

研究院語料庫-現代漢語詞類與標記原則

邱智銘

中央研究院語言學研究所

E-mail: henning@hp.iis.sinica.edu.tw

詞類標記

- 詞類標記集
 - 詞類標記與功能
 - 特徵標記集
 - 詞類標記原則及範例
 - tagtool for Windows 版使用介紹
 - 文本統計、修改與收集工具
 - 現代漢語語料庫數位處理流程圖
 - 中研院現代漢語平衡語料庫
-

詞類標記集

□ 八大類別

- **A** : 非謂形容詞 **C**: 連接詞 **D**: 副詞 **N**: 體詞 (名詞)
- **I** : 感嘆詞 **T**: 語助詞 **P**: 介詞 **V**: 述詞 (動詞)

□ 178個詞類 (詞庫小組1993)

□ 43個簡化標記

□ 3個特殊標記

詞類標記集

- 非謂形容詞(A): 主要是作名詞的修飾語，不具謂語作用，是純粹的形容詞。

 - 非謂形容詞(A)類型：
 1. 以名詞成分為基礎：空心
 2. 以動態述詞性成分為基礎：平裝 野生 新興
 3. 以狀態不及物的形容成分為基礎：大紅 上好 全盛
 4. 其他：真正 共同 有機
-

詞類標記集

- 連接詞(C): 主要是在連接兩個或兩個以上的語言單位，組成較大的語言單位。
 - 連接詞(C)類型:
 1. **Ca**: 並列連接詞
 - 1.1. **Caa**: 和 跟
 - 1.2. **Cab**: 等 等等 之類
 2. **Cb**: 關聯連接詞
 - 2.1 **Cba**(移動性前繫連接詞): 因為 即使
 - 2.2 **Cbb**(非移動性前繫連接詞): 就是 不但
 - 2.3 **Cbc**(後繫連接詞): 那麼 而且
-

詞類標記集

- 語助詞(T): 是一種後置成份，必須附在句子或詞組之後，藉以修飾句子或詞組，為表示說話者的語氣
 - 語助詞分類: (以出現先後次序分)
 1. **Ta**: 了 的
 2. **Tb**: 沒 而已 罷了 也好
 3. **Tc**: 啊 哇 呢 耶 喔
 4. **Td**: 了嗎 而已嗎 與否 哉
-

詞類標記集

- 感歎詞(1): 表說者情緒或態度，是永遠獨用的語式，一般出現在句字前，有時也在句後

 - 感歎詞為列得完的一類，依情緒分如下：
 1. 表驚訝或感嘆: 哎呀 哇
 2. 表悲痛或痛惜: 嗚呼 唉
 3. 表憤怒或鄙斥: 哼 呸
 4. 表懊悔或惋惜: 咳
 5. 表疑惑: 咦 哦
 6. 表稱讚: 妙哉 嘿
 7. 表了解: 喔 噢
 8. 表否定: ?
 9. 表應諾: 欸 嗯
 10. 表招呼: 喂 嗨 哈囉
 11. 表警語: 噓
-

詞類標記集

- 介詞(P): 介詞在漢語中屬於前置詞(preposition) , 同時也是功能詞的一種, 其判斷標準為:
 1. 介詞必須引介一論元, 且此論元成分不可省略。
 2. 介詞不做謂語中心。
 3. 介詞沒有時態(aspect) , 沒有嘗試貌。

 - 介詞為一封閉的集合, 詞庫小組依介詞的語法表現和扮演的語意角色歸類66組介詞(P01 ~ P66)
-

詞類標記集

- 副詞(D): 主要是當作謂語或句子的修飾語，副詞在句中出现的位置常在主語和述詞之間，部分可出現在句首，極少數可出現在述詞後。
 - 副詞的次分類是依語意判斷的，可分以下11類:
 1. Da (數量副詞): 一共 都
 2. Dba (法相副詞): 可能 應該
 3. Dbb, Dbc (評價副詞): 居然 難怪
 4. Dc (否定副詞): 未曾 從不
 5. Dd (時間副詞): 時常 漏夜 近來
 6. Df (程度副詞): 非常 很 極了
-

詞類標記集

7. Dg (地方副詞): 當街 一路 處處
 8. Dh (方式副詞): 變相 私自 千方百計
 9. Di (標誌副詞): 著 過 起
 10. Dj (疑問副詞): 為何 幹啥 是否
 11. Dk (句副詞): 總之 老實說 據說
-

詞類標記集

- 體詞(N)：一般而言，名詞在句中充當主語或賓語，只有少數名詞才會充當謂語。
 - 體詞的分類：
 1. Na(名詞)
 2. Nb(專有名稱): 人名 部落名 歷史事件
 3. Nc(地方名詞): 地方名稱 行政單位
 4. Nd(時間名詞): 季節 朝代
 5. Ne(定詞): 指示定詞 特指定詞 數詞定詞 數量定詞
 6. Nf(量詞): 計量的單位詞
 7. Ng(方位詞)
 8. Nh(代名詞)
-

詞類標記集

- 述詞(V): 為一個句子的中心語，其所承載的訊息，包括述詞必要的論元個數、述詞論元的詞組形式、論元的語意角色及語意限制，都是述詞分類架構的依歸。
 - 述詞分類:
 1. VA(動作不及物述詞): 只需一個論元作主語 /坐/睡/進駐
 2. VB(動作類單賓述詞): 需兩個參與論元，且賓語不能直接出現在述詞之後 /求婚/洗塵/拜年
 3. VC(動作單賓述詞): 需兩參與論元，且皆為名詞組 /檢查/學
 4. VD(雙賓述詞): 一個述詞，後接兩個賓語。需三個論元來滿足其語意表現 /送/交
 5. VE(動作句賓述詞): 接句子論元的動作述詞。需二或三個論元 /自言自語/大聲急呼
-

詞類標記集

6. VF(動作謂賓述詞): 以述詞組為其賓語的及物述詞，需二或三個論元 /打算/勸
 7. VG(分類述詞): 連結客體(THEME)與範圍(RANGE)兩個角色，需二或三個論元 /稱呼/等於
 8. VH(狀態不及物述詞): 只有一必要論元 /動聽/瀟灑
 9. VI(狀態類單賓述詞): 需兩參與論元，但賓語不能直接出現在述詞之後 /鍾情/失信
 11. VJ (狀態單賓述詞): 在語意上要求兩個必要的名詞組論元 /代表
 12. VK(狀態句賓述詞): 接句子論元之狀態述詞，需兩論元 /了解/不滿
 13. VL(狀態謂賓述詞): 以述詞組為其賓語的狀態述詞，需二或三個論元 /輸/讓
-

詞類標記集

□ 3個特殊標記:

1. **DE**: 的, 之, 得, 地
 2. **SHI**: 是
 3. **FW**: 外文標記
-

詞類標記與功能

- 詞類給定的的原則，理論上是一個詞一個類，但就語法功能來說，一個標記不一定只代表一個功能，所以詞類標記可分為兩大類，一是單一功能標記，另一是多功能標記。
 - 單一功能標記:
Caa、Cab、Cba、Cbb、Dfa、Dfb、Di、Dk、D、Nf、Ng、Neu、Nes、Neqb、P、I、T。
 - 多功能標記:
A、Da、DE、SHI、N*(Nf、Ng、Nd、Nep、Neqa....)、V*(VH、V_2....)。
-

特徵標記集

□ 除了標記詞類外，詞庫小組也為某些特殊句法表現做標記，目前使用的特徵標記共九個，包括：

1. 動補式特徵標記
 2. 動賓式特徵標記
 3. 合併詞中插特徵標記
 4. 外來語特徵標記
 5. 名物化特徵標記
 6. 專有名詞特徵標記
-

特徵標記集

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

詞類標記原則及範例

□ 詞類標記原則

1. 詞類標記應符合它在語境中所扮演的語法功能。
 2. 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標記。
 3. 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。
-

詞類標記原則及範例

□ 範例：過(Di, Dfa, VH, VCL)

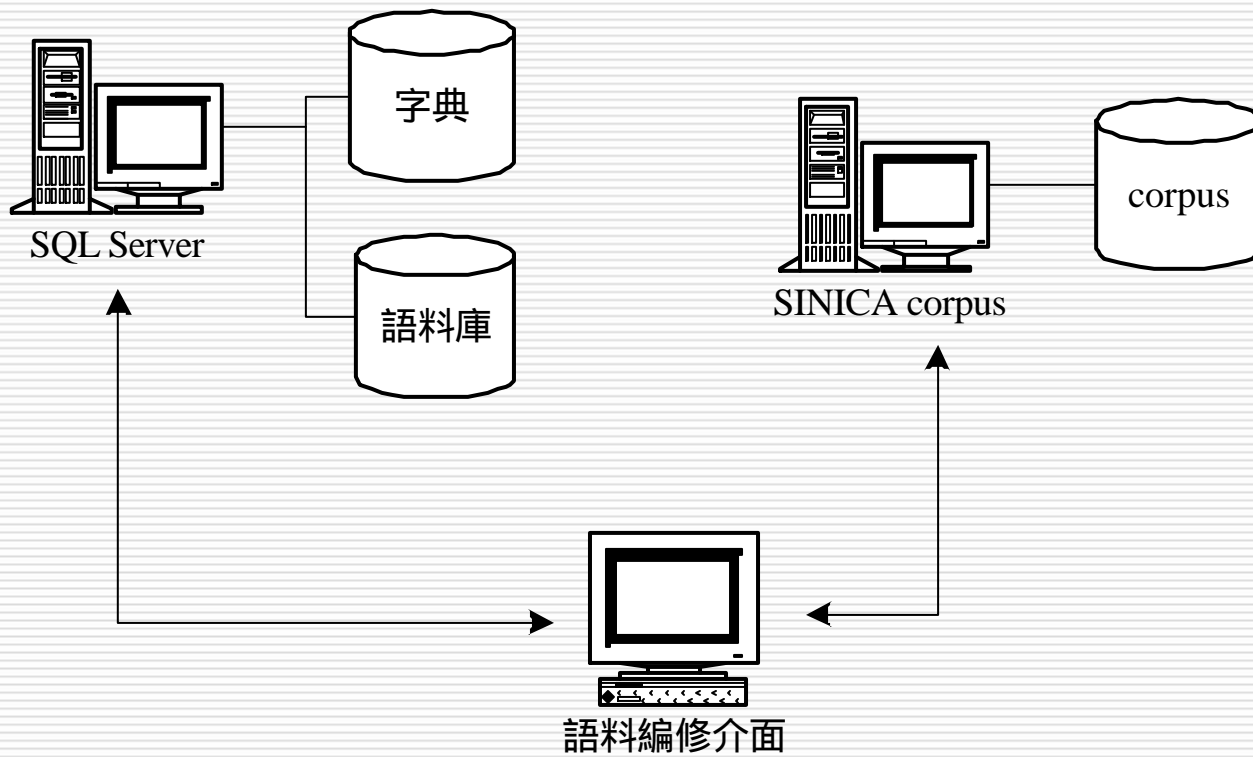
1. 我看過(Di)這本雜誌
 2. 他的體重是過(Dfa)重了點
 3. 他的檢定考沒過(VH)
 4. 時間真的過(VC)的太快了！
 5. 他走過(VCL)這座橋。
 6. 這座橋，他走過(Di)/走過(VCL)。
-

tagtool for Windows 版使用介紹

- 本系統在輔助處理斷詞標記後的檢驗動作，以提高語料庫的標記品質。使用者可以在任何一台電腦上，透過區域網路的連線，連上後端語料庫及詞典伺服器，來選取欲編輯的文本與查詢某詞的詞典資料等。
-

tagtool for Windows 版使用介紹(續)


□ 整個系統的連結如下圖所示：



tagtool for Windows 版使用介紹(續)

- 系統操作流程大致上可分為4個部份：
 1. 語料選取
 2. 未知詞擷取與斷詞標記
 3. 人工檢驗
 4. 語料修改、輸出與統計。
-

tagtool for Windows 版使用介紹(續)

- 語料選取：使用者欲檢驗文本時，在主畫面中點選 [檔案] / [載入標記檔] 或按 。本介面以條件式的查詢方式從語料庫伺服器中取出。操作畫面如下：



載入標記檔

其它功能

編號 範圍 文類 文體 語式 本機斷詞

5903 主題 主題大分類 媒體 字典 取未知詞 不取未知詞

tableArticle 列出 取出

articleNo	PublishDate	Title
5903	2003.1.29	靈鷲山泰國講堂開光典禮圓滿

選取編號 收集者 斷詞者 編輯列 狀態 未看

5903 henming 刪除本篇 -1 0

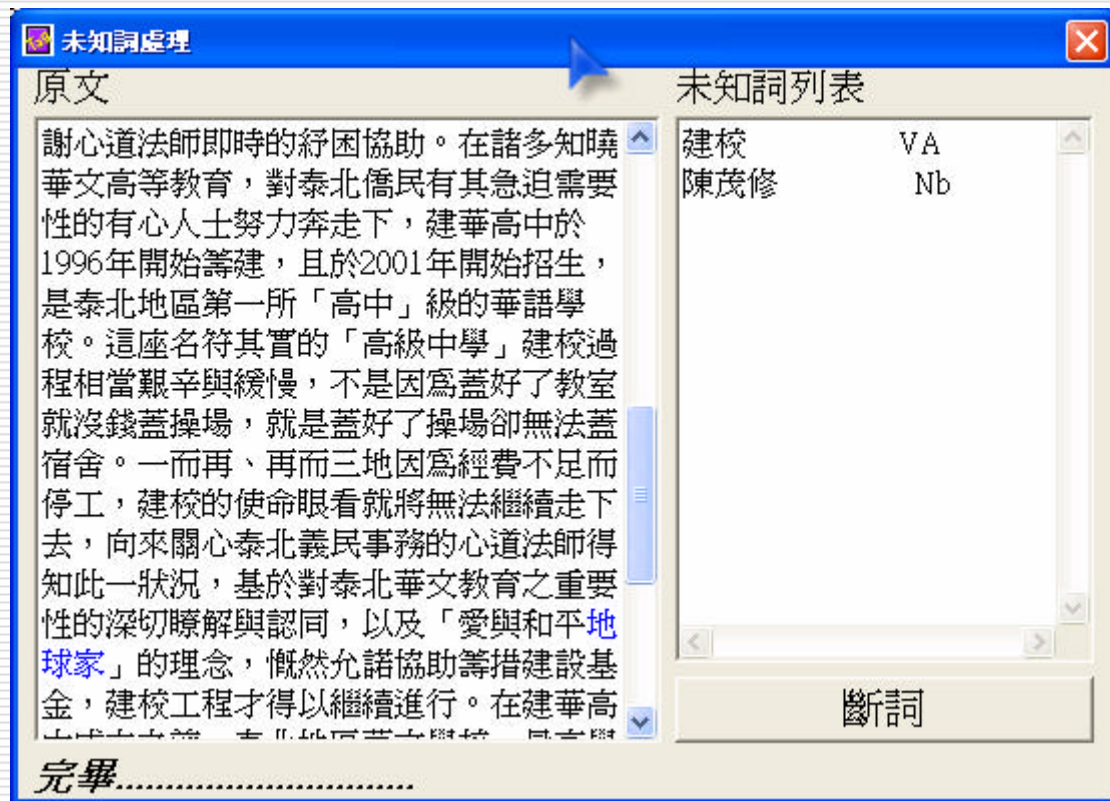
編輯狀態 收集時間 最後標記時間 0 字 / 0 詞

0 society 2003/3/7 下午 04:30:07 1900/1/1 上午 12:00:01

雖然再過一天就是中國的農曆新年，也是僑民最忙碌的時候，但是昨(29日)天由心道法師親臨主持的靈鷲山泰國講堂開光典禮仍湧進大批信眾，專程來聽心道法師開示。在開光典禮中，心道法師表示，靈鷲山能在泰國成立講堂，內心感到相當歡喜，希望藉由講堂的成立，為泰國僑民提供一處心靈庇護所。在昨天的典禮中同時舉行-「散播愛的種子」-靈鷲山捐贈泰北建華高中建校捐贈儀式，由泰北建華高中的董事長陳茂修將軍代表接受。在-「散播愛的種子」-捐贈典禮中，當心道法師將捐款親手交給陳茂修將軍時，陳將軍紅著眼眶感性地緊握住心道法師的手，激動哽咽地無法說出話來。內心對心道法師真摯的感謝溢於言表，現場氣氛溫馨感人。陳將軍表示如果沒有心道法師的協助，完成建華高中建校工程恐怕是遙遙無期。全校師生為了感念心道法師，建華高中特別在男教員宿舍牆面上以

tagtool for Windows 版使用介紹(續)


□ 未知詞擷取與斷詞標記



tagtool for Windows 版使用介紹(續)

- 人工檢驗主畫面如下圖，



- 快速按鈕介紹：
 - 1) 載入標記檔
 - 2) 儲存
 - 3) 查辭典
 - 4) 上一句
 - 5) 下一句
 - 6) 全文瀏覽
 - 7) 批次更新
 - 8) 提醒
 - 9) 查看規則檔。

tagtool for Windows 版使用介紹(續)

- 語料統計:在主畫面下，點選 [管理] / [統計]，選擇文章的範圍。



文本統計、修改與收集工具

1.6-tableNew Article

語料

文本類別修改 | 統計 | 問題處理 | 設定 | 語料收集 | 條件找尋 | 使用者管理 | 條件統計

中時電子報(即時) | 運動

8公尺59 菲立浦斯 首跳定金牌

http://news.chinatimes.com//Chinatimes/newslist/newslist-content/0,3546,130512+132

12年前，因1場車禍，醫生說他將從此半身不遂；但現在，他贏得雅典奧運男子跳遠金牌，美國菲立浦斯用行動告訴世人：只要不放棄，成功最後會屬於你。
菲立浦斯26日在奧運男子跳遠項目，以8公尺59的成績封王，而隊友莫菲特也以8公尺47的個人最佳成績拿下第2名，使美國包辦金銀牌。菲立浦斯可說是1跳定天下，他第1跳就躍出8公尺59的成績，底定大局。他後來試著要挑戰奧運與世界紀錄，可惜急於求表現，反而越跳越差，在第2、3跳犯規，第4、5跳則放棄，第6次的最後1跳成績只有8公尺35，最後就以首跳成績拿下金牌。
4年前雪梨奧運，菲立浦斯僅獲第8名，使美國自1980年以來首度未獲男子跳遠獎牌，但雪梨奧運後，菲立浦斯的成績大躍進，他在去年贏得室內與室外田徑賽冠軍後，又在今年8月初曾跳出8公尺60個人最佳成績，寫下4年來的世界最佳成績。

字數：
文本載入
網址輸入
存檔 更新
 標題相似
0.5
文章編號
清除畫面
刪除 入庫
條件找尋

Category

標題： 8公尺59 菲立浦斯 首跳定金牌

文類： 報導 | 文體： 記敘 | 語式： written

主題： 生活 | 體育 | 媒體： 報紙

Publication

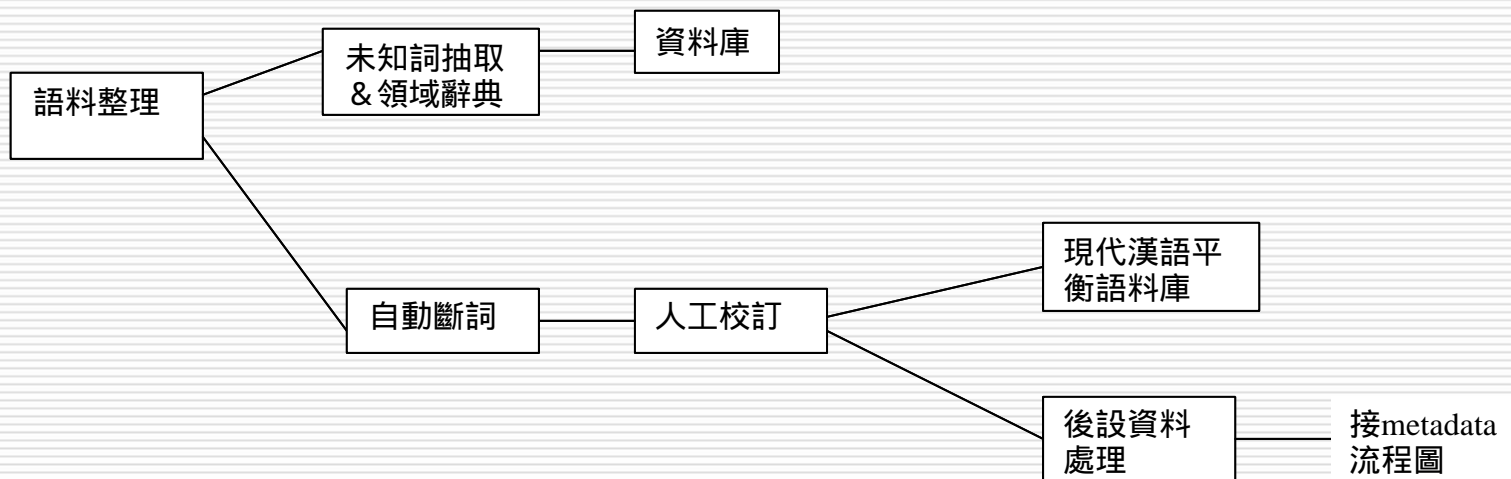
姓名： 楊欣怡 | 出版商： 中時晚報

性別： 女 | 出版地： 台灣

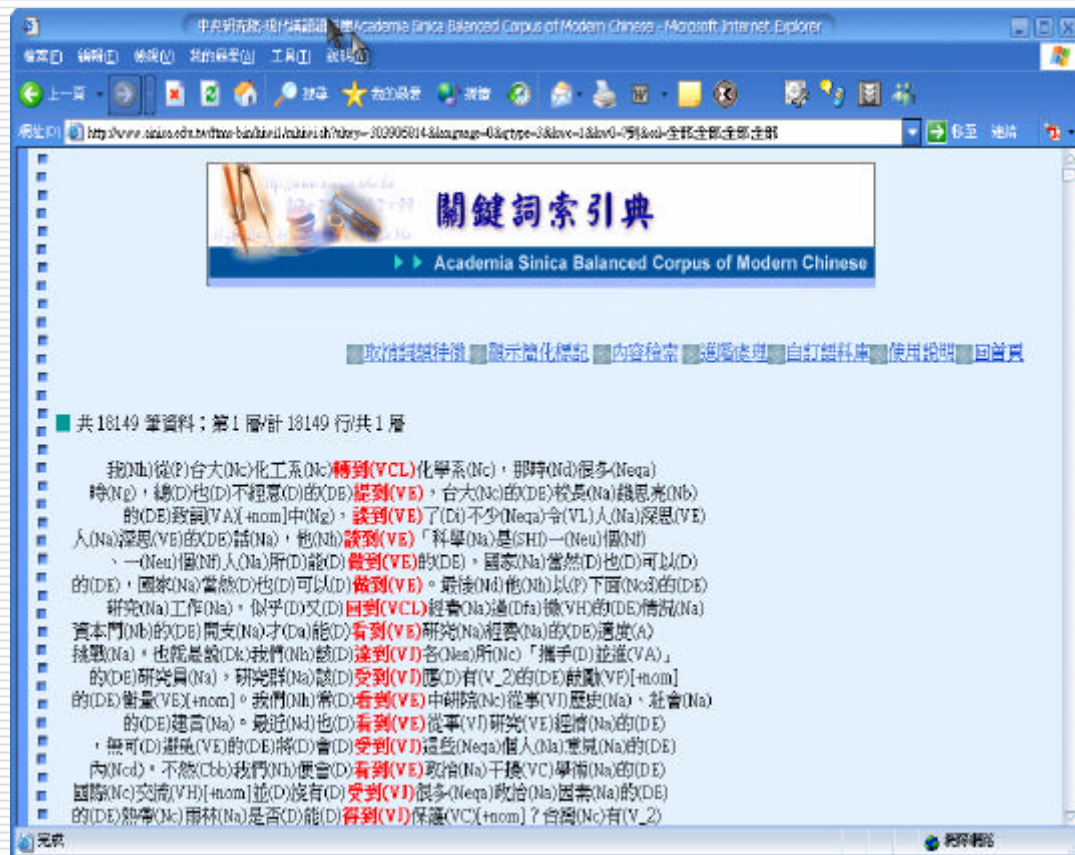
國籍： 中華民國 | 出版日期： 2004.08.27 | 2002.10.12

母語： 中文 | 版次： 綜合報導

現代漢語語料庫數位處理流程圖



中研院現代漢語平衡語料庫



中央研究院現代漢語平衡語料庫 Academia Sinica Balanced Corpus of Modern Chinese - Microsoft Internet Explorer

http://www.sinica.edu.tw/tw/index.html?key=203905814&lang=zh-tw&type=0&nav=1&nav0=0

關鍵詞索引典

Academia Sinica Balanced Corpus of Modern Chinese

取消關鍵字 顯示簡化標記 內容檢索 調整速度 自動換行 使用說明 回首頁

■ 共 18149 筆資料；第 1 層/計 18149 行/共 1 層

我(Nh)從(P)台大(Nc)化工系(Nc)轉到(VCL)化學系(Nc)，那時(Nd)很多(Neqa)時(Ng)，總(D)也(D)不經意(D)的(DE)提到(VE)，台大(Nc)的(DE)校長(Na)錢思亮(Nb)的(DE)政詞(WA[+nom])中(Ng)，談到(VE)了(Di)不少(Neqa)令(VL)人(Na)深思(VE)人(Na)深思(VE)的(XE)話(Na)，他(Nh)談到(VE)「科學(Na)是(SH)一(Neu)個(Nf)一(Neu)個(Nf)人(Na)所(D)能(D)做到(VE)的(DE)，國家(Na)當然(D)也(D)可以(D)的(DE)，國家(Na)當然(D)也(D)可以(D)做到(VE)。最後(Nd)他(Nh)以(P)下面(Nc)的(DE)研究(Na)工作(Na)，似乎(D)又(D)回到(VCL)經濟(Na)過(Dfa)渡(VH)的(DE)情況(Na)資本門(Nb)的(DE)開支(Na)才(Da)能(D)看到(VE)研究(Na)經費(Na)的(DE)進度(A)摘要(Na)，也就是說(Dk)我們(Nh)總(D)達到(VI)名(Nca)所(Nc)「攜手(D)並進(VA)」的(XE)研究員(Na)，研究群(Na)談(D)受到(VI)應(D)有(V_2)的(DE)鼓勵(VF[+nom])的(DE)衡量(VE[+nom])。我們(Nh)常(D)看到(VE)中研院(Nc)從事(VI)歷史(Na)、社會(Na)的(DE)建言(Na)。最近(Nd)也(D)看到(VE)從事(VI)研究(VE)經濟(Na)的(DE)，無可(D)避免(VE)的(DE)將(D)會(D)受到(VI)這些(Neqa)個人(Na)意見(Na)的(DE)內(Ncd)，不然(Cbb)我們(Nh)便會(D)看到(VE)政治(Na)干預(VC)學術(Na)的(DE)國際(Nc)交流(VH)[+nom]並(D)沒有(D)受到(VI)很多(Neqa)政治(Na)因素(Na)的(DE)的(DE)熱帶(Nc)雨林(Na)是否(D)能(D)得到(VI)保護(VC[+nom])？台灣(Nc)有(V_2)

<http://www.sinica.edu.tw/SinicaCorpus/>

參考文獻

- 詞庫小組。1993。中文詞類分析技術報告no.93-05。
 - 詞庫小組。1996。「搜」文解字 – 中文詞界研究與資訊用分詞標準技術報告no.96-01。
 - 詞庫小組。1998。中央研究院平衡語料庫的內容與說明技術報告no.95-02/98-04。
-