河南省软科学研究计划项目（142400411235）

语料库语言学与机器翻译导论

郑茗元　常霜林　著

中国水利水电出版社
www.waterpub.com.cn

河南省软科学研究计划项目（142400411235）

语料库语言学与机器翻译导论

郑茗元　常霜林　著

中国水利水电出版社
www.waterpub.com.cn

内 容 提 要

计算机信息技术的迅猛发展为语言的研究和应用提供了广阔的平台，而同时又拓展了计算机信息技术研究的疆域。而语料库语言学的问世促进了文理学科之间的渗透与相通，也昭示了机器翻译的美好前景。鉴于此，本书以系统介绍为主，以具体实用为本，探讨了人类是如何利用语料库的计算机技术和程序模型来处理、分析、实现自然语言的信息工程，旨在使读者能之、好之、乐之。

本书适合外国语言学及应用语言学专业的研究生、口笔译翻译行业的专职人士、机器翻译研究机构的初级研究者们阅读使用。

图书在版编目（CIP）数据

语料库语言学与机器翻译导论 / 郑茗元，常霜林著
. -- 北京 : 中国水利水电出版社，2015.12
ISBN 978-7-5170-3822-1

Ⅰ. ①语… Ⅱ. ①郑… ②常… Ⅲ. ①语料库－语言学－研究②机器翻译－研究 Ⅳ. ①H0

中国版本图书馆CIP数据核字(2015)第270641号

书　　名	**语料库语言学与机器翻译导论**
作　　者	郑茗元　常霜林　著
出版发行	中国水利水电出版社 （北京市海淀区玉渊潭南路1号D座　100038） 网址：www.waterpub.com.cn E-mail：sales@waterpub.com.cn 电话：（010）68367658（发行部）
经　　售	北京科水图书销售中心（零售） 电话：（010）88383994、63202643、68545874 全国各地新华书店和相关出版物销售网点
排　　版	中国水利水电出版社微机排版中心
印　　刷	北京九州迅驰传媒文化有限公司
规　　格	184mm×260mm　16开本　10.25印张　243千字
版　　次	2015年12月第1版　2015年12月第1次印刷
定　　价	**38.00**元

凡购买我社图书，如有缺页、倒页、脱页的，本社发行部负责调换

前言

计算机翻译是语料库语言学应用的主要学科内容之一。它是一门用计算机技术来研究和处理自然语言的新兴应用语言学学科，涉及语言学、信息科学、数学、心理学等多个领域。其目的是通过建立形式化的语料库资源模型，凭借计算机程序来分析、处理和实现自然语言的转换，从而实现用计算机信息技术来模拟人类的部分乃至全部语言能力的目的。利用机器进行翻译的梦想肇始于古希腊时代。到20世纪40年代，计算机的问世昭示了机器翻译的美好前景。在经历了60多年的曲折发展后，机器翻译已取得了长足的进步，研究院所和开发机构林林总总，各种溢美之词不绝于耳。但由于自然语言的复杂性，全自动的机器翻译译文质量仍难以令人满意，计算机翻译系统商品化的艰难进程也印证了这种欲进不能的尴尬状况。

本书系统地介绍了语料库语言学在机器翻译研究中的应用，内容涉及语料库与语料库语言学的关系、语料库语言学的概念、自然语言处理技术的应用与语言信息工程、机器翻译的原理和实现过程，以及机器翻译评价的复杂性和主要方法、机器翻译的难点解析等，还预测了计算机翻译的未来研发方向。总起来看，这部著作系统性强，且深入浅出、循序渐进，知识性与趣味性乃至哲理性相辅相成，相信在普及语料库语言学与机器翻译知识方面是会起到一定的作用的。

本书是2014年度河南省科技厅软科学计划项目“实用机器翻译软件的语料库数据驱动技术应用研究”（142400411235）的研究成果。

书中多处引用了前辈及时贤的学术见解、观点及相关论述，均按照学术研究规范和惯例进行了标注，特此说明，并在此向这些前辈们的辛勤耕耘表示深深的谢忱和诚挚的敬意。

本书由华北水利水电大学外国语学院郑茗元副教授、常霜林讲师撰写。囿于作者自身的学术视野和认知能力，相信尚有许多富有新意的学术动态研究成果没有得到引述，祈望见谅。

郑茗元　常霜林

2015年4月

于郑州龙子湖畔

目录

第一章

语料库语言学

语料库是载有语言信息的大量语言资料的集合。语料库中的语料可以是为了特定目的而收集的语言资料，也可以是某一特定范围的书面材料，还可以是为了一般语言研究的目的而收集的语言资料。语料库语言学就是以语料库为基本知识源来研究自然语言规律的一门学科。它不仅仅是研究方法论的一个重要突破，更孕育着对语言描述框架乃至语言观念的不断更新（冯志伟，2002）。本书首先介绍了语料库与语料库语言学之间的密切关系，然后简述了后者的发展历史。随后较详细地描述了语料库的种类、加工和分析方法。最后，展现了语料库语言学在诸如词汇研究、外语教学、口语研究、修辞与文学、语域研究和翻译研究以及自然语言处理等多方面的应用前景。

第一节　语料库与语料库语言学

顾名思义，语料库（corpus/corpora）就是存放语言材料的仓库。或者说，语料库是为专门目的、按照明确设计标准收集的文章的集合（陈群秀，2007）。这个定义强调了三个方面：①语料库是为专门目的而建；②语料库有明确的设计标准；③语料库是文章的集合。McEnery 和 Wilson（2001）则对计算机时代的语料库给出了三个层次的定义，这三个层次由宽泛到精确，分别是：①语料库是任何文本的集合（any body of text）；②语料库是任何可以机读的文本的集合（a body of machine-readable text）；③语料库是一定量的可以机读的文本的集合，取样的文本在最大程度上代表一种语言或变体（黄河燕，2002）。

简单地说，语言材料堆积到一定的量或规模，就可以构成一个语料库。或者说，为了语言研究按照一定的原则收集和组织的真实的自然语言作品（书面的和口头的）的集合就可称之为语料库。现在人们谈起语料库，不言而喻是指存放在计算机里的原始文本或经过加工后带有语言学信息标注的语料文本。但从现代语言学的意义上，对语料库的认识至少还应包括如下三点。

（1）语料库中存放的是在语言的实际使用中真实出现过的语言材料。

（2）语料库是以计算机为载体承载语言知识的基础资源。

（3）真实语料需要经过分析、加工、处理，才能成为有用的基础资源，即加工后的语料库才能成为熟语料库（吕雅娟，2003）。

任何一种语言的语料都是无限的，要将其全部存入计算机是不可能的。但是为了研究一种语言，可以根据统计学的原理把这种语言的语料按一定的原则抽样（sampling）存入计算机，把它作为这种语言的代表而进行统计分析。这样分析得出的结果，只要抽样的原则合理，存入的语料有足够的量，结果也是相当可靠的（侯敏，1999）。现在世界上已经

有了不少规模较大的语料库，有些是国家级的，有些由大学和词典出版商联合建设。另外，由于个人电脑的迅猛发展和存储数据的硬盘价格持续下降，不少研究者也开始建立适合于自己研究的小型语料库。

那么什么是语料库语言学（corpus linguistics）呢？虽然语料库语言学的研究已经有一段历史，但还没有一个公认的定义。下面是三个较为常见的定义。

（1）根据篇章材料对语言的研究，称为语料库语言学。

（2）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。

（3）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。

因此，我们有理由认为“语料库语言学就是以语料库为基本知识源来研究自然语言规律的一门学科”（张祖建，1999）。

由此可以看出，语料库语言学这个术语有两层主要含义：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”不是一个新学科的名称，而仅仅反映了一个新的研究手段；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，只有在这个意义上，“语料库语言学”才是一个新学科的名称。

因此，语料库语言学不是语言学的一个新的分支，而是一种以语料库为基础的语言研究方法。它主要包括两个方面：一是对自然语料进行加工，研究语料库加工的理论、方法和工具；二是基于语料库的知识获取已加工语料库的利用方法。作为一种方法，它不仅可以用于研究语言系统的各个层面，而且可以应用于语言学之外的其他领域。

语料库语言学是20世纪80年代才崭露头角的一门交叉学科，它研究自然语言文本的分类、采集、存储、加工、统计分析和应用，目的是凭借大规模语料库提供客观翔实的语言证据来从事语言学研究和指导自然语言信息处理系统的开发和应用（辜正坤，2004）。

“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越来越大的影响”（刘颖，2002）。这是美国人汤姆斯（John Tomch，1962—　）等人于1996年为祝贺语料库语言学的主要奠基人和倡导者英国语言学家杰弗里·利奇（Geoffrey Leech，1928—　）60诞辰而编撰的语料库语言学研究论文集的开场白。近年来，对语料库语言学的类似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。

20世纪80年代以来，语料库语言学的崛起和迅速发展令世人耳目一新。人们希望通过大规模真实语料的调查来获取自然语言的各种语言事实及语言规律，多方面、多层次描写语言并验证各种语言理论和假设，甚至建立新的语言理论和语言观。许多国家相继建立了各种语料库，数量以数百计，库容规模也跃升到数亿词级。语料库建设正朝着扩大库容、国际化和多元化的方向发展。

第二节　语料库语言学的发展历史

运用语料库进行语言研究的历史可以追溯到19世纪末，许多语言学者（J. Svarabhakti、T. McEnery、黄昌宁等）都对语料库语言学进行过论述。现在一般以乔姆斯基的

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成形期、发展期和腾飞期三个时期。

一、成形期

早期语料库语言学是指20世纪50年代中期以前，在美国诺姆·乔姆斯基博士（Avram Noam Chomsky，1928— ）提出转换生成语法理论（transformational－generative grammar）之前的所有基于语言材料的语言研究。在这个时期，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（language acquisition）是较早大量采用语料进行研究的一个领域。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，许多父母对其子女话语发展的观察日记成为了研究素材。这些日记作为原始资料，不仅是当时Preyer和Stern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（黄曾阳，1998）。自20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. 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（1992）所说，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。

3. 句法和语义

一些语言学家用语料库研究语言的描述。例如，结构主义语言学家弗里斯（C. C. Fries，1952）在语料库调查的基础上建立了英语的描写语法。这项工作比20世纪80年代后期语言学家夸克（R. Quirk）等用语料库方法编撰的《英语语法大全》（A Comprehensive Grammar of English）早了30年。

4. 音系研究

利用自然语料开展音系（phonology）研究，在西方当首推早期的结构主义语言学家，如博厄斯（Franz Boas）和萨巫尔（Edward Sapir）等人。他们强调语料获取的自然性和语料分析的客观性。

20世纪50年代中前期，在实证主义（positivism）和行为主义（behaviorism）思潮的影响下，总的来说，经验主义在语言研究中占主导地位。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（荣晶，2000）。

二、发展期

1957年，乔姆斯基的转换生成语法开始兴起。《句法结构》（Syntactic Structures）等论著的发表，从根本上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，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与生俱来的，是遗传决定的。与

之相对的是，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。人并非生来具有一套有关语言的原则和处理方法。由于在语言学研究中，乔姆斯基的内在语言或语言能力说被广泛接受，从20世纪60—70年代这长达20年的时间里，实际上理性主义方法主宰了欧美众多国家的语言学研究（陈原，2003）。

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。

第一，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。他认为，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做出部分解释，因而语料不应当是语言学家从事语言研究的得力工具。

第二，语料的不充分性。乔姆斯基（1965）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（recursion）。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（陈群秀，2007）。

三、腾飞期

20世纪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。世界各地都开始建设自己的语料库并且开始跨国联合建立国际性的语料库。在这个时期，我国首批建成的语料库中就有广州石油大学石油英语语料库GPEC（Guangzhou Petroleum English Corpus，50万词次）（朱德熙，1985）和上海交通大学100万词次的科技英语语料库JDEST（杨惠中，1986）。国外具有代表性的是20世纪90年代问世、拥有1亿词次的英国国家语料库BNC（The British National Corpus），包含有9000万词次的书面语和1000万词次的口头语。还有由Greenbaum（1991）主持建立的国际英语语料库ICE（The International Corpus of English），它汇集了全球20个国家和地区的英语语料，仅每个子库就已经收有书面和口头语料各50万词次。

20世纪90年代中期以来，语料库语言学发展具有如下三个方面的特点：①建设了大规模、多品种的语料库；②对语料进行了深加工研究；③语料库在与语言相关的各个领域中得到广泛应用。由于计算机等电子传播和储存技术的快速发展带动了语料库的建设规模迅速扩大，20世纪60—70年代建立的第一代电脑语料库，如拥有百万词次的BROWN和LOB等，如今看来已算不上大的语料库。20世纪90年代建立的BNC就已经达到1亿词次，而21世纪初已实现通过因特网检索的在线语料库（WEBCORP）更是达10亿～50亿词次，平均每天更新的新闻语料网页就达200万词次。可见，建立大型语料库已经不再是很困难的事。更重要的工作是，如何建立各种有语域（register）、语体（style）乃至语篇（discourse）特色的大型语料库以及如何对现有的语料库进行多层次的开发与研究。例如对原始语料作各层面信息的自动附码处理、对多功能语料库检索工具的开发研制、对自然语言的自动识别和电脑化处理系统研究等（刘卓，2002）。

第三节 语料库的种类

语料库因所承载的语言知识的性质不同，就构成了不同类型的语料库，如有的语料库收集的是书面文本语料，有的语料库是专门收集口语语料，有的是单语语料，即只收集一种语言材料，如汉语或英语（如英国国家语料库 Brrc）；有的语料库是双语（指两种语言的文本构成的语料库，如英汉对照的语料）或多语语料等。有些语料库是属于国家的语言资源档案库，有些语料库在解决了语料版权问题之后可以成为流通的文化商品，但是更多的语料库是用于单位和个人的学术研究。建立电脑语料库可以在全世界、全国或某地区某单位乃至个人的能力范围内实施，现时国内外流行的电脑语料库类型可以包括以下几点。

1. *原始语料库*（Raw Corpora）

将现实中使用的口头和笔头语用文字形式收集起来，按一定原则语体（style）、历时（diachronic）、共时（synchronic）等归类汇编起来的各种语料库。例如，1999 年挪威 Bergen 大学新版的国际现代英语计算机档案 ICAME（International Computer Archive of Modern English）中的 Collection of English Language Corpora 光盘就包含了世界各地的英语口语和书面语、成人语和青少年语、当代语和古代语等 21 个语料库，共 1700 万词次的语料。其中，有 20 世纪 90 年代后期建立的，与先前的 Brown 和 LOB 作历时性对比的新版 FLOWN 和 FLOB[1] 以及 Wordsmit 和 TACT（Transient Area Control Table）等一批检索工具。

2. *附码语料库*（Annotated Corpora）

对原始语料进行了词性、语法、语音、语义或语篇乃至语用标记附码的语料库，如已作词性附码的 BROWN 和 LOB、已作语法附码的 ICE－GB（国际英语语料库-英国子语料库）和已作语音韵律附码的 LLC（The London-Lund Corpora）和英语口语语料库（Spoken English Corpora，SEC）等。

3. *平行语料库*（Parallel Corpora）

平行语料库是指两种或多种语言的对译语料在句子乃至单词短语层面上实现了对齐的语料库，如英语与法、德、西班牙等语种的平行语料库 CRATER 和北京外国语大学汉英对应语料库等。

4. *学习者语料库*（Learners Corpora）

学习者语料库即非母语学习者的口头语和书面语语料库。其中包括标注有学习者拼写和语法差错标记以及改错提示的语料库，如 ICLE（International Corpus of Learner English，国际英语学习者书面语料库）和 CLEC（Chinese Learner English Corpora，中国英语学习者书面语料库）等。

5. *网格式语料库*（Lattice Corpora）

这是指对自然语言（包括口语和手写语）进行自动语音和手写体识别处理之后生成的语料库。此外，Donald E. Walker（1990）在“The Ecology of Language”中，根据内容

[1] FLOWN 和 FLOB 语料库是 BROW 和 LOB 语料库的更新版，其内容含量分别高达 100 万词次。

把语料库划分为以下四种：

（1）异质型（heterogeneous）语料库。这种语料库收集不同种类的文本，没有事先确定的选材原则和特定的标准，范围广泛，并且主要以原貌形式存在。

（2）同质型（homogeneous）语料库。这种语料库收集同类语料，内容具有同一属性，如德国波恩大学 Karat 的语料库，只收集德国作家 Karat 的语料。

（3）系统型（systematic）语料库。它强调选材的系统性、均匀性、合理性，力求使选材具有广泛的代表性，以便真实反映特定语种、特定范围的语言事实的全貌，如英国的 BNC、欧洲的文本语料库网（The Network of European Textual Corpora）。

（4）专用型（specialized）语料库，即为某种特定用途而收集建立，如美国为研究儿童心里语言学建立的 Childes 语料库。

另外，语料库根据它所包含语言种类的数目可以分为单语语料库或多语语料库。而多语（双语）语料库大致上还可分为平行语料库（parallel corpora）和比较语料库（comparable corpora），即由不同语言的文本或同一语言不同变体的语素所构成的两个或两个以上的语料库。平行语料库指的是语料库中的文本构成译文关系，比较语料库指的是将表述同样内容的不同语言文本收集到一起的平行语料库，这些不同语言文本之间并不构成翻译关系。双语平行语料库在机器翻译、双语词典编纂等领域有着广泛的应用，而双语比较语料库则主要用于语言对比研究，如研究在同样情景下不同语言表达方式的差异等（冯志伟，1995b）。

目前，国际上著名的单语语料库有美国布朗大学的 Brown 语料库、柯林斯一伯明翰大学的 COBUILD（Collins Birmingham University International Language Database）国际语料库（2 亿词次的英语）、美国宾州大学为句法分析而设计的树库（Penn Treebank）等。国内也有北京大学计算语言学研究所等单位研制的基于《人民日报》的“汉语词性标注语料库”（杨惠中，2002）；国家语委和几家高校、科研机构在“863”项目支持下正在建设的巧亿字超大规模的平衡语料库；台湾学术研究机构的带有词性标注的 200 万词次汉语平衡语料库以及在“973”项目支持下清华大学建设的汉语句法树库（吕雅娟，2003）和汉语单语语料库等。

双语平行语料库中两种语言的文本构成互译关系，而多语平行语料库可以看作是多个双语平行语料库的集合。可见，平行语料库是双语多语语料库的一种，只要两种语言的语料达到篇章级的译文对应即可称为平行语料。此外，双语语料库还包括其他两种类型：一种是对语料库中的文本有更高的要求，如双语平衡语料库就要求题材分布等也要大致平衡，而双语对齐语料库则要求平行文本内的句子、短语或词汇需对齐之后才能处理；另一种是两种语言的文本只要是表述同样的内容就可以，不要求有互译关系，这样的语料可称为“双语比较语料”（俞士汶，2003）或“双语相同事件文本对”（许超，2005）。与单语语料库相比，双语平行语料库的建设起步较晚，数量也较少。最著名的平行语料库当属加拿大的议会会议录（Canada Hansards），该会议录同时用英、法两种语言记录而成。国内有前面提到的北京外国语大学汉英对应语料库。此外，20 世纪 90 年代建立的英语-挪威语双语语料库、英语-意大利语双语语料库，以及英国曼彻斯特大学科技学院翻译研究中心的翻译语料库（TEC）等也都很著名。而包含语言最多的平行语料库是圣经语料库，

由马里兰大学的 Resnik 等人构建，包含了 9 种语言（英语、法语、丹麦语、芬兰语、希腊语、瑞典语、拉丁语、西班牙语和越南语）的语料。

第四节　语料库的收集与加工

语料库的设计和建设是一件费时、费力的工作。著名的语言学家 Leech（1998）曾中肯地指出，只有对收集与建立计算机语料库有第一手实践经验的人，才能充分理解建库过程中的艰苦。建立一个质量、设计标准等恰当的语料库，比起预先估计的复杂程度，可能会多花费一倍的时间，有时甚至多花费十倍的努力。只有语料库大量的基础性工作做完之后，才会有语料库使用者的大丰收。此外，还要注意以下几点。

1. 语料的平衡

由于当前电子文本已经随处可见，收集语料特别是单语语料相当容易，足够自然语言处理的研究使用。在此情况下，需要对语料进行适当的选择和整理，主要的要求是在选材时注意语料的广泛性和代表性，要选择不同题材和体裁的语料。题材应包括国内外新闻、政治、经济、科普、各行各业、军事、体育、文艺等；体裁应包括记叙文、说明文、议论文、应用文等。

2. 语料的管理

在收集到的大规模语料基础上进行随机抽样，建立不同的可控语料库，以供各种研究使用，这是语料库语言学研究的一大优势。抽样一般采取自动方式，这要求对原始语料进行适当的管理，需要知道语料的类型、版权、名称以及内部篇章结构等信息。此外，还可以对原始语料进行自动分类，以便更好地利用。

3. 语料库的大小

语料库越大，覆盖面越广，并且按照实际使用比例进行选取，就越具有代表性。这就要求语料的收集应最大限度地代表被收集语言的使用情况。因此，语料库语言学中的量化比其他经验语言学中的量化更有意义，这样的语料库能使我们了解语言的多样性，而不是仅仅被分析的样品。

4. 加工的深度

仅仅把原始语料按照某种原则收集和组织起来，对于语料库的深入研究还是不够的。因为没有对语料的深加工，就不可能把真实文本中所蕴含的语言知识挖掘出来。因此，必须对语料库进行不同层次的加工或标注（annotation 或 markup）。Meyer（2002）将单语语料库的加工层次分为三种：结构标注、词性标注和语法标注。此外，Meyer（2002）还提出了其他种类的标注，如语义标注（semantic tagging），即用详细指明语料库中词汇意义的各种特征的标记标注；语篇标注（discourse tagging），即语篇标注是指把一个文本的特征标注出来，以便分析者可以了解该语篇的结构；以问题为主的标注，这种标注方法需要分析者定义被使用的标记，并要人工标注待分析的语料库。此外，还包括词汇语言标注、句法标注（syntactic tagging）等。对双语语料库的加工则主要集中在句子和词汇级的对齐（alignment）上（黄曾阳，1998）。

目前，对语料进行词性标注比较普遍，英语和汉语都有各种已经标注了词性的语料

库，有的还达到了相当的规模。对语料进行词性标注，是对语料进行句法分析前重要的一步。

汉语词汇语义的标注工作主要是利用汉语《同义词词林》中的语义分类对每个汉语词进行语义标注。对于汉语语料，分词（segment）是一道必要的工序。北京大学计算语言学研究所、人民日报标注语料库样例和所用部分标记见表 1－1 和表 1－2。

表 1－1　　北京大学计算语言学研究所、人民日报标注语料库样例

历史/n 将/d 铭记/v 这个/r 坐标/n；/w 北纬/b 41.1/度/q；/w 东经/b 114.3/m 度/q；/w 人们/n 将/d 铭记/lv 这一/r 时刻/n；/w 1998 年/t 1 月/t 10 日/t 11 时/t 50 分/t；/w

表 1－2　　北京大学计算语言学研究所、人民日报标注语料库所用部分标记

标　记	含　义	标　记	含　义
b	区别词	d	副词
n	名词	r	代词
m	数词	q	量词
v	动词	w	标点符号
t	时间词		

带句法标注的语料库的代表是英语 Penn Tn. Bank，在词汇级以上给出了 40 多种的标注符号，句法标注已经相当深入，如每个被省略词汇的句法作用都被标注了出来，如表 1－3 所示。表 1－4 则是对其中的标记的含义的解说。

表 1－3　伦敦-隆德英语口语语料库 LL（London－Lund Corpus of Spoken English）样例

```
ˆ what a _ bout a cigarl \ ette#. /
* ( (4 sylls)) * /
* [ˆw \ on't have one th/anks# * ——— /
ˆ aren't you . going to sit d/own# — /
ˆ[/ \ m] #. — /
ˆ have my _ coffee in p=eace# ——— /
ˆ quite a nice. Room to! S/it in ( (actually)) #. /
* ˆ \ isn't * it /
* ˆy/ \ es# * ——— /l
```

表 1－4　　伦敦-隆德英语口语语料库所用部分标记

标　记	含　义
#	End of the tone group 语调群的结束
ˆ	Onset 语音开始
/	Rising nuclear tone 上升型核心语调
\	Falling nuclear tone 下降型核心语调
ˆ	Rise—fall nuclear tone 先升后降型核心语调
_	Level nuclear tone 平型核心语调

续表

标　记	含　义
[ ]	Enclose partial words and phonetic symbols 标记不完整的词语和音节符号
.	Normal sues：标准重音
!	Booster：higher pitch than preceding prominent syllable 音高高于前一个音节的重音
=	Booster：continuance 音高跟前一个相当的重音
(( ))	Unclear 不清晰的音节
* *	Simultaneous speech 同步发音
—	Pause of one stress unit 一个重音单位的停顿

从上面两个简单的样例可以看出，语料库中的语料在做了规范的加工标注后才真正显示出了其价值所在。

第五节　语料库语言学的应用

语料库语言学的研究和应用主要有这样几个方面：词汇研究（Lexical Studies）、在外语教学中的应用、在其他领域的应用等。

一、词汇研究

传统的词汇学主要研究词汇的各种意义以及同义词之间的区别。利用语料库的研究方法，可以拓宽词汇学研究的领域，归纳起来主要有以下三个方面。

1. 语义、词语搭配和主题词研究

对词语语义、语用的研究一直是词汇学研究的主要领域。利用语料库中真实的语料，研究者能对单个词语的意义和语用功能做出更为客观的描述，将研究成果将对学习者深入全面了解词的意义和实际面貌很有帮助。

词语搭配研究越来越得到人们的重视。语言学家 J. R. Firth（1957）有一句名言："观其伴，而知其意。"一个词的词义只能通过与之相伴出现的搭配才能加以辨识。从这一观点出发，无论是要识别一个词的不同词义，还是学会这个词的用法，都必须普遍调查词语的搭配关系和用法模式。换句话说，词的含义与上、下面有极其密切的关系，即词的含义服从与这样的一般规则：一个词用于一种新的语境时，就具有了新的含义。通过语料库我们可以在自然语境下观察词语的搭配行为和类连接，进而促进词汇教学（黄昌宁等，2002）。

例如，冯跃进和陈伟（1999）研究了含有 3.2 亿词次的 Bank of English 中对于表达汉语中"副职"的英语语料后发现：与 deputy 搭配的词汇较多，有 minister、leader（议长）、chairman、director、mayor、editor、manager、secretary-general 等；与 associate 搭配的词汇较少，主要是 professor、editor、director 等与学术相关的副职；与 vice 搭配的词汇相对固定，主要是 president、chairman、chancellor 等；与 assistant 搭配的主要涉及立法、执法方面；与 under 搭配的词汇只 secretary-general、secretary（次秘书长）；与 sub 搭配的词汇只有 dean、agent、deacon（补祭，助祭）、prefect（县长）。

他们根据这个调查，对汉语职务中副职的汉译英提出了这样的建议：将行政事业、商业、企业、公司的副职译为 deputy；将学术团体、学校、院所、报社杂志社的副职译为 vice、associate 和 deputy 其中 deputy 指业务副职，vice 和 associate 指职称性和学术性的副职（陈群秀，2007）。

利用语料库研究主题词，分析主题词与主题表达之间的关系，关键主题词与它们的联想词以及搭配词之间的相互关系，可以拓宽词汇研究的方法。

2. 词语语义韵律研究

词语的语义韵律（semantic prosody）是一个词语与语言中其他词语反复联系而获得的连续的意义氛围，它通常表达某种态度意义。使用语料库，通过检索节点词（node），以及检索上下面或临近若干个词而组成的并置结构（collocates），并将出现这些节点词或并置结构的句子进行比较分析，就能揭示用常规方法很难发现或很难确定的语义特征——语义韵律。

3. 词典编撰

字典编撰者用语料库来编撰词典是语料库运用的又一个方面。语料库与词典学的关系和对词典学的贡献，在国内外辞书出版界可谓人人皆知。计算机语料库给出了关于某一词或词语的所有用法举例，使词典的编撰与修改速度空前加快。同时，语料库中大量的自然语言例证使词的定义更加完整、精确。词或词语在真实语料中的前后搭配语境更清楚地显示该词或词语的语义特征、使用频率和语用特点等。这一切都使词典的编撰更趋科学化。

二、在外语教学中的应用

利用语料库对英语语言作多方面研究，还能进一步揭示语言规律，有助于英语的教和学。对教师课堂用语的研究可以提高教师对自己使用英语的认识和敏感性；对学习者中介语（interlanguage）的研究可以帮助教师认识外语学习的规律，采取科学合理的教学方法。利用语料库辅助写作教学和翻译教学也取得很好的进展。语料库在外语教学中的应用可以转变教学思想，改进教学方法，具有重要意义。

1. 语言教学

语料库的研究成果在语言教学中的运用是多方面的。参考语料库语言学对英语语言的描述，人们可以更科学地制定和修订教学大纲，更合理地编写教材，更准确地制定教字词表。1994 年 9 月开始实行的《大学英语教学大纲通用词汇表（1－4 级）》就是利用上海交通大学的 JDEST 语料库提供的科技英语词汇表和其他词汇表进行定量分析而制定的。语料库用于课堂教学有助于改进教学方法。一个重要的例子是基于语料库的数据驱动学习（data-driven learning）。这种新的教学模式鼓励学生自己积极主动地从真实语料中去观察语言现象，发现语言规律。

2. 教师语言和中介语的研究

基于语料库对英语教师话语的研究成果有助于教师对其教学用语的认识并提高其应用教学用语的敏感意识，为英语教育和教学改革提供实证性参考依据。何安平（2003）调查和分析了英语课堂教学语料库中教师话语的部分语言特征，探讨了国内高中、初中和小学英语课教师话语中的认知思维导向特点及其教育教学功能。利用语料库对学生中介语的研究有利于教师在教学过程中采取更合理的教学方法，提高教学效果。李文中（2004）的基

于大学英语学习者语料库（COLEC），分析和研究了学生在写作中主要搭配类型的基本中介语特征和学生采取的策略，对词汇教学具有积极意义。

3. 写作教学和翻译教学

利用语料库研究写作教学，可以通过分析学生作文中出现的错误，或是通过与本族语学习者语料库比较来了解母语写作能力对英语写作能力的影响，为写作教学提供建议和参考依据，也可以利用语料库改进写作教学的评估模式。娄宝翠（2001）利用中国学习者英语语料库（Chinese Learner English Corpus，CLEC）中的“大学英语学习者作文子语料库”研究中国学生的造词现象，提出了外语教师对造词现象应采取的态度以及在教学中应采取的相应措施。

语料库对翻译教学的研究基本上还是理论层面的探讨，在翻译教学中的具体运用还有待进一步的探究。于连江（2004）归纳了语料库在翻译教学中的运用。王克非（2004）探讨了平行语料库在翻译教学中的应用价值。

4. 口语教学的研究

利用口语语料库对学生口语的研究，包括对小品词的研究、韵律特征研究、学生交际策略的研究、某些句式在口语中的语用功能的研究等。对口语多方面的研究能更好地指导教师的口语教学，培养和提高学生的口语能力。何安平（2003）等采用国内外多个口语语料库，研究英语语篇标识语中的口语小品词在各类语料库中的分布频率和类型，语篇和语用功能等。这对改进英语口语教学，尤其对口试评估具有启发意义。何莲珍等（2004）利用大学英语四、六级考试口语考试（CET SET）语料库研究非英语专业大学生在大学英语口语考试中使用交际策略的情况。该研究发现，口语水平对交际策略的观念和使用影响显著（何安平，2004）。

三、在其他领域的应用

1. 利用语料库检索研究修辞和文学

用语料库研究修辞和文学可以将定性与定量研究方法相结合，使研究更科学可信。周江林等（2003）使用英国国家语料库（BNC）检索 high 和 low 两个词，研究英语的空间隐喻，从语义的角度，通过分析共现于同一语境中的有关词项的语义特点，可以开辟一个新途径来理解英语的修辞手段及其效果。

语料库在文学领域的研究可以通过对文学作品文本总体特征的描述来分析文学文本、作家的写作技巧、语言风格等。杨建玫（2002）通过对《警察与赞美诗》文本的统计数字推断出文本的难易程度，篇幅长短。通过统计出来的高频词可以了解文本的大意，通过对关键人物的检索和对其上下面搭配结果可以进一步分析作品的详细情节，揭示作品的主题思想，体会本课题组的写作技巧。

2. 对英语在不同语域的研究

通过对不同语域语料库的调查研究，可以了解在不同语域中英语使用的一些特点。余千华等（2001）以科技英语语料库中统计的模糊限制语使用频率作为参照标准，研究中外重要英语科技期刊上的论文中模糊限制语的使用情况，分析说明了中外科技工本课题组在用英语写作科技论文时使用模糊限制语习惯上的一些异同点，给中国科技工本课题组用英语写科技论文时提供了借鉴。