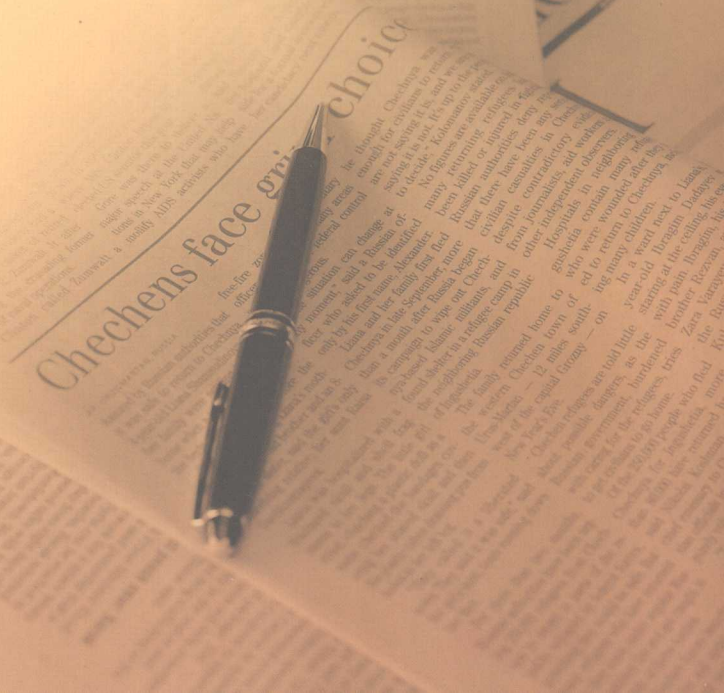
郑志恒 著

# 报刊英语语料库概论

## An Introduction to the Newspaper English Corpus

南京大学出版社



# An Introduction to the Newspaper English Corpus

郑志恒 著

# 报刊英语语料库概论

南京大学出版社

**图书在版编目(CIP)数据**

报刊英语语料库概论/郑志恒著. —南京:南京大学出版社,2009.6
ISBN 978-7-305-05800-4

Ⅰ.报… Ⅱ.郑… Ⅲ.报刊—英语—研究 Ⅳ.H31

中国版本图书馆 CIP 数据核字(2009)第 038587 号

出 版 者 南京大学出版社
社 址 南京市汉口路 22 号 邮 编 210093
网 址 http://press.nju.edu.cn
出 版 人 左 健

书 名 **报刊英语语料库概论**
著 者 郑志恒
责任编辑 裴维维 编辑热线 025-83686029

照 排 南京南琳图文制作有限公司
印 刷 南京紫藤制版印务中心
开 本 787×1092 1/16 印张 11.5 字数 243 千
版 次 2009 年 6 月第 1 版 2009 年 6 月第 1 次印刷
ISBN 978-7-305-05800-4
定 价 28.00 元

发行热线 025-83594756
电子邮箱 sales@press.nju.edu.cn(销售部)
njupress@gmail.com(编辑部)

---

# 序

我从40余年高校英语教学实践中深切体会到，英语报刊是十分理想的教学资料。报刊具有贴近时代、贴近大众、贴近现实、贴近生活的特点。与其他教学资料相比，英语报刊具有内容新颖、语言现代、资料丰富、词语实用的显著优势。伴随着我国对外交流的拓宽，我国高校英语教学改革的深化，外报外刊课程的价值得到更加广泛的认可，越来越多的高校为英语专业和非英语专业学生开设了这门课程。

报刊英语的深入研究可以大大增进对报刊语言特色的了解，从而促进报刊英语教学和研究水平的提高。目前国外已经出版了多部关于报刊语言研究的专著。虽然国内的报刊语言和报刊教学研究领域内硕果累累，但基于语料库的报刊语言和教学的相关研究还在起步阶段。语料库语言学采用数据驱动的实证主义研究方法，采用定性和定量相结合的分析手段，在大量真实语境中的文本信息的基础上，对语言现象进行描述、统计、检验，揭示现代英语的语言使用规律，给报刊语言研究带来了新的理念。

郑志恒博士的主要研究方向是传媒英语，曾在系列《美英报刊阅读教程》中担任多部教程的编著者，对报刊英语语言具有很好的驾驭能力，同时他在语料库语言学和统计学等方面做过长期研究，积累了丰富的理论知识和实践经验。

《报刊英语语料库概论》是他运用多种学科理论，长期以来艰辛努力所研究出的成果。这一成果填补了我国报刊英语语料库领域的空白，具有较好的学术价值。

本书既是多种理论在语料库建设中实际应用经验的汇编，也是报刊语言研究基于语料库的探索成果。我认为他的语料库研发理论和经验在语料库设计和建设方面具有较好的实际参照性，书中对报刊英语语言现象的展现和各领域理论的融合应用体现了较强的科学性。我确信《报刊英语语料库概论》一书对英语教育、英语语言学、媒介研究等专业的教师和学生具有良好的指导作用。

**端木义万**

**2009年6月**

---

端木义万教授系本书作者的博士生导师，南京国际关系学院英语资深教授，全国优秀教师，享受政府特殊津贴专家，全国报刊英语教学和研究的创始人，从事外报外刊教学与研究42年。其撰写的系列著作多次重印，经久不衰。

# 前　言

语料库语言学是基于大量机读文本数据，采用定性和定量分析相结合的方法，对语言的形态、意义和功能进行系统分析的新兴学科。在过去的二十余年，语料库越来越受到语言学界的重视，成为语言研究的一个重要方法论，它所带来的影响已经波及到外语教学和研究的各个学科领域。

在外语教学界，作为新闻英语主干的报刊英语已经成为一个热门教学课程并孕育着十分广阔的发展前景。但这种最能体现时代语言特色的报刊语言，迄今仍然缺少科学的语料库研究基础。笔者三年来对国内外语料库研制与应用状况做了全面的调研，在此基础上依据语料库语言学理论、统计学理论和新闻学理论，经过长期艰苦工作，建成首个专为研究报刊英语量身定制的百万词级的报刊英语语料库(Newspaper English Corpus，简称 NEC)。

NEC 语料库的特点在于：(1) 严格依据统计学理论设计语料抽样原则，语料库具有很好的代表性和平衡性；(2) 按照报刊英语的语体特点全库分为四个子库，即英国报刊纯新闻报道文本子库、英国报刊意见性报道文本子库、美国报刊纯新闻报道文本子库、美国报刊意见性报道文本子库；(3) 采用 SGML 置标语言，可以为 WordSmith Tools 等语料库分析和检索软件识别处理；(4) 语料库的结构设计和标识设置利于开展关于报刊英语语体、语域、标题、导语、主体、文本结构、作者风格、语言变体等五十多项参数的单项统计研究；(5) 全库做了词性赋码处理，蕴含更丰富的语言学信息，便于掌握用词、句法等报刊语言规律。

本书记载了笔者多年来在语料库语言学和报刊英语研究这两个领域的主要研究心得，介绍了基于新闻英语的最新语料库发展和应用成果，并且凭借已建成的 NEC 语料库，在实践经验的基础上深入浅出地介绍了什么是语料库、什么是报刊英语语料库、语料库的种类、语料库的数理统计方法、语料库建设的理论和方法、如何自行设计和建立报刊英语语料库、报刊英语语料库对英语教学和报刊语言研究有何用途等内容。本书不仅是探讨语料库语言学应用于报刊英语教学研究的理论和实践相结合的一本书，而且可以为建设其他类型的语料库提供一定的理论依据和方法参照。

本书的特点在于：(1) 解剖大批国外新闻英语语料库，介绍其结构及实际操作过程，为语料库建设提供典型参考；(2) 聚焦语料库设计、建设、加工、标识等实

际操作步骤和具体方法，帮助读者自行创建语料库；(3) 读者着手语料库研究的起点就是如何借助语料库软件进行文本分析和检索，本书专门给予了详细的介绍和分析，帮助读者尽快入门开展语料库研究；(4) 本书通过NEC语料库的真实数据和语料库的数理统计方法揭示语料库研究的意义、模式和NEC在报刊英语教学研究方面的价值；(5) 书中多处提供网址和语料库相关资源信息，方便读者寻找语料库资源并下载相关语料库工具，鼓励读者自己动手研究。

语料库语言学在英语教学研究领域内方兴未艾，报刊英语语料库作为语料库一个全新的分支，尚有许多亟待解决的问题，本书所论述的创建报刊英语语料库的理论和方法仍有许多不成熟之处。由于笔者功力不深、锤炼不足，书中难免出现疏漏和不尽准确之处。笔者恳请广大高校英语教师和读者提供宝贵意见，并期待各位专家、同行批评指正。

本书在撰写过程中得到了许多人士的指导和鼓励。特别感谢笔者的博士生导师端木义万教授，恩师渊博的知识、严谨的作风、宽广的胸怀给了笔者深刻的影响。教授在工作学习上给予耐心的指导，在思想精神上给予极大的鼓励，带领笔者走出低谷，指明努力的方向。笔者所获得的每一丝成绩都凝聚着恩师的心血，谨以此书致谢端木义万教授所赐予的无私帮助和栽培。

另外，完成《英汉双向口译虚拟教学系统》和《基于平行语料库的〈汉英词典〉的研编》两项国家社会科学基金项目的李德俊教授在语料库建设方面给予了悉心的指导，笔者表示真诚的感谢。

本书从选题到完稿过程中，北京对外经济贸易大学王立非教授，广州中山大学黄国文教授，南京师范大学辛斌教授，南京大学陈新仁教授，南京解放军国际关系学院李战子教授、张辉教授、陈开顺教授都提出了很多宝贵意见，在此谨表诚挚的谢意。

南京大学出版社的杨金荣编审对本书的出版给予了宝贵的支持，笔者表示诚挚的感谢。

郑志恒

2009年6月

# 目　录

# 第一章　语料库概论

世界上以英语为母语的人口仅有3.8亿，而全球有三分之一的人口在使用这种语言，也就是说，大约有20亿人在工作、生活或者学习中离不开英语。人类现代文明的载体，从书籍文献到报纸杂志，从收音机电波到卫星电视节目，英语无疑占据了主导地位。因特网上有80%的信息以英语这种语言存在，但上网用户中说英语的人数仅有56%。一切事实数据都说明这种语言的使用需求量在急剧膨胀。本书在此姑且不去谈论英语的“语言帝国主义”，而是揭示研究这种语言的一个新思路——语料库语言学方法。

# 第一节 语料库的定义和类型

## 一、语料库的定义

语言学有长期的描写性研究(descriptive,相对于规约性研究 prescriptive)传统,语言学家的论述都建筑于语言文本的基础之上,但是存在四点缺陷。首先,研究所用的文本几乎都用于证实某项语法规则或某个语言学原理,而这些文本数据在实质上仅仅是真实语言使用范畴的一部分。语言学家们通过内省法(introspection)得到的语言结构变异案例不足以展现语言的实际使用全貌,而且容易出现任意性偏态。其二,研究的文本趋于陈旧,导致研究结果缺乏时新性。其三,传统研究缺乏系统性,一般关注的对象皆为语言中出现频率高、使用较普遍的结构和类型,也夹杂少数使用频率很低的语言现象,这势必导致容易忽略一些语言特点。这些描写性研究缺少频数和分布的信息,仅限于大致印象性陈述。其四,虽然英语语言的传统研究在主体上是描写性的,但传统语法书目经常夹杂着作者的规约性痕迹,告诉读者什么是合理的语言使用,而不是真实状况下的语言使用,这容易给学习者造成一定的理解偏移。

例如 Quirk, *et al*. 关于英语中省略情况的论述:

However, within the textually recoverable category there is an even stronger criterion, which distinguishes [1] from [2]:

She might *sing tonight*, but I don't think she will (sing tonight). [1]

She rarely *sings*, so I don't think she will (sing) tonight. [2]

The ellipted expression in [1] is an exact copy of the antecedent (sing tonight), while in [2] the ellipted verb is morphologically different from its antecedent. ... But it is important to recognize that [1] and [2] illustrate what, for most grammatical purposes, is the same kind of ellipsis: it remains true, in particular, that the ellipsis of sing is "precisely recoverable" in the sense of 12.33.

(Quirk, *et al.* 1985: 887)

现代计算机技术的发展特别是计算语言学(Computational Linguistics)的相关研究成果使英语语言研究的突破成为可能。计算语言学是"通过建立形式化的数学模型来分析、处理自然语言,并在计算机上用程序来实现分析和处理的过程,从而达到以机器来模拟人的全部或者部分语言能力的目的。"(俞士汶 2004:2)计算语言学其中的一个分支通过计算机设备对语言进行全面、系统地描述和研究,成为描写性语言研究传统的延伸,这就是自 20 世纪 60 年代以后发展起来的语料库语言学(Corpus

Linguistics)。

语料库,顾名思义,就是存储语言材料的仓库。20世纪早期的语料库都是手工收集的记录着作品中段落的文本卡片,用来对某种语言结构进行阐释和示范,同现阶段的语料库有着本质的区别。20世纪60年代早期布朗语料库(Brown Corpus)的出现,成为语料库发展史上的一个里程碑,标志着语料库语言学进入了一个新阶段。现代语料库包含了大规模的书面语和抄录后的口语文本,以机器可读的方式(machine-readable form)存储。所以,更确切地说,语料库的定义是:语料库是以机器可读的方式存储,用来代表特定语言或语言变体的自然发生的语言材料的集合。由此,对语料库的概念至少包含以下三点:

(1) 语料库的收集、存储、处理和使用都以计算机为主要工具(computer-mediated);

(2) 语料库必须要有代表性(representativeness),前提就是具有一定的规模;

(3) 语料库同传统语言研究的内省法不同(注重 competence output),其样本全部是真实使用的语言(研究 performance record)。

语料库的出现为语言学研究"不仅提供了一种全新的研究方法,而且开发了一项新的研究事业,拓展出一个新的哲学思路。计算机作为技术工具,使这种新兴语言学领域的诞生成为可能。所以技术手段在这个领域内比其他辅助研究更加重要:我把它看作获得新知识的基本途径,看作是通往语言研究新思路所需要的'芝麻开门'的秘诀。"(Leech 1992: 106)

## 二、语料库的类型

迄今,计算机科学和计算语言学的发展为语料库建设提供了原始驱动力,诸多大中型语料库纷纷建成,若干超大规模语料库正在建设当中。从宏观上看,语料库仅有两种类型。一种是相比之下规模不大的语料库,但一系列现代计算机的自然语言处理技术(Natural Language Processing, 简称 NLP),如词类赋码(tagging)和句法赋码(parsing)等处理步骤,使库蕴涵了相当丰富的语言信息和细节,例如国际英语语料库(International Corpus of English, 简称 ICE)和作为英国17世纪和18世纪语言镜像的 Lampeter Corpus。另外一类是富含各种语言变体,包括口语体的大型和超大型语料库。最典型的莫过于总容量达1亿词次、口语体容量达1 000万词次的英国国家语料库(British National Corpus, 简称 BNC)和规模在不断扩大中的英语文库(Bank of English, 简称 BOE)。但是这种大规模语料库的语言标注信息并不是特别丰富,例如 BNC 虽然全库都已经做了词类赋码处理,但仅有一小部分经过句法分析处理。

从微观上看,语料库的类型颇为多样。

以语料库建设的采样过程来分,可以把语料库分为四种类型:(1) 异质语料库(heterogeneous corpus),语料采样不受原则约束,广泛收集,原样存储,不经加工处理;(2) 同质语料库(homogeneous corpus),语料采样只局限于同一类型;(3) 系统语料库(systematic corpus),在语料采样时考虑到语料库的平衡性和系统性,按照一定

的抽样原则，并遵照语言样本的规律进行收集，这种语料库能够代表某一范围内的语言事实；(4) 专用语料库(specialized corpus)，语料采样时仅收集用于某一特定用途的语料。

如果按照语料的处理方式来分，有原始语料库(raw corpus)和赋码语料库(annotated corpus)两种。前者未经任何标注和赋码，又称生语料库，例如挪威 ICAME Corpus Collection 光盘中注明"untagged"的 BROWN 和 LOB 都是生语料库。后者经过了词性、语法、语篇等的标记，又称标注语料库或熟语料库，BNC 和标注过的 BROWN，LOB 都为熟语料库。

按照语料的时效性分为共时语料库(synchronic corpus)和历时语料库(diachronic corpus)两种。LOB 语料库收集了 1961 年出版的英国书面英语 500 篇文本，每篇 2 000 词，总计 100 万词次。该语料库收录某特定时间内的语言样本，所以被称为共时语料库。Lampeter Corpus 收录了自 1640 年至 1740 年 100 年间的 120 篇文本，总计 1 172 102 词次，每 10 年一个单元，每单元含有由宗教、政治、经济、科学、法律、杂烩六个语域(domain)构成的两篇文本。该语料库收录了一段时间内的语言样本，可以对语言的历时发展进行研究，被称为历时语料库(见图 1.1)。1991 年建成的赫尔辛基历时语料库(Helsinki Diachronic Corpus)跨度从 1500 年至 1710 年，总计 551 000 词次，同样也为历时语料库。

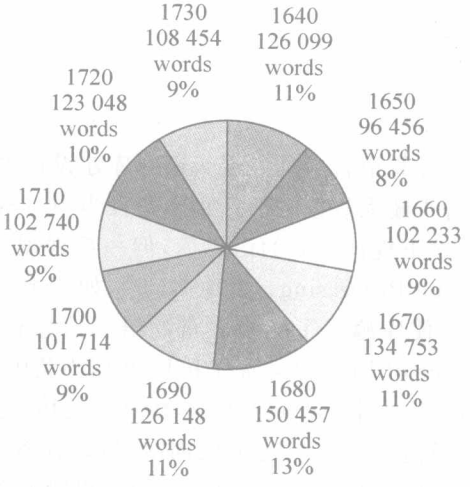



**图 1.1 Lampeter Corpus 的内部时间结构**

按照语料的语种来分，可以分为单语语料库(monolingual corpus)、双语语料库(bilingual corpus)和多语语料库(multilingual corpus)。单语语料库只收集一种语言的语料，如 BROWN，LOB 和 BNC 等。双语语料库和多语语料库指由两种或者两种以上语言的语料构成的语料库，按照语料的组织形式，这两种语料库可以往下继续分为平行语料库(parallel corpus)和类比语料库(comparable corpus)。平行语料库也叫对应语料库或对齐语料库，其实是双语语料库，其结构由原文和译文两种语言对应组成，多用于机器翻译、双语词典的编撰等领域。类比语料库又称对比语料库或可

比语料库，为双语或者多语语料库，由不同语言的文本或者同一语言的多种变体组成一一对应的关系，多用于语言对比研究。两者的区别在于平行语料库的原文和译文间存在对译关系，而类比语料库的两种或者多种语言之间并不存在对译关系。例如英语—挪威语平行语料库(ENPC)中的语料就是对译关系，而国际英语语料库(ICE)则是典型的类比语料库。

另外根据语料库的应用可以分为通用语料库(general corpus)和专用语料库(specialized corpus)。通用语料库包含多种语域、语体的语言样本，而专用语料库仅关注特定的语体(register)。如 BNC 包括多种语域，有口语体和书面语体，为通用语料库。CANCODE (Cambridge and Nottingham Corpus of Discourse in English)只收录英国英语中的非正式语体的样本，MICASE (Michigan Corpus of Academic Spoken English)只收录美国学术圈内交流的口语体样本，这些都是专用语料库。

## 第二节　语料库的发展历史

### 一、早期阶段(20 世纪 60 年代之前)

语料库的历史发展阶段一般从 20 世纪 60 年代初开始算起，确切地说，1961 年至 1964 年间，世界上第一个通过系统采样原则建设完成的布朗语料库(Brown Corpus)标志着语料库语言学这门新兴学科的诞生。在此之前，语料库已经在语言研究上有着广泛的应用，但与其说是语言学中的一个学科分支，不如说是一种研究手段，语料的收集和处理全部手工操作，记载于卡片上，效率低下。比如 1928 年编撰完成的《牛津英语词典》(*Oxford English Dictionary*，简称 OED)中 400 多万条引证语料来源于 1 100 多万张卡片语料库。

1957 年，语言学大师乔姆斯基(N. Chomsky)发表了《句法结构》(*Syntactic Structures*)一书，书中对语料库这种研究方法进行了批判。乔姆斯基认为语言学家研究语言只能根据本身的母语知识和直觉，采用内省式的理性主义研究方法。他指出，语言研究的最终目的是研究人类的语言能力(competence)，而不是语言行为(performance)。本族语的人其语言能力可以产生和理解数之不尽的语言现象，而语言行为是受外界影响的，不能反映语言的实际能力。因此，研究语言只能用内省法，通过对本族语的人的语言能力进行考察得出相应语言理论，而不是通过语料库方法去搜集无法反映语言能力的实际使用中的语料，语料库充其量只不过是语言能力的外在表现。乔姆斯基又指出，语料库永远无法解决无尽的语言事实和有限的语料样本之间的矛盾，而语料库研究方法就是用有限的语料代表整体语言事实。

乔姆斯基的批评对语料库的发展无疑是致命的，语料库方法一度被冷落，理性主义盛行于学术界。今天，即使现代计算机技术可以帮助研究者搜集超大规模的语料，

建成远远超出乔姆斯基当时规模的语料库，他的批评观点并没有过时，对语料库建设仍然具有很大的指导意义。我们并不需要根据他的观点，摒弃语料库方法，坐在椅子上苦思冥想用语言直觉去考察说话人的语言能力(armchair linguist)，我们需要牢记的是致力于建设尽最大可能代表语言整体的语料库，使其能够为语言研究描绘出一幅完整图案。

虽然20世纪50年代至80年代一直处于乔姆斯基转换生成语法的影响之下，语料库已经被排斥出主流语言学(mainstream linguistics)，但对语料库的研究一直没有停止。首先于1959年，R. Quirk提出建立"英语用法调查"项目(Survey of English Usage，简称SEU)。开始SEU一直用"6×4"规格的卡片存档，1975年之后其中的口语部分开始转成机读语料(见下文London-Lund Corpus)。SEU所采集的样本包括了大量"教养语言"的书面语和口语变体。"教养语言"指的是"以英国英语为母语的受过教育的说话者的所有作品，从他们所写的情书、演讲稿到公共讲坛的发言、社交宴会上轻松的私人谈话等一系列语言活动。"(Quirk & Svartvik 1979：204)著名的《现代英语语法》(*A Grammar of Contemporary English*，1972)和《英语语法大全》(*A Comprehensive Grammar of the English Language*，1985)就建立在这个语料库的基础之上，对英语语言学界产生了深远影响。

## 二、第一阶段(20世纪60年代至70年代)

接下来1964年布朗语料库(Brown Corpus)的建成标志着语料库进入了现代语料库发展的第一阶段，即第一代电子语料库阶段，时间跨度从20世纪60年代至70年代。布朗语料库在语料库发展史上具有重大意义。其一，把语料库研究带入了电子化时代，利用计算机强大的信息处理能力，对语料进行存储、检索，极大地丰富了语料库研究的内涵，为语料库向电子化方向发展奠定了基础。其二，在20世纪60年代初语言学界理性主义(rationalism)革命的笼罩下，布朗语料库为经验主义(empiricism)研究带来了曙光。其三，布朗语料库引入TAGGIT词性赋码系统，为语料库的赋码处理开创了先河。在这阶段，除布朗语料库外，另外还建成两个颇有影响力的语料库，兰开斯特—奥斯陆/卑尔根语料库(The Lancaster-Oslo/Bergen Corpus，简称LOB)[1]和伦敦—隆德英语口语语料库(London-Lund Corpus of Spoken English，简称LLC)。

1970年，英国兰开斯特(Lancaster)大学的著名语言学家G. Leech倡议建立兰开斯特—奥斯陆/卑尔根语料库，就是LOB语料库，该库由挪威奥斯陆(Oslo)大学的S. Johansson主持，于1978年完成。LOB语料库与Brown语料库相对应，收集了1961年出版的英国书面英语，同样包含500篇文本，每篇约2 000词，共计100万词次，主要代表当代英国英语。LOB语料库首次采用概率统计思维，UCREL小组(Unit for Computer Research on the English Language)用概率的方法根据LOB语料库所提供的133×133个标注过渡矩阵而编制的CLAWS程序，标注精确率达96%—97%。CLAWS不断更新，目前它的第四版吸收了以规则为基础的标注程序

的优点，用来处理1亿词的BNC，其错误率(error-rate)是1.5%，歧义率(ambiguity-rate)是3.3%。

1975年，R. Quirk的学生，瑞典隆德大学的J. Svartvik开始把SEU语料库中2 000个小时的谈话和广播等英语口语素材的书面材料转化成机读形式，总计87篇口语文本，43.5万词次。他又增加了13篇文本，对口语语料进行了细致的音韵标注，最终建成伦敦—隆德英语口语语料库，简称伦敦—隆德语料库(London-Lund Corpus，简称LLC)，包含100篇文本，每篇5 000词，共计50万词次，成为英语口语语料库建设的一个参照标准。

语料库发展第一阶段内出现的这三个典型语料库现都收录于1999年挪威卑尔根(Bergen)大学ICAME国际学术组织(International Computer Archive of Modern English)出版的ICAME Collection of English Language Corpora之中，该光盘收录总计1 700万词次的21个语料库。

## 三、第二阶段(20世纪80年代至90年代)

20世纪80年代至90年代为语料库发展的第二阶段——第二代电子语料库阶段。自20世纪80年代开始，新兴计算机技术使语料库的规模不断扩大，由此语料库的代表性得到显著提高。随着语料库的深度加工技术日趋成熟完善，新开发的自动赋码(tagging)和句法分析(parsing)计算机软件揭开了数据驱动(data-driven)的实证主义研究方法，语料库的建设进入了高速发展期。20世纪80年代之后，第二代千万级以上甚至上亿级别的语料库进入历史发展舞台。语料库语言学迅速成为语言学研究的一门新兴分支学科，“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。”(杨惠中 2002：4)

1980年，英国柯林斯出版社同伯明翰大学开始合作开发COBUILD语料库，全称柯林斯伯明翰大学国际语言资料库(Collins Birmingham University International Language Database)，亦称伯明翰语料库，由伯明翰大学J. Sinclair负责。其中口语语料占25%，英国英语语料占70%，美国英语语料20%，其他英语变体10%。书面语部分来源于通俗作品，75%的内容选自男性作家的作品，25%的内容选自女性作家的作品。COBUILD语料库在1991年被柯林斯和伯明翰扩展成为英语文库(The Bank of English，简称BOE)[2]，目前BOE全库已达5.24亿词次。英语文库现已成为监控语料库(monitor corpus)，其规模仍在不断扩大之中，其语料主要来源于英美两国，是迄今规模最为宏大的语料库。COBUILD语料库是第一部为词典编撰服务的语料库，在该库的基础之上诞生了著名的COBUILD系列词典、语法书等英语学习参考用书。COBUILD语料库的最大成功之处在于语料库的开发同商业用途紧密结合，超大规模的语料为其在全球诸多词典、语法参考书的巨大销量奠定了扎实的基础。

20世纪80年代后期朗文语料库网络(The Longman Corpus Network)开始创

建，由英国三大语料库组合而成，即朗文/兰开斯特英语语言语料库(Longman/Lancaster English Language Corpus，简称 LLELC)、朗文口语语料库(Longman Spoken Corpus，简称 LSC)和朗文学习者英语语料库(Longman Corpus of Learners' English，简称 LCLE)。该语料库的主要目的是编撰英语学习词典，库容量达 5 000 万词级。

目前全球流传最广、应用最多的超大规模语料库就是英国国家语料库(British National Corpus，简称 BNC)[3]。BNC 由英国牛津大学出版社、朗文出版公司、钱伯斯—哈洛普出版公司、牛津大学计算机服务中心、兰卡斯特大学英语计算机研究中心以及大英图书馆等联合开发，由 G. Leech 等学者主持建立，自 1991 年开始到 1994 年建成。BNC World Edition 的两张 CDROM 光盘压缩了总共 3.73G 容量的语料库文件，总规模达 100 467 090 词次[4]，其中书面语 9 000 万，口语 1 000 万，是迄今已建成的最具代表性的现代英语语料库之一(见表 1.1)。

**表 1.1 英国国家语料库的组织结构(BNC World Edition)**

| Text type | Texts | Kbytes | W-units | S-units | Percent |
|---|---|---|---|---|---|
| Spoken demographic | 153 | 4 206 058 | 4.30 | 610 563 | 10.08 |
| Spoken context-governed | 757 | 6 135 671 | 6.28 | 428 558 | 7.07 |
| All Spoken | 910 | 10 341 729 | 10.58 | 1 039 121 | 17.78 |
| Written books and periodicals | 2 688 | 78 580 018 | 80.49 | 4 403 803 | 72.75 |
| Written-to-be-spoken | 35 | 1 324 480 | 1.35 | 120 153 | 1.98 |
| Written miscellaneous | 421 | 7 373 707 | 7.55 | 490 016 | 8.09 |
| All Written | 3 144 | 87 278 205 | 89.39 | 5 013 972 | 82.82 |

- Texts: number of distinct samples not exceeding 45 000 words.
- W-units: number of ⟨w⟩ elements identified by the CLAWS system (more or less equivalent to words).
- S-units: number of ⟨s⟩ elements identified by the CLAWS system (more or less equivalent to sentences).

BNC 建设项目由五个项目小组各自负责一个环节分工完成，五个环节相互联系，构成一个完整的语料库建设体系，成为设计、建设语料库的标准规划参照。

(1) 授权(permissions)

语料库建设者事先设计一封标准请求获取语料使用许可的信件，逐个联系语料的版权所有者，获取语料的授权许可。

(2) 采样(sampling)

语料库建设者制定语料采样标准，纳入符合标准的文本类型，并确定各文本类型在语料库中所占比例。

(3) 赋码(annotation)

使用 CLAWS4 赋码软件对语料库文本根据上下文语境进行语言学赋码。

(4) 标识(markup 或 encoding)

作为标准通用置标语言 SGML 的一个子集,XML 集 SGML 和 HTML 的优势于一身,目前已经成为语料库建设的主流标注体系,但当时建成的 BNC 还是采用了 SGML 标记体系,其标注工作量少于 XML。[5]

(5) 检索(retrieval)

研发识别 SGML 标注体系的语料库检索软件 SARA(SGML Aware Retrieval Application),使标注体系具备实际应用价值。

1999 年 12 月 BNC World Edition[6] 得以正式出版并在全球范围内发行,产生了远远超出建设者预期的广泛影响,成为计算语言学和应用语言学研究的可靠语料来源。

1988 年,英国伦敦大学的 S. Greenbaum 倡议建立一个对世界范围内的地区英语变体的书面语和口语进行比较分析和研究的语料库——国际英语语料库(International Corpus of English,简称 ICE)。截至 1991 年,全球 20 个国家和地区参与此项计划。该库的语料抽样依旧遵照了 BROWN 和 LOB 的经典抽样方式,即每个参与的国家或地区建立各自的子库,每个子库由 500 篇 2 000 词左右的书面语和口语文本组成,达到 100 万词级。ICE 是迄今参与建设的国家和地区数目最多、收录英语地区变体数目最多的一个语料库(见表 1.2)。

**表 1.2 国际英语语料库组织结构**

<table>
<tr><td colspan="2">Spoken(300)</td></tr>
<tr><td>DIALOGUE (180)</td><td>MONOLOGUE (120)</td></tr>
<tr><td>Private (100)</td><td>Unscripted (70)</td></tr>
<tr><td>direct conversations (90)<br>distanced conversations (10)</td><td rowspan="2">spontaneous commentaries (20)<br>unscripted speeches (30)<br>demonstrations (10)<br>legal presentations (10)</td></tr>
<tr><td>Public (80)</td></tr>
<tr><td rowspan="2">class lessons (20)<br>broadcast discussions (20)<br>broadcast interviews (10)<br>parliamentary debates (10)<br>legal cross-examinations (10)<br>business transactions (10)</td><td>Scripted (50)</td></tr>
<tr><td>broadcast news (20)<br>broadcast talks (20)<br>speeches (not broadcast) (10)</td></tr>
<tr><td colspan="2">Written (200)</td></tr>
<tr><td>NON-PRINTED (50)</td><td>PRINTED (150)</td></tr>
</table>