

语料库语言学

道格拉斯·比伯 苏珊·康拉德 兰迪·瑞潘 著
刘颖 胡海涛 译

清华大学出版社

语料库语言学

道格拉斯·比伯 苏珊·康拉德 兰迪·瑞潘 著
刘颖 胡海涛 译

清华大学出版社
北 京

内 容 简 介

本书是一部介绍语料库语言学的经典书籍。语料库语言学是利用计算机和大规模语料库对语言进行词汇、语法、语义、语篇、语域变异、语言习得、语言历时发展和语言风格进行研究的学科。它利用计算机来标注、检索和统计语料,并对检索的语言实例和统计的语言数据从功能上进行语言学解释。运用语料库的实证研究,能为语言学难以解决的问题提供新的研究办法。本书每章都集中探讨一个语言学问题,用实例详细说明了对其进行定量分析和定性分析的过程。

本书可作为中文、外语等专业高年级本科生和研究生教材,也可供从事语料库语言学、计算机辅助语言研究和自然语言处理的研究者参考。

版权所有,侵权必究。侵权举报电话:010-62782989 13701121933

图书在版编目(CIP)数据

语料库语言学/(美)比伯(Bibe, D.), (美)康拉德(Conrad, S.), (美)瑞潘(Reppen, R.)著;刘颖,胡海涛译. —北京:清华大学出版社, 2012. 10

书名原文:Corpus Linguistics

ISBN 978-7-302-30192-9

I. ①语… II. ①比…②康…③瑞…④刘…⑤胡… III. ①语料库—语言学 IV. ①H0

中国版本图书馆 CIP 数据核字(2012)第 230452 号

责任编辑:马庆洲

封面设计:傅瑞学

责任校对:王荣静

责任印制:王静怡

出版发行:清华大学出版社

网 址: <http://www.tup.com.cn>, <http://www.wqbook.com>

地 址:北京清华大学学研大厦 A 座 邮 编:100084

社 总 机:010-62770175 邮 购:010-62786544

投稿与读者服务:010-62776969, c-service@tup.tsinghua.edu.cn

质 量 反 馈:010-62772015, zhiliang@tup.tsinghua.edu.cn

印 装 者:三河市金元印装有限公司

经 销:全国新华书店

开 本:185mm×230mm 印 张:12.5 字 数:254 千字

版 次:2012 年 10 月第 1 版 印 次:2012 年 10 月第 1 次印刷

印 数:1~3000

定 价:30.00 元

产品编号:039891-01

译者序

道格拉斯·比伯,苏珊·康拉德和兰迪·瑞潘所著的《语料库语言学》是一部经典专著,它介绍了如何运用语料库方法对语言本体及语言应用进行研究。该书由英国剑桥大学出版社于1998年出版。该书是目前为止全面介绍如何用语料库方法研究语言各个层面(词汇、句法、语义、语篇)和语言应用(语域变异、风格和语言习得)的一本书,每章以英语为例,但这里提供的方法可以应用到其他语言。

语料库语言学是利用计算机强大的检索、统计和处理语料的能力,从大规模的语料库中检索符合研究问题的实例,对其进行统计。在大量实例和统计数据的基础上,对研究问题进行定性分析,从功能上对其进行语言学解释。利用计算机和语料库可以对语言各个层面的特征(单个特征或多个特征)进行分析和研究。进行语料库语言学研究需具备两个基本条件:(1)有代表性的大规模语料库;(2)有语料库处理、检索和统计软件。除了可利用免费软件外,最好还应该学会编程。

语料库语言学的优势在于:(1)可以利用计算机的强大功能,进行快速、准确地分析;(2)语料库规模大,所包括的语域全面,文本量大,语言信息范围广;(3)既有定量分析,又有定性的功能解释,对语言的描写全面。(4)语料库方法与以往的方法相比能做出更概括和更全面的调查。因此,基于语料库的方法可以扩大以往调查的范围和调查语言的新应用。

语料库语言学已经成为语言研究的主流,它正对许多语言研究领域产生愈来愈大的影响。

本书的重心是介绍如何利用语料库来对语言现象进行定量分析和定性解释。它没有详细介绍使用的语料库,也没有详细介绍如何收集语料和标注语料,只是简单地列出了语料库和获取语料库的地址。没有详细介绍如何使用检索软件和统计工具。也没有介绍如何编程来处理、检索和统计语言。而是把重点放在对语言实例和语言应用的处理过程、结果和分析上,这也是我翻译此书的主要原因。对语料库的详细介绍和检索软件的使用细节介绍可以参考其他书籍。本书涉及的语言分析之广泛,过程和分析之细致,对于任何语

料库语言学初学者或对语料库语言学感兴趣的读者,都是一本不可多得的好书。

本书章节主要内容如下:

第1章介绍了语料库语言学的研究中心是语言用法研究。通过调查联结模式研究语言使用的特征。给出了定量分析和定性分析的作用,定量分析表示语言特征与语境之间的联结程度,定性分析则对语言用法作出功能解释。本章还给出了语料库语言学的研究方法和应用领域。介绍了本书使用的语料库和分析工具,概述了本书的结构与各章节的主要内容。

第2章介绍基于语料库的词典编纂,研究对象集中在单个词上。主要研究 deal 的词义、词频、搭配、语法功能、在不同语域中的分布以及同义词 big, large 和 great 在语料库中的用法和分布等。其他词的调查采用与 deal 相同的方法,就可以编写一部基于语料库的词典。

第3章介绍基于语料库的语法研究,调查形态特征(以名词化结尾为例)、词类(以名词和动词为例)、句法结构(以 that 补语从句和 to 补语从句为例)以及同义语法结构(以 that 主语从句与其后置从句为例)在不同语域中的分布,了解其用法和功能。

第4章调查词汇和语法结构的联结。本章首先以两个近义形容词 little 和 small,以及两个近义动词 begin 和 start 为例,从它们的语法联结偏好不同来区分它们。其次,通过调查 that 补语从句和 to 补语从句的不同词汇联结来区分这两个近义语法结构。

第5章分析语篇结构。语篇分析是研究超出句子范围的语言特征。语篇特征分析比词汇和语法特征分析难得多,但对描写语言学和应用语言学十分重要。本章主要分析了不同语域中名词和代词的使用情况及其影响因素,统计了其频率和分布。还调查了语篇中动词时态和语态的使用情况,比较不同时态和语态在实验科学论文中引言、方法、结果和结论部分的频率和分布,从而了解实验科学论文的语篇展开模式。

第6章研究了语域变异和特殊用途英语。首先分析了关系从句、原因状语从句和补语从句在不同语域中的分布,这与语域的交流功能有关。重点调查了不同语域(以英语口语和书面语为例)的差别、特殊用途英语(以生物学和历史学论文为例)在使用上的差异以及同一个语域(以生物学论文为例)内部各部分(引言、方法、结果和结论)的变异,使用多维度分析方法,运用大规模语料库,调查多个语言学特征的共现和变异模式,来对这些领域进行研究。

第7章,研究本族语儿童和第二语言学习者语言的习得与发展。本章从研究个别语言特征,到运用多维度分析方法,来调查不同层次本族语学生的语言发展趋势,并对小学生和成人语言进行比较。针对第二语言习得,调查非本族语学生说英语时所犯的错误。

第8章,语言的历时发展与单个作者风格研究。本章运用语料库分析技术主要调查了在过去三个世纪,情态动词和半情态动词的使用形式变化、书面语和口语的语言学特征变化、男性和女性的语言变化。还调查了某一作家特定作品的语言风格。

第9章回顾了语料库方法对语言学调查的贡献,语料库方法可用于研究单个词汇、语法结构、语篇模式、各类文本、语言习得、历时比较和文体风格比较等。语料库方法的研究广度远远超过了本书所给出的实例范围,本书每章给出实例就是为了举一反三,其他实例的分析乃至各个层面的语言学研究都可以按照类似的方式进行。

附录1包含的10个方法箱分别介绍了语料库研究的一些重要方法。进一步帮助读者了解代表性的语料库、语料库赋码、语料库检索和统计、词汇统计量度、显著性检验以及多维分析等。

本书由刘颖和胡海涛翻译,并对全书进行统一修改、审阅并定稿。李惠、余畅、王曼和刘江洋参与了部分章节的初译,在此表示感谢!在翻译过程中,译者力求忠实于原著,并尽量用通俗的语言把原文准确、清晰地表达出来。

由于译者水平有限,翻译中难免会出现一些不妥之处,译文中也存在许多生硬之处,欢迎广大读者批评指正。

刘 颖 胡海涛

2012年8月31日

目 录

译者序	I
前 言	1
第 1 章 语料库语言学目的和方法	2
1.1 研究语言:结构与应用	2
1.2 什么是基于语料库的方法	3
1.2.1 基于语料库分析的特点	4
1.2.2 语言使用的联结模式	4
1.2.3 定量分析和功能解释的作用	6
1.2.4 基于语料库的方法与其他方法的比较	7
1.2.5 语料库语言学研究的领域	7
1.3 语料库和语料库分析工具	8
1.3.1 语料库	8
1.3.2 分析工具	9
1.4 本书概要	11
1.4.1 章节概要	11
1.4.2 本书包含的内容和未包含的内容	12
第 2 章 词典编纂	13
2.1 调查词典编纂的问题	13
2.2 词义的调查	16
2.3 调查词频	17
2.4 不同语域 deal 的分布	20

2.5 不同语域词义分布	22
2.5.1 DEAL 做名词	22
2.5.2 作为动词的 deal	26
2.6 同义词的分析	27
2.6.1 big, large, great 的频率分布	27
2.6.2 紧跟 big, large 和 great 的右搭配	28
2.6.3 large 的远距离搭配	32
2.7 结论	34
第3章 语法	35
3.1 语法的研究	35
3.2 形态特征的调查	37
3.2.1 名词化的分布和功能	37
3.2.2 名词化结尾的分布和功能	40
3.3 语法类的分布	42
3.3.1 语法类的计数	42
3.3.2 名词动词比率在不同语域中的比较	43
3.4 句法结构分布和功能	44
3.4.1 that 和 to 补语从句的识别和赋码	45
3.4.2 that 和 to 补语从句在不同语域的分布和功能	46
3.5 影响选择同义结构变异的因素	48
3.5.1 影响主语和后置 that 从句选择的因素	49
3.5.2 与 ESL 课本的比较	51
3.6 结论	52
第4章 词汇语法	53
4.1 词汇语法问题调查	53
4.2 意义相近词汇的语法联结: little 和 small	54
4.2.1 自动分析与人工分析技术结合	55
4.2.2 对 little 和 small 的语法联结模式的解释	59
4.3 同义词 begin 和 start 的语法联结	60
4.4 几乎等价的语法结构 that 从句和 to 从句的词汇联结	63
4.5 结论	66

第 5 章 语篇结构特征研究	68
5.1 语篇特征研究	68
5.2 口语语域和书面语语域的指称类型	69
5.2.1 指称语的特征	70
5.2.2 互动式分析技术:指称语的赋码	71
5.2.3 口语和书面语语域中指称语的使用模式	74
5.3 修辞结构的语言关联:动词时态和语态的语篇地图	78
5.4 结论	83
第 6 章 语域变异与特殊用途英语	84
6.1 语域变异研究	84
6.2 口语和书面语语域中的从句	86
6.3 口语和书面语语域的差异	89
6.3.1 语言特征中的共现模式	89
6.3.2 英语的多维度分析法	90
6.3.3 口语和书面语语域的比较	96
6.4 ESP 中的语域特征需求	97
6.4.1 调查生物学和历史学学术论文之间的差异	98
6.4.2 学术论文各部分间的关系	103
6.5 结论	105
第 7 章 语言的习得与发展	107
7.1 引言	107
7.2 学生语言的语料库	109
7.3 从单个语言特征研究写作的发展	110
7.4 小学生语言发展的多维调查	112
7.4.1 小学生演讲和写作的多维分析	112
7.4.2 从多维角度考虑写作发展	117
7.4.3 小学生和成人维度变化的比较	120
7.5 母语和非母语的错误模式	123
7.6 结论	125

第 8 章 语言的历时和风格研究	127
8.1 风格和历时研究	127
8.2 语法和词汇特征的历时变化	128
8.3 书面语和口语的历时演变	132
8.4 方言变化:男性和女性语言的历时变化	135
8.5 作者风格的分析	140
8.6 结论	143
第 9 章 结论	145
9.1 基于语料库方法的贡献	145
9.2 其他基于语料库的研究	146
9.3 基于语料库的方法和语言教育	147
9.4 参与基于语料库的研究	149
附录 1 方法箱	151
方法箱 1 语料库设计要解决的问题	152
方法箱 2 历时语料库设计	155
方法箱 3 检索程序包与语料库分析程序	157
方法箱 4 标注语料库的特征	159
方法箱 5 标注过程	162
方法箱 6 频率统计的标准化	164
方法箱 7 词汇联结的统计测量	165
方法箱 8 语料库研究的分析单位	168
方法箱 9 显著性检验和统计报告	172
方法箱 10 因素负荷与维度分数	174
附录 2 商业语料库和分析工具	176
语料库	176
分析工具	179
在线资源	180
参考文献	182

前言

开始写此书时,我们可以选择不同的目的和重点。我们可以介绍语料库的研究历史,调查到目前为止已开展的语料库研究。可以介绍以语料库为基础的研究,帮助读者展开自己的研究。还可以介绍语料库的分析方法和技巧。本书综合考虑这些方面,揭示语料库方法如何把难以处理的语言学问题进行解决的过程,为语言学提供新的研究办法。

最近几年,在线语料库和分析工具越来越容易获得,基于语料库的研究越来越普遍。然而,对于语料库研究,仍有许多方面对读者来说很神秘,因为论文对研究方法和分析过程描述得不够详细。本书介绍了如何利用具有代表性的语料库,给出利用语料库方法能进行哪些有趣的语言应用调查,提供了基于语料库研究的详细过程和分析方法细节。

本书中的许多研究都受国家科学基金和 Addison-Wesley Longman 的慷慨资助和支持。由国家科学基金资助的 ARCHER 项目,分析英语历时语域变异。项目编号:BNS-9010893,由 Doug Biber 和 Ed Finegan 负责,研究助教:Dwight Atkinson, Jena Burges, Dennis Burges, Randi Reppen 和 Ann Beck。朗曼 ELT 部门提供了许多语料库和本书实例分析所使用的计算机。在此表示感谢。

除此之外,我们还要感谢北亚利桑那大学和爱荷华州立大学的同事们,在与他们进行篇章和语言应用合作研究时,他们给予了我们友谊和鼓励。尤其感谢:Jena Burges, Carol Chapelle, Bill Grabe, John Hagge, Marie Helt 和 Susan Wright。我们要特别感谢最近两年参与我们的硕士 TESOI 项目和博士生课程的学生们,他们对我们的草稿提供了许多建议。最后,我们要感谢三位校阅者,他们对本书的草稿提供了详细的、有益的反馈意见。

本书调查了人们在口语和写作中使用语言的方式。基于语料库的方法是对存储在计算机中大规模真实语言实例进行的分析。每章集中讨论一个不同的语言学领域,包括:词典编纂,语法,篇章,语域变异,语言习得和历时语言学。每章讨论的过程相同,首先说明研究的语言问题,其次阐述用语料库方法分析的步骤,最后对研究结果给出语言学解释。用实例具体描述了语料库研究方法和优点。

10 个方法箱简洁清晰地解释语料库研究的重要方法。附录中给出基于语料库调查的资源。本书清晰全面的介绍适合本科生和专业研究者。

第 1 章 语料库语言学目的和方法

1.1 研究语言:结构与应用

语言研究可以分为两个主要领域:语言结构研究和语言用法研究。传统上,语言分析强调结构——确定语言的结构单位与类别(如:词素、单词、词组、语法类)并描述较小的单位如何结合成较大的语法单位(如:单词如何组成词组,词组如何组成句子,等等)。

强调语言用法研究是本书的中心,这是一种不同的研究角度。从这个角度,我们可以看出演说家和作家怎样利用他们的语言资源。我们研究在自然真实的文本中实际应用的语言,而不是研究在理论上它可能是什么。

许多语言应用研究集中在特殊的语言结构,调查相似的语言结构在不同的语境中的使用和功能。例如,英语 that 动补从句和 to 动补从句在结构上有相似的特征,意义也相近,像这些句子:

(1) I hope that I can go.

(2) I hope to go.

另外,that 从句中可以省略 that。

(3) I hope I can go.

结构分析会描述这三个句子中的语法异同点。这三个句子以相同的语法结构来完整地表达动词的意思。不过,如果要问为什么语言中有许多意义和语法功能都相似的结构,这种语言应用分析就超出了传统的语法描述。

回答这个问题需要考虑一系列因素。比如:口语语域和书面语语域是否有不同的偏好?这种形式是否经常用于不同的动词后面?这种形式是否用于不同的特定意义中?有些问题能在研究中找到答案。事实上,本书第 3 章和第 4 章的实例分析中,你会看到优先使用哪些结构与联结模式有关。

另外,可以针对一篇文章或一组说话人/写作者的语言来分析一个语言

结构的语言应用模式。比如:单个作者的写作风格或不同社会群体使用的语言是否有共同的动机。可以思考这样的问题:与同时代使用的语言相比,一个特定的作者是怎样使用语言的?女性用语与男性用语如何不同?

比较不同文本或文本组的语言同样重要。每天当我们处于不同的场合都会使用不同的语言变体——从与家庭成员谈话,阅读报纸,写信给朋友,到阅读学术性文章。不同语境中使用的语言变体,被称为“语域”。描述这些语域的特征是研究的一个重要领域,也是很复杂的一个领域,这涉及许多不同的语法和词汇特征的选择。我们如何发现在会话、新闻、学术论文、私人信件中的模式呢?这些问题也是应用研究的重要方面,本书中第二部分将讨论这部分内容。

在所有的应用研究中,分析家都试图揭示语言的典型使用模式而不是对其是否符合语法作出判断。在这些分析中有两个主要的研究目标:①估计某一使用模式的范围。②分析产生变异的语境条件。比如:在分析 that 从句和 to 从句时,我们可能会考虑说话者是否喜欢用一种从句而写作者喜欢用另一种从句。进而我们可能会考虑到一系列的语境条件,如每个从句类型使用的典型动词。

发现使用模式和分析语境条件对我们提出了挑战。因为我们在寻找典型模式,不能依靠直觉语感和零星的例子。在很多情况下,人们往往注意到异常的实例而不是典型的实例,因此基于直觉所得出的结论是不可靠的。进一步,我们需要分析由许多说话者收集的大量语言,以保证我们的结论不是基于少数说话者的个性语言而作出的。然而,处理大量语料费时费力,而且难以弄清大量的语境因素。如果你想比较会话和学术文章中的语言,假如每个语域各包含十篇文章,考虑二十种不同的语言结构,先不考虑这些结构与语境的联系,在这些语料库中查找这些结构就很困难。

因此,长期以来,大规模的语言用法的调查难以开展。语料库语言学不仅能处理大量语料,而且能够弄清许多语境因素,从而为语言研究开辟了新道路。

1.2 什么是基于语料库的方法

到底什么是基于语料库的方法?它跟其他语言分析方法有什么不同?下面几部分将说明这些问题。1.2.1节给出基于语料库分析的主要特点。基于这些特点的研究是语言应用的新方法。语料库语言学通过考察相关的联结模式研究语言使用的特征。这个概念形成了本书后面分析的基础,1.2.2节将介绍这个概念。联结模式即表现为量的关系,也表现为质的关系。定量分析表示语言特征及不同形式与语境之间的联结程度,定性分析则对此作出功能解释。在任何基于语料库的研究中,定量分析和定性分析都是非常重要的一步,因此我们在1.2.3节讨论定量分析与定性分析的关系。在1.2.4

节和 1.2.5 节,我们比较了基于语料库的研究与其他语言分析方法,并给出基于语料库方法适合的研究领域。

1.2.1 基于语料库分析的特点

基于语料库分析的主要特点是:

- 具有实验性,分析自然语言文本中语言使用的实际模式。
- 它收集大量的真实文本,被称作语料库。以语料库作为分析基础。
- 使用计算机的自动与交互技术进行分析。
- 使用定量与定性分析的技术。

大规模的语料库和计算机的应用使得分析具有可靠性,否则是不可能的。基于语料库的几个优点与计算机应用有关。计算机使得识别和分析复杂的语言使用模式成为可能。因为计算机可以存储和分析大规模的自然语言语料库。而人工分析很难完成。而且,计算机可以对语料库提供一致和可靠的分析,不会在分析时改变想法和变得疲惫。计算机也可以与人互动使用,让分析家做繁难的语言学判断,而计算机进行记录和运算。

最后一点也很重要,基于语料库的分析绝不仅仅是语言特征的简单计数。可以对定量模式进行定性的功能解释。本书每章,你都会发现在定量分析中发现的模式基础上会有大量篇幅解释、列举和说明这些模式。基于语料库研究的目的是不是简单地进行定量分析,而更重视定性的功能解释。

1.2.2 语言使用的联结模式

许多早期的语料库语言研究都只是简单地统计语言条目的出现频率。比如,一些词汇研究比较特定单词或 2 字符、3 字符、4 字符单词的频率。一些语法研究统计名词、动词、和形容词的出现频率。这种研究可以提供参考资料,如识别 50 个最常用的单词。也可以提供简单的风格研究,如一篇文章中名词和动词的相关频率。

但是,如果一个代表性的语料库能被正确地利用,那么它能提供多种语言应用方面的信息。基于语料库的分析可以识别和分析复杂的联结模式:语言使用的联结模式系统地反映了一些语言特征与其他的语言学特征之间以及与非语言学特征之间的关系。

联结模式分为两类,一类是集中在词项或语法结构的使用上;另一类集中在文本的语言学特征上。如表 1.1 所示。传统的语言学分析只研究某一语言特征、单词或者语法结构。但是通过考虑与其他特征的联系,我们可以对这些特征的应用做进一步的研究。有两种主要的联结很重要:语言上的联结和非语言上的联结。

表 1.1 语言应用中的联结模式

-
- A. 调查语言特征(词汇或语法)的使用
- (i) 特征的语言学联结
- 词汇联结(跟特殊单词联系)
 - 语法联结(跟特殊语法结构联系)
- (ii) 特征的非语言学联结
- 语域分布
 - 方言分布
 - 时代分布
- B. 调查语域或文本(如, 语域、方言、历时)
- (i) 语言联结模式
- 单个语言学特征或特征类
 - 语言学特征的共现模式
-

语言联结有两种主要类别:

1. 词汇联结——研究语言学特征是怎样系统地与特殊单词联系的;
2. 语法联结——研究语言学特征是怎样系统地与相关语境中的语法特征联系的。

在第2章通过分析单词 big, large 和 great 对词汇联结进行了说明。这种分析特别考虑到这三个单词的搭配——即那些倾向于与各个目标单词共现的单词。比如, big 一般与 toe 共现, 而 large 一般与 number 共现。虽然这三个单词分开来看几乎是同义的, 然而这三个词与不同的单词搭配使用。因此, 通过对“词汇——词汇”的联结模式分析, 发现它们的词汇联结是不同的。

第4章研究词汇——语法的联结。比如, 我们比较几乎同义的形容词 small 和 little, 看它们作为定语和作为表语时有什么不同(如, the small boy 和 the boy is small)。我们比较了经常与 that 从句连用和经常与 to 不定式连用的动词(如与 that 从句连用的 think 和与 to 不定式连用的 want)。

除了语言学联结, 在非语言学联结中也可以研究语言学特征的应用。这里有三个主要因素是相关的: ①词汇或语法结构在不同语域中分布如何不同。②词汇或语法结构在不同语域变异(方言)中分布如何不同。③词汇或语法结构在不同时期分布如何不同。第3章调查了名词化在学术语域和对话语料中分布如何不同? 这是语法特征(名词化)与非语言学特征(语域)之间联系的实例。

语言学和非语言学联结模式不是独立的, 它们互相影响。因此, 本书中最典型的分析包括两种联结模式。比如, 当我们考虑 big, large 和 great 的词汇——词汇联结时, 我们也会考虑它们在不同语域的应用。

语料库语言学不只可以来研究特殊的语言学特征, 也可以从联结模式来描述文本或变体的特征(表 1.1 的 B 部分)。在这种情形中, 基于语料库的研究试图寻找语域、方言、风格和个人文学作品中的语言学联结模式。这些语言学联结可以是单个特征也可以是特

征类。第6章,我们通过三种从句的使用来刻画不同的口语语域和书面语语域的特征。

为了更全面刻画语域变异,另一种语言联结模式很重要:语言学特征组经常在语料库中共现。比如:名词,介词,长词和定语形容词倾向于在某种语域中共同出现。为什么会这样?这些特征有什么功能?当这些特征很少的时候哪些其他的特征会在文本中出现?这些问题都能在本书中后半部分找到答案。

虽然基于语料库方法能研究很多不同种类的联结模式,这些模式都有一个共同的重要特征:它们表示连续变化的关系。也就是说,这些模式在语言应用中都不会总是发生,或决不会发生,而是在不同程度上会发生。我们经常或偶尔想起某些模式——但是“经常”或“偶尔”又代表什么?联结模式间的比较更精确地刻画出不同模式存在的程度,即定量的度量。下一部分讨论基于语料库研究中的定量分析,还有定性分析和功能解释所起的补充作用。

1.2.3 定量分析和功能解释的作用

上一部分我们了解了不同种类的联结模式,它们都可以用基于语料库方法来进行研究,并注意到这些联结都是连续变化的结构。因此,定量分析对基于语料库的研究是至关重要的。比如,如果你想比较 big 和 large 的语言应用模式,就需要知道在语料库中它们各出现多少次,有多少不同单词经常与其共现(词的搭配),这些搭配有多频繁。这都是量的方法。

在本书的所有实例分析中,你都可以找到定量分析。很多例子,尤其是在书中的前几章,我们都只是列出了频率数据——某种模式与其他的相关模式共同出现多少次。很多情况下,我们可以通过这些频率直接观察固定模式,为了让这些例子简单易懂,能为广大的读者所接受,我们没有采用统计学程序。

在本书后几章中的例子中,统计学程序对调查复杂的联结模式很重要。有些分析还包括统计学显著性检验。显著性检验用来显示定量结果偶然出现的可能性。因此在基于语料库的研究论文中总有显著性检验。但是,教你怎样进行统计学检验并不是我们的目的。例子中用到的统计学概念已给出介绍。方法箱部分给出了方法细节。本书的讨论将让你明白使用统计学程序的目的和重要性,但我们并不打算全面地讨论统计学技术。你可以在统计学书籍中找到完整介绍。

另外,当你阅读本书中的实例分析时,你还会发现比量化和统计学结果多得多的内容。基于语料库方法的一个重要部分是除了定量统计外还要给出功能解释,功能解释说明这些模式为什么会存在。因此,基于语料库研究中功能解释和定性分析也是非常关键的部分。书中由于篇幅限制,不可能给每个分析提供完整的功能解释。但是,我们将突出这些功能解释的主要部分,强调在基于语料库分析中这一步的重要性。

1.2.4 基于语料库的方法与其他方法的比较

到目前为止,我们已经强调了基于语料库方法与众不同的特点。特别强调了它在研究语言应用中的重要作用。我们已经注意到全面研究不能凭感觉和零星的例子;需要大量语料库来进行经验性分析。

基于语料库的分析应当看作是传统方法的补充,而不是唯一正确的方法。简单地确定出基于语料库方法和其他方法的相互关系是很有用的,但列出所有的语言学研究方法并不是我们的目的。

基于语料库分析以多种方式与其他方法相联系。首先,许多问题是由早先的结构分析产生的。比如,对相关结构的好奇——如 that 从句与 to 从句的比较——首先来源于知道存在这些相似结构。其次,研究问题也可能是由假设或理论框架产生的。比如,当我们与别人交流时,由于缺乏时间进行计划和组织,使得口语不可能像书面语一样结构复杂。在第6章我们调查了一个与这种假设相关的问题,调查口语和书面语语域中不同种类的复杂结构的频率和分布——调查结果令人吃惊。类似地,理论上可以列出不同年龄段孩子所获得的语言学特征,这可以用基于语料库的方法来进行检验。

直觉语感和零星例证也可以引出有趣的基于语料库的调查。例如:大学生经常有这种印象,不同的学术研究成果是用不同的方法写的。第6章用语料库方法研究了生物和历史学术论文的不同。类似地,有些初等学院的老师感觉他们的非英语母语学生始终在作文里犯很多语法错误,甚至学了好几年英语之后仍然如此;第7章比较了非英语母语与英语母语初等学生的错误。因此,语言经验和以往的语言学研究都可以用基于语料库方法来研究。

最后,对特殊文本中特殊的语言特征进行详细分析,也能补充从大规模语料分析得出的发现。例如,为信息顺序(如,新信息之前是已知信息)提供证明的早期研究是基于对简单文本透彻的定性分析基础上得出的。类似地,对话中小片段交互的分析,也能对语言应用提供不同的研究角度,基于语料库方法中没有包含这些角度。通过本书,我们认为语言应用的全面分析需要基于语料库的方法,但是这种研究经常建立在对单个文本分析产生的构想和假说基础上来进行。

1.2.5 语料库语言学研究的领域

基于语料库的方法几乎可以应用于语言学研究的所有领域。前面已经提到许多研究——单个词语,语法特征,男性和女性用语,儿童语言习得,作者风格和语域类型。事实上,基于语料库的方法几乎可以应用到语言学的所有领域。

语言学的核心领域,如词典编纂(词语研究)和语法,可以利用基于语料库的方法进行