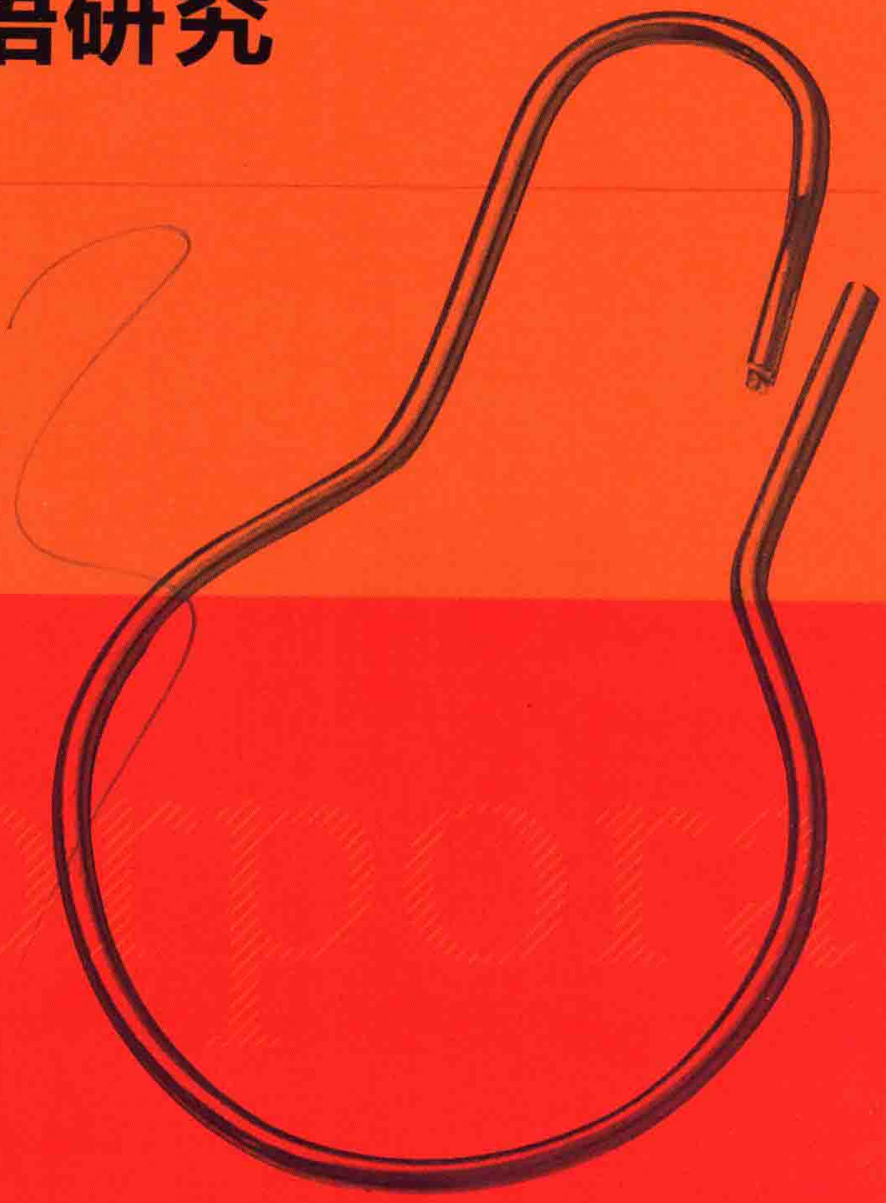
外研社语料库研究系列

Corpora and Chinese Learners' Spoken English

语料库与中国学习者英语口语研究

许家金 著



外语教学与研究出版社
FOREIGN LANGUAGE TEACHING AND RESEARCH PRESS

外研社语料库研究系列

Corpora and Chinese Learners' Spoken English

语料库与中国学习者英语口语研究

许家金 著

外语教学与研究出版社
FOREIGN LANGUAGE TEACHING AND RESEARCH PRESS
北京 BEIJING

图书在版编目（CIP）数据

语料库与中国学习者英语口语研究 / 许家金著. -- 北京 ：外语教学与研究出版社，2020.6
（外研社语料库研究系列）
ISBN 978-7-5213-1874-6

Ⅰ. ①语… Ⅱ. ①许… Ⅲ. ①语料库－应用－英语－口语－教学研究－中国
Ⅳ. ①H319.9

中国版本图书馆 CIP 数据核字（2020）第 108308 号

出 版 人　徐建忠
项目负责　李晓雨
责任编辑　解碧琰
责任校对　段长城
封面设计　袁　凌　吴德胜
出版发行　外语教学与研究出版社
社　　址　北京市西三环北路 19 号（100089）
网　　址　http://www.fltrp.com
印　　刷　北京虎彩文化传播有限公司
开　　本　650×980　1/16
印　　张　13
版　　次　2020 年 7 月第 1 版　2020 年 7 月第 1 次印刷
书　　号　ISBN 978-7-5213-1874-6
定　　价　49.90 元

购书咨询：（010）88819926　电子邮箱：club@fltrp.com
外研书店：https://waiyants.tmall.com
凡印刷、装订质量问题，请联系我社印制部
联系电话：（010）61207896　电子邮箱：zhijian@fltrp.com
凡侵权、盗版书籍线索，请联系我社法律事务部
举报电话：（010）88817519　电子邮箱：banquan@fltrp.com
物料号：318740001

记载人类文明
沟通世界文化
www.fltrp.com

教育部人文社会科学重点研究基地重大项目子课题
“大数据视野下的外语及外语学习研究”
（17JJD740003）成果

序

本书主要是本人十年左右学习者语料库研究的一个小结。我自博士研究开始即以口头话语为主要研究语料。我有关中国学习者英语中介语的实证研究十之八九也都聚焦于口语，而口语数据较难收集，基于口语的语言研究始终未得到足够重视。希望本书能成为学习者口语语料库研究的铺路石。

书中研究案例有以下特点：1）较多关注话语层面的语言现象。例如，第三章的互动话语词块研究、第九章的人物指称研究、第十章的话语评价研究、第十一章的英语情境惯用语研究。此类研究选题中，相关的语言特征常常无法通过语料库软件自动提取，因此，需要经过大量手工标注后才可进行量化统计分析；2）关注动词相关的构式习得。例如，第四章的词体研究、第五章的及物性研究、第六章的述补行为研究、第七章的动词方位构式研究、第八章的复合运动事件表达研究；3）研究方法较多采用综合对比的方法，即将中介语对比分析法与英汉对比相结合，尝试考察中介语中可能存在的母语影响。部分研究中，我们还对比了学习者口语语料库和书面语语料库，以探究中国学习者在二语习得中可能存在的语体差异；4）较多将叙事性口语纳入学习者语料库研究；5）注重将语言学理论与语料分析相结合。例如，第五章和第十章借用了系统功能语言学中的及物性分析和评价理论。第七章和第八章借用了论元结构分析和认知语言学中的图形－背景理论。概言之，本书研究案例在相关语言学理论的指导下，通过对比分析法，考察了学习者英语口语中的话语层面和动词习得情况。

本书涉及的研究选题仍很有限，尚有很多重要的语料库语言学和二语

习得理论和方法未纳入讨论之中。希望在今后的研究中，我能与学界同仁携手，对中国学习者口语语言特征作更多深入探讨。

我国学习者语料库研究至今已有二十多年，本领域的发展似乎进入平台期。希望语料库研究同行能协力在新的学习者语料库的创建、新的学习者语料库研究方法、新的研究选题以及新的研究理念挖掘方面能有所突破，使学习者语料库的相关研究在我国进入新一轮快速发展期。

书中相关实证研究的完成离不开各位合作者的全力支持。本书除第一章、第二章、第十章和第十二章外，其余章节均为合作成果。其中第三至第九章分别与许宗瑞、欧群超、张明芳、陈哲、刘洁琳、刘洋、刘霞合作完成，第十一章与赵珺合作完成。在相关研究的酝酿、实施阶段，我得到了北京外国语大学北京高校高精尖学科“‘外语教育学’建设项目”的支持。我还特别受到北京外国语大学中国外语与教育研究中心的诸位前辈和同事的点拨和指导。此外，在数据标注过程中还得到过龙满英的帮助。运用于多项研究的数据标注工具 BFSU Qualitative Coder 软件由贾云龙负责开发。没有上述各位的鼎力相助，相关课题难以顺利完成。张懂同学在本书定稿之前，花费大量心力帮忙校改文字。在此，对给予我帮助的所有合作者致以诚挚的谢意。

本书中所举案例，过半数是基于“中国英语学习者梨子故事综合语料库”开展的研究。该语料库有幸在本书定稿之前最终建成。在此，我也要向所有协助收集语料的英语老师表达由衷的感谢。最后要特别感谢 1,000 多位大学生认真完成语料产出任务，并完成初步转写工作。希望我们基于各位同学的英语产出的相关研究，能为中国学习者英语口语的提高指明一些改进的方向，更希望“梨子故事”能成为更多中国学习者英语口语研究的基础资料。

在“中国英语学习者梨子故事综合语料库”建设过程中，我夫人帮忙联系了十多所高校的老师，收集了近千个学生的梨子故事样本。她无疑是梨子故事语料库项目的首功之臣。在课题开展及书稿撰写过程中，她及年迈的岳父岳母承担了所有的家庭事务，特别是悉心照顾幼子，使我能安心科研与写作。对其无私付出，我的感激之情无以言表。希望将拙作献给他们。

许家金
北京外国语大学中国外语与教育研究中心
人工智能与人类语言重点实验室

目 录

序 I

第一章 绪论 1

第二章 学习者英语口语语料库研究方法 7
2.1 引言 7
2.2 “梨子故事语料库”的创建 8
2.2.1 梨子故事 8
2.2.2 梨子故事叙事语料的采集过程 10
2.2.3 语料文本的转写与标注 13
2.3 本研究中学习者英语口语中介语研究方法 18

第三章 英语口语中的互动话语词块研究 21
3.1 研究背景 21
3.2 研究设计 23
3.2.1 研究方法 23
3.2.2 研究问题 23
3.2.3 语料及工具 23
3.2.4 词块的界定 24
3.2.5 互动词块 25
3.3 数据分析及讨论 26
3.3.1 互动词块的类型和数量 26
3.3.2 中国大学生互动词块的形式特点分析 27
3.3.3 中国大学生互动词块的语用功能表现 29
3.4 小结 32

第四章 英语口头叙事中动词体的研究 33
4.1 引言 33
4.2 文献综述 34
4.2.1 基于情状体假说的相关研究 34

4.2.2 以往有关母语对体习得影响的研究.....35
4.2.3 研究空间.....36
4.3 研究设计和研究方法.....37
4.3.1 受试.....37
4.3.2 语料采集.....37
4.3.3 语料标注.....37
4.4 结果与讨论.....38
4.4.1 重新审视情状体假说.....39
4.4.2 母语对英语体习得的影响.....44
4.5 小结.....52

第五章 英语口头叙事话语的及物性研究 55
5.1 引言.....55
5.2 文献回顾.....55
5.2.1 二语学习者口头叙事话语评价研究.....56
5.2.2 二语学习者口头叙事话语结构研究.....56
5.3 研究设计.....57
5.3.1 研究问题.....57
5.3.2 中介语拓展对比分析法.....58
5.3.3 语料来源.....58
5.3.4 语料标注.....59
5.4 结果与讨论.....60
5.4.1 6 种及物性过程在 3 种语料中的分布.....60
5.4.2 高级英语学习者与英语本族语者的英语口头叙事话语的及物性过程对比.....62
5.4.3 英语本族语者与高级英语学习者英汉口头叙事话语的及物性过程对比.....64
5.4.4 高级英语学习者英汉口头叙事话语的及物性过程对比.....65
5.5 小结.....66

第六章 英语口语中不定式和动名词述补行为研究 67
6.1 引言.....67
6.2 研究综述.....68

6.3 研究方法 70
6.4 结果与讨论 73
6.4.1 对中外语言使用者均有显著影响的因素 76
6.4.2 中外语言使用者述补行为存在差异的因素水平 78
6.5 小结 80

第七章 英语口头叙事中的动词方位构式研究 83
7.1 引言 83
7.2 文献综述 84
7.3 研究设计 85
7.3.1 研究问题 85
7.3.2 研究方法 85
7.3.3 语料收集与标注 86
7.4 结果与讨论 87
7.4.1 动词方位构式的论元结构 87
7.4.2 动词方位构式的图形 – 背景构造 90
7.5 小结 96

第八章 英语口头叙事中复合运动事件表达研究 99
8.1 引言 99
8.2 文献综述 99
8.2.1 运动事件 99
8.2.2 研究空缺 101
8.3 研究方法 102
8.3.1 研究问题 102
8.3.2 研究对象和实验材料 102
8.3.3 数据收集 103
8.3.4 数据标注 103
8.3.5 结果 104
8.4 讨论 115
8.4.1 位移动词的使用 115
8.4.2 运动事件的背景信息提供 116
8.4.3 词汇化模式的异同 116
8.4.4 中国学习者语料与英语母语者语料的关系 117

8.5 小结 117

第九章 英语口头叙事话语中的人物指称研究 119
9.1 引言 119
9.2 文献综述 120
9.3 研究设计 121
9.3.1 研究问题 121
9.3.2 分析框架 122
9.3.3 语料收集与标注 123
9.4 结果与讨论 124
9.4.1 人物指称在中国英语学习者和英美本族语者英语口头叙事中的分布及对比 124
9.4.2 中国英语学习者与英美本族语者英语口语中人物回指的分布差异与人物和情节的关系 126
9.4.3 人物指称在中国英语学习者英语口头与书面语叙事中的对比 128
9.5 小结 129

第十章 英语口头叙事中的话语评价研究 131
10.1 研究背景 131
10.2 研究框架 132
10.3 研究设计 135
10.3.1 研究问题 135
10.3.2 分析方法 135
10.3.3 语料采集与标注 136
10.4 结果与讨论 136
10.4.1 话语评价在中外语料中的分布及对比 136
10.4.2 中外话语评价的概念系统差异 138
10.4.3 中外话语评价的语言表述差异 139
10.5 小结 142

第十一章 英语情境惯用语的语用知识及加工研究 143
11.1 引言 143
11.2 文献综述 144

11.3 研究方法.....147
11.3.1 研究对象.....147
11.3.2 实验材料.....148
11.3.3 研究设计.....149
11.3.4 实验设备.....149
11.3.5 实验程序.....150
11.3.6 数据分析.....151
11.4 研究结果.....151
11.4.1 自然度判断.....151
11.4.2 反应时.....152
11.5 讨论.....155
11.5.1 自然度判断.....155
11.5.2 反应时.....157
11.6 小结.....161

第十二章 学习者英语口语语料库研究展望.....163
12.1 学习者英语口语语料库的建设.....163
12.2 学习者英语口语语料库的研究选题.....164
12.3 学习者英语口语语料库的研究方法.....167

参考文献.....169
附录.....189

表格目录

表 2.1 CLIPS 语料库构成情况 11
表 2.2 CLIPS 语料库中学习者语料类型分布 12
表 2.3 CLIPS 语料库文件名命名规则 14
表 3.1 COLSEC 和 ICE-GB-Spoken 语料库库容 24
表 3.2 互动词块的类型和数量（3—6 词的前 50 词块） 26
表 3.3 同源互动词块数量比较 27
表 3.4 COLSEC 中 I think 族互动词块（部分） 28
表 3.5 ICE-GB-Spoken 中 I think 族互动词块 28
表 3.6 COLSEC 和 ICE-GB-Spoken 互动词块功能分布及比较 30
表 4.1 标注方案 38
表 4.2 语法体的分布 39
表 4.3 情状体的分布 39
表 4.4 中介语中的情状体分布 40
表 4.5 本族语者的英语口语中情状体分布 41
表 4.6 中国学生英语口语和英语本族语者英语口语中的情状体和语法体交互 42
表 4.7 动词 – 小品词构式的情状体分布 45
表 4.8 中国英语学习者中介语中常出现的小品词 48
表 4.9 汉语口语中的动结式 48
表 4.10 中国英语学习者中介语和本族语叙事口语中的常用小品词 49
表 5.1 标注方案一 59
表 5.2 标注方案二 60
表 5.3 6 种及物性过程在 3 种语料中的分布 61
表 5.4 高级英语学习者与英语本族语者的英语口头叙事话语的及物性过程对比 63
表 5.5 高级英语学习者与英语本族语者的英语口头叙事话语的关系过程对比 63
表 5.6 英语本族语者与高级英语学习者英汉口头叙事话语的及物性过程对比 65

表 5.7 高级英语学习者英汉口头叙事话语的及物性过程对比 65
表 6.1 标注的因素及因素水平 71
表 6.2 具有统计显著性的解释变量和交互项 74
表 6.3 中外语言使用者在某些交互因素水平上呈现一致性 77
表 6.4 中外语言使用者存在显著差异的因素水平 79
表 7.1 中国学生英语口语与英美本族语者英语口语中的动词方位构式的论元角色对比 87
表 7.2 中国学生汉语口语与英美本族语者英语口语中的动词方位构式的论元角色对比 89
表 7.3 中国学生英语口语与英美本族语者英语口语中的动词方位构式所用介词对比 91
表 7.4 中国学生英语口语与英美本族语者英语口语中的“动词 +off/away”构式 93
表 7.5 中国学生英语口语与英美本族语者英语口语中的“动词 +by”构式 94
表 8.1 中国学习者英语语料中的动词及频数 104
表 8.2 英语母语者语料中的动词及频数 106
表 8.3 中国学习者汉语语料中的动词及频数 107
表 8.4 SPEN 与 NSSP 中位移动词种类对比 108
表 8.5 SPEN 与 SPCH 中位移动词种类对比 108
表 8.6 NSSP 与 SPCH 中位移动词种类对比 109
表 8.7 “梨子故事”语料库中不同组别使用位移动词比例 110
表 8.8 SPEN 与 NSSP 中位移动词数量对比 110
表 8.9 SPEN 与 SPCH 中位移动词数量对比 110
表 8.10 NSSP 与 SPCH 中位移动词数量对比 110
表 8.11 位移动词结构在 SPCH 语料库中的分布情况 111
表 8.12 “梨子故事”语料库中带背景小句比例 111
表 8.13 三组数据中带背景信息成分比例对比 112
表 8.14 “梨子故事”语料库中复合事件不同连接方式的比例 112
表 9.1 人物指称在 3 种口头叙事语料中的分布和 T 检验结果 124
表 9.2 中国英语学习者与英美本族语者口头叙事中人物回指在不同人物上的分布及对比 126

表 9.3　英美本族语者和中国英语学习者口头叙事中对主要人物的名词回指在三个语步上的差异 128
表 9.4　中国英语学习者口头与书面语中人物指称的对比 128
表 10.1　标注方案及示例 133
表 10.2　话语评价在各类语料中分布 136
表 10.3　中国学生英语口语与本族语者英语口语中的话语评价对比 137
表 10.4　中国学生汉语口语与本族语者英语口语中的话语评价对比 138
表 10.5　各组语料中聚焦评价的具体语言表述（部分）....140
表 11.1　各水平组受试的具体信息 147
表 11.2　各水平组受试对（非）惯用语接受率的描述性统计结果 151
表 11.3　一般线性模型 – 重复测量方法对于（非）惯用语接受度的统计结果 152
表 11.4　不同水平组对（非）惯用语接受度的单因素方差分析统计结果 152
表 11.5　“关键词反应时”和“阅读判断总反应时”在（非）惯用条件下的描述性统计结果 153
表 11.6　一般线性模型 – 重复测量方法对（非）惯用条件下两组反应时的统计结果 154
表 11.7　单因素方差分析对不同水平组（非）惯用条件下的两组反应时统计结果 154

图目录

图 2.1 利用 Sub-corpus Creator 创建 CLIPS 子语料库 16
图 2.2 利用 BFSU Qualitative Coder 标注文本 17
图 2.3 统计的结果 18
图 3.1 互动词块功能表征示意图 29
图 4.1 中介语对比分析扩展模型 37
图 4.2 off 的原型意象图式 51
图 4.3 away 的原型意象图式 51
图 5.1 中介语拓展对比分析法 58
图 7.1 中介语对比分析扩展模型 86
图 7.2 ride away 的意象图式 93
图 7.3 ride off 的意象图式 93
图 7.4 fall over 的意向图式 96
图 7.5 fall down 的意向图式 96
图 7.6 fall off 的意向图式 96
图 9.1 中介语对比分析扩展模型 122
图 10.1 中介语对比分析扩展模型 135
图 10.2 中国学生话语评价与英美本族语者的类别差异分布图 138

第一章　绪论[1]

基于书面语的语言研究在理论语言学领域始终占据主导地位（参阅Linell 2005）。在应用语言学领域，情形大体相同。基于口语语料的（应用）语言学研究之所以稀缺，主要原因在于口语数据的匮乏，并非学界无视口语研究的价值。

据Ballier & Martin（2015：107）不完全统计，各语种学习者口语语料库与书面语语料库之比约为1：3。而根据比利时鲁汶天主教大学“全球学习者语料库一览”（Learner corpora around the world）网站信息，在181个学习者语料库中，有61个口语学习者语料库（截至2020年5月3日），口语学习者语料库约占三分之一。

到目前为止，我国公开发布的英语学习者语料库有如下一些：桂诗春、杨惠中主持创建的“中国学习者英语语料库”（Chinese Learner English Corpus，CLEC）；稍晚，杨惠中、卫乃兴等主持开发的国内首个英语学习者口语语料库，即“大学英语学习者英语口语语料库”（College Learners' Spoken English Corpus，COLSEC）（杨惠中、卫乃兴 2005）；文秋芳团队创建的“中国学生英语口笔语语料库”（Spoken and Written English Corpus

1　本书另有配套网页：http://corpus.bfsu.edu.cn/info/1069/1632.htm，书中提到的相关语料库、视频、工具、文献等资源在该网页有更详细的介绍。

of Chinese Learners，SWECCL）和“中国大学生英汉汉英口笔译语料库”（Parallel Corpus of Chinese EFL Learners，PACCEL）（文秋芳等 2005，2008；文秋芳、王金铨 2008）；上海外国语大学筹建的“中国高校外语专业多语种语料库”，其中主要的英语语料库子项目（Corpus for English Majors，CEM）内含英语专业学生的作文和翻译两类书面语文本（参阅戴炜栋、冯辉 2008）；许家金主持创建的“中国学生万篇英语作文语料库”（Ten-thousand English Compositions of Chinese Learners，TECCL）（许家金 2016）。其中，COLSEC为我国最早的学习者口语语料库，SWECCL1.0/2.0和PACCEL中包含学习者口语语料子库。这些学习者口语语料库往往都是在考试环境下收集的。语料库收集的任务类型包括独白式的朗读、复述、口头叙述、看图说话、口头论说、口译，以及对话式的考官-考生会话、学生-学生两人会话、多人讨论等。

国际范围内的学习者口语语料库情况也大体类似。其中较有影响的两个口语语料库是“鲁汶国际英语口语中介语数据库”（Louvain International Database of Spoken English Interlanguage，LINDSEI）（Gilquin *et al*. 2010）和“三一学院兰卡斯特语料库”（Trinity Lancaster Corpus）（McEnery 2016）。这两个语料库与我国学习者口语语料库的主要区别在于，LINDSEI和Trinity Lancaster Corpus都属于多母语背景的综合学习者口语语料库。LINDSEI语料库中包含保加利亚、中国、荷兰、法国、德国、希腊、意大利、日本、波兰、西班牙和瑞典等11个国家的学习者英语口语语料；Trinity Lancaster Corpus语料库中考生来自意大利、西班牙、墨西哥、阿根廷、巴西、中国、印度、斯里兰卡和俄罗斯等9个国家。不同母语背景的学习者语料有助于进行学习者组别之间的比较，从而更好地探究学习者英语特点是否与母语相关，抑或具有一定的中介语共性。然而，即便上述两个知名学习者口语语料库，其规模也只是百万词级。LINDSEI总共大约100万词次，Trinity Lancaster Corpus大约350万词次。与如今动辄上亿，乃至百亿词次的书面语语料库相比，百万词级的口语语料库确实相形见绌。

近年来，随着语料采集技术，特别是语音识别技术的进步，利用口语语料开展研究，较之以往已有了长足进步。然而，学习者英语口语语料库的总库容仍然明显不足，口语语料库的主要形式多是由音频转写后的文本