臺灣客語斷詞前導研究與模型建立
The Pilot Study and Model Construction for Word Segmentation in Taiwan Hakka

葉秋杏 Chiou-shing Yeh*
國立政治大學英國語文學系
Department of English
National Chengchi University
csyeh.corpus@gmail.com

賴惠玲 Huei-ling Lai
國立政治大學英國語文學系
Department of English
National Chengchi University
hllai.nccu@gmail.com

劉吉軒 Jyi-Shane Liu
國立政治大學資訊科學系
Department of Computer Science
National Chengchi University
jyishane.liu@gmail.com

摘要

斷詞是自然語言處理以及資料檢索查詢的關鍵角色，本文《臺灣客語語料庫》的斷詞前導研究，乃是運用華語斷詞系統 Stanford CoreNLP，以華客對應轉換方式為客語文本進行斷詞與標記。然而，斷詞效能不盡理想，因華客之間有許多字詞難以對譯，且臺灣客語次方言之間詞彙與語音也存在差異。有鑑於此，臺灣客語語料庫提出客語專屬的斷詞模型，建構客語詞庫，以六腔分列詞目，並採用長詞優先演算法以及動態規劃演算法設計。斷詞效能評估測試結果顯示，詞庫查找與詞頻統計（透過長詞優先演算法以及 N-gram 語言模型）兩者相輔相成，無論是斷詞效能或是斷詞準確率皆有著明顯提升。

Abstract

Word segmentation plays a key role in natural language processing and data retrieval queries. The pilot study employed Stanford CoreNLP, a word segmentation system for Chinese, for segmentation and tagging of Hakka texts in Taiwan Hakka Corpus. Nevertheless, the performance was unsatisfactory due to the intractable correspondent translations between Mandarin and Hakka and the lexical and phonetic varieties among the six dialects of Hakka. In view of these reasons, a tailor-made Hakka segmentation model is constructed that encompasses Hakka lexicon with six accents and that applies Maximum Matching Algorithm (MM) and dynamic programming algorithm in the system. The segmentation performance evaluation test results show that combining lexicon lookup and word frequency statistics (with Maximum Matching Algorithm and N-gram Language Model) significantly improves both segmentation performance and accuracy.

關鍵字：客語斷詞、長詞優先演算法、N-gram 語言模型、詞頻統計、臺灣客語語料庫

Keywords: Hakka word segmentation, Maximum Matching Algorithm (MM), N-gram Language Model, word frequency statistics, Taiwan Hakka Corpus

1 緒論

在自然語言處理中，斷詞是一個基礎且至關重要的角色。相較於拉丁語系（如英語）詞彙之間以空白字元做為區隔，可將每個詞彙斷開，漢語的詞彙缺乏詞間空格，詞彙之間的邊界模糊，因此漢語的自然語言處理更加不易。臺灣客語除了用字體系仍未穩健外，使用客語用字書寫之文本數量與臺灣強勢語（華語）更是有著明顯落差，資源較為稀缺匱乏不易取得，部分用字仍存紛雜未定，均對人工或者機器進行客語斷詞帶來諸多挑戰（謝杰雄，2006；江俊龍，2010, 2013；黃豐隆，2015）。相比而言，華語斷詞系統的發展日趨成熟，其中較廣為人知的包含中央研究院 CKIP 中文斷詞系統、美國史丹佛大學之 Stanford CoreNLP 與國家教育研究院之國教院

分詞系統。臺灣客語語料庫團隊於斷詞系統建置初期，遂藉由中文斷詞系統的輔助達成前導斷詞程式的設計，並希冀透過前導斷詞程式的測試結果來評估臺灣客語斷詞系統的建置方向。然而，經過前導斷詞程式的測試後，顯示出將中文斷詞系統套用於臺灣客語斷詞系統的建置上所遭遇到的瓶頸，包含客語特殊字問題，以及華客對譯字數無法完全對應、一對多或多對一、翻譯不易等狀況，因此客語斷詞及詞性標注系統的獨立開發成為一件必要工程，且須同時投入系統研發技術以及具語言學知識背景的人力資源，方可將臺灣客語斷詞系統穩健地建置起來。以下將分別闡述前導斷詞程式的試驗結果與評估，以及客語斷詞及詞性標注系統第一階段與第二階段的內容及歷程。

2 臺灣客語斷詞前導研究

華語的漢字系統穩定悠久，加上華語斷詞系統較為蓬勃發展，因此在建立客語斷詞系統之初，臺灣客語語料庫（THC）團隊借助可將華語文本依詞切開之華語斷詞系統，嘗試以華客翻譯的方式做為客語斷詞系統開發的前導測試。THC 團隊首先參考了世界知名大型語料庫所開發之華語斷詞系統，並於計畫初始階段採用客家委員會發布之客語認證詞彙做為前導斷詞程式的詞庫基礎，進行程式開發作業。以下將說明如何選用華語斷詞系統架構，並介紹底層資料之建置方式，以及前導斷詞程式的斷詞及詞性標注流程。

2.1 來源語—華語之斷詞系統架構選用

較著名的四個華語斷詞系統，為中央研究院 CKIP 中文斷詞系統、中國結巴斷詞系統（Jieba）、CQPweb，以及美國史丹佛大學之 Stanford CoreNLP – Natural language software（以下簡稱 Stanford CoreNLP）。THC 於 2018 開始規劃斷詞系統建置，當時 CKIP 中文斷詞系統（Chen, 1992）僅提供線上斷詞服務，無開放原始碼，故斷詞只能單純發送文本並取回斷詞結果，無法瞭解其內部運作機制，亦無法對於斷詞結果進行修正。[1]而中國結巴斷詞系統（Jieba）雖是一個開放原始碼的程式，然模型搭建時所使用的語料為簡體中文，儘管此程式亦支援繁體中文，斷詞精確度可能仍有其侷限。國教院分詞系統（柯華葳等人，2016）所採用的 CQPweb 則是檢索語法細緻複雜，也因此對於使用者而言較有難度，操作上也較不直觀，對於系統執行任務而言，較複雜的查詢也就需耗費較長的時間處理。至於 Stanford CoreNLP (Manning et al., 2014) 是在 GitHub 上的開放程式碼專案，支援多種語言（包含繁體中文），另提供線上服務的版本（http://corenlp.run/）。綜合考量前導斷詞程式與後續客語斷詞系統的開發彈性，THC 團隊選擇使用 Stanford CoreNLP 做為前導斷詞程式之基礎。

2.2 目標語—客語之斷詞底層資料建置

由於 Stanford CoreNLP 只支援華語詞集，針對臺灣客語的詞彙及語句無法進行有效的斷詞判讀，例如以客語文本「佢舖娘今晡日愛轉外家」進行斷詞（圖 1），Stanford CoreNLP 無法判讀客語「佢舖娘」包含兩個詞彙（華語為「他」及「太太」），故須透過臺灣客語之底層資料建置，來處理詞彙的客華對譯及轉換。亦即，將客語語句轉換為華語語句，並以華語語句來進行斷詞，例如以對譯後之華語文本「他太太今天要回娘家」進行斷詞（圖 2），而後再將華語語句之斷詞結果對應回客語語句，並顯示出相對應的客語詞類標記。

圖 1. Stanford CoreNLP 客語斷詞結果

圖 2. Stanford CoreNLP 客轉華之斷詞結果

因此，THC 團隊將 Stanford CoreNLP 專案下載後，進行編譯並部屬執行於位處 linux 虛擬主

[1] 中研院後於 2019 年正式開源釋出中文斷詞程式 ckiptagger（Li et al., 2020），程式碼與相關操作方式存放於 GitHub 平臺供使用者研究運用。

機上的 Java 環境。客語轉成華語的關鍵，便是底層資料（data base）之建置，底層資料主要包含客語詞表及客語斷詞標記對應表兩大部分。在詞表方面，由於客語次方言腔調存在用字差異，往往造成客華對譯及詞類標注之衝突。以客語「討」字為例，在客語各腔皆有表「求取」之意，而在詔安腔獨有表情態「要」之用法（如：「你討做麼个？」華語為「你要做什麼？」），故為避免客華字詞轉換錯誤，THC 團隊先行以單一腔調且少量樣本進行前導測試。考量四縣腔為客語中最多人使用的腔調，語料及詞彙數量相對充足，因此選用四縣腔做為前導測試資料，詞表內容則是採用客家委員會（2018）《107 年度客語能力認證初級詞彙（四縣腔）》以及《107 年度客語能力認證中級詞彙（四縣腔）》。THC 團隊擬透過單一腔調的測試結果，進而觀察斷詞判斷效能，若斷詞正確性在一定程度上，即可逐一類推至其他客語腔調之斷詞。選定腔調後，即是詞目彙整與資料前處理，建立客華對應轉換用之詞表。前處理之項目包含：（一）若客語欄位及華語欄位內容完全相同（如：客語「字帖」對應華語為「字帖」），則不列入客華轉換用的詞表當中（亦即不需轉換，因此須於轉換表中刪除）；（二）刪除詞組（如「嗷無目汁」）；（三）拆分詞目，例如主詞目原為「等路【妄想】」（華：禮物），須拆分為兩個獨立詞目「等路」、「妄想」）；（四）若華語欄位中出現描述性之文字，則由系統移除字樣，例如客語「儕」的華語對譯「位（量詞）」，須將「（量詞）」刪除。

整理後的客華對應轉換詞表，條目數量為 2018（民 107）年客家委員會認證詞彙初級與中級合計 1,452 筆。檔案為 excel 格式，而後轉成 CSV（Comma-Separated Values），接著再轉換成 JSON（JavaScript Object Notation），以利客語詞表與 Stanford CoreNLP 相互運作，讀取客語與華語欄位資訊。THC 團隊係透過 REST API (Representational State Transfer Application Programming Interface) 將詞表匯入底層資料庫中（請見圖 3）。

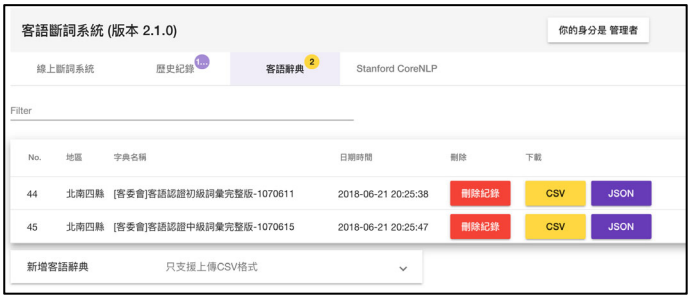

圖 3. 前導斷詞程式管理介面

在詞性標記方面，客語語料庫在前導斷詞程式開發階段所採用之詞性標記為 17 類。[2]於 Stanford CoreNLP 使用的 33 項賓州樹庫詞類標記（The Treebank Part-of-Speech Tagset (Xia, 2000)）之中，有 28 類可歸納對應至客語 17 類詞性標記，另外 5 類（BA、FW、JJ、LB、SB）則回歸到詞彙之典型詞類，例如華語「男/JJ 的/DE」以典型標記標示為「男/N 的/DE」。客語詞性標記 17 類分別為：AD（Adverb，副詞）、AS（Aspect marker，時態）、C（Conjunction，連詞）、DE（「个」（的）、「得」（得）、「个」（地））、DET（Determiner，限定詞）、IJ（Interjection，感嘆詞）、M（Measure word，量詞）、N（Noun，名詞）、NEG（Negative，否定詞）、ON（Onomatopoeiae，擬聲詞）、P（Preposition，介詞）、PN（Pronoun，代名詞）、PRT（Particle，助詞）、PU（Punctuation，標點）、NR（Proper Noun，專有名詞）、V（Verb，動詞）、VC（Copula Verb，繫動詞）。

在「詞表」及「詞性標記對應表」皆建立完畢後，遂可應用至客華文本轉換機制之中。

2.3 客華文本轉換機制

客華文本之轉換機制，係藉由程式中之底層資料將使用者輸入的客語文本以字串搜尋與文本取代的方式將客語文本轉成華語文本，而後再將斷詞後的華語文本轉換回客語文本。若詞表的條目數量越多，轉換出來的華語文本就會越適合進行斷詞處理。

以客語文本「就係恁仔，大家莊頭莊尾識透透。」為例，使用者在前導斷詞程式介面輸

[2] 斷詞前導實驗於計畫第一年執行，當時制訂的斷詞標記共 17 項。基於語言共性以及客語特殊性，計畫期間團隊與顧問委員多次進行討論與修訂，終以 24 類斷詞標記為定。《臺灣客語語料庫》正式版於 2022 年 10 月上線，標記及其示例可詳見「語料庫元資訊」之「臺灣客語語料庫斷詞標記表」：https://corpus.hakka.gov.tw/#/corpus-info。

入文字後，生成的華語文本為「就是這樣，大家莊頭莊尾識透透。」（詳見圖 4）。由此可知，客語「係」轉換為華語「是」，客語「恁仔」轉換為華語「這樣」，其餘未替換之字詞，則包含客華用字相同（如：大家），或是客語詞表中未收錄對應條目而無法轉換之情形（如：識透透）。

圖 4. 客語轉換成華語範例

當客語文本轉換為華語文本後，便可透過 Stanford CoreNLP 進行斷詞並標示賓州斷詞標記，再藉由其提供的 API 取得斷詞結果。Stanford CoreNLP 視覺詞性標記呈現如圖 5：

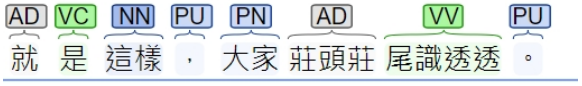

圖 5. 視覺詞性標記

接著，由程式將華語文本之賓州詞性標記，重新標示成客語斷詞標記，如圖 6 所示：

就(AD) 是(HE) 這樣(N)，(PU) 大家(PN) 莊頭莊(AD) 尾識透透(VA)。(PU)

圖 6. 標示為客語斷詞標記

最後，華語文字由程式轉換回客語文字，其斷詞結果請見圖 7：

就(AD) 係(HE) 恁仔(N)，(PU) 大家(PN) 莊頭莊(AD) 尾識透透(VA)。(PU)

圖 7. 華語轉換回客語文字

前導斷詞程式之斷詞結果，須由人工進行檢核修正。THC 斷詞原則主要係參考中央研究院詞庫小組（1998）《中央研究院平衡語料庫的內容與說明（修訂版）》以及 Huang et al. (2017) 的 *Mandarin Chinese words and parts of speech: A corpus-based Study* 來進行分合判斷，將程式斷詞結果以手動方式修正為「就(AD) 係(HE) 恁仔(VS) ，(PU) 大家(PN) 莊頭(N) 莊尾(N) 識透透(VS) 。(PU)」（詳見圖 8），並將結果複製至 WORD 檔，後續提供工程師進行斷詞正確性測試。

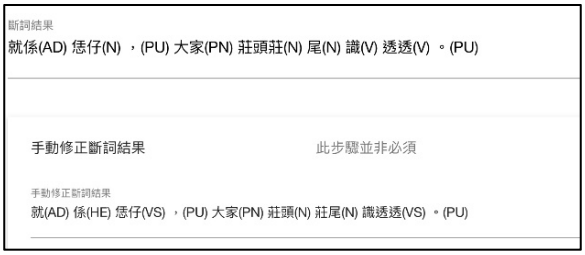

圖 8. 人工手動修正斷詞結果

2.4 前導研究成果評估

前導斷詞程式一共進行了 1,000 句四縣腔語句之斷詞測試，斷詞總字數為 9,318 字，研究成果之檢驗方式有兩種，包含「斷詞結果比較」以及「斷詞正確率評估」。

第一種檢驗方式「斷詞結果比較」，係以相同之客語例句，透過「前導斷詞程式」與「CKIP 中文斷詞系統」生成斷詞結果，再與人工修正之結果所進行比較，範例內容請見表 1。

客語例句	恁樣个日仔實在還快樂哪！
CKIP 中文斷詞系統之斷詞結果	恁樣(Na)&#(FW)２００１０(Neu);(SEMICOLONCATEGORY) 日(Nc)仔(Na)實在(D)還(D)快樂(VH)哪(T)！(EXCLAMATIONCATEGORY)
臺灣客語前導斷詞程式之斷詞結果	恁樣(V) 个(DE) 日仔(N) 實(N) 在(P) 還(RN) 快樂哪(N) ！(PU)
專家人工修正結果	恁樣(V) 个(DE) 日仔(N) 實在(AD) 還(AD) 快樂(V) 哪(PRT) ！(PU)

表 1. 各斷詞結果比較表範例

儘管同屬漢語，CKIP 中文斷詞系統是專為臺灣華語而設計，因此在處理客語文本斷詞時，即會因為客語的特殊用字、構詞或語法差異，出現一些字型判讀或顯示的問題。例如臺灣客語的「个」被 CKIP 中文斷詞系統轉換成字元參照值（numeric character reference, NCR）之編碼「&#２００１０;」，導致無法正確斷詞。

第二種檢驗方式「斷詞正確率評估」，則是計算客語前導斷詞程式之斷詞正確率，計算公式為：（總客語文本長度 - 斷詞錯誤文本長度）/（總客語文本長度）取百分比之值即為正確率。依照前述計算公式，可得出表 4 客語文本「恁樣个日仔實在還快樂哪！」透過前導斷詞程式斷詞的正確率為 50%，計算如下：

總客語文本長度 = 12
斷詞錯誤文本長度 = 6
（錯誤：實(N) 在(P) 還(RN) 快樂哪(N)）
斷詞正確率為 (12 - 6) / 12 = 50%

經過前導斷詞程式的測試後，1,000 句客語斷詞（帶斷詞標記）平均正確率為 37.5%。
儘管臺灣客語前導斷詞程式在客語斷詞成效上已經優於 CKIP 中文斷詞系統與 Stanford CoreNLP，然也遭遇到下述幾項困境：

(1) 客華對譯問題：
例如客語詞目「反躁」，其華語欄位文字為「精神亢奮而失眠」。「反躁」之斷詞標記為 N（名詞），但若直接使用華語欄位文字透過 Stanford CoreNLP 進行斷詞，會對應到 4 個詞彙「精神/N 亢奮/V 而/C 失眠/V」，因此客華對譯須力求字數對應，然客華的語言與文化差異，仍難免造成詞彙對譯上的困難。

(2) 客華一對多或多對一問題：
多（客）對一（華）的狀況較好處理，系統只要在取得華語對譯詞標記後，還原為原本的客語詞目即可；然一（客）對多（華）則較為棘手，例如客語詞目「妄想」為多義詞，可表華語動詞「妄想」或名詞「禮物」，然前者標記為 V，後者為 N，Stanford CoreNLP 無法判別與選擇，導致斷詞標記可能產生錯誤。

(3) 詞表條目數量不足：
未收錄的字詞無法進行客華轉換。

除了前述已知的問題之外，還有客語六腔差異性的議題尚待克服，若仍持續以客華翻譯之框架進行客語斷詞及詞性標注系統的建置，其成效將會相當侷限。綜合評估下，客語斷詞系統必須為客語量身打造專屬此語言的斷詞模型，讓機器直接學習客語的語言結構，而非藉由其他語言翻譯（例如客華對譯）方式進行斷詞，因此客語斷詞與詞性標注系統之獨立開發以及底層模組建構遂成為必要的方向。

3 臺灣客語斷詞系統模型建立

客語斷詞及詞性標注系統的建置，首先要建立一套客語詞庫，並持續擴充條目數量。而在系統開發進程方面，客語斷詞及詞性標注系統第一階段係採用詞庫查找、長詞優先（Maximal Matching Algorithm）及動態規劃演算法（Dynamic Programming Algorithm）設計，透過運算找出與詞庫中匹配之斷詞及詞性標記之組合，並進行斷詞標注。其中又可依照詞庫的建置進程分成兩階段，分別為第一階段之詞表式詞庫，以及第二階段之詞庫及語料詞彙篩選系統。以下將介紹詞庫資料模型之建立，並分述各階段的建置歷程。

3.1 第一階段：詞表式詞庫

臺灣客語專屬之斷詞系統開發，須從語料庫最底層的資料模型（Data Model）開始規劃，其中「詞庫」即是舉足輕重的核心角色。為了有效提升詞庫查找的正確性，充足的語料量及詞彙量是首要條件，因此在詞庫的資料來源方面，採用詞彙量較為充足之教育部《臺灣客家語常用詞辭典》[3]詞目。THC 團隊在參考《臺灣客家語常用詞辭典》之辭典資料欄位設計後，因應語料庫之斷詞需求所設計的詞庫欄位，分別為「字詞主資訊」、「單字輔助資訊」、「華語資訊」與各腔音讀資訊等頁籤。其中，「字詞主資訊」頁籤中包含「客語詞目」、「斷詞標記」等，屬於條目的核心資訊，亦是做為斷詞依據的重要資料；而「單字輔助資訊」頁籤則是專屬

[3] 在底層資料模型建置初期，語料庫團隊向教育部取得授權的辭典版本為 2018 年試用版，其後教育部陸續公告修正用字，因此目前語料庫所收錄之教育部詞目已更新至 2022 年 4 月（可參見教育部（2019），網址為 https://hakkadict.moe.edu.tw/cgi-bin/gs32/gsweb.cgi/ccd=V4R_Z9/newsearch?&menuid=gsnews）。

於單字條目的資訊，包含「部首」、「單字部首外筆畫數」與「單字總筆畫數」；至於「華語資訊」頁籤則包含了「對譯詞」、「釋義」等欄位，其資訊可延伸做為未來華客對譯應用之基礎。[4]

詞庫欄位建立好之後，則須進行詞目彙整以及編輯修訂。由於《教育部客家語常用詞辭典》所包含的資料內容相當豐富且龐雜，而且一個詞目往往同時帶有多種詞性，如名詞、動詞、副詞、介詞等，故為了系統斷詞需求，每筆教育部詞目須仰賴人工逐條檢查與修訂，除了用字勘誤外，也須將詞目依不同的詞性獨立拆分。除此之外，部分詞目為客語難字，網頁上係以圖片顯示，因此工作人員依據 Unicode 擴展漢字的資料（The Unicode Consortium, 2022），一一比對後將所有圖片改為可被檢索之文字。其他如主詞目為詞組或俚諺語者，也須予以刪除。詞目的拆分與彙整，須仰賴專家人工的分類及整理，因此在語料庫建立初期，著實投入了大量的人力及時間，致力於資料清理與統整工作。

詞表完成後，即匯入於系統後端。詞表來源為教育部《臺灣客家語常用詞辭典》，經彙整後共計 21,617 筆。

此時期的斷詞系統採用長詞優先法及動態規劃演算法機制，即時與 THC 詞庫連線取得各詞條基礎資料，透過運算找出匹配之斷詞及詞性標記之組合，並進行斷詞標注。斷詞及詞性標注器如圖 9 所示，中間欄位可貼入文字，按下 Analyze 鍵之後，系統即會自動斷詞。標注方式係在字詞後方加上左右圓括號，標記出斷詞標記，例如「自家(N)」。藍色字即為已入詞庫帶標記的詞彙，黑色字即屬尚未收錄於詞庫的字詞。

圖 9. 斷詞及詞性標注器（第一階段）

經過系統斷詞後的文本，還需經由人工進行手動斷詞修訂，例如將這些黑色字入庫，或是若此字是與前後字元組成的詞彙，即須合詞並入庫。然而這些被標注詞性或修正過後的字詞，無法反饋或收錄至詞庫之中，工作人員在不同文本中發現同一詞彙時，都仍必須再人工重複修訂一遍，因此除了繼續優化客語斷詞及詞性標注器功能外，建構可直接於網頁操作的詞庫也是勢在必行。

在第一階段的斷詞系統效能評估方面，係以 1,000 句專家修訂之客語例句（帶斷詞標記）及 21,617 筆的詞庫數量來進行評估測試，評估方式採用 Levenshtein Distance（LevDis）及 Longest Common Subsequence（LCS）演算法。Levenshtein Distance 為一種量化指標演算法，稱為「編輯距離（Edit Distance）」，係兩字符串之間，由一個轉換為另一個字符串所需的最少編輯操作次數。編輯允許以下三種操作：刪除一個字符（deletions）、插入一個字符（insertions）、將一個字符替換成另一個字符（reversals）（Levenshtein, 1966）；而 Longest Common Subsequence 概念與 Levenshtein Distance 相似，唯獨編輯操作不將一個字符替換成另一字符，做為另一評估指標。一般來

[4] THC 計畫囿於時間與人力限制，詞庫以建置「字詞主資訊」為優先，其餘欄位內容將於後續列入排程。

說，編輯距離越小，兩字符串相似度越大。物理意義上來說，由於會評估字符是否為插入、取代、刪除，因此連續位移之字符串並不會視為是連續錯誤，倘若斷詞結果越相近，則得出距離越近、數值越小，反之則數值越大。若考量以詞性標記加入詞彙一起評估下，則詞性標記應視為單一字符處理。

以上兩種評估模式（LevDis、LCS），客語斷詞及詞性標注系統獲得 43%～45% 之正確率（請見圖 10），且在與 Stanford Parser 及 CKIP 中文斷詞系統之比較中，正確率皆具有兩倍以上之改良幅度，說明客語斷詞及詞性標注系統使用詞庫來斷詞的實例中，著實發揮顯著功效（請見表 2）。

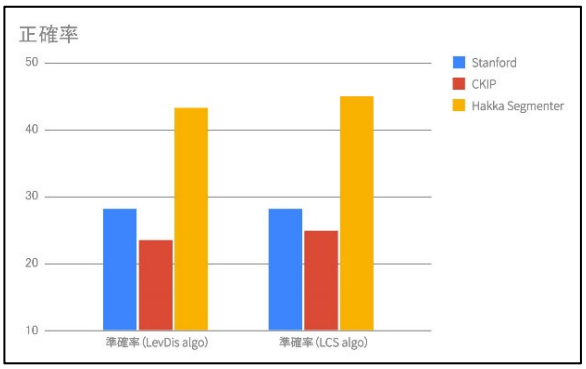

圖 10. 斷詞效能評估（帶斷詞標記）

斷詞系統 / 項目	Stanford Parser	CKIP 中文斷詞系統	客語斷詞及詞性標注系統 (Hakka Segmenter)
準確率(LevDis)	28.21	23.62	43.35
準確率 (LCS)	28.38	25.02	45.01
有效測試組數	882	999	999
詞數	8341	11912	12724
Levenshtein 平均距離	10.27	11.03	8.2

表 2. 斷詞效能評估資料（帶斷詞標記）

綜整以上所述，儘管在斷詞系統第一階段的斷詞正確率（43%～45%）相較於前導斷詞程式時期（平均 37.5%）已有小幅度提升，但仍存在諸多問題。首先，詞庫僅能由工作人員彙整出條目列表，請工程師於系統後端匯入，因此無法即時針對詞庫條目的斷詞標記進行適切的修正。如何有效率地增加詞庫條目，以因應大量語料文本的斷詞並提升斷詞正確性，亦至關緊要。再者，此階段的斷詞及詞性標注系統為獨立頁面，尚未與語料庫後臺連動。工作人員在處理斷詞時，先將客語語料貼入此系統頁面斷詞，若該詞彙為詞庫內詞目，斷詞後即會於詞彙後方顯示其斷詞標記（如「跲/VA」），未收錄於詞庫的未知詞後方則會以空白顯示（如「餡　」）。隨後，工作人員將系統斷詞結果貼入 WORD 檔，並判定未知詞於此語句中的詞性後，於此字後方加上「/」以及斷詞標記（如「餡/N」）。完成所有未知詞標記後，再將文檔交由工程師匯入語料庫。有鑑於此，詞庫的編輯功能、未知詞分析與審核機制以及斷詞及詞性標注系統與語料庫後臺連動之開發設計，尤為迫切。

3.2 第二階段：詞庫及詞彙篩選系統（N-gram）

隨著語料持續匯入，未知詞（Out of Vocabulary, OOV，語料庫稱之為待決詞，Pending Words）也不斷增加。為快速辨識出這些未知詞並判讀是否成詞，列入詞庫擴充詞庫數量且提升機器斷詞正確率，「詞彙篩選系統」應運而生。因此，客語斷詞及詞性標注系統之開發邁入第二階段，以長詞優先法與 N-gram 模型為基礎，並以詞庫查找的方式進行斷詞標記。此時期的詞庫已可由語料庫後臺介面進行編輯，包含條目新增或刪除，以及條目欄位內容編修等。

直接建構於語料庫後臺的語料詞彙篩選系統，其運作機制係透過待決字詞推薦功能，亦即基於數據統計的語言模型演算法，將語料文本內容進行文字分割，依照所選取相鄰字數當作條件機率計算，形成長度為 N 的字詞（N=1~5）。斷詞系統首先會經由長詞優先比對詞庫，掃描這些長度為 1~5 的字詞片段序列，若 N-gram 分析後該字詞片段不存在於詞庫中，則將這些字詞片段自動顯示於待決字詞清單，再由人工審核是否成詞。語料詞彙篩選系統會定期統計所有已收語料之待決字詞出現頻率，並於管理後臺功能中顯示頻率及當前收錄狀態（如圖 11）。此外，操作介面中還提供跳轉至語料前端檢索介面之功能，供詞庫組對照實際文本以評估是否將待決字詞收錄進詞庫之中（如圖 12）。

圖 11. 語料詞彙篩選系統操作介面

圖 12. 透過介面開啟之關鍵詞檢索（非成詞）頁面

斷詞及詞性標注器也持續進行改良（如圖 13）：

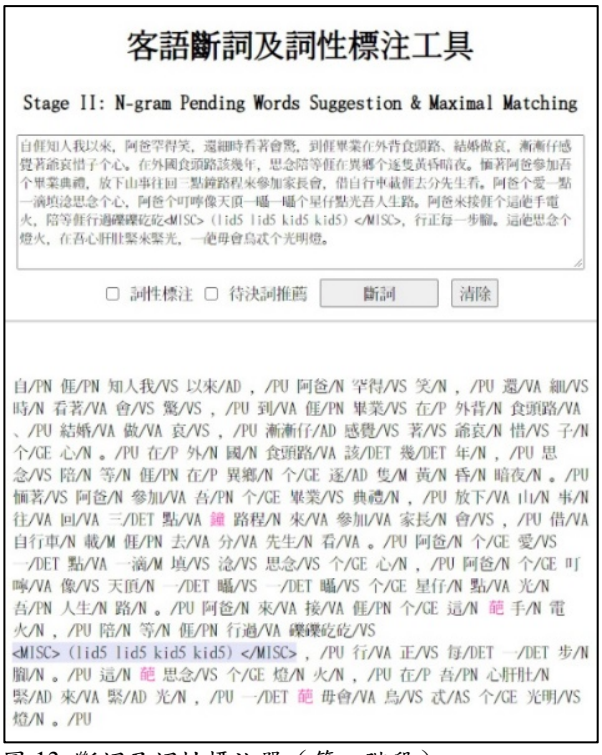

圖 13. 斷詞及詞性標注器（第二階段）

標注方式改為在字詞後方加上半形斜線並給予斷詞標記，例如「知人我/VS」。而後依據動態規劃演算法之標記結果，經由文本後處理器將資訊（如：斷詞標記、文字顏色、斷詞區段區隔）附加於斷詞後之文本，並呈現於斷詞結果介面。字型顯示為黑色者，即表示系統已建立在詞庫中並可辨識的字詞，若顯示為桃紅色則屬於尚未收錄至詞庫的字詞，而灰色底色則表示轉寫標記，可被系統辨識而不被斷詞。[5]

至於斷詞標記則修訂為 24 類，分列如下：AD、AS、B（Bound Morpheme，附著語素）、BUN（「分」）、C、DED（「得」）、DET、DO（「到」）、GE（「个」（的））、HE（「係」，Copula，繫動詞）、IJ、LAU（「摎」）、M、N、NEG、P、PN、PRT、PU、RN（Proper Noun，專有名詞）、SYM（Symbol，符號）、TUNG（「同」）、VA（Action Verb，行動動詞）、VS（State Verb，靜態動詞）。

客語斷詞及詞性標注器後亦歷經多次修正，介面設計也有所更新，現已佈線於語料庫前臺介面（見圖 14），同時提供工作人員以及一般使用者執行客語斷詞。

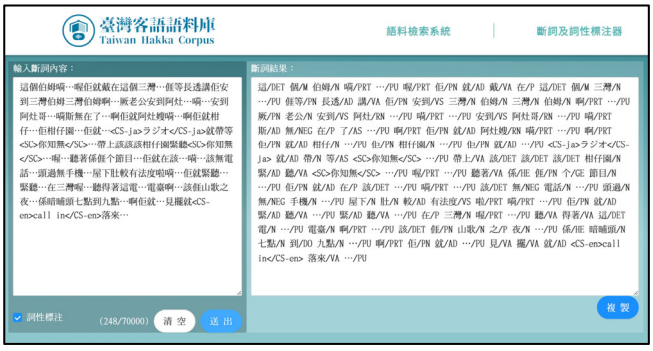

圖 14. 斷詞及詞性標注器（第二階段，語料庫網頁版）

語料庫後臺的書面文本後設欄位中，包括「文本內容（前臺顯示用）」以及「文本內容（含斷詞結果）」，工作人員會將文本內容（前臺顯示用）貼入斷詞及詞性標注器左欄並一鍵斷詞，右欄即會顯示系統自動斷詞後結果。若出現未帶標記的詞目，工作人員即

[5] 臺灣客語語料庫採用「轉寫標記」標示非客語字，例如書面文本中的拼音、其他語言文字（如出現日文時，標記為<CS-ja>ラジオ</CS-ja>），或是口語自然語流中穿插使用其他語言的現象等，目的係供系統辨識之用，可以有效降低斷詞錯誤。詳見語料庫網頁：https://corpus.hakka.gov.tw/#/corpus-info。

至詞彙篩選系統搜尋此未知詞，並將該詞選為「已決詞」，選取斷詞標記後，將此詞彙儲存，系統即會同時將此詞彙加入於詞庫當中。所有未知詞加入詞庫後，工作人員會再將「文本內容（前臺顯示用）」再次斷詞，此時該文本中所有詞彙皆帶有斷詞標記，因此即可將此結果貼入「文本內容（含斷詞結果）」，儲存入庫。

對於第二階段的斷詞系統效能評估，係以兩種差異量（一致性）評估模式 Levenshtein distance（LevDis） 及 Longest Common Subsequence（LCS）進行實驗組（機器斷詞結果）與控制組（文本斷詞及詞性標計資料）的比較實驗。實驗設計說明如下，測試結果請見表 3。

(1) 控制組：文本斷詞及詞性標計資料
  (a) 書面文本 6,015,180 字（4,281,654 詞）
  (b) 口語文本 404,282 字（300,734 詞）

(2) 實驗組：機器斷詞及詞性標記器
  (a) CKIP 中文斷詞系統
  (b) Stanford Parser
  (c) 客語斷詞及詞性標注系統

(3) 差異評估方法
  (a) Levenshtein Distance（LevDis）：1 – Levenshtein（實驗組結果，控制組資料）
  (b) Longest Common Subsequence（LCS）：（實驗組交集控制組詞數）／控制組詞數

(4) 一致性比較配置設計
  (a) 書面文本斷詞及詞性標記資料 vs. 機器斷詞及詞性標記器結果
  (b) 口語文本斷詞及詞性標記資料 vs. 機器斷詞及詞性標記器結果

	書面文本	口語文本
CKIP 中文斷詞系統	LevDis: 30.7 LCS: 33.5	LevDis: 28.6 LCS: 29.9
Stanford Parser	LevDis: 42.1 LCS: 43.2	LevDis: 39.3 LCS: 39.8
客語斷詞及詞性標注系統	LevDis: 88.3 LCS: 89.6	LevDis: 86.2 LCS: 87.3

表 3. 斷詞效能評估結果

經測試後，客語斷詞及詞性標注系統在書面文本斷詞方面達到 88% 以上的一致性，口語文本則達到 86%以上的一致性。至於在 CKIP 中文斷詞系統及 Stanford Parser 的比較中，因其模型本身是用來處理華語，詞性標記系統亦不相同，因此平均而言產生的一致性皆較低，僅能達到 28%至 43%的一致性（如表 3 所示）。這也正說明，儘管臺灣客語與華語同屬漢語系，兩者之語言表現除了共同性外，也存在著各自的獨特性（如構詞方式或語法結構）。此外，文本中出現客語特殊字時，CKIP 中文斷詞系統及 Stanford Parser 皆會因為字串處理問題而產生亂碼，被判定為標記錯誤；THC 所使用的 Binary/UTF-8-mb4 底層資料模型，可正確比對並完整重現正確斷詞後樣貌於斷詞及詞性標記結果中。

4 結論

臺灣客語語料庫的現階段的斷詞系統主要為詞庫查找匹配法以及詞頻統計法併用。詞庫查找係基於人工事先建立好的詞庫，並採用長詞優先演算法以字符串匹配原理，將匯入語料庫之文本進行斷詞以及詞性標注。詞頻統計演算法法則是利用詞彙篩選系統的 N-gram 語言模型，計算出字串組合的出現次數，提供給語料庫工作人員參考與判讀是否為詞。接下來的目標，則是擬發展以 sequence-to-sequence 為基礎的深度學習更進一步優化斷詞系統，希冀可以解決長詞優先法較無法克服的斷詞錯誤以及歧異性問題。

參考文獻

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations,* pages 55-60.

Chu-Ren Huang, Shu-Kai Hsieh, and Keh-jiann Chen. 2017. *Mandarin Chinese words and parts of speech: A corpus-based Study.* New York: Routledge.

Fei Xia. 2000. The Part-of-Speech Guidelines for the Penn Chinese Treebank (3.0) *IRCS Report 00-07*, University of Pennsylvania.

Keh-Jiann Chen. 1992. Design Concepts for Chinese Parsers. Computational Linguistics and Chinese Language Processing, 1(1):183-204.

Peng-Hsuan Li, Tsu-Jui Fu, and Wei-Yun Ma. 2020. Why Attention? Analyze BiLSTM Deficiency and Its Remedies in the Case of NER. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI/arXiv)*.

The Unicode Consortium. 2022. *The Unicode Standard (Version 15.0.0)*. Mountain View, CA: The Unicode Consortium.

Vladimir Iosifovich Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics-Doklady*, 10(8): 707-710.

江俊龍。2010。《臺灣客家語語料庫之建置及應用》。行政院國家科學委員會成果報告（編號：97-2410-H-153-007-MY2）。臺北：行政院國家科學委員會。

江俊龍。2013。《東勢客語故事採集整理暨『臺灣客家語語料庫』的增建》。行政院國家科學委員會成果報告（編號：99-2410-H-194-136-MY2）。臺北：行政院國家科學委員會。

客家委員會。2018a。《107 年度客語能力認證初級詞彙（四縣腔）》。https://elearning.hakka.gov.tw/hakka/files/downloads/43.ods.

客家委員會。2018b。《107 年度客語能力認證中級詞彙（四縣腔）》。https://elearning.hakka.gov.tw/hakka/files/downloads/202.xls.

柯華葳、林慶隆、張俊盛、陳浩然、高照明、蔡雅薰、張郁雯、陳柏熹、張莉萍。2016。《華語文八年計畫「建置應用語料庫及標準體系」105 年工作計畫【期末報告】》。新北：國家教育研究院。

教育部。2019。《臺灣客家語常用詞辭典》。https://hakkadict.moe.edu.tw/.

詞庫小組。1998。《中央研究院平衡語料庫的內容與說明（修訂版）》。詞庫小組技術報告（編號：95-02/98-04）。臺北：中央研究院。

黃豐隆。2015。《中文與客語文句斷詞處理之研究》。104 年客家委員會獎助客家學術研究計畫成果報告書。新北：客家委員會。

謝杰雄。2006。《語料庫的建置與台灣客家語 VP 研究》。國立新竹教育大學台灣語言與語文教育研究所碩士論文。新竹：國立新竹教育大學。