载 Journal of Chinese Language and Computing 11：2，127~136 , 2002
http://cslp.comp.nus.edu.sg/cgi-win/journal/paper.exe

中国语料库研究的历史与现状

冯志伟
教育部语言文字应用研究所
朝内南小街 51 号
100010 北京，中国
e-mail: zwfengde@public.bta.net.cn

2001 年 11 月 25 日提交，2002 年 7 月 25 日修改

摘要

本文首先简要回顾了国外语料库的概况，然后，比较详细地介绍中国语料库的发展情况，包括早期的语料库、国家级语料库、大规模真实文本语料库、口语语料库、双语语料库、少数民族语言语料库等，接着介绍语料库的各种加工技术，如自动切分、自动词类标注、自动短语结构标注、自动双语对齐等，使我们对于语料库研究得到一个鸟瞰式的认识。最后讨论了当前语料库研究中的一些问题，如语料库的规范和标准问题，语言资源共享问题、知识产权问题等。

关键词

语料库；大规模真实文本；口语语料库；双语语料库；少数民族语言语料库；自动切分；自动词类标注；自动短语结构标注；双语对齐

语言学的研究必须以语言事实作为根据，必须详尽地、大量地占有材料，才有可能在理论上得出比较可靠的结论。传统的语言材料的搜集、整理和加工完全是靠手工进行的，这是一种枯燥无味、费力费时的工作。计算机出现后，人们可以把这些工作交给计算机去作，大大地减轻了人们的劳动。后来，在这种工作中逐渐创造了一整套完整的理论和方法，形成了一门新的学科 -- 语料库语言学（corpus linguistics），并成为了自然语言处理的一个分支学科。

语料库语言学主要研究机器可读自然语言文本的采集、存储、检索、统计、语法标注、句法语义分析，以及具有上述功能的语料库在语言定量分析、词典编纂、作品风格分析、自然语言理解和机器翻译等领域中的应用。多年来，机器翻译和自然语言理解的研究中，分析语言的主要方法是句法语义分析。因此，在很长一段时间内，许多系统都是基于规则的，而根据当前计算机的理论和技术的水平很难把语言学的各种事实和理解语言所需的广泛的背景知识用规则的形式充分地表达出来，这样，这些基于规则的机器翻译和自然语言理解系统只能在极其受

限的某些子语言（sub-language）中获得一定的成功。为了摆脱困境，自然语言处理的研究者者们开始对大规模的非受限的自然语言进行调查和统计，以便采用一种基于统计的模型来处理大量的非受限语言。不言而喻，语料库语言学将有可能在大量语言材料的基础上来检验传统的理论语言学基于手工搜集材料的方法所得出的各种结论，从而使我们对于自然语言的各种复杂现象获得更为深刻全面的认识。

本文首先简要介绍国外语料库的发展情况，然后，比较详细地介绍中国语料库的发展情况和主要的成绩，使我们对于语料库研究得到一个鸟瞰式的认识。

1. 国外语料库概况

现在，美国 Brown 大学建立了 BROWN 语料库（布朗语料库），英国 Lancaster 大学与挪威 Oslo 大学与 Bergen 大学联合建立了 LOB 语料库。欧美各国学者利用这两个语料库开展了大规模的研究，其中最引人注目的是对语料库进行语法标注的研究。他们设计了基于规则的自动标注系统 TAGGIT 来给布朗语料库的 100 万词的语料作自动标注，正确率为 77%. 他们还设计了 CLAWS 系统来给 LOB 语料库的 100 万词的语料作自动标注，根据统计信息来建立算法，自动标注正确率达 96%, 比基于规则的 TAGGIT 系统提高了将近 20%. 最近他们同时考察三个相邻标记的同现频率，使自动语法标注的正确率达到 99.5%。这个指标已经超过了人工标注所能达到的最高正确率。

现在，国外的主要语料库还有：

- London-Lund 口语语料库：收篇目 87 篇，每篇 5000 词，共为 43.4 万词，有详细的韵律标注(prosodic marking)。
- AHI 语料库：美国 Heritage 出版社为编纂 Heritage 词典而建立，有 400 万词。
- OTA 牛津文本档案库（Oxford Text Archive）：英国牛津大学计算中心建立，有 10 亿字节。
- BNC 英国国家语料库(British National Corpus)：1995 年正式发布，使用 TEI 编码(Text Encoding Initiative)和 SGML 通用标准置标语言的国际标准(The Standard Generalized Mark up Language, ISO 8879, 1986 年公布)。
- ACL/DCI 美国计算语言学学会数据采集计划：美国计算语言学学会(The association for Computational Linguistics, ACL)倡议的数据采集计划(Data Collection Initiative, DCI)，其宗旨是向非赢利的学术团体提供语料，以免除费用和版权的困扰，用标准通用置标语言 SGML 统一置标，以便于数据交换。
- LDC 语言数据联合会 (Linguistic data Consortium): 设在美国宾州大学，实行会员制，有 163 个语料库 (包括 Text 的以及 speech 的)，共享语言资源。
- RWC 日语语料库：日本新情报处理开发机构 RWCP 研制，包括《每日新闻》4 年的全文语料，语素标注量达 1 亿条。
- 亚洲各语种对译作文语料库：日本国立国语研究所研制，中野洋主持，北京外国语大学参加。

为了推进语料库研究的发展，欧洲成立了 TELRI 和 ELRA 等专门学会。TELRI 是跨欧洲语言资源基础建设学会(Trans-European Language Resources

Infrastructure)的首字母缩写，John Sinclair 担任主席，由欧洲共同体提供经费，其目的在于建立欧洲诸语言的语料库，现已经建成柏拉图(Plato)的《理想国》(Politeia)多语语料库，建立了计算工具和资源的研究文档 TRACTOR (Research Archive of Computational Tools and Resources)，正在语料库的基础上建立欧洲语言词库 EUROVOCA。TELRI 每年召开一次 Seminar 。最近的一次 Seminar 在 Lubljana, (Slovenia)召开(22.September – 26.September.2000)，主题是从语料库中自动抽取知识（Automatic knowledge extraction）。ELRA 是欧洲语言资源学会 (European Language Resources Associationi)的首字母缩写，由 Zampolli 担任主席， ELRA 负责搜集、传播语言资源并使之商品化，对于语言资源的使用提供法律支持。ELRA 建立了欧洲语言资源分布服务处 ELDA (European Language resources Distribution Agency)，负责研制并推行 ELRA 的战略和计划。ELRA 还组织语言资源和评价国际会议 LREC (Language Resources & Evaluation Congress), 每两年一次。第一次会议于 1998 年在西班牙的 Grenade 举行；第二次会议在 Athens（Greece）召开（31.May – 02.June.2000），第三次会议于 2002 年在西班牙的 Las Palmas de Gran Canaria 召开(27.May – 02.June 2002)。

2. 我国语料库的发展概况

2.1 早期的汉语语料库

2.1.1 我国语料库研究的先河

在我国，从 20 世纪 20 年代开始，就有学者建立文本的语料库，采用统计的方法来研究汉字的频率，其目的在于制定基础汉字的字表。当然，这样的语料库不是机器可读的，规模也很小，它是现代语料库的雏形，开我国语料库研究的先河，在我国语料库的发展史上是功不可没功的。著名教育学家陈鹤琴为了教学的目的，在对语料统计的基础上，编写了《语体文应用字汇》，于 1925 年完成，于 1928 年由商务印书馆出版，陈书前有“绪论”，说明“中文应用字汇”曾有多种，其中包括 P.克仑茨(Pastor P. Kronz)的研究和他自己的编写的《常用四千字表》。陈鹤琴做过两次统计，第一次统计使用了六种材料，包含 554,478 个汉字的语料，得不同汉字 4261 个；第二次使用包含 34,818 个汉字的语料，得出与 4261 个汉字相异的汉字 458 个。第二次统计所得的成果毁于战火，在《语体文应用字汇》中印出的只是第一次统计的结果。

陈鹤琴用的语料分如下六类：

1. 儿童用书：127,293 字；
2. 报刊（以通俗报刊为主）：153,344 字；
3. 妇女杂志：90,142 字；
4. 小学生课外作品：51,807 字；
5. 古今小说：71,267 字；
6. 杂类：60,625 字。

书末附有“字数次数对照表”，这是按汉字在语料中出现的绝对频率排列的字表。

我国著名教育家陶行知先生为《语体文应用字汇》写了序言。序言中说:“他们（指“近代教育家”）对于一门一门的功课，甚至一篇文章，一个算题，一项运动，都要依据目标去问他们的效用。他们的主张是要所学的，即是所用的。.....到了后来他们连学生学的字也要审查起来了。学生现在所学的字，个个字都是有用的字吗？自从这个问题发生就有好几位学者开始研究应用字汇。我国方面也有几位先生研究这个问题，其中以陈鹤琴先生的研究最有系统。他和他的助理九人先后费了二三年工夫，检查了几十万字的语体文，编成这本《语体文应用字汇》。这册报告未付印以前已经做了《平民千子课》用字的根据。将来小学课本用字当然也可以拿他来做一个很好的根据。虽然不能十分完备，但我想这本字汇对于成人及国民教育一定是有很大的贡献的。”（见陈鹤琴《语体文应用字汇》，商务印书馆，1928 年）。

2.1.2 早期的机器可读语料库

从 1979 年以来，中国就开始进行机器可读语料库的建设，早期在中国建立的主要的机器可读语料库有：

- 汉语现代文学作品语料库（1979 年），527 万字，武汉大学。
- 现代汉语语料库（1983 年），2000 万字，北京航天航空大学。
- 中学语文教材语料库（1983 年），106 万 8 千字，北京师范大学。
- 现代汉语词频统计语料库（1983 年），182 万字，北京语言学院。

我们以北京语言学院的汉语词频统计语料库来说明早期语料库的情况。

1979 年，北京语言学院（现在改名为“北京语言文化大学”）针对对外汉语教学的特点，把“现代汉语词汇统计研究”作为重点科研课题，开始进行规模较大的汉语单词的频率统计研究。

这项研究工作，采用人工与计算机相结合的方式，对 179 篇样文、182 万字的语料进行了词语切分、词频统计和数据分析的工作，统计的总词汇量为 1,315,752 词次，含不同单词 31,159 个，其中包括十年制语文课本（52 万字，374,654 词次）的字频和词频的定量分析，统计结果编成《现代汉语频率词典》出版。

他们选取的语料可以分为如下四类：

1. 报刊政论：44 万字，占语料总量的 24.4%。
2. 科技和科普文章：29 万字，占语料总量的 19.8%。
3. 口语材料：20 万字，占语料总量的 11.1%。
4. 文学作品：89 万字，占语料总量的 48.7%。

整个语料共 182 万字。这样容量的语料，在当时已经是比较大的语料库了。

根据数理统计的原理，所统计的语料的总体个数必须达到一定足够的数量，才能保证统计结果符合客观实际。《现代汉语频率词典》的编者认为，如果常用词的出现频率不低于百万分之一，也就是在一百万次的场合，常用词的出现机会至少应该有一次，就可以保证统计结果的客观性。《现代汉语频率词典》实际上统计了 182 万个汉字的语料，因此，其抽样是合理的、经济的、适度的。

但是，国外在 1971 年进行英语词频统计时，所用语料量有 5,088,721 个词，

包含不同单词 86,741 个，统计规模比《现代汉语频率词典》大得多。由于语料库语言学的发展，语料库的容量不断扩大，现在，数千万词甚至于数亿词的语料库已经不算少见。与当前语料库的容量比较起来，《现代汉语频率词典》所依据的语料规模是小了一些。不过，尽管这样，《现代汉语频率词典》在词频统计方面取得的成绩仍然是很大的。

这次词频统计得出了如下词表：

1. 按字母音序排列的频率词表：共列出常用词 16,593 个，按音序排列，从中可以看出：

- 汉语中以 Z、S、J、Y 开头的词较多：以 Z 开头的词有 1457 个，占 8.78%；以 S 开头的词有 1327 个，占 7.99%；以 J 开头的词有 1243 个，占 7.49%；以 Y 开头的词有 1205 个，占 7.26%。
- 汉语中以 E、O 开头的词很少：以 E 开头的词只有 64 个，占 0.38%；以 O 开头的词只有 13 个，占 0.07%。

2. 按频率递减的顺序排列的词表：在词表中，最常用词的使用频率相当高，前 100 个词占了语料总量的 40%以上，前 500 个词占了语料总量的 70%以上，前 2562 个词占了语料总量的 85%，词表共有不同单词 31,159 个，这些词占了语料总量的 100%。从前 100 个词到前 500 个词，不同的单词数增加了 400 个，百分比就增加了 30%，而从前 2562 个词到前 31,159 个词，不同单词数增加了 30,597 个，百分比材增加了 15%。由此可见，高频词对于百分比的增加有着很大的作用，而低频词对于百分比的增加，其作用是微乎其微的，往往要大量的低频词，才能使百分比增加一点点。

3. 按使用度递降顺序排列的词表：

使用度是 1954 年尤兰德(Juilland)和洛德西盖(Chang-Rodsiguez)在计算西班牙语的词汇频率时提出的一个新概念,他们并且也提出了计算使用度的数学公式，根据这个使用度公式计算出的使用度，可以综合地反映单词在出现频率和分布率两方面的情况。

他们根据使用度的计算公式，计算了单词的使用度，并给出了按使用度递降顺序排列的词表。这个词表又分为两个表：使用度较高的前 8000 词的词表，使用度较低的词语单位表。

在使用度较高的前 8000 词的词表中，使用度在 20 以上的词共 4186 个，其词次累计占了全部语料（314,404 词次）的 90.1%。这说明，《现代汉语频率词典》所统计的语料中，有十分之九是用这 4186 个词写成的，这些词可以成为“常用词”的候选对象。

在使用度较低的词语单位表中，收入了使用度为 5 及小于 5 的词 22,446 个，这些词一般也都是低频词。在这种情况下，如果有的词的使用度和频率相匹配，则说明这些词的分布还是比较均匀的，这些词可以作为“通用词”的候选对象。

4. 按语体分类的高频词表，又可再分为 4 个表：

a. 报刊政论语体的前 4000 词的词表：本表共统计 34 种语料，29 万词次(44 万字)，有不同词条数 12,107 个。前 4000 个词累计频率 94.77%。其中一些政治词语，如“唯心、党派”等，在本表中出现频率都比较高，反映了政论语体的特点。

b. 科普语体的前 4000 词的词表：本表共统计 21 种语料，20 万词次(29 万字)，有不同词条 12,364 个。前 4000 个词累计频率 92.27%。其中一些科技用语，如“纤维、合成”等，在本表中出现频率都比较高，反映了科普语体的特点。

c. 生活口语中前 4000 词的词表：本表共统计 18 种语料，16 万词次(20 万字)，有不同词条 8263 个。前 4000 个词的累计频率为 96.65%。从统计数字可以看出，口语语体的用词量比前两种语体要少三分之一，但高频词出现的词次却相当多，前 1000 个高频词的出现频率比 a 表高出 6%，比 b 表高出 12%。这意味着，口语语体的用词量虽然不大，但是它们的出现次数对语料的覆盖面却相当大。

d. 文学作品类前 4000 高频词的词表：本表共统计 106 种语料，66 万词次(89 万字)，有不同词条 23,622 个。前 4000 个高频词累计频率为 90.63%。这说明文学作品的用词量大，但是为了追求用词的多样化，即使是高频词的出现频率也比较低，这反映了文学作品词汇丰富多采的特点。

早期的这些语料库的具有如下特点：

① 多数是采用手工键入的方式建立的，耗时耗力，缺乏规范，规模较小，重用性差。为了建设这样的语料库，需要付出艰辛的劳动，著名专家刘源教授（北京航空航天大学计算机系教授）在 2000 万字的语料库建设中积劳成疾，健康受到严重的损害。我国语料库的早期建设者的敬业精神是值得我们尊敬的。

② 发现了汉语文本切分歧义的两种类型：北航和北语的语料库进行了词频统计，北航还进行了自动分词研究，发现了两种不同的分词歧义字段（Ambiguous Segmentation Strings, ASSs）：交集型歧义字段和多义组合型歧义字段。

- 交集型歧义切分字段：例如：“地面积”可能切为“地面”或“面积”，“面”成为交段，从而产生歧义。
- 多义组合型歧义切分字段：例如：“马上”本身是一个词，但也可以切为“马”+“上”两个单词，而“马上”与“马”+“上”的含义不同。

梁南元（1987）对一个 48092 字的自然科学、社会科学样本进行了统计：交集型切分歧义 518 个，多义组合型切分歧义 42 个。据此推断，中文文本中切分歧义的出现频度约为 1.2 次/100 字，交集型切分歧义与多义组合型切分歧义的出现比例约为 12:1。

③ 建立了初步的分词规范：1990 年 10 月，在计算机界和语言学界的共同努力下，我国制定了国家标准 GB-13715《信息处理用现代汉语分词规范》，这个国家标准提出了确定汉语单词切分的原则，是汉语书面语自动切词的重要依据。

2.2 国家级语料库的建设

1991 年，国家语言文字工作委员会开始建立国家级的大型汉语语料库，以推进汉语的词法、句法、语义和语用的研究，同时也为中文信息处理的研究提供语言资源，计划其规模将达 7000 万汉字，当时宣称，这将成为世界上最大的汉语语

料库。这个语料库是均衡语料库。其语料要经过精心的选材，语料的选材应受到如下限制：

① 时间的限制：语料描述具有历时特征，着重描述共时特征。选取从 1919 年到当代的语料（分为 5 个时期），以 1977 年以后的语料为主。
② 文化的限制：主要选取受过中等文化教育的普通人能理解的语料。
③ 使用领域的限制： 语料由人文与社会科学类、自然科学类和综合类 3 大部分，人文和社会科学再分为 8 大类 29 小类，自然科学再分为 6 大类，综合类再分为 2 大类。主要选取通用的语料，优先选取社会科学和人文科学的语料。

这个语料库现在只完成了 2000 万字语料的输入和校对工作，尚未进行进一步的加工，还是“生语料库”，因而还不能提供社会使用。由于主要靠手工录入，人工劳动的成本很高，据说单是建立生语料库，耗资约 200 万人民币。

为了加工这个国家级语料库，国家社科基金设立了社科重大项目“信息处理用现代汉语词汇研究”，希望利用该项目的成果来加工这个语料库。该课题分 10 个子课题：

① 信息处理用现代汉语分词词表
② 歧义切分与专有名词识别软件
③ 词的构造研究
④ 现代汉语词类及标记集规范
⑤ 汉语词类兼类研究
⑥ 现代汉语的语法属性描述研究
⑦ 现代汉语述语动词机器词典和槽关系研究
⑧ 汉语知识词典建立及词汇内部语义网络描述研究
⑨ 汉语文本短语结构的人工标注
⑩ 常用动词语义特征及词义搭配研究

现在，该课题已经结项，国家语委语言文字应用研究所成立了“汉语语料库深加工”的课题组，准备对国家级语料库的 2000 万字的核心语料进行深加工，逐步把这个生语料库变为熟语料库。

2.3 大规模真实文本语料库

1992 年以来，大量的语料库在中国研究中文信息处理的单位建立起来，语料库成为了研究中文信息处理的基本语言资源。没有语料库的支持，中文信息处理的研究将会寸步难行。建设大规模真实文本语料库的单位有：

- 《人民日报》光盘数据库
- 北京大学计算语言学研究所
- 北京语言文化大学
- 清华大学
- 山西大学
- 上海师范大学
- 北京邮电大学

- 香港城市理工大学
- 东北大学
- 哈尔滨工业大学
- 中国科学院软件研究所
- 中国科学院自动化所
- 北京外国语大学日本学研究中心
- 台湾中央研究院语言研究所（筹备处）

下面分别加以介绍。

2.3.1 《人民日报》光盘数据库

收集该报 48 年的全部文字和图像内容，公开发行。

2.3.2 北京大学计算语言学研究所

该研究所建立了现代汉语标注语料库，与富士通公司（Fujitsu）合作，加工 2700 万字的《人民日报》语料库，加工项目包括词语切分、词性标注、专有名词（专有名词短语）标注。还要对多音词注音。

示例 1：古城/n 虽/c 遭/v 破坏/v ，/w 但/c 它/r 留下/v 了[le5]/u 契丹族/nz 和[he2] 各[ge4]/r 民族/n ，/w 特别/d 是/v 汉族/nz 劳动/vn 人民/n 共同/d 开拓/v 祖国/n 北疆/s ，/w 创造/v 我国/r 历史/n 文明/n 的[de5]/u 足迹/n 。/w
示例 2：19970310-01-002-0020/m [全国/n 人大/j]nt 代表/n 、/w [陕西/ns 西安/ns 美术/n 学院/n]nt 名誉/n 院长/n 刘/nr 文西/nr 利用/v 会议/n 休息/vn 时间/n 创作/v 了/u 邓/nr 小平/nr 画像/n 《/w 与/p 人民/n 同/d 在/v 》/w 。/w 画像/n 表现/v 了/u 邓/nr 小平/nr 同志/n 祝愿/v 祖国/n 繁荣/a 、/w 人民/n 幸福/a 的/u 伟大/a 胸怀/n 。/w （/w 新华社/nt 记者/n 齐/nr 铁砚/nr 摄/v ）/w
示例 3：19970310-01-003-0020/m 世纪/n 之/u 交/Ng ，/w 中华/nz 民族/n 正/d 迎来/v 前所未有/i 的/u 发展/vn 机遇/n 。/w 十几/m 年/q 来/f ，/w 改革/v 开放/v 的/u 不断/d 深入/v ，/w 党/n 的/u 民族/n 政策/n 的/u 贯彻/vn 落实/vn ，/w 全国/n 人民/n 的/u 大力/d 支援/v ，/w 使/v 我国/r 民族/n 地区/n 经济/n 和/c 社会/n 发展/vn 步伐/n 大大/d 加快/v 。/w 加倍/d 珍视/v 和/c 继续/vd 发展/v 这种/r 好/a 的/u 局面/n ，/w 巩固/v 发展/v 各/r 民族/n 大/a 团结/an ，/w 成为/v 全国/n 各族/r 人民/n 的/u 共同/b 愿望/n ，/w 也是/v 在/p 京/j 参加/v “/w 两会/j ”/w 的/u 代表/n 和/c 委员/n 的/u 一致/a 心愿/n 。/w

经富士通公司检验，标注的正确率很高。

他们制订《现代汉语语料库加工手册——词语切分与词性标注》。切分规范中，主要规定现代汉语的切词原则，即什么样的汉字组合可以为一个切分单位。他们

采用切分和标注相结合的原则来建立规范 ，在汉语中，像“双音节动词＋单音节名词”通常构成新的名词，对于这个新的名词，即使在词典中没有登录，也应该把它们处理为一个切分单位。因此，在该规范中，给出了一些基于词性描述的构词规律，规定了什么样的组合可以处理为一个切分单位，并给出了新组合的词的词性标记。在标注规范中，规定了一般词性的标注规范和专有名词的规范。

此外，他们还建立了一个小型汉语树库：与新加坡国立大学计算机系合作，内容为新加坡中学语文教材（1995 年），所有的句子都分析为树形图。

示例：

[zj [dj 富士山/n [vp 是/v [np 日本/n 的/u [np [mp 一/m 座/q] 活火山/n]]]] 。/w]

[zj [fj [fj [dj 山峰/n [vp 终年/d 积雪/v]] ，/w [dj 云雾/n 围绕/v]] ，/w [vp 只有/d [vp [pp 在/p [np [dj 空气/n 干燥/a] 的/u [np [np 秋/n 冬/n] [np 两/m 季/Ng]]]] ，/w [vp 才/d [vp 能/v [vp [vbar 看/v 清/a] [np 它/r 的/u 全貌/n]]]]]] 。/w]

[zj [fj [dj [np [vbar 多/d 变/v] 的/u 气候/n] ，/w [vp 更/d [vp [pp 为/p 它/r] [vp [vbar 增添/v 了/u] [np 神秘/a 的/u 色彩/n]]]] ，/w [vp 甚至/d [vp 使/v 它/r [vp [vbar 孕育/v 了/u] [np 许多/m [np 美丽/a 的/u 神话/n]]]]] 。/w]

[zj [dj [np 富士山/n 的/u 景色/n] ，/w [dj 四季/t 不同/a]] 。/w]

[zj [fj [fj [fj [fj 春天/t ，/w [fj [dj 山顶/s [vp 还/d [vp [vbar 戴/v 着/u] [np 雪/n 帽子/n]]]] ，/w [fj [dj [dj [np 山腰/n 的/u 雪/n] [vp 却/d 溶化/v]] 了/y] ，/w [fj [dj [np 细碎/a 的/u [np 小/a 花/n]] [vp 开遍/v 山坡/n]] ，/w [vp [vbar 远/a 看/v] [vp 象/v [np [mp 一/m 片/q] [np 紫色/n 的/u 海洋/n]]]]]]] ，/w [fj 夏天/t ，/w [fj [dj [np [np 残/Vg 雪/n] 与/c [np 山/n 花/n]] [vp 倒映/v [sp 湖/n 中/f]]] ，/w [vp 充满/v 诗情画意/n]]]] ，/w [fj 秋天/t ，/w [fj [dj [np [np [np 满/a 山/n] 红叶/n] 与/c [np 雪/n 影/Ng]] 辉映/v] ，/w [vp 象/v [np 个/q [np 娇羞/a 的/u 姑娘/n]]]]] ，/w [fj [dj 冬天/t [dj 则/c [vp 是/v [np [ap 纯/a 白/a] 的/u [mp 一/m 片/q]]]] ，/w [ap 庄严/a 而/c 圣洁/a]]] 。/w]

北大语料库研究的特色是：

① 规模大：加工成的熟语料已经达到 2000 万字，不久将达到 2700 万字，国内尚无先例。

② 加工深：不仅做了切分和词性标注，而且部分语料还进行了短语结构分析，建立了树库。在大规模的语料库中，地名和专有名词都进行了短语结构标注。

③ 覆盖面广：人民日报的语料不仅包括新闻，还包括各种题材、各种风格、各种语体的文章，涉及社会科学和自然科学多种领域，有很广泛的覆盖面。

④ 正确率高：在自动加工的基础上进行了大量的人工加工，采用人机结合

的策略，是语料库加工的正确率达到了国内最高水平，在国际上也是罕见的。

⑤ 无著作权纠纷：与《人民日报》达成协议，没有著作权问题。

2.3.3 北京语言文化大学

该校计算机系宋柔在远景校对系统的研究、开发和测试过程中一直注重采用大规模真实语料进行各种语言现象的统计、分析、检索、归纳。为此，他们与一些报社、出版社合作，收集、整理了一批综合性、规范性的电子文档资料，建立了一个大型的中文语料库（共约 5 亿字）。在获取语料后，又专门用工具软件或人工加工清理了语料，分别建立了 10 个语料库。各语料库情况如下：

- ✧ 《当代中国 丛书》：150 卷（约 6 千万汉字）
- ✧ 《中华人民共和国年鉴》：1997 年语料（约 200 万汉字）
- ✧ 《新闻出版报》：1988 年语料（约 260 万汉字）
- ✧ 《辉煌五十年 湖南卷》：1949-1999 年语料（约 70 万汉字）
- ✧ 《人民日报》：1993-2000 年七年语料（约 2 亿字）
- ✧ 《人民日报 市场报》：2000 年语料（约 1400 万汉字）
- ✧ 《人民日报 华南新闻：2000 年语料（约 600 万汉字）
- ✧ 《人民日报 华东新闻》：2000 年语料（约 500 万汉字）
- ✧ 《经济日报》：1992 年语料（约 1820 万字）
- ✧ 《新华社》：1994-1996 年三年语料（约 3793 万字）

宋柔还建立了面向语言学研究的汉语语料库检索系统 CCRL，可以让用户使用自己的生语料库和词典生成语料索引，进行检索。

此外，北京语言文化大学还建立了如下的语料库：

- ■ 当代北京口语语料库（1992 年）
- ■ 现代汉语语法研究语料库（1995 年）
- ■ 现代汉语句型语料库（1995 年）
- ■ 现代汉语语料库（1998 年，与香港理工大学中文及双语学系联合建立）
- ■ 现代汉语语料库（1998 年，与清华大学联合，为国家自然科学基金重点项目“语料库语言学研究的理论、方法和工具”而建立）

2.3.4 清华大学

该大学也建立了现代汉语语料库：1998 年建立了 1 亿汉字的语料库，着重研究歧义切分问题。现在生语料库已达 7-8 亿字。

他们对于分词技术进行了深入研究，发现了伪歧义，提高了分词精度: 计算机系孙茂松、左正平（1998）指出，切分歧义应进一步区别“真切分歧义”和“伪切分歧义”。譬如：同属交集型，“地面积”为真歧义（“这几块 | 地 | 面积 | 还真不小”“地面 | 积 | 了厚厚的雪”），“和软件”则为伪歧义（虽然存在两种不同

的切分形式“和软 | 件”和“和软 | 件”，但在真实文本中，无一例外地应被切分为“和 | 软件”)；同属组合型，“把手”为真歧义，“平淡”则为伪歧义。

他们还编制了信息处理用现代汉语分词词表，作为分词最重要的语言资源。

中文系罗振声建立了现代汉语句型研究语料库，从中总结出 209 种汉语句型。

清华大学智能技术与系统国家重点实验室与北京语言文化大学语言信息处理研究所联合研发的人工标注语料库 HuaYu。这个语料库区别于其它类似语料库的特点:是：分布平衡，不仅仅限于新闻报纸。

HuaYu 的分布见表 1：

分类	篇数	汉字数	比例	标点符号数	词次数	比例
文学	295	880,057	44%	148,453	760,337	48%
新闻	376	600,490	30%	86,163	438,095	28%
学术	29	402,623	20%	52,823	278,728	18%
应用文	258	119,488	6%	28,727	91,929	6%
合计	958	2,002,658	100%	316,116	1,569,089	100%

表 1　Hua Yu 语料库的分布

其中文学语料的分布见表 2：

分类	篇数	汉字数	百分比	标点符号数	词次数
小说	199	648,796	32.5%	112,749	566,730
散文	37	80,067	4%	10,347	65,453
回忆录	29	50,401	2.5%	6,908	38,338
报告文学	13	50,019	2.5%	8,225	40,386
剧本	17	50,774	2.5%	10,224	49,430
合计	295	880,057	44%	148,453	760,337

表 2　文学语料的分布

他们对这个语料库进行了切分和标注。

语料示例如下：

我|rn 认识|vgn 王眉|npc 的|usd 时候|ng ，|， 她|rn 十|mw 三|mx 岁|qnm ，|， 我|rn 二|mx 十|mw 岁|qnm 。|。 那时|t 我|rn 正|dr 在|pza 海军|ng 服役|vgi ，|， 是|vi 一|mx 条|qns 扫雷舰|ng 上|f 的|usd 三七|ng 炮手|ng 。|。 她|rn 呢|y ，|， 是|vi 个|qng 来|vgn 姥姥|ng 家|ng 度假|vgi 的|usd 中学生|ng 。|。 那|rn 年|qt 初夏|t ，|， 我们|rn 载|vgn 着|utz 海军|ng 学校|ng 的|usd 学员|ng 沿|pg 漫长|a 海岸线|ng 进行|vf 了|utl 一|mx 次|qv 远航|vgx 。|。 到达|vgn 了|utl 北方|s 著名|a 良港|ng 兼|vgn 避暑|vgp 胜地|ng ，|， 在|pza 港|ng 外|f 和|pg 一|mx 条|qns 从|pg 南方|s 驶来|vgi 满载|vgn 度假者|ng 的|usd 白色|ng 客轮|ng 并行|vgi 了|utl 一|mx 段|qns 时间|ng 。|。 进|vgn 港|ng 时|ng 我|rn 舰|ng 超越|vgn 了|utl 客轮|ng ，|， 很|dd 亲近|a 的|usd 擦|vgn 舷|ng 而|c 过|vgi 。|。 兴奋|a 的|usd 旅游者|ng 们|ki 纷纷|dr 从|pg 客舱|ng 出来|vgi ，|， 挤|vgi 满|a 边舷|ng ，|， 向|pg 我们|rn 挥|vgn 手|ng 呼喊|vgi ，|， 我们|rn 也|dr 向|pg 他们|rn 挥|vgn 手|ng 致

意|vgi 。|。

清华大学智能技术与系统国家重点实验室最近对 HuaYu 完成了语块库人工标注（200 万汉字左右），并根据句子中动词类型和句子长度等因素，从中随机选取了 20 万词，进行语块标注，并进一步完成了完整的句法树标注。
语块标注示例如下：
[从/p [他/rN 的/u 身上/s] ，/， [我们/rN [看/v 到/vB] 了/u] [一/m 位/qN] [跨越/v [中国/nS {CS 近代/t 和/c 现代/t } 的/u [共产主义/n 战士/n] [光彩照人/iV 的/u 楷模/n] 。/。]
句法树标注示例如下:
[zj [dj [pp 从/p [sp 他/rN 的/u 身上/s]] ,/， [dj 我们/rN [vp [vp [vp 看/v 到/vB] 了/u] [np [np [mp 一/m 位/qN] [np [vp 跨越/v [tp 中国/nS [tp 近代/t 和/c 现代/t]]] 的/u [np 共产主义/n 战士/n]]] [np 光彩照人/iV 的/u 楷模/n]]]]] 。/。

2.3.5 山西大学

他们根据不同的需要，建立如下不同的语料库，主要有：

- 汉语新闻语料库（1988 年），250 万字，山西大学，包括 4 部分：《人民日报》：150 万字，《北京科技报》：20 万字；《电视新闻》（CCTV）：50 万字；《当代》（杂志）：30 万字。
- 标准语料库（2000 年）：.以分词规范和语委词类标记集为标准的综合语料：70 万字，其中 10 万字标到了小类。
- 特定加工的语料：
  -- 标注中国地名:标注了 280 万字的语料库(含不同地名 1793 个，共出现 11590 次)，建立中国地名库，共收 88026 个地名，同时建立中国地名用字库（3685 个）和中国地名用词库（3917 个）。地名识别正确率为 86.7%，召回率为 93.8%。
  -- 标注中国人名：标注了 300 万字的语料库，建立了姓氏人名库,得出了姓氏用字频率表(729 个)和名字用字频率表(3345 个)，中国人名识别的正确率为 87.31%，召回率为 95.23%。
  -- 标注西文译名：标注了 250 万字的语料库，西文译名资源来自《英语姓名译名手册》和《新英汉词典》的《常见英语姓名表》，正确率为 63%，召回率为 98%。。
  -- 标注中文组织机构名：标注了 50 万字的语料库，对组织机构名的结构进行分析，使用基于规则的方法。
  -- 标注网上语料新词语：标注了 150 万字的语料库。

为了标注交集型歧义字段，建立了 7.8 万字的交集型歧义字段库，已标注 510 万字语料，分词正确率达 97%以上，同时证实了交集型歧义字段只有惟一可能的正确切分结果的伪歧义，占歧义字段总数的 94%。
他们还建立多义组合型歧义字段库：收录了 133 条多义组合型歧义字段，比

较全面的反映了组合型歧义字段的实际情况。

2.3.6 上海师范大学

该校建立了 3000 万字的生语料库；根据北大的标注规范建立了 300 万字的标注语料库。他们还建立了 100 万字《作家文摘》的标注语料库，选取 1997 年的《作家文摘》，题材包括传记文学、历史故事、记实文学、人物特写、小说、散文、评论等，依靠手工进行标注，不仅完成了切词和词性标注，还完成了短语结构关系和结构功能的标注。加工层次深。

标注示例：

[zw 他/rp [db[zc 期望/vz 着/ut]vp[db 打/vs [dz[sl 一/mx 个/qi]mp[dz[zc 漂亮/ax 的/us]np[dz 大/ax 胜战/ng]np]np]vp]vp]jp 。/w

其中的 zw （主谓结构）、db（动宾结构）、 dz（定中结构）、sl（数量结构）等都是结构功能的标记。

2.3.7 北京邮电大学的树库

他们在美国LDC的汉语句法树库的基础上进行自动获取语法规则的研究。LDC的树库包含新华社1994到1998年的325篇文章，包含4185颗树，10万个词。他们对LDC树库进行了改造，语法规则和分析模型参数都是通过LDC树库统计和训练得到。在抽取规则之前，进行了如下的预处理工作：

- 删除所有空的单词；
- 去掉所有的非终结符的功能标记；
- 去掉哪些只有一个孩子结点，且此孩子结点是非终结符的结点。

在此基础上进行规则的自动获取，采用改进的CYK算法自动获取了3690条规则，形式如下：

$parent\_symbol|current\_symbol \rightarrow RHS_1 \ldots RHS_n \quad log\_probability$

比如：*NP|NP -> NN NN NN -0.879602*

2.3.8 哈尔滨工业大学机器翻译实验室(MT-Lab)的汉语语料库

容量约 1GB。

2.3.9 香港城市理工大学的对比语料库

该大学语言资讯科学研究中心建立了 LIVAC(Linguistic variety in Chinese communities)语料库，其宗旨在于研究使用中文的各个地区使用语言的异同。这个语料库从 1993 年开始策划，在香港、澳门、上海、新加坡和台湾五个不同的地区，每日选定一天的报纸摘录其部分资料入库，资料的内容包括社论、第一版的全部新闻和文章、国际版、地方版、特写、评论等。每天收集的份量约两万字，如果已经达到两万字，不太重要的资料就只好割爱。从 1995 年 7 月到 1997 年 6

月的两年内，该语料库所收集的资料总字数为 15,234,551 字,经过自动切词和人工校对之后总词数约为 8,869,900 词。

统计结果表明，中文各地区所使用的词语，以双音节为最多，其次是三音节，再其次是四音节，再再其次是单音节，但是，单音节词语的使用频度却比较高，仅次于双音节词语的频度，而且远远超出其他音节词语频度之总合。

统计结果还表明，香港和澳门的用词相同率最高，香港与台湾、香港与新加坡的用词相同率居第二，香港与上海的用词相同率最低。从历史背景和社会情况来看，这个数字是可以接受的。因为香港与澳门距离很近，又都长期被欧洲国家管制，香港与台湾和新加坡的商务情况和社会结构之间的相同点都比香港与上海之间多，这种情况，在词语中必定会反映出来。

统计结果还表明，新加坡所用词语比较少，而上海的特有词语比较多，这似乎可以从新加坡华语并非当地社会生活的唯一语言，而上海在中国的特殊地位和经济活动非常活跃有关。

2.3.10 台湾的语料库

台湾建立了平衡语料库（Sinica Corpus，中央研究院）和树图语料库（Sinica Treebank，中央研究院）。两个都是标记语料库，有一定加工深度。语料库规模约 500 万字。

2.4 口语语料库

2.4.1 中国社会科学院语言所

他们建立了现代自然口语语料库，包括一个旅馆预定口语语料库，搜集了 2 小时电话的对话，对话人数 200 人以上，进行韵律切分和句法标注，是 wav 文件，用 SAMPA-C 标音，C-ToBI 2.0 标注韵律，并转写成汉字文本；还包括一个无限制的自然对话语料库：14.2 小时的对话，对话人数 22 人，进行韵律切分和句法标注，是 wav 文件，用 SAMPA-C 标音，C-ToBI 2.0 标注韵律，并转写成汉字文本。

语言所还正在建立现代汉语方言自然口语语料库，设计了 1500 种引导话题和多种采集自然口语的交际环境，其中，采用话题引导的方式采集的话题语料占 60%，在说话人不知道的情况下现场采集的口语语料占 40%。

2.4.2 中国科学院自动化所

该所建立了一个旅游咨询口语对话语料库和一个旅馆预定口语对话语料库，可以用于限定领域的口语理解模型、口语对话管理模型、基于统计的口语翻译技术等研究。

2.5 双语语料库的建设

2.5.1 英汉双语语料库

- 北大计算语言学研究所的双语语料库，英汉对齐的句子已有 5 万多对，并开发了相应的对齐工具和双语语料库管理软件。正在此基础上做汉英对照短语库，预计规模将达数十万条。
- 哈尔滨工业大学的英汉双语语料库：1998 年有 3 万句子对，已经进行了词性标注，正在扩充为 40-50 万句子对，在句子、短语、词汇三级实现双语对齐。
- 东北大学的英汉双语语段库：在双语语料库基础上，建造双语语段库，1999 年构造了 10 万双语语段库，进行了基于语段的英汉机器翻译实验，正在以“机获人校”的办法建造 100 万双语语段库，拟扩充到 500 万双语语段库，进一步建造具有 1000 万语段的大容量网上英汉语段电子词典，研究电子词典中搭配短语获取算法，建造大容量网上电子英汉搭配词典。
- 外语教学与研究出版社：
  -- 英汉文学作品语料库
  -- 冯友兰《中国哲学史》汉英对照语料库
  -- 李约瑟(Joself Needham)《中国科学技术史》英汉对照语料库
- 国家语言文字工作委员会语言文字应用研究所建立了英汉双语语料库，其中包括一个计算机专业的双语语料库和一个柏拉图(Plato)哲学名著《理想国》（Politeia）的双语语料库。在这些双语语料库上，他们进行了汉字极限熵的测定和双语对齐的研究。
- 中国科学院软件研究所的英汉双语语料库：进行双语对齐算法研究。现有 15 万对英汉双语对齐句子库，已经切分和标注。
- 中国科学院自动化研究所的英汉双语语料库：购买 LDC 香港新闻英汉双语对齐语料 36294 段以及香港法律英汉双语对齐语料 31 万句子对，并从英汉双解词典中摘取例句 25000 个句子对。

2.5.2 日汉对译语料库

北京外国语大学的北京日本学研究中心建立汉语和日语并行语料库，内容以中日文学名著为主，兼收剧本、散文、政论文，原文和译文全文收录，部分名著收入多个译本。2000 万字。进行自动切分和词性标注，部分文本进行语法和语义标注，采用 SGML 国际标准。

2.5.3 德汉双语语料库

山东海洋大学语言文学院研制的《蝴蝶》（王蒙小说）德汉对照语料库，用于德汉翻译对比研究，完全采用手工方式排比语料，主要比较了汉语的“了”与德语动词完成式的关系。规模很小。

2.5.4 汉日英分类熟语料库

复旦大学计算机系建立了容量为 1GB 汉日英分类熟语料库，包含数千个类别，数十万篇文章。

2.6 少数民族语言语料库

2.6.1 维吾尔语语料库

新疆师范大学建立了 200 万词的维吾尔语语料库，拟发展到 300 万词。

2.6.2 藏语语料库

中国社会科学院民族研究所建立了 500 万藏语字符的藏语语料库，拟进行切分和标注的研究。

2.6.3 蒙古语语料库

内蒙古大学建立了蒙古语语料库，进行了初步的切分和标注。

3 语料库的加工技术

3.1 自动切分

在自动切分方面，提出的切分歧义技术有：“松弛法”（Fan C.K., Tsai W.H. 1988），“扩充转移网络”（黄祥喜 1989），“短语结构文法”（梁南元 1990；姚天顺、张桂平等 1990；Yeh C.L., Lee H.J. 1991；韩世欣、王开铸 1992），“专家系统”方法（徐辉、何克抗等 1991），“神经网络”方法（徐秉铮、詹剑等，1993），“有限状态自动机”方法（Sproat R., Shih C.L *et al*. 1996），“隐 Markov 模型”（Lai B.Y., Sun M.S. *et al* 1997；沈达阳、孙茂松等 1997a；孙茂松、左正平等 1999），“Brill 式转换法”（Palmer D.D. 1997）等。

此外还研究了人名识别技术、地名识别技术、机构名识别技术、新词语识别技术。

3.2 自动标注

在自动标注方面，基于规则的方法主要解决标注中的兼类词问题；基于统计的方法主要有 CLAWS 算法、VOLSUNGA 算法、HMM（隐马尔可夫模型）、TBED 法（Transformation-Based Error-Driven，Eric Brill 于 1993 年提出的方法）

3.3 自动短语结构标注

短语结构标注的结果，可以用短语结构语法树 (P-Tree)来表示，也可以用依存树 （D-Tree）来表示，有的系统采用了从 P-Tree 到 D-Tree 的转换技术，有的系统采用 CYK 算法进行短语结构分析。

3.4 双语对齐技术

主要采用基于长度的方法、基于词典的方法以及把这两种方法结合起来的混合方法。

4 语料库建设中的若干问题

4.1 语料库的规范与标准

我国中文信息界从 1988 年开始研制《信息处理用现代汉语分词规范》的国家标准，根据科学性、严谨性、稳定性、通用性、实用性和完整性（规范对现代汉语语言现象的覆盖率应该达到 99%以上），经过三年时间的研究，七易其稿，于 1992 年批准为国家标准，标准号为 GB/T13715-92。这个规范的主体结构分为主题内容与适用范围、引用标准、术语、概述和具体说明五个部分。由于汉语中语素、单词和词组的界限不够清晰，分词规范中除了基本上采用了《暂拟汉语教学语法系统》中词的定义，把词定义为“最小的独立运用的语言单位”之外，还特别地提出了“分词单位”的概念，把“分词单位”定义为“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”，并且指出，分词单位“包括本规范限定的词和词组”。“分词单位”的提出，巧妙地避开了关于词的定义的争论，协调了当时学术界的矛盾。

我国还研制了《信息处理用现代汉语常用词表》。由于汉语语言现象的极端复杂性，几乎每条规则都会出现例外，因此，分词规范提出了“结合紧密，使用稳定”的原则作为判定一个符号串是否可以作为分词单位的准则。但是，这个原则不够具体，实行起来往往见仁见智，从而造成不同系统中分词单位的不一致。所以，后来有的学者建议在规范之外，还应该根据规范提出一个词表来作进一步具体的说明，以利规范的实施。采用“规范+词表”的策略，这是很有远见的做法。1994 年，该规范的主要制定者刘源教授等人根据现代汉语词频统计的结果，公布了一个《信息处理用现代汉语常用词表》，收词 43570 条，可惜，这个词表对于规范中的一些难点，仍然没有作出很好的处理，权威性不够。

台湾研制了一个《资讯处理用中文分词规范》。台湾的计算语言学会在 1995 年提出了《资讯处理用中文分词规范》，这个规范提出三条基本原则：1.分词单位必须符合语言学理论的要求；2. 在信息处理上确实可行；3. 能确保真实文本处理的一致性。另外还制定了一些辅助原则（合并原则、切分原则），以决定合并还是切分。该规范按照分词的难易程度，把分词规范分为信、达、雅三个不同的等级。信级标准是基本资料交换的标准；达级标准是机器翻译、情报检索等自然语言处理的标准；雅级标准则是分词的理想境界。这种分等级的做法有利于处理难易程度不同的分词作业。

我国还研制了《信息处理用现代汉语规范词表》。国家语言文字工作委员会在 1995 年提出研制《信息处理用现代汉语规范词表》的任务，目的在于从政府的角度，研制规范的现代汉语通用词表，以便作为大家遵循的、统一的通用词表，词表的规模大约 6 万至 8 万条，这项工作还未完成。现代汉语的词汇是一个复杂的

体系，除了通用词之外，还有术语、方言词语、文言词语、专名词语（包括人名、地名、机构名等）、各种熟语（包括成语、惯用语、歇后语、谚语、格言等）。词语是不断发展变化的，随着社会的发展，还会出现大量的新词新语。信息处理会涉及到上述各种词语，因此，词表的制定，除了制定通用词表之外，还应该制订不同专业的术语词表、方言词表、文言词表、专名词表、熟语词表、新词语词表等。这是一项庞大的工程。这些词表的制订和规范化，对于我国计算语言学的进一步发展有着深远的影响。

我国还研制了《信息处理用现代汉语词类标记集规范》：该规范由教育部语言文字应用研究所计算语言学研究室研制，包括 18 个大类。信息处理用现代汉语词类标记集规范制订的主要原则有三个：① 语法功能原则。语法功能是词类划分的主要依据。词的意义不作为划分词类的主要依据，但有时也起某些参考作用。② 允许有兼类。根据各种统计研究，现代汉语的某些词具有多种语法功能，但这多种功能的分布概率不同。在信息处理用现代汉语词类体系中，各词类的确立要根据词的主要语法功能。 ③词类标记集中的大类应能覆盖现代汉语的全部词。这个规范正在考虑提升为国家标准。

我国某些学者关注到国际上关于通用置标语言的进展。由计算机和人文科学学会（ACH）、计算语言学学会(ACL)和文学与语言计算学会(ALLC)联合提出了 TEI（Text Encoding Initiative，文本编码倡议，1998 年），其目标是为电子文本制订一套统一的编码规范，以推动语料存储格式的标准化，实现语料的交换和共享。

由欧洲 MULTEXT、EAGLES 和 VASSAR/CNRS collaboration 联合提出 CES(Corpus Encoding Standard，语料库编码标准)，可广泛应用于语料库的研制与开发。

1986 年 ISO 正式发布了国际标准 SGML(Standard Generalized Markup Language，标准通用置标语言)，标准号是 ISO8879-1986.我国于 1995 年也把 SGML 语言作为国家标准，标准号为 GB 14814。冯志伟在《当代语言学》(1998 年，第 4 期)的《标准通用置标语言 SGML 及其在自然语言处理中的应用》一文详细介绍过 SGML 语言。

XML(eXtensible Markup Language，可扩充置标语言)是 SGML 的一个子集，被广泛地用做语料库标注的元语言，通过 DTD(Document Type Definition，文件类型定义)和 Schema 来规范 XML 文件，从而使表现与内容分离，规范与实现分离，具有良好的扩缩性。

我国语料库的建设将一定会采用通用置标语言作为描述语料库的元语言。

4.2 语料库的资源共享

语料库的资源共享的方式有如下几种：作为产品出售；实行会员制；授予使用许可权；给非赢利目的的学术结构提供无偿使用。

4.3 语料库的知识产权

随着语料库的广泛使用，语料库的知识产权问题越来越尖锐，以正式出版物为资源的语料库面临版权的问题。建议政府有关部门建立关于语料库资源的版权

法规，建议中国中文信息学会出面协调。

4.4 语料库加工中的统计垃圾

由于电子文本的普遍使用，语料资源的获取变得越来越容易，我国大规模的真实文本语料库其规模已经达到 5 亿字。美国计算语言学会的 ACL/DCI 数据采集计划指出，如果以文本形式存储语料，语料库的容量一般可以为 1 亿词次以上，将来可以达到万亿词次的数量级。随着语料库容量的不断增大，语料统计中的数据稀疏现象会越来越严重。宋柔在统计语料库中的词语接续对时发现：“随着语料库规模的增大，新增加的接续对中的垃圾逐渐会占大部分甚至绝大部分。垃圾主要分布在统计到的低频度接续对中，主要来源是分词中专名识别错误。”应该看到，在统计垃圾中蕴藏着许多正在萌芽的新的语言现象，如“喷塑、蒜农、危改、市话、高检”等低频度的新词语，由于在词典中没有存储，都可以在统计垃圾中找到。如何真确地对待统计垃圾，避免统计中的数据稀疏现象，变垃圾为有用的语言资源，是大规模真实文本处理的一个新课题。

参考文献

J. Sinclair, Reflections on computer corpora in English language research [M], 1982.
北京语言学院语言教学研究所，汉语词汇的统计与分析[M]，外语教学与研究出版社，1985 年。
冯志伟，计算语言学基础[M]，商务印书馆，2001 年。
冯志伟，语料库语言学与机器翻译[M]，《信息网络时代与日本研究》，山东大学出版社，1999 年。
刘开瑛，中文文本自动分词和标注[M]，商务印书馆，2000 年。
孙茂松等，高频最大交集型歧义切分字段在汉语自动切分中的作用[J]，中文信息学报，第 13 卷，第 1 期，1999 年。
俞士汶、朱学锋、段慧明，大规模现代汉语标注语料库的加工规范[J]，中文信息学报，第 14 卷，第 6 期，2000 年。

Evolution and present situation of corpus research in China

Zhiwei FENG
Institute of Applied Linguistics, Ministry of Education
Chaonei Nanxiaojie 51
100010 Beijing, China
e-mail: zwfengde@public.bta.net.cn

Submitted on 25. November. 2001 Revised and Accepted on 25. July. 2002

Abstract

In this paper, the author shortly reviews the development of corpus research abroad. Then he introduces in detail the development and present situation of corpus linguistics in China: earlier corpus, large-scale & authentic text corpus, national corpus, speech corpus, bilingual corpus and corpus of minority languages in China. The various processing techniques for corpus are also introduced: automatic word segmentation of Chinese text, automatic POS tagging, automatic tagging of phrase structure and automatic alignment of bilingual corpus. This paper is a bird's –eye view of corpus linguistics of China. At last, the author discusses several problems in present corpus research: standardization of corpus specifications, commonly sharing of language resources, knowledge properties, etc.

Keywords

Corpus; large-scale & authentic text; speech corpus; bilingual corpus; corpus of minority languages in China; automatic word segmentation; automatic POS tagging; automatic tagging of phrase structure; automatic alignment of bilingual corpus