「中文分詞語料庫」說明

（一）簡介

中文分詞語料庫為一包含兩百萬詞、不含詞類標記的語料庫，每個文句皆根據分詞原則來斷 詞。而此分詞原則，乃是中華民國計算語言學學會在經濟部中央標準局委辦的「資訊處理用 中文分詞規範調查研究及草案研擬」計畫中所訂定的。本語料庫來源包括書面語和口語兩部 分，其中資訊類佔 21％。

（二）內容說明

1. 檔案說明

共有 94 筆 *.seg 檔案，每個檔案為 UTF-8 編碼。

2. 語料庫樣本內容

64. 。 我 真正 認識 錢 院長 還 是 在 一九八三年 被 選為 中央 研究院 院士 之後 。
65. 。 那時 為 了 在 中研院 籌設 新 的 原子 與 分子 科學 研究所 ，
66. ， 我們 曾 有 過 頻繁 的 接觸 與 深入 的 討論 ，