中研院漢語對話語音語料庫 (Sinica MCDC8)

一 語料庫內容

聲檔大小： 2.5 GB (.wav)

聲檔長度： 8 小時

中文字數： 12 萬 (122K)

文字轉記： PRAAT 格式 (.TextGrid)

二 語料錄製

本語料庫為接近日常生活之對話口語資料。在錄製時，從錄音地點到錄音設備，盡量以語者在對談過程中感覺自然為目標。並明確強調無須特別注意發音的正確與否。語者人選是中央研究院調查研究工作室由台北市市民中隨機抽樣選出後配對錄音。數位錄音採用 SONY TCD-D10 Pro II DAT 錄音機，Audio-Technica ATM 33a 手持式麥克風。兩位語者分錄於左右聲道。錄音地點為普通房間。由於雙方是第一次見面，為確保不陷入無話可談的窘境，首先請他們自我介紹以熟悉彼此。提供談話主題列表供語者參考。語者可以任意選擇列表的主題或是自行決定話題。不限定單一主題，可以隨時轉移任何話題。平均每個對話約一小時。語者資料與談話主題表列於下。

對話編號	性別(年齡)		主題
	左聲道	右聲道	
mcdc-01	女(29)	男(25)	工作、休閒活動、經濟、開車
mcdc-02	女(37)	男(35)	休閒活動、經濟、工作、性別、政治
mcdc-03	女(16)	女(17)	家庭、學校、購物、生涯規劃、明星
mcdc-05	男(40)	女(46)	工作、家庭、社會階層、保險、歷史、省籍情結、名人
mcdc-09	女(30)	女(35)	工作、旅行、生活態度、環保、健康
mcdc-10	男(35)	男(23)	電影、政治、軍隊、捷運、學校、經濟
mcdc-25	男(43)	女(45)	交通、工作、小孩、旅行、電腦、管理
mcdc-26	女(37)	男(24)	工作、求職、家庭、車禍、休閒活動、學英文、婚姻、軍隊

三 語料庫格式

檔案命名原則： 對話編號_聲道_個別語者說話輪序號

檔案分割單位： 語者說話輪。

標記層名稱： IPU-Hanzi。

非漢字標記內容： (uncertain)：無法明確辨識內容之語音。

(para)：非關語言內容之言語現象，例如停頓或吸氣聲。

FW：非中文內容。

WHITE_NOISE：個人隱私之內容，其語音清空為空白。

FILLER：填充詞。以大寫英文字母轉寫。

PARTICLE：語氣詞。不論該語氣詞是否有慣用的漢字，

所有語氣詞皆以大寫英文字母轉寫。

PARTICLE_M：Mandarin 的語氣詞。

PARTICLE_S：Southern Min 的語氣詞。

MARKER：語言內容為漢字可轉寫之內容，例如「那個」，但其作用主要為言談功能者以大寫英文字母標示，NA GE。

FILLER, PARTICLE, MARKER 列表：

FILLER	PARTICLE_M	PARTICLE_S	MARKER
MHM	A	EIN	GE
MHMHM	AI	HAN	ME
MHMHMHM	AN	HEIN	NA
MHMHMHMHM	AU	HEN	NA GE
MHMHMHMHMHM	BA	HO	NE
MHMHMHMHMHMHM	E (EP)	HON	NE GE
MHMHMHMHMHMHMHM	EI	HYO	NEI
MHMHMHMHMHMHMHMHMHMHM	EN	MEI	NEI GE
MHMHMM	HA	NEIN	SHE
MHMM	HAI		SHE ME
NHN	HE		SHEN
NHNHN	HEI		SHEN ME
NHNHNHN	HWA		ZHE
NHNHNHNHN	LA		ZHE GE
NHNHNHNHNHN	LEI		ZHEI
NHNHNHNHNHNHN	LO		ZHEI GE
NHNHNHNHNHNHNHN	MA		
NHNN	NO		
UHM	NOU		
UHMHM	O		
UHMHMHM	ON		
UHMHMHMHM	OU		
UHMHMHMHMHM	SAI		
UHMHMHMHMHMHM	WA		
UHMHMHMHMHMHMHM	WEI		
UHMM	YA		
UHN	YE		
UHNHN	YEI		
UHNHNHN	YI		
UHNHNHNHN	YOU		
UHNHNHNHNHN			
UHNHNHNHNHNHN			
UHNHNHNHNHNHNHN			
UHNN			

四文獻引用

使用本語料庫所獲致之研究成果，應引用以下文獻。

Tseng, S.-C. 2013. Lexical coverage in Taiwan Mandarin conversation. *International Journal of Computational Linguistics and Chinese Language Processing* 18(1): 1-18.