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(I)：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，副詞在句中出現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/瀰漫
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 /輪/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	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)

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)

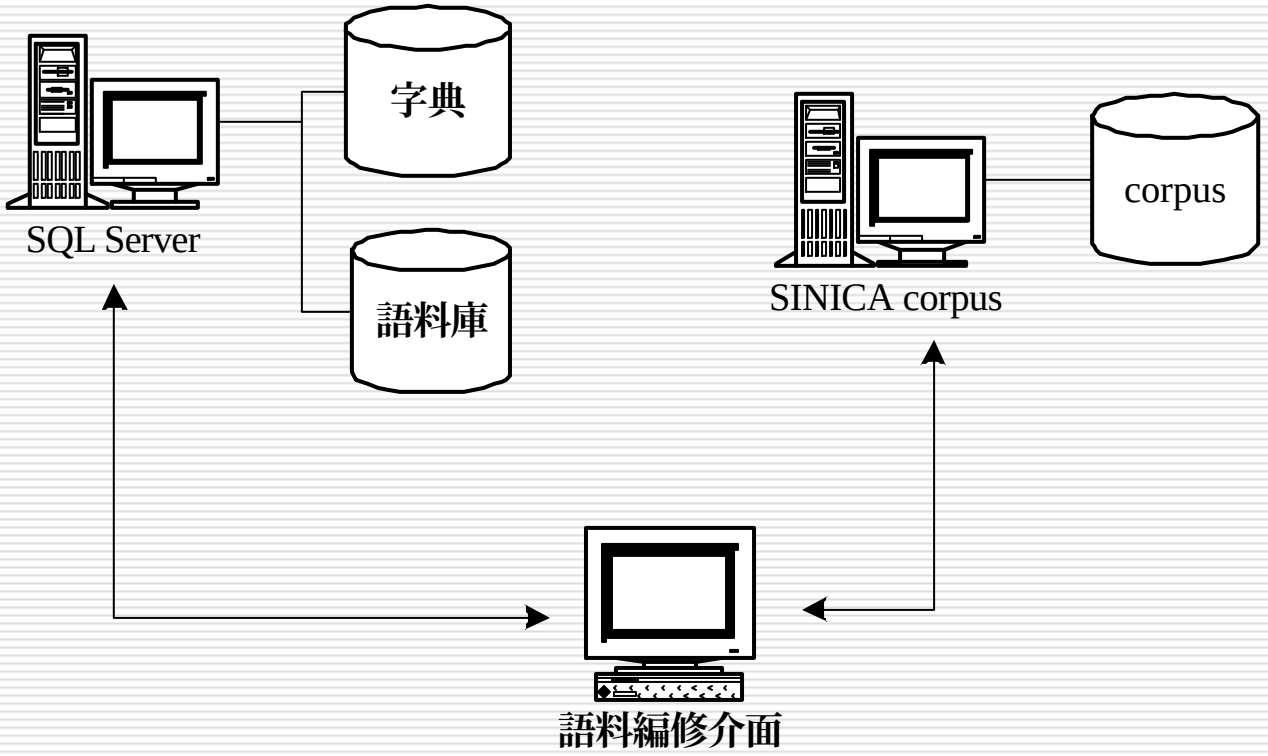
1. 我看過(Di)這本雜誌
2. 他的體重是過(Dfa)重了點
3. 他的檢定考沒過(VH)
4. 時間真的過(VC)的太快了！
5. 他走過(VCL)這座橋。
6. 這座橋，他走過(Di)/走過(VCL)。

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：

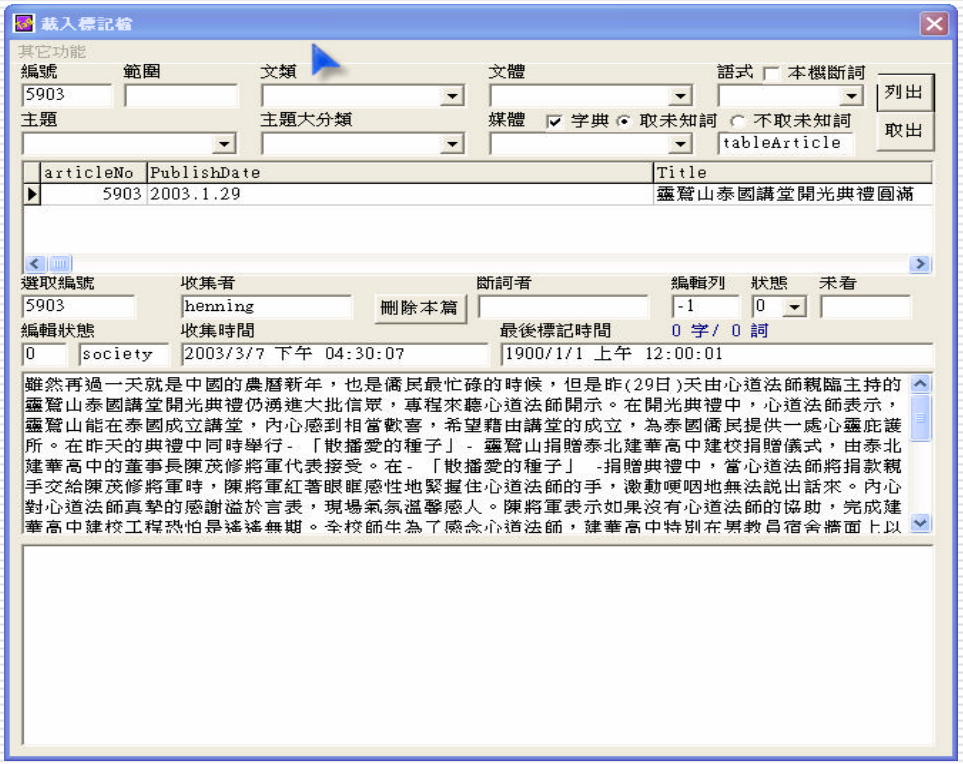


tagtool for Windows 版使用介紹(續)

- 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。

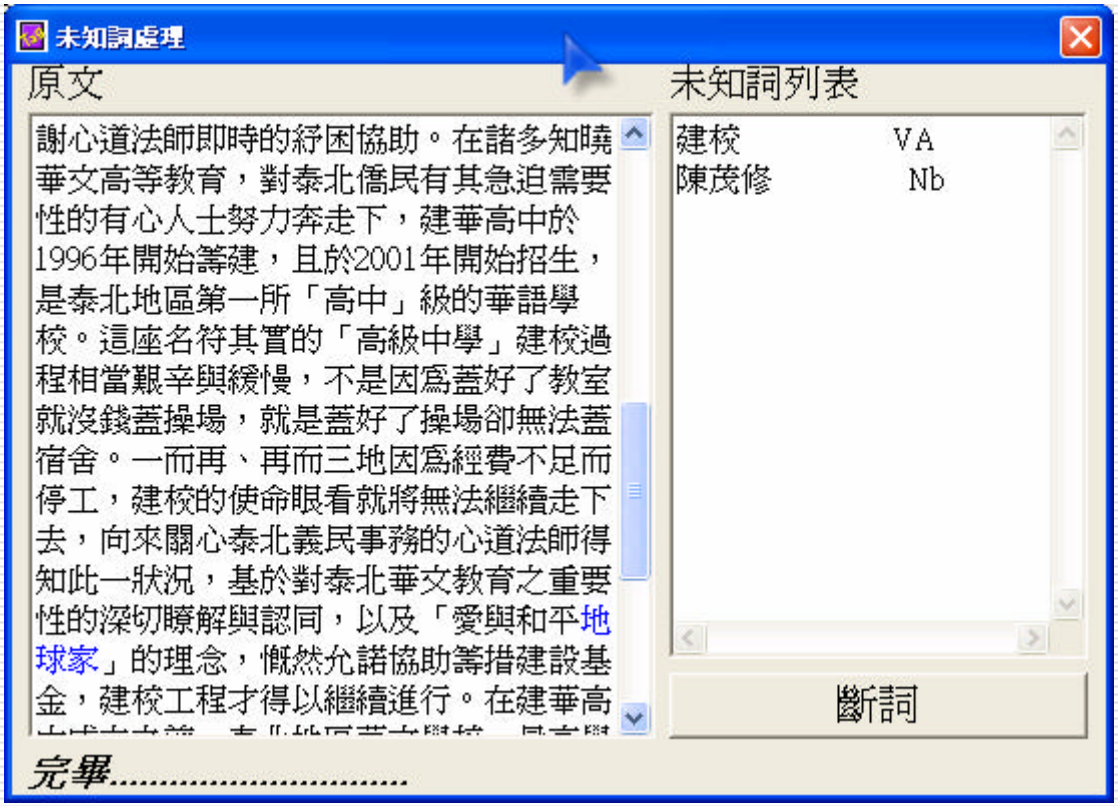
tagtool for Windows 版使用介紹(續)

- 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：



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

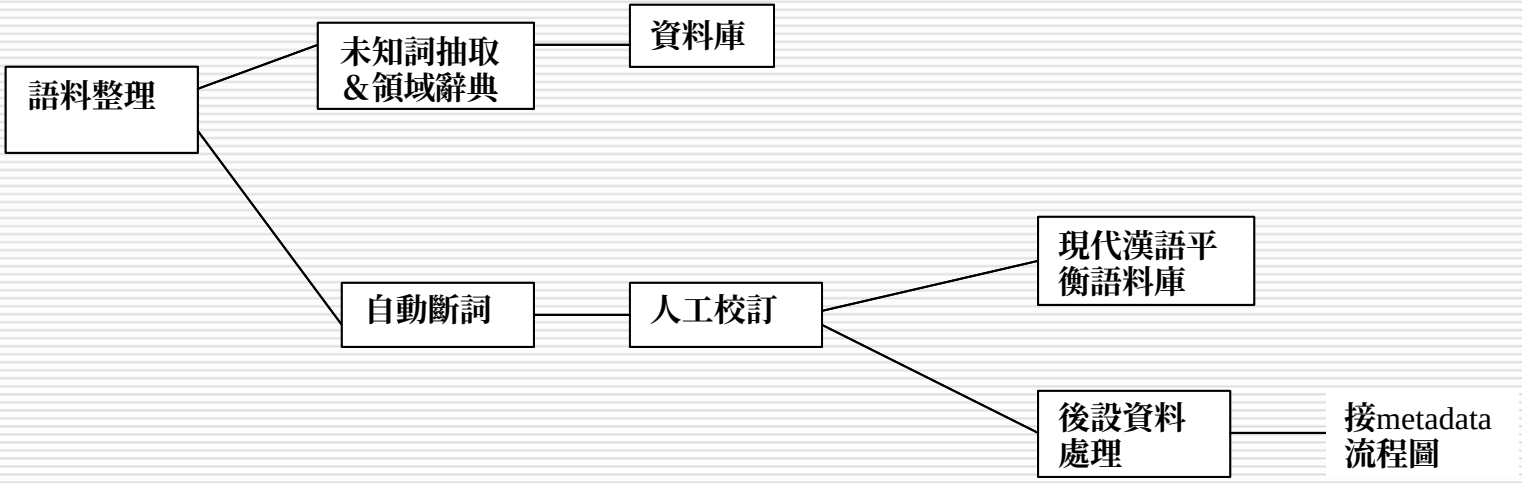
5)下一句 6)全文瀏覽 7)批次更新 8)提醒 9)查看規則檔。

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

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

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 。