

外语课堂教学多模态语料库 MCCT 的创建

李玉香

(同济大学 外国语学院, 上海 200092)

摘要: 教育信息化 2.0 时代, 利用多模态语料库语言学方法, 通过语料库建设及其数据挖掘进行教学音视频流反思, 已成为课堂教学和教师发展研究的重要手段。在遵循语料库建设基本原则基础上, 聚焦“教师自身发展”专门用途外语课堂教学多模态语料库 MCCT 的创建, 强调建库目的。与此同时, 介绍建库过程, 概述文字转写、语料切分与标注等具体做法, 总结深加工后 MCCT 特点及其应用开发。目的是为高校教师利用多模态语料库方法进行教学、科研和自身发展提供理念思路、技术参考和典型样例。

关键词: 课堂教学; 多模态语料库; 教师自身发展; 专门用途语料库; 商务英语; 切分与标注

中图分类号: H 319 **文献标志码:** A **文章编号:** 1009-895X(2019)01-0001-11

DOI: 10.13256/j.cnki.jusst.sse.2019.01.001

Constructing Foreign Language Multimodal Corpus of Classroom Teaching (MCCT)

LI Yuxiang

(School of Foreign Languages, Tongji University, Shanghai 200092, China)

Abstract: The era of education informationization 2.0 highlights the multimodal corpus linguistic approach of corpus constructing and relevant data-mining in classroom teaching and teacher development research. This paper focuses on the construction of Foreign Language Multimodal Corpus of Classroom Teaching (MCCT) based on the basic corpus constructing principles, emphasizing the specific purpose of “teacher self-development”. In order to provide the conceptual framework, technological parameters, and a typical example for classroom teaching and teacher self-development research using multimodal corpus linguistic approach, this paper covers MCCT research aim, the construction process, detailed transcription procedures, data segmentation and annotation, as well as the specific features and future applications of the processed MCCT.

Keywords: classroom teaching; multimodal corpus; teacher self-development; specialized corpus; business English; segmentation and annotation

收稿日期: 2018-09-10

基金项目: 2017 年度教育部人文社会科学研究专项任务项目(工程科技人才培养研究)“工科学术英语论文多模态语料库建设与其在国际化工程科技人才培养中的应用研究”(17JJDGC034); 同济大学“欧洲研究”一流学科建设“欧洲思想文化与中欧文明交流互鉴”子项目资助课题(180200); 同济大学中央高校基本科研业务费资助项目(20160613); 同济大学中央高校基本科研业务费资助项目(221220170446)

作者简介: 李玉香(1967-), 女, 副教授。**研究方向:** 功能语言学、语料库语言学、多模态话语分析、应用语言学、学术英语、商务英语。E-mail: 2010liyuxiang@tongji.edu.cn

当今“互联网+”环境下,计算机软硬件技术,特别是人工智能、大数据、区块链等技术发展迅猛,教育信息化快速步入2.0新时代,高校师生不仅要增强信息技术应用能力,更要提升信息素养,让信息化理念、认识、思维方式和行动方法“内化于心,外化于行”^[1]。

教育信息化2.0时代,“数据”的基础性和重要性日益突显,数据建设、数据思维、数据挖掘成为高校教师教学、科研、管理和自身发展必须掌握的基本信息素养的重要内容。利用语料库语言学方法,通过课堂教学语料库建设、语料切分标注和数据挖掘,对课堂教学音视频流进行反思,已成为信息时代高校教师教学、科研和自身发展的重要方法。如何以课堂教学音视频流为原始语料,进行课堂教学多模态语料库建设,也成为一项十分紧迫的任务。

本文主要探讨“教师自身发展”专门用途外语课堂教学多模态语料库(Multimodal Corpus of Classroom Teaching,简称MCCT)的创建,聚焦语料库创建过程中的技术参数,包括建库目的、原始语料采集、样本选取、语料元信息、语料预处理、语料切分与标注规范、具体标注顺序,并简述深加工后语料库特点及其应用开发,旨在为高校教师利用多模态语料库方法进行教学、科研和自身发展提供理念思路、技术参考和典型样例。

一、文献综述

教学音视频流相关研究表明,教学音视频流分析能够促进教师教学反思,利用教学音视频流进行教学反思已成为教师改进教学的有效反馈策略,推动了教学改革和教师发展研究^[2-5]。随着音视频流在教师教育和职业发展中应用研究的指数增长^[6],发现音视频流标注(video annotation)方法对教师进行复杂自身教学视频流分析和反思具有巨大指导作用^[5,7-8];但教师专业见解的提升,需建构和分享基于音视频流证据的良好教学方法和实践的论点,而这些论点可以通过音视频流分析工具,如Studio Code,ELAN等来获得^[7,9]。可见,利用教学音视频流分析进行课堂教学反思,促进教师教育和职业发展的应用实践和理论研究,急需多模态语料库语言学研究方法的指导。

本文中多模态语料库是指包含文字、静态图像、动态图像(如视频流)和音频流多种模态数

据的多媒体资源库^[10]。多模态语料库语言学研究方法,就是根据研究目的,利用一定的工具,通过多模态语料库建设、切分与标注等加工,进行相关语料数据挖掘和分析统计,就相关问题进行实证研究,或者提出理论假设和理论建构的方法。

在学习科学领域,关于视频研究的理论架构逐渐成熟^[11-13]。随着音视频流语料相关研究的迅速发展,多模态语料库语言学研究方法在话语分析、教学研一体化发展等研究中的价值日益彰显^[11,14-17]。

在教师课堂教学反思方面,目前教学音视频流个案分析^[18-19]、教学竞赛案例分析^[20]、教师自我发展策略研究^[21]、通过多模态语料库研究路径将课堂教学科研化的方法论研究^[22]等不断发展,但是通过实际构建具有一定规模的课堂教学多模态语料库,并对课堂教学音视频流语料进行切分、标注和数据挖掘的研究尚不多见。

利用多模态语料库语言学方法,通过多模态语料库建设、加工和数据挖掘,基于坚实论据构建良好教学实践的论点^[9],已成为利用教学反思进行课堂教学和教师自身发展研究领域一项十分迫切的任务。

多模态语料库的建设与研究,在最近十几年,随着数字视频技术的普及逐步增加^[23]。

多模态语料库,依据一定的研究目的,按一定的标准收集,以计算机为基础,含有文字、图像(静态或动态)和音频流多媒体资源,且多模态语料库的音视频材料应同时伴有相关转写和标注信息^[24]。

像单模态语料库一样,多模态语料库按用途也分成:为一般用途而建的通用语料库(general corpus),如1991年建立的第一个含音频的BNC(British National Corpus,英国国家语料库)的口语语料库;为特殊用途而建的专门用途语料库(specialized corpus),如辛普森(R. C. Simpson)等人在1997至2002年间所建的总时长达200小时的密歇根学术英语口语语料库(Michigan Corpus of Academic Spoken English,简称MICASE)^[25]。

随着多模态理论研究的日渐成熟,多模态语料库建设成为多模态实证分析的必经之路^[26-27],而课堂教学研究与教材分析越来越成为热门话题^[27]。国外的多模态语料库有“新加坡教育专用语料库”^[28]“美国学术英语和职业英语专用语料库课堂互动多模态语料库”^[29],国内有“现场即席话语语料库”^[30]

“英语教育教学语料库”^[31-32]“理工科大学生英语口语多模态语料库”^[33]“多模态即席话语语料库”^[34]“多模态口译语料库”^[35]“当代英语教材专用语料库”^[36]等等。

但经过深度加工并具有一定规模的“教师自身发展”专用外语课堂教学多模态语料库尚不多见^[23,37-38]。

本文所建课堂教学多模态语料库 MCCT 为“教师自身发展”专门用途多模态语料库,原始语料为研究生公共英语选修课程《商务英语》课堂教学音视频流,共 64.5 小时,MCCT 语料有效时间共 16.48 小时,深加工语料时间为 10.9 小时。MCCT 深加工语料为三轮三类课堂教学音视频流,其中的“三轮”是指同一课程中同一教育目标类型的三轮次不同课堂教学录像,“三类”是指与课程教学内容相关的三种课堂教育目标类型,包括陈述性知识类、程序性技能类和问题解决类。

MCCT 语料库的标注体系(annotation scheme),涵盖课堂教学设计、组织、实施和效果评估等不同视角的 12 个分析维度,称为 12 个母标注层(annotation tier),具体有:文字转写、教育目标类型、教学任务分析、课堂组织架构、课堂教学事件、课堂活动类型、课堂师生互动模式、课堂师生话语、PPT 支撑的媒体、课堂教学用语类型、衔接与转换和学生注意力分析。每一个标注层都有各自的标签集(tag set),包括标签的工作定义和标注样例。根据研究需要的母标注层又包括相应的次级标注层。因本文篇幅所限,有关 MCCT 语料库的含 12 个维度的标注体系,包括其构建的理论基础、具体标签的设定、对前人研究的继承和自身的创新等,将撰文另叙^[39]。

基于 12 个维度切分与标注深度加工后的 MCCT 语料统计数据,可用于课堂教学研究和在职教师教学研一体化自身发展研究,此类“教师自身发展”专门用途外语课堂教学多模态语料库在国内外尚不多见。

二、“教师自身发展”专门用途 MCCT 的创建

本文所建语料库虽为多模态语料库,但其建设基本过程与单模态语料库的大致相同,首先要进行总体规划和准备,包括建库目的、语料库规模、语

料选取来源及采样原则、语料加工程度、语料采集及加工所用软硬件配备和建库所需人员及资金等。之后着手语料库的实际建设,具体过程包括语料获取及处理、语料库标注体系的制定、语料库的深度加工(即切分与标注)。

(一) 遵循多模态语料库建库基本原则

本文所建外语课堂教学多模态语料库 MCCT 为“教师自身发展”专门用途多模态语料库,建库时遵循一般语料库的建库原则,同时也遵循多模态语料库的建库原则。一般来说,语料库建库时,都应遵循以下五大原则^[40-41]:1) 目的原则,即计划用所建语料库来研究什么问题^[42];2) 代表性原则,即指语料样本在多大程度上包括了所研究对象总体变异的全部范围,且所包括的变异性范围包括语言学特征和非语言学特征(即情景特征)两个方面^[43-44];3) 数据采集原则,就是自然性(naturalness 或 spontaneity)原则,即收集语料时要尽量选择自然发生的语料,保持其原来的自然状态^[23];4) 原始数据编码原则,即标准化:不论是文本、音频、视频数据,还是语料库的元数据(metadata),其编码格式都要做到标准化,即讲求通用性、可共享性,这是不同语料库数据整合共享的基础;5) 切分与标注原则,即讲求效度、信度和一致性:建库目的决定语料切分与标注方案,但任何标注体系的制定和方案的实施,都要保证标注体系自身效度(validity)以及应用过程中标注的信度(reliability)和一致性(consistency),还要遵循里奇(G. Leech)提出的语料库标注的基本准则^[45]。

本文所建课堂教学多模态语料库 MCCT 的标注体系有 12 个分析维度,该标注体系经过了专家效度验证、试标注信度验证、试标注的专家标注信息反馈和标注者内部一致性验证,验证结果良好。因本文篇幅所限,MCCT 标注体系的具体信度、效度和一致性验证,也将撰文另叙。

(二) 课堂教学多模态语料库 MCCT 的创建

1. 建库目的:教师自身发展

本文所建课堂教学多模态语料库 MCCT 为“教师自身发展”专门用途多模态语料库,其总体设计、原始语料的采集、深加工语料的取样策略、语料库规模、语料库标注体系的制定等都根据教师自身发展这一研究目的而定,主要探讨“如何利

用课堂教学音视频流进行教师自身发展研究”这一核心问题,包含两个子问题,即对课堂教学音视频流如何处理和如何利用这些语料进行教师自身发展研究。

2. MCCT 原始语料采集

MCCT 原始语料的采集来自同济大学外国语学院汇文楼 418 多媒体-多模态实验室,实验教师就是笔者。笔者利用这个实验教室,对自己的课堂教学,即非英语专业硕士研究生的《商务英语》选修课程,进行跟踪录音录像,历时三年。

实验教室采集的内容同时包括三套音视频流:1) 教师的课堂教学实时行为(音视频流文件名均为 Teacher);2) 学生课堂学习的实时行为(音视频流文件名均为 Student);3) PPT 的演示活动(音视频流文件名均为 VGA)。采集语料的摄像头三年一直固定在实验教室的墙壁上。

录制这三套音视频流时,设备会同时将其自动放在同一个文件夹内,文件夹可以根据需要命名,可以包括录制时间及地点、课程名称、教师姓名、性别及职称等语料元信息。

现已录制三轮《商务英语》选修课程的教学录像,每轮 17 周,每周 2 学时,每学时为 45 分钟,每学期的最后 1 或 2 周安排的是课堂考试,有时因放假和设备原因无录像。实际教学录像情况具体为:第一轮为 31 学时,有 35 段教学录像;第二轮为 25 学时,有 25 段教学录像;第三轮为 30 学时,有 34 段教学录像。三年共跟踪课堂教学 86 学时,录制教

学录像 94 个时段(有时一个学时可能录制成 2 个时段;每个时段的录像均包括教师行为、学生行为和 PPT 演示 3 个音视频流),共 64.5 小时。

笔者对每学期的教学录像都按时间顺序进行了编号:第一轮为 2010 年春季学期,编号为 2010S_01 ~ 35;第二轮为 2011 年秋季学期,编号为 2011A_01 ~ 25;第三轮为 2012 年秋季学期,编号为 2012A_01 ~ 34。

这是本文所建课堂教学多模态语料库 MCCT 的原始资料来源。

3. MCCT 语料样本选取

MCCT 建库目的是进行教师自身发展研究,所选语料应足能代表这一现象。为保障语料的代表性,采取了分层分类抽样方法进行语料选取。

首先利用多模态标注工具 ELAN 软件对每个时段的教学录像进行课堂教学的教育目标类型标注。

根据教育目标分类学,特别是信息加工心理学的知识分类^[51]和加涅认知领域的学习结果分类^[52],将与课程教学内容相关的课堂教育目标分为三大类型,即陈述性知识类(Declarative Knowledge,简称 DK)、程序性技能类(Procedural Skill,简称 PCS)和问题解决类(Problem Solving,简称 PBS);将与每种课堂教学内容均相关的,学期开始时的课程介绍、期中考试、学期结束时的期末考试等归为通用类(Commons)。课堂教学的教育目标类型(Teaching Objective Type,简称 TOT)的标注体系,或称标签集,详见表 1 所示。

表 1 课堂教学教育目标类型(TOT)标签集
Tab.1 Teaching Objective Type(TOT) tag set

标签(Tag)	标签定义(Tag definition)	说明(Note)
Declarative Knowledge (DK)	陈述性知识类,要掌握涉及“是什么”的知识,包括术语、事实、事件、概念、原理等。如“定价策略”“公司的四种形式”等。	标注标签后可跟下划线再加上关键内容提示。
Procedural Skill(PCS)	程序性技能类,要习得一套程序性知识,即涉及“怎么做”的知识,并具备能按这一套程序去办事的能力。如“如何做口头演讲”“如何参加工作面试”等。	同上。
Problem Solving(PBS)	问题解决类,掌握“是什么”和“怎么做”的知识,并能按照既定程序去做,且在遇到全新情况和问题时,知道应该综合运用哪些知识和哪种程序去解决新问题。如“如何进行案例分析”等。	同上。
Commons	通用类,指那些所有类型的教育目标课堂教学都通用的,如学期开始时的课程介绍,结束时的期末考试等。这不是本论文的研究内容。	同上。

具体做法是,先将每个时段课堂录像文件夹里的教师课堂行为(Teacher)录像的音频流,利用 Ulead Video Studio 软件,或其他能从音视频流提取

音频流的软件,将音频流提取出来(每个文件夹里的 3 个音视频流是同步录制的,所以录制的音频是一样的内容),以 WAV 格式存放在这个时段的文

文件夹中; 然后利用 ELAN 软件将这一文件夹中的教师行为录像、学生行为录像、PPT 演示活动录像和课堂的音频流 WAV 文件, 导入新建的 EAF 标

注文件, 进行同步, 建立这个时段的 EAF 标注文件; 之后就可以用 ELAN 进行切分与标注, ELAN 标注界面如图 1 所示。



图1 利用 ELAN 工具进行语料切分与标注的界面

Fig.1 Data segmentation and annotation interface in ELAN

利用 ELAN 工具对三个学期的教学录像进行课堂教学教育目标类型 (TOT) 的初步切分与标注, 从标注结果中整理出 MCCT 多模态语料库样本的选取框, 即三轮课堂教学教育目标类型标注结果。

随后, 本文采纳分类抽样方法, 从 MCCT 样本选取框中, 选取具代表性的需进一步深入加工的语料。

语料选取的具体方法是从三轮教学录像中选取陈述性知识类 (DK)、程序性技能类 (PCS) 和问题解决类 (PBS) 三类课堂各一个话题的完整教学片段, 要求该话题的教学内容和教学录像均完整, 且每类课堂教育目标类型在三学期中一致, 这样每类课堂教育目标类型都有两轮提升。并且对教育目标类型的标注进行了精细化处理和完善, 如将紧随陈述性知识类、程序性技能类和问题解决类三类课堂教育目标之后的作业布置、休息宣布、下课宣布教学时段归属于该部分教学, 不再作为 Commons 归属处理, 这样完善了课堂教育目标类型的标签处理。

所选出的语料包括: 1) 陈述性知识类, 话题为“定价策略” (Pricing strategies), 三轮教学录像中共有 4 个教学片段; 2) 程序性技能类, 话题为“如何做有效口头演讲” (How to make an effective oral presentation), 三轮教学录像中共有 7 个教学片段; 3) 问题解决类, 话题为“如何做案例分析” (How to do case study), 三轮教学录像中共有 10 个教学片段。上述所选 21 个教学片段的语料编号详见下页表 2 所示, 教学视频持续总时间为 16.48 小时 (989 分钟), 占总原始语料的 25.55%。

本研究对这些语料中的陈述性知识类、程序性技能类和问题解决类三种教育目标类型的课堂教学音视频流进行深加工, 以探讨其在教师自身发展中的应用。

4. MCCT 语料预处理

笔者对所选 21 个教学片段的音视频流文件夹里的 EAF 标注文件进行了重新统一命名, 命名规则是: 课堂教育目标类型_学年学期_标注文件编

号.eaf,如 DK_2010S_01 代表陈述性知识类课堂教学类型的 2010 年春季学期的第 1 号标注文件。

21 个教学时段的音视频流的所有 EAF 标注文件

清单如表 2 所示,其中的 PBS_2010S_03 标注文件有两个(因录像操作原因):一个是 PBS_2010S_03(1).eaf,另一个是 PBS_2010S_03(2).eaf。

表 2 三类课堂教育目标的教学视频编号、EAF 标注文件、深入切分与标注视频持续时间清单
Tab.2 Video NO., EAF file name, and deep-processed data duration of three teaching objective types

轮次	陈述性知识类(定价策略)			程序性技能类(如何做口头演讲)			问题解决类(如何做案例分析)		
	视频编号	EAF 文件名	深入切分标注视频持续时间	视频编号	EAF 文件名	深入切分标注视频持续时间	视频编号	EAF 文件名	深入切分标注视频持续时间
第一轮	2010S_15	DK_2010S_01	00:12:53.857	2010S_07	PCS_2010S_01	00:17:58.772	2010S_27	PBS_2010S_01	00:19:46.735
							2010S_29	PBS_2010S_02	00:15:45.009
							2010S_31	PBS_2010S_03	00:45:10.370
第二轮	2011A_20	DK_2011A_01	00:06:27.783	2011A_03	PCS_2011A_01	00:44:03.446	2011A_16	PBS_2011A_01	00:51:22.227
							2011A_17	PBS_2011A_02	00:09:38.855
							2011A_18	PBS_2011A_03	00:58:04.772
第三轮	2012A_27	DK_2011A_01	00:39:33.041	2012A_01	PCS_2012A_01	00:34:24.064	2012A_19	PBS_2012A_01	00:34:19.429
							2012A_20	PBS_2012A_02	00:25:28.876
							2012A_21	PBS_2012A_03	00:43:36.786
							2012A_22	PBS_2012A_04	00:45:17.990

5. MCCT 的建立

经过对三年课堂教学录像的分层分类抽样,选出 21 个课堂教学片段,共计 16.48 小时,成为本文所建的课堂教学多模态语料库(MCCT)。其中第一轮课堂教学视频持续时间为 3:45:28.788,占该语料库的 22.75%;第二轮为 5:51:01.664,占 35.49%;第三轮为 6:52:35.153,占 41.76%。

在上述 16.48 小时的 MCCT 语料库中,本研究对陈述性知识类(定价策略),程序性技能类(如何做口头演讲)和问题解决类(如何做案例分析)三类教学内容的课堂教学时段进行深入加工,即 12 个标注维度的同步切分与标注,得到 MCCT 的深度加工部分,根据表 2 计算得知,MCCT 进行深层切分与标注的语料总时间为 10.9 小时(654 分钟),这是 MCCT 语料库内容的核心。语料中的其余 335 分钟,为正式上课前或正式下课后的视频,不在本文研究范围。

MCCT 多模态语料库整体建设过程如图 2 所示。

从表 2 可以计算出 MCCT 语料库深度加工部分的三轮教学语料中的每一轮语料的总时间,第一轮深度加工语料总时长为 1:51:34.743(约 111 分

钟),第二轮为 3:47:41.140(约 227 分钟),第三轮为 5:15:57.604(约 316 分钟)。MCCT 多模态语料库深加工部分中三轮语料所占时间比如图 3 所示。

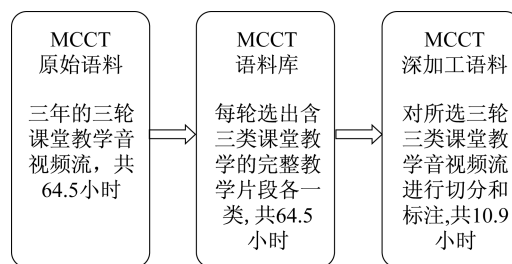


图 2 MCCT 多模态语料库建库过程

Fig.2 MCCT construction process

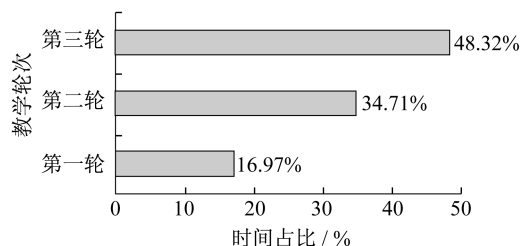


图 3 MCCT 深加工部分三轮语料所占时间比

Fig.3 Time percentage of three semester deep-processed MCCT data

同样,从表 2 可以计算出 MCCT 语料库深度

加工部分的三类课堂教学目标中每一类的教学视频总时间,其中陈述性知识类课堂教学共持续时间为1:08:16.502(约68分钟),程序性技能类课堂教学为3:58:33.936(约238分钟),问题解决类为5:48:31.049(约348分钟)。MCCT多模态语料库中深加工部分的三类教育目标课堂教学时间比如下图4所示。

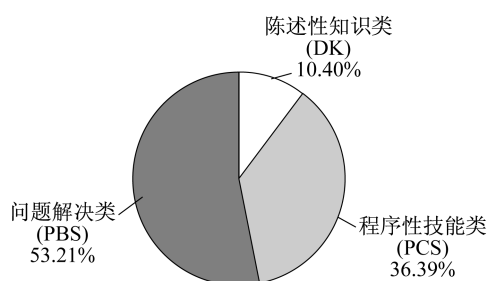


图4 MCCT深加工部分三类教育目标语料时间比

Fig.4 Three Teaching Objective Type time percentage of deep-processed MCCT data

(三) 课堂教学多模态语料库 MCCT 的深度加工

MCCT的深度加工,就是对MCCT的三轮教学中陈述性知识类(话题为“定价策略”)、程序性技能类(话题为“如何做口头演讲”)、问题解决类(话题为“如何做问题案例分析”)三种教育目标类型的课堂教学视频流的12个维度的切分与标注。本节主要介绍文字转写具体做法、切分与标注的具体规范、内容和顺序。

1. 文字转写

本文MCCT多模态语料库的核心内容,是由三轮完整的三类课堂教学内容,即陈述性知识类(“定价策略”)、程序性技能类(“如何做口头演讲”)和问题解决类(“如何做问题案例分析”)构成的,总时长为654分钟的教学音视频流。本研究将对这一部分语料进行切分与标注深度加工,以探讨教师自身发展问题。

对MCCT语料库进行深度加工(即切分与标注)的第一步是进行音频流的文字转写。

文字转写包括课堂教师话语和学生话语的文字转写。教师话语标注在一个标注层(T_Talk_Behavior);学生话语又按学生所在小组(老师一般会 把班级分成6个学习小组)和讲话发生的先后顺序进行编号,每位学生的课堂话语标在同一个标注层

内,有几个学生发言,就建几个标注层,如G1La_Talk_Behavior, G2Lb_Talk_Behavior, G4Ld_Talk_Behavior等。所以每个教学片段的文字转写部分的总标注层数,从一到十几层不同。

本研究中的文字转写是指课堂上教师和学生话语中的文字和数字部分的转写,不包含语调、音高、重音等信息,大小写和标点符号为标注者根据自己理解所添加。师生话语的文字转写,按语料的音频流内容忠实转写,发音和用词错误均不做修改,话语前添加讲话者的身份编码和冒号,冒号后空一个英文字符,句子开头单词首字母大写,如G2lb: So, that's all.

师生话语文字转写与相应的音频和视频流在时间上做同步处理,是分析师生课堂言语行为的重要依据之一。

2. 切分与标注规范

教育目标类型标注和文字转写后,其他各标注维度的具体切分与标注,依据MCCT多模态语料库制定的标注体系中各标注层的具体标签集进行,若有任何变化、歧义、不确定和修订的地方,都要立即做好文字记录,以备后查。

具体标注时的切分单位(unit of segmentation),不受师生话语长短的限制,而是根据某个词或词组、某一句话、几句话或一段话的实际功能和在教学过程中的作用来切分。每一标注层内,用不同的标签来标注和解释这一切分段(segment)的教学功能和作用。

如:PBS_2012A_04语料中,教师话语中的前两句话为:

T: OK. Let's continue.

T: Let's come to the analysis, come to the analysis of the problem.

这两句话,根据其不同功能,在不同标注层的具体切分与标注有所不同,如:1)在课堂组织架构(Lesson Organization)标注层中,这两句话为一个切分段,标注具体标签为Introducing(导入);2)在课堂教学事件(Lesson Instructional Event)标注层中,第一句话为一个切分段,标注的具体标签为Gain Attention(获取注意),第二句话和后面的相关内容一起被切分与标注。同样,该例在课堂活动类型、课堂师生互动模式等其他标注层中,都有相应不同的处理。

本研究所建MCCT语料库研究的核心问题是

“如何利用课堂教学音视频流进行教师自身发展”，所以语料切分与标注的重点是教师“教行为”的组织和实施的言语及其教学行为功能的分析。

本研究中的课堂师生话语(Classroom Talk)分析维度又包括教师话语(Teacher Talk)和学生话语(Learner Talk);教师课堂话语包括教师课堂管理话语、教师教学话语和教师人际交往话语。其中教师教学话语标注层又包括组织、讲解、引发、回应与反馈四种教学行为类(Pedagogical Act Type, 简称PAT)的次级标注层。

上述PBS_2012A_04语料中的例子,在教师教学话语中的组织教学行为类标注层(PAT_Structuring)中,第一句为一个切分段,标注具体标签为Call Attention(唤起注意);第二句话为一个切分段,标注具体标签为Organize Learning Activity(组织学习活动)。

课堂中教师的某一句话,可能同时具有多种教学功能,其标注可能会在两个甚至多个标注层中存在。从不同视角的标注,同一句话其功能可能有所异同,甚至在同一标注层,某一句话也可能有多种教学功能,但这时,标注的标签只是这句话最凸显的教学功能。

MCCT多模态语料库建库的主要目的是进行教师自身发展研究,教师的课堂教学方法是分析的重要内容,该语料库对教师话语里的组织、讲解、引发、回应与反馈四种教学行为类进行切分与标注,因为这是课堂教学内容和方法的重要和具体体现;而对教师课堂管理话语、教师人际交往话语、教师教学话语里的混杂教学行为类未做具体切分与标注。

MCCT多模态语料库中,12个维度的多视角、多层次、同步具体切分与标注,就是对教师的“教”行为和学生的“学”行为进行的深入剖析。对教师来说,这就是对自身教学过程进行深入剖析与反思的过程,这正是教师自身发展的核心环节。

3. 具体切分与标注顺序

MCCT多模态语料库语料的标注过程,是按照课堂教育目标、课堂教学组织及实施、课堂教学行为、课堂教学效果的检测等逻辑顺序进行的。具体切分与标注按下列步骤进行。

(1) 课堂教育目标是一切课堂教学组织和活

动安排及行为的指导,本文是最先对其进行切分与标注的。

(2) 课堂话语是课堂教学组织和实施的具体行为表现,是分析课堂师生言语行为的重要依据,本文第二个切分与标注的内容是课堂师生话语的文字转写。

(3) 实际课堂教学的组织架构以及教学具体实施的内容和手段,体现在课堂组织架构、课堂教学事件、任务分析、活动类型、互动模式、PPT支撑的媒体、衔接与转换和语言类型标注层,这也是分析课堂师生话语的重要背景参考,所以是标注的第三个步骤。

(4) 课堂师生话语的分析是语料切分与标注的核心部分,这部分的标注结果是课堂教学方法的重要体现。教师课堂话语分析是一种综合分析,要充分考虑话语发生的语境因素,教师在何种情况下,通过什么手段,想表达什么意图,话语的字面意义与教师的真正意图是不是一致,教师的语言行为和非语言行为是否一致,如不一致则真正的意图是什么,等等。本研究对课堂教师的非语言行为,如目光注视、手部姿势、头部姿势、身体移动等没做具体分析和标注,但在标注教师言语行为时要充分考虑这些因素,参照言语行为发生时教学音视频流的具体情况,进行综合考虑。

(5) 学生注意力分析这一维度是最后被切分与标注的,这是检测课堂教学效果的重要参数之一。

利用ELAN软件完成对MCCT多模态语料的切分与标注后,ELAN软件本身还具有相关标注信息的提取和统计分析功能。

(四) 深加工后课堂教学多模态语料库MCCT的特点

1. 专用性和代表性

MCCT建库目的是研究教师自身发展,其中包含陈述性知识类、程序性技能类和问题解决类三种课堂教育目标类型的三轮完整教学录像,MCCT是能代表教师自身发展的专门用途多模态语料库。

2. 具有一定规模

MCCT总有效时间为989分钟,深加工语料总

持续时间为 654 分钟, 对于深度加工的多模态语料库来说, 考虑到其深度加工的时间和人财物等成本, 可以说已具有一定规模。

3. 历时性

MCCT 时间跨越三年, 包括 2010 年春季学期、2011 年秋季学期和 2012 年秋季学期共三学期的课堂教学资料。

4. 标注维度和内容

MCCT 深加工语料总时长 654 分钟, 具体的切分与标注有文字转写、教育目标类型、教学任务分析、课堂组织架构、课堂教学事件、课堂活动类型、课堂师生互动模式、课堂师生话语、PPT 支撑的媒体、课堂教学用语类型、衔接与转换和学生注意力分析 12 个维度。其中课堂师生话语分析维度中, 对教师教学话语的标注包含 4 个次级标注层^[39-40], 具体如下。

(1) 组织教学行为类标注层, 该层的标签集包括安排学生做好学习准备、告知学习目标、陈述教学步骤、布置学习任务、组织课堂学习活动、安排学习环境、告知学习材料位置、分配任务时间、提出任务要求、各种衔接与转换、布置课后作业等 16 种具体标签。

(2) 讲解教学行为类标注层, 该层的标签集包括概览、阅读、翻译、释义、解释要点与难点、拓展、总结、联想、比较、对比、举例、提取先行知识、修辞性问句、强调重点等 23 种具体标签。

(3) 引发教学行为类标注层, 该层的标签集包括指示学生回答问题、指示做出汇报、启发学生思维、向学生提出问题、检查学生是否理解、向学生确认教师对学生话语的解释是否正确、寻求学生同意、挑战学生观点等 8 种具体标签。

(4) 回应与反馈教学行为类标注层的标签集包括教师对学生话语的评价反馈或修订, 澄清学生话语的确切含义, 学生回答后教师对其回答的重复、翻译、修正、拓展、解释、举例等, 教师对学生发言的总结以及教师对学生发言的关注的言语表示等 17 种具体标签。

5. 效度和信度验证

MCCT 所含 12 个维度的标注体系经过专家效

度验证, 语料试标注经过信度验证, 其标注信息有效且可靠。

6. 可统计信息和数据

MCCT 多模态语料的切分与标注均使用多模态标注工具 ELAN 软件进行, 可统计信息和数据包括: 命中率、频率、最小时长、最大时长、平均时长、总标注时长、标注时长比率、潜伏时间、传统文字转写文本等。

三、外语课堂教学多模态语料库 MCCT 的应用

课堂教学多模态语料库 MCCT 为研究“教师自身发展”专门用途而建, 相关研究可以包括: 1) 同一教育目标类型的课堂教学, 在不同轮次教学中的教学方法的异同分析; 2) 同一轮次课堂教学中, 不同教育目标类型的课堂教学的异同分析; 3) 同一教学时段, 在 12 个不同标注维度的标注结果的异同分析; 4) 不同教育目标类型的课堂教学, 教师教学话语中组织、讲解、引发、回应与反馈等教学行为类的实现手段的异同分析; 5) 陈述性知识类课堂教学的典型教学方法和策略分析; 6) 程序性技能类课堂教学的典型教学方法和策略分析; 7) 问题解决类课堂教学的典型教学方法和策略分析; 8) 学生注意力程度较高和较低的时段和原因分析; 9) 师生话语中何时使用母语, 何时使用目标语, 原因是什么; 等等。

除应用于上述教师自身发展类型的研究外, 多模态语料库 MCCT 还可进行基于课堂教学多模态语料库的相关研究, 如课堂教学教育目标类型分析、课堂教学事件分析、师生互动模式分析、PPT 制作与讲解策略分析、教学任务分析、课堂教师“教”行为及其效果分析、课堂学生“学”行为及其效果分析、师生语言行为与非语言行为对比分析、不同教育目标类型的有效教学策略分析、师生话语的单词数、词汇表及类符形符比统计等等。

基于 MCCT 多模态语料库的研究, 还可以与其他相关数据, 如教师网上评学、学生网上评教、教学督导听课、课堂教学实时评估、教师教学日记及反思、学生学习日记及反思等等, 相结合进

行互证研究,使研究论据更加详实,研究结论更加可靠。

四、结束语

本文所建外语课堂教学多模态语料库 MCCT 为“教师自身发展”专门用途语料库,遵循建库目的性、代表性等基本原则,其中深加工语料为针对陈述性知识类、程序性技能类和问题解决类三种不同课堂教学目标的历时三年的三轮次课堂教学音视频流,经过文字转写、教育目标类型、课堂教学事件、课堂师生话语等 12 个维度的同步切分与标注,并通过信度和效度验证,深加工后可用于课堂教学、教师发展和课堂话语分析等研究,为高校教师利用多模态语料库方法进行教学、科研和自身发展提供理念思路、技术参考和典型样例。

致谢: MCCT 语料库建设一直得到同济大学兼职博士生导师中国社会科学院顾曰国教授指导,本文写作过程中得到南京大学陈新仁教授关于主题聚焦的热情指导,谨致谢忱。

参考文献:

- [1] 雷朝滋. 教育信息化:从 1.0 走向 2.0——新时代我国教育信息化发展的走向与思路[J]. 华东师范大学学报(教育科学版),2018,36(1):98-103.
- [2] Stigler J W, Hiebert J. Understanding and improving classroom mathematics instruction: an overview of the TIMSS video study[J]. The Phi Delta Kappan, 1997, 79(1):14-21.
- [3] Fuller F F, Manning B A. Self-confrontation reviewed: a conceptualization for video playback in teacher education[J]. Review of Educational Research, 1973, 43(4):469-528.
- [4] Penny A R, Coe R. Effectiveness of consultation on student ratings feedback: a meta-analysis[J]. Review of Educational Research, 2004, 74(2):215-253.
- [5] Tripp T, Rich P. Using video to analyze one's own teaching[J]. British Journal of Educational Technology, 2012, 43(4):678-704.
- [6] Koh K. The use of video technology in pre-service teacher education and in-service teacher professional development[C] // Ng S F. Cases of Mathematics Professional Development in East Asian Countries: Using Video to Support Grounded Analysis. Singapore: Springer, 2015:229-247.
- [7] Rook M M, McDonald S P. Digital records of practice: a literature review of video analysis in teacher practice[C] // Proceedings of 2012 Society for Information Technology & Teacher Education International Conference. Austin: AACE, 2012:1441-1446.
- [8] Gaudin C, Chaliès S. Video viewing in teacher education and professional development: a literature review[J]. Educational Research Review, 2015, 16:41-67.
- [9] McDonald S. Building a conversation: preservice teachers' use of video as data for making evidence based arguments about practice[J]. Educational Technology, 2010, 50(1):28-31.
- [10] 顾曰国. 论言思情貌整一原则与鲜活话语研究——多模态语料库语言学方法[J]. 当代修辞学, 2013(6):1-19.
- [11] Goldman R, Pea R, Barron B, et al. Video Research in the Learning Sciences[C]. Mahwah: Lawrence Erlbaum Associates, 2007.
- [12] Roth W M. Epistemic mediation: video data as filters for the objectification of teaching by teachers[C] // Goldman R, Pea R, Barron B, et al. Video Research in the Learning Sciences. Mahwah: Lawrence Erlbaum Associates, 2007:367-382.
- [13] Derry S J. Video research in classroom and teacher learning(Standardize that!)[C] // Goldman R, Pea R, Barron B, et al. Video Research in the Learning Sciences. Mahwah: Lawrence Erlbaum Associates, 2007:305-320.
- [14] Gu Y G. Multimodal text analysis: a corpus linguistic approach to situated discourse[J]. Text and Talk, 2006, 26(2):127-167.
- [15] Gu Y G. From real-life situated discourse to video-stream data-mining[J]. International Journal of Corpus Linguistics, 2009, 14(4):433-466.
- [16] Gu Y G. A chess master model for classroom teaching and teacher/researcher development[J]. Chinese Journal of Applied Linguistics, 2012, 35(1):5-23.
- [17] 李玉香. 教师发展“大师复盘模式”:以商务英语教师发展为例——顾曰国和 Evan Frendo 访谈录[J]. 中国外语, 2012(6):81-85.
- [18] 张莲. 问答、交互与课堂话语——一位高校英语专业教师的课堂话语个案分析报告[J]. 中国外语教育, 2009, 2(2):56-64.
- [19] 张莲, 王艳. 通过课堂话语分析促进外语教师学习:一

- 项实证案例研究[J]. 外语与外语教学, 2014(3): 36 - 41.
- [20] 朱彦, 杨红燕, 束定芳. 外语课堂教学话语有效性的多维度评析——试析第四届“外教社杯”全国高校外语教学大赛教学案例[J]. 外语教学, 2016, 37(1): 53 - 57.
- [21] 李玉香. 高校在职外语教师自我发展策略[J]. 上海理工大学学报(社会科学版), 2013, 35(3): 269 - 273.
- [22] 李玉香. 如何通过多模态语料库研究路径将课堂教学科研化[C] // 张德禄, 黄立鹤. 多模态与外语教育研究. 上海: 同济大学出版社, 2018: 111 - 222.
- [23] Knight D. The future of multimodal corpora: o futuro dos corpora modais [J]. Revista Brasileira de Linguística Aplicada, 2011, 11(2): 391 - 415.
- [24] Allwood J. Multimodal corpora [C] // Lüdeling A, Kytö M. Corpus Linguistics: An International Handbook. Berlin: Mouton de Gruyter, 2008: 207 - 225.
- [25] Simpson R C, Briggs S L, Ovens J, et al. The Michigan Corpus of Academic Spoken English [M]. Ann Arbor: The University of Michigan, 2002.
- [26] Bateman J A. Using multimodal corpora for empirical research [C] // Jewitt C. The Routledge Handbook of Multimodal Analysis. London: Routledge, 2014: 238 - 252.
- [27] 冯德正, Low F. 多模态研究的现状与未来——第七届国际多模态会议评述[J]. 外国语, 2015, 38(4): 106 - 111.
- [28] 吴福焕, 林进展, 周红霞. 新加坡教育专用语料库的建设与应用[J]. 华文教学研究, 2016(3): 36 - 45.
- [29] Amory M, Kissilev O. The annotation of gesture designed for classroom interaction [C] // Proceedings of International Conference on Language Resources and Evaluation; Multimodal Corpora: Computer Vision and Language Processing. Portorož, 2016.
- [30] 顾曰国. 北京地区现场即席话语语料库取样与代表性问题[C] // 中国社会科学院世界经济研究中心. 全球化与 21 世纪: 首届“中法学术论坛”文集. 北京: 社会科学文献出版社, 2002: 484 - 500.
- [31] 何安平. 语料库在外语教育中的应用——理论与实践[M]. 广州: 广东高等教育出版社, 2004.
- [32] 何安平, 等. 英语教育语料库教学光盘[CD]. 广东省音像教材出版社, 2000.
- [33] 刘芹, 潘鸣威. 理工科大学生英语口语多模态语料库构建研究[J]. 现代教育技术, 2010, 20(4): 69 - 72, 119.
- [34] 张佐成, 邢建玉, 王彦, 等. 多模态即席话语研究[M]. 广州: 世界图书出版广东有限公司, 2014.
- [35] 刘剑, 胡开宝. 多模态口译语料库的建设与应用研究[J]. 中国外语, 2015(5): 77 - 85.
- [36] 胡开宝, 李翼. 当代英语教材语料库的创建与应用研究[J]. 外语电化教学, 2016(3): 34 - 39.
- [37] Wolf J C, Bugmann G. Linking speech and gesture in multimodal instruction systems [C] // Proceedings of the 15th IEEE International Symposium on Robot and Human Interactive Communication. Hatfield: IEEE, 2006.
- [38] Knight D, Evans D, Carter R, et al. Redrafting corpus development methodologies: blueprints for 3rd generation “multimodal, multimedia” corpora [J]. Corpora, 2009, 4(1): 1 - 32.
- [39] 李玉香, 顾曰国, 吕培明, 等. 外语课堂教学音视频流切分与标注综合架构——以“教学反思”为导向的多模态语料库视角[R]. 上海: 上海外国语大学“语言·认知·话语第一届语言学国际学术研讨会”, 2017.
- [40] 李玉香. 课堂教学多模态语料库建设与应用研究——以教师自身发展为例[D]. 上海: 同济大学, 2015.
- [41] 李玉香. 多模态语料库建设的基本原则和实现路径[R]. 同济大学第二届多模态与特殊人群话语多学科研究求索论坛多模态主要研究方法专家讲座, 2018.
- [42] Biber D, Conrad S, Reppen R. Corpus Linguistics [M]. Beijing: Foreign Language Teaching and Research Press, 2000.
- [43] McEnery T, Xiao R, Tono Y. Corpus-Based Language Studies: An Advanced Resource Book [M]. New York: Routledge, 2006.
- [44] Biber D. Representativeness in corpus design [J]. Literary and Linguistic Computing, 1993, 8(4): 243 - 257.
- [45] Leech G. Corpus annotation schemes [J]. Literary and Linguistic Computing, 1993, 8(4): 275 - 281.

(编辑: 朱渭波)