北京市高等教育精品教材立项项目

计算机语料库的建设与应用

王建新　编著

清华大学出版社

计算机语料库的建设与应用

王建新　编著

清华大学出版社
北京

内容简介

计算机语料库是可以用计算机处理的电子文本库，是提高自然语言处理系统性能的重要工具，又是新兴的语料库语言学的研究基础，它对信息产业、词典出版、外语教学与研究等领域的发展影响巨大，因而日益受到重视。本书介绍如何收集建立计算机语料库和在诸多领域如何开发利用语料库，可作为英语、计算机、中文信息处理、信息与网络管理等专业的研究生和高年级本科生相关课程的教材，也可作为相关专业的研究生和毕业生选择与确定科研与毕业论文题目的参考书，亦可供信息产业的技术和管理人员、高校相关专业的教师学习参考。

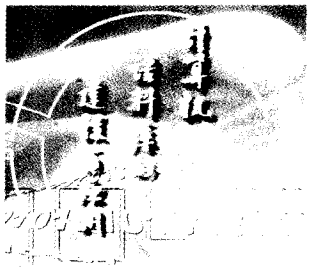
图书在版编目(CIP)数据

计算机语料库的建设与应用/王建新编著．—北京：清华大学出版社，2005.9
ISBN 7-302-10878-1

Ⅰ.计…　Ⅱ.王…　Ⅲ.计算机应用－英语－语言－教学－研究生－教材　Ⅳ.H319

中国版本图书馆 CIP 数据核字(2005)第 037988 号

出 版 者：清华大学出版社　　**地　　址**：北京清华大学学研大厦
http://www.tup.com.cn　　**邮　　编**：100084
社 总 机：010-62770175　　**客户服务**：010-62776969
责任编辑：陈国新
印 刷 者：北京密云胶印厂
装 订 者：三河市新茂装订有限公司
发 行 者：新华书店总店北京发行所
开　　本：185×260　**印张**：20　**字数**：474 千字
版　　次：2005 年 9 月第 1 版　2005 年 9 月第1 次印刷
书　　号：ISBN 7-302-10878-1/TP·7236
印　　数：1～2500
定　　价：39.00 元



前　言

在信息产业界、语言工程界、词典出版界、外语教学与研究界，计算机语料库的巨大作用正日益显现，越来越得到普遍重视。计算机语料库是自然语言处理中统计方法的基础。基于语料库的统计方法，已经成为提高语言自动处理系统性能的突破口。计算机语料库又是新兴的语料库语言学的研究基础。近年来语料库语言学发展十分迅速，硕果累累。

本书讲述什么是语料库，什么是计算机语料库，语料库有哪些种类，目前世界上有哪些著名的语料库，语料库在历史上起过什么作用，对语言学研究有何用途，对语言工业有何用途，对英语教学有何用途，如何设计和收集建立语料库，如何开发利用语料库，如何上网利用现成的语料库。

本书是为英语教育、英语语言学、计算机信息处理、信息与网络管理等专业研究生编写的教材，也可供拟报考以上专业研究生的高年级本科生、高校的中青年教师、从事或有志从事有关信息技术工作的人员参考。相关专业的硕士研究生与博士研究生，在选择与确定毕业论文的题目和科研项目时，也可以参考本书。

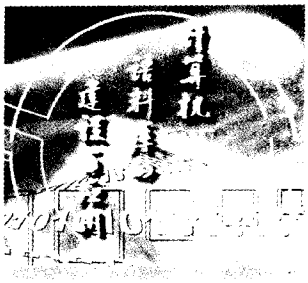
全书分为上下两篇。上篇介绍计算机语料库的建设与加工，下篇讨论计算机语料库的开发与应用。其中上篇第 1 章是对计算机语料库的用途与意义的总述；第 2～7 章介绍语料库的分类与发展；第 8～13 章讨论语料库的设计、建设、标注与加工方法。下篇第 14 章介绍开发语料库的主要软件，特别是索引软件的功能；第 15～24 章探讨语料库在语言学、应用语言学、词汇学、英语教学、辞书编撰等方面的应用；第 25～27 章介绍语料库在自然语言处理方面的应用；第 28 章讨论计算机语料库的发展趋势。附录 1～4 供感兴趣的读者进一步参考，附录 5 列出了一些有用的网址。书中的一些重点内容用黑体显示，以方便读者阅读和查找相关内容。为了避免误译，书中的外国人名一般使用原文。

本书的出版得到北京市高等教育精品教材建设立项项目和北京邮电大学语言学院配套经费的资助；书中引用了多位作者的研究成果，作者一并深表谢意。

虽然积累素材近 10 年，但本书是作者在繁重教学之余挤时间写成的，因时间仓促，书中难免有疏漏之处，衷心欢迎批评指正。

王建新

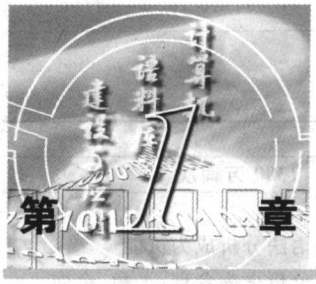
2005 年 4 月



目　录

上篇

计算机语料库的建设与加工



第1章 计算机语料库的重要意义

计算机语料库普遍受到重视 自20世纪80年代以来，计算机语料库的建设与应用，引起了许多国家信息技术领域和语言学界日益浓厚的兴趣。计算机语料库对自然语言处理的不同方面（如话语识别、人机对话、信息提取、网页分类、机助翻译、文字处理等）的重要性和蕴藏的潜力，得到了国际计算语言学界的广泛认可。作为信息工业的重要部分，计算机语料库的研制与建立，在发达国家已经受到高度重视并积累了丰富的经验。例如，英国就把建设大规模计算机语料库作为发展信息技术的战略任务，并投放巨资加以实施。同时，计算机语料库以其语料的充分性、客观性、可验证性、使用的便捷性，正在使语言学的研究与应用产生量与质的飞跃。新兴的语料库语言学，就是围绕着建设计算机语料库与利用它进行研究和应用的一门学科，是将语言学与计算机技术整合起来的一门新兴交叉学科。计算机语料库所提供的丰富鲜活的真实语料与强有力的开发软件相结合，已经使词典编纂发生了革命性的变化，使语言学研究产生了丰硕的成果，使词汇学从过去的次要地位上升到主流地位。在外语教学方面，计算机语料库可以使学生自主地去学习和探索，从大量的语言现象中体验到对所学语言更深切的认识。

Stubbs（1996：231～232）把利用计算机语料库进行研究的意义，比作显微镜和远望镜对科学家的意义。显微镜和远望镜的发明，使科学家无论从微观角度还是从宏观角度，都能够观察到过去从来没有观察到的事物。将计算机、软件和大语料库相结合，已经使语言学家看到了过去从未看到或想到过的现象，发现了过去从未留意过的类型。

Stubbs的观察在Renouf所举的实例中得到了印证。Renouf（1987）把计算机语料库的用途与摄影领域中照片处理的分辨率（resolution）加以对比。在照相时，如果要得到更清晰、更详尽的图像，需要高分辨率的镜头。用语料库分析语言，为了获得对一个词形（word form）的更清楚与更有区分性的了解，所研究的资料必须足够充分，才能使研究可靠。她举出的实例见表1-1。

因此，基于语料库的研究方法，已经成为当代语言学和应用语言学研究中的一种主流方法。

语料库对自然语言研究必不可少 计算机语料库普遍受到重视绝非偶然。这与它在自然语言处理和语言学研究中的重要性密切相关。要使自然语言处理系统成功地为某一目的处理某一语言，必须把它放在语言是怎样被真实使用的证据上。对语料库的分析正是获

表 1-1 用语料库分析语言的一个实例

词 语	语料库规模	
	730 万词次	1300 万词次
faddish, moot, off-key	无证据	出现证据
advisedly, accede	极少例证	例证增加使其语法分析成为可能
faggot, vainly 等		充分的例证使分析更为合理、平衡、全面
sorely 等		充分的例证提供了其大量搭配方面的用法

(据 Renouf 1987)

得这种证据的主要方法。自然语言处理系统应用方面的实践表明,有必要利用亚语言(sub languages)的有关特征与性质来扩大计算机的应用范围,提高系统的效率。亚语言指的是同一语言被不同使用者为不同交际目的在不同交际场合中的不同使用情况。它提供了若干类语言现象在范围与分布上相联系的区别。如果能客观地辨认不同的亚语言,充分地描述其相关的特点,其区别与具体的特点就能在自然语言处理中被有效利用。而语料库分析则是能辨别与描述亚语言的惟一已知的工具。语料库是能提供各种语言、亚语言的统计数据的惟一可能的资料来源。统计数据包括音素、字母、词、类、结构、搭配、系列、各种关系以及其他特征的频次与分布。而自然语言处理中最新、最成功的系统正是特别依据统计的证据与方法的。

因此,精心设计的具有代表性的分为不同层次与类别的语料库,对研制与测评自然语言处理系统具有基础性的意义。

在计算语言学与语料库语言学长足发展的基础上,1986 年在欧共体的 Tours 会议上,正式提出了语言工业的提法。语言工业包括:

(1) 计算机对应用语言学的支持。例如,对词汇学、翻译、语言教学以及其他方面的支持。

(2) 在自然语言处理基础上所研制的计算机系统。例如,话语(speech)的分析与合成、自动加索引与提供文摘(abstracting)、办公自动化、机器翻译以及对通信的支持等。

语言工业的产品,从拼写检查软件到信息提取软件乃至机器翻译软件,都需要强有力的自然语言处理成分,以便能够处理真实篇章。20 世纪 80 年代随着计算机功能的巨大提高,使研制充分大的文章库和大规模的词汇库作为主要的研究与发展领域成为可能。在这种研究中,计算语言学加入到语言学、词汇学、心理语言学及各种人文学科的行列中来。文章库被认为是描述自然语言在各种交际环境中被真实使用的基础性的资料来源。为了使自然语言处理系统能应用于真实的语言环境,它必须能处理几十万、几百万的词汇项,因而研制大型的文章库和词汇库就成为语言工业中最紧迫的任务。Zampolli(1994:22 ~ 23)指出:自然语言处理与计算机语料库语言学在几个方面是互补的,例如,前者有助于提出分析语法的形式、词类与句法标注的工具与类型,后者则发展了语料库的收集与处理语料的方法、统计学的语言分析法、亚语言的辨认与描述方法。

语料库方法成为自然语言处理的重要方法 计算机语料库在自然语言处理中日益受到重视的程度,体现在相关领域的国际会议上。第 36 届计算语言学会(ACL)年会和第 17

届国际计算语言学大会(COLING)于1998年8月10日~14日在加拿大的蒙特利尔大学召开,出席者超过800人,会议论文等反映出的主要动向是:

(1) 统计模型和语料库方法占据文本和语音处理技术的主导地位。以语言学知识指导统计语言的建模,已被证明是进一步提高自然语言处理系统全面性能的关键。

(2) 重视语义学研究,尤其是词汇语义学资源的建设,以及语料库的语义标注、文本的词义排歧。期盼语义和句法的紧密结合使单纯句法分析走出困境。

(3) 社会需求推动,使机器翻译、信息检索、信息抽取、语音输入和文-语转换等应用领域市场广阔,推动了理论研究。体现词汇主义(lexicalism)思想的语法理论,如依存语法(dependency grammar)、范畴语法(categorical grammar)和链语法(link grammar)受到计算语学界的重视。对语言这种多层次复杂对象的建模理论和涉及语言知识自动获取的计算学习理论等,普遍受到关注(黄昌宁,孙茂松 1999:54~57)。

在新加坡召开的主题为"最新技术进展与应用"的中文计算语言学国际会议ICCC'96(International Conference on Chinese Computing)上,入选大会论文集的70篇论文中,有关语料库建设和应用的论文达16篇,占入选论文总数的23%,其中包括语料库建设与统计方面的文章8篇,语料库应用方面的文章8篇。这反映出语料库方法在中文信息界也开始受到普遍关注(黄昌宁,孙茂松 1996:44~48)。

冯志伟先生(1996:511~512)指出,近两年来国际计算语言学越来越注意未经编辑的、非受限的大规模真实文本的处理,语料库语言学在自然语言处理中的地位越来越重要,语言知识的颗粒度正日趋精细,对语料库中的非受限文本的自动标注已经取得了令人鼓舞的成绩。国际计算语言学界把这种基于语料库和词库的经验主义方法,确立为未来一个时期内计算语言学发展的战略目标。令人高兴的是,我国在汉语真实文本自动语义标注方面已经取得了突破性的进展,引起了世界计算语言学界的关注。这种经验主义的研究方法有助于全面地观察语言现象,克服传统语言研究的局限性和片面性。但是,在采用这种经验主义方法的同时,我们不能忽视理性主义的方法,即基于规则的自动句法——语义分析方法,建立一些实用的自然语言处理系统。

计算机语料库使语言学研究产生了深刻变化 Sture Allen(1997:304~305)认为,自然语言处理特别是语料库、词汇学和研制软件工具方面的重要工作,很显然帮助了将语言学从丧失它的立足点挽救过来的过程。人们都曾经目睹过一些很大胆的思索,却难以经受住真实语言的检验。这就是为什么我们呼唤真实语料的原因。另一方面,语料库工作和词汇学工作,作为科学的范畴,毫无疑问都涉及理论的思考。实证与理论二者应该相辅相成,相得益彰。

传统语言学多数依靠单个的研究者所能经历与记住的资料,没有直接的观察与测量。由于缺少足够的资料,语言学变得几乎全部是内省式的(introverted)。流行的方式是向内心寻求答案,而不是向社会寻求答案。直觉成了关键,强调的是把语言结构比作各种抽象的形式和模型,几乎不涉及语言的交际作用。

由于语言学的内省阶段与计量数学的兴趣热同时出现,一种很有限的计量语言学流行

起来,在许多地方仍旧享有正宗的地位。总的来说,过去的语言学似乎集中在过窄的研究范围中。首先,由于跃过了语言的许多具体表现,过快地抽象给语音学和话语识别技术带来很大困难。第二,语言学又不够抽象,它在仅仅比所确立的单位的初始阶段抽象一步的程度上,就想解释语言中的所有含义。第三,语言学从很小的单位,如音素和字母,自下而上解释,很难触及有长度有难度的全篇。即使触及了全篇,语言学也似乎无法在大、小单位之间保持联系(Sinclair 1991:1,14)。

Knowles(1996)回顾了传统语言学的发展历程,比较了早期英语语法、拉丁语法、结构语言学、转换生成语言学和语料库语言学的特点,指出任何一种语言学理论,都得对存在的许多不同类型的资料的关系做出形式上的解释。从本质上说,这些关系与现代的数据库所处理的关系相似,因而有理由从数据库角度来考虑理论语言学的关系。

Knowles 认为,语言在文章中得到清楚的表现。传统上语言学把对文章的概括性看法存储在词典中,把对词典中条目的概括体现在语法中。从这一角度观察,过去几个世纪对语言学的研究途径表面上差异很大,实际上异途同归。

单词与范式(word-and-paradigm)语法假定,词形学(morphology)是语言学结构的中心。传统的英语语法家认为词典是中心,这一观点仍然是人们对语言的普遍看法。结构主义语言学家认为音位学(phonology)是中心,用准确符号(narrow transcription)转写发音,研究真实或想象的话语。生成语言学家认为句法是中心。这些观点的共同之处,是它们全都有不言自明的任意性。

当语言被实际使用时,它是在社会的环境中用于交际目的的。在交际活动中,正常表现出来的语言结构的惟一方面,是所用的语篇文章(text)。主流的理论语言学恰恰始终把语言的正常表现形式推到了背景之中。

而语料库语言学由于从真实的文章开始,有可能形成一种新的语言学,具有更好的理论基础。计算机语料库潜在的一种可能性,是削弱有关语言学理论的一些传统假设。

计算机语料库能够拓宽语法的描述范围。Aarts(1999:3 ~ 5)认为,20 世纪对英语语言研究的历史告诉人们,描述语言的用法(language use)始终是语法学家的主要目标。这些语法学家的鸿篇巨著被称为伟大的传统语法(Great Tradition)。代表人物有 Kruisinga、Poutsma、Jespersen、Quirk 等。然而这批伟大的传统语法对英语语言的描述,实际上只是对一种方言、一种社会文体、一种媒介的描述,即对"标准英语"的描述。更具体地说,是对英国东南部受过良好教育的人们所用的书面英语的描述,用这一种变体(variety)来代表其他,描述偏重于受过教育者的英语用法和书面语用法。

20 世纪末出现的巨大的英语语料库,如含有 1 亿词次的英国国家语料库 BNC(the British National Corpus),含有苏格兰英语同步对话的 HCRC Map Task 语料库(Human Communication Research Centre Map Task Corpus),使"标准英语"变得越来越不明显。这些语料库使研究者对语言形式的种种变化有了空前的认识。对语言用法的研究,已经成为对语言变体的研究。这对描述语言用法的语言模型和工具,都提出了新的要求。

语料库中的资料被称为"真实的资料"、"实际使用的语言"等,这些语料都是行为资料

(performance data)。大多数语料库语言学家,不会主要对语言行为感兴趣,原因是其中包括太多的非语言因素。他们会对语言使用更感兴趣。

语言使用可以作为语言行为的第一个层次的抽象,该层次排除了所有非语言的行为,以及一切和语言分析无关的行为。它指的是语言学的产物,即口语或书面语的文本,而不是产生的过程或语言学行为。

Stubbs(1996:233)指出,语料库常被认为仅仅是"语言行为的资料"。实际上语料库是人们语言的集合,不管是口语的还是书面语的,因而是实际语言行为的取样,是对语言行为的记录。记录与行为本身不同,其结果至关重要。可以用这种记录来研究人类的语言行为。他指出 Popper 曾以温度与气温为例,说明了语料库与语言的关系。气象学家可以收集温度变化的记录来研究气温的变化。气温是自然界物质的属性,而温度记录则是研究自然界的人为安排。气温的一系列变化无法直接研究,但研究温度记录则可以确定气温的变化类型。通过人为的设计,自然界的物质属性可以被转变为一种公共的知识。这种温度的记录不但可以用来研究大气的局部变化,而且可以用来研究大气的长远变化。局部变化是可以粗略观察到的,而长期的变化当然是无法直接观察到的。

按照索绪尔的观点,言语(parole)有任意性,有个人嗜好。这些是实情,但不能全面说明语言的现状。正像混沌理论中的混沌也有秩序一样,言语是有内在秩序的。这种秩序只有在检验大量的资料后,才能显现出来。言语的真实本质,不能仅仅通过咨询自己的直觉就能得到。只有通过观察分析大量的自然出现的语言,才能真正理解人际交流的本质。这种交流指任何有意义的书面的和口头的交流。

Hasan(1992:257 ~ 259)指出,语言的真实使用,会不可避免地涉及社会环境,也会自然地显示我们对内在的语言系统的感觉。因此,以语料库为基础的语言学,使我们能从两个方面探讨语言:一是将语言作为社会符号学来研究,一是将语言作为言语过程的产物即语言来研究。

语料库之所以成为语言学研究的方法与基础,是与下列原因密不可分的(Gerbig 1997:43 ~ 44):

(1) 核实与复现是科学研究当中的标准要求,单靠内省和诱导(elicitation)是做不到的。

(2) 对词汇学而言,语料库提供了几乎取之不尽、用之不竭的实例来源,提供了具体的词在实际上下文中使用情况的统计证据。

(3) 对语法学而言,对语料库的研究是对语言结构进行量化的概率调查的惟一途径。

(4) 对文体学而言,一般性的和专门化的语料库为比较语言学特征提供了广泛的取样和背景资料。

(5) 利用计算机的处理与提取能力和收集在语料库中的自然语言资源,能够对所调查研究的语言学特征做到完全的描述。只要指令得当,计算机就能找到资料中每一个相关的例子,避免了人为性与片面性。

(6) 由于有了大量精选的、有代表性的语言资料,对语言学特征和文体类型之间关系的讨论,就有了更坚实的基础。

计算机语料库催生了语料库语言学 Granger(1998:3)指出,自从20世纪60年代计算机语料库首次出现以来,它已经渗透到和语言研究相关的各个领域:从词汇学到文学批评,到人工智能,到语言教学。这种计算机语料库的广泛使用,已经导致了一门新学科的生成,称为语料库语言学。这个术语,不仅是指一种新的基于计算机的方法,而且如Leech所说,是指一种"全新的研究事业",是对语言全新的思考。这种新的思考,正在对我们有关语言最根深蒂固的认识构成挑战。语料库语言学的焦点是语言行为,而不是语言能力,是语言描述,而不是描述语言的共同点。它既强调量化,又强调质的分析,这种学派与乔姆斯基的方法形成了鲜明的对照,然而这两种方法并非水火不相容。正如Fillmore指出的,这两种语言学家相互需要,更准确地说,在可能之处,这两类语言学家存在于一体之中。

Leech(1992:105～106)认为,语料库语言学不是指研究的领域或范围,而是指进行语言学研究的方法与基础。从原理上和实践中,语料库语言学与其他的语言学分支很容易结合在一起。可以通过语料库来研究语言学、句法,研究社会语言学以及语言学的其他方面。这样做是将基于语料库的方法和技术与语言学、句法学、社会语言学等内容结合起来。

计算语言学是通过计算机来研究语言,也指的是工具或方法,而不是内容。这与语料库语言学相似。现在二者互相重叠。因此语料库语言学应该被称作计算机语料库语言学。它不仅是正在出现的研究语言的新方法,而且是新的产业、新的哲学方法。强有力的计算机已经不仅仅是起计算、辅助作用,而且是获取新知识的根本性的途径,起到对语言从新的角度加以理解的"开门芝麻"的作用。

随着英语语料库的增加,以其为基础对英语所做的研究与日俱增。其迅速增加的趋势,可以从表1-2中看出。

表1-2 基于英语语料库的研究统计

年代	研究规模/项
—1965	10
1966—1970	20
1971—1975	30
1976—1980	80
1981—1985	160
1986—	320

(据Johansson 1991)

表1-2清楚地显示出自1965年以来,基于语料库的研究大约以每5年增加一倍的速度在快速增长。这种增加表明,语言学领域的工作者越来越认识到,需要大量容易得到的真实资料,而不能只靠内省和引证,才能更好地研究英语。同时,在信息科学的每一个分支里,也有越来越多的人逐步认识到,语料库作为活的语言的代表,能利用先进的计算机开拓研究与应用的新天地(Svartvik 1991:8)。

利用大规模多样化的语料库和量化计算的工具,为分析多种语言的结构和用法提供了广阔的天地和新的见解。在Altenberg 1991年所汇集的参考书目中,就列出了大约650项以

语料库为基础对具体语法结构的形式和作用的研究。近年来，英语中这类的专门研究更加五彩纷呈，各种专著也纷纷出版。20世纪80年代以来有关计算机语料库和语料库语言学的专著已经出版了不下几十种。例如S. Johansson 1982年编写的"Computer Corpora in English Language Research"，J. M. Sinclair 1991年编写的"Corpus, Concordance, Collocation"，S. Armstrong 1993年编写的"Using Large Corpora"，A. Renouf 1998年编写的"Explorations in Corpus Linguistics"，都对推动语料库语言学的发展发挥了重要作用。特别是1992年Svartvik编写的"Directions in Corpus Linguistics"更是语料库语言学发展史当中里程碑式的著作，它标志着语料库语言学这门新兴交叉学科的独立与成熟。

计算机语料库在词典出版界掀起了一场革新风暴　基于计算机语料库编写的词典，已经成为上世纪末英语词典出版业中的时尚与主流。特别是英国的辞书出版界，为外国学习英语的学生编写了相当数量的基于语料库的单语或双语词典。尤其是1995年以来，在前期探索的基础上出版或再版了不少高质量的在国际上影响巨大的此类词典。例如：

- 柯林斯COBUILD英语词典(Collins COBUILD English Dictionary)。
- 朗文当代英语词典(英语版)(Longman Dictionary of Contemporary English)。
- 剑桥国际英语词典(Cambridge International Dictionary of English)。
- 麦克米伦高阶英语词典(英语版)(MACMILLAN English Dictionary for Advanced Learners)。
- 柯林斯COBUILD英语学习词典(Collins COBUILD Learner's Dictionary)。
- 朗文联想词典(Longman Language Activator)。
- 钱伯斯基础英语词典(Chambers Essential English Dictionary)。

对这些基于语料库的英语词典的特色与作用，本书中有专门的讨论。

计算机语料库对汉语的研究也正在发生巨大的影响　例如，经过语言学家多年的辛苦努力，于根元先生主编的《现代汉语新词词典》1994年由北京语言学院出版社出版。这部人工编辑的新词词典共收录了1978—1990年12年间的新词语3710条。

北京工业大学计算机学院人工智能研究室，从1991—1997年7年间的《人民日报》、《经济日报》、《新华社电讯稿》中，收集了约2亿字的电子版语料。以含有5万个2字词的大约6万条词语作为"启动知识"对语料进行了处理，将自动处理与人工甄别相结合，共选出新的2字词组11万条，加入到原有的5万条2字词中，形成了含有16万条2字词组的词典。筛选出的新的2字词组中包括了大量实用的新词，像"喷塑"、"蒜农"、"危改"、"市话"、"高检"等(张普　1999:24～33)。

人工收集的汉语新词，12年间共收集了3710条，而计算机语料库对7年之间部分报纸2亿词次语料的分析，辅以人工甄别，却产生了11万条2字新词或词组。语料库反映语言现象变化的全面性和客观性得到了充分显示。

另一方面，把这种基于对电子语料分析收集到的2字词语作为电子词典，无疑会为自然汉语的分析提供强有力的对照词典，对汉语的自动分词、句法分析等提供有力的支持。相形之下，原来的汉语新词词典，就显得不够完备，以其作为参考的电子辞典，势必会将许多新词

视为“未登录词”。

计算机语料库使词汇学研究地位升华 计算机语料库和功能强大的索引软件等工具相结合,使词汇学研究(lexis)从过去在语言学中所占的低于语法和句法的次要地位,猛然上升到语言学研究中的首要地位。这在形式语言学和功能语言学中都是如此。基于语料库对词汇之间搭配的研究,发现了许多过去没有注意到的词语的组合与多词的单元。基于语料库对词汇和语法之间的关联研究,使人们注意到除了语法的层面、词汇的层面之外,还存在着词汇-语法的层面。在这个层面上词汇和语法互相制约,相辅相成。词汇学研究不再杂乱无章、盲人摸象,它已经在三个方面被赋予全新的含义。首先,研究者开始达成共识,词汇学和语法是互相依存的,这在新型的语法书中得到了体现。例如 Biber 等 1999 年编写的“Longman Grammar of Spoken and Written English”就特别强调和着意描述了词汇和语法结构之间的密切制约关系。第二,词汇学研究的重心已经从注重纵向的单词个体的研究,转为注重横向的词汇之间的同现与搭配上来。第三,人们对词汇学在区分不同文体中的作用,有了比以前更充分的认识。计算机和语料库使词汇学发生了脱胎换骨的变化,现在的词汇学和语言的其他层面尤其是语法层面有了更多的联系,研究的内容既有单词也有搭配和多词的单元,范围既有纵向也有横向(Altenberg,Granger 2002:3～5,39)。

计算机语料库使乔姆斯基的观点进一步受到质疑 1950 年以来,以乔姆斯基为首的转换生成语法学家,对探索人头脑中的语言机制,寻求英语的句法规律,做出了杰出的贡献。转换生成语法垄断了语言学界几十年,并且成为早期用计算机处理自然语言时所普遍采用的语法体系。但转换生成语法的盛行,也使语言学界产生了一种错觉,似乎只有转换生成语法才是真正的科学,而对语言资料的实证性研究,则被认为是不全面的。乔姆斯基反对把语言行为资料作为语言学的研究内容,认为语言行为资料受到记忆局限、干扰分心、注意力转移以及兴趣和错误的影响。Stubs(1996:233)指出,这一批评,已经被计算机语料库改变得不再适用。计算机处理的语料库,其资料之巨大,计算之迅速,使量化的研究有了质的飞跃。基于语料库能研究更多的资料、新类型的资料以及过去无法构想与研究的类型。语料库研究中的主要类型是重复出现的搭配,是词法与语法的反复同现。很难想象这些反复出现的语言行为都是错误的。

语言能力和语言行为能否分开值得怀疑。任何能力都以行为的形式表现出来,没有离开形式的空洞的能力。非要把人的语言能力从语言行为中剥离出来加以研究,这本身就意味着大量的概括和抽象,意味着把大量丰富多彩的语言现象排除于研究的范围之外。这样的研究结果,尽管有高度的概括性,却难以全面代表丰富多彩的语言实际。况且,语言学家的语言知识、语言直觉不会是先天而来,而只能是接受周围语言环境影响的产物。个人的语言只是社会语言总体的一部分。尽管人有创造性,能说出从未听别人说过的句子,这种句子本质上还是基于头脑中的语法规则的。而这种规则是建立在言语的一部分现象的基础之上的,是不全面的,带有因人而异的特点。因此,基于语言学家内省的研究,只是对语言现象研究的一种途径,是不全面的,需要用其他的方法加以补充。

从索绪尔到乔姆斯基,都把语言(language)或能力(competence)看作是系统的,是研究

的惟一真实目标。但二者都是抽象的、无法观察到的。言语(parole)或语言行为(performance)尽管是真实的,却被认为是零碎的、不系统的、任意的、因人而异的,因此至多只能作为研究的边缘。20世纪语言学的主流,将自己用这种双重定义(dualism)加以界定,导致研究的对象都是观察不到的,使语言学变得玄而又玄,脱离语言实际。而计算机辅助对大语料库研究的意义,就在于有可能打破这种死胡同的局面。索引软件与其他软件,使查找几百万词次的语料中的词语和类型如同探囊取物一般。在没有这些工具的帮助之前,这是难以想象的。

语料库的证据也给生成语法带来麻烦 Sampson(1987:220～226)从LOB语料库中取出含有39969词次的语篇,从其中所含的8328个名词词组中归纳出747个不同的名词短语类型。分析表明,其中名词短语类型的分布很不均匀。最常见的类型是冠词+单数名词,占14%,共有1135个标形。而468种不同的名词短语类型,则只由一个标形所代表。

这种现象对使用综合的生成语法来处理真实的自然语料提出了疑问。这是因为在一个有一定规模且有代表性的真实英语的取样中,用一种很粗略的分类方法进行语法分析后所辨明的语法区别,必然远远少于任何语言处理系统所要面临的实际语言处理要求。即使在这种情况下,其中近2/3的名词词组也仅由一个例子代表。这就意味着在实际中,以生成语法为基础的语言处理系统,能不能辨认输入的语篇中的结构,只是一种偶然性。既然在这个有一定规模的语言取样中有这么多只出现一次的结构,有理由推断:一定还有许多其他的结构,恰巧在取样中没有出现,但这些结构在作为整体的语言中会同样普通。

问题的关键是在"名词短语"这一范畴中,如何划分合乎生成语法的扩展与不合乎生成语法的扩展之间的界限。当大量语法扩展现象的出现频次变得非常稀少的时候,二者之间的界限就很难划分。Sampson对取样中的名词短语的分类研究表明,名词短语不能分为一组是高频出现的、牢固确立的结构,另一组是少见的变异结构。相反,名词短语的类型连续分散地分布于频次带的范畴之中。

根据Sampson的观察,在取样的名词短语中,若把高频次短语的出现频次视为单位1(39969词中的1135例),较低频次的短语的出现频次是小数,则低频次与高频次之比是0.4:1。也就是说,在所有的标形中,出现频次低的名词短语占到4/10。如果将最常见的名词短语的1/100视为罕见短语类型,则每6个标形中有1个就是罕见的类型。若定位1/1000,则每16个标形中有1个代表罕见的类型。

相信综合的生成语法的研究者似乎假定,在名词短语中低频次短语与高频次短语的连续分布会在某一处中断,从而形成可以清楚界定的两部分。对上述实例的观察,没有提供任何支持这一假定的证据。相反,构成名词短语的标形与名词短语的类型之间存在成正比的斜线关系 $y=0.4x$。这一现象是不利于生成语法的证据。

计算机语料库将有力地推动应用语言学的发展 应用语言学建立在描述语言学的基础之上。应用语言学的很多工作都立足于理论研究发现,以便对真实世界中涉及到与语言有关的作法提出具体的建议。如果作为基础的理论研究缺少足够的实证性的基础,对研究的应用也就会相应地先天不足。一种出路就是直接研究语言在真实世界范畴中的结构与用