DMC（DIG Mandarin Conversations）

会话分析汉语普通话语料库章程

* 使用本语料库的语料，请务必引用：
Guodong Yu, Yaxin Wu & Chase Wesley Raymond. forthcoming. The DIG Mandarin Conversations (DMC) Corpus: Mundane Telephone Calls in Mandarin Chinese as Resources for Research and Teaching. *Chinese Language & Discourse*.

* 欢迎为本语料库提供语料，与同行共享。共享语料请发至yuguodong@ouc.edu.cn

DMC（DIG Mandarin Conversations）会话分析汉语普通话语料库，是于国栋教授与吴亚欣教授带领的会话分析团队建立的，供国内外会话分析学者交流、共享的会话分析汉语普通话语料库平台。本章程旨在为语料库中语料收集、转写、提交等事宜提供统一的规范，并保障语料提供者与交际参与者的知情同意权。语料转写是会话分析学科中重要的一环，是对每一位会话分析学者的基本要求。语料转写中使用的每一个符号及其位置，包括空格或声音的标注，都需要严谨规范。本章程适用于参与到本语料库建设的语料提供者。本章程包含以下几个方面：**1.**语料收集；**2.**语料转写；**3.DMC** 转写符号；**4.DMC** 转写符号说明；**5.**语料互检、终检与提交；语料共享知情同意书。

1. 语料收集

语料提供者在收集语料时需要注意以下两个方面：

1) 语料提供者在录音、录像前，需征得交际参与者的知情同意，并签署语料共享知情同意书，可使用电子签名。（请确保交际参与者明确知晓语料共享知情同意书的各项条款）。

2) 语料提供者要确保所录音频、视频的质量，确保言谈应对清晰可辨，较少杂音，画面清晰等。

2. 语料转写

语料提供者在转写语料时需要注意以下四个方面：

1) 语料转写需严格采纳 **DMC** 转写符号中规定的符号，并严格按照 **DMC** 转写符号说明中描述的情况使用转写符号。

2) 语料转写中的时间可以使用打节拍的方式计算，例如，使用 **MI SSI SSI PPI ONE**（Mississippi one）来计算时间，每个音节占 0.2 秒；也可使用电脑软件来计算时间。常

用的处理音频、视频语料的软件有 ***Audacity***（仅限于处理音频语料，可免费下载，操作简单便捷，完全满足音频语料的转写需求，但该软件无法处理视频语料。）；***Adobe Premiere Pro***（音频、视频语料都可处理，操作相对复杂，需要购买使用。）等其他软件。需要注意的是转写时间的计算，使用电脑软件计算得出的时间需要减去 **0.2** 秒，以便更加客观体现停顿、沉默等现象。

3) 使用以下网站（无需注册或购买），或其他可以转换汉语拼音的网站或软件，为语料添加带声调的拼音话轮：https://www.lddgo.net/convert/pinyin; 遇到多音字或拼音不确定的地方（例如："喂"的拼音是"wèi"，表示再见的"拜拜"的拼音是"báibái"等）可查询现代汉语词典或百度汉语：https://hanyu.baidu.com。

3. DMC 转写符号

DMC 转写符号源自 Gail Jefferson（2004）转写体系，并根据汉语语料转写的实际需求进行了相应增加与修改，DMC 转写符号均采用英文半角符号，且字体为 **Times New Roman**。DMC 转写符号包括以下三个部分：

1) 与时间、序列相关的符号

(0.6)	表示时长为 0.2 秒及 0.2 秒以上的沉默。
(.)	表示时长在 0.1 秒至 0.2 秒之间的沉默。
=	表示紧随话语。
[ ]	表示重叠话语或同步话语。
	[表示重叠话语或同步话语的起始位置，
	]表示重叠话语或同步话语的结束位置。

2) 与话语表达相关的符号

. ? ,	表示末尾声调。点号表示降调，问号表示升调，逗号表示平调或微升调。
Cu-	表示话语突然停止。
s:::	表示拖音，每个冒号表示 0.1 秒至 0.2 秒。
> <	表示语速加快。
< >	表示语速减慢。
º º	表示音量低。
underline	表示强调、重读或音量高。
↑	表示语调突然升高。
↓	表示语调突然降低。
.hhh	表示吸气音，每个"h"表示 0.1 秒至 0.2 秒。

hhh　　表示呼气音，每个“h”表示 0.1 秒至 0.2 秒。

hehehe　　表示笑声，每个“he”表示 0.1 秒至 0.2 秒。

¥　¥　　表示笑着说的话语。

heeheehee　　表示哭声，每个“hee”表示 0.1 秒至 0.2 秒。

%　%　　表示哭着说的话语。

3)　其他符号

荣:　　表示交际参与者，使用一个汉字指代。

(word)　　表示录音中可能发生的话语，

(　)　　表示录音中无法确认的话语，或无法确认的交际参与者。

((description))　表示对某个现象的描述，而非转写本身。
也可用在面对面交谈的语料中，用来描述某个动作或表情等。

pinyin　　不加声调的小写汉语拼音，用于不完整的发音、无法确认的话语或声音、以及无法找到对应汉字的方言中。

4.　**DMC** 转写符号说明

为更好地说明 DMC 转写符号的各种使用情景与方式，以下语料已进行改编。小号字体的双括号内容是转写符号的使用说明，按照 **DMC** 转写符号中符号出现的顺序进行解释说明。

[OUC-DMC-HQX_山间小路_0000-0034]

((标题采用[OUC-DMC-HQX_山间小路_0000-0034]的加粗形式。标题起始部分是 OUC-DMC 与每位语料提供者名字缩写的结合，名字缩写也可以是大家愿意采用的名字代号，但需使用三个大写字母来表示；标题中间部分使用三到五个汉字，表示语料主题；标题结尾部分为语料截取时间。))

((汉字及中文全角符号（占两个字节）使用华文仿宋字体，英文、拼音及英文半角符号（占一个字节）使用 Times New Roman字体，字号选用五号，行距为 1.0。行号使用：01，02，03 模式，汉字话轮与拼音话轮开端对齐（不需要每个汉字与拼音一一对齐），使用同一个行号，若汉字话轮有多个重叠部分，则对应的拼音话轮中每个重叠部分的开端都需与汉字话轮相应部分对齐，若汉字话轮与拼音话轮内容相同，则省略拼音话轮。除语料背景及汉字话轮中的双括号说明，使用中文全角符号，且符号字体为华文仿宋外，其他任何地方都使用英文半角符号，且符号字体为 **Times New Roman**。推荐一个与字体及半角、全角相关的操作方法：在实际转写中，我们可以暂时不必管字体问题，汉字话轮输入中文时使用全角，添加符号时 shift 切换到英文半角（语料背景及汉字话轮中的双括号说明除外，不需要 shift 切换)。拼音话轮则直接从转换拼音的网页粘贴，修改空格及多音字检查等即可。最后使用全选（全文选择)，先选择华文仿宋字体，之后再全选，选择 Times New Roman 字体。这样语料中，全部汉字、语料背景及汉字话轮中的双括号说明的符号是华文仿宋字体，其他的英文、拼音及符号，包括行号等全部是 Times New Roman字体。P.S. 采用这样的方式，是为了实际操作中便于后期统一格式。例如，转写好语料之后，修改某些转写符号时，可以不用去管，符号是什么字体状态下输入的，修改符号完成之后，按照上面的步骤，全选，再次统一字体即可。))

((每份语料需要添加页码，使用处于页面底端的加粗显示的 X / Y 形式，参见本章程页码形式。))

语料背景：儿子威（**Wēi**）给母亲荣（**Róng**）打电话，询问路程情况，其中飞（**Fēi**）和伟（**Wěi**）分别是威的妹妹、妹夫。

((语料背景为加粗形式，其中涉及的交际参与者指称需要添加首字母大写的拼音。语料背景内容若超过一行，第二行及之后的行（若有）顶格开始，不使用悬挂、缩进等。))

01 (0.6)

(((0.6)表示时长为 0.2 秒及 0.2 秒以上的沉默。若时长恰好为整数，则添加小数点后的零，例如两秒的表示方式为：(2.0)。))

02 荣：喂?

wèi?

03 (.)

(((.)表示时长在 0.1 秒至 0.2 秒之间的沉默。))

04 威：>诶.< 妈?=

>éi.< mā?=

05 荣：=诶.

=éi.

((=表示紧随话语。在实际转写过程中，等号还可以用来衔接三人及三人以上的紧随话语，以及同一个交际参与者的多个话轮。例如，在同一个交际参与者占据多个话轮时，为了保持每一行长度的适中性（汉字话轮通常最多包含十到十五个汉字），可选择合适位置转行，若恰好一个 TCU 分别出现在上下两行，那么这个 TCU 的两个部分的后与前分别用等号衔接。此外，等号不可以用在同一交际参与者的同一个话轮中，若要表示同一话轮中不同 TCU 之间的紧随，则两者之间不加点号逗号问号等，并取消空格直接衔接。例如，示例语料可改为：>诶<妈?=，相应的拼音话轮取消末尾声调后，不取消空格：>éi< mā?=。此外等号与其他符号之间无空格。拼音话轮同样适用。))

05 (0.6)

07 威：[你们到了没有?]

[nǐmen dào le méi yǒu?]

08 荣：[怎么啦，儿子.]

[zěn me la, érzi.]

(([ ]表示重叠话语或同步话语。在实际转写过程中，可能是部分重叠或同步。此外若同一话轮中存在多处重叠部分，则每个重叠部分都需要与其相应的前后话轮内的重叠部分对齐。拼音话轮同样适用))

09 (0.7)

10 荣：>刚:到:.<

>gāng: dào:.<

((.表示末尾降调。在实际转写过程中，如果遇到[ ]; > <; < >; ° °; ￥ ￥; % %; ↑↓; ( )等符号组合，应根据具体情况，将点号、问号、逗号等放置在以上符号组合的内部或外部。若符号组合内的内容与前面相衔接，则点号、问号、逗号应放置于符号组合的外部。例如，示例语料若改为：我们>刚:到:<.，那么点号则放置于符号组合的外部，此时点号表示整个句子的声调特点；否则则为示例语料中的情况（符号组合内的内容与前后内容不衔接），放置于符号组合的内部。拼音话轮同样适用。))

11 (.)

12 威: 噢:. 我看见飞飞发的那个-

ō:. wǒ kànjiàn fēifēi fā de nàgè-

((Cu-表示话语突然停止。在实际转写过程中，若存在无法确定的话语内容，但该话语内容在末尾时的突然停止却可以判断，那么该现象要标注为( )-。拼音话轮同样适用。))

13 (0.3) ￥那个:朋友圈,￥

(0.3) ￥nàgè: péngyǒuquān,￥

((s:::表示拖音，每个冒号约为 0.1 秒。TCU 内部的冒号后不需要空格。拼音话轮不适用。))

14 >￥看见你们这-￥<

>￥kànjiàn nǐmen zhè-￥<

((> <表示语速加快。在实际转写过程中，若遇到以上示例语料中出现多个转写符号组合的情况，例如: [ ]; > <; < >; ° °; ￥ ￥; % %; ↑↓; ( )等，那么> <; < >包含° °; ￥ ￥; % %; 。￥ ￥; % %包含° °。转写过程中要避免符号组合嵌套的情况，例如，可将示例语料改为: >￥看见你们￥< ￥这-￥，而不可以采用: >￥看见你们< 这-￥。这样处理是为了每一个 TCU、semi-TCU、abandoned TCU 等内容的语言表达特点都可以独立且清晰地得到传达。此外，若同一交际参与者占据多个话轮，且多个话轮都有加速、减慢、降低音量、笑着说等情况，则需要每个话轮都使用完整的符号组合，而不可以在第一个话轮前使用一个符号，而多个话轮后使用另一个符号。< >用法相同，不再赘述。拼音话轮同样适用。))

15 (0.4) 呃:没找 °见路:°?

(0.4) è: méi zhǎo °jiàn lù:°?

((° °表示音量低; underline 表示强调、重读或音量高。在实际转写过程中，使用网页或软件转换拼音内容时，下划线会在拼音内容中消失，其他符号及内容不会（但需要注意是否需要取消或增加空格的问题），因此需要手动添加下划线。在拼音话轮添加下划线时，除了名词、代词、副词、动词短语等的拼音需衔接在一起外，其他汉字的拼音是分开的，因此添加下划线时，不要将空格或除冒号外的其他符号下也添加下划线。))

15 (0.5)

17 荣: ↑噢:↓

((↑表示语调突然升高; ↓表示语调突然降低。在实际转写过程中，末尾的降调用箭头表示后，若后面声音已停止，则不需要再加点号。例如以上转写实例。若后面声音仍未停止，则需要再加相应的符号，例如，↑噢:↓::,。若箭头出现在话语前面，且话语前面紧

随其他话语，则需要在箭头前空格，以独立展现这一语言现象，例如：嗯 ↑噢:↓。拼音话轮同样适用。))

18 荣: .hh hh hehehehe=

((.hhh 表示吸气音，每个“h”表示 0.1 秒；hhh 表示呼气音，每个“h”表示 0.1 秒；hehehe 表示笑声，每个“he”表示 0.1 秒；heeheehee 表示哭声，每个“hee”表示 0.1 秒。相同类型之间不空格，不同类型之间需要空格，见示例语料。此外笑声、哭声也可使用相应汉字来表示。例如，嘿嘿、呵呵、哈哈、嘻嘻、嘤嘤、哇哇等。若无法使用对应汉字表示笑声、哭声，则必须使用 he 或 hee 来表示。此外需要注意的是，若以上符号结合><、<>时，那么相应语言现象持续时间也会发生变化，转写符号可以灵活应用，但仍要注重使用规范。))

19 威: =he ¥啊?¥ (0.4) he

=he ¥a?¥ (0.4) he

((¥ ¥表示笑着说的话语；% %表示哭着说的话语。))

20 (0.2)

21 荣: 噢.

ō.

((荣:表示交际参与者，使用一个汉字指代。交际者代称加冒号之后，空一个 TAB 的距离，之后添加话轮。若同一交际参与者有多个话轮，那么之后的话轮显示行号后需要空一格，才能使用 TAB，之后使用 TAB 纵向对齐话轮。若某一话轮与其前、后话轮存在重叠话语或同步话语现象，则尽量使用 TAB 之后再使用空格键对齐重叠的开端部分。))

22 跟的那(大车)走的,

gēn de nà (dàchē) zǒu de,

(((word)表示录音中可能发生的话语，此处填写录音中可可能出现的话语，无需提供多个可能话语。同时要注意，因单括号内为可能话语，若其内容与前或后的内容为同一 TCU，那么衔接部分不需要空格，但拼音话轮依然按照拼音特点分开标示，除名词、代词、副词、动词词组等。))

23 ( ): 那么多( )大[车.

nàme duō ( ) dà [chē.

(( )表示录音中无法确认的话语，或无法确认的交际参与者。中间空格根据时长，一个空格表示一个字符。需要注意的是，单括号内是不确定的内容，单括号外是确定的内容。例如，转写时可能无法识别单括号内的具体内容，但可识别出是笑着说的，且末尾是突然停止的，那么可以标示为：¥( )-¥。))

24 飞: [头几次没走[对. ((边吃东西边说话。))

[tóu jǐcì méi zǒu [duì. ((Talking while eating.))

((((description))表示对某个现象的描述，而非转写本身。需要注意的是，双括号前必须有解释说明的对象，不可使用双括号代替话轮，若双括号前无话语，则单括号展示时间后再加双括号。例如，汉字话轮为：(1.2) ((背景噪音))，相应的拼音话轮为：(1.2) ((Background noise))。双括号内若为词汇则无需添加标点，若为句子，则汉字内容使用中文全角标点，英文翻译则添加英文半角标点，

参见示例语料。此外英文翻译尽量简洁明了，使用名词、动名词来描述。若使用句子描述，或其他情况导致话轮过长，则可以将双阔号中内容移到下一行。同时语料中不使用脚注，统一使用双括号来解释说明。))

25 伟: [( )-

25 荣: 后来又走的山间小路过来 di.

hòulái yòu zǒu de shānjiān xiǎolù guò lái di.

((pinyin 不加声调的小写汉语拼音，用于不完整的发音、无法确认的话语或声音、以及无法找到对应汉字的方言中。此处的拼音不加声调，用以区别拼音话轮中的其他拼音。))

27 (0.4)

28 威: 是:? 那你们玩哇.

shì:? nà nǐmen wán wa.

((汉字话轮通常最多包含十到十五个汉字。))

((汉字话轮与拼音话轮的区别在于，汉字话轮中的 TCU、semi-TCU、abandoned TCU 等内容的字与字之间是衔接的、无空格的，拼音话轮则需要拼音与拼音分开（除名词、代词、副词、动词词组等。），但 TCU、semi-TCU、abandoned TCU 等内容之间确是各自独立，需要空格，不能无间隔，以便清晰表达每一组语言内容的特点。))

29 吃点东西再玩儿吧. 哈.

chī diǎn dōngxī zài wánr ba. hā.

((儿化音的处理，是在出现儿化音的词的拼音后直接加 r。))

30 (0.6)

31 荣: .tch

((若汉字话轮中并无汉字出现，则无需添加拼音话轮。))

32 (0.2)

33 威: 噢. 那就这哇.

ō. nà jiù zhè wa.

34 (0.4)

35 荣: 噢. 行哓.

ō. xíng lǎo.

35 威: 我就问问. °哈.° BYEBYE.

wǒ jiù wènwèn. °hā.° BYEBYE.

((汉字话轮中出现英文单词时，或使用英文字母代替的人名、地名等时，一律使用全部大写来标示。这一点是为了与不加声调的小写汉语拼音相区别。))

37 荣：°嗯.° (0.2) °嗯.°

°ṅ.° (0.2) °ṅ.°

5. 语料互检、终检与提交

1) 语料提供者在完成语料转写后，需要进入分组互检环节。互检的重点在于语料是否客观、准确将音频、视频的内容进行了转写。

2) 完成互检之后，每组组长进行终检。终检是语料提交的最后阶段，各组组长以抽查的方式进行，若抽查过程中发现任何转写准确度的问题，则语料提供者需重新进行自检，之后再次提交。

3) 每人提供的语料，分别放入单独的文件夹中，每个文件夹包括：语料音频或视频、语料 **WORD** 文档、语料 **PDF** 文档、及语料共享知情同意书。除语料共享知情同意书命名为，例如，语料提供者的知情同意书标题为：00_CONSENT-[OUC-DMC-HQX_山间小路_0000-0034]（其他交际参与者序号依次为 01，02，03 等），其余三个文件以语料标题命名，例如：[OUC-DMC-HQX_山间小路_0000-0034]。最后，所有语料的相关文档放入一个文件夹中，压缩后发送邮箱。注意音频格式为 MP3、视频格式为 MP4（推荐使用格式工厂转换音频或视频格式，可免费下载，操作简单便捷）。知情同意书电子版提供 PDF 版本。

中国海洋大学会话分析研究团队

2022 年 **11** 月 **16** 日

语料共享知情同意书

研究负责人：于国栋教授、吴亚欣教授

单位：中国海洋大学

联系方式：**13834561566** （于国栋教授）

语料用途：该语料将用于有关互动交际的语言学基础研究，涉及互动中交际者所使用的语言与非语言资源，以及整体的会话结构等。接受本协议则表明您同意 DMC（DIG Mandarin Conversations）语料库将您提供的音频、视频等材料进行转写，并用于语言学研究、以及研究成果的发表与出版。

风 险：本协议不会对交际参与者造成任何可预见的风险，若您希望在相关材料或语料转写中删除某些信息，可在语料共享前要求检查并进行删除。

受 益：本语料将收录于 DMC 语料库中，并无偿提供给国（境）内外语言学研究者，我们不会向您提供任何物质补偿。

隐私保护：相关语料将用于建立一个开源性的音频、视频等语料档案，并将永久免费用于语言学研究及成果的发表与出版。同意本协议则表明：您同意我们在后续的研究及成果的发表与出版过程中使用相关语料，且不再需要征得您的额外许可。我们保证在语料的收录和使用过程中将隐去交际参与者的私人及敏感信息，如隐去真实姓名，对视频进行人像模糊处理等，但不排除交际参与者的身份会通过其声音被识别。

详情咨询：如果您需要对相关研究或语料库进行咨询，请联系研究负责人。

退 出：您可以在语料共享之前的任何时间申请退出本协议，且不会因此遭受任何损失。需要特别注意的是：一旦语料在互联网共享，将不再移除。

我确认已知晓该语料的用途，并已知悉上述条款。

签名：_____（交际参与者☐ / 语料提供者☐）

日期：_____