103 年度客家委員會獎助客家學術研究計畫

成果報告書

計劃主持人: 黃豐隆 博士

國立聯合大學資工系 副教授

http://web.nuu.edu.tw/~flhuang/nlp

中華民國 103年 11月 25日

本研究計畫由客家委員會獎助部份經費完成

特此致謝

目　錄

摘要(Abstract)

電腦的作業系統有 Windows，Linux，Mac OS 以及手機 Mobile cell 等，對任何電腦平台之使用者而言，語言文字(Texts)之輸入是非常重要的工作。因此，一個具有效率、快速與使用上正確之輸入法(Input Method)均是必要、重要的人機介面。

依客家委員會統計資料，臺灣客家族群約佔總人口的六分之一，約有四百萬人，為閩南語族群外第二大之族群。根據臺灣的客語使用狀況調查報告指出，阻礙客語傳承之主因是:「不太會講」，由於台灣教育學習環境使然，導致連客籍家庭的學童亦少能以客語說話、交談，具有聽、說客語能力者逐年下降，能說客語的人口大量減少，台灣出現了客語失聲、客家文化失傳之危機。

這幾年來，客家委員會大力倡導客家文化，在語言方面，實施客語能力之認證並在各級學校推動客語拼音學習等工作，均是提升口說客語的有效方式，漸有不錯之成效。如何使客家人可以經由語言認識客家文化，進而認同、傳承文化是當前推廣客家文化的重要工作之一。此外，如何使非客籍人士，如台灣閩南族群，甚至是其他國家的人士，可以有效學習客語進而認識客家文化，均是傳續、推廣客家文化的工作中不可或缺的。因此，有效的書寫客語客語文章的之輸入法將是十分重要的，如果再結合客語之語音，應可提升客語之「聽與說」能力。

經過我們實際操作後發現，現有的教育部官方版本與民間研發較普遍使用的方法，均存在許多值得改進之處。基於前述的原由，本計劃主要目的即:研發具有創新與智慧式功能之客語拼音輸入法，我們提出許多改進之處，改善現有方法並研發創新性、更具智慧與實用的功能。

本計劃主要的目標功能如下:

1. 相容多種拼音輸入功能。
2. 可自動擴增詞典之詞目。
3. 往後預測出現候選詞目。
4. 具備容錯能力(Fault Tolerance)。
5. 可自動校正輸入之錯誤拼音(Error corrections)。
6. 具有語音(Speech)輸出，結合語言數位學習之功能。

本研究計劃提出具有強鍵性(Robust) 的方法「**優客輸入法(U-Hakka Input Method)**」，易於擴充至其它語言，除了客語語文之外，亦能擴展至其它的臺灣本土語文。因此可提供一種具有智慧、高效率與實用的客語拼音輸入法 (以四縣與海陸腔兩種為主)，並具有客語語言之數位學習功能(E-learning)，對客家語言之使用與推廣具有助益。

第一章　研究內容概述

電腦在現在已經越來越普及，無論是家裡用的桌上型電腦或者是隨身攜帶的手機，對於使用中文的我們，系統裡一定都會使用到輸入法來作為輸入中文的工具，因此如何提供有效率及方便的輸入法是值得討論的議題。

語言是文化傳承與推廣的首要工具，任何一種語言均有保存之必要性，要能讓語言容易傳承與推廣即是要能讓人容易理解與學習此種語言的文字，若要在電腦上輸入各種語言的文字必須要使用輸入法，然而目前客語的輸入法並不常見，且目前有的客語輸入法功能也較為基本。因此本論文研究與實作出一種能讓使用者方便且快速的輸入客語文字的輸入法。

1.1　研究動機

根據行政院客家委員會在 2010 年至 2011 年的調查[1]，台灣的客家佔了 18.1%，約有 419.7 萬人，為閩南族群外的第二大族群，但不同於閩南語，會聽及說客語的人口比例遠不及閩南語。根據 2004 年客語使用狀況調查報告指出[2]，阻礙客語傳承的主要原因為不會講，且由於我們的教育學習環境影響，導致連客籍家庭的學童也很少能使用客語說話、交談的機會，具有聽、說客語能力者逐年下降。而客語文字不夠通行的原因之一是沒有統一的文字。

雖然目前客委會已有建議用字，但仍能看到許多用字的不同，一些不常見的字甚至用看的也不知道其意義；另外就是輸入的困難，因為大多數人對於客語的拼音系統不熟悉，即使會說客語也無法正確拼出。因此根據上述原因我們希望能設計出一個能讓不熟悉客語拼音系統的使用者都能夠易於使用的客語輸入法，而且能夠打出讓別人知其義的客語句子，讓輸入法不光只是輸入，甚至提供數位學習的功能。

1.2　研究方向

本研究所提出「**優客輸入法(U-Hakka)**」的輸入方式，是根據教育部制定的客家語拼音方案[3]所制定的拼音來進行輸入，且我們選擇兩種最常見的客語腔調，四縣腔及海陸腔來實作，其拼音是使用二十六個羅馬字母表示，因此在我們現在一般使用的 QWERTY 配置的鍵盤都能輕易地進行輸入，由於不須使用額外的符號來表示拼音，因此對

於外國的使用者也較容易來學習與使用。由於客語是屬於使用聲調的語言,每個客語文字都有其聲調,且當兩個以上的客語字組成詞句時,可能會產生變調的情形，例如：在四縣腔「音」這個客語字原本讀 im´,但是「音樂」這個詞一起讀時,「音」會產生變調而改讀成 im˘。考慮到變調造成使用者在輸入時的困擾,且對於音調不熟悉的使用者在音調輸入上的困難,因此我們提出的輸入法輸入的拼音採取不需要輸入聲調的方式進行輸入。且考慮到客家語拼音方案所制定的客語拼音會有版本更改的問題,因此我們還設計了拼音錯誤提示的功能,目的即為使用者輸入拼音時,若是拼音輸入錯誤隨即產生目前輸入的拼音還有哪些可能的正確拼音可以選,設法提供給客語拼音的初學者友善的輸入環境。

除了基本的拼音輸入之外，希望能提供使用者快速且便利的輸入方式，因此加入了便捷輸入的功能，其功能提供了自訂輸入、音首輸入、縮寫輸入及英文詞輸入。此模式能有效的減少輸入時按下鍵盤的次數，以節省輸入所耗費的時間，例如：利用音首輸入只需要輸入「GLGD」即可得到客語詞「緊來緊多」、利用縮寫輸入只需要輸入「NUU」即可得到「聯合大學」、利用英文詞輸入只需要輸入「Golf」即可得到「高爾夫」,而自訂輸入可以讓使用者自行去新增、刪除及修改自訂辭典，以自己想要的拼音來輸入自訂詞。

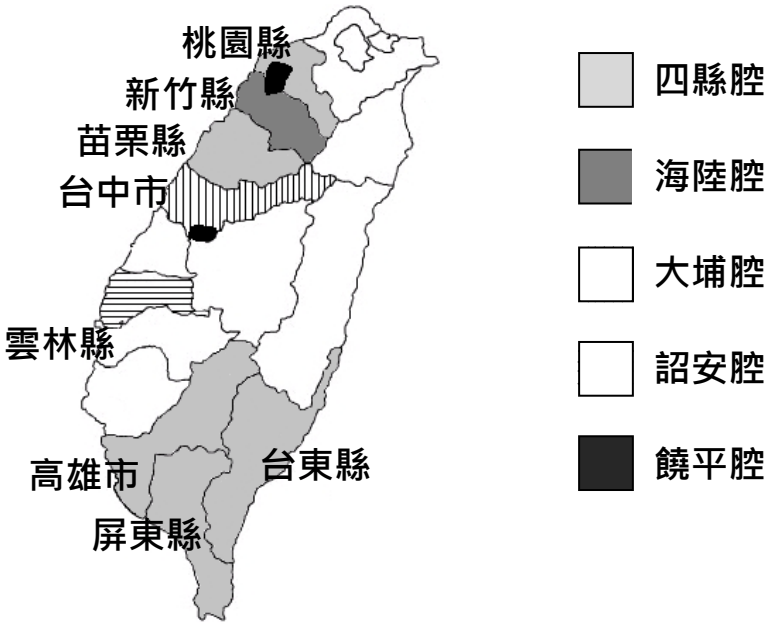
此外，希望輸入法能結合語言數位學習的功能，我們考慮到現有客語文章常會發生客語詞不易理解的情況,因此我們加入了加註國語詞與拼音的選項，讓輸入得到的客語文章更具有可讀性，且加註國語詞的功能也能提供使用者做為輸入參考,用來判斷是否為使用者所想的詞。以及我們提供了在輸入時，唸出正在輸入的拼音，讓使用者不只是用看的來得知是否輸入正確，還能使用聽的方法來得知。輸入後會以詞為單位讓使用者選擇是否唸出此詞,讓使用者在輸入的過程中還能學習客語的發音。

第二章　客語四縣腔與海陸腔介紹

客語為台灣客家族群所使用的語言，客語主要源於中國大陸粵東，因地區不同造成了許多不同的客語腔調，在台灣也有十多種客語腔調之分，但大多數的腔調只有在少部份地區，甚至只有在自己家族裡使用。目前客語五個主要使用的腔調為四縣、海陸、大埔、饒平、詔安，又以四縣及海陸腔最通行，佔了客語使用人數的九成以上。因此我們在這章節主要介紹這兩個本論文提出的輸入法所使用之客語腔調。

2.1　台灣客語分佈

目前台灣主要使用的客語腔調為四縣、海陸、大埔、饒平、詔安，其中前二者「四縣客語和海陸客語」是使用最多的二種客語腔，客語腔調的地區分佈如圖一所示。



圖一、台灣客語腔調分佈

四縣客語是指大陸廣東省之蕉嶺、平遠、五華、興寧四個客家地區通行的語言。台灣的客語以四縣腔為客家話中通行最廣的語言，全台客家人的地方，大部分都能用四縣話溝通、或聽得懂四縣之客家話。苗栗縣是客語四縣腔的大本營，還有北部有桃園縣中壢市、平鎮市、龍潭鄉等地。苗栗縣除了靠海的苑裡鎮、通霄鎮、竹南鎮、後龍鎮與山區的三灣鄉、卓蘭鎮等地以外，其它山線的鄉鎮大都通行客家話。

南部高雄縣的美濃鎮，屏東縣的長治、新埤、萬巒、竹田、佳冬、高樹等地屬於南部四縣腔。

海陸客語則是指通行於廣東省海豐縣、陸豐縣地區的方言，由於兩地通行語相差不多，故合稱命名海陸腔客語。由於海陸客語是台灣客語的少數族群，其流通範圍以新竹縣為主，略見於桃園，較少見於苗栗。故海陸客語通行地區以新竹縣為大本營，其餘分散地區：北部有桃園縣的觀音鄉、新屋鄉、楊梅鎮。新竹縣則有竹東鎮、橫山鄉、關西鎮、新埔鎮、湖口鄉、芎林鄉與寶山鄉等地區。

2.2 客家語拼音方案

基於歷史因素，台灣近 50 年來幾乎沒學習與使用羅馬拼音，而在 2001 年九年一貫新課程啟動後，課程綱要中明訂音標系統應用能力為台語、客語的基本能力指標。

客家語拼音方案為教育部所公告的台灣客家語羅馬字拼音方案[3]，最近一次的更新為中華民國 101 年 9 月 12 日的修正公告。因為 2009 年政府將漢語拼音訂為國家標準，因此客家語拼音方案可能會持續修正，使其符合漢語拼音。因此我們實作的四縣與海陸腔客語輸入法之拼音系統以最新公告的客語拼音方案為基礎。

客家話的音節結構和其他漢語方言(如閩南語)一樣，可以分為兩大部分：聲母和韻母及聲調，聲母是指音節的第一個輔音，而韻母又可分為韻頭、韻腹與韻尾。在音節結構中，只有聲調和韻腹是不可或缺的要素，其它則可有可無。表一為客家語拼音方案中的聲母符號表、表二為客家語拼音方案中的韻母符號表。

表一、客語聲母符號表

客語拼音	b	p	m	f	v	bb(註 1)
國際音標	[p]	[p^h]	[m]	[f]	[v]	[b]
注音符號	ㄅ	ㄆ	ㄇ	ㄈ	ㄪ	ㄋ
客語拼音	d	t	n	l	r(註 2)	g
國際音標	[t]	[t^h]	[n]	[l]	[j]	[k]
注音符號	ㄉ	ㄊ	ㄋ	ㄌ		ㄍ
客語拼音	k	ng	h	j(註 3)	q(註 3)	x(註 3)
國際音標	[k^h]	[ŋ]	[h]	[tɕ]	[$tɕ^h$]	[ɕ]
注音符號	ㄎ	ㄫ	ㄏ	ㄐ	ㄑ	ㄒ
客語拼音	z	c	s	zh(註 4)	ch(註 4)	sh(註 4)
國際音標	[ts]	[ts^h]	[s]	[tʃ]	[$tʃ^h$]	[ʃ]
注音符號	ㄗ	ㄘ	ㄙ	ㄓ/ㄐ	ㄔ/ㄑ	ㄕ/ㄒ
客語拼音	rh(註 4)					
國際音標	[Z]					
注音符號	ㆎ					

附註：

1. bb 可用於雲林詔安腔、南投國姓鄉及部分南部客家地區。
2. r 為摩擦音，僅用於部分南四縣腔。
3. j、q、x 可用於四縣腔及南四縣腔。
4. zh、ch、sh、rh (ㄓ、ㄔ、ㄕ、ㆎ)用於海陸、饒平、詔安等三腔。
   zh、ch、sh、rh (ㄐ、ㄑ、ㄒ、ㆎ)用於大埔腔。

表二、客語韻母符號表

客語拼音	ii	i(註 1)	e	ee(註 2)	a	o
國際音標	[ɨ]	[i]	[e]	[ɛ]	[a]	[o]
注音符號	ㄇ	ㄧ	ㄝ		ㄚ	ㄛ
客語拼音	oo(註 2)	u(註 1)	er(註 3)	m(註 4)	n(註 4)	ng(註 4)
國際音標	[ɔ]	[u]	[ə]	[m]	[n]	[ŋ]
注音符號		ㄨ	ㄜ	ㄇ	ㄋ	ㄫ
客語拼音	b(註 4)	d(註 4)	g(註 4)	nn(註 5)	m(註 6)	n(註 6)
國際音標	[p]	[t]	[k]	[~]	[m̩]	[n̩]
注音符號	ㄅ	ㄉ	ㄍ		ㄇ	ㄋ
客語拼音	ng(註 6)					
國際音標	[ŋ]					
注音符號	ㄫ					

附註：

1. i、u 可用於韻頭、韻腹及韻尾。
2. ee、oo 僅用於詔安腔。
3. er 用於部分海陸、饒平。
4. -m, -n, -ng 用於陽聲韻尾(鼻音韻尾)；-b, -d, -g 用於入聲韻尾(塞音韻尾)。
5. nn 一般僅用於詔安腔，但其他腔亦偶可見，如：歪 uainn+(大埔)。
6. 輔音 m, n, ng 可視為韻腹，自成音節。如：(四縣)魚 ngˇ。

2.3 四縣與海陸腔聲調與連音變調

國語與客語都是聲調語言，同樣的拼音配上不同聲調會產生不同的意義，國語使用了五種聲調，而客語四縣腔使用了六種聲調，分別為：陰平、陽平、上聲、去聲、陰入、陽入，客語海陸腔則比四縣腔再多使用一種聲調陽去。下表三為客語四縣腔聲調符號表、下表四為客語海陸腔聲調符號表。

表三、客語四縣腔聲調符號表

調類	陰平	陽平	上聲	去聲	陰入	陽入
調值	24	11	31	55	21	5
調型	fu˴	fu˅	fu˪	fu	fug˪	fug
例字	夫	扶	虎	富	福	服
近似國語聲調	2 聲ˊ	3 聲ˇ	4 聲ˋ	1 聲		
音檔調號	2	3	4	1	2	5

表四、客語海陸腔聲調符號表

調類	陰平	陽平	上聲	陰去	陽去	陰入	陽入
調值	53	55	24	11	33	5	2
調型	fu˪	fu	fu˴	fu˅	Fu⁺	fug	fug˪
例字	夫	扶	虎	富	護	福	服
近似國語聲調	4 聲ˋ	1 聲	2 聲ˊ	3 聲ˇ			
音檔調號	4	1	2	3	5	5	2

同樣的客語也會有連音變調的問題，四縣腔可以歸納出三種連音變調的規則，如表五所示；而海陸腔則可以歸納出兩種連音變調的規則，如表六所示。

表五、客語四縣腔連音變調規則

規則1：由兩個陰平字構成的字彙，讀時前字變調讀陽平 陰平（ˊ） + 陰平（ˊ） → 陽平（ˇ） + 陰平（ˊ）			
範例	詞彙	變調前之拼音	變調後之拼音
	新衫	xinˊsamˊ	xinˇsamˊ
	買新衫	maiˊxinˊsamˊ	maiˇxinˇsamˊ
規則2：陰平與去聲構成的詞彙，讀時前字變調讀陽平 陰平（ˊ） + 去聲 → 陽平（ˇ） + 去聲			
範例	詞彙	變調前之拼音	變調後之拼音
	針線	ziimˊxien	ziimˇxien
	拿針線	naˊziimˊxien	naˇziimˇxien
規則3：陰平與陽入字構成的詞彙，讀時前字變調讀陽平 陰平（ˊ） + 陽入 → 陽平（ˇ） + 陽入			
範例	詞彙	變調前之拼音	變調後之拼音
	音樂	imˊngog	imˇngog
	聽音樂	tangˊimˊngog	tangˇimˇngog

表六、客語海陸腔連音變調規則

規則1：上聲變調 即低聲調上聲（ˊ）後面不論接什麼調時，皆要變為中平調陽去（⁺）。			
範例	詞彙	變調前之拼音	變調後之拼音
	打球	daˊkiuˊ	da⁺kiuˊ
	解決	gaiˊgied`	gai⁺gied`
規則2：陰入聲變調 即高入調陰入聲後面不論接什麼調時，皆要變為低入調陽入聲（`）。			
範例	詞彙	變調前之拼音	變調後之拼音
	目珠	mug zhu`	mug`zhu`
	八字	bad sii⁺	bad`sii⁺

第三章　詞典、語音檔及語言模型介紹

詞典對於輸入法而言是最基礎且不可或缺的部分，且若是讓輸入法自動選字，詞典是直接影響音轉字正確率的部分。在這一章節中我們分別介紹輸入法所使用的國客語對照詞典、輸入法邊打邊唸功能所使用的語音檔、及輸入音轉字過程所使用到的語言模型。

3.1　國客語對照詞典

輸入法所使用的詞典為國客語對照的詞典，每一筆客語詞都有其對應的國語詞。拼音的部份我們則是使用教育部所制定客語拼音方案提出的四縣腔與海陸腔的拼音為標準，再額外加入客委會辭典中使用到不包含在客語拼音方案的拼音。最後使用的四縣腔拼音總共有 688 種，而海陸腔拼音總共有 789 種。因發音功能中的唸出客語詞部分需要使用到有聲調的拼音，因此詞典中的拼音需要包含聲調的部份。去掉有亂碼部分的詞，四縣腔的部份總共收錄了約三萬三千個詞目，海陸腔部分則收錄了約兩萬七千七百個詞目。其中同一客語詞有多種拼音的部份則會分別收錄成多筆的詞目來儲存。。表八為四縣腔各字詞的詞數分布。

表七、國客語對照詞典範例

欄位	內容
客語詞	暗晡
國語詞	夜間
拼音	am1 bu2
分數	319

表八、客語詞數分布

字詞	四縣腔個數	海陸腔個數
1 字詞	4952	5522
2 字詞	18043	14399
3 字詞	6175	4654
4 字詞	3948	2856
5 字詞	275	208
6 字詞	80	51
7 字詞	67	33
8 字詞	15	5
總計	33555	27728

3.2 客語語音檔

我們設計的輸入法具有邊打邊唸的功能及唸出客語詞，因此我們需要使用到客語四縣及海陸腔的語音檔。四縣腔的語音檔為實驗室委託熟悉客語四縣腔的陳婷芳老師(台中市北屯區陳平國小)錄製的客語四縣腔基本合成單元，以客語單音節為單位，包含了四縣腔的六種聲調，總共錄製了 2427 個基本合成單元。海陸腔的語音檔則是委託熟悉海陸腔的詹益雲老師(新竹縣芎林鄉五龍國小退休)錄製，同樣也是以客語單音節為單位，包含了海陸腔的七種聲調變化，總共錄製了 3005 個基本合成單元。

因為本論文提出的輸入法以無聲調進行輸入，因此邊打邊唸功能，四縣部份我們只需使用去聲聲檔，某些沒有去聲的拼音則使用陽入聲檔；海陸部分我們使用陽平聲檔，沒有陽平的拼音則使用陰入聲檔。而唸出客語詞的功能則會將變調後的客語詞唸出來，因此我們需要使用到所有錄製的合成單元。四縣腔與海陸腔總共使用了 5432 個音檔，錄製格式為：11025Hz、16bits，儲存成 Windows PCM 格式(wav 檔)。

3.3 語言模型

我們設計的客語輸入法是會讓輸入法在音轉字過程中自動選字的，因此這一部分需要使用到語言模型來決定要選哪個字詞。且因為

在客語語料缺乏的情況下，我們沒有足夠的語料來訓練客語的語言模型。考量到我們的詞典是國客對照辭典及中文語料充裕的情況下，因此我們的做法是使用客語詞所對應的國語詞到中文語料中去訓練 Uni-gram 的客語語言模型。

關於訓練客語語言模型所使用到的中文語料其來源有二：

(1)中文 13 萬詞詞典：其詞典是以中研院的八萬詞(ASCED)為基礎，再從中研院的平衡語料庫(ASBC,3.0)加入未收錄在八萬詞中的詞。

(2)Chinese GigaWord 3.0：此語料庫是從許多單位收集新聞語料而來的，並且由 Linguistic Data Consortium (LDC)發行。我們使用的版本為 2007 年所發行的第三版，其新聞語料主要來源為：Central News Agency of Taiwan、Xinhua News Agency of Beijing、Zaobao Newspaper of Singapore。因為我們需要繁體中文的語料，所以我們只使用到台灣中央通訊社的部份，總共約五千多萬行句子。

透過上述的兩種中文語料，蒐集客語詞對應的國語詞之詞頻後，因為兩種語料大小差異很大，為了平衡兩種語料的影響，因此我們先將統計出來的詞頻加一，再取 log 以二為底，最後將兩個分數相加起來乘十再將此分數無條件進位取整數。因為我們再計算分數時需要將分數相乘，因此分數不能有零分的情況，所以再將加起來的分數全部都加一。

第四章　優客輸入法(U-Hakka)實作

我們要實作出的「**優客輸入法(U-Hakka)**」，所研究開發平台係以OpenVanilla[2]架構為基礎，開發工具則是使用 Visual Studio 2010版，將分別簡介OpenVanilla、本輸入法所要設計的功能及介面介紹、以及與其他目前存在的客語輸入法比較與分析。

4.1　系統架構

關於輸入法的系統架構，我們延續過去幾位相關研究開發台語無聲調輸入法[10]、國台語無聲調拼音輸入法[11]以及中文無聲調拼音輸入法[12]所使用的架構，使用的輸入法架構是以 OpenVanilla 作為基礎 OpenVanilla 為一套輸入法及文字輸出的處理架構，且為開放原始碼的自由軟體。

但是，OpenVanilla 官方已經不在繼續提供及維護 Windows 的版本了，只有繼續維護 Mac 的版本，不過還是能找得到過去在 Windows 上開發的原始碼及 OpenVanilla 程式。因為現在大家普遍使用 64 位元的作業系統，因此我們不使用過去學長使用的 OpenVanilla 0.7.2 版，改使用了能在 Windows 32 位元及 64 位元作業系統上穩定使用的 OpenVanilla 0.8.1 版。

OpenVanilla 主要組成的三個元件為：(1)與系統平台溝通的模組載入器(loader)，這部份再安裝 OpenVanilla 時即會安裝在系統上，輸入法的開發者無需去更動此部份。(2)輸入法與文字輸出模組(module)，輸入法即是我們要開發的部份，針對按下什麼按鍵組字窗及候選詞窗要做出什麼資訊做設計；而文字輸出模組則是輸入法將文字送入到程序之前可以額外選擇是否要呼叫的模組，例如：繁體轉簡體的模組。(3)輸入法的偏好設定程式(preferences)。

上述這部份開發輸入法時也需要去設計，可以透過偏好設定程式來改變輸入法的一些參數設定，例如：是否開啟某項功能，或者調整門檻值的動作。這些即是組成 OpenVanilla 的主要元件。

4.2　輸入法功能與介面

在這一節中，我們將介紹本論文提出的輸入法之各項功能與在系統上介面的呈現方式，也會說明我們設計這項功能的想法。以及將輸入法流程以流程圖表示。

- **拼音輸入**

組字窗與候選詞窗(圖二)是組成一個輸入法最基本的元件，我們的客語輸入法在輸入拼音與音轉字的過程都會在組字窗中進行，如同我們常使用的新注音輸入法一樣，要按下 Enter 鍵後才會將組字窗的內容送到程序中。但是不同的地方在於新注音需要輸入聲調，且輸入聲調的動作會隨即將此拼音與聲調送入音轉字；而我們的輸入法考量到大多數使用者對客語聲調不熟悉，因此我們在輸入拼音時不需要輸入聲調，且會在每輸入一個拼音字母的動作中進行音轉字，會這樣做是考量到使用者可能對客語拼音較不熟悉，若使用者需要按下空白音轉字後才能得知此拼音能得到什麼字，對於不熟悉客語拼音的使用者較不友善。而我們的輸入法在輸入拼音時列出的候選字，可以供使用者參考再按下空白鍵選取，讓輸入法後續會拿此拼音音轉字(第五章會有音轉字過程介紹)，或者直接按下對應的數字選取，輸入法之後就不會自動更動此字。圖二為輸入客語拼音「ab」後組字窗與候選字窗的內容。

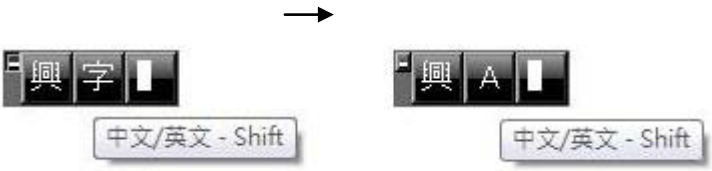
圖二、輸入客語拼音「ab」後組字窗內容

(上)和候選詞窗內容(下)

- **切換拼音與英數輸入**

因為我們輸入法在拼音輸入時，是以小寫英文字母來進行拼音輸入，而若是第一字母為大寫則是要啟動便捷輸入模式來進行輸入。因此我們必須考量到使用者要如何不切換輸入法來輸入英文字母，關於這一部分我們使用 OpenVanilla 架構下提供以 Shift 鍵切換拼音與英數輸入(圖三)的功能，如同我們常使用的新注音輸入法，按下 Shift

鍵後即會把組字窗內容送出以及切換到英文輸入模式下。圖三為切換拼音與音數輸入的工具列。



圖三、切換拼音與英文輸入

- **標點符號輸入**

標點符號的輸入對於使用中文輸入法而言也是必要的功能，而我們的做法是，按下對應的按鍵就能直接輸入全形的標點符號。因為我們輸入法使用的拼音為英文字母，因此在一般使用的鍵盤符號上皆沒有衝突的問題。但考慮到「・」這符號在輸入中文時較不常使用，因此實作上我們將按下此按鍵輸入的符號改為中文較常使用到的「。」句號。而除了直接按下按鍵輸入符號外，也提供了按下 Ctrl+Alt+, 來呼叫完整標點符號清單(圖四)的功能，熱鍵部分同新注音呼叫標點符號鍵盤，較能符合常使用新注音使用者的習慣。

圖四、標點符號清單

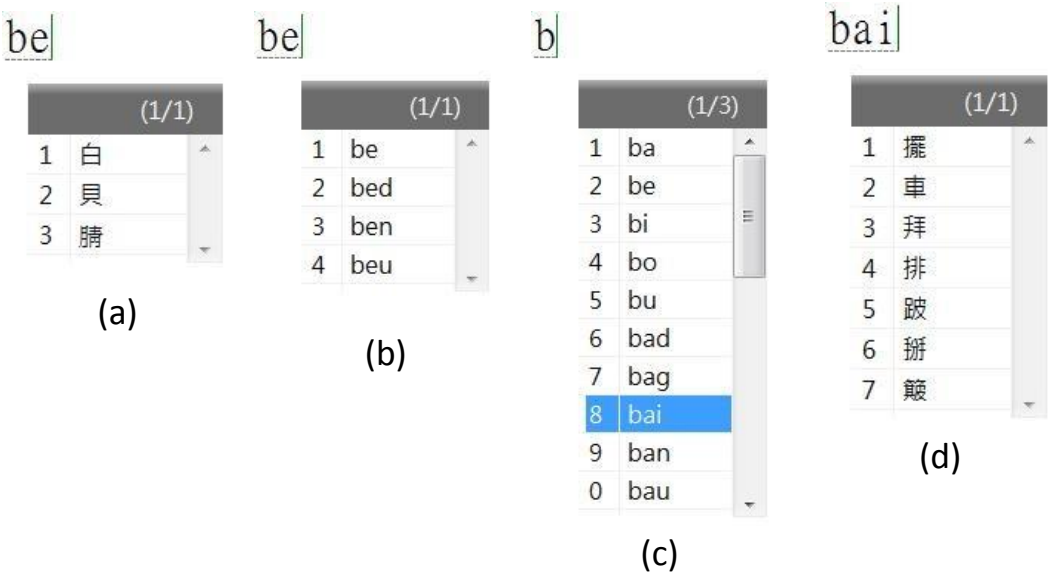
- **便捷輸入模式**

除了一般的拼音輸入模式外，輸入法另外提供一種方便且快速的輸入模式，稱為「便捷輸入模式」。我們是利用輸入拼音時，第一個字母為大寫還是小寫來區分便捷輸入模式與拼音輸入模式。在此模式

中提供了四種輸入不同的輸入方式分別為：自訂輸入、音首輸入、縮寫輸入及英文詞輸入。自訂輸入為讓使用者能快速輸入自己在自訂辭典中加入的詞，音首輸入為只需要輸入客語詞拼音各音節中的第一個字母即可得到該客語詞，縮寫輸入為輸入組織、學校或公司的縮寫即可得到該組織、學校或公司的名稱，而英文詞輸入即為輸入英文詞，輸入法會顯示對應的中文詞供使用者選取。詳細的輸入流程我們在第六章會介紹。

- **拼音錯誤提示**

考量到大多數的使用者對客語拼音可能不是很熟悉，而且客語拼音方案可能也會持續更新，因此我們試圖讓使用者在輸入拼音時，能得到輸入法的額外輔助拼音的輸入。除了前面提到在輸入每個拼音時，若音節為合法的客語拼音，候選詞窗就會列出可能字外，另一項功能為拼音錯誤提示。拼音錯誤提示會在使用者按下錯誤的拼音時，產生提示聲且將還有哪些可能的客語音拼顯示在候選詞窗中，供使用者尋找是否有要的拼音來選取。圖五為使用者欲輸入詞的拼音為「bai」，但使用者記錯成「bei」，圖五(a)為輸入至「be」時候選詞窗顯示「be」的候選詞、圖五(b)為繼續輸入「i」造成拼音錯誤，呼叫錯誤提示功能將「be」的後續拼音列在候選詞窗中供使用者選取或參考、圖五(c)呼叫錯誤提示功能後，按下 Backspace 刪除一個拼音，候選詞窗會顯示目前拼音的後續拼音、圖五(d)使用者以選取後續拼音的方式，將組字窗改為拼音「bai」。



圖五、拼音錯誤提示功能的實例

- **發音功能**

  我們的輸入法提供了兩種的發音功能，如下表示：

  1. **拼音邊打邊唸**：我們希望讓使用者不只是用看的來得知是否輸入錯誤，也能用聽的來得知是否有輸入錯誤。因此在輸入拼音時，若輸入音節為合法客語拼音，即會將此拼音唸出。

  2. **唸出客語詞**：我們希望輸入法也能提供數位學習的功能，因此當使用者在組字窗尾端輸入完拼音後，隨即會呼叫出選單供使用者選取是否唸出此詞。詳細流程於第七章介紹。圖六為使用者輸入完客語詞後呼叫出唸出此詞的選項供使用者選取。

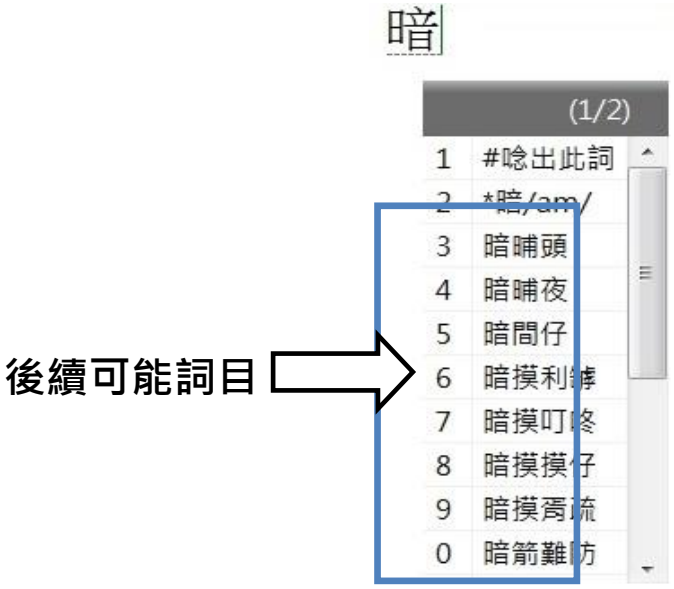
圖六、唸出客語詞選項

- **提高常用詞之優先順序**

輸入法在自動選字時，會依據訓練的語言模型結果，來將詞作排序的動作，且會選取分數最高的詞。但我們訓練模型的方式是利用客語詞對應的中文詞去訓練，效果必定會比用客語詞在大量的客語語料中訓練來得差。因此使用者修改的結果是很重要的資訊，我們會將使用者選取的詞分數進行調整，使得使用者常使用的候選詞能夠顯示在候選詞列的前面甚至第一位。分數調整的流程於第七章介紹。

- **往後預測可能詞目**

我們希望在輸入的過程中，能列出後續可能的詞目供使用者選取，若是使用者欲輸入較長的詞，直接透過後續詞目預測的結果選取，則可以省下很多輸入的時間。而預測的門檻值我們設定為多出 2 字後的詞，因為若列出 1 字後的詞供使用者選取，對於節省的時間幫助不大。使用者也可以透過輸入法的偏好設定，依照自己的習慣來設定門檻值，對客語較不熟的使用者可以將門檻設為 1，這樣可以得到較多的候選詞目。往後預測可能詞目詳細的流程於第七章介紹。圖七為往後預測詞目且門檻值設為 2 的情形。



圖七、往後預測可能詞目

- **國語與拼音選項**

考慮到客語中有很多平常國語不常見到的字，而且客語使用的字也沒有訂定的標準字，對於不是以客語為母語的人甚至是會說客語但

不常閱讀客語文章的人，閱讀非常不易。圖八為擷取自教育部電子報「閱讀越懂閩客語」客語文章中的一段。而我們希望能讓輸入法寫出一篇可讀性較高的客語文章，能讓較不熟悉客語字詞的使用者也能看懂與學習。

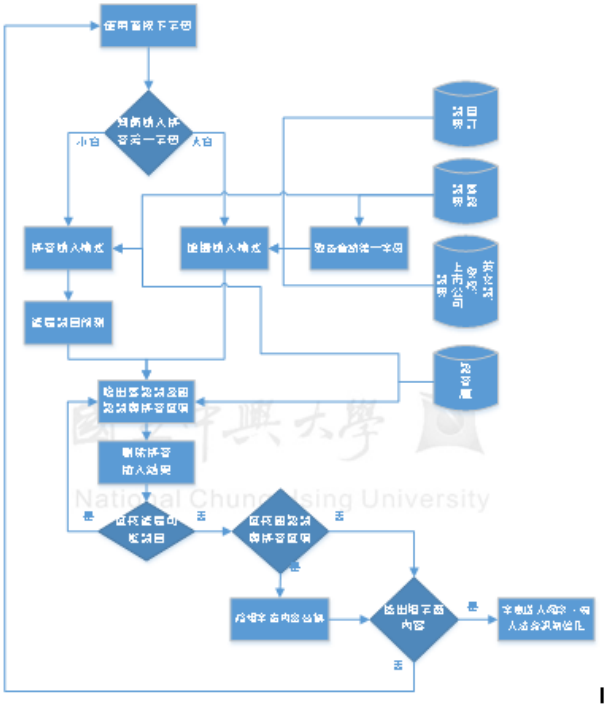
平常時佢無麼个肯細人仔食糖仔，一來驚蛀牙，二來驚食飯毋落。毋過，見擺佢兜兩子阿公去街路寮轉來，就攞到大包細包。細人仔有阿公好靠勢，嘴項食等糖仔、面項笑微微仔，連行路就風神挨挨。

圖八、「閱讀越懂閩客語」客語文章中一段

因此我們的做法是在輸入時，能讓輸入法加註國語與拼音，如此一來就能讓客語文章更具可讀性且不需要額外再解釋某些用詞的意義。我們在第七章中會有詳細的介紹及使用此功能將客語文章重新輸入的例子。圖九為客語詞「暗晡」的加註國語與拼音選項。

圖九、加註國語與拼音選項

- 優客語輸入法流程圖(如下圖所示)



4.3 臺灣現有客語輸入法簡述

與閩南語輸入法不同，現有客語輸入法非常的少，目前現有的客語輸入法只有二種，一為信望愛客語輸入法，另一為教育部台灣客家語拼音輸入法。在這一節中我們會先分別介紹這兩種客語輸入法，最後在與我們再輸入法做比較與討論。

4.3.1 信望愛客語輸入法

信望愛客語輸入法[5]為財團法人信望愛資訊文化藝術基金會所開發的客語輸入法，使用的拼音系統為台羅拼音或教羅拼音。信望愛客語輸入法能以數字鍵 1-9 選擇詞及以 Shift 鍵+1-9 選對應的台羅或教羅拼音，也有提供自訂辭典、音首輸入及加註拼音的功能。

拼音輸入的部分，可以一次輸入一個單字或者一次輸入一個詞，但不支援輸入法自動選字，且單字輸入需要加上聲調。而輸入一個詞時拼音不能有任何錯誤，否則就無法正確音轉字，因此使用者需要猜此詞是否存在於辭典中，且辭典內大部分的詞都用不到，使用者可能使用到的只有少部分的常用詞。

4.3.2 教育部台灣客家語拼音輸入法

教育部台灣客家語拼音輸入法[6]為教育部委託廠商製作的客語輸入法，使用的拼音系統為遵照客家語拼音方案制定的各腔調拼音。且收錄六種腔調的拼音輸入，包含了四縣、海陸、大埔、饒平、詔安、南四縣。也有提供自訂辭典的功能以及供使用者修改辭典。但輸入法只有基本的功能，較沒有提供額外的功能。

拼音輸入的部分，如同舊版注音輸入法，只能輸入單字的拼音來得到客語單字，在輸入完單字後會列出後續詞目。在輸入單字拼音時需要加上聲調，也可以以調號 0 號作為查詢功能。

4.3.3 比較與討論

這裡我們將本論文提出的輸入法與現有的客語輸入法做詳細比較與討論，如表九所表示。

表九、客語輸入法比較

	信望愛客語輸入法	教育部台灣客家語拼音輸入法	本研究輸入法
輸入拼音	台羅、教羅	教育部客家語拼音	教育部客家語拼音
輸入聲調	需要	需要(附註 1)	不需要
輸入方式	一次一字或詞	一次一次或詞	組字視窗自動選字
自訂詞典	有	有	有
音首輸入	有	無	有
縮寫及英文詞輸入	無	無	有
詞組輸入	有	無	無(附註 2)
拼音輸入錯誤提示	無	無	有
邊打邊念	無	無	有
提高常用詞之優先順序	有	無	有
往後預測可能候選詞目	無	有	有
選擇國語詞彙	無	無	有
加註拼音	有	無	有
萬用字元	有	有	無(附註 3)

附註:

1. 教育部台灣客家語拼音輸入法提供聲調 0 作為查詢模式
2. 因為詞組輸入時使用者需要猜此詞是否在詞典內,我們認為能被音首模式取代
3. 我們提供的拼音錯誤提示功能,可以取代萬用字元來輔助拼音不熟悉的使用者進行輸入

關於詞組輸入的部分,我們提出的輸入法沒有加入此功能的原因是,我們認為此功能可以被音首輸入取代。因為在詞組輸入時,需要連續輸入多個音節,而輸入長詞時,其中一個拼音打錯了,會造成整個拼音都錯誤,會耗費許多時間。而另一個更重要的原因為詞組輸入時,使用者需要猜此詞是否存在於辭典內,而既然要猜此詞是否在辭典內,倒不如使用音首輸入來尋找即可,音首輸入也可以避免因為某個拼音字母錯誤,而無法正確音轉字的情形。

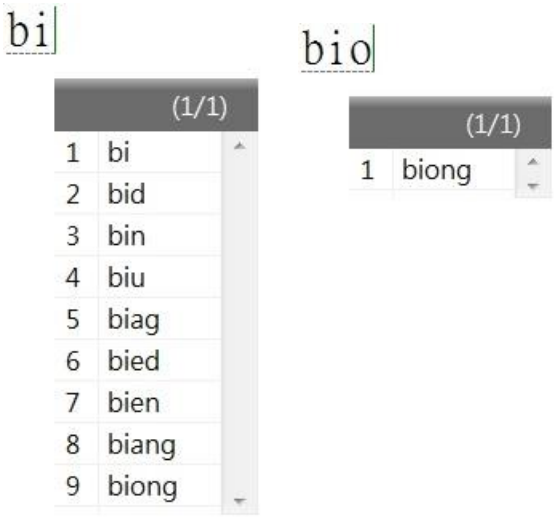
而萬用字元為輸入拼音時,可以以「*」符號來表示接任意拼音皆可的功能,例如輸入「a*」會列出所有以 a 開頭的客語單字。我們認為我們的輸入法的輸入模式及拼音錯誤提示功能,能夠輔助使用者來選取拼音,且因為每個拼音對應到的字數已經不少了,再將範圍擴大對使用者來說尋找要的字會更困難。因此我們認為列出可能的拼音給使用者參考比起列出所有的字還來得有效果。

第五章　拼音輸入模式

在這一章節中，我們會詳細介紹本輸入法在拼音輸入模式下拼音錯誤提示功能的流程、音轉字過程使用的少詞優先演算法、以及輸入法自動選字後，使用者對於結果不滿意而去修改結果的流程。

5.1　拼音錯誤提示

拼音錯誤提示功能會在輸入拼音時出錯時啟動，而拼音輸入指的就是在組字窗以小寫字母進行輸入。假設目前候選字窗中拼音資訊$P = C_1, C_2, \cdots, C_n$，其中$C_1, C_2, \cdots, C_n$為小寫英文字母，使用者接著輸入$C_m$至P中，此時輸入法會將目前拼音資訊$P = C_1, C_2, \cdots, C_n, C_m$送去檢查使否為合法拼音，若P為合法拼音則候選字窗會更新為詞典中與P相同的單字詞；若P不為合法拼音則此時會呼叫拼音錯誤提示功能，且此時會有兩種情況，第一種情況為：若是接著輸入也沒有合法拼音，錯誤提示功能會先產生系統音效「叮」來提示使用者，且造成錯誤的拼音字母C_m會被刪除，候選窗會更新為合法拼音中前 n 個字母與目前輸入的拼音資訊$C_1, C_2, \cdots, C_n$相符的拼音，且這些拼音會由短至長、及字母順序排序，以供使用者參考或選取；第二種情況為：若接著輸入還會有合法拼音，此時並不會產生「叮」也不會將使用者輸入的C_m刪除，但也會將候選窗更新為合法拼音中前 n+1 個字母與目前輸入的拼音資訊$C_1, C_2, \cdots, C_n, C_m$相符的拼音，讓使用者可以快速來選取。圖十左圖為拼音資訊$P = bi$時，再按下造成錯誤的拼音字母「m」後，呼叫拼音錯誤提示功能的組字窗與候選詞窗情形；圖十右圖為拼音資訊$P = bi$時，再按下造成錯誤的拼音但有後續合法拼音的拼音字母「o」，此時並不會將「o」刪除和產生「叮」聲，但候選窗同樣會列出所有可能拼音。



圖十、拼音錯誤提示的兩種情況

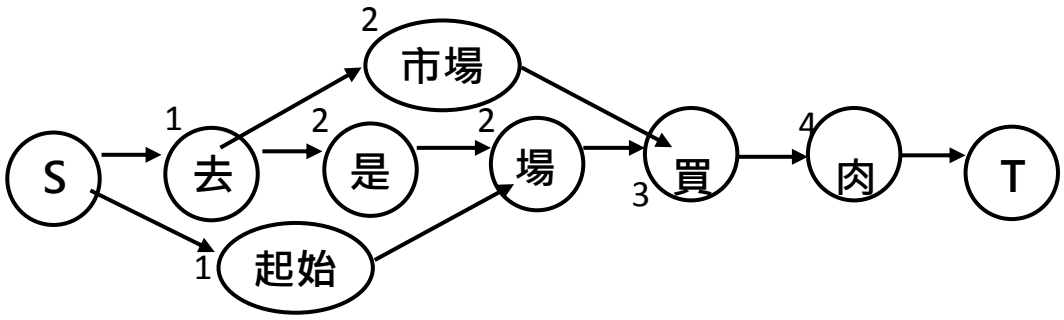
5.2 音轉字處理

音轉字處理是輸入法的核心，關係到輸入法自動選字的正確率，音轉字處理指的就是將使用者輸入的拼音字串轉成對應的客語詞輸出的過程。而每個音節對應到的客語字數，對於音轉字正確率有直接的影響。客語四縣腔音節總數有 672 個，客語海陸腔總共則有 749 個音節，我們詞典中使用到的客語字數約有 7500 個客語字，平均一個音節會對應到約 10 個字。因此對自動選字的輸入法而言，音轉字的處理是最重要的部分。

我們輸入法在音轉字處理中所使用的演算法為少詞優先演算法，搭配 Uni-gram 語言模型來使用，演算法的詞數限制為三個詞，也就是說當音轉字結果出現四個詞時，第一個詞在之後就不會再被送入音轉字演算法中，也可稱為三個詞的少詞優先演算法。我們會將詞數限制為三個詞的原因在第八章音轉字實驗中會提到。少詞優先演算法即為選擇輸入拼音能組合出最少詞的那條路徑，若是有詞數相同的情況時，則依靠 Uni-gram 語言模型計算哪條路徑的機率較高，最後選擇分數最高的路徑作為音轉字結果。計算公式在下方表示，其中W_i為拼音對應到的詞彙，$P(W_i)$為該詞彙在模型中的機率，且因在 Uni-gram 語言模型中的機率即為該詞的詞頻除以總詞數，因此我們可以將分母省去直接相乘即可。

$$\text{路徑分數} = \prod_{i=1}^{n} P(W_i)$$

這裡以一個實際的例子來說明少詞優先演算法的流程，我們將四縣拼音「hi siicongmaingiug」送入音轉字，演算法會先將拼音可能組出的所有詞找出來，然後列出所有可能的路徑，接著找出從 S 到每個節點的最短路徑也就是最少詞的情況，如圖十一所示。此例子計算到「買 mai」這一節點時，可以看出最少詞數的路徑有兩條，分別是「去 市場」及「起始 場」同樣都是兩個詞，因此這時要靠 Uni-gram 語言模型計算分數，比較這兩條路徑的分數後，最後選擇分數高者「去市場」作為到走到「買」的路徑，以此方式繼續走到結點 T 為止，即可得到最少詞且分數最高的路徑作為結果。



圖十一：少詞優先演算法例子

少詞優先演算法的虛擬碼如下表示：

先宣告一結構 Word，其中包含了 string word 用來存放客語詞，string pinyin 用來存放客語詞對應的拼音，double score 用來存放客語詞的分數。

假設 n 為送入音轉字的音節數,接著依照音節個數建置一個 n*n 二維的 Word 矩陣 Graph，表示某音節至某音節是否能組成一個詞。因為音轉字是有次序的由左至右轉換，也就是單向且沒有迴圈的圖，因此矩陣只會使用到右上角三角形部分，其餘部分填入 0 值表示沒有路徑。而右上角三角形的內的值則搜尋客語辭典中，是否有音節相符的客語詞，若有存在則找出分數最高的客語詞，並將分數填入矩陣中，若不存在則填入 0 值。

宣告長度為 *n* 的一維矩陣紀錄每個音節的最短詞數 shortestWords、最高分數 greatestScore、及從哪一音節走來 shortestPath。儲存最高分數的資料型態需宣告為 double 以避免 overflow，其餘則使用 int。

宣告一個 Queue 存放要尋找路徑的音節。

接著由第一個音節開始，將第一音節加入 queue 中，且將開始音節的 shortestWords 設為 0、greatestScore 設為 1。

```
while Queue != empty
    q 為 Queue 中要處理的音節
    for each path in Graph[q] //尋找陣列 Graph,找出此音節可能的路徑
        if path.score> 0 //代表此音節至該音節能組成詞
            if ( shortestWords[q]+1 <shortestWords[path] ) OR( shortestWords[q]+1 = shortestWords[path] AND greatestScore[q]*path.score>greatestScore[path] )
                shortestWords[path] = shortestWords[q]+1
                shortestPath[path] = q
                greatestScore[path] = greatestScore[q]*path.score
                Queue.push(path)
```

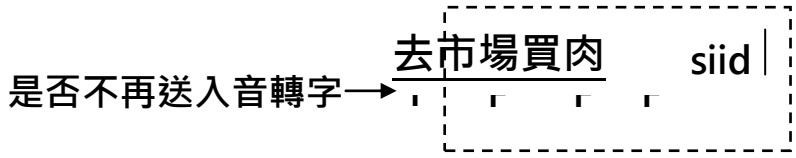
宣告一個 stack 用來儲存音轉字結果

```
While shortestPath[n] != 0
    stack.push(Graph[shortestPath][n].word)
    n=shortestPath
```

上述表示「少詞優先演算法」之運算過程，進一步需要考慮送入詞數的問題，因此以輸入法記錄下的組字窗內容結構來實作三個詞的少詞優先演算法。組字窗中的內容雖然看不出斷詞的狀況，但輸入法是以詞為單位表示組字窗內容的，且每個詞都會由一個布林值記錄此詞是否會被音轉字演算法自動修改。因此我們的做法是在組字窗尾端輸入拼音按下空白音轉字時，會由最後一個標記 true 的詞(若無標記 true 的詞則將所有拼音送入)往後拿出所有詞的拼音與現在輸入的拼音一起送入音轉字。

若音轉字結果出現四個詞時，會將第一詞設為 true，也就是不

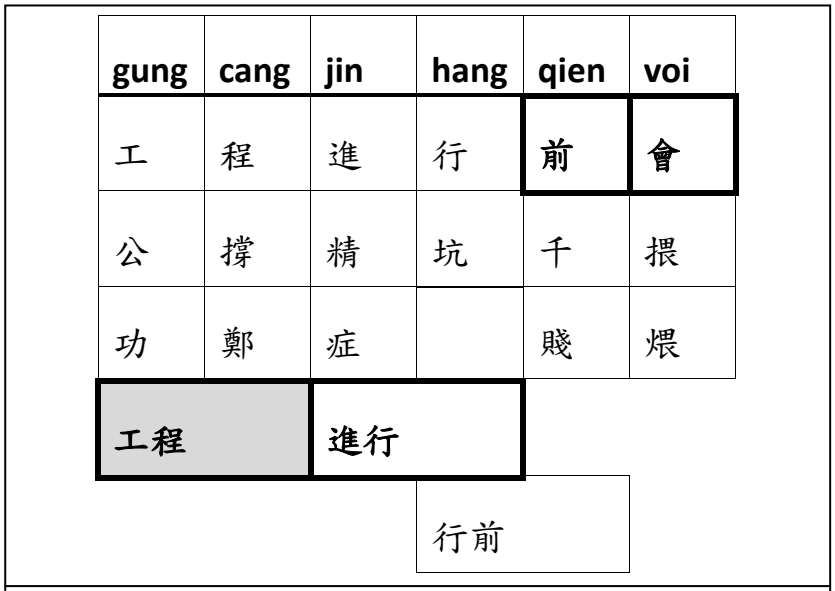
再送入音轉字，且把音轉字結果加回組字窗尾端。圖十二以剛剛「hi siicongmaingiug」「去市場買肉」為例子，音轉字的結果有四個詞，因此第一詞之後不再送入音轉字。此時輸入 siid 送入音轉字時，會由標記 true 的詞「去」之後的拼音「siicongmaingiug」與「siid」一起送入音轉字。



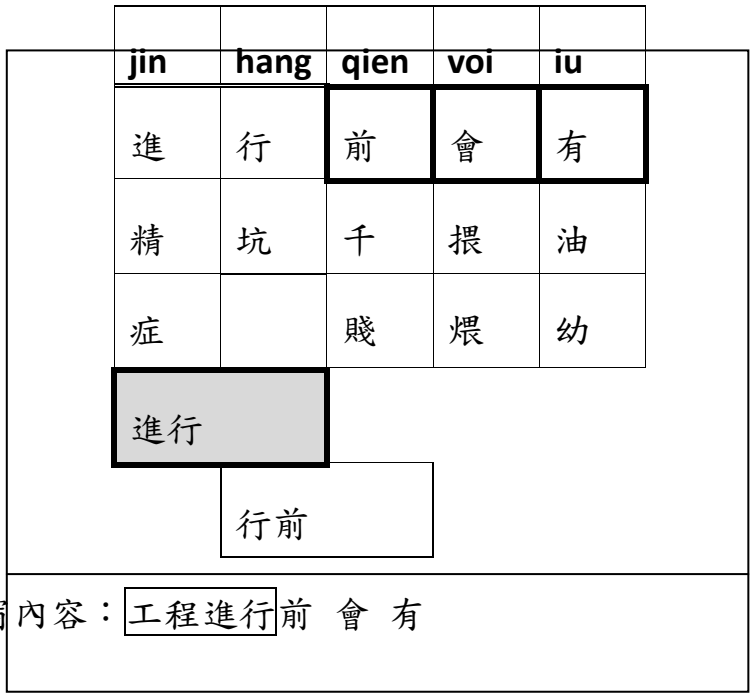
圖十二、三詞少詞優先例子

標點符號在音轉字的部分也會標記是否不再送入音轉字，因為字與字之間有隔著標點符號，可以得知這兩個字不會是一個詞，因此標點符號都會標記成 true，表示不再送入音轉字。

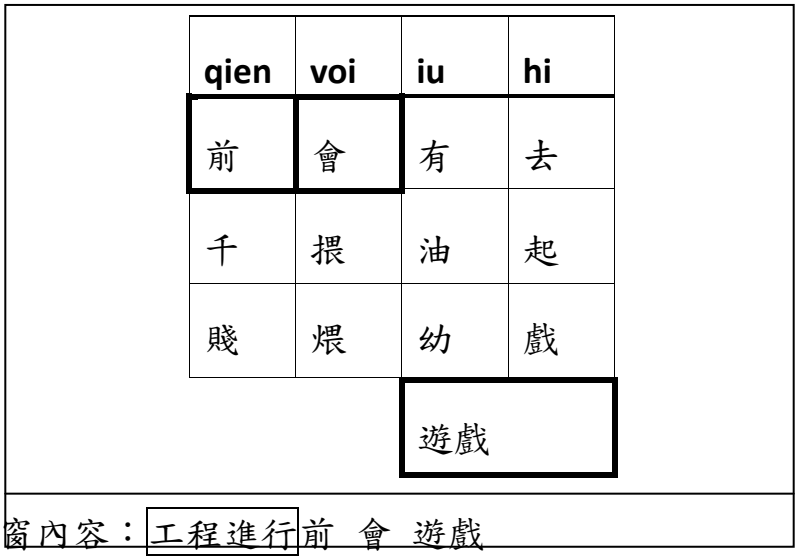
這裡我們再舉一個完整的例子，來模擬使用者輸入的情況。假設使用者欲輸入一短句「工程進行前會有起工儀式」其拼音為「gung cangjin hang qienvoiiu hi gung ngisiid」，我們跳過前五個拼音輸入的過程，因為輸入前五個拼音還沒達到音轉字演算法的詞數上限，因此從拼音 voi 開始，其輸入過程如圖十三至十八表示。



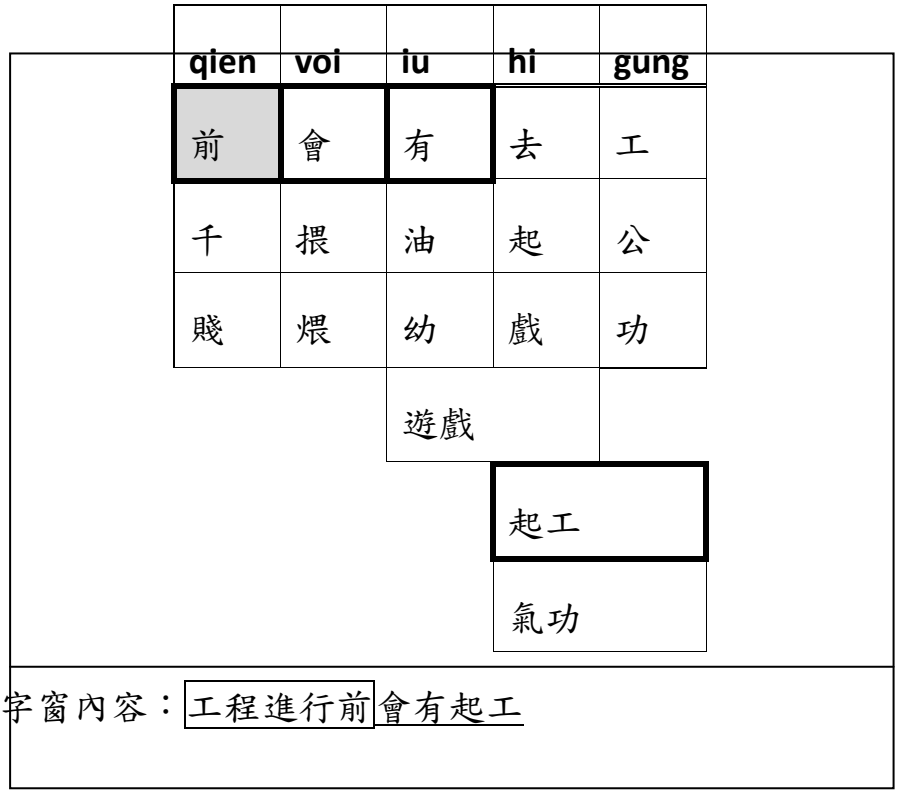
圖十三、音轉字過程一：當輸入到拼音 voi 時，音轉字的結果為「工程進行前會」四個詞，因此「工程」這一詞會固定下來。



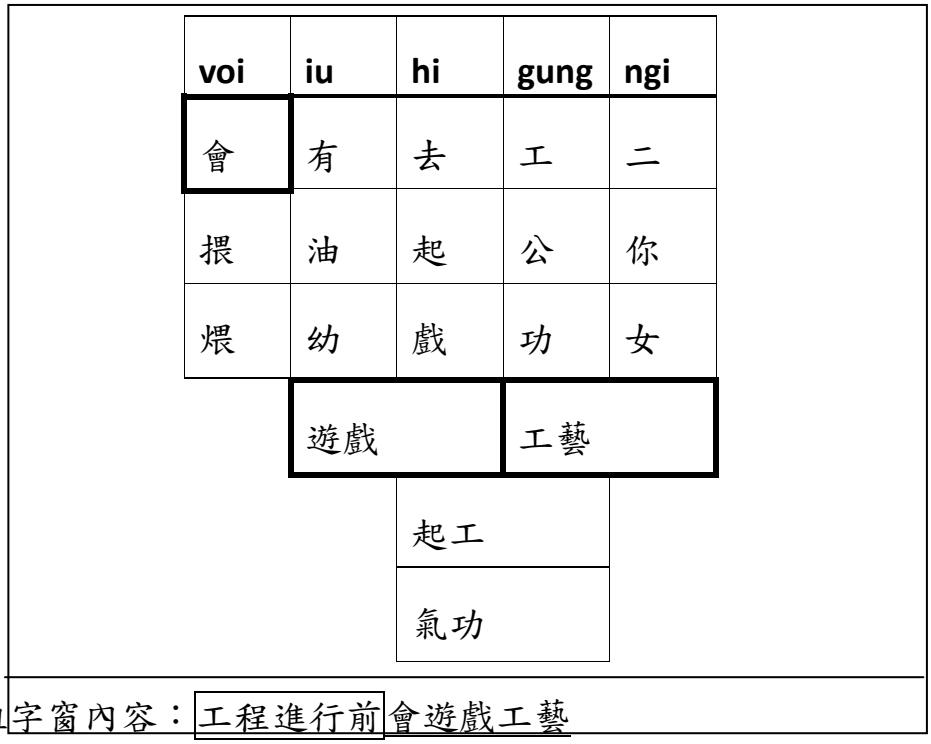
圖十四、音轉字過程二：當輸入到拼音 iu 時，音轉字的結果為「進行前會有」四個詞，因此「進行」這一詞會固定下來。



圖十五、音轉字過程三：當輸入到拼音 hi 時，音轉字的結果為「前會遊戲」三個詞，因此並不會將「前」固定。

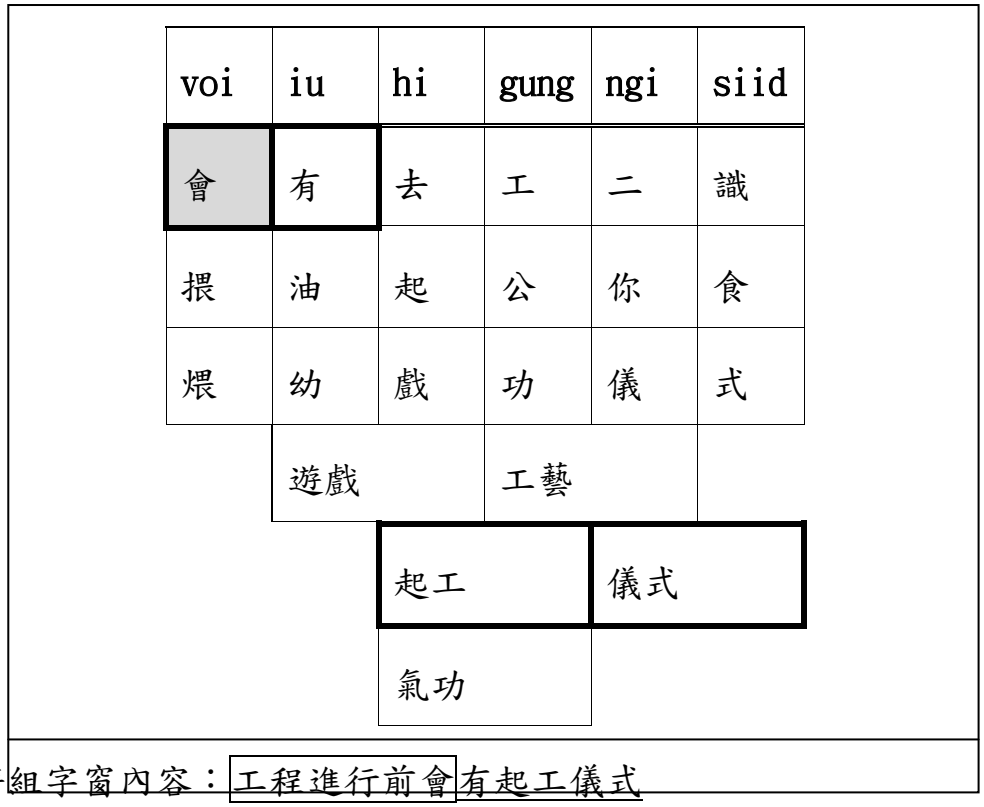


圖十六、音轉字過程四：當輸入到拼音 gung 時，音轉字的結果為「前會有起工」四個詞，因此「前」這一詞會固定下來。



此時組字窗內容：工程進行前會遊戲工藝

圖十七、音轉字過程五：當輸入到拼音 ngi 時，音轉字的結果為「會遊戲工藝」三個詞，因此並不會將「會」固定。



此時組字窗內容：工程進行前會有起工儀式

圖十八、音轉字過程六：當輸入到拼音 siid 時，音轉字的結果為「會有起工儀式」四個詞，因此「會」這一詞會固定下來，此時使用者也已經輸入

完畢。若是此時按下 Enter 即會將組字窗內容「工程進行前會有起工儀式」送出至使用輸入法的程序中。

5.3 使用者修改功能

在輸入法音轉字的過程中，使用者不一定會對所有音轉字的結果感到滿意，因此必須讓使用者有修改組字窗內容的功能。我們實作的方法為，使用者可以將指標移到欲修改字的前端按下鍵(若欲修改最後一字在組字窗尾端按下鍵也可)因為最長詞的拼音數為八，按下鍵後會由指標處往後找一至八個字的拼音來做音轉字找出所有以第一字開始的可能詞。結果會顯示在候選詞窗中，由詞長至短、分數高至低排序。圖十九為組字窗為「客家花布(hag-ga-fa-bu)」時，將指標移至「客」字前按下下鍵的情形，此時候選詞窗會列出所有可能組成詞，且由長至短、高分至低分排序。

圖十九、使用者按下鍵修改音轉字結果

且修改後若是將原本的長詞拆開，此詞會將拆開後的長詞拼音及往後搜尋至是否送入音轉字被設為 True 的拼音為止，將這些拼音重新送入音轉字。如圖二十所示，將前面三個音轉字結果改成「客家話」後，最後一個拼音也會重新送入音轉字，得到分數最高的單字詞「部」。

圖二十、使用者修改後拼音重新送入音轉字

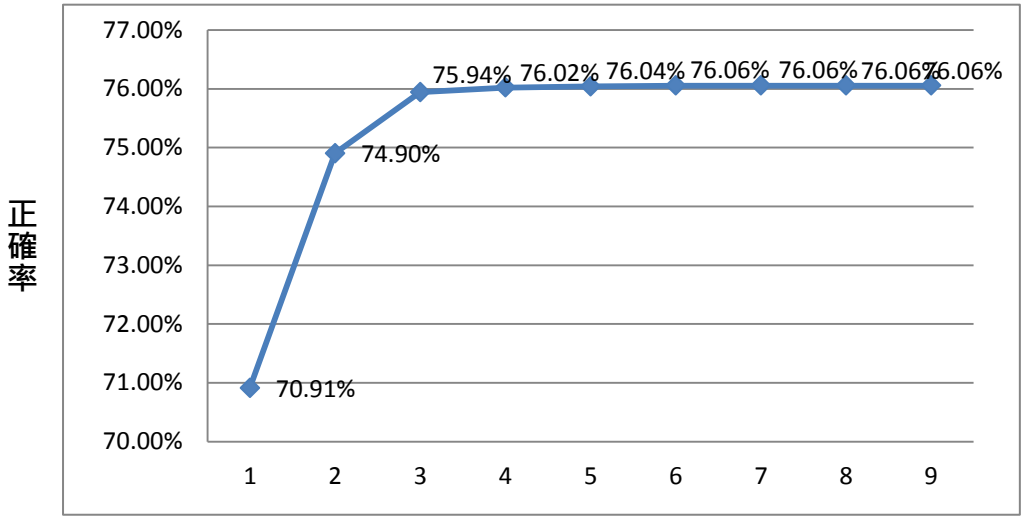
5.4 音轉字實驗

在這一節中，我們將對音轉字所使用的少詞優先演算法進行實驗，目的為找出一個正確率較佳的詞數門檻。

實驗語料的部分我們使用客委會的四縣腔例句以及 101 年客語能力認證基本詞彙-中級、中高級暨語料選粹[7]，總共蒐集了 9309 句四縣腔例句。因為這些例句並不包含對應的拼音，因此我們必須先對這些例句做字轉音的動作，也就是要進行斷詞及標上拼音的動作。我們使用長詞優先方法來進行斷詞，此方法可以以「由前往後」及「由後往前」來斷詞，其結果不一定會相同

有些情況下很難去分辨哪個是正確的斷詞結果，且斷詞不同可能會造成拼音不同的問題。因此我們將「由前往後」及「由後往前」斷詞結果不相同的例句去除，希望降低斷詞造成正確率的誤差。最後我們留下 7697 句四縣腔例句總共有 127885 字，且將拼音標上後，來進行音轉字的實驗。

實驗的流程我們希望模擬如使用者在輸入一般，因此會逐字的進行音轉字，直到輸入完最後一個拼音按下 Enter 送出組字窗內容為止。例如一客語例句：「三層肉(samcenngiug)鹹菜(ham coi)煮(zu)湯(tong)，(punc)味緒(mi si)盡好(qin ho)。(punc)」，將斷詞結果及標點符號標上拼音後，我們會從拼音「sam」開始輸入，直到最後一個「punc」表示標點符號的拼音輸入完後即將結果輸出。其中標點符號也會給音轉字提供資訊，因標點符號前後不可能組成一個詞，因此會做斷開兩邊拼音的動作。且因為標點符號是完全正確不會輸入錯誤的，因此標點符號的字元不列入音轉字結果中正確率的計算。實驗的結果如圖二十一所示。



圖二十一、少詞優先演算法正確率

由上圖的結果可以看出，正確率最高為使用六至九詞少詞優先演算法，正確率皆為 76.04%，不會在提升的原因是我們詞典中收錄最長的詞為八字詞，因此六詞的少詞優先演算法已經幾乎涵蓋了詞典內的詞。而我們實作的輸入法選擇使用三詞少詞優先的原因為：三詞少詞優先已經能應付詞典中大部份的詞，正確率已經達到 75.94%與最高的 76.06%只相差了 0.12%，可能造成錯誤的只有比較長的長詞。
，三詞少詞優先在輸入後面那句「毋好三心打兩意」時，因為輸入到「三心打兩」時每個字皆為單字詞，因此第一字「三」即被固定詞數送入音轉字，也就組不出「三心打兩意」這五字詞了。因此使用者必須將指標移至「三」前面來進行修改得到正確的結果。

但是五字詞中每個字都無法組成詞的情況並不常見，且詞典中五字詞以上的詞目數量也並不多，且我們不希望輸入法修改離組字窗指標處太遠之前的結果，避免拿更多的詞來做音轉字，造成使用者需要移動指標到很前面的結果重新修正的情形。因此我們決定使用三個詞的少詞優先。

第六章　便捷輸入模式

在這一章節中，我們將會介紹本輸入法提出的一種額外輸入模式，我們將它稱為便捷輸入模式。此輸入模式可以提供方便、快速的輸入方式，包含了自訂、音首、縮寫及英文詞，四種輸入方式。為了跟拼音輸入作區隔，我們是以正在輸入的拼音資訊$P = C_1, C_2, \cdots, C_n$的開頭字母$C_1$為大寫還是小寫作為判斷依據，若$C_1$為大寫則呼叫便捷輸入模式；$C_1$為寫小則是一般的拼音輸入模式。此外此輸入模式與一般拼音輸入不同，不須額外按下空白鍵來音轉字得到結果的動作，而是在每次輸入一個字母時，就會將所有前 n 個字母符合拼音資訊$P = C_1, C_2, \cdots, C_n$的詞放入候選詞窗，供使用者來尋找需要的詞做選取。

6.1　自訂輸入

自訂輸入為提供使用者自行去設定任意自訂拼音轉出任意自訂詞的模式，使用者可以透過輸入法偏好設定中優客語輸入法的模組設定頁面，來自訂拼音與詞。自訂輸入的候選詞會放在候選詞窗中的最上端。圖二十二為輸入法偏好設定中，自訂詞典的介面及展示。

圖二十二、自訂詞典介面及自訂輸入展示

6.2 音首輸入

音首輸入的意思是，使用者可以透過輸入詞典內客語詞的各個音節中第一個字母來快速得到此客語詞。這裡我們先觀察詞典中各字詞的音首種類及對應總詞數，結果如表十表示。由表十可以看出詞長為1的單字詞，因為平均每種音首對應到的詞數實在在多，平均需要23頁的候選詞頁才能顯示完畢，因此我們音首輸入模式並不提供單字詞供使用者選取。而詞長為2的兩字詞雖然需要六頁的候選詞頁才能顯示完畢，但我們認為對於較不熟悉客語拼音的使用者而言，以音首輸入來尋找兩字詞是有幫助的，因此我們音首輸入從兩字詞的客語詞開始顯示。

表十、音首平均對應詞數

詞長	種類	總詞數	平均對應詞數	平均所需候選詞頁
1	22	4952	225	23
2	355	18043	51	6
3	2662	6175	2	1
4	3206	3948	1	1
5	256	275	1	1
6	78	80	1	1
7	60	67	1	1
8	15	15	1	1

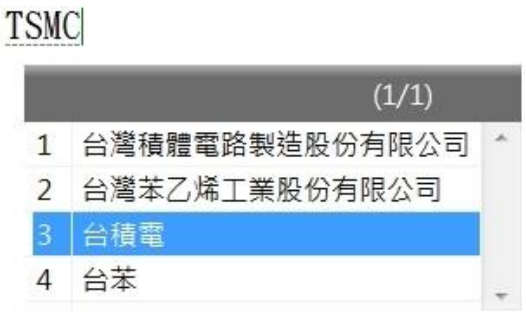
這裡實際舉一個例子：「緊來緊多」其對應拼音為「gin loi gin do」，若使用拼音輸入總共需要按下鍵盤「ginSpaceloiSpaceginSpacedoSpace」共15次，才能得到「緊來緊多」這個四字詞，而若使用者以音首進行輸入只需要輸入「GLGD」四字個字母即可在候選詞窗中找到「緊來緊多」這個詞。甚至使用者只需要輸入「GLG」三個字母，即可找出所有以「緊來緊」為首的詞，列在候選詞窗中，如圖二十三所示。

圖二十三、音首輸入「GLGD」及「GLG」後的候選詞窗內容

6.3 縮寫輸入

考慮到學校或公司名稱往往很長一串，使用拼音來輸入需要耗費較多的時間，因此輸入法提供了讓使用者以縮寫來輸入組織名、我國大學、上市公司的功能。圖二十四為使用者欲輸入「中興大學」，只需要輸入「NCHU」即可在候選詞窗中找到此詞。圖二十五為使用者欲輸入「台積電」，只需要輸入「TSMC」即可在後選詞窗中找到公司完整的名稱或者是縮寫。

圖二十四、縮寫輸入「NCHU」後的候選詞窗內容



圖二十五、縮寫輸入「TSMC」後的候選詞窗內容

6.4 英文詞輸入

對於某些對英文較熟悉的狀況，使用者可能以英文拼出這些詞，會比使用客語拼音拼出來還來得容易。因此我們使用了約 17 萬詞的英中對照詞典，來提供使用者以英文詞輸入得到中文詞的功能。例如使用者想輸入「高爾夫」這個詞，但客語詞典中不一定會有收錄此詞，若沒收錄的話使用者可能必須以單字音節的拼音將此詞打出來，但若是使用者知道此詞的英文拼音「Golf」，則使用者就能以英文詞輸入輕鬆將此詞打出來。且我們會列出前四個字母都符合「Golf」的詞供使用者選取，也就是說某些英文詞較長的單字，使用者只需輸入前幾個字母字，即可從候選窗中挑選到想要的詞，如圖二十六所示。

圖二十六、英文詞輸入「Golf」後的候選詞窗內容

第七章　結論與未來改進方向

本研究實作的無聲調客語拼音輸入法，提出了拼音錯誤提示功能，讓客語拼音的初學者能較快上手。且輸入法具有往後預測可能後選詞目的功能，可以讓較不熟悉客語詞彙的使用者直接選取。對於熟練客語拼音的使用者而言，輸入法的輸入方式是以組字窗自動選字，因此熟練的使用者可以連續輸入多個客語拼音來自動組成客語詞。而自動選字的音轉字演算法為三個詞的少詞優先，搭配以客語詞對應的國語詞訓練出來的模型，能提供約 75.94%的正確率。除了基本的拼音輸入模式還提供了便捷輸入模式的功能，能提高使用者的輸入效率。

我們希望讓使用者輸出得到一篇更具可讀性的客語文章，因此我們在音轉字得到詞後，候選詞窗會列出最後一個詞的國語詞與拼音選項供使用者選取。以在客語詞旁加註的方式，能讓不常讀客語文章的讀者，較快速的看懂整篇客語文章的內容。

且輸入法提供發音的功能，讓使用者在輸入時能聽見自己打了什麼客語音節，還能讓使用者去聽客語詞的唸法，讓使用者在輸入的過程中去學習正確客語詞彙的發音。

至於未來可以改進之方向有幾項如下:

一、語料(Corpus)問題:若是能收集大量的客語語料，來訓練出 bi-gram 的客語語言模型，對於音轉字的正確率應能有效的提升。

二、另外就是詞典收錄的詞目數量:目前的詞目數量並不算多，若能擴大詞典收錄的詞目數，對正確率也會有直接的影響。

三、輸入法功能方面:往後預測可能的詞目這項功能，我們目前是以一個詞為單位來預測，將來可以改為以多個字來進行預測，或許能更貼近使用者想要輸入的詞，以便提高使用者輸入的效率。

四、輸入拼音部份:可以加入相容拼音的功能，讓不衝突的拼音例如四縣腔中:輸入bao也能對應到bau、輸入bian也能對應到bien，讓使用者慣用輸入的那些拼音也能對應到正確的拼音。

附錄一：參考文獻與相關網站

1. 99 年至 100 年全國客家人口基礎資料調查研究, http://www.hakka.gov.tw/dl.asp?fileName=1521131271.pdf
2. 93 年度台灣客家民眾客語使用狀況, http://www.hakka.gov.tw/ct.asp?xItem=6921&ctNode=522&mp=1&ps=
3. 客家語拼音方案, http://www.edu.tw/pages/detail.aspx?Node=3653&Page=15592&Index=7&WID=c5ad5187-55ef-4811-8219-e946fe04f725
4. OpenVanilla 香草輸入法, http://openvanilla.org/
5. 信望愛台語客語輸入法 3.1.0 版, http://taigi.fhl.net/TaigiIME/
6. 教育部台灣客家語拼音輸入法, http://www.edu.tw/userfiles/url/20130116154410/moe_hkim_download.pdf
7. 101 年客語能力認證基本詞彙-中級、中高級暨語料選粹, http://elearning.hakka.gov.tw/Kaga/Kaga_QDMiddle.aspx
8. 教育部台灣客家語常用詞辭典, http://hakka.dict.edu.tw/hakkadict/index.htm
9. 哈客網路學院, http://elearning.hakka.gov.tw/default.aspx
10. 劉昭甫，台語無聲調輸入法的實作及改良，中興大學資訊科學與工程學研究所碩士論文，2010。
11. 蔡承融，國台語無聲調拼音輸入法實作，中興大學資訊科學與工程學研究所碩士論文，2008。
12. 羅火嵐，中文無聲調拼音輸入法及其實作，中興大學資訊科學研究所碩士論文，2006。
13. 羅丞邑，以資料探勘之技術解決線上客語語音合成系統中多音字發音歧義之研究，中興大學資訊網路與多媒體研究所碩士論文，2011。
14. 自然輸入法, http://www.iq-t.com/PRODUCTS/going9_01.asp
15. 微軟新注音輸入法, http://office.microsoft.com/zh-tw/help/HA010212138.aspx
16. Yabin Zheng, Lixing Xie, Zhiyuan Liu, Maosong Sun, Yang Zhang & Liyun Ru, 2011“Why Press Backspace? Understanding User Input Behaviors in Chinese Pinyin Input Method”, HLT '11 Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics Volume 2 pp. 485-490.
17. Yabin Zheng, Chen Li & Maosong Sun, 2011“CHIME: An Efficient Error-Tolerant Chinese Pinyin Input Method”, IJCAI'11 Proceedings of the Twenty-Second international joint conference on Artificial Intelligence - Volume

Three pp. 2551-2556.

18. Ming-Shing Yu, Feng-Long Huang and Piyu Tsai, 2006, Statistical Behavior Analysis of Smoothing Methods for Language Models of Mandarin Data Sets, to appear on Lecture Notes on Computer Science (LNCS), Springer, 2006.
19. Feng Long Huang, Neng-Huang Pan, Ming-Shing Yu, Jun-Yi Wu, 2011, Break Prediction of Prosody for Hakka's TTS Systems Based on Data Mining Approaches, IEEE International Conference on Machine Learning and Cybernetics (2011-ICMLC), Guangxi, China, Jul 10-13 (EI).
20. Feng-Long Huang, 2011, A Novel Approach for Disambiguating Chinese Homographs, 2011 International Conference on Materials Science and Information Technology (MSIT 2011),(EI).
21. Feng Long Huang, Neng-Huang Pan, Ming-Shing Yu, Jun-Yi Wu, 2011, Break Prediction of Prosody for Hakka's TTS Systems Based on Data Mining Approaches, IEEE International Conference on Machine Learning and Cybernetics (2011-ICMLC), Guangxi, China, Jul 10-13 , pp. 51-55 (EI).
22. Neng-Huang PAN, F.-L. HUANG, Chun-Hsien HO, Xin-Wei LIN and Shu-Hau SHIU, 2010, Constructing Online Audio Dictionaries for Bilingual Mandarin-Taiwan Dialects Based on Web 2.0 Concept, International Symposium on Chinese Spoken Language Processing (ISCSLP-2010), NCKU, Tainan, Taiwan, Nov. 29 - Dec.3, 2010. (EI)