宁波工程学院学术著作出版基金资助

语料库语言学的研究与应用

RESEARCHES AND APPLICATIONS OF CORPUS LINGUISTICS

余国良 著

四川大学出版社

前 言

语料库语言学与计算机语言学、自然语言处理、计算机科学等不同学科有着密切的联系，它们在各自的发展过程中也是相互影响、相互促进的。但不可否认的是，语料库语言学是一门完全独立的学科，因为它有自己的哲学基础和工作目标。它的方法论认为，个人的语言直觉是不完备的、个性化的，因而也是不可靠的，尽管语料库所提供的丰富的、系统的、综合的外部证据可以用于证实和修正现有的语言学理论，但更为重要的是，作为一种由语料库驱动的“自下而上”的研究方法，在开辟崭新的、具有变革意义的语言研究方面，它的作用是其他学科不能替代的。

语料库语言学以真实语言使用中的语言事实为基本证据，采用定量与定性相结合的方法，对语言、语言交际及语言学习的行为规律进行多层面和全方位的研究，为语言学理论做出了多方面的贡献。第一，语料库研究发现，自然语言的口语活动所使用的许多范畴同样可以用于书面语描述，反之则不然；因而语言研究应当更多地注重口语，而不是将书面语的模式强加于口语。第二，研究发现，语言的向心式结构与离心式结构是两种基本的意义组织形式。这些组织既是意义的，又是结构的，因此单纯用本

体学模型、逻辑学模型和指称模型等形式化的体系来解释意义是行不通的。第三，语料库研究还表明，语法与词汇在实现意义时是交织在一起的，必须整合描述。意义的基本单位是由多个单词构成的词项，语法则扮演着管理意义、组合成分和构筑词项的角色，因此两者在意义的表达上是相互依存的。第四，语料库研究为语言教学提供了许多有益的启示。语言学习者除了一般的听、说、读、写技能之外，还需要掌握另一套技能，即将话语切分为有意义成分的能力，区分向心式结构与离心式结构的能力，使用语言对语言进行认识、讨论、重组的能力以及正确释义的能力。

Sinclair 在 2003 年语料库语言学国际会议上预言：语料库研究的发现极有可能导致语言学理论的深刻变革。事实上，我们现在已经切切实实地感受到了语料库语言学研究成果正在积极地推动和促进语言学及其他多种学科领域的研究，并得到了广泛的应用。

为了使国内更多的人了解语料库语言学，并投入到相关的研究和应用中，我们采用理论阐释和研究实例分析相结合的方法，撰写了《语料库语言学的研究与应用》一书。

本书共分七章。第一章对语料库语言学进行概述，明确了其作为语言学分支的学科定位；第二章与第三章分别对语料库的建设和发展、加工和利用进行了介绍；第四章至第七章则分别阐述了语料库语言学在外语教学和翻译、机器翻译、文体学研究等领域的具体应用。

在本书的写作过程中，我们参考了国内外学者的多种著作，在此谨致以诚挚的谢意！我们也热忱欢迎读者朋友们不吝批评、指正！

余国良

2008 年 12 月

目 录 CONTENTS

第一章 语料库语言学概述 …… 1
1.1 语料库语言学的学科定位 …… 1
1.2 语料库语言学的研究方法 …… 4
1.3 语料库语言学的研究目标与内容 …… 7
1.4 语料库语言学的发展趋势 …… 16

第二章 语料库的建设和发展 …… 19
2.1 语料库的定义 …… 19
2.2 语料库的发展阶段 …… 21
2.3 语料库的分类 …… 24
2.4 国外重要语料库介绍 …… 27
2.5 国内语料库的建设和发展 …… 29
2.6 语料库建设的几个问题 …… 33
2.7 语料库的发展方向 …… 39

第三章 语料库的加工和利用 …… 40
3.1 语料标注的原则 …… 40
3.2 语料标注的模式 …… 42

3.3 语料标注的类型 …… 44
3.4 语料库的数据管理方式 …… 52
3.5 语料库查询的基本功能 …… 53
3.6 语料库应用软件的开发和使用 …… 58

第四章 语料库语言学与外语教学 …… 64
4.1 对外语教学观念的影响 …… 64
4.2 对教学大纲设计的影响 …… 66
4.3 对教学内容的影响 …… 68
4.4 对外语测试命题的影响 …… 75
4.5 对外语学习模式的影响 …… 77
4.6 对中介语研究的影响 …… 80

第五章 语料库语言学与翻译研究 …… 84
5.1 翻译语料库的种类 …… 85
5.2 语料库在翻译理论研究中的作用 …… 88
5.3 语料库在翻译实践研究中的作用 …… 95
5.4 关于翻译语料库的几个问题 …… 103

第六章 语料库语言学与机器翻译 …… 106
6.1 机器翻译的定义 …… 106
6.2 机器翻译的发展历程 …… 107
6.3 机器翻译的方法分类 …… 110
6.4 机器翻译系统的双语对齐 …… 117
6.5 机器翻译的实现过程 …… 120
6.6 机器翻译题材的可译度 …… 124
6.7 基于语料库机器翻译的开发工具 …… 125
6.8 机器翻译的发展方向 …… 127

第七章 语料库语言学与文体学 …… 130
7.1 文学文体学研究 …… 131
7.2 语篇结构与文体关系研究 …… 136
7.3 口语体的语言特征研究 …… 137
7.4 学术语言的文体研究 …… 139
7.5 话语标记研究 …… 143
7.6 词汇的文体意义研究 …… 146
7.7 其他相关研究 …… 148

附录 …… 151
参考文献 …… 168

第一章　语料库语言学概述

语料库语言学至今已走过了40余年的发展历程，国内的相关研究也有20年的历史了。正如J. Thomas等人（1996）所指出的那样："语料库语言学已经成为语言研究的主流。基于语料库的研究不再是计算机专家的独有领域，它正在对语言研究的许多领域产生愈来愈大的影响。"那么，究竟什么是语料库语言学？语料库语言学是否真的算得上是语言学的一个分支学科？如果是的话，它所担负的研究任务又有哪些？这些理论问题是我们必须首先解决的。

1.1　语料库语言学的学科定位

什么是语料库语言学（corpus linguistics）？我们现引述西方几位语言学家对它的定义：

（1）根据篇章材料对语言的研究称为语料库语言学（K. Ailmer & B. Altenberg，1991：1）。

（2）基于现实生活中语言运用的实例进行的语言研究称为语料库语言学（T. McEnery & A. Wilson，1996：1）。

(3) 以语料为语言描述的起点或以语料为验证有关语言的假说的方法称为语料库语言学 (D. Crystal, 1991: 86)。

中国语言学家对语料库语言学的定义是:

"语料库语言学是 80 年代才崭露头角的一门计算机语言学的新的分支学科。它研究机器可读的自然语言文本的采集、存储、检索、统计、语法标注、句法语义分析, 以及具有上述功能的语料库在语言定量分析、词典编纂、作品风格分析、自然语言理解和机器翻译等领域中的应用。" (黄昌宁, 1990)

上述几个定义尽管表述方式有所不同, 但我们可以据此得出语料库语言学的两层主要含义: 一是利用语料库对语言的某个方面进行研究, 即语料库语言学主要是指一种新的研究手段。二是依据语料库所反映出来的语言事实对现行语言学理论进行批判, 提出新的观点或理论 (顾曰国, 1998)。实际上, 正是基于对这两层含义的把握和理解程度的不同, 使得国内外相关研究者在对语料库语言学的学科定位的问题上产生了不同的看法。

一种观点是: 语料库语言学不能称作独立的学科领域, 它只是一种基于语料库的语言研究方法而已。

Tognini-Bonelli 指出, 语料库语言学并不是一个真正意义上的科学研究领域, 只不过是为语言研究提供了一种方法论基础, 同时它又给语言学的研究提供了新的哲学思路, 所以它是介于理论与方法论之间的一种东西 (Tognini-Bonelli, 2001)。

国内持相同观点的代表人物有丁信善、黄昌宁、许家金等。

丁信善认为, 语料库语言学的研究范围主要包括以下两个方面: 一是对自然语料进行标注; 二是对已经标注的语料进行利用和研究, 它所研究的并非是语言本身的某个方面, 因此从方法论角度看, 语料库语言学不仅可以用于研究语言系统的各个方面, 而且可以应用于语言学之外的其他领域 (丁信善, 1998)。

黄昌宁也提出: "作为一个学科的名称, '语料库语言学'

与‘语法学’或‘语义学’不同，它不属于语言自身某个侧面的研究，而是一种以语料为基础的研究方法。”（黄昌宁，2002）

许家金从理论构架和研究方法两个方面进行了阐述，认为语料库语言学尽管为语言学的研究提供了新的哲学思路，但它在语言研究方法论上的意义更加深远，因此，“基于语料库的研究方法这一提法倒是更能准确地反映语料库语言学的性质和定位”（许家金，2003）。

另一种观点是：语料库语言学是一门新兴而独立的语言学分支学科。其代表人物有杨惠中、潘永樑等。

杨惠中在其主编的《语料库语言学导论》一书中明确指出，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。但是因为语言现象涉及人类活动的一切方面，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、语用学等众多跨学科的研究领域，而“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”（杨惠中，2002）。

潘永樑的看法也与此类似。他认为，语料库语言学虽然不像社会语言学或心理语言学那样研究社会或心理与语言之间的关系，而是以语料库为手段来研究语言，但它发展至今，在语言研究中的地位已得到了广泛的承认，成为一门独具特色的语言研究学科（潘永樑，2001）。

我们认为，语料库语言学为语言研究提供了一种全新的研究思路，以大量真实的语言使用实例为研究对象，借助于统计学手段和方法得出客观、可靠的语言数据，从而寻找语言使用的规律，并对先前的语言理论进行验证或修改。这一研究迄今已取得了令人瞩目的研究成果，因此可以说已经成为现代语言学的一个重要分支。

1.2 语料库语言学的研究方法

尽管语言学界对语料库语言学究竟能否称为一门独立的语言学分支学科还莫衷一是，但是在关于它给语言研究提供了新的方法和思路这个问题上的看法却是一致的。那么语料库语言学的研究方法到底新在何处？这种研究方法与以前所使用的方法到底有何不同？

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

1.2.1 内省法

内省法主要是转换生成语言学家所采用的研究方法，他们以语言学家本人为资料提供人，根据自己的语感或称直觉（tuition）作为判断语言现象歧义、正误、可接受性的出发点和依据。

20世纪初期，现代语言学之父Ferdinand de Saussure提出在研究语言时，应该排除差异、找出共性，即语言学研究的是同质的、抽象的语言形式，是凌驾于个人和社会之上的一个具有高度独立性的抽象的符号系统。这种结构主义观点将语言视为一个相互限定的存在体系统，倡导根据系统和结构来描述语言特点的方法，此观点渗透到许多语言学流派，成为传统语言学的基础。但是与传统语法一样，这种方法的焦点仍然集中在语言的表层语法结构上，由此遭到了由Noam Chomsky开创的转换—生成语法（transformational-generative grammar，简称TG Grammar）学派的批评和抨击。

转换—生成语言学家认为，人的语言能力（competence）是一种天赋的生理和心理现象。根据这种能力，操某种语言的人不

但能够识别某一个句子是否符合语法，而且可以生成符合语法规则的数量无限的句子。转换生成语言学家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、语义和音位结构等。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，等等）：

NP → Det Noun

VP → Verb NP

VP → Verb NP PP

Det → the

Noun → girl，dog

Verb → chased

上述语言研究的内省法可以看做是着眼于语言能力的研究方法（杨惠中，2002）。

1.2.2 诱导法

诱导法是一种调查方法，指通过实地或问卷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，通常采取控制的方法诱导出受试者（informant）对句子或句子中某个成分的判断，要求他们确定句子中有没有错误、句子的可接受程度、句子的可理解程度以及提供其他类似的数据和信息，主要有会话完成任务（discourse completion task，简称 DCT）、角色扮演（role play）和真实话语观察（authentic speech observation）等不同方式。

但是，这种方法也存在某些不足，正如 Beebe & Takahashi 所指出的那样："……自然数据因说话者、听话者及发生语境的不同而失去可比性，除非我们能够限定话语的语境，而这样的限定又产生了其他的局限性（... natural data give us lots of examples that are not all comparable in terms of speakers，hearers，and social

situations, unless one or two situations are selected, and this poses other limitations.)。(Beebe & Takahashi, 1989: 120) 此外，社会调查的规模不可能很大，受试者的主观判断容易受到实验者提示的干扰，而且不同受试者的学历水平也参差不齐，等等，诸如此类的因素都会影响调查结果的有效性。

诱导法既依靠客观调查，又依靠受试者的主观判断，因此可以说诱导法是部分着眼于语言能力、部分着眼于语言运用（performance）的研究方法。

1.2.3 基于语料库的方法

基于语料库的研究方法也是从调查真实的语言材料出发，但与诱导法不同的是，它所处理的是收集于语料库之中的不受限制的真实的语言材料。其基本手段是采用概率分析法，即通过对语料库中的语料进行观察、比较，在统计分析的基础上得出语言运用的概率信息，再反过来以概率信息为依据分析真实的语言材料。由于语料库中的语言材料都来源于人们的实际使用，因此语料库语言学的研究结果具有可靠性，并且在调查过程中排除了研究者和受试者的主观判断的影响，其研究结果又具有客观性。

语料库语言学所收集的是语言事实，即人们实际使用的语言素材，所以语料库语言学方法可以说是着眼于语言运用的研究方法。

Biber 等在《语料库语言学》一书中指出，以语料库为基础的研究方法具有以下主要特点：(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) 以语料库作为分析研究的基础；(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 (Biber et al, 1998)。

从下面的列表中可以看出上述三种研究方法各自所具有的长处及存在的不足：

表 1–1

特　点	语言学研究的语言数据		
	I	E	C
1. 是否简单易行，速度和费用如何	+	−	−
2. 是否受到被试者态度的影响	+	+	−
3. 能否提供概率信息	−	−	+
4. 是否客观	−	+/−	+
5. 能否收集大规模数据	−	+/−	+
6. 非母语者能否使用	−	+	+
7. 能否涉及不同文体或语域	−	−	+
8. 会否受到疲倦或犹豫不决的影响	−	−	+
9. 能否进行历时研究	−	−	+
10. 是否为真实的、实际的语言运用	−	−	+

I = 内省法　E = 诱导法　C = 语料库方法

（Svartvik，1996；转引自杨惠中，2002）

潘永樑认为，语料库语言学无疑会广泛地丰富语言学的理论和方法。从目前的情形来看，它不仅发展了对实际语料的研究方法，而且也不排除内省的方法。例如，实际语料常被用来验证关于语言的一些假设，这些假设可以是从语料中归纳出来的，也可以是研究者内省的结果，或者两者兼而有之。因此可以说，语料库语言学有助于形成兼有早期美国结构主义语言学语料归纳法的优点和乔姆斯基内省法长处的综合的研究方法（潘永樑，2001）。

1.3 语料库语言学的研究目标与内容

王克非等对语料库语言学研究的主要目标给出了明确的界

定："以计算机储存大量的真实语料，对语料做各种带有研究目的的加工标注，利用研制的检索工具对语料进行快捷的搜索和分类，以发现并分析以往因条件所限而未能注意的语言现象。"（王克非等，2004）

语料库语言学研究包括两个大的方面：一个是对于语料库本身的建设与加工。它包括针对语料库的以下几方面的研究：设计与建设、标注与规范、词类标注、句法标注、双语库的对齐加工及检索用的软件工具的开发，以及语料库资源的共享平台构建等。另一个是在语言学上的应用。语料库语言学发展至今，几乎可用于与语言研究相关的所有领域。下面对其应用范围进行简要列举（对于部分重要领域的应用情形本书将分章阐述，详见第四章至第七章）。

1.3.1 词典编纂

大型语料库对于词典编纂无疑极有用处。Biber 等（1998）认为语料库可以回答词典编纂所面对的六大问题：（1）根据词汇在大量自然语境中的使用情况，决定其意义；（2）决定词频，从而编制常用词表与非常用词表；（3）决定某个词汇具有哪些非语言的联结（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；（4）决定词项的搭配及其在不同语域中的分布；（5）决定某词的义项及其用法分布；（6）决定同义词的使用与分布，从而了解语境对词义的选择、搭配与语域的关系。

《朗文英语词典》在编纂时所依据的语料库网共包含以下三个子语料库：一是朗文/兰卡斯特语料库，收集了英美各种类型的书面语 3000 万词；二是朗文学生语料库，是世界上唯一的收集了各国英语学习者书面语的语料库；三是英语口语语料库，含世界上第一个日常英语会话语料库。该词典正文中的词义解释、所用例句和词语使用频率标记等都得益于朗文/兰卡斯特语料库。词

典编纂者利用计算机终端可以轻松地从数百万甚至数千万词次的语篇语料中调出某个单词或短语的用法实例，不仅使得编纂和修订速度大大加快，而且能够及时得到新的语言信息，大量自然语言实例的使用无疑会使词的释义更加完整和确切。

英国考林斯出版社和伯明翰大学合作编辑出版的《国际通用词典》在很大程度上打破了词典编纂的传统，从词条的选定、用法到释义的先后顺序等都依据了由 2 亿词次的 COBUILD 语料库中统计出的频率。由于语料库的素材来自实际使用的语言，因此以其作为词典援引的例句无疑更具真实性和准确性。

总之，利用语料库编纂词典，不仅能够反映语言的真实变化，而且编纂周期也会大大缩短，从而提高工作效率。

1.3.2 语言研究

除了词汇研究之外，语料库的检索功能和统计手段同样给口语（言语）研究、句法研究、语义研究、语用研究与话语分析、语篇分析等提供了极大的便利。

（1）词汇研究。借助语料库方法，使得传统词汇学的范畴得到了扩展。词汇研究主要包括：词语搭配、词义及词的用法分布、单词拼写错误和语义韵等方面。

词语搭配是语料库语言学研究活动中最为活跃的研究领域，处于中心位置（卫乃兴，2001）；对词义及词的用法分布的研究主要关注词的不同意义和用法；词汇拼写错误的研究主要是从对比分析、错误分析及心理语言学等方面对语言（包括外语）学习者的拼写错误进行分析，探究错误的原因；语义韵研究不仅涉及特定词语或短语本身，还涉及两者的习惯搭配、相互作用而产生联想意义的微妙关系等。

（2）口语（言语）研究。利用口语语料库研究言语，主要集中在韵律层面。这类研究大致可以分为三种情形：一是探究韵律的实质以及言语的韵律成分如何与其他语言层面相联系，如用归

纳法从语料中生成假说并验证关于语调群的分界问题等；二是探究韵律标注的基础，如研究不同的语言学家对同一语调群的感知差异等；三是从韵律的角度探究语篇的类型，如研究影响语调模式的各种因素等。不过，目前对于口语体的语言特征研究也引起了不少研究者的关注。

(3) 句法研究。句法及语言结构分析对于语言教学、文本分类、机器翻译及信息检索等都具有重要作用。借助于语料库，研究者能够统计出各种语言结构的频率分布；调查语法结构与语言各种分析层次之间，以及语言因素与非语言因素之间的关系；解释讲话者选择特定形式的语言表达的原因（Biber，2000），因此，句法是利用语料库开展研究最多的语言层面之一。自 20 世纪 80 年代以来，出现了大批难以简单地归于唯理语法或描写语法的语法研究者，他们既不是通过内省的方法构建语法理论，也不是通过描写归纳生成新的语法理论，而是致力于利用语料库来验证唯理派的语法学说，探查出唯理语法能够在多大程度上解释语料库数据以及要完全解释这些数据需对其做多少修改。

词语的歧义问题一直是句法研究的“瓶颈”问题，只有把静态分析和动态知识结合起来，才能比较有效地提高句法研究的正确性。

(4) 语义研究。语义分析的语言学基础是语义学理论，其主要任务是得出语言文本的词汇语义单元和它们之间的依赖关系。

语料库用于语义研究主要体现在以下两个方面：①语料库可用来为词项赋义提供客观标准。语义区别是与词法、句法及韵律等上下文的语篇相关的，通过语料库来考察这些相关成分，可以找到特定语义区别的客观指示。②语料库有助于建立语义的模糊范畴的梯度概念。语义区别作为一种认知范畴，是存在模糊界的，这就意味着义项之间并非简单的包容与非包容的关系，或是非此即彼的关系，而是一种与包容比例有关的梯度关系。语料库

对于判断和揭示这种比例关系的存在及其大小具有重要作用。

范云等（2005）在《汉英平行语料库双语语义对应空位研究》一文中指出，由于一种语言的词汇同义现象非常普遍，语料库程序通常难以作出准确的辨认和处理。就这一问题，常见的处理办法是模糊检索，就是语料库向检索者提供一系列符合检索要求的某个或某些字段的检索结果，供检索者进一步筛选，这在很大程度上降低了检索失败的概率，但检索者还要再花精力去进行主观取舍（因为很多时候能匹配部分字段的检索结果非常多，有的还很牵强），因此只能是部分地解决了问题，而要更进一步解决难题，则应在语料收集和加工上做文章。语义学告诉我们，很多同义或近义词固然很容易混淆，但大多有不同的感情色彩和不同的使用场合以及不同的搭配群组，在整理语料时，如果充分考虑这些问题，某些字段的参数将会更加丰富，也会更加系统，这实际上回到了细分语料范畴的问题上。在此基础上，检索页面可以对用户的检索要求提出更细致的范畴询问，检索的成功率就有了提高的可能。当然，要完成这么细致的检索任务则需要一套更加成熟的算法支持。

（5）语用研究与话语分析。这方面的研究对语料库的标注、赋码和数据分析都提出了较高的要求，除了可以分析口语中的词汇和句法等书面文字特征外，还能够对话语中所使用的短语结构进行语用功能分析。

意大利 Verona 大学的 R. Facchinetti 通过对英国口语语料库和青少年口语语料库中情态助动词的社会变体和语用功能分析，发现 can，would 和 will 的使用最为广泛，must 和 shall 的使用逐渐减少，不过 will 和 may 在私人会话中明显减少，would 则在公众演说中使用频率最高。还有一些年轻人普遍使用情态助动词来表达试探性语气，尤其喜欢用 might 和 should。然而在非试探性的语境场合，女孩子却比男孩子更多地使用 can，shall，will 和