詞　庫　小　組

技 術 報 告 95-02/98-04

Technical Report no. 95-02/98-04

中央研究院平衡語料庫的內容與說明

(修訂版)

出 版 處：台 北 ， 南 港

中央研究院資訊科學研究所
中 文 詞 知 識 庫 小 組

中央研究院平衡語料庫的內容與說明

目　錄

圖表目錄

(中研院平衡語料庫3.0版網址 **http://www.sinica.edu.tw/SinicaCorpus**)

序

「中央研究院平衡語料庫」簡稱為「中研院平衡語料庫」是第一個帶有詞類標記的漢語平衡語料庫。這個平衡語料庫雖比第一個英語平衡語料庫足足晚了卅年，卻仍領先世界上絕大多數的語言。這卅年的時間也正好提供了許多資訊科技進行的空間，特別是中文資訊從無到有日趨成熟，以及語料庫與自然語言處理上學理與技術的突飛猛進。我們相信「中研院平衡語料庫」不但可以滿足提供語料作為觀察、驗證、測試的基本功能；更可配合各種搜尋檢索，及統計工具，協助使用者找到語言規律與自然語言處理的良好方法。

當然，作為首次測試中的語料庫，「中研院平衡語料庫」並非完美無缺。特別是詞類標記原則，及平衡語料選取原則均是理論上尚無定論且可深入研究的問題。因此也可以由學界繼續討論中得到助益而改進的。我們期盼這顆小石頭所牽動的小漣漪，可以更進一步推動語言學者的討論，對漢語詞類的分類原則得到更完善的理論與實際分析。在標記的過程中，詞庫小組力求的最高原則是一致性。我們難免犯錯，但幾次校定，希望錯誤率已降到最低。至於標記原則上或許與某學派或某特定分析研究不相容，但是詞類標記（tag）可以看作是一種方便檢索分類的標籤，只要所指的類大致合理，便能找到有意義的語言分佈現象。

我們要感謝所有提供語料的單位與個人，在技術報告中已有說明，在此不一一列舉。我們更要感謝所有實際進行標記的工作伙伴們的辛勞。他們都有良好的語言學素養，而願意進行此勞心勞力又極煩瑣的工作，實是為以後的研究者造福。張莉萍負責詞類標記集的規劃及詞庫工作之總協調；許蕙麗負責語料庫搜集及標記工作之協調，實際進行標記的有洪偉美、張麗麗、魏文真、藍素禎、周芸青、萬怡君、黃惠婷、高照明、王寧馨、詹曉蕙、漆聯成、劉淑梅、陳鳳儀等人。程式方面，劉興寰、陳大業設計並維護自動分詞標記系統，謝明華設計了搜尋程式及介面。我們更要感謝中研院「中文資訊跨所研究群」及國科會「漢語平衡語料庫」兩個計劃經費的支持。

「中研院平衡語料庫」是中文標記平衡語料庫的開端，我們希望詞庫小組的努力可以促進更多的研究發展，更希望在使用中所發現尚可改進之處，請毫無保留賜告，以使以後的版本更臻完善。

黃居仁，陳克健
謹誌於中研院詞庫小組
一九九五年八月十二日

修訂版說明

「研究院語料庫」（Sinica Corpus） 3.0 版的五百萬詞語料庫於 1997 年 10 月完成，並開放上線使用（**http://www.sinica.edu.tw/SinicaCorpus**）。這表示詞庫小組七年來構建語料庫的工作有了一個完整的段落；漢語語言學研究也有一個大規模標記語料庫可以使用。

因為語料庫內容的更新，說明也需要隨之更動。本次修訂主要更新的內容如下：

1. 語料庫內容分佈統計（表一~表五）
2. 第三章之分詞標準的更新：詞庫受中央標準局委託之「中文資訊處理分詞規範」，幾經修訂，已成國家標準草案，語料庫之分詞及本書說明均根據最新的標準草案更正。
3. 「緣起及內容」（第二頁）部分更正。
4. 增加+prop 特徵（30 頁及 32 頁）。

修訂過程特別感謝語料庫全體同仁標記校對的辛勞，特別是陳鳳儀實際負責大部分修訂工作。

在修訂語料庫說明的同時，我們也出版了一部根據研究院語料庫統計出的詞頻詞典。這是第一部有詞類標記的中文詞頻詞典，希望可以提供計量研究的重要基本資料。

黃居仁、陳克健

1998.8.19 于 中研院

一、前言

「中央研究院平衡語料庫」簡稱「中研院平衡語料庫」（Sinica Corpus），是世界上第一個有完整詞類標記的漢語平衡語料庫。由於加詞類標記的漢語語料庫是史無前例的嚐試，第一步先以較小規模（但仍大於較早英語語料庫的一百萬詞規模），於1994年公開提供給國內外學術研究使用；以期在使用過程中得到回饋，在完成目標規模前可以做必要的修正。1997年開放的研究院語料庫3.0版已達到五百萬目詞的預計規模。

1.1 建立平衡語料庫的動機

語料庫為本（corpus-based）的研究是近年來語言學及計算語言研究的一個重要發展〔Svartvik 1992, Church and Mercer 1993, 陳克健 1994, 黃居仁 1995〕，其影響更遠及文學及社會學的計算研究。在語言研究的前提下，語料庫為理論語言學或自然語言處理研究所擔負的功能是在無窮衍生的語言事實中抽出一個具代表性的樣本來。這個樣本不能太大，否則失去了抽樣的意義與優點。又不能太小，否則無法提供足夠的訊息，也無法提供大量素材作統計研究或作測試語料。因此語料庫構建的第一個大問題是如何在有限的語料中代表複雜的當代語言全貌。舉世聞名的布朗語料庫（Brown Corpus）〔 Krucera and Francis 1967〕在一九六０年代中期構建時即是以解決這個問題為目標。他們的想法很簡單，即是一個具代表性的平衡語料庫必須包含各種不同的文體。他們根據抽樣調查決定了一個他們認為英文平衡語料庫應有的分布，再根據此一分布收集了百萬詞的語料，並加上詞類標記，輸入電腦。建構成了第一個機讀語料庫，也是第一個平衡語料庫。儘管由現在理論及技術的水準看來，布朗的資料及平衡方式略嫌粗糙，可是這個語料庫一直是（英語）平衡語料庫的標準，甚至到了八十年代新構建的英語平衡語料庫如LOB（Lancaster-Oslo/Bergen，英國英文）及London-Lund（英語口語），都還遵循布朗語料庫的架構。足見這種平衡語料庫在各種語言學研究上有其不可取代的價值。可惜的是在國際間我們很難得看到其他語言的平衡語料庫，更不用提中文平衡語料庫了。

平衡語料庫中最重要的訊息，也是關鍵性的特色，便是每個詞上的詞類標記。簡單說來，若把語料庫看成是幾萬個詞的排列組合，則其規律性及相關訊息極複雜而不易掌握，但若把其內部關係化簡成（幾十個到上百個）詞類間的關係，則其規律性將較明顯易掌握，統計上也較易處理。當然，每個詞上有意義的標記（tag），並不一定是詞類，也可以是語義、語音、筆劃等。可是只有詞類可以算是（所有語言）的基本架構單位，是語言學家公認建構語法的基礎，也是不論對語言從事何種研究都可能用得到的訊息。因此為增加平衡語料庫的活用性（versatility）及其所承載的訊息，詞類標記是必要的。近五年來中文語料庫的搜集構建雖然已經開始〔黃居仁 1995〕，進行詞類標記者則仍尚未有。

1.2 中研院平衡語料庫的源起

中央研究院詞知識庫小組，自一九九０年前後便開始致力於中文語料庫的收集〔Huang & Chen 1992〕，截至目前止已收集有近二千萬字之現代漢語語料及超過五百萬字之古代漢語語料〔Huang 1994〕。由於有了處理中文語料庫的經驗，及大量處理電子詞庫中詞條的經驗〔陳克健等 1991, Chen 1994〕，我們覺得有足夠的實質與人力條件來進行耗時費力的漢語平衡語料庫建構。在一九九四年分別得到了中央研究院「中文資訊」跨所研究群之專案計劃及國科會計劃補助，乃開始著手進行。為兼顧理想與實用性，初步目標定為兩百萬詞，為傳統小規模平衡語料庫之兩倍，到今日的五百萬詞，這個最終目標則接近目前計算語言學常用之規模。

平衡語料之抽取以自中央研究院詞庫小組現有之語料中取得為優先，但也同時透過不同管道取得不同文體、內容之語料。以下依來源之不同種類大致列舉，並向提供語料之單位致謝。

(一) 交換取得之語料：此項包括經由合作計劃交換取得的，如中國時報，洪建全基金會，師大國語中心。或是由計算語言學會內部之語料作共同體（consortium）間交換語料而得，如由致遠科技及台大取得。

(二) 直接向版權所有單位取得：慷慨提供我們版權語料做學術研究用的有：天下雜誌社，國語日報社，資訊傳真雜誌社，「女人女人」製作單位，「伴我成長」製作單位，「我們一家都是人」製作單位以及許多中研院內的單位等。另有舊金山州立大學畢永峨，清大郭賽華，交大劉美君，輔大楊承淑等多位教授提供他們轉寫（transcribe）的口語資料。

(三) 由公共區域取得的公共資料：大部份由電子佈告欄（BBS）或蕃薯藤等萬維網中取得。

1.3 中研院平衡語料庫的設計理念

研究院語料庫因為中文的特性，也因為我們觀察語料的經驗及研究語料庫語言學的結果，有以下幾個重要的設計理念（Design Features），這些設計理念中有不少是我們所獨創的，希望能使得研究院語料庫成為科學研究漢語不可或缺的利器與基本材料。

(一) 遵循計算語言學學會的分詞標準

分詞（或稱斷詞）是中文自然語言處理的先決條件，但因中文詞的分界在實際書寫上不標明，在理論上亦有爭議；故一直很難標準化。目前國內有中華民國計算語言學學會受中央標準局委辦研擬「中文資訊處理分詞規範」〔黃居仁等 1997〕，並已完成國家標準草案。我們依此標準分詞不但可以有助於資源分享，對語料庫分詞結果之回饋也可成為爾後修定國家標準草案的依據。

(二) 裁文是以文章（text）的自然段落為準，而非以文章長度為準

布朗語料庫的設計特色之一，是為了求數字上的平衡，故每篇文章只不多不少取兩千詞即截斷。這在使用上造成文章內容不完整，偏取各種文章之起頭部份等缺點。而且我們認為文章長短其實也是各種不同文體的一個重要特色；若裁成長短如一反而失去了這個特色。因此，我們雖然仍避免過短或過長的文章，但在選取文章後，便隨其自然段落截取。也因為如此，我們的平衡語料庫無法達到如布朗語料庫等的完整小數點。可是我們認為我們的設計理念可以取得更完整不偏頗的語言訊息內容。

(三) 語料庫多重分類原則分類

我們認為布朗語料庫傳統下以文體單一特徵來界定平衡語料庫是不足的。理由很簡單，因為影響整個語言全貌的內在因素實在太多了。為了突破這種過於單純化的線性描述；我們把所有語料都給了五個不同特徵的值：(1) 文類 (2) 文體 (3) 語式 (4) 主題 (5) 媒體。目前初步雖然仍以主題為主軸來進行語料庫的平衡。理想上是希望有了更多研究的結果之後，可以同時利用一個以上的軸來定義更完善的平衡語料庫（見Hsu and Huang 1995）。

具有五個軸的多重分類，另一個立即的好處是研究上的活用性（versatility）增加了許多。研究者可任選其中特徵的組合，定義自己的次語料庫（sub-corpora）；也可以在次語料庫間作比較研究。舉例說明研究者可以比較報紙中的論說文與學術期刊中的論說文、用詞語法有何不同；或代名詞「我」在口語會話及劇本中出現頻率的不同等。

這個多重分類原則也有利於以後平衡語料庫的更新。比如說一般認為愈正式的書面語變化愈慢，而日常的口語變化愈快。因此在有監看語料庫（monitor corpus）的前提下，我們可以隨時抽換平衡語料庫中某個符合一組特徵條件的次語料庫，以保證平衡語料庫仍忠實代表當代語言的真實現況。

1.4 中研院平衡語料庫的構建過程

語料庫語料的來源已在1.2節敘述過。實際上，要建構一個平衡帶詞類標記的語料庫，收集語料只是第一個起步工作。接下來是語料整理的工作，包括語料清潔、為語料分類、加詞類標記等等〔陳克健 1994〕。

以下就構建一個中文的帶詞類標記的平衡語料庫需要考慮的三個中心問題分三章依次說明：

第二章談平衡語料的分類與選取，如何為語料做分類，分類的標準以及各類的比例。

第三章是中文的斷詞問題，中文基本上以小句為單位，從來源處得到的資料，並無標示詞的訊息，但是切分詞的結果也直接或間接影響到詞類標記的判定及句子的分析。

第四章討論如何訂出詞類標記集，詞類標記的原則以及每一個標記所代表的涵義。

建立帶詞類標記的平衡語料庫是一個浩大的工程，但也是自然語言研究的基礎工程（infrastructure）。其效應可由現存語料庫，如布朗，LOB，London-Lund等所衍生的大量研究成果得到證明，語料庫所憑藉的是提供大量真實的語料作為研究素材，但它也忠實的反映了人們使用語言的一個事實－那就是難免要犯錯。即使是經過了將近卅年斷續的修正，學者一般估計布朗語料庫其中詞類標記尚有百分之二左右的錯誤。當然，理論分析的不同，更會導致標記上看法的分歧。但這並未損及布朗等語料庫之研究價值。

研究院語料庫的構建並不是要立下顛撲不滅的真理。相反的，我們相信在這五百萬詞的分詞與標記中，必定有些不一致處，而可能有更多的爭議。我們希望這個百分之九十幾正確的資料，可提供學界作更進一步研究發展的基礎。更希望這百分之個位數的爭議能讓我們深入的思考，因而解決中文語言學中的一些疑難問題。至少，使用者的回饋可使研究院語料庫更接近完善！

二、平衡語料庫的分類標準與選取結果

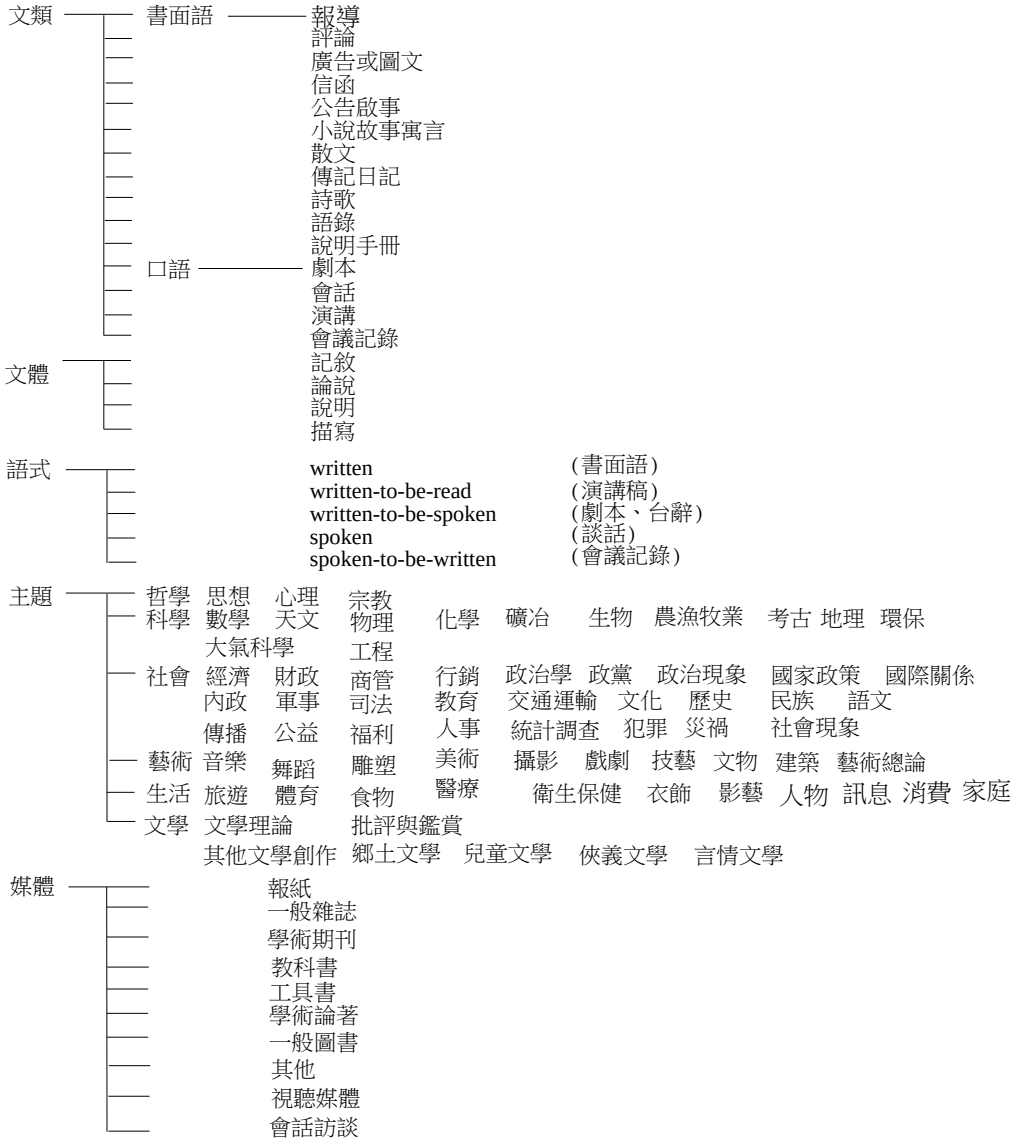
為了妥善管理以及選取平衡語料庫的內容，將收集來的語料做分類。在每篇文章前頭標示它們的文類、文體、媒體、語式、主題等，如圖一所示。這些屬性的訂定過程及其詳細內容將在2.1節說明。

```
%% 文類=散文
%% 文體=描寫
%% 語式=written
%% 主題=兒童文學
%% 媒體=教科書
%% 姓名=
%% 性別=
%% 國籍=中華民國
%% 母語=中文
%% 出版單位=國立編譯館
%% 出版地=臺灣
%% 出版日期=
%% 版次
%% 標題=星光
我永遠忘不了小的時候，
依偎在母親身邊的情景唉，
童年的回憶中，
        …
```

圖一、中研院平衡語料庫篇章標記

2.1 屬性特徵的訂定

我們參考了LANCASTER-OSLO/BERGEN (LOB)語料庫、美國布朗大學布朗語料庫、英國伯明罕大學COBUILD Project 語料庫的管理經驗，然後再參考圖書館的圖書分類，制定出一套分類中文語料的屬性特徵。這些屬性用來說明文檔的來源出處、寫作的方式、以及談論的內容。主題標示了文檔的內容，文類、文體、和語式說明了文檔呈現的型式，而出處則由媒體、作者、出版三項屬性來標示。媒體說明了文檔的出處來源。姓名、性別、國籍、母語標示了和作者有關的訊息，出版單位、出版地、出版日期、版次則記錄了和出版有關的資料同時採用了階層管理的方式在三大屬性下描述更多的屬性，如圖二所示。細類的說明將在以下各小節詳述。


文類
書面語
報導
評論
廣告或圖文
信函
公告啟事
小說故事寓言
散文
傳記日記
詩歌
語錄
說明手冊
口語
劇本
會話
演講
會議記錄
文體
記敘
論說
說明
描寫
語式
written (書面語)
written-to-be-read (演講稿)
written-to-be-spoken (劇本、台辭)
spoken (談話)
spoken-to-be-written (會議記錄)
主題
哲學 思想 心理 宗教
科學 數學 天文 物理 化學 礦冶 生物 農漁牧業 考古 地理 環保
大氣科學 工程
社會 經濟 財政 商管 行銷 政治學 政黨 政治現象 國家政策 國際關係
內政 軍事 司法 教育 交通運輸 文化 歷史 民族 語文
傳播 公益 福利 人事 統計調查 犯罪 災禍 社會現象
藝術 音樂 舞蹈 雕塑 美術 攝影 戲劇 技藝 文物 建築 藝術總論
生活 旅遊 體育 食物 醫療 衛生保健 衣飾 影藝 人物 訊息 消費 家庭
文學 文學理論 批評與鑑賞
其他文學創作 鄉土文學 兒童文學 俠義文學 言情文學
媒體
報紙
一般雜誌
學術期刊
教科書
工具書
學術論著
一般圖書
其他
視聽媒體
會話訪談

圖二、語料分類屬性階層

2.1.1 主題

主題是依照文檔內容，討論重點而定。大體上我們是參考圖書館的分類方法來定主題的屬性。以下是對主題之下各細類的說明。

哲學：

- 思想：理論、學說、主義、觀念、見解。如沙文主義、儒家思想。或是道德方面，良知、正義、貞潔、美德、婦德，如國人公德心的探討等。
- 心理：認知心理學、發展心理學、生理心理學、變態心理學、超意識心理學、人格心理學；心理衛生：諮商輔導、人生觀；價值觀；人際關係：五倫關係、人際交往藝術。如讀者投書裡的感情困擾、婆媳相處之道、如何做個成功的領導人、靈異、超能力現象等。
- 宗教：各宗教之教義、經典、組織、教派、神祇；術數：占卜、命相、紫微、風水、陰陽五行。如佛教戒律、天主教教義。

科學：

- 數學：數學總論、算術、代數、幾何、三角、應用數學。
- 天文：天文學總論、天象、太空科學、歲時、曆法。
- 物理：物理學總論、力學、熱學、光、電、磁學、現代物理。
- 化學：化學總論、固態化學、普通化學、有機、無機、定量分析、結晶學。
- 礦冶：地質、礦物、冶金、採礦、煉油。
- 生物：生命科學、植物、動物。
- 地理：地理學總論、區域地理、人文地理。環境地理學、自然資源與利用、地理探險與發現。
- 農漁牧業：農藝、森林、畜牧、漁獵。
- 考古：古生物與考古學。
- 環保：有關環保之政策、活動、理論等。
- 大氣科學：大氣圈之各種變化、氣象、氣壓、氣流等。
- 醫學：基礎醫學：醫用一般科學、生物醫學工程、生理學、病理學、醫學心理學..；臨床醫學、中西醫治療法、臨床各科治療、臨床診斷、急救、飲食療法、內、外科學、婦產科學、兒科學、腫瘤學、精神病學與神經病學、皮膚、眼耳鼻喉、口腔學、腦神經學、藥理學、中醫學。
- 工程：電子、資訊、核子、土木、機械工程等。

社會：

- 經濟：經濟制度、政策、理論、貿易問題。如中東戰事爆發對台經濟之衝擊、中日貿易逆差、中小企業外移等。
- 財政：銀行、證券、期貨、匯率、貨幣政策、賦稅、金融。如民營銀行開放、股票、公債發行、台幣匯率調高、稅捐問題、關稅調整、黃金市場震盪等。
- 商管：企業團體經營、管理狀況、經營理念、財務狀況。如台塑經營理念、鴻源財務管理不善、中興紡織赴大陸設廠等。
- 行銷：有關商品推銷的方法、市場調查、廣告、商品形象等，以賣方立場為主。如法國香檳酒來台的宣傳及進攻消費市場的策略、日本新上市聲控玩具的介紹等。
- 政治學：政治學理論、思想、國家政策改革。如統獨之爭、總統制、台灣在國際的定位問題等。
- 政黨：政黨運作組織、政治團體、次級團體、選舉。如兩黨問題、新國民黨連線、集思會、立委選舉、總統大選等。
- 政治現象：國家政策、人權運動、政治衝突、影射政壇現象。如修憲、老立委退職、解嚴、台獨、異議人士、二二八事件、國會亂象、議事、政治寓言等。
- 國際關係：兩國以上交流的各種經貿、政治、軍事、外交等關係。如蘇韓雙方進行經濟合作、
- 國家政策：台海兩岸交流相關問題、具全國性影響的國家制度。如中共犯台的可能性、開放大陸觀光、返鄉探親、兩岸聯姻問題等。
- 內政：一切地方行政，地方性的政務及決策，如地方建設、治安問題、議會決策、地方施政、地方活動（如鄰里清潔比賽）、水利事業、地方糾紛（垃圾傾倒、揭發官商勾結）、移民、外籍勞工問題等。
- 軍事：一切與軍事相關的國防、軍備、限武談判、戰事等。
- 司法：法律、訴訟程序、判決書、法律透視等。如通過兩岸關係人民條例。
- 教育：教育理論、政策、學制、學校、教師、師範教育。如自願升學方案的探討、森林小學的創設理念、才藝教育、交通安全教育。
- 交通運輸：除了一切與交通運輸有關的事業外，還包括郵政、電信、電力事業。如捷運、高鐵工程、計程車費率調漲。
- 民族文化：各國文化習俗、禮俗，描寫各民族生活習性等。如相親方式、原住民、少數民族的生活方式。
- 歷史：各類事物發展的記錄。如文化史、教育史、基督教發展史。
- 語文：語言與文字。如外語學習的方法、台語語文的探討。
- 傳播：大眾媒體、廣電、新聞學。
- 人事：人事薪資、升遷、培訓、考績、轉調等事項。組織介紹，有哪些成員。如組閣名單、何謂公務員、公務員的權利義務。
- 統計調查：一切統計及調查數據結果。如民意調查結果、人口統計數據。
- 公益：一切公眾利益事項。如急難救助、募款、義診、器官捐贈。
- 福利：有關工作福利的。如年終獎金、勞農公保及其他保險、興建勞工住宅、工時的制訂。
- 犯罪：犯罪事件。如兇殺竊盜、擄人、吸食違禁藥品等犯案。
- 災禍：天災人禍。如火災、旱災、颱風。
- 社會學：社會學理論。如法蘭克福學派之理論、社會組織架構理論。
- 社會現象：收納各類無法歸入其他主題的社會百態。如謊報事件、六合彩、非政治目的的抗議活動、自殺等。

藝術：

- 藝術總論：藝術通論，如藝術與人生、美學。
- 音樂：與音樂相關之報導與評論。如流行音樂、樂器。
- 舞蹈：與舞蹈相關之報導與評論。
- 雕塑：與雕刻、捏陶相關的。
- 美術：與繪畫、書法、版畫相關的。例如如何寫書法、如何欣賞書畫之美。
- 攝影：有關攝影的報導評論。
- 戲劇：有關戲劇、電影、舞台劇等的評論報導。如地方戲、話劇、影評。
- 技藝：民俗文化有關的雜技。如皮影戲、扯鈴、麵包花、中國結。
- 文物：陶、磁、銅器、清朝文物等。
- 建築：建築之美、與建築有關的報導評論。

生活：

- 旅遊：旅遊、遊記、休閒、娛樂、風景。遊記，如威尼斯之旅、如何度假。
- 體育：體壇消息、職棒、奧運。
- 食物：食品烹調、食物營養、食譜、健康食品、美食介紹。
- 衛生保健：公共衛生、環境衛生、食品衛生、一般衛生保健法、健康常識、醫藥常識、疾病預防、防疫接種，健康教室的設立。
- 衣飾：衣服、飾物、服飾。
- 影藝：藝壇消息、藝人生活、影片花絮。
- 人物：對於人物的簡介、評論、專訪，如總統下鄉巡訪、清潔工的一天、我的母親。
- 訊息：一般訊息，各種活動舉辦的消息及內容簡介，如停水、停電、藝術展、某軟體已安裝了可供大家使用、某人來訪、會議通告。
- 消費：以買方為出發的報導或評論。如消費者權益、買電器應有的認識。
- 家庭：無法歸入其他類的雜類。如親子交流、住家設計、瓦斯安全、如何省電、婚姻的各種問

文學：

- 文學通論：文學理論、比較文學及其他。如文學的特質、文學與人生、文藝美學。
- 批評與鑑賞：文學批評、賞析、與寫作。如書評。
- 鄉土文學：取材自鄉土、使用鄉土語言的創作。
- 兒童文學：特地為兒童創作或為兒童寫的作品。
- 俠義文學：武俠小說、推理小說。
- 言情文學：有關愛情的作品。
- 其他文學創作：無法歸入其他類的文學創作。

2.1.2 文類

文類是說明文檔的呈現方式，可分為報導、評論、廣告圖文、信函、公告啟事、小說故事寓言、散文、傳記日記、詩歌、語錄、說明手冊、劇本、會話、演講、會議記錄。其中的語錄都是來自報刊邊緣的小語錄，數量很少。信函約有三類來源，報章雜誌的讀者投書，教科書裡的書信範例，以及電子佈告版裡的書信往來。劇本都是來自小學課本，都是記敘文，主題為兒童文學，語式為written-to-be-spoken。演講包括三民主義演講稿，以及一些集成書，或刊於期刊中的演講。

2.1.3 媒體

媒體是根據資料來源分類。大體上書面語和口語會有不同的來源，書面語的來源大致可分期刊、圖書、書信、視聽媒體、會議、其他；視聽媒體包括了「女人女人、我們一家都是人」電視節目的台詞，還有一些電子佈告版裡的文章，電子佈告版對大量語料庫的建立極有幫助，我們不必費時取得版權，也沒有修改亂碼取得造字檔的問題，可以在其中收集到多樣化的文檔。如果電子佈告版裡的文章標明了原來出處，我們就依照其出處歸類到其他媒體來源。其他就是用來標示不能歸類於任何一種媒體的文檔。我們的期刊類分為報紙、學術期刊、一般雜誌；圖書分為教科書、工具書、學術論著、和一般圖書。報紙包括中國時報、自由時報、兒童日報、中央研究院計算中心通訊等。一般雜誌包括天下雜誌、光華雜誌、海天遊蹤、翰林雜誌、世界電影雜誌、空間雜誌、華夏文摘、資訊傳真等；學術期刊包括生醫簡訊、民族所集刊。教科書有小學國語課本和師大國語中心提供的國語實用會話；工具書則收了詞庫小組的技術報告。學術論著是我們收集到的一些論文。一般圖書包括了三民主義演講稿、洪建全基金會的大眾心理類書籍、時報出版的兩本書等。口語語料來自大陸民運人士訪談，及大陸留美學生的日常對話。

2.1.4 文體

文體是文檔的寫作方式，分為記敘、論說、說明、描寫。記敘是將人、物的狀態、性質、動作、變化等記錄下來，一般記事敘述、訊息報導的文章都屬於記敘文。在我們所收集的文檔裡，記敘是最常用的寫作方法。論說是提出自己的主張、意見、以得到他人認同、說服他人，一般評論的文章都是論說文。說明文的功用主要是分析事物的結構、現象、道理，使人獲得某方面的知識和道理。所以僅以客觀的文字說明事物的功能性質、形狀等的文字屬於說明文。描寫是對人、物、事或景等做深刻的描繪，可能運用到比喻、修飾、排比、象徵等多種描寫技巧，來突出他的性質、特點，加深他人的印象。我們的描寫文包括抒發心靈感觸的抒情文章，如描述景物的遊記也多半是散文。

2.1.5 語式

語式標示文檔的呈現方式，是以書面語或口語的方式表達就大有不同。我們把語式分為 written 、 written-to-be-read 、 written-to-be-spoken 、 spoken 、 和 spoken-to-be-written 。 written即一般的書面語，也是我們語料庫裡收集最多的文檔；written-to-be-read是指演講稿之類，寫下了讓人唸出來的，因為是經過審慎思考的文稿，所以和一般口語的鬆散大不相同；written-to-be-spoken是指劇本、台辭等，寫了讓人在模擬現實會話情境下講的，因為是以事先預想演練過的方式表現出來，所以還是和實際的口語不盡相同；spoken即指一般的口語談話，這類資料的整理較不容易，所以在目前的語料庫裡尚佔少數。spoken-to-be-

written 是指會議記錄之類的文檔，由於還有修改整理的機會，可能去除了許多冗雜的部份，因此，值得另分一類，以和真正的口語、書面語區別。

2.2 語料的選取與分佈比例

目前，我們以主題為準，訂出平衡語料庫的內容比例為：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，根據此參考值為基準選取語料。結果在兩百萬的語料中，各類主題實際分佈狀況，如表一所示。為了研究主題的分佈和文類、媒體、文體、語式彼此的相關性，我們也統計後四者各小類在五百萬語料庫中所佔的百分比，請見表二至表五。

表一、中央研究院平衡語料庫 3.0 版各主題分佈比例（單位：萬）

主題	哲學	科學	社會	藝術	生活	文學	總計
平衡語料庫百分比	10%	10%	35%	5%	20%	20%	100%
現有 CORPUS 字數總計	68.53	10.24	276.13	73.22	141.20	127.85	789.27
現有 CORPUS 詞數總計	45.17	67.50	182.03	48.27	93.08	84.28	520.28
實際百分比%	8.68	12.97	34.99	9.28	17.89	16.20	100

表二、中央研究院平衡語料庫 3.0 版各文類分佈比例（單位：萬）

文類	報導	評論	廣告圖文	信函	公告啟示	小說寓言故事	散文	傳記日記	詩歌	語錄	說明手冊	劇本	演講	會話	會議記錄
CORPUS 字數	443.94	78.97	4.68	10.17	5.79	79.85	66.93	3.94	2.31	0.23	15.98	0.43	64.61	10.57	0.84
CORPUS 詞數	292.64	52.06	3.08	6.71	3.82	52.64	44.12	2.60	1.52	0.15	10.54	0.29	42.60	6.97	0.56
百分比%	56.25	10.01	0.59	1.29	0.73	10.12	8.48	0.50	0.29	0.03	2.03	0.05	8.19	1.34	0.11

表三、中央研究院平衡語料庫 3.0 版各媒體分佈比例（單位：萬）

媒體	報紙	一般雜誌	期刊	教科書	工具書	學術論文	一般圖書	視聽媒體	會話訪談	演說	其他
CORPUS 字數	246.89	230.28	5.49	32.23	1.06	10.71	66.70	180.20	12.90	2.00	0.81
CORPUS 詞數	162.57	151.80	3.62	21.25	0.70	7.06	43.96	118.80	8.50	1.32	0.53
百分比%	31.28	29.18	0.70	4.08	0.13	1.36	8.45	22.83	1.63	0.25	0.10

表四、中央研究院平衡語料庫 3.0 版各語式分佈比例（單位：萬）

語式	書面語	演講稿	劇本台詞	會話	會議記錄
CORPUS 字數	711.47	10.93	6.45	57.55	2.87
CORPUS 詞數	469.00	7.20	4.25	37.94	1.89
百分比%	90.14	1.38	0.82	7.29	0.36

表五、中央研究院平衡語料庫 3.0 版各文體分佈比例（單位：萬）

文體	記敘文	論說文	說明文	描寫文
CORPUS 字數	557.72	96.60	112.62	22.34
CORPUS 詞數	367.64	63.68	74.24	14.72
百分比%	70.66	12.24	14.72	2.83

三、分詞標準

語料選取完畢，接下來的工作是標記詞類，但是在這之前，還要先為語料做斷詞工作，唯有每個詞區隔非常明確之後，才能標記詞類。目前機器自動斷詞正確性，在不統計專有名稱與複合詞的前提下，可達99%左右〔Chen & Liu 1992〕。基本上，自動斷詞的步驟是以中研院辭典中的八萬目詞為基礎，切分為一個一個獨立的詞。沒列在辭典中的成分，則以字為單位，一一切分開。然後佐以構詞律對衍生性強的詞綴及數字組合成分進行結合詞彙的工作。而目前分詞的原則是採用中央標準局委託中華民國計算語言學學會研擬的「中文資訊處理分詞規範」國家標準草案的原則切分。

3.1 分詞原則

定義

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。這些基本詞類中，前五者，尤其是動詞和名詞的判定較複雜。原因有三：一、動詞和名詞皆另有詞組形式，便有區分複合詞和詞組的問題。另外副詞、定詞、和量詞也有類似的困擾。二、動詞、名詞是個開放性詞集，隨時都有新詞產生。三、一些結構複雜的字串，像是中插結構「洗了澡」或合併結構「中小學」，也需要細則來規範其分合標準。

因此除了定義外，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。

(1) 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）[1]…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。

(2) 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。

(1) 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[2]

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）

(2) 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。

(3) 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]。

(4) 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。

(5) 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。

[3] 因此我們需要一部標準辭典作為區分詞和非詞的依據。

(6) 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
   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）
   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
4. 正反問句：喜歡 不 喜歡、參加 不 參加
5. 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：

(1) 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

(2) 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：

(1) 有明顯分隔標記應該切分之。 切分原則

(2) 附著語素盡量和前後詞合為一個分詞單位。 合併原則

(3) 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

(4) 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

(5) 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

(6) 內部結構複雜之詞盡量切分之。 切分原則

3.2 範例與說明

依照上述分詞原則，我們對每種不同詞性、不同結構的字串便有了較一致、明確的分合標準。以下我們將依類舉例，標示出各式字串的分合情況，並說明所引用之分詞原則。舉例多半是具爭議性之例子，以便參考。

1.動詞：

並列結構：若符合基本原則(1)(2)或輔助原則(2)(3)任何一項則合併，否則予以切分。

例：讀誦文章、擴建完畢　　符合輔助原則(2)

叮 咬 不停　　不符合上述原則

偏正結構：若符合基本原則(1)(2)或輔助原則(2)(3)(4)任何一項，則合併。

例：改祭瓜果、大笑不已　　符合輔助原則(4)

高奏凱歌　　符合輔助原則(3)

主謂結構：若符合基本原則(1)(2)或輔助原則(2)(3)任何一項，則合併。

例：陷入價 升 量 減的走勢　　不符合任何一項

動賓結構：若符合基本原則(1)(2)或輔助原則(2)(3)任何一項，則合併。又中插情形依輔助原則(1)切分。

例：騙人、關門、洗衣、拔草、卸貨　　符合輔助原則(3)

騙了人、洗了一個澡　　符合輔助原則(1)(6)

述補結構：依細則基本原則(1)(2)一律合併。唯當補語是結果補語且是雙音節時，依輔本原則(6)切分。又中插情形依輔本原則(1)切分。

例：哭濕枕頭、爬上山頭、走進去、看 清楚、清洗 完畢

到：接觸到、認知到、跑到　　述補結構合併

聊 到半夜、走 到腿酸、加 到十萬　　非述補結構

為：譯為、流為、批評為、選拔為　　述補結構合併

成：擠成、剪成、歸劃成、堆積成　　述補結構合併

作：鑄作、換作、署名作、轉變作　　述補結構合併

動補中插：打得破、打不破　　符合輔助原則(1)

重疊結構：若符合基本原則(1)則合併。唯中插情形依輔助原則(1)切分

例：嘗試貌：談談、研究研究　　符合基本原則(1)

說說 看、說 看看　　符合輔助原則(1)

暫時貌：坐坐就走、擦擦即可　　符合基本原則(1)

程度貌：胖胖的、辛辛苦苦、慢吞吞　　符合基本原則(1)

其　它：坐 坐 坐　　不符合任何一項

重疊中插：笑 了 笑、哭 一 哭　　符合輔助原則(1)

帶詞綴：依輔(2)應合併。唯當動詞詞幹是動賓、述補結構依輔切分。

例：送給、贈送給、批發給　　符合輔助原則(2)

分紅 給、取出 給、退回去 給　　依輔助原則(6)

收有、列印有　　符合輔助原則(2)

正反問句結構：完整形式依輔助原則(1)將之切分，不完整形式依輔助原則(2)合併。另外若不完整形式有中插，則依輔助原則(6)把結構複雜者切分。

例：喜不喜歡　盜不盜壘　開不開刀　　符合輔助原則(2)

喜歡 不 喜歡　　符合輔助原則(6)

開 不 開 他 的 玩笑　　符合輔助原則(6)

合併結構：依基本原則(1)應合併。唯中插時依輔助原則(1)切分。

例：上下學，入出境，上、下課，入、出境

中插結構：依輔助原則(1)必須切分。

例：動賓、述補交互中插：幫 得 上 忙、喘 不 過 氣 來

2. 普通名詞：

並列結構：若符合基本原則(1)(2)、輔助原則(2)(3)任何一項則合併。

例：春夏秋冬、輕重緩急、男女、花草　　符合輔助原則(3)

偏正結構：若符合基本原則(1)(2)、輔助原則(2)(3)任何一項則合併。

例：大雨、象牙　　符合輔助原則(3)

公職 人員、財務 報表、公共 設施　　依輔助原則(6)分

重疊結構：依基本原則(1)應合併。

例：一隻狗狗、長痘痘、小車車

帶衍生詞綴、接頭/接尾詞：依輔助原則(2)(5)應合併。唯當詞組帶詞綴時，依輔助原則(6)應切分。

例：電腦室、業務部、太空 計畫 室、國際 關係 組

簡稱：依基本原則(1)應合併。

例：男單、女網、空姐、影視、化工、音像

合併結構：依基本原則(1)應合併。唯帶專名之合併結構不符合基本原則(1)不需合併。

例：詞頭合併：高中職、國內外

詞尾合併：父母親、公私立

套裝合併：事務局長、台北 市長、新竹 縣政府

中插結構：依輔助原則(1)應切合。

例：並列中插：春、夏、秋、冬、

男、女、老、少

3. 專有名詞：依基本原則(1)應一律合併。唯依輔助原則(6)有幾種結構複雜之專有名詞將不予合併。

例：單純詞：胡適、桂林、布農、貝多芬、克寧、阿爾及利亞

專名＋普名：普名是接尾詞：阿美族、光復橋、竹聯幫

普名是自由語素：胡 先生、平漢 鐵路 、二二八 事變

縮寫：勞基法、奧申委、文建會、台三線、中常會

複雜詞：台北市 第一 信用 合作社、省 自來水 公司

詞組或句子：鯨魚 的 生 與 死（書名）、

那 一 年 我們 都 很 酷（戲劇名）

4. 定量式

定詞：依定義應予以切分。唯數詞依基本原則(1)一律合併。

例：三十五，八萬零二十點七，三又二分之一，百分之四十，

三八,０００，２・３、２０％

量詞：依定義應予以切分。唯重疊結構依基本原則(1)一律合併。

例：片片、個個

定量詞：依定義定詞和量詞應切分。重疊結構依基本原則(1)予以合併。

又表時間、地點之定量詞依基本原則(1)應合併。

例：一 片、一 個	依定義切分
一片片、一個個	符合基本原則(1)具泛指功能
一 片 一 片、一 個 一 個	不符合基本原則(1)
八十四年 九月 一日 三時 二十分	符合基本原則(1)

七十巷 二十號 之一 三樓　　符合基本原則(1)

5.副詞：唯有符合細則基本原則(1)(2)、輔助原則(2)(3)任何一項才予以合併。又重疊結構若符合基本原則(1)則應予以合併。

例：暫不、既已　　符合輔助原則(3)

不過、要不是、或早或晚　　符合基本原則(1)

不料、不便　　符合輔助原則(2)

偷偷、悄悄　　符合基本原則(1) 或輔助原則(2)

叮噹 叮噹、砰 砰、咻 咻 咻　　不符合任何一項

6.成語、諺語：成語依基本原則(1)合併，諺語則依輔助原則(6)將成分作切分。

例：陰錯陽差、貌合神離、一不做二不休、一而再再而三

話 不 投機 半 句 多、虎 落 平陽 被 犬 欺

四、詞類標記

分詞工作完成後，接下來是為每一個成分標記詞類。用人力一個個去標記詞類，太耗時費力，目前用機器自動標記的準確率已能達到96%左右〔Chen et al. 1994〕，人所要做的是後處理工作，包括訂正自動標記錯誤的詞類，修正斷詞錯誤成分和指定詞類給這些新詞的工作〔Chang & Chen 1995〕。這些工作的前提就是要有一套完整的標記集及標記原則。

4.1 詞類標記集

我們採用的標記基本上是由詞庫小組八萬目詞辭典中的178個詞類〔詞庫小組 1993〕（可參見附錄二）經簡化後所得到的43個標記，另外加上3個特殊標記，共46個標記。如表六所示。這個對應的表，左邊所顯示的就是平衡語料庫所用的詞類標記，右邊則是相對應的辭典中的詞類，其後都附有簡單說明或例子。除了Daa, Dab, Neu, Nep, Nes, Neqa, Neqb, 這七個新增詞類外[4] ，其餘每個詞類所代表的意涵，可參照詞庫小組技術報告#93-05。簡單的說，C開頭的代表連接詞類、V代表動詞類、N代表體詞類、A代表非謂形容詞、D代表副詞類、P代表介詞、I代表感歎詞、T代表語助詞。簡化為46個標記的原則，是去除因語意區別而產生的細類，純粹就語法行為不同以及和其它類詞具區別性功能簡化而來的詞類。

也因為這個標記集是簡化而來的，要更粗糙或更精細，可以應使用者的要求很容易地做調整。

---

4 Daa和Dab是由Da（數量副詞）細分而來的。差異處在於前者可以直接修飾名詞組，後者不可以。

另外，原Ne（定詞）依語意和語法特性，細分為以下五類：

數詞定詞（Neu）：純數字、純數字組合及序數數詞，例：三、三千五百、幾、好幾，七百五十萬、第一、三十（好）幾、甲。

特指定詞（Nes）：具有特指（Specific）意義的定詞，不能單獨出現，可以直接修飾名詞，例： 某、該、本、同。

指代定詞（Nep）：除了定語，另有代詞功能。例：這、那、哪、什麼、啥、其

數量定詞（Neqa）：除了修飾中心語外，還可出現在論元位置（因中心語省略）、狀語位置（主語和謂詞之間）。例：許多、百分之五十、三分之一、五成三。少部分還可當補語，例： 她漂亮了許多。

後置數量定詞（Neqb）： 嚴格說，這類詞並不符合定詞的定義，只是它的用法都接在量詞之後，暫時將它們放在此類。例：三點正、五十歲出頭、兩丈許。

表六、中研院平衡語料庫詞類標記集

簡化標記	對應的CKIP詞類標記[5]	
A	A	/*非謂形容詞*/
Caa	Caa	/*對等連接詞，如：和、跟*/
Cab	Cab	/*連接詞，如：等等*/
Cba	Cbab	/*連接詞，如：的話*/
Cbb	Cbaa, Cbba, Cbbb, Cbca, Cbcb	/*關聯連接詞*/
Da	*Daa*	/*數量副詞*/
Dfa	Dfa	/*動詞前程度副詞*/
Dfb	Dfb	/*動詞後程度副詞*/
Di	Di	/*時態標記*/
Dk	Dk	/*句副詞*/
D	*Dab*, Dbaa, Dbab, Dbb, Dbc, Dc, Dd, Dg, Dh, Dj	/*副詞*/
Na	Naa, Nab, Nac, Nad, Naea, Naeb	/*普通名詞*/
Nb	Nba, Nbc	/*專有名稱*/
Nc	Nca, Ncb, Ncc, Nce	/*地方詞*/
Ncd	Ncda, Ncdb	/*位置詞*/
Nd	Ndaa, Ndab, Ndc, Ndd	/*時間詞*/
Neu	*Neu*	/*數詞定詞*/.
Nes	*Nes*	/*特指定詞*/
Nep	*Nep*	/*指代定詞*/
Neqa	*Neqa*	/*數量定詞*/
Neqb	*Neqb*	/*後置數量定詞*/
Nf	Nfa, Nfb, Nfc, Nfd, Nfe, Nfg, Nfh, Nfi	/*量詞*/
Ng	Ng	/*後置詞*/
Nh	Nhaa, Nhab, Nhac, Nhb, Nhc	/*代名詞*/
I	I	/*感嘆詞*/
P	P*	/*介詞*/
T	Ta, Tb, Tc, Td	/*語助詞*/
VA	VA11,12,13,VA3,VA4	/*動作不及物動詞*/
VAC	VA2	/*動作使動動詞*/
VB	VB11,12,VB2	/*動作類及物動詞*/
VC	VC2, VC31,32,33	/*動作及物動詞*/
VCL	VC1	/*動作接地方賓語動詞*/
VD	VD1, VD2	/*雙賓動詞*/
VE	VE11, VE12, VE2	/*動作句賓動詞*/
VF	VF1, VF2	/*動作謂賓動詞*/
VG	VG1, VG2	/*分類動詞*/
VH	VH11,12,13,14,15,17,VH21	/*狀態不及物動詞*/
VHC	VH16, VH22	/*狀態使動動詞/
VI	VI1,2,3	/*狀態類及物動詞*/
VJ	VJ1,2,3	/*狀態及物動詞*/
VK	VK1,2	/*狀態句賓動詞*/
VL	VL1,2,3,4	/*狀態謂賓動詞*/
V_2	V_2	/*有*/
DE	/*的, 之, 得, 地*/	
SHI	/*是*/	
FW	/*外文標記*/	

5 斜體詞類，表示在技術報告#93-05中沒有定義，即後來增列的。

4.2 詞類標記所代表的功能

詞類標記的目的是標示一個詞在句子中的語法功能。而這個系統所採用的46個詞類標記基本上是由詞庫小組辭典而來。辭典中詞類給定的原則，理論上是一個詞一個類，相近語義不多重分類。對於某個詞類中大部份的詞都可以扮演其它相同的語法功能，也不另外多重分類。例如狀態不及物動詞（VH）在句子中除了當主要謂語，也可以是狀語或名詞修飾語。我們在做詞類標記（tagging）時，也是秉持這個原則，只給VH。表示它可以是主要謂語、狀語或名詞修飾語。也就是說在中研院平衡語料庫中一個標記不一定只代表一個功能，因此在這一節裏必須先介紹每個標記可以扮演的語法功能。

一般而言，這些標記可以分為兩大類，一類是單一功能的標記，一個標記只代表一個語法功能，出現的環境很固定；另一類是多功能的標記，例如上述的VH。以下簡述哪些是屬於單一功能的標記，哪些是屬於多功能的標記，以及這些多功能標記所代表的涵意與功能。

單一功能的標記有：

Caa	對等連接詞。例：張三和李四、大又圓
Cab	列舉連接詞。例：身分證、戶口名簿等證件
Cba	句尾關聯連接詞。例：你不來的話，我也不來
Cbb	關聯連接詞。例：雖然他很聰明，但是不用功
Dfa	程度副詞，大都緊接在狀態動詞前。例：萬分難過
Dfb	程度副詞，緊接在狀態動詞後。例：難過萬分
Di	時態標記，大都緊接在動詞後。例：做起事來、看了書
Dk	句副詞，大都出現在句首。例：總而言之，你不對
D	一般副詞（包括法相詞），在動詞和主語之間。例：他可能回去了
Nf	量詞，多緊接在定詞後。例：一枝筆、這天
Ng	後置詞，出現在詞組尾的帶論元功能詞。例：三年來、理論上
Neu	數詞定詞，可以單獨出現，可以直接修飾量詞或名詞。 例：年滿三十、兩輛
Nes	特指定詞，不能單獨出現，直接修飾定量詞或名詞。例：某（位）名人
Neqb	後置數量（定）詞，只出現在定量詞之後，不會出現在量詞前的。 例：五十歲開外
P	介詞（或前置詞），是一種帶論元的功能詞。例：他從家裡來
I	感歎詞，一般出現句子前。例：哦，我知道了
T	語助詞，幾乎都出現在句尾。例：你來嗎

多功能的標記，我們簡單的用V來表示這個標記有謂詞的功能，ADV表示狀語功能或說可以出現在狀語位置，N-modifier表示可以修飾一般名詞或說緊跟在名詞前面。“是”

和“有”語法行為比較複雜，自成一類〔魏文真等 1991a, b〕。DE這個標記則是針對“的/之、地、得”等附加語扮演的特殊功能所設計的，目前暫不區分其間的差異。

多功能的標記有：

Da	ADV、N-modifier　例：僅(Da)知道；僅(Da)三人
N*-Nf-Ng	N、N-modifier（“N*”表示名詞類的集合） 例：坐公車(Na)；公車(Na)司機
Ncd	N、N-modifier、locative marker 例：前(Ncd)有樹；前(Ncd)院；公園前(Ncd)
Nd	N、N-modifier、ADV 例：在暑假(Nd)；暑假(Nd)作業；他暑假(Nd)不回家
Nep	N(h)、N-modifier　（Nh表代名詞） 例：這(Nep)代表什麼；這(Nep)車子
Neqa	N、N-modifier、ADV 例：吃了全部(Neqa)；全部(Neqa)學生；他們全部(Neqa)走了
V*	V、N-modifier　例：主辦(VC)奧運；主辦(VC)單位
VH	V、N-modifier、ADV[6] 例：很努力(VH)；努力(VH)方式；他努力(VH)工作
V_2	有　例：有(V_2)人走來；他有(V_2)書
SHI	是　例：他是(SHI)很認真；他是(SHI)老師；他全身是(SHI)傷
A	N-modifier、ADV　例：天生(A)好手；他天生(A)脾氣壞
DE	nominal marker、adverbial marker、complement marker 例：我的(DE)書；高興地(DE)笑；玩得(DE)高興

由於中文詞彙活用的現象很普遍，辭典不可能盡列每一個詞的每一種用法。例如：“八股”在辭典登錄的是Na（普通名詞）；但是在“他很八股”中做主要謂語時，則應該標記為VH，因為Na並無謂 詞的功能。“善意”在辭典中登錄的是Na，當它在“他善意規勸”中應保留Na標記，抑是副詞(D)標記？這些入句結果在我們後處理標記的過程中，都曾是疑問，現今依照每個標記所對應的功能得知Na無狀語功能，此處應標記為D。因此，目前採取這樣的方式，對每一個詞類標記都有明確的功能規定，期望能儘量做到一致性。

6　VH可以有狀語功能，但只限於表方式的狀語。

4.3 詞類標記的原則與範例

既然詞類標記的目的，是標示一個詞在句子中的語法功能，詞類標記最基本的原則，應該就是這個詞的詞類標記符合它在語境中所扮演的語法功能。所以，如果一個詞在辭典中的標記並不符合他在語境中所扮演的功能，應該要另外指定詞類標記給它[7]，如“八股”、“善意”。

另外一種情況也是機器自動標記易出錯的，像同形詞的問題，例如：“想”有二個標記，一個是VE， 一個是VJ，想(VE)是認為或思考的意思，想(VJ)是想念的意思。這時人要去判斷在語境中這個“想”究竟是何種語意，才能給予適當的標記[8]。所以，詞類標記的第二個原則是一個字串在辭典中有一個以上的標記，依它在語境中的語意及語法功能給予適當的標記。

以下要說明的是在實際標記詞類的過程，會遇到一個詞（可能是同形詞，也可能是同義詞）有兩個以上的標記，而這些標記中有部分的功能重疊，此時，最易造成標記者的困擾，這一部分可能也是做得最不一致的，因為有些結論是在持續討論後才獲得最後結果。在此，只能以舉例的方式來說明對某一多重標記類型的標記原則或對某個別詞的標記原則。期望從這些範例與說明中讓使用者能更清楚我們的標記標準。因此，第三個標記原則是一個字串在辭典中有一個以上的標記，且標記間有功能重疊之處，則依各類型的規範處理。

範例一：來（VA,Ng,D,T），去 (VCL, VC, D, T)

(1) 我們(Nh)到(VCL)海邊(Nc)去(D) 玩(VC)。

(2) 我們(Nh)去(VCL)海邊(Nc)玩(VC)。

(3) 我們(Nh)到(VCL)海邊(Nc)去(T)。

(4) 她(Nh)把(P)蘋果(Na)去(VC)皮(Na)了(T)。

(5) 他(Nh)來(VA)了(T)。

(6) 三(Neu)年(Nf)來(Ng)。

(7) 他(Nh)唱(VC)起(Di)歌(Na)來(T)了(T)。

說明：這個例子主要是要說明“來、去”(V)和(D)的區別。在動作語意是否虛化，如(1)和(2)的對比。另外是(T)的指定，只有在它們出現在句尾或在“V起N來”的語境下給予。

---

7 一個標記可以扮演的語法功能，請見4.2節。至於入句結果或稱活用結果目前暫未登錄在辭典中，以後將利用統計方法一併處理。如“善意”的例子，如果發現它在實際使用中狀語的頻率比名詞高，就可以把這種用法也登錄在辭典中

8 在研究院語料庫看到的詞類標記似乎是語法標記，然而這些標記在辭典中代表的涵意是語意加上語法功能的標記。

範例二：(Ncd, Ng, Nes)　　如：上，下，前，後

(8)一千四百(Neu)年(Nf) 前(Ng) ；院子(Nc)前(Ncd)

(9)前(Ncd)有(V_2)小(VH)河(Na)，後(Ncd)有(V_2)小(VH)溪(Na)

(10)前(Ncd)院(Na);後(Ncd)殿(Nc)

(11)前(Nes)三(Neu)天(Nf)

說明：位置詞（Ncd）和方位詞（以後將正名為後置詞，Ng）的功能有重疊的地方。以往兩者都可以出現在地方詞組的最後，現在我們對兩類詞的定義有所改變，如果一個詞的作用是表方位的後置詞，仍給予Ncd，非表方位的則給Ng，如(8)。故位置詞(Ncd)除了本有的地方名詞性質，也有類似後置詞功能；後置詞(Ng)，是功能詞，和前接論元組合而成的詞組成分角色，由其論元成分決定。

另，位置詞也有定語的功能，如(10)所示。而我們對於某些位置詞仍給予特指定詞(Nes)的標記，一方面是因為它們的語意已經轉為順序關係，而非方位，一方面是便於語法上的表達，定詞和量詞或數詞先組合成定量式詞組修飾名詞，如(11)所示。

範例三：(P, Caa)　　如：和、跟、與

(12) 他(Nh)喜歡(VK)蔬菜(Na)和(Caa)水果(Na)

(13) 他(Nh) 和(Caa) 我(Nh) 去(D) 打球(VA)

(14) 他(Nh) 昨天（Nd) 和（P) 我（Nh) 去(D) 打球(VA)

(15) 這(Nep)件(Nf)事(Na) 和(P) 他(Nh) 沒關係(VH)

(16) 他(Nh)和(P)她(Nh)一樣(VH)高(VH)

說明：對等連接詞(Caa)和介詞(P)都是功能詞，這類詞它所以仍保留歧義的這兩個標記，實在是這兩個標記的功能是截然不同的，如(12)的“和”不可能是介詞。只是在某些語境下因語意無明顯差異而難以判定，如(13)可以是Caa，也可以是P，給Caa是強調兩人一起去，給P是著重在“伴隨、跟著他”的意思。在這種歧義情況下會優先給予Caa標記。但在兩個成分間如果中插有時間詞或副詞則優先給予P標記，如(14)。

另外，這類詞在“伴隨、比較”語境下，都是介詞標記，如(15)和(16)。如果(15)給予Caa，則語意將是這件事與他一起（或都）沒關係，實際上不是這樣的。

範例四：為(P, VG, P, VJ)

(17) 他(Nh)為(P)人(Na) 所(D)殺(VC)。(為(P) =被)

(18) 為(P) 提高(VC)工作(Na)效率(Na)。(為(P) =為了)

(19) 連戰(Nb)為(VG)行政院長(Na)。(為(VG) =是)

(20) 我(Nh)買(VC)這(Nep)棟(Nf)房子(Na)，是(SHI)為(VJ)了(Di)你(Nh)。

說明：某些詞有動詞也有介詞標記，語意類似，只是在語境中扮演的語法功能不同，如(18)和(20)。(20)標VJ，是因為人覺得“為”在這裡是謂詞，語意已完整，在某些語境下(18)也會標為VJ。所以，這一部份可能因標記者語感的不同而有不一致處。(17)和(18)中的“為”都標記為P，實際上是同形詞。

範例五：過(Di,Dfa,VH,VCL)

(21) 我看過(Di)這本書。

(22) 車身過(Dfa)高。

(23) 我英文過(VH)了。

(24) 過(VCL)聖誕節是小孩最喜歡的。

(25) 過(VCL)橋

說明：這一個例子要講的是對於某些常用詞，語意上做了一些轉喻的用法，如果不怎麼影響語法行為，我們並不另外給標記，如(24)和(25)。(25)中的“過”是空間上的意義，(24)則是轉為時間上的意義。(23)則是及格意義，已經轉成一個新語意，又是不及物的用法，因此，當一個新詞處理。同理“經過”也是一樣處理，無論是經過什麼地方，或經過這些事，“經過”都只標記為VCL。

範例六：(Neqa, VH)　　如：多、很多、不少、…

(26) 很多(Neqa)位(Nf)學生(Na)都(D)看到(VC)。

(27) 很多(Neqa)都(D)被(P)他(Nh)看到(VC)。

(28) 客人(Na)走(VA)了(Di)很多(Neqa)。

(29) 民進黨(Nb)很多(Neqa)退出(VCL)會場(Nc)。

(30) 他(Nh)吃(VC)得(DE)多(VH)

(31) 她(Nh)漂亮(VH)很多(Neqa)。

(32) 他(Nh)很少(D)開口(VA)。

(33) 錯誤(Na)不少(VH)。

說明：基本上，VH也可以修飾名詞，也可以當狀語，但是當它是數量語意，又有定語功能時，我們會優先給予Neqa的標記，因為(26)～(29)就是數量定詞(Neqa)常見的語法行為。對於這一類詞如果出現在狀態不及物動詞後面，雖有程度

意涵，仍然標記為Neqa，如(31)。如果語意已轉為頻率時間副詞，則給予D，如(32)所示，畢竟只有少數這類詞有這樣轉喻。如果在語境中，是動詞功能，自然標記為動詞，如(33)。至於(30)之所以給VH標記，是因為得字補語的特性，給動詞較合理，況且也不能說“他吃得很多東西”，所以，“很多”出現在(30)句，並非其後省略被修飾名詞的定語角色，而是不折不扣的動詞。

範例七：有些、有點、有一些、有一點

(34) 他(Nh)有(V_2)些(Nf)錢(Na)。

(35) 他(Nh)有些(Dfa)難過(VH)。

(36) 有(V_2)一些(Neqa)人(Na)不(D)相信(VK)。

(37) 他(Nh)有一些(Dfa)瘋狂(VH)。

(38) 他(Nh)一點(Neqa)也(D)不(D)知道(VK)。

說明：“些”和“點”基本上是量詞，故(34)的“些”標記為Nf，但是在“一些、一點”出現而且是“Some”語意時，我們則不將它們切分開[9]，而將它們視為詞彙化的一個成分，如(38)所示。它們的語法行為也大致和數量定詞（Neqa）吻合。

範例八：了（T, Di）、的（T, DE）

(39) 她(Nh)穿(VC)新(VH)衣服(Na)了(T)。

(40) 她(Nh)吃(VC)了(Di)。

(41) 她(Nh)是(SHI)很(Dfa)高興(VK)的(T)。

(42) 她(Nh)是(SHI)賣(VD)菜(Na)的(DE)。

說明：如果是以句尾的位置來判斷是否該標記T（語助詞）並不正確。對於“了”而言，如果是緊接在動作動詞的後面，則一定標記為Di，所以(40)中的“了”即使出現在句尾，也不是標記為T。在(39)情況則一定是T，其它情況則看自動標記做出來的結果是什麼就保留什麼。在句尾的“的”通常為T，除了出現在表示所有格、副詞標誌或名詞中心語省略的位置，如(42)所示。

4.4 特徵標記集

除了標記詞類以外，我們也為某些特殊句法表現做標記，目前使用的特徵標記共9個，是針對動補式動詞和動賓式動詞的可拆（separable）現象、合併詞中插現象〔計算語言學通訊 1995〕、名物化結構、外來語和專有名詞所設計。特徵標記集如表七所示。

9 但是在“有一（或兩）點污點在他臉上”時，“點”則須和“一”切分開當量詞處理。因為這時“點”是個體量詞“個”的意思。

表七、中研院平衡語料庫特徵標記集

特徵標記	使 用 情 況	例　　子
+vrv	V of a separable VR compound	叫Vc[+vrv]不醒
+vrr	R of a separable VR compound	叫不醒 Vc[+vrr]
+spv	V of a separable V N compound	吃Vc[+spv]了他的虧
+spo	N of a separable V N compound	吃了他的虧 Na[+spo]
+p1	the first part of a separated compound	初(Nc)[+p1]、高中(Nc)
+p2	the second part of a separated compound	星期六(Nd)、日(Nd) [+p2]
+fw	the feature of a foreign word	卡拉OK(Na)[+fw]
+nom	the feature for verbal nominalization	他的不講理(VA)[+nom]
+prop	the feature for proper nouns	人本(A)[+prop]基金會(Nc)

A. 動補式特徵標記

動補式動詞依分詞標準應為一個單位，當作一般詞處理，給予適當的詞類標記。但是我們也知道某些動補式動詞具有可拆的語法行為，也就是在中間可以加入“得”或“不”成分。這時候這個詞在形式上就被切分為動、補兩個成分了，如(43)所示。為了保持原詞彙的語法類與語意，在語料庫中標記的方式如(44)所示，給予動詞成分和補語成分原動補動詞的詞類標記，另外，在兩個成分上各加上[+vrv]和[+vrr]的語法特徵。

(43) 我 叫 不 醒 他。

(44) 我(Nh) 叫(VC)[+vrv] 不(D) 醒(VC)[+vrr] 他(Nh)。

附帶一說的是，像(45)句中的“上不了”是當一個詞處理，因為並沒有“上了”這樣一個動補式動詞。“忍不住”也是當一個詞處理，而不切分，原因是它的語法行為和“忍住”不同，可視為一個詞彙化的成分。

(45) 他(Nh)上不了(VCL)車(Na)。

(46) 他忍不住(VH)哭了。

B. 動賓式特徵標記

可拆的動賓式動詞其間中插的成分更多樣，可以是時態副詞、所有格、定量詞或其他修飾語。表面上符合詞組律結構（動詞組→動詞　名詞組），語意卻不是可以從這個動詞組得來，如(47)和(48)中的“吃虧”和“吃醋”。在語料庫中對這種現象的標記是給動詞部份適當的詞類標記，並標示[+spv]的特徵，而賓語部份給Na（一般名詞）的詞類標記及[+spo]的特徵，如(47)-(51)。

(47) 他(Nh) 吃(VC)[+spv]了(Di) 很(Dfa) 大(VH) 的(DE) 虧(Na)[+spo]。

(48) 吃(VC)[+spv]你(Nh)的(DE)醋(Na)[+spo]

(49) 睡(VA)[+spv]了(Di)一(Neu)覺(Na)[+spo]

(50) 走(VA)[+spv]起(Di)路(Na)[+spo]來(T)

有時候因為動補式和動賓式互相中插的結果，使得標記情況更複雜，如(51)-(53)所示。(51)中“幫上”是一個動補式，“幫忙”是一個動賓式，因此在“幫”和“上”上分別有[+vrv] [+vrr]的特徵，而“幫”和“忙”上分別有[+spv] [+spo]的特徵。

(51) 幫(VC)[+spv][+vrv]得(DE)上(VC)[+vrr]忙(Na)[+spo]

(52) 使(VC)[+spv][+vrv]不(D)上(VC)[+vrr]勁(Na)[+spo]

(53) 喘(VC)[+spv][+vrv]不(D)過(VC)[+vrr]氣(Na)[+spo]來(T)

C. 合併詞中插特徵標記

合併詞或合併詞中插的現象在計算語言學通訊的「搜文解字」專欄（1995年3月）有清楚的描述。由於詞（或成分）的完整性被中插成分破壞，造成標記詞類的困擾，我們目前處理的方法說明如下。

如果有標點或連接詞等中插的情形，每一成分都標記原來未省略時的詞類，然後在省略的那一部份標特徵[+p1]或[+p2]，表示他和整個詞的密切關係。而且此特徵只加在那些被切開的獨立部份是附著語素或本身詞類和整個合併詞的意思、詞類不同時，如(54)-(56)。

(54) 冠(Na)[+p1]、亞(Na)[+p1]、季軍(Na)

(55) 初(Nc)[+p1]、高中(Nc)

(56) 高中(Nc)、職(Nc)[+p2]

並列結構的定量式、定名式等，可用構詞律組合成的詞，因有規律可循，只要把每個成分分開，獨立給詞類，不需標示此特徵。如：一(Neu)、二(Neu)壘(Na)，上(Nes)、下(Nes)半(Neqa)場。

D. 外來語特徵標記

在中研院平衡語料庫中，對於外來語（或字串）處理有兩種標記，一個是詞類標記（FW），一個是特徵標記（+fw）。給FW詞類或特徵[+fw]的判斷標準在於，能否判斷出現的外文字串是什麼詞類，如果根本不知道那是什麼就給FW的詞類標記，否則，就給適當詞類加上[+fw]的特徵。如 KTV，及卡拉OK等已是大眾熟悉的外來語，而且已知所指為何，如果它是一個地方名詞就給Nc的標記，如果是一般名詞就給Na的標記，另外加註+fw的特徵。如，KTV(Nc)[+fw]；卡拉OK(Na)[+fw]。

E. 名物化特徵標記

中文動詞名物化的現象非常普遍〔葉美利等 1992〕，通常動詞（組）出現在主語位置或某些動詞（有人稱為虛化動詞）的賓語論元就感覺動詞事物化了，如(57)-(58)。而我們反而沒有在這些現象上標記[+nom]的特徵，主要原因是這些將來可能都可以靠語法結構得知。我們會標註[+nom]特徵的只有在動詞出現在名詞片語結構的中心語或修飾名詞時，主要是為了在日後自然語言處理時減少語法結構的歧義，如(59)-(62)都是名詞片語結構。

(57) 進行(VC)調查(VE)

(58) 維持(VJ)清潔(VH)

(59) 學生(Na)的(DE)不(D)合作(VH)[+nom]

(60) 他(Nh)對(P)國家(Na)的(DE)認同(VJ)[+nom]

(61) 主辦(VC)[+nom]單位(Nc)

(62) 吵架(VA)[+nom]方式(Na)

但是像(63)中的動詞“犯”雖然也在名詞片語結構中，卻不必給[+nom]特徵，原因是“犯”在此處仍保有所有動詞的特性，應和“錯誤”先組合成動詞組，此處如果加特徵[+nom]反而給了錯誤的訊息。

(63) 他(Nh)的(DE)頻(D)犯(VJ)錯誤(Na)

F. 專有名詞特徵標記

中文專有名詞的用法非常普遍，雖然已經有專有名詞專屬的標記 Nb，但是中文任何一類的詞皆有可能作為專有名詞，如(64)-(68)，為了保留原有的詞類，又不失掉其專有名詞的特性，除了原有的詞類外，另外標上[+prop]，表示專有。

(64) 香港(Nc)中文(Na)[+prop]大學(Nc)、

(65) 歐威爾(Nb)一九八四(Nd)[+prop]小說(Na)的(DE)楔子(Na)

(66) 九十年代(Nd)[+prop]月刊(Na)總編輯(Na)李怡(Nb)

(67) 人本(A)[+prop]基金會(Nc)

(68) 上揚(VH)[+prop]有聲(A)出版社(Nc)

附錄一、詞綴列舉表

(A) 詞頭：<副>廠長、<非>假日、<多>功能、<老>王、<小>林、<高>收入、<低>姿態、<超>低溫、<零>風險、<單>音節、<雙>淘汰、<前>首相、<準>女婿、<總>支出、<主>幹線、<代>校長、<阿>忠、<第>一。

(B) 詞尾：

ㄅ-ㄈ　澳<幣>、出勤<簿>、姪孫<輩>、得分<榜>、當權<派>、西瓜<棚>、女士<們>、丁<某>、美<方>、安裝<費>、劫機<犯>、雞<販>、幅射<波>、會長<盃>、體育<版>、職業<別>、擊球<棒>、武打<片>、洗澡<盆>、合成<品>、換算<表>、國畫<班>、豌豆<苗>、歌劇<迷>、麵包<坊>、倒閉<風>、罷免<法>、經濟<面>。

ㄉ-ㄌ　洩洪<道>、飽和<度>、信號<彈>、芒果<凍>、書報<攤>、緝私<艇>、祖孫<倆>、打氣<筒>、兄弟<檔>、蛋<農>、交易<量>、曝光<率>、河川<地>、反對<黨>、申報<單>、平衡<點>、救援<隊>、主播<台>、塑膠<桶>、購買<力>、東湖<里>、舉手<禮>、候車<亭>、特赦<令>、鋼鐵<類>、北方<佬>、仁愛<路>。

ㄍ-ㄏ　韻律<感>、鳳梨<乾>、糖果<罐>、油漆<工>、偷竊<狂>、宇宙<觀>、水泥<塊>、槍枝<庫>、透明<化>、拆除<戶>、工人<荒>、國光<號>、投機<股>、工程<款>、侵略<國>、會員<卡>、金屬<桿>、實習<課>、香水<盒>、說明<會>、搶手<貨>、廣東<話>、水果<行>、感謝<函>、蒸<鍋>、油漆<料>、踢<給>、踢踢<看>、踢<看看>。

ㄐ-ㄒ　陪產<假>、電焊<匠>、音樂<季>、健保<局>、李<君>、提貨<券>、週轉<金>、伊<軍>、國家<級>、收盤<價>、攪拌<器>、黑暗<期>、山東<腔>、生活<圈>、備詢<席>、迷你<型>、季節<性>、劉<姓>、吳<嫌>、新竹<訊>、孔子<像>、研磨<機>、生髮<劑>、日本<籍>、古裝<劇>、婦女<節>、老鼠<夾>、油畫<家>、銀行<界>、展示<區>、編輯<群>、採訪<權>、指定<曲>、防盜<險>、嘉義<線>、泰山<鄉>、雲林<縣>、ㄒ字<形>、動物<學>、政治<秀>、行李<箱>、歷險<記>、化工<系>。

ㄓ-ㄖ　成交<值>、車<主>、回顧<展>、鋸齒<狀>、偵測<站>、追求<者>、藏匿<處>、荷花<池>、曬穀<場>、營養<師>、藝術<史>、休息<室>、邵<氏>、羊<舍>、成衣<商>、歡呼<聲>、子女<數>、擒拿<術>、復古<式>、兩黨<制>、鄭<宅>、教務<長>、電腦<桌>、橡皮<章>、識別<證>、加工<廠>、鳳林<鎮>、雲南<省>、嘉義<市>、診斷<書>、溜冰<社>、娛樂<稅>、狙擊<手>、車<身>、投資<人>、收藏<熱>。

ㄗ-ㄟ　電玩<族>、台中<仔>、年齡<層>、兒童<餐>、延長<賽>、強盜<罪>、西瓜<子>、新坪<村>、玫瑰<色>、採訪<組>、採購<案>、扣除<額>。

ㄧ-ㄦ　尾牙<宴>、水泥<業>、樣品<屋>、助選<員>、西班牙<語>、森林<浴>、攻擊<慾>、污染<源>、銅錢<味>、麻醉<藥>、德<裔>、潛水<衣>、高腳<椅>、鞋<印>、抵押<物>、印度<文>、扇子<舞>、螞蟻<窩>、草莓<園>、列印<有>、鳥<兒>。

＊“給”作詞綴，請參考Huang (1992)。

＊ 以上所列詞頭和詞尾有些看起來像是個自由語素，像是“盆”“線”都有獨用情形。但是“洗澡盆、嘉義線”中，我們認為“盆、線”仍是附著語素，因為它們和前者結合緊密，且具有衍生性功能，可組合成許多複合詞，像是“花盆、臉盆、鋼盆、聚寶盆、玉盆、火盆、金盆、果盆、…”“台三線、港澳線、花蓮線、北迴線、成渝線、湘黔線、日本線、黃金線、山線、平原線…、”。

＊ 其中“單、主”可兼作詞頭及詞尾的功能。

附錄二、中文詞類分析總表

一、述詞，是謂語中心。（依動作/狀態、及物性、論元個數以及述詞後接成分的詞組形式分為十二大類）

VA 動作不及物述詞，這類述詞只需要一個名詞組參與論元即可。（依論旨角色、語意特性的不同分為四類）

VA1 表移動或存在的述詞，論旨角色為客體（theme）。（依語意及內部結構的不同分為三類）

VA11 表移動或行動的述詞，可後接地方成分，有主語倒置的現象。
例：跑、飛、走。

VA12 表存在或靜態的述詞，可後接地方成分，有主語倒置的現象。
例：坐、躺。

VA13 內部結構為述賓結構且賓語表地方成分的行動述詞。
例：逛街、上臺、出場。

VA2 作格述詞，論旨角色為客體（theme），述詞前可有一個肇始者（causer），原來出現在述詞前的客體移到賓語的位置。例：出動、轉。

VA3 氣象述詞，論旨角色為客體（theme）。
例：下雨、颱風、打雷。

VA4 一般的動態述詞，論旨角色為主事者（agent）。
例：違規、謀生、開會。

VB 動作類單賓述詞，語意上需要兩個參與論元，但它的賓語不能直接出現在述詞後，而以介詞引介或賓語前提的方式出現。（依論旨角色的不同分為兩類）

VB1 賓語為動作施行的對象，其角色為終點（goal）。（依句型的不同分為兩類）

VB11 終點一定要以介詞引介出現在述詞前或後。例：求婚、拜年。

VB12 終點可以是名詞組出現在主語位置。例：立案、整容、解體。

VB2 賓語的角色為客體（theme）。例：充公、除名、送醫。

VC 動作單賓述詞，語意上需要兩個參與論元。（依論旨角色的不同分為三類）

VC1 表移動或存在的述詞，主語為客體（theme），賓語為表地方的終點（goal）。
例：進、闖入、經過、逃離、住、世居。

VC2 以主事者（agent）為主語，終點（goal）為賓語。
例：打、學、訪問、使用、破壞、照顧。

VC3 以主事者（agent）為主語，客體（theme）為賓語。（依句型的不同分為三類）

VC31 述詞後除了賓語外不需再接一個地方成分。
例：買、賺、吃、生產、組織、收取、洩露。

VC32 述詞後除了賓語外，通常還接一個由介詞「到」引介的地方詞。
例：走私、引渡、調遣、押送、發射、搭載。

VC33 述詞後除了賓語外，通常還接一個由介詞「在」或「到」引介的地方詞，而且有地方詞倒置的現象。
例：放、埋、懸掛、儲存、搭建、囚禁。

VD 雙賓述詞，這類述詞在語意上有傳遞事物的動作訊息，需要三個參與論元。（依間接賓語的論旨角色的不同分為兩類）

VD1 表將一事物傳遞給對方的述詞，主事者（agent）具有起點特徵（+source），間接賓語是終點（goal）。例：寄、送、捐。

VD2 表向對方取得一事物，主事者（agent）具有終點的特徵（+goal），間接賓語是起點（source）。例：搶、敲詐、索取。

VE 動作句賓述詞，後接句賓語的動作及物述詞。（依論元個數的不同分為兩類）

VE1 三元述詞。（依語意上的不同分為「問類」及「說類」兩類）

VE11 問類，以主事者（agent）為主語，以終點（goal）為間接賓語，客體（theme）為直接賓語 （句賓語），句賓語為疑問句式，且疑問範圍只到包接句。 例：責問、詢問。

VE12 說類，和VE11的論旨角色相同，不同的是：

VE12的句賓語不限於疑問句。

句賓語的疑問範圍不限於包接句。

主語或間接賓語與句賓語之主語間可有共指關係。

例：提示、許諾、指引。

VE2 二元述詞，以主事者（agent）為主語，終點（goal）為句賓語，語意多為表語言行為之述詞。
例：悲歎、自誇、下令、研究、討論、探索、反省、強調、猜想、說、提到。

VF 動作謂賓述詞，後接述詞組賓語的動作及物述詞。（依論元個數的不同分為兩類）

VF1 二元述詞，以主事者（agent）為主語，終點（goal）為賓語，語意多含有「打算」之意。例：企圖、想、打算。

VF2 三元述詞，以主事者（agent）為主語，後帶一個終點（goal）的名詞組賓語，再帶一個表客體（theme）的述詞組賓語。其中這個名詞不但是主要述詞的賓語，也是述詞組賓語的主語，是一般所謂的「兼語式」述詞，此類述詞語意多表「鼓勵」、「命令」、「強迫」、「請求」。例：任用、勸。

VG 分類述詞，擔任主語和補語間連結的角色。（依論元個數的不同分為兩類）

VG1 三元述詞，這類述詞帶有主事者（agent）、客體（theme）和範圍（range）三個論元。例：稱呼、喊、命名。

VG2 二元述詞，典型的分類述詞，帶客體（theme）和範圍（range）兩個論元。例：姓、當。

VH 狀態不及物述詞，用以描述事物所呈現的某種狀態，這類述詞只需要一個參與論元即可。（依論旨角色的不同分為兩大類）

VH1 論旨角色為客體（theme）。（依句型的不同分為七類）

VH11 一般的不及物述詞。例：動聽、浪漫、特別。

VH12 能夠後接定量詞表示量度的述詞。例：入超、增值、淨重。

VH13 能夠後接比較對象及兩者差額的述詞。例：大、高、慢。

VH14 可以後接地方成分，有地方詞倒置句型。例：瀰漫、矗立。

VH15 可以以句子作為主語，且可將句子移至述詞後。例：值得、夠、適合。

VH16 作格述詞，述詞前可有一個肇始者（causer），原來述詞前的客體（theme）移到一般賓語的位置。例：辛苦、豐富、穩固。

VH17 述詞前可有一個接受者（recipient），是客體（theme）的擁有者。例：丟、瞎、斷。

VH2 論旨角色為經驗者（experiencer）。（依句型的不同分為兩類）

VH21 非作格述詞。例：心酸、想不開。

VH22 作格述詞，述詞前可有一個肇始者（causer），原來述詞前的經驗者（experiencer）移到賓語的位置。例：震驚、為難、急煞、感動。

VI 狀態類單賓述詞，語意上需要兩個參與論元，但它的賓語不能直接出現在述詞後，而以介詞引介或賓語提前的方式出現。（依論旨角色的不同分為三類）

VI1 以經驗者（experiencer）為主語，終點（goal）為賓語，表心靈狀態。例：心動、灰心、傾心。

VI2 以客體（theme）為主語，以終點（goal）為賓語。

例：內行、不利、為例。

VI3 以客體（theme）為主語，以起點（source）為賓語。

例：受教、取材、取決。

VJ 狀態單賓述詞，這類述詞在語意上需要兩個參與論元。（依論旨角色不同分為三類）

VJ1 以客體（theme）為主語，以終點（goal）為賓語。

例：迎合、代表。

VJ2 以經驗者（experiencer）為主語，終點（goal）為賓語，表心靈狀態。例：景仰、惦念、嫌忌。

VJ3 以客體（theme）為主語，以範圍（range）為賓語。

例：長達、剩餘。

VK 狀態句賓述詞，後接句賓語的狀態及物述詞。（依照主語論旨角色的不同分為兩類）

VK1 以經驗者（experiencer）為主語，以終點（goal）為賓語，表心靈狀態。 例：不滿、嫌惡。

VK2 以客體（theme）為主語，以終點（goal）為賓語。

例：反應、在於、干係。

VL 狀態謂賓述詞，後接述詞組的狀態及物述詞。（依照主語論旨角色或論元個數的不同分為四類）

VL1 以經驗者（experiencer）為主語，終點（goal）為賓語的二元述詞，表心靈狀態而其語意多表「意願」。例：樂於、甘願。

VL2 以客體（theme）為主語，終點（goal）為賓語的二元述詞。其語意多表「專門」之意。例：擅長、專門、擅於。

VL3 不帶主語的二元述詞，後接一個終點（goal）和一個表客體（theme）的述詞組論元，例：輪、該。 其中表客體的述詞組中賓語部分常常會移到輪、該等主要述詞前面的位置。

VL4 使役述詞，帶肇始者（causer）、終點（goal）、客體角色（theme）的三元述詞。例：使、讓。

二、體詞（N），體詞通常出現在主語或賓語的位置。（依其語意、作用分八類）

Na 名詞（下分五類）

Naa 物質名詞，是不可數的實體名詞。例：泥土、鹽、水、牛肉。

Nab 個體名詞，是可數的實體名詞，可受個體量詞修飾。

例：桌子、杯子、衣服、刀。

Nac 可數抽象名詞，是可數的非實體名詞。例：夢、符號、話、原因。

Nad 抽象名詞，是不可數的非實體名詞。例：風度、香氣、愛心、馬後砲。

Nae 集合名詞：這類名詞　不能指涉個體，只能指涉複數，且　不可以受個體量詞修飾，又依是否受定量式複合詞修飾分二類。

Naea 不能加任何定量式詞組來修飾的集合名詞。
例：三餐、五臟六腑、四肢。

Naeb 可用定量式詞組來修飾。
例：車輛、船隻、夫妻。

Nb 專有名稱（下分兩類）

Nba 正式專有名稱，包含時間、地方以外的專有名稱。
例：吳大猷、余光中、詩經、雙魚座。

Nbc 姓氏。例：張、王、李。

Nc 地方名詞（下分五類）

Nca 專有地方名詞，特指某一地方、行政單位或機構，通常不能用定量式複合詞來修飾。例：西班牙、台北。

Ncb 普通地方名詞，可以用定量式詞組來修飾。

例：郵局、市場、學校、農村。

Ncc 名方式地方名詞。例：海外、身上、腳下。

Ncd 表事物相對位置的地方詞, 大部分由獨用的方位詞或方方式或定方式複合詞構成（下分二小類）。

Ncda 單音節位置詞，其後不能加"的"。例："上"有天堂。

Ncdb 雙音節位置詞。例：上頭、中間、左方、西北。

Nce 定名式地方名詞，例：四海、當地。

Nd 時間名詞（下分三類）

Nda 時間名稱（下分兩類）

Ndaa 歷史性的時間名稱（下分四小類）

Ndaaa 特指的時代名稱。例：洪荒時代、五０年代。
Ndaab 朝代名稱。例：唐朝、西漢。
Ndaac 歷代帝王的年號名稱。例：乾隆、光緒、天寶。
Ndaad 年份名，用以計數年份的紀元。例：公元、西元。

Ndab 可循環重複的時間名稱（下分六小類）

Ndaba 年稱。例：今年是"辛未"年。
Ndabb 季節，即春、夏、秋、冬四季。例：今年"春天"雨水多。
Ndabc 月份名稱。例："十二月"又叫"臘月"。
Ndabd 日期。例：三月"六日"、冬至。
Ndabe 日以內的時間名稱。例：傍晚、大清早。
Ndabf 時期，指一段時間。例：寒假、年假、春節。

Ndc 名方式時間名詞，由附著語位的時間成分加上方位詞複合而成。
例：年底、週末、日後。

Ndd 副詞性時間詞（以下分三類）

Ndca 表過去的副詞性時間詞。例：過去、從前、當初。

Ndcb 表將來的副詞性時間詞。例：以後、後來、將來。

Ndcc 表現在及其他的副詞性時間詞。例：現在、當今、眼前、近來。

Ne 定詞，用以表示物品的指涉或物品的數量。例：這、哪、少許。

Nf 量詞，用以計量的連用語位，常和定詞構成定量式詞組。

Nfa 個體量詞，表示每一個名詞所屬的專門單位詞。
例：一"張"桌子、一"個"杯子、一"件"衣服、一"把"刀子。

Nfb 跟述賓式合用的量詞，放於述詞與賓語之間。
例：下一"盤"棋、寫一"手"好字、說一"口"標準國語。

Nfc 群體量詞，語義上能標示出一組或一群的物體。
例：一"對"夫妻、一"雙"筷子、一"副"耳環、一"群"鴨子。

Nfd 部分量詞，表示事物的部分而非整體的概念。
例：一"部分"原因、一"節"甘蔗、一"段"文章、一"點"事情。

Nfe 容器量詞，用器皿式的名詞來作量詞，表示概括性的容量。
例：一"箱"書、一"櫃子"衣服、一"盤"水梨、一"碗"飯。

Nff 暫時量詞，是以名詞作量詞，加在定詞後面。
例：一"肚子"牢騷、一"頭"秀髮、一"鼻子"灰、一"地"落葉。

Nfg 標準量詞，是正規的量詞，為名副其實的量詞。包括：
長度單位。例：尺、寸、丈。
面積單位。例：頃、畝。
重量單位。例：公斤、磅。
容量單位。例：升、斗。
時間單位。例：分、秒、時。
錢幣單位。例：元、法郎、先令。
數量單位。例：刀、令。
能量單位。例：馬力、燭光、卡路里。

Nfh 準量詞，由名詞轉化而來的單位化量詞，是獨立自主的，它不是後頭名詞的量詞。例：國、面、撇。

Nfi 述詞用量詞，是動作述詞的量詞，表示動作發生的次數。
例：看一"遍"、摸一"下"。

Ng 後置詞。它是一個附著成分，前接一個詞組形成時間成分或表情況的成分。例：睡覺"之前"、夜"裡"、三百人"以上"。

Nh 代名詞（下分三類）

Nha 人稱代名詞（下分三小類）

Nhaa 常用的人稱代名詞，是我、你、他及其複數式、同義詞。

Nhab 一般的人稱代名詞，可與第一、二、三人稱同位並列。例：自己。

Nhac 特別的人稱代名詞，有所專指的代名詞。例：您、足下、令尊、本人、賤內、小犬。

Nhb 疑問代名詞，包括誰、什麼及其別體甚麼、啥等。

Nhc 泛指代名詞，可通用於人、物的代名詞。例：之、其。

三、介詞（**P**），用以引介一個角色，作述詞的修飾成分或必要論元。（依介詞組所可能表示的角色、介詞對其論元之語意及語法限制的不同分為六十五類）

四、副詞（**D**），主要用作謂語的修飾語。（依語意下分十類）

Da 表範圍和數量的副詞。例:只、僅僅。

Db 表示評價的副詞。（下分三類）

Dba 法相副詞。例:也許、大概、一定。

Dbaa 推測用法。例:也許、大概、可能、一定。

Dbab 義務用法。例:必須、可以、得。

Dbb 表示說話者的評斷的副詞。例:幸虧、果然。

Dbc 由"-起來"與述詞組成的評價詞。例:這條路"看起來"很平直。

Dc 表否定的副詞。包括:未、沒有、沒、不。

Dd 時間副詞。例:先、立刻。

Df 程度副詞。（下分兩類）

Dfa 述詞前程度副詞。例:很、非常。

Dfb 述詞後程度副詞。例:得很、之至。

Dg 地方副詞。例:處處、到處。

Dh 方式副詞。例:逐一、從頭、一起。

Di 標誌副詞，幾乎都緊接在述詞之後，表現時態。例:過、著。

Dj 疑問副詞。例:為什麼、幹麼。

Dk 句副詞，有轉變或連接語氣的功能。例:總之、據說。

五、連接詞（C），用以表示並列關係或標明兩分句關係的詞。（依連接成分組合關係的不同下分兩類）

Ca 並列連接詞，連接兩個詞性相似的成分形成向心式結構，其中每一個成分的功能都跟整個結構相同。（下分兩類）

Caa 這類連接詞多半同時具有介詞的特性。例：和、跟。

Cab 連接兩個同類的成分,後一成分常可省略。包括：等、等等、之類。

Cb 關聯連接詞，能夠把幾個分句連成複句形式的連接詞。（下分三類）

Cba 移動性前繫連接詞，語意上具起頭作用，後面常須接一個分句，其所在分句可能移位至複句的後半段。（下分兩類）

Cbaa 偏正句移動性連接詞。例:雖然、因為、即使、只有。

Cbab 偏正句句尾連接詞。這一類只有"的話"和"起見"。

Cbb 非移動性前繫連接詞。語意上具起頭作用，後面常須接一個分句，位置固定在前一分句。（下分兩類）

Cbba 偏正句非移動性前繫連接詞。例:雖、既、就是。

Cbbb 聯合句前繫連接詞。例:不但、一來。

Cbc 後繫連接詞，能將一個分句聯繫於前一個句子的連接詞。（下分兩類）

Cbca 偏正句後繫連接詞。例:可是、所以、那麼、否則。

Cbcb 聯合句後繫連接詞。例:而且、二來。

六、語助詞（**T**），附加於詞組或句子後的連用詞。（依語助詞間共存的次序分為四類）

Ta 了、 的。

Tb 沒、沒有、而已、罷了、也好、也罷、云云、等等、之類、爾爾、來哉、著。

Tc 啊、呀、哇、哪、吶、呢、哩、喲、唷、嘛、嘿、麼、哦、喔、嘔、詼、耶、囉、嘍、吧、罷、啦、咧。

Td 了嗎、了否、而已嗎、啦云云、咧云云、嗎、否、不、與否、哉、耶、矣、唄。

如果有一個以上的語助詞一起出現，其先後的順序依序為：Ta，Tb，Tc。Td 不與前三類共存。

七、感歎詞（**I**），表示說話者的口氣或態度的獨用語式。例:啊、喂、唉。

八、非謂形容詞（**A**），是純粹的形容詞，不具謂語作用。例:公共、共同。

附錄三、中央研究院平衡語料庫 WWW 版檢索系統使用說明

歡迎您進入
「中央研究院平衡語料庫」
（Sinica Corpus）
檢索系統

網址：**http://www.sinica.edu.tw/SinicaCorpus**

「中央研究院平衡語料庫」是第一個開放使用的現代漢語語料庫。所謂語料庫顧名思義就是存放許多文句的資料庫，透過檢索介面，便可對文句資料進行各種搜尋統計的工作。

這個語料庫是專門針對語言分析而設計的，是現代漢語無窮多的語句中一個代表性的樣本。

本語料庫適合從事語文分析的人士使用，像是語言學家、語言心理學家、語文學系的教授與學生、辭典或語文教材編輯人員…等。

在使用此語料庫以前建議您先參看「使用說明」。

「中央研究院平衡語料庫」（Sinica Corpus）
檢索系統使用說明

版權聲明

中央研究院擁有「中研院平衡語料庫」的相關智慧財產權，包括介面程式、語料組成方式、詞類標記、分詞標準及詞彙集等。使用者僅能使用介面系統檢索資料以及利用檢索結果進行研究，但是不得隨意擷取、修改、出版檢索結果。語料的版權仍歸原著作人或提供單位，他人不得轉載、抄襲或侵犯任何語料的智慧財產權。

使用限制

中研院語料庫WWW版所有功能均開放使用，但為防主機資源耗用過劇及顧及資料傳輸之實際限制，暫以檢索結果為限制的條件：院內檢索限兩萬行資料，院外檢索限二千行資料。

「中央研究院平衡語料庫」（Sinica Corpus）簡介

「中央研究院平衡語料庫」是專門針對語言分析而設計的，每個文句都依詞斷開，並標示詞類。語料的蒐集也盡量做到平衡分配在不同的主題和語式上，是現代漢語無窮多的語句中一個代表性的樣本。

這個語料庫是由中央研究院詞庫小組完成的。該小組由陳克健（資訊所）、黃居仁（語言所籌備處）兩位教授主持，自一九九零年前後便開始致力於漢語語料的蒐集。一九九一年得到蔣經國基金會補助，開始構建語料庫；並於一九九四年分別得到中央研究院「中文資訊」跨所研究群專案計畫及國科會計畫補助，正式開始進行語料標記。一九九五年七月完成第一版（兩百萬詞），一九九六年十一月開放WWW版供各界使用。並於一九九七年完成 3.0 版，約五百萬詞。

如欲更進一步了解語料庫的內容，請參考中央研究院詞庫小組所編技術報告第 95-02/98-04 號「中央研究院平衡語料庫的內容與說明」。技術報告及其他參考資料之取得請參看詞庫首頁：

詞庫首頁(http://godel.iis.sinica.edu.tw/CKIP/index.htm)

「中央研究院平衡語料庫」（Sinica Corpus）介面系統使用說明

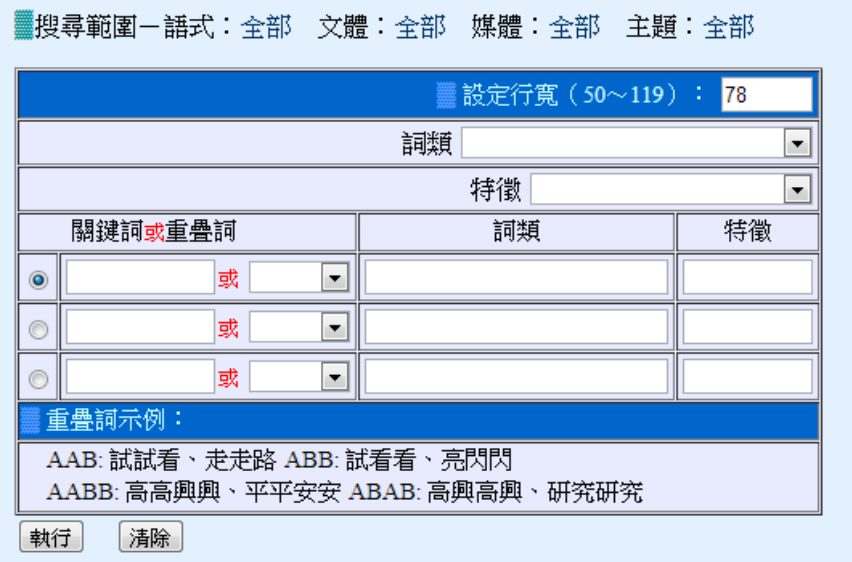
透過本語料庫的介面可以進行下列幾項工作：一、檢索：檢索詞項、檢索詞頭詞尾、檢索詞類…等；二、顯示：將檢索到的資料依句顯示在螢幕上；三、過濾：依照使用者設定的條件篩選語料；四、詞類累計：統計每個詞類出現的數量；五、統計共現率（collocation）；六、排序：針對使用者設定的條件將語料依序排列。

本使用說明共分三個部份：**「主畫面」使用說明**

「再處理」使用說明

「顯示畫面」使用說明

「主畫面」使用說明



本介面僅提供檢索功能：在語料庫中將所有包含檢索詞的文句挑出並顯示。檢索的方式分兩種：一、單項條件檢索方式：一次設定一項檢索條件；二、多項條件檢索方式：一次設定多項檢索條件，包括「and 檢索條件」，以及利用「續設條件」設定「or 檢索條件」。檢索的對象分四種：1. 關鍵詞；2. 重疊詞；3. 詞類；4. 特徵。

一、單項條件檢索：一次設定一項檢索條件

1. 關鍵詞： 將滑鼠移到「**關鍵詞**」的框框內，鍵入欲搜尋的關鍵詞，再將滑鼠移到「**執行**」按下。

關鍵詞可以由下列幾種符號組合而成：

中文字

？ ：表示一個任意字元

＊ ：表示零至無限多個任意字元

範例：鍵入「電話」，會搜盡包含「電話」的文句。

鍵入「電＊」，會搜盡包含以「電」開頭的詞（單字詞、雙字詞、多字詞都包括在內，如：電、電話、電視機）的文句。

鍵入「電？」，會搜盡包含以「電」開頭的雙字詞（如：電話、電腦、電子）的文句。

鍵入「＊電」，會搜盡包含以「電」結尾的詞（單字詞、雙字詞、多字詞都包括在內，如：電、觸電、無線電）的文句。

鍵入「電？？」，會搜盡包含以「電」開頭的三字詞（如：電擊棒、電療法、電纜車）的文句。

鍵入「＊電＊」，會搜盡出現過含有「電」的詞（單字詞、雙字詞、多字詞都包括在內，如：電、電扇、靜電、電壓計、無線電）的文句。

鍵入「？電？」，會搜盡包含將「電」置於中央的三字詞（如：心電圖、手電筒）的文句。

鍵入「？電＊」，會搜盡包含將「電」置於第二字的詞（雙字詞、多字詞都包括在內，如：充電、核電廠）的文句。

鍵入「？？？？」，會搜盡含有任何四字詞的文句。

可以隨時按**「清除」**欄，清除方才所設定的資料。

2. 重疊詞：將滑鼠移到**「重疊詞」**的框框內，鍵入欲搜尋的重疊詞種類，再將滑鼠移到**「執行」**按下。

或將滑鼠移到「重疊詞」框框旁的箭頭按一下，即出現四種重疊詞種類，在所欲搜尋的重疊詞種類按一下，再將滑鼠移到**「執行」**按下。

重疊詞種類共分四種，如介面所示：

重疊詞ＡＡＢ　　如：試試看、走走路

重疊詞ＡＢＢ　　如：試看看、亮閃閃

重疊詞ＡＡＢＢ　　如：高高興興、平平安安

重疊詞ＡＢＡＢ　　如：高興高興、研究研究

範例：鍵入「ＡＡＢ」，會搜盡含有任何ＡＡＢ型重疊詞的文句。

鍵入「ＡＡＢＢ」，會搜盡含有任何ＡＡＢＢ型重疊詞的文句。

可以隨時按**「清除」**欄，清除方才所設定的資料。

3. 詞類：將滑鼠移到**「詞類」**的框框內，鍵入欲搜尋的詞類，再將滑鼠移到**「執行」**按下。

或者將滑鼠移到**「詞類選單」**框框旁的箭頭，按一下，即出現四十六種詞類，在所欲搜尋的詞類按一下，再將滑鼠移到**「執行」**按下。

詞類共分四十六種，一律以英文符號表示，如介面「詞類選單」所示。如欲對詞類更進一步瞭解，請參看附錄一，或中央研究院詞庫小組所編技術報告第 95-02 號「中央研究院平衡語料庫的內容與說明」以及技術報告第 93-05 號「中文詞類分析」。

在詞類欄可鍵入表示一個詞類的英文字串，或是表示一個總類的英文字元。

範例：鍵入「ＶＡ」，會搜盡含有任何ＶＡ類詞（動作不及物動詞）的文句。

鍵入「V*」，會搜盡含有任何Ｖ類詞（包含所有以Ｖ開頭的詞類：ＶＡ、ＶＢ、ＶＣ、ＶＤ、ＶＥ、ＶＦ、ＶＧ、ＶＨ、ＶＩ、ＶＪ、ＶＫ、ＶＬ，即所有的動詞）的文句。

鍵入「Ｎｅ*」，會搜盡含有任何Ｎｅ類詞（包含所有以Ｎｅ開頭的詞類：Ｎｅｐ、Ｎｅｑａ、Ｎｅｑｂ、Ｎｅｓ、Ｎｅｕ，即所有的定詞）的文句。

可以隨時按**「清除」**欄，清除方才所設定的資料。

4. 特徵：將滑鼠移到**「特徵選單」**框框旁的箭頭按一下，即出現九種特徵，在所欲搜尋的特徵按一下，再將滑鼠移到**「執行」**按下。

特徵共九種，一律以英文符號表示，如介面「特徵選單」所示。如欲對特徵更進一步瞭解，請參看附錄二，或中央研究院詞庫小組所編技術報告第 95-2 號「中央研究院平衡語料庫的內容與說明」。

在特徵一欄可以鍵入表示一個特徵的英文字串。

範例：鍵入「vrv」，會搜盡含有任何帶有「vrv」（動補動詞中的動詞成分）的詞的文句。

可以隨時按**「清除」**欄，清除方才所設定的資料。

二、多項條件檢索：一次設定兩項以上檢索條件

1.「and 檢索條件」：每一項條件都符合才檢索出來

「關鍵詞/重疊詞」、**「詞類」**、**「特徵」**這三種條件可以同時設定，會檢索出同時符合三項條件的詞。

範例：在「關鍵詞」欄鍵入「？？」，在「詞類」欄鍵入「ＶＣ」，在「特徵」欄鍵入「vrr」，會搜盡含有任何雙音節帶有「vrr」特徵標記的ＶＣ類詞（動作及物

動詞）的文句。

可以隨時按「**清除**」欄，清除方才所設定的資料。

2.「or 檢索條件」：一次檢索兩個以上的對象

「續設條件」：可以一次檢索多個對象。將第一個對象的各項條件設定完成後，將滑鼠移到**「續設條件」**前的框框按下，框內顯示叉號，再移到**「執行」**按下，主畫面會重新跳出；然後再設定第二個對象的各項資料。如此步驟可以一再重複，「續設條件」可以連續使用十次以下。

範例：

在「關鍵詞」欄鍵入「高興」，到「續設條件」按一下，再按「執行」；

回「關鍵詞」欄鍵入「快樂」，到「續設條件」按一下，再按「執行」；

回「關鍵詞」欄鍵入「開心」，到「續設條件」按一下，再按「執行」；

回「關鍵詞」欄鍵入「歡喜」，到「續設條件」按一下，再按「執行」；

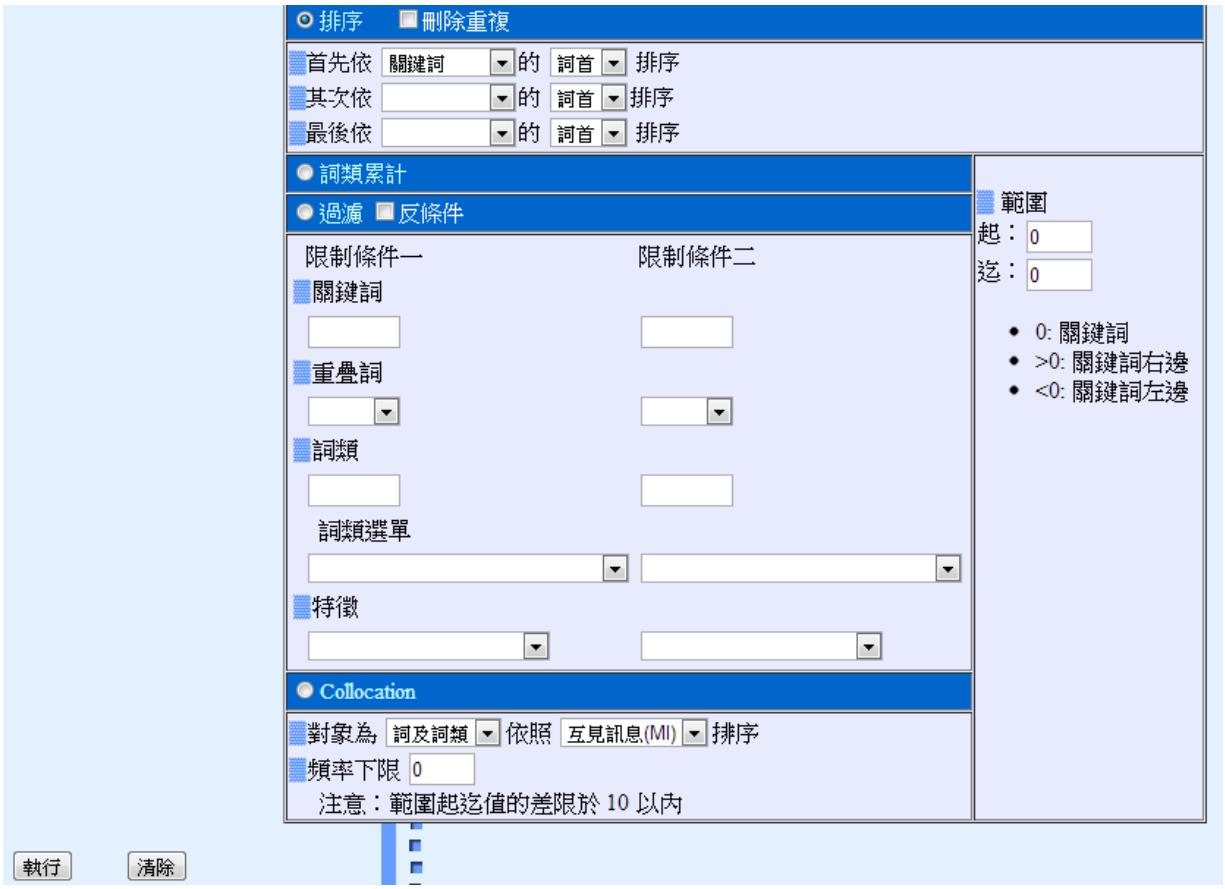
會搜盡含有「高興」、「快樂」、「開心」、「歡喜」這四個詞之一的所有文句。

可以隨時按**「清除」**欄，清除方才所設定的資料。

三、查詢使用說明

隨時將游標移至帶著問號標誌的求助框按下，即進入線上使用說明。使用說明包含了：版權聲明、使用限制、語料庫簡介、檢索系統使用簡介、「主畫面」使用說明、「再處理」使用說明、顯示畫面使用說明、畫面轉換使用說明、附錄一：詞類標記表、附錄二：特徵標記表。

「再處理」使用說明



本介面提供一些功能將檢索出來的資料作處理，而且處理出來的語料可以再次處理，形成一層又一層的語料。

本介面所提供的功能有以下五種：一、過濾：檢索出來的資料可能很龐雜，可以利用本功能作更進一步的篩選；二、詞類累計：若想對語料的語法特性作更進一步的觀察，可以利用此功能統計每種詞類出現的次數；三、collocation 統計：本功能提供統計數據來表示關鍵詞和前後的詞或詞類一起出現的機率；四、排序：檢索出來的資料是依照在語料庫中的次序來排列，透過排序功能可以調整語料排列順序，更方便觀察或比較；五、畫面切換：可以隨時跳換到主畫面重新檢索，或進入每一層語料顯示畫面。

每一次跳入本介面都是以「過濾」為預設功能。

一、過濾：以本次檢索結果為範圍，依照設定條件和範圍作過濾。

條件設定： 和檢索條件相同，包含**詞**、**重疊詞**、**詞類**、**特徵**四種條件。請參看「主畫面使用說明」。

反條件：可以依照設定的條件「挑出」語料或依照設定的條件「去除」語料。

預設條件是依照設定的條件「挑出」語料。如果要依照設定的條件「去除」語料，則將滑鼠移向**「反條件」**前的方框框按下，框框內出現叉號即可。

範圍設定：過濾可以針對關鍵詞本身進行，也可以針對關鍵詞的前後幾個詞作。範圍可大可小，在**「範圍 起：」**以及**「範圍 迄：」**兩欄設定。預設值為關鍵詞本身。若果要放寬過濾範圍，將滑鼠移到「範圍 起：」以及「範圍 迄：」右方的框框內，輸入數字（0 表示關鍵詞本身，-1 表示關鍵字左方一個詞，+1 表示關鍵字右方一個詞，範圍以不超過十個詞為限）。

可以同時給兩種不同的過濾條件，分別在**「條件一」**和**「條件二」**設定。

如果設定的範圍是「0」到「0」，即關鍵詞本身，則直接在**「條件一」**設定關鍵詞本身的條件。

如果設定的範圍是「-x」到「0」，則**「條件一」**是關鍵詞左邊諸詞的條件，而**「條件二」**是關鍵詞本身的條件。

如果設定的範圍是「0」到「x」，則**「條件一」**是關鍵詞本身的條件，而**「條件二」**是關鍵詞右邊諸詞的條件。

如果設定的範圍是「-x」到「y」，則**「條件一」**是關鍵詞左邊諸詞的條件，而**「條件二」**是關鍵詞右邊諸詞的條件。

條件和範圍設定好了之後，將滑鼠移到**「執行」**按下。

可以隨時按**「清除」**欄，清除方才所設定的資料。

二、詞類累計：以本次檢索結果為範圍，統計關鍵詞及其前後語境內不同詞類出現次數。

將滑鼠移到**「詞類累計」**前的圈圈，按一下。

範圍設定：詞類累計可以針對關鍵詞本身進行，也可以針對關鍵詞的前後幾個詞作。範圍可大可小，在**「範圍 起：」**以及**「範圍 迄：」**兩欄設定。預設值為關鍵詞本身。若果要放寬或改變語境範圍，將滑鼠移到「範圍 起：」以及「範圍 迄：」右方的框框內，輸入數字（0 表示關鍵詞本身，-1 表示關鍵字左方一個詞，+1 表示關鍵字右方一個詞，範圍以不超過十個詞為限）。

範圍設定好了之後，將滑鼠移到**「執行」**按下。

可以隨時按**「清除」**欄，清除方才所設定的資料。

三、collocation 統計：以本次檢索結果為範圍，計算關鍵詞與其語境內所含詞或詞類間的共現率（collocation）。共現率是以互見訊息值（ＭＩ值：mutual information value）為準則。互見訊息值表示兩個單位同時出現的機率，值越高，表示共現率越高；反之，值愈低，表示共現率越低。

※collocation 統計只有在沒有過濾任何資料的清況下才能做出有效數值。所以作 collocation 統計前，不能進行過濾的工作。因此若要作 collocation，必須在「主畫面」就指定所有限制。

將滑鼠移到「**collocation**」前的圈圈，按一下。

條件設定：包含「詞及詞類」、「詞」、「詞類」三種條件。

「**詞及詞類**」視扮演不同詞類的同一個詞為不同的統計單位。

「**詞**」表示只統計詞和關鍵詞之間的互見訊息值，相同的詞但是扮演不同詞類也視為同一個統計單位。

「**詞類**」表示只統計詞類和關鍵詞之間的互見訊息值，相同的詞類但是詞不同也視為同一個統計單位。

範圍設定： collocation 統計是針對關鍵詞和關鍵詞的前後幾個詞作。範圍不可只設定在關鍵詞（即「範圍 起：」以及「範圍 迄：」兩欄不可以都設定為0），也不可以不包括關鍵詞（即「範圍 起：」和「範圍 迄：」兩欄所設數值必須橫跨 0）。範圍在「**範圍 起：**」以及「**範圍 迄：**」兩欄設定。預設值為關鍵詞本身，一定要更動。將滑鼠移到「範圍 起：」以及「範圍 迄：」右方的框框內，輸入數字（0 表示關鍵詞本身，-1 表示關鍵字左方一個詞，+1 表示關鍵字右方一個詞，範圍以不超過十個詞為限）。

排序設定：統計的結果可以依照「**互見訊息值**」排列或依照「**詞頻**」排列。

頻率下限：設定至少出現幾次以上的詞才作 collocation 統計。預設值為兩次。

條件、範圍和排序都設定好了之後，將滑鼠移到「**執行**」按下。

可以隨時按「**清除**」欄，清除方才所設定的資料。

ＭＩ的計算：$I(x,y) = \log P(x,y)/P(x)P(y)$

$$= \log \frac{f(x,y)/N}{f(x)/N \cdot f(y)/N}.$$

Ｉ：mutual information

Ｐ：probability

Ｎ：size of the corpus

freq（x）：關鍵詞在整個語料庫中出現的次數

freq（y）：該單位在整個語料庫中出現的次數

freq（x，y）：關鍵詞和該單位在本次範圍內出現的次數

ＭＩ的意義：如果ＭＩ大於零，表示關鍵詞和該單位在所設定的範圍寬度間傾向一起出現，值越大，「共現率」越高。

如果ＭＩ小於零，表示關鍵詞和該單位在所設定的範圍寬度間傾向「不」一起出現，負值越大，「互斥率」越高。

四、排序：以本次檢索結果為範圍，依照設定對象及依據將資料依次排列。

將滑鼠移到**「排序」**前的圈圈，按一下。

設定對象：排序的對象只有三種：關鍵詞、關鍵詞左邊、關鍵詞右邊。

關鍵詞：依照關鍵詞排序。

關鍵詞左邊：依照關鍵詞左邊第一詞排序。

關鍵詞右邊：依照關鍵詞右邊第一詞排序。

設定對象可以有三項，並有先後順序，分別鍵入**「首先依」**、**「其次依」**、**「最後依」**右方的框框內。

排序依據：排序的依據只有三種：詞首、詞尾、詞類。

詞首：依照設定對象的詞首排序。

詞尾：依照設定對象的詞尾排序。

詞類：依照設定對象的詞類排序。

取消重複：依設定對象和依據排序後，取消上下行重複的資料。

範例：

如果設定「依關鍵詞的詞首/尾排序」，則相同的關鍵詞會排在一起。「取消重複」若設定，則相同的關鍵詞只取一筆；

如果設定「依關鍵詞的詞類排序」，則關鍵詞帶有相同詞類會排在一起。「取消重複」若設定，則關鍵詞每一種詞類都只取一筆；

如果設定「首先依關鍵詞的詞首/尾排序」及「其次依關鍵詞右邊的詞首/尾排序」，則相同的關鍵詞且右邊詞也相同會排在一起。「取消重複」若設定，則每一組排在一起的相同資料只取一筆；

如果設定「依關鍵詞的詞類排序」及「依關鍵詞右邊的詞類排序」，則帶有相同詞類的關鍵詞及右邊詞都帶有相同詞類會排在一起。「取消重複」若設定，

則每一組排在一起的相同資料只取一筆；

如果設定「依關鍵詞的詞首/尾排序」及「依關鍵詞右邊的詞類排序」，則相同關鍵詞且右邊詞帶有相同詞類會排在一起。「取消重複」若設定，則每一組排在一起的相同資料只取一筆。

對象和依據都設定好了，也決定要不要取消重複之後，將滑鼠移到**「執行」**按下。

可以隨時按**「清除」**欄，清除方才所設定的資料。

五、畫面切換

第一次由主畫面檢索出來的語料屬於第一層語料。如果再處理將該語料縮減，則會形成第二層語料。第二層語料還可以再處理，形成第三層語料。以此類推，最多可以記錄十層語料。

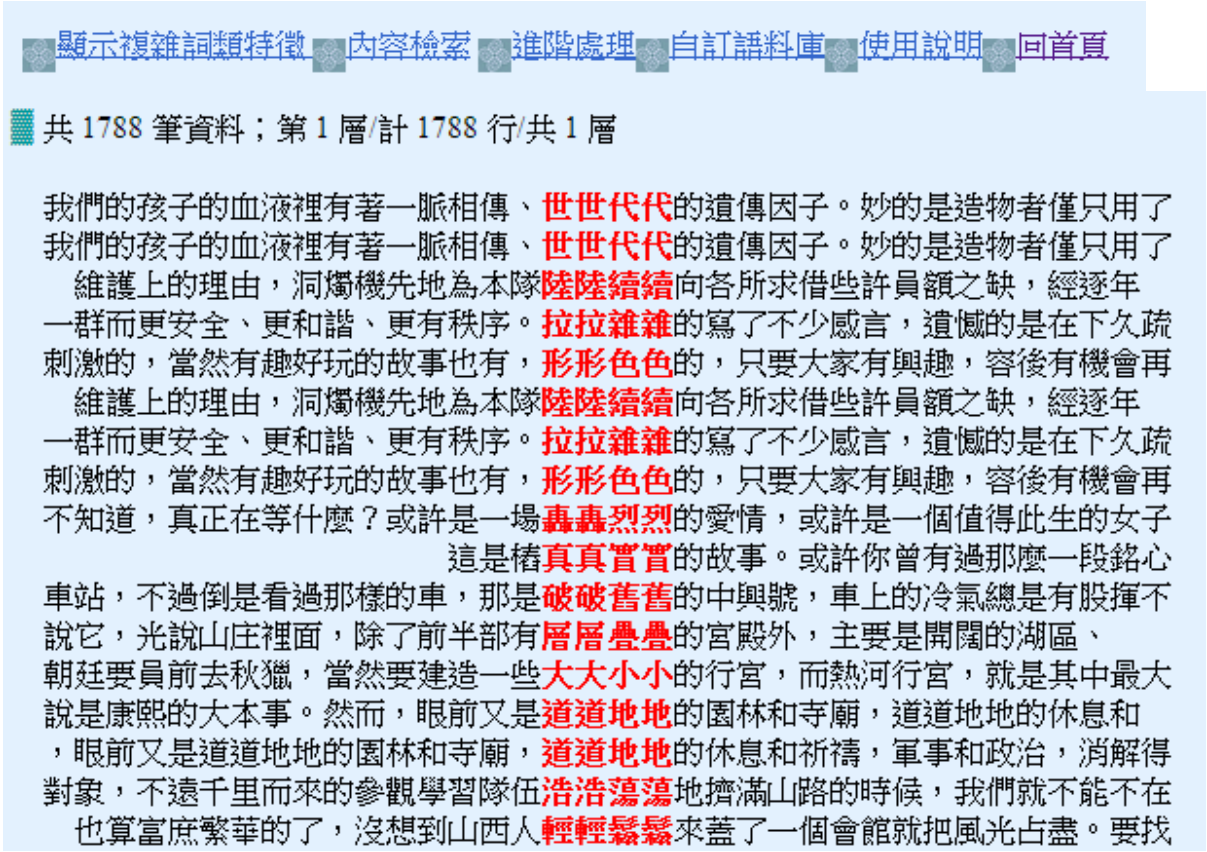
畫面上方會有**「主畫面」**。將滑鼠移到「主畫面」按一下，就會進入主畫面的介面。進入「主畫面」表示要重新檢索資料，目前的語料將全部取消。

畫面上方有**「上層結果」**、**「本層結果」**或**「下層結果」**。將滑鼠移到「上層結果」按一下，就會進入上層的語料顯示畫面；將滑鼠移到「本層結果」按一下，就會進入這一層語料顯示畫面；將滑鼠移到「下層結果」按一下，就會進入下層的語料顯示畫面。

六、查詢使用說明

隨時將游標移至帶著問號標誌的求助框按下，即進入線上使用說明。使用說明包含了：版權聲明、使用限制、語料庫簡介、檢索系統使用簡介、「主畫面」使用說明、「再處理」使用說明、顯示畫面使用說明、畫面轉換使用說明、附錄一：詞類標記表、附錄二：特徵標記表。

「顯示畫面」使用說明



顯示畫面分兩種：語料顯示畫面及統計資料顯示畫面。語料顯示畫面提供下列三種功能：一、選擇文句包含詞類或不包含詞類；二、畫面切換；三、層次說明。統計資料顯示畫面則提供後兩種功能。

一、選擇文句包含詞類及特徵或不包含詞類及特徵

只有語料顯示畫面有這項功能。將滑鼠移到畫面上方「**顯示詞類及特徵**」按一下，畫面上每一個詞後即會顯示詞類及特徵標記。若要消除詞類及特徵標記，同樣地，也是將滑鼠移到畫面上方「**取消詞類及特徵**」按一下，畫面上每一個詞後的詞類及特徵標記就會消失。

二、畫面切換

畫面上方有「**主畫面**」、「**再處理**」。將滑鼠移到「主畫面」按一下，就會進入主畫面的介面。進入「主畫面」表示要重新檢索資料，目前的語料將全部取消；將滑鼠移到「再處理」按一下，就會進入再處理的介面。進入「再處理」表示要根據目前的語料再作處理。

畫面上方可能有「**上層結果**」、「**本層結果**」或「**下層結果**」。將滑鼠移到「上層結果」按一下，就會進入上層的語料顯示畫面；將滑鼠移到「本層結果」按一下，就會進入本層的語料顯示畫面；將滑鼠移到「下層結果」按一下，就會進入下層的語料顯示畫面。語料最多可以有十層，利用「**上層結果**」、「**下層結果**」可以連續向上層或向下層切換。各語料的分層方式，請參看下一節「層次說明」。

三、層次說明。

畫面上方有（**第__層/計_____行/共__層**）的訊息，是用來說明該次顯示畫面是屬於第幾層語料、共含有幾筆資料、整個過程中一共建立了幾層語料。

第一次由主畫面檢索出來的語料屬於第一層語料。如果再處理將該語料縮減，則會形成第二層語料。第二層語料還可以再處理，形成第三層語料。以此類推，最多可以記錄十層語料。

附錄一：「中研院平衡語料庫」詞類標記表

簡化標記	對應的 CKIP 詞類標記	
A	A	/*非謂形容詞*/
Caa	Caa	/*對等連接詞，如：和、跟*/
Cab	Cab	/*連接詞，如：等等*/
Cba	Cbab	/*連接詞，如：的話*/
Cbb	Cbaa, Cbba, Cbbb, Cbca, Cbcb	/*關聯連接詞*/
D	*Dab*, Dbaa, Dbab, Dbb, Dbc, Dc, Dd, Dg, Dh, Dj	/*副詞*/
Da	*Daa*	/*數量副詞*/
DE		/*的, 之, 得, 地*/
Dfa	Dfa	/*動詞前程度副詞*/
Dfb	Dfb	/*動詞後程度副詞*/
Di	Di	/*時態標記*/
Dk	Dk	/*句副詞*/
FW		/*外文標記*/
I	I	/*感嘆詞*/
Na	Naa, Nab, Nac, Nad, Naea, Naeb	/*普通名詞*/
Nb	Nba, Nbc	/*專有名稱*/
Nc	Nca, Ncb, Ncc, Nce	/*地方詞*/
Ncd	Ncda, Ncdb	/*位置詞*/
Nd	Ndaa, Ndab, Ndc, Ndd	/*時間詞*/
Nep	*Nep*	/*指代定詞*/
Neqa	*Neqa*	/*數量定詞*/
Neqb	*Neqb*	/*後置數量定詞*/
Nes	*Nes*	/*特指定詞*/
Neu	*Neu*	/*數詞定詞*/.
Nf	Nfa, Nfb, Nfc, Nfd, Nfe, Nfg, Nfh, Nfi	/*量詞*/
Ng	Ng	/*後置詞*/
Nh	Nhaa, Nhab, Nhac, Nhb, Nhc	/*代名詞*/
P	P*	/*介詞*/
SHI		/*是*/
T	Ta, Tb, Tc, Td	/*語助詞*/
VA	VA11,12,13,VA3,VA4	/*動作不及物動詞*/
VAC	VA2	/*動作使動動詞*/
VB	VB11,12,VB2	/*動作類及物動詞*/
VC	VC2, VC31,32,33	/*動作及物動詞*/
VCL	VC1	/*動作接地方賓語動詞*/
VD	VD1, VD2	/*雙賓動詞*/
VE	VE11, VE12, VE2	/*動作句賓動詞*/
VF	VF1, VF2	/*動作謂賓動詞*/
VG	VG1, VG2	/*分類動詞*/
VH	VH11,12,13,14,15,17,VH21	/*狀態不及物動詞*/
VHC	VH16, VH22	/*狀態使動動詞/
VI	VI1,2,3	/*狀態類及物動詞*/
VJ	VJ1,2,3	/*狀態及物動詞*/
VK	VK1,2	/*狀態句賓動詞*/
VL	VL1,2,3,4	/*狀態謂賓動詞*/
V_2	V_2	/*有*/

附錄二：中研院平衡語料庫特徵標記表

特徵標記	標 記 定 義	例 子
+fw	the feature of a foreign word	卡拉 OK (Na [+fw])
+nom	the feature for verbal nominalization	他的不講理(VA[+nom])
+p1	the first part of a separated compound	初 Nc[+p1]、高中(Nc)
+p2	the second part of a separated compound	星期六(Nd)、日(Nd[+p2])
+prop	the feature for proper nouns	蘋果(Na[+prop])電腦
+spv	V of a separable V N compound	吃 VC[+spv]了他的虧
+spo	N of a separable V N compound	吃了他的虧 Na[+spo]
+vrv	V of a separable VR compound	叫 VC[+vrv]不醒
+vrr	R of a separable VR compound	叫不醒 VC[+vrr]

相關文獻

詞庫小組，1993，中文詞類分析，中文詞知識庫小組技術報告 # 93-05，南港，中央研究院。

詞庫小組，1996，『搜』文解字:中文詞界研究與資訊用分詞標準，中文詞知識庫小組技術報告#96-01，南港，中央研究院。

陳克健，中文詞知識庫小組，1991，中文詞知識庫計劃與中文電子辭典，中日雙邊資訊研討會論文集，pp.19-37，台灣，台北。

陳克健，1994，素材語言學與文本處理，發表於ICCL-3會議，一九九四年七月，香港。

黃居仁，1995，科際整合與整合科技－談計算語言學與語料庫語言學之角色與發展。「語言學研究之現況與發展」研討會，七月十五日，國立台灣師範大學。

黃居仁，陳克健，陳鳳儀，魏文真，張麗麗 ，1997，資訊用中文分詞規範設計理念及規範內容。*語言文字應用學刊*。第一期。92-100頁。

魏文真，葉美利，莫若萍，1991a，「有」的語法表達模式，民國八十年國科會報告。

魏文真，陳克健，1991b，「是」的語法表達模式，民國八十年國科會報告。

魏文真，陳克健，1991c，連接詞的語法表達模式－以中文訊息格位語法（ICG）為本的表達模式，第四屆計算語言學研討會論文集，pp. 79-95，台灣，屏東（墾丁）。

葉美利，湯志真，黃居仁，陳克健，1992，漢語的動詞名物化初探－漢語中帶論元的名物化派生詞，第五屆計算語言學研討會論文集，pp.177-193，台灣，台北（劍潭）。

張麗麗，黃居仁，1995，漢語數量詞後置，NACCL論文集。

劉源、譚強、沈旭昆，1993，信息處理用現代漢語分詞規範及自動分詞方法。北京，清華大學出版社。

劉興寰，1994，中文語料詞類自動標記，清華大學碩士論文。

Chang, Li-ping and Keh-jiann Chen, 1995. *The CKIP Part-of-speech Tagging System for Modern Chinese Texts*. Proceedings of ICCPOL'95. Hawaii.

Chen, Keh-jiann, Shing-huan Liu, 1992. *Word Identification for Mandarin Chinese Sentences*. Proceedings COLING'92, pp.54-59.

Chen, Keh-jiann, Shing-huan Liu, Li-ping Chang and Yeh-Hao Chin, 1994. *A Practical Tagger for Chinese Corpora*. Proceedings of ROCLING VII, pp.111-126.

Church, K. W. and R. L. Mercer, 1993. *Introduction to the Special Issue on Computational Linguisitcs Using Large Corpora*. Computational Linguistics, Vol.19, No.1, pp.1-24.

Huang, Chu-Ren, Keh-jiann Chen and Li-Li Chang. 1996. *Segmentation Standard for Chinese Natural Language Processing.* Proceedings of the 1996 International Conference on

Computational Linguistics (COLING 96). August. Copenhagan, Denmark.

Huang, Chu-Ren, 1994. *Corpus-based Studies of Mandarin Chinese: Foundational Issues and Preliminary Results*. In Matthew Chen and Ovid Tzeng Eds. In Honor of William S-Y. Wang: Interdisciplinary Studies on Language and Language Change. pp. 165-186. Taipei: Pyramid.

Huang, Chu-Ren, and Ruo-ping Mo, 1992. *Mandarin Ditransitive Constructions and the Category of Gei*. In the Proceedings of the Berkeley Linguistics Society Annual Meeting (BLS 18), pp. 109-122. Berkeley: BLS.

Huang, Chu-Ren and Keh-jiann Chen, 1992. *A Chinese Corpus for Linguistics Research*. In the Proceedings of the 1992 International Conference on Computational Linguistics (COLING-92). pp.1214-1217. Nantes, France.

Huang, Chu-Ren, 1987. Mandarin Chinese NP *de*: A Comparative Study of Current Grammatical Theories. Special Publications No.93 of the Institute of History & Philology, Academia Sinica, Taipei.

Hsu, Hui-li and Chu-Ren Huang, 1995. *Design Criteria for a Balanced Modern Chinese Corpus*. Proceedings of ICCPOL'95, Hawaii.

Kucera, H. and W. N. Francis, 1967. *Computational Analysis of Present-Day American Englis*h. Providence: Brown University Press.

Sproat, R. and Shi C. (1990) *A Statistical Method for Finding Word Boundaries in Chinese Text*, Computer Processing of Chinese & Oriental Languages, Vol. 4.

Svartvik, Jan, 1992. *Ed. Directions in Corpus Linguistics*. Proceedings of Nobel Symposium 82, 4-8 August 1991. Trends in Linguistics Studies and Monographs 65. Berlin: Mouton.