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。

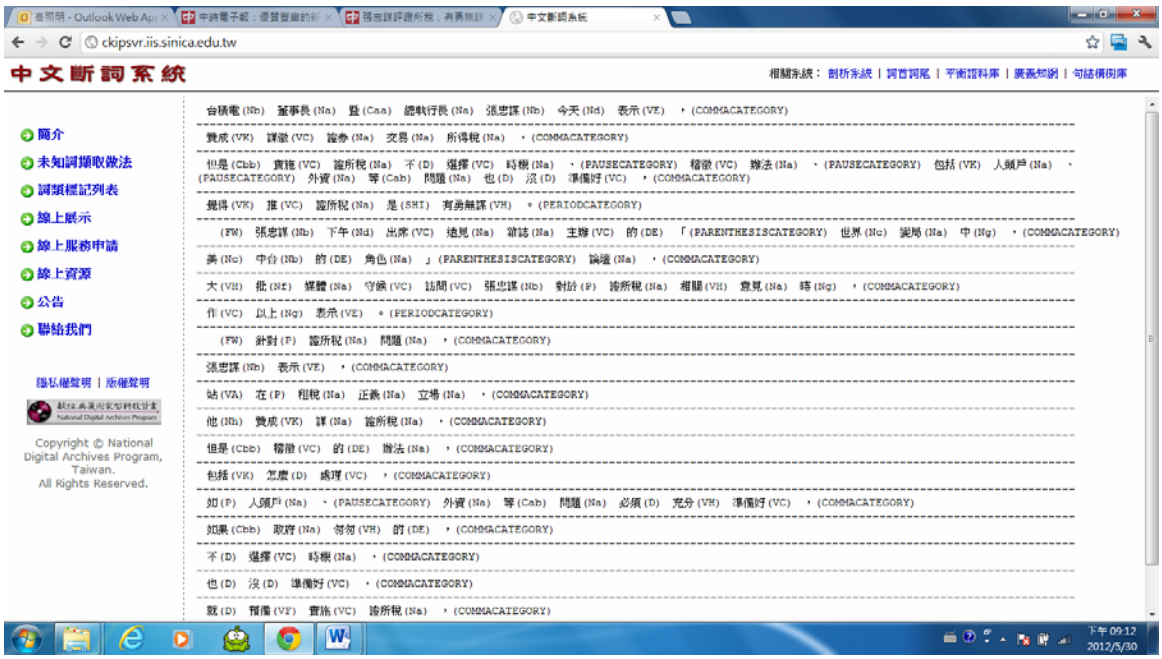
四、中文分詞程式：輸入一個句子，程式自動找到詞與詞的界線並將詞分開。由於人名，地名，及具有衍生性的詞無法全部列舉在辭典中，在加上分詞程式無法完全解決岐義的問題，中文分詞程式的準確率大約只有 90%到 97%。中文最簡單的分詞演算法是長詞優先，但如下例有時會造成錯誤。

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。