高等院校英语语言文学专业研究生系列教材

总主编　戴炜栋

语料库语言学导论

An Introduction to Corpus Linguistics

杨惠中　主编

外教社

SHANGHAI FOREIGN LANGUAGE EDUCATION PRESS

上海外语教育出版社

高等院校英语语言文学专业研究生系列教材

总主编　戴炜栋

语料库语言学导论

An Introduction to Corpus Linguistics

主编：杨惠中

编著：卫乃兴　李文中

濮建忠　雷秀云

外教社

上海外语教育出版社

SHANGHAI FOREIGN LANGUAGE EDUCATION PRESS

图书在版编目(CIP)数据

语料库语言学导论 = An Introduction to Corpus Linguistics / 杨惠中主编. - 上海:上海外语教育出版社,2002
高校英语语言文学专业研究生系列教材
ISBN 7-81080-373-5

Ⅰ. 语… Ⅱ. 杨… Ⅲ. 词语-研究-研究生-教材 Ⅳ. H03

中国版本图书馆 CIP 数据核字(2001)第 097439 号

出版发行：上海外语教育出版社
（上海外国语大学内） 邮编：200083
电　　话：021-65425300（总机），65422031（发行部）
电子邮箱：bookinfo@sflep.com.cn
网　　址：http://www.sflep.com.cn　http://www.sflep.com
责任编辑：王彤福

印　　刷：上海锦佳装璜印刷发展公司
经　　销：新华书店上海发行所
开　　本：890×1240　1/32　印张 13.125　字数 348 千字
版　　次：2002 年 7 月第 1 版　2002 年 7 月第 1 次印刷
印　　数：3 500 册

书　　号：ISBN 7-81080-373-5 / H · 150
定　　价：18.60 元

高等院校英语语言文学专业研究生系列教材

编委会主任：戴炜栋

编　委　（以姓氏笔划为序）

总　序

近年来，随着我国经济的飞速发展，社会对以研究生为主体的高层次人才的需求日益增长，我国英语语言文学专业的研究生教学规模也在不断扩大。各高校在研究生培养方面，形成了各自的特色，涌现出一批学科带头人，开设出自己的强项课程。但同时我们也认识到，要使研究生教育持续健康地发展，要培养学生创新思维能力和独立研究与应用能力，必须全面系统地加强基础理论与基本方法方面的训练。而要实现这一目标，就必须有一套符合我国国情的、系统正规的英语语言文学专业研究生主干教材。

基于这一认识，我们邀请了全国英语语言文学专业各研究领域中的知名专家学者，编写了这套《英语语言文学专业研究生系列教材》，旨在集各高校之所长，优势互补，形成合力，在教材建设方面，将我国英语语言文学专业的研究生培养工作推上一个新的台阶。我们希望通过这套教材的出版，来规范我国的英语语言文学专业的研究生课程，培养出更多基础扎实、知识面广、富有开拓精神、符合社会需要的高质量研究生。

在内容上，本套系列教材覆盖了英语语言文学专业各学科的主要课程。我们总的编写指导思想是：**结合我国英语语言文学专业研究生教学的实际情况与需要，强调科学性、系统性、先进性和实用性；力求做到理论与应用相结合，介绍与研究相结合，中与外相结合，史与论相结合；广泛搜集资料，全面融会贯通，使每一本教材都能够反映出该研究领域的新理论、新方法和新成果。**本套教材的这些特点，使其有别于单纯引进的国外同类原版教材，是国外教材所不可取代的。同时，两者的作用是相辅相成的。正是由于这些特点，本套教材不仅可以作为我国英语语言文学专业研究生的主干教材，也可作为中国语言文学专业的教师与学生的参考用书。

在编写体例上，我们参照了国家标准局的有关标准以及国际上的通行做法，制定了统一的规范。每章后面，都列出了思考题和深

入阅读书目，以便启发学生思考和进一步深入研究。

教材建设是学科建设的一项重要基本建设，对学科发展有着深远的影响。我们相信，正如国外剑桥大学和牛津大学出版社出版的语言学和应用语言学教材和丛书对推动国际语言学和应用语言学的发展起了巨大作用一样，在世纪之交推出的这套系列教材，也必将大大推动我国21世纪英语语言文学专业研究生教育事业的发展，促进我国英语语言文学研究水平的提高。

戴炜栋

2000年9月

前　言

语料库语言学的发展迄今已三十年余。从我国第一个语料库JDEST语料库在上海交通大学建成至今，也有近二十年了。语料库语言学以真实语言使用中的语言事实为基本证据，凭借现代计算机技术，采用数据驱动的实证主义研究方法，对语言、语言交际和语言学习的行为规律进行多层面和全方位的研究，从而给语言学工作者带来了一种新的理念，揭示了一种新的研究方法，开辟了一个新的研究领域。语料库语言学的数据不同于以往研究中采用的"直觉性数据"和"诱导数据"，对实际使用中的语言事实进行定性定量的描写和概括，使研究更具科学性和准确性。由数百万、数千万乃至数亿词的连续文本构成的大型语料库给语言学工作者提供了强大、丰富的数据，加上一套科学的研究手段和方法，研究人员已能够对真实语言交际的各个方面，包括词汇的、句法的、语义的、语用的、语篇的，进行深入的探讨和全面的描写；运用学习者语料库数据，研究者可对学习者的典型行为进行研究，调查不同类型和不同背景的学生的不同语言特征，发现他们语言能力发展的具体过程与阶段，探讨学生的学习策略；通过监控语料库，人们还可对语言在历史进程中的动态变化进行实时监控和观察；在语言教学领域，语料库在制定教学内容、教学目标等方面为决策者提供坚实可靠的决策依据，建立在科学依据基础上的大纲设计、教材开发和语言测试将会使教学更加有效。语料库研究还为自然语言处理、机器翻译和诸多语言工程项目提供了重要的理论启示与参照数据。近二十年来语料库语言学的发展以及所取得的丰硕成果已经引起人们愈来愈广泛的重视。语料库语言学的发展迎来了新的高潮。

通过几年的研究和探索，我的几位学生写成了这本介绍语料库语言学的小书，真诚地将其奉献给国内的语言学工作者、语言学研究生和语言教师，以期对大家了解、认识和利用语料库进行研究和教学能有所帮助。在本书的第一部分，我们对语料库语言学的一些

基本概念、语料库的种类、语料库语言学的发展概貌、语言数据的性质、一般的研究方法等问题进行了概括的描述；在第二部分里，我们讨论了语料库建设的一般原则、具体方法和程序、常用的统计手段和索引工具；第三部分是我们用语料库方法开展的部分研究。

无须讳言，本书描述和探讨的决不是语料库语言学内容的全部。由于我们的研究经验和学识水平所限，书中难免挂一漏万和不尽准确之处。因此，我们恳请广大同仁不吝指正。另一方面，语料库语言学正在蓬勃发展，尚存在许多不确定的因素，某些概念和方法会发生变化，并逐步完善。我们愿意随时就该领域里的发展与大家一道切磋。

本书各章的编著分工如下：杨惠中：第一章；卫乃兴：第三、六、八章；李文中：第二、五章；濮建忠：第四章；雷秀云：第七章；第九章由辛克莱教授撰写，由卫乃兴译成中文。全部书稿由我审阅，如有错误不妥之处，理应由我负责。

上海外语教育出版社的同志为本书的出版花费了大量精力和时间，给了我们宝贵支持。对此，我们表示诚挚的感谢。

杨惠中

2001 年 7 月

目　　录

第一部分

理论研究

第一章　语料库语言学概述

1　一种全新的研究思路

人类科学的发展从综合向高度专门化进行分化，然后又在更高的层次上进行新的综合。早期的学术研究并没有严格的学科划分，18世纪的工业革命要求各学科有深入的发展，于是分化成数学、物理学、化学、天文学、生物学等等。每一学科都有自己的研究对象和研究方法，有自己的学科界定。随着人们对客观世界的深入理解，学科进一步细分。力学从物理学中分出来并且进一步分化为普通力学、材料力学、结构力学、量子力学等等。高度专门化使学科得到深入的发展，但缺点是各学科之间缺乏联系，不符合客观世界万事万物是相互联系的这一事实。进入20世纪中叶以后人们越来越重视跨学科的研究，实际上许多重要的进展恰恰是在学科交叉的边缘上取得的。在自然科学领域中这种例子举不胜举，如物理化学、数学物理学、生物力学等。在语言学领域，现代语言学从20世纪初诞生起一直以研究语言体系为自己的学科方向。但是语言现象涉及人类活动的一切方面，要深入理解语言的本质、理解人类的言语机制、理解人类言语活动的心理本质、理解语言与社会的关系，语言研究就必然涉及心理学、社会学、神经生理学等等，于是出现了心理语言学、社会语言学、神经生理语言学、语言哲学、语用学等众多崭新的学科；关于语言体系本身的研究也早已突破了句子的界限，在纵深方向上出现了语义学，在高于句子的层次上出现了篇章语言学。这些五彩缤纷的众多交叉学科从不同的侧面揭示语言的本质，使人们对人类的言语机制有了更多的了解，使语言学能更好地为语言实

践服务，包括语言使用、语言教学、语言工程等等。这种从综合到分化，再从分化到新的综合的过程是螺旋式进行的，新的综合是在更高层次上的综合。计算机的出现为各行各业提供了强大的研究手段。语料库语言学就是出现在语言学、计算机科学、认知语言学和应用语言学边缘上的一门新的交叉学科。

语料库语言学为语言学研究提供了一种全新的研究思路，它以真实的语言数据为研究对象，从宏观的角度对大数量的语言事实进行分析，从中寻找语言使用的规律；在语言分析方面采用概率法，以实际使用中的语言现象的出现概率为依据建立或然语法进行语法分析。语料库语言学从一个新的角度揭示自然语言的复杂性。

重视对真实语言事实的研究一直是语言学研究中的优秀传统，20 世纪 60 年代初以夸克(R. Quirk)等为主进行的“英语用法调查”(“Survey of English Usage”) 就通过系统的调查建立了现代英语语料库，不过这一项艰巨的工程在当时是用手工完成的。夸克等以这一语料库为基础完成的《现代英语语法》(*A Grammar of Contemporary English*) 和《英语语法大全》(*A Comprehensive Grammar of the English Language*)，对现代英语进行了全面的、系统的描写，在英语语言学界产生了广泛影响。最早的计算机语料库出现在 20 世纪 60 年代初，是由纳尔逊 (F. Nelson) 和库切拉 (H. Kuçéra) 建立的 BROWN 美国英语语料库。这是一个小型语料库，总容量 100 万英语词，收集了 60 年代有代表性的美国英语语料，语料取材自各种出版物，选材时考虑到各种不同文体的平衡，选材又严格按照随机原则，因此迄今为止 BROWN 语料库仍被视为标准语料库，是现代语料库语言学的发端，对语料库语言学的发展产生了重要影响。

在 60 年代初美国语言学界占主导地位的是乔姆斯基(N. Chomsky)的转换生成语法，他认为语言学的研究对象应当是人脑的语言机制，即为什么操本族语的人有能力生成和理解无限数量的合乎语法的句子并且有能力识别不合语法的句子。语言学应当研究人脑的这种语言机制而不是语言的具体运用，他把前者叫做语言

能力（competence），把后者叫做语言运用（performance），并认为语言学的中心任务是前者而不是后者，语言学的任务就是要解释这种能力的本质。由于语言能力存在于理想的本族语者的头脑中，无法直接观察。能够直接观察的是语料，是语言运用结果，而语言运用受到各种其他因素的干扰，不能用来揭示语言的本质，因此反对研究具体语料。乔姆斯基认为，“任何自然语料都是偏颇的，对其描述不过是列举一张清单而已”（1962:59）。从20世纪60年代起转换生成语言学在语言学界占主流地位，语料库语言学因此一度进入低谷。70年代和80年代，从事语料库语言学研究的学者主要集中在欧洲，包括英国、瑞典、挪威、芬兰等国家。国际语料库语言学会议也主要是人数不多的语料库语言学家的聚会。但是由于语料库语言学代表了一种崭新的研究思路，并且其研究成果在辞典编纂、语言教学、自然语言处理等方面得到了实际应用，因此很快显示出强大的生命力。语料库语言学研究于是出现了新的高潮。继BROWN美国英语语料库以后，由英国Lancaster大学、挪威Oslo大学和Bergen大学等校的学者合作建立了完全与之对应的LOB英国英语语料库，以后相继出现了COBUILD语料库、英国国家语料库等容量达到上亿英语词的大型英语语料库；其他语言的语料库以及各种专门用途语料库也相继问世，各种研究工具也不断开发。可以说20世纪末语料库语言学研究进入了一个新的高潮，许多国际计算语言学学术会议上，语料库语言学方面的论文要占到一半以上。这种马鞍型的发展说明了语料库语言学的强大生命力。

计算机化的语料库是语言研究的强大工具，可以服务于范围广泛的研究领域，在语言描写方面包括词汇、句法、语篇等各种层次的语言研究。语料库语言学提供了强有力的手段，用来对自然语言进行定量分析。这种从真实语料出发，对大数量的语料进行宏观研究的做法有可能产生新的研究方法、建立新的语言学模型，因此计算机语料库又是各种语言学模型的试验田，可用来检验语言学模型的正确性。

语料库语言学有着广阔的应用领域，包括词典编纂、语言教学、机器翻译、人机互动查询、自动文摘等等。20世纪末叶以来，语料

库语言学的研究和应用正方兴未艾。

2 语言学研究的三种方法

语言学研究必然涉及语言材料，根据采集和使用语言材料的途径不同，现代语言学研究的基本方法主要有三种，即内省法（introspection）、诱导法（elicitation）和基于语料库的方法（corpus-based approach）。

2.1 内省法

内省法主要是转换生成语言学家采用的研究方法，他们以语言学家本人为资料提供人（informants），依靠自己的语感（intuitions）作为判断语言现象歧义、正误、可接受性等的依据。这是转换生成语言学家获得语料的主要来源。他们认为传统语言学、历史比较语言学、结构主义语言学只是分类之学，不能解释人类语言的本质。转换生成语言学家认为具体的语言事实是无限的、收集不完的，仅仅收集语言事实并加以分类并不能说明人类语言为什么具有创造性。既然人脑具有理解和生成无限数量的句子的能力，能够区分什么是合乎语法的，什么是不合乎语法的，语言学的任务就是要描写支配这种语言能力的规则。语言学的研究对象就不应当是具体的语料，而只能是理想的本族语者头脑中的语感。这种心灵主义的（mentalist）研究方法通常采用通过内省法自造的句子，如

Flying planes can be dangerous.
Sincerity may scare the boy.
John is eager to please.
John is easy to please.

在研究方法上，转换生成语言学家认为人的语言能力可以归结为一套公理系统，按照这套公理系统操某种语言的人可以生成和识别合乎语法的句子而不会生成和识别不合语法的句子。他们通常采用形式化的方法和专门的符号系统来揭示人的语言能力，揭示句子的句法、语义和音位结构，例如下面这一组简单的改写规则，可以生成若干合乎语法的英语句子，如 the dog chased the girl, the girl chased the dog 等：

NP → Det Noun
VP → Verb NP
VP → Verb NP PP
Det → the
Noun → girl, dog
Verb → chased

语言研究的内省法可以称作着眼于语言能力的研究方法(competence-based approach)。

2.2　诱导法

诱导法(elicitation)是一种调查方法，通过实地调查来收集人们对实际使用的语言材料的看法和人们对语言材料的心理反应，通常采用有控制的方法诱导出被试者对句子或句子中的某个成分的判断，要求被试者确定句子中有没有错误、句子的可接受程度、对句子的理解程度、以及其他类似的有关数据。采用调查方法可以使结论带有某种程度的客观性和可靠性，对某个语言事实的可接受性能得到一定的量的判断。但是语言虽然是一种社会现象，语言的使用则是一种心理现象。对某种语言形式的可接受性，人们往往没有统一的认识。个别人的语言直觉可能是不可靠的，这必然影响到调查结果的可靠性。

表 1.1 是对 76 名受过高等教育的美国人进行调查的结果，对

这些被试者来说英语是他们的母语。供调查的实验句共有五个，表中“+”号表示被试者认为实验句可以接受，“-”号表示不可接受，“?”号表示没有把握。

表 1.1　诱导法调查实例

实验句	-	?	-
1. He wants some cake.	76	0	0
2. Neither he nor they know the answer.	53	16	7
3. The old man chose his son a wife.	31	24	21
4. They aren't very loved.	4	20	52
5. A nice little car is had by me.	1	2	73

(From J. Svartvik)

由表 1.1 可以看出，对于某种语言形式的可接受性，人们往往没有统一的认识。

在研究方法上，诱导法采用问卷调查和实地调查的方法，费时费力，由于受到客观条件的限制，采用这种方法，规模不可能很大。另一方面，社会调查虽然可以提供一定的量的信息并且调查结果也有一定的客观性，但是被试者的主观判断容易受到试验者提示的干扰，例如对某个句子被试者认为不可接受，但是如果试验者多问几遍“真的不可这样说吗?”“这样说一定不行吗?”被试者就可能改变想法，或变得模棱两可，没有把握。这就会影响调查结果的有效性。

诱导法既依靠客观调查，又依靠被试者的主观判断，因此可以说是一种部分着眼于语言能力部分着眼于语言运用的研究方法(partly competence-based and partly performance-based approach)。

2.3　语料库方法

语料库语言学的研究方法以语料库为基础。所谓语料库是指