Modern Linguistics 现代语言学, 2024, 12(8), 536-542
Published Online August 2024 in Hans. https://www.hanspub.org/journal/ml
https://doi.org/10.12677/ml.2024.128721

Hans汉斯

日常话语研究的优质原料：CED语料库介绍

付晓丽

河北师范大学外国语学院，河北 石家庄

收稿日期：2024年7月10日；录用日期：2024年8月19日；发布日期：2024年8月28日

摘　要

随着汉语日常话语研究的日益深入，相关的语料库建设工作愈发重要。自然发生的、人际互动过程中的语言事实，是支撑科学研究的优质基础原材料。对此类语料进行详尽描述和深入探讨，可发现并揭示汉语日常话语的使用规律和本质特征。本文介绍已建成并投入使用的“汉语日常话语语料库”(简称CED)。文章介绍“日常话语”的工作定义，报告该语料库建设的指导原则、研究方法，数据采集标准、标注方案制定及语料构成情况，希望更多的研究者从中得到些许启发。

关键词

日常话语语料库，汉语口语，语料采集，标注方案，语料构成

High-Quality Raw Materials for Everyday Discourse Research: An Introduction to the CED Corpus

Xiaoli Fu

College of Foreign Studies, Hebei Normal University, Shijiazhuang Hebei

Received: Jul. 10th, 2024; accepted: Aug. 19th, 2024; published: Aug. 28th, 2024

Abstract

With the deepening of the study of daily Chinese discourse, the construction of related corpora is becoming more and more important. Linguistic facts occurring naturally in human interaction are the high-quality basic raw materials that underpin scientific research. A detailed description and in-depth discussion of such corpus can discover and reveal the rules and essential features of daily Chinese discourse. This article introduces the “Corpus of Everyday Chinese Discourse” (CED) that has been built and put into use. This article introduces the working definition of “everyday discourse”

文章引用: 付晓丽. 日常话语研究的优质原料: CED 语料库介绍[J]. 现代语言学, 2024, 12(8): 536-542.
DOI: 10.12677/ml.2024.128721

and reports the guiding principles, research methods, data collection standards, annotation scheme formulation and corpus composition of the corpus, hoping more researchers can get some inspiration from it.

Keywords

Corpus of Everyday Discourse, Spoken Chinese, Corpus Collection, Annotation Schemes, Corpus Composition

1. 引言

随着语料库语言学的兴起，语料库建设工作越发受到学界重视。话语研究(discourse studies)是语言学的重要分支，它尤其强调语料的真实性。由自然发生的语言事实所累积的语料库，是开展话语研究的必要资源。研制一个立意独特的语料库，是话语研究的原始创新得以实现的有利条件。但言谈话语的听觉表现形式，无法直接作为研究材料使用，必须转录成书面文字，还要对相关语言特征进行标注。此类语料库的建设过程难度较大，耗时、费力。它是一项比较艰巨的任务，非一日之功。因为有难度，所以国内以日常话语为素材的语料库目前还比较少。近年来，笔者主持建设了一个中型日常话语语料库，本文对此专门介绍。希望该项基础性工作，为汉语日常话语研究的原始创新工作添砖加瓦。

2. “日常话语语料库”简介

“日常话语语料库”(英文名称 Corpus of Everyday Discourse，缩略为 CED)是河北师范大学 2016 年立项的人文社会科学基金交叉协同科研项目。CED 建设历时四年，其中 2016 年是筹备阶段，2017~2019 年是正式实施阶段。这期间，课题组先后举办了数十场专题研讨会和培训工作坊，对语料库建设中的各种问题，进行了深入、细致的研讨。CED 以日常汉语交流为研究对象，近 70 万字。语料库所有语料，均是自然情境中发生的汉语日常话语交流。

2.1. “日常话语”概念及语料库定位

课题组把“日常话语”定义为“那些聚焦大众话题，且一般没有事先计划的、人际间即兴的语言表达”(付晓丽、荣红、董东、宋文辉，2019 [1])。以此概念冠名的语料库，在搜集语料方面，主导思想是强调语料的“自然”“真实”和“不刻意”“无准备”。

优质的语料库，在内容选择上，应具有代表性和平衡性(郑家恒等，2010 [2])。CED 语料选择符合“代表性”的要求。鉴于言语是语言表达的基础媒介(Dash & Arulmozi, 2018 [3])，CED 以言语语料库(a speech corpus)为首要建设目标，选择非正式的话语交流。因为非正式的言语交流更可靠、更生动，它们在自然性(naturalness)方面，比正式的口语表达(如演讲)更有代表性，更接近自然语言的核心特征。正如著名的“莫拉维克悖论”(Moravec’s paradox)所描绘的：越是无意识的、直觉的技能，越需要更强的算力。就话语表现而言，越是人们无意识、随便说的话，就越值得研究。这些随意的语言交流，具有其他言语变体无可比拟的优势(Dash & Arulmozi, 2018 [3])，能充分揭示人类语言奥秘。在语料库的平衡性上，课题组顶层设计也有兼顾。允许少量的书面交流的话语文本(如微信聊天)被接受入库，以反映新媒体时代人们日

常话语交流的状况。

CED 重点采集人际互动话语，话题比较广泛，涉及购物、理发等多个生活场景。话语交流类型比较丰富：既有二人对话，也有三人及以上会话；既有公开的话语交流，也有私密的话语交流；既有以任务为导向的(task-oriented)、目标明确的话语，也有目的不明确、显性任务缺席的日常闲谈。

2.2. 研究方法

CED 使用民族志研究范式，结合会话分析理论，对语料进行搜集和整理。在私人语料报送过程中，严格遵守科研伦理规范，参考国际学界(如 José & Teixeira, 2013 [4]; Keel, 2016 [5]; Tannen, Kendall & Gordon, 2007 [6])的做法，采用正式的书面授权和非正式的口头同意这两种授权方式。CED 中涉及私人话语的部分，均经过语料提供者本人同意。操作流程是先用录音、截屏把自然发生的语言使用记录下来，制成 word 文档，再进行赋码和标注。word 语料文档完成后，再经过多遍复查确定无误，最终定稿入库。以下是具体的做法。

2.3. 数据采集

这个环节包括两项任务：1) 采集话语交流信息；2) 记录并报告话语发生的语境情况。

2.3.1. 采集

CED 所搜集的语料，主要有公共话语和私人话语这两类。公共话语的语料相对便于搜集和整理。难度较大的是私人话语的搜集工作。CED 采集的语言信息包括三大类，一是言语类，即口头表达的音频文件；二是文字类，即书面表达的视觉信息；三是混合模式，既包含语音类，也包含文字类。这三种情况，课题组采取不同的方式来处理。

为 CED 直接贡献语料的人，被称作语料库文本语料会话人。CED 对语料真实性尤其重视，要求贡献者报告现实生活中真实发生的话语，不得转录影视剧节目的话语，也不得编造、虚构话语。音频语料的录音工作在 2017~2019 年间。私人话语部分，报送语料者都是土生土长的中国人。会话人被要求尽可能地说普通话，如有微量方言土语，也照常收录。需要说明的是，有一些报料人不仅复杂报料，还负责转写。这种情况下，转写者是“内部人”(insiders)，对话语发生的语境情况清楚，对当时的谈话内容了解，能高质量地完成转录文本工作。这种做法虽然涉及“观察者悖论”问题，但由于报料人提前经过培训被告知：所采集的语料就是他们的日常话语，他们平时怎么说话，被录音时就还怎么说话，越自然、越随意越好，事实上“观察者悖论”对 CED 语料采集的真实性没产生什么影响。据多位报料人反映，在录音开始时，他们还有一点紧张，但随着话语交流的展开，他们便忘了自己说话“被录音”这件事，很快进入到自然的话语交谈状态中。

语料库建设涉及诸多细节问题：其一是个人数据隐匿。个人数据是个人数据保护法中的核心概念，其界定的关键是可识别特定自然人的数据，包括直接识别与间接识别(王融，2016 [7])。数据隐私(data privacy)和去识别化(de-identification)都是语料库建设的重要环节。CED 把真实的人名及敏感信息(如住址)都进行隐藏处理；其二是数字的转录。尽量使用文字而非数字来表示。比如口头说的“1972”，语料库文档会呈现为“一九七二”，而非数字“1972”；其三是表情符的处理。先把表情符分为两类：一种是意义清楚明确的，另一种是意义模糊的。对意义明确的表情符(如高兴)如实转录，用括号文字加注的方式标注；对意义不清的表情符则忽略不计。

2.3.2. 会话日志记录

私人话语的语料采集完成后，报料人要对所采集的语料进行简要描述，撰写会话日志(conversation log)，讲明话语发生的语境情况。实验研究后，课题组发现报告行文有差异。于是课题组提供 AB 两种方

式，供报送者自由选用。

方式 A：填充信息，即以填空的形式，报告语境情况。不确定的信息，就不填。例如：

1) 话语发生时间：_____

2) 发生地点：_____

3) 参与人数：_____

4) 参与者人际关系：_____

5) 参与者年龄：_____

6) 传播方式：_____

7) 言语事件：_____

方式 B：自由撰写，即以文字形式，对语境情况进行详细说明。例如：

“2018 年 6 月 8 日，我想买一双鞋，我闺蜜陪我去逛商场。我看中了某品牌的一款样式，想着试一试鞋子。这个录音就是当时我、我闺蜜、售货员这三个人所说的话。”

2.4. 转录及标注

Table 1. CED transcription scheme
表 1. CED 转写规范

类别	符号	意义	例子
语言特征标注(注：凭转写者的直觉来判断)	,	表示一个意群的停顿，言说者还未完成完整意义的表述，一般为平调或者轻微的升调	甲：今天的主要任务，其实很简单
	。	表示言说者完成一个完整意义的表述，一般为降调	甲：今天的主要任务，其实很简单，就是复查语料。
	?	表示言说者的疑问，一般为升调	甲：你今天有课吗？ 乙：没课
	!	表示言说者的强烈情绪	甲：九点了，你还来吗？ 乙：我不去了 甲：你可真行！
	[ ] [2 2]	表示两个或两个以上的人同时说话时的重叠 [表示重叠起点] 表示重叠终点 [2 表示同一语轮中的第二次重叠的起点 2] 表示同一语轮中的第二次重叠的终点	甲：[他们学 乙：我跟你讲]他们比咱们[2 紧张 甲：是啊 2]
	： ：： ：：：	表示某个音节的延长，每增加一个冒号，就表示多延长一拍 (注：重读音节，要尽可能地按照这种方式去标注。)	你们谁去：：：那个：：邮电局不去。
	-	表示言说者所做的短暂停顿，多为自我修补	中国的儿童是比外-美国的儿童聪明呀
	^	表示打断对方，突然插话	甲：我想是 乙：^你别你想，你就说，你做了什么？ 甲：哦，我直接去了她家
	= =	表示一方刚说完，另一方就紧接着说	甲：我就没想= 乙：=你没想，你干嘛不想！
	(……)	表示无法辨识的话语	都是母亲在(……)

续表

	(文本)	表示貌似某种话语表达，即听起来像是这样的话	那么在美国(的话)那些孩子吧：：：是华裔美国华裔美国人
	〈字〉	表示勘误，以防误解	她带〈待〉几天
	(E 文本)	表示各种表情符号	收到(E 微笑)
	(EP 文本)	表情包	收到了，(EP 谢谢)
	(短默)	表示较长时间的沉默	甲：你觉得怎么样？ 乙：(短默)也就那样
	(长默)	表示较短时间的沉默	甲：你觉得他好吗？ 乙：(长默)怎么说呢？
	文本〈转(具体语种)〉	表示语码转换为其他语言 (注：仅在词语后面标注)	1) 这款式，你 hold 〈转英〉 住吗？ 2) 饿了，咱们去米西米西 〈转日〉 3) 你问 moi 〈转法〉 啊？
	〈方省/市〉 文本〈方省/市〉	表示普通话里面出现方言 (注：允许笼统判断区域)	我就想知道， 〈方四川〉 吃啥子不得长肥 〈方四川〉 ？
	@@	表示笑	你来啦@@
	@文本@	表示话中带笑，即笑着说	我@一直这个样啊@
其他行为/事件标注 〈 〉	〈鼓掌〉	表示鼓掌	现在欢迎王主任讲话〈鼓掌〉
	〈发网址〉	表示会话人发了网址	甲：你把网址给我吧。 乙：〈发网址〉 发过去了。 甲：看到了，谢谢。
	〈发图片〉	表示会话人发了图片	甲：给你看看我家新开的海棠花。 乙：好啊。 甲：〈发图片〉 乙：好漂亮！
	〈发语音〉	表示会话人发了语音	甲：给你看看我家新开的海棠花。 乙：好啊。 甲：〈发语音〉你看看！ 乙：好漂亮！
	〈其他具体行为/情况〉	表示会话人的其它行为或其它语境线索	甲：我给你留个电话号码 乙：嗯，好 〈甲在纸上写电话号码〉 甲：给 乙：太谢谢啦 甲：就这样吧 乙：好的 〈大约三分钟后，二人又开始交谈〉 甲：唉，我说，你那本书买了吗？ 乙：哪本书？

这方面的具体做法是，语料转录人两人一组，交替工作，相互检查。在音频录制完成后，先用科大讯飞软件对音频素材先做简单转录，再经过多遍细听，对文本内容进行修订。完成转录后，再对语料文本进行附加标注。这期间有多轮复查工作。课题组对照学界影响较大的口语语料库(如 Santa Barbara Corpus of Spoken American English; English as Lingua Franca of Academic Discourse)的转写规范，并借鉴 Atkinson & Heritage (1984 [8])、Jefferson (2004 [9])、刘虹(2004 [10])的标注方法，本着“简洁”“实用”的原则，研发了适合 CED 语料特点的转写方案。详见表 1。

为保证语料库标注质量，采用双重标注的方法。第一重标注由语料转写者及语料文档的复查人完成；第二重标注由课题组经验丰富的教师来确定。CED 标注的 IAA (inter annotator agreement)指数高于 90%，符合相关要求(如 Palmer, Gildea & Kingsbury, 2005 [11])。

2.5. 语料库的构成

CED 语料库涵盖“公共话语”和“私人话语”两大类话语别。公共话语都是电话形式的交流，包括“情感咨询”和“事务咨询”两个子类别；私人话语包括“学生话语”和“亲子话语”两个子类别。详细情况请见表 2：

Table 2. The composition of CED
表 2. CED 构成

类别			篇数	字数(千)
公共话语	情感咨询	郑州新闻广播：《今夜不寂寞》 辽宁经济广播：《心有千千结》 济南新闻广播：《金山夜话》 网络电台：《峰人学院》	100	258
	事务咨询	河北交通广播：992 大家帮	52	72
私人话语	学生话语	研究生面对面闲谈	115	155
		研究生微信互动	84	45
		本科生交流	22	26
	亲子话语	母亲与孩童交流	157	135
总计			530	691

需要说明的是，字数总计是保守的数据统计。语料库的实际体量，要大于这个数值。CED 语料的选择有独特之处：“情感节目话语”既是日常话语，又是机构话语，篇幅较长，话语结构复杂，叙事视角不断转换，引人入胜；“992 话语”是生活中常见的咨询话语，话语参与者多、话轮短、话语重叠度高；亲子话语的主要参与方是年轻母亲与未成年孩童，孩童的年龄从两岁到十一岁不等。

3. 结语

汉语口语研究，宛如一个富矿，吸引着越来越多的国内外学者去挖掘、探索和发现。CED 近年被我校师生使用，已产出本科生优秀毕业论文、硕士学位论文及原创研究多项(如杨换丽，2019 [12]；谷伟明，2021 [13]；张丽媛，2024 [14])。CED 内的语料是自然的语言使用，是发生在我们身边的、最真实的汉语日常表达。在这些真实的语言使用中，有话语中断、倾听者反馈、重叠话语、错误开始、自我纠正、停顿、不连贯等，这些看似“坏的语法”(bad grammar)的语言现象，却是口语语法的鲜明特征。通过仔细

观察并记录这些非常规信息，研究人员对口语语言使用的洞察力，会得到大幅提升(Norrick, 2000 [15])。CED 所承载的是鲜活的语言使用的事实情况，希望它的建成，为蒸蒸日上的汉语话语研究提供新的素材和视角。

基金项目

本文为河北师范大学人文社会科学基金交叉协同科研项目(S2016JC06)的相关成果。

参考文献

[1] 付晓丽, 荣红, 董东, 宋文辉. 汉语日常话语语料库建设研究报告[R]. 石家庄: 河北师范大学, 2019.

[2] 郑家恒, 张虎, 谭红叶, 钱揖丽, 卢娇丽. 智能信息处理——汉语语料库加工技术及应用[M]. 北京: 科学出版社, 2010.

[3] Dash, N.S. and Arulmozi, S. (2018) History, Features, and Typology of Language Corpora. Springer. https://doi.org/10.1007/978-981-10-7458-5

[4] José de São, J. and Teixeira, A.R. (2013) At the "Ethical Crossroads" of Ethnography: Observing the "Care Encounter" at the Elderly Person's Home. In: Isabella, P., Maria, I. and Fernanda, M., Eds., *Practices of Ethics*: *An Empirical Approach to Ethics in Social Sciences Research*, Cambridge Scholars Publishing, 43-64.

[5] Keel, S. (2016) Socialization: Parent-Child Everyday Interaction. Routledge. https://doi.org/10.4324/9781315609706

[6] Tannen, D., Kendall, S. and Gordon, C. (2007) Family Talk: Discourse and Identity in Four American Families. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195313895.001.0001

[7] 王融. 数据匿名化的法律规制[J]. 研究与开发, 2016(4): 38-44.

[8] Atkinson, J.M. and Heritage, J. (1984) Structures of Social Action: Studies in Conversation Analysis. Cambridge University Press.

[9] Jefferson, G. (2004) Glossary of Transcript Symbols with an Introduction. In: Lerner, G.H., Ed., *Conversation Analysis*: *Studies from the First Generation*, John Benjamins, 13-31. https://doi.org/10.1075/pbns.125.02jef

[10] 刘虹. 会话结构分析[M]. 北京: 北京大学出版社, 2004.

[11] Palmer, M., Gildea, D. and Kingsbury, P. (2005) The Proposition Bank: A Corpus Annotated with Semantic Roles. *Computational Linguistics Journal*, **31**, 71-106. https://doi.org/10.1162/0891201053630264

[12] 杨换丽. 基于语料库的情感咨询中自我修正研究[D]: [硕士学位论文]. 石家庄: 河北师范大学, 2019.

[13] 谷伟明. 情感咨询人际互动中话语标记"然后"功能探析[J]. 现代语言学, 2021, 9(3): 737-747.

[14] 张丽媛. 电台咨询节目中建议拒绝言语行为[D]: [硕士学位论文]. 石家庄: 河北师范大学, 2024.

[15] Norrick, N.R. (2000) Conversational Narrative: Storytelling in Everyday Talk. John Benjamins. https://doi.org/10.1075/cilt.203