

目 录

致谢.....	xi
第 1 章 词与词义：另辟蹊径.....	1
1.1 遣词造义.....	1
1.2 规则支配行为中的语言能力.....	6
1.3 通过比较建构意义.....	7
1.4 常态用法的拓展.....	8
1.5 开放选择与习语约束.....	15
1.6 一个基于词汇的语言学理论.....	17
1.7 本体.....	18
1.8 证据与直觉.....	20
1.9 本书概览.....	22
1.10 小结.....	23
第 2 章 什么是词？.....	24
2.1 “词”的概念之争.....	24
2.2 语言词库：有限与无限之辨.....	28
2.3 齐夫定律.....	30
2.4 动态词库.....	31
2.5 专有名词.....	32
2.6 新术语诞生的路径探寻.....	40
2.7 科学家用语.....	43
2.8 语境锚定.....	49
2.9 多词表达.....	50
2.10 启示.....	63
2.11 小结.....	64
第 3 章 词义是否真实存在.....	65
3.1 引发深思的词义谜题.....	65
3.2 常识.....	66
3.3 奥卡姆剃刀.....	70
3.4 不兼容成分的和谐共生.....	71
3.5 意义事件与意义潜势.....	73

3.6 小句结构与远距语境	74
3.7 语料库分析的局限性	80
3.8 启示	82
3.9 小结	82
第4章 原型与常态	84
4.1 对传统观点的质疑	84
4.2 意义事件与意义信念：格赖斯的“含意”	86
4.3 常态用法的甄别与判定	90
4.4 意义潜势与短语学	93
4.5 意义、偏好语义学和原型理论	96
4.6 “climb”：实证分析的范例	99
4.7 启示	101
4.8 小结	102
第5章 语境依存与词汇集	109
5.1 模式辨识	109
5.2 用法常态与信念：动词	111
5.3 用法常态与信念：名词	128
5.4 意义潜势在句法中的投射	131
5.5 特定领域的常态	133
5.6 没有释义的词典	135
5.7 创新创造与陈词滥调	135
5.8 启示	136
5.9 小结	137
第6章 因时而变的常态	138
6.1 不朽的碑文	138
6.2 意义与用法的关联常态：以“enthusiasm”为例	139
6.3 常态的拓展与替换：以“enthusiasm”为例	142
6.4 否定和疑问之困	144
6.5 简·奥斯汀作品中“enthusiasm”的意义	145
6.6 简·奥斯汀作品中“condescension”的意义	150
6.7 常态、交互信念与社会地位	155
6.8 意义变化的常见例证	156
6.9 多个本义的词	157
6.10 小结	158

第 7 章 三种替换类型.....	159
7.1 语义核心区	159
7.2 词汇替换	160
7.3 语义类型替换	162
7.4 句法替换	171
7.5 启示	197
7.6 小结	199
第 8 章 拓展.....	200
8.1 拓展的定义	200
8.2 拓展的类型	204
8.3 修辞格是否皆为拓展	214
8.4 双关	225
8.5 理解复杂的拓展	227
8.6 模式歧义的拓展	228
8.7 旧词拓展造新词	230
8.8 普通语境消歧	231
8.9 低频拓展词与偶然拓展现象	233
8.10 误用与拓展的区分	234
8.11 启示	237
8.12 小结	238
第 9 章 互文性：文学与常态的拓展.....	240
9.1 词汇的互文性	240
9.2 古代寓言和民间故事的生命力	241
9.3 通俗文本的互文性	246
9.4 诗歌词汇常态的拓展	248
9.5 莎士比亚的影响	252
9.6 《圣经》的影响	256
9.7 文字游戏	258
9.8 拓展的极端形态及其他	262
9.9 拓展的终极形态	264
9.10 语言证据、戏剧和小说	268
9.11 小结	271
第 10 章 词与模式的意义：复杂的语言格式塔.....	272
10.1 语义与组合关系的复杂性：程度之辨	272
10.2 拓展用法成为衍生常态的轨迹	277

10.3 拉丁语和希腊语的衍生常态可为英语的原生常态.....	292
10.4 小结.....	293
第 11 章 意义、语言哲学与人类学.....	294
11.1 优先性：证据先于理论.....	295
11.2 亚里士多德：词汇语义学与定义.....	298
11.3 启蒙时期：威尔金斯到韦日比茨卡.....	300
11.4 维特根斯坦：词义的可变性.....	311
11.5 奥格登与理查兹：语义三角理论.....	316
11.6 格莱斯与奥斯汀：日常语言哲学.....	317
11.7 罗什和普特南：词汇分析中的原型和刻板印象.....	320
11.8 小结.....	330
第 12 章 词库在语言学理论中的作用.....	332
12.1 语言学的理论流派.....	332
12.2 欧洲结构主义语言学中的词库.....	333
12.3 俄罗斯研究传统.....	339
12.4 生成语言学中的词库.....	348
12.5 普斯特约夫斯基：生成词库理论.....	360
12.6 认知语言学中的词库.....	363
12.7 菲尔墨：框架语义学和框架语义网.....	366
12.8 构式语法.....	371
12.9 弗斯学派.....	373
12.10 结论.....	386
12.11 小结.....	388
第 13 章 未来研究展望.....	389
13.1 使用词汇，创造意义.....	389
13.2 常态与拓展理论总结.....	390
13.3 语言规则与语言数据.....	393
13.4 理论与应用.....	396
13.5 结论.....	405
注释.....	407
参考文献.....	414

第 1 章

词与词义：另辟蹊径

1.1 遣词造义

人们如何遣词造义？自柏拉图时代起，这个问题便激发了不少智者对人类语言本质的好奇，同时亦使其为之困惑不已。虽然迄今已有种种答案，但是无一令人满意。若能找到令人满意的答案，我们便可通过计算机编程精准处理自然语言的意义，实现地道流畅的机器翻译，还能直接教人正确运用外语。然而，出于种种原因，这些艰巨挑战至今依旧棘手，本书会涉及其中部分原因。所以，参与到这场纷争里，于笔者而言着实如履薄冰、如临深渊。毕竟，古往今来众多名垂青史的人物在这一问题上都鲜有建树，笔者又凭什么能拨云见日、有所突破呢？即便如此，还是值得一试。原因如下：

首先，如今我们有可能重新审视语义实证分析的理论基础，重点聚焦于词汇层面。这是因为在过去二十年里，出现了新的证据形式（比如语料库、互联网），而且新的统计分析及推理方法也已应运而生。遗憾的是，这些新方法仍未得到充分利用，人们对其也缺乏透彻的理解。即便到现在，也鲜有语言学家试着去深入探究有关**搭配和原型（prototype）**的统计概率问题，而我们将会了解到，搭配和原型恰恰是语义的基础。就连部分语料库语言学家也易受此批评。简言之，搭配是指文本中语词之间的共现关系。具有统计学意义的搭配关系，其出现频率高于随机概率。这种共现现象需要通过大型语料库的计算分析来度量。Church & Hanks (1989) 展示了使用点间互信息 (pointwise mutual information) 进行分析的方法。此后，涌现出诸多关于语料库搭配研究的论著，其中部分还采用了其他各类统计方法，这些研究对应用语言学各领域均产生了一定影响。

“**认知原型（cognitive prototypes）**”这一概念由 Rosch (1973a, 1973b) 首次提出，其目的在于取代针对自然语言概念进行“**定义（definition）**”的传统做法。

本书的观点是，许多认知原型均可有效地映射为短语原型，即词汇共现频率超出随机概率的短语。

不过，也有语言学家觉得，词义根本无法成为科学研究的对象。例如，杰弗里·桑普森（Geoffrey Sampson）在一部有关语言数据统计分析的深奥著作里断言：“在科学与人文的分野中，词义更偏向人文学科。”倘若果真如此，那情况恐怕会更糟，毕竟表达意义是语言的核心功能，理应对其展开严谨的科学研究。（Sampson 2001）

仔细分析就能发现，Sampson 反对的是两点：一是对词义成分分析的某些简单化假设，二是用准数学符号对词义做形式化分析的做法，这其中包括 Katz & Fodor（1963）所提出的“语义标记”。作为词汇分析专家，笔者认同桑普森的反对意见，但笔者的结论并不悲观。正如桑普森在著作最后一章中所指出的那样：

词义同人类创新思维存在紧密联系，这一情况显示出词义具有不可预测性，所以它难以被纳入科学实证研究的范畴……试图对人类语言的意义作出严格科学的阐释，或许就如同追逐彩虹一般，只是徒劳之举。（Sampson 2001）

他提议采用一套名为“意义假设”（*meaning postulates*）的推理规则，以此替代对词义的形式化表征。比如，依据这些规则，从“A buy B from C”能够推断出“C sell B to A”，反之亦然。这体现出意义假设依赖于组合常态（*syntagmatic norm*），也就是搭配和句法依存关系。而组合常态是可以借助语料库数据进行统计测量与推测的。不过，有时这种推论并不成立。例如，在“Geoffrey doesn't buy generative grammar”这句话里，并不意味着“generative grammar”[生成语法]属于一种商品，也不存在相应的支付行为。恰恰相反，此处的“generative grammar”并非买卖活动中常见的产品或原材料，所以这个句子很可能另有他意，实际上它表达的是“Geoffrey doesn't believe in generative grammar”。“生成语法”是一个命题系统，通常情况下，当“buy”后面接一个表示命题的词作为直接宾语时，便不涉及买卖行为，而是关乎一种信念的激活，这是动词“buy”的衍生常态之一。

通过上述讨论，我们可得出如下结论：桑普森正确地指出，试图用类似数学符号对词义进行严谨的形式化表征，不过是徒劳之举，但他不应就此停步。桑普森所倡导的、用于其他语言分析类型的统计方法，理应拓展至人类语言的意义领域。虽说这项工作不像追逐彩虹那般虚幻无果，可确实需要在严谨性方面作出一定妥协，这与其他形式的语言统计分析情况类似。尽管词义具有动态性，但这并不意味着它无法被测量。测量的途径包括研究大量实证证据，记录其实际使用情

况，以及基于使用模式进行推理等。所以，词义研究应被视作新兴实证语言学的一部分。

认知语言学 (Langacker 1987, 1991a, 1991b, 1999, 2008; Croft & Cruse 2004) 和构式语法 (Fillmore et al. 1988; Goldberg 1995, 2006) 都将心理词库和词义置于核心位置，这与早期美国语言学研究方法形成了鲜明对比。诚如 Jackendoff (2002) 所说，早期那些研究方法有着“句法中心论”的偏好。不过，尽管认知语言学极为重视“规约”这一概念，却没能充分阐释如何识别规约用法以及它们具有的特性。本书旨在填补这一空白，或者至少提供一种填补的思路。要识别规约用法，就得去探寻不同说话者和写作者所共有的词汇使用模式，而这属于一项社会语言学任务，唯有通过比较不同文本里每个词项（即概括词）的大量用例，方可达成目标。好在 20 世纪末语料库语言学技术的发展，让这项任务具备了实现的可能。

当下，诸多理论问题的研究路径仍依赖编造的例证，这些例证常基于随意构想且脱离实际的非规约性情景与结构¹。这是因为遵循传统路径的学者试图找寻一种能解释所有可能句子及意义的统一理论，却对常态和可能的区分含混不清。本书基于对实际使用证据的全面调研，系统地区分了常态用法与非常态用法。语言中，常态与非常态用法之间并无明确界限，而是呈现渐变关系，即部分用法相较其他更为常见。可能与不可能用法之间同样不存在严格边界：凭想象得出的可能性会愈发离奇，直至违背某些核心语言原则，沦为不被使用的语言。这种模糊地带意味着，我们尤其需要研究的不只是真实用法，还应涵盖那些经反复使用得以强化的真实用法。总体来讲，传统语言学在这方面做得并不充分。20 世纪后期的语言学研究大多依靠编造例证，通过母语者对这些例证的接受程度来划分合乎语法与不合语法的界限。然而如今我们明白，这样的界限其实并不存在，或者至少难以清晰划定。常态与非常态的语言用法（涵盖符合语法和不符合语法的情况）之间是渐变关系，而非界限分明。

传统上依靠编造语言数据来进行思辨研究的方法着实令人遗憾。这种依赖编造例证且聚焦边缘情况的做法，让语言学陷入了诸多虚构可能性的泥沼，同时却忽视了语言系统中那些核心、典型的常态用法，甚至觉得这些用法太过浅显，不值得深入探究。正如 Sinclair (1991) 在点评 1987 年之前的语言学研究时所说：“因缺乏充足的数据，语言学陷入了停滞状态。”要知道，编造的数据说到底并非真实的数据，仅仅是一种推测罢了。

在语言分析中，诸多看似显而易见的现象，其实只有一部分是真实的。确立词汇的常态用法颇具挑战性，若缺乏充足的实证证据以及关于原型常态的理论支撑²，这项任务几乎无法完成。语料库中偶尔会出现一些看似古怪的语句，虽说令人难以置信，但它们确实是真实存在的。

- (1) Always vacuum your moose from the snout up, and brush your pheasant with freshly baked bread, torn not sliced. (Example cited by Judy Kegl <personal communication, c. 1986, on first looking into a corpus>, from *The Massachusetts Journal of Taxidermy*, quoted in an article in a corpus of Associated Press newswire texts) [用吸尘器吸麋鹿时，务必从其鼻孔处往上吸；用刚出炉的面包擦拭野鸡肉时，要用撕开的面包，别用切片面包。]

例(1)源于笔者的记忆，因无法重新查阅早期语料库，只能凭借记忆来描述这一例子了。显然，该例在多个方面都偏离了常态用法，比如，“moose”并非“vacuum”常见的直接宾语。这类零星的怪异引文，令词典编纂者与语言学家在面对某些看似显而易见的问题时慎之又慎，例如：在英语中，“vacuum”通常与哪些宾语搭配？或者，该动词本身是否常用作不及物动词？

对此，本书有以下两点思考：

1. 真实的经验证据虽是语言分析的基础，但仅有真实性还不够，我们还需依靠规约化的证据展开全面分析。

2. 笔者仍清楚记得 25 年前听到的这个句子，这意味着它以某种方式在笔者的记忆中留下了深刻印记，也就是具有**认知突显性 (cognitively salient)**。笔者认为，至少从认知角度来讲，笔者和其他人一样，都会对这个句子难以忘怀。它之所以令人印象深刻，正是因其不同寻常。尽管 1986 年笔者接触过成千上万个日常语句，可记住的却没几个。大部分日常语言，那些频繁出现且具有**社会突显性 (socially salient)**的模式，往往呈现常态性，反倒不容易被记住。这表明认知突显性和社会突显性是两个相互独立（甚至可能是相反）的变量。倘若这一观点成立，那么也就解释了编造的例证具有误导性的另一个缘由。

在词典学领域，类似的问题也屡见不鲜。例如，有些词典（无疑参考了若干真实引文）收录了“to newspaper”这一动词短语，释义为“阅读报纸”或“从事报纸工作”。但出人意料的是，它实际表达的意思是“用报纸覆盖（物体表面）”。这类词典的编辑并未深入思考如何区分常态用法与所有可能用法，也没有探讨记录所有可能用法的可行性，更不曾研究若将**常态用法**作为词典编纂目标时，该如何定义和识别它。

我们之所以采用语料库驱动的方法来重新审视语言理论与词汇规范，另一个原因在于：即便在语料库语言学领域，常态用法与真实存在的非常态用法之间，始终缺乏令人满意的界定。语料库语言学的先驱约翰·辛克莱 (John Sinclair) 早在 1966 年就预见到了词库实证研究的发展前景，他指出：

如果我们想要研究词汇组织的“形式”方面，就会面临诸多复杂问题，而倘若不借助大型计算机的帮助，这些问题几乎是无法解决的。
(Sinclair 1966)

如今，我们所拥有的计算机设备早已远超 1966 年时所能想象的程度，不过在词汇组织形式的实证研究上，我们仍处于起步阶段。部分原因在于，“形式”这一概念往往局限于遵循严格句法规则的框架内，从而忽视了概率因素的作用。辛克莱本人在词汇搭配研究领域有着开创性贡献，他通过创立 COBUILD 项目，推动了词汇计算领域的发展。该项目的成果之一便是 1987 年出版的《柯林斯 COBUILD 英语词典》(*Collins Cobuild English Language Dictionary*)³ 及相关论文集 (Sinclair 1987)。搭配分析以统计显著性作为基础这一理念，正是由辛克莱率先提出的。此后，Church & Hanks (1989) 以及众多后来学者借助计算启发式算法，进一步发展并完善了这一理念。然而，辛克莱并未将词汇语义学的概念全面整合到搭配研究当中，也未曾提出系统区分常态搭配模式和创新用法模式的标准，而这恰恰是本书所聚焦的核心问题。

在辛克莱所开创的传统之下，Hunston & Francis (2000) 提出的模式语法 / 型式语法 (pattern grammar) 是一项重要进展。不过，在此框架中，词汇搭配在决定“何为模式”的过程里并未发挥核心作用。Hunston 和 Francis 对“模式”这一概念的理解偏于传统，与 Hornby (1954) 的看法相符，他们觉得模式主要由词类和功能词构成，却忽视了搭配的统计显著性。因而，当整体句法结构一致时，模式语法便难以区分单个词的不同意义。例如，在模式语法框架下，“firing a gun” [开枪] 和 “firing an employee” [解雇员工] 在结构上并无差别，因为二者都契合动词 “fire” 的及物用法，模式均为 “V + n”。要理解这两个短语之间的意义差异，听者或读者就得依靠各自的搭配信息，也就是要知晓 “gun” 和 “employee” 所属的类别。

对于 Herbst 等学者 (2004) 以及布拉格学派所倡导的配价分析，我们同样可以提出类似批评。这些学者将配价理论与依存语法发挥到了极致 (像 Daneš 1964; Panevová 1974–75; Sgall, Hajičová & Panevová 1986; Skoumalová, Lopatková & Žabokrtský 2001; Hajičová 2006)，然而，他们对语义类型 (semantic type) 的关注依旧欠缺。尽管他们在句法结构分析方面 (例如针对状语论元选择的研究) 给出了重要见解，可在显著词汇搭配方面的探讨却不够深入，也没有充分阐释一个词的语义类型是怎样影响另一个词的意义。

不同语言流派的学者之所以忽视词汇搭配，“可变性” (variability) 的存在是主要原因之一。语言学家往往追求确定性，可搭配的本质恰是具有可变性，而且词汇搭配的可能性近乎无限。当我们试着罗列某个动词所有可能的直接宾语，或是列

举所有能修饰某个名词的形容词时，简单的回答或许就是：“理论上讲，任何词汇都有可能。”不过，倘若我们进一步追问，通常哪些名词会充当某个动词的直接宾语，或者哪些形容词通常用来修饰某个名词，那我们便能开启对语言中丰富却又可控的语言模式及变体的挖掘之旅，而这些模式与意义的实现息息相关。因此，本研究得出的一项重要结论是：问题的关键在于传统的提问方式，也就是将“can”解读为“所有可能性”。那些试图一次性解释所有可能性的语言学家已然失败，并且注定会失败。若想在这一领域有所突破，我们首先应聚焦于确定常态的、核心的、典型的用法。这一目标可通过分析母语者在大型语料库中产出的书面语或口语数据来达成。接下来，我们要认真对待那些偏离任何标准模式的用法，不能简单地将其视作“错误表达”。对于这些例外情况，应当用另一套规则来解释，这套规则不仅能阐释有意的偏离，还能揭示语言的创造性。

因此，我们需要一种以词汇为基础、由语料库驱动的理论路径。它有助于我们深入探究词汇经组合产生意义的方式，进而颠覆传统的句法导向研究路径。若语料库驱动的搭配分析路径始终依据实际语言行为中的模式证据，那它将有力补充认知语言学、构式语法、启动理论（priming theory）以及模式语法等理论，并为其筑牢根基。事实上，该理论路径的作用不止于此，它不但为检验和测试语言学方法中的理论假设提供了新手段与新标准，在部分情况下，还能对这些理论加以改进。而且，它也会与计算语言学、词典编纂以及语言教学等诸多领域密切相关。

曾有一些著名语言学家断然否定语料库证据与语言学理论之间存在关联性。近年来，也不乏部分学者将语料库当作“鱼塘”，只挑选那些能支撑自身既有假设和理论的“鱼儿”，对与理论不符的则弃之不理⁴。但本书的目标并非如此，而是旨在识别这个“鱼塘”中“鱼儿”的类型。本书后续的理论与实践建议源于对语料库证据的深入剖析，且不依赖于先入为主的假设。关于词库的讨论，是建立在作者毕生从事词典释义撰写与编辑工作的经验基础之上，也就是处理词义解读及相关问题的实际经验。另外，从最早的大型语料库项目——1983年启动的 Cobuild 项目（当时 730 万词就已被视作“大型”），到由 Hanks 开发且目前仍处于研发阶段的《英语动词模式词典》项目，这些项目都针对包含数十亿词的语料库证据展开了深入的词汇分析。Cinková, Holub & Smejkalová (2010) 还对后者进行了初步评估。

1.2 规则支配行为中的语言能力

本书提出的理论着重指出，在习得自然语言时，人和机器都得掌握两种相互关联、受规则支配的行为系统能力。首先要掌握词汇的常态用法与习语用法。而语

言使用者一旦掌握某词的“常态”(norm)用法,往往就会通过各种方式对其进行“拓展”(exploit),使得语言使用变得更为复杂。人们对词汇游戏的探索似乎无穷无尽,更关键的是,语义的可拓展性赋予了语言使用者借现有规则表达新事物的能力。比如,我们可将新体验比作熟悉的体验,让交流对象自行体会两者的相似性。

为词提供精确且毫无歧义的定义属于规定性过程,并非描述性过程。但这一规定性定义过程难免要用专业术语界定该词,使其脱离自然语言中歧义带来的创造潜能。本书旨在通过描述常态用法,编纂具备自然语言词汇显著特性的清单或词典,来阐释怎样“精确地处理模糊性”。这些特性不仅具有可拓展性,还呈现出拓展的方式。正如 Wierzbicka (1985) 所讲:

对于歧义概念的恰当定义必须追求模糊性,而非精确性:这一定义必须要准确反映概念本身的模糊程度。(Wierzbicka 1985)

本书旨在探索一种兼具便于机器处理以及适用于教学特点的精确性。把模糊性视作自然语言的基本属性,这一观点可追溯至维特根斯坦的思想。

当我想给某人拍照时,我可能会说:“大概站这里”。意即只要对方站在我预设范围内,我便会满意。然而,对方或许会站在我预期范围之外,抑或我后续发觉原本觉得合适的区域实际并不理想。而且,即便尝试给出更具体的指示,如“站在这片草叶上”,也难以彻底消除此类不确定性,仍可能出现不尽如人意的情况。(Wittgenstein 1953)

Wilks (1971) 认为,如果存在限定意义的规则,那么必然存在支配限定空间扩展的元规则。

1.3 通过比较建构意义

人们在运用语言描述新情景或陌生现象时,常借助比较的方式建构意义,以下例证可予以说明。例(2)源自《改造旧建筑》一书,书中描述了一种既非门、亦非窗,还非百叶窗,但具备这三者共同特点的事物。此处加引号的介词“like”表明,“door”[门]和“shutter”[百叶窗]两词属拓展用法,并非字面意义,该拓展在限制原词本义的同时,凸显了其他特征。由此,作者无需为这一人工制品创造新术语或精确定义,仅通过拓展现有术语,便能在特定语境中传达其认为必要的意义。

- (2) ...the new infilling panel of clear double-glazing, set well back from the external face of the wall, helps to light a sitting room while a “door”, which can be swung like a shutter to seal this opening, is fixed back against the face of the adjoining flintwork. (BNC) [……新的透明双层玻璃填充板被安置在远离墙体外立面的地方，以提高起居室的采光效果，而那扇具备百叶窗特性的“门”，可像百叶窗一样来回摆动以封闭门口，并被安装固定在相邻的燧石墙面上。]

另一个例证来自《卫报》科学栏目的记者，他在描述专为火星探索机器人设计的新型交通工具“Bridget”时写道：

- (3) She looks like a cross between a remote-controlled tank and Johnny Five, the irritating star of 80s robot movie *Short Circuit*. (<http://www.guardian.co.uk/space/article/0,,1855981,00.html>) [她看起来就像是遥控坦克和80年代机器人电影《霹雳五号》中那个令人讨厌的角色“强尼五号”的结合体。]

这里使用的比喻标记“like”表明“坦克”(tank)和“强尼五号”(Johnny Five)这两个词在此处是被拓展使用，但只激活了最显著的部分特征。例如，虽然Bridget有着坦克的外形，但它并没有坦克的火力，而且装甲非常薄弱。对于那些看过《霹雳五号》(*Short Circuit*)的人来说，“强尼五号”作为双重比喻的第二层意思是对一个特定术语(专有名词)的拓展。然而，对于没有看过该电影的人而言，这个比喻可能会让他们不知所云。优秀的记者深知，他们的读者并不一定具有相同经历，因此，在这种情况下，记者会通过附加注释来提供更多信息。比如，提到这个角色名时，可能会补充说明“《霹雳五号》中那个令人讨厌的角色”，以帮助那些未看过这部电影的读者理解这个比喻。

Hanks (2005) 探讨了介词“like”在意义构建中的重要性。本书认为，如果自然语言在很大程度上是一个由类推、概率和偏好(preference)构成的系统，那么那些最常用于类推和比较的词汇无疑起着核心作用。

1.4 常态用法的拓展

在深入探讨之前，本章将借助一个更为详尽的例证，来阐明词汇的常态用法及其拓展方式。比如，我们可以这样提问：“英语动词‘hazard’[尝试/冒险]通

常是如何使用的？”依据语料库的证据，在包含该动词的英语句子里，超过半数是以“People hazard guesses”[人们大胆猜测]这种形式呈现的⁵。当然，读者或许会马上联想到一些反例。然而，按照当今语料库分析的原则，我们不宜将虚构的例证和反例当作主要数据来依赖。所以，我们会深入钻研那些容易获取的实际数据，以便探寻动词“hazard”的其他用法。基于从英国国家语料库（BNC50，详见附注5）中抽取的一半数据，本研究获得了如下发现：

我们借助 Sketch Engine 中的词性标注工具（Kilgarriff et al. 2004），在这半个语料库中找到了 50 条呈现“hazard”动词用法的语料。其中有六条语料里该词的词性标注出现了偏差，即将其名词或形容词用法误标成了动词。诸如“hazard assessment”[风险评估]、“hazard control”[风险控制]以及“hazard studies”[风险研究]等短语，实际属于“修饰语+名词”的结构，而非“动词+名词”的结构。尽管这些例子中的词性标注存在错误，但这并不代表这些用法缺乏合理性或不常见。通常而言，词性标注中 12% 的错误率是处于可接受范围的。在样本里余下的 44 个动词“hazard”的用例中，有 23 例是以“guess”或“guesses”作为直接宾语的。也就是说，“hazard a guess”无疑是该动词在英语中的原型常态用法，其占比约为总数的 50%。那么，另外 21 例“hazard”的动词用法又是怎样的情况呢？以下便是其中的部分例证：

- (4) Although Phillips **hazarded** some theoretical conjectures concerning the relationship he had uncovered, it was left to others...to provide the analytical underpinning for this astonishing statistical regularity. [尽管菲利普斯对他所揭示的关系**贸然提出**了一些理论性推测，但却把为这一惊人的统计规律提供分析基础的工作留给了其他人……]
- (5) Perhaps we can now **hazard** an attempt at defining “a good reader”. [或许我们现在可以**大胆尝试**给“优秀读者”下个定义。]
- (6) He does not **hazard** any opinions on how costs depend on the complexity of the service! [对于成本如何依赖于服务的复杂性，他并未**贸然发表**任何意见！]
- (7) In the spirit of such theorising, should we limit our visual world to one of rapidly succeeding shapes and colours from which we **hazard** the inference that a leaping dog is in the vicinity? [本着这样的理论化精神，我们是否应当将自己的视觉世界局限于一个快速更迭的形状和颜色之中，并据此**贸然推断**附近有一只跳跃的狗呢?]
- (8) I would **hazard** that the ratio of real balances to total private sector net

worth is less than 1%... [我敢冒昧地说，实际余额与私营部门净资产总额的比率不到1%……]

- (9) “My uncle,” said Wendy, expanding further on her family, “was Provost of Dumfries; he had a rather odd name — ‘Chicken’.” “Not Hen Chicken?” I **hazarded**, as this humorous diminutive was part of my family mythology. [“我的叔叔，”温迪在进一步谈论她的家人时说道，“曾是邓弗里斯的教务长；他有一个颇为古怪的名字——‘鸡’。”“不是‘母鸡’吗？”我试探着问道，因为这个诙谐的昵称是我家族传说的一部分。]
- (10) Although one may from time to time admire people who **hazard** their entire company on one major throw, it has to be a risk that for most of us would be, hopefully, both an unnecessary one and certainly an undesirable one. [尽管人们偶尔会钦佩那些敢于孤注一掷将整个公司押上的人，但对我们大多数人而言，这种冒险想必既不必要，也肯定不可取。]
- (11) Without a clear military objective, and a principle strong enough to **hazard lives** for, America cannot hope to lead by hassling. [倘若没有明确的军事目标，也没有足以让人舍生忘死的坚定原则，美国就别指望通过骚扰来担当领导角色。]

在对动词“hazard”用法的分析中，例（4）至例（9）能够被看作是对该动词原生意义的“拓展”（**exploitation**），与之相对，例（10）和例（11）则呈现出另一种截然不同的、基于常态用法的拓展形式。那么，为何会得出这样的判断？又是基于文本中的哪些具体事实依据呢？

在例（4）至例（9）所构成的语料情境中，“hazard”一词的直接宾语与它的原型直接宾语“guess”在语义类型方面呈现出高度的一致性。举例而言，“conjecture”[推测]一词在语义上与“guess”近乎等同，二者均指代说话者思维活动所产生的言语行为表现。类似的，“an attempt at defining”[尝试定义]与“a guess”[猜测]、“conjecture”[猜想]具有相似性，同属于一种言语行为范畴，只不过相较于口语表达，“an attempt at defining”更倾向于书面语的使用场景。此外，“opinion”[意见]和“inference”[推断]也都归属于代表思维活动的言语行为类别。像例（8）中的“that”从句，其所表达的是一种言语行为或者该言语行为背后所蕴含的主张；而例（9）中的直接引语形式，则清晰明确地展现出一种主张。特别需要指出的是，在例（9）中之所以特意选用“I hazarded”这样的表达，而非相

对更为中性的“*I said*”，其目的在于借助拓展的语义内涵来暗示所表述的内容仅仅是一种猜测性质的言语行为，而非确凿无疑的事实陈述。由此可见，原型常态对于拓展语义的解释具有至关重要的作用和价值。从例（4）至例（9）的一系列语料实例中能够明显察觉，其语义值呈现出从原型常态“*hazard a guess*”逐步向外拓展延伸的趋势和迹象。值得一提的是，部分美式英语的使用者反馈，在例（9）中这种涉及动词“*hazard*”与直接引语搭配使用的方式并非他们日常习惯采用的表达形式。尽管在英式英语的实际运用中，这种拓展方式也并不常见，但恰恰是这种相对罕见的特性，在特定的语境下反而能够使沟通交流更加富有成效。

相比之下，例（10）和例（11）中的直接宾语隶属于与之前例证全然不同的语义范畴，这种差异激活了动词的另一种语义内涵，此内涵与 Fillmore & Atkins (1992) 深入研讨的“*risk*”[风险]的意义存在相似之处。借助对话料库证据的进一步剖析，本书发现“*hazard*”的宾语范畴不仅涵盖生命、公司或商业企业，甚至能够拓展至财富、名誉以及其他具有价值的实体领域。这就意味着“*hazard*”的宾语往往是具有一定价值的事物，而实施这一行为的目的通常在于试图获取某种利益回报。

这两种常态用法之间并非毫无关联、完全割裂。从历时性的视角来看，“*hazard* + 某一言语行为”这一概念其实是从“*hazard* / 赌博 + 掷骰子游戏中的金钱等有价物”这种用法逐步拓展演变而来的。在社会文化层面上，讲真话的能力曾被当作一种具有价值的物品来看待。然而，这种历史文化层面的意义与该词在现代英语中的语义和实际用法已不存在直接关联。在众多情形下，一个词的现代语义是从其早期语义逐渐发展演变而成的。有时，这新旧两种词义会并行存在达数个世纪之久；而在另外一些情况下，新的词义则有可能彻底取代旧有的词义，从而成为该词在现代语境中的主导语义。

1.4.1 词汇集与语义类型

在词汇的语义与用法体系中，众多词汇的大多数用法都较为清晰直观，能够被明确地划分至常态用法的范畴之内，而另有部分用法则显著呈现出对常态用法的拓展。不过，在实际的语言运用过程中，常常会出现处于常态与拓展这两种状态之间的模糊区域，这种情况无疑使得语言分析工作变得更为错综复杂。诚如本书后续章节将会深入探究的那般，不同的意义或者使用模式之间的界限往往也并非清晰可辨，而是呈现出一种模糊的状态。

其核心要点在于，一个词可能具备多个不同的意义，并且每种意义都与一种或者多种特定的使用模式相互关联。进一步来讲，每一种意义或者使用模式均有可

能以多样化的方式进行拓展，而且在这些分析层面上所涉及的各类边界通常都缺乏明确的界定。从某种程度上来说，分析人员能够自主选择运用何种分析工具，并且有权决定在何处划定边界，以此来实现不同程度的精确性。因而，就“hazard”这个词而言，可以按照前文所阐述的方式，极为严格地将其常态用法限定为“hazard a guess”这一具有典型代表性的短语⁶。当然，也可以采取一种更为宽泛的界定方式，把常态用法定义为“hazard + [[言语行为]]”⁷。不过，这种归类方式或许会由于追求更广泛的包容性，而在一定程度上牺牲掉部分语义的精确性。尽管存在着多种处于边界地带的可能性情况，但这些边缘情况的存在并不意味着二者在语义方面毫无差异。

存在另外一些动词与“hazard”情况类似，即它们通常与某个特定的名词紧密相连，作为典型的直接宾语，而大部分动词则能够与多个不同的名词短语产生关联，这些不同的名词短语往往会激活该动词的不同义项。例如，动词“fire”在某一特定的义项中，与一系列指代枪械的词汇存在密切的联系，像“guns”[枪]、“rifle”[步枪]、“pistol”[手枪]、“revolver”[左轮手枪]、“machine gun”[机枪]、“Kalashnikov”[卡拉什尼科夫枪]等等。这样一组在语义上紧密相关的词汇，共同构成了与动词“fire”相对应的“词汇集”(lexical set)，并且它们共享着一个统一的“语义类型”(semantic type)，也就是说，这些词汇都归属于枪械类。这样的词汇集通常会被赋予统一的语义类型名称，诸如[[枪械]]和[[实物]]等，并且按照惯例采用双方括号的形式进行标记。正如 Hanks & Ježek (2010) 所明确指出的那样，语义学家致力于将这些纷繁复杂的语义类型纳入到一个被称为“本体”(ontologies)的分层结构体系之中。因此，从这一结构体系来看，[[枪械]]隶属于[[武器]]这一类别，而[[武器]]又归属于[[人工制品]]，进而[[人工制品]]属于[[实物]]等层级关系。这些特定的词汇集及其所代表的语义类型，与动词“fire”的其他用法形成了鲜明的对比。例如，在“[[人类]] fire [[人类]]”这样的结构中，“fire”所表达的意思是“解雇”；而在“[[人类]] fire up [[机器]]”的表达里，“fire”则具有“启动”的含义。

就词汇集与语义类型的关联而言，其外在表象往往给人以清晰易懂之感，然而深入探究后便会发现，二者间的内在联系实则可能错综复杂、千头万绪。本书后续章节将会从多元的视角出发，并结合多样化的语境，对此主题展开反复且深入的研讨剖析。需要明确的是，本书的核心诉求并非对分层本体论的理论进行详尽阐释，而是着力于构建一种全新的理论架构体系，以此来解读词汇于实际运用情境下所蕴含的意义，以及这些意义之间的相互作用关系。在针对词义类型与本体论展开探讨的进程中，当前唯一能够确凿认定的是，尽管这些概念在初次接触时，可能会被视作简单直观的常识性内容，但是当将它们切实应用于诸如词义消歧这类实际任

务时，截至目前为止所进行的种种尝试，尚未能够达成预期的理想成效。这种状况的出现，至少在一定程度上是由于我们对不同类别之间所存在的模糊边界缺乏足够的认知与重视，严重低估了其可能带来的复杂影响⁸。

1.4.2 极端拓展的典型案例

在即将完成本书基于语料库的初次拓展性分析之际，不妨让我们再次将目光汇聚到动词“hazard”之上，进而考察一个更为繁杂的拓展用例。

- (12) I **hazarded** various Stuartesque destinations like Florida, Bali, Crete and Western Turkey. [我试着猜测了几个斯图尔特式的目的地，比如佛罗里达、巴厘岛、克里特岛以及土耳其西部。]

研究发现，众多读者，尤其是那些从事计算语言学或者具有逻辑学倾向的读者，当他们脱离文本的上下文语境，仅仅单独地去审视这一特定句子时，常常会产生一种感觉，即认为这句话荒诞无稽、毫无实际意义且严重违背常理，甚至会觉得莫名其妙、难以理解。然而，这种现象从本质上来说，恰恰凸显出了语言学家和逻辑学家在通常情况下所进行的研究工作，以及语料库语言学家在特定情境下的研究活动，都在一定程度上存在非自然性。在正常情况下，读者极少会全然不顾文本的上文内容，而将某一个句子硬生生地从整体文本当中孤立出来进行单独的考量。毕竟，一个完整的文本大体上都会包含开头、主体以及结尾这三个基本组成部分。例(12)实际上是出自朱利安·巴恩斯(Julian Barnes)在1991年出版的小说《尚待商榷的爱情》(*Talking It Over*)。朱利安·巴恩斯乃是一位以文风优雅而广受赞誉的作家，例如《完整评论》(*The Complete Review*)在对这部小说进行评价时，就称赞他为“一位相当出色的文体作家”。所以，我们在理解例(12)时所遭遇的困难，大概率并非是由于作者自身的表达不清或者知识欠缺所导致的。实际上，如果我们能够将该句子放置于其所处的上下文中去进行综合解读，那么它就会显得极为合乎情理、顺理成章，而这种情况也只能通过对常态用法的拓展这一理论来加以阐释说明。关于其详细的上下文内容，可参见例(13)。

- (13) Stuart needlessly scraped a fetid plastic comb over his cranium. “Where are you going? You know, just in case I need to get in touch.” “State secret. Even Gillie doesn’t know. Just told her to take light clothes.” He was still smirking, so I presumed that some juvenile guessing game was

required of me. I **hazarded** various Stuartesque destinations like Florida, Bali, Crete and Western Turkey, each of which was greeted by a smug nod of negativity. I essayed all the Disneylands of the world and a selection of tarmacked spice islands; I patronised him with Marbella, applauded him with Zanzibar, tried aiming straight with Santorini. I got nowhere. [斯图尔特毫无必要地拿着一把散发着恶臭的塑料梳子在自己脑袋上胡乱梳了几下。“你要去哪儿？要知道，万一我需要联系你呢。”他说道。“这可是国家机密。就连吉莉都不知道呢。我只是告诉她带上轻便的衣物。”他仍旧挂着那副讥笑的神情，所以我推测他是想让我参与到他那幼稚的猜谜游戏当中去。于是，我试着猜测了几个斯图尔特式的目的地，比如佛罗里达、巴厘岛、克里特岛以及土耳其西部，可每一次猜测都换来他那带着得意劲儿的否定式点头。我又列举了全世界所有的迪士尼乐园以及一些铺设了柏油路面的香料岛屿；我提到马贝拉试图讨好他，提到桑给巴尔来附和他，甚至还试着直截了当地猜测是桑托里尼岛。然而，我却一无所获。]

在这个段落中，作者至少运用了两种语言拓展手法，即省略法和构词成分的创新使用。

首先，作者巧妙运用了省略法。就拿“I hazarded various Stuartesque destinations”这个句子来说，它实则是“I hazarded a guess at various Stuartesque destinations”的一种简略形式。鉴于前文已经提到“some juvenile guessing game was required”，所以在此处便无需再重复使用“guess”一词了。紧接着，后续五个分句都极为迅速且精准地遵循了这一独特的省略规则，它们借助不同地点或者地点类型的名词，像“Disneylands”“spice islands”“Marbella”“Zanzibar”“Santorini”等来展现关于地点的言语行为。其次，该段落还呈现出了构词成分创新使用的特点。例如，“tried aiming straight”便涉及一种较为复杂的衍生拓展情况。通常情况下，我们会习惯性地使用“aim a gun straight at something”[直接用枪瞄准某物]或者“aim (or fire) a question at someone”[向某人提问]这样的表达方式，但是却很少会用到“aim straight at a destination”[直接瞄准某个目的地]这般表述。不过，只要设定好了相应的上下文语境，即便存在这样的文体复杂性，也并不会对读者理解文本的内容以及作者想要传达的意图造成干扰。一般来说，读者面对这样的表达并不会感到困惑不解。而本书的目的之一，正是要阐释清楚为何会出现这种现象。

为了帮助读者更好地理解例(12)中提到的第二种语言拓展形式，本研究将深入探讨构词成分在造词过程中的创新使用情况。经过多方面语料的细致检索发

现，无论是在该小说的其余部分，还是英国国家语料库（BNC）里的其他语料当中，甚至是规模达数亿词的更大体量的语料库内，均未发现“Stuartesque”一词的使用踪迹。尽管在网络文本里偶尔能够看到该词用于表示“斯图尔特（Stuart）式的”之意，但每次所指代的“斯图尔特”都有所不同。然而，这并不意味着该词在上下文中毫无意义或难以理解。在小说中，“Stuart”是一个角色的名字。在英语里，词尾“-esque”常附于专有名词之后，使其转类为形容词，进而用于描述那些与所命名的人物或地点存在相似之处的事物，而且通常这类事物带有某些显著的，甚至是颇为古怪的特征。像这样的创新用法在牛津英语语料库（Oxford English Corpus）中不乏大量实例，比如“Kafkaesque”[卡夫卡式的]、“Bergmanesque”[伯格曼式的]、“Monroesque”[梦露式的]、“Hollywoodesque”[好莱坞式的]、“Dylanesque”[迪伦式的]、“Jaggeresque”[贾格尔式的]、“Caravagguesque”[卡拉瓦乔式的]等表达。

本章之所以选定对动词“hazard”展开详尽深入的分析，其关键缘由在于该动词的词义呈现出一种在简单性与复杂性之间徘徊游移的独特状态，而这种特性恰好能够极为有效地阐释与支撑本书所提出的核心论点。诚如后续章节将会更为深入地探究与揭示的那般，许多其他词汇在语义方面所展现出的复杂程度远超“hazard”。

1.5 开放选择与习语约束

在着手开展词汇数据的实证分析工作时，我们不可避免地需要借助一些复杂的工具，之所以如此，主要是因为这一过程涉及多个不同维度的变量。在这些变量中，习语原则（the principle of idiomaticity）与开放原则（the principle of openness）二者之间所存在的张力表现得尤为显著。诚如 Sinclair（1991）所阐述的那样，这两个原则形成了颇为鲜明的对比。其中，“开放选择原则”（the open-choice principle）秉持这样的观点：

语言乃是大量复杂选择的结果，即在语言的每一个完整单元里，无论是词、短语还是分句，都存在极为广泛的选择空间，而唯一对其起到限制作用的是语法规则。

相对而言，“习语原则”（the idiom principle）则主张：

许多语言选择与外部世界之间的联系微乎其微。……语言使用者能够运用大量的半预构短语，而这些短语共同构成了一种单一选择。

正如前文所述，部分词汇在习语搭配方面展现出较为狭窄的偏好范围，动词“hazard”便是其中的典型例证。与之形成鲜明对比的是，像动词“accumulate”[堆积/积累]等其他词汇，则呈现出极为宽泛的搭配偏好，几乎不存在显著的搭配偏好限制。基于对语料库的搭配偏好所展开的统计分析结果来看，似乎更多地呈现出的是语料库文本自身的主题内容，而非词汇在语言体系中的实际行为模式。除此之外，还有一些词汇处于上述两种极端情况之间。以动词“switch”[转换]为例，它明显倾向于和小品词“on”以及“off”搭配使用，而且常常会以某类机械装置作为其直接宾语。然而，当该词用于表达“用一件物品替换另一件物品”这一主要语义时，其常规的句法结构以及结构内部名词的使用范围就会变得十分广泛。诸如“switch from one thing to another”[从一种物品换到另一种物品]、“switch things around”[调换物品的位置]、“switch between things”[在物品之间互换]，以及“switch”作为不及物动词使用时所表达的“two things can switch”[两物可互换]等表达方式，都属于该词的常态用法。这也体现出了交互动词(reciprocal verb)的一个共性特征。但即便从交互动词的角度来审视，“switch”的直接宾语范围仍然显得格外宽泛。尽管传统观点认为许多(甚至是大多数)动词的直接宾语都具有类似的广泛语义类型，但就实际情况而言，“switch”的这种用法更倾向于一种例外情况，而非普遍的常态用法。

对于那些行事不够审慎的分析者来说，还存在一个常见的误区，那就是他们往往期望语言现象具备同质性。就拿之前提及的两个例子来讲，我们并没有充足的理由去认定那些适用于动词语义分析的工具，同样能够适用于名词的语义分析，反之亦然。此外，在借助配价结构(valency structure)分析动词意义时，我们常会遇到两种情况：有时若缺少某个特定论元，动词的意义就会发生变化；而有时即便省略该论元，动词的意义依然保持不变。例如，“to treat someone”这一表达通常意味着实施治疗行为，然而当它与“with respect”[尊重地]、“with contempt”[轻蔑地]、“respectfully”[恭敬地]、“contemptuously”[轻蔑地]等方式状语搭配时，如“to treat someone with respect”[待人以尊重]或者“to treat someone with contempt”[以轻蔑的态度对待某人]，原有的“治疗”之意便不复存在。在这种情况下，方式状语(如“with respect”)的运用就决定了动词所表达的特定意义。然而，对于众多其他动词而言，方式状语的存在或者缺失，对其意义所产生的影响并不显著。

诚如本书前文所述，一个词的常态用法与它的拓展用法之间的界限常常是模糊不清的。在着手进行语料分析时，首要必须保证所分析的语料库内的全部文本要么是由母语使用者撰写创作的，要么是由他们口头讲述的，之后再依据语料的权重情况来判定哪些表达应当被看作是常态。那些频繁现身且多次重复出现的短语，无

疑必须被归到常态的类别之中，即便这些短语在历史渊源上有可能是从某个年代更为久远的常态短语衍生而来的，又或者被一些秉持语言纯粹主义观点的人视作为“错误”的表达。从另一方面来看，对于某些呈现形式极为独特的短语，分析者需要审慎地思考，这究竟是一种罕见但属于常态的具体例证，抑或仅仅是一种简单的拓展情况，而不应轻易认定为后者。例如，“The strain was beginning to tell” [压力开始显现] 这样一个句子，它代表了一种颇为罕见的模式，在此句中，动词短语“begin to tell”属于不及物的用法，而且句子的主语是名词“strain”。在牛津英语语料库所涵盖的15亿词汇量里，这种模式的例证仅仅只有六个（此外还存在几处对这种模式的拓展），但它们绝不能被草率地归类为其他常态短语的拓展形式，这是因为此处的动词“tell”具有独特的含义，即表示某种压力开始对人造成明显的不利影响。所以，这样的用法理应被视作是一种较为罕见的常态。实际上，许多习语虽使用频率较低，但在认知层面却有较强的突显性，并且在语义与句法方面也都呈现出独特的性质。

1.6 一个基于词汇的语言学理论

尽管在实际运用过程中面临着一定的难度，但“常态与拓展理论”（**the theory of norms and exploitations**）至少在词汇分析方面展现出了一项显著优势。该理论为我们提供了一条颇具原则性的途径，借助这一途径，我们能够对词汇的拓展用法加以识别，并且将其同与之相关的常态短语区分开来，进而凸显出常态短语的特有模式。倘若意义（抑或是翻译、含意）能够与上下文中的词相互关联，也就是说，和短语模式联系起来，而非仅仅与孤立的词产生联系，那么诸多广为人知的语言学问题便能在很大程度上得以解决。例如，词义消歧程序长期以来所面临的一大难题便是各种可能性的组合呈现出急剧增多的态势。要是学习者能够辨别出一个词的常态用法，又或者自然语言处理程序在生成文本或执行基于知识的信息检索等依赖词义的任务之前，能够把句子中的每个词都映射到特定的搭配或者短语模式之上，那么组合激增的问题将会得到有效缓解。即便这一过程仅仅只能取得部分成效，该问题的严重程度也会大幅度降低。所以，基于常态与拓展理论所构建的、适用于英语及其他语言的模式词典，理应成为语言教学以及自然语言处理领域重点关注的范畴。就目前的情况来看，自然语言处理领域现有的可用资源往往只是基于对字面意义以及词汇语义关系的猜测与假设，缺乏由深入研究和精细的词汇分析所提供的支撑。截至2010年，知识富集型自然语言处理方法通常收效甚微，而与之相对的，知识稀疏型的统计方法却在过去的二三十年里取得了颇为显著的成绩。这一现象意味着，知识富集型路径或许包含了对语言处理而言看似丰富但实则有误的知识，又

或者是建立在错误的假设基础之上，再不然就是缺乏坚实的根基。不过，可以预见的是，统计方法终究会遇到发展的瓶颈，而这极有可能促使人们重新回归到人工智能领域，去探索如何运用复杂的词汇网络来对世界、实体以及事件进行表征。常态与拓展理论展示出了将统计方法同语言知识以及世界知识的表征相结合的可能性，并且代表了一种以词汇为基础、受语料库驱动、自下而上的语言学理论。

自然语言处理与人工智能的应用固然具有至关重要的地位，然而本书的涵盖范畴并不囿于这一特定领域。实际上，本书为语言教师、语言学习者、认知科学家、语言学家以及语言哲学家提供了一种基于实证的研究方法，助力他们更为深入地探究语义的奥秘。对于语言教师、教材编撰者，以及那些期望将重点聚焦于如何以正常、地道的方式运用目标语言词汇，而非徒劳地试图掌握这些词汇的所有潜在用法的学习者而言，他们能够在本书中寻觅到一种理论方法以及一种分析工具，从而得以实现这一目标。这一点显得尤为关键，原因在于语料库语言学在其发展进程中的一项关键发现是：一个词的^{可能}用法范畴远远超出了在缺乏语料库支撑的情况下我们所作出的预测。前文所提及的“hazarding a destination”便是一个生动的例证，而且类似的实例在本书中可谓俯拾皆是。更为关键的是，这一现象普遍存在于任何语料库或者自然语言文本的集合当中。换言之，语料库语言学揭示了这样一个事实：与强调“创造性”的推测性语言学理论预估的情况相比，语言使用中的规律性其实更为显著，而某些非规律性现象则比这些理论预测的更偏离常规。

1.7 本体

在之前的讨论及后续章节中，本书频繁地提及了语义类型及其在分层概念本体中的表征。“**本体**”（ontology）这一术语是从希腊语动词的现代分词衍生而来，其原义为“存在”。在计算语言学领域，该术语至少存在着三种不同的含义：

1. 依据传统哲学的观点，本体论起源于亚里士多德针对科学研究对象所阐述的概念。它所囊括的范围极为广泛，包含了整个世界以及世界当中形形色色的万事万物，其中既有处于物理层面的实体，也涵盖了形而上学领域的抽象存在形式。一言以蔽之，本体论所涉及的就是所有存在着的事物集合。

2. 在计算语言学中，“本体”通常用于指代某一语言的全部词汇，其中至少涵盖实词部分。WordNet⁹ 便是一个典型范例，它构建了英语词汇与概念间的语义关系网络。WordNet 项目已拓展为涵盖欧洲多种语言的 EuroWordNet¹⁰ 项目，当前还催生出全球 WordNet 协会，将其延伸至汉语、韩语、波斯语、阿拉伯语以及诸多欧洲语言领域。WordNet 类型的本体着重呈现词项间的语义网络联系，尤其是层级

关系，即所谓的“IS-A 层级”。该层级在表征自然产物、人工制品等术语时颇为有效。例如：

金丝雀 (canary) 归属于雀类 (finch) 范畴，雀类又隶属于鸟类 (bird)，鸟类属于生物 (living being) 这一类别，生物进一步被纳入有形实体 (physical object) 的范畴，而有形实体则属于实体 (entity) 这一宽泛的类别。

同样地，枪 (gun) 被认定为一种武器 (weapon)，武器属于人工制品 (artifact) 类别，人工制品属于有形实体 (physical object)，有形实体则归属于实体 (entity) 类别¹¹。

不过，针对抽象词汇，表示部分、特征或属性的词汇，以及某些特定的动词类别而言，IS-A 层级的应用效果欠佳，此时便需要运用其他方法来进行处理。

3. 在语义网 (Semantic Web) 这一领域中，本体目前被用于表征实际实体的组织架构与安排形式，诸如人名、地点、地址、日期、事件、日程安排、商品以及产品描述等等，而并非针对词与词义进行表述。

尽管亚里士多德着力于借助语言来描绘宇宙，涵盖了具象化和抽象化这两个维度，但他未曾尝试将语言里的所有词汇以及概念整合并组织成一个层级体系。实际上，真正开创分层概念本体论的既非 Roget (1852)，也不是 Miller 和 Fellbaum (WordNet; Miller 1985; Fellbaum 1998)，而是威尔金斯 (Wilkins 1668)。威尔金斯作为欧洲启蒙运动中的关键人物，其终极目标是构建一种通用语言。更确切地说，他旨在打造一个通用书写系统，该系统中的符号在任何语言里都能依据该语言的发音规则读出，就像汉字在不同汉语方言中，虽读音不同，但写法一致且表意相同。威尔金斯的工作起始于“广泛梳理所有知识，从而确立所有理性存在者共同秉持的观念” (Eco 1995)。在本书的第 11 章将会详尽地介绍威尔金斯其人，以及他对莱布尼茨等哲学家所产生的影响，还会深入阐述他与常态与拓展理论之间的内在关联。

书中屡次提及的证据有力地表明，尽管分层概念本体论 (例如 WordNet) 在充当词项和概念集合的工具时具备一定的优势，然而在将意义投射到实际任务的过程中，其成效却并不尽如人意。造成这种状况的原因之一在于存在过度的还原论思维倾向。具体而言，这种理论将意义与词汇 (或词汇的同义词集合) 孤立地联系起来，而不是结合词汇在短语语境中的使用情况。借助次范畴化和选择限制等消除歧义的标准来精确辨别不同词义的方法，在实际应用中常常难以奏效，因为该方法一般只适用于少数精心挑选或人为编造的例子。对于数量众多的词汇，尤其是动词和形容词而言，更为行之有效的途径是去理解不同的意义与各自所具有的搭配偏好以及句法模式之间的内在联系。本书的第 4、5 和 7 章将会围绕此方面展开全面且深入的探讨与分析。

1.8 证据与直觉

如同许多语言学和语义学著作一样，在奋力探究自然语言的极端复杂性时，本书亦难以做到尽善尽美，不可避免地会存在一些欠缺之处，甚至可能会踏入某些误区。其中，有一个误区是本书竭尽全力想要规避的，那就是效仿 20 世纪语言学中常见的做法，即仅仅转述前人针对某个词汇所发表的所有观点，却全然不去探究该词汇在实际语言运用中的真实情形。与之相反，本书所选用的方式是首先对词汇的实际使用状况展开深入调查。这并不意味着要彻底舍弃必要的概括以及理想化的处理手段，然而描述的充分性原则明确要求，任何理想化的操作都必须与实际的数据相互匹配、相互印证，而绝不能仅仅凭借毫无根据的臆想与虚构。

同样不可取的是为佐证观点而编造示例，这无异于一种循环论证。在语言学的范畴中，直觉是一个极易引发激烈争议的议题，争议的双方都存在不少的误解。一方坚称在语言学的科学探究进程中绝不能依赖直觉，而另一方则认定唯一确凿可信的证据便是母语使用者的直觉。鉴于此，有必要在此清晰地表明本书的立场：凭借直觉去解读数据与依靠直觉来虚构数据，二者之间有着霄壤之别。每一位科学家都会通过自省来阐释数据，但除语言学领域外，未曾有哪位声名卓著的科学家会通过虚构数据来解释已有数据。过去，人们曾长期认为语言学有其特殊性，可以不受常规约束。然而，只要将语言学家虚构的示例与语料库中的真实用法相比较，便会发现这种虚构示例的做法根本站不住脚。Sinclair (1984) 探讨了当时流行的英语教材（部分教材至今依然流行）中虚构示例的习语性以及自然性方面的问题。例如，像“*Prince Charles is now a husband*” [查尔斯王子现在是一个丈夫] 这样的句子，其作者全然没有意识到在没有物主限定词或形容词性修饰语的情况下使用“husband”是极为罕见的情形。在这类陈述句中，英语的语法规则要求明确查尔斯王子是谁的丈夫，或者他属于何种类型的丈夫。语言学家和哲学家们臆造的众多示例以及设定的场景，往往与实际情况脱节，显得荒诞不经，这无疑使得此类问题变得更为复杂棘手。他们之所以采取这样的做法，是因为试图探寻可能与不可能的句法结构或语义之间的界限，只是在这一过程中不经意地违背了自然性以及文本良构性。

辛克莱曾明确指出，“语言学家对于异常现象的容忍程度高得惊人”。这一现象在一定程度上源于人类自身并不擅长精准地报告自身的行为表现。在本章的前文部分（另可参阅 Hanks 1990），本书提出了这样一个观点：在词库研究领域，“社会突显性”（以使用频率作为衡量标准）与“认知突显性”（依据回忆的难易程度来衡量）属于两个独立变量，甚至这两者之间有可能呈现出负相关的关系。也就是说，一个词条的使用频率越高，人们就越难以精确地回忆起并详尽地讨论它的

所有常态用法（可参考前文的例（1））。这个假设的相对弱化版本则认为，社会突显和认知突显仅仅是两个独立变量。为了以一种非正式的方式对这些假设进行验证，读者可以尝试开展一些思维实验。例如，试着凭借直觉写出“chew”“bite”和“swallow”这三个动词的所有常态用法的例句。随后，再写出它们在习语以及字面用法中的例句。接着，请参照第 10.1.2 节中基于语料库数据统计所呈现出的这些动词的语言结构整体的非正式表征情况。届时，你便能发现自己遗漏了多少种用法。鉴于这些动词的使用频率并不高（相较于“take”和“make”等轻动词而言），按常理来说它们的用法应该相对比较容易回忆。倘若你能够凭借直觉准确地总结出这些动词的各类常态用法，那么可以尝试一下更为复杂的动词，比如第 10.2 节中所讨论的“scratch”。通过这样的实验，应该能够使你确信，无论你自身的语言运用能力有多么出色，你的意识都无法全面地捕捉到你在日常使用的所有常见词汇的细微之处和具体细节。

另一个可行的实验是挑选一对近义词，比如“broad”和“wide”，先凭借直觉去判定这些词最为突显的搭配形式。接着，运用诸如 Sketch Engine 之类的工具，在英国国家语料库或者其他大型英语语料库中进行检索，以此来核实这些词在实际运用中的社会突显性状况究竟如何。

最后，你不妨邀请几位英语母语者，做一个非正式实验。让他们对单词“total”进行思考，并通过造句来展示该词最为典型的动词用法。之后，把他们所造的句子与语料库中的实际证据相互对照比较。这些句子是不是真正体现了该词的原型动词用法呢？倘若不是，那背后的原因又是什么呢？关于这一问题更为深入细致的探讨，可查阅本书的第 10 章内容。

需要注意的是，并非所有的词汇都会存在社会突显和认知突显二者之间的矛盾状况。在某些情形下，你对某个词的运用方式也许会和其实际的使用情况毫无二致。而且，有一部分人可能会比其他人更擅长回忆以及描述自己日常的语言使用行为¹²。然而，至少对部分人来说，要完整且清晰明确地描述他们在日常生活中所使用的一些词汇的原型用法，是存在一定困难的，这一点就足以说明，我们有着充分且合理的理由去拒绝仅仅将直觉当作是原型用法的唯一证据来源。

然而，本书在某些情境下也会运用编造的例子以实现对比或阐释的目的。举例来说，为了清晰地阐述致使/起始替换这一语言现象，呈现诸如“*Someone broke the window*”[某人打破了窗户]和“*The window broke*”[窗户坏了]这样的对比性示例，似乎是顺理成章且合理的做法，即便其中某一个示例有可能是虚构的，并非直接源于实际的语言用法。当然，这些被编造的例子务必契合实际使用中所呈现出的模式，而绝不能仅仅依据某些母语使用者主观的认知或者语言学家的个人观点来随意编造。

另外一个特殊情况是反例的运用，反例主要用于表示语法错误的用法。这是因为无论拥有再多的语言数据，都无法确凿地证明一个句子或者其他语言现象绝对不可能发生。

1.9 本书概览

本书聚焦于词汇及其使用方式，其核心方法在于利用大型语料库提取词汇使用模式，以此开创一条分析语言数据与文本的新路径，并深入挖掘词义与用法之间的内在关联。通过对词汇的常态用法与那些有意为之的非常态用法（即“拓展”用法）进行细致对比，本书正式提出了常态与拓展理论。该理论有着明确而重要的目标，旨在为语料库语言学、认知语言学、计算语言学、历史语言学、语言哲学以及语言教学等诸多与语言密切相关的领域开展实证研究筑牢坚实的根基。此外，该理论还为日常语言的意义计算、词典编纂以及教材编写等实际应用注入了新的理论动力。

常态与拓展理论将关注点主要放在对语言实际使用情况的阐释方面，也就是深入探究在认知以及社会这两个层面上，人们究竟是怎样运用语言的。所以，该理论着重聚焦于模式、概率以及预测这些关键要素，而对于那些仅仅可想象的可能性则只是处于次要关注的位置。也正因如此，书中所列举的大部分例证都是源自真实的文本或者文本集合（也就是语料库）。回溯到 20 年前，像这样的研究方法是难以实现的，毕竟当时的语料库所提供的证据远远不足以支撑并验证部分语言学家（尤其是弗斯、辛克莱和韩礼德等）所怀疑可能存在的使用模式。有时候，出于更好地解释某些语言现象的目的，编造一个能够与实际例证形成鲜明对照的例子似乎是一种较为妥当的做法。在本书当中，此类编造出来的例子在编号呈现时会采用斜体格式，而在正文中出现时，则是以斜体并且外加双引号的形式展现，同时还会依据语料库证据对这些例子进行严格且必要的核查，其目的就是要确保相关的描述能够与基于语料库所获取的常态用法证据保持高度的契合。

本书将研究范畴限定于英语这一对象语言，书中所有示例均取材于英语资料。然而，一个极具探讨价值的经验性问题随之而来，即其他语言是否同样受到常态使用规则与拓展常态规则的关联组合所支配，以及这种支配作用的程度究竟怎样。从目前的情况来看，在这一方面，英语似乎并非孤立的特殊案例。建立语言使用的常态并在此基础上进行拓展（有时某些拓展常态的行为会逐渐常规化，进而演变为新的常态），这极有可能是人类语言行为所共有的普遍性特征。但我们也能够设想，在某些特定的文化背景下，对于词汇创造性（也就是对常态的拓展行为）或许存在着更为强烈的贬损态度和观念，甚至会将其直接判定为“不正确”的语言运用方

式。所以，未来语言学领域的一项重要研究任务，便是深入探究常态与拓展理论是否能够有效地应用于对英语之外的其他语言的分析工作中。

1.10 小结

- 对意义的任何恰当解读都必须基于词汇使用方式的相关证据。
- 这样的解读可从大型文本语料库中的搭配模式以及句法依存关系的证据中推导出来。
 - 仅有真实用法的证据并不足够，还需常态用法的证据。
 - 规约性属于程度问题；不存在任何启发式方法能在常规用法与异常用法之间，或者在语法结构正确与语法结构错误的表达之间，划出清晰界限。
 - 真实但罕见的用法常常源于对常态用法模式的有意创造性拓展。
 - 拓展受规则支配。其支配规则与常规用法的支配规则类型不同。
 - 自然语言的使用受两种规则的制约：一是生成常态、常规、语法正确的话语的规则，二是拓展常态的规则。
- 因此，要揭开语义的真相，尚需一个有关常态与拓展的理论。