

中央研究院平衡語料庫詞集及詞頻統計

Word List with Accumulated Word Frequency in Sinica Corpus 3.0

http://www.aclclp.org.tw/use_wlawf.php

詞頻統計表是繼詞庫技術報告 93-02 新聞語料統計表後根據新的五百萬語料庫所計算出的詞統計資料。本表共收錄五萬五千三百零三筆詞，收錄標準為所有在語料庫中出現頻率在 4 以上的詞項。本表為檢索方便及版面簡潔之故，每欄固定列 50 筆詞項，並於欄底標註以 50 為基底的詞頻序。每一筆詞有四欄資料，第一欄為詞項，第二欄為詞類，第三欄為此詞在語料庫中出現的次數，最後一欄為對依詞頻排序至此詞的累積頻率（100 為總頻率和）
詞類分類請參考詞庫小組技術報告 93-05 中文詞類分析。

檔案的樣本內容如下：

```
\documentclass[twocolumn]{book}
\usepackage{isolatin1}
\usepackage{fancyheadings}
\usepackage{supertabular}
\usepackage{a4wide}
\textheight 10.6in
\begin{document}
\twocolumn
\pagestyle{fancy}
\lhead{Accumulated Word Frequency in CKIP Corpus}
\rhead{\thepage}
\lhead[Accumulated Word Frequency in CKIP Corpus]{Technical Report 98-02}
\rhead[\thepage]{\thepage}
\cfoot{}
\chmagnification 1050
\begin{supertabular}{llrr}
的 & DE & 285826 & 5.82\\
是 & SHI & 84014 & 7.53\\
一 & Neu & 58388 & 8.72\\
在 & P & 56769 & 9.88\\
有 & V_2 & 45823 & 10.81\\
個 & Nf & 41077 & 11.64\\
\end{supertabular}
\end{document}
```

或參考 http://www.aclclp.org.tw/doc/wlawf_abstract.pdf 的說明