语料库研究与应用综述

目录

语料库研究与应用综述
(1998-2003)

傅爱平

一 概述

语料库通常指为语言研究收集的、用电子形式保存的语言材料，由自然出现的书面语或口语的样本汇集而成，用来代表特定的语言或语言变体。经过科学选材和标注、具有适当规模的语料库能够反映和记录语言的实际使用情况。人们通过语料库观察和把握语言事实，分析和研究语言系统的规律。语料库已经成为语言学理论研究、应用研究和语言工程不可缺少的基础资源。

语料库有多种类型，确定类型的主要依据是它的研究目的和用途，这一点往往能够体现在语料采集的原则和方式上。有人曾经把语料库分成四种类型：（1）异质的（Heterogeneous）：没有特定的语料收集原则，广泛收集并原样存储各种语料；（2）同质的（Homogeneous）：

只收集同一类内容的语料；（3）系统的（Systematic）：根据预先确定的原则和比例收集语料，使语料具有平衡性和系统性，能够代表某一范围内的语言事实；（4）专用的（Specialized）：只收集用于某一特定用途的语料。除此之外，按照语料的语种，语料库也可以分成单语的（Monolingual）、双语的（Bilingual）和多语的（Multilingual）。按照语料的采集单位，语料库又可以分为语篇的、语句的、短语的。双语和多语语料库按照语料的组织形式，还可以分为平行（对齐）语料库和比较语料库，前者的语料构成译文关系，多用于机器翻译、双语词典编撰等应用领域，后者将表述同样内容的不同语言文本收集到一起，多用于语言对比研究。

语料库建设中涉及的主要问题包括：

（1） 设计和规划：主要考虑语料库的用途、类型、规模、实现手段、质量保证、可扩展性等。

（2） 语料的采集：主要考虑语料获取、数据格式、字符编码、语料分类、文本描述，以及各类语料的比例以保持平衡性等。

（3） 语料的加工：包括标注项目（词语单位、词性、句法、语义、语体、篇章结构等）标记集、标注规范和加工方式。

（4） 语料管理系统的建设：包括数据维护（语料录入、校对、存储、修改、删除及语料描述信息项目管理）、语料自动加工（分词、标注、文本分割、合并、标记处理等）、用户功能（查询、检索、统计、打印等）。

（5） 语料库的应用：针对语言学理论和应用领域中的各种问题，研究和开发处理语料的算法和软件工具。

我国语料库的建设始于 80 年代，当时的主要目标是汉语词汇统计研究。进入 90 年代以后，语料库方法在自然语言信息处理领域得到了广泛的应用，建立了各种类型的语料库，研究的内容涉及语料库建设中的各个问题。90 年代末到新世纪初这几年是语料库开发和应用的进一步发展时期，除了语言信息处理和言语工程领域以外，语料库方法在语言教学、词典编纂、现代汉语和汉语史研究等方面也得到了越来越多的应用。

语料库与语言信息处理有着某种天然的联系。当人们还不了解语料库方法的时候，在自然语言理解和生成、机器翻译等研究中，分析语言的主要方法是基于规则的（Rule-based）。对于用规则无法表达或不能涵盖的语言事实，计算机就很难处理。语料库出现以后，人们利用它对大规模的自然语言进行调查和统计，建立统计语言模型，研究和应用基于统计的（Statistical-based）语言处理技术，在信息检索、文本分类、文本过滤、信息抽取等应用方向取得了进展。另一方面，语言信息处理技术的发展也为语料库的建设提供了支持。从字符编码、文本输入和整理，语料的自动分词和标注，到语料的统计和检索，自然语言信息处理的研究都为语料的加工提供了关键性的技术。

下面先简要叙述 1998 年到 2003 年中国语料库建设的基本情况，然后介绍语料库的加工、管理和规范问题，最后谈谈语料库方法在语言研究和语言工程等方面的应用。由于以前的《中国语言学年鉴》很少谈及语料库问题，为了尽可能全面地反映我国语料库研究和应用的情况，必要时会将时间上限向前延伸几年。

二 中国语料库建设的基本情况

90 年代末到新世纪初这几年投入建设或开始使用的语料库有数十个之多，不同的应用目的使这些语料库的类型各不相同，对语料的加工方法也各不相同。下面是其中已开始使用并且具有一定代表性的语料库。

（一）现代汉语通用语料库

这是一个由国家语言文字工作委员会主持建立、面向全社会应用需求的大型通用语料库，从 90 年代初开始建设，计划规模 7000 万字，主要应用目标是语言文字信息处理、语言

文字规范和标准的制定、语言文字的学术研究、语文教育、以及语言文字的社会应用。

这个语料库收录的语料以书面语为主、以书面语转述的口语为辅。语料来源是 1919 年至今，主要是 1977 年至今出版的教材、报纸、综合性刊物、专业刊物和图书。在设计原则上，讲求通用性、描述性、实用性和抽样的科学性。在语料分类方面，以“门类为主，语体为辅”为原则制定三个大类：

第一类：人文与社会科学类（包括 8 个次类、30 个细类）

1．政法类：　哲学　政治　宗教　法律
2．历史类：　历史　考古　民族
3．社会类：　社会学　心理　语言文字　教育　文艺理论　新闻　民俗
4．经济类：　工业经济　农业经济　政治经济　财贸经济
5．艺术类：　音乐　美术　舞蹈　戏剧
6．文学类：　小说　散文　传记　报告文学　科幻　口语
7．军体类：　军事　体育
8．生活类

第二类：自然科学类（包括 6 个次类）

1．数理类
2．生化类
3．天文地理类
4．海洋气象类
5．农林类
6．医药卫生类

第三类：综合类（包括 6 个次类，30 多个细类）

1．行政公文类：　请示　报告　批复　命令　指示　布告　纪要　通知等
2．章程法规类：　章程　条例　细则　制度　公约　办法　法律条文等
3．司法文书类：　诉讼　辩护词　控告信　委托书等
4．商业文告类：　说明　广告　调查报告　经济合同等
5．礼仪辞令类：　欢迎词　贺电　讣告　唁电　慰问信　祝酒词等
6．实用文书类：　请假条　检讨　申请书　请愿书等

在不同类别、不同来源、不同时期的语言材料中，按照不等密度的思路确定合适的语料选取比例，从共时和历时两个角度保证入选语料的平衡性，是这个语料库的特点。譬如，在语言材料的年限方面，选材比例是：

1919 年 – 1925 年　5%	1926 年 – 1949 年　15%
1950 年 – 1965 年　25%	1966 年 – 1976 年　5%
1977 年以后　50%	

在语言材料的门类、语体和来源方面，选材比例是：

人文与社会科学类占 59.6%。其中各个次类在本大类中的比例是：

政法　12.7%　历史　8.4%　社会　14.0%　经济　9.8%
艺术　6.7%　文学　44.9%　军体　2.3%　生活　1.4%

自然科学类占 17.24%。其中各个次类在本大类中的比例是：

数理　17.2%　生化　19.1%　天文地理　14.1%
海洋气象　9.1%　农林　22.8%　医药卫生　17.7%

综合类占 9.36%。其中各个次类在本大类中的比例是：

各类应用文　91.1%　其他　8.9%

报纸类占 13.79%。其中各个次类在本大类中的比例是：

全国性报刊　25%　省市报刊　75%

这个语料库在选材过程中收集和记录语料的有关描述信息，为每个语料样本设立了 20 个描述项目：总号、分类号、样本名称、类别、作者、写作时间、书刊名称、编著者、出版者、出版日期、期号（版面号）、版次（初版日期）、印册数、总页数、开本、选样方式、样本起止页数、样本字数、样本总数、繁简字。用户可以利用这些语料描述标记根据各自的需要进行各种方式的检索。语料库的建库工作分为两步，第一步先建立核心语料库（由 7000 万字

的语料中筛选出 2000 万字语料组成）。到 90 年代末，完成了 2000 万字生语料的收录工作。从 2001 年开始，对 2000 万字核心语料进行分词和词性标注加工。

（二）《人民日报》标注语料库

《人民日报》标注语料库由北京大学计算语言学研究所和日本富士通公司合作，从 1999 年开始，到 2002 年完成，原始语料取自 1998 年全年的《人民日报》，共约 2700 万字，到 2003 年又扩充到 3500 万字，是我国第一个大型的现代汉语标注语料库。这个语料库加工的项目有词语切分和词性标注，还有专有名词（人名、地名、团体机构名称等）标注、语素子类标注、动词、形容词的特殊用法标注和短语型标注。下面是一段语料标注的示例，对于 1998 年 1 月 1 日第 5 版第 1 篇文章的第 11 段：

我国的国有企业改革见成效。位于河南的中国一拖集团有限责任公司面向市场，积极调整产品结构，加快技术改造和新产品研制步伐。图为东方红牌履带拖拉机生产线。（赵鹏摄）

标注后的形式是：

19980101-05-001-011/m 我国/n 的/u 国有/vn 企业/n 改革/v 见/v 成效/n 。/w 位于/v 河南/ns 的/u [中国/ns 一拖/j 集团/n 有限/a 责任/n 公司/n]nt 面向/v 市场/n ，/w 积极/ad 调整/v 产品/n 结构/n ，/w 加快/v 技术/n 改造/vn 和/c 新/a 产品/n 研制/vn 步伐/n 。/w 图/n 为/v 东方红牌/nz 履带/n 拖拉机/n 生产线/n 。/w （/w 赵/nr 鹏/nr 摄/Vg ）/w

在每一个切分出来的词和标点符号后面，是该词语的标记。譬如词性标记（n，v，a，u，m，w 等），专有名词标记（nr，ns，nz 等），语素子类标记（Vg 等），动词和形容词特殊用法标记（vn，ad）。所有的标记都是以北京大学的《现代汉语语法信息词典》为基础词库，在一个加工规范的指导下标注的。

利用《人民日报》标注语料库，人们可以从各个角度考察和分析语言事实，统计各种语言单位出现的频率，譬如，词语或词类的分布、搭配和共现，专有名词的结构方式、兼类词在句子中的表现，语素字的使用情况，等等。也可以从语料里提取各种语言单位或语句片段作为研究实例。与仅仅以汉字串的形式表示的“生语料”相比，经过标注的“熟语料”显然含有更多的语言学特征信息，对汉语词汇研究、语法研究和汉语信息处理系统来说是更好的语言知识资源。

《人民日报》标注语料库中一半的语料（1998 年上半年）共 1300 万字已经通过《人民日报》新闻信息中心公开提供许可使用权。其中一个月的语料（1998 年 1 月）近 200 万字在互联网上公布，供自由下载。

（三）用于语言教学和研究的现代汉语语料库

建立现代汉语语料库的主要目的之一是对外汉语教学和现代汉语研究，可以分为书面语语料库和以文本形式表示的口语语料库两类。前者如北京语言大学的汉语中介语语料库、现代汉语研究语料库，后者如中国社会科学院语言研究所的北京地区现场即席话语语料库。

汉语中介语语料库的建设目标是为对外汉语教学、中介语研究、偏误分析和汉语本体研究提供资源，因此它的语料来源很有对外汉语教学的特点。作者先在北京和其他省市的 9 所高等院校里，从来自 96 个国家和地区的 1635 位外国留学生那里收集了成篇成段的汉语作文或练习材料 5774 篇，共 3528988 字。再从中抽取了 740 人的 1731 篇语料，共有 44218 句，1041274 字。全部语料都记录了学生姓名、性别、年龄、国别、是否华裔、第一语言、文化程度、所学主要教材、语料类别、写作时间、提供者等 23 项属性。然后对这 104 万字的语料进行词语切分、词性标注以及一些专用的语言学特征标注。例如，标出了字、词、句、篇等不同的层次，对语料的非规范形式（例如：错字、别字、繁体字、拼音字、非规范词等）做出索引标记，记录其对应的规范形式。这个语料库的管理系统有语篇属性登录、文本过滤、文字预处理信息登录、语料抽样、断句、分词、词性辅助标注、自动标注以及语料的主题检索、全文检索和数据浏览等各种功能，分别处理语料库的建立、管理和维护，以及用户浏览、查询和检索等。与人工收集的学生病句卡片资料相比，中介语语料库能够更好地反映学生学

习汉语的情况，帮助教师更加全面地观察他们的学习过程，了解影响学习和习得的各种因素。在汉语作为第二语言的教学中，为教材编写、课堂教学、测试等环节提供依据。

现代汉语研究语料库的建设目标是为语言学家提供一个研究平台，由 2000 万字的粗语料库和 200 万字经过分词和词性标注的精语料库两个部分组成。粗语料库收录的语料样本中绝大部分是九十年代的出版物，有《人民日报》1000 万字，《中国新闻》500 万字，各种书籍 250 万字，文学作品 150 万字，准口语材料（书面形式的对话、独白）100 万字。精语料库的 200 万字语料样本是从粗语料库中按照规定的比例由计算机随机抽取的，有书面语语料 160 万字，准口语语料 40 万字，是从语体、题材、体裁三个方面均衡选取的平衡语料库。为了对这些语料进行词语切分和词性标注，作者制定了词语切分的细则和词性标记体系的原则，采用了一个含有 112 个词类标记的标记集，确定了兼类词的处理方法。这个语料库的管理系统具有建库、检索、浏览、统计、输出等功能，可以按词或词类检索，统计出词的频率、词类频率、词类共现频率、平均词长、平均句长等结果。这个语料库建成以后，很快应用在现代汉语语法、汉语教学和汉语信息处理的研究中，研究内容涉及现代汉语的插入语、汉语句子的主题-主语标注、V+N 序列实验分析、词性标注中词语归类问题、动宾组合的自动获取与标注，等等。

建设北京地区现场即席话语语料库的目的是，通过收集大量的现场即席话语语料研究现场即席话语的各种动态机制，以揭示现场即席话语的使用规律。这个语料库的研究策略和取样方法很有特点，首先是严格区分资源库和语料库，资源库收集符合现场即席话语定义的录音材料，语料库收录按照一定标准从资源库提取出来的材料；另外在语料采样前先做摸底性研究，通过研究对现场即席话语的真实情况有所了解，确定取样域，再定取样范畴，然后根据取样范畴去录现场典型材料，这是一种层次范畴化的取样方法。这个语料库目前正在建设之中，已经取得了近 600 小时的录音材料和 50 多小时的录象材料。

在用于汉语研究的语料库中，讲究选材均衡，注重语料加工，同时也提供公开服务的，当数台湾中央研究院历史语言研究所的现代汉语平衡语料库(简称 Sinica Corpus)。这个语料库的规模为 500 万个词，每个句子都依词断开，标示词类标记，并且配备了检索系统，在网上开放供大家使用。根据自己制定的一套汉语文本属性特征为语料分类，在不同的类别上尽量均衡地采集语料，是这个语料库的特点之一。文本属性用来说明文档的呈现方式、文章的写作方式、文章写作的内容和文档的来源出处，包括 7 类，每类下设若干小类：

文类（文档的呈现方式）
　　报导、评论、广告图文、信函、公告启事、小说故事寓言、散文、
　　传记日记、诗歌、语录、说明手册、剧本、会话、演讲 、会议记录
文体（文章的写作方式）
　　记叙、论说、说明、描写
语式（文档的呈现方式）
　　书面语、演讲稿、剧本/台辞、口语谈话、会议记录
主题（文章写作的内容）
　　哲学、科学、社会、艺术、生活、文学
媒体　报纸、一般杂志、学术期刊、教科书、工具书、学术论著、一般图书、
　　书信、视听媒体、其它
作者　姓名、性别、国籍、母语
出版　出版单位、出版地、出版日期、版次

不同研究目的的语言学者可以自己按语式、文体、媒体和主题的小类选取不同类别的语料，组成“自订语料库”，在“自订语料库”的范围内进行语料的检索和统计。除了通常的按词语、词类的检索和统计以外，这个语料库的管理系统还提供了一种“进阶处理”功能，对检索出来的数据作进一步处理，对处理的结果还可以再次处理，形成多层的检索结果。

（四）面向语言信息处理的现代汉语语料库

90 年代中后期，面向语言信息处理的现代汉语语料库开始建立并投入应用。其中最早开发的是清华大学用于研究和开发汉语自动分词技术的现代汉语语料库，经过几年的积累已达到 8 亿多字生语料。在这个语料库的支持下，用统计语言模型的方法研究了汉语自动分词中的理论、算法和技术，编制了总数为 9 万多个词语的《信息处理用现代汉语分词词表》。这些研究工作体现了我国汉语自动分词技术的发展水平，词表被许多汉语自动分词系统作为底表使用，是不可缺少的基础资源。

TH 通用语料库系统是清华大学建立的另一个现代汉语语料库。这个语料库有两个特点，一是语料库管理系统根据不同的加工深度，分四个等级管理语料。第一级是生语料分库，有 4 千余万字；第二级以上都是加工程度不同的熟语料库，其中第二级存放经过自动分词并由人工校对过的初加工语料 500 余万字；第三级存放经过词性标注和人工校对的语料约 300 万字；第四级是经过句子成分标注和人工校对的语料。每个分库又按语料的来源分成一般书籍、报纸、杂志、论文和工具书五类子库。不同等级的语料可以为不同的应用目标服务。第二个特点是在这个语料库的支持下，进行了汉语信息处理技术的研究。譬如，采用以谓语为中心的句型成分分析与语料统计相结合的方法，自动分析汉语的句型，提出了一个“汉语句型频度表”；在汉语文本中自动标注句子成分和句型成分的边界；根据指定的句型在语料库里搜寻句子实例，等等。

HuaYu 人工标注语料库是清华大学和北京语言大学合作建立的一个现代汉语平衡语料库。这个语料库按文学、新闻、学术、应用文四个大类收录了 200 余万字语料。它的特点是讲究加工的深度，除了词语切分和词性标注以外，还根据语句中动词的类型和句子的长度进行“语块”标注和“句法树”标注，目的是为建立汉语短语分析或句法分析的语言模型获取统计数据提供资源。下面分别是语块标注和句法树标注的示例。

对句子“自古以来，人类就重视档案的保存和利用，设置馆库、选派专人进行管理。”进行语块标注以后得到的是一个无嵌套的线性序列，其中 S 是主语语块，P 是述语语块，O 是宾语语块：

[D 自/p 古/t 以来/f ，/， [S 人类/n [D 就/d [P 重视/v [O 档案/n 的/u 保存/vN 和/c 利用/vN ，/， [P 设置/v [O 馆库/n 、/、 [P 选派/v [O 专人/n [P 进行/v [O 管理 v。

对句子“我哥哥送给我一本很漂亮的书。”进行句法树标注以后，得到的是一个与树形结构等价的线性序列：

[zj-XX [dj-ZW [np-DZ 我/ rN 哥哥/n] [vp-PO [vp-PO [vp-SB 送/v 给/v] 我/rN] [np-DZ [mp-DZ 一/m 本/qN] [np-DZ [ap-ZZ 很/d 漂亮/a] 的/u 书/n]]] 。/w]

（五）用于开发特定语言分析技术的专用语料库

这类语料库是针对汉语信息处理技术的需要专门建立的。例如山西大学的专有名词标注语料库和分词与词性标注语料库。

分词与词性标注语料库，规模为 500 万字，带有分词标记、词性标记和句法标记。标注时依据《信息处理用现代汉语分词规范》和《信息处理用现代汉语词类及标记集规范》。在这个语料库的支持下，开发汉语自动分词和词性标注软件，研究自动分词和词性标注的评测技术。为了解决汉语自动分词中的切分歧义问题，还建立了交集型歧义字段库和组合型歧义字段库，专门收集这两种类型的歧义切分实例。前者有 7.8 万字，后者收录了 140 多条。并且在分词和词性标注语料库里作了这两类切分歧义的标注。利用这些语料调查交集型歧义当中的“伪歧义”现象（既切分结果只可能有唯一选择的那些交集型歧义切分字段），发现这种现象在歧义切分字段中很普遍，可以达到 90%以上。

专有名词标注语料库用于研究汉语自动分词中专有名词的识别算法。其中包括标注了中国地名的语料 280 万字，标注了中国人姓名的语料 300 万字，标注了西文姓名的语料 250 万字，标注了汉语机构名称的语料 50 万字，还有标注了网络新词语的语料 150 万字。利用这些语料，建立了中国地名用字、用词库，姓氏人名库，姓氏用字频率表，名字用字频率表等，用统计语言模型的方法识别专有名词。

（六）双语语料库

基于实例的机器翻译（Example-based）需要大规模的双语平行语料库来支持。语料库里的源语和目标语实例要按照相同级别的翻译单位一一对齐。目前已有的双语平行语料库主要是汉语和英语的，语料对齐的单位有句子级的、子句级的、短语级的，也有词汇级的。机器翻译系统把要翻译的句子与语料库里的源语实例进行对比，分析相似程度，找到最适合的源语实例，再参照与它对齐的目标语实例生成译文。用于这类机器翻译系统的双语语料库必须有一定的规模，用人工做语料对齐的工作显然很难满足要求。这就使文本自动对齐成为建立双语语料库的关键技术。

在目前已有的双语语料库中，哈尔滨工业大学的汉英平行语料库已经直接用来开发英汉双向机器翻译系统。这个语料库有 6 万个汉语和英语的句子，使用多级对齐加工技术，分别按照句子、短语结构和词一一对齐。中国科学院计算技术研究所的汉英双语语料库有 20 万个句对，也完成了句子一级的对齐，并在网上提供查询服务。北京大学、中国科学院软件研究所等单位也建立了按句对齐的汉英双语语料库。除此之外，还有以语段或短语为单位收集的汉英双语语料库，譬如中国科学院自动化研究所的汉英双语短语库，有 3~5 万对已对齐的汉语和英语短语。东北大学的英汉双语语段库，用来帮助建立电子版的英汉搭配词典。

（七）面向汉语史研究的语料库

面向汉语史研究的语料库建设是从搜集汉语史文献资料开始的。台湾中央研究院历史语言研究所从 90 年代初期就开始了这项工作，他们先收集上古汉语的语料，然后扩展到中古汉语和近代汉语。90 年代中后期逐步开始上古汉语语料和近代汉语的标注，在该院信息研究所和计算中心的协助下进行标注技术和检索技术的开发。根据是否经过分词处理和词性标注，台湾中央研究院的古汉语语料库和近代汉语语料库可以分成两类：生语料库和标记语料库。目前生语料库收集的语料已涵盖上古汉语（先秦至西汉）、中古汉语（东汉魏晋南北朝）、近代汉语（唐五代以后）的大部分重要文献资料，并已陆续开放使用。在标记语料库方面，上古汉语及近代汉语都已有部分语料完成标注工作，也逐步提供网上检索。2001 年底，开放了近代汉语标记语料库ＷＷＷ版供各界使用，首先提供查询的文献是《红楼梦》及《三遂平妖传》。在查询方面，除了常用的功能以外，还可以在显示词项及词类的同时给出例句的出处，便于历史语法的研究者使用。

多年来中国社会科学院语言研究所也一直在致力于文献资料的建设，搜集整理了近代汉语书面语语料 150 万字，中古近代汉语语料约 1 千万字，部分语料已作了标注。目前已经完成了一个小型语料库，包括：敦煌变文集、祖堂集、三朝北盟汇编、碧岩录、朱子语类、刘知远诸宫调、西厢记诸宫调、元刊全相平话五种、元典章 刑部、老乞大谚解、朴通事谚解、孝经直解、鲁斋遗书、经筵讲义等十余种文献，成为汉语史和语言学理论研究的重要资源。此外，语言研究所的先秦专书电子文档有 4 部文献，共约 120 万字，并且已由古汉语学者逐篇逐句标注了语法信息。

上海师范大学、浙江师范大学、四川大学等学校也依据各自汉语史研究的方向，建立了历史文献语料库。四川大学的中古汉语语料库有 1 亿字的中古汉语语料和有关中古汉语研究的资料。浙江师范大学的楚辞语库、前四史语库、六朝语库、太平广记语库、唐诗语库、宋词语库，已用于"前四史"语言研究和唐宋诗词语词研究。

目前历史文献语料库建设的特点是依托学科建设和研究方向，广泛收集资料，注重校勘精审。随着汉语史研究和语料库应用的发展，资源共享和语料加工将得到越来越多的重视。历史文献资源共享，首先要避免语料的重复收集，还要采用国际通用的标准处理语料文本，使语料能够准确、方便地交换和使用。语料加工则是充分发掘语料应用价值的基础工作，从收集历史文献的电子文档，到建成一个具有必要的语言学标记信息、合理的逻辑结构和方便的检索功能的语料库，语料的加工是不可或缺的一步。

（八）比较语料库

为了研究汉语在不同地区的使用情况，香港城市大学建立了 LIVAC 共时语料库（Linguistic Variation in Chinese Speech Communities）。语料来自香港、台湾、北京、上海、澳门及新加坡六地有代表性的中文报纸，以及电子媒介上的新闻报道。自 1995 年 7 月开始，每四天一次，收集这六个地区的对等书面语文本，每次约两万字。内容包括新闻、特写、评论等文章。到 2003 年上半年，已收集了 1 亿 1 千多万字、超过 56 万个词条。计划收集到 2005 年 6 月，囊括新旧世纪交接点前后各五年各地华语社区有代表性的重要语言数据，供汉语的各种共时比较研究使用。

在语料的组织和加工方面，这个语料库用计算机自动分词，再经人工校对分类，可以依字、词、句为基础进行检索，提供字、词配搭、分布等数据，有统计功能。语言学家能通过这个语料库考察上述六地出现的新词、词义有所发展或转移的旧词、以及有地方特色的词语，还可以对具体字或词的频率作统计比较，对字词的差别作计量分析。对研究华人社区的文化、社会、语言差异也有作用。这个语料库的一部分已经在网上提供服务。

（九）少数民族语言语料库

新疆大学从 2002 年起开始建设现代维吾尔语语料库系统，计划包括 5 个部分：语料库、电子语法信息词典、规则库、统计信息库和检索统计软件包。其中语料库部分又分成生语料库（经初步整理的原始语料）和加工语料库（经过标注和校对的语料）。目前已有生语料 800 万词。另外，新疆大学也正在以新闻领域的维汉-汉维机器翻译为目标，建设双语平行语料库。内蒙古大学的中世纪蒙古文语料库收集了《元朝秘史》、《黄金史》、《回鹘蒙古文文献集》等历史文献。他们还建立了 500 万词的现代蒙古语语料库，研究了蒙古文附加成分的自动切分、复合词的自动识别和语料的词性标注，获得了词频统计、音节统计、词类统计、附加成分统计等数据。西北民族大学建立了 1 亿 3 千万字节的大型藏文语料库，用于藏文词汇频度和通用度的统计。中国社会科学院民族学与人类学研究所建立了 500 万藏语字符的藏语语料库，进行词语切分和标注的研究。新疆师范大学也建立了 200 万词的维吾尔语语料库。

与汉语语料库相比，少数民族语料库的建设还需要解决一些特殊的问题，譬如拼音文字转写的标准和规范，词语分类体系及其标记集等。

到 2003 年，已建和在建的各种文本语料库还有很多（包括书面语语料库和以文本形式表示的口语语料库），以上提到的只是有代表性的一部分。与文本语料库相对的，是语音语料库。语音语料库不仅记录语图、声学参数等语音学数据，还有句法、韵律等各种语言学信息标记和副语言学信息标记，可以在语音识别与合成系统中用来建立语音模型，用于语音研究、语音工程开发和汉语普通话教学等领域。语音技术是当前信息技术和通讯领域里最具潜力的发展方向之一，语音语料库在科研和工程上有很高的使用价值。关于语音语料库的详细情况，请见“语音学和言语工程研究综述”。

三 语料库的加工、管理和规范

（一）语料的加工

一个计算机语料库的功能主要与三个因素有关，一是语料库的规模，二是语料的分布，三是语料的加工程度。规模的大小关系到统计数据是否可靠，语料的分布涉及统计结果的适用范围，语料加工的深度则决定这个语料库能为使用者提供什么样的语言学信息。

加工语料主要指文本格式处理和文本描述两项工作，前者是对采集的语料文本进行整理，转成统一的电子文本格式，例如数据库格式、XML 文本格式等。后者是描述每一篇语料样本的属性或特征，包括篇头描述和篇体描述。篇头描述说明整篇语料样本的属性，例如语体、内容所属的领域、作者、写作时间、来源出处等等，篇体描述是在文本里添加各种语言学属性标记，对于汉语书面语语料库来说，常见的是词语切分标记、词性标记、专有名词

标记，还有某些语法特征如短语标记、子句标记，或语义信息标记，等等。对汉语书面语语料的加工一般是从词语切分、词性标注，到语法、语义属性标注，按顺序进行。标注的信息逐步增多，语料加工的深度也就逐渐增加。人们通常把没有篇体描述信息的语料叫做生语料。对汉语的生语料只能以字为单位进行检索和统计。经过词语切分处理的语料，就能以词为单位进行检索、统计和定量分析。如果还作了词性标记，那么可以获得的语言学信息就更多了。语料的标注如果由人来做，当然能够保证准确性，但是人工标注对处理大规模的语料显然不够现实。所以几乎每一个大规模语料库的加工都需要借助自动化的手段，词语自动切分、词性自动标注等就成为备受关注的语料加工技术。

自动分词是我国最早开始研究的汉语信息处理技术之一。语料库的建设开始以后，自动分词技术在语料加工中又得到了应用和发展。自动分词和词性自动标注一般都需要一个词典，作为分词和词性标注的基础。这个词典与常用的语文词典相比，收录的词目不大一样，包括了语言学家认可的词，以及一些比词小的单位（如语素字、词缀等）和一些比词大的单位（如成语、习语、简称略语等）。词典中也包括词类信息和其他语法信息。目前的自动分词技术是基于字符串匹配原理的，有正向最大匹配、逆向最大匹配等基本算法。在切分过程中会出现歧义现象，如何处理歧义是自动分词研究的重点之一，在这方面投入的研究也最多，先后提出了“短语结构法”、“专家系统法”、“隐马尔科夫模型”、“串频统计和词匹配”等辩识歧义的方法。识别未登录词是自动分词研究的第二个重点。未登录词指没有被分词底表收录的词语，包括人名、地名、机构名等专有名词和新出现的词语。对未登录词的识别一般以基于语料库的统计语言模型方法为主。

词性自动标注通常与自动分词同时进行，根据带有词类信息的分词词典，给切分出来的词语标上初始的词类标记。对于兼类词，必须在句子里判断类别。因此需要分析兼类词语在上下文中的分布特点和语法功能，并用形式化的方式表达出来，作为词性标注系统排除兼类的规则。近年来，已经有几个自动分词和词性自动标注系统投入了应用，其中北京大学用自己研制的系统为《人民日报标注语料库》做分词和词性标注的初加工，北京语言大学的自动分词系统也成为其《面向语言教学研究的汉语语料检索系统》中的关键技术。此外，经过十几年的研究和实践，2001 年发布了收录 9 万多词语的《信息处理用现代汉语分词词表》和《现代汉语词类及标记集规范》。对于 1993 年制定的国家标准《信息处理用现代汉语分词规范》的可操作性问题，也进行了积极的讨论和实验，提出了有效的解决方法。关于自动分词和词性自动标注的详细情况，请见“计算语言学和自然语言信息处理研究综述”。

经过分词的语料，除了标注词性以外，还可以进一步标注其他语言学属性，譬如韵律、语调、短语结构、句法结构、语义关系等等。句子的语法结构需要有形式化的方式来表达，大多数语料库或者采用短语结构树，或者采用依存语法树的方式，这样标注过的语料库就成为短语树库或句法树库。一般情况下，在词性标注的基础上再作进一步的语法标注加工，多以人工为主，也有关于自动短语定界和句法信息自动标注的研究和实验。目前已有的汉语短语库、句法树库规模都不大，至多百万词级。

在双语语料库的建设中，除了上述语料加工项目以外，还有一项不可缺少的语料加工任务：双语语料对齐。语料对齐分为段落、句子、子句、短语和词语几个不同的层次。如果考虑用计算机程序做自动对齐，不同的层次要解决的问题各不相同。每种语言的段落都有可识别的标志，因此段落的对齐最容易实现，句子的对齐在印欧语言之间比它们和汉语之间要容易，词语的对齐需要借助词典，句子内的各种结构要自动对齐则是最难的。目前双语自动对齐技术的研究主要是针对句子和句子内的结构，采用的方法有基于长度的、基于词典的，或者是这两种方法的混合策略。

（二）语料库管理系统

经过科学选材和标注、具有适当规模的语料库，还应该有一个功能齐备的管理系统，包括数据维护（语料录入、校对、存储、修改、删除及语料描述信息项目管理）、语料自动加工（分词、标注、文本分割、合并、语料对齐、标记处理等）、用户服务功能（查询、检索、

统计、打印等）。其中数据维护部分主要涉及汉字字符处理、文本处理、文件管理等计算机程序设计技术。语料自动加工部分的主要内容是自动分词、各种语言学属性的标注技术，已经在前面专门介绍过了。这里主要谈谈面向用户的语料检索、统计和分析技术。

语料检索是一种全文检索技术，但是也有自己的特点，仅用普通的全文检索技术还不能满足语料检索的需要。这是因为，全文信息检索关心的是检索目标的意义，不是检索目标的语言表述形式。而面向语言研究的语料检索则特别注重语言的表述形式，它既需要按照字、字串和词检索，也需要把词语的语言学属性作为检索的目标和约束条件，还要求把检索的结果或目标的出处按照研究的需要排序、输出。除此之外，还要有字频、词频和特定语言形式出现频率的统计功能。

对汉语生语料的检索和统计是以字或字串为单位进行的。这一类检索系统主要以单字索引和字符串匹配为关键技术，由于把词语当作字串来检索，所以检索结果中经常出现“非词”的问题。例如要查找“出警”，检索结果中除了“迅速出警”、“拒绝出警”、“出警次数”等实例以外，“发出警告”、“放出警犬”等也混在其中。为了解决这些问题，常常需要为字符串匹配的检索表达式另外设置限制条件。这些限制条件大多是个性的，只能排除一部分“非词”的实例。要想从根本上解决这个问题，就必须对语料作词语切分。经过词语切分处理的熟语料，能以词为单位进行检索、统计和定量分析。但是熟语料库的加工代价很高，而且对于语料的词语切分和词性标注，目前还没有既成熟又便于操作的规范，所以近年来，面向生语料库的检索技术一直在广泛应用，并且在用户功能方面不断发展。譬如，可以对用户给出的任何生语料快速生成索引；可以使用具有复合逻辑关系的检索表达式；可以按照汉字、拼音、笔画对检索结果的上下文自动排序；可以提供检出实例的来源、出处；可以按字频统计的数据排序；检索结果和统计结果既可以按文本形式输出，也可以按数据库形式输出；还可以通过网络支持多用户远程检索。

对于经过词语切分处理和词性标注的熟语料库，除了所有生语料的检索功能以外，语料检索系统还可以把词语或词性作为检索的关键字或限制条件，得到关于这些语言学属性的检索和统计结果，并按各种排序和输出形式的提供给用户。语言学属性来自语言学家对汉语的研究，研究过程中有各种观点和认识，从词的定义到词类的确定，一直还没有统一的意见。另一方面，人们检索语料时的目的也各不相同，有的关心词汇问题，有的关心语法现象，还有的目标是汉语信息处理的应用问题。因此对于熟语料库检索来说，一个好的检索系统应该能够包容各种不同的语言学观点，可以用于不同的检索目的。

为了做到这一点，通常采用的办法是，把用于语料库自动分词的底表和附着于底表的词性、构词等属性都看作语言学属性表，使这个属性表与检索系统的程序相互独立，检索系统只把属性标记作为抽象的字符串处理，而把建立属性表的工作交给用户。以北京语言大学的《面向语言教学研究的汉语语料检索系统》为例，它的自动分词词表、词属性集和每个词的属性标记都由用户提供，提供的方式是把词目和它的属性标记登记在数据库里。检索系统使用用户提供的这个属性表对生语料自动分词，并生成索引，供给用户检索。检索系统对属性表没有任何限制，规模可大可小，表中的词目也可以跟通常认为的词没有关系，属性可以是语法的，也可以是构词的、语义的、语音的，等等。这样用户就能根据自己的需要检索和研究各种字串在语料中的表现。

把语料加工技术集成在检索系统里面，是语料库检索系统的另一个特点。语料加工技术一般指词语自动切分和词性自动标注。在北京语言大学的语料检索系统中，未登录词的自动识别技术比较有特点。它可以识别各种数字串、中西人名、中西地名、机构名、后缀短语等，并为它们建立索引，供用户检索和统计。

（三）语料库的规范问题

语料库的规范问题主要是对语料加工而言的。汉语语料库首先遇到的规范问题是词语切分。我国 90 年代初发布了国家标准《信息处理用现代汉语分词规范》（标准号为

GB/T13715-92）。这个规范基本上采用《暂拟汉语教学语法系统》中的观点，把词定义为“最小的独立运用的语言单位”。针对汉语语素、词和词组界限不够清晰的问题，还特别提出了“分词单位”的概念。把“分词单位”定义成“汉语信息处理使用的具有确定的语义或语法功能的基本单位”，并且用“结合紧密、使用稳定”的原则作为判断分词单位的标准。这样做的目的是避免关于如何界定词的争论。但是“结合紧密、使用稳定”的原则缺少可操作性，对于自动分词研究中的具体问题常常难有定论。于是就有了根据规范制定一个词表，用“规范+词表”的办法指导分词的建议。这样在 90 年代中期和末期，分别提出了收词 43570 条的《信息处理用现代汉语常用词表》和收词 9 万多条的《信息处理用现代汉语分词词表》。其中后者是在 8 亿字的大规模语料库支持下，采用“串频”、“互信息”、“相关度”等计算统计方法，依据定量的数据分析结果辨识“分词单位”的。与此同时，语言学家也参与了制定这个词表的工作，他们提出的各种语言学规则，从定性分析的角度与统计数据相互作用，最后经过人工审定，确定了 92843 个词目，其中一级常用词 56606 个，二级常用词 36237 个，成为目前许多自动分词系统使用的词表。

90 年代中期，台湾的计算语言学会也提出了一个《资讯处理用中文分词规范》。这个规范有三条基本原则，一是分词单位必须符合语言学理论的要求；二是在信息处理上切实可行；三是能够确保真实文本处理的一致性。它把分词规范分成信、达、雅三个不同的等级，“信”级是基本资料交换的标准，“达”级是机器翻译、情报检索等自然语言处理的标准，“雅”级则是分词的最好结果。这样可以根据不同的应用目的做难易程度不同的分词处理。

词语切分以后，下一个规范问题就是词性标注。经过十多年的词性标注研究和实践，教育部语言文字应用研究所于 2001 年提出了《信息处理用现代汉语词类标记集规范》。这个规范吸收了语言学家的研究成果，也兼顾了已有的各个用于语言信息处理的词类系统，制定了标记现代汉语书面语词类的符号集，使各种汉语信息处理应用系统能够尽量使用统一的词类标记，有助于信息交换和资源共享。

标注短语和句子结构是语料库进一步深加工的内容，虽然目前尚处于起步阶段，但已经在标注的同时考虑了规范的问题。清华大学提出的《汉语句子的句法树标注规范》，主要包括句法标记集的内容描述、句法树的划分规定、歧义结构的处理、结构分析的方向性等问题。上海师范大学根据自己制定的《汉语文本短语结构人工标注规范》，对 100 万字的 1997 年《读者文摘》进行了分词、词性标注和人工标注短语的试验。哈尔滨工业大学采用包含 23 个短语符号的标记集合，开发了一个 8000 个句子的汉语树库。清华大学还建立了一个基于语义依存关系的语料库，也涉及到标注体系的选择和标注关系集的确定。这些工作规模都不大，在规范方面还处于各自为政的状态。随着语料的进一步深入加工，统一规范将成为不可避免的问题。

北京大学的《人民日报》标注语料库是目前规模最大的汉语基本标注语料库。在它的开发过程中，各种加工规范起了关键的作用。在这些加工规范中，有词语的切分规范，主要规定把句子的汉字串形式切分为词语序列的原则；有现代汉语词类及标记集规范，规定切分出来的词语、短语、标点符号的类别和标识符号；有切分和标注相结合的规范，规定语素构成合成词的方式（重叠、附加和复合）；有标注规范，规定词性标注与词库的关系，主要解决如何在上下文环境里确定兼类词的词性；还有收词 7 万余条的词库《现代汉语语法信息词典》。加工大规模的语料是一项浩大的语言工程。语料标注的准确性和一致性需要靠完善、合理的词库和严谨、实用的加工规范来保证。《人民日报》标注语料库的加工规范和《现代汉语语法信息词典》是语言学家和信息处理专家合作，在汉语语法研究的理论和方法指导下，根据汉语信息处理的实际需要制定和开发的。在标注大规模语料的实践中，又得到了验证和完善。

除了语料加工以外，语料库还应该在语料的采集和存储格式上有所规范。对于平衡语料库来说，采集规范主要是为了保证语料的平衡性，而类别分布和时间分布是语料平衡的两大要素。每个语料库都要对语料进行分类，分类的原则各不相同。有的根据内容涉及的主题分

类，有的根据语体分类。在众多平衡语料库当中，台湾中央研究院的现代汉语平衡语料库的分类标准很值得注意。这个语料库的研制者认为，用传统的文体单一特征来界定平衡语料库不足以反映影响整个语言全貌的内在因素。因此他们采用的是多重分类原则：把所有语料都标上五个不同特征的值：(1) 文类 (2) 文体 (3) 语式 (4) 主题 (5) 媒体。利用以主题为主的五个特征的多重分类来进行语料库的平衡。这样做还使研究者能够任选其中几个特征的组合，定义自己的次语料库（sub-corpora），也可以在次语料库间作比较研究。另外，多重分类原则也有利于以后平衡语料库的更新。语料存储格式的规范一般指采用统一的编码规范为电子文本作标记，目前可扩充置标语言 XML 被广泛地用作语料库标注的元语言，存储格式的标准化有助于语料的交换和共享。

四 语料库在语言研究中的的应用

在语言研究中，语料库方法是一种经验的方法，它能提供大量的自然语言材料，有助于研究者根据语言实际得出客观的结论，这种结论同时也是可观测和可验证的。在计算机技术的支持下，语料库方法对语言研究的许多领域产生了越来越多的影响。各种为不同目的而建立的语料库可以应用在词汇、语法、语义、语用、语体研究，社会语言学研究，口语研究，词典编纂，语言教学以及自然语言处理、人工智能、机器翻译、言语识别与合成等领域。我国在语料库的应用上还处于起步阶段，在计算语言学和语言信息处理领域，语料库主要用来为统计语言模型提供语言特征信息和概率数据，在语言研究的其他领域，多使用语料的检索和频率统计结果。

语料库与自然语言信息处理有着相辅相成的关系，大规模的语料库是用统计语言模型方法处理自然语言的基础资源。然而统计语言模型本身并不关心其建模对象的语言学信息，它关心的只是一串符号的同现概率。譬如 N 元语法模型，它只关心句子中各种单元（比如字、词、短语等）近距离连接关系的概率分布，而对于许多复杂的语言现象，它就无能为力了。在统计语言建模技术最先得到成功应用的自动语音识别领域，语料库的开发和建设受到格外的重视，标注语料库成为不可缺少的系统资源，就是因为，要想改进 N 元语法的建模技术，必须利用语料库引入更多的语言特征信息和统计语言数据。同样，在书面语语言信息处理领域里，语料库提供的语言知识也越来越多地用在统计语言模型方法中。除了词语自动切分、词性自动标注、双语语料对齐等语料加工技术以外，人们还在语料库的支持下，建立有关语法、语义的语言知识库，开发信息抽取系统、信息检索系统、文本分类和过滤系统，并且把基于统计或实例的分析技术集成到机器翻译系统里面。

近年来在语料库的支持下，从信息处理的角度研究汉语词汇、语法和语义问题的报告也日渐增多。这些研究包括：根据逐词索引作汉语词义的调查；对词语搭配进行计量分析；利用量词--名词的搭配数据研究汉语名词分类问题；进行现代汉语句型的统计和研究；做短语自动识别（例如基本名词短语、动宾结构）和自动句法分析的试验；研究在句子里为词语排除歧义的算法；分析和统计汉语词语重叠结构的深层结构类型及产生方式；等等。

对于词汇学、语法学、语言理论、历史语言学等研究来说，语料库的作用目前大多还是通过语料检索和频率统计，帮助人们观察和把握语言事实，分析和研究语言的规律。语料库方法的发展会使这种仅起辅助作用的手段逐步变成必备的应用资源和工具。利用语料库，人们可以把指定的语法现象加以量化，并且检测和验证语言理论、规则或假设。

在少数民族语言和方言调查研究方面，比较有代表性的工作是“藏缅语语料库及比较研究的计量描写”。它建立了我国境内藏缅语族五大语支 82 个语言点 16 万词条的词汇语音数据库，对藏语方言的音节、音位、声母、韵母、声词、词素、构词能力和语音结构等 10 余项特征作了分布和对比分析。对藏语 15 个方言点作了语音对应关系和音系对比关系的量化描述，并且在这个基础上做出具有历时和共时比较研究意义的相关分析，得出了语言分类的相关矩阵和聚类分析图表。

在应用语言学领域，词典编纂和语言教学同是语料库的最大受益者。目前已有多部词典在编纂或修订过程中，不同程度地使用语料库或电子文档收集词语数据，用于收词、释义、例句、属性标注等。南京大学近年来开发了 NULEXID 语料库暨双语词典编纂系统，涉及英汉两种语言，在《新时代英汉大词典》的编纂过程中起了重要作用。从词典编纂的整体情况看，我们还缺少充分的语料资源和有效的分析工具，很多有意义的事情还做不了。譬如，分析语料中显现的词语搭配现象，利用语料库进行词语意义辨析，在动态的语料库中辅助提取新词语，等等。把语料库用于语言教学的一个例子是上海交通大学的 JDEST 英语语料库，利用这个语料库，通过语料比较、统计、筛选等方法为中国大学英语教学提供通用词汇和技术词汇的应用信息，为确定大学英语教学大纲的词表提供了可靠的量化依据。这个语料库也在英语语言研究中发挥了作用，支持基于语料库的英语语法的频率特征、语料库驱动的词语搭配等项研究。2003 年，中国学习者英语语料库由上海外语教育出版社正式发行。这个语料库是一个 100 多万词的书面英语语料库，涵盖我国中学生、大学英语 4 级和 6 级、英语专业低年级和高年级的学习内容，并对所有的语料作了语法标注和言语失误标注。根据这个语料库得到了词频排列表、拼写失误表、词目表、词频分布表、语法标注频数表、言语失误表等，还把这些数据与一些英语本族语语料库（如 BROWN，LOB，FROWN，FLOB）进行了某些比较。这个语料库为词典编纂、教材编写和语言测试提供了必要的资源。目前上海交通大学正在建设大学英语学习者口语英语语料库。

在几年来语料库建设和应用的基础上，2003 年国家“973”计划开始支持中文语言资源联盟（Chinese Linguistic Data Consortium，简称 ChineseLDC）的建立。ChineseLDC 是吸收国内高等院校、科研机构和公司参加的开放式语言资源联盟。其目的是建成能代表当今中文信息处理水平的、通用的中文语言信息知识库。ChineseLDC 将建设和收集中文信息处理所需要的各种语言资源，包括词典、语料库、数据、工具等。在建立和收集语言资源的基础上，分发资源，促成统一的标准和规范，推荐给用户，并且针对中文信息处理领域的关键技术建立评测机制，为中文信息处理的基础研究和应用开发提供支持。

几年来在计算语言学和语言信息处理领域的学术会议上，语料库的建设和应用一直是重要论题之一。讨论的重点集中在基于语料库的语言分析方法，以及语料的标注、管理和规范等问题上。语言学家更多关心的是语料库的规划和建设，语料库方法在语言研究和教学中的应用。近年来语言学界也召开有关语料库的专门学术会议，譬如 2001 年由中国社会科学院语言研究所主办、在清华大学召开的语料库语言学与计算语言学研究与实践研讨会（主要讨论了语料库的建设和应用、语言信息处理等问题）；2003 年由上海交通大学等单位主办、在上海交通大学召开的语料库语言学国际研讨会（会议主题是语料库研究与外语教学）。

下面是有关的参考文献以及部分公开发布的语料库的网址（有的互联网网址可能会随时间而有所变动）。

陈小荷等 1996　关于建立大规模汉语树库的设想，《计算机时代的汉语和汉字研究》，罗振声、袁毓林主编，北京：清华大学出版社

冯志伟 2002　中国语料库研究的历史与现状，《汉语语言与计算学报》，Vol.12，Num.1

顾曰国 1998　语料库与语言研究，《当代语言学》，第 1 期

顾曰国 2001　北京地区现场即席话语语料库的取样与代表性问题，《首届中法学术论坛论文集》

黄昌宁等 2002　《语料库语言学》，北京：商务印书馆

黄居仁等 1997　《国语日报量词典》，台北：国语日报社

教育部语言文字应用研究所计算语言学室 2001　信息处理用现代汉语词类标记集规范，《语言文字应用》第 3 期

靳光瑾 2003　谈语料库建设与规范标准问题，《中文信息处理若干重要问题》，徐波等主编，北京：科学出版社

雷秀云等 2001　基于语料库的研究方法及 MD/MF 模型与学术英语语体研究，《当代语言

学》，第 2 期
刘开瑛 2003 基于互联网的多层次汉语语料库构建研究，《中文信息处理若干重要问题》，徐波等主编，北京：科学出版社
刘连元 1996 现代汉语语料库研制，《语言文字应用》第 3 期
卢亚军等 2003 基于大型藏文语料库的藏文字符、部件、音节、词汇频度与通用度统计及其应用研究，《西北民族大学学报（自然科学版）》，第 24 卷，第 2 期
罗振声 1996 清华 TH 语料库的结构、功能与应用，《计算机时代的汉语和汉字研究》，罗振声、袁毓林主编，北京：清华大学出版社
孙茂松等 1997 汉语搭配定量分析初探，《中国语文》，第 1 期
孙茂松等 2001 信息处理用现代汉语分词词表，《语言文字应用》，第 4 期
卫乃兴 2002 基于语料库和语料库驱动的词语搭配研究，《当代语言学》第 2 期
邢红兵 2000 汉语词语重叠结构统计分析，《语言教学与研究》，第 1 期
杨惠中主编 2002 《语料库语言学导论》，上海：上海外语教育出版社
尤 方等 2003 基于语义依存关系的汉语语料库的构建，《中文信息学报》，第 1 期
俞士汶 2002 北京大学现代汉语语料库基本加工规范，《中文信息学报》，第 5，6 期
俞士汶 2003 语料库与综合型语言知识库的建设，《中文信息处理若干重要问题》，徐波等主编，北京：科学出版社
张 普 2003 关于汉语语料库的建设与发展问题的思考，《中文信息处理若干重要问题》，徐波等主编，北京：科学出版社
赵 军等 2003 中文语言资源联盟的建设和发展，《中文信息处理若干重要问题》，徐波等主编，北京：科学出版社
郑玉玲等 1996 藏缅语语料库及比较研究的计量描写，《中文信息学报》，第 2 期
邹嘉彦等 2003 汉语共时语料库与信息开发，《中文信息处理若干重要问题》，徐波等主编，北京：科学出版社

北京大学《人民日报》标注语料库：http://www.icl.pku.edu.cn/
北京语言大学的语料库：http://www.blcu.edu.cn/kych/H.htm
清华大学的汉语均衡语料库 TH-ACorpus：http://www.lits.tsinghua.edu.cn/ainlp/source.htm
山西大学的语料库： http://www.sxu.edu.cn/homepage/cslab/sxuc1.htm
台湾中研院的语料库：
现代汉语平衡语料库：http://www.sinica.edu.tw/SinicaCorpus
或 http://www.sinica.edu.tw/~tibe/2-words/modern-words/
或 http://www.sinica.edu.tw/ftms-bin/kiwi.sh
近代汉语标记语料库：http://www.sinica.edu.tw/Early_Mandarin/
古汉语语料库：http://www.sinica.edu.tw/ftms-bin/ftmsw3
或 http://www.eastasian.ucsb.edu/projects/scriptasinica/cgi-bin/ghy/kiwi.cgi
或 http://www.sinica.edu.tw/~tibe/2-words/old-words/
台湾南岛语典藏：http://www.ling.sinica.edu.tw/Formosan/
闽南语典藏：http://southernmin.sinica.edu.tw/
汉籍电子文献：http://www.sinica.edu.tw/~tdbproj/handy1/
或 http://www.sinica.edu.tw/ftms-bin/ftmsw3
香港城市大学的 LIVAC 共时语料库：http://www.rcl.cityu.edu.hk/livac/
或 http://www.LIVAC.org
浙江师范大学的历史文献语料库：http://lib.zjnu.net.cn/xueke/hyywzx/xkjj.htm
中国科学院计算所的双语语料库：http://mtgroup.ict.ac.cn/corpus/query_process.php
中文语言资源联盟：http://www.chineseldc.org/xyzy.htm