语音研究的新平台:中国英语学习者语音数据库*

陈 桦 文秋芳 李爱军
(南京大学, 南京 210093; 北京外国语大学, 北京 100089;
中国社会科学院, 北京 100732)

提 要: 本文简要介绍中国英语学习者语音库 ESCCL(English Speech Corpus of Chinese Learners)建设的起因、方法及意义。基于语料库进行大规模实证研究已成为语言学研究的主流方法, 但是现存的中国英语学习者口语语料库均为文本转写格式的语料库; 同时, 录音时没有控制噪音, 不宜进行语音学研究。因此, ESCCL的建设以方言区为点、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而完成, 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。语音库的建成势必为二语习得研究提供又一个平台, 为我国的英语教学与研究服务。

关键词: 英语语音库; 学习者
中图分类号: H319.3 文献标识码: A 文章编号: 1000-0100(2010)01-0095-5

A Learner Corpus) ESCCL

Chen Hua Wen Qiu2fang Li Ai2jun
(Nanjing University, Nanjing 210093, China; BFSU, Beijing 100089, China;
Chinese Academy of Social Sciences, Beijing 100732, China)

This paper describes the reason, the design and the implication of compiling a learner corpus) ESCCL (English Speech Corpus of Chinese Learners). As for the main reason, the existing spoken corpora of Chinese EFL learners in China are completely text2based, and not suitable for phonetic analysis because of the poor quality of the recordings. The subjects at four different educational backgrounds were asked to fulfill two tasks) reading2aloud and topic2based spontaneous dialogue. The recordings were collected from different parts of China and dialectal areas. The annotation system employed in the corpus combines the British system and the American one. The corpus2based research findings have important implications for China. s EFL pedagogy, and will be helpful for the improvement of rating rubrics for China. s oral English tests.

Key words: English speech corpus; learner

1 必要性

随着计算机技术的飞速发展, 利用语料库(corpora)进行语言学研究应运而生(Biber等 1998; Chafe 1992; Johansson 1982; 李爱军 2001)。语料库是应用计算机技术对海量自然语言材料进行处理、存储, 以供检索(retrieval)、索引(concordance)和统计分析的大型资料库, 它是按照明确的设计标准为某一具体目标而建立的语言资料库(Armstrong 1993; Granger 1998; 李文中 1999)。因此, 国内建成了两个大型的学习者口语语料库: 由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、六级考试语料为主要来源的 CLEC(Chinese Learners. English Corpus)(口语部分为 50万词的 COLSEC); 由南京大学创

* 本文系教育部人文社科基金项目/构建中国英语学习者语音库的模式研究0(06JA740031)的阶段性成果。第一作者为北京外国语大学中国外语教育研究中心兼职研究员, 第二作者为北京外国语大学中国外语教育研究中心专职研究员。

建的英语专业学生四级口试语料组成的 SWECCL (Spoken and Written English Corpus of Chinese Learners) (口语部分为 100万词的 SECCL) (王立非 孙晓坤 2005)。基于这些文本格式的语料库, 大量的研究论文面世, 使人们对学习者具有口语特征的词汇、句法、语篇、语用和其他方面有了一定的认识。

但是, 现有的学习者英语口语语料库存在以下不足: (1)学习者口语特征不仅指词汇、句法等方面, 而且应该研究学习者的英语语音特征, 才能对学习者的英语口语有全面了解; (2)针对学习者口语特征的现有研究几乎都是错误分析, 缺少对学习者英语口语特征的全面、客观的描述; (3)现有语料库均为文本转写格式, 在其上进行文本标注, 缺少韵律标注; (4)语音分析须要依赖声学指征, 因而对所分析语料的录音质量要求高。现有的学习者语料库创建时并非针对语音研究, 并未控制噪音, 因此不宜进行语音学研究 (见图 1); (5)研究中国英语学习者语调模式发现, 我国英语学习者在使用英语语调传递信息时存在一定问题, 不仅会影响本族语听者的理解, 更可能产生误解, 直接导致交际困难或交际失败 (陈桦 2006a, 2006b, 2006c)。

语音研究难度系数大, 费时费力。在录音、标注、提取、分析整个流程中涉及到计算机技术、声学知识以及语言学知识 (王建新 2005)。在第二语言习得的研究领域, 语音习得的研究在国际上已经具有较为成熟的研究模式: 实证性方法、仪器性手段、定量结果 (王韫佳 李吉梅 2001)。目前, 我国英语学习者的语音研究逐渐增多, 但大多数研究仍然采用非常原始、笨拙的方法; 受试者人数偏少、标注不规范、人工计数代替批量提取, 都会对研究结果的客观性和有效性产生影响, 质量不能令人满意; 还没有英语作为二语的学习者语音库, 更没有进行韵律标注的二语语音库, 这在一定程度上阻碍了二语口语研究的进程 (杨军 2005)。

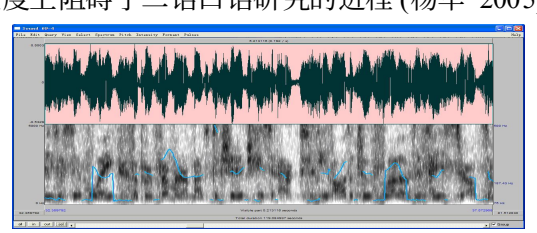

图 1　国内现有学习者口语语料库的录音质量

鉴于上述原因, 有必要创建中国英语学习者语音库, 为后续大规模语音研究搭建平台。

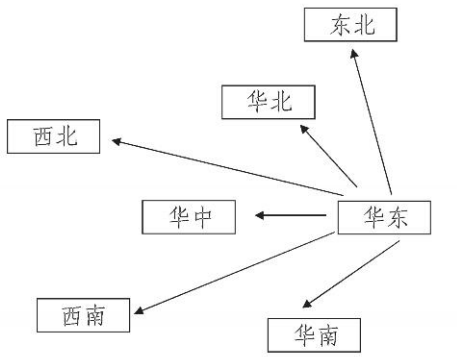

图 2　录音人的地域分布

2　语音库的相关变量

2 1受试者

选择发音人时, 两个最重要的参考因素是母语特点和目的语水平 (王韫佳 李吉梅 2001)。因此, 在创建语音库时, 录音人的确定主要考虑下面几个因素: 地域、方言以及受教育层次。

从地域分布上看, 语音库的录音人来自全国 7个大的自然地理区域 (见图 2)。从方言划分上, 根据中国社会科学院方言地图划分 (李荣 熊正辉 张振兴 1988), 语音库的录音人覆盖除少数民族语言以外的 10大方言: 北方方言 (主要分布在东北、华北和西北地区)、吴方言 (主要分布在华东地区)、湘方言 (主要分布在华中地区)、赣方言 (主要分布在华中地区)、客家话 (主要分布在华南地区)、粤方言 (主要分布在华南地区)、闽方言 (主要分布在华南地区)、晋语 (主要分布在华北)、平话 (主要分布在华南)和徽方言 (主要分布在华东地区)。

从受教育层次上看, 语音库包括我国正规英语教育的 4个层次: 初中生、高中生、英语专业大学学生以及英语专业硕士生。每两个层次间隔约 3年。选择录音人时, 充分考虑男女比例及学习者英语水平等因素。

2 2任务

录音时, 4个受教育层次的学习者应邀请朗读若干组对话 (中学生 16组对话、大学生和研究生朗读 10组对话); 大学生另外还须完成 2分钟

左右的自主对话任务。学习者从 14个题目中随机抽取一个, 准备 10分钟, 两人一组展开对话并且录音。

朗读语料浅显易懂, 内容是录音人十分熟悉的日常会话, 难度符合初中学生的平均英语水平; 对话包含各种基本句型和丰富的韵律考查点 (详见表 1)。录音前, 登记录音人的姓名、性别、籍贯或中学受教育地点、是否作过教师 (针对研究生)、代码等信息。

录音在语言实验室内进行; 录音软件为 Cool Edit Pro 2 1; 录音采样率为 16000(16kHz, 16 bit mono PCM) (祖漪清 1998)。录音时, 室内仅录音人及录音操作者, 由于严格控制噪音, 可以确保录音效果 (见图 3)。

特　征	实　例
简单陈述句	Something wrong with my computer
宾语从句	I've heard that that film...
定语从句	... the shop where you bought it
选择问句	Do you want to..., or have you..
一般疑问句	Will they make the man sick?
特殊疑问句	What do you want to buy?
附加疑问句	..., didn't they?
祈使句	Close the window
列举句式	spring, summer, autumn and winter
句末报告短语	/...0, Betty asked
句末呼语	..., Mummy
时间 /地点	... last night ... over there
对比特征	... clever... foolish..
强调特征	Mary does like swimming
否定表达	... don't
复合名词	New York
名词短语	a famous tragedy

表 1　朗读语料中涵盖的句式及韵律考查点

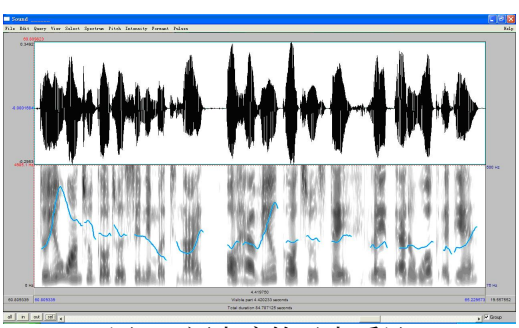

图 3　语音库的录音质量

对录制好的声音文件, 根据地域分布、方言区、受教育层次、任务类型等进行分类。

利用软件 Praat (http: //www. praat. org), 对朗读语料进行多层语音标注。

语音库朗读部分的标注结合英国 (调冠 - 调头 - 调核 - 调尾)和美国 (ToBI)两大标注系统。标注层级为 6层 (见图 4)。

3　语音标注

3. 1音段标注

第一、二层为语音库的音段标注层, 对应于 ToBI中的正则层 (orthographic tier)。第一层描述学习者的实际发音, 以英语单词为单位体现; 第二层为发音的标准层, 以音节为单位体现。

音段层的标注代码以 A. C. Gimson所创立的 de facto standard为基准, 借助 Praat自带的键盘符号实现。

在标准层音节划分的判断上, 参照现任国际语音学会主席 J. C. Wells(2005)编著的 5朗文英语发音词典 6以及 Wells(1990)教授提出的 5大音节切分原则。音位标注均采用强势发音, 不考虑语境对发音强弱的影响。

空白段位于每两组对话间, 标为 sil; 位于对话内部的空白段标为 pau; 单词起首爆破音的成阻和持阻段仍然标为 pau, 但由于时长不足 100毫秒, 统计时可以不计入结果。

3. 2超音段标注

语音库标注的第三层等同于 ToBI系统的间断指数层 (break index tier)。其中, 4代表语调短语边界 (intonation phrase boundary); 3代表中间短语边界 (intermediate phrase tier); 1代表韵律词边界 (prosodic word boundary); 0代表粘着语素 (clitic group)。ToBI系统中边界指数层的 2表示音调标记的边界和声学边界停顿的不确定性。由于本朗读语音库主要为英语教学和研究服务, 并非进行非常细致的声学标注, 所以不标出。

第四层主要标识句中的重读音节, H^* 表示位于发音人高域的重读音节, L^* 表示位于发音人低域的重读音节。对于一个语调单位内部最突显音节 (焦点所在), 标以 H^{**}。

第五、六两层对应于 ToBI系统中的语调层 (tone tier)。第五层为英式语调模式层。英式传

统将英语语调描述成调冠 - 调头 - 调核 - 调尾连续体, 调核音调包括调核至调尾的部分 (见表 2)。英式语调模式共有 2种调冠、7种调头、7种调核音调 (陈桦 2006d)。

调核前段 Pre2tonic segment		调核 nucleus
调冠 pre2head	调头 head	
A	dog is a man. s best	FRIEND

表 2　英式语调模式 (Tench 1996: 12)

第六层为美式的 ToBI层。美式语调模式非常重视对每一个音高事件的描述; 同时, 还关注各单位边界的音高走势, 即语调短语边界和中间短语边界。美式语调模式中对音高事件的描述类型分为 7种; 对边界调的描述主要划分为高、低两种模式。

美式	英式	本语音库代码
H* L L%	High Fall	< NTHF>
H* L H%	Fal2Rise	< NTFR>
H* H H%	High Rise	< NTHR>
L* L L%	Low Fall	< NTLF>
L* H H%	Low Rise	< NTLR>
L+ H* L L%	Rise2Fall	< NTRF>
L+ H* H H%	High Rise (with low head)	<HL> < NTHR>
L* + H L L%	Rise2Fall (emphatic)	< sliding> < NTRF>
H+ L* L L%	Low Fall (with high head)	<HH> < NTLF>
H+ L* L H%	Fal2Rise (with high head)	<HH > < NTFR>
H* + L H L%	Fal2Rise (/ calling contour0)	H% <HH > < NTFR>

表 3　英、美语调模式对照

英式语调模式注重对调群连续体的描述, 而美式语调模式特别关注对音高事件和边界调的描述, 因此在本语音库中, 两种语调模式的标注虽然在大多数情况下具有很高的一致性, 但并不具有绝对的互推性 (见表 3)。例如, 在一个调群中, 调核自高域向低域的调尾发生变化, 英式传统将这一语调模式描述成 < NTHF > (高降调型); 但由于重读音节的元音部分有自低向高的走势, 因此美式传统将其描述成 L+ H* L L%; 从标识上看, 似乎应该属于升降调。这时, 就存在着描述的不吻合性。

4　意义

作为一种专用语料库, 学习者语音库除了对二语习得的应用语言学研究具有重要意义以外, 对外语教学也具有同样重要的实际意义。现在概述如下。

1)创建学习者英语语音库, 同时增加基础语音标注, 可以与国内现有的学习者口语库优势互补。目前, 口语库的文本标注层面和切入点有限, 且完全依赖口语文本进行研究, 研究者无法如语音标注般直接根据声学频谱等指征对标注结果进行验证。因此, 建立学习者语音库的意义在于为学习者语音研究搭建平台, 让更多的研究者和学习者了解学习者英语的特点, 并使其与本族语者口语库 (如 LONDON2LUND 库)的对比研究和学习成为可能 (何安平 2004)。

2)语音语料库与一般纯文本口语语料库的区别在于: 一般的语料库以文本形式与用户见面, 而语音库以文本 (标注文本)、声音和声学参数三种形式与用户见面 (刘岩 2006)。永久的录音可以证实语言事实, 使没有参加田野工作的研究者能够相信并验证描述结果的正确性。同时, 可以依据声学参数对语音进行量化分析, 如比较语言内部或语言间某一成分的时长差异。

3)借助实验语音学研究方法, 基于语音库的研究不仅能够对学习者的英语语音特征和语音发展进行全面、系统描述和对比分析, 从中揭示中国英语口语教学中的薄弱环节; 问题的发现, 有助于英语教师更加重视和加强相关方面的教学, 帮助学生克服这些学习难点, 对提高我国英语教学的有效性有指导和借鉴作用 (何安平 2004)。同时, 可以发现学习者某一阶段的个体特征和共性 (卫乃兴 2004), 从而为我国英语教学大纲的设计、英语教材的开发与编写、英语口语测试标准的完善与细化等提供一定依据, 使其更加适应学习者的需求 (杨惠中 2002; 杨惠中 桂诗春 2005)。

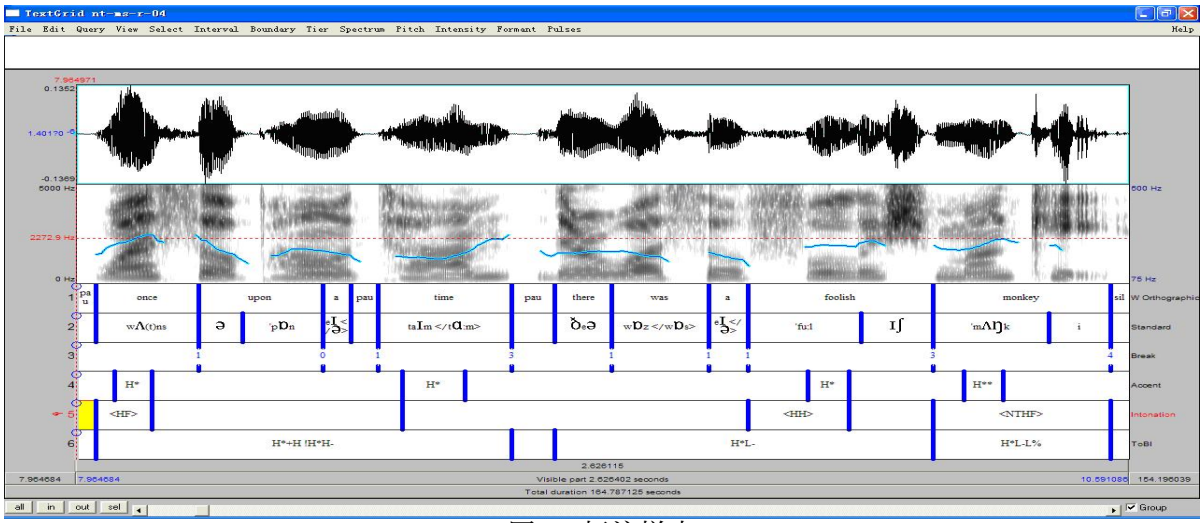

图 4　标注样本

参考文献

陈　桦. 中国学生朗读口语中的调群切分模式研究 [J]. 外语教学与研究, 2006a (5).

陈　桦. 英语学习者朗读口语中的调核位置 [J]. 解放军外语学院学报, 2006b (6).

陈　桦. 中国学生朗读口语中的英语调型特点研究 [J]. 现代外语, 2006c (4).

陈　桦. 英语语调模式及其声学实现 [J]. 外语研究, 2006d (5).

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.

李爱军. 口语对话语音语料库 CADCC 和其语音研究 [A]. 第 5届全国语音学会论文集 [C]. 2001.

李荣熊 正　辉 张振兴. 中国语言地图集 [M]. 香港: 朗文出版社, 1988

李文中. 语料库、学习者语料库与外语教学 [J]. 外语界, 1999(1).

刘　岩. 关于中国少数民族濒危语言语音语料库的设计 [J]. 中央民族大学学报, 2006(4).

王建新. 计算机语料库的建设与应用 [M]. 北京: 清华大学出版社, 2005

王立非 孙晓坤. 国内外英语学习者语料库的发展: 现状与方法 [J]. 外语电化教学, 2005(105).

王韫佳 李吉梅. 建立汉语中介语语音语料库的基本设想 [J]. 世界汉语教学, 2001(1).

卫乃兴. 中国学习者英语口语语料库初始研究 [J]. 现代外语, 2004(2).

文秋芳 王立非 梁茂成. 中国学生英语口笔语语料库 [M]. 北京: 外语教学与研究出版社, 2005

杨惠中. 语料库语言学导论 [M]. 上海: 上海外语教育出版社, 2002

杨惠中 桂诗春等. 基于 CLEC 语料库的中国学习者英语分析 [M]. 上海: 上海外语教育出版社, 2005

袁家骅. 汉语方言概要 [M]. 北京: 语文出版社, 2004.

祖漪清. 实现语音数据库科学性的重要环节 [J]. 语言文字应用, 1998(1).

Armstrong S. Using Large Corpora [M]. London: A. Bradford Book, MIT Press, 1993.

Biber, D., Conrad, S. & R. Reppen. Corpus-based Approaches to Issues in Applied Linguistics [J]. Applied Linguistics, 1998(2).

Chafe, W. The Importance of Corpus Linguistics to Understanding the Nature of Language [A]. In Svartvik J. (ed.) Directions in Corpus Linguistics [C]. Berlin / New York: Mouton de Gruyter, 1992.

Granger, S. Learner English on Computer [M]. London / New York: Longman, 1998.

Johansson, S. Computer Corpora in English Language Research [M]. Bergen: Norwegian Computer Center for the Humanities, 1982.

Tench, P. The Intonation Systems of English [M]. Cassell, 1996.

Wells, J. C. Studies in the Pronunciation of English [M]. London & New York: Routledge, 1990.

Wells, J. C. Longman English Pronunciation Dictionary [M]. London: Longman, 2005.

收稿日期: 2009-03-13　　=责任编辑　李洪儒>

99