

当代国外语言学与应用语言学文库

# Corpus Linguistics

## 语料库语言学

Douglas Biber  
Susan Conrad  
Randi Reppen

外语教学与研究出版社

Foreign Language Teaching and Research Press

大学出版社

Cambridge University Press



当代国外语言学与应用语言学文库

# Corpus Linguistics

## 语料库语言学

*Douglas Biber,*  
*Susan Conrad and Randi Reppen* 著

潘永樑 导读

外语教学与研究出版社  
剑桥大学出版社

**(京)新登字 155 号**

**京权图字: 01 - 2000 - 0164**

**图书在版编目(CIP)数据**

语料库语言学/(美)拜伯(Biber, D.)等著;潘永樑导读. - 北京:外语教学与研究出版社, 2000. 8

ISBN 7 - 5600 - 1961 - 7

I . 语… II . ①拜… ②潘… III . 描写语言学 - 基本知识 IV . H0

中国版本图书馆 CIP 数据核字(2000)第 67487 号

English edition © Douglas Biber, Susan Conrad and Randi Reppen 1998

All rights reserved. No part of this publication may be reproduced, stored or transmitted by any means without the prior permission of the publishers.

This edition of Corpus Linguistics by Douglas Biber, Susan Conrad and Randi Reppen is published by arrangement with the Syndicate of the Press of the University of Cambridge, Cambridge, England. It is for sale in the People's Republic of China only. Not for export elsewhere.

本书由剑桥大学出版社授权外语教学与研究出版社出版

**版权所有 翻印必究**

**语料库语言学**

Douglas Biber 等 著

潘永樑 导读

\* \* \*

责任编辑: 刘相东

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.cn>

印 刷: 北京外国语学院印刷厂

开 本: 650 × 980 1/16

印 张: 21.5

版 次: 2000 年 8 月第 1 版 2000 年 8 月第 1 次印刷

印 数: 1—5000 册

书 号: ISBN 7 - 5600 - 1961 - 7/G · 858

定 价: 27.90 元

\* \* \*

如有印刷、装订质量问题出版社负责调换

# 当代国外语言学与 应用语言学文库



## 专家委员会

主 任 王宗炎

副主任 (以姓氏笔画为序)

|     |     |     |     |     |
|-----|-----|-----|-----|-----|
| 刘润清 | 吴一安 | 李朋义 | 沈家煊 | 陆俭明 |
| 陈国华 | 胡文仲 | 胡壮麟 | 徐烈炯 | 桂诗春 |
| 顾曰国 | 戴炜栋 |     |     |     |

委 员 (以姓氏笔画为序)

|     |     |     |     |     |
|-----|-----|-----|-----|-----|
| 文秋芳 | 方 立 | 王才仁 | 王立弟 | 王克非 |
| 王初明 | 王逢鑫 | 王嘉龄 | 史宝辉 | 宁春岩 |
| 田贵森 | 申 丹 | 刘世生 | 朱永生 | 何兆熊 |
| 何自然 | 张绍杰 | 张柏然 | 张德禄 | 李宇明 |
| 李延福 | 李行德 | 李筱菊 | 杨永林 | 杨信彰 |
| 杨惠中 | 杜学增 | 汪榕培 | 邵永真 | 陈治安 |
| 周流溪 | 林连书 | 罗选民 | 姚小平 | 祝畹瑾 |
| 徐盛桓 | 秦秀白 | 贾玉新 | 顾 阳 | 高 远 |
| 高一虹 | 黄国文 | 惠 宇 | 董燕萍 | 蒋祖康 |
| 韩宝成 | 蓝 纯 | 潘永樑 |     |     |

策 划 霍庆文

## Preface by Halliday

Foreign Language Teaching & Research Press is to be congratulated on its initiative in making these publications in linguistics available to foreign language teachers and postgraduate students of linguistics in China.

The books are a representative selection of up-to-date writings on the most important branches of linguistic studies, by scholars who are recognized as leading authorities in their fields.

The availability of such a broad range of materials in linguistics will greatly help individual teachers and students to build up their own knowledge and understanding of the subject. At the same time, it will also contribute to the development of linguistics as a discipline in Chinese universities and colleges, helping to overcome the divisions into “English linguistics”, “Chinese linguistics” and so on which hinder the progress of linguistics as a unified science.

The series is to be highly commended for what it offers to all those wanting to gain insight into the nature of language, whether from a theoretical point of view or in application to their professional activities as language teachers. It is being launched at a time when there are increasing opportunities in China for pursuing linguistic studies, and I am confident that it will succeed in meeting these new requirements.

M. A. K. Halliday  
Emeritus Professor  
University of Sydney

# 王宗炎序

近年来，国际交往日益频繁，国际贸易急速发展，出现了一种前所未有的现象：学外语、教外语、用外语的人多了；研究语言学和应用语言学的人多了；开设这方面专业的高校也多了，语言学硕士生和博士生也多了。就是不以此为专业，学习语言学和应用语言学的也不乏其人。为了给从事这个专业的师生提供便利，同时又帮助一般外语教师、涉外工作者以及汉语研究者开阔思路，扩大视野，提高效率，我们献上这套内容崭新而丰富的丛书——英文版《当代国外语言学与应用语言学文库》。

文库首批推出 54 部外国英文原著，它覆盖了语言学与应用语言学 28 个分支学科。这批书是我们与各地有关专家教授反复研究之后精选出来的。出版这样大规模的语言学与应用语言学丛书，这在我国语言学界和外语教学界是破天荒第一次。

我们这样做，抱着什么希望呢？总的说来，是遵循教育部关于加强一级学科教育的指示，在世纪之交，推出一套书来给中国的外语教育领航，同时也给一般外语工作者和汉语研究者提供信息，拓宽思路。

我们希望这个文库能成为进一步带动外语教学改革和科研的发动机；我们希望它能成为运载当代外国语言学理论、语言研究方法和语言教学方法来到中国的特快列车；我们希望，有了这套书，语言学与应用语言学专业师生就能顺利地进行工作；我们希望，通过读这套书，青年外语教师 and 外语、汉语研究者能迅速把能力提高，把队伍不断扩大。

以上是我们的愿望，可是从广大读者看来，这个文库是否真的有出台的有必要呢？我们想，只要大家看一下今天的客

观情况，就知道这套书有填空补缺的作用，是让大家更上一层楼的扶梯。

我们跟许多人一样，认为国内的外语教学和语言学与应用语言学研究是成绩斐然的，但是某些不足之处也无庸讳言。

在语言研究方面，有大量工作还等着大家去做。汉语语法研究，过去由于结构主义的启示，已经成绩卓著，可是现在虽则引进了功能主义，还看不出什么出色的成果。语料语言学是新兴学科，在我国刚刚起步，机器翻译从 50 年代就有人搞，然而其进展至今不能令人满意。

在语言理论方面，我们不时听到一些片面的、所见不全的论调。有人说，1957 年前西方根本没有什么理论语言学，其创始者是 Chomsky；也有人说，语言纯属社会文化范畴；还有人说，搞语言研究只有量化方法才是科学方法，定性方法不值得一提。

谈到外语教学，某些看法做法是分明不值得赞许的。有人以为交际教学只管听说，不管读写，也有人以为教精读课就是教阅读，不管口语。在分析课文时老师满堂灌，学生开口不得，是常见的；教听力课时老师只管放录音，对学生不给半点提示点拨，也并非罕有现象。

上述这些缺点，我们早有所知，现在我们更加明白，必须力图改进，再也不能安于现状了。为了改进，我们就得参考国外的先进理论，借鉴国外的有效措施。眼前这个文库，就是我们上下求索的结果。

在编辑这个文库时，我们在两方面下了功夫。

一方面，在选书时，我们求全，求新，求有代表性和前瞻性。我们不偏爱一家之言，也不只收一家外国出版社之书。语言学与应用语言学的主干学科固然受到了应有的重视，分支学科可也不忽视。语料语言学、语言统计学是新兴学科，我们收入了专著；句法学、语义学久已有人研究，我们也找到了有关的最新著作。

另一方面，我们邀请了国内知名的博士生导师、硕士生导师为各书撰文导读，为读者铺平道路。语言学和应用语言学专著包罗宏富，初学者读起来可能觉得茫无头绪。为了助他们一臂之力，本文库中每一种书我们都请专家写了一万字左右的导读材料。哪怕书中内容比较陌生，谁只要在读书前看一下导读材料，读书后把材料再看一遍，一定能弄清脉络，掌握要点。

在结束本文时，我们想向爱好泛读的人们提个建议。语言和社会生活息息相关；我们靠语言与他人协作；通过语言继承传统文化，接受外国先进思想和科学知识；利用语言来教育下一代，帮助他们创造美好的未来；语言又反过来表达着我们的个性和我们充当的各种角色。学一点语言学和应用语言学，有助于增强我们的语言意识，对我们的工作和生活都是有利的。我们不妨把此事作为一个项目，列入自己的日程。持之以恒，必有所获。

王宗炎

中山大学教授  
博士生导师

# 导 读

Douglas Biber 等三位学者所著的《语料库语言学》是一部很有实用价值的介绍语料库语言学研究方法的专著。1998 年,该书由英国剑桥大学出版社出版,是该社有名的语言学丛书中的一种。

本书介绍在语料库基础上进行语言研究的基本方法,介绍如何使用计算机强大的检索和统计手段处理语料库所提供的大量语料,进行词汇、语法、语篇、用法变异、语言习得、及语言的历时发展等各种课题研究。一本介绍语料库研究方法的书可以有不同的目的和重点,它可以介绍语料库研究的历史与现状,可以介绍以语料为基础的研究课题,它也可以介绍语料库研究的具体方法与分析技术。本书综合论及所有这些方面,但始终把重点放在说明语料库的实证性研究可以为语言学中难以解决的课题提出新的解决办法,说明语料库语言学的方法和分析技术可以有效地应用于语言使用的研究。本书的主要章节以课题研究的实例说明语料库研究的原则与方法。其实用性也体现在章节的结构上,每一章都先说明研究的主题,再阐述用语料库的方法进行研究的步骤,最后对有关的语言现象做出解释。书末总结了十个研究方法框架,简明清楚地解释了语料库研究中的重要方法与学习者应当注意的事项。书后附有大型语料库与语料检索工具的简介。本书简明实用,对大学学生和专业研究者都有很大的参考价值。

本书介绍的原理和方法反映了过去 20 多年来语料库语言学的长足发展。为了更好理解本书的内容,在分章介绍和述评本书内容之前,首先概括一下语料库语言学发展中的几个重要问题。

语料库语言学是一门独具特色的语言研究学科。语料库语言学以语料库为手段来研究语言。语料库是作为信息载体的大量语言资料的集合。语料库中的语料可以是为了特定目的而收集的语言资料(如采访记录,或特定范围内的书面材料),也可以是为了普遍的语言研究目的而收集的语言资料(如大量自然会话的转写资料和各种报刊杂志书籍的文字资料)。传统的结构主义语言学很早就重视专门语料的收集,并使用它们来描写音位和单句。然而,语料库语言学的方法与传统的方法不太一样,而且研究语料主要的目的也不是描写音位和单句

用法，而是为了描写和解释语言中的词汇和句法的各种问题，以及处理自然语言的各种课题。目前，计算机语料库可以贮存和处理亿万字符的语料。语言研究者可以根据研究课题的要求，使用赋码手段（tagging，或译为“加标、标注”，即给语料的词汇和其他形式加上识别和分类标记）和检索工具（concordancing programs），分析语料库中的语料，对语言现象进行定量与定性的分析。

随着计算机技术的进步，语言研究中以实证和统计资料为基础的研究方法得到了新的肯定。这明显体现在语言研究的一些最新成果上。例如，在词汇研究方面，1995 年同时出版了四部主要为英语学习者编纂的英语大词典：《朗文当代英语词典》（第 3 版）、《牛津高级学生词典》（第 5 版）、《科林斯科贝尔特英语词典》（第 2 版）和《剑桥国际英语词典》。为了编写这些颇有影响的英语大词典，使用的英语语料库至少有一亿多英语单词。在语言结构研究方面，20 世纪 90 年代末，根据语料库的研究的最新成果编写的英语语法专著相继问世，其中最重要的当推 Douglass Biber 等五位学者所著的《朗文英语口语与笔语语法》（Douglas Biber, et al. Longman Grammar of Spoken and Written English. Pearson Education Limited. 1999）。这些成果说明，在语言研究中以自然语言文本所组成的语料库已从边缘地带走到了中心位置，成为语言研究资料的基本来源。语言研究中重新肯定以实证和统计资料为基础的研究方法，是语言学的一个重要进展。回顾 60 年代到 70 年代，还只有少数学者在语料库领域拓荒，当时他们的困难不只是语料处理技术上的局限，更重要的是他们的理论和方法与语言学界占主导地位的乔姆斯基的语言观不一致，因此他们的实证原则与方法常常被唯理论的主流语言学视为不合时宜的徒劳。而今，随着计算机信息处理技术的迅猛发展与语料库语言学成果的不断出现，面貌已完全改观。在语料库语言学的发展中，虽然在语料库的最佳规模、结构组成以及语料的代表性等问题上还有很多争论，但语料库语言学在语言研究中的重要地位已得到了广泛的承认，已经不是一个需要争论的问题了。

**语料库的分类** 已经建立和正在建立的语料库大致可以分为下列几大类型。（1）以语料所代表的媒体形式分类：可分为书面文字材料的电脑文本语料库、经过转写的口语语料库、视频语料库、及上述几种形式的混合语料库。其中，口语语料库必须先把声音资料转写为书面文本资料，这中间涉及理论性很强、极其繁复的转写操作，难度比

建立书面语语料库大得多。(2)以语料库设计结构分类:可分为均衡结构语料库、无结构的随机开放式语料库、及由若干子语料库叠加而成语料库网。(3)以语料的来源分类:有单语种语料库与多语种语料库之分,原文语料库与翻译语料库之分,母语语料库与外语学习者语料库之分。(4)以语料的时效分类:有共时语料库与历时语料库之分。(5)以语料的处理方式分类:有未经赋码的文本语料库与经过赋码的文本语料库的区别。在最后这一类中,前者是自然语料的文本,即未经任何赋码或标注处理的原始语料文本;后者是根据研究课题的需要对文本经过赋码处理的语料库。本书语言分析所使用的语料库大致不超出这些范围。

**语料库建设中遇到的几个主要问题** 现代计算机具有巨大的存贮和处理信息的能力,因此贮存和检索大量文本材料和转写的口语语料现在已不成问题。但语料库在理论和实际的操作中还有不少问题尚待进一步解决,最主要的是以下几个问题:

一是语料库的设计目的。建立语料库可以只是为个人的研究课题服务,也可以是为了给广大的语言研究者提供语料资源。建库的直接目的可以是为了编词典、排词频、可以是为了研究某些语言现象、研究某个作家的语言或某个时期的语言、也可以是为了研究语言习得、研究外语教学、或是设计母语教学的课程,等等。语料库的目的决定收集语料的范围与规模,决定是建立抽样语料库,还是建立兼收并蓄的全面语料库。总之先要按建库的目的,进行语料库建设的可行性论证。

二是收集语料时的文本形式、文本所有权和版权的问题。

三是需要不断更新应用软件以满足语料库的发展需要。为了使新的软件与原来的软件兼容,就必须尽量做到赋码符号的标准化。

四是语料库的规模问题。为了研究词汇和编纂词典,比较各种文本的语言结构,一般来说,语料库越大越好。著名的布朗语料库与伦敦—奥斯陆/卑尔根语料库(LOB)都是100万字,已不敷各种语言研究所需。1987年出版的《科林斯科贝尔特英语词典》(第一版)就是以2000万字的伯明翰英语文本语料库为基础的。1995年出版的四部英语大词典就是几家主要的词典出版商与语言学学者进行大规模的语料库研究的最著名成果。语料库大,对词典编纂与语言结构研究无疑是有益的。可是,当语料库越建越大时,语料中大约有一半的字型只会出现一次,因此,扩大语料库的规模实际上只是增加了词频为1

的词的数量。这样就产生两个问题：第一是一个语料库不可能收集到所有词频为 1 的词，第二是收集到的这些词在目标语中的出现频率很低，因此它们在语料库中的频度可能被大大高估，不能准确反映它们在语言中的实际使用频率。这个问题对字典编纂的影响也许还不大，但是，赋码器（tagger）等自然语言处理器的设计与使用很大程度上依赖于词频统计的精度，因此对词频为 1 的词汇的过高评估会降低这些处理系统的效度和精度。

五是语料库的代表性问题。为了使在语料分析基础上得出的结论可靠，语料库就必须具有代表性。然而，绝对的代表性是不可能的，设计者只能在取样过程中尽量兼顾语料的多样性和比例。代表性对于同类语料所组成的语料库并不造成多大问题（如法律英语语料库或新闻英语语料库）。可是，要建立一个普遍用途的大型语料库时，语料的代表性问题就十分突出。设计者必须正确估计各种类型的语料在目标语言中的比例，从而分配它们在语料库中应占的份额。然而，对于什么类型的语料可以入选、以及占有多少比例，学者的认识并不总是一致。况且，还有很多语料难于归类。因此，往往只能暂且采取传统的语料分类方法进行取样，这样就可能出现语料库的代表性问题。然而，目前还没有更可靠、更有效的方法可以代替这种以语料类别来给语料进行有比例抽样的方法。此外，在检索分析中还会遇到一些与语料库的代表性有关的其他问题。例如，有时语料库并不含有研究者所需要的足够的研究资料，在这种情况下，虽然研究者想要调查的是与某一语言现象有关的统计数字，而实际得到的只是计算机所能检索到的有关现象的统计数字。

六是语料的时限问题。这个问题涉及语言的历时发展。布朗语料库和伦敦-奥斯陆/卑尔根语料库早已有 30 多年历史，即使它们能够代表 60 年代的英语，也不等于就能代表 90 年代的英语。因此，更新语料库的语料就成了大问题。设计者面临设计新的、可以不断更新、不断扩大的开放式语料库的任务。

七是由于语料库须要不断更新，检索统计手段须要不断改进，因而人们对语料库的标准化提出了更高要求。

**语料库对语言学理论和研究方法的影响** 语料库的建设和应用已对语言学研究产生重大影响，有可能引起语言学理论和方法的重要变革。使用计算机技术建立语料库和处理大批语料，使研究者能够进行以前不可能进行的大规模的语料分析和统计。如今，以语料库为基础

的语言研究已经揭示了许多用原有语言理论不能充分解释的语言现象，这客观上就要求研究者用新的理论来描写与解释它们。例如，近年来根据口语语料库的研究建立起来的语法是一种与现有语法不同的新语法。语料库语言学的新发现不仅对现有的语言学理论和方法产生冲击，而且在发展语料库研究方法的同时，并不排除内省的演绎的方法。例如，语料研究的结果常被用来验证关于语言理论的一些假设，这些假设可以是大量语料分析中归纳出来的，也可以是语言学家内省的结果，或者二者兼有。可以说，语料库语言学正在逐步形成兼有美国结构主义语言学语料归纳法的优点、又有乔姆斯基学派内省法长处的综合研究的方法。在语言学研究中，语料库（尤其是口语语料库）将以其代表性、真实性、机器可处理性、兼容性、资源的多元利用性、语料库的开放性、以及分析的客观性等明显优点得到语言研究者的高度重视。

## 第一章 语料库语言学的目的和方法

本章是全书的总论。首先概括语言用法研究的特点，指出语料库语言学特别适合于语言用法的研究。然后介绍语料库语言学的主要研究方法和应用领域。接着介绍本书使用的语料库与语料分析工具。最后概述本书的结构与分章内容。本书每一章都围绕样本分析展开，都有研究课题的介绍及其意义的说明，然后是研究方法、结果与讨论。读者由此可以了解到以语料库为基础的各种研究的步骤以及分析过程中的重点，并按照自己的研究需要设计相似的研究题目。只要使用恰当的检索工具和语料库，就可以开展研究。本章介绍的语料库语言学的目的与方法可以概括如下。

**语言用法研究的特点** 语言研究可以分为两大领域：语言结构研究与语言用法研究。语言用法研究的重点不是研究语言在理论上的可能性，而是研究语言在文本中的实际用法，研究讲话者和作者如何使用语言。语言用法研究主要做下列工作：（1）调查相似的语言结构在不同语境中的使用与功能，以及解释多个语言结构具有相似意义与语法功能的原因。（2）研究某一文本的语言或某一群说话人或作者的语言，如作者的个人语言风格、不同社会集团中语言的使用，不同作者之间的语言风格比较，以及不同性别之间语言使用情况比较。（3）比

较不同文本的语言，研究不同语境中使用的语言变体，称为“语域研究”。这涉及语法和词汇形式的选择，如会话、报纸、论文、私人信件的不同语言模式，以及这些不同语域的语言特征。(4) 揭示语言的典型使用模式，如估计某一使用模式的范围和分析产生变异的语境条件，而不是对其是否符合语法作出判断。研究语言用法和分析语境因素不能依靠直觉语感和零星的例子，而必需分析许多语言使用者提供的大量语料。然而，处理大量语料费时费工，而且难以弄清大量的语境因素。因此，长期以来，大规模的语言用法的研究难以开展。语料库语言学不仅能处理大量的语料，而且能够同时弄清许多语境因素，从而为语言研究开辟了新的道路。

**语料库语言学的主要研究方法** 语料库语言学与传统的语言结构研究的最大不同，在于通过考察相关的“联结模式”(association patterns)，来研究语言使用的特征。联结模式既表现为量的关系，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，定性分析则对此做出功能解释。以语料库为基础的研究方法的主要特点是：(1) 具有实验性，分析自然语言文本中语言使用的实际模式；(2) 以语料库为分析研究的基础；(3) 使用计算机的自动与互动技术进行分析；(4) 使用定量与定性分析的技术。

显然，研究规模的扩大与信度的提高有赖于计算机的使用。计算机可以识别、分析和贮存大量复杂的语言使用模式，是过去人工分析做不到的。同时它也可以与人工互动地结合，即用机器纪录资料数据，让分析者做出有关的判断。如今，语料库的方法已经走出了简单统计语言特征的范围，可以有效地对定量模式做出定性的功能解释。因此，本书每章都不仅有语言使用的定量分析，而且重视定性的功能解释。

语言使用的联结模式系统反映了一些语言特征与另一些语言特征之间、以至非语言特征之间的关系。联结模式大致分为两类，一类是集中在词项或语法结构的使用上，另一类是集中在文本或不同类型文本特征上。传统的分析往往只研究某一语言特征，而语料库的方法系统探讨这些特征与其他特征之间的联结关系。联结模式的表述不是以传统的绝对形式，而是用一定的连续关系表达，即表述为某种模式出现的可能性程度。对语料进行定量分析和定性的功能分析，就是正确反映这种可能性的有效方法。语料库研究方法最关键的部分，是在定量分析的基础上对用法模式进行功能解释。

以语料库为基础的研究方法与语言研究的其他方法有着互补关系。本书介绍的研究个案有力地说明，语料库研究中不少课题是在早先的语言结构的研究中产生的，现在可以用语料库的方法进行验证；直觉语感与零星例证也可以引出语料库的大规模调查；对具体文本中语言个别特征的详细调查，也能补充大规模语料研究中得出的发现。

**语料库语言研究的领域** 以语料库为基础的研究方法几乎可以应用于语言学研究的所有领域，如语言结构的核心领域（如词汇学与语法学等），语言学的其他领域（如社会语言学、方言和语域研究、语言教学、文体学、历时文本的比较、语言模式的变异研究等）。语言学的所有领域几乎都可以从语言使用的角度进行研究，因此都是语料库语言学的用武之地。

**本书使用的语料库和语料分析工具** 本书除了不同课题中使用了一些特别的语料库以外，主要使用了4个著名的公开语料库，它们是伦敦-隆德语料库（London-Lund），兰开斯特-奥斯陆/卑尔根语料库（LOB），英国国家语料库（BNC）和朗文-兰开斯特语料库。分析工具主要是两类：一类是现成的工具包，如最通用的文本分析计算机工具（TACT）与词汇分析（Lexa）工具包，可以分析词汇搭配（即联结模式）与词频；另一类是语料库课题的研究者根据需要自己设计的工具。

**思考题：**

- (1) 语言用法研究有哪些主要特点？
- (2) 语料库为基础的语言研究有哪些主要特点？
- (3) 本书与语料库语言学其他入门书相比有什么独到之处？

## 第二章 词典编纂

词典编纂是本书第一部分“研究语言特征的使用情况”的最重要的四个课题之一。由于计算机技术的发展，现在的词典编纂有过去无法想象的三大有利条件：大型语料库、更高的代表性和完善的分析工具。语料库可以回答词典编纂中六个大问题：

- 根据词汇在大量自然语境中的使用情况，决定其意义。
- 决定词频，从而编制常用词表与非常用词表。

- 决定某个词汇具有什么非语言的联结（如语域、历史阶段与方言等），从而了解不同类型语言中用语的特征。
- 决定词项的搭配及其在不同语域中的分布。
- 决定某词的义项及其用法的分布。
- 决定同义词的使用与分布，从而了解语境对词义的选择、及搭配与语域的关系。

这六个问题就是本章的论题。语料取自朗文－兰开斯特语料库中两个语域的语料各一百万字。

本章首先以英语词 *deal* 为例研究义项、词频、搭配及其分布。词义是词汇学的中心问题。可以用检索工具列出目标词在样本中的所有使用位置及其上下文，从而决定其义项。这里，检索关键词（KWIC）与词频是决定词义的重要步骤。然后，考察了 *deal* 在不同语域中的分布。统计说明，该词用作名词或动词时在不同语域中用法有差异。最后，调查了 *deal* 的义项在不同语域中的分布。这里，采取了检索目标词搭配（collocates）的办法。调查结果不仅说明词义是由上下文语境决定的，而且纠正了许多词典中关于该词常见义项的误解。

接着，本章以英语同义词 *big*, *large* 和 *great* 为例说明如何用语料库对它们的使用模式进行比较。语料检索的结果说明，这三个同义词的词频在两种不同的语域中差异很大；它们的右向搭配也说明，表面上同义的词项在实际使用中意义是不同的，这三个形容词有自己的最常用的搭配、不同的优先义项、不同的义项分布。此外，对与 *large* 并不直接相连接的词汇搭配（collocates at a distance）的调查也说明这些同义词之间具有很多的用法差别。

本章论及一些重要的技术细节，如原始统计数据与标准统计数据、词频分布的交叉验证、语料库规模的控制、以及检索工具（如 TACT 与 Lexa）的使用等。这些都是语料库语言学的基本常识，读者应当注意。

### 思考题：

- (1) 计算机技术的发展给词典编纂提供了什么有利条件？
- (2) 词汇研究中哪些常见问题可以通过语料库的方法解决？
- (3) 您手头有什么可以使用的词汇索引检索工具与可供使用的语料库？请参照本章介绍的方法，做一次相似的词汇研究。

### 第三章 语法

本章的重点是语法的实证研究，讨论如何在不同的层次上用语料库技术研究语法，包括形态特征、语类与句法结构。语料统计说明语法特征在不同语域中有着系统的分布特点。语法研究没有词汇研究那样的长期的实证研究的传统。但是，由于大型语料库的建立和计算机工具的使用，研究者现在可以调查各种语言结构的频率分布；调查语法结构与语言的其他分析层次之间的关系，以及语言因素与非语言因素之间的关系；解释讲话者选择一定语言形式的原因。用实证的方法对语法资源的各种使用模式做出定量定性的分析，这在过去是做不到的。本章着重考察其中4个问题：

- 怎样通过分析不同语域中形态特征的分布来解释它们的用法和功能？
- 怎样通过分析语类在不同语域中的分布来解释它们的用法与功能？
- 怎样通过分析不同语域中句法结构的分布和联结模式来解释它们的功能？
- 什么样的语言特征和非语言特征决定了语言使用者总要从多个同义的语法结构形式中选择使用其中某一种形式？

现在，只要使用词汇检索软件检索未经赋码的语料，即可调查出不同的形态特征在不同语域中的分布，所以形态研究可以用作语料库方法研究语法的入门课。本章选择了学术文章、小说与口语语篇三种语域，调查词缀在动词和形容词两种词类的名词化过程中的作用，以及名词化词尾的分布和功能。在词类分布的研究中，主要是考察了名词、动词和形容词在不同语域中的频率和分布。统计结果表现出有规律的趋势。但这些趋势必须与许多共现的特征结合起来才有意义。在句法结构的研究中，主要比较了具有相近意义的两种结构：that-引导的名词性从句与to引导的不定式。比较了它们在会话与学术文章两种语域中的分布与功能，调查了它们与一些常用动词的联结模式，以及这些句子中形式主语的使用情况。结果发现这些语法联结模式与不同语域中一些其他用语的模式有着系统的关联。

是什么决定了讲话者选择某种表达方式，而不用另一种虽然语义相似、但结构却不同的表达方式？本章还是以这两种结构为例，使用