詞　　庫　　小　　組

技術報告　02-01

Technical Report no. 02-01

# 現代漢語口語對話語料庫標註系統說明

曾淑娟、劉怡芬



出版處：台北,南港
中央研究院 資訊科學研究所 中文詞知識庫小組
中央研究院 語言學研究所籌備處

# 現代漢語口語對話語料庫

# 標註系統說明

曾淑娟，劉怡芬

二００二年九月

中央研究院語言學研究所籌備處

# 目錄

## 1 引論

這份技術報告是從二０００到二００二在中央研究院語言學研究所籌備處所執行的的現代漢語口語對話語料庫計劃的執行成果，其中包含語料庫的建立、語料的處理和語料的標記。 語料的收集無疑地是語言學研究不可或缺的一項工作。即便是理論的研發，若無實際語料爲支持，怕所產出的結果也只是象牙塔裡的產物。 在大家認同科技無遠弗屆的今天，語料收集的工具已不再限於紙和筆，而是數位聲音與影像錄製、自動文件擷取、自動分類，以便能快速計算與檢索的電腦輔助工具。 在口語語料的收集與標記的方法上，我們有一些較爲不同的作法。 發音人是由籍設台北市的市民中按年齡分層隨機抽樣選取、語料的收集由數位錄音機 (DAT) 錄製、語料的標記配合由我們自己開發的輔助電腦介面進行、語料標記集系統化的定義，最重要的是所標記完的語料內容可以再藉由程式自動轉換爲資料，進行事後的語料分析和檢索。 以下各點都將分別介紹，但本技術報告重點將是語料標記系統的建立。 經過將近一年密集的討論和修正，我們的口語標記系統雖然一定不完美，而且不同研究目的導引下標記集的需求也不同，但至少能提供同行一個具系統性且有嚴格定義的標記集。中央研究院資訊所王新民老師在試用我們的標記集的同時也提供了不少寶貴的意見。 我們希望同行在使用我們的標記系統時，若有任何批評或建議，請不吝賜教。 以作爲日後修訂時的參考。

## 2 現代漢語口語對話語料庫(Mandarin conversational dialogue corpus)

現代漢語口語對話語料庫收集於二００一年夏天。 我們在此感謝所有不辭辛勞到中央研究院跑一趟的發音人。

### 2.1 語料庫架構說明(corpus design)

計劃的長程目標是要蒐集多樣的現代漢語口語語料，一方面以數位錄音方式這個時代人們口語語言的使用，另一方面也提供語言研究的素材。 就現階段來說，所蒐集的語料只限於日常生活的對話。 由錄音地點到錄音設備，我們盡量使發音人對談過程感覺更自然，更貼近日常對話。 由於發音人雙方是第一次見面，爲確保不陷入無話可談的窘境，我們首先請他們先自我介紹，至於談話主題方面，我們也預先列出一些主題供發音人參考。 發音人可以任意選擇我們所提供的主題或是任何其他的話題與對方聊天。 不限定要繞著同一主題，可以隨時轉移到其他的主題上。唯一一項要求是發音人雙方必須在整個談天的過程中提出幾個有關路徑的問題。 限定一個明確的路徑主題是爲了希望不只蒐集到

多樣性的日常生活主題語料，還能蒐集到一部分描述詳盡且主題明確的語料。因此，事前我們告訴發音人雙方要精確地說明路徑，直到彼此清楚對方所描述的路徑爲止。 如此一來，我們可觀察到現代人們在詢問以及回答方向、位置時，他們的語言使用情形。

## 2.2 語料蒐集(data collection)

### 2.2.1 選取發音人(subject selection)

發音人是中央研究院調查研究工作室依據 16-25 歲、26-35 歲以及 36-45 歲三大年齡層由台北市市民中隨機抽樣選出。 抽樣結果取得 1080 位候選人後，再寄出邀請函詢問是否有意願前來參與錄音。 而由於計劃初階段只預計蒐集 30 個對話，因此我們從回覆同意前來參與錄音的人之中依他們的回覆順序取前 60 位，共 37 位女性，23 位男性。 而年齡分布於 16-25 歲的有 20 位，26-35 歲的有 19 位，36-45 歲的有 21 位。 發音人職業分布的統計圖表如下：

表一：發音人職業

| **職業** | **女性發音人** | **男性發音人** | **發音人總數** |
|---|---|---|---|
| 學生<br>高中<br>大學<br>研究生 | <br>8<br>4<br>0 | <br>0<br>1<br>2 | 15 |
| 商 | 9 | 5 | 14 |
| 待業 | 5 | 5 | 10 |
| 教師 | 2 | 2 | 4 |
| 資訊科技業 | 1 | 3 | 4 |
| 醫療保健 | 3 | 0 | 3 |
| 自營業 | 1 | 2 | 3 |
| 銀行 | 0 | 2 | 2 |
| 公務員 | 2 | 0 | 2 |
| 義工 | 2 | 0 | 2 |

| | | | |
|---|---|---|---|
| 娛樂業 | 0 | 1 | 1 |
| 總數 | 37 | 23 | 60 |

2.2.2 錄音過程說明與指示(instructions)

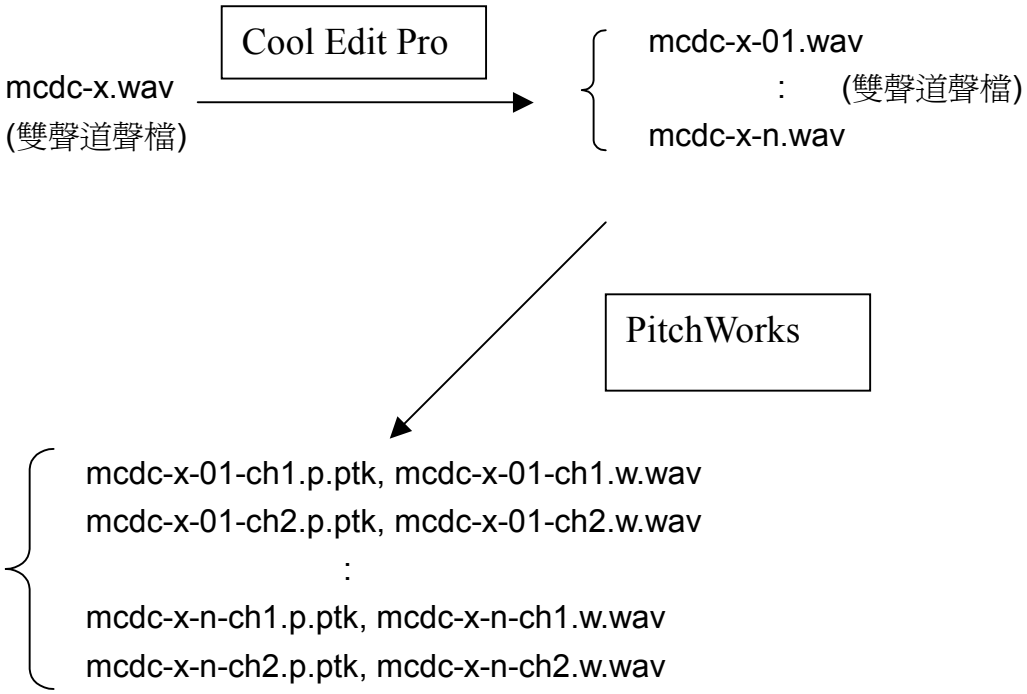
在正式錄音之前，我們先向發音人說明蒐集語料的目的以及整個計劃的研究目標。 然後是錄音過程說明。 待他們瞭解同意後，請他們簽署同意書、填寫基本資料和語言使用的問卷，最後再請他們閱讀過程說明，直到他們清楚瞭解才開始錄音。 並且我們也告訴發音人儘可能的自然地談話，不必特別注意句法或發音，因爲計劃的目的是希望能蒐集到不同的說話腔調及風格。 由於發音人雙方在錄音前是未曾謀面的，因此一開始他們需先自我介紹，而後才開始選定主題聊天，聊天的主題並不限定只能是說明書上有的，像旅行、購物、食物、音樂、工作、家庭、政治或經濟，也可以是他們自己想要講的任何主題。 此外，我們希望發音人雙方輪流向對方詢問路徑，並請對方具體詳盡地描述所詢問的路徑，如地標、方向。

2.2.3 錄音設備與格式(digital recording)

數位錄音採用 SONY TCD-D10 Pro II DAT 的數位錄音機，使用 Audio-Technica ATM 33a 手持式麥克風。 以取樣本率 48 kHz 將兩位發音人的語料分別錄於左右聲道。 錄音地點爲普通房間。

2.2.4 錄音資料處理(audio files)

將 30 個對話的錄音資料轉成數位聲檔，每個對話儲存一個聲檔。 檔名以 mcdc-01 到 mcdc-30 記錄。 存取地點路徑爲 /dialogue/mcdc-x.wav。 另外，因電腦的記憶體跟速度有限，爲了處理語料內容更有效率，我們利用軟體 Cool Edit Pro 將它們分割成更小的雙聲道聲檔，方式是在長度約三分鐘左右找到一個清楚可辨的停頓切開，存放的路徑及檔名爲 /stereo/mcdc-x/mcdc-x-n.wav，x 變數指的是 1-30 的對話聲檔，而 n 指的是 1-20 個單一對話中個別的三分鐘聲檔。 爲了之後要分析語音以及進行切音工作，而大多數語音處理軟體只能處理單聲道的檔案。 因此在儲存語料時，我們另外利用語音軟體 PitchWorks 將雙聲道的聲檔個別存成.ptk 跟.wav 的單聲道格式。 錄音資料轉成聲檔的流程與資料庫如下。




2.2.5 對話內容整理(recorded dialogues)

我們總共錄下 30 個對話，共 25.6 個小時。 平均每個對話大約是五十分鐘，其中主題包羅萬象，有些取自說明書，有些則是發音人自己想談的主題。 以下是對話內容的總表：

表二：對話主題

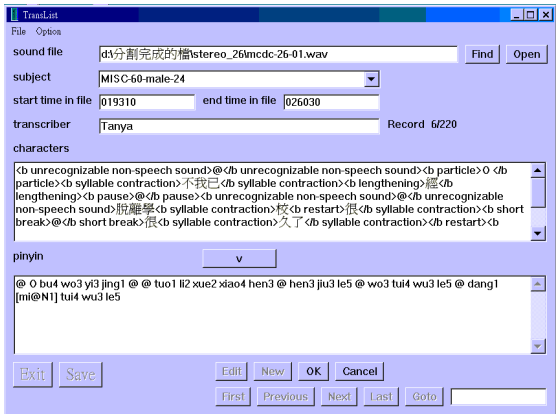
| 對話序號 | 長度(分) | 發音人：性別(年齡) | 對話主題 |
|---|---|---|---|
| mcdc-01 | 61 | 女(29),男(25) | 工作、休閒活動、經濟、開車 |
| mcdc-02 | 63 | 女(37),男(35) | 休閒活動、經濟、工作、性別、政治 |
| mcdc-03 | 61 | 女(16),女(17) | 家庭、學校、購物、生涯規劃、明星 |
| mcdc-04 | 65 | 男(44),女(21) | 代溝、家庭、工作、程式設計、客戶、旅行 |
| mcdc-05 | 63 | 男(40),女(46) | 工作、家庭、社會階層、保險、歷史、省籍情結、名人 |

| | | | |
|---|---|---|---|
| mcdc-06 | 61 | 男(33),男(28) | 工作、經濟、交通、考試、人際關係、教育、學校 |
| mcdc-07 | 57 | 男(40),男(20) | 工作、旅行、童年、交通、食物、電視節目、明星、政治 |
| mcdc-08 | 50 | 女(30),女(36) | 教育、工作、家庭、生活經驗、結婚、休閒活動、學校 |
| mcdc-09 | 66 | 女(30),女 35) | 工作、旅行、生活態度、環保、健康 |
| mcdc-10 | 54 | 男(35),男(23) | 電影、政治、軍隊、捷運、學校、經濟 |
| mcdc-11 | 49 | 女(25),男(26) | 旅行、休閒活動、食物、音樂、童年、台北、電腦遊戲、匯率、頭髮 |
| mcdc-12 | 53 | 女(34),女(20) | 嗜好、旅行、工作、開車、寵物、小孩、教育 |
| mcdc-13 | 54 | 女(43),女(39) | 工作、小孩、網咖、捷運、省籍情結、個性、教育 |
| mcdc-14 | 31 | 女(19),女(17) | 休閒活動、旅行、生涯規劃、打字、電影、網路、打工、學校、學英文 |
| mcdc-15 | 59 | 男(29),男(21) | 旅行、學校、入學考試、工作、音樂、網路、留學、電影、軍隊、鬼 |
| mcdc-16 | 51 | 女(28),男(35) | 工作、旅行、電視節目、電影、方言、休閒活動、嗜好 |
| mcdc-17 | 51 | 男(29),女(21) | 學校、入學考試、留學、鬼、休閒活動、暑假、電腦遊戲、音樂、KTV、電影 |
| mcdc-18 | 56 | 女(41),女(45) | 休閒活動、視力、童年、旅行、工作、同事 |
| mcdc-19 | 57 | 女(18),女(18) | 休閒活動、音樂、電影、家庭、學校、購物、電視節目、考試、食物、打工、明星 |
| mcdc-20 | 60 | 男(44),女(25) | 旅行、生涯規劃、家庭、童年、政治、教育、學英文、工作、軍隊、台灣社會、婚姻 |
| mcdc-21 | 39 | 男(39),男(31) | 工作、休閒活動、電影、音樂、家庭、 |

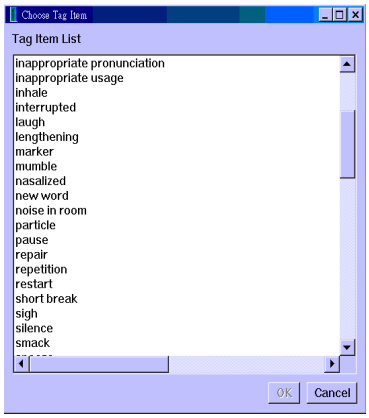
| | | | |
|---|---|---|---|
| | | | 食物 |
| mcdc-22 | 46 | 女(27),女(27) | 工作、家庭、讀書、小孩、旅行、氣功、休閒活動 |
| mcdc-23 | 51 | 女(31),女(20) | 學校、休閒活動、閱讀、旅行、工作、購物、家庭、運動、健康 |
| mcdc-24 | 47 | 女(39),男(40) | 運動、食物、體重、抽煙、工作、拖吊、生活態度、電視節目、視力、中國大陸、旅行、家庭 |
| mcdc-25 | 55 | 男(43),女(45) | 交通、工作、小孩、旅行、電腦、管理 |
| mcdc-26 | 46 | 女(37),男(24) | 工作、求職、家庭、車禍、休閒活動、學英文、婚姻、軍隊 |
| mcdc-27 | 51 | 女(39),女(18) | 教育、家庭、休閒活動、學英文、閱讀 |
| mcdc-28 | 48 | 女(42),男(28) | 工作、旅行、黑道幫派、賭博、中國大陸、政治、運動、游泳、家庭、電腦遊戲 |
| mcdc-29 | 51 | 女(37),女(19) | 旅行、考試、拼圖、藝術、動畫、設計、學校 |
| mcdc-30 | 49 | 男(45),女(25) | 旅行、網咖、麻將、生涯規劃、經濟、生活經驗 |

### 2.3 轉寫介面與資料庫管理 (interface and data management)

**TransList** 是由我們自行開發的轉寫工具（見圖一），欄位有聲音檔案、發音人、標記員、檔案起訖時間、語料中文內容與語料漢語拼音內容。 所有欄位的資料都將自動轉入資料庫中，方便建立日後的後設資料 (metadata)。 標記集可依照不同的需求訂定（圖二）。 在轉寫的同時也可以隨時加入標記。 所加入的標記程式會自動整合入資料庫中。 詳細的資料處理與管理，將另行出版介紹。




圖一：轉寫介面 **TransList**




圖二：可選取之標記集

處理好的標記語料，再經由 **TransList** 自動轉爲資料庫。 訂定的標記集，會自動編號，語料中有加以標記的部分，欄位就會給定該編號爲欄位的值。 格式見圖三。

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 18304 | 073050 | 105588 | 68 | 蓋 | gai4 | MISC-08-male-25 | Fen | mcdc-01-2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 18305 | 073050 | 105588 | 69 | 章 | zhang1 | MISC-08-male-25 | Fen | mcdc-01-2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 18306 | 073050 | 105588 | 70 | 認 | ren4 | MISC-08-male-25 | Fen | mcdc-01-2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 18307 | 073050 | 105588 | 71 | 可 | ke3 | MISC-08-male-25 | Fen | mcdc-01-2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 18308 | 073050 | 105588 | 72 | 的 | [le5] | MISC-08-male-25 | Fen | mcdc-01-2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 18309 | 073050 | 105588 | 73 | @ | @ | MISC-08-male-25 | Fen | mcdc-01-2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 18310 | 073050 | 105588 | 74 | 只 | zhi3 | MISC-08-male-25 | Fen | mcdc-01-2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 18311 | 073050 | 105588 | 75 | 有 | you3 | MISC-08-male-25 | Fen | mcdc-01-2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 18312 | 073050 | 105588 | 76 | 三 | san1 | MISC-08-male-25 | Fen | mcdc-01-2 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 18313 | 073050 | 105588 | 77 | 分 | [men1] | MISC-08-male-25 | Fen | mcdc-01-2 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 18314 | 073050 | 105588 | 78 | 之 | zhi1 | MISC-08-male-25 | Fen | mcdc-01-2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 18315 | 073050 | 105588 | 79 | 一 | yi1 | MISC-08-male-25 | Fen | mcdc-01-2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 18316 | 073050 | 105588 | 80 | @ | @ | MISC-08-male-25 | Fen | mcdc-01-2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 18317 | 073050 | 105588 | 81 | NA | NA | MISC-08-male-25 | Fen | mcdc-01-2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 18318 | 073050 | 105588 | 82 | 其 | qi2 | MISC-08-male-25 | Fen | mcdc-01-2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 18319 | 073050 | 105588 | 83 | 它 | ta1 | MISC-08-male-25 | Fen | mcdc-01-2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 18320 | 073050 | 105588 | 84 | 的 | de5 | MISC-08-male-25 | Fen | mcdc-01-2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 18321 | 073050 | 105588 | 85 | @ | @ | MISC-08-male-25 | Fen | mcdc-01-2 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 |
| 18322 | 073050 | 105588 | 86 | @ | @ | MISC-08-male-25 | Fen | mcdc-01-2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 10 |
| 18323 | 073050 | 105588 | 87 | 三 | san1 | MISC-08-male-25 | Fen | mcdc-01-2 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 18324 | 073050 | 105588 | 88 | 分 | [men1] | MISC-08-male-25 | Fen | mcdc-01-2 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 18325 | 073050 | 105588 | 89 | 之 | zhi1 | MISC-08-male-25 | Fen | mcdc-01-2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 18326 | 073050 | 105588 | 90 | 二 | er4 | MISC-08-male-25 | Fen | mcdc-01-2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 18327 | 073050 | 105588 | 91 | 是 | shi4 | MISC-08-male-25 | Fen | mcdc-01-2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 18328 | 073050 | 105588 | 92 | @ | @ | MISC-08-male-25 | Fen | mcdc-01-2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 18329 | 073050 | 105588 | 93 | 警 | jing3 | MISC-08-male-25 | Fen | mcdc-01-2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

圖三：轉換後之資料庫格式

## 3 語音部份口語標註(speech sounds)

因爲所收集的語言材料是口語形式，就免不了要作標音的工作。 除了下一節起要介紹的語言學標記系統外，實際發音的重現是口語語料整理分析的一大課題。 國際音標(IPA)固然具有公認的學術使用地位而且標音準確，但由於字型少見，大多數的文書編輯，語音處理軟體都無法使用。 臺灣慣用的注音符號，

除了有字型的問題，還有無法準確標音的缺點。 中國大陸使用的以拉丁字母爲主的漢語拼音，雖然沒有字型問題，但也有一字表多音的毛病。 因此綜合這幾套標音系統，我們決定採用目前國際間計算語言學界找到的妥協方式，也就是以國際音標爲基底，再由一般鍵盤中能找到的符號作一對一的對應，一般稱爲 SAMPA (Speech Assessment Methods Phonetic Alphabet)。每一個語言都必須個別訂定自己的音標系統。 漢語部分，目前有 SAMPA-T 與 SAMPA-C 兩套系統分別使用於臺灣與中國大陸。 因爲在本計劃裡所處理的是自發性口語材料，語音包羅萬象，有閩南、客、日、英語，也有無法歸類於某一種語言的語音，因此我們的作法是將音標的系統性重新分析歸類，以現代漢語的語音爲主軸，稱之爲 SAMPA-M (M 代表 Mandarin)。 現代漢語所指的是臺灣所謂的國語，中國大陸所謂的普通話。 之後再配合我們在標記語料時的需求，加入其他語音。

### 3.1 漢語口語語音標音系統(phonetic transcription)

漢語口語語音標音系統主要有兩部分：現代漢語與其他語音。其他語音可以是漢語方言裡出現的語音，也可以是其他非漢語語族像是英語或日語的語音。

#### 3.1.1 現代漢語輔音(Mandarin consonants)

現代漢語裡共有二十一個輔音。 注音、漢語拼音、國際音標與 **SAMPA-M** 的對照，見表三。

表三：現代漢語輔音

| **注音（Zhuyin）** | **漢語拼音（Pinyin）** | **國際音標（IPA）** | **SAMPA-M** |
|---|---|---|---|
| ㄅ | b | p | p |
| ㄆ | p | p’ | p_h |
| ㄇ | m | m | m |
| ㄈ | f | f | f |
| ㄉ | d | t | t |
| ㄊ | t | t’ | t_h |
| ㄋ | n | n | n |
| ㄌ | l | l | l |

| | | | |
|---|---|---|---|
| ㄍ | g | k | k |
| ㄎ | k | k’ | k_h |
| ㄏ | h | x | x |
| ㄐ | j | tɕ | t6 |
| ㄑ | q | tɕ’ | t6_h |
| ㄒ | x | ɕ | 6 |
| ㄓ | zh | tʂ | TS |
| ㄔ | ch | tʂ’ | TS_h |
| ㄕ | sh | ʂ | S |
| ㄖ | r | ʐ | Z |
| ㄗ | z | ts | ts |
| ㄘ | c | ts’ | ts_h |
| ㄙ | s | s | s |

3.1.2 現代漢語元音(Mandarin vowels)

現代漢語裡共有十七個元音。 注音、漢語拼音、國際音標與 **SAMPA-M** 的對照，見表四。

表四：現代漢語輔音

| **注音（Zhuyin）** | **漢語拼音（Pinyin）** | **國際音標（IPA）** | **SAMPA-M** |
|---|---|---|---|
| ㄚ | a | a | a |
| ㄛ | o | o | o |
| ㄜ | e | ə,ɣ | @ |
| ㄝ | e | e | e |

| | | | |
|---|---|---|---|
| ㄞ | ai | ai | ai |
| ㄟ | ei | ei | ei |
| ㄠ | ao | au | au |
| ㄡ | ou | ou | ou |
| ㄢ | an | an | an |
| ㄣ | en | ən | @n |
| ㄤ | ang | aŋ | aN |
| ㄥ | eng | əŋ | @N |
| ㄦ | er | ɚ | 2 |
| ㄧ | i | i | I |
| ㄨ | u | u | U |
| ㄩ | y | y | y |
| ㆩ | i | ˥,˩ | I_i, I_u |

### 3.1.3 其他語音(other phonemes)

其他非現代漢語的輔音與元音的符號隨語料標記的進行加入，所以在此不詳列。

## 3.2 特殊音韻現象(pronunciation variation)

以下介紹的標記集將分別依照定義、標記原則與標記實例三方面作說明。 所有的標記實例都取自於現代漢語口語對話語料庫，並註解其聲音檔案、筆數與起訖時間點。

### 3.2.1 拖長音(lengthening)

音節拖長現象，不限定位於音節的哪個位置。

一、標記原則：

標記有拖長音的音節。

二、標記實例：(*取自 mcdc-01-01.wav, record 24, 079804-084323*)

原始句例：我目前是從事外貿 (*"事"的[ˉ]有拖長現象*)

漢字轉記：我目前是從<b lengthening>事</b lengthening>外貿

拼音轉記：wo3 mu4 qian2 shi4 cong2 shi4 wai4 mao4

### 3.2.2 音的同化(assimilation)

字的發音受到相鄰音發音部位或方法的影響而改變本身發音的同化現象。 同化現象可以是因受相鄰音影響而增加的音，也可以是受相鄰音的影響而使原本的發音改變成與其相鄰音發音部位或方法相近或一致的音。

一、標記原則：

標記範圍包括被同化的字與使其發音產生變化的相鄰字。 轉記漢字時以標準發音的漢字轉寫，拼音部分被同化字的字音則以實際發音轉寫，實際發音採用 SAMPA-M（*參考 3.1 漢語口語語音標音系統*），並置於中括號[ ]內。 若有音節省略現象時，漢字與拼音相差的音節則以 % 標示。

二、標記實例：(*取自 mcdc-01-01.wav, record 23, 077458-080547*)

原始句例：賴先生呢您從事什麼工作 (*"呢"受到"您"的影響，在音節末增加[n]的音*)

漢字轉記：賴先生<b assimilation>呢您</b assimilation>從事什麼工作

拼音轉記：lai4 xian1 shen1 [n@n2] nin2 cong2 shi4 shen2 me5 gong1 zuo4

### 3.2.3 音節合併(syllable contraction)

說話者說得太快或不清楚時出現的音節合併現象。 合併現象有三，一是清楚可辨的音節短少，像是從原本正常的三個字三個音節變成三個字兩個音節，或者是兩個字兩個音節變成兩個字一個音節，例如："我們"的實際發音變成[om]；二是音節雖無短少，但卻都連在一起，難以切割，例如："就是"的實際發音變成[tɕio˥]；三是音節無短少且音節可切割，只是音節結構有變，例如："誇張"的實際發音變成[k'ua1] [aŋ1]。

一、標記原則：

標記範圍包括所有音節合併的字。 拼音部份仍以標準發音的漢語拼音轉寫，而非以實際發音轉寫。

二、標記實例：(*取自 mcdc-01-03.wav, record 117, 000000-015968*)

原始句例：但是相對跟淡水啊那種什麼木柵那邊比就少很多了 (*"那種"實際發音爲[noŋ]*)

漢字轉記：但是相對跟淡水 A<b syllable contraction>那種</b syllable contraction>什麼木柵那邊比就少很多了

拼音轉記：dan4 shi4 xiang1 dui4 gen1 dan4 shui3 A na4 zhong3 shen2 me5 na4 bian4 bi3 jiu4 shao3 hen3 duo1 le5

註：出現在原始句例中的"啊"爲有相對國字感嘆詞，標記員須在漢字與拼音轉記中轉寫爲大寫英文拼音"A"，有關感嘆詞的標記請參考 3.6.2。

### 3.2.4 鼻化音(nasalized)

標準字音中無任何鼻音，但實際發音卻出現鼻化音現象。 與音的同化現象不同的是，鼻化音現象並非受相鄰鼻音影響去改變本身的發音部位或方法，而只是整個標準字音充斥著鼻音而已。

一、標記原則：

標記範圍包括所有帶鼻音的字。 拼音部份仍以標準發音的漢語拼音轉寫，而非以實際發音轉寫。

二、標記實例：(*取自 mcdc-01-03.wav, record 139, 100158-109932*)

原始句例：室內就是一小間一小間嘛那露天就是大家一起啊 (*"家"實際發成[tɕia]整個帶有鼻音*)

漢字轉記：室內就是一小間一小間 MA NA 露天就是大<b nasalized>家</b nasalized>一起 A

拼音轉記：shi4 nei4 jiu4 shi4 yi4 xiao3 jian1 yi4 xiao3 jian1 MA NA lu4 tian1 jiu4 shi4 da4 jia1 yi4 qi3 A

註：出現在原始句例中的"嘛"與"啊"爲有相對應國字感嘆詞，標記員須在漢字與拼音轉記中轉寫爲大寫英文拼音"MA"與"A"，有關感嘆詞的標記請參考 3.6.2。另外原始句例中的"那"，標記員判斷其爲語助詞，因此根據 3.6.1，語助詞在漢字與拼音轉記方面，須轉寫爲大寫英文拼音"NA"。

### 3.2.5 發音偏差(inappropriate pronunciation)

說話者發音偏離原字詞標準發音，但標記員依據談話內容，仍可辨識出原字詞爲何，且其母音、子音部分須清楚可辨。

一、標記原則：

標記範圍爲發音偏差字詞本身，轉記漢字時以標準發音的漢字轉寫。 漢語拼音部分則以[*實際發音*]轉寫，呈現出該字詞的偏差發音，若聲調亦可辨識出，一併轉記在[*實際發音*]內。 漢字與拼音相差的音節則以 % 標示。

二、標記實例：(*取自 mcdc-01-02.wav, record 103, 115815-129795*)

原始句例：我比較喜歡從事一些球類運動啦 (*"比"的實際發音爲[pu2]*)

漢字轉記：我<b inappropriate pronunciation>比</b inappropriate pronunciation>較喜歡從事一些球類運動 LA

拼音轉記：wo3 [pu2] jiao4 xi3 huan1 cong2 shi4 yi4 xie1 qiu2 lei4 yun4 dong4 LA

註： 出現在原始句例中的"啦"爲有相對應國字感嘆詞，標記員須在漢字與拼音轉記中轉寫爲大寫英文拼音"LA"，有關感嘆詞的標記請參考 3.6.2。

## 3.3 無法或難以辨識的語音(unintelligible speech sound)

### 3.3.1 喃喃自語(mumble)

說話者不是在回應對方以接續話題的語流，而是他無意讓對方聽見而小聲的喃喃自語。 標記爲喃喃自語的語言內容必須是清楚可辨的，若是語言內容無法辨識則標記爲無法辨識的語音或不確定的字音(*參考 3.3.2 或 3.3.3*)。

一、標記原則：

標記範圍即是說話者喃喃自語的內容。

二、標記實例：(*取自 mcdc-03-02.wav, record 89, 079120-081141*)

原始句例：都在賺錢喔賺錢 (*最後"賺錢"二字爲說話者小聲的喃喃自語*)

漢字轉記：都在賺錢 O<b mumble>賺錢</b mumble>

拼音轉記：dou1 zai4 zhuan4 qian2 O zhuan4 qian2

註： 出現在原始句例中的"喔"爲有相對應國字感嘆詞，標記員須在漢字與拼音轉記中轉寫爲大寫英文拼音"O"，有關感嘆詞的標記請參考 3.6.2。

### 3.3.2 無法辨識的語音(unrecognizable speech sound)

確屬人所發出之語音，但標記員無法辨認何字何意何音。

一、標記原則：

由於辨認不出何字何意何音，所以並無語言內容可轉記，因此以'@'標示轉

寫內容。

二、標記實例：(*取自 mcdc-01-02.wav, record 76, 033838-035657*)

原始句例：因爲…太貴了 (*"…"爲辨識不出的語言內容*)

漢字轉記：因爲<b unrecognizable speech sound>@</b unrecognizable speech sound>太貴了

拼音轉記：yin1 wei4 @ tai4 gui4 le5

### 3.3.3 不確定字/音(uncertain)

不確定字/音可標記的現象有二類：一、標記員根據前後語意，可以猜測出大概的語意內容，但無法百分之百確定。 二、標記員無法根據語意猜測出對應字詞，但可清楚記錄出其發音。

一、標記原則：

依上述標記情形不同，標記原則也區分成二類：一、根據前後語意猜測得出大概符合語意的對應字詞時，標記內容即所猜測的漢字與其標準拼音。二、無法根據前後語意猜測出對應字詞，但可辨識出清楚的發音時，漢字與拼音的標記內容都記爲[*實際發音*]。 若聲調亦可辨識出，也一併標記。

二、標記實例：

1. (*取自 mcdc-01-14.wav, record 580, 073420-122030*)

原始句例：至少我對我自己的車子有有一個瞭解程度吧 (*就聽到的語音，不確定是爲"有"字，但根據後面的語言內容"有一個瞭解程度吧"可猜測出*)

漢字轉記：至少我對我自己的車子<b uncertain>有</b uncertain>有一個瞭解程度 BA

拼音轉記：zhi4 shao3 wo3 dui4 wo3 zi4 ji3 de5 che1 zi5 you3 you3 yi2 ge5 liao2 jie3 cheng2 du4 BA

註： 出現在原始句例中的"吧"爲有相對應國字感嘆詞，故標記員須在漢字與拼音轉記中轉寫爲大寫英文拼音"BA"，有關感嘆詞的標記請參考 3.6.2。

2. (*取自 mcdc-01-03.wav, record 117, 000000-015968*)

原始句例：[fa1]因爲大概離台北市區比較遠一點所以人不會那麼多 (*在明確的語言內容前有一個不確定音[fa1]*)

漢字轉記：<b uncertain>[fa1]</b uncertain>因爲大概離台北市區比較遠一點所以人不會那麼多

拼音轉記：[fa1] yin1 wei4 da4 gai4 li2 tai2 bei3 shi4 qu1 bi3 jiao4 yuan3 yi4 dian3 suo3 yi3 ren2 bu2 hui4 na4 me5 duo1

## 3.4 不順暢的語流(disfluency)

### 3.4.1 語流中斷(prosodic disfluency)

#### 3.4.1.1 沉默(silence)

對話者因話題銜接不上而無法維持正常接話速度所產生的沉默。

一、標記原則：

標記沉默的記錄共兩筆，起始與結束的時間相同，兩個對話者各記一筆。 標記內容均以無語言內容的符號'@'標示，並忽略其他口腔發出無法辨識的聲音。

二、標記實例：(*取自 mcdc-01-05.wav, record 233/234, 097944-099626*)

第一筆記錄 (Speaker: MISC-08-male-25)：

原始句例：(*沉默 1570 毫秒*)

漢字轉記：<b silence>@</b silence>

拼音轉記：<b silence>@</b silence>

第二筆記錄 (Speaker: MISC-07-female-29)：

原始句例：(*沉默 1570 毫秒*)

漢字轉記：<b silence>@</b silence>

拼音轉記：<b silence>@</b silence>

#### 3.4.1.2 停頓(pause)

說話者在自身的語流中產生的停頓，標記員依說話者的速度判斷，若有明顯的中斷，即爲停頓，一般情況下約 600 毫秒以上。 因較長的呼吸聲所產生的語流停頓則以呼吸聲爲標記。

一、標記原則：

在說話者自身語流中的停頓處插入此標記，因無語言內容，所以以'@'符號標示。

二、標記實例：(*取自 mcdc-01-09.wav, record 392, 098046-119985*)

原始句例：然後黃線好像是九百然後有的開到一千二 (*在第二個連接詞"然後"之前有一個停頓*)

漢字轉記：然後黃線好像是九百<b pause>@</b pause>然後有的開到一千二

拼音轉記：ran2 hou4 huang2 xian4 hao3 xiang4 jiu3 bai3 @ ran2 hou4 you3 de5 kai1 dao4 yi4 qian1 er4

註： 數字轉寫成漢字時，一律以國字大寫，而非以阿拉伯數字。

### 3.4.1.3 短停頓(short break)

說話者在自身的語流中產生的短停頓，標記員依說話者的速度判斷，若有較不明顯的中斷，即爲短停頓，在大部份的情況下，短停頓不會影響語流的順暢度，且一般情況是介於 200~400 毫秒之間。 因呼吸聲所產生的語流停頓則以呼吸聲爲標記。

一、標記原則：

在說話者自身語流中的短停頓處插入此標記，且因無語言內容，所以以'@'符號標示。

二、標記實例：(*取自 mcdc-01-08.wav, record 367, 158225-162650*)

原始句例：那邊你要是熟就要鑽到吳興街那邊算近的了 (*在第一句話"那邊你要是熟"結束後，有一個短停頓*)

漢字轉記：那邊你要是熟<b short break>@</b short break>就要鑽到吳興街那邊算近的了

拼音轉記：na4 bian1 ni3 yao4 shi4 shou2 @ jiu4 yao4 zuan1 dao4 wu2 xing1 jie1 nei4 bian1 suan4 jin4 de5 le5

### 3.4.1.4 字詞片段(word fragment)

根據前後文內容知道說話者要說的是哪一個字，但說話者實際上只發了部分的音。

一、標記原則：

標記範圍爲發音不完整字本身，轉記漢字時以完整發音的漢字標記。 拼音轉記部分則以 *實際發出的部份音-與該不完整字詞的後半部* 標記。

二、標記實例：(*取自 mcdc-01-01.wav, record 25, 083970-087268*)

原始句例：外貿啊是進口嗯出口嗎 (*進口的"口"字只發了部份音**[k']***)

漢字轉記：外貿 A 是進<b word fragment>口</b word fragment>EN 出口嗎

拼音轉記：wai4 mao4 A shi4 jin4 k-o EN chu1 ko3 ma5

註： 標記員從語言內容中發現到一個詞語更正，範圍從"進口嗯出口"，而出現在詞語更正中的感嘆詞"嗯"，亦爲更正插語，但此節著重字詞片斷的標記，相關詞語更正與感嘆詞的標記請參考 3.4.3.3 與 3.6.2。

3.4.1.5 口吃(stutter)

說話不流暢，語言遲滯，有時重複字音。

一、標記原則：

標記範圍以詞爲界線，重覆的字以漢字轉寫。 若只是重覆部分字音，就以[*實際發音*]表示。

二、標記實例：(*取自 mcdc-01-01.wav, record 42, 143163-151691*)

原始句例：其實沒什麼影響因爲那個價格跟外外國人的那些商人都已經 講好了 (*說話者在講到"外國人"時，因口吃而重複"外"的部份字音**[u5]**，之後才又清楚地重講*)

漢字轉記：其實沒什麼影響因爲那個價格跟<b stutter>[u5]外國人</b stutter>的那些商人都已經講好了

拼音轉記：qi2 shi2 mei2 shen2 me5 ying3 xiang3 yin1 weii4 ne4 ge5 jia4 ge2 gen1 [u5] wai4 guo2 de5 nei4 xie1 shang1 ren2 dou1 yi3 jing1 jiang3 hao3 le5

3.4.2 不完整句法結構(lexico-syntactic disfluency)

3.4.2.1 不適當用法(inappropriate usage)

當語言內容語意大致完整，但不符合句法時，則以 *不適當用法* 標記。

一、標記原則：

標記範圍以一個主題語意爲單位。

二、標記實例：(*取自 mcdc-01-02.wav, record 107, 138553-148589*)

原始句例：可是烏來也很塞 EI 上次是我們去也是一路塞上去然後再塞下來

（*在感嘆詞"**EI**"之後的句子中有兩個"是"，造成句法不對*）

漢字轉記：可是烏來也很塞 EI <b inappropriate usage>上次是我們去也是一路塞上去然後再塞下來</b inappropriate usage>

拼音轉記：ke3 shi4 wu1 lai2 ye3 hen3 sai1 EI shang4 ci4 shi4 wo3 men5 qu4 ye3 shi4 yi2 lu4 sai1 shang4 qu4 ran2 hou4 zai4 sai1 xia4 lai2

註： 出現在原始句例中的"EI"爲無相對應國字感嘆詞，故標記員須在漢字與拼音轉記中轉寫爲大寫英文拼音"EI"，有關感嘆詞的標記請參考 3.6.2。

3.4.2.2 被對方打斷(interrupted)

當說話者還沒結束說話輪，說話權就被另一方搶走，造成句子被迫中斷。

一、標記原則：

標記範圍爲整個被中斷的不完整句子。

二、標記實例：(*取自 mcdc-01-02.wav, record 105/106, 128058-130411 / 129810-142717* )

原始句例：Speaker MISC-07-female-29: 喔去山上繞一繞是 (*句子在此被打斷*)

Speaker MISC-08-male-25: 像譬如說會去烏來啊

漢字轉記：Speaker MISC-07-female-29: O 去山上繞一繞<b interrupted>是</b interrupted>

Speaker MISC-08-male-25: 像譬如說會去烏來 A

拼音轉記：Speaker MISC-07-female-29: O qu4 shan1 shang4 rao4 yi2 rao4

Speaker MISC-08-male-25: xiang4 pi4 ru2 shuo1 hui4 qu4 wu1 lai2 A

3.4.2.3 句子中斷(abridged)

說話者本身在語法未完整前即中斷句子，並且重新開始新句。

一、標記原則：

標記範圍爲整個語法不完整的句子。

二、標記實例：(*取自 mcdc-01-03.wav, record 143, 114700-125838*)

原始句例：它有一個天呃那邊有個天籟渡假村嘛 (*說話者未講完第一句，即放棄重講*)

漢字轉記：<b abridged>它有一個天</b abridged>E 那邊有個天籟渡假村 MA

拼音轉記：ta1 you3 yi2 ge5 tian1 E ne4 bian1 you3 ge5 tian1 lai4 du4 jia4 cun1 MA

註： 出現在原始句例中的"呃"爲有相對應國字感嘆詞，故標記員須在漢字與拼音轉記中轉寫爲大寫英文拼音"E"，有關感嘆詞的標記請參考 3.6.2。

3.4.2.4 語誤(error)

標記凡可確定爲某一字詞詞彙、語法、成語或諺語之錯誤使用，但不包括語音上的錯誤。 所有語音偏差現象包括語音錯誤，皆以發音偏差(inappropriate pronunciation) 標記。

一、標記原則：

標記範圍爲錯誤字詞。

二、標記實例：

1. 詞彙錯誤 (*取自 mcdc-01-20.wav, record 826, 000000-034150*)

原始句例：你也不知道是誰開車的啊對不對你就開這張車子而已 (*修飾"車子"的量詞不正確*)

漢字轉記：你也不知道是誰開車的 A 對不對你就開<b error>這張車子</b error>而已

拼音轉記：ni3 ye3 bu4 zhi1 dao4 shi4 shei2 kai1 che1 de5 A dui4 bu2 dui4 ni3 jiu4 kai1 zhe4 zhang1 che1 zi5 er2 yi3

3.4.3 詞語修補(repair)

3.4.3.1 重覆(repetition)

說話者完整地重覆詞語一次以上，即以重複標記之。 它在句中出現的語法位置並沒有限制。 合乎語法的重覆詞語則不在此類標記範圍內(例如: 大大的)。

一、標記原則：

標記範圍爲所有完整重覆出現的詞語。

二、標記實例：(*取自 mcdc-01-13.wav, record 521, 000176-014600*)

原始句例：啊要處理可是又有又有這個情理法法理情 (*說話者重複了"又有"兩字*)

漢字轉記：A 要處理可是<b repetition>又有又有</b repetition>ZHE GE 情理法法理情

拼音轉記：A xiang3 yao4 chu3 li3 ke3 shi4 you4 you3 you4 you3 ZHE GE qing2 li3 fa3 fa3 li3 qing2

3.4.3.2 部分重覆(restart)

因他人插話被打斷或因說話者自身的緣故而重覆詞語的片斷，與完整的詞語重覆不同。 它在句中出現的語法位置並不限定。

一、標記原則：

標記範圍爲重複的詞語片斷與標記員認爲是此片斷的完整詞語。

二、標記實例：(*取自 mcdc-01-07.wav, record 294, 000000-033787*)

原始句例：真的是稍微動用一下就覺得很很不很不夠用這樣子(*說話者重複了完整詞語"很不夠用"的片斷*)

漢字轉記：真的是稍微動用一下就覺得<b restart>很很不很不夠用</b restart>這樣子

拼音轉記：zhen1 de5 shi4 shao1 wei2 dong4 yong4 yi2 xia4 jiu4 jue2 de5 hen3 hen3 bu2 hen3 bu2 gou4 yong4 zhe4 yang4 zi5

3.4.3.3 詞語更正(repair)

說話者一自覺到已說出的話不適當，就立即更正說話內容。 詞語更正包含三部分：一、需要被更正的詞語；二、更正插語；三、更正後的詞語，更正插語可有可無。 詞語更正包括四種型態：一、語意更正；二、語音更正；三、聲調更正；四、詞語更正。

一、標記原則：

$X_1$ X2…$X_n$ reparandum $M_1$ $M_2$…$M_m$ $X_1$ $X_2$…$X_n$ alteration $M_1$ $M_2$…$M_m$

$X_i$：同時出現在 reparandum 之前與 alteration 之前的字串

$M_i$：同時出現在 reparandum 之後與 alteration 之後的字串

標記範圍以 $X_1$ 之前的短語(phrase)爲起始界線，$M_m$ 之後的短語爲結束界線。只標記立即更正的詞語更正，用於補述細節而且自成完整句子的語流則不列入。

二、標記實例：

1. (*取自 mcdc-01-01.wav, record 20, 067738-074274*)
   原始句例：你您的住處就是在永春站那附近就對了 (*第二人稱從"你"更正爲有禮貌的說法"您"*)
   漢字轉記：<b repair>你您的住處</b repair>就是在永春站那附近就對了
   拼音轉記：ni3 nin2 de5 zhu4 chu4 jiu4 shi4 zai4 yong3 chun1 zhan4 na4 fu4 jin4 jiu4 dui4 le5

2. (*取自 mcdc-09-03.wav, record 75, 001720-174560*)
   原始句例：當時我才反應到我才意識到說其實愛是需要填補的 (*"反應到"更正爲"意識到"*)
   漢字轉記：當時<b repair>我才反應到我才意識到</b repair>說其實愛是需要填補的
   拼音轉記：dang1 shi2 wo3 cai2 fan3 ying4 dao4 wo3 cai2 yi4 shi4 dao4 shuo1 qi2 shi2 ai4 shi4 xu1 yao4 tian2 bu3 de5

3.4.3.4 更正插語(editing term)

更正插語可能出現在詞語更正中被更正詞語與更正後詞語之間，或是出現在完整重覆或部分重覆中的兩個重覆詞語之間。

一、標記原則：

標記範圍爲更正插語本身。

二、標記實例：

1. 詞語更正(repair) (*取自 mcdc-01-01.wav, record 25, 083970-087268*)
   原始句例：外貿啊是進口嗯出口嗎 (*說話者在將錯誤的詞語"進口"更正爲正確的"出口"前，有一個更正插語"嗯"*)
   漢字轉記：外貿 A 是進口<b editing term>EN</b editing term>出口嗎
   拼音轉記：wai4 mao4 A shi4 jin4 [k-o] EN chu1 ko3 ma5
   註：標記員從語言內容中發現到一個詞語更正，範圍從"進口嗯出口"，而出現在詞語更正中的感嘆詞"嗯"，亦爲更正插語，此節著重更正插語的標記，相關詞語更正與感嘆詞的標記請參考 3.4.3.3 與 3.6.2。另外，原始句例中"進口"的"口"字爲字詞片斷，相關之標記請參考 3.4.1.4。

2. 部分重覆(restart) (*取自 mcdc-10-01.wav, record 2, 082805-133980*)
   原始句例：我是直嗯直升機飛行員 (*在詞語片斷"直"之後，先有一更正插語"嗯"，*

*再重講該片斷的完整詞語*)

漢字轉記：我是直<b editing term>EN</b editing term>直升機飛行員

拼音轉記：wo3 shi4 zhi2 EN zhi2 sheng1 ji1 fei1 xing2 yuan2

註： 原始句例中出現的"嗯"是介於詞語片斷與該片斷之完整詞語中間的更正插語，亦爲有相對國字感嘆詞，標記員須在漢字與拼音轉記中轉寫爲大寫英文拼音"EN"，有關感嘆詞的標記請參考 3.6.2。

## 3.5 受外語或方言影響(socio-linguistic phenomena)

### 3.5.1 語言轉換(code-switching)

當說話者使用漢語以外的語言，即語言轉換現象。 對此現象，標記員須標記語言轉換，之後再針對使用的語言做標記，因此會有兩層標記，一、外層的語言轉換標記；二、內層的語言標記。

一、標記原則：

標記範圍即各語言之語言內容。 無論漢字轉記或拼音轉記，內容都以其語言慣用之書寫方式轉記。 唯獨語言內容無法以其語言慣寫的文字轉記時，如閩南語，則暫時漢字轉記部份以可翻譯成相對應的漢字爲轉記內容，拼音轉記部分則以[*實際發音*]標示。

二、標記實例：

1. 閩南語 (*取自 mcdc-01-04.wav, record 167, 033790-040690*)

原始句例：它有一個很大的看板會 (*"看板"閩南語的實際發音爲[k'aŋ2] [paŋ4]*)

漢字轉記：它有一個很大的<b code switching><b Min-Nan>看板</b Min-Nan></b code switching>會

拼音轉記：ta1 you3 yi2 ge5 hen3 da4 de5 [k_haN2] [paN4] hui4

註： 說話者尚未說完，說話權即被另一方搶走，句子被迫到"會"這個字就中斷，所以須加註被對方打斷的標記，有關此標記請參考 3.4.2.2，此節將重點著重於語言轉換。

2. 英語 (*取自 mcdc-01-09.wav, record 376, 024395-029698*)

原始句例：真正通化街那一條不是有 HANGTEN NA GIORDANO 那一些 (*說話者在講到該衣服品牌時是用該品牌的英文名字作爲指稱*)

漢字轉記：真正通化街那一條不是有<b code switching><b English>HANGTEN</b English></b code switching>NA<b code switching><b English>GIORDANO</b English></b code switching>

拼音轉記：zhen1 zheng4 tong1 hua4 jie1 nei4 yi4 tiao2 bu2 shi4 you3

HANGTEN NA GIORDANO nei4 yi4 xie1

註： 出現在原始句例中的"NA"爲無相對應國字感嘆詞，故標記員須在漢字與拼音轉記中轉寫爲大寫英文拼音"NA"，有關感嘆詞的標記請參考 3.6.2。

### 3.5.2 受閩南語影響的發音(Min-Nan-influenced pronunciation)

受閩南語影響的發音，而經由觀察可歸納出變化的規則，列如：f-h (即表示國語中[f]受閩南語影響變成[x])、v-m (英語[v]的音受閩南語影響變成[m])。 臺灣國語現象需要兩層標記，一、外層的台灣國語標記；二、內層的發音變化規律標記。

一、標記原則：

標記範圍爲受影響的字詞

二、標記實例：(*取自 mcdc-01-03.wav, record 146, 123005-139327*)

原始句例：不然泡在室內不會很溫暖 E 只是很悶而已因爲真的很熱 (*"熱"的實際發音爲[lɔ4]*)

漢字轉記：不然泡在室內不會很溫暖 E 只是很悶而已因爲真的很<b Taiwanese-influenced pronunciation><b r-l>熱</b r-l></b Taiwanese-influenced pronunciation>

拼音轉記：tong1 chang2 wo3 hui4 zai4 shi4 zheng4 fu3 xia4 che1 ran2 hou4 zai4 zhuan3 cheng2 gong1 che1 qu4 shang4 ban1

### 3.5.3 alternative-約定俗成讀音

當出現的讀音是被廣爲使用，但尚未被收錄在辭典時，便以 ***alternative-****約成俗成讀音* 標記。 是否被收錄於辭典中，我們以教育部重編國語辭典與現代漢語詞典爲參照標準。 目前觀察發現有約定俗成讀音的字詞，例如：1."符"字在字典中只有[fu2]這個讀音，沒有[fu3]這個讀音，故以 ***alternative-fu3*** 來標記； 2."嗎"字在字典中有[ma2]、 [ma3] 和[ma5]這個讀音，沒有[ma1]這個讀音，故以 ***alternative-ma1*** 標記。

一、標記原則：

標記範圍即爲該漢字，但在拼音轉記部份還是以其相對應的標準讀音爲轉記內容。 如此一來我們可得知該字音由哪個標準讀音轉變成哪個約定俗成讀音。

二、標記實例：(*取自 mcdc-01-02.wav, record 108, 142920-160763*)

原始句例：真的在那邊車子開個八九十沒有關係啊 (*"係"的實際發音爲[ci1]*)

漢字轉記：真的在那邊車子開個八九十沒有關<b alternative-xi1>係</b alternative-xi1>A

拼音轉記：zhen1 de5 zai4 ne4 bian1 che1 zi5 kai1 ge5 ba1jiu3 shi2 mei2 you3 guan1 xi4 A

3.6 其他(others)

3.6.1 語助詞(marker)

說話者本身在語流中慣用的插入語，這些習慣插語有其基本詞彙意義。 但在語流中習慣插語已不保有其原有的完整語意，而較具語用功能。 例如，作用於口語中說話者意欲保有其說話權且又需緩衝時間去思索組織其想說的句子，此時習慣插語 NA 便常被使用。 語流中常出現的語助詞包括有 NA、NE、NA GE、NE GE、NEI GE、SHEN ME、ZHE GE。

一、標記原則：

標記範圍爲習慣插語本身。 爲與其標準詞彙使用區別，以大寫漢語拼音方式轉記，不寫出漢字。

二、標記實例：(*取自 mcdc-01-03.wav, record 139, 100158-109932*)

原始句例：室內就是一小間一小間嘛那露天就是大家一起啊 (*"那"屬語流中習慣插語，已不保有指涉某物的意義*)

漢字轉記：室內就是一小間一小間 MA <b marker>NA</b marker>露天就是大家一起 A

拼音轉記：shi4 nei4 jiu4 shi4 yi4 xiao3 jian1 yi4 xiao3 jian1 MA NA lu4 tian1 jiu4 shi4 da4 jia1 yi4 qi3 A

註： "大家"的"家"有鼻化音現象，其標記請參考 3.2.4，此節只探討語助詞標記。

3.6.2 感歎詞(particle)

不具標準語意的感嘆詞，其語用成份居多如回應或同意。 語流中出現的感歎詞有四類，一、有相對應國字的感歎詞，例如 A、AI YA、AI YOU、BA、E、EP、EN、HAI、HE、HEI、HWA、LA、LIE、LEI、LO、MA、NOU、NO、O、OU、WA、YE、YI、YOU；二、無相對應國字的感歎詞，例如 AI YE、EI、HEN、HON、NA、NE、NEI、ON；三、源於台語的感歎詞，例如 EIN、HAN、HEIN、HO；四、其他的感歎詞(Fillers)，例如 UHN、UHNN、UHNHN、UHM、UHMM、

UHMHM、NHN、NHNN、NHNHN、MHM、MHMM、MHMHM、MHMHMHM、MHMHMHMHM。

一、標記原則：

標記範圍爲感嘆詞本身，即使有相對應的國字，也以大寫漢語拼音方式轉記，不寫出漢字。

二、標記實例：

1. 有相對應國字的感嘆詞 (*取自 mcdc-01-03.wav, record 117, 000000-015968*)

原始句例：去什麼富基漁港啊那些 (*專有名詞"富基漁港"後面跟著一個不具標準語意的感嘆詞"啊"*)

漢字轉記：去什麼富基漁港<b particle>A</b particle>那些

拼音轉記：qu4 shen2 me5 fu4 ji1 yu2 gang3 A na4 xie1

2. 無相對應國字的感嘆詞 (*取自 mcdc-01-01.wav, record 1, 000000-009514*)

原始句例：EI 你好我姓賴請問一下貴姓 （*句子一開始，即以一個不具標準語意的感嘆詞"EI"作爲起頭*）

漢字轉記：<b particle>EI</b particle>你好我姓賴請問一下貴姓

拼音轉記：EI ni2 hao3 wo3 xing4 lai4 qing3 wen4 yi2 xia4 gui4 xing4

## 4 非語音部份口語標註(Non-Speech Sounds)

4.1 人聲(human voice)

凡非語音但確定由人所發出的聲音，包括笑聲、咳嗽聲、呼吸聲、吐氣聲、吸氣聲、咂嘴聲、嘖舌聲、嘆氣聲、打嗝聲、噴嚏聲、哈欠聲、吞口水聲、清喉嚨聲 和其他口腔發出無法辯識的聲音等等。

標記規則與實例：

一、伴隨語言內容之人聲，其人聲標記之內容部份爲語言內容。

1. 笑聲(laugh)：(*取自 mcdc-01-05.wav, record 205, 000000-016204*)

原始句例：我覺得今天我少吃一點花個三百塊跟直接投資三萬塊這差很多 (*一邊笑一邊講"差很多"*)

漢字轉記：我覺得今天我少吃一點花個三百塊跟直接投資三萬塊這<b laugh>差很多</b laugh>

拼音轉記：wo3 jue2 de5 jin1 tian1 wo3 shao3 chi1 yi4 dian3 hua1 ge5 san1 bai3 kuai4 gen1 zhi2 jie1 tou2 zi1 san1 wan4 kuai4 zhe4 cha1 hen3 duo1

二、無伴隨語言內容之人聲，其人聲標記內部份以'@'符號標記。

1. 笑聲(laugh)：(*取自 mcdc-01-09.wav, record 382, 044700-048531*)

原始句例：大概是我們的運氣不好　　(*講完，即笑*)

漢字轉記：大概是我們的運氣不好<b laugh>@</b laugh>

拼音轉記：da4 gai4 shi4 wo3 men5 de5 yun4 qi4 bu4 hao3 @

4.2 非人聲(non human sound)

4.2.1 室內雜音(noise in room)

非語音且確定非人所發出的聲音，包括雨聲、手機聲、搓揉紙聲等等。

標記規則與實例：

一、伴隨語言內容之非人聲，其非人聲標記之內容部份爲語言內容。

1. 下雨聲：(*取自 mcdc-26-03.wav, record 106, 044505-086450*)

原始句例：像我工作就是在那邊去看到的 (*說話同時，背景帶有雨聲*)

漢字轉記：<b noise in room>像我工作就是在那邊去看到的</b noise in room>

拼音轉記：xiang4 wo3 gong1 zuo4 jiu4 shi4 zai4 na4 bian1 kan4 dao4 de5

二、無伴隨語言內容之非人聲，其非人聲標記內部份以'@'符號取代之。

1. 麥克風聲：(*取自 mcdc-10-12.wav, record 401, 153244-160305*)

原始句例：NHN (*在說話者發出感嘆詞"NHN"之前，有一個敲到麥克風的聲音*)

漢字轉記：<b noise in room>@</b noise in room>NHN

拼音轉記：@ NHN

## 5 其他

5.1 符號 >

同一筆語音內容，因聲音檔切割緣故分別出現在兩個聲音檔的最後及最前，因此原轉記在同一筆記錄的語音內容被迫分成兩筆記錄，而爲了讓程式能辨別此

兩筆記錄實爲同一筆語音內容，故將此 > 符號置於第二筆記錄語音內容最前面，當程式讀取時，碰到時間點爲 000000 以及 > 符號，便可得知此筆記錄內容與相同發音人的前一筆記錄爲同一發言輪之語言內容。

## 6 參考文獻


Janet A. Edwards, Martin D. Lampert, 1993. *Talking Data: Transcription and Coding in Discourse Research*. Lawremce Erlbaum Associates Inc.

C. Barras, E. Geoffrois, Z. Wu and M. Liberman. 2001. *Transcriber: Development and Use of a Tool for Assisting Speech Corpora Production*. Speech Communication. 33/1-2:5-22.

S. Bird and M. Liberman. 2001. *A Formal Framework for Linguistic Annotation*. Speech Communication. 33/1-2:23-60.

J. Carbonell and P. Hayes. 1983. *Recovery Strategies for Parsing Extragrammatical Language*. American Journal of Computational Linguistics. (3/4):123-146.

Y. R. Chao, 1968. *A Grammar of Spoken Chinese*. Berkeley: University of California Press.

K.-W. Chui, 1996. *Organization of Repair in Chinese Conversation*. Text 16/3, pp. 343-372.

CKIP, 1995. 中央研究院平衡語料庫的內容與說明, Technical Report no. 95-02/98-04.

Janet A. Edwards and Martin D. Lampert. 1993. *Talking Data: Transcription and Coding in Discourse Research.* Lawremce Erlbaum Associates Inc.

J. J. Godfrey, E.C. Holliman, and McDaniel, 1992. *SWITCHBOARD: Telephone Speech Corpus for Research Development*. In Proc. of the IEEE Conference on Acoustics, Speech and Signal Processing, pp. 517-520.

D. Gross, J. Allen, and D. Traum. 1993. *The TRAINS 91 Dialogues*. Technical Report 92-1, Dept. of Computer Science. University of Rochester.

Guoyucidian, 1995. 教育部重編國語辭典. 臺灣商務印書館.

C.T. Huang, 1982. *Logical Relations in Chinese and the Theory of Grammar*. PhD Thesis, MIT.

D. Hindle. 1983. *Deterministic Parsing of Syntactic Non-fluencies*. In

Proc. of ACL ’83. 123-128.
E. Hovy and D. Scott. (eds.) 1996. *Computational and Conversational Discourse*. Burning Issues - An Interdisciplinary Account. Springer.
D. Jurafsky, E. Shriberg, and E. Fox, 1998. *Lexical, Prosodic, and Syntactic Cues for Dialog Acts*. In Proc. of ACL-COLING 98 Workshop on Discourse Relations and Discourse Markers.
J. C. Kowtko and P.J. Price, 1989. *Data Collection and Analysis in the Air Planning Domain*. In Proc. of the DARPA Speech and Natural Language Workshop, pp. 119-125.
Y.-S. Lee, and H.-H. Chen, 1997. *Using Acoustic and Prosodic Cues to Correct Chinese Speech Repairs.* In: Proceedings of EUROSPEECH’97. Rhodes, Greece. pp. 2211-2214.
W. J. Levelt, 1989. *Speaking: From Intention to Articulation*. MIT Press, Cambridge, MA.
C. Li, and S. Thompson, 1981. *Mandarin Chinese: A Functional Reference Grammar*. Berkeley: University of California Press.
MADCOW, 1992. *Multi-Site Data Collection for a Spoken Language Corpus*. In Proc. of Speech and Natural Language Workshop, pp. 7-14.
C. Nakatani, J. Hirschberg, and B. Grosz. 1995. *Discourse Structure in Spoken Language: Studies on Speech Corpora*. Presented at the AAA-I-95 Spring Symposium on Empirical Methods in Discourse Interpretation and Generation.
S.G. Nooteboom, 1980. *Speaking and Unspeaking: Detection and Correction of Phonological and Lexical Errors in Spontaneous Speech*. In Errors in Linguistic Performance. Ed. Fromkin. Academic Press.
J. Pustejovsky, and B. Boguraev, 1993. *Lexical Knowledge Representation and Natural Language Processing*. Artificial Intelligence 63;193-223.
J. Pynte, and B. Prieur, 1996. *Prosodic Breaks and Attachment Decisions in Sentence Parsing. Language and Cognitive Processes* 11:1;165-192.
P. Roach, G. Knowles, T. Varadi, and S. Arnfield, 1993. *Marsec: A Machine-Readable Spoken English Corpus*. Journal of the International Phonetic Association, 23(1):47-53.
G. Sagerer, H. Eikmeyer, and G. Rickheit, 1994. *"Wir bauen jetzt ein*

*Flugzeug": Konstruieren im Dialog*. Arbeitsmaterialien. Technical Report, SFB 360 "Situierte künstliche Kommunikatoren", Universität Bielefeld.
L. Stirling, J. Fletcher, I. Mushin and R. Wales. 2001. *Representational Issues in Annotation: Using the Australian Map Task Corpus to Relate Prosody and Discourse Structure*. Speech Communication. 33/1-2:113-134.
A. Syrdal, J. Hirschberg, J. McGory and M. Beckman, 2001. *Automatic ToBI Prediction and Alignment to Speech Manual Labeling of Prosody*. Speech Communication. 33/1-2:135-151.
D. Traum and P. Heeman. 1997. *Utterance Unit in Spoken Dialogue*. In Dialogue Processing in Spoken Language Systems. Maier/Mast/Luper Foy (eds.). Lecture Notes in Artificial Intelligence. Springer Verlag.
Xiandaihanyu Cidian, 2001. 現代漢語詞典.商務印書館.

## 附錄一、標記系統總表

| 標記標籤 | 標記現象 | 標記原則 | 原始語流<br>=> 標記範例 |
|---|---|---|---|
| **laugh** | 笑聲 | 一、伴隨語言內容之笑聲，其笑聲標記之內容部份爲語言內容<br><br>二、無伴隨語言內容之笑聲，其笑聲標記內部份以'@'符號標記 | 一、我覺得今天我少吃一點花個三百塊跟直接投資三萬塊這差很多(*一邊笑一邊講"差很多"*)<br>=> 我覺得今天我少吃一點花個三百塊跟直接投資三萬塊這<b laugh>差很多</b laugh><br><br>二、大概是我們的運氣不好 (*講完，即笑*)<br>=> 大概是我們的運氣不好<b laugh>@</b laugh> |
| **cough** | 咳嗽聲 | 同笑聲 | |
| **breathe** | 呼吸聲 | 同笑聲 | |
| **exhale** | 吐氣聲 | 同笑聲 | |
| **inhale** | 吸氣聲 | 同笑聲 | |
| **smack** | 咂嘴聲 | 同笑聲 | |
| **click** | 嘖舌聲 | 同笑聲 | |
| **sigh** | 嘆氣聲 | 同笑聲 | |
| **hiccup** | 打嗝聲 | 同笑聲 | |
| **sneeze** | 噴嚏聲 | 同笑聲 | |
| **yawn** | 哈欠聲 | 同笑聲 | |
| **swallow** | 吞口水聲 | 同笑聲 | |
| **clear throat** | 清喉嚨聲 | 同笑聲 | |
| **unrecognizablenon-speech sound** | 其他由人發出非語音，而且無法辨識的聲音 | 同笑聲 | |
| **lengthening** | 音節拖長現象，不限定位於音節的哪個位置 | 標記有拖長音的音節 | 我目前是從事外貿 (*"事"的[]有拖長現象*)<br>=>我目前是從<b lengthening>事</b lengthening>外貿 |

| | | | |
|---|---|---|---|
| **assimilation** | 字的發音受到相鄰音發音部位或方法的影響而改變本身發音的同化現象。 同化現象可以是因受相鄰音影響而增加的音，也可以是受相鄰音的影響而使原本的發音改變成與其相鄰音發音部位或方法相近或一致的音 | 標記範圍包括被同化的字與使其發音產生變化的相鄰字。 轉記漢字時以標準發音的漢字轉寫，拼音部分被同化字的字音則以實際發音轉寫，實際發音採用 SAMPA-M（*參考 3.1 漢語口語語音標音系統*），並置於中括號[ ]內。 若有音節省略現象時，漢字與拼音相差的音節則以 % 標示 | 賴先生呢您從事什麼工作 (*"呢"受到"您"的影響，在音節未增加[n]的音*)<br>=> 賴先生<b assimilation> 呢您</b assimilation>從事什麼工作 |
| **syllable contraction** | 說話者說得太快或不清楚時出現的音節合併現象。 合併現象有三，一是清楚可辨的音節短少，像是從原本正常的三個字三個音節變成三個字兩個音節，或者是兩個字兩個音節變成兩個字一個音節，例如: "我們"的實際發音變成[om]；二是音節雖無短少，但卻都連在一起，難以切割，例如: "就是"的實際發音變成[tɕio˥]；三是音節無短少且音節可切割，只是音節結構有變，例如："誇張"的實際發音變成[k'ua1] [aŋ1] | 標記範圍包括所有音節合併的字。 拼音部份仍以標準發音的漢語拼音轉寫，而非以實際發音轉寫 | 但是相對跟淡水啊那種什麼木柵那邊比就少很多了 (*"那種"實際發音爲[noŋ]*)<br>=> 但是相對跟淡水 A<b syllable contraction>那種</b syllable contraction>什麼木柵那邊比就少很多了 |
| **nasalized** | 標準字音中無任何鼻音，但實際發音卻出現鼻化音現象。 與音的同化現象不同的是，鼻化音現象並非受相鄰鼻音影響去改變本身的發音部位或方法，而只是整個標準字音充斥著鼻音而已 | 標記範圍包括所有帶鼻音的字。 拼音部份仍以標準發音的漢語拼音轉寫，而非以實際發音轉寫 | 室內就是一小間一小間嘛那露天就是大家一起啊 (*"家"實際發成[tɕia]整個帶有鼻音*)<br>=> 室內就是一小間一小間 MA NA 露天就是大<b nasalized>家</b nasalized>一起 A |

| | | | |
|---|---|---|---|
| **silence** | 對話者因話題銜接不上而無法維持正常接話速度所產生的沉默 | 標記沉默的記錄共兩筆，起始與結束的時間相同，兩個對話者各記一筆。 標記內容均以無語言內容的符號'@'標示，並忽略其他口腔發出無法辨識的聲音 | 第一筆記錄 (Speaker: MISC-08-male-25)：<br>原始句例：(*沉默 1570 毫秒*)<br>漢字轉記：<b silence>@</b silence><br>拼音轉記：<b silence>@</b silence><br><br>第二筆記錄 (Speaker: MISC-07-female-29)：<br>原始句例：(*沉默 1570 毫秒*)<br>漢字轉記：<b silence>@</b silence><br>拼音轉記：<b silence>@</b silence> |
| **pause** | 說話者在自身的語流中產生的停頓，標記員依說話者的速度判斷，若有明顯的中斷，即爲停頓，一般情況下約 600 毫秒以上。 因較長的呼吸聲所產生的語流停頓則以呼吸聲爲標記 | 在說話者自身語流中的停頓處插入此標記，因無語言內容，所以以'@'符號標示 | 然後黃線好像是九百然後有的開到一千二 (*在第二個連接詞"然後"之前有一個停頓*)<br>=> 然後黃線好像是九百 <b pause>@</b pause>然後有的開到一千二 |
| **short break** | 說話者在自身的語流中產生的短停頓，標記員依說話者的速度判斷，若有較不明顯的中斷，即爲短停頓，在大部份的情況下，短停頓不會影響語流的順暢度，且一般情況是介於 200~400 毫秒之間。 因呼吸聲所產生的語流停頓則以呼吸聲爲標記 | 在說話者自身語流中的短停頓處插入此標記，且因無語言內容，所以以'@'符號標示 | 那邊你要是熟就要鑽到吳興街那邊算近的了 (*在第一句話"那邊你要是熟"結束後，有一個短停頓*)<br>=>那邊你要是熟<b short break>@</b short break>就要鑽到吳興街那邊算近的了 |
| **error** | 標記凡可確定爲某一字詞詞彙、語法、成語或諺語之錯誤使用，但不包括語音上的錯誤。 所有語音偏差現象包括語音錯誤，皆以發音偏差 (inappropriate pronunciation) 標記 | 標記範圍爲錯誤字詞 | **詞彙錯誤：**你也不知道是誰開車的啊對不對你就開這張車子而已 (*修飾"車子"的量詞不正確*)<br>=>你也不知道是誰開車的 A 對不對你就開<b error>這張車子</b error>而已 |

| | | | |
|---|---|---|---|
| **inappropriate pronunciation** | 說話者發音偏離原字詞標準發音，但標記員依據談話內容，仍可辨識出原字詞爲何，且其母音、子音部分須清楚可辨 | 標記範圍爲發音偏差字詞本身，轉記漢字時以標準發音的漢字轉寫。 漢語拼音部分則以[*實際發音*]轉寫，呈現出該字詞的偏差發音，若聲調亦可辨識出，一併轉記在[*實際發音*]內。 漢字與拼音相差的音節則以 % 標示 | 我比較喜歡從事一些球類運動啦 (*"比"的實際發音爲[pu2]*)<br>=>我<b inappropriate pronunciation>比</b inappropriate pronunciation>較喜歡從事一些球類運動 LA |
| **inappropriate usage** | 當語言內容語意大致完整，但不符合句法時，則以 *不適當用法* 標記 | 標記範圍以一個主題語意爲單位 | 可是烏來也很塞 EI 上次是我們去也是一路塞上去然後再塞下來（*在感嘆詞"EI"之後的句子中有兩個"是"，造成句法不對*）<br>=> 可是烏來也很塞 EI <b inappropriate usage>上次是我們去也是一路塞上去然後再塞下來</b inappropriate usage> |
| **interrupted** | 當說話者還沒結束說話輪，說話權就被另一方搶走，造成句子被迫中斷 | 標記範圍爲整個被中斷的不完整句子 | SpeakerMISC-07-female-29: 喔去山上繞一繞是 (*句子在此被打斷*)<br>Speaker MISC-08-male-25: 像譬如說會去烏來啊<br>=>SpeakerMISC-07-female-29: O 去山上繞一繞<b interrupted>是</b interrupted><br>Speaker MISC-08-male-25: 像譬如說會去烏來 A |
| **abridged** | 說話者本身在語法未完整前即中斷句子，並且重新開始新句 | 標記範圍爲整個語法不完整的句子 | 它有一個天呃那邊有個天籟渡假村嘛 (*說話者未講完第一句，即放棄重講*)<br>=> <b abridged>它有一個天</b abridged>E 那邊有個天籟渡假村 MA |

| | | | |
|---|---|---|---|
| **repair** | 說話者一自覺到已說出的話不適當，就立即更正說話內容。 詞語更正包含三部分：一、需要被更正的詞語；二、更正插語；三、更正後的詞語，更正插語可有可無。 詞語更正包括四種型態：一、語意更正；二、語音更正；三、聲調更正；四、詞語更正 | $X_1$ X2…$X_n$ **reparandum** $M_1$ $M_2$…$M_m$ $X_1$ $X_2$…$X_n$ **alteration** $M_1$ $M_2$…$M_m$<br>$X_i$：同時出現在 **reparandum** 之前與 **alteration** 之前的字串<br>$M_i$：同時出現在 **reparandum** 之後與 **alteration** 之後的字串<br><br>標記範圍以 $X_1$ 之前的短語(phrase)爲起始界線，$M_m$ 之後的短語爲結束界線。只標記立即更正的詞語更正，用於補述細節而且自成完整句子的語流則不列入 | 你您的住處就是在永春站那附近就對了 (*第二人稱從"你"更正爲有禮貌的說法"您"*)<br>=> <b repair>你您的住處</b 	repair>就是在永春站那附近就對了<br><br>當時我才反應到我才意識到說其實愛是需要填補的 (*"反應到"更正爲"意識到"*)<br>=> 當時<b repair>我才反應到我才意識到</b repair>說其實愛是需要填補的 |
| **editing term** | 更正插語可能出現在詞語更正中被更正詞語與更正後詞語之間，或是出現在完整重覆或部分重覆中的兩個重覆詞語之間 | 標記範圍爲更正插語本身 | **詞語更正**(repair)：外貿啊是進口嗯出口嗎 (*說話者在將錯誤的詞語"進口"更正爲正確的"出口"前，有一個更正插語"嗯"*)<br>=>外貿 A 是進口<b editing term>EN</b editing term>出口嗎<br><br>**部分重覆**(restart)：我是直嗯直升機飛行員 (*在詞語片斷"直"之後，先有一更正插語"嗯"，再重講該片斷的完整詞語*)<br>=>我是直<b editing term>EN</b editing term>直升機飛行員 |
| **restart** | 因他人插話被打斷或因說話者自身的緣故而重覆詞語的片斷，與完整的詞語重覆不同。 它在句中出現的語法位置並不限定 | 標記範圍爲重複的詞語片斷與標記員認爲是此片斷的完整詞語 | 真的是稍微動用一下就覺很很不很不夠用這樣子(*說話者重複了完整詞語"很不夠用"的片斷*)<br>=> 真的是稍微動用一下就覺得<b restart>很很不很不夠用</b restart>這樣子 |
| **repetition** | 說話者完整地重覆詞語一次以上，即以重複標記之。 它在句中出現的語法位置並沒有限制。合乎語法的重覆詞語則不在此類標記範圍內(例如: 大大的) | 標記範圍爲所有完整重覆出現的詞語 | 啊要處理可是又有又有這個情理法法理情 (*說話者重複了"又有"兩字*)<br>=> A 要處理可是<b repetition>又有又有</b repetition>ZHE GE 情理法法理情 |

| | | | |
|---|---|---|---|
| **word fragment** | 根據前後文內容知道說話者要說的是哪一個字，但說話者實際上只發了部分的音 | 標記範圍爲發音不完整字本身，轉記漢字時以完整發音的漢字標記。 拼音轉記部分則以 *實際發出的部份音-與該不完整字詞的後半部* 標記 | 外貿啊是進口嗯出口嗎 (*進口的"口"字只發了部份音[k']*)<br>=> 外貿 A 是進<b word fragment>口</b word fragment>EN 出口嗎 |
| **stutter** | 說話不流暢，語言遲滯，有時重複字音 | 標記範圍以詞爲界線，重覆的字以漢字轉寫。 若只是重覆部分字音，就以[*實際發音*]表示 | 其實沒什麼影響因爲那個價格跟外外國人的那些商人都已經 講好了 (*說話者在講到"外國人"時，因口吃而重複"外"的部份字音[u5]，之後才又清楚地重講*)<br>=> 其實沒什麼影響因爲那個價格跟<b stutter>[u5]外國人</b stutter>的那些商人都已經講好了 |
| **marker** | 說話者本身在語流中慣用的插入語，這些習慣插語有其基本詞彙意義。 但在語流中習慣插語已不保有其原有的完整語意，而較具語用功能。 例如，作用於口語中說話者意欲保有其說話權且又需緩衝時間去思索組織其想說的句子，此時習慣插語 NA 便常被使用。 語流中常出現的語助詞包括有 NA、NE、NA GE、NE GE、NEI GE、SHEN ME、ZHE GE | 標記範圍爲習慣插語本身。 爲與其標準詞彙使用區別，以大寫漢語拼音方式轉記，不寫出漢字 | 室內就是一小間一小間嘛那露天就是大家一起啊 (*"那"屬語流中習慣插語，已不保有指涉某物的意義*)<br>=> 室內就是一小間一小間 MA<b marker>NA</b marker>露天就是大家一起 A |

| | | | |
|---|---|---|---|
| **particle** | 不具標準語意的感嘆詞，其語用成份居多如回應或同意。 語流中出現的感歎詞有四類，一、有相對應國字的感歎詞，例如 A、AI YA、AI YOU、BA、E、EP、EN、HAI、HE、HEI、HWA、LA、LIE、LEI、LO、MA、NOU、NO、O、OU、WA、YE、YI、YOU；二、無相對應國字的感歎詞，例如 AI YE、EI、HEN、HON、NA、NE、NEI、ON；三、源於台語的感歎詞，例如 EIN、HAN、HEIN、HO；四、其他的感歎詞 (Fillers)，例如 UHN、UHNN、UHNHN、UHM、UHMM、UHMHM、NHN、NHNN、NHNHN、MHM、MHMM、MHMHM、MHMHMHM、MHMHMHMHM | 標記範圍爲感嘆詞本身，即使有相對應的國字，也以大寫漢語拼音方式轉記，不寫出漢字 | 一、有相對應國字的感嘆詞：去什麼富基漁港啊那些 (*專有名詞"富基漁港"後面跟著一個不具標準語意的感嘆詞"啊"*)<br>=> 去什麼富基漁港<b particle>A</b particle>那些<br>二、無相對應國字的感嘆詞：EI 你好我姓賴請問一下貴姓（*句子一開始，即以一個不具標準語意的感嘆詞"EI"作爲起頭*）<br>=><b particle>EI</b particle>你好我姓賴請問一下貴姓 |
| **mumble** | 說話者不是在回應對方以接續話題的語流，而是他無意讓對方聽見而小聲的喃喃自語。 標記爲喃喃自語的語言內容必須是清楚可辨的，若是語言內容無法辨識則標記爲無法辨識的語音或不確定的字音(*參考 3.3.2 或 3.3.3*) | 標記範圍即是說話者喃喃自語的內容 | 都在賺錢喔賺錢 (*最後"賺錢"二字爲說話者小聲的喃喃自語*)<br>=> 都在賺錢 O<b mumble>賺錢</b mumble> |
| **unrecognizable speech sound** | 確屬人所發出之語音，但標記員無法辨認何字何意何音 | 由於辨認不出何字何意何音，所以並無語言內容可轉記，因此以'@'標示轉寫內容 | 因爲...太貴了 (*"..."爲辨識不出的語言內容*)<br>=> 因爲 <b unrecognizable speech sound>@</b unrecognizable speech sound>太貴了 |

| | | | |
|---|---|---|---|
| **uncertain** | 不確定字/音可標記的現象有二類：一、標記員根據前後語意，可以猜測出大概的語意內容，但無法百分之百確定。二、標記員無法根據語意猜測出對應字詞，但可清楚記錄出其發音 | 依上述標記情形不同，標記原則也區分成二類：一、根據前後語意猜測得出大概符合語意的對應字詞時，標記內容即所猜測的漢字與其標準拼音。二、無法根據前後語意猜測出對應字詞，但可辨識出清楚的發音時，漢字與拼音的標記內容都記爲[*實際發音*]。 若聲調亦可辨識出，也一併標記 | 一、至少我對我自己的車子有有一個瞭解程度吧 (*就聽到的語音，不確定是爲"有"字，但根據後面的語言內容"有一個瞭解程度吧"可猜測出*)<br>=> 至少我對我自己的車子<b uncertain>有</b uncertain>有一個瞭解程度 BA<br><br>二、[fa1]因爲大概離台北市區比較遠一點所以人不會那麼多 (*在明確的語言內容前有一個不確定音[fa1]*)<br>=> <b uncertain>[fa1]</b uncertain>因爲大概離台北市區比較遠一點所以人不會那麼多 |
| **noise in room** | 非語音且確定非人所發出的聲音，包括雨聲、手機聲、搓揉紙聲等等 | 一、伴隨語言內容之非人聲，其非人聲標記之內容部份爲語言內容<br><br>二、無伴隨語言內容之非人聲，其非人聲標記內部份以'@'符號取代之 | 一、像我工作就是在那邊去看到的 (*說話同時，背景帶有雨聲*)<br>=> <b noise in room>像我工作就是在那邊去看到的</b noise in room><br><br>二、NHN (*在說話者發出感嘆詞"NHN"之前，有一個敲到麥克風的聲音*)<br>=> <b noise in room>@</b noise in room>NHN |
| **alternative-<br>約定俗成讀音** | 當出現的讀音是被廣爲使用，但尚未被收錄在辭典時，便以 *alternative-約成俗成讀音* 標記。是否被收錄於辭典中，我們以教育部重編國語辭典與現代漢語詞典爲參照標準。 目前觀察發現有約定俗成讀音的字詞，例如：1."符"字在字典中只有[fu2]這個讀音，沒有[fu3]這個讀音，故以 *alternative-fu3* 來標記； 2."嗎"字在字典中有[ma2]、 [ma3] 和[ma5]這個讀音，沒有[ma1]這個讀音，故以 *alternative-ma1* 標記 | 標記範圍即爲該漢字，但在拼音轉記部份還是以其相對應的標準讀音爲轉記內容。 如此一來我們可得知該字音由哪個標準讀音轉變成哪個約定俗成讀音 | 真的在那邊車子開個八九十沒有關係啊 (*"係"的實際發音爲[ɕi1]*)<br>=> 真的在那邊車子開個八九十沒有關 <b alternative-xi1> 係 </b alternative-xi1>A |

| | | | |
|---|---|---|---|
| **code switching** | 當說話者使用漢語以外的語言，即語言轉換現象。 對此現象，標記員須標記語言轉換，之後再針對使用的語言做標記，因此會有兩層標記，一、外層的語言轉換標記；二、內層的語言標記 | 標記範圍即各語言之語言內容。 無論漢字轉記或拼音轉記，內容都以其語言慣用之書寫方式轉記。 唯獨語言內容無法以其語言慣寫的文字轉記時，如閩南語，則暫時漢字轉記部份以可翻譯成相對應的漢字爲轉記內容，拼音轉記部分則以[*實際發音*]標示 | |
| **Min-Nan** | 閩南語 | | 它有一個很大的看板會 (*"看板"閩南語的實際發音爲[k'aŋ2] [paŋ4]*)<br>=>它有一個很大的<b code switching><b Min-Nan>看板</b Min-Nan></b code switching>會 |
| **Hakka** | 客家話 | | |
| **English** | 英語 | | 真正通化街那一條不是有 HANGTEN NA GIORDANO 那一些 (*說話者在講到該衣服品牌時是用該品牌的英文名字作爲指稱*)<br>=>真正通化街那一條不是有<b code switching><b English>HANGTEN</b English></b code switching>NA<b code switching><b English>GIORDANO</b English></b code switching>了 |
| **Japanese** | 日語 | | |
| **Cantonese** | 廣東話 | | |
| **>** | 同一筆語音內容，因聲音檔切割緣故分別出現在兩個聲音檔的最後及最前，因此原轉記在同一筆記錄的語音內容被迫分成兩筆記錄，而爲了讓程式能辨別此兩筆記錄實爲同一筆語音內容，故將此 **>** 符號置於第二筆記錄語音內容最前面，當程式讀取時，碰到時間點爲 000000 以及 **>** 符號，便可得知此筆記錄內容與相同發音人的前一筆記錄爲同一發言輪之語言內容 | | |

## 附錄二、TransList 標記須知

1. 插入標記的方式是將所要標記的內容範圍反白再按滑鼠右鍵選擇所要插入的標記。
2. 當沒有辦法利用滑鼠右鍵來複製與貼上文字轉記的內容時，可以利用鍵盤上 Ctrl+C 來複製，Ctrl+V 來貼上。
3. 漢字轉記部份除了漢字以及@符號外，其餘用英文轉記的無論是語助詞、感歎詞或是實際發音一律在其後面空一格。