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 `nnet` 程序包 …… 313
13.4.2.2 使用 `polytomous` 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

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。这些书也是我们的重要参考文献，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

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，语言学研究中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）

（续表）

时代及特点	研究定位	语料分析方法
协同多特征的“后经典时代”	文内文外共选生义	1. 综合考察文内特征（如词汇、短语、语法、话语、语义）和文外特征（语用、社会语言学变量）的共选机制，揭示形义对应关系； 2. 多项语言特征归集成组，有助于将同一个语义单位作细分类别探讨，常采用主成分分析、因子分析、对应分析、多维尺度分析、聚类分析等； 3. 多项语言特征协同解析两个或多个近义单位的选用机制，常采用逻辑斯蒂回归、条件推断树、随机森林、线性判别分析、支持向量机、朴素贝叶斯模型等。

注：（1）理论上，几乎所有统计方法都可用于语料库研究。本书聚焦计算语言特征出现频率和分布规律的典型统计学方法。

（2）语言现象是一种社会化现象，其本质是多元协同的。多元一体、互动协同的观点并不新。从经典时代到后经典时代的发展，主要体现在技术和方法的进步上。可以说，经典时代对单变量的关注，是因为技术和方法不够成熟，是不能而非不为。在后经典时代，多因素、可视化技术极大地提高了语料库研究结果的可解读性，为语言现象及其使用规律的概念化提供了更直观的工具。工具带来方法的革新，从而加深我们对语言的理解。从经典时代到后经典时代的发展，一方面明显推动了语料库文本意义研究技术的进步，另一方面，也使研究者对语言理论有了更深的认识。

（3）语言单位间相互关联和制约，而非相互独立。在统计上，语言现象的这种关联性可通过聚类、分类、回归等方法加以分析。例如，在筛选出强相关变量的基础上，回归分析有希望捕捉到变量间可能存在的因果关系。通过将具体指标进行聚类和分类，（探索性）因子分析可测量某些抽象的概念，如语域。相应工具和方法论的使用能够精准地描写语言单位间的多元协同效应，可以更好地贯通语言文本意义研究的点、线、面、体。

（4）语料库分析中也经常见到语言单位长度（如词长、句长、语言单位间的距离、词汇句法复杂度、可读性等）和文本得分（如作文成绩、难度评分）等连续变量。这类数据并非语料库语言学研究的最典型数据类型，因此不作重点介绍。在语料库相关的应用语言学实证研究中，对于连续变量的统计并不少见。常用的研究方法包括 T 检验、方差分析、线性回归、结构方程模型等。这些方法和工具进一步深化了应用语言学的理论与实践研究。

本书区分了经典时代和后经典时代，主要是为方便梳理几十年间语料库研究方法的发展历程，但这并不意味着经典的方法已弃而不用。从流行时间和功能定位两个维度来看，表 1.1 中“局部语法”“搭配构式分析”“多维分析”这

三种研究方法其实出现于经典时代向后经典时代过渡的时期。另外，很多如今流行的统计方法，也并非最近一二十年才出现。这些统计方法只是近期逐渐进入语料库研究领域，并发挥积极作用。例如，心理学家 Louis Leon Thurstone 于 1935 年提出“多元因子分析法”（multiple factor analysis），但这一方法成为语料库研究的关键分析方法，主要还是 Douglas Biber 的功劳（Biber 1984，1988）。近年多变量统计在实证语言研究中受到重视，更激活了因子分析这类统计方法在语料库研究中的应用。

1.3.1　经典时代的语料库研究方法简述

经典时代的语料库研究方法包括词频表（word frequency list）、词丛表（word cluster list）、索引分析（concordance analysis）、搭配分析（collocation analysis）、主题词分析（keywords analysis）和多维分析（multi-dimensional analysis）等。其中手工编制的词频表和索引分析的历史相当悠久。20 世纪中叶，随着电子化语料库的产生，利用计算机手段获取词频表和建立索引成为早期的文本分析手段。这两种功能可以进一步归结为“检索”和“查询”，是经典时代语料库研究方法的基础。经典时代的分析方法依赖于在大量文本中查询特定字词或短语结构用法，并给出相应的使用频率。在无特定检索目标的情况下，语料库软件可对所有文本进行穷尽式检索，并枚举文本中的所有词汇，附上频数信息。如有确定的检索目标，研究者可针对检索词，观察其在语料库中的上下文语境、社会文本情境，从而全面了解其用法。

在词频表和索引分析的基础上，词丛表、主题词分析、搭配分析逐渐发展起来。词丛有词块、词簇、N 元组等不同称谓，其实质是多词词频表。主题词分析则是将两个编制好的单词词表或多词词表中的词汇或短语逐个进行比对，其中统计学上存在显著差异的词汇或短语即为主题词或主题词丛。由此得到的词汇或短语列表往往能揭示文本的主题内容，“主题词分析”也因而得名。搭配分析基于索引分析的检索结果，对检索词和语境共现词之间的依存关系进行分析，进而帮助我们理解短语意义。

在语料库研究方法的经典时代，词丛表和搭配分析是“短语学”（phraseology）

（Sinclair 1991，2004a）研究的重要技术手段。主题词也可以理解为语料库文本中的高频共现词语（Scott 1997），可以用于建构话语主题（McEnery 2006）。在经典时代，操作流程上最为复杂的研究方法当属 Biber（1984，1988）提出的多维分析。该方法主要是在检索词汇、语法特征的基础上，通过因子分析法将几十乃至上百个语言特征自动归结为几个共现特性大类（即多个维度），从而支撑对某类特定文本语域的识别和探讨。以上相关技术在当前的语料库研究中仍被广泛应用，并有所拓展和升级。比如：

（1）基于语料库的“批判话语分析”（Critical Discourse Analysis）是当代语料库研究分支中成果最为丰富的一派，以兰卡斯特大学相关学者为代表。从语料库分析技术层面来看，批判话语分析采用的仍是经典时代的研究方法，本质上立足于词汇的共选。

（2）“局部语法”（Local Grammars）（Gross 1993；Barnbrook 2002）通过对索引行进行多层级、细颗粒度的标注，得出构型和功能的共选模式。

（3）多维分析最初用于英语的语域研究，现不断拓展至多个语种、各类子语域。

（4）在短语学之外产生了融合构式思想的“搭配构式分析”（Collostructional Analysis）（Stefanowitsch & Gries 2003）。

总的来说，经典时代的语料库研究方法以索引行观察、频数计算与对比、文本主题词分析为主，描述统计是重要手段，短语学研究是重要内容。经典时代后期的批判话语分析、局部语法和多维分析在内容和方法上丰富了短语学研究和文本解读的维度，呈现出向细分变体或细分领域发展的趋势。这些经典方法是语料库语言学的重要基础，在后经典时代依然被广泛应用。

1.3.2 后经典时代语料库研究方法的特色

后经典时代的语料库研究方法表现出明显的整体观和协同性。这一时期的研究方法关注整体语境，注重言内（语言特征）和言外（社会文化特征）语境

的相互协同，从而综合考察语言运用，揭示语言形式和意义间的对应关系。在具体分析技术上，后经典时代的语料库研究方法主要有两种取向。第一种是基于特征异同将不同实例归集成组，可称为“聚类型方法”（clustering）[1]。一般来说，聚类型方法整合归并的是语言实例，如包含某个词汇的所有语句。但是，主成分分析、因子分析这类降维方法也被用来分析词汇语法特征，此时语言特征就成了语言实例。第二种研究取向是“分类型方法”（classification），指研究者事先知晓存在多种语言范畴（如图 1.1 中的圆形、三角形和正方形），继而通过统计手段将这些范畴尽量区分开来。这种方法可用于解析和界定语言使用中的近义范畴。事实上，聚类型方法和分类型方法的共同点在于分类，聚类型方法适用于语言实例所属的范畴不清或界限不明的情况，而分类型方法适用于类别已知的情况。究其根本，语言研究重在解决语言范畴划分问题，或语言使用者舍此取彼的语言运用机制问题。

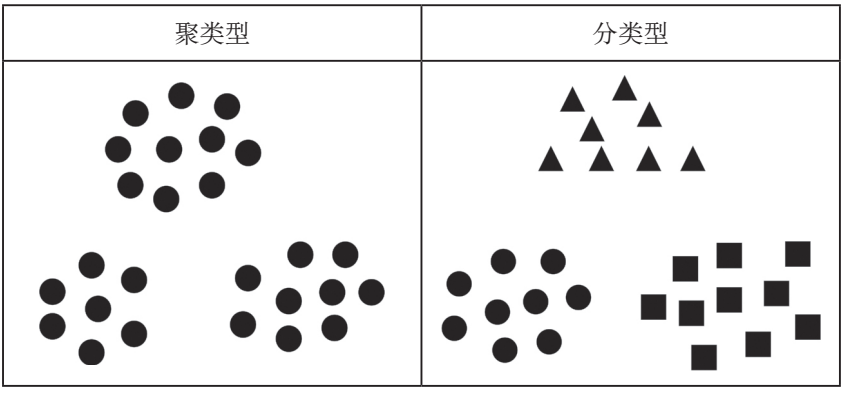

图 1.1　聚类型和分类型方法的示意图 [2]

1　这里的聚类型方法是上位概念，其中包含“聚类分析”这一方法。本文对语料库研究方法的区分采用了多变量统计和机器学习中聚类和分类的表述，但主要还是从语言实践中的多项语言特征共选生义和共选辨义的角度进行概括的，并非统计术语的简单借用。统计学中的回归和降维，在本书中也按其实现的语言目标被归入聚类和分类。例如，逻辑斯蒂回归因其解决的核心问题是分类，故而未单列出一类回归算法，而是将其并入分类型方法。

2　该示意图的绘制受到 https://www.atmosera.com/blog/supervised-learning-with-k-nearest-neighbors/ 页面图片启发。

第一种取向的相关方法可整合归并众多语言特征。归并而成的数据组合可更好地概括某个不易分解或较抽象的语言范畴。聚类型方法常包括主成分分析、因子分析、对应分析、多维尺度分析、聚类分析等。根据研究需要，相关统计方法有时也会组合使用。例如，语言研究中的“语域”（register）这一概念，相对不容易界定和判定。Biber（1988）在其多维分析框架中，采用因子分析这一方法分析 67 个词汇的语法特征，归纳并概括出交互性 / 信息性、叙述性、指称明晰性等 7 个维度。这种自下而上的方法，将看似纷繁复杂的语域实践条分缕析、化繁为简。据此得出的多维分析思路被广泛应用于学术语言（Biber *et al*. 2007）、英语变体（Kruger & van Rooy 2018；Bohmann 2019）、翻译语言（Hu *et al*. 2019）、学习者语言（Larsson *et al.* 2021）、语域演变（许家金、李佳蕾 2022）、著者身份（Grieve 2023）等方面的研究，成为文本语域变异研究的重要手段。

对于隐喻、转喻这样的抽象概念议题，也有越来越多的研究采用后经典时代的研究方法进行分析。例如，Glynn（2015）采用层次聚类和多重对应分析对 19 世纪到 20 世纪美式英语中“家”（home）这一概念的多项语义属性进行了挖掘。研究发现，在两个世纪的语料中，“家”的核心含义，如住宅（house）、土地（land）、国家（nation）、精神家园（abstract place），从众多属性中被分离出来。另外，有别于 19 世纪，20 世纪的语料中表示抽象含义的“精神家园”属性更为凸显。Dylan Glynn 推断，随着时间的推移，人们的出行、迁徙变得更为频繁，进而引发“家”的概念表征发生变化。聚类和对应分析一类的方法往往能以可视化手段直观呈现研究发现。Szmrecsanyi（2013）和 Grieve（2016）则更进一步，根据研究需要，利用地理信息可视化技术，将英式英语和美式英语在两国国内的词汇语法使用异同以热力图和散点图的方式呈现在英国和美国的地图上。

后经典时代语料库研究方法的第二种取向，即分类型方法，在构式交替（construction alternation）研究中得到广泛应用。这类研究主要关注两个或多个近义构式的选用理据，如 Bresnan（2007），Bresnan & Ford（2010），房印杰（2017a），张炜炜、王芳（2017），张懂、许家金（2019）有关英语和汉语中构式交替现象的研究。其中较有代表性的研究是 Bresnan & Ford（2010）。

这项研究探讨了双宾与格构式（如 Mary gave John the book）和介词与格构式（如 Mary gave the book to John）的区分依据。通过对多个潜在因素进行手工标注（如与事和受事成分是否是代词，与事和受事的生命性，名词短语的长度等），该研究利用二分类逻辑斯蒂回归模型，在概率上确定了影响两种近义构式变换使用的依据（例如，有生或无生、有定或无定、与事或受事是否为代词等），并知晓了相关因素的影响权重。Zhang & Xu（2023）采用类似方法对汉语与格的变换作了探讨。Gries & Bernaisch（2016）以类似的方法和思路对亚洲各种英语变体中的与格交替现象作了比较。Gast（2015）运用多分类逻辑斯蒂回归方法对比了德语和英语中非人称用法的选用情况。Kruger（2019）和 Liu（2023）分别用多因素统计建模对笔译中关系代词 that 的使用与省略，以及汉英口译中 I 和 we 的选用规律作了深入分析。Dubois *et al.*（2023）对学习者英语中的属格交替现象（即 of 和 's 构式的选用情况，例如 the tail of the dog/the dog's tail）进行了深入探讨。更多的研究选题和方法介绍可参阅 Gries（2018）、Speelman *et al.*（2018）、De Sutter & Lefer（2020）、许家金（2020a）及李元科等（2022）。

除此之外，后经典时代采用的“混合效应回归模型”（mixed-effects regression model）（Speelman *et al.* 2018）还能很好地分析语言特征和语言使用者的个体差异，这是经典时代的研究方法所不擅长之处。这一研究方法既能触及语言的规约性，又能考察语言的创造性（Goldberg 2019），为语料库研究注入了新的活力。

后经典时代的语料库研究充分利用推断统计、多因素分析、文本聚类、可视化呈现等现代技术，这极大地推动了语言研究方法的创新，使得短语学、话语分析、语言对比与翻译等主题的研究持续深化。因而后经典时代的语料库研究方法成为当前语料库研究的主要方法论。

1.3.3　两个时代的语料库研究方法评述

概言之，经典时代和后经典时代的语料库研究方法的区分与联系可概括如下：

（1）都关注真实文本中语言使用的共选特征，前者多为词语、语法的典型使用特征描写，后者为不同语境参数中词语、语法、文本类型等要素的局部使用特征描写；

（2）都重视真实语言的使用概率，前者多采用索引行分析和频数百分比等描写统计手段，后者则发展为使用多因素分析、文本聚类与分类、主题建模等相对复杂的推断统计方法；

（3）都以语言意义描写为主要任务，前者多在通用语料库中进行词语和语法使用的互文描写与解释，后者开始注重受特定语境限制的词语和语法局部意义或功能；

（4）都以多维分析和局部语法为发展过渡，经典时代的语料库研究方法在语言描写和语言理解方面为后经典时代的研究方法打下了重要基础，后经典时代的研究方法是经典时代研究方法的自然发展，两者有机地联系起来。

就后经典时代的方法而言，两类主要方法都指向语言运用中的“选择”（choice）和“概率”（chance）问题（Herdan 1966）。从用法本位（usage-based）的观点来看，语言运用可归结为语境制约下的意义和功能实现及其变异。在语言层面具体体现为话语使用者如何在特定形态句法选项中作出最优选择。这样的选择机制基于人们的长期语言实践，受到多种语境因素共同制约，是有概率基础的。语境特征的概率性共选可以看作语言运作的一种本体性特征。共选论横跨经典时代与后经典时代，而第二个阶段的研究方法总体上体现出建模思维。后经典时代的分析方法主攻多项语言特征的整体考察和协同分析。这些语言特征的共选会使人们在选用语言形式、语义和功能范畴时产生不同的行为倾向。

如果用机器学习的概念来解释聚类型方法和分类型方法，那么前者指所得到的语言现象类别是事先不确定的，偏向于无监督式的学习算法；后者所关注的类别是事先给定的，属于监督式的学习算法。前者偏向探索性，后者偏向验证性。研究者采用聚类型方法或分类型方法，取决于他们是否已事先明确语言范畴的分类。虽然前文只是通过案例重点介绍了因子分析和逻辑斯蒂

回归两种方法，但根据研究实际，表 1.1 中涉及的后经典时代分析方法，都可以适时采用，且可以在一项研究中采用多种方法，甚至还可以在概率统计的基础上，结合反应时、眼动、脑电等心理神经实验方法进行交叉验证。

后经典时代研究方法的关键数据是与所研究语言范畴相关的语言特征集（feature set 或 feature catalogue），即对选用的特定范畴有潜在影响的词、句、篇、语义、语用、认知乃至社会文化方面的因素。通过在语料库中检索得到相关特征的频数后，可选用因子分析、聚类分析、多维尺度分析、逻辑斯蒂回归、条件推断树、随机森林等一个或多个方法对语言实例进行分析。后经典时代的语料库分析数据在形态上的突出特色是以行列式呈现，通常来说，每一列对应一个语言特征或语境变量，而每一行记录的是某一个文本出现相应语言特征的频数或词句长度等信息。这样的数据格式正与整体语言观和协同语言观相适应。

前文介绍两类分析方法时涉及的语言现象以二分类型案例居多。实际上，语言中的多分类现象并不少见。例如，本书介绍的回指语名词、代词和零形式的选用问题，可以通过先行语和回指语相关的语境变量，如先行语和回指语之间的距离、语境中与先行语存在竞争关系的其他名词性成分的数目等，构建回归模型，从而推断出选用名词、代词和零形式三种回指类型的关键制约因素及其影响权重。前文提及的很多统计方法也适合对语言中的多分类现象进行研究。此外，多分类现象往往也可以约减为二分类现象。例如，徐秀玲（2020）就将回指归并为显性回指（包括名词和代词）和隐性回指（零形式）两大类。

从拓展研究视野和解决语言应用问题的角度来看，对比研究设计至关重要，可分为横向对比和纵向对比两个维度。在横向维度上，我们可以通过选取汉语和外语、学习者和本族语者、翻译和原创、男性和女性、口语和书面语、文科和理工科、新媒体和传统纸媒等方面的语料，开展汉外对比研究、二语习得研究、翻译研究、社会语言学研究、话语研究、媒体研究等。这些横向维度的对比，在操作层面，可以分别检索两组语料，之后进行统计差异检验；也可以将语料差异作为一个变量，与统计模型结合，让算法自动进行分类比较。前一种对比在经典时代的语料库研究方法中更常见，后一种在后经典时代是主流。在纵向维度上，研究者可以基于不同历史时期的语料开展语言演变研究，

也可以对学习者不同阶段的语料开展语言发展研究。

本书重点介绍后经典时代的语料库研究方法。对经典时代的语料库研究方法只作一般性介绍，此方面文献资料较为充分。

1.4 语料库研究方法的理论贡献

在经典时代，John Sinclair 通过词语搭配、扩展意义单位、局部语法、线性单位语法等路径，构建了以词汇为中心的意义理论，形成了自成一体的短语学或“词汇语法”（Lexical Grammar）。这一研究传统对理论语言学和应用语言学都产生了不小的影响，在相当长的时期内，学界对语块、构式一类的意义单位的研究热度空前高涨。短语学的理念在理论上消除了词汇和语法之间的界限。词汇—语法连续统观产生于语料库语言学起步时（McIntosh 1961；Halliday 1966；Sinclair 1966）。虽无法断言功能语言学、认知语言学中的类似观点是否源于语料库语言学，但至少在方法论层面，大型语料库的出现以及词语搭配、短语框架自动抽取技术的发展，使我们比以往任何时候都更易于充分观察和描写词汇和语法之间的语言单位，进而有可能在此基础上，提炼出基于短语的新语言学理论。Sinclair 开创的理论之路，在后经典时代仍在延续，虽从者不多，但影响不小。目前，语料库语言学研究中常用的语义韵、话语韵、局部功能等术语大多来自 Sinclair 的扩展意义单位模型和局部语法理论。

发轫于经典时代、成长于后经典时代的语料库语言学理论探索以 Biber 开创的“语域研究”发展最快，现已完成理论体系建设，并通过“口笔语语法”（Biber *et al*. 1999，2021）进入主流教学语法中。此外，同时期的局部语法研究则另辟蹊径，从微观层面出发，为凝练语言单位的用法构型提供行之有效的路径。当前，局部语法与言语行为理论的结合，从理论和实践上扩充了语料库语用学的研究内容。

在后经典时代，语料库建设的深化和多变量统计的普及，使得 Quirk（1960）提出的“穷尽阐释”（total accountability）原则和 Labov（1972）倡导的“可阐释性原则”（principle of accountability）得到越来越充分的体现。穷尽阐释原则强调语料的代表性，据此可以实现更高程度的描写充分性。在后经典

时代，语料库方法多元且全面，一定程度上已经触及语言使用的机制和规律。语料库语言学家从关注特定语言单位的多用和少用入手，然后拓展至探索多项语言特征在语境中如何使用，进而逐步探究“为什么如此使用”这样的深层次问题。

综观近六十年来的语料库语言学研究，工具技术、方法设计与语言理论紧密联系。在持续更新的技术支持下，语言研究的方法与设计日趋精准化、多元化，这使得语言描写的维度广泛而细致，全面而深入，在很大程度上深化了我们对语言的理论认识。此处可以用一个形象的类比：将搜索项视为一个点，索引行看作一条线，同主题文本中该搜索项的所有索引行构成一个面，若干主题文本中的词语和语法使用则勾勒出真实文本使用的体。在“点—线—面—体”的研究推进中，结合从简单到相对复杂的统计技术，语料库研究方法已经丰富了我们对语言的理论认识，将来也会持续为语言研究带来新的方法论和理论贡献。

综合来看，从经典时代到后经典时代，语料库语言学的发展走过了从点到线，从线到面，从面到体的历程。这两个时期的语料库研究，在词语和语法范畴频数统计的基础上，都注重从语言特征的共选规律上探讨意义和功能，这或许是今后语料库语言学理论构建的重要立足点之一，出现多因素语法或共选语法，也不无可能。

1.5　结语

本书所谈的语料库研究方法包括以下几部分：(1) 语料库建设方法，解决的是文本从哪里来的问题。(2) 语料库检索、统计、对比、建模、可视化等分析方法。这部分是全书的主体，也是语料库研究方法的核心内容。(3) 理论阐释的方法。例如，搭配分析、局部语法、多维分析、概率语境共选分析以及其他语言学理论的阐释方法。这部分有的单独成章，有的融于其他章节。语言本身是复杂的，本应以全息式视角和整体观进行综合考察。后经典时代的语料库研究方法已经能够比较充分地实现统筹兼顾，我们应该勇于尝试和采纳相关的新思路和新方法。

目前看来，后经典时代研究方法存在的弊端是在操作上有把简单问题复杂化的现象。在数据结果呈现方面仍需简明扼要，读者友好，同时也要有利于后续的理论阐释。另外，也必须清醒地认识到，复杂的算法不一定优于个人的语言直觉。语言实践是检验语言研究方法有效性的基本准绳。

近些年，受文本挖掘相关技术的影响，存在将语言看作“一袋子词”（a bag of words）的趋势。这种“词袋模型”（bag-of-words model）将语言使用视为书写符号的机械物理过程，这在某种程度上背离了语言用以达意传情的基本特性。语言是“有序的异质体”（Weinreich 1968），存在一定的自组织规律，即存在语言要素、语言与语境要素相互依存、共选共生的机制。语料库语言学应更加关注语境中的意义表征和形义共选现象。立足形义对应关系，回归意义研究，才应该是语料库语言学需要始终恪守的。母语习得之易，二语习得之难，均在于具体使用情境中的形义对应这一问题上。

从经典时代到后经典时代，用于研究的核心数据呈现出从线性文本到行列式数据的转变。今后语料库研究方法的发展或将进入以行列式数据为主，结合线性文本语境化分析的新发展阶段。

更进一步，我们也要注意到工具技术、方法手段与语言解释的辩证关系。以新近流行的 ChatGPT 为例，该人工智能机器人背后的大数据和神经网络算法，可比作语料库研究中的工具与方法，它产出的文本可看作语料库研究中使用各种方法而输出的数据。人工智能的输出结果，仍然需要人来解读。换言之，方法工具只是人们理解语言的辅助手段，对语言本质的深入探究，仍是语言学家的根本任务。