**********

**社團法人中華民國計算語言學學會**—ACLCLP

The Association for Computational Linguistics and Chinese Language Processing

**********

No.: 2024-005

**Subject: 臺灣客語語音資料庫 (Hakka Across Taiwan)**

---

# 臺灣客語語音資料庫 (Hakka Across Taiwan)

客家委員會建置了一個指標性的語料庫：「臺灣客語語音資料庫」。此舉不僅能保存豐富多樣的客家語言，更是推動客語語音 AI 應用進步的核心力量。語料庫積極蒐集包含四縣腔、海陸腔在內的各種客語語音資源，後續規劃進一步拓展至大埔腔、饒平腔以及詔安腔等等，致力於建立一個全面而多元的客語腔調語料庫。如今，這些寶貴的語音資料授權予社團法人中華民國計算語言學學會發行，促進語音合成、語音辨識技術的研究，以及媒體內容的創新應用。期許本語料庫建置能為客語的語音 AI 應用奠定堅實的基礎，並透過各界不斷努力投入研發，邁向更卓越的客語語音應用技術。

有關客語語料各類型內容說明如下：

- 語音辨識語料

  本語音資料庫為客語朗讀語料 ( reading speech )，是以原生客語文本，收集來自臺灣各地不同腔調的客語語音，並同時以 8 支麥克風進行錄製完成。錄製的客語語音，經由兩次人工校正文本後，整理成可供語音辨認技術研究與開發使用之語音語料庫。
- 語音合成語料

  本語音資料庫為客語朗讀語料 ( reading speech )，是以原生客語文本，男女各 1 位發音員每人錄製約 30 小時。主要目的是要建置客語語音合成語料庫，作為研發人工智慧之基礎建設，尤其是針對基於深度學習之語音合成研究資料之要求。
- 媒體語音語料

  蒐集廣播電臺/電視臺提供客語音檔增加自對話，刪除音檔中空白及背景音樂部分，續以校正完成之媒體音檔及逐字稿轉換至客語語音資料庫格式，產出具 metadata 之標準化電子語音語料庫格式，音檔標註 json 檔 ( Unicode 編碼，儲存 metadata)， 約 280 小時口語語音資料，可以使

小時口語語音資料，可以使人工智慧開發人工智慧開發團隊或廠商，開發出更貼近生活用語之語音辨識。

- 平行辭庫
  將華語辭典對譯為客語，建立人工智慧使用之華客語平行辭庫內容收錄教育部國語辭典部分用詞、客語能力中高級認證、 臺灣街道名及知名地標、常見生活詞彙單位、日期、時間、新聞常見用詞共計提供約 50,000 詞華語翻譯客語漢字、約 50,000 詞客語漢字四縣腔拼音、約 50,000 詞客語漢字海陸腔拼音。

工本費:

- 四縣腔(含南四縣)語音辨識語料 NT$1,000 元
- 海陸腔語音辨識語料 NT$1,000 元
- 四縣腔語音合成語料 NT$1,000 元
- 海陸腔語音合成語料 NT$1,000 元
- 四縣腔媒體語料 NT$1,000 元
- 四縣腔客華平行辭庫 NT$1,000 元
- 海陸腔客華平行辭庫 NT$1,000 元

申請文件與相關資訊請參閱學會網頁

https://www.aclclp.org.tw/use_mat_c.php#hat

**********
社團法人中華民國計算語言學學會—ACLCLP
Tel. : +886-2-27881638
Fax.: +886-2-26519386
Email: aclclp@aclclp.org.tw
**學會網頁**
**********