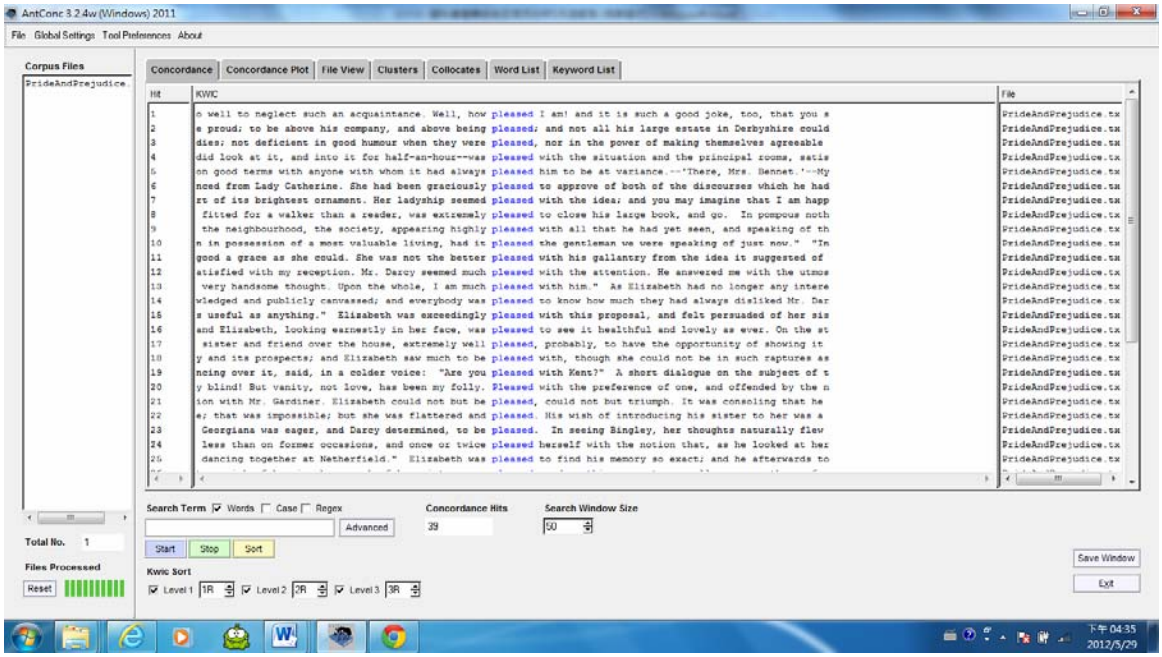
中文句法樹庫（Penn Chinese Treebank ）(http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2004T05)。兩者在語言，語料來源，語料庫大小，標記集，標記單位，標記訊息，及依據的語言學理論都不相同。

Sinica Treebank 與 Penn Chinese Treebank 最大的差別在於結構樹的語法單位不同。前者以標點符號作為分隔不同結構樹的單位，因此一個結構樹很多時候只是一個詞組（如 PP, NP）而不是一個完整的句子。而後者除小部分結構樹是句子的片段（以 FRAG 標示）大部分的結構樹是完整的句子(sentence)(以 IP 標示)。另外 Sinica Treebank 語法結構採取中心語主導原則 (Head-Driven Principle)，註明中心語(Head)和其他成分（如附加語）的語法和語意訊息，表達出句子中詞和詞之間的語法結構和語意角色關係，而 Penn Chinese Treebank 並沒有中心語與語意角色的訊息，而是在詞組上加註如主詞 SBJ 受詞 OBJ 等語法功能的方式來取代。

由於 Sinica Treebank 有未簡化標記，簡化標記及精簡標記三種標記集，相較於 Penn Treebank 只有一種標記集，Sinica Treebank 的三種不同的標記集可以作為不同的特徵。除此之外只有 Sinica Treebank 有標示語意角色的訊息，Penn Chinese Treebank。由 Linguistic Data Consortium (LDC)所發行，其中標示語意角色的 Penn Chinese Treebank 稱為 Chinese Proposition Bank。

參、語料庫語言學的工具

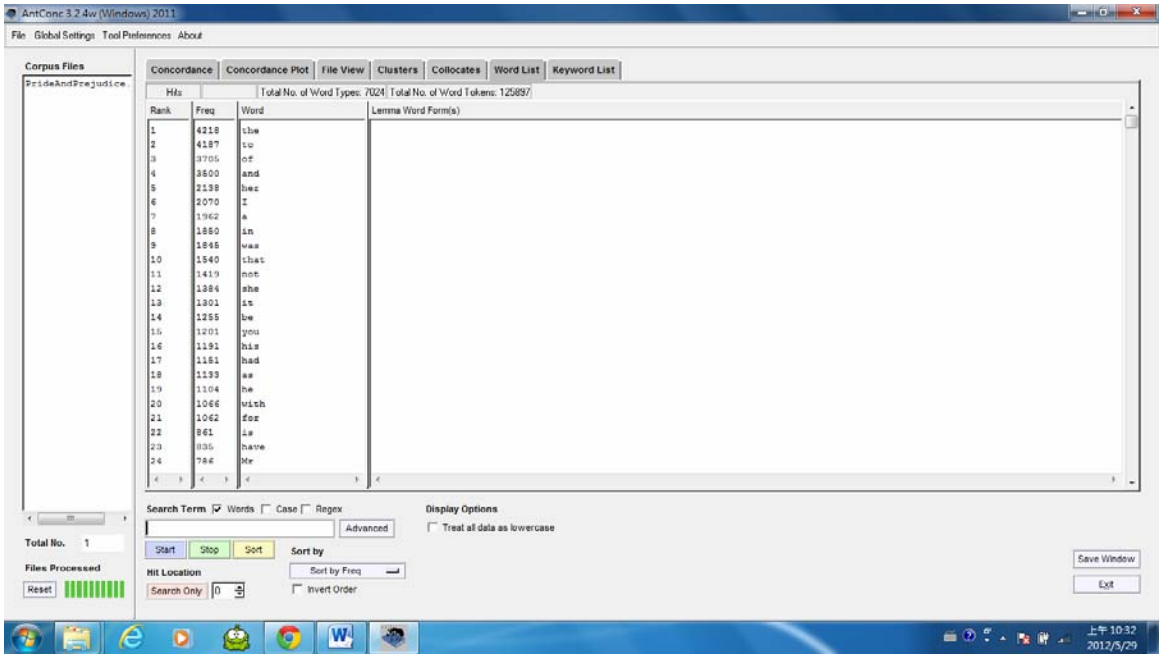
一、關鍵詞前後文程式(concordancer)：輸入一個關鍵詞或字串，程式自動將語料庫中所有包含這個詞或字串例子找出來置中並顯示前後語境。Antconc 是一個免費軟體，可以計算語料關鍵詞的頻率，並檢索關鍵詞以及搭配語。下面的畫面擷取自 Antconc 關鍵詞上下文檢索程式 concordancer 的功能。



圖一 AntConc 關鍵詞前後文排序程式

http://www.antlab.sci.waseda.ac.jp/antconc_index.html

二、詞頻程式：計算某個特定的字串或每個出現在語料庫中的詞的頻率。如上面 Antconc 內建 concordancer 功能，搜尋某一個關鍵詞時，下方 Concordance hits 會顯示這個關鍵詞在這個語料庫出現幾筆。如下圖，點選 Antconc 上方 Wordlist 即可計算每個出現在語料庫中的詞的頻率，且會依照頻率高低排序。



圖二 AntConc 詞頻排序程式

http://www.antlab.sci.waseda.ac.jp/antconc_index.html

三、英文還原詞原型程式(lemmatizer)：輸入一個英文詞，程式自動將句中的每一個詞詞轉為原形。

四、中文分詞程式：輸入一個句子，程式自動找到詞與詞的界線並將詞分開。由於人名，地名，及具有衍生性的詞無法全部列舉在辭典中，在加上分詞程式無法完全解決岐義的問題，中文分詞程式的準確率大約只有 90%到 97%。中文最簡單的分詞演算法是長詞優先，但如下例有時會造成錯誤。

例如輸入：把手舉起來。

輸出：把手　舉　　起來。

最具代表性的正體字分詞程式是中研院詞詞知識庫小組的分詞程式。利用機器學習演算法發展出來且可以自由下載的簡體字中文分詞程式有 LingPipe http://alias-i.com/lingpipe/demos/tutorial/chineseTokens/read-me.html 以及史丹福大學的 Chinese Word Segmenter http://nlp.stanford.edu/software/segmenter.shtml。若要使用簡體字中文分詞程式處理正體字需先轉成簡體字，程式處理完再轉回正體字，在繁簡繁三道轉換過程，有些字可能會轉錯。

五、詞類標記程式(part-of-speech tagger):程式自動將輸入的句子的每一個詞標上詞類。目前英文的詞類標記程式可達到 98%以上的正確率,如 Stanford Parser。繁體中文的詞類標記程式以中研院詞庫小組以最具代表性。中研院詞詞知識庫小組的分詞程式以及史丹福大學的 Chinese Word Segmenter 都可以同時處理分詞和詞性標記，但兩者的分詞標準和詞性標記集(tagset)不同。

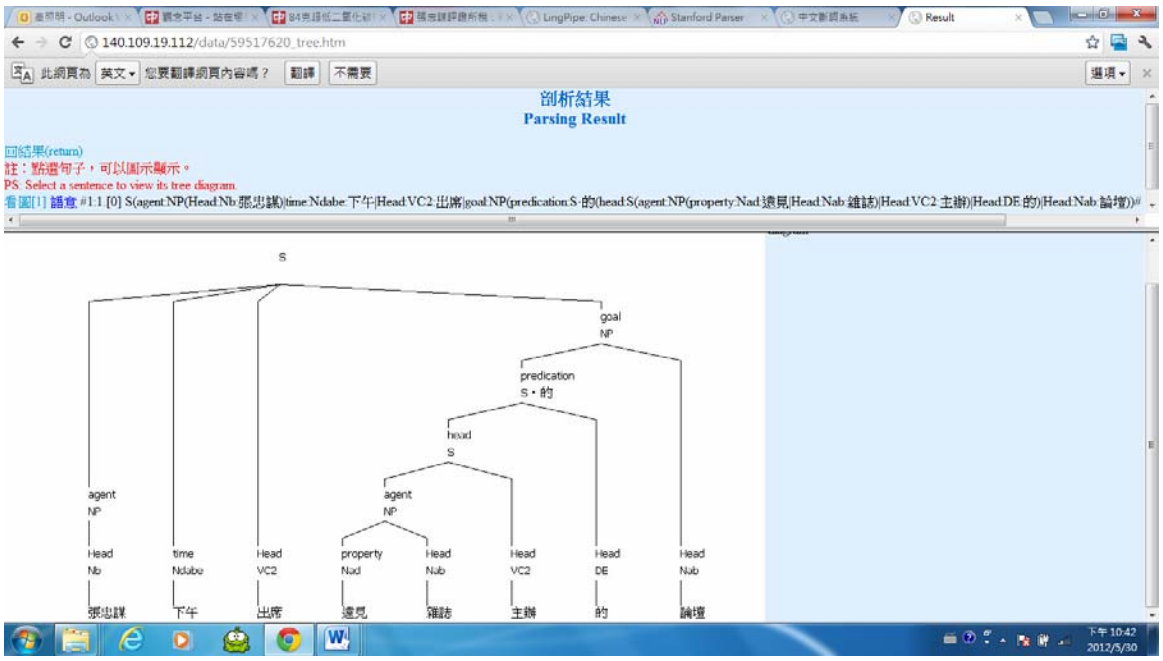
圖三 中研院詞詞知識庫小組的分詞和詞性標記程式
http://ckipsvr.iis.sinica.edu.tw/

六、語法剖析器(parser)：程式自動將輸入的句子的句法層次結構標示出來。語法剖析器可以分成兩種，完全剖析和部分剖析(partial parse)。近年來興起能判斷依存關係的語法剖析器,如英文的 Minipar 及 Stanford Parser。瑞典 Lund 大學以 Mate-tool 為基礎發展簡體中文的語法剖析器提供程式碼供研究人員下載。

Sinica Treebank 與 Penn Chinese Treebank 最大的差別在於結構樹的語法單位不同。前者以標點符號作為分隔不同結構樹的單位，因此一個結構樹很多時候只是一個詞組（如 PP, NP）而不是一個完整的句子。而後者除小部分結構樹是句子的片段（以 FRAG 標示）大部分的結構樹是完整的句子(sentence)(以 IP 標示)。另外 Sinica Treebank 語法結構採取中心語主導原則（Head-Driven Principle），註明中心語(Head)和其他成分（如附加語）的語法和語意訊息，表達出句子中詞和詞之間的語法結構和語意角色關係，而 Penn Chinese Treebank 並沒有中心語與語意角色的訊息，而是在詞組上加註如主詞 SBJ 受詞 OBJ 等語法功能的方式來取代。

如（圖五）所示，中研院的中文句法樹庫的 terminal node 是詞，詞上方有詞性標記和中心語（head）這類的語法訊息，構成詞組的結點(node)有詞組標記和語意角色等語意訊息。

圖四　中研院詞詞知識庫小組中文剖析器的輸入介面

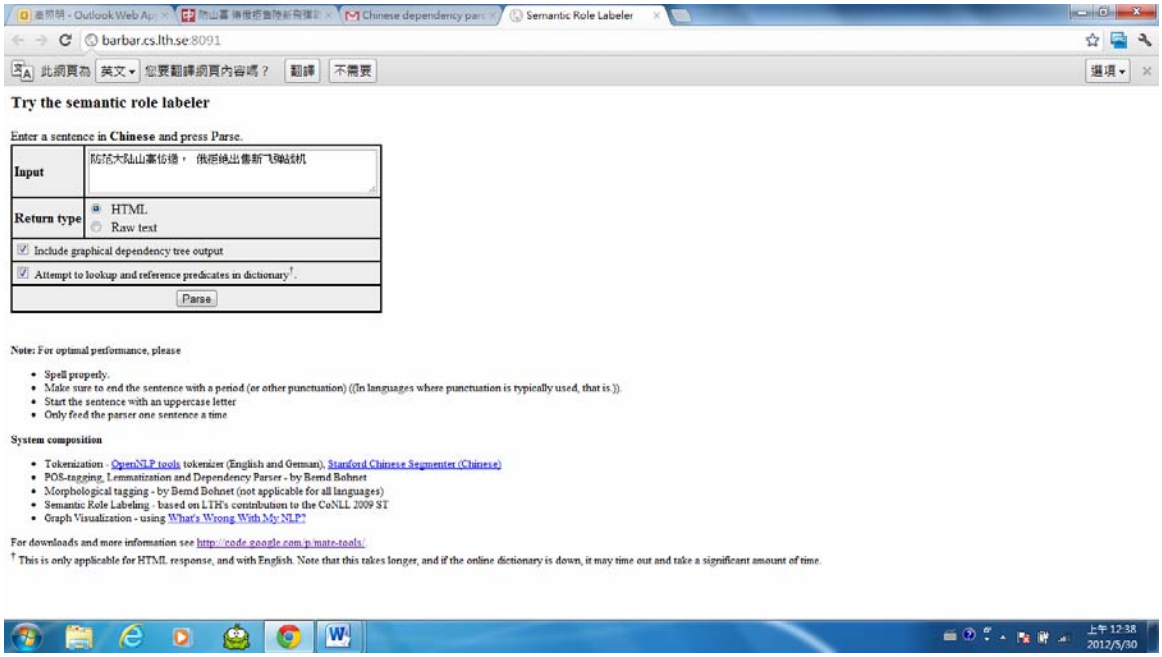


圖五　中研院詞知識庫小組中文剖析器的輸出介面

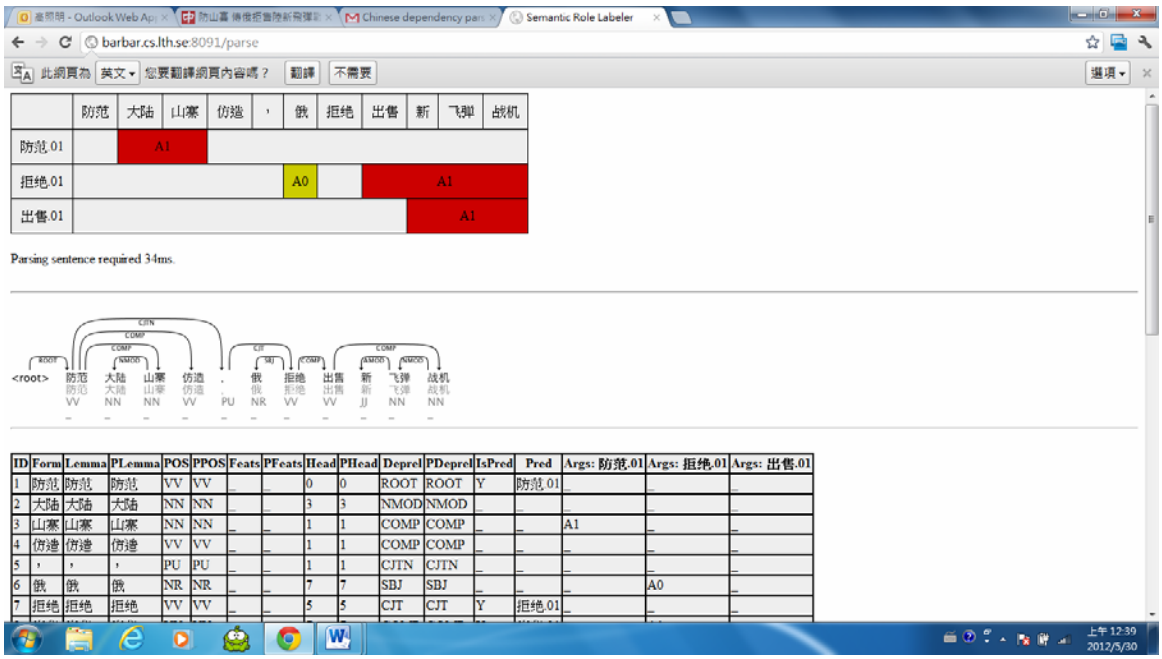
圖六是 Stanford Parser 中文剖析器的輸入介面。圖七 Stanford Parser 中文剖析器的輸出最重要的部分是句子的語法結構樹和語法依存關係。語法結構樹顯示詞組之間的語法關係如 NP 是名詞組，VP 是動詞組，IP 是句子。圖七最下方顯示詞與詞與詞之間的語法依存關係。

圖六 Stanford Parser 中文剖析器的輸入介面
http://nlp.stanford.edu:8080/parser/

圖七 Stanford Parser 中文剖析器的輸出介面
http://nlp.stanford.edu:8080/parser/



圖八 瑞典 Lund 大學以 Mate-tool 為基礎發展簡體中文的語法剖析器的輸入介面
http://barbar.cs.lth.se:8091/



ID	Form	Lemma	PLemma	POS	PPOS	Feats	PFeats	Head	PHead	Deprel	PDeprel	IsPred	Pred	Args: 防范.01	Args: 拒绝.01	Args: 出售.01
1	防范	防范	防范	VV	VV	_	_	0	0	ROOT	ROOT	Y	防范.01	_	_	_
2	大陆	大陆	大陆	NN	NN	_	_	3	3	NMOD	NMOD	_	_	_	_	_
3	山寨	山寨	山寨	NN	NN	_	_	1	1	COMP	COMP	_	_	A1	_	_
4	仿造	仿造	仿造	VV	VV	_	_	1	1	COMP	COMP	_	_	_	_	_
5	，	，	，	PU	PU	_	_	1	1	CJTN	CJTN	_	_	_	_	_
6	俄	俄	俄	NR	NR	_	_	7	7	SBJ	SBJ	_	_	_	A0	_
7	拒绝	拒绝	拒绝	VV	VV	_	_	5	5	CJT	CJT	Y	拒绝.01	_	_	_

圖九 瑞典 Lund 大學以 Mate-tool 為基礎發展簡體中文的語法剖析器的輸出介面
http://barbar.cs.lth.se:8091/

七、線上機讀辭典：教育部國語會出版的國語辭典，除了解釋，並有例句，相似詞，相反詞。

圖十 教育部重編國語辭典修訂本檢索介面
http://dict.revised.moe.edu.tw/

圖十一 教育部重編國語辭典修訂本檢索結果
http://dict.revised.moe.edu.tw/

八、詞彙知識庫：大陸董振東先生獨力發展出來的知網 Hownet 是一個非常重要的詞彙知識庫(參考 Dong and Dong (2006)。知網 Hownet 包含的訊息相當的多，是一個雙語的知識庫，可以表達概念的語意成分，概念之間的語意關係，概

念之間的常識關係。例如醫生在 Hownet 裡面有三個英文翻譯 doctor, surgeon, doctor，它們的義元表示都是{human| 人 :HostOf={Occupation| 職位},domain={medical|醫},{doctor|醫治:agent={~}}}。義元是一種表達語言知識的 meta language,醫生的義元表示醫生是一個人，具有職位，是醫學領域，且是醫治事件裡面扮演主事者的語意角色。 而醫療這個詞只有一個義元{doctor|醫治}。中研院詞庫小組將 HowNet 擴充成為 E-HowNet。

圖十二 詞庫小組的 E-HowNet 檢索結果
http://ehownet.iis.sinica.edu.tw/

「同義詞詞林」是另一個具有中文語義的資料來源。「同義詞詞林」編排的方式是按照語意階層由大類到小類分類。如下列例子所顯示，A 大類的詞都與人有關，Ae142 都與裁判有關而 Ae151 都是醫師。可惜的是，它的分類方式仍然不夠詳細。同義近義語意上下位詞的區分不夠。

Ae142","裁判員" "Ae142","裁判" "Ae142","公正人" "Ae142","國際裁判" "Ae142","國家裁判" "Ae142","巡邊員" "Ae142","記分員" "Ae142","計時員" "Ae151","醫生" "Ae151","醫師" "Ae151","醫" "Ae151","大夫" "Ae151","郎中" "Ae151","醫官" "Ae151","先生" "Ae151","白衣戰士" "Ae151","國醫" "Ae151","中

醫" "Ae151","良醫" "Ae151","賢醫" "Ae151","名醫" "Ae151","神醫" "Ae151","太醫" "Ae151","御醫" "Ae151","法醫" "Ae151","仵作" "Ae151","世醫" "Ae151","儒醫" "Ae151","庸醫" "Ae151","西醫" "Ae151","牙醫" "Ae151","獸醫" "Ae151","軍醫" "Ae151","廠醫" "Ae151","校醫" "Ae151","赤腳醫生" "Ae151","主任醫師" "Ae151","副主任醫師" "Ae151","主治醫師" "Ae151","住院醫師" "Ae151","醫士"

肆、語料庫與計算語言學

利用語料庫與統計是近年來計算語言學研究的主要趨勢。無論對語音辨識，語法剖析，岐義的解決，機器翻譯，與詞彙知識的自動取得在在都需要大型語料庫。語音辨識是將語音訊號轉變成文字，基本上可以分成前處理與後處理。前處理是從聲波的物理性質猜測最有可能的母音與子音組合。而後處理則從最有可能的音中選出最有可能的詞，無論是語音辨識的前處理或後處理或解決詞類岐義目前最常用的統計理論是隱式馬可夫模型(Hidden Markov Model 簡稱 HMM)。而解決結構岐義則常利用語法樹庫計算某一詞出現在某一種結構的機率，例如 John saw the man with a telescope.其中的介詞組 with a telescope 可以修飾名詞組 the man(約翰看見一個帶望遠鏡的人)也可以修飾動詞 saw(約翰用望遠鏡看見一個人)，造成岐義的現象。英文這種所謂 PP attachment 結構岐義的問題跟詞彙的語義與語用有關，過去計算語言學家嘗試用規則來處理效果不好，目前改以語法樹庫計算介詞組內的名詞分別跟受詞與動詞的相關性機率，從而預測介詞組究竟修飾受詞或動詞。統計演算法不需大量的人力來撰寫語言規則或編纂語言知識，可以從大型語料庫中直接抽取諸如同義詞，反義詞，搭配語等語言知識。統計方式也可以自動抽取部分中文詞彙。