[文章编号]1009-3729(2013)04-0093-04

地方口音英语学习者语音库建设构想

陈文凯

(郑州轻工业学院 外语学院, 河南 郑州 450002)

[摘 要]我国地域辽阔,地区口音差异显著,建设针对地方口音英语学习者的语音库显得十分必要和重要。建设此语音库须科学划分方言区、确定发音人及语料采集地点、根据发音人的特点确定任务类型及发音语料、科学制定录音方案以确保录音的高保真、根据音系学理论和地方口音特征开发语音标注系统。其中,语音标注是语音库建设的核心环节。语音库可采用ToBI及IViE标注系统,对音段、韵律及非语言信息进行全面标注。该语音库既能为学习者的二语语音习得提供有益参照,又能展示英语学习者语音习得的区域性特点,有助于勾勒中国英语学习者语音习得全貌。

[关键词]地方口音;语音库建设;语音标注

[中图分类号]H311 [文献标志码]A [DOI]10.3969/j.issn.1009-3729.2013.04.017

随着计算机与语音技术的飞速发展,语音学在研究方法上已从传统"口耳之学"的质性研究发展为质性研究与量化研究相结合,语音库建设也应运而生并呈繁荣发展的趋势。语音库是为特定目的而建立的有关语音语料及其标注的集合,与传统的语音材料相比有明显的优势。首先,语音库能提供真实语音材料,对学习者更有吸引力和说服性,是传统教材配套录音材料所不可比拟的。其次,语音库提供的语音材料包含不同语音变体,有助于学习者熟悉各种语音变体,消除因口音问题而造成的理解障碍。基于此,在拥有大量英语学习者且方言区众多的中国,有必要进一步研究地方口音英语学习者的语音库建设问题。本文拟在梳理分析国内外口语语料库和语音库建设现状的基础上,阐述建设地方口音英语学习者语音库的必要性和重要性,并以河南口音英语学习者语音库建设为例提出建设构想。

一、国内外口语语料库和语音库建设现状

近年来,国内外在口语语料库和语音库建设方面取得了显著的成果。国外口语语料库及语音库建设已经比较成熟。1970年代初,英国建成第一个英语口语语料库——伦敦-兰德英语口语语料库(London-Lund Corpus of Spoken English),该语料库设计严密,质量上乘,进行了严格的语调、重音、停顿等韵律标注。[1]比利时Louvain大学S. Granger教授负责创建的鲁汶国际英语口语中介语数据库(简称LIDSEI:Louvain International Database of Spoken English Interlanguage)是一个包括日本、瑞典、西班牙、意大利、保加利亚和中国等国家英语学习者的中介语子库。目前,国际上涵盖英语语音变体最全面的英语口音语音库当属美国乔治梅森大学的Steven H. Weinberger教授主持建设的Speech Accent Archive,它包含不同母语背景者的英语口音语料。截至2013年4月17日,该语音库已有1 734个样本,包含英语母语者和非母语者朗读同一段文字录音,并且设计者对朗读者的录音进行了详细转写和标注,可用于对比分析不同母语背景人士的英语语音特征。此外,由英国牛津大学和剑桥大学联合开发并于2002年建成的IViE Corpus(Intonational Variation in English Corpus),设计科学、语料丰富、标注全面,堪称口音语音库建设的典范。

[收稿日期]2013-05-15

[基金项目]河南省哲学社会科学规划项目(2012BYY015);河南省软科学项目(122400450039)

[作者简介]陈文凯(1966—),女,河南省遂平县人,郑州轻工业学院副教授,主要研究方向:应用语言学、二语习得。

国内已建成两个大型的英语学习者口语语料库:2002 年上海交通大学与广东外语外贸大学联合创建的以大学英语四、六级口语考试录音为语料的中国学习者英语语料库(CLEC:Chinese Learners' English Corpus,口语部分为 50 万词的 COLSEC)和 2005 年南京大学建成的以英语专业四级口试录音为语料的 SWECCL(Spoken and Written English Corpus of Chinese Learners,口语部分为 100 万词的 SECCL)。近年,国内英语语音库建设紧跟国际潮流,2008 年陈桦等主持建立的中国英语学习者语音库 ESCCL(English Speech Corpus of Chinese Learners)以方言区为点、以地域分布为面、以国内 4 个不同层次受教育群体(初中、高中、英语专业本科、英语专业硕士)作为录音对象、以朗读和自主对话为任务,并结合英美标注系统对学习者录音进行多层音段及韵律标注[2];纪晓丽等主持建立的多口音英语学习者语音库 CELSCOM(Chinese EFL Learners' Speech Corpus with Multi-accents)收集了中国不同方言区英语学习者(这些学习者是来自北方地区的以母语为普通话的员工)的英语口语语音语料[3]。

这些国内外已建成的英语口语语料库及语音库,一方面为语音学、音系学及二语语音习得研究提供了全新的研究视角和研究平台;另一方面又提供了有关学习者语言发展的全面信息,便于研究者对大量口语事实及语音现象进行分析,进而寻找语言使用规律,对语音及口语研究和习得有着极大的促进作用。然而,国外建设的英语口语语料库及语音库几乎不涉及中国英语学习者,而国内已建成的中国英语学习者口语语料库及语音库(如 ESCCL)尽管覆盖的地域较广,却不能兼顾所有地区细微的区域性差异,并且未直接使用国外已建成的母语语音库中的发音语料,不利于研究者及二语学习者进行语音对比。因此,在借鉴现有口语语料库及语音库建设成果的基础上,建设真正具有区域特色的英语学习者语音库显得十分必要和重要。

二、地方口音英语学习者语音库的建库构想与设计

以河南地方口音英语学习者语音库的构建为例,河南省地域辽阔,地区口音差异显著,而目前尚未建有专门的基于河南口音的英语学习者语音库。河南口音英语学习者语音库 He′nan EFL(English as Foreign Language) Learners' Speech Corpus with Multi-accents(简称 HELSCOM)正是在此背景下启动研究与建设的。其建库总体原则是:以河南口音英语学习者为研究对象,不仅要建设学习者的英语口音语音库,而且要建设在河南方言语境下英语学习者的汉语发音字库(附有发音人方言录音),便于研究不同方音对学习者英语语音学习的影响。另外,语音库整体上与 IViE 在语料上保持一致。

"任何基于语料库的研究均始于语料库建库,关于要选什么样的语料入库以及有序地筛选语料几乎决定了后续要做的一切相关研究"[4]。因此,建设语音库前首先要做好以下工作:科学划分河南方言区、根据方言区的划分确定发音人及语料采集地点、根据发音人的特点确定任务类型及发音语料、科学制定录音方案以确保录音的高保真、根据音系学理论和地方口音特征开发语音标注系统。

1. 科学划分方言区

科学划分河南方言区对 HELSCOM 建库非常重要,是建库的前提。河南方言是在漫长的演变过程中受语言本身及社会、历史、地理等诸方面因素的影响逐渐形成的。关于河南方言区的划分及河南省境内中原官话区的划分,专家意见不一。根据《河南省志·方言志》,河南方言分为两大类:一类属晋语,主要包括豫北的焦作、新乡、安阳等所辖县市;一类是中原官话,根据其内部语音差异,又可分为 5 个片。全省的方言片整体上有 6 个:郑汴片、洛嵩片、蔡汝片、信潢片、陕灵片、安沁片。[5]根据《中国语言地图集》,河南境内的中原官话区包括 9 个片:兖菏片、郑开片、安沁片、洛嵩片、南鲁片、漯项片、商阜片、信蚌片、汾河片。[6]贺巍[7]在此基础上根据若干语音特点,对中原官话做了调整,删掉了其中的"安沁片",把中原官话分为 8 个片,以求在区域分布上更趋合理。

在对河南省境内不同口音英语学习者的语音状况进行综合调研的基础上,结合上述有关河南方言区及河南省境内中原官话区的划分,本语音库所选发音人拟从郑开片、洛嵩片、南鲁片、漯项片、商阜片、信蚌片、汾河片、安沁片 8 个片中选取,同时兼顾行政区划分,主要体现在所选的具体县市上。受普通话推广及城市化的影响,一些重要的方言语音特征在城市人群中尤其是在年轻人口中已经很难听到,所以建库时主要选择来自方言区农村乡音比较浓重的发音人。

2. 确定发音人及语料采集地点

参照 IViE Corpus 的做法,从上述 8 个方言区中选取 96 名河南口音英语学习者,每个区域 12 名,男

女各 6 名,使其涵盖我国英语教育的 6 个层次:小学生、初中生、高中生、非英语专业大学生、英语专业大学生、英语专业研究生。

3. 确定发音素材

为了对比研究不同方言(口音)英语学习者的英语发音与标准发音的差异,以进一步研究其英语中介语系统,本语音库的录音材料涉及发音者方言录音及其英语录音,前者为发音者就某个话题用方言讲述 2 分钟,后者的语料主要基于 IViE Corpus,并根据不同发音人的英语水平对语料进行筛选和微调。参照 IViE Corpus 的录音素材,发音人需完成多种类型的任务,包括句子朗读(含陈述句、一般疑问句、特珠疑问句、祈使句、感叹句等)、故事复述、访谈以及口头交流,以全面客观地反映学习者的中介语语音状况。朗读的句子取自复述故事素材中出现的简单句,针对不同层次的朗读者,朗读的句子难易度不同;故事复述素材取自《灰姑娘》《黑骏马》《爱丽丝漫游奇境记》等经典儿童文学作品或其改编版;访谈及口头交流话题为《普通高中英语课程标准》(实验)所列常用话题,如天气、购物、旅游等。

4. 录音及其整理

录音在专业语言实验室里进行,使用索尼 ICD - SX712(2G)录音笔(带指向性功能的麦克风)进行录音,确保录音音质的高保真;使用专业的音频编辑器 Cool Edit Pro 2.0 对录音进行编辑整理,并根据任务类型和录音人情况(受教育层次、性别、地区分布等)对编辑好的语音文件进行分类整理。

5. 语音标注

语音标注是语音库建设的核心环节,指的是使用 Pratt 软件对编辑整理好的录音进行详细的语音标注,包括音段标注、韵律标注及非语言信息的标注。

6. 语音库后期管理

语音库后期管理的主要任务有:上传初始语料,即按照学习者所在方言区对语料进行分类;语音库录音备份和数据统计;语音库查询;语音库对比学习界面设计等。

在 HELSCOM 建库过程中,主要难点是确定发音人、进行韵律标注。在筛选发音人时,要充分考虑发音人的口音特征和背景,这是建库中的一大难点;韵律标注较为复杂,要求高(语音库建设需要既精通技术又懂语音学的专门技术人员),难度大,做起来费时费力。

三、HELSCOM 语音标注方案

HELSCOM 标注方案主要借鉴 ToBI 标注系统及 IViE 标注系统,并结合本语音库实际情况进行了适当调整。下面以句子 This makes life difficult for those who prefer to use their left hands. 的标注为例(见图 1),详细阐述 HELSCOM 标注具体方案。

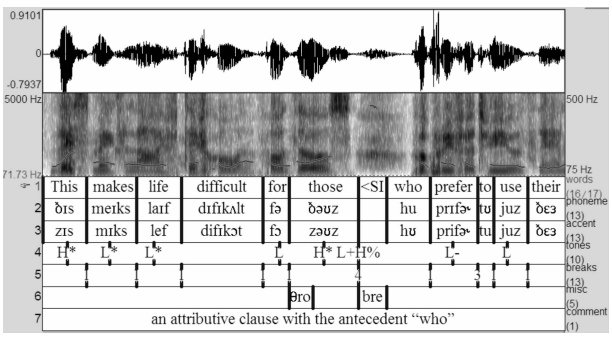

图 1　HELSCOM 标注实例

1. 音段标注

音段标注实质上是文本信息与声音信息之间的转换,即将表义的文本形式转化为表音的文本形式。HELSCOM 的音段标注包括:(1)文本层,将声音信息转写为文本形式,如图 1 中的第 1 层;(2)音位层,将声音信息转写成相应的国际音标形式(如图 1 中的第 2 层);(3)口音层,将学习者带有地方口音的实际发音转写成相应国际音标形式(如图 1 中的第 3 层),以便于分析和研究地方口音英语学习者的发音特征及习得策略。这也是本语音库的特色之一,它可从口音层标注中找出学习者的发音错误,如/ð/ - /z/,/eI/ - /I/,/aI/ - /e/,/u/ - /ʊ/等。

2. 韵律标注

韵律标注即对语音信号中具有语言学意义的韵律特征进行的质性描写。HELSCOM 的韵律标注包括 2 个标注层级:(1)语调层,对讲话人的重音及语调的变化进行描述(如图 1 中的第 4 层)。H 表示升调、L 表示降调,H * 表示升调的重读音节位置、L * 表示降调的重读音节位置,L + H% 表示呈降 - 升边界调,L - 表示降调短语。此外,在进行语调层标注时,还会用到其他标注符号,如 L + H * 、L * + H、! H * 、H + ! H * 等。(2)停顿指数层,对讲话人的停顿进行描述。如图 1 中的第 5 层中,1 表示短语中词与词之间的停顿,3 表示固定短语中两个词之间的停顿,4 表示一个语调群的结尾等。

3. 非言语信息及其他信息的标注

交际中除了语言信息之外,还有很多副语言和非语言信息。HELSCOM 的非言语信息标注即杂类层:标注各种副语言现象(如拉长、喘气、含混音、哈欠、清嗓、喷嚏、哭声、咳嗽等)和非语言学现象(如嗓音等),如图 1 第 6 层所示。在标注过程中可根据情况采用不同的标注方式,如 bre 表示喘气;θro 表示嗓音等。此外,HELSCOM 的标注还包括对有关语料的注释即评论层的标注,如图 1 第 7 层所示,标注样本为一个 who 引导的定语从句。

四、结语

具有区域特色的英语学习者语音库建设具有重要的应用价值:(1)便于国内外语言学习者、研究者进行语音资料的查询、检索、统计和研究,便于对比研究方言区不同口音的英语学习者与英语母语者的发音,更深层、更系统、更客观地研究中国英语学习者的中介语系统;(2)从语音技术方面为英语语音识别提供区域性的训练音库;(3)便于对学习者的英语语音特征和语音发展进行全面而系统的描述和分析,为二语语音习得及研究和口语教学及研究提供一定的依据,有助于改善英语学习者的语音面貌及口语交际能力;(4)展示国内英语学习者语音习得的区域性特点,与其他中国英语学习者语音库互补,有助于勾勒中国英语学习者语音习得全貌。

河南口音英语学习者语音库建成后将是一个有母语语音库参照的,附有详细音段标注、韵律标注和非言语信息标注的具有地域特征和方言特征的英语学习者语音库,包含文本、语音和声学参数三种形式的界面,兼具查询、检索和对比学习等功能。可为英语语音识别提供区域性的训练音库,为二语语音教学、测评及学习者的二语语音习得提供参照,有助于改善英语学习者的语音面貌及口语交际能力。更重要的是,该语音库可为河南口音英语与标准英语的语音特性差异研究和基于语音库的学习者中介语语音习得实证研究提供平台,推动中国英语学习者中介语研究及二语语音习得理论的构建。

[参 考 文 献]

[1] Svartvik J, Eeg-Olofsson M. Tagging the London-lund corpus of spoken English[C]//Computer Corpora in English Language Research. Bergen: Norwegian Computing Centre for the Humanities, 1982:85.

[2] 陈桦,文秋芳,李爱军. 语音研究的新平台——中国英语学习者语音数据库[J]. 外语学刊,2010(1):95.

[3] 纪晓丽,孙佳,李爱军,等. 多口音英语学习者口语语音库[A]. NCMMSC,2009:284.

[4] John Sinclair. Corpus, Concordance, Collocation[M]. Oxford: Oxford University Press, 1991.

[5] 邵文杰. 河南省志·方言志[M]. 郑州:河南人民出版社,1995.

[6] 中国社会科学院,澳大利亚人文科学院. 中国语言地图集[M]. 香港:朗文出版(远东)有限公司,1987.

[7] 贺巍. 中原官话分区(稿)[J]. 方言,2005(2):136.

(上接第 87 页)

[3] Glosten L P Milgrom. Bid, ask, and transaction prices in a specialist market with heterogeneously informed traders [J]. Journal of Financial Economies, 1985, 14(1):71.

[4] Mankiw N. The equity premium and the concentration of aggregate shocks [J]. Journal of Financial Economies, 1986(17):211.

[5] Campbell J C, Kyle A. Smart money, noisy trading and stock price behavior [J]. Review of Economic Studies, 1993 (60):1.

[6] Wang J. A model of inter-temporal asset prices in securities markets[J]. Journal of Review of Economic Studies, 1993(60):249.

[7] Wang J. A model of competitive stock trading volume[J]. Journal of Review of Economic Studies, 1994(102):127.

[8] Fama E. Market efficiency, long-term returns, and behavioral finance [J]. Journal of Financial Economies, 1998 (49):283.

[9] 唐伟敏,邹恒甫. 一种不完全信息下的资产定价模型[J]. 经济学(季刊),2008(1):309.

[10] 梁玉梅,李红刚. 信息不对称框架下资产均衡定价模型分析[J]. 北京师范大学学报:自然科学版,2006 (4):437.

[11] 李平,曾勇,唐小我. 有限记忆对金融资产短期价格的影响分析[J]. 系统工程学报,2004(4):408.

[12] 李平,曾勇. 基于不完全理性学习的资产价格行为分析[J]. 电子科技大学学报,2005(6):857.

[13] 宋军,吴冲锋. 金融资产定价异常现象研究综述及其对新资产定价理论的启示[J]. 经济学(季刊),2008 (1):701.

[14] 袁子甲. 不完全信息下的投资组合选择与资产定价[D]. 广州:中山大学,2009.