中央研究院平衡語料庫簡介

黃居仁，陳克健，張莉萍，許蕙麗

摘　要

中央研究院平衡語料庫（Academia Sinica Balance Corpus，簡稱研究院語料庫 Sinica Corpus）是第一個有完整詞類標記的中文語料庫。這個語料庫由中央研究院詞知識庫小組蒐集標記完成。其測試版（Sinica 1.0）共計兩百萬詞，將於一九九五年九月公開開放給學術研究界使用。

帶詞類標記的平衡語料庫是計算語言學及語料庫語言學研究必需的資料。但中文一直缺乏這個基礎研究工具。中央研究院平衡語料庫構建的目的即在彌補這個研究基礎架構上的空缺。這個語料庫是以五百萬詞的平衡語料爲目標，第一階段完成二百萬詞。

除了說明語料來源、文體、語式等基本統計計算資料外；重點在解說此平衡語料庫中所採用的分詞標準及標記集（tagset）。分詞標準是採用計算語言學學會的分詞標準，將向中央標準局提出爲資訊用分詞國家標準草案。標記集是根據中研院詞庫小組的詞類分析簡化而成，共有四十六個標記。

一、中央研究院平衡語料庫的構建：動機，源起與設計理念

本文介紹的「中央研究院平衡語料庫」簡稱「研究院語料庫」（Sinica Corpus），是世界上第一個有完整詞類標記的漢語平衡語料庫。研究院語料庫最終目標是要建立五百萬詞的平衡語料庫〔Huang 1994〕。但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），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。研究院語料庫1.0版提供的規模是二百萬目詞。

1.1 建立平衡語料庫的動機

語料庫爲本（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，語料庫爲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爲目標。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爲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。因此爲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 研究院語料庫的源起

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。爲兼顧理想與實用性；我們把初步目標定爲兩百萬詞，爲傳統小規模平衡語料庫之兩倍，而最終目標定爲五百萬詞，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爲優先，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，「女人女人」製作單位，「伴我成長」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）中取得。

1.3 研究院語料庫的設計理念（Design Features）

研究院語料庫因爲中文的特性，也因爲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爲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，在理論上亦復有爭議；故一直很難標準化。目前國內至少有中華民國計算語言學學會的分詞標準〔計算語言學通訊 1992〕，而且在會員中通行，也將透過中央標準局研究計劃成爲「中文資訊用分詞國家標準草案」的藍本。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爲爾後修定國家標準草案時的依據。

(二) 裁文是以文章（text）的自然斷落爲準，而非以文章長度爲準

布朗語料庫的設計特色之一，是爲了求數字上的平衡，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。而且我們認爲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。也因爲如此，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。可是我們認爲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

我們認爲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，因爲影響整個語言全貌的內在因素實在太多了。爲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爲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。比如說一般認爲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眞實現況。

1.4 簡介的內容

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在文中分三節依次說明：

第二節談平衡語料的分類與選取，如何爲語料做分類，分類的標準以及各類的比例。

第三節是中文的斷詞問題，中文基本上以小句爲單位，從來源處得到的資料，並無標示詞的訊息，但是切分詞的正確率也直接影響到詞類標記的準確度。

第四節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。

二、語料的分類與選取：

爲了妥善管理以及選取平衡語料庫的內容，在每篇文章前頭會標示它們的文類、文體、媒體、語式、主題等，如圖一所示。這些屬性是如何得來的呢？將在2.1節說明。

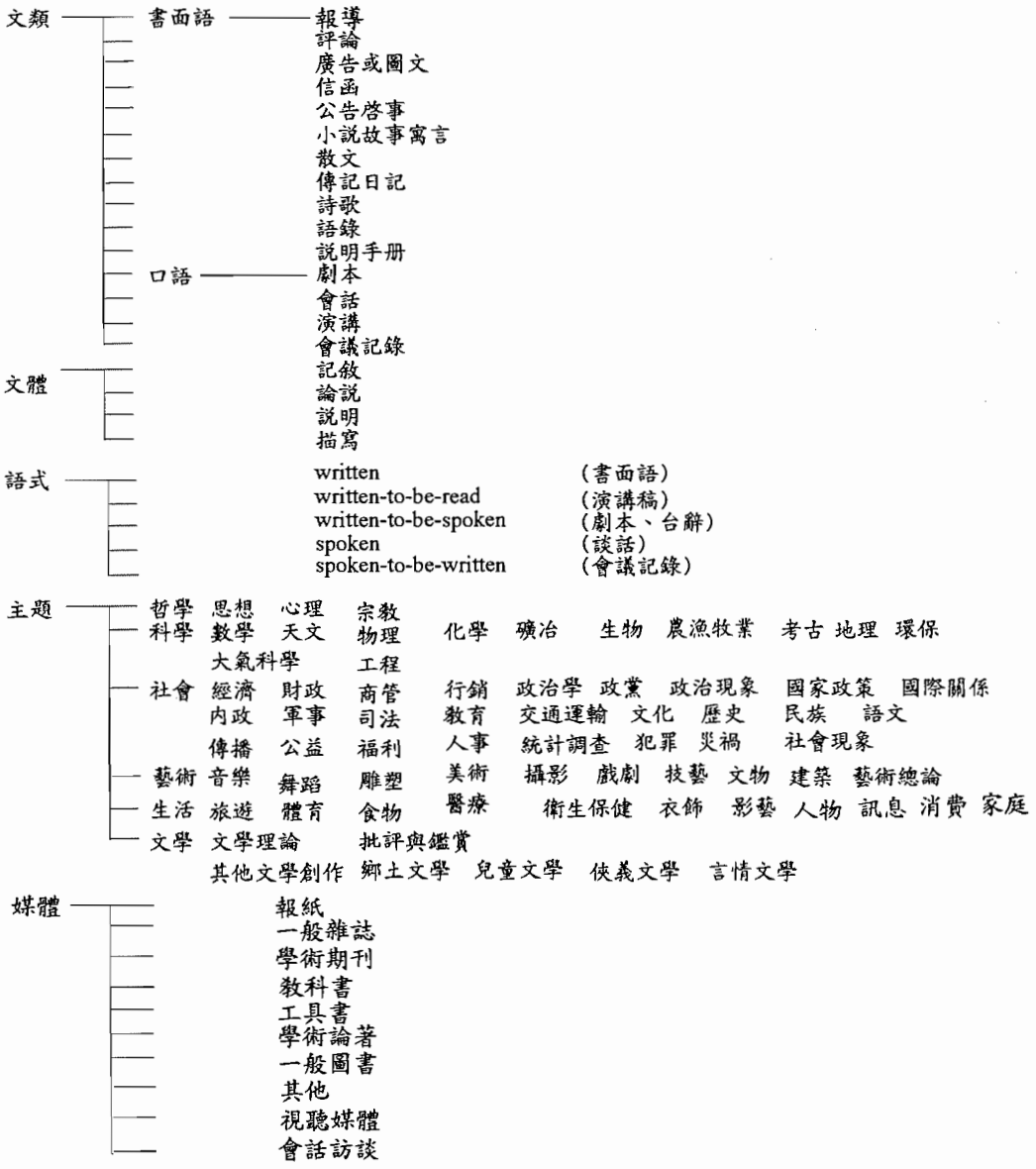
圖一

```
%% 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，
      ⋮
```

2.1 屬性特徵的訂定

我們參考了LANCASTER-OSLO/BERGEN (LOB) CORPUS、美國BROWN 大學布朗語料庫、英國BIRMINGHAM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簡述。

圖 二



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，以賣方立場爲主。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、美希望中共釋放政治犯等。
- 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、郵寄、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：以買方爲出發的報導或評論。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、婚姻的各種問題。

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：特地爲兒童創作或爲兒童寫的作品。
- 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，可分爲報導、評論、廣告圖文、信函、公告啓事、小說故事寓言、散文、傳記日記、詩歌、語錄、說明手册、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，主題爲兒童文學，語式爲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。我們的期刊類分爲報紙、學術期刊、一般雜誌；圖書分爲教科書、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。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；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、洪健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，分爲記敘、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，如描述景物的遊記或哈佛散記之類的也多半是散文。

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。我們把語式分爲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，因爲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，因爲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，以和眞正的口語、書面語區別。

2.2 語料的選取與分佈比例

目前，我們以主題爲準、訂出平衡語料庫的內容比例爲：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，根據此參考值爲基準選取語料。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。爲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200百萬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第1.0版各主題所佔比例統計表 (單位:萬)

主　題 (平衡百分比目標值)	哲學 10%	科學 10%	社會 35%	藝術 5%	生活 20%	文學 20%	總計 100%
現有CORPUS字數總計	29.21	29.05	113.25	14.29	57.91	61.47	305.18
現有CORPUS詞數總計	19.22	19.11	74.51	9.40	38.10	40.44	200.78
實際百分比	9.57%	9.52%	37.11%	4.68%	18.97%	20.14%	100%

表　二

文類	報導	評論	廣告 圖文	信函	公告 啓事	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
CORPUS字數	171.49	29.46	1.60	3.36	0.58	32.83	44.07	2.62	1.70	0.15	4.84	0.43	5.85	5.68	0.51
CORPUS詞數	112.82	19.38	1.05	2.21	0.38	21.60	28.99	1.72	1.12	0.10	3.18	0.28	3.85	3.74	0.34
百分比%	56.19	9.65	0.52	1.10	0.19	10.76	14.44	0.86	0.56	0.05	1.59	0.14	1.92	1.86	0.17

表　三

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
CORPUS字數	215.09	27.52	5.38	8.87	0.68	3.91	31.55	0.80	8.50	2.88
CORPUS詞數	14.15	18.10	3.54	5.84	0.45	2.57	20.76	0.53	5.59	1.90
百分比%	70.48	9.02	1.76	2.91	0.22	1.28	10.34	0.26	2.78	0.94

表 四

文體	記敘文	論說文	說明文	描寫文
CORPUS字數	215.70	38.14	34.35	17.00
CORPUS詞數	141.90	25.09	22.60	11.18
百分比%	70.68	12.50	11.26	5.57

表 五

語式	書面語	演講稿	劇本台詞	談話	會議記錄
CORPUS字數	292.14	5.38	0.46	2.88	4.32
CORPUS詞數	192.20	3.54	0.30	1.90	2.84
百分比%	95.72	1.76	0.15	0.94	1.42

由表三可得知，這個語料庫的內容來源主要來自報紙，佔了百分之七十。詳細的來源及數量，如下所示（以字爲單位）：

1. 報紙	中國時報	500,556
	自由時報	1,258,334
	兒童日報	299,260
	中央研究院計算中心通訊	95,774
2. 一般雜誌	天下雜誌	61,944
	光華雜誌	29,840
	海天遊蹤	138,462
	世界電影雜誌	14,869
3. 學術期刊	中央研究院民族所集刊	11,225
	中央研究院生醫簡訊	39,507
4. 教科書	國民小學國語教科書十二册	88,744
5. 工具書	中研院資訊所詞庫小組的技術報告	6,842
6. 學術論著	論文	39,076
7. 其他	無法歸入其他媒體的檔案	8,011
8. 圖書	洪健全基金會的大眾心理叢書八本	216,721
	時報出版的巴西狂歡節	89,592
9. 視聽媒體	台灣學術網路裡刊登的文章	103,955
10. 會話訪談	民運人士的訪談記錄及大陸留美學生日常會話	28,831

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，還要先爲語料做斷詞工作，唯有一個一個成份非常明確了，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詞典中的九萬目詞爲基礎，切分爲一個一個獨立的詞。沒列在詞典中的成份，則以字爲單位，被一一切分開。然後佐以構詞律對衍生性強的詞綴及數字組合成份進行結合詞彙的工作。

由於中文並沒有一個統一的分詞標準，在我們分詞的過程中，因爲沒有依循的標準而難以規範，什麼情況下字串應該合成一個單位，什麼時候應該分，例如，動詞後接“給、到、於、爲”等像介詞性的成份，究竟該不該合爲一個單位？這些問題經過討論彙整後，反映給中華民國計算語言學學會，期望能訂定出一個中文的分詞標準，讓大家能有效地共享語料庫。因此，目前分詞標準是採用學會向中央標準局提出的「中文資訊用分詞國家標準草案」的原則切分。基本的大原則如下，細則可參見附錄一。

1. 有獨立意義的語法類可依類爲一分詞單位。
2. 慣用的語言成分依人的使用習慣切分。
3. 語意失去組合性，或語法起變化，得合爲一分詞單位。
4. 有明顯的分隔標記時得切分之。
5. 同形異構的成分依實際語境切分。
6. 原則互有衝突時由計算語言學會統籌協議之。

四、詞類標記

分詞工作完成後，接下來是爲每一個成份標記詞類。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，修正斷詞錯誤成份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個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詞典中的178個詞類經簡化後所得到的43個標記〔詞庫小組 1993〕，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，右邊則是相對應的詞典中的詞類，其後都附有簡單說明或例子。至於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。簡化爲46個標記的原則，是去除因語意區別而產生的細類，純粹就語法行爲不同以及和它類詞具區別性功能簡化而來的詞類。

也因爲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。

表六、平衡語料庫詞類標記集

簡化標記	對應的CKIP詞類標記	
Caa	Caa	/*和、跟*/
Cab	Cab	/*等等*/
Cba	Cbab	/*的話*/
Cbb	Cbaa, Cbba, Cbbb	/*可在subj之後*/
Cbc	Cbca, Cbcb	/*在句首*/
Da	Daa	/*後可接名詞*/
Dfa	Dfa	/*後爲VH～VL*/
Dfb	Dfb	/*接在V之後*/
Di	Di	/*在V之後*/
Dk	Dk	/*句首*/
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, Ndba, Ndbb, 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*/
P	P*	/*介詞*/
VA	VA11,12,13,VA2,VA3,VA4	/*動作不及物動詞*/
VB	VB11,12,VB2	/*動作類及物動詞*/
VC	VC1, VC2, VC31,32,33	/*動作及物動詞*/
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*/
A	A	/*非謂形容詞*/
I	I	/*感嘆詞*/
T	Ta, Tb, Tc, Td	/*語助詞*/
Str	Str	/*字串*/
De	/*的, 之, 得, 地*/	
SHI	/*是*/	
FW	/*外文標記*/	

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CKIP詞典而來。詞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、狀語或修飾語。也就是說一個標記不一定只代表一個功能，至於詳細的標記原則，請見CKIP技術報告#95-02。

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，以及多功能標記所代表的功能。

單一功能的標記有：Caa、Cab、Cba、Cbb、Cbc、Dfa、Dfb、Di、Dk、D、Nf、Ng、P、I、T、Str、Nes、Neq　（限於篇幅，每一類的功能請參考CKIP技術報告#93-05）

多功能的標記有：

Da	ADV（ADV表示可以出現在狀語位置）、N-modifier（N-modifier表示可以緊跟在名詞之前）　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表代名詞功能） 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
VA～VG	V、N-modifier　例：主辦(VC)奧運；主辦(VC)單位
VH～VL	V、N-modifier、ADV(manner)　（manner爲表方式的狀語） 例：很努力(VH)；努力(VH)方式；他努力(VH)工作
V_2	有(V)　例：有(V_2)人走來；他有(V_2)書
SHI	是　例：他是(SH1)很認眞；他是(SH1)老師；他全身是(SH1)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

這些詞類標記所代表的功能，事實上，有些也是經驗累積的結果。由於中文詞彙活用的現象很普遍，詞典不可能盡列每一個詞的每一種用法。例如：“八股”在詞典登錄的是Na（普通名詞）；但是在“他很八股”中做主要謂語時，則應該標記爲VH。“善意”在詞典中登錄的是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。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。

4.3 特徵標記集

除了標記詞類以外，我們也爲某些特殊句法表現做標記，目前使用的特徵標記共8個，是針對動補式動詞和動賓式動詞的可折（separable）現象、合併詞中插現象〔計算語言學通訊 1995〕、名物化結構和外來語所設計。特徵標記集如表七所示。

表七、特徵標記集

特徵標記	使 用 情 況	例　子
+vrv	V of a separable VR	叫不醒 Vc[+vrv]
+vrr	R of a separable VR	叫不醒 Vc[+vrr]
+spv	V of a separable V N compound	吃了他的虧 Vc[+spv]
+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]

五、結語

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眞實的語料作爲研究素材。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顚撲不滅的眞理。相反的，我們相信在這兩百萬詞的分詞與標記中，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，使用者的回饋可使下版的研究院語料庫更接近完善！

參考文獻：

計算語言學通訊，1992，分詞標準。

計算語言學通訊，1995，「搜」文解字。

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。

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。

詞庫小組，1995，研究院語料庫的內容及說明，中文詞知識庫小組技術報告#95-02，南港，中央研究院。

Chang, Li-ping and Chen Keh-jiann 1995. *The CKIP Part-of-speech Tagging System for Modern Chinese Texts.* To be presented at ICCPOL,'95 Conference, 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 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 Keh-jiann Chen. 1992. *A Chinese Corpus for Linguistics Research.* In the Proceedings of the 1992 International Conference on Computational Linguistics (COLING-92). 1214-1217. Nantes, France.

Hsu, Hui-li and Chu-Ren Huang, 1995. Design Criteria for a Balanced Modern Chinese Corpus. To be presented at ICCPOL'95 conference, Hawaii.

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

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.

附錄一：　　分詞細則

1. 複合詞：

1.1 定量式複合詞：

a. 數詞　　合

例：一千八百、百分之三十、三十%、三又二分之一、六十六點五、五成三。

b. 表時間、地點之定量詞　　合

例：一九九五年　三月　六日　二點　二十分

c. 普通定量詞　　分

例：三 位、五十二 隻、三又二分之一 打、七十餘 位、七十 位 餘

1.2 複合動詞：

a. 並列式		合	例：發交相關單位
偏正式		合	例：增蓋、冷燻、勇奪
動賓式		合	例：洗澡、回國
主謂式		合	例：頭昏
動補式	方向補語	合	例：跑上來
	結果補語補語是單音節	合	例：打壞
	補語是雙音節	分	例：打掃　乾淨

b. V-給、V-到、V-於、V-有、V-爲、V-成、V-作　另行規定

V-給：合；但動詞本身是動賓或動補結構：分

例：批發給、寫信 給、分紅 給、取出 給、退回去 給

V-到：合；但後接時間，補語和數量詞時：分

例：接觸到、聊 到 半 夜、走 到 腿酸、加 到 兩百萬

V-於：分；但動詞是附著詞、合併後意義改變、表示比較：合

例：生 於 台北、吝於、有感於、大於、優越於

V-有：合

V-爲：合

V-成：合

V-作：合

1.3 偏正式複合名詞：

簡單式：合

帶詞綴	如：修路費、準女婿
語意無組合性	如：土包子、鐵公雞
專指	如：白菜、黑板
不能中插	如：太陽眼鏡、旅行支票
使用頻率高	如：牛肉麵
含有附著成分	如：奇案、勇將

複雜式：簡短、常見式：合　例：借書證、租車費

冗長、少見式：分　例：欲偷渡到美國者

1.4 複合副詞：合 例：竟是、並非、暫不、既已…

1.5 名方式複合詞：

a. 語意失去組合性	合	例：手上、腦中、眼裡
b. 有清楚的指涉	合	例：目前、月底
c. 其他	分	

2. 定語+的、狀語+的/地　分

例：詞+的/地　我的書、代理的人員

詞組+的/地　很神氣的天鵝、聲音平緩的讀著稿

3. 簡稱：合

例：男單、女網、空姐、影視、化工、音像

4.合併詞：

無中插：詞頭合併		合	例：國內外、高中職
詞尾合併		合	例：父母親、公私立
頭尾合併		合	例：中山南北路
套裝合併	前面是專名	分	例：台北　市長、新竹　縣政府
	前面是其他	合	例：事務局長、體育司長
有中插：		分	例：初、高中

5. 重疊詞：合

動詞：嘗試貌：談談、想想、研究研究、說說看

暫時貌：坐坐 就 走、擦擦 就 可

程度貌：胖胖 的、辛辛苦苦

名詞：車車、狗狗、痘痘

量詞：片片、一 片片　　（但：一 片 一 片、一 片 又 一 片）

擬聲詞：叮叮噹噹、乒乒乓乓　　（但：嘩啦 嘩啦、哈 哈 哈）

6. 專有名詞：

單純詞：合　例：胡適、布農、貝多芬、阿爾及利亞、宇宙光

專名+普名：普名是詞綴：合　例：阿美族、光復橋、竹聯幫、桃園廠、王董

普名是自由詞素：分　例：胡 先生、二二八 事變、永新 加油站

縮寫：合　例：勞基法、奧申委、文建會、臺三線、北二高

複雜詞：分　例：第一 信用 合作社、省 自來水 公司、

詞組或句子：分　例：前世 今生、鯨魚 的 生 與 死

7. 中插詞：分

動賓中插：分　　例：洗 了 一 個 澡

動補中插：分　　例：打 得 破、打 不 破

重疊中插：分　　例：笑 了 笑、哭 一 哭

8. A-not-A：分

喜歡 不 喜歡　　喜 不 喜歡