Chinese Word Sketch Engine操作手冊

1. 網址

http://wordsketch.ling.sinica.edu.tw/

2. 登入

3. 檢索首頁

檢索關鍵詞或詞組的語料

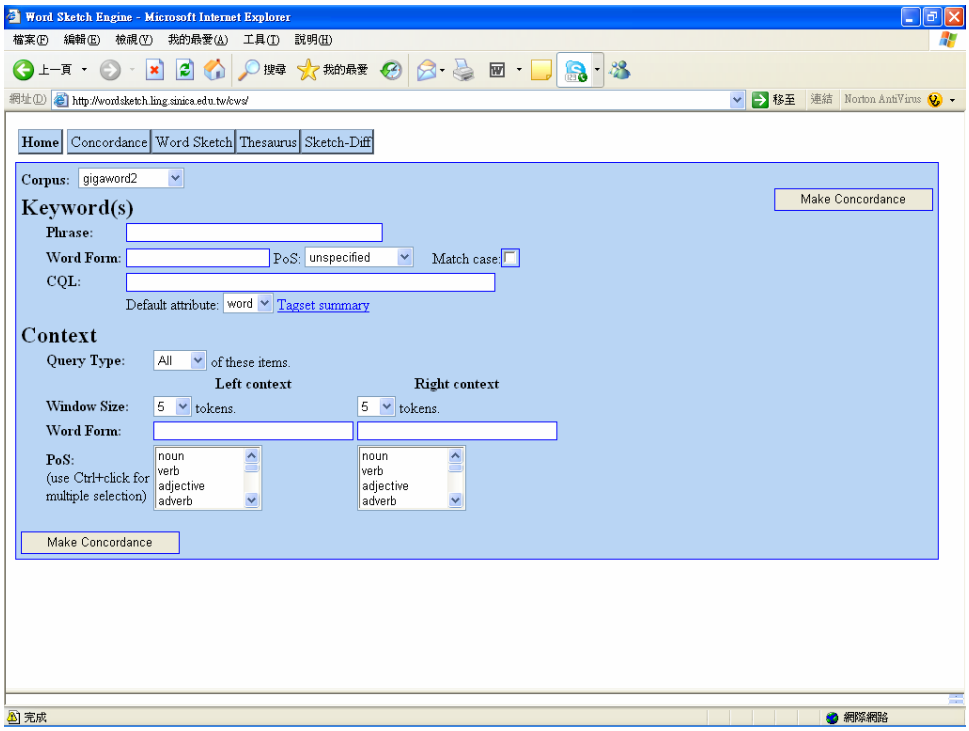
對語料庫的進階處理

檢索詞彙的進階相關訊息

其他相關訊息與使用說明

4. 檢索關鍵詞或詞組（Concordance）的語料

點選 New Query 後進入以下頁面：



4.1 上方選項

點選上方 Home、Concordance、Word Sketch、Thesaurus、Sketch-Diff 等五選項，可連結至主頁(Home)與其他檢索功能頁面。

4.2 語料庫（Corpus）選單

此下拉選單提供繁體字版與簡體字版語料，以及次語料庫檢索設定選項：

(1) 設定 gigaword2_cna，可查詢繁體字版的 CNA 中央通訊社語料。
(2) 設定 gigaword2_xin，可查詢繁體字版的 XIN 新華社語料。
(3) 設定 gigaword2_zbn，可查詢繁體字版的 ZBN 新加坡聯合早報語料。
(4) 設定 sinica，可查詢繁體字版的中央研究院平衡語料庫 5.0 版語料。
如果已經建有其他次級語料庫，次語料庫的名稱亦會出現在下拉選單中。

4.3 關鍵詞（Keyword(s)）設定

可輸入單一詞彙（word），含二個或二個緊鄰詞彙的詞組(phrase)，或者以語料庫檢索語言（Corpus Query Language （CQL））來設定檢索的關鍵項目。
(1) 詞組 （Phrase）設定
每個詞彙之間應空一個半形空白間隔，例如：「覺得 說」、「我 想 你」、「一般 認爲 ， 這 是」等等。
(2) 詞彙(word)
可輸入所要查詢的關鍵詞彙。例如：「覺得」、「想」、「認爲」。

(3) 語料庫檢索語言（Corpus Query Language （CQL））

設定細則請參看語料庫檢索語言說明：
http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/CQPSyntax.html

4.4 語境（Context）設定

本項目提供使用者進一步設定關鍵詞前後搭配的詞語環境。使用者可以檢索到非緊鄰的詞彙搭配語料。

4.4.1 檢索類型（Query Type）

下拉選單中提供三種進階設定方式：

(1) All (of these items)：擷取同時包含以下欄位中關鍵詞左、右二詞彙的語料。

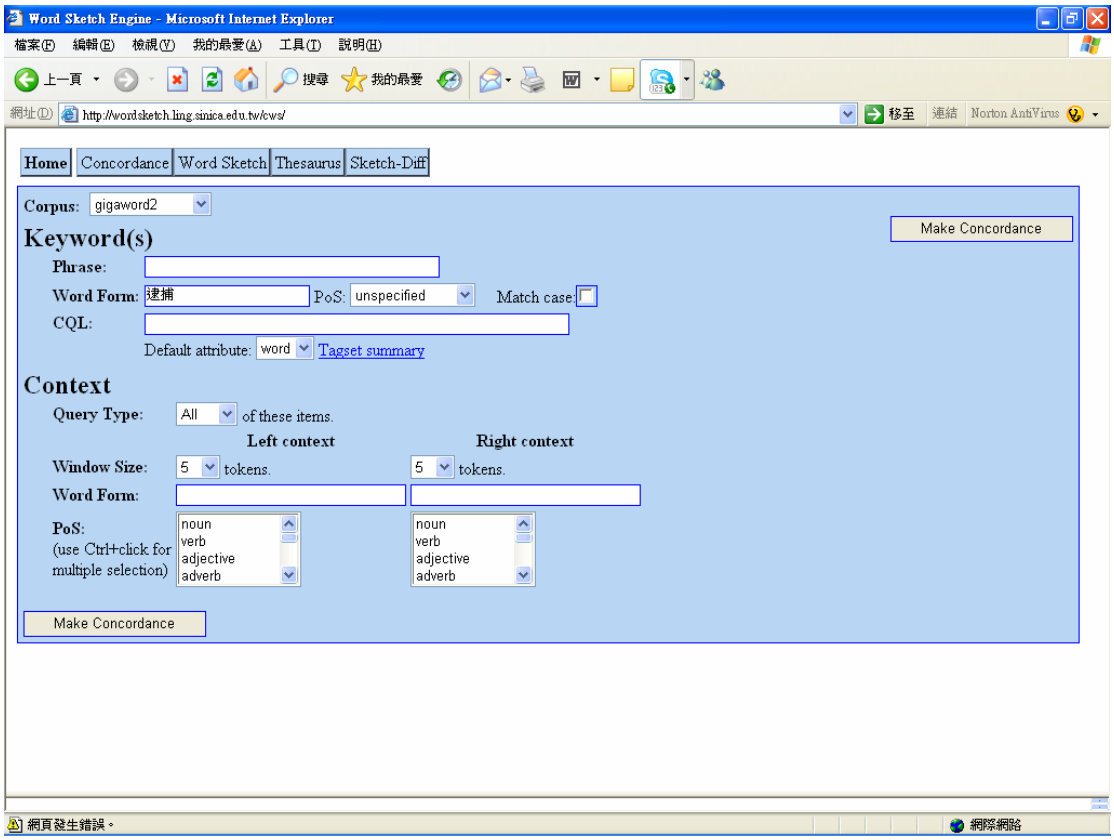
(2) Any (of these items)：擷取含以下欄位中關鍵詞左側或右詞彙之一的語料。

(3) None (of these items)：過濾掉同時含以下欄位中關鍵詞左、右二詞彙的語料。

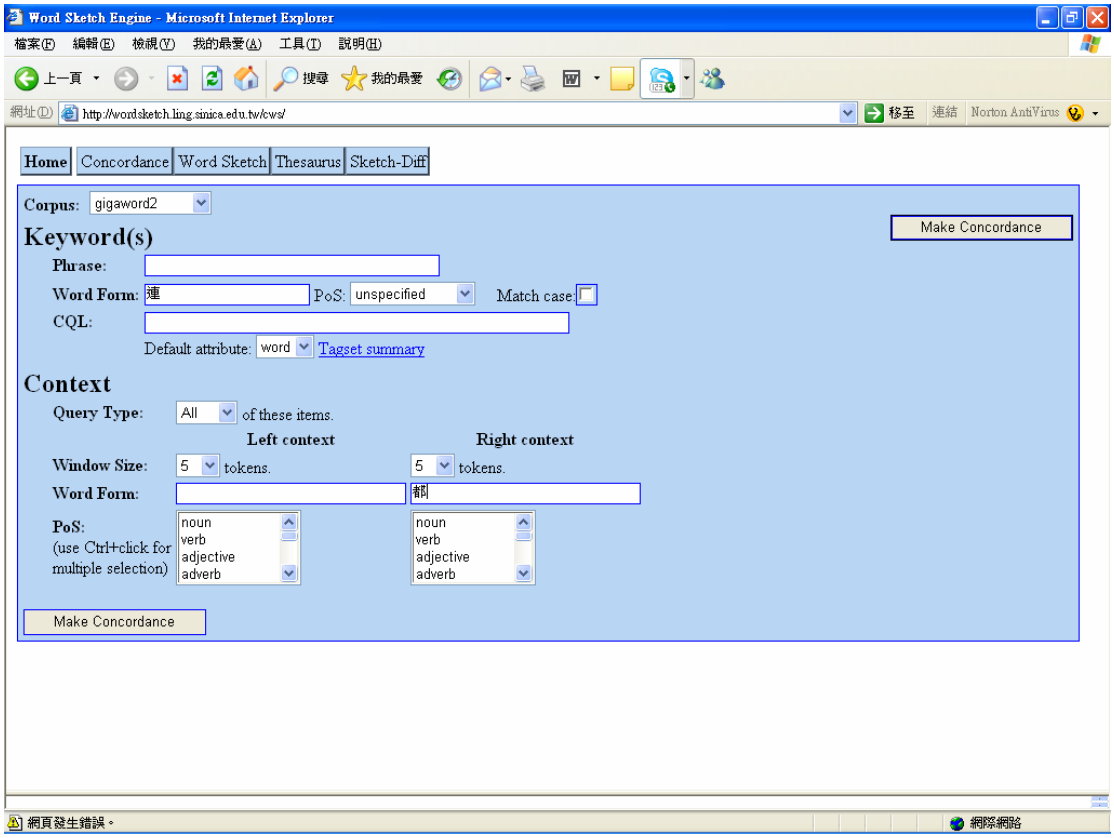
4.4.2 設定與關鍵詞的距離（Window Size）

下拉選單中提供 1 到 10 個詞(token)的設定方式。預設值爲 5，表示在和關鍵詞距離 5 個詞彙之內的範圍。

範例一：欲檢索含「逮捕」的語料。可如下設定：

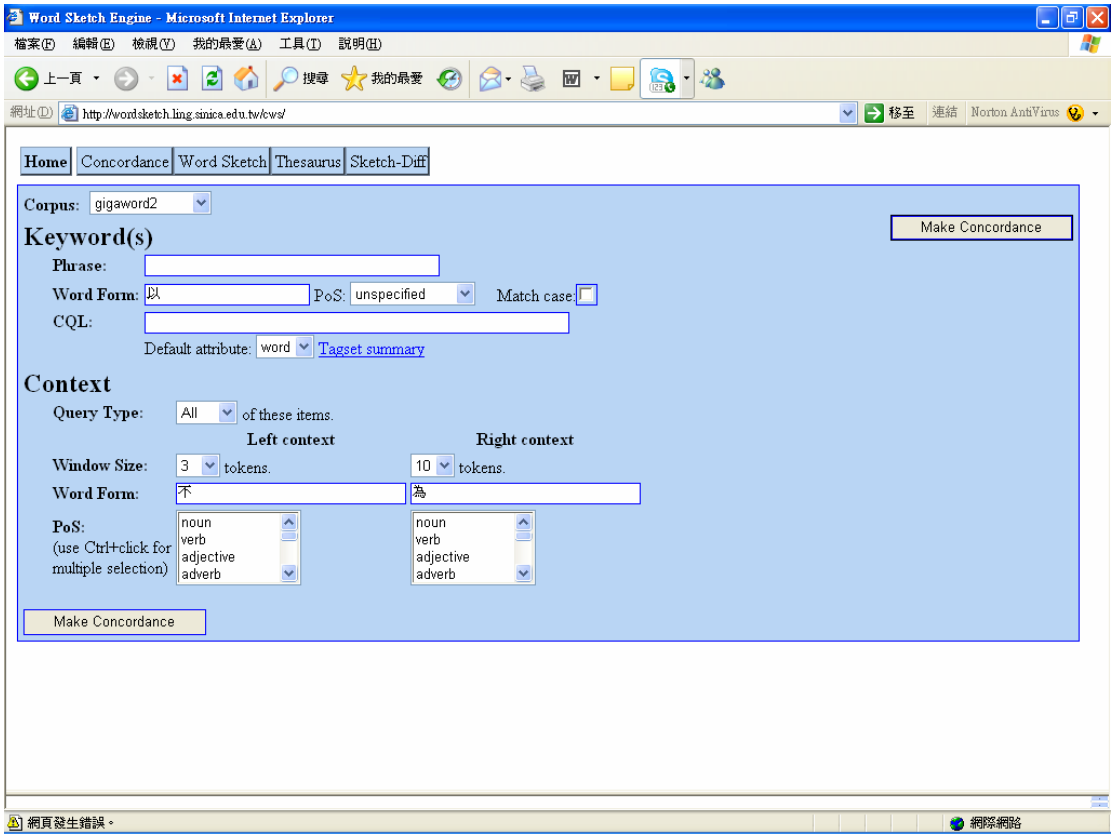


範例二：欲檢索含「連…都」的語料。若如下設定：



在關鍵詞(keyword(s))的 word form 中輸入「連」，
在語境（Context）的 Query Type 中設定 All，
在右側語境（Right context）中輸入「都」，
在右側語境，距離範圍（Window Size）中設定 5 個詞長。
則表示：「擷取關鍵詞「連」右側五個詞的範圍內，出現「都」的語料。」

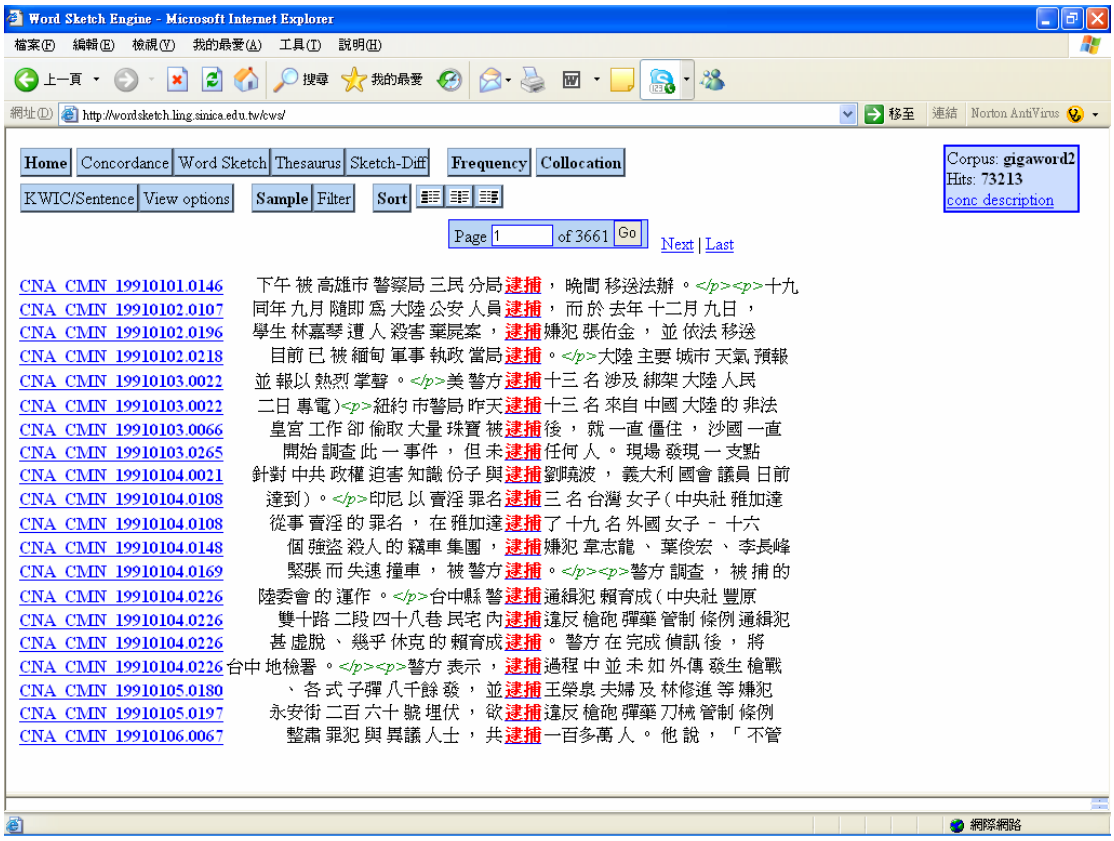
範例三：檢索「以」，且欲過濾掉「不…以…爲」的語料，若如下設定：



在關鍵詞(keyword(s))的 word form 輸入「以」
在 Query Type 中設定 None，
在語境（Context）的左側語境（Left context）中輸入「不」，
在右側語境（Right context）中輸入「爲」，
在左側語境，距離範圍（Window Size）中設定 3 個詞長，
在右側語境，距離範圍（Window Size）中設定 10 個詞長。
則表示：「擷取「以」的語料，但刪除關鍵詞「以」左側三個詞的範圍內，出現「不」，或右側十個詞的範圍內，出現「為」的語料。」

5　關鍵詞(組)（Concordance）語料檢索結果頁面操作

以檢索「逮捕」爲例，關鍵詞或詞組（Concordance）語料檢索結果頁面可能如下：



5.1 左上方第一排五選項

點選上方 Home、Concordance、Word Sketch、Thesaurus、Sketch-Diff 等五選項，可連結至主頁(Home)與其他檢索功能頁面。

5.2 右上方訊息

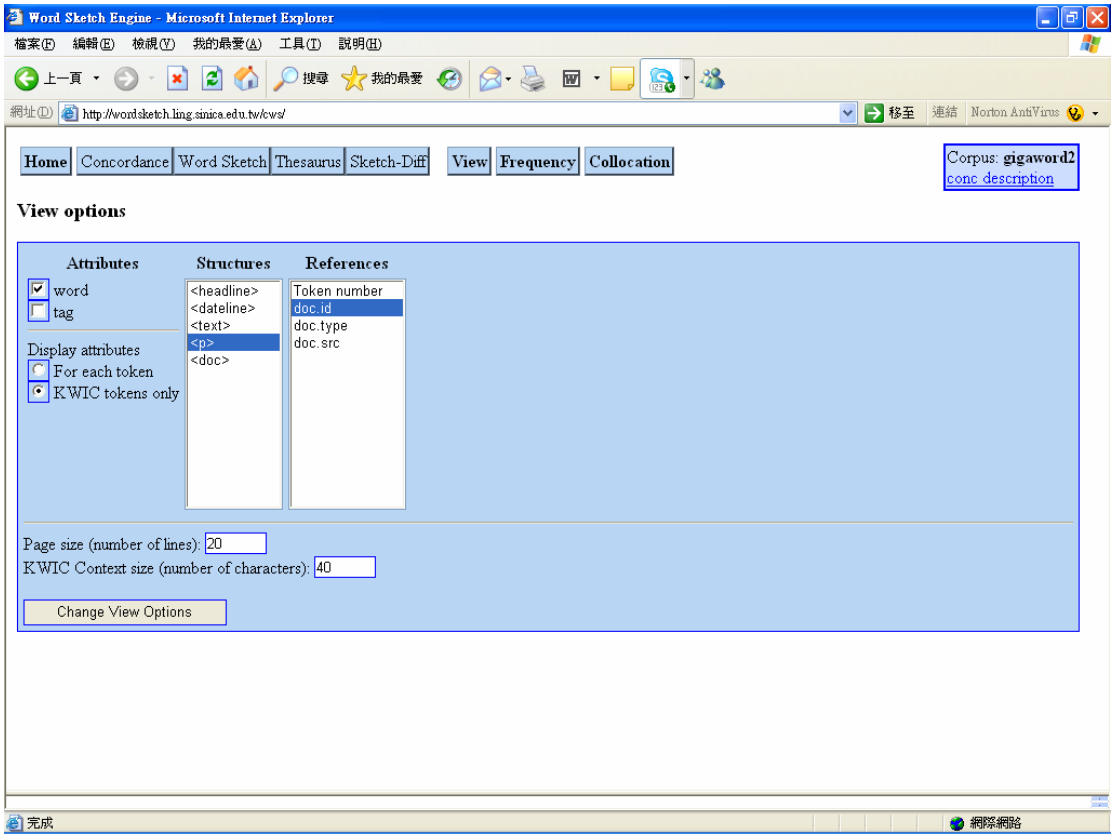
右上方藍色方框內，顯示檢索的語料庫（例如，爲 chinese_giga_trad），以及檢索所得關鍵詞（組）的總筆數，（如，共有 73213 筆「逮捕」語料）。

5.3 左上方第二排選項：**KWIC/sentence**

提供檢索結果顯示的不同選擇，預設值爲「以關鍵詞爲中心」(key word in context （kwic）)的排列方式，呈現檢索結果。如上 5 中之圖所示。
點選「**KWIC/sentence**」，可切換至以整句的方式呈現檢索結果。

5.4 左上方第二排選項：顯示選項(View Options)

提供其他檢索結果顯示的選擇。點選此項，將會出現另一頁面，提供使用者更改結果顯示的設定，包括：



5.4.1 標記 (Attributes)

提供「詞」(word)與「標記」(tag)二種可複選的顯示選項。
在此二選項下可進一步設定所有詞彙(For each token)或者只有關鍵詞(KWIC tokens only)才顯示前述選項。

預設值未點選「標記」(tag)，則語料呈現時不顯示詞類標記。
例如：

今天 下午 被 高雄市 警察局 三民 分局 逮捕 ， 晚間 移送法辦 。

可更改設定選項，如：
(1)若點選「詞」(word)、「標記」(tag)，以及「只顯示關鍵詞標記」(KWIC tokens only)，則會顯示關鍵詞的詞類標記。
例如：

今天 下午 被 高雄市 警察局 三民 分局 逮捕/VC31 ， 晚間 移送法辦 。

(2)若點選「詞」(word)、「標記」(tag)，以及「顯示每個詞的標記」(For each token)，則會顯示句中所有詞彙的詞類標記。
例如：

高雄市/Nca 警察局/Ncb 三民/Nc 分局/Ncb 逮捕/VC31 ，
/COMMACATEGORY 晚間/Ndc 移送法辦/VB12

5.4.2 結構 (Structures)

結構 (Structures)欄提供標記開頭(<)與結尾(/>)的結構顯示，如：
(1) 標題(headline)，例如：<*headline*>大陸 主要 城市 天氣 預報</*headline*>

(2) 時間(dateline) ，例如：*<dateline>*(中央社 紐約 二日 專電)*</dateline>*
(3) 文本(text) ，例如：

*<text>*針對 中共 政權 迫害 知識 份子 與 逮捕 劉曉波 ，
翁山蘇姬 目前 已 被 緬甸 軍事 執政 當局 逮捕 。*</text>*

(4) 段落(paragraph) ，例如：

* *警方 表示 ， 逮捕 過程 中 並 未 如 外傳 發生 槍戰 事件 。
， 今天 下午 被 高雄市 警察局 三民 分局 逮捕 ， 晚間 移送法辦 。* *

(5) 檔案(doc) 例如：

， 並 報以 熱烈 掌聲 。*</doc><doc>*美 警方 逮捕 十三 名 涉及 綁架 大陸 人民 (中央社 紐約

5.4.3 參考資料 (References)

參考資料 (References)顯示在結果頁面的左方，以藍色字體呈現，預設值爲該語料出現的檔案代號(doc.id)，如**CNA19910102.0196**。
除檔案代號(doc.id)外，參考資料 (References)欄提供語料來源的相關訊息顯示設定，如檔案類型(doc.type)，以及檔案來源(doc.src)等等。

5.4.4 頁面長度 (Page Size)

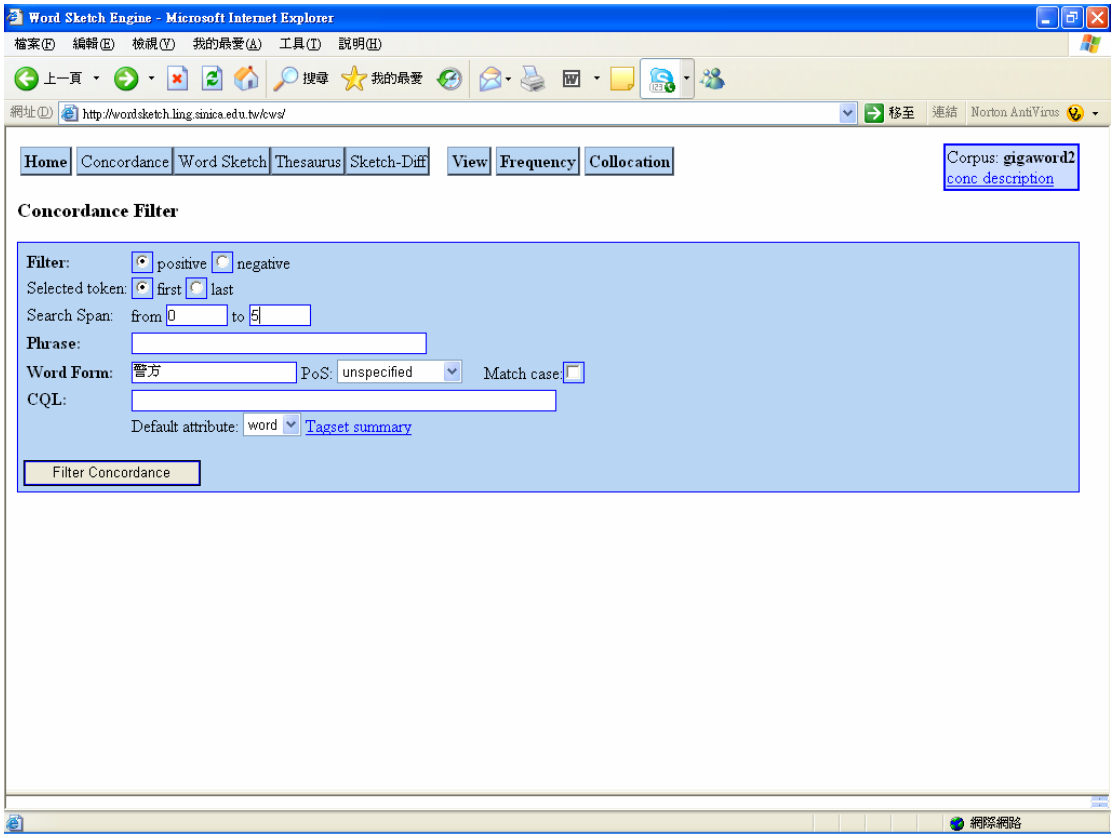
頁面長度(Page size)欄位，提供自由的(行數 number of lines)設定，預設值爲每頁顯示二十行，但亦可設定成每頁顯示十行或者五百行。（每頁顯示行數越多，檢索速度可能越慢。）

5.5 左上方第二排選項：樣本(Sample)

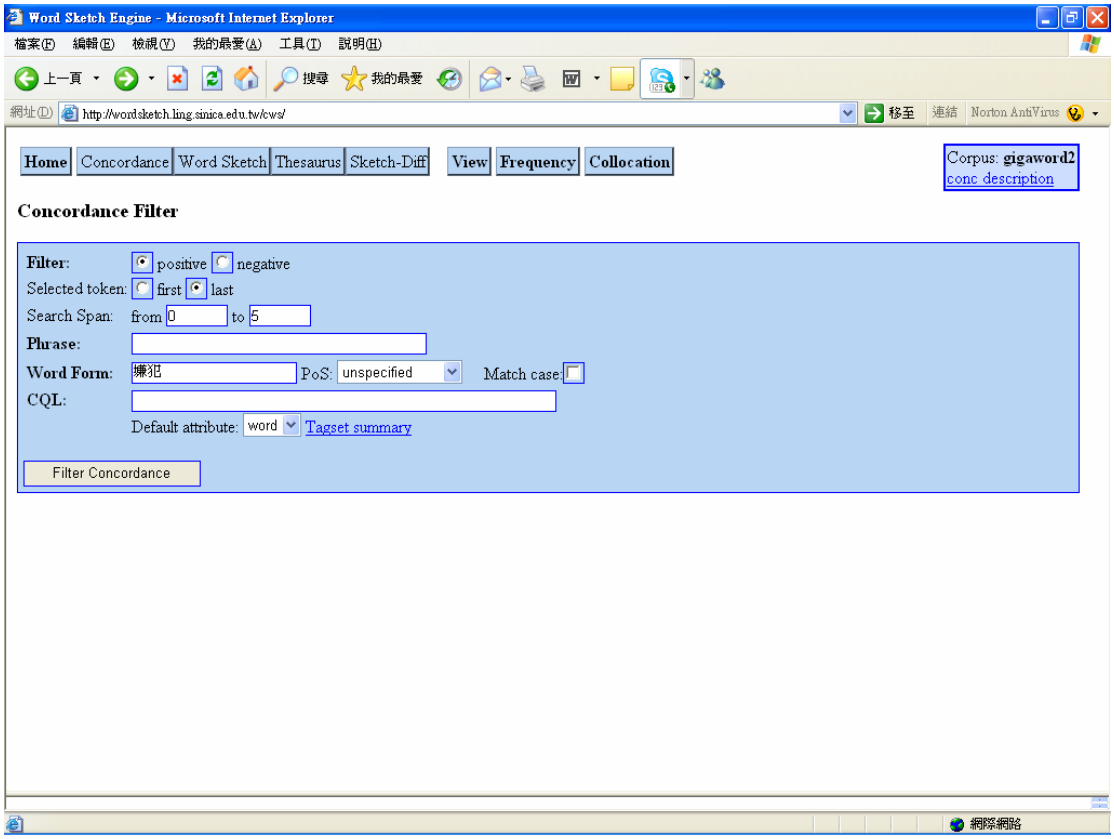
樣本(Sample)功能，提供從多量結果語料中，隨機擷取少量語料的功能。
例如，「逮捕」共有 73213 筆，可點選樣本(Sample)鍵，在「樣本行數」(Number of lines in the sample)中設定 250 或者 500 等，隨機抽取 250 筆或 500 筆語料，進行觀察與分析。

5.6 左上方第二排選項：過濾(Filter)

過濾(Filter)提供多重過濾功能。可設定「擷取」(positive)或者「過濾」(negative)，搭配詞語(Selected token)出現在關鍵詞之前(first)或之後(last)。
搜尋範距(Search Span)亦可自由設定。可輸入與關鍵詞搭配的詞組(phrase)或者詞(word)等目標檢索項目。亦可以語料庫檢索語言（Corpus Query Language（CQL））檢索。例如，若如下設定：

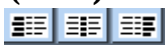


則表示從「逮捕」73213 筆中抽取在關鍵詞「逮捕」左側 5 個詞的範圍內，出現「警方」的語料。則會出現含「警方.....逮捕」的語料 2595 筆。若再如下設定：



則表示從含「警方…..逮捕」的語料 2595 筆中，再抽取在關鍵詞「逮捕」右側 5 個詞的範圍內，出現「嫌犯」的語料。點選「Filter Concordance」後則會出現含「警方…..逮捕….嫌犯」的資料 245 筆。

5.7 左上方第二排選項：排序(Sort)

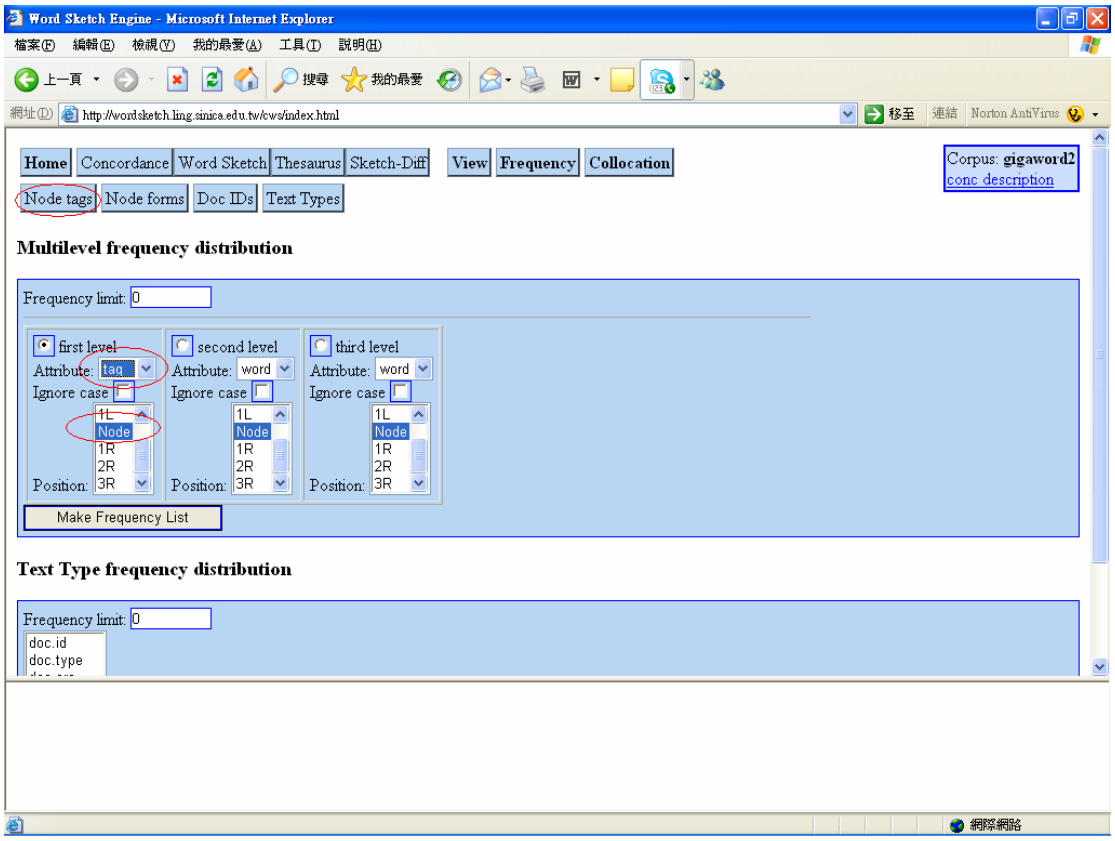
可利用排序(Sort)左方三鍵，  ，簡便設定「依關鍵詞左邊」、「依關鍵詞」、「依關鍵詞右邊」的詞首排序。亦可點選排序(Sort)鍵進階設定排序條件。可依照詞類排序，可設定排序詞數，並可做多重排序選項的設定。

5.8 頻率（Frequency）

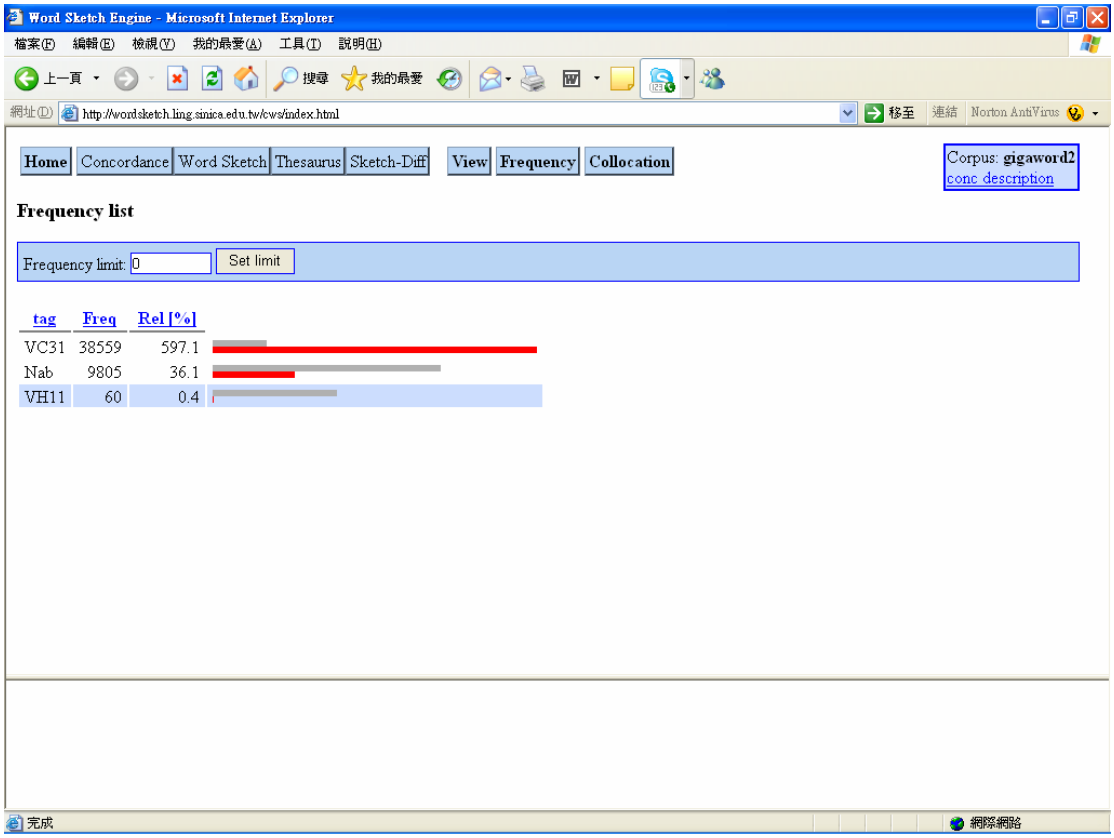
提供二種相關頻率訊息的設定：

(1) 多層頻率分佈(Multilevel frequency distribution)設定

提供關鍵詞在語料庫中的詞類分佈，與前後詞語搭配的頻率等。例如檢索關鍵詞「花」，進入「頻率」（Frequency）設定頁面，點選「關鍵詞詞類」(Node tags)鍵，或者在「多層頻率分佈」(Multilevel frequency distribution)第一層，屬性(Attribute)下拉選單中選「詞類」(tag)，位置(Position)關鍵節點(node)，設定如下：

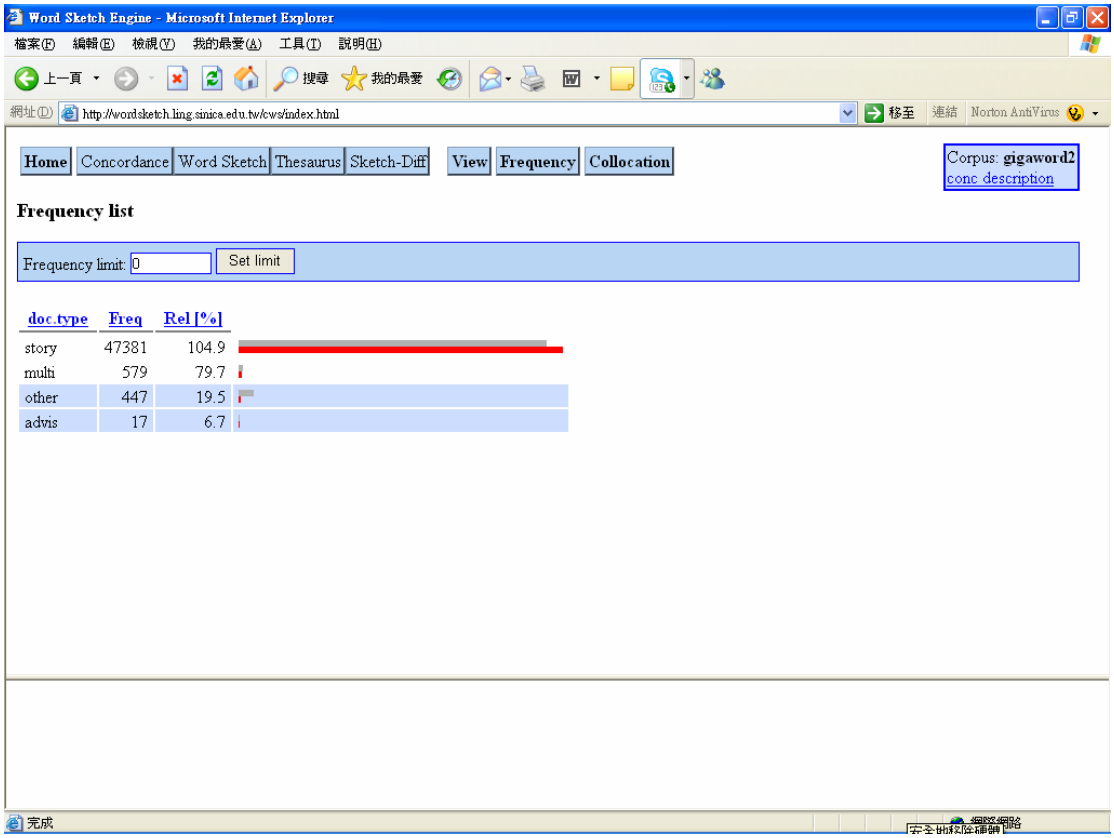


即可得到「花」以下的詞類分佈訊息：



(2) 文類頻率分佈(Text Type frequency distribution)設定

提供檢索的關鍵詞(組)在文本檔案中分佈的頻率訊息，例如檢索關鍵詞「花」，點選「檔案類型」(doc.type)，則可得出如下「花」在不同文類中出現的分佈訊息：



5.9 共現訊息（Collocation）

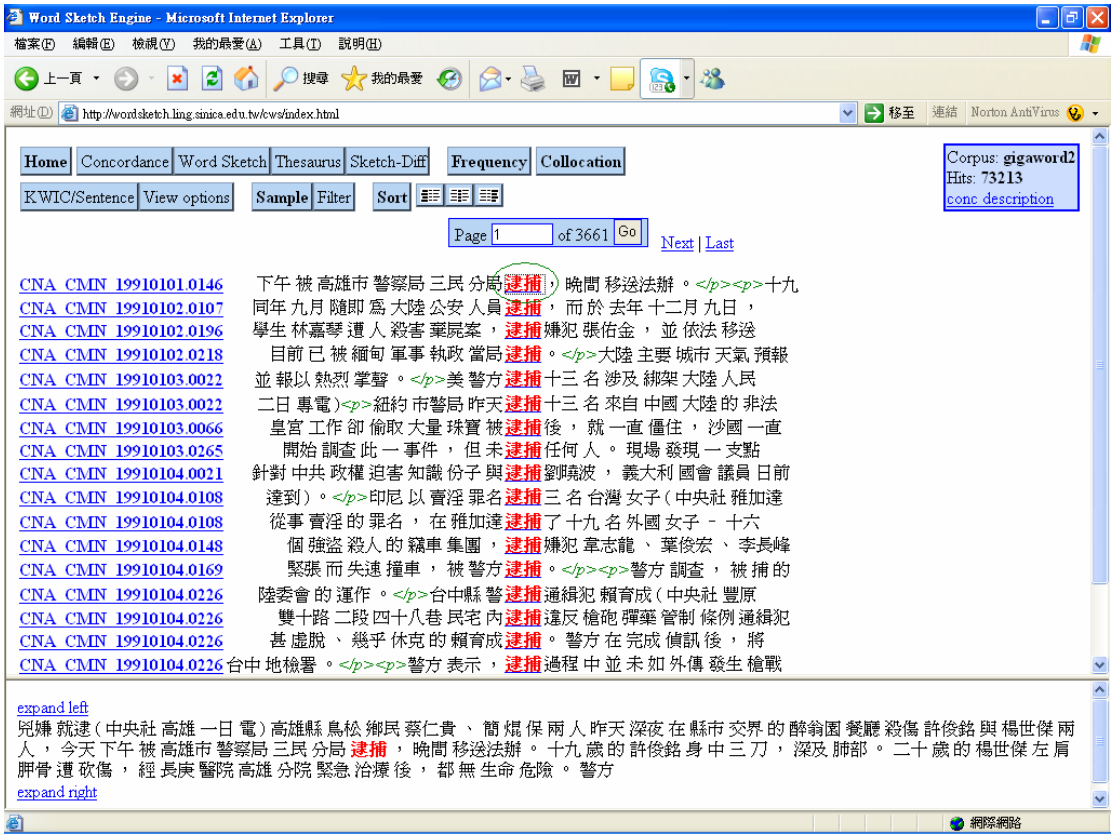
提供與關鍵詞(組)所搭配詞彙的相關共現訊息，包含 T-score、MI、MI3、log likelihood、min.sensitivity，以及 sailence 等值的設定。

5.10 關鍵詞(組)（Concordance）結果頁面的設定

欲檢視其他關鍵詞(組)（Concordance）結果頁面，可在「頁數」(page) 欄位中指定欲檢視的頁數，按「前往」(go)鍵，或者點選「下一頁」(next)、「最後一頁」(last)、「第一頁」(first)，以及「前一頁」(previous)等鍵移動頁面。

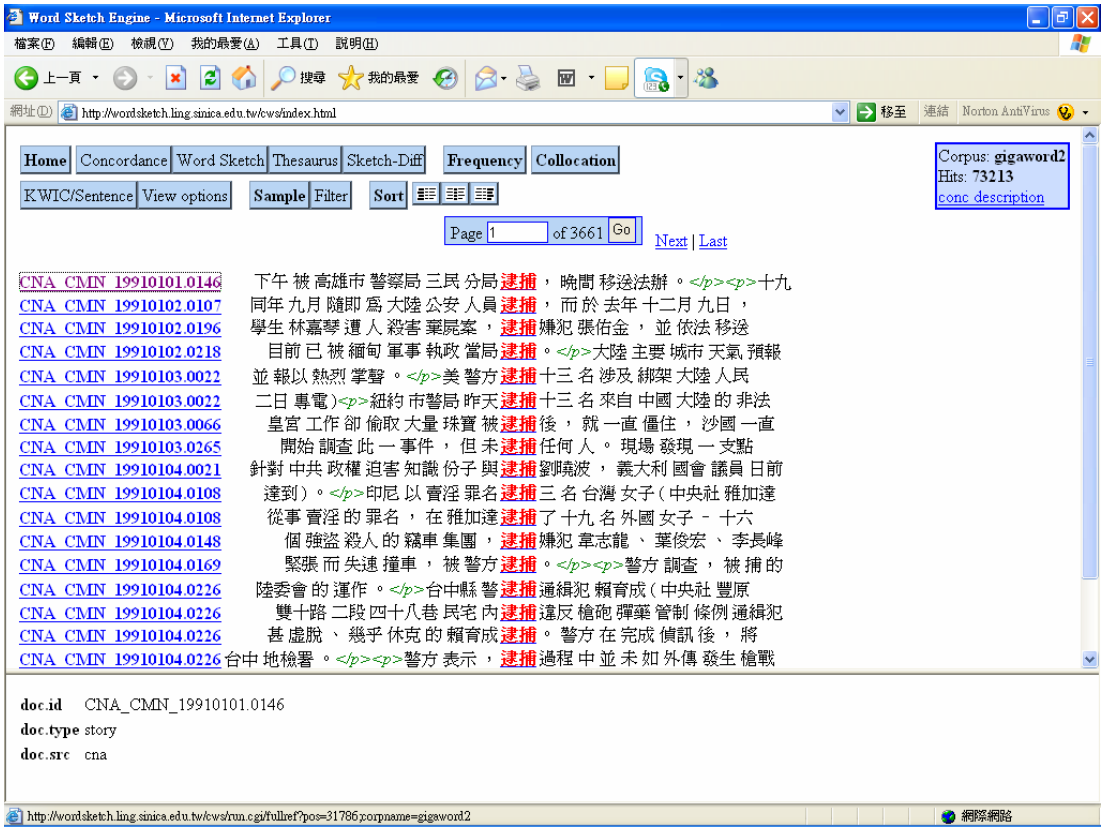
5.11 察看某行關鍵詞(組)（Concordance）結果

點選結果頁面中紅色關鍵詞(組)，例如點選第一行「逮捕」，會在頁面下方顯示更多的語境，如下：



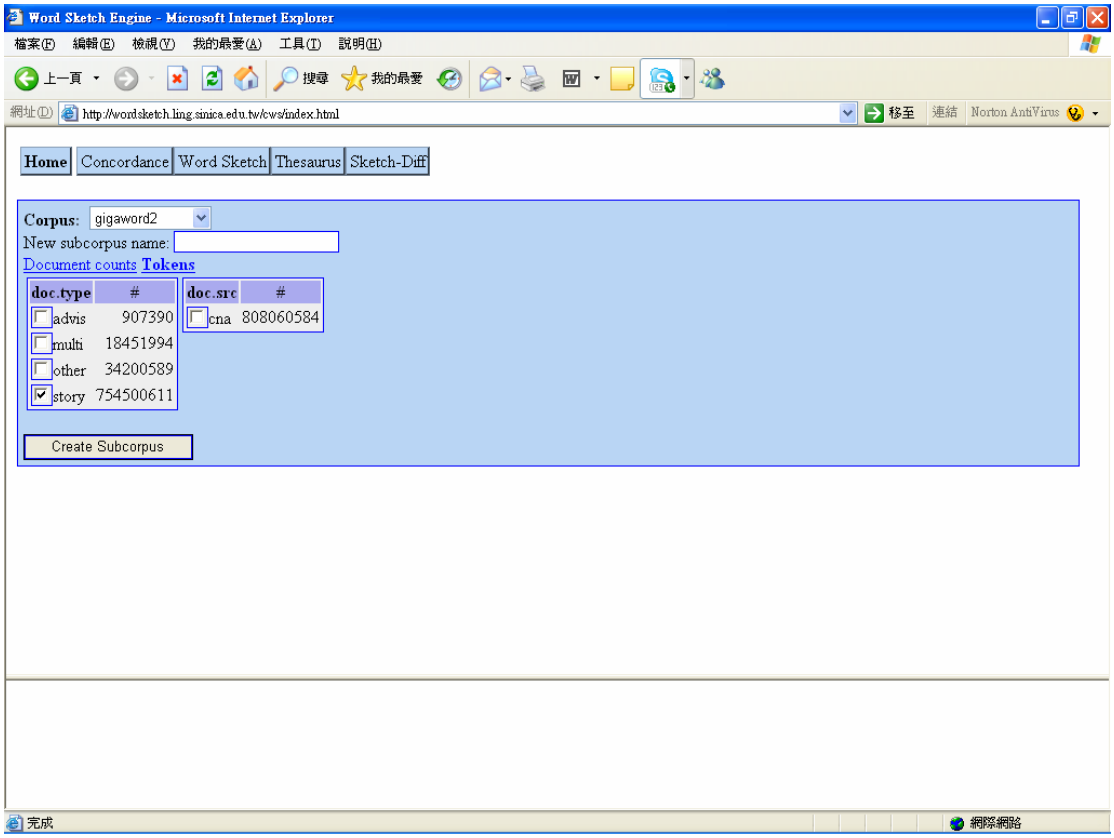
並且點選「擴展左側」(expand left)鍵，或「擴展右側」(expand left)鍵，可以繼續擴展左側或右側語境。

欲知某一筆語料的文本來源訊息，可點選結果頁面左側藍色字體有關文件檔案訊息的部分，例如在以上檢索「逮捕」的頁面中，點選第一行語料左側的CNA19910101.0146，則會在頁面下方，顯示檔案代號(doc.id)、檔案類型(doc.type)，以及檔案來源(doc.src)等訊息如下：



6 建立次級語料庫(Creating a subcorpus)

如果您想檢索某個語料庫的一部份，（例如繁體字版的 gigaword2，只想檢索 story 類型的文本），可在檢索主網頁中，（或者點選「首頁」(Home)鍵，回到主頁面），再點選「建立次級語料庫」(**Create subcorpus**)，在以下視窗中點選，並輸入該次級語料庫的名稱即可。

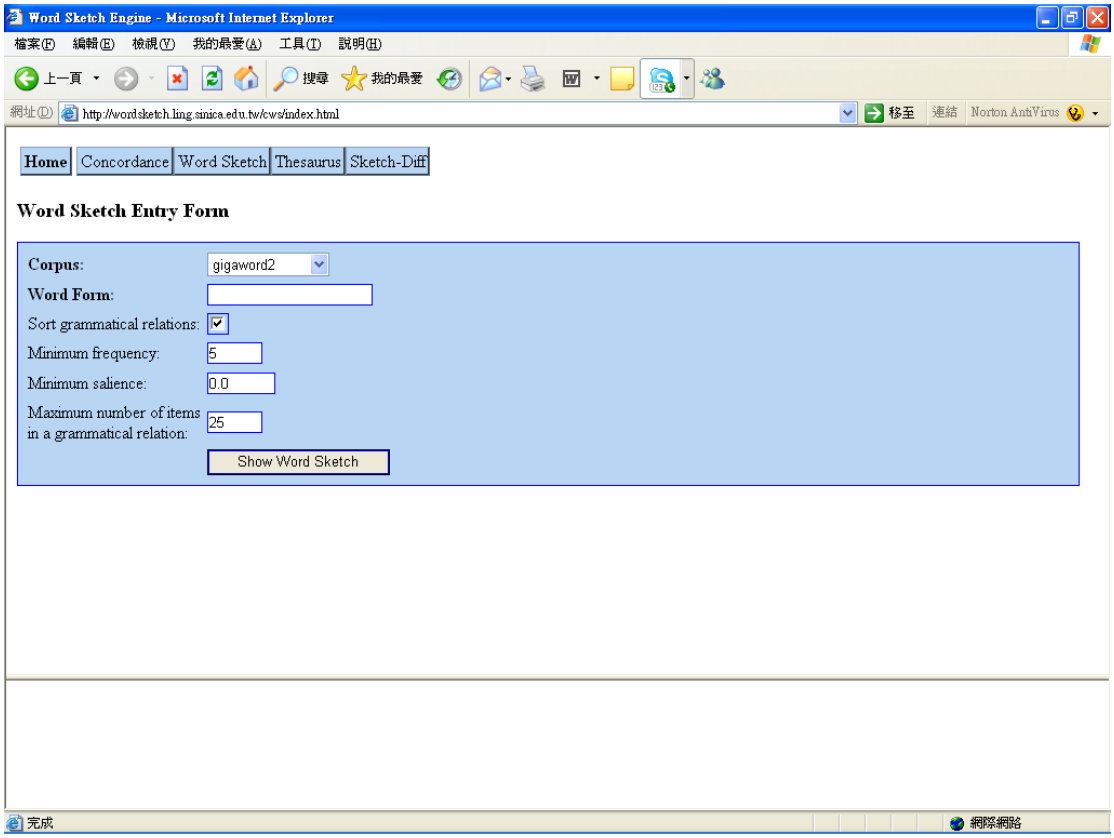


如欲檢索該次級語料庫，可在檢索頁面的「語料庫」(Corpus) 下拉選單中，選取該語料庫，進行檢索。

如欲刪除該次級語料庫，可在檢索首頁中（點選「首頁」(Home)鍵，回到主頁面），點選「刪除次級語料庫」(Delete subcorpora)，再點選欲刪除的語料庫名稱，進行刪除即可。

7　詞彙特性速描（Word Sketch）功能

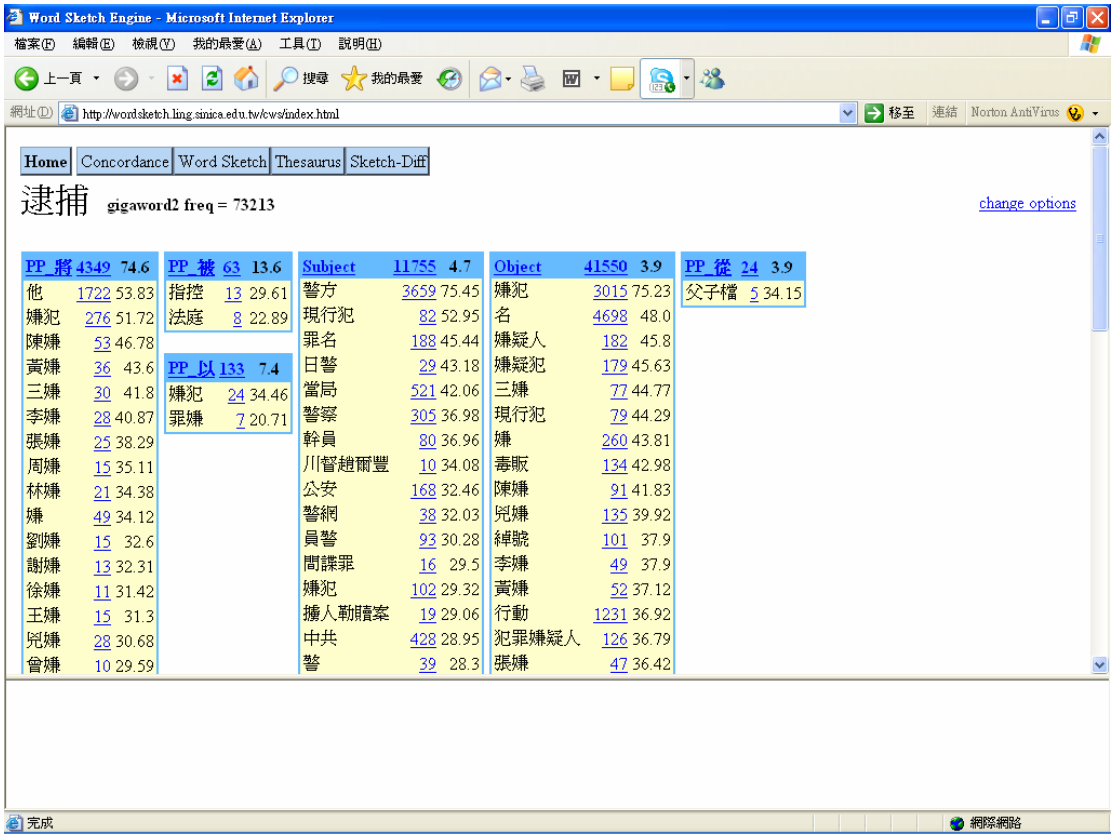
「詞彙特性速描」（Word Sketch）是以語料庫爲本對某個詞彙的語法以及和其他詞語搭配表現的一覽表。點選首頁中的「詞彙特性速描」（Word Sketch），會出現以下頁面：



可在「語料庫」（Corpus）的下拉選單中，點選所欲察看的語料庫，在「詞彙形式」(Word Form)中輸入欲察看的詞彙，可選擇是否按照語法關係排序（Sort grammatical relations）、設定搭配詞語出現最低的頻率限制（Minimum frequency）、最低的顯著性（Minimum Salience），以及某個語法關係中最大的詞項數量（Maximum number of items）限制。

此功能適用於大型語料庫，若語料庫太小，則不適用此功能。

若在預設值設定下，輸入「逮捕」查詢，則會出現以下結果頁面：



以上頁面中，顯示「逮捕」的語法搭配關係，如賓語(object)、主語「subject」、修飾語（modifier），以及搭配的「以」字介詞組等。每個欄位下，顯示搭配詞，與搭配詞出現的總頻率，與搭配詞出現的顯著性。例如「嫌犯」出現當「逮捕」的賓語共有 276 筆，其顯著性爲 51.72。以上結果頁面，按照搭配詞的顯著性，由高至低排序。

「詞彙概述」（Word Sketch），由統計方式自動抽取搭配詞關係，如其他檢索系統一樣，允許某些程度的噪音(noise)，即可能擷取到不正確的結果。

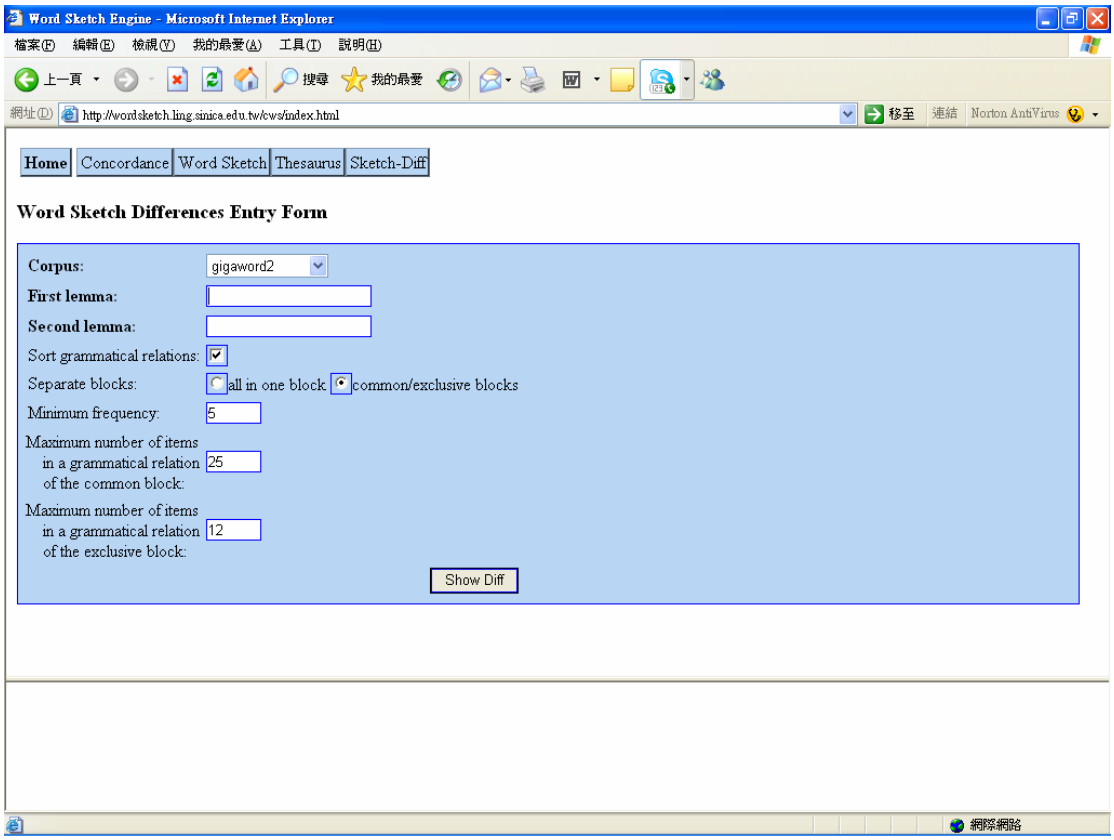
使用者可以隨時點選頁面上排的功能鍵，轉換到其他功能檢索頁面。

8 同義辭典(Thesaurus)功能

提供在搭配詞語上，與關鍵詞表現相近的詞彙訊息。

9 詞彙素描異同（Word Sketch Difference）功能

提供二個詞彙的語法功能、搭配詞語異同的訊息。並且以顏色顯示此二比較詞彙，在搭配某個詞彙上的顯著程度。可點選此功能鍵，進入以下頁面：



在「第一個詞彙」(the first lemma)，與「第二個詞彙」(the second lemma)欄位中輸入所欲比較詞彙，可細部設定所欲查詢的條件，以及顯示的方式。

- **網路資源：**

The Word Sketch Engine: http://www.sketchengine.co.uk/
Chinese Gigaword 訊息:
http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2003T09