语言与计算机丛书

语料库语言学

黄昌宁 李涓子 著

商务印书馆

语言与计算机丛书

语料库语言学

黄昌宁　李涓子　著

商务印书馆
2002年·北京

图书在版编目(CIP)数据

语料库语言学/黄昌宁,李涓子著.—北京:商务印书馆,2002
(语言与计算机丛书)

ISBN 7-100-03364-0

Ⅰ.语… Ⅱ.①黄…②李… Ⅲ.计算机应用-语言学-研究 Ⅳ.H0

中国版本图书馆 CIP 数据核字(2001)第 057971 号

语言与计算机丛书
YǓLIÀOKÙ YǓYÁNXUÉ
语料库语言学
黄昌宁 李涓子 著

商务印书馆出版
(北京王府井大街36号 邮政编码100710)
商务印书馆发行
北京民族印刷厂印刷
ISBN 7-100-03364-0/H·850

2002年4月第1版 开本 787×1092 1/32
2002年4月北京第1次印刷 印张 9⅛

定价:14.00元

目　录

第1章 绪 论

“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”。这是汤姆斯(Thomas)等人1996年为祝贺语料库语言学的主要奠基人和倡导者里奇(Leech)六十诞辰而编纂的语料库语言学研究论文集的开场白。近年来，对语料库语言学类似的说法频频见于导论和方法论的专著及教科书中，它不仅仅是语料库语言学家的自誉，而且正在成为整个语言学界的共识[1]。

第一节 什么是语料库？什么是语料库语言学？

语料库(corpus)顾名思义就是存放语言材料的仓库(或

数据库)。传统上,语言学家用语料库这个术语表示可作为语言研究基础的、大量自然出现的语言数据。这些语料库可以由书面语和(或)口语的样本组成,并通常被用来代表一种特定的语言或语言变体。在计算机出现之前,研究者——特别是词典编纂者,也有语料库,只是规模小、范围窄,因而难以在学术界形成气候。近 40 年以来,语料库这个术语通常指以电子形式保存的语言材料,并被广泛用于语言研究和语言工程。随着计算机功效的成倍增长,语料库在规模、多样性和使用方便等方面都发生了剧烈的变化。与此同时,为了存取和加工语料库所拥有的信息,已经开发了大量专用的软件。计算机语料库迅速成为语言研究的一种普遍资源,现在世界上已经建立了许多规模较大的语料库,有些是国家级的,有些是大学和词典出版商联合研制的。另外,由于个人电脑的迅猛发展,存储数据的硬磁盘造价持续下降,研究者个人也开始建立适合自己研究兴趣的小型语料库。

虽然语料库语言学研究已经历了不短的历史,但还没有一个公认的定义。下面引述两个见诸书本的定义:

定义 1:以现实生活中人们运用语言的实例为基础进行的语言研究,称为语料库语言学。(McEnery & Wilson, 1996)[2]

定义 2:以语料为语言描写的起点,或以语料为验证有关语言假说的方法,称为语料库语言学。(Crystal, 1991)[3]

从上述两个定义可见,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,而是一种以语料库为基础的语言研究方法。它实际上包括两方面的内容:一是对自然语料进行加工、标注,二是用已经标注好的语料进行语言研究和应用开发。

第二节 语料库语言学的发展历史

语料库语言学作为一种语言研究的方法,可以追溯到上个世纪,甚至更为久远。文献[1]对语料库语言学进行了论述,在此现在一般以乔姆斯基(N. Chomsky)转换生成语法的兴衰史为参照点,将语料库语言学的发展历史分为如下三个时期[1]。

一、早期的语料库语言学

早期语料库语言学是指20世纪50年代中期以前,即以乔姆斯基提出转换生成语法理论之前的所有基于语言材料的语言研究。在50年代,语料库在语言研究中曾被广泛使用,主要集中体现在以下几个方面。

1. 语言习得

语言习得是较早普遍用语料为研究方法的一个领域。19世纪70年代,在欧洲兴起了儿童语言习得研究的第一个高

潮。当时许多研究素材来自父母对其子女话语发展的观察日记。据悉,这些日记作为原始资料,不仅是当时 Preyer[4] 和 Stern[5] 等人提出理论假说的依据,而且时至今日仍是许多学者的研究材料之一。自本世纪 30 年代以来,语言学家和心理语言学家提出了许多关于儿童在不同年龄段的语言发展模式,这些模式大都建立在对儿童自然话语的大量观察材料的基础上。

2. 方言学

方言学从其产生以来,就与语料结下了不解之缘。在西方,方言学脱胎于 19 世纪的历史比较语言学,最初的兴趣主要是,运用直接法所获得的有关单音不同分布的事实来绘制方言地图。"方言研究者手持笔记本,后来是手提录音机,记下或录下他所遇到的一切方言材料。此种采样方式至今仍为某些业余研究者所沿用,它对于研究方言词汇的分布有一定价值"(Francis,1980)[6]。在我国,运用语料的研究方法可追溯至周秦。据南朝应劭《风俗同义序》"周、秦常以岁八月遣辖轩之使,求异代方言。"我国汉语方言学第一部著作《方言》就是这种方法的产物。据载,扬雄非常喜爱方言,他利用考廉(略等于后代的举人)和士兵集中在首都的方便,普遍地进行走访,不断积累材料,坚持编撰整理,经过 27 年的艰苦努力,终成《有辖轩使者绝代语释别国方言》。

3. 语言教学

Fries、Traver 和 Bonger (1947)是使用语料研究外语教学

法的语言学家。正如 Kennedy 所说(1992)[7],在 20 世纪的前 50 年中,语料库与外语教学有着密切的联系。外语教学中使用的词汇表,往往都是直接从语料中统计的。它对控制外语的学习过程十分重要。

4. 句法和语义

一些语言学家用语料库研究语言的描述。如:语言学家 Fries(1952)在语料库调查的基础上建立了英语的描写语法[8]。这项工作比 80 年代后期语言学家夸克(Quirk)等用语料库方法编撰的《英语语法大全》,早了 30 年。

5. 音系研究

利用自然语料开展音系研究,在西方当首推早期的结构主义语言学家,如 F. Boas 和 E. Sapir 等人。他们注重"野外工作",强调语料获取的自然性和语料分析的客观性。这些都为后来的语料库语言学所继承和发展。

20 世纪 50 年代中前期,在实证主义和行为主义思潮的影响下,总的来说在语言研究中占主导的是经验主义。这种气氛无疑促进了对语料的重视,使其成为当时的热点之一。特别是在美国,以哈里斯(Harris)等人为代表的后布龙菲尔德结构主义语言学家,视语料为语言学的唯一研究对象。在他们看来,直觉证据是第二位的,是靠不住的,应该扬弃。

二、乔姆斯基的转换生成语法时期

1957年乔姆斯基《句法理论》[9]及其以后一系列论著的发表,根本改变了语料库语言学的上述发展状况。语言学研究的主流方法也随之从经验主义(empiricism)转向理性主义(rationalism)。在这段时期中,笛卡儿的理性主义占据了主导地位,经验主义几乎无立足之地,被视为经验主义产物的各种语料库自然被完全否定。

理性主义研究方法认为,人的很大一部分语言知识是生来俱有的,是遗传决定的。而与理性主义相对峙的经验主义则认为,人的知识只是通过感官输入,并经过某些简单联想与一般化的操作而得到。人并非生来俱有一套有关语言的原则和处理方法。由于在语言学中乔姆斯基的内在语言或语言能力说被广泛接受,理性主义方法,从20世纪60年代到70年代长达20年的时间里,实际上主宰了欧美众多国家的语言学研究。

乔姆斯基及其转换生成语法学派批判早期语料库研究方法的主要论点如下:

1. 基于语料库的研究方法有误。乔姆斯基区分了语言能力(language competence)和语言使用(language performance)这两个概念。认为,语言研究的主要目标是建立一种能够反映说话人心理现实的语言认知模式,也就是语言能力模

式。因为只有语言能力才能对说话人的语言知识作出解释和描述。语言使用只是语言能力的外在证据,往往会因超语言因素的影响而发生变化,因此,后者不能确切反映语言能力。乔姆斯基还认为,语料从本质上只是外在化话语的汇集。基于语料的研究所建立的经验模式,充其量只能对语言能力作出部分解释。因而语料不应当是语言学家从事语言研究的得力工具。

2. 语料的不充分性。乔姆斯基在《句法理论》一书中首次发现英语短语结构规则具有递归性。这种递归性表明:自然语言句子的数量是无限的,是任何有限的语料所不可能穷尽的。换言之,语料永远是不完整的,不充分的。

转换生成语法学派的上述批评从根本上改变了50年代结构主义语言学的研究方向。在此后的近20年里,整个语言学界几乎唯直觉是从,唯思辨独尊,语料库方法几乎名誉扫地。但是即使在这种情况下,语料库的研究从未完全终止。许多语言学家凭着非凡的学术勇气,顶着无形的压力,继续不懈地从事语料库语言学研究,并不断取得进展。1959年,夸克着手建立"英语用法调查"语料库(Survey of English Usage)。与此同时,Francis 和 Kucera 开始了建造在现在非常著名的布朗语料库的工作,这项工作从开始到最终语料库的建成,花费了近20年的时间。此外,在1975年,Jan Svartvik 在前两项的工作基础之上,开始建造伦敦—隆德语料库(London-Lund Corpus),并且最终实现了计算机上的 SEU 语料

库。

对此,里奇(Leech,1991)认为:“作为英语口语研究的语料资源,它至今仍无以伦比”。另外,以弗朗西斯(N.Francis)和 Kucera 为首的一批语言学家和计算机专家汇集在美国的布朗大学合力攻关,于 1961 年建立了当今最早的机读语料库—布朗语料库(Brown Corpus)。布朗语料库以共时原则采集不同主题的英语样本,总规模为一百万词次,目的是研究美国英语。这两个语料库可以说是现代语料库的开山鼻祖,并共同为 80 年代语料库语言学的复苏奠定了基础。

三、语料库语言学的复苏时期

80 年代以来,语料库语言学在相对沉寂了近 20 年后开始复苏,并得到迅速发展。主要表现在下面几个方面。

1. 第二代语料库相继建成

80 年代以来,以柯林斯—伯明翰英语语料库(COBUILD)为代表的一大批语料库相继建成。这些机读语料库尽管规模、设计和研究目的各异,但大多采用了较新的 KDEM(Korowai Data Entry Machine)光电字符识别技术,从而使语料的编码和编辑得以从繁重的人工键盘输入中解脱出来。这个时期建设的语料库,不仅规模上有数以十倍的增长,而且建设速度大大加快了,故称第二代语料库。根据美国加州大学伯可莱分校的语言学家爱德华兹(Edwards)在 1993 年

的不完全统计,80年代以来建成并投入使用的各类语料库达50余个,按语种分布如下:

英语	24	法语	4	意大利语	2	丹麦语	2
德语	7	西班牙语	2	芬兰语	2	瑞典语	2

此外,葡萄牙语、南斯拉夫语等语种也相继建立了自己的语料库。在这些语料库中,规模大且特色鲜明的有:

(1)兰卡斯特－奥斯陆－卑尔根语料库(Lancaster-Oslo-Bergen Corpus, 简称LOB语料库)。70年代,在英国兰卡斯特大学著名语言学家里奇的领导下,以研究英国英语为目的,采用与布朗语料库相同的语料分布和采样原则来设计,1983年建成。语料库由五百个样本组成,每个样本约两千词次,总规模也是一百万词次。由于这些特点,人们通常把LOB语料库和布朗语料库视为姐妹库,以进行英国英语和美国英语的对比研究。

(2)法语语料库(Tremor de la Language Françoise, 简称TLF语料库)。该语料库是法国国家科学研究中心与美国芝加哥大学的合作项目。语料的跨度从7世纪至20世纪,包括书面法语各种文体的两千个样本,总规模达1.5亿词次。有关数据已制成光盘,并可通过UNIX操作系统查阅。

(3)赫尔辛基历史英语语料库(The Helsinki Corpus of Historical English)。这个语料库是以罗西尼(Roseanne)等为

首的一批语言学家在赫尔辛基大学建立的。语料包括从公元850年到1720年这一时期的各类英语语篇，并以每百年分段，总库容量达1600万词次。作为第一个历时的英语语料库，它对于从社会语言学、方言学及语用学角度来研究英语的变迁，具有重要价值。

(4)国际英语语料库(The International Corpus of English，简称ICE)。该库于1988年由伦敦大学英语系承建，旨在为世界范围内英语民族变体的对比研究提供数据。语料分别取自所有英语国家，并采用统一的分类和编码系统。每个国家的语料限定为一百万词次，口语和书面语各占一半。语料采样时间限定在1990年至1993年之内。采样对象为18岁以上接受英语教育成长起来的成年人。

在这个时期内，我国语料库的建设队伍也在逐渐壮大，利用语料库进行语言调查和研究的学者也在不断扩大。如利用大规模语料库对汉字和词语的使用频率进行统计调查。见到的主要成果有《现代汉语常用字表》和《现代汉语频率词典》等。本书以后的章节中还将对各种语料库的建设和使用情况进行详细描述，在此不再赘述。

2．基于语料库的研究项目增多

大批语料库的建成极大地推动了基于语料库的研究项目。下表的统计数字充分说明了这一点。

1959—1991 年语料库研究项目统计表(Johansson, 1991)

起止年限	研究项目数
1959—1965	10
1966—1970	20
1971—1975	30
1976—1980	80
1981—1985	160
1986—1991	320

事实表明,计算机语料库是开展大范围语言研究的极好资源,因为它所提供的语料较之先前的材料更具有真实性,其层次结构更加明晰,因而更有助于对语言的不同层面进行描写,对不同语体开展对比研究,进而实现语言的计量研究。

这期间许多研究项目取得了重要成果,有的深化了原有的研究,有的则拓宽了原有的研究领域。如 Halliday (1991)[10]和 Svartvik(1992)[11]等人的概率语法研究;Dottie (1991)的英国英语和美国英语话语风格研究,以及 Sinclair (1985)等人关于英语搭配的计量研究等。

对 80 年代以来英语语料库语言学复苏的原因,近年来多有评说。概而言之,主要有如下两条:

(1)计算机科学的飞速发展与计算技术的迅速普及和应用,为语料库语言学的复苏提供了强大的物质基础。80 年代以来,语料库的发展进入了一个良性循环:计算机的运行速度和存储能力的成倍增长加快了语料库的建设,提高了语料库

的处理能力和处理层次。同时,大量经过标注的语料又反过来促进了语料库的研究和利用。在此期间,诞生了更为先进的研究方法和语言模型,许多先前需要人工处理的标注和统计工作,现在可以通过计算机软件自动或半自动地完成。在这一循环中,计算机显然是一个举足轻重的环节。

(2)转换生成语言学派对语料库语言学的批判,经过20年的实践已证明,有的是错误的,如指责计算机是伪技术;有的是片面的,如对语料库的全盘否定;有的则是正确的,如乔氏关于自然语言句子的数量无限性的观点。对于乔氏倡导的理性主义方法,人们经过跟从、应用和反思之后,也逐渐发现其不足,如不可验证性等。因此,80年代以来语料库语言学的复兴,在很大程度上反映了语言学界的一种普遍心态,即想要恢复语言研究中人工数据和自然数据的平衡。既然语料库研究方法和基于内省的唯理主义方法各有长短,为什么不能让二者共存或结合,以充分发挥其互补的优势呢?为了达到这种有益的平衡,许多语言学家发出呼吁,如:

"语料研究在语言的理论探索中具有中心位置,对语料的开发途径很多,……并非只有一种"。(Halliday 1991)[10]

"从科学方法的角度,语料库方法是一种更为强有力的研究方法,因为其结果是可以验证的"。(Leech 1993)[12]

即使像C. Fillmore这样曾经对语料库语言学有诸多批评的语言学家,也对语料库语言学作出了公允的表述:

"我不认为有这样的语料库:它能包括我要探究的英语词