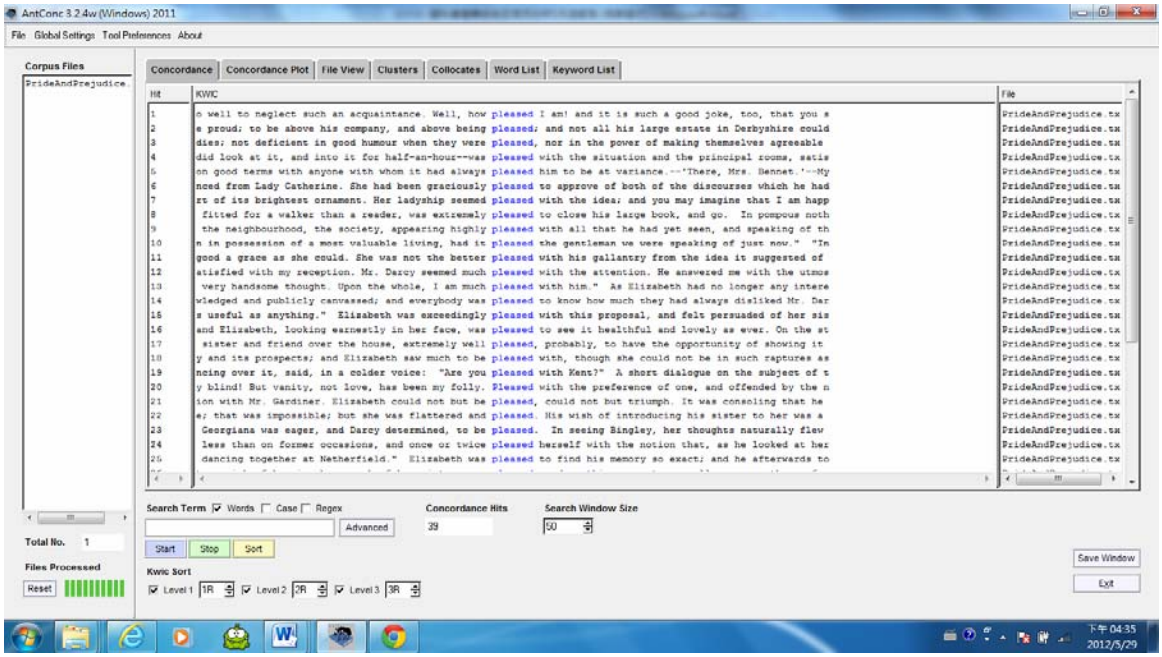
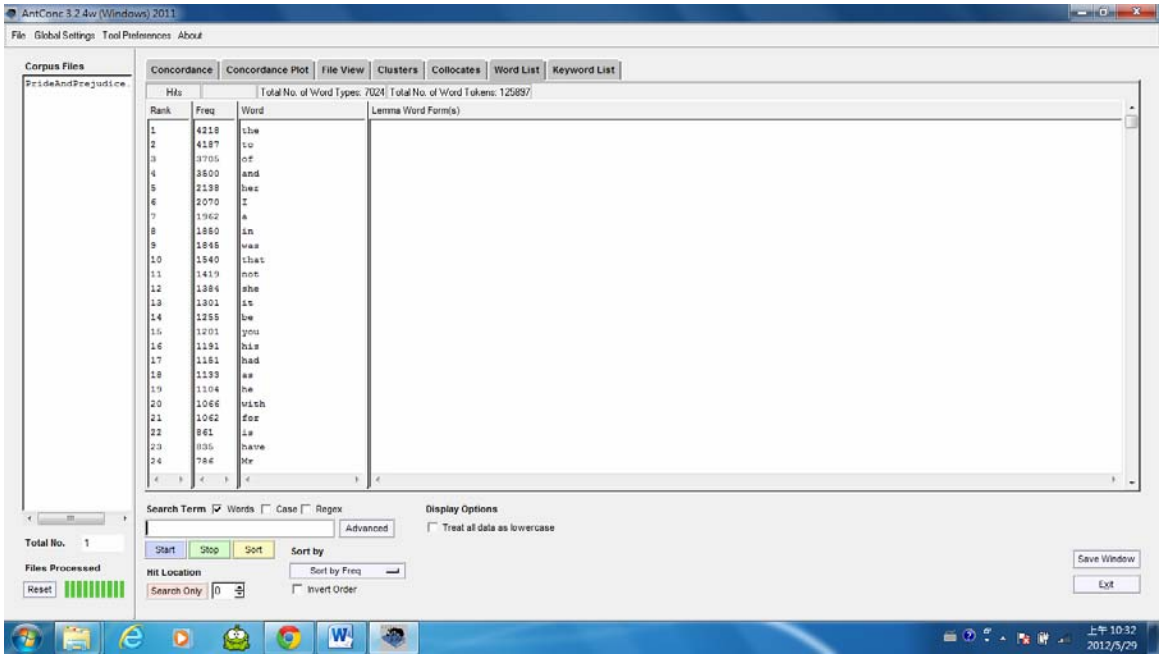


圖一 AntConc 關鍵詞前後文排序程式

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。