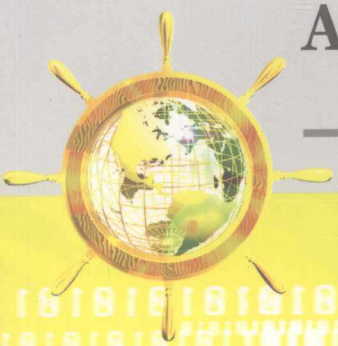
西方语言学前沿书系
语料库与计算语言学研究丛书 05



Applied Corpus Linguistics
A Multidimensional Perspective

应用语料库语言学的多维视角

[美] Ulla Connor
Thomas A. Upton 编

世界图书出版公司

西方语言学前沿书系
语料库与计算语言学研究丛书 05

Applied Corpus Linguistics
A Multidimensional Perspective

应用语料库语言学的多维视角

[美] Ulla Connor
Thomas A. Upton 编

王立非 导读

世界图书出版公司
北京·广州·上海·西安

图书在版编目(CIP)数据

应用语料库语言学的多维视角 = Applied Corpus Linguistics: A Multidimensional Perspective: 英文/［美］康奈尔（Connor, U.），厄普顿（Upton, T. A.）编；王立非导读. —北京：世界图书出版公司北京公司，2009.9
（西方语言学前沿书系·语料库与计算语言学研究丛书 05）
ISBN 978-7-5100-0547-3/H·1037

I. 应… II. ①康… ②厄… III. 语言教学—国际学术会议—文集—英文 IV. H09-53

中国版本图书馆 CIP 数据核字(2009)第 111974 号

应用语料库语言学的多维视角

Applied Corpus Linguistics: A Multidimensional Perspective

编　　者：［美］Ulla Connor, Thomas A. Upton
导　　读：王立非
责任编辑：王晓燕
封面设计：春天书装图文设计工作室

出版发行：世界图书出版公司北京公司　　http://www.wpcbj.com.cn
地　　址：北京市朝内大街 137 号（邮编 100010，电话 010-64077922）
销　　售：各地新华书店及外文书店
印　　刷：三河市国英印务有限公司

开　　本：711mm×1245mm　1/24
印　　张：14
字　　数：420 千
版　　次：2009 年 9 月第 1 版　2009 年 9 月第 1 次印刷
书　　号：ISBN 978-7-5100-0547-3
版权登记：京权图字　01-2007-3745
定　　价：35.00 元

序

“语料库与计算语言学研究丛书”旨在向国内读者推荐语料库语言学与计算语言学这两个学科最新和最经典的外文著作。语料库语言学的语料要建立在计算机上，而计算语言学则专门研究自然语言的计算机处理，这两个学科都要使用计算机，都与计算机有着不解之缘。这篇序言主要介绍了这两个学科的学术背景，以及它们之间的关系，以方便读者的阅读和理解。

1. 语料库语言学研究简介

语料库是为一个或多个应用目标而专门收集的、有一定结构的、有代表性的、可被计算机程序检索的、具有一定规模的语料的集合。

语料库应该按照一定的语言学原则，运用随机抽样方法，收集自然出现的连续的语言运用文本或话语片段来建立。从其本质上讲，语料库实际上是通过对自然语言运用的随机抽样，以一定大小的语言样本来代表某一研究中所确定的语言运用总体。

语料库一般可分为如下类型：

- 按语料选取的时间划分，可分为历时语料库（diachronic corpus）和共时语料库（synchronic corpus）。
- 按语料的加工深度划分，可分为标注语料库（annotated corpus）和非标注语料库（non-annotated corpus）。
- 按语料库的结构划分，可分为平衡结构语料库（balance structure corpus）和自然随机结构的语料库（random structure corpus）。
- 按语料库的用途划分，可分为通用语料库（general corpus）和专用语料库（specialized corpus）。专用语料库又可以进

一步根据使用的目的来划分，例如，又可以进一步分为语言学习者语料库（learner corpus）、语言教学语料库（pedagogical corpus）。

- 按语料库的表达形式划分，可分为口语语料库（spoken corpus）和文本语料库（text corpus）。
- 按语料库中语料的语种划分，可分为单语种语料库（monolingual corpus）和多语种语料库（multilingual corpus）。多语种语料库又可以再分为比较语料库（comparable corpus）和平行语料库（parallel corpus）。比较语料库的目的侧重于特定语言现象的对比，而平行语料库的目的侧重于获取对应的翻译实例。
- 按语料库的动态更新程度划分，可分为参考语料库（reference corpus）和监控语料库（monitor corpus）。参考语料库原则上不作动态更新，而监控语料库则需要不断地进行动态更新。

早在1897年，德国语言学家Kaeding就使用大规模的语言材料来统计德语单词在文本中的出现频率，编写了《德语频率词典》（J. Kaeding, Häufigkeitswörterbuch der deutschen Sprache, Steglitz: published by the author, 1897）。由于当时还没有计算机，Kaeding使用的语言材料不是机器可读的（machine readable），所以他的这些语言材料还不能算真正意义上的语料库，但是Kaeding使用大规模语言资料来编写频率词典的工作，是具有开创性的。

1959年，英国伦敦大学教授Randolph Quirk提出建立英语用法调查的语言资料库，叫做SEU（Survey of English Usage）。由于当时技术条件的限制，SEU是用卡片来建立的，也不是机器可读的。后来Quirk把这些语言资源逐步转移到计算机上，使之成为机器可读的语料库，并根据这个语料库领导编写了著名的《当代英语语法》[①]。

---

① R. Quirk, Towards a description of English usage, *Transactions of the Philological Society*, pp. 40-61, 1960.

1964 年，A. Juilland 和 E. Chang-Rodriguez 根据大规模的西班牙语资料编写了《西班牙语单词频率词典》[①]。在收集语言资料时，注意到了抽样框架、语言资料的平衡性、语言资料的代表性等问题。

1979 年，美国 Brown 大学的 Nelson Francis 和 Henry Kucera 在计算机上建立了机器可读的 BROWN 语料库（布朗语料库）。这是世界上第一个根据系统性原则采集样本的平衡结构语料库，规模为 100 万词次，并用手工作了词类标注（part of speech tagging）。BROWN 语料库是一个代表当代美国英语的语料库[②]。

接着，英国 Lancaster 大学的 Geoffrey Leech 教授提出倡议，挪威 Oslo 大学的 Stig Johansson 教授主持完成，最后在挪威 Bergen 大学的挪威人文科学计算中心联合建立了 LOB 语料库（LOB 是 Lancaster，Oslo 和 Bergen 的首字母缩写），规模与 Brown 语料库相当。这是一个代表当代英国英语的语料库。

欧美各国学者利用 BROWN 和 LOB 这两个语料库开展了许多大规模的研究，取得了引人注目的成绩。

从 20 世纪 90 年代初、中期开始，语料库逐渐由单语种向多语种发展，多语种语料库开始出现。目前多语种语料库的研究正朝着不断扩大库容量、深化加工和不断拓展新领域等方向继续发展。随着从事语言研究和机器翻译研究的学者对多语种语料库重要性的逐渐认识，国内外很多研究机构都致力于多语种语料库的建设，并利用多语种语料库对各种各样的语言现象进行了深入的探索。

近年来，语料库语言学的研究硕果累累，关于这些研究成果，我在《应用语言学中的语料库》（世界图书出版公司，2006）一书的导读中已经作过介绍，有兴趣的读者可以参看。

在建设或研究语料库的时候，我们应当注意语料库的代表

---

① A. Juilland and E. Chang-Rodriguez, *Frequency Dictionary of Spanish Words*, The Hague, Mouton, 1964.

② W. Francis, Problems of assembling, describing and computerizing large corpora, Scripter Verlag, pp. 110-123, 1979.

性、结构性和平衡性，还要注意语料库的规模，并制定语料的元数据规范。下面分别讨论这些问题。

首先讨论语料库的代表性。

语料库样本的有限性是无法回避的，所以在语料的选材上，要尽量追求语料的代表性，使有限的样本语料尽可能多地反映无限的真实语言现象的特征。语料库的代表性不仅要求样本取自于符合语言文字规范的真实的语言材料，而且要求样本要来源于正在“使用中”的语言材料，包括各种环境下的、规范的或非规范的语言应用。语料库的代表性还要求语料具有时代性，能反映语言的发展变化和当代的语言生活规律。只有通过具有代表性的语料库，才能让计算机了解真实的语言应用规律，才有可能让计算机不仅能够理解和处理规范的语言，而且还能够处理不规范的但被广泛接受的语言、甚至包含有若干错误的语言。

再来讨论语料库的结构性。

语料库是有目的地收集的语料的集合，不是任意语言材料的堆积，因此要求语料库具有一定的结构。在目前计算机已经普及的技术条件下，语料库必须是以电子文本形式存在的、计算机可读的语料集合。语料库的逻辑结构设计要确定语料库子库的组成情况，定义语料库中语料记录的码、元数据项、每个数据项的数据类型、数据宽度、取值范围、完整性约束等。

接着讨论语料库的平衡性。

平衡因子是影响语料库代表性的关键特征。在平衡语料库中，语料库为了达到平衡，首先要确定语料的平衡因子。影响语言应用的因素很多，如：学科、年代、文体、地域、登载语料的媒体、使用者的年龄、性别、文化背景、阅历、语料的用途（公函、私信、广告）等。一般根据实际需要，即平衡语料库的用途选取其中的一个或者几个重要的指标作为平衡因子，最常用的有学科、年代、文体、地域等。

在建设语料库时，还应当考虑语料库的规模。

大规模的语料库对于语言研究，特别是计算语言学的研究具有不可替代的作用。但随着语料库的增大，垃圾语料带来的统计垃圾问题也越来越严重。而且，当语料库达到一定的规模后，语

料库的功能并不会随着其规模同步地增长。我们应根据实际的需要来决定语料库的规模，语料库规模的大小应当以是否能够满足其需要来决定。

我们还应当考虑语料库的元数据（metadata）问题。

语料库的元数据对语料库研究具有重要的意义。我们可通过元数据了解语料的时间信息、地域信息、作者信息、文体信息等各种相关信息；也可通过元数据形成不同的子语料库，满足不同兴趣研究者的研究需要；还可通过元数据对不同的子语料库进行比较，研究和发现一些对语言应用和语言发展可能有影响的因素；元数据还可记录语料的知识版权信息、语料库的加工信息和管理信息。

关于语料库的标注（annotation）问题，学术界存在不同看法。由于在汉语书面文本中词与词之间没有空白，不便于计算机处理，因此，汉语书面文本的语料库一般都要做切词和词性标注。有的学者主张对语料进行标注，认为标注过的语料库具有开发和研究上的方便性、使用上的可重用性、功能上的多样性、分析上的清晰性等优点。有的学者则对语料库标注提出批评，批评主要来自两方面：一方面认为，语料库经过标注之后失去了客观性，所得到的标注语料库是不纯粹的，带有标注者对于语言的主观认识；另一方面认为，手工标注的语料库准确性高但一致性差，自动或半自动的标注一致性高但准确性差，语料库的标注难以做到两全其美，而目前大多数的语料库标注都需要人工参与，因而很难保证语料库标注的一致性（J. Sinclair, Corpus, Concordance, Collocation, Oxford University Press, 1991）。我们认为，不论标注过的语料库还是没有标注过的语料库都是有用的，其中都隐藏着丰富的语言学信息等待我们去挖掘，我们甚至可以使用机器学习的技术，从语料库中自动地获取语言知识。

近年来，在语料库的建立和开发中逐渐创造了一些独特的方法，提出了一些初步的原则，并且对这些方法和原则在理论上进行了探讨和总结，逐渐形成了“语料库语言学”（corpus linguistics）。由于语料库是建立在计算机上的，因此，语料库语言学是语言学和计算机科学交叉形成的一门边缘学科。目前语料库语言

学主要是利用语料库对语言的某个方面进行研究，是一种新的研究手段，同时也逐步建立了自己学科的理论体系，正处于迅速的发展过程中。

语料库语言学是一种新的获取语言知识的方法，它提倡建立语料库，在计算机的辅助下，使用统计的方法或机器学习的方法，自动或半自动地从浩如烟海的语料库或因特网中获取准确的语言知识，其中包括经过标注的结构化的语言数据和未经过标注的非结构化的语言数据。这是语言学获取语言知识方式的巨大变化，在语言学的发展史上具有革命性的意义。

语料库语言学也为语言研究人员提供了一种新的思维角度，辅助人们的语言“直觉”和“内省”判断，从而克服语言研究者本人的主观性和片面性。我们预计，语料库方法将会逐渐成为语言学研究的重要方法，受到语言研究者的普遍欢迎。

目前，语料库语言学主要研究机器可读自然语言文本的采集、存储、检索、统计、自动切分、词性标注、语义标注，并研究具有上述功能的语料库在词典编纂、语言教学、语言定量分析、词汇研究、词语搭配研究、语法研究、多语言跨文化研究、法律语言研究、作品风格分析等领域中的应用，已经初步展现出这门新兴学科的强大生命力，并且也影响和推动了计算语言学的发展。

2. 计算语言学研究简介

1946 年美国宾夕法尼亚大学的 J. P. Eckert 和 J. W. Mauchly 设计并制造出世界上第一台电子计算机 ENIAC。电子计算机惊人的运算速度，启发人们开始思考传统翻译技术的革新问题。为了探索如何用计算机来改进翻译技术，1952 年在美国的 MIT 召开了第一次机器翻译会议，1954 年美国乔治敦大学在国际商用机器公司（IBM）的协同下，用 IBM-701 计算机，进行了世界上第一次机器翻译试验，把几个简单的俄语句子翻译成英语，拉开了人类历史上使用计算机来处理自然语言的序幕。接着，苏联、英国、日本也进行了机器翻译试验，机器翻译出现热潮。

为了推动机器翻译的研究，1954 年美国出版了第一本机器翻

译的期刊 Machine Translation（《机器翻译》）。1962 年美国成立了“机器翻译和计算语言学学会”（Association for machine Translation and Computational Linguistics），为使期刊名与学会名称保持一致，1965 年 Machine Translation 杂志改名为 Machine Translation and Computational Linguistics（《机器翻译和计算语言学》）。在杂志的封面上，首次出现了“Computational Linguistics”这个新学科的名字，但是“and Computational Linguistics”这三个单词是用特别小号的字母排印的，说明当时学者们对于“计算语言学”是否能够算为一门真正的独立的学科还没有确实的把握。根据这些史料，我们认为，早在 1962 年，就出现“计算语言学”这个学科了，尽管刚出现时还“犹抱琵琶半遮面”，但现在，它已登上了庄严的学术殿堂。

40 多年来，计算语言学发展迅速，逐渐建立了完整的理论和方法，成为一门独立的学科，取得了很大成绩，在当代语言学中引人注目。

计算机的速度和存储量的增加，使得计算语言学在语音合成（speech synthesis）、语音识别（speech recognition）、文字识别（character recognition）、拼写检查（spelling check）、语法检查（grammar check）这些应用领域，都进行了商品化的开发。除了早期就开始的机器翻译（machine translation）和信息检索（information retrieval）等应用研究进一步得到发展之外，计算语言学在信息抽取（information extraction）、问答系统（question answering system）、自动文摘（text summarization）、术语的自动抽取和标引（term extraction and automatic indexing）、文本数据挖掘（text data mining）、自然语言接口（natural language interaction）、计算机辅助语言教学（computer-assisted language learning）等新兴的应用研究中，都有了长足的进展。计算语言学的技术在多媒体系统（multimedia system）和多模态系统（multimodal system）中也得到了应用。

3. 语料库语言学与计算语言学之间的关系

在过去 40 多年间，从事计算语言学应用系统开发的绝大多数

学者，都把自己的研究局限于某个十分狭窄的专业领域之中，他们采用的主流技术是基于规则的句法—语义分析，尽管这些应用系统在某些受限的“子语言”（sub-language）中也曾获得一定程度的成功，但是，要想进一步扩大这些系统的覆盖面，用它们来处理大规模的真实文本，仍然有很大的困难。因为从计算语言学应用系统所需要装备的语言知识来看，其数量之浩大和颗粒度之精细，都是以往任何系统所远远不及的。而且，随着系统拥有的知识在数量上和程度上发生的巨大变化，系统在如何获取、表示和管理知识等基本问题上，不得不另辟蹊径。这样，在计算语言学中就提出了大规模真实文本的自动处理问题。

1990 年 8 月在芬兰赫尔辛基举行的第 13 届国际计算语言学会议（即 COLING'90）为会前讲座确定的主题是：“处理大规模真实文本的理论、方法和工具”，这说明，实现大规模真实文本的处理已经成为计算语言学在今后相当长时期内的战略目标。为了实现战略目标的转移，计算语言学需要在理论、方法和工具等方面实行重大的革新。1992 年 6 月在加拿大蒙特利尔举行的第四届机器翻译的理论与方法国际会议（TMI-92）上，宣布会议的主题是“机器翻译中的经验主义和理性主义的方法”，所谓“理性主义”，就是指基于规则（rule-based）的方法；所谓“经验主义”，就是指以大规模语料库的分析为基础的方法，也就是基于语料库（corpus-based）的方法。语料库的建设和语料库语言学的崛起，为计算语言学战略目标的转移提供了语言资源方面的保证。随着人们对大规模真实文本处理的日益关注，越来越多的学者认识到，基于语料库的方法至少是对基于规则的方法的一个重要补充。因为从“大规模”和“真实”这两个因素来考察，语料库才是最理想的语言知识资源。在每两年召开一次的“自然语言处理中的经验主义方法会议”（Empirical Methods in Natural Language Processing，简称 EMNLaP）上，基于语料库的机器学习方法成为了会议的主流议题。计算语言学和语料库语言学发生了鱼水难分的密切联系。

在 21 世纪，这种基于语料库的机器学习方法在计算语言学中进一步以惊人的步伐加快了它的发展速度。我认为，计算语言学的

加速发展在很大程度上受到下面三种彼此协同的因素的推动。

第一个因素是带标记语料库的建立。在语言数据联盟（Linguistic Data Consortium，简称 LDC）和其他相关机构的帮助下，计算语言学的研究者可以方便地获得口语和书面语的大规模语料库，而且其中还包括数量可观的标注过的语料库，如宾州树库（Penn Treebank）、布拉格依存树库（Prague Dependency Tree Bank）、宾州命题语料库（PropBank）、宾州话语树库（Penn Discourse Treebank）、修辞结构库（RST Bank）和 Time Bank。这些语料库是带有句法、语义、语用、修辞结构等不同层次标记的标准文本语言资源。这些标注语料库的存在使得计算语言学的研究可以使用“有监督的机器学习方法”（supervised machine learning）来处理那些在传统上非常复杂的自动句法分析和自动语义分析等问题。这些标注语料库也推动了计算语言学中有竞争性的评测机制的建立，不再采用传统的人工评测方法而采用机器自动评测方法，评测的范围涉及到自动句法分析、信息抽取、词义排歧、问答系统、自动文摘等领域。

第二个因素是统计机器学习技术的成熟。对机器学习日益增长的重视，导致了计算语言学的研究者与统计机器学习的研究者更加频繁地交流，彼此之间互相影响。支持向量机技术（support vector machine）、最大熵技术（maximum entropy）、多项逻辑回归（multinomial logistic regression）、图式贝叶斯模型（graphical Bayesian models）等统计机器学习技术在计算语言学中得到了普遍的应用，深受计算语言学研究者的欢迎。

第三个因素是高性能计算机系统的发展。高性能计算机系统的广泛应用，为机器学习系统的大规模训练和效能发挥提供了有利的条件，这在上一个世纪是难以想象的。

进入 21 世纪以来，除了有监督的机器学习方法之外，大规模的“无监督统计学习方法”（unsupervised statistical machine learning）在计算语言学中也得到了广泛的关注。机器翻译（machine translation）和主题模拟（topic modeling）等领域中统计方法的进步，说明了在计算语言学中也可以只训练完全没有标注过的语料库来构建机器学习系统，这样的系统也可以得到有成效的应用。

由于建造可靠的标注语料库要花费很高的成本，建造难度很大，在很多问题中，这成为使用有监督的机器学习方法的一个限制性因素。因此，今后在计算语言学研究中将会更多地使用无监督的机器学习技术。我们相信，计算语言学和语料库语言学的联系将会更加密切，进一步发展到水乳交融的程度。

世界图书出版公司北京公司为引进国外关于语料库语言学和计算语言学的专著和论文集，出版了这套“语料库与计算语言学丛书”。这套丛书可以帮助读者更好地了解这两门新兴学科的发展概貌，扩大读者的语言学视野。

这套“语料库语言学与计算语言学丛书”现已收入7本国外有关语料库语言学和计算语言学研究的论文集和专著，今后还会不断引进其他最新的相关著作，力求反映当前语料库语言学和计算语言学的研究成果和发展动向。

《语料库语言学的进展》（Advances in Corpus Linguistics）是第23届国际英语语料库语言学年会的论文选集，包括22篇论文，反映了语料库语言学的最新发展情况。论文中心内容是讨论理论、直觉和语料的关系以及语料库在语言学研究中的作用。大多数论文是关于英语某个特定方面的经验研究，从词汇和语法到话语和语用，涉及面很广。此外，还讨论了语言变异、语言发展、语言教学、英语与其他语言的跨语言比较、语言研究软件工具的研制等问题。论文作者中有许多著名的语言学家，如M. A. K. Halliday、John Sinclair、Geoffrey Leech和Michael Hoey等。本文集既注意理论，又注意方法，清楚地显示了在经验主义方法的影响下语料库语言学这个新兴学科正在稳步地发展中。

《通过语料分析进行教与学》（Teaching and Leaning by doing corpus analysis）是第四次教学与语言语料库国际会议文集（2000年7月19—24日在Graz举行）。该文集反映了在语言教学中应用语料库取得的进展，不论把语料库作为一种资源还是作为一种方法，它对于语言的教学或研究都有积极的作用。文集强调了“发现式学习”（discovery learning）的重要性，指出发现式学习在课堂教学和课外研讨中都有很好的效果。文集还强调了在使用中学习口语和书面语的重要性，提出要充分利用现代的语料库来学

习、翻译和描述语言。文集主张以学生为中心，以基于语料库的语言调查为手段来进行语言教学。作者们描述了他们使用语料库来教学的实践与担心，成功与失败，让读者来分享他们的教学经验。文集所收的文章既有回顾，也有前瞻。

《语言学中的数学方法》（Mathematical Methods in Linguistics）是一本关于计算语言学的专著。全书包括A，B，C，D，E五篇。A篇讲述集合论，B篇讲述逻辑和形式系统，C篇讲述抽象代数，D篇讲述作为形式语言的英语，E篇讲述形式语言、形式语法和自动机。如果读者从A篇开始，一篇一篇地仔细阅读，反复推敲，认真做练习，逐步深入下去，就可以了解语言学研究中使用的主要的数学方法。本书是专门为语言学工作者编写的，讲数学问题时都紧紧扣住语言，深入浅出，实例丰富，作者还精心设计了大量练习，书末附有练习答案选，正好满足了语言学工作者更新知识的迫切需要，是一本不可多得的优秀读物。

《超句法表示结构的形式与功能》（Form and function of parasyntactic representation structure）根据真实的语料数据，从功能的视角来研究韵律和句法之间的相互作用。作者介绍了Halliday关于声调是一个信息单位的解释，Halford关于从韵律方面和句法方面定义“谈话单位”（talk unit）的思想，Esser关于抽象表达结构的概念，在这些理论的基础上，作者建立了一个“修正的谈话单位模式”（modified talk unit model）。这种谈话单位模式是一种“超句法的模式”（parasyntactic unit），既要进行定量的分析，也要进行功能的分析，并在声调单位的边界处来研究韵律状态和句法状态的相互作用。这项研究的数据是从London Lund英语口语语料库中采集的，样本包含50000个单词。研究结果表明，使用韵律和句法之间的相互作用，可以更有效地对语言信息进行结构化的描述。本研究应用了语料库语言学的方法来分析谈话单位在风格和语用方面的潜在特征，对英语口语进行功能主义和经验主义的分析，具有开创性。

《应用语料库语言学：多维视角》（Applied Corpus Linguistics：A Multidimensional Perspective）是美国印第安纳大学跨文化交流中心第四届北美研讨会的文集（2002年11月在Indiananpolis举

行)，作者来自美国、比利时、中国、法国、德国、爱尔兰、荷兰、西班牙等8个国家，内容涉及基于语料库的课堂教学、口语话语分析、书面语话语分析、网络话语分析等。整个文集分为两部分：第一部分是语料库语言学在口语话语分析和书面语话语分析中的应用；第二部分是语料库语言学在直接教学法中的应用。

《拓展基于语料研究的范围》(Extending the Scope of Corpus-based Research) 是北亚利桑那大学现代英语和中古英语计算机文档国际会议的文集，该会议于2001年在Arizona举行。这次会议的主题是“对语料库语言学的新挑战”。这种新挑战包括：改进语料库语言学的方法论标准，划清基于语料库的研究与理论语言学之间的界限，进一步探讨语料库语言学在语言教学中的应用。文集中的文章清楚地显示了基于语料库的研究正在迎击这样的挑战。

《应用语言学中的语料库》(Corpora in Applied Linguistics) 以丰富而有趣的实例说明了语料库在应用语言学中的作用，本书广泛地使用了COBUILD“英语银行”(Bank of English) 语料库中丰富的语言材料，把应用语言学与语料库密切地结合起来，对于如何在应用语言学中发挥语料库的作用，提出了许多独到的见解。本书还讨论了语料库对应用语言学的重要性和它的局限性。

世界图书出版公司北京公司出版的这套“语料库与计算语言学丛书”内容丰富而新颖，是反映这两个学科当前发展情况的一面镜子。读者可以通过这面镜子，对当前的语料库语言学和计算语言学有一个鸟瞰式的认识。希望广大读者喜爱这套丛书，从阅读中开阔眼界，获得新知。是为序。

冯志伟

2008.12

《应用语料库语言学的多维视角》导读

王立非

一　语料库语言学的发展现状

1.1　什么是语料库语言学

语料库语言学（corpus linguistics）可以定义为运用语料开展语言学研究的学科。语料库是载有大量语言信息的海量数据集合，可以具有特定目的（如科技英语语料），也可以是一般的语言资料（如报刊杂志书籍的自然语言）。语料库语言学利用语料库对语言进行研究，不仅是一个新手段，它还根据语言事实，对现行语言学理论进行批判，提出新观点或新理论。

语料库语言学发展至今已有近 50 年的历史，已经得到越来越多的认可。从语言分析、语言教学、词典编撰到人工智能等领域都开始应用语料库。语言研究者越来越重视对语料库作不同层次的标注，如：语音、构词、句法、语义以及语用等层次的标注。这种进步主要因为两方面的原因：一是基于计算机处理的多媒体技术的飞速发展；二是社会语言学、语用学、会话分析、人类语言学、计算语言学、人机对话研究、语音识别与合成等研究取得令人瞩目的成就。不论在理论上，还是技术上，语料库语言学都已趋于成熟。

1.2　语料库的发展现状

1.2.1　自然语言语料库

最早的语料库要追溯到 1936 年 Edward L. Thorndike 等人提出的 2000 词表。早期计算机语料库则以 1963—1964 年建成的 Brown 语料库为标志。它的全称是《布朗大学当代美国英语标准语料库》（The Brown

University Corpus of Present Day American English），有代表性地选取了500篇每篇约2000词的文本，包含100万1961年前后的书面英语。

其后20多年间涌现出了LOB，COBUILD，BNC，ICE，ACL/DCI，NERC等语料库。其中，LOB（The Lancaster-Oslo/ Bergen Corpus of British English）语料库由Johansson和Leech共同领导，于1978年编制完成，含100万1961年前后的书面英国英语。

大型计算机语料库COBUILD由Sinclair主持完成，是“英语语料库”的前身，1990年，它被扩展为The Bank of English语料库。这个语料库始建于1980年，Collins出版社与伯明翰大学合作，把它建成了世界上第一个大型英语语料库。该语料库反映了当代英语的现状，主要服务对象是英语学习者、教师、语言学家。语料库里25%是口语，75%为书面语。1987年Cobuild Dictionary出版时，Cobuild语料库的主体部分有130万词，另外还有保留语料库。到1997年，这个语料库的规模已经达到3亿词，其中口语语料库为7500万词，包括各种口语语料，而且全部标注。

BNC（The British National Corpus）语料库由英国政府、科研机构、出版商共同投资建设，项目开始于1991年，1995年完成。牛津大学出版社、Longman出版集团、Chambers Harrap出版集团、大不列颠图书馆、牛津大学计算中心和Lancaster大学英语计算研究中心合作参与了英国国家语料库（BNC）的建设。BNC收词1亿，共有4124个语篇，9000万词是书面语料，1000万词为口语语料。建立该语料库的目的是为了编写词典、语法参考书和为自然语言处理服务。英国国家语料库利用SGML语言（Standard General Markup Language）建立了一种编码系统，还利用Lancaster大学开发的语法标注器CLAWS进行了自动语法标注。

ICE（The International Corpus of English）语料库是在伦敦大学英语用法调查研究所第3任所长Greenbaum的倡导下于1988年开始建立的，这是世界上英语对比研究最庞大的计划。ICE语料库建立了20个子语料库，收录了英国、美国、加拿大、澳大利亚、新西兰等以英语为第一语言以及印度、尼日尔、新加坡、加勒比地区等以英语为官方语言或者第二语言的国家的各种书面语和口语语料，以便研究英语在世界不同地区的变体。所收材料为同一时期、同一题材的样本，参与的20多个国家和地区各编制一个含1990—1993年间的100万词的核