汉语中介语语料库建设与应用研究 第一辑

张宝林　靳继君　胡楚欣◎主　编

国家社会科学基金项目资助

中国书籍出版社
China Book Press

图书在版编目（CIP）数据

汉语中介语语料库建设与应用研究. 第一辑 / 张宝林, 靳继君, 胡楚欣主编. --北京：中国书籍出版社, 2021.8
ISBN 978-7-5068-8660-4

Ⅰ. ①汉… Ⅱ. ①张… ②靳… ③胡… Ⅲ. ①汉语—中介语—语料库—研究 Ⅳ. ①H1

中国版本图书馆CIP数据核字（2021）第177911号

汉语中介语语料库建设与应用研究. 第一辑

张宝林　靳继君　胡楚欣　主编

责任编辑　王志刚
责任印制　孙马飞　马　芝
封面设计　中尚图
出版发行　中国书籍出版社
地　　址　北京市丰台区三路居路 97 号（邮编：100073）
电　　话　（010）52257143（总编室）（010）52257140（发行部）
电子邮箱　eo@chinabp.com.cn
经　　销　全国新华书店
印　　刷　天津中印联印务有限公司
开　　本　710 毫米 × 1000 毫米　1/16
字　　数　413千字
印　　张　26.5
版　　次　2021 年 8 月第 1 版　2021年 8 月第 1 次印刷
书　　号　ISBN 978-7-5068-8660-4
定　　价　89.00 元

目录

国内外主要英语口语学习者语料库概述 *

许家金　董通

内容提要　本文综述了国内外代表性英语口语学习者语料库的建设与研究现状，其中涉及LINDSEI语料库、TLC语料库、COLSEC语料库、SWECCL 1.0/2.0语料库、PACCEL语料库等。文中介绍了相关语料库的建库方案和加工规范。概言之，英语口语学习者语料库主要仍以转写文本库为典型形态，少有录音文本同步库、语音标注库或多模态视频库。相关研究主要以词汇、短语、句型分析为主，针对语音或其他口语特征的研究不足。今后英语口语学习者语料库研究可多关注口语中介语的动态发展，更加关注语音、互动、语用等方面的中介语特征。

关键词　英语口语学习者语料库；建库方案；加工规范；中介语

一、引言

学习者语料库（learner corpus）主要收集大学生或中学生的非母语写作或口语产出，是用以分析中介语的专用数据集。本文主要关注英语口语学习者语料库。因采集、转写、标注难度较大，与写作语料库相比，英语口语学习者语料库的建设与研究相对滞后。Sinclair（1991：16）认为，“从我个人经验来看，即兴会话是无可替代的”（In my own experience, there is no substitute for impromptu speech）。英语学习者的口语产出也是最能反映口语中介语的研究材料。缺少口语中介语的相关研究，学习者语料库研究对英语教学的指导意义也必然大受影响。以下本文拟对英语口语学习者语料库的建设与研究作简要梳理。

*　基金项目：本文系北京市社科基金项目“语料库语言学史”（20YYB013）的阶段性成果。

二、英语口语学习者语料库建设现状

口语语料库建设的总体进展一直不及书面语语料库，英语口语学习者语料库落后于书面语学习者料库的情况也同样存在。即便如此，自20世纪90年代开始，国内外也陆续建成若干极具影响力的英语口语学习者语料库。例如，LINDSEI（Louvain International Database of Spoken English Interlanguage）语料库（Gilquin、De Cock、Granger，2010）、TLC（Trinity Lancaster Corpus）语料库（Gablasova、Brezina、McEnery，2019）、COLSEC（College English Learners' Spoken English Corpus）语料库（杨惠中、卫乃兴，2005）、SWECCL（Spoken and Written English Corpus of Chinese Learners）1.0/2.0语料库（文秋芳、王立非、梁茂成，2005；文秋芳、梁茂成、晏小琴，2008）等。这些语料库的推出，至今仍是英语口语中介语研究领域的关键资源。

2.1 英语口语学习者语料库建设的总体情况

截至2021年3月，比利时鲁汶天主教大学“英语语料库语言学中心”（Center for English Corpus Linguistics, CECL）世界学习者语料一览网页上列有185个学习者语料库，其中有61个口语库，占比约为32%。这一数据直观说明了英语口语学习者语料库数量远不及书面语的现状。我国英语口语学习者语料库的情形大抵相同。这其中的原因是多方面的，有对口语的认识局限，认为口语是不规范的、芜杂的。当然，最为重要的原因是口语语料库的建设难度大，特别是转写工作的费时耗力。这也从另一角度突显了现有这些英语口语学习者语料库的价值。

2.2 知名英语口语学习者语料库

2.2.1 LINDSEI语料库

LINDSEI语料库由比利时鲁汶天主教大学Sylviane Granger主持建设。该语料库是ICLE（International Corpus of Learner English）作文语料库的姊妹库。LINDSEI收录的是不同母语背景学习者的英语口语语料。Granger团队首先开

发了母语背景为法语的英语学习者的口语库，之后在与国际同人合作下继续建设，最终涵盖了保加利亚、中国、荷兰、法国、德国、希腊、意大利、日本、波兰、西班牙、瑞典共11个国家的学习者口语语料。为便于对比，Granger团队还建设了名为LOCNEC的语料库，其中包含与LINDSEI设计相同的英语本族语者口语语料。

LINDSEI共有130小时录音，库容逾100万词，每个子库约有10万词。其中学习者英语口语约有80万词左右，其余为与之对话的教师语料。所含的11个子库取样方案一致，即包含50组口语任务，每组任务包含3个子任务。它们分别是：（1）口头作文；（2）教师与学生的对话；（3）看图说话。转写文本都被赋予了23项元信息属性，包括性别、年龄、母语、在校英语学习时长，这些是研究学习者英语口语的重要外部变量。LINDSEI语料库以CD-ROM形式存储并发售，随盘附赠专用的检索和分析工具。该检索工具可按三个子任务分别检索，并能限定只检索学生语料或教师语料。相关统计数据可按23个变量分别呈现。

2.2.2 TLC语料库

TLC语料库是由英国兰卡斯特大学“社会科学语料库研究中心”（Center for Corpus Approaches to Social Science，简称CASS）团队主持建设。其核心成员包括Dana Gablasova、Vaclav Brezina与Tony McEnery等。该库语料全部来自伦敦圣三一学院（Trinity College London）的英语口语等级考试（Graded Examinations in Speakers of Other Languages，简称GESE）。TLC语料库的规模为420万词次，是当前世界范围内规模最大的英语口语学习者语料库。语料的时间跨度为2012—2018年。该语料库为2000名考生与考官的口头对话，共涉及2~4个任务。考生的母语有35个之多。考生年龄从9~72岁不等。

TLC语料库提供了免费的在线检索平台。该网络平台可根据母语背景、性别、年龄等筛选检索结果，并能将检索所得频数数据可视化。

2.2.3 COLSEC语料库

COLSEC语料库是国内首个英语学习者口语语料库，由杨惠中、卫乃兴等统筹建成。该库收集了全国大学英语四、六级考试（CET-SET）中口语部分的内容，转录了约70万词，涉及“考官—学生问答”和“学生—学生讨论”。

该库根据考生的地区、专业、考试成绩、话题等按比例随机抽取语料。语料库标注了话轮、发音错误、停顿、犹豫、打断、非言语交际等信息。图1为COLSEC语料库的文本片段。其中的<interlocutor>…</interlocutor>框定的是考官话语，<sp1>…</sp1>框定的则是考生1的讲话内容。语误则以［ ］表示，如其中的［Wd-t］、［Pu-r］、［Pl-r］、［Ws-s］、［M2t］分别表示考生将/d/音误发成了/t/音、字母“u”之后加了卷舌音、字母“l”音之后加了卷舌音、/z/音误发成了/s/音以及第二个字母t吞音。

<interlocutor>Ok. Now would you please briefly introduce yourselves to each other? And don't en remember, you should not mention the name of your university, Ok? </interlocutor>↓
<sp1> Yeah. From me? <sp1>↓
<interlocutor> Yeah. </interlocutor>↓
<sp1> Good morning, everyone. My name is ***, and [Wd-t] my major is International Economics. And now I'm a junior student. I have learned [Wd-t] English about eight years and I am interest in English. I think in my view English is not only a tool for us [Pu-r] to communicate with the outside world, but also [Pl-r] provides us a a kinds of culture which is totally different with ours [Ws-s]. I interested [M2t] in English. </sp1>↓

图1　COLSEC语料库文本片段

COLSEC语料库与CLEC（Chinese Learner English Corpus）中国学生英语语料库的团队成员有很大交集。两个语料库的共同之处是都比较多地依赖全国性的大规模英语考试，都比较关注错误标注。

2.2.4　SWECCL1.0/2.0语料库

SWECCL语料库1.0版由文秋芳、王立非、梁茂成等学者在南京大学建成。SWECCL2.0由文秋芳、梁茂成、晏小琴等在北京外国语大学建设。SWECCL2.0为全新语料库，并非SWECCL1.0修正或简单升级。两个版本的SWECCL都分为书面语（WECCL）与口语（SECCL, Spoken English Corpus of Chinese Learners）两个子库。本节重点介绍SECCL子库。

SECCL1.0的语料取自全国英语专业四级口语考试。考生为英语专业二年级学生。该语料库转写了1 141位考生11 410多分钟的录音，语料产生的时间从

1996年到2002年共7年，SECCL1.0语料库的库容约为100万词。任务类型包括3类，分别是复述故事、口头作文与考生对谈。复述故事即听300词左右的故事两遍后进行复述；口头作文则是围绕给定的话题，准备3分钟，用英语讲述3分钟；考生对谈部分要求考生根据给定的议题交换意见或辩论，准备3分钟，之后交谈4分钟。

SECCL1.0在每个文本的开头提供了8个方面的元信息（如图2），<SPOKEN>指口语，<TEM4>指英语专业四级考试，<GRADE2>指二年级，<YEAR00>指2000年，<GROUP065>指第65组，<TASKTYPE1>指口试的任务类型1，<SEXM>指考生性别为男，<RANK27>指该考生在小组排名第27。除此之外，自我重复、非流利性停顿、发音错误、语法错误等也都有所标注，比如图2中“…”意为停顿，<in>、<cross>等是语误。

<SPOKEN> <TEM4> <GRADE 2> <YEAR00> <GROUP065>
<TASKTYPE 1> <SEXM> <RANK27>

Task 1

Mr. Hall was a rich business…man. He lived <in> a very big house near a river. The river is frozen in the winter. Err…so by Chri…Christmas Day, the river always covered…eh…very thick ice. Some one even crossed <cross> the river on foot, and some brave men even could <can> cross the river by motorcars. On the sight of…eh…the…scenery, Mr. Hall had <have> an idea. He wanted <want> to hold a party on Christmas Day on ice. En …he put…err…all his furniture <furnitures>, and carpets <capitries> and a beautiful ice to hang over the river and…sent out all the invitations to…the… his…important friends. On Christmas Day, his friends…

图2　SECCL1.0某语料库文本片段

作为更新版，SECCL2.0中选取的是本科扩招后2003年到2007年间的英语口语语料，另外还增加了全国英语专业八级口语考试的内容，因此任务类型更为丰富，增加了根据指定话题进行评论的内容。

2.2.5　PACCEL语料库

PACCEL（Parallel Corpus of Chinese EFL Learners）语料库是由北京外国语大学文秋芳、王金铨等学者创建，分为口译平行语料库（PACCEL-S）和笔译

平行语料库（PACCEL-W）两个子库。其中的PACCEL-S子库收录的是英语专业学生的口译语料。

PACCEL-S语料来源为全国英语专业八级口语考试的任务1英译汉与任务2汉译英两部分，库容为496 177词次，包含汉译英的11 425词与英译汉的318 752字，时间跨度为2003年到2007年。相关口译语料进行了句级对齐。在标注上，PACCEL-S借鉴了SECCL的语误标注方案。

2.2.6 ESCCL语料库

ESCCL（English Speech Corpus of Chinese Learners）语料库由南京大学陈桦、北京外国语大学文秋芳、中国社会科学院李爱军团队创建。该语料库不同于前文综述的转写文本形式的英语口语学习者语料库。严格来说，ESCCL语料库应称为英语中介语语音库。该库利用Praat软件对语料进行了多层语音标注。Praat标注中的第一层以英语单词形式出现，又称正则层（orthographic layer）；第二层为发音的标准层，以音节为单位体现；第三层为间断指数层；第四层主要标注重读音节；第五层为英式语调模式层；第六层则为美式语调模式层。

ESCCL的语料主要是来自初中生、高中生、英语专业本科生以及硕士生的英语朗读以及自主对话任务。语料收集时根据学生的生源地、方言区、受教育层次、任务类型等进行分类。

2.3 小结

综上所述，在语料来源上，英语口语学习者语料库既有考试场景下的英语口语数据，也有非考试场景下的命题口语表述，以及课堂英语交流等。任务类型则涵盖朗读、复述、口头叙述、看图说话、口头作文、口译等形式，以及考官–考生会话、学生–学生两人会话与多人讨论等互动形式。我国多数英语口语语料库来自各类标准化的考试，较少收集学习者的日常英语口语产出。这无疑减少了语料收集的工作量，保证了语料库的规模，还可以增加历时的语料，并提供口语产出者的详细信息，为相关研究带来便利。但是，这样的语料也有明显的缺陷，由于场景与任务类型的限制，无法收集到学习者自然的日常口语产出。

在语料库的形态上，根据有无原始录音可以分为两大类，一类是转写文本库，也被称为无声口语库（Ballier & Martin 2015：111），依赖于成熟的书面语语料库技术，该类型语料库可用来开展词汇、短语、句型、语言点等的分析。另一类为语音库，又可进一步细分为单纯的录音加文本库（如SECCL语料库）、录音文本同步库（time-aligned）、多模态视频库（multimodal）以及语音标注库（phonetically-annotated）几小类。

在转写技术上，此前的英语口语学习者语料库基本采用的都是人工转写。随着语音识别技术的进步，相信今后的中介语口语库可以实现较大程度的自动转写。在现阶段我们主张可以采用念转法（林纾式转写）（许家金，2020a：16），即在录音质量或者学习者口语差以至无法被语音识别软件识别时，由转写人员将所听到的学习者的口语念出来，从而让语音识别软件去识别转写，提高转写效率，再经进一步的人工校对，以确保转写语料的质量。

在语料的标注上，主要涉及学生个人信息（性别、年级、学英语的时长、国别等）、声学标注、发音错误标注、词汇语法错误标注、词性标注等。标注的格式主要有分离式与XML两种，前者顾名思义，即语料与标注单独存放，不在同一个文本中，而XML则一般放在文本的开头，作为可扩展的标记语言，格式要求很严格，以尖括号<>表示，且需要前后围堵起来。至于偏误标注，原则是能不标则不标，应由学者根据自己的研究目的决定标什么和怎么标。另外，口语语料中有很多停顿、重复等不流利的现象，因此也不能简单地将笔语标注的方案套用在口语语料标注上。

三、英语口语学习者语料库相关研究

国内外基于英语口语学习者语料库的研究有多个切入点。例如学习者英语口语中的回应语、口语程式语、高频口语词汇、口语流利性研究、口语重复研究、口语副词等。除了这里提及的口语特色词汇短语研究外，语音语调错误以及话语语用研究等尤其值得关注。对于不同母语背景、不同语言水平、各种任务类型等对英语学习者在口语产出上的影响，可以拓宽我们的分析维度。最后，还可开展追踪式口语语料的收集，从而进行中介语的动态发展研究。

然而，总体来说，英语口语学习者语料库研究存在简单移植和套用书面语研究模式的问题，应努力摆脱书面语句子切分的做法，根据学习者语言实际，识别和划分英语口语中介语的分析单位。另外，研究选题还应更加关注语音、互动、语用等方面。

四、下一代英语学习者口语语料库的建设

在大数据时代，面对海量的数据，口语语料库的建设也呈现出了新的模式，具体表现为：文字转写整合语音识别技术与人工校对，语料规模和建库效率大幅提高；依托覆盖全国的在线学习平台，语料取样更平衡多样，更具代表性；语料规模将持续扩增，可以动态监控中国学习者英语口语表现。在与语言教学的深度融合下，我国应能建设规模更大、质量更高、利用率更高的学习者口语文本库、语音和文本时间轴同步双模态语料库、含语音错误标记的语料库以及在线部署可供检索的语料库。类型丰富，代表着英语口语学习者语料库建设的未来方向，对于提高我国英语口语学习者语料库在国际上的影响力具有重要意义。就个人研究者而言，应鼓励根据特定研究目的，设计建设10万词至100万词规模的“小而精”英语口语学习者语料库（参见许家金，2020a中有关“梨子故事语料库”的设计）。在研究方法上，多模态、多变量统计分析应是未来趋势（许家金，2020b）。

参考文献

陈桦、文秋芳、李爱军（2010）语音研究的新平台：中国英语学习者语音数据库,《外语学刊》第1期。

王立非、孙晓坤（2005）国内外英语学习者语料库的发展：现状与方法,《外语电化教学》第5期。

文秋芳、王立非、梁茂成（2005）《中国学生英语口笔语语料库（1.0 版）》，北京：外语教学与研究出版社。

文秋芳、王金铨（2008）《中国大学生英汉汉英口笔译语料库》，北京：外语教学与研

究出版社。

文秋芳、梁茂成、晏小琴（2008）《中国学生英语口笔语语料库（2.0 版）》，北京：外语教学与研究出版社。

许家金（2020a），《语料库与中国学习者英语口语研究》，北京：外语教学与研究出版社。

许家金（2020b），多因素语境共选：语料库语言学新进展，《外语与外语教学》第3期。

杨惠中、卫乃兴（2005）《中国学习者英语口语料库建设与研究》，上海：上海外语教育出版社。

Ballier, N. & P. Martin (2015) Speech annotation of learner corpora. In Granger S., F. Meunier & G. Gilquin (eds). *The Cambridge Handbook of Learner Corpus Research*, 107-134. Cambridge: CUP.

Gablasova, D., V. Brezina & T. McEnery (2019) The Trinity Lancaster Corpus: Development, Description and Application. *International Journal of Learner Corpus Research*, 2:126-158.

Gilquin, G., S. De Cock & S. Granger (2010) *Louvain International Database of Spoken English Interlanguage*. Louvain-la-Neuve, Belgium: Presses Universitaires de Louvain.

Sinclair, J. (1991) *Corpus Concordance Collocation.* Oxford: Oxford University Press.

作者简介

许家金，博士，教授，博士生导师。现任职于北京外国语大学中国外语与教育研究中心。研究方向：话语研究、二语习得、语言对比与翻译、语料库语言学。

董通，北京外国语大学中国外语与教育研究中心博士研究生，曲阜师范大学外国语学院讲师。研究方向：语料库语言学。