30，完成語料時間約 8 小時。

2. 娘家

戲劇娘家總集數為 415 集，從中挑選 16 集，挑選之集數為 3、13、23、33、43、53、203、223、243、253、263、333、363、383、393、413 集，完成之語料時間為 26.8 小時。

（四） 談話語料

所謂談話語料，在我們的研究中，指的是廣播節目中主持人與來賓之間，依固定主題做廣泛的交談，這種交談，通常沒有文稿，不做事先演練，幾乎完全是當場隨興的自由發揮，說話的形式無拘無束，語句也不見得很流暢或合乎文法規則。目前，我們有 1 套談話語料，是來自網路電台「夢中的國家」[9]，簡介如下：

1. 夢中的國家

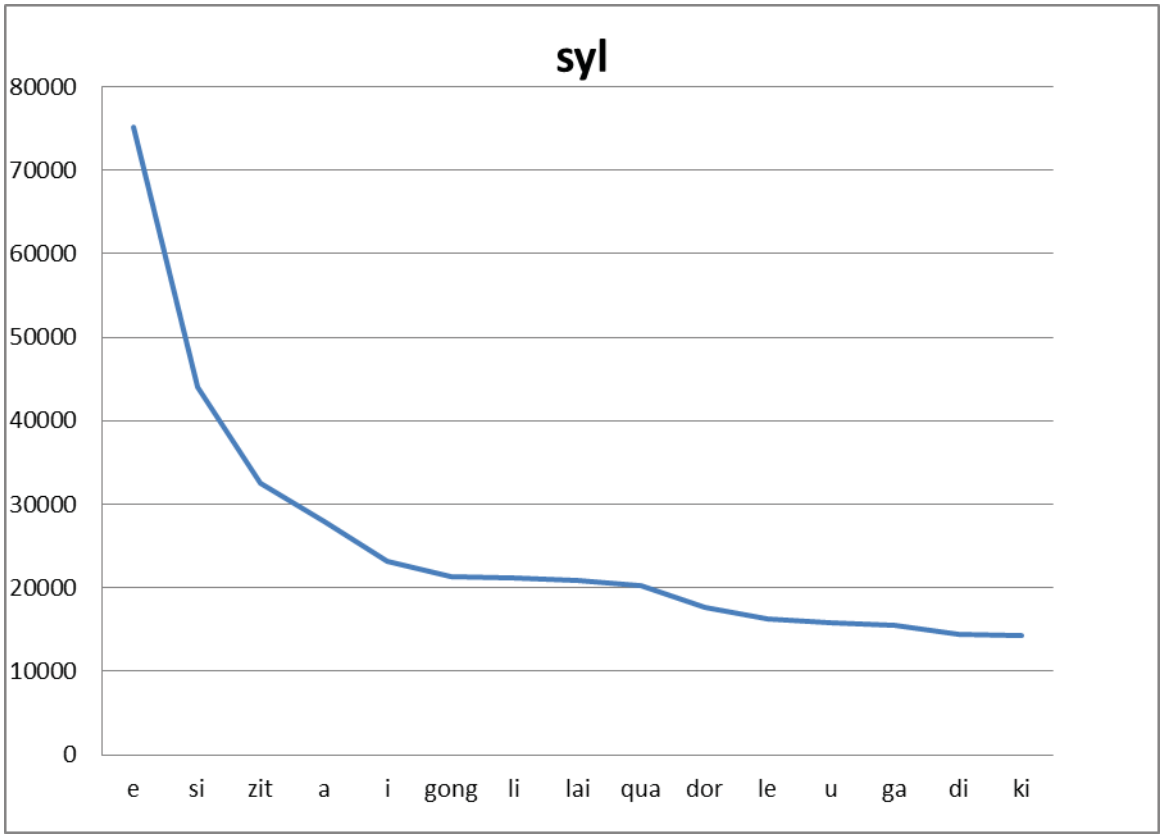
由張素華小姐所主持的台語談話性節目，談話主題遍佈生活、健康、教育、政治、財經等，內容非常多元。目前完成之語料有 29 集共 25.8 小時。

以上所描述的語料，以下表做一個綜合性的呈現：

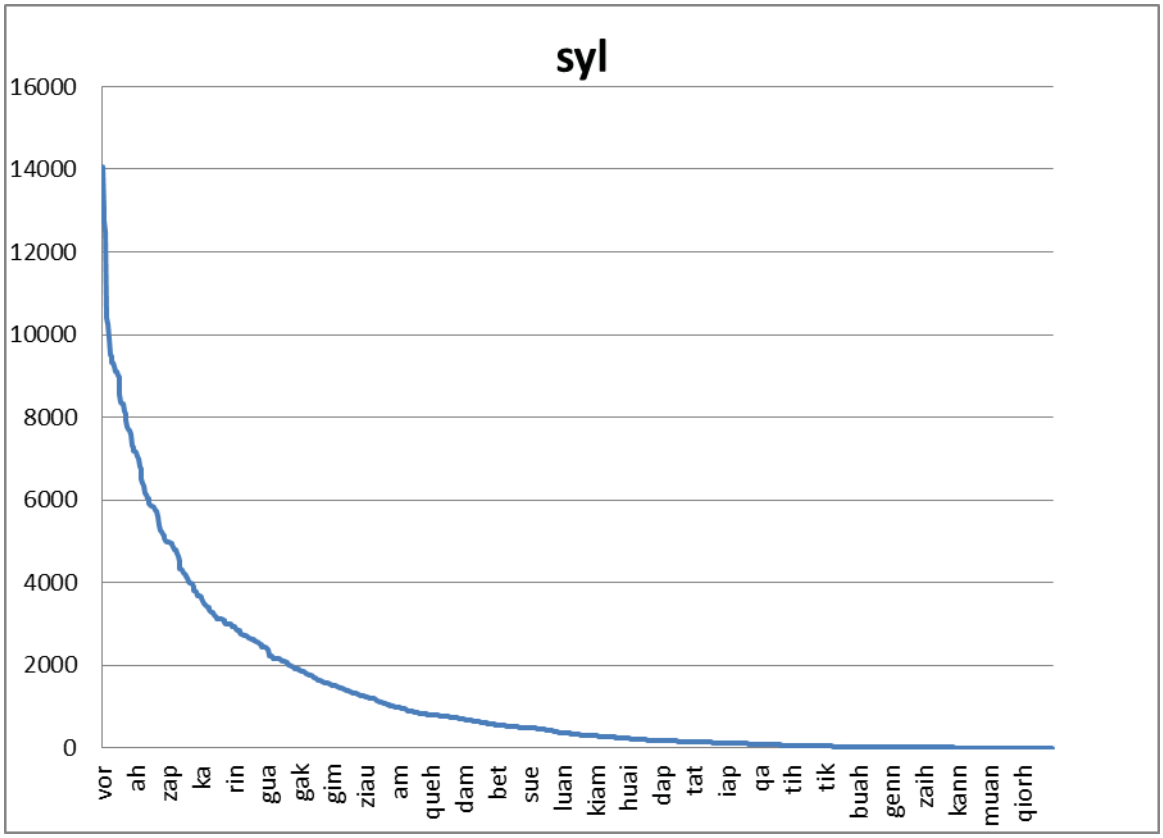
表一、語料庫統計

類型	名稱	時間(分)	時間(小時)	檔案數	音節數	相異音節	音素數	相異音素	語者數
朗讀	TW03-GS	1421.53	23.69	14	312299	749	592616	139	4
朗讀	EUM	679.74	11.33	14	113838	681	219719	134	1
新聞	FTVN-1	332.34	5.54	6	102706	604	197191	130	484
新聞	FTVN-2	399.57	6.66	8					
新聞	FTVN-3	766.22	12.77	21	328995	712	643584	132	142
新聞	FTVN-4	716.74	11.95	25					86
新聞	PTSN	656.92	10.95	11	75004	577	145412	128	303
戲劇	LTS	474.86	7.91	9	38753	476	71672	125	85
戲劇	MH	1609.76	26.83	15	204194	656	374021	134	166
談話	DRM	1549.21	25.82	58	290524	684	547477	134	59
	total	8606.89	143.45	181	1472457	884	2804459	147	1252

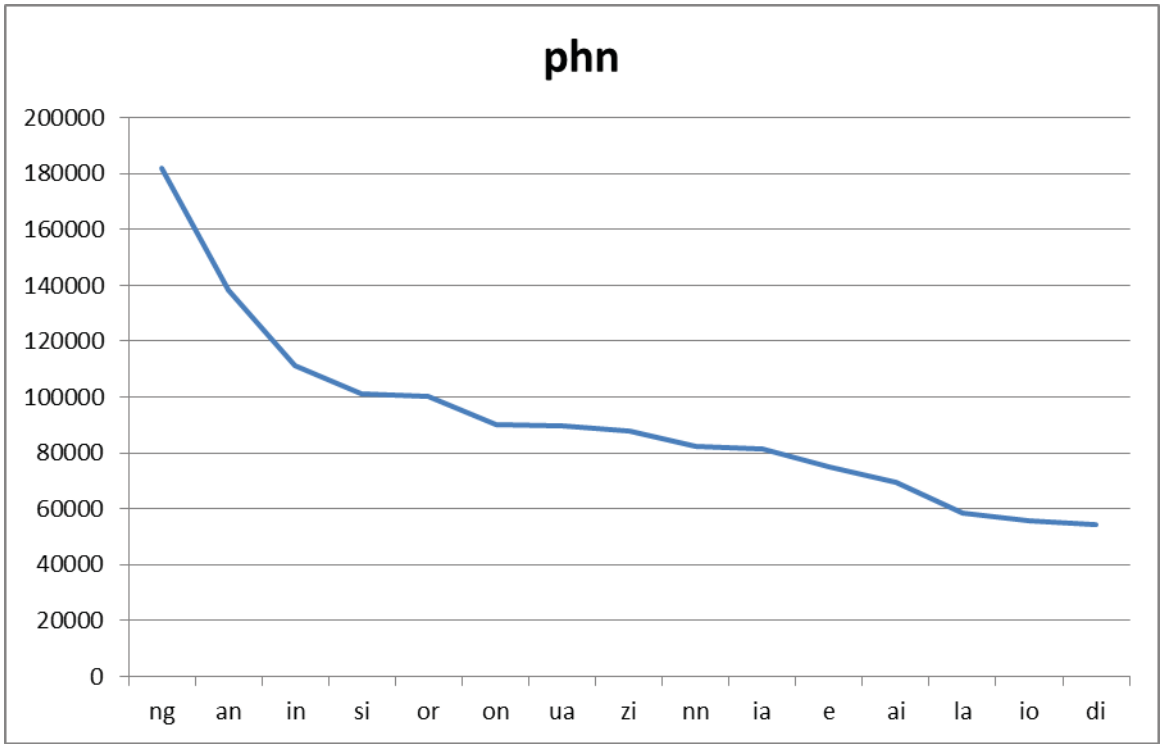
同時將內容做統計與分析，統計所有音節數和音素數的出現次數並製作成統計圖如下：



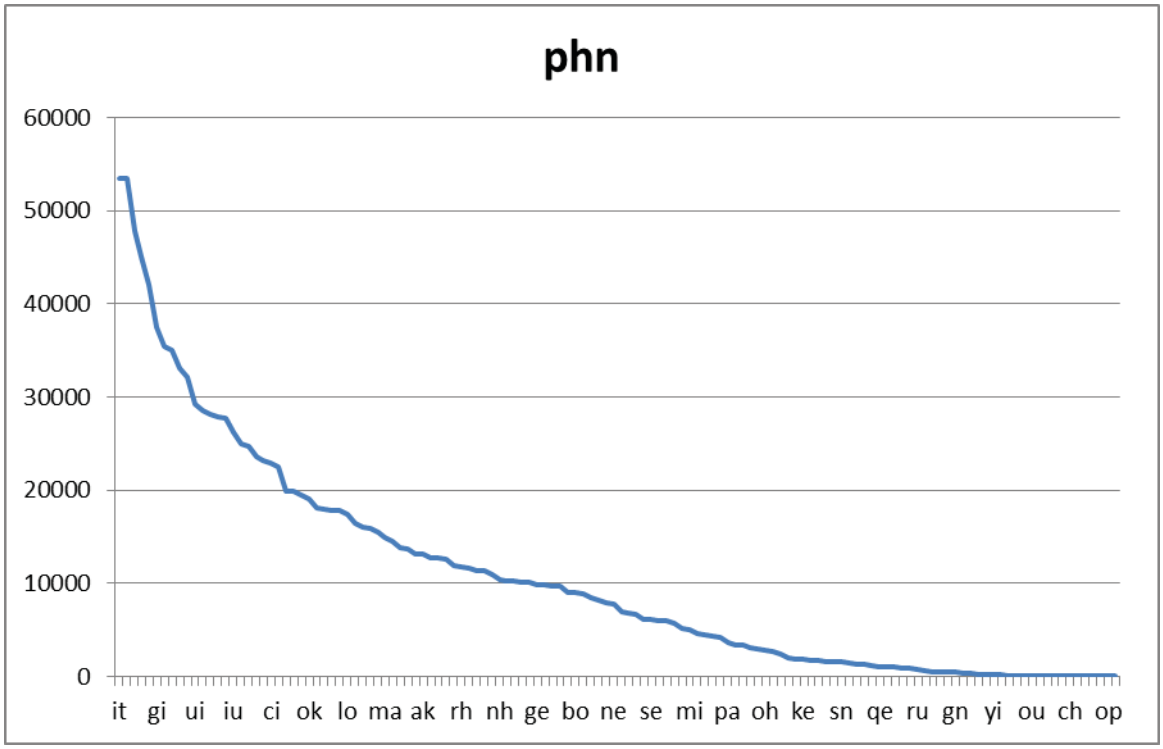
圖二 音節統計圖(1~15)



圖三 音節統計圖(16~885)



圖四 音素統計圖(1~15)



圖五 音素統計圖(16~148)

四、結論

目前語料庫蒐集進度尚未達到預期設定的數量，未來會加快腳步進行語料蒐集的工作，期望在計畫進行時間內完成。為使語料庫更方便使用，接下來會將語料庫整理成不

同格式方便作使用與檢索以提供更簡單更大眾化的語料庫工具，讓語料庫不只能夠做語音相關研究也能夠作其他用途。目前計畫將所有聽打之內容與聲音製作為 EPUB 電子書並美化其排版與內容與建構語料庫檢索網頁提供線上檢索語料庫內容。

參考文獻

[1] http://www.ethnologue.com/14/show_language.asp?code=CFR

[2] Wikipedia. (2006). Available from http://en.wikipedia.org/wiki/Demographics_of_Taiwan

[3] Trainscribe http://www.seventhstring.com/

[4] 全國語文競賽台灣閩南語朗讀參考資料 http://140.111.34.54/MANDR/minna/first.html

[5] FTVN(民視新聞網)http://news.ftv.com.tw/

[6] PTSN 公視新聞網 http://news.pts.org.tw/

[7] LTS(浪淘沙)http://program.ftv.com.tw/Drama/TDoctress/

[8] MaternalHome(娘家)http://program.ftv.com.tw/Drama/Momshouse/

[8] Dream_State(夢中的國家)http://sowhuaa.ning.com/