



全国高等学校外语教师丛书 · 科研方法系列

**Research Methods
in Corpus Linguistics**

语料库研究方法

许家金等 著

非外借

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





全国高等学校外语教师丛书·科研方法系列

Research Methods in Corpus Linguistics

语料库研究方法

许家金等 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

语料库研究方法 / 许家金等著. — 北京: 外语教学与研究出版社, 2023.10
(全国高等学校外语教师丛书·科研方法系列)
ISBN 978-7-5213-4862-0

I. ①语… II. ①许… III. ①语料库—语言学—研究方法 IV. ①H0

中国国家版本馆 CIP 数据核字 (2023) 第 202162 号

出版人 王 芳
项目负责 段长城
责任编辑 解碧琰
责任校对 陈 阳
封面设计 覃一彪 彩奇风
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 <https://www.fltrp.com>
印 刷 北京九州迅驰传媒文化有限公司
开 本 650×980 1/16
印 张 24.75
版 次 2023 年 12 月第 1 版 2023 年 12 月第 1 次印刷
书 号 ISBN 978-7-5213-4862-0
定 价 99.90 元

如有图书采购需求, 图书内容或印刷装订等问题, 侵权、盗版书籍等线索, 请拨打以下电话或关注官方服务号:

客服电话: 400 898 7008

官方服务号: 微信搜索并关注公众号“外研社官方服务号”

外研社购书网址: <https://fltrp.tmall.com>

物料号: 348620001



记载人类文明
沟通世界文化
www.fltrp.com

总 序

“全国高等学校外语教师丛书”是外语教学与研究出版社高等英语教育出版分社近期精心策划、隆重推出的系列丛书，包含理论指导、科研方法、教学研究和课堂活动四个子系列。本套丛书既包括学界专家精心挑选的国外引进著作，又有特邀国内学者执笔完成的“命题作文”。作为开放的系列丛书，该丛书还将根据外语教学与科研的发展不断增加新的专题，以便教师研修与提高。

编者有幸参与了这套系列丛书的策划工作。在策划过程中，我们分析了高校英语教师面临的困难与挑战，考察了一线教师的需求，最终确立这套丛书选题的指导思想为：想外语教师所想，急外语教师所急，顺应广大教师的发展需求；确立这套丛书的写作特色为：突出科学性、可读性和操作性，做到举重若轻，条理清晰，例证丰富，深入浅出。

第一个子系列是“理论指导”。该系列力图为教师提供某学科或某领域的研究概貌，期盼读者能用较短的时间了解某领域的核心知识点与前沿研究课题。以《二语习得重点问题研究》一书为例，该书不求面面俱到，只求抓住二语习得研究领域中的热点、要点和富有争议的问题，动态展开叙述。每一章的写作以不同意见的争辩为出发点，对取向相左的理论、实证研究结果差异进行分析、梳理和评述，最后介绍或者展望国内外的最新发展趋势。全书阐述清晰，深入浅出，易读易懂。再比如《认知语言学与二语教学》一书，全书分为理论篇、教学篇与研究篇三个部分。理论篇阐述认知语言学视角下的语言观、教学观与学习观，以及与二语教学相关的认知语言学中的主要概念与理论；教学篇选用认知语言学领域比较成熟的理论，探讨应用到中国英语教学实践的可能性；研究篇包括国内外将认知语言学理论应用到教学实践中的研究综述、研究方法介绍以及对未来研究的展望。

第二个子系列是“科研方法”。该系列介绍了多种研究方法，通常是一本书介绍一种方法，例如问卷调查、个案研究、行动研究、有声思维、语料库研

究、微变化研究和启动研究等。也有的书涉及多种方法，综合描述量化研究或者质化研究，例如：《应用语言学中的质性研究与分析》《应用语言学中的量化研究与分析》和《第二语言研究中的数据收集方法》等。凡入选本系列丛书的著作人，无论是国外著者还是国内著者，均有高度的读者意识，乐于为一线教师开展教学科研服务，力求做到帮助读者“排忧解难”。例如，澳大利亚安妮·伯恩斯（Anne Burns）教授撰写的《英语教学中的行动研究方法》一书，从一线教师的视角，讨论行动研究的各个环节，每章均有“反思时刻”“行动时刻”等新颖形式设计。同时，全书运用了丰富例证来解释理论概念，便于读者理解、思考和消化所读内容。凡是应邀撰写研究方法系列的中国著作人均有博士学位，并对自己阐述的研究方法有着丰富的实践经验。他们有的运用了书中的研究方法完成了硕士、博士论文，有的采用书中的研究方法从事过重大科研项目。以秦晓晴教授撰写的《外语教学问卷调查法》一书为例，该书著者将系统性与实用性有机结合，根据实施问卷调查法的流程，系统地介绍了问卷调查研究中问题的提出、问卷项目设计、问卷试测、问卷实施、问卷整理及数据准备、问卷评价以及问卷数据汇总及统计分析方法选择等环节。书中各个环节的描述都配有易于理解的研究实例。

第三个子系列是“教学研究”。该系列与前两个系列相比，有两点显著不同：第一，本系列侧重同步培养教师的教学能力与教学研究能力；第二，本系列所有著作的撰稿人主要为中国学者。有些著者虽然目前在海外工作和生活，但他们出国前曾在国内高校任教，也经常回国参与国内的教学与研究工作。本系列包括《英语听力教学与研究》《英语写作教学与研究》《英语阅读教学与研究》《英语口语教学与研究》《翻译教学与研究》等。以《英语听力教学与研究》一书为例，著者王艳副教授拥有十多年的听力教学经验，同时听力教学研究又是她博士论文的选题领域。《英语听力教学与研究》一书，浓缩了她多年来听力教学与听力教学研究的宝贵经验。全书分为两部分：教学篇与研究篇。教学篇中涉及了听力教学的各个重要环节以及学生在听力学习中可能碰到的困难与应对的办法，所选用的案例均来自著者课堂教学的真实活动。研究篇中既有著者的听力教学研究案例，也有著者从国内外文献中筛选出的符合中国国情的听力教学研究案例，综合在一起加以分析阐述。

第四个子系列是“课堂活动”。该系列汇集了各分册作者多年来的一线教学经验，旨在为教师提供具体、真实、具有较高借鉴价值的课堂活动案例，提高教师的课堂教学能力。该系列图书包括《英语阅读教学活动设计》《英语听力课堂活动设计》《英语合作学习活动》等。以《英语阅读教学活动设计》一书为例，阅读教学是学生学习语言知识和教师培养学生思维的重要途径和载体。该书第一作者陈则航教授多年来致力于英语阅读教学研究，希望通过该书与读者分享如何进行具体的阅读教学活动设计，探讨如何在课堂教学中落实阅读教学理念。该书包括三个部分。第一部分介绍在阅读前、阅读中和阅读后这三个不同阶段教师可以设计的阅读教学活动，并且介绍了阅读测评的目的、原则和方式。第二部分探讨了如何通过阅读教学促进学生思维发展。第三部分展示了教师在阅读课堂中的真实教学案例，并对其进行了分析与点评，以期为进一步改进阅读教学活动设计提供启示。

教育大计，教师为本。“全国高等学校外语教师丛书”内容全面，出版及时，必将成为高校教师提升自我教学能力、研究能力与合作能力的良师益友。编者相信本套丛书的出版对高校外语教师个人专业能力的提高，对教师队伍整体素质的提高，必将起到积极的推动作用。

文秋芳

北京外国语大学

中国外语与教育研究中心

2011年7月3日

前言

自 2013 年起,我开始在北京外国语大学讲授“语料库研究方法”研究生课程,早有将教学内容编为教材的计划,但着实难有余暇。当前,语料库研究方法快速发展,本书既是对十年教学的整理,更是对前沿方法的学习和追赶。

本书得以付梓多半归功于我这些既能干又可爱的学生。若没有他们的参与,再过三五年《语料库研究方法》恐怕也难以面世。推却所有工作,专注于一部书稿,已成奢侈之事。看到学生成长之快,学习能力之强,让我心生合著此书的念头。感谢他们帮我达成夙愿。面对学生,我不断强调,弟子不必不如师。敢于超越老师,方能进步,修得正果。于是找来几位学生,大家一拍即合,在确定全书框架后分工撰写。本书是团队协作的成果,是特殊的纪念,可庆可贺。

此外,由衷感谢外语教学与研究出版社高等英语教育出版分社副社长段长城老师的邀约和敦促。没有段老师的不断跟进,本书可能还会拖延一段时间。

语料库研究方法,搭着计算机科学的顺风车,不言一日千里,也是日新月异。若再没有这样的中文专论,恐怕只得求助外文书籍了。最近三五年,国内同行出版了几本基于 R 或 Python 开展语言学研究的书籍。这些书也是我们的参考文献,是我们学习的榜样。

本书将语料库研究方法的发展历程划分为“经典时代”和“后经典时代”两个阶段。前一个阶段主要包括索引、词表、搭配、主题词一类的方法,后一个阶段主要涵盖聚类、条件推断树、随机森林、逻辑斯蒂回归等多变量统计方法。本书重点关注后经典时代的语料库研究方法,以充分反映本领域近十几年的新进展。另外,共选思维、建模思维、对比思维,是本书倡导的

语料库方法论导向，特别是共选思维，即共选论。本书采用共选论，试图将全书所涉及的语料库研究方法集于一个统一的思路下，权作试水，留待读者斧正。

本书致力于从我国语言学及应用语言学研究者的实际需求出发。在研究方法的实操环节，所用案例多为我国学者实际发表的科研成果。在研究选题方面，我们特别关注汉语研究、英语研究、多语种研究、翻译研究、口语研究、中介语研究、话语研究等，以贴近我国学者的研究需求。在操作方法的介绍中，考虑到有不少读者使用 macOS 操作系统，我们对 macOS 版本的分析软件也作了一些特别介绍和说明。

书中涉及较多语料、数据、代码、工具、文献，我们特创建配套网页 (<http://corpus.bfsu.edu.cn/info/1084/1873.htm>) 来存储相关资料，供读者读取、使用。若有更新或勘误，也请读者访问该配套网页。若遇网址变动，可通过网络查询“《语料库研究方法》配套网页”搜索该页面。

本书具体分工如下：许家金撰写第一章、第二章、第十四章，孙铭辰撰写第三章和第五章，赵冲撰写第四章，刘运锋撰写第六章，郝美佳撰写第七章和第八章，李佳蕾撰写第九章和第十章，李维静撰写第十一章，周顾盈和李维静撰写第十二章，周顾盈和张懂撰写第十三章。许家金负责统稿。

本书系教育部人文社会科学重点研究基地重大项目“基于多语种语料库的外语及外语教育研究”（编号：22JJD740012）的阶段性成果。

特别感谢卫乃兴教授、梁茂成教授、吴淑琼教授、唐美华博士、王冰昕博士同意我们使用他们研究中的部分数据。

在本书的写作过程中，我从我的学生那里学到了很多。他们教会了我很多新的概念和方法，与我分享了很多令人兴奋的研究，指出了我的很多错漏。由衷向他们表示感谢。

本书的责任编辑李晓雨老师，以其专业的编辑能力和丰富的语料库语言学知识，对书稿进行了全面的审查和修改。她提出的详细修改建议，使得书稿的结构更加清晰，概念表述更加准确，文献条目更加规范，格式更加统一。在此，我们向李晓雨老师表达最诚挚的感谢和深深的敬意。

全书由我统稿，学识所限，加之语料库研究方法更新迭代超乎想象，书中舛误概由我来负责，敬请各位赐教。

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

2023 年 3 月

目 录

总 序	文秋芳	vii
前 言	许家金	x
第一章 概论		1
1.1 引言		1
1.2 语料库研究方法中的文本		1
1.3 语料库研究方法中的工具		3
1.3.1 经典时代的语料库研究方法简述		5
1.3.2 后经典时代语料库研究方法的特色		6
1.3.3 两个时代的语料库研究方法评述		9
1.4 语料库研究方法的理论贡献		12
1.5 结语		13
第二章 语料库建设的一般方法		15
2.1 建库原则		15
2.2 文本规格		19
2.3 语料标注		33
第三章 语料库建设的网页爬取方法		39
3.1 R 语言静态网页爬取		39
3.2 R 语言动态网页爬取		44
3.2.1 微博搜索话题博文的爬取		44
3.2.2 微博博主博文的爬取		49
3.3 基于爬虫软件的爬取		54

第四章 语料检索及语料清洗	60
4.1 检索语料	64
4.2 语料清洗	77
第五章 数据可视化	90
5.1 数据分布	90
5.1.1 箱线图	92
5.1.2 小提琴图	93
5.1.3 密度图	94
5.1.4 峰峦图	95
5.2 数据比较	97
5.2.1 条形图	99
5.2.2 棒棒糖图	100
5.2.3 雷达图	101
5.3 数据演变	102
5.3.1 折线图	104
5.3.2 面积图	105
5.3.3 动态图	108
5.4 数据关系	109
5.4.1 散点图	110
5.4.2 气泡图	112
5.4.3 热力图	113
5.5 其他数据	114
5.5.1 饼图	114
5.5.2 网络图	117
5.5.3 词云图	119
第六章 短语学分析	122
6.1 语料库短语学研究实践	122

6.1.1 搭配基础研究	123
6.1.2 搭配扩展研究	125
6.1.3 局部语法研究	128
6.2 语料库短语学的分析方法	133
6.2.1 搭配分析	133
6.2.2 扩展意义单位描写	136
6.2.3 局部语法构型描写	141
第七章 搭配构式分析	149
7.1 共现词分析	149
7.2 显著共现词分析	156
7.3 共变共现词分析	161
第八章 对应分析	166
8.1 简单对应分析	166
8.2 多重对应分析	170
第九章 主成分分析和因子分析	184
9.1 主成分分析和因子分析的基本概念	184
9.2 主成分分析和因子分析在语言学领域的应用	184
9.3 主成分分析和因子分析的相似性和差异性	187
9.4 案例分析：一百多年间汉语书面语的语域演变研究	188
第十章 多维尺度分析	198
10.1 多维尺度分析的基本概念	198
10.2 多维尺度分析在语言学领域的应用	198
10.3 多维尺度分析、因子分析和聚类分析之间的相似性和差异性	201
10.4 案例分析：现代原创汉语和翻译汉语的分期和历时演变	201

第十一章 聚类分析	209
11.1 聚类分析综述	209
11.2 聚类分析种类	211
11.3 案例实操	215
11.3.1 层次聚类分析: 数值数据	215
11.3.1.1 案例背景介绍	215
11.3.1.2 操作及代码	216
11.3.2 层次聚类分析: 分类数据	222
11.3.2.1 案例背景介绍	222
11.3.2.2 操作及代码	225
11.3.3 划分聚类分析: 分类数据	234
11.3.3.1 设定聚类数量	235
11.3.3.2 选择划分方式	235
11.3.3.3 选择距离度量	236
11.3.3.4 确定最优分类	236
11.3.3.5 操作及代码	236
第十二章 决策树和随机森林	240
12.1 决策树	241
12.1.1 经典决策树	243
12.1.2 条件推断树	244
12.1.2.1 算法介绍	244
12.1.2.2 语言研究中的运用	245
12.1.2.3 案例实操	246
12.2 随机森林	251
12.2.1 基于 CART 的(经典)随机森林	254
12.2.2 基于 CTREE 的条件推断森林	260

第十三章 逻辑斯蒂回归	267
13.1 逻辑斯蒂回归模型基本介绍	267
13.1.1 模型原理	267
13.1.2 前提假设和数据要求	269
13.1.3 两种推广	273
13.1.3.1 加入随机效应的逻辑斯蒂回归	273
13.1.3.2 多分类逻辑斯蒂回归	275
13.2 逻辑斯蒂回归在语言学领域的应用	275
13.3 二分类任务：英语中的与格交替	280
13.3.1 语料获取与处理	280
13.3.2 固定效应模型	286
13.3.2.1 模型拟合与解读	286
13.3.2.2 模型优度评估	294
13.3.2.3 模型诊断	299
13.3.3 混合效应逻辑斯蒂回归模型	303
13.4 多分类任务：德语中的三种回指形式选择	308
13.4.1 语料获取与处理	309
13.4.2 两种方式拟合多分类模型	313
13.4.2.1 使用 <code>nnet</code> 程序包	313
13.4.2.2 使用 <code>polytomous</code> 程序包	321
第十四章 语料库研究方法展望	327
14.1 语料库建设展望	327
14.2 语料库分析技术展望	329
14.3 语料库语言学理论建构展望	331
参考文献	333
附录 A：英汉双语语料库语言学学科术语表	363
附录 B：常用正则表达式小结	379

第一章 概论

1.1 引言

“语料库语言学”这一术语的中心语虽为“语言学”，但多数学者倾向于把语料库语言学当作方法论，而将其视为学科理论体系的学者不过十之一二。本书无意参与语料库语言学是方法论还是学科理论体系的争辩。实际上，理论与方法密不可分，语料库语言学具有理论和方法的双重特点。语料库语言学方法优势突出，其研究方法中蕴含着诸多理论关切。本书聚焦语料库语言学研究方法，旨在对经典和前沿的研究方法作系统梳理，进一步探讨语料库研究方法的发展对语料库语言学理论建构的反哺作用。

本书将语料库语言学的核心要义概括为“3个T”：Texts（文本）、Tools（工具）、Theories（理论）。其中，工具是桥梁，联通文本和理论。语料库研究正是借助分析工具从语言事实（文本）中探求语言规律和机制（理论）。这里的工具不单指语料库软件，也泛指语料库分析方法。

本书试以“共选论”（许家金 2014a, 2017, 2020a）统摄语料库研究中的文本收集、工具运用、理论阐释环节。共选论强调语义的传达和识解受制于多重语境因素的协同作用。具体而言，语境中的词汇、语法范畴等的概率分布及其共现可有效区分词语意义、探究句式选用机制、裁定语域类别等。从应用语言学的视角来看，通过语境共选分析可以解决词不达意、句不合规、话不中用等语言使用不当问题。

1.2 语料库研究方法中的文本

首先探讨文本。如今百亿词级乃至更大规模的语料库已不鲜见。然而，语言学研究使用的语料库库容仍在亿级以下，多为千万或百万词级，或更小规模的语料库。大规模的巨量文本数据主要源自自动爬取的网络语料。这类语

料取之不竭，量不封顶。但若不加筛选，泥沙俱下的网络语料则难以直接为语言学学者所用。一般来说，用于语言学和语言教学的语料库对语料的规范性和文本产生的语境要求更高。语料库文本取样的代表性，语料产生的时间、地点，说话人的职业、性别等因素，是语言理论探究的基石，影响着理论探索的方向。可以说，在语言研究中，语料不可靠，结论必不牢。因此，从文本数据的质量要求来看，人工智能领域的语料库和语言研究领域的语料库是明显不同的。前者以量取胜；后者要求质和量兼顾，质优者胜。两者对语料库文本的不同理解主要归因于研究目的的差异。人工智能研究依靠大数据优化算法、辅助决策，因而重“量”；语言研究者通过大数据考察词汇、语法等形式特征与说话人观点和立场之间的对应关系，“质”和“量”缺一不可。除了两者存在的差异，我们还要注意到当前人工智能研究进展对语料库语言学“真实文本”质量的影响。例如，利用 ChatGPT 这类工具得到的人工智能生成内容（AIGC），是否算真实发生的自然语言。换言之，人工智能能否算作与人等同的话语参与者。尽管这些问题涉及人工智能与人类智能的边界争议，一时不易作出裁断，但势必会对语料库研究中的文本概念和文本选择产生一定影响。

在语言教研实践中，我们一般会选用规模较大的权威语料库，如当代美国英语语料库（Corpus of Contemporary American English，简称 COCA）和英国国家语料库（British National Corpus，简称 BNC）。很多情况下，我们也提倡研究者自建语料库，或运用百万词级规模、小而精的语料库，即包含丰富的文本语境且社会文化语境信息的取样广泛多元的语料库。这样的语料库更便于后续研究考察语言运用的制约因素。在建设语料库时，研究者就应将文本产生的关键语境因素尽力记录在案。例如，英国国家语料库中存在几十项语境因素记录，包括作者和说话人的性别、年龄、社会阶层等。总之，文本是语料库研究的重要载体，它的量与质同等重要。在两者的统筹兼顾中，较为关键的是注重文本使用的社会、文化、情景等相关参数以及语音、体势、语气等多模态变量。这是文本意义研究和多因素、多变量分析的基础。

1.3 语料库研究方法中的工具

本节介绍语料库研究方法中的工具。本书中的“工具”主要指：(1) 语料分析方法；(2) 语料库研究设计。从语料库研究方法的发展历程来看，可粗略以 2000 年为界，将其前后分为语料库研究方法的“经典时代”和“后经典时代”。这两个时期并非截然分开，而是呈交叠之势：在 21 世纪初，“后经典时代”的语料库研究方法逐渐兴起，“经典时代”的语料库研究方法依然流行。表 1.1 呈现了两个时代的主要语料分析方法。

表 1.1 语料分析方法概览

时代及特点	研究定位	语料分析方法
聚焦单特征的 “经典时代”	语言特征 计量对比	1. 单词频数统计，如词频表、形次比； 2. 多词频数统计，如词丛表、短语框架频数表； 3. 词性、句法单位频数统计，如名词及名词短语、T 单位、小句等语言单位的频数计算； 4. 话语、语义、语用、隐喻、偏误等语言特征的频数统计； 5. 频数差异计算，如通过卡方检验、对数似然率、费舍尔精确检验等方法，对比相关语言特征在不同语料库的频数，其中包含主题词表、主题短语表、主题性短语框架表的抽取。
	词汇语法 共选生义	1. 索引分析：通过频数、互信息、卡方检验、对数似然率等词语搭配算法描写和解析意义单位； 2. 局部语法：通过具体词形和细颗粒度的功能标注分析特定意义 / 功能与词汇语法的对应关系； 3. 搭配构式分析：通过 ΔP、费舍尔精确检验等方法分析词语与构式间的相互吸引和排斥关系； 4. 多维分析：通过多项词汇、语法特征在文本中的共现模式，分析语域的变异情况。

(待续)