漢語平衡語料庫

本院覽號

05T-890902

公告日期

智財權狀態

know-how

摘要

中央研究院漢語平衡語料庫(簡稱Sinica Corpus)第4.0版，為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。所蒐集的文章為1981年到2007年之間的文章。語料庫得中央研究院及中華民國行政院國家科學委員會補助，由中央研究院中文詞知識庫小組執行、研究，並授權中華民國計算語言學學會發行。

技術優勢

本詞集整理完善，為繁體中文語言處理技術之基礎資料

應用範圍

資料搜尋，中文語言內容處理

創作人

陳克健、馬偉雲

中央研究院
ACADEMIA SINICA