現代漢語平衡語料庫數位化工作流程簡介

更新日期：2005/10/30

計畫單位：中央研究院語言學研究所

計畫名稱：語言典藏計畫

計畫簡介：

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，否則易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進行統計研究或作爲測試語料。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。

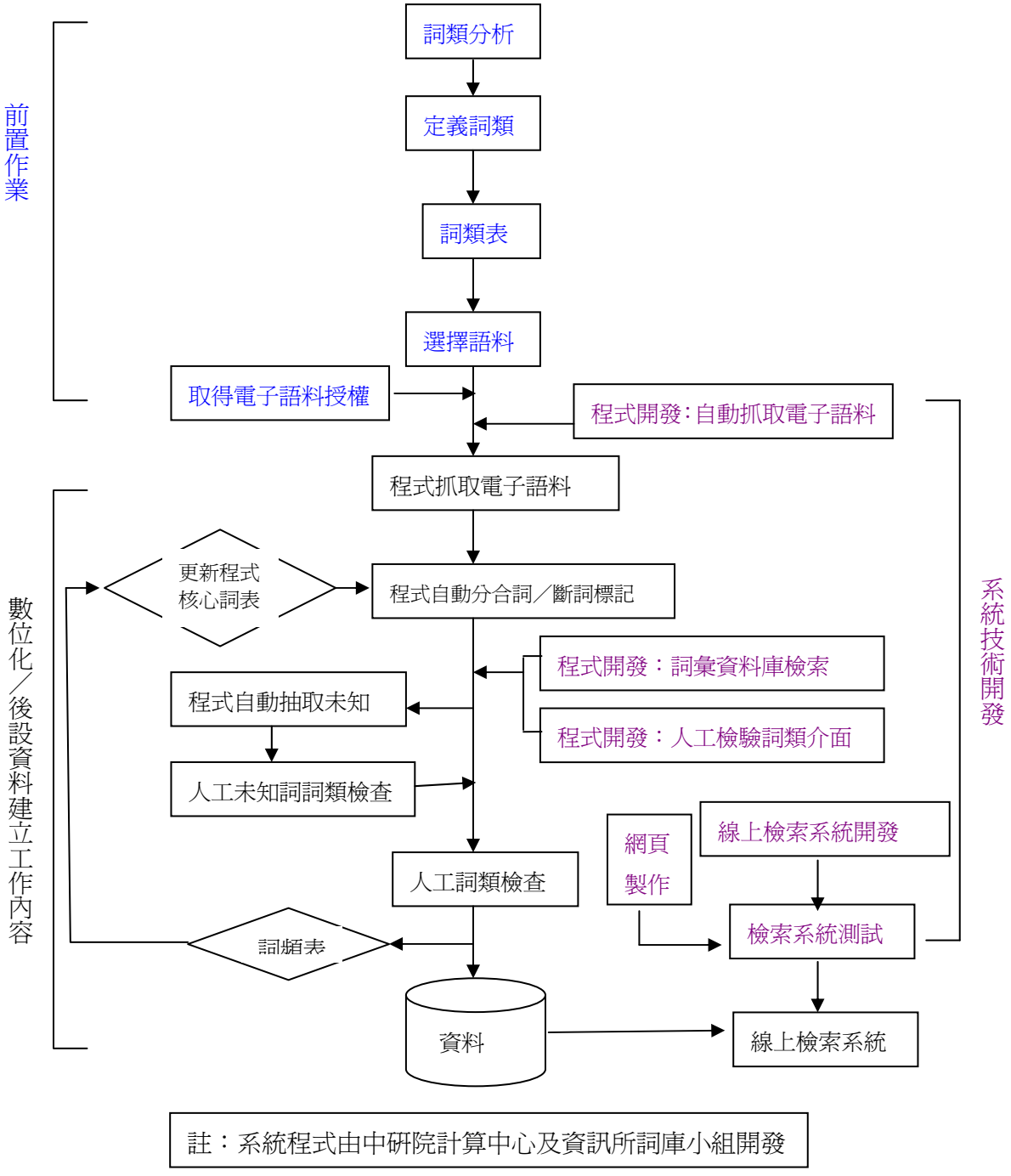
「中央研究院現代漢語平衡語料庫」簡稱「研究院平衡語料庫」（Sinica Corpus ），是世界上第一個有完整詞類標記的漢語平衡語料庫。由於加詞類標記的漢語語料庫是史無前例的嚐試，這個語料庫是由中央研究院資訊所、語言所共同指導的詞庫小組完成的。該小組由陳克健(資訊所)、黃居仁 (語言所) 兩位研究員主持，自 1990 年前後便開始致力於中文語料庫的收集(Huang & Chen 1992)，至 1994 年止已收集有近二千萬字之現代漢語語料及超過五百萬字之古代漢語語料(Huang 1994)。由於有了處理中文語料庫的經驗，及大量處理電子詞庫中詞條的經驗(陳克健等 1991, Chen 1994)，中央研究院詞知識庫小組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。

因此，在 1994 年分別得到了中央研究院「中文資訊」跨所研究群之專案計畫及國科會計畫補助，乃開始著手進行現代漢語平衡語料庫的建構。爲兼顧理想與實用性，初步目標定爲兩百萬詞，爲傳統小規模平衡語料庫之兩倍，1996 年經計算中心設計規劃完成 WWW版，開放供各界使用，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。2001 年國家型數位典藏科技計畫展開，詞庫小組認爲應持續收集近年之語料，使語料樣本能完整呈現二十世紀臺灣使用漢語的全貌，因此以新五百萬詞爲目標進行知識典藏工作，目前介面已升級至 4.0 版，提供更完整的語料條件檢索功能。

數位化工作流程說明：

該計畫的數位化作業，大致依照下列六項步驟進行，依序分別爲：一、詞類分析、定義及確定；二、選擇語料文本來源；三、程式抓取電子語料；四、程式自動分合詞及詞類標記；五、人工詞類檢查；六、匯入語料庫。茲分別介紹如次。

現代漢語平衡語料庫工作流程圖：



圖片提供者：

中央研究院語言學研究所　盧秋蓉小姐

一、詞類分析、定義及確定

分詞規範的研擬分爲兩種方式進行，一方面是邀請台灣知名的學者專家召開討論會，就其專業領域的角度，對分詞規範的大方針進行討論；另一方面則是中央研究院詞庫小組根據分詞規範，實際從事語料分析，從上百萬的語料中，整理出分詞標準的細節規定。然後，於1998年舉行分詞規範公聽會，1999年中文分詞原則正式通過爲國家標準，編號CNS14366[1]。中央研究院詞庫小組再依此規範進行詞類分析、定義及確定之工作。

《資訊處理用中文分詞規範》有下列兩個突破：(1)提出分級的觀念及確立信、達、雅三級的標準。最容易達到的信級訂爲基本資料交換的標準；以技術上較難，但自動分詞程式仍可達到的達級作自動翻譯、資訊檢索等自然語言處理的標準；至於最需要人工分詞才能達到的雅級則視爲電腦處理、理解中文之最高目標。(2)把分詞規範分成不變核心（分詞單位定義及基本原則），以及可變準則（輔助原則）。在確定分詞規範架構後，只要定時更新基本詞庫或特殊領域的專門詞庫，便可維持分詞規範的不變性[2]。

二、選擇語料文本來源

平衡語料之抽取以自中央研究院詞庫小組現有之語料(近二千萬字之現代漢語語料)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。

(一)交換取得之語料：此項包括經由合作計畫交換取得的，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：慷慨提供該計畫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：大部份由聯合新聞網、中時電子報及電子佈告欄（BBS）或蕃薯藤等萬維網中取得。

---

[1]「資訊處理用中文分詞原則」國家標準 ，檔案編號CNS14366，內容詳附錄一。

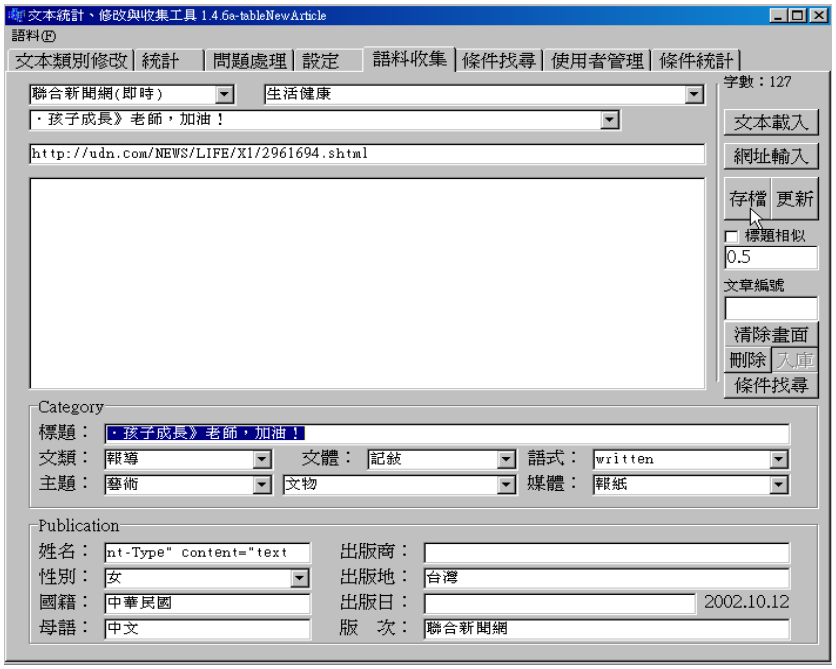
[2]「資訊處理用中文分詞規範」設計理念及規範內容

三、程式抓取電子語料

使用程式 CKIP Corpus&Spider1.4.6a 抓取線上電子語料。助理使用電子語料抓取程式，需先選擇語料來源，再選取欲匯入語料庫之文章。由於語料來源媒體之分類並不一致，，而現代漢語平衡語料庫分類爲 6 類：文學、哲學、藝術、科學、社會、生活，故需將文章重新分類，以便匯入。(圖二、圖三、圖四)

目前，以主題爲準，訂出平衡語料庫的內容比例爲：文學 20%、哲學 10%、藝術 5%、科學 10%、社會 35%、生活 20%，根據此參考值爲基準選取語料。

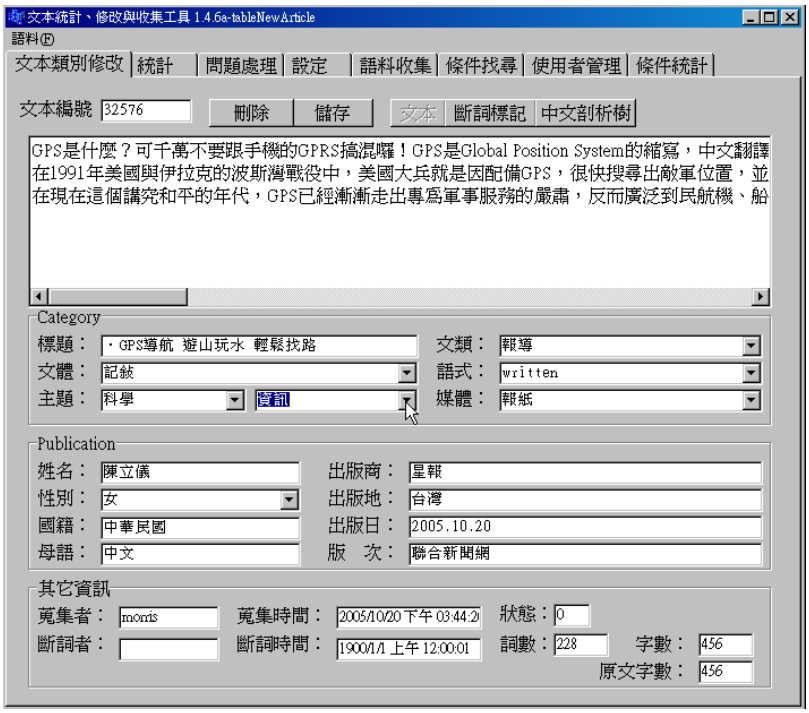
圖一：抓取電子語料工作畫面。(示範者：邱智銘)



圖二：語料收集畫面

圖三：確認需匯入之語料庫位置

圖四：目前語料庫收集文章情形



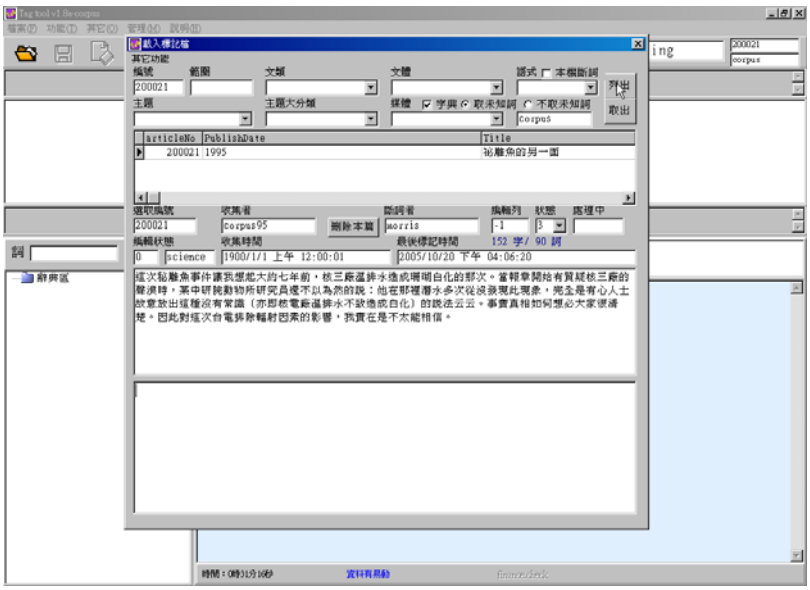
圖五：語料修改的畫面

四、程式自動分合詞及詞類標記

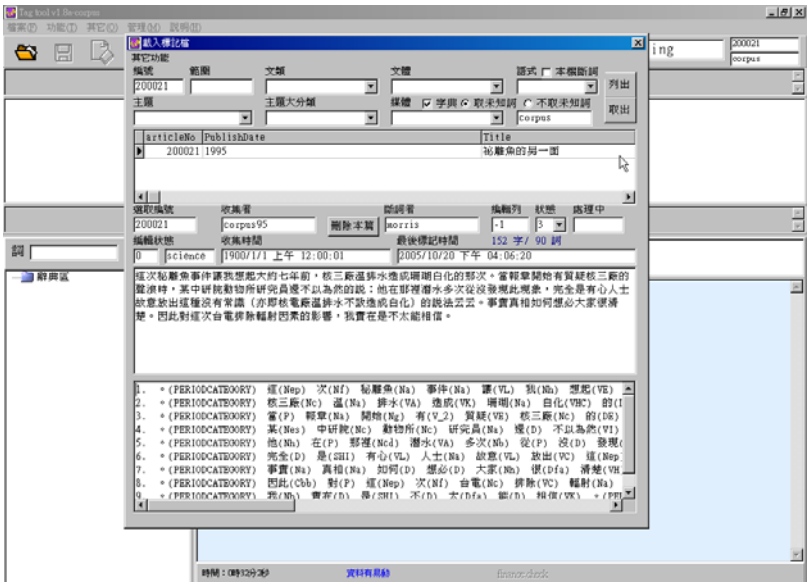
語料選取完畢，接下來的工作是標記詞類，但是在這之前，還要先爲語料做斷詞工作，唯有每個詞區隔非常明確之後，才能標記詞類。目前機器自動斷詞的正確性約達 95%。

基本上，自動斷詞的步驟是以中研院辭典中的八萬目詞爲基礎，切分爲一個一個獨立的詞。未列在辭典中的成分，則以字爲單位，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。

機器自動斷詞是使用 CKIP Tag Tool V1.8a 系統，該程式即是一個協助詞類標記檢查的輔助工具，輸入欲執行自動斷詞之語料的文本編號，執行自動斷詞後，程式會將斷詞後之語料顯現於語料本文下方欄位。（圖六、圖七）



圖六：選取欲執行自動斷詞之語料

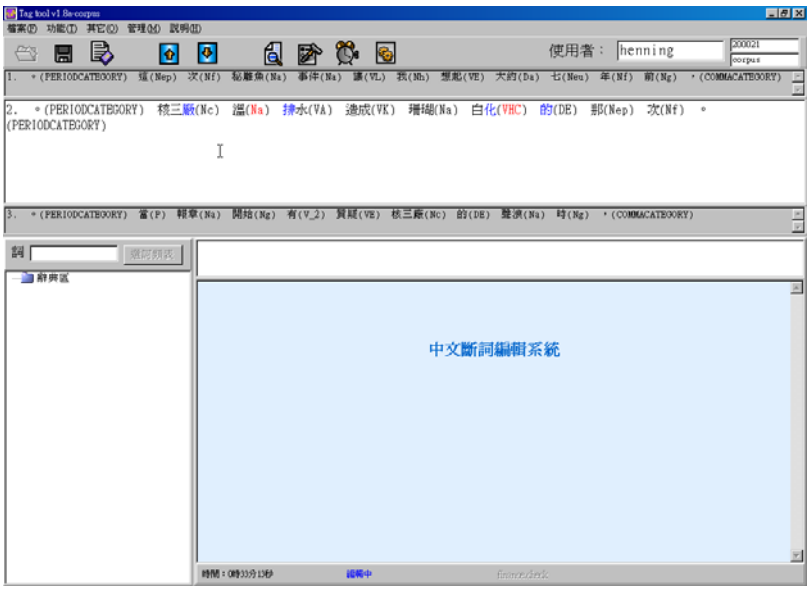


圖七：自動斷詞執行完畢

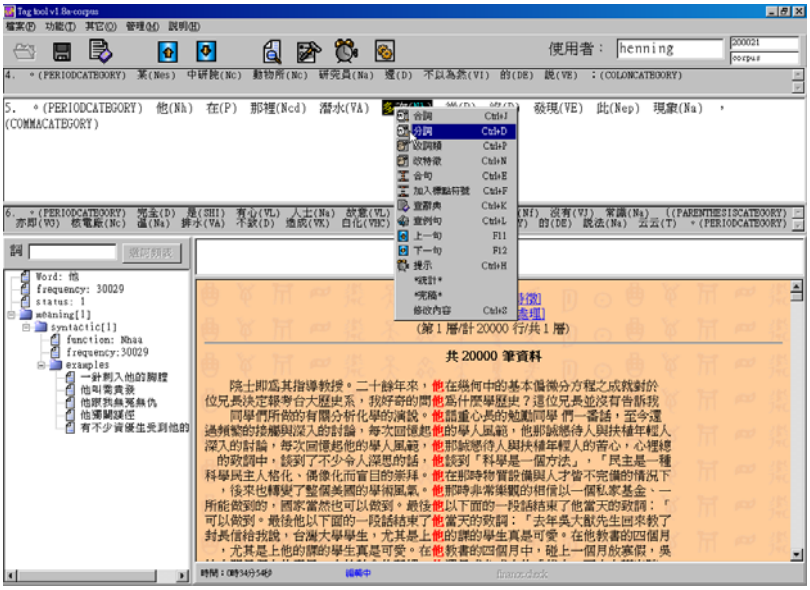
五、人工詞類檢查

使用程式自動斷詞及詞類標記後，由於斷詞會因文章內容而導致詞彙不同的切斷方式，故爲避免斷詞與文義不符，再由助理以人工方式作詞類檢查。

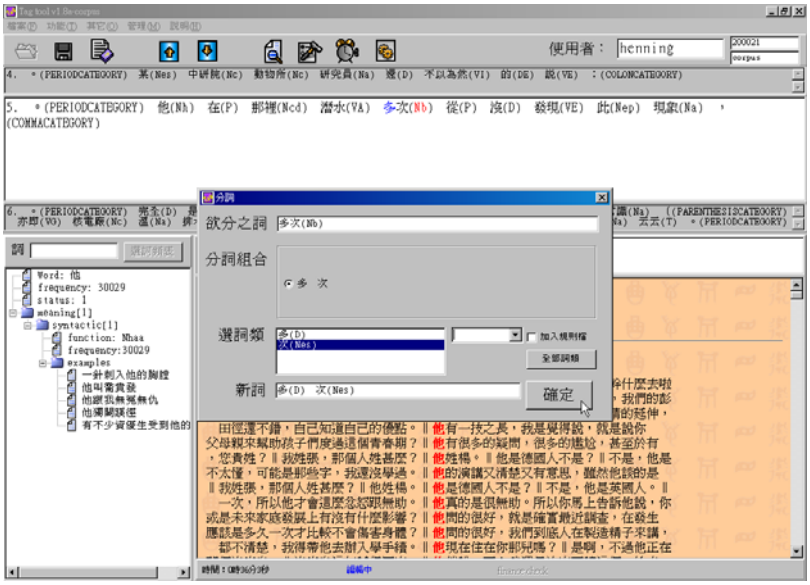
在進行人工確認時，會利用中文斷詞編輯介面。系統進行人工確認，每次以一句爲單位，並列出上下句供參考。確認無誤後，再以上下鍵移動繼續進行人工詞類檢查之工作。（圖八）



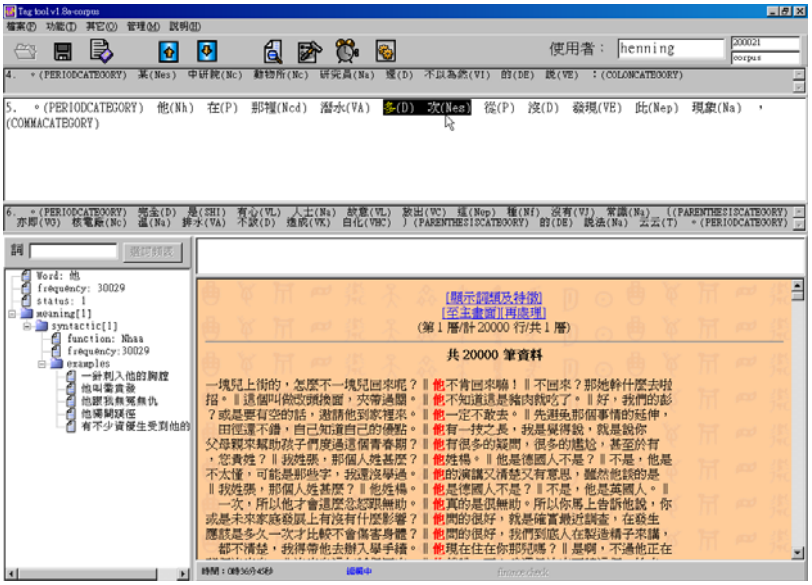
圖八：中文斷詞編輯系統



圖九：修正斷詞畫面



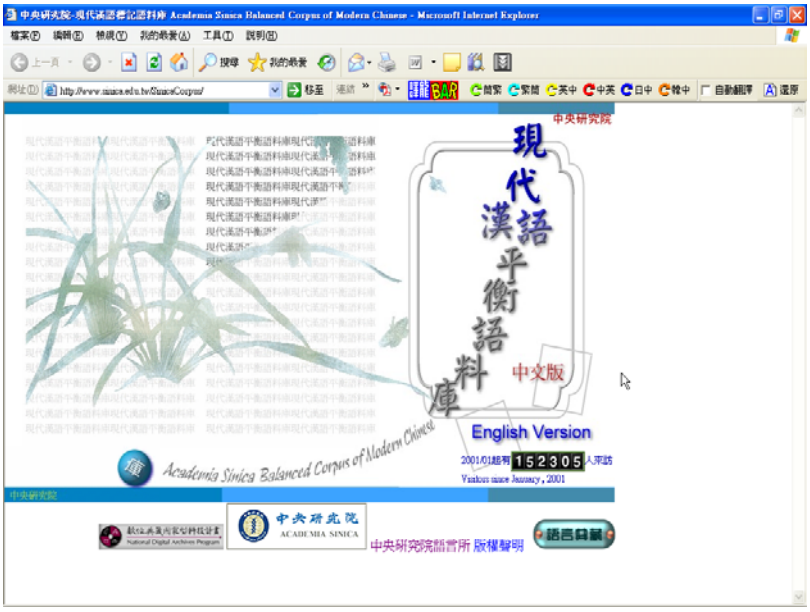
圖十：輸入欲修正之詞彙及斷詞方式



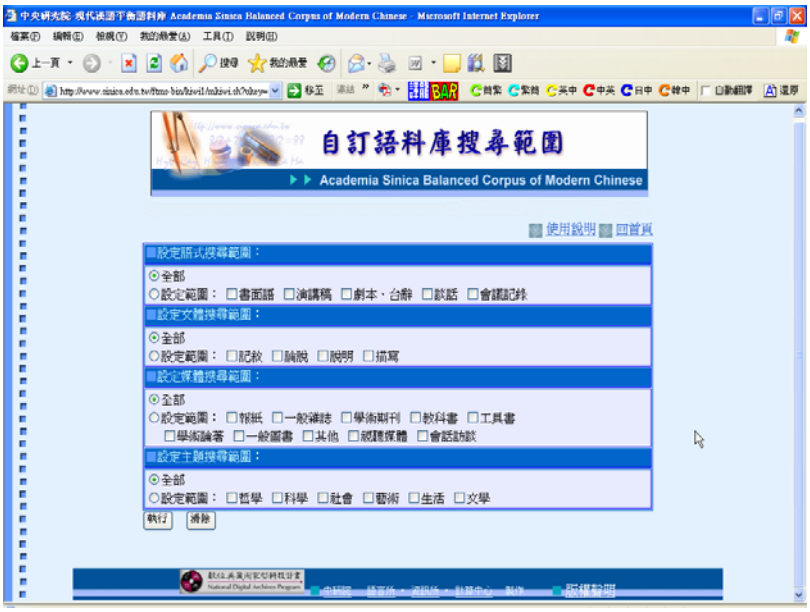
圖十一：斷詞修正完畢

六、匯入語料庫

人工詞類檢查進行完畢後，再將完成詞類斷詞和標記之語料，以網路傳送至中央研究院計算中心，再由該單位匯入現代漢語標記語料庫。



圖十二：現代漢語平衡語料庫網頁



圖十三：現代漢語平衡語料庫搜尋網頁

※ 製作單位：數位典藏國家型科技計畫---內容發展分項計畫
中央研究院語言學研究所---語言典藏計畫
※ 文字撰寫：數位典藏國家型科技計畫---內容發展分項計畫
語言主題小組助理 賴佳旻
中央研究院語言學研究所語言典藏計畫助理 盧秋蓉、邱智銘
※ 圖片拍攝：數位典藏國家型科技計畫---內容發展分項計畫
語言主題小組助理 賴佳旻、林淑惠
※ 圖文編輯：數位典藏國家型科技計畫---內容發展分項計畫
語言主題小組助理 賴佳旻、陳秀華
※ 感謝中央研究院語言學研究所---【語言典藏】之計畫主持人 鄭錦全院士與共同主持人 黃居仁老師及 陳克健老師和助理盧秋蓉小姐、邱智銘先生撥冗指教及協助拍攝與提供資料，特別致謝。