客家研究 第五卷第一期
2012年6月 頁 149-186

研究論文

客語語料庫分詞原則探析

鍾屏蘭
國立屏東教育大學中國語文學系

摘 要

語料庫的建置是語言學近年來的重大發展，建立一個語料庫，並進行詞頻的統計研究，不論在華語或客語的處理過程中，最棘手也是最重要的工作就是如何分詞的問題，本論文主要便是探討客語語料庫的分詞原則。要確立分詞原則，首先不能不解決什麼是「詞」的問題，所以本論文先從客語中「字」、「詞」、「詞組」的區分進行釐清探討；接著進一步探究並建立客語的分詞原則。此一部分是參考有嚴謹學術理論基礎的中央研究院及教育部國語會的漢語分詞原則，經比較其異同後，再考慮客語本身的特性，最後加以修訂成適合客語的分詞原則。

本文所確立的客語分詞原則，主要將「詞」定義為：「語句中具有完整概念且能獨立自由運用的基本單位為詞。」並將詞的切分歸納成合併原則、切分原則及補充說明三大類，其中合併原則共十條細則；切分原則共八條；補充說明共四條，做為切分的依歸。

關鍵詞：客語、語料庫、分詞原則、詞、詞組

Analysis of Word Segmentation Principles in Hakka Corpus

Ping-Lan Chung

Department of Chinese Language and Literature
National Pingtung University of Education

Abstract

Corpus as the most important development in linguistics nowadays, is always troubling researchers whiling constructing, no matter in Mandarin or Hakka, for the word segmentation. The purpose of the thesis is discussion in word segmentation principles in Hakka Corpus . To define the word segmentation principles, the works of the idea clarification of word, term, and phrase must be done at the very beginning; and then, researching and discovering word segmentation principles in Hakka. By referencing Chinese word segmentation systems of theory-based Academia Sinica and Chinese Committee of Ministry of Education, rethinking and comparing Hakka Characteristics, the word segmentation principles in Hakka has been acquired.

The principle was ensured by defining term as the basic element with clear meaning and functioning independently. There are three principles for word segmentation: combination, chopping and supplementation. There are 10 detail principles for combination; 8 for chopping and 4 for

supplementation. Hope the principles founded as the thesis will dedicate Hakka Corpus in near future.

Keywords：Hakka, Corpus, Word segmentation principle, Term, Phrase

壹、前言

族群的維繫、認同、保存、發展，從語言文化著手是最直接最根本的方法。所以不論政府機關或民間團體，皆不斷蒐集文化傳承材料，以編輯成各式教材或字辭典。然而在客語方面，不論是在撰寫教材或編輯詞目的過程中，始終缺乏科學客觀的客語字頻、詞頻數據可供參考依循。因此，運用科學客觀的方法建立一個客語語料庫，並進行字頻、詞頻的統計，以得出客語常用的高頻字、高頻詞，提供客語教材撰寫或編輯詞目之參考，成為亟待研究解決之問題[1]。

語料庫的建置是語言學近年來的重大發展，語料庫一般可分為書面語語料和口語語料，也可根據語料的年代、地區、文體類型、使用對象來區分[2]。至於字頻、詞頻統計是就語料庫中的語言材料去進行累計「單字」字彙，以及詞彙出現的頻次，並以此為基礎去觀察語言的脈動及語言內部屬性結構分布的情形。此種研究法結合了語言學、統計學及心理學等相關領域知識，若透過不同角度去了解運用，字

---

1 賴惠玲：〈客語語法研究議題的開發：以語料庫為本〉，《96 年補助大學校院暨獎助客家學術研究計畫成果發表會論文集》，台北：行政院客家委員會，2008 年，頁 153-164。鍾屏蘭：〈從語料庫的開發探討客語教材的編輯與出版〉，《屏東教育大學學報—教育類》第 36 期，2011 年，頁 347-370。

2 語料庫(corpus)的建立和研究在國外已經行之有年，在眾多語料庫中，英語語料庫建立最早，而且類型最多。另外世界上其他許多語言也都建有語料庫。賴惠玲在〈客語語法研究議題的開發：以語料庫為本〉一文中，對各國語料庫的建置有概略的介紹：「除英語外，許多其他語言也紛紛建立語料庫，包括德語、義大利語、西班牙語、瑞典語、葡萄牙語、俄語、荷蘭語、威爾斯語、波斯尼亞語、保加利亞語、以色列語、塞爾維亞語、日語、泰語，和中文。同樣有以書寫語料為主的，例如義大利文語料庫 CORIS/CODIS；也有以口語語料為主的，像保加利亞語語料庫 A corpus of spoken Bulgarian；也有包括詞頻和語法分析的，如德語語料庫 NEGRA Corpus。」同時誠如賴惠玲所言，語料庫(corpus)的建立和研究在國外已經行之有年，蒐集大量的語言資訊可提供教學、比較語言學、自然語言處理等各種不同學科或跨領域的研究。

頻、詞頻統計結果具有多方面的參考價值[3]。尤其拜當今電腦進步之賜及統計學觀念的運用，將字頻、詞頻統計的方法應用在常用字詞的統計上，便能對語言的學習應用傳承，提供更為廣大而無可限量的價值。尤其對教學上階梯教材的編輯影響非常大，學前及低中高年級教材的用字用詞，應如何訂出標準，適當的頻率調查就是一個重要的憑據[4]。

建立一個語料庫，並進行字頻、詞頻的統計研究，不論在華語或客語的處理過程中，最棘手也是最重要的工作就是如何分詞的問題[5]。本論文主要的研究目的便是探討能提供客語教材撰寫或編輯詞目的常用高頻字、高頻詞的客語平衡語料庫分詞原則[6]。要確立分詞原則，首先不能不解決什麼是「詞」的問題，所以本論文先從釐清客語中「字」、「詞」、「詞組」的區分進行探討；接著承繼探討結果，進一步探討並建立客語的分詞原則。此一部分本研究擬參考有嚴謹學術理論基礎的中央研究院及教育部國語會的漢語分詞原則，經比較其異同後，再考慮客語本身的特性，最後加以修訂成適合客語的分詞原則。

---

[3] 參見曾榮汾：〈字頻統計法的實例－國小常用字彙統計析述〉，《警學叢刊》第 27 期，頁 83。

[4] 參見曾榮汾：〈字頻統計法及學術應用〉，《警學叢刊》第 25 卷第 2 期，》頁 32。吳敏而《國民小學兒童常用字詞詞彙資料庫之建立與初步分析(Ⅲ)》(台北：台灣省國民學校教師研習會研究室，1998 年)，頁 3。劉杰〈漢語超高頻詞分類統計與分析〉(收於胡盛侖主編：《語言學與漢語教學》，北京：北京語言學院出版社。1990 年)，頁 266-278。

[5] 客語語料庫的建置問題包含語料收集的對象與方式，口語及書面語，及口語如何轉寫成文字等問題，另外還有客語語料庫如何才能達到「平衡語料庫」等。以上參見鍾屏蘭：〈從語料庫的開發探討客語教材的編輯與出版〉，《屏東教育大學學報—教育類》第 36 期，2011 年， 頁 347-370。

[6] 平衡語料庫如台灣方面自一九八六年開始，由中央研究院完成了《現代漢語平衡語料庫》，語料庫共收約五百萬詞，語料主題內容比例為：哲學百分之十、科學百分之十、社會百分之三十五、藝術百分之五、生活百分之二十，文學百分之二十。資料取得可分為書面及口頭資料兩類，但目前語料來源偏重於書面語，約佔百分之七十。

貳、客語中「字」、「詞」、「詞組」的區分

確立分詞原則，首先必須要分清楚什麼是「詞」的問題。語言文字是溝通的工具，由於在漢語中一個句子才能表達完整的意思，所以句子是漢語中最大的語法單位。但漢語中每個句子都是方塊漢字寫出來的，「字」、「詞」、「詞組」的界線並不明顯。在這裡首先要探討的是「字」和「詞」的區分，其次則是「詞」和「詞組」的區分。

一、 「字」和「詞」的區分

字和單音詞在形體上都是一個方塊漢字，但他們在文法上卻有完全不同的性質與功能。王力說：「字在語法上是沒有地位的...若說它們是一種形體單位，那它只是文字上的單位，適足以證明語言本身無所謂字[7]。」張壽康也說：

> 構詞的單位是詞素，不是字。...有時寫出來的一個形體，這個形體正好是紀錄了語言中的一個詞，或是語言中具有意義的一個詞素。比如『火』在文字上講，它是一個象形字，在語言上這個形體紀錄了語言中的一個詞（一個詞素構成），或者紀錄了『火車』中的『火』（一個構詞單位，詞素）。可見詞素和字是不容相混的。」詞素就是「構成詞的具有意義的構詞單位[8]。

羅肇錦教授則說：

---

[7] 王力：〈詞和仂語的界線問題〉，《中國語文》（北京：北京大學中國語言文學系，1953年3月號）頁 3-8。

[8] 張壽康：〈略論漢語構詞法〉，《中國語文》（北京：北京大學中國語言文學系，1957年6月號）

> 詞的範疇與字最大不同的是它可以單獨表達完整概念，縱使是單音節和字一樣的形式，但已代表了完整的概念，就算是詞不是字了。譬如廈門、潮州的〔tsai〕雖然只有一個音節，但它已完整地表達了〔知道〕或〔曉得〕的概念，它就是道道地地的詞，而不是字[9]。

歸納上述說法，字和單音節詞雖有重疊，但嚴格來說，「字」是形體的單位，是在書寫上可以單獨出現或獨立使用的單位；「詞」則是意義的單位，一個詞可能是一個字，例如「手」，也可能是兩個以上的字，例如「火車」。因此「字」是屬於文字學的單位，應將「字」在文字學去探討。但是若從語言學的角度來看，從語言文字的溝通功能來看，「詞」的意義遠大於「字」，則應該從詞彙學的角度來研究具有意義的「詞」。

(一)　「詞」與「詞組」的區分

何謂「詞」？根據教育部國語推行委員會「八十七年常用語詞調查報告書」說：「語句中具有完整概念且能獨立自由運用的基本單位為詞[10]。」和這個相近的說法，有葉蜚聲、徐通鏘的：「詞是造句的時候能夠獨立運作的最小單位[11]。」這個定義強調"獨立運用"和"最小"兩層意思。所謂獨立運用，是它在造句中能夠到處作為一個單位出現；所謂最小，就是說不能擴展，或者說就是中間不能插入別的成分。如"朋友"裏面的"朋"和"友"不能獨立運用，"朋友"是現代漢語裏能夠獨立運用的最小單位，是一個詞[12]。

至於「詞組」，是由「詞」組成，比詞大一級的單位。根據中國

---

[9] 羅肇錦：《客語語法》（台北市：學生書局，1988 年），頁 420。
[10] 引自教育部國語推行委員會：《常用語詞調查報告書》（台北市：教育部，1998 年）。
[11] 引自葉蜚聲、徐通鏘：《語言學綱要》（台北：書林出版社，1993 年）頁 103 。
[12] 參見葉蜚聲、徐通鏘：《語言學綱要》頁 103 。

國家標準 GB/T13715-92《信息處理用現代漢語分級規範》對「詞組」的定義是：「由兩個或兩個以上的詞，按一定的語法規則組成，表示一定意義的語言單位[13]。」葉蜚聲、徐通鏘也有近似的說法：「詞組是詞的組合，它是句子裏面作用相當於詞，而本身又是由詞組成的大於詞的單位[14]。」而句子裏的絕大部分詞組是根據表達的需要臨時作出的組合。這類詞組，一般按照語法規則把有關的詞組織起來，稱之為自由詞組。另外語言中也有不少必須完整記住詞的固定組合，這類詞組叫做固定詞組，如"北京大學"等[15]。從上面的解釋，可以瞭解「詞」與「詞組」的區分。

以上是有關漢語對「字」、「詞」、「詞組」的區分[16]，客語屬漢語的一環，書寫文字和語法結構不脫漢語的體系，是以進行客語的詞頻研究，一方面要藉助漢語對「字」、「詞」、「詞組」的區分來從事斷詞工作，一方面也要以客語為對象，深入瞭解客語的用字、構詞、語法，建立屬於客語為中心的分詞原則[17]。

參、客語分詞原則的文獻探討

現代漢語的詞頻語料庫建置，大陸方面早有嘗試，由北京語言學院執行，從一九七九年十一月開始至一九八五年七月為止，歷時五年

---

13 引自傅永和《中文信息處理》附錄 5（廣州：廣東教育出版社，1999 年）。

14 引自葉蜚聲、徐通鏘：《語言學綱要》頁 102 。

15 參見葉蜚聲、徐通鏘：《語言學綱要》頁 103 。

16 有關「字」、「詞」、「詞組」的區分，許多語法書籍皆有深入之析論，如：
趙元任著，丁邦新譯：《中國話的文法》。香港：香港中文大學出版社，1980 年。
竺家寧：《漢語詞彙學》，台北市：五南圖書公司，1999 年。

17 可參考羅肇錦：《客語語法》，台北：學生書局，1988 年再版。
鍾榮富：《福爾摩沙的烙印—台灣客家話導論（上冊）》，台北市：行政院文化建設委員會，2001 年。 鍾榮富：《福爾摩沙的烙印—台灣客家話導論（下冊）》，台北市：行政院文化建設委員會，2001 年。

零八個月，最後完成了《現代漢語頻率詞典》。台灣方面，則自一九八六年開始，由中央研究院資訊所與語言所跨所合作的一「中央研究院資訊科學研究院中文詞知識庫小組」，完成了中文詞頻語料庫的研究。其中包括《現代漢語平衡語料庫》、《近代漢語語料庫》、《上古漢語語料庫》等，提供了極具學術參考價值之資訊。其中的《現代漢語平衡語料庫》，是中文語料庫一個新的里程碑，是世界上第一個有完整詞類標記的漢語平衡語料庫。另外在我國教育部方面，亦自一九九五年起，針對國內語言環境逐年進行「常用語詞調查」工作，亦建置有語料庫進行語料蒐集分析，研究成果包括〈八十七年常用語詞調查報告書〉、〈八十七年口語語料調查報告書〉、〈八十七年口語問卷調查報告書〉等，對語文教育的推展有相當貢獻。

本論文主要在建立客語的分詞原則，此一部分本研究基本上參考有嚴謹學術理論基礎的中央研究院及教育部國語會的漢語分詞原則[18]，經比較其異同後，再考慮客語本身的特性，最後加以修訂成適合客家語的分詞原則，因此以下先依序介紹上述兩種研究的分詞原則。

一、中央研究院漢語平衡語料庫分詞標準

中央研究院漢語平衡語料庫的研究小組，所訂定分詞的基本原則

[18] 在國內涉及分詞標準的研究，多運用於詞頻統計方面，國語方面的有
國立政治大學教育學系：《兒童常用詞彙研究》(1982)
劉英茂，莊仲仁，吳瑞屯：《中文詞及敘述單位分析原則》(1987)
葉明德：〈華語文常用詞彙頻率等級統整研究〉(1997)
吳敏而：《國民小學兒童常用字詞彙資料庫之建立與初步分析》(1998)
在鄉土語言方面的有
陳惠玉：《臺灣臺中市何仁里臺語常用詞彙調查與詞頻之初步研究－－並與台中縣新里里做比較研究》(2003)
劉秀珍：《客家語教科書常用詞彙與詞頻之初步研究－以高市版為例》(2006)
謝杰雄：《語料庫的建置與台灣客家語 VP 研究》，(2006)
以上各項研究，首先皆不免觸及分詞原則，然各家分詞原則不一，並未有共同標準，尤其鄉土語言部分還相當混亂，但還是提供後來研究者相當的學術奠基工作。

規範如下[19]：

基本原則

(一)語意無法由組合成分直接相加而得到之字串應該合為一分詞單位。

合併原則

(二)詞類無法由組合成分直接得到，應該合為一分詞單位。 合併原則

輔助原則

(一)有明顯分隔標記應該切分之。 切分原則

(二)附著語素盡量和前後詞合為一個分詞單位。 合併原則

(三)使用頻率高或共現率高的字串盡量視為一個分詞單位。 合併原則

(四)雙音節結構之偏正式動詞盡量視為一個分詞單位。 合併原則

(五)雙音節加單音節之偏正式名詞盡量視為一個分詞單位。 合併原則

(六)內部結構複雜之詞盡量切分之。 切分原則

詳細的說明及例證如下：

訂定分詞標準的首要工作是定義切分字串的基本單位。因此我們定義一個具有獨立意義，且扮演特定語法功能的字串應視為一個詞。根據定義，動詞、名詞、副詞、定詞、量詞、介詞、方位詞、連接詞、語助詞、感歎詞皆可依類一一斷開。

除了定義外，必須另有原則規範分詞，我們提出兩條基本原則以及六條輔助原則，以求在語料庫的斷詞部份能達到一個符合語感、分

[19] 摘自《中央研究院/現代漢語平衡語料庫》，網址〈http://dbo.sinica.edu.tw/SinicaCorpus/〉。

析一致、並具語言學專業要求的水準。

基本原則

基本原則是從語意與語法兩方面來說明分詞單位。以基本原則作為指導原則，我們便可以在語言學理論上找到分詞依據，使分詞標準有執行的歸依。

(一) 語意無法由組合成分直接相加而得到之字串應該合為一分詞單位。合併原則

這是一條很重要的分詞細則，凡是組合後意義起變化的字串皆應視為一個詞。試舉一例："撞期" 依此原則必須視為一個詞，但是「撞山」仍可保持斷開，視為動詞加賓語之動詞組。此原則的適用面很廣。即便是一個字串表面有明顯的詞組甚至句子的構造，但凡意義失去組合性時亦應合為一個詞。因此下列字串皆應視為一個分詞單位，例如：飛黃騰達（成語），撞期、吃醋（動詞組），或多或少（副詞片語），十二萬分（定量結構），五月（定名結構，不是五個月）、三樓（定名結構，不是三層樓），談談（重疊結構，表嘗試）、「坐坐」就走（重疊結構，含短暫貌）、辛辛苦苦（重疊結構，表程度加強）、片片、一片片（重疊結構，具泛指意涵）、「好好」孝順父母（重疊結構，表盡力）...等。

合併結構，像是「上下課、高中職、中山南北路」，依此原則也應該合併為一個詞。因為該字串的意義並非「上」加「下課」、「高中」加「職」，「中山南」加「北路」，而是「上課」加「下課」、「高中」加「高職」、「中山南路」加「中山北路」，可見合併結構的意義不等於組合意義，故應合併。唯帶專名之合併詞，像是「台北市長」（「台北市」加「市長」）、「新竹縣政府」（「新竹縣」加「縣政府」），因切分後前方的專名和後方的名詞皆可獨用，意義可以組合成，故仍予以切分。

(二) 詞類無法由組合成分直接得到，應該合為一分詞單位。合併原則

此原則分兩部份：一、該字串之語法功能不符合組合結果。例如：動作及物動詞「喝、吃、聽」前面加「好」構成「好喝、好吃、好聽」，不能再加賓語，成為不及物，且能被程度副詞「很、十分、非常」修飾，與原來的語法特性不同，故可視為一個分詞成分。二、該字串之內部結構不符合語法規律。例如：「那隻狗不會游水」中「游水」指的是「在水裡游」，但「游」是不及物動詞，不可直接後接名詞。因此，「游水」不符合動詞「游」的語法規律，故應合併之。

輔助原則：

除了基本的理論性原則外，我們也必須有操作性原則，視分詞的實際狀況設定分合的依據。相對於基本原則的不變性，輔助原則富於彈性，可能依時代的演變或視情況的需要而有所增減。

(一) 有明顯分隔標記應該切分之。 切分原則

一個詞可能中插別的成分，或是一個詞、一個標點符號，或是英文等外來語，在此情況下，不得不將之斷開。例子有：

動賓中插：洗了一個澡

述補中插：打得破、打不破

交互中插：彎下腰去、喘不過氣來

合併中插：動詞：上、下課

1.當重疊結構之意義未失組合性，則不予合併。例如「坐坐坐、哈哈、叮噹叮噹」不須組合成一個詞，因該字串之語意可從每個成分組合而成，並無多出的詞意。

2.但像「養得起、養不起」、「處得來、處不來」因無相對應之「養起」、「處來」，所以視為一分詞單位，不予切分。

名詞：父、母親，高中、職，中山南、北路

定量：本（二）月，七、八月，1995、6 年，三 到 四月

外來語：BBS 站、user 們、txt 檔

數詞及表時間、地點或編號之詞雖含有標點符號，但是我們認為這些符號不具標點符號功能，所以不算是中插，故下列情形仍維持合併。

七、五００，三・六，2/28（二月二十八號），3：30（三點三十分），二０～一號（門牌號碼），AB-8888（車牌號碼）

(二) 附著語素盡量和前後詞合為一個分詞單位。 合併原則

附著語素指的是有獨立意義卻無法獨立扮演一個語法功能的語素。例如：「立」可分為三個語素：一、表「站立」，是不及物動詞；二、表「建立」，是及物動詞；三、表「立刻」，是附著語素，多半出現在「立刻」「立即」的詞中。由於書面語文白夾雜，常常可見附著語素獨用情形，如「情勢立告逆轉」。此例中，我們依此原則將「立告」合為一個偏正式複合動詞。又例如「吝」也是個附著語素，多半出現在「吝嗇」「吝惜」中，但依此原則「不吝」「吝於」也會被合併成一個動詞。不過，我們也可能遇到附著語素無法和前後詞合成一個語言成份的情況，如「為什麼還吝而不做呢？」我們也只好將附著詞「吝」斷開，依其在該句中所扮演的功能給予詞類。

現代漢語中有許多詞具詞綴特色，常用來和其它詞結合，具有一致的意義，並往往決定該組合詞之詞類（詞頭多半無此功能，但詞尾多半都有）。詞綴也是附著語素，因此帶詞綴之字串也應合為一詞。例如：「演員、救生員、隊員、查哨員、技術員、組成員、督導員、郵務員...」「現代化、合理化、泛政治化、民營化、地下化、本土化、小丑化、多元化...」。這些詞在詞典中收不勝收，必須藉構詞律由電腦自動結合成詞。但是從電腦處理的角度來看，在初步的處理時並不容易達成自動合詞的目標，必須依不同層次分階段達成，因此依附著

詞結合難易的程度分為詞綴及接頭/接尾詞。目前我們挑選出衍生性強的接頭詞及接尾詞作為分詞的參考依據，請見附錄 1。此外，「的、地、之」雖通常被視為詞綴，但是由於下列兩個理由我們不將它們當作詞綴處理。一、它們所附著之詞幹無詞類限制，無論名詞、動詞、副詞、數量詞甚至句子皆能帶這些詞綴，這和一般詞綴表現不一；二、它們常和詞組結合，如「常常和官員打交道的記者」「欲退出選委會之人」，這點也和一般詞綴的衍生方式不同，所以這三個詞將和前後詞一律斷開。

(三) 使用頻率高或共現率高的字串盡量視為一個分詞單位。合併原則

有些字串因為常常一起出現，所以其結合較緊密，較少見中插情形。縱使這些字串完全不符合上述三條原則，即它們的語意、語法功能未失組合性、也不含附著語素，仍可因此原則合為一個詞。例子有：

動詞：並列結構：進出、收放、……

偏正結構：大笑、改稱、……

動賓結構：關門、洗衣、卸貨、……

名詞：並列結構：春夏秋冬、輕重緩急、男女、花草、……

偏正結構：象牙、……

副詞：並列結構：暫不、既已、不再、……

這條原則有兩個難處，在於如何得出副詞：並列結構：暫不、既已、不再、……等之使用頻率，以及區分值應該設在何處。這不是個容易解決的問題，在沒有一套可遵循的標準法則時，對於一些字串此原則是否適用就成了見仁見智的情形，因此這條原則可視為一條參考原則 3。

(四) 雙音節結構之偏正式動詞盡量視為一個分詞單位。 合併原則

當一個字串具有動詞之語法功能，若符合雙音節結構，且是偏正

結構，即可視為一個分詞單位。因此，在「緊追其後」中的「緊追」雖然語意、語法功能未失組合性，不含附著語素，也不是常見字串，仍可依此原則合併之。此原則並不用於動賓及主謂式複合動詞。所以「警察無故擒 人」「股市陷入價 升 量 減的走勢」中「擒人」和「價升量減」不會因此原則合併。

(五) 雙音節加單音節之偏正式名詞盡量視為一個分詞單位。合併原則

有些單音節的名詞本身可獨立成詞，但是常與前面的雙音節成分結合緊密，可視為一分詞單位。例如：「線、權、車、點」所構成的成分「防衛線、捷運線、木柵線、平均線；監護權、領導權、使用權、發言權、優先權；垃圾車、交通車、宣傳車、娃娃車；著眼點、立足點、共同點、爭議點」。從與其他成分結合的觀點來看，這些單音節名詞也可視為接尾詞，與衍生性附著語素並列在接尾詞之列。因此我們需要一部標準辭典作為區分詞和非詞的依據。

(六) 內部結構複雜之詞盡量切分之。 切分原則

這是一條暫行原則。下列結構雖然依前述五條細則是應合為一個詞，但由於合併起來過於冗長，故不予合併。

1. 詞組帶接尾詞：太空 計劃 室、塑膠 製品 業
2. 動詞帶雙音節結果補語：看 清楚、討論 完畢
3. 專有名詞：專名帶普名：胡 先生、平漢 鐵路、二二八 事變、永新 加油站
4. 詞組或句子之專名，最常見為書名、戲劇名、歌曲名：

鯨魚 的 生 與 死（書名）、那 一 年 我們 都 很 酷（戲劇名）

5. 複雜結構：省 自來水 公司、台北市 第一 信用 合作社

輔大 景觀 設計 系、中文 分詞 規範 研究 計畫

6. 正反問句：喜歡 不 喜歡、參加 不 參加
7. 動賓結構、述補結構之動詞帶詞綴時，不予合併。
   例：寫信 給、分紅 給、取出 給、退回去 給

綜合上述，分詞原則共有定義、兩條基本原則、以及六條輔助原則。

定義：具有獨立意義，且扮演固定詞類的字串視為一分詞單位。

基本原則：

(一) 語意無法由組合成分直接相加而得到之字串應該合為一分詞單位。合併原則
(二) 詞類無法由組合成分直接得到，應該合為一分詞單位。合併原則

輔助原則：

(一) 有明顯分隔標記應該切分之。 切分原則
(二) 附著語素盡量和前後詞合為一個分詞單位。 合併原則
(三) 使用頻率高或共現率高的字串盡量視為一個分詞單位。合併原則
(四) 雙音節結構之偏正式動詞盡量視為一個分詞單位。 合併原則
(五) 雙音節加單音節之偏正式名詞盡量視為一個分詞單位 合併原則
(六) 內部結構複雜之詞盡量切分之。 切分原則

二、教育部國語會《八十七年常用語詞調查報告書》分詞標準[20]

(一)、基本原則

1、詞的定義：語句中具有完整概念且能獨立自由運用的基本單位為詞。

2、使用頻率高或連用程度強的字串應視為一分詞單位。

如：動詞"關門、知道、進出、密談"等。
名詞"個人、手錶、筆墨、白酒"等。

[20] 摘自《教育部國語推行委員會/八十六年常用語詞調查報告書》，網址〈http://www.edu.tw/files/site_content/m0001/86news/index.htm〉。

形容詞"火紅、淡綠、飛快、膚淺"等。

介詞"關於、對於、按照、根據"等。

副詞"必會、實應、仍然、既已、不再、也罷"等。

＊正反義及相對詞素結合詞因連用程度強，不予切分。

如：大小、長短、好壞、內外、真假、男女、是非、動靜　等。

＊形容詞與名詞連用無特定意義者，予以切分。

如："新　衣服"、"小　花朵"等。

但"新人類"、"大人物"等詞具有特定意義，屬偏正式名詞，不予以切分。

3、諺語、名言、口號、歇後語等常連用但內部結構複雜者儘量予以切分。

如：一　動　還　不如　一　靜

幾　家　歡樂　幾　家　愁

保密　防諜　人人　有　責

啞巴　吃　黃蓮　有　苦　說　不　出

4、語詞可以其在句子中的功能判定是否為一分詞單位。如：

「之前」－表「助詞」＋「方位詞」時切分，如：「在此之前」。

表時間功能時不切分，如：「我之前說過」。

「以後」－表「助詞」＋「方位詞」時切分，如：「在此以後」。

表時間功能時不切分，如：「他以後不敢了」。

「最近」－表「副詞」＋「形容詞」時切分，如：「距離最近」。

表時間功能時不切分，如：「我最近買了一隻錶」。

「最後」－表「副詞」＋「形容詞」時切分，如：「落在最後」。

表時間功能時不切分，如：「最後，他終於明白了」。

(二)、處理原則

1、專有名詞和固定語視為一分詞單位。

(1)專有名詞

人名、地名、國名、公司行號、機關學校、行政區域、產品名、書名、電影名稱等專有名詞特有所指，應視為一分詞單位，不予切分。

如：梁啟超、九份、美利堅合眾國、大同公司、北京大學、臺北市、文心貴族、日清杯麵、紅樓夢、亂世佳人等。

＊「陳　太太」、「王　太」、「陳　家」、「余　家」予以切分。

專科語詞視為一分詞單位。

如：光學式自動識別計測系統、卡波西氏肉瘤等。

器官名稱，若左右器官的功能有所不同，一律視為一分詞單位。

如：左腦、右腦外來音譯語詞視為一分詞單位，不予切分。譯字不同，視為不同語詞。

如：瑪麗蓮夢露、麥當勞、香奈兒等。

(2)固定語

成語：具有文獻典故來源，且具多層表義效果的固定語。

如：對牛談琴、人去樓空、紙上談兵、五十步笑百步

慣用語

A. 一般口語習用，表示特定語義的固定語。

如：敲竹槓、吃豆腐、吃醋、灌水、吹牛、翹辮子、老掉牙、落湯雞

B. 正反問慣用語。

如：好不好、要不要、可不可以、喜不喜歡

＊"喜歡不喜歡"一句為完整結構，應切分為"喜歡　不喜歡"。

C. 四字格慣用語。結構為四個音節的固定語，形式類似成語，但不具文獻典源，又無多層表義效果。因其為慣用語的一類，故不予以切分，如：寶裡寶氣、馬馬虎虎、亂七八糟。

2、詞綴和前後詞合為一分詞單位。

詞綴為附加在詞根上的構詞成份，詞義虛化，構詞能力強，

有前綴、中綴、後綴三種。

(1)前綴詞：有“老、阿、小”。如：老虎、阿婆、小張。

(2)中綴詞：有“里”，如：糊里糊塗。

(3)後綴詞：有“子、兒、頭、巴、麼、們”。如：房子、花兒、罐頭、泥巴、這麼、我們”。

3、有固定意義之重疊詞視為一分詞單位。

詞素或詞的重疊有特定的語法功能而不是修辭上的反覆時，因其通常具有擴大語義的效果，故視為一分詞單位。重疊詞分為下列幾個形式：

(1)詞素重疊

AA 式。如：看看、想想、走走、剛剛。

AAB 式。如：嚐嚐看、散散步、聊聊天、幫幫忙。

ABB 式。如：香噴噴、病懨懨、水汪汪、一些些。

AABB 式。如：乾乾淨淨、快快樂樂、吵吵鬧鬧。

(2)鑲嵌重疊

A 一 A 式。如：看一看、想一想、做一做、聽一聽。

A 了 A 式。如：看了看、笑了笑、問了問、算了算。

A 了一 A 式。如：看了一看、笑了一笑。

A 來 A 去式。如：走來走去、想來想去。

(3)全詞重疊

ABAB 式。如：研究研究、討論討論。

4、簡稱與結合語視為一分詞單位。

(1)簡稱

縮語：就原詞抽取關鍵詞重組，如：文建會、證交稅。

略語：說原詞截取部分詞素而成，如：公賣局、文化大學。

統稱：以數字概括幾項有關的內容，如：三軍、十大建設。

簡代詞：用一語素來取代全稱，如：臺、閩、港。

(2)結合語：由兩個或兩個以上的詞併合並加以節縮而成的詞，如：

入出境、中小學、工商業、中山南北路。

＊套裝合併之形式視為專有名詞，亦不予以切分。如：台北市長、新竹縣政府、教育部長。

5、所有由數字組合成之定語，不論以國字或阿拉伯數字表達均視為一分詞單位。如：100、一千兩百。

(1)有關時間的詞目，如：

“一九九八年二月十五日”，切分為“一九九八年　二月　十五日”。

＊因其為數名結構，是“二月”，不是“二個月”。

(2)數字定語與量詞的組合，如：

“三百兩”、“100 個”，切分為“三百　兩”、“100　個”。

(3)地址予以切分，如：

“臺北市　信義路　二段　15 號　3 樓”。

6、語詞合併用後詞性改變，且能被程度副詞（十分、非常、很）修飾，可視為一分詞單位。

如：「好＋動詞」－好吃、好看（副詞＋動詞→形容詞）

＊「有＋名詞」之形式，名詞為雙音節切分；單音節時不分。

7、補語結構與前詞視為一分詞單位，唯當述語或補語為雙音節時切分。

(1)述補結構之述語為雙音節時切分，單音節不分。

如：到－接觸　到、認知　到

為－譯為、流為、選拔　為

成－擠成、形成、堆積　成

作－鑄作、換作、轉變　作

(2)補語為結果補語且是雙音節時切分；單音節時不分。

如：哭濕　枕頭、爬上　山頭

看　清楚、清洗　完畢

(3)補語為趨向補語且是雙音節時切分；單音節時不分。

如：走　回來、挽救　回來

＊述補中插切分：打　得　破、打　不　破

此二詞對應「打破」一詞，而「得」字視為助詞、「不」字視為否定之交互運用，均予以切分。但「處得來」、「處不來」等詞無相對應之「處來」，視為一分詞單位，不予切分。

8、否定語料於《重編國語辭典修訂本》中有收錄者，視為一分詞單位；未收錄者予以切分。如：

「不」－不但、不然、不好、不夠、不變、不想　等。

「沒」－沒有、沒事、沒想到　等。

「無」－無法、無非、無妨、無聊、無情　等。

以上各詞因《重編國語辭典修訂本》收錄視為一分詞單位。

＊所有否定語料另置於附錄，以供參考。

9、附著於詞根表時態或特定語法功能之詞素，予以切分。如：

「去　過」－表過去

「看　了」－表動作之完成態

「燃燒　著」－表動作之持續態

10、某些特定詞素做為起始字時，因可任意組合，視為詞組予以切分。如：

(1)雖：“雖　大”、“雖　有”

(2)未：“未　遠”、“未　減”

(3)正：“正　濃”、“正　香”

(4)很：“很　好”、“很　美”

(5)已：“已　到”、“已　成”

(6)本：“本　指”、“本　篇”

(7)可：“可　改變”、“可　吃”

(8)非常：“非常　好”、“非常　差”

＊“未來”做名詞用時表達獨立概念，不予以切分。

11、方位詞與前後詞切分。

(1)單音節方位詞：上、下、前、後、裏、內、外、中 等。

(2)雙音節方位詞：上面、東邊、裏頭、內部 等。

如："睡 前"、"屋 裏"、"街 上"、"討論 中"、"事實 上"、"三 天 後"、"桌子 上面"、"在 我 之後"。

12、"的、地、之、得"等助詞與前後詞切分。

如："我 的 志願"、"迅速 地 蔓延"、"遊行 之 人"、"看得 很 清楚"。

13、介詞與前後詞切分。

介詞為位於名詞或名詞性詞組之前，合起來表示方向、對象、時間、處所等的虛詞。如：『死於安樂』的『於』。常用的介詞有：把（拿）、比、並、方、打、由、對、替、連、給、跟、管、和、往（望）、從、就（依照）、對、於、向、自、讓（被）、叫（被、受）、在、為、對於、關於、由於、至於 等。

14、其他

(1)英文、日文或中英合併詞予以保留，成果以附錄方式呈現。

(2)詞素中有統一用字者，因二字皆可通用，不予統一，各自視為獨立單位。如：部分（部份）、鞭炮（鞭砲）、分布（分佈）。

三、中央研究院與教育部分詞差異討論

以上中央研究院《漢語平衡語料庫》分詞標準，教育部國語推行委員會《八十七年常用語詞調查報告書》分詞標準，可以說均是根據語言學的理論基礎進行詞的切分。僅有極小部份之差異，茲分析於下：

（一）、專有名詞之處理

這裡所謂專有名詞之處理，兩者不同處在於中央研究院《漢語平衡語料庫》分詞標準是：

帶專名之合併詞，像是「台北市長」（「台北市」加「市長」）、「新竹縣政府」（「新竹縣」加「縣政府」），因切分後前方的專名和後方的名詞皆可獨用，意義可以組合成，故仍予以切分。

這種情形在教育部國語推行委員會《八十七年常用語詞調查報告書》中則是：

套裝合併之形式視為專有名詞，亦不予以切分。如：台北市長、新竹縣政府、教育部長。

這是兩者在套裝合併形式的專有名詞有不同的處理方式，一從分，一從合。對此問題，本研究認為如要凸顯客語的語言文化特色，從合的方式較能顯現，故本研究類似此種情形擬採取教育部的分詞原則。

（二）、否定詞之處理

中央研究院《漢語平衡語料庫》對否定詞並未有特別的規定，而是在副詞的地方加以討論：

使用頻率高或共現率高的字串盡量視為一個分詞單位。合併原則

如並列結構：暫不、既已、不再、.....這條原則有兩個難處，在於如何得出副詞使用頻率，以及區分值應該設在何處。這不是個容易解決的問題，在沒有一套可遵循的標準法則時，對於一些字串此原則是否適用就成了見仁見智的情形，因此這條原則可視為一條參考原則。

至於這種情形在教育部國語推行委員會《八十七年常用語詞調查報告書》中則是：

否定語料於《重編國語辭典修訂本》中有收錄者，視為一分詞單位；未收錄者予以切分。如：
「不」－不但、不然、不好、不夠、不變、不想　等。
「沒」－沒有、沒事、沒想到　等。
「無」－無法、無非、無妨、無聊、無情　等。
以上各詞因《重編國語辭典修訂本》收錄視為一分詞單位。
＊所有否定語料另置於附錄，以供參考。

可以說規範的相當明確，但是國語會以辭典收詞為依據，處理棘手問題看似容易，可惜並無特別敘明理由。

否定詞之處理上，由於中研院及教育部未能一致，因此在客語的分詞上形成難以處理的問題。由於客語對應國語的否定詞主要是「毋」（不）及「無」（沒、沒有）兩個字，如「毋係」、「毋要」、「毋知」、「毋曉得」、「毋中意」...，「無錢」、「無人」、「無影」、「無相關」、「無要緊」...本研究衡諸各種情況後發現，有些是已然構成單一副詞，如「無論」、「毋使」，或已然構成單一連接詞者，如「毋過」等，應合併為「詞」；有些則是去掉否定詞素後原來的詞變成沒有意義，如「無採」（可惜之意）、「無膽」（膽小之意）、「無錢」（貧窮之意），也應合併為「詞」。為了分詞的簡便不易出現不一致之情形，採取的方式是：觀察去除否定詞素後所接之詞，若為雙音節詞時則視為詞組予以切分，如「毋 曉得」、「毋 中意」、「無 相關」、「無 要緊」。去除否定詞素後所接之詞若為單音節時則不分。如「毋係」、「毋要」、「毋知」、「無錢」、「無人」、「無影」都視為一詞不予切分。

（三）、正反問慣用語的處理

正反問慣用語的處理，中央研究院與教育部處理方式不太一致。中央研究院《漢語平衡語料庫》是採用：

正反問句：喜歡 不 喜歡、參加 不 參加

教育部國語推行委員會《八十七年常用語詞調查報告書》中則是：

正反問慣用語。
如：好不好、要不要、可不可以、喜不喜歡
＊"喜歡不喜歡"一句為完整結構，應切分為"喜歡　不喜歡"。

本研究則認為，正反問慣用語應視其有無縮略來分別處理較適當。如國語的「好不好」、「要不要」「喜歡不喜歡」、「參加不參加」，可切分成「好 不 好」、「要 不 要」「喜歡 不 喜歡」、「參加 不 參加」。至於「可不可以」、「喜不喜歡」，這類有所縮略的正反問慣用語，則不予切分。同樣的，客語中有同樣的情形，如「合毋合意」(中意否)「合意毋合意」(中意否)。這時前者因有縮略的情形，故不切分；後者則可切分為「合意 毋 合意」。

(四)、通用字的處理：

教育部特別在「其他」項中列一條說明，主張像部分（部份）、鞭炮（鞭砲）、分布（分佈）等，詞素中有統一用字者，因二字皆可通用，不予統一，各自視為獨立單位。至於中研院對此並無主張。然而本文主張採標準用法，予以統一，即以前三例而言，皆採前者為標準。

本文主張採標準用法，予以統一的原因，主要是由於客語長期以來並未發展出標準化、規範性的文字化文本，部份客語有音無字，撰寫者多以自行造字或假借方式處理，各自為政的結果，便形成同音同義卻不同字的情形。如相當於國語的「玩」的意思的客語，即有「聊」、「料」、「尞」、「燎」、「嫽」等不同寫法；相當於國語的「邊」的意思的客語，亦有「唇」、「脣」、「漘」、「滆」等不同寫法，相當於國語的「不」這個意思的，則更有「毋」、「m」、「不」、「冇」、「唔」、「莫」、

「無」等混用情形。又再加上這些字與其他詞素結合成為另一個詞，情形就更形複雜。另外在複詞方面，如國語的「斗笠」一詞，客家語便有做「笠母」、「笠麻」「笠蔴」、「笠嫲」各種不同寫法。上述這種文字書寫的異用情形，往往也形成一堆怪字及生難字，相當不利於語料庫的建置、研究及推廣，也會導致研究軟體設計及統計上的困擾，尤其勢必影響字頻詞頻統計之精確度，更有可能使高頻字詞變成非高頻字詞，因此加以統一應該是較為妥適的處理方法。本研究建議的作法是將生語料保持原樣建檔登錄，再將這種情形予以統一。至於統一的原則，則容後敘明。

肆、本研究的分詞原則

上述討論的分詞原則皆為中央研究院及教育部國語推行委員會的分詞標準，也可以瞭解國語的詞頻研究在這兩個單位的努力之下，已有豐碩的研究成果。相對的，客語的語詞語法都在啟蒙階段，所以本研究進行有關客語的詞頻統計，必須訂定明確的客語分詞原則。

本文客語分詞原則的擬定，是基於國語、客語皆為漢語系統的一環，所以除了參考前述國語詞頻統計的分詞原則外，還根據詞彙學的原理，以及客語的語言特性和目前所蒐集到的客語語料等，擬訂本研究的客語分詞原則。茲陳述說明於下：

詞的定義：語句中具有完整概念且能獨立自由運用的基本單位為詞。

(一)、合併原則

1、詞意無法由組合成分直接相加而得到的字串應合為一分詞單位

（1）字串組合後的意思已改變或不限原來的意義。

詞例：「食晝」、「做得」、「恁仔細」、「三月」、「五

樓」。

說明：「食晝」是「吃午飯」的意思，詞意並非由「食」和「晝」相加而得。「做得」是「可以」的意思，詞意並非由「做」和「得」兩個字直接相加而得。「恁仔細」是「謝謝」的意思，字串組合後已不限原來的意義。「三月」特指三月分，不是三個月。「五樓」不是五層樓，故應予以合併。其它如「三叔」、「滿妹」（最小的女兒）、「食清早」（吃早飯）都為分詞單位，以符合客家語使用的認知與習慣。

（2）合併結構。

詞例：「上下課」、「國中小」、「中山南北路」、「客委會」。

說明：「上下課」並非「上」加「下課」，而是上課加下課。「國中小」也非「國中」加「小」，而是指國中和國小。「中山南北路」不是「中山南」加「北路」，而是「中山南路」加「中山北路」。「客委會」則是「行政院客家事務委員會」合併的簡稱。

2、詞類無法由組合成分直接得到，應合為一分詞單位。

詞例：「好看」、「好食」、「好搞」。

說明：「好看」為「副詞＋動詞→形容詞」，「看」原是及物動詞，不能被程度副詞修飾，與詞素「好」合併後，詞類改變為形容詞，且能被程度副詞（如盡、蓋）修飾，如「生著蓋好看」（長得很好看）、「電影盡好看」（電影很好看）、「面怕粄盡好食」（粄條很好吃）、「跳索仔盡好搞」（跳繩很好玩）。

3、附著詞素盡量與前後詞合為一分詞單位。

如前後詞綴，前綴如「老、阿」，詞例：「老虎」、「老妹」、「阿姆」、「阿德」。後綴如「頭、嫲、公、仔」，詞例：「鑊

頭」、「膝頭」、「笠嫲」、「舌嫲」、「碗公」、「蝦公」、「貓仔」、「細人仔」。

4、偏正式結構的雙音節字串盡量視為一分詞單位。

詞例：紅豆、甜粄、新娘、舊曆、豬肉、魚塘。

5、並列式結構的雙音節字串盡量視為一分詞單位。

詞例：大小、多少、圓扁、左右、鬧熱、歡喜。

6、主謂式結構的雙音節字串盡量視為一分詞單位。

詞例：地動、冬至、 天光、肚渴、肚饑。

7、動賓式結構的雙音節字串盡量視為一分詞單位。

詞例：關門、捲手、寫字、掩水、食茶。

說明：動賓式結構的雙音節字串視為一分詞單位，有些待討論的空間。主要是動賓式結構的雙音節詞，在做名詞用時兩個詞素間的結合比較緊密，通常塞不進其他的字，如「主席」、「司機」。做動詞用時，兩個詞素間往往可以塞進旁的字，[21]如「關門」一詞，可以插入別的詞成為「關上門」、「關一下門」。雖然如此，不論是在中央研究院或教育部國語會的處理方式都是從合，中央研究院的原則是「使用頻率高或共現率高的字串盡量視為一個分詞單位」，教育部國語會的原則是「使用頻率高或連用程度強的字串應視為一分詞單位」。但是在客家語中，是否連用程度強，由於缺少足夠的詞頻統計，所以只能主觀的認定，難以客觀的證明。其他的例詞也有同樣的情形，這裡本文主張從寬認定，惟音節數限定在雙音節。因此如為「食飯」、「食茶」是詞，採合併方式；「食水果」就是詞組，須切分成「食 水果」。

8、動補式結構的雙音節字串盡量視為一分詞單位。

---

[21] 2004，竺家寧，《漢語詞彙學》。台北：五南出版公司，頁 74。

詞例：食飽 、擺好、做忒、養成、捨得、捉著。

說明：動補結構的補語成分，必須是單音節，與前面的動詞性詞素合併歸入動詞。如果是雙音節以上，則是詞組。如「看清楚」、「企 起來」。

9、專有名詞（含套裝合併形式）、固定語（成語、慣用語）和外來音譯語詞視為一個分詞單位。

（1）專有名詞（含套裝合併形式）

詞例：十八尖山、桃園縣、家扶中心、薑絲炒豬腸、米篩目、西遊記、新竹縣政府、迪士尼樂園、雪霸國家公園。

說明：包含人名、地名、國名、公司行號、機關學校、行政區域、產品名、書名、電影名稱等專有名詞特有所指，應視為一分詞單位，不予切分。但這裡不含一般稱謂，如「陳 太太」、「劉 先生」，應予以切分。至於所謂套裝合併形式，如「新竹縣政府」、「雪霸國家公園」。

（2）固定語（成語、慣用語）

詞例：點石成金、節節高升、桃園三結義、誤入歧途、打嘴鼓、打門敘、缺牙耙、料天穿、矮嬤車、打溜崎、平安順序、恭喜發財、汗流脈落、咬薑啜醋、面紅濟借。

說明：結構為四個音節的固定語，形式類似成語，但不具文獻典源，又無多層表義效果，因其屬慣用語的一類，故不予以切分。

（3）外來音譯語詞

詞例：梵谷、福爾摩沙、愛迪生、萊特兄弟、阿彌陀佛。

10、有固定意義的重疊詞視為一分詞單位。

（1）詞素重疊

AA 式。詞例：憨憨、遽遽、慢慢、彎彎、直直。

AAB 式。詞例：樣樣會、低低飛、叭叭跌、呀呀叫、弄弄行。

ABB 式。詞例：嘴擘擘、頭纍纍、看現現、分淨淨、圓滾滾、白雪雪。

AABB 式。詞例：鬧鬧熱熱、彎彎斡斡、上上下下、四四方方、辛辛苦苦。

（2）鑲嵌重疊

詞例：行來行去、飛上飛下、翻來翻去、晃來晃去、彎來彎去。

（3）全詞重疊

詞例：畏羞畏羞、烏金烏金、頭擺頭擺、本成本成。

說明：詞素或詞的重疊有特定的語法功能而不是修辭上的反覆時，因其通常具有擴大語義的效果，故視為一分詞單位。

(二)、切分原則

1、諺語、俗語、歇後語等常連用但內部結構複雜者儘量予以切分。

如：「一 樣 米 降 百 樣 人」

「盲 食 五月 粽 襖婆 毋 人 甕」

2、所有由數字組合成之定語，不論以國字或阿拉伯數字表達均視為一分詞單位。

（1）有關時間的詞目：

如：「一八二八年四月二十二日」，切分為「一八二八年 四月 二十二日」。

（2）數字定語與量詞的組合：

如：「一撮」、「一隻」、「一公斤」、「3505 公尺」，切分為「一 撮」、「一 隻」、「一 公斤」、「3505 公尺」

（3）地址予以切分：

如：「桃園縣新屋鄉過嶺里 15 號」切分成「桃園縣 新屋鄉 過嶺里 15 號」

3、附著於詞根表時態或特定語法功能之詞素，予以切分。

如：「看 咧」－表動作之完成態。「去 咧」、「來到 咧」、「吃

咧」、「咬 等」－表動作之持續態。「拿 等」、「穿 等」、「看 等」、「去 過」－表過去。「看 過」、「食 過」、「歇 過」、「參觀 過」

說明：相當於國語「了、著、過」動態助詞的時態標記，客家語中有「咧、等、過、 吔」 等字。

4、某些特定詞素做為起始字時，因可任意組合，視為詞組予以切分。

(1) 盡：「盡 好」、「盡 大」、「盡 認真」

(2) 蓋：「蓋 多」、「蓋 靚」、「蓋 辛苦」

(3) 當：「當 好」、「當 會」、「當 暢」

(4) 異：「異 少」、「異 高」、「異 大」

(5) 恁：「恁 多」、「恁 少」、「恁 會」

(6) 正：「正 知」、「正 係」、「正 會」

(7) 本：「本 人」「本 班 」

(8) 可：「可 用」、「可 吃」

(9) 一等：「一等 會」、「一等 快樂」、「一等 靚」

5、方位詞與前後詞切分。

(1) 單音節方位詞：項、頂、肚、唇、背、底。

(2) 雙音節方位詞：底背、東片、西片、頂項、外背。

如：「街 項」、「樹 頂」、「屋 肚」、「路 唇」、「屋 背」、「桌 底」、「肚屎 底背」、「屋 頂項」、「在 外背 打球」、「學校 東片」

但若是像「海唇」、「唇項」、「手邊」等已具有引伸義者，則不予切分。

6、「之、得 、个」等結構助詞與前後詞切分。

如： 「不孝 之 人」

「上課 个 地方」

「時間 過 得 還 遽」

7、介詞與前後詞切分。

介詞為位於名詞或名詞性詞組之前，合起來表示方向、對象、時間、處所等的虛詞。

常用的介詞有： 在、分、到、也、像、佇、對、從、比、適、畀、向、將、共、讓、從來、在於。

8、英文、日文或中英合併詞予以保留

例詞：e-mail

(三)、補充說明

1、動賓中插切分

動賓結構的賓語成分必須是單音節，如「食飯」、「食茶」，如果是雙音節以上，則是詞組，要切分，如「食 水果」、「食 豆奶」。另外若動賓結構有相當於國語「了、著、過」動態助詞的時態標記，客家語中有「咧、等、過、吔」 等字的，則賓語即使是單音節，仍應切分。如「食 等 飯」、「啉 吔 酒」「掛 忒 紙」。

2、動補中插切分

動補結構的補語成分，必須是單音節，與前面的動詞性詞素合併歸入動詞，如「捨得」、「食飽」；如果是雙音節以上，則是詞組，要切分，如「看 清楚」、「企 起來」。 但是動補結構中如果中插結構助詞「得」，或否定副詞「毋」，則補語即使仍是單音節，仍應切分。如「食 得 飽」 、「食 毋 飽」、「聽 得 識」、「聽 毋 識」、「買 得 到」、「買 毋 到」。

3、否定語料的處理

由於客語對應國語的否定詞主要是「毋」（不）及「無」（沒、沒有）兩個字，如「毋係」、「毋要」、「毋知」、「毋曉得」、「毋中意」…，「無錢」、「無人」、「無影」、「無相關」、「無要緊」…本研究衡諸各種情況後發現，有些是已然構成單一副詞者，如「無論」、「毋過」、「毋使」等，應合併為「詞」；有些則是去掉否定

詞素後原來的詞變成沒有意義，如「無採」（可惜之意）、「無膽」（膽小之意）、「無錢」（貧窮之意），也應合併為「詞」。為了分詞的簡便不易出現不一致之情形，採取的方式是：觀察去除否定詞素後所接之詞，若為雙音節詞時則視為詞組予以切分，如「毋 曉得」、「毋 中意」、「無 相關」、「無 要緊」。去除否定詞素後所接之詞若為單音節時則不分。如「毋係」、「毋要」、「毋知」、「無錢」、「無人」、「無影」，都視為一詞不予切分。

4、通用字的處理

客語由於長期以來並未發展出標準化、規範性的文字化文本，部份客語有音無字，撰寫者多以自行造字或假借方式處理，各自為政的結果，便形成同音同義卻不同字的情形。有關文字書寫的異用情形，為便於研究軟體設計及詞頻統計上的精確度，本文主張加以統一。

統一的原則，第一，依照教育部公布的「臺灣客家語書寫推薦用字」作為標準予以統一，如「聊」、「料」、「寮」、「燎」、「嫽」等統一成「寮」。又如「毋」、「m」、「不」、「冇」、「唔」、「莫」、「無」等混用情形，依文意內容，統一成「毋」。第二，不在教育部公布推薦用字之列的字，則依客委會初、中高級檢定考試的用字為準，如「唇」、「脣」、「漘」、「滆」等不同寫法，統一成「唇」;「笠母」、「笠麻」、「笠蔴」、「笠嫲」等統一成「笠嫲」。若兩者皆無，則以教育部國語會的「台灣客家語常用詞典」蒐尋結果予以統一，如「一個」、「這個」的「個」，又如「食忒」、「放忒」、「除忒」的「忒」等。其餘則依電腦輸入能處理、用字普遍通用及請教學者專家等原則來選定處理。

伍、結論

運用科學客觀的方法建立一個客語語料庫，並進行字頻、詞頻的統計，以得出客語常用的高頻字、高頻詞，提供客語教材撰寫或編輯

詞目之參考，是傳承客家語言文化，極重要也亟待處理的重大問題。

建立一個語料庫，並進行字頻、詞頻的統計研究，在客語的處理過程中，最棘手也是最重要的工作就是如何分詞的問題。本論文主要的研究目的便是探討客語的分詞原則。由於漢語中每個句子都是方塊漢字寫出來的，「字」、「詞」、「詞組」的界線並不明顯，所以本論文先從客語中「字」、「詞」、「詞組」的區分進行釐清探討；接著進一步探究並建立客語的分詞原則。此一部分是參考有嚴謹學術理論基礎的中央研究院及教育部國語會的漢語分詞原則，經比較其異同後，再考慮客語本身的特性，最後加以修訂成適合客語的分詞原則。

本文所確立的客語分詞原則，主要將「詞」定義為：「語句中具有完整概念且能獨立自由運用的基本單位為詞。」並將詞的切分歸納成合併原則、切分原則及補充說明三大類，其中合併原則共有十條細則；切分原則共八條；補充說明共四條，做為切分的依歸。希望經由本文所建立的客語分詞原則，能為建置客語平衡語料庫略盡棉薄之力，亦請各位方家不吝指教。

參考書目

王力，1996，〈詞和仂語的界線問題〉，《中國語文》，北京：北京大學中國語言文學系。

王建新，2005，《計算機語料庫的建設與應用》，北京：清華大學出版社。

北京語言學院，1990，《現代漢語詞頻率詞典》，北京：北京語言學院。

台北市客家事務委員會，2004，《現代客語詞彙彙編》，台北：台北市政府。

吳敏而，1998 年，《國民小學兒童常用字詞詞彙資料庫之建立與初步分析（Ⅰ）》（國科會研究成果報告），台北：臺灣省國民學校教師研習會研究室。

_____，1998，《國民小學兒童常用字詞詞彙資料庫之建立與初步分析（Ⅱ）》（國科會研究報告），台北：臺灣省國民學校教師研習會研究室。

_____，1998，《國民小學兒童常用字詞詞彙資料庫之建立與初步分析(Ⅲ)》（國科會研究成果報告）。台北：臺灣省國民學校教師研習會研究室。

呂叔湘、胡繩等，1996，《現代漢語詞典》，北京：商務印書館。

呂叔湘，1963，〈現代漢語單雙音節問題初探〉，收於《中國語文》第 1 期，北京：北京大學中國語言文學系。。

竺家寧，2004，《漢語詞彙學》，台北：五南出版公司，頁 74。

國立政治大學教育學系，1982，《兒童常用詞彙研究》，台北：政治大學。

張雁雯 2008，《台灣四縣客家話構詞研究》，國立台灣大學中國文學研究所碩士論文。

張壽康，1957，〈略論漢語構詞法〉，收於《中國語文》，北京：北京大學中國語言文學系。

教育部國語推行委員會，1998，《常用語詞調查報告書》，台北市：教育部。

陳惠玉，2003，《臺灣臺中市何仁里臺語常用詞彙調查與詞頻之初步研究──並與台中縣新里里做比較研究》，新竹教育大學台灣語言與語文教育研究所碩士論文。

傅永和，1999，《中文信息處理》附錄 5，廣州：廣東教育出版社。

曾榮汾，1994，〈字頻統計法及學術應用〉，《警學叢刊》第 25 卷第 2 期。

_____，1977，〈字頻統計法的實例－國小常用字彙統計析述〉，《警學叢刊》第 27 期。

湯廷池，1982，〈國語詞彙學導論：詞彙結構與構詞規律〉，收於《教學與研究》第 4 期。

_____，1985，《國語語法研究論集》，台北：台灣學生書局。

程祥徽、田小琳，1992，《現代漢語》，台北：書林出版有限公司。

黃宣範，1988，〈台灣話構詞論〉，收於鄭良偉、黃宣範 編《現代台灣話研究論集》。台北：文鶴出版有限公司。

葉明德，1997，〈華語文常用詞彙頻率等級統整研究〉，《華文世界》第 85 期。

葉蜚聲、徐通鏘，1993，《語言學綱要》，台北：書林出版社。

趙元任著，丁邦新譯，1980，《中國話的文法》，香港：香港中文大學出版社。

劉秀珍，2006，《客家語教科書常用詞彙與詞頻之初步研究－以高雄市版為例》，國立高雄師範大學台灣語文及教學研究所碩士論文。

劉杰，1990，〈漢語超高頻詞分類統計與分析〉，收於胡盛侖主編：《語言學與漢語教學》，北京：北京語言學院出版社。

潘文國、葉步青、韓洋，1993，《漢語的構詞法研究》，台北：台灣學生書局。

賴惠玲，2008，〈客語語法研究議題的開發：以語料庫為本〉，收於《96 年補助大學校院暨獎助客家學術研究計畫成果發表會論文集》，台北：行政院客家委員會，頁 153-164。

謝杰雄，2006，《語料庫的建置與台灣客家語 VP 研究》，新竹教育大學台灣語文與文學教育研究所碩士論文。

鍾屏蘭，2011，〈從語料庫的開發探討客語教材的編輯與出版〉，收於《屏東教育大學學報—教育類》第 36 期。

鍾榮富，2004，《台灣客家話導論》，台北：五南圖書公司。

_____，2001，《福爾摩沙的烙印—台灣客家話導論（下冊）》，台北：行政院文化建設委員會。

_____，2001，《福爾摩沙的烙印—台灣客家話導論（上冊）》，台北：行政院文化建設委員會。

羅肇錦，1988，《台灣的客家話》，台北：台原出版社。

_____，2000，《台灣客家族群史〔語言篇〕》，南投：台灣省文獻委員會編印。

_____，1998，《客家化字詞與音義析論》，台北：紅葉文化事業公司。

_____，1998，《客語語法》，台北：學生書局。

網路資料

〈中央研究院/現代漢語平衡語料庫〉，網址：http://dbo.sinica.edu.tw/SinicaCorpus/。

〈台灣大學客家社/客語小辭典〉，網址：http://www2.ee.ntu.edu.tw/~r8921044/hakdict/hakquery.htm。

〈國科會數位博物館先導計畫/搜文解字〉，網址：http://words.sinica.edu.tw/。

〈教育部國語推行委員會/八十七年口語語料調查報告書〉，網址：http://www.edu.tw/files/site_content/M0001/87oral/index.htm。

〈教育部國語推行委員會/八十七年常用語詞調查報告書〉，網址：http://www.edu.tw/files/site_content/M0001/87news/index.htm。

〈教育部國語推行委員會/臺灣客家語常用詞辭典（試用版）〉，網址：http://hakka.dict.edu.tw/。

鍾屏蘭

國立屏東教育大學中國語文學系

屏東市民生路 4-18 號

pin@mail.npue.edu.tw