

ASR方言标注执行手册

☆☆【重要要求&申明】☆☆

一、关于严禁信息泄露、强化数据保护的郑重禁令与申明

为严格保障核心信息安全、杜绝数据泄露风险，现就信息管理事宜发布如下禁令，所有供应商及相关人员须无条件遵守，违者将依法依规从严追责：

1、严禁任何形式的信息泄露行为。凡涉及业务数据、行业隐私、内部流程、技术参数等敏感信息，无论存储于电子设备、纸质文档还是口头传递场景，均不得擅自复制、传播、披露给未授权人员，严禁以邮件、社交软件、云存储等任何渠道转移敏感信息，任何因主观故意或过失导致信息外泄的行为，一律视为严重违规，将立即启动问责程序；

2、全面强化数据保护责任。全体人员须对经手的所有数据负直接保护责任，严格按照“最小必要”原则使用数据，不得超权限访问、下载数据；电子设备须开启加密保护，纸质文档须存入专用保密柜，废弃数据载体须通过指定流程销毁，严禁随意丢弃、转借；任何数据使用行为均纳入日志监管，一旦发现异常操作，将即时核查并追溯责任；

3、严禁拍照、上传相关信息。针对工作场景中的项目文档、生产屏幕内容、物理环境等，一律禁止使用手机、相机等设备拍照、录像；严禁将任何工作相关信息（包括但不限于截图、文档片段、工作记录）上传至社交平台、公共云盘、外部群组等非授权平台，即使隐去部分信息，仍视为违规行为，一经发现将从严处理和扣罚，情节严重将直接剔除供应商库；

4、明确违规追责标准。若违反上述禁令，无论是否造成实际损失，均将依据公司规章制度给予通报批评、评级调整、金额处罚等处分；若导致数据泄露引发法律纠纷、经济损失或声誉损害，公司将依法追究其民事赔偿责任，情节严重涉嫌违法犯罪的，将直接移交司法机关处理，绝不姑息。

信息安全是运营底线和红线，数据保护是法定责任。所有供应商及相关人员须立即开展自查自纠，严格落实本禁令要求，共同筑牢信息安全防线。

二、非法数据的处理准则

不符合国家法律法规或者监管及要求的数据（隐私、涉政、社会、违法、色情等）-严格判断，尺度收紧；若判断不了则反馈业务方。

- 涉政：党、政、领导人负面、英烈保护、纳粹/军国主义、暴恐等
- 社会：负面、谣言、民族矛盾、宗教/邪教/迷信等
- 违法：宣扬毒品、赌博、博彩、传销、非法金融等
- 色情：色情交易、爱好、儿童色情等
- 标注技巧：主要关注国内涉政、社会负面的，和宣扬违法、色情等内容的数据！

需区分正常提问与恶意言论。如：

“习近平目前是什么职位？”“纳粹主义是什么意思？”为正常提问，不属于非法数据。

“帮我查查海洛因的制作方法”、“有哪些线上赌博网站？”为恶意言论，属于非法数据。

上述信息严格受到《关于严禁信息泄露、强化数据保护的郑重禁令与申明》的管控，禁止一切外泄操作

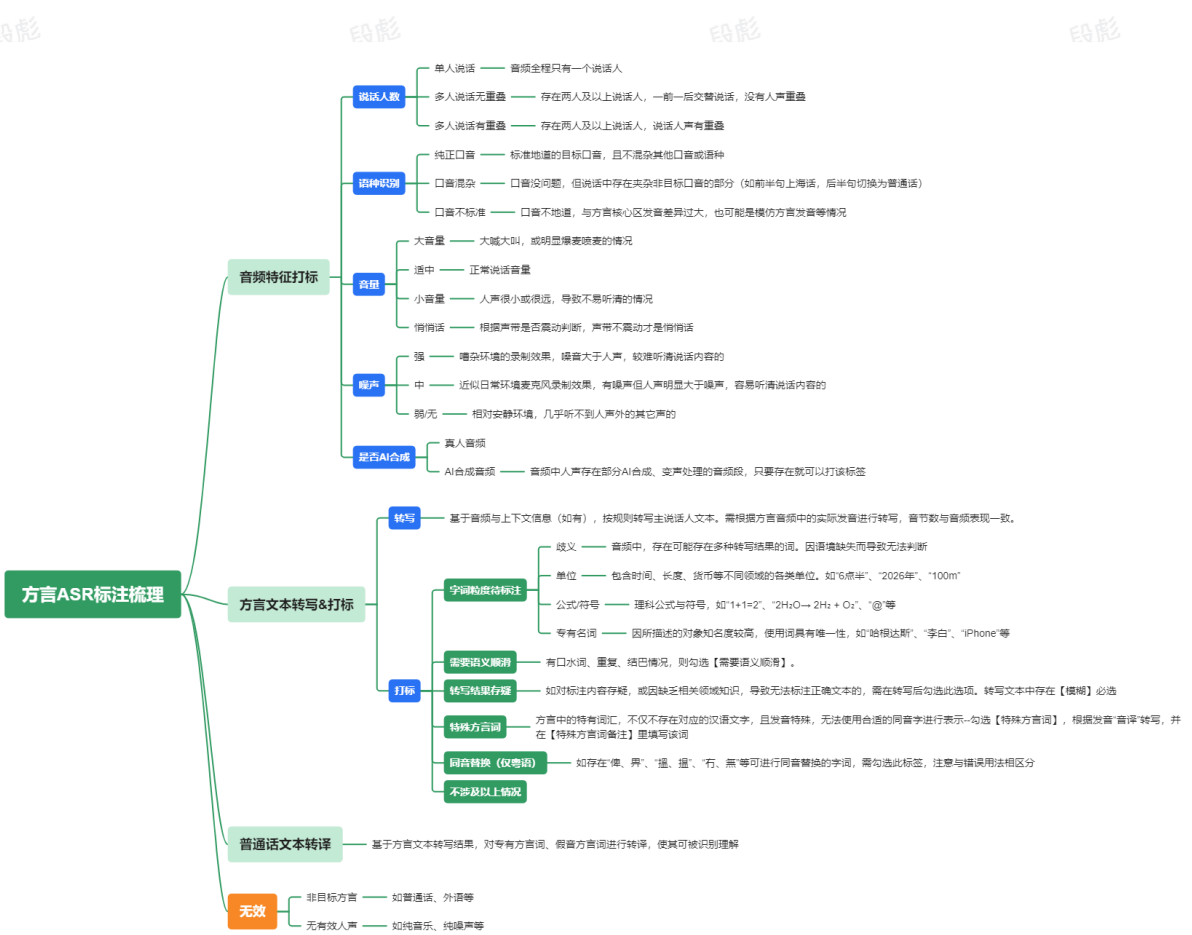
培训录频：

https://meeting.tencent.com/wework/cloud-record/share?id=ffffe216-f6ff-4273-a4ec-72b390fe5bf9&record_type=2&hide_more_btn=true&from=qywx

访问密码：7XV3

一、标注概览

 方言ASR标注梳理



二、方言用字规范


1、词汇分类方法

各《方言用字规范》中，工作表名称应当与下表中的“方言词汇类型”一致。

方言词汇类型	定义	示例
通用词汇	在普通话语境中存在的方言词汇，可被直接理解。往往是由于某种方言的影响力较大，使其用词习惯被推广到了普通话中。	如“忽悠、唠嗑、嘚瑟、晓得”等。
书面词汇	仅存在于普通话中的词汇，用方言的发音被读出，且在方言语境中的用词与普通话不同。常见于南方话中。	如用粤语读“雨伞”，但方言语境中应读为“遮”；用四川话读“蜻蜓”，但方言语境中被读为“丁丁猫儿”。
专有方言词	是对应方言中的特有词汇，但存在对应的汉字文字。与通用词汇的差异在于该词汇是否广泛存在于普通话语境中。	如粤语中的“中意”；四川话中的“嬢嬢”；上海话中的“欢喜”。

假音方言词	是对应方言中的特有词汇，但不存在对应的汉语文字，往往采用同音字进行表示。可能由于各地区发音习惯差异，导致存在多种写法。	如上海话中的“阿拉”；闽南语中的“虾米”；四川话中的“嘎嘎”
特殊方言词	方言中的特有词汇，不仅不存在对应的汉语文字，且发音特殊，无法使用合适的同音字进行表示。	/

2、方言用字规范

- 粤语  粤语方言用字规范表
- 西南官话（成渝片）
- 中原官话（郑州话）
- 东北官话

3、词表查询&新增方法

- 3.1 常用词（标红）：方言常用字+方言常用音对应规范字，要求供应商在正式作业前培训学习掌握
- 3.2 非常用词：要求供应商在作业过程中必须查询词表确定规范用字，并强调应根据词汇的“意思”而非“读音”进行检索，以提高效率和准确性。
- 3.3 无法识别/词表未覆盖的词汇：
 - 1) 方言用字规范不包含但可以转写的词：（可先暂时跳过）-->供应商通过内部机制内部统一结果-->上报标管后标管同算法共同分析确认-->由标管新增词条（**注意不能生造**）-->供应商执行
 - 2) 方言用字规范不包含且无法转写的词：根据发音“音译”转写，并在【特殊方言词备注】里填写该词

三、标注模板&流程

1、标注模版



2、标注流程

任务一：音频特征打标

标签1：说话人数

- 单人说话：音频全程只有一个说话人
- 多人说话无重叠：存在两人及以上说话人，一前一后交替说话，没有人声重叠
- 多人说话有重叠：存在两人及以上说话人，说话人声有重叠

标签2：语种识别

- 纯正口音：标准地道的目标口音，且不混杂其他口音或语种
- 口音混杂：口音没问题，但说话中存在夹杂非目标口音的部分；（如前半句上海话，后半句切换为普通话）
- 口音不标准：口音不地道，与方言核心区发音差异过大，也可能是模仿方言发音等情况

标签3：音量

- 大音量：大喊大叫，或明显爆麦喷麦的情况
- 适中：正常说话音量，按此标注
- 小音量：人声很小或很远，导致不易听清的情况
- 悄悄话：根据声带是否震动判断，声带不震动才是悄悄话

标签4：噪声

- **强**：嘈杂环境的录制效果，噪音大于人声，较难听清说话内容
- **中**：近似日常环境麦克风录制效果，有噪声但人声明显大于噪声，容易听清说话内容
- **弱/无**：相对安静环境，几乎听不到人声外的其它声音

注意：如噪声仅部分时间出现，参考噪声最大的部分进行标注

标签5：是否含有AI合成

- **真人音频**：真人讲话音频
- **AI合成音频**：音频人声存在部分AI合成、变声处理的音频段，只要存在就可以打该标签

任务二：方言文本转写及内容特征

定义：基于音频与上下文信息（如有），按规则转写主说话人文本。需根据方言音频中的实际发音进行转写，音节数与音频表现一致。

用途：用于语音模型训练，要求音字对齐。

转写规则：

- 根据“所听即所得”的原则，保留语气词、重复、口误特征，不进行语义顺滑。
- 方言特有词，应当严格区分词汇类型，参考《方言用字规范表》进行转写。
可被规范表达的词汇（通用词汇、书面词汇、专有方言词），不允许用同音字（假音方言词）替代。
假音方言词的使用，应当参考对应方言的用字规范，确保标注结果一致。
- 若存在规范未收录的方言词，则需根据语意，检索该方言的用字习惯，并沉淀至《方言用字规范表》。检索无果的，标记为特殊方言词。
- 专有名词，需根据语境或上下文确定写法。如：“腾讯的马化腾今年多大年纪？”，此处“马化腾”为知名人物，不可写为同音其他人名。“美心的月餅真是好好吃。”此处“美心”为知名企业名称，不可写为同音其他名称。
- 如存在个别无法辨认的字词，可使用【**模糊**】作为占位符表示，代替听不清的文字。

标签6：方言文本转写特征

- **字词粒度待标注**：转写文本存在歧义、单位、公式/符号、专有名词时，需勾选此标签
- **需要语义顺滑**：有口水词、重复、结巴情况，需勾选此标签
- **转写结果存疑**：如对标注内容存疑，或因缺乏相关领域知识，导致无法标注正确文本的，需在转写后勾选此标签；如转写文本中存在【模糊】则必选
- **特殊方言词**：方言中的特有词汇，不仅不存在对应的汉语文字，且发音特殊，无法使用合适的同音字进行表示--勾选【特殊方言词】，根据发音“音译”转写，并在【特殊方言词备注】里填写该词
- **同音替换（仅粤语）**：如存在“俾、昇”、“搵、搵”、“冇、無”等可进行同音替换的字词，需勾选此标签。注意与错误用法相区分
- **不涉及以上情况**

任务三：普通话文本转译

定义：基于方言文本转写结果，对专有方言词、假音方言词进行转译，使其可被识别理解。

用途：用于语义解析、NLP、通话场景，要求模型能正确理解方言语意。

转写规则：

- 确保转写后的文本可被识别理解，消除因方言俚语导致的理解障碍。
- 按照语义等价原则进行转译，一般不改变其语序和语法，除非转译后仍不可被识读。如“晓不晓得”（西南官话）应当转写为“知不知道”。
- 转译后的文本字数，应当尽量与方言文本保持一致。如“瓜娃子”（西南官话）应当转写为“傻孩子”、“笨孩子”、“笨小孩”。
- 专有名词，需根据语境或上下文确定写法。如：“腾讯的马化腾今年多大年纪？”，此处“马化腾”为知名人物，不可写为同音其他人名。“美心的月餅真是好好吃。”此处“美心”为知名企业名称，不可写为同音其他名称。
- 如存在个别无法辨认的字词，可使用【模糊】作为占位符表示，代替听不清的文字；

无效处理

- 非目标方言，如普通话、外语等；
- 无有效人声，如纯音乐、纯噪声等；

如遇到以上情况，标注平台内点无效，并勾选对应标签

标记无效

无效原因

无效标签

非目标方言 无有效人声

四、详细标注规则

1 主说话人判断规则

如存在多说话人，且人声存在重叠，则仅需转写主说话人的内容，判断规则如下：

step1：是否存在人声音重叠：

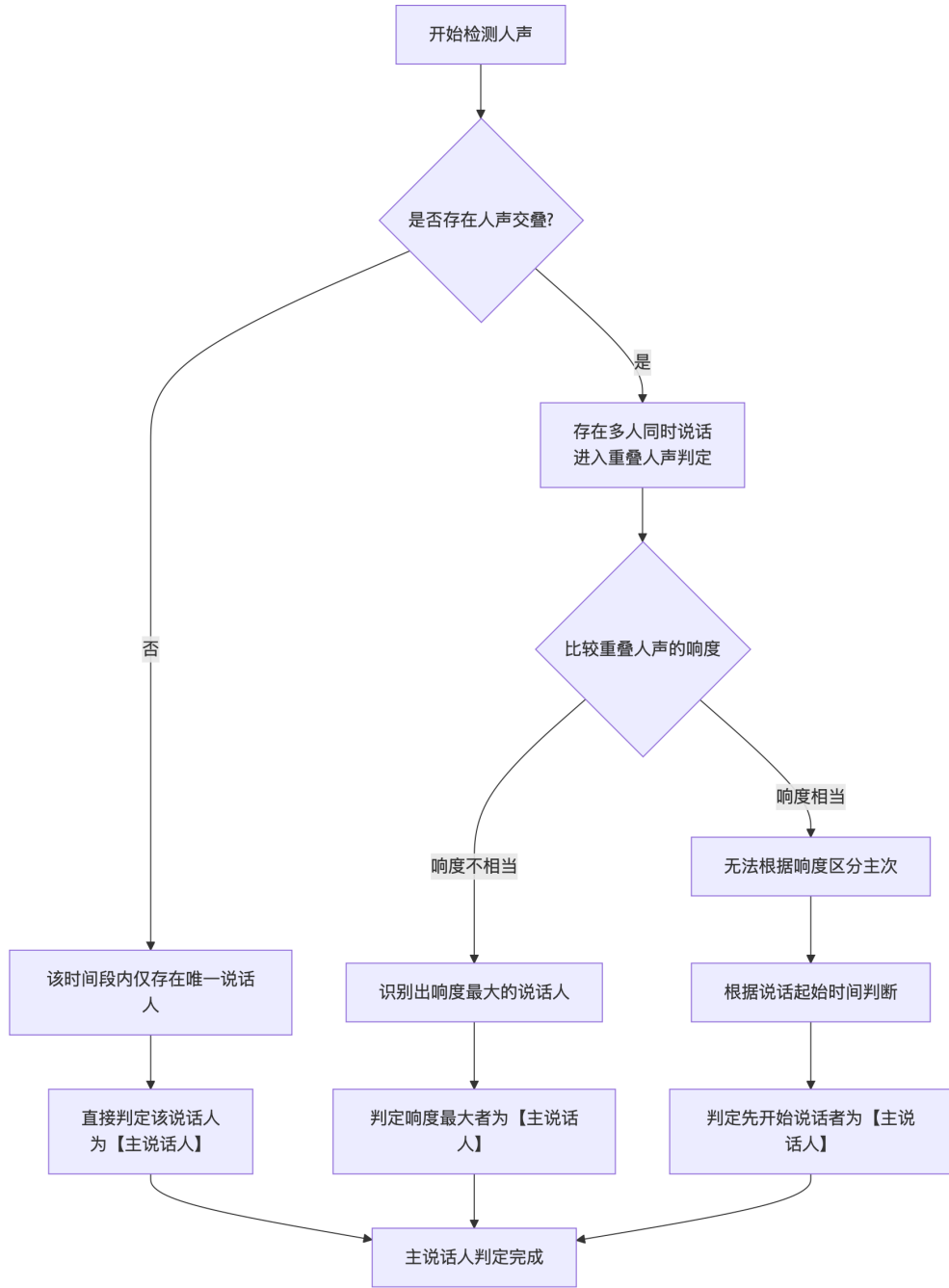
1. 无重叠：该时间段内唯一的说话人即为【主说话人】。
2. 有重叠：进入下一步判断。

step2：判断音量大小：

3. 音量不一致：音量明显更大、更清晰的说话人为【主说话人】。
4. 音量差不多：进入下一步判断。

step3：判断开口顺序：

时间轴上最先开口说话的人为【主说话人】。



2 字词粒度待标注

当文本中存在以下词汇时，需勾选【字词粒度待标注】。

字词类型	定义
歧义	音频中，存在可能存在 多种转写结果 的词。因语境缺失而导致无法判断。
单位	包含时间、长度、货币等不同领域的各类单位。如“6点半”、“2026年”、“100m”。

公式/符号	理科公式与符号，如“1+1=2”、“2H O—2H + O ”、“@”等。
专有名词	因所描述的对象知名度较高，使用词具有唯一性。 如“哈根达斯”、“李白”、“iPhone”等。

3 符号用法

3.1 特殊符号

转写类型	转写规则	示例
符号格式	1.当语音表达内容为结构化/公式化语意时， 禁止逐字汉字转写 ，应直接转写为 标准符号、标准格式 。 2.无法用文字表示的特殊符号，如【@】。	$x^2+y^2=1$ 你群里@我一下。
文字格式	若只是 零散数字/单词 嵌入自然语句，一般按普通口语转写为文字。	你连一加一等于几都不知道？ 文件名前面加个星号标记。

3.2 标点符号

转写文本中，标点符号的用法应当符合国标GB/T 15834-2011（[标点符号用法](#)）。常见规则如下：

标点符号用法：

<http://www.moe.gov.cn/ewebeditor/uploadfile/2015/01/13/20150113091548267.pdf>

- 中文、方言、日语、韩语字符统一采用全角符号，其他语种统一采用半角符号。
- 省略号，写作：“……”，不可以用英文逗号“.”代替。
- 符号使用应当与音频表现完全相符。如：
 - 疑问语气句尾使用问号【？】；
 - 感叹语气句尾使用感叹号【！】；
 - 语义未结束的停顿处使用逗号【，】；
 - 语义结束的停顿处使用句号【。】；
 - 转述与发言使用冒号与双引号结合【XX：“xxx。”】。
- 影视、文学作品、引述文章、有声书等作品名称，需要添加书名号。如《红楼梦》、《流浪地球》等。
- 整段音频的句尾必须使用【。】、【？】、【！】。

4 语气词用法

严格按照实际音频表现与语义使用语气词，不允许出现语气词混用的情况。若语气词存在不同

情绪，则应当参考“标点符号用法”，正确搭配标点符号。

常见语气词与释义如下：

语气词	易错字	释义	例句
了	啦	1. 动态助词 ：表动作完成。 2. 语气词 ：表事态变化或新情况。	下雨了。
啦	子	表示 完成兼感叹、告知、提醒 。语气较强。	饭好啦！
的	得、地	语气词 ：表肯定、确认，强调事实。	我会去的。
地	的、得	用于状语之后，连接修饰语与中心语。	天渐渐地亮了。
得	的、地	结构助词 ：用于补语前，不单独作语气词。	他跑得快。
吗	嘛	疑问语气词 ：用于疑问句结尾。	你知道吗？
嘛	吗	陈述语气词 ：表道理明显、建议、期望。	别这样嘛。
呢	呐	1. 疑问语气 ：用于特指问、选择问。 2. 陈述语气 ：表持续状态或指明事实。	我钥匙呢？ 我正忙呢。
呐	呢	感叹语气 ：用于加重语气，抒发感情。	加油干呐！
吧	把、罢	揣测、商量、建议语气 ，略带不确定。	我们走吧。 这样行吧？
啊	阿	无实义 ，用于表示 迟疑、同意、疑问 等情绪。音调不同，语气不同。	啊……这样啊。 啊？这样啊。 啊！这样啊。 啊，这样啊……
诶	欸、哎	无实义 ，用于表示 招呼、应答、诧异 。音调不同，语气不同。	诶，你好！ 诶？怎么会这样？
唉	哎	叹息、伤感、无奈或应答 。	唉，真可惜。
哎	唉、诶	惊讶、提醒、不满或打招呼 。发音短促，语气较直接。	哎，你小心点！
嘞	勒	方言常用词，无实义。	我在吃饭嘞。
哦	噢、喔	理解、醒悟、赞叹、答应 。音调不同，语气不同。	哦！我懂了。

5 字母转写规则

字母与文字之间，需要添加空格。如“我喜欢 Apple 手机”。（字母汉字混合词除外）

字母转写规则如下：

转写类型	转写规则	示例
专有名词	如商品、品牌、专业词汇、单位、化学成分等。保留其官方或通用的大小写及格式，不翻译、不添加额外字符。	iPhone、ChatGPT、QQ、Windows 11、C++、NaCl、
缩略词	保留其官方或通用的大小写及格式，不翻译、不添加额外字符。	USB、SUV、GDP、AI、APP
字母汉字混合词	字母部分保留其常见写法（通常大写），与汉字直接连接， 不加空格或连接符 。	T恤、B超、X光、AA制、维生素C

6 数字转写规则






核心原则：**转写为最符合中文阅读习惯和口语表达的阿拉伯数字或汉字大写形式。**

转写类型	转写规则	示例
数字与单位	4位数及以下可直接转写为数字，4位数及以上需转写为数字+单位。当单位为字母时，不允许使用大写数字。	300米，3千米，三千米，3km、20万、二十万。
序数词	一般使用汉字大写数字表示。	第一名、第三。
小数、分数、百分数	小数 ：只允许使用阿拉伯数字。 分数 ：非公式场景下，仅允许使用大写数字。 百分数 ：大写数字或阿拉伯数字均可。	3.14、三分之二、50%，百分之五十。
比例	使用冒号“:”+阿拉伯数字。	16:9、1:10000
日期	一般用阿拉伯数字+年月日表示，农历日期或节日名用大写数字。	2024年2月28日、2024年2月、正月十五、五一劳动节
时间	口语表达用汉字，时分秒用标准格式。	两点半，2:30。
货币金额	除口语表达外，一般使用数字后直接连接货币单位的格式。口语“块”可转写为“元”。	100元、\$199.99、五毛钱、3万美元、三万美元
模糊概述	必须使用汉字大写数字。	七八个、五六个、二十来个、三四



十岁




五、转写样例

1、粤语标注样例






方言音频	方言文本	普通话文本
 粤语1.wav	係喇，噉所以如果能夠同屋企人一齊， 誼誼誼一齊去食餐飯，跟著呢一齊去講 啲恭賀說話，誼我都幾鐘意嘅一個嘅傳 統節日嚟嘅。	是的，那所以如果能够同家里人一起， 诶诶诶一起去吃餐饭，然后呢一起去讲 点吉祥的话，诶我都很喜欢的一个的传 统节日来的。
 粤语2.wav	噉但係呢，如果有需要嘅時候呢，我都 可以煮到一餐飯出嚟嘅，譬如話我可以 煮到嘅，小菜譬如話係啊菜心炒牛肉？	那但是呢，如果有需要的时候呢，我都 可以煮到一餐饭出来的，比如说我可以 煮到的，小菜比如说是啊，菜心炒牛 肉？
 粤语3.wav	其實亦都唔一樣，所以無論你係窮光蛋 定係，唔係窮光蛋，誼其實無論做嘅事 情都唔需要介意別人點樣去諗。	其实也都不一样，所以无论你是穷光蛋 还是，不是穷光蛋，诶其实无论做的事 情都不需要介意别人怎样去想。
 粤语4.wav	噉誼噉，澳洲呢畀人個感覺亦都係好舒 服啦，但誼同大自然亦都係誼扯上一個 好緊密嘅一個關係啦。	那.....诶噉，澳洲呢给人的感觉也都是好 舒服啦，它诶和大自然也都是诶车上 一个好紧密的一个联系啦。
 粤语5.wav	跟著呢就嚟入面一路去累積經驗到到而 家，所以呢，喺呢一個人力資源部入面 嘅一個行業入面呢，其實都已經累積咗 一定嘅經驗。	然后呢，就在里面一直去累积经验到到 现在，所以呢，在那一个人力资源部里 面的一个行业里面呢，其实都已经积累 了一定的经验。

2、东北官话标注样例

方言音频	方言文本	普通话文本
 东北话1...	点上一锅那排骨炖豆角儿哇，是吧？再 贴上那一圈儿烋的溜软的玉米饼子啥 的，妈我现在跟你说，我都觉会流口水 了，那香味儿都能顶嗓眼儿，一直暖到 心窝子里头。	点上一锅那排骨炖豆角儿哇，是吧？再 贴上那一圈儿煮的溜软的玉米饼子啥 的，妈我现在跟你说，我都觉得流口水 了，那香味儿都能顶嗓眼儿，一直暖到 心窝子里头。
 东北话2...	头顶零下二三十度的大冷天儿，身子泡 的热气腾腾的温泉里头，哎妈呀那反差 儿啊，想想都老得劲儿了。我现在都觉 得我身上都热乎儿了。	头顶零下二三十度的大冷天儿，身子泡 的热气腾腾的温泉里头，哎妈呀那反差 儿啊，想想都老得劲儿了。我现在都觉 得我身上都热乎儿了。






 东北话3...	而且呢我跟你讲，草原有啥呢？你到了天气最热的时候儿，你往那树荫儿下一躲，那小凉风儿一吹，哎妈呀别提多舒服儿了。	而且呢我跟你讲，草原有啥呢？你到了天气最热的时候儿，你往那树荫儿下一躲，那小凉风儿一吹，哎妈呀别提多舒服儿了。
 东北话4...	再吹上那草原的风，哎呀那凉爽的小风儿啊，让你觉得微风拂面也不过就是这样儿了，所以说呀，夏天也得来东北瞅瞅。	再吹上那草原的风，哎呀那凉爽的小风儿啊，让你觉得微风拂面也不过就是这样儿了，所以说呀，夏天也得来东北看看。
 东北话5...	大姑娘、小媳妇儿的，还有那些个大叔、大婶子啥的，哎呀穿得那花花绿绿的那大衣裳，手里，拿着那绸子呀、扇子呀啥的还得装点上，哎那头上，带着那个blingbling。	大姑娘、小媳妇儿的，还有那些个大叔、大婶子什么的，哎呀穿得那花花绿绿的那大衣裳，手里，拿着那绸子呀、扇子呀啥的还得装点上，哎那头上，带着那个blingbling。

3、西南官话标注样例（成渝片）

方言音频	方言文本	普通话文本
 四川话1...	然后在北京吃过哪些火锅啊？咱们把这个北京吃过嘞火锅改成在四川吃过哪些火锅嘛，最喜欢哪家嘞火锅？	然后在北京吃过哪些火锅啊？咱们把这个北京吃过的火锅改成在四川吃过哪些火锅嘛，最喜欢哪家的火锅？
 四川话2...	差不多这儿今天再录到个录到个五点钟吧，差不多再录到个五点钟今天，然后我们嗯进行下一个话题，我看看看到哪儿了哈。	差不多这儿今天再录到个录到个五点钟吧，差不多再录到个五点钟今天，然后我们嗯进行下一个话题，我看看看到哪儿了哈。
 四川话3...	但是我不就会是一直就逮到这家吃，我会再去搜刮一哈有啥子没有吃过嘞，没有吃过嘞餐厅或者没有吃过嘞菜系。	但是我不就会是一直就逮着这家吃，我会再去找一下有什么没有吃过的，没有吃过的餐厅或者没有吃过的菜系。
 四川话4...	嗯我自己做的时候我一般是买速溶自己冲泡，但是说那种东西自己冲多了也不好，毕竟是速溶的嘛，肯定没得那门健康。	嗯我自己做的时候我一般是买速溶自己冲泡，但是说那种东西自己冲多了也不好，毕竟是速溶的嘛，肯定没有那么健康。
 四川话5...	所以说你要说我希望改变啥子，我其实第一个想到嘞很突出的一个观点就是我想去长高，就是这样子长高。	所以说你要说我希望改变什么，我其实第一个想到的很突出的一个观点就是我想去长高，就是这样子长高。

4、中原官话标注样例（郑州片）

方言音频	方言文本	普通话文本
------	------	-------

<p> 河南话1...</p>	<p>恁可以对俺说：喵喵喵，体检，吃饭，洗澡，去睡觉；直接对我说：退出，可以结束游戏。</p>	<p>你可以对我说：喵喵喵，体检，吃饭，洗澡，去睡觉；直接对我说：退出，可以结束游戏。</p>
<p> 河南话2...</p>	<p>客人点了一份西红柿炒鸡蛋。赶快完成恁嘞魔鬼烹饪吧！请对俺说点火儿。说点火儿开始炒菜。</p>	<p>客人点了一份西红柿炒鸡蛋。赶快完成你的魔鬼烹饪吧！请对我说点火儿。说点火儿开始炒菜。</p>
<p> 河南话3...</p>	<p>幼猫六十日龄以上才能进行疫苗注射。除此之外，无论幼猫成猫，都要保证猫咪身体健康嘞前提下再打疫苗。</p>	<p>幼猫六十日龄以上才能进行疫苗注射。除此之外，无论幼猫成猫，都要保证猫咪身体健康的前提下再打疫苗。</p>
<p> 河南话4...</p>	<p>猫咪现在比较困，活力指数为六十点儿零零。让猫咪去睡觉。现在对俺说去睡觉，可以让猫咪去休息。</p>	<p>猫咪现在比较困，活力指数为六十点儿零零。让猫咪去睡觉。现在对我说去睡觉，可以让猫咪去休息。</p>
<p> 河南话5...</p>	<p>惨绿少年嘞释义是惨绿：浅绿，指服色。原指穿浅绿衣服嘞少年。后指讲究服饰嘞青年男子。</p>	<p>惨绿少年的释义是惨绿：浅绿，指服色。原指穿浅绿衣服的少年。后指讲究服饰的青年男子。</p>