The 2015 Conference on Computational Linguistics and Speech Processing
ROCLING 2015, pp. 27-42

以自然語言處理方法研發 智慧型客語無聲調拼音輸入法

Smart Toneless Pinyin Input Method for Hakka Based on Natural Language Processing

余明興[1] 黃豐隆[2] 魏俊瑋[3] 林昕緯[4]

[1,3,4] Department of Information Science, National Chung-Hsing University, Taichung 40227, Taiwan.
msyu@nchu.edu.tw

[2] Department of Computer Science and Information Engineering, National United University, MiaoLi, Taiwan,
flhuang@nuu.edu.tw

摘要

本論文研發的「好客拼音輸入法」以自然語言處理的方法為基礎，實作一套支援四縣及海陸腔的智慧型客語無聲調的智慧型拼音輸入法，使用者能夠快速且方便地輸入客語文句。輸入拼音時，以無聲調(Toneless)的方式進行，使用者不必考慮連音變調的問題，同時本輸入法提供當錯誤出現時之提示功能，以輔助不熟悉客語拼音的使用者。除基本的拼音輸入外，此輸入法還提供便捷輸入模式，包含四種輸模式：(1)快速輸入常用字串的自訂輸入(2)以客語詞各音節字首快速輸入的音首輸入(3)以學校、上市公司與組織名縮寫進行輸入得到該單位完整名稱的縮寫輸入和(4)以英文詞輸入得到對應中文詞的英文詞輸入。

此外，本輸入法提供將客語詞轉換成對應的國語詞選項，或者是以加註國語詞或拼音的方式表示。讓使用者輸入得到更具可讀性的客語文章。還有，提供發音的功能，當使用者在輸入拼音時，可以將客語音節正確讀出來，讓使用者在輸入時除了用看的也能用聽的來得知自己輸入的拼音是否有誤。以及提供唸出此客語詞的選項，讓使用者使用輸入法時還能學習客語詞彙的正確唸法。

因為客語語料不足，因此音轉字使用的語言模型是以客語詞對應的國語詞去建置。音轉字使用三個詞的少詞優先演算法搭配此語言模的情況下，有接近 76% 的正確率。

關鍵詞：好客拼音輸入法、音轉字、便捷輸入模式、少詞優先演算法，語言模型。

一、 前言

目前客語輸入法並不常見，市面上的客語輸入法只有財團法人信望愛所開發的信望愛客語輸入法[4]及教育部公告的客家語拼音輸入法[5]。而客語文字不夠通行的原因之一是沒有統一的文字用語，雖然客委會有建議用字，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期望，能夠研發出給不熟悉客語拼音系統的使用者都易於使用的「好客拼音輸入法」，而且能夠打出讓別人知道其義的客語客語句子，讓輸入法不光只是輸入，進一步可提供客語數位學習的功能。

二、 客家語拼音方案

本論文中，我們所採用的四縣與海陸腔客語拼音方案，為教育部所公告的台灣客家語羅馬字拼音方案[2]最新的客語拼音方案為基礎。最近一次的更新為中華民國 101 年 9 月 12 日的修正公告。下表為聲調符號表以及我們使用的音檔對應的調號。其中表一為客語四縣腔的部份，表二為海陸腔的部份。

表一：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

表二：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

客語如同中文，同樣也有連音變調(Tone Sandhi)的問題。對於客語的四縣腔，可歸納出三種連音變調規則；而海陸腔則歸納出兩種規則，如下表所示。其中表三為四縣腔的變調規則，表四為海陸腔的變調規則。

表三：客語四縣腔連音變調規則

<table>
<tr><td colspan="4">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（˴） + 陰平（˴） → 陽平（˅） + 陰平（˴）</td></tr>
<tr><td rowspan="4">範例</td><td>詞彙</td><td>變調前之拼音</td><td>變調後之拼音</td></tr>
<tr><td>新衫</td><td>xin˴sam˴</td><td>xin˅sam˴</td></tr>
<tr><td>買新衫</td><td>mai˴xin˴sam˴</td><td>mai˅xin˅sam˴</td></tr>
<tr><td colspan="4">規則 2：陰平與去聲構成的詞彙，讀時前字變調讀陽平</td></tr>
</table>

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

表四：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⁺

三、 使用之語料與語音庫

表五：客語詞數分布統計

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

我們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。在我們的詞典中，總共收錄約三萬三千個四縣腔詞目，海陸腔部分則收錄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，表五為四縣腔各字詞的詞數分佈。

- **客語語音庫(Hakka's SpeechBase)**

我們所設計的輸入法具有單字(Character)邊打邊唸的功能及唸出客語詞(Word)，因此我們需使用客語四縣及海陸腔的語音庫。四縣腔的語音檔為由熟悉客語四縣腔的老師錄製的基本合成單元，以客語單音節為單位，包含四縣腔的六種聲調，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新竹海陸腔的老師錄製，同樣也是以客語單音節為單位，包含海陸腔的七種聲調變化，總共錄製 3005 個基本合成單元。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)。

- **語言模型(Language Models)**

我們應用客語語言模型來作為音轉字的自動選字之決策依據，因考量到詞典是國客對照辭典及中文語料充裕的情況下，我們使用客語詞所對應的國語詞到中文語料中來訓練 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三：

(1) 中研院八萬詞(ASCED)

(2) 中研院平衡語料庫(ASBC3.0)

(3)Chinese GigaWord 3.0 繁體中文部分

透過上述的中文語料統計出詞頻，因中文及客語的語料規模差異很大，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。因為我們在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。

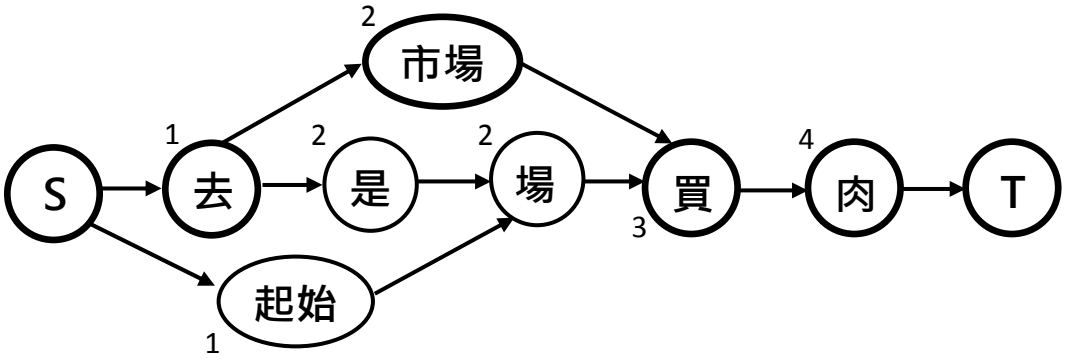
四、 音轉字處理之理論基礎

音轉字處理(Pinyin to Character)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。本論文採用三個詞的少詞優先演算法，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。我們會將詞數限制為三個詞的原因，將在實驗部份進一步說明。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。由於大量客語語料收集不易，運用 Bigram 時將產生嚴重的資料稀疏(Data Sparseness)問題[14]，故目前本研究僅使用 Uni-gram。
計算公式如(1)所示:

$$\text{Path} - \text{score} = P(W_1) * P(W_2) * \ldots\ldots * P(W_n) = \prod_{i=1}^{n} P(W_i) \quad (1)$$

這裡以一個實際的例子來說明少詞優先演算法的流程，假設四縣拼音「hi sii cong main giug」這一字串為少詞優先音轉字演算法的輸入部分，而輸出會得到分數最高且詞數最少的客語字串。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，如下圖所示。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，如圖一所示。

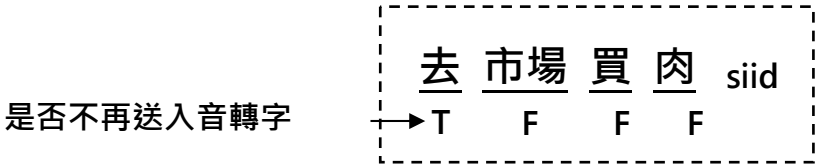


圖一：少詞優先演算法例子

前面表示「少詞優先演算法」的運算過程，但還需要進一步考慮送入詞數的問題，我們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，但輸入法係以詞為單位表示組字窗內容，每個詞都會由一個布林值記錄著，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。

下圖以剛剛「hi sii cong main giug」「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，作音轉字處理。接著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[6]，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「由前往後(Forwarding)」及「由後往前(Backwarding)」來斷詞，其結果不一定會相同，例如：一客語例句：「行政院長親自頒獎」斷詞結果分別為：

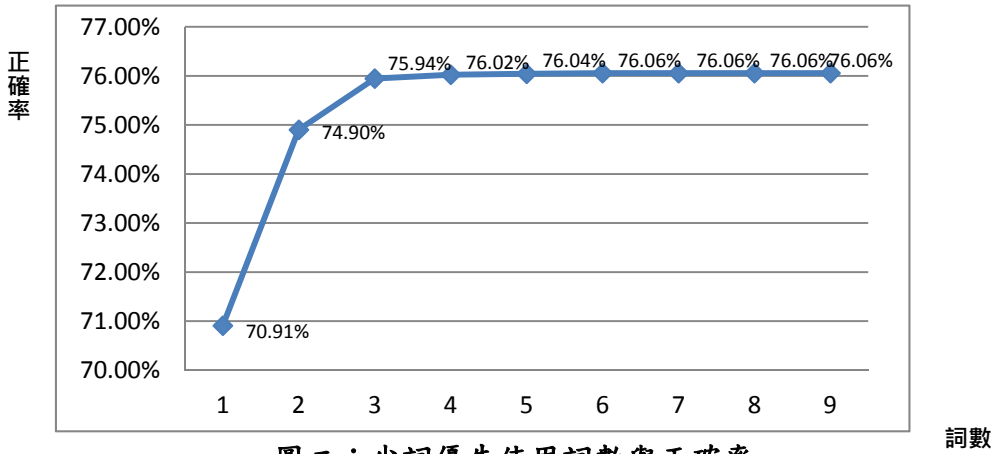
由前往後：行政院(hang ziin ien)長(**cong**)親(qin)自(cii)頒獎(ban jiong)。(E1)
由後往前：行政(hang ziin)院長(ien **zong**)親(qin)自(cii)頒獎(ban jiong)。(E2)

斷詞產生不同的結果，可能造成不同拼音輸出。為改善此問題，我們採用「由前往後」及「由後往前」斷詞，將二者結果不相同的例句移除，希望可以降低斷詞對正確率的影響。最後，只保留四縣腔 7697 個例句，共計 127885 字，並且將拼音標上後，以進行音轉字實驗。

實驗的流程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，下列客語例句：

三層肉鹹菜煮湯，味緒盡好。(E3)
(samcenngiug)(ham coi) (zu) (tong)(punc)(mi si)(qin ho) (punc)

實驗的結果如圖二所示。



圖二：少詞優先使用詞數與正確率

由結果可以看出，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，而會造成六詞的少詞優先演算法錯誤的八字詞皆為單字詞的可能性也不高。且一句話通常是由六詞以下組成的，因此六詞少詞優先之後正確率不會再提升。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，例如此例句「愛 就愛遽啊決定，毋好三心打兩意。」，音轉字結果分別為：

三詞少詞優先：愛就愛遽阿決定，毋好三心打涼椅。(E4)
六詞少詞優先：愛就愛遽阿決定，毋好三心打兩意。(E5)

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」即被固定不再送入音轉字，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。再看另一個錯誤的例子「山苦瓜苦丟丟仔」，在三詞少詞優先的情況中會轉錯成「珊瑚跨苦丟丟仔」，原因為輸入到第四個拼音時選擇的最少詞會得到「珊瑚」「寡婦」，而繼續輸入到第二個「丟」的拼音時，第一個詞「珊瑚」即被固定下來，因此輸入到最後雖然已經組出「苦丟丟仔」這樣的詞，但第一詞已經被固定，因此不會修正最前面的詞。

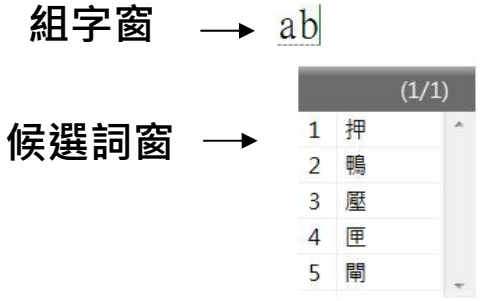
但是五字詞中每個字都無法組成詞的情況並不常見，且詞典中五字詞以上的詞目數量也並不多，且我們不希望輸入法修改離組字窗指標處太遠之前的結果，避免拿更多的詞來做音轉字，造成使用者需要移動指標到很前面的結果重新修正的情形。因此我們決定使用三個詞的少詞優先。

五、好客拼音輸入法

我們實作的好客拼音輸入法是以 OpenVanillla 香草輸入法[3]架構為基礎。接著我們會分別介紹輸入法的各項功能，以及與其它的客語輸入法比較。

● 拼音輸入

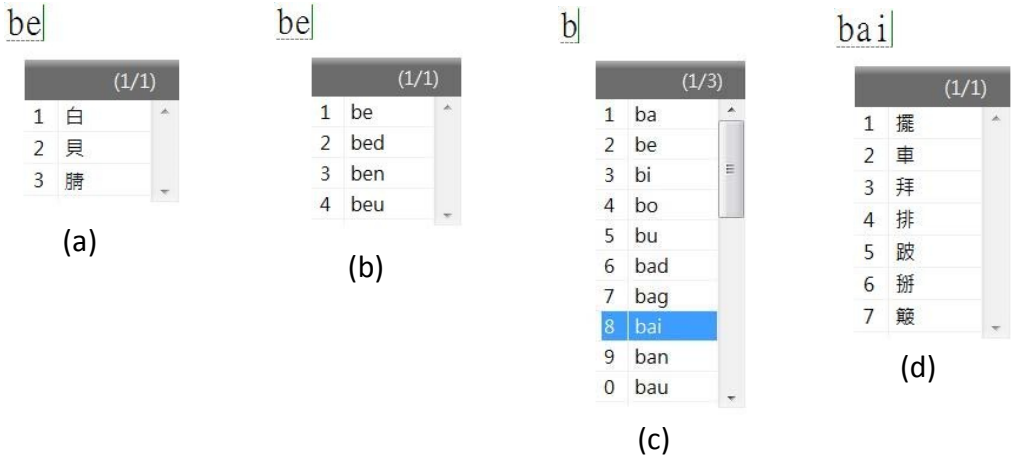
組字窗與候選詞窗是組成一個輸入法最基本的元件，我們的客語輸入法在輸入拼音與音轉字的過程都會在組字窗中進行，如同我們常使用的新注音輸入法[13]一樣，要按下 Enter 鍵後才會將組字窗的內容送到程序中。但是不同的地方在於新注音需要輸入聲調，且輸入聲調的動作會隨即將此拼音與聲調送入音轉字；而我們的輸入法考量到大多數使用者對客語聲調不熟悉，因此我們在輸入拼音時不需要輸入聲調，且會在每輸入一個拼音字母的動作中進行音轉字，會這樣做是考量到使用者可能對客語拼音較不熟悉，若使用者需要按下空白音轉字後才能得知此拼音能得到什麼字，對於不熟悉客語拼音的使用者較不友善。圖三為輸入客語拼音「ab」後組字窗(上)與候選詞窗(下)的內容。



圖三：輸入客語拼音「ab」後組字窗與候選詞窗的內容

- **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。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。圖四為使用者欲輸入詞的拼音為「bai」，但使用者記錯成「bei」，(a)為輸入至「be」時候選詞窗顯示「be」的候選詞、(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、(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、(d)使用者以選取後續拼音的方式，將組字窗改為拼音「bai」。



圖四：錯誤拼音提示功能

- **便捷輸入模式**

便捷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。

- **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好客客語拼音輸入法的模組設定頁面，來自訂拼音與詞。圖五為輸入法偏好設定中，自訂詞典的介面及展示。

圖五：自訂詞典界面及展示

- **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，結果如表六所示。

表六：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

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。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，然而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。如「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「gin <Space> loi <Space> gin <Space> do <Space>」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，即可在候選詞窗中立即取得「**緊來緊多**」詞。

圖六：音首輸入「GLGD」及「GLG」結果

- **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。如下圖為使用者欲輸入「中興大學」，只需要輸入「NCHU」即可在候選詞窗中找到此詞。

圖七：縮寫輸入「NCHU」及「TSMC」結果

- **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。如下圖所示。

圖八：英文詞輸入「Golf」結果

● 語音功能

以我們實驗室過去建置的客語語音合成(HTTS)系統為基礎[10, 15]，我們希望輸入法也能夠如 TTS 一般將客語的字與詞唸出。因此，我們在兩個部份加入了唸出客語的功能，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，如現有的國語自然輸入法[11]會念出輸入的注音及聲調。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。而唸出客語詞的選項我們會加入在兩個部分(1)在組字窗尾

端將拼音音轉字後，抓取最後一個詞，讓使用者選擇是否唸出 (2)對音轉字結果不滿意時，將指標往前移做修改時，修改的結果會讓使用者選擇是否唸出。發音時我們會根據詞典內客語詞與其聲調，經過客語連音變調規則後，將此詞以有聲調的方式唸出自然語音。

下圖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。

圖九：唸出客語詞選項

- **提高常用詞之優先順序**

我們可以透過調整使用者選擇的客語詞分數，來使語言模型更逼近客語文句的情況，也會使模型更符合使用者最近輸入的情況，進而提高輸入法音轉字的正確率。

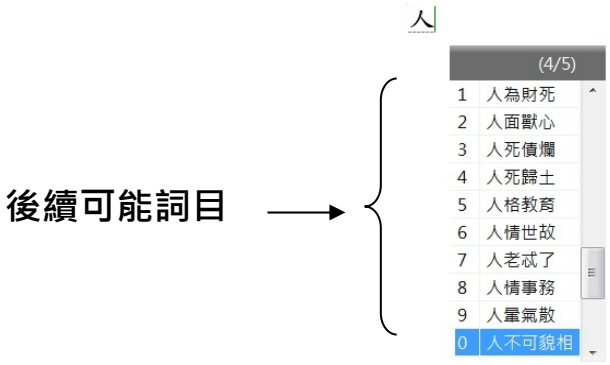
假設使用者選取的候選詞Candi_n在相同拼音的候選詞中次序為n，其分數為$\text{Candi}_\text{n}.\text{Score}$，若是$n \neq 1$時，我們就會將它的分數做調整。調整後分數$\text{Candi}_\text{n}.\text{Score}_\text{after} = \text{Candi}_\text{n}.\text{Score}_\text{pre} \times 2$，此時根據調整後分數$\text{Candi}_\text{n}.\text{Score}_\text{after}$在同拼音的後選詞之次序不同，會有三種不同的情況：

1. 若是$\text{Candi}_\text{n}.\text{Score}_\text{after} \leq \text{Candi}_{10}.\text{Score}$，也就是說調整後次序仍不在候選詞第一頁中，我們則將分數調整為$\text{Candi}_\text{n}.\text{Score} = \text{Candi}_{10}.\text{Score} + 1$，也就是強制使其出現在候選詞第一頁中。
2. 若是$\text{Candi}_\text{n}.\text{Score}_\text{after} \geq \text{Candi}_1.\text{Score}$也就是說調整後次序已經變成第一位，我們則將分數調整為$\text{Candi}_\text{n}.\text{Score} = \text{Candi}_1.\text{Score} + 1$。這樣對於 double 資料型態而言，不太可能發生溢位情形。
3. 若是$\text{Candi}_\text{n}.\text{Score}_\text{after} \leq \text{Candi}_{\text{n}-1}.\text{Score}$，也就是說調整後次序沒變化，則我們將分數調整為$\text{Candi}_\text{n}.\text{Score} = \text{Candi}_{\text{n}-1}.\text{Score} + 1$，也就是強制將次序上升一位。

- **往後預測可能詞目**

為了讓五字詞以上的長詞更有用處，因此我們加入了往後預測可能候選詞目的功能。其作法是當使用者在組字窗尾端進行音轉字之後，候選窗會顯示出組字窗最後一個詞後續還有哪些可能的詞目，假設最後一詞字數為 n，我們預設則是會列出字數為 n+2 的客語詞。我們的想法是因為，讓使用者選取只比目前輸入詞

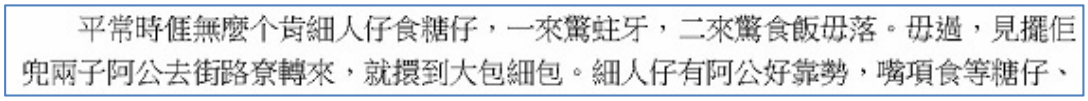
多一字的客語詞對於輸入的效率並沒有太大幫助，因此將門檻設為 2。對於較不熟悉客語或不想要往後預測詞目功能的使用者，也可以到偏好設定中自行調整門檻值或關閉此功能。下圖為輸入單字詞「人」之後，候選詞窗往後預測可能詞目的結果。



圖十：單字詞「人」往後預測可能詞目結果

- **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。下圖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。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。下圖為客語詞「暗晡」的加註國語與拼音選項。


平常時偃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、

圖十一：客語文章「鼻空向往下」一段

圖十二：客語詞「暗晡」加註拼音與國語詞選項

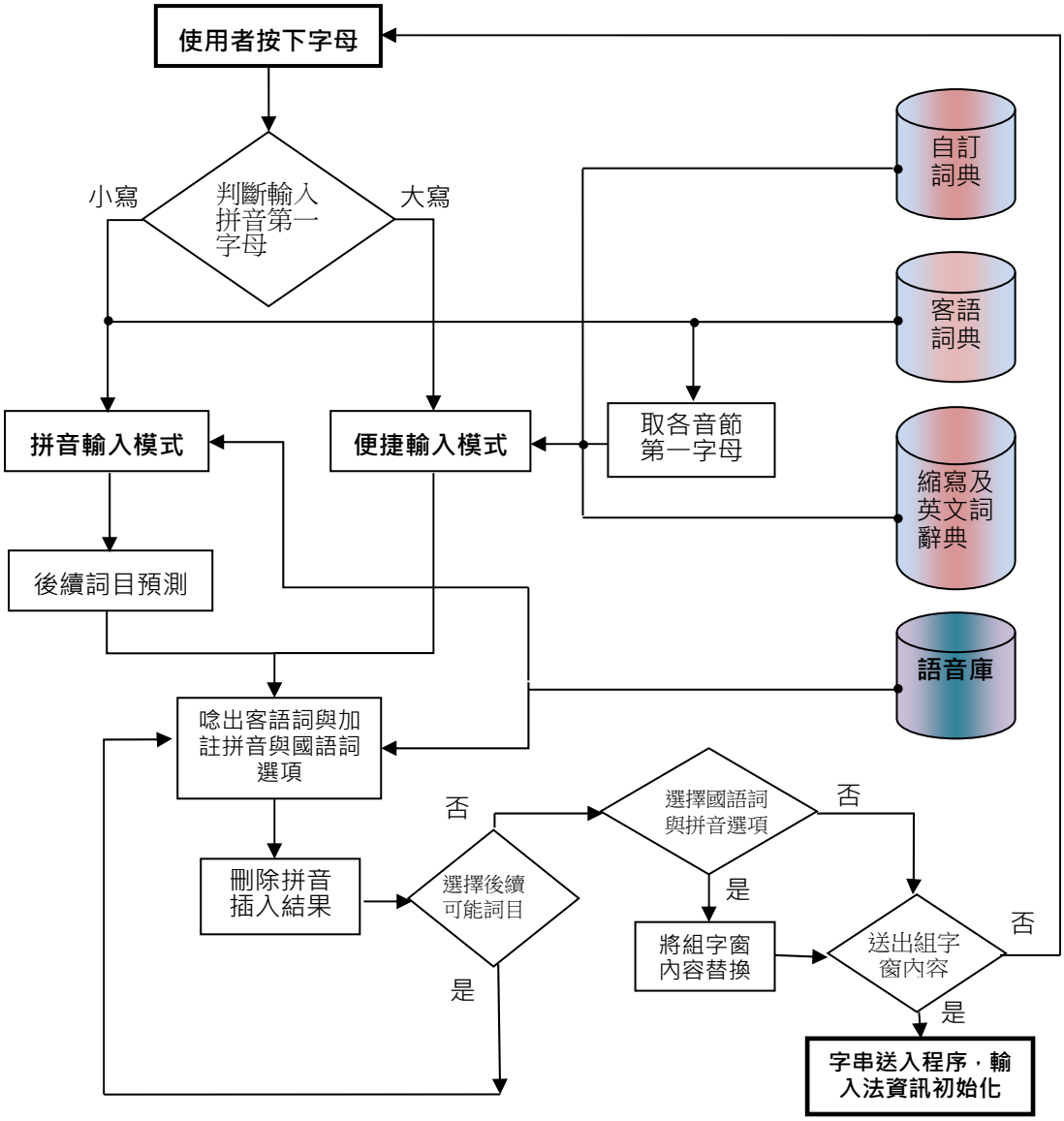
透過這些選項，即可將輸入得到的客語文章更具有可讀性，下圖為將上面那篇客語文章以輸入法加註拼音與國語詞後的結果。

平常時偃(我)無麼个(不太)肯細人仔(小孩子)食糖仔(糖果)，
一來驚(害怕)蛀牙，二來驚(害怕)食毋落(吃不下)飯。

毋過(但是)，每次/gien-bai/佢兜(他們)/gi-deu/兩子阿公(祖孫倆)
去街路(街道)寮/liau/轉來(回來)，就攞(提)/kuan/到大包細(小)包。

圖十三：客語文章加註拼音與國語詞後

- **輸入法流程圖**



圖十四:輸入法之流程圖

- **比較與討論**

本論文提出的輸入法功能與現有的客語輸入法做詳細比較與討論，如表七所表示。

表七：客語輸入法比較

	信望愛客語輸	教育部台灣客家	本論文輸入法

	入法	語拼音輸入法	
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，必須要加入[13]所提出的容錯拼音，才能夠改善此問題。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。基本上，我們所提出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提供使用者參考，比起列出所有的字還來得有效。因此綜合上述說明，本輸入法在這項功能上，比其它兩種輸入法更具具效益。

六、結論與未來研究

本論文的重點在於研究具有智慧功能之「好客拼音輸入法」，其中有多項具智慧性與創新的作法。我們提出拼音錯誤提示的功能，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便捷輸入模式的功能，能提高使用者的輸入效率。

為了輸出一篇更具可讀性的客語文章，在音轉字得到客語詞之後，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，這項功能對於推廣客語文字化將有很大的助益。

此外，本輸入法結合客語語音的功能，使用者輸入時能聽見自己鍵入的客語音節，還能讓使用者去聽客語詞的唸法，在輸入的過程中可以學習正確客語詞彙的發音，亦可作為客語數位學習。

關於進一步改進方向，首先的問題就是採集語料的問題，未來若是能收集大量的客語語料，訓練 bi-gram 客語語言模型，對於音轉字之正確率應可有效提升。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。而輸入法功能方面，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。且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。

參考文獻

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. 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
3. OpenVanilla 香草輸入法, http://openvanilla.org/
4. 信望愛台語客語輸入法 3.1.0 版, http://taigi.fhl.net/TaigiIME/
5. 教育部台灣客家語拼音輸入法, http://www.edu.tw/userfiles/url/20130116154410/moe_hkim_download.pdf
6. 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

7. 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。
8. 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。
9. 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。
10. 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。
11. 自然輸入法, http://www.iq-t.com/PRODUCTS/going9_01.asp
12. 微軟新注音輸入法, http://office.microsoft.com/zh-tw/help/HA010212138.aspx
13. YabinZheng, 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.
14. 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.
15. 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.