# 「西班牙語詞語搭配工具」Spanish Collocation Tool 簡介

「西班牙語搭配詞工具」由成功大學資工所「網路探勘暨跨語知識系統實驗室」支援技術所開發，目的是為了協助西班牙文語搭配詞的研究分析。本工具是以C#程式語言所撰寫，可選擇計算的統計法及設定被搭配成分的距離範圍。不同於一般檢索軟體，該工具可檢索的被搭配元素除了單詞外，還包括詞類、字根其彼此的組合。

## Spanich Collocation Tool 安裝與使用說明

### (一). 安裝步驟

**1.** 安裝 **Microsoft .NET Framework 3.5 版或 3.5 版之後**

- 點選下面網址,請按下載並安裝,

- 連結網址: http://www.microsoft.com/zh-tw/download/details.aspx?id=21

**2.** 安裝 strawberry.perl

- 點選下面網址,請依照個人電腦的系統選擇 **32bit** 或 **64bit** 的版本

(若不清楚電腦的系統,請對「**我的電腦**」按右鍵選內容,64bit 會看到「x64 Edition」字樣,若無出現則是 32bit)

- 安裝時軟體會自動放進(c: )資料夾
- 安裝結束後請到桌面,對「**我的電腦**」點擊右鍵選「**內容**」再點選「**進階**」進入「**環境變數**」中,並在「**系統變數**」中找「**path**」**(點兩下)**按「**編輯**」後,加入 「**C:\strawberry\perl\bin**」

- 連結網址: http://strawberryperl.com/

**3.** 安裝 TreeTagger

- 請點選下面網址,進入後下拉網頁,找到粗體標題 **Download**

並在第 14 行 **"1.Download the tagger package for your system"**的後面點選 PC-Linux 下載 (電腦系統為 64bit 的使用者,請選擇 PC-Linux (64 Bit)。)

- 下載後,請解壓縮該檔案,解壓縮後會出現 **TreeTagger 資料夾**,請將它移至(c: )資料夾
- 繼續在同樣的網址上找粗標題「**Parameter files (for PC)**」
- 選擇「**Spanish parameter file**」下載,解壓縮後將檔案放到剛剛安裝的 **TreeTagger** 的「**Lib 資料夾**」中。

- 連結網址:

**已註解 [u1]:** 步驟 1,2 是環境; 步驟 3,4 是工具(tool)中的一部分

**已註解 [u2]:** 通常安裝後都會成功, 若不確定是否安裝, 可至"控制台"-> 點"系統及安全"->找"檢視已安裝的更新"確認!

**已註解 [u3]: 方法 1:** 若此程式在其他電腦中已經有下載好的檔案,當以後需要安裝時,請直接至有此程式的電腦中的 C 槽將 TreeTagger 的資料夾複製到要安裝的電腦中的 C 槽即可! 所以就可以不用再按照下列步驟去下載此程式。

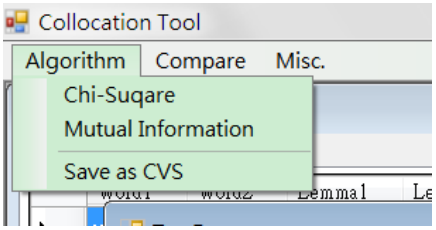
http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html#Linux

**4.** 在安裝完後上述的軟體或程式後，再將 **Collocation Tool** 的資料夾存至電腦中即可使用 SCT。

(二).操作說明

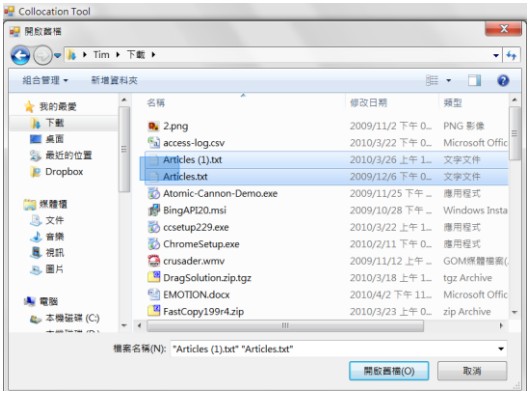
•點選進入 Collocation Tool 資料夾中的"Collocation"

- 點選右上角的 **Algotithm**,再按 **Chi-Suquare**




- 選取要匯入的檔案

使用者可以同時匯入多個語料,方便之後在 KL-divergence 比較兩個語料庫搭配清單的差異性




* 如果 SCT 有成功安裝，匯入 txt 檔後，會跑出 collocation 的列表

若有列表跑不出來的情況，可能的原因大概有兩種：

1. 匯入的 txt 檔案太大，所以需要一段時間跑出列表。
2. 中間的安裝步驟可能沒有成功，需要重頭檢查看看程式下載的過程中是否有地方出錯或遺漏了。

(1) 針對 Microsoft .NET Framework 3.5 版或 3.5 版之後，可再按照註解[u2]，確認是否有成功安裝?

**已註解 [u4]: 方法 2:** 直接連結網址下載此程式,若電腦屬於 Windows->請連結此網址!

http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/#Windows

➔找到粗體標題"**Windows version**"

➔點選"**here**"

➔下載後請解壓縮該檔案

➔解壓縮後會出現 TreeTagger 資料夾,請將它移至(c: )資料夾

➔繼續在同樣的網址上找粗標題"**Parameter files (for PC"**

➔選擇「**Spanish parameter file** 」下載,解壓縮後將檔案放到剛剛安裝的 TreeTagger 的「**Lib** 資料夾」中即可

#要特別注意的是在下載 TreeTagger 後,若發現其解壓縮後的檔案資料夾名稱不是 TreeTagger,請自行將檔名改成 **TreeTagger**! (因為若資料夾檔案名稱不是 TreeTagger 的話,在使用 SCT 時就會遇到無法開啟使用的情況。)

**已註解 [u5]:**

(2) 針對 strawberry.perl，請確認 **Strawberry** 的資料夾有放進 c 槽中，並確定檔案名稱是 **Strawberry**？ 接著，再到桌面對「我的電腦」點擊右鍵，選「內容」，再點選「進階」(或是「進階系統設定」)進入「環境變數」中，並在「系統變數」中找「path」(點兩下)按「編輯」後，加入「C:\strawberry\perl\bin」(即重做一次 "SCT 使用說明"檔中的步驟)。

(3) 針對 Tree Tagger，請確認 **TreeTagger** 的資料夾有放進 c 槽中，並確定檔案名稱為 **TreeTagger**？

• 語料庫匯出畫面後,可以在畫面的右邊「**Filter**」的地方出入指令並按「**Apply**」即可顯示結果

指令舉例:

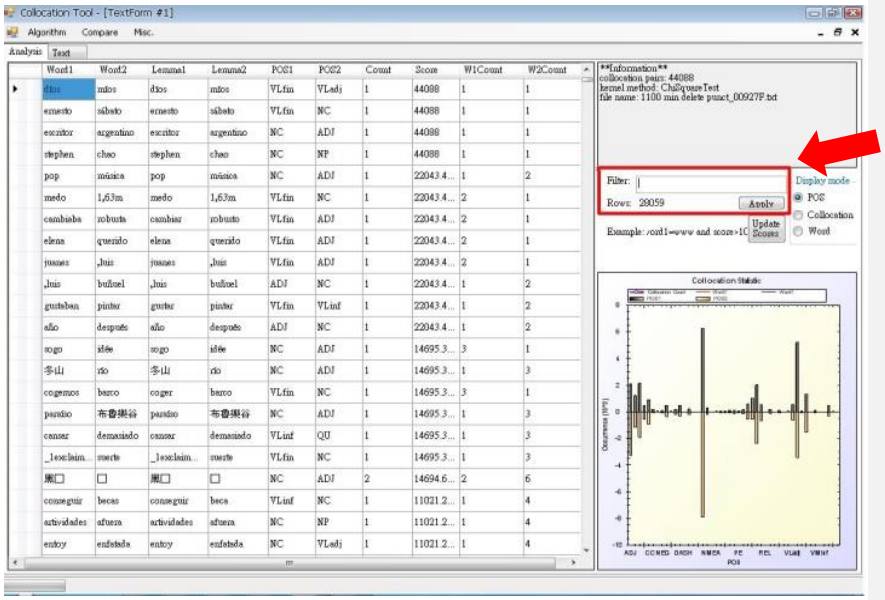
Lemma2=Yo and Score>2

W1count>2 or w2count>2

Pos1=NC

Pos1=NC and Pos2=NP and word1=ahora

(圖一)




**畫面介紹:**

上圖左邊部分是 collocation 的列表，每個 collocation 是由 Word1 和 Word2 所組成，lemma1 和 lemma2 分別為其原形，POS1 和 POS2 分別為其詞類，Count 代表該 collocation 在語料庫中出現的次數，Score 為演算法的結果，W1Count 和 W2Count 是該字在語料庫中出現的次數。

右上方的部分，分別記載了偵測到的 collocation 的個數(即左邊 Count 欄位的總和)、使用的演算法和載入的語料庫檔案。

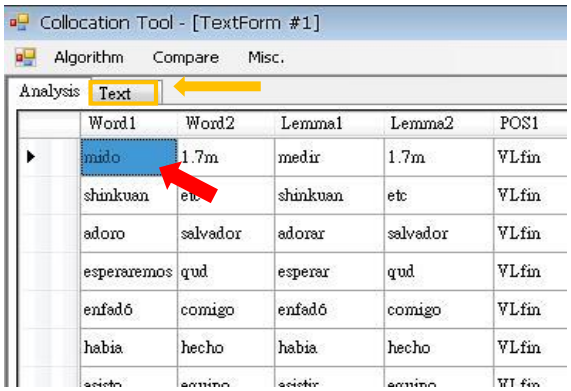
右邊中間的部分，可以用來過濾或篩選 collocation，

(三).其他功能說明

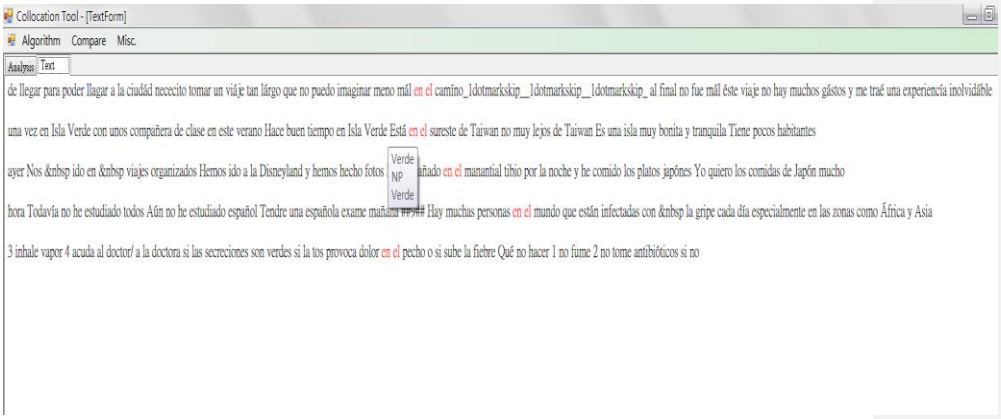
**1. 詞彙與原文對照**:

- 選定畫面 Analysis 中 word1 或 word2 裡的其中一個單字,連續點擊滑鼠左鍵兩次之後(圖二紅箭頭的地方),再點選左上角的 Text(圖示一黃箭頭的地方)。
- 點進之後會顯示剛剛點選的單字(紅字)在文章中出現的地方(圖三), 將滑鼠移到字上面可以觀看其 lemma 和 POS。

(圖二)

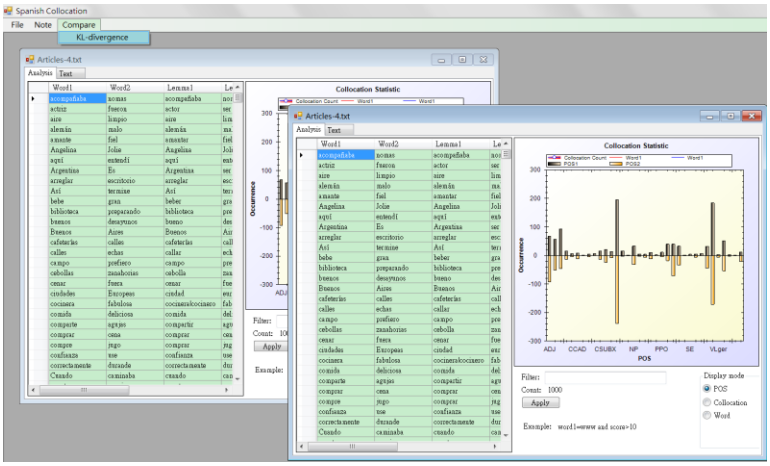



(圖三)

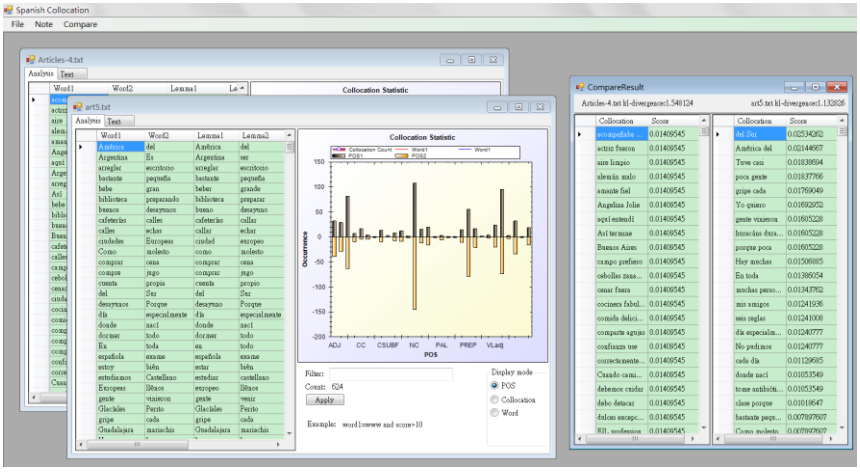


**2. KL Value:** 使用 KL value 做比較時,要先確定 collocation tool 裡面至少已經有兩筆資料,這樣才可以進行做比較的動作。
點選(圖四)最上排的第二個選項 **Compare** ,接著點擊 **KL-Divergence**。
請稍後幾秒,畫會跳出另一個視窗

(圖四)



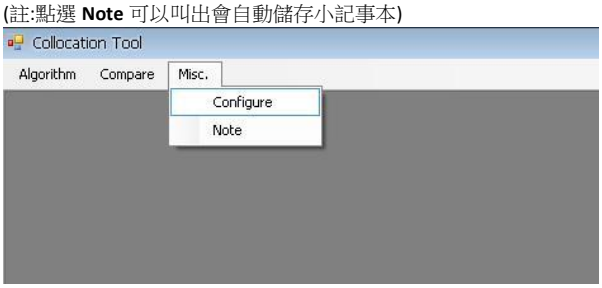

左方跟右方分別是兩個語料庫的搭配清單，依照分數排序，分數(的絕對值)越高代表差異性越顯著

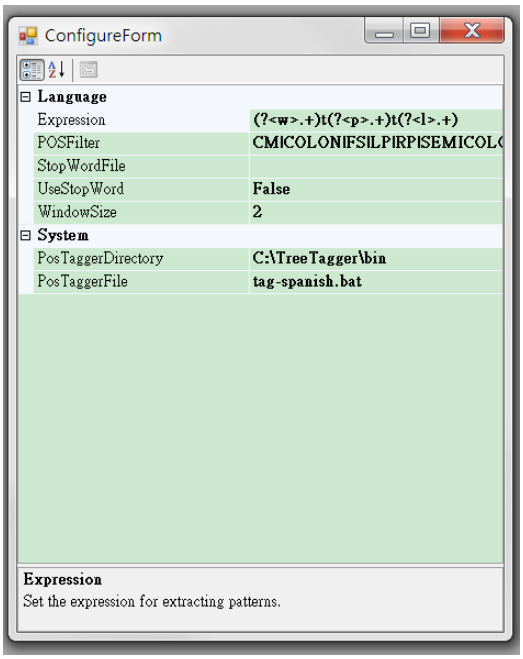



### 3. 參數設置

設定參數點選最上排的 **Misc.**再選 **Configure**

(註:點選 **Note** 可以叫出會自動儲存小記事本)






• Expression 為正規表示式,斷詞工具書出的結果有關,其中<w>,<p>和<l>三個偵測出來的 pattern 分別是 Word、POS 和 Lemma。
• POSFilter 是代表欲過濾掉的詞性,StopWordFile 和 UseStopWord 目前保留(尚未實作)。
• WindowSize 為 Collocation 擷取時在一個句子內可以橫跨的字數。
• PosTaggerDirectory 是斷詞工具的執行檔目錄，PosTaggerFile 是斷詞工具的執行檔名稱。

**4. 儲存檔案**

使用者可以將資料庫匯出的檔案存成 Excel 檔

點選左上角的 Algorithm 之後選擇 Save as CVS 即可完成存檔的動作