中央研究院漢語平衡語料庫 簡介

中央研究院漢語平衡語料庫(簡稱 Sinica Corpus)第 4.0 版，爲一包含一千多萬目詞的帶標記平衡語料庫。本語料庫中每個文句都依詞斷開，並標示詞類標記。語料的蒐集也盡量做到平衡分配在不同的主題和語式上，是現代漢語無窮多的語句中一個代表性的樣本。所蒐集的文章爲 1981 年到 2007 年之間的文章。

總文章數：19,247

總句數：1,396,133

總詞數：11,245,932 個詞 (word token)

總詞形：239,598 個詞形 (word type)

文章共分 6 類

==========

1378 哲學

2525 文學

5461 生活

7466 社會

1482 科學

935 藝術

==========

平衡語料庫內容樣本如下，

```
<?xml version="1.0" encoding="UTF-8" ?>
- <corpus>
- <article no="200705">
<genre>報導</genre>
<style>記敘</style>
<mode>written</mode>
<topic>社會現象</topic>
<class>社會</class>
```

<medium>視聽媒體</medium>
- <author>
<name>黃學碩</name>
<gender>男女</gender>
<nationality>中華民國</nationality>
<nativelang />
</author>
<publisher>政治大學</publisher>
<publishlocation>台灣台北市</publishlocation>
<publishdate>1995</publishdate>
<edition />
<title>請願卻遭驅離留置 原住民抗議處理不當</title>
- <text>
<sentence>針對(P) 二十三日(Nd) ，(COMMACATEGORY)</sentence>
<sentence>原住民(Na) 代表(Na) 三十二(Neu) 人(Na) 前往(VCL) 陽明山(Nc) 中山樓(Nc) 遞交(VD) 「(PARENTHESISCATEGORY) 台灣(Nc) 原住民族(Na) 憲法(Na) 條款(Na) 」(PARENTHESISCATEGORY) ，(COMMACATEGORY)</sentence>
<sentence>卻(D) 遭(P) 警察(Na) 強制(D) 驅離(VC) ，(COMMACATEGORY)</sentence>
<sentence>並(Cbb) 隔離(VC) 偵訊(VC) 達(VJ) 六(Neu) 小時(Na) 一(Neu) 事(Na) ，(COMMACATEGORY)</sentence>
<sentence>民進黨(Nb) 原住民族(Na) 委員會(Nc) 二十六日(Nd) 在(P) 立法院(Nc) 第二(Neu) 會議室(Nc) 召開(VC) 記者會(Na) ，(COMMACATEGORY)</sentence>
<sentence>說明(VE) 整個(Neqa) 事件(Na) 的(DE) 經過(Na) 。(PERIODCATEGORY)</sentence>