《当代语言学》第3卷2001年第1期1-21页，北京

汉字和汉语的计算机处理

冯志伟　国家语言文字工作委员会

提要 本文从汉字输入与汉语语料库、语料库中汉语书面文本的自动切词、语料库中汉语书面文本的词性(POS)自动标注、语料库中汉语书面文本的自动短语定界与句法标注、机器词典的建造、术语数据库的建造、机器翻译、计算机辅助文本校对、情报自动检索、汉语语音自动识别、汉语语音自动合成、汉字自动识别等12个方面，介绍我国汉字和汉语的计算机处理的研究情况。

关键词 中文信息处理　语料库　自动切词　自动标注　机器翻译

我们正在进入一个信息革命的新时代，这个信息时代的显著特点，是计算机在人类生活的各个方面，起着越来越大的作用。

自然语言是人们最重要的交际工具，它与信息处理有着十分密切的关系。在信息时代，只有四十年历史的计算机向拥有六千年历史的汉字提出了严峻的挑战。

汉语是汉藏语系最重要的语言。现在，世界上约有九亿四千万人以汉语为母语。不仅中国的汉族讲汉语，新加坡和马来西亚也有不少人讲汉语，汉语还是联合国的工作语言之一。

汉字是记录汉语的符号集。在世界各种书面文字中，汉字是最大的符号集。拉丁字母有26个符号，斯拉夫字母有33个符号，亚美尼亚字母有38个符号，泰米尔字母有36个符号，缅甸字母有52个符号，泰文字母有44个符号，老挝字母有27个符号，藏文字母有35个符号，韩文字母有24个符号，日文假名有48个符号。而汉字符号是最多的。

在汉字从古到今的发展过程中，汉字的总数在不断增大。下面列举的是在不同的历史时代的辞书中所包含的汉字的数目：

字书名	编者	字数	年代
《苍颉篇》	李斯	3300	秦代
《训篡》	杨雄	5340	汉代，1-5年
《续训篡》	班固	6180	汉代，60-70年
《说文解字》	许慎	10516(重文1163字)	汉代，100年
《广雅》	张揖	16150	汉代
《声类》	李登	11520	魏代，230年
《字林》	吕忱	12824	晋代，400年
《字统》	杨承庆	13734	北魏，500年
《玉篇》	顾野王	16917	南梁，534年
《切韵》	陆法言	12158	隋代，601年

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

《韵海镜源》	颜真卿	26911	唐朝，753年
《龙龛手鉴》	释行均	26430	辽代，997年
《广韵》	陈彭年	26194	宋朝，1008年
《字汇》	梅膺祚	33179	明朝，1615年
《正字通》	张自烈	33440	明朝，1675年
《康熙字典》	陈廷敬	47043	清朝，1716年
《大汉和辞典》	诸桥辙次	49964	1959年
《中文大辞典》	张其昀	49888	1971年
《汉语大字典》	徐中舒	54678	1990年
《中华字海》	冷玉龙	86000	1994年

面对这样的大字符集，如何用计算机来进行汉字的信息处理？这是对计算语言学和语料库语言学的一个挑战。

远在40多年前的1956年，著名学者丁西林就提出了设计中文电动打字机的建议。另一位学者钱文浩在《科学通报》发表了《文字与通讯》一文，开始讨论汉字编码的问题。1959年，一些中国学者设计了俄汉机器翻译系统(RC-59)，这是中文信息与计算机最早的结合。1969年9月，邮电科学研究院试制成功中国第一台电子式中文电报快速收报机，揭开了用计算机技术处理汉字信息的序幕。1974年8月9日，中国科学院、一机部、四机部、新华社和国家出版局向国家计委和国务院提出《研制汉字信息处理系统工程的请示报告》。9月24日，国家计委批准把汉字信息处理系统列为1975年国家科技发展计划。这就是有名的“七四八”工程。“七四八”工程把能够输出高质量汉字的汉字照相排版编辑系统作为重点攻关项目。研究人员经过二十多年的艰苦奋斗，取得了令人瞩目的成就。我国已经以计算机激光汉字编辑排版系统改造了传统的铅字排版，在印刷技术上结束了“铅与火”的时代，在推广应用上达到了普及的程度。我国省以上的大报已采用了计算机激光汉字编辑排版技术，百分之五十左右的地区一级的报刊以及部分书刊印刷也跨入了这一技术改造的行列。1989年，我国自行研制的计算机激光编辑排版系统开始出口海外。现在，香港、澳门、马来西亚的大多数中文报刊，美国、加拿大、澳大利亚、法国、巴西、印度尼西亚、泰国、菲律宾和台湾地区的一些中文报刊也先后采用了这个计算机激光编辑排版系统。我国自行研制的计算机彩色制版系统已成为商品推向市场。北京、上海、广州、香港、澳门和美国的部分中文报刊，已经采用了中国自行开发的彩色图片与汉字合一处理和整页编辑排版的系统，印出的彩色汉字报纸十分精美。

随着计算机汉字输入输出问题的解决，我国的中文信息处理技术得到了多方面的发展，诸如汉字信息压缩、汉字信息通讯、汉字输入与汉语语料库、语料库中汉语书面文本的自动切词、语料库中汉语书面文本的词性(POS)自动标注、语料库中汉语书面文本的自动短语定界与句法标注、机器词典的建造、术语数据库的建造、机器翻译、计算机辅助文本校对、情报自动检索、汉语语音自动识别、汉语语音自动合成、汉字自动识别等多项技术，均取得了显著的成就。中文信息处理已经不再停留在处理汉字上，而且还处理汉语，其中包括解决自动切词、自动标注、自动分析等更深层次的问题。其实其中的大部分问题，早在50年代初期就有研制机器翻译系统的

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

语言学工作者提出了，只不过在80年代以后，需要从工程的角度更加系统地来研究它们而已。可以说，我国早期的中文信息处理研究是从机器翻译系统中自动句法分析和生成的研究开始的，以“七四八”工程为代表的汉字信息处理系统的研究比机器翻译系统的研究要晚得多。

现在，中文信息处理系统技术和产品的开发和生产、销售、服务，已经成为我国计算机信息产业的重要组成部分。在“七四八”工程中诞生的北京大学方正集团和华光集团1993年的年营业额已达到十亿元人民币(一亿两千万美元)，年出口创汇近两千万美元。长城集团的CCDOS、四通集团的中英文打字机、联想集团的联想汉卡、巨人集团的汉卡，都成为这些著名企业集团赖以发展的重要产品。此外，在中国还有数以百计的企业从事中文信息处理产品的开发、生产和经营。

在本文中，我们将介绍汉字和汉语计算机处理的主要成就，分12个方面来介绍：

1）汉字输入与汉语语料库

2）语料库中汉语书面文本的自动切词

3）语料库中汉语书面文本的词性(POS)自动标注

4）语料库中汉语书面文本的自动短语定界与句法标注

5）机器词典的建造

6）术语数据库的建造

7）机器翻译

8）计算机辅助文本校对

9）情报自动检索系统

10）汉语语音自动识别系统

11）汉语语音自动合成系统

12）汉字自动识别系统

1. 汉字输入与汉语语料库

现代汉字是以形声字为主要构造方式的表义兼表音的文字，这种文字体系不是纯表义的，但有相当的表义功能。这种文字体系也不是纯表音的，但在很多字形中又包含着表音的成分。表义成分和表音成分在现代汉字中互相制约、互相补充，却又很不完备。

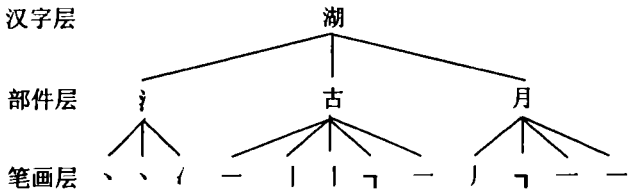
由于我国简化了一大批汉字，现代汉字字形构造已不再纯粹按照“六书”方式，它已经明显地分成了两大类：一类仍然按照传统的“六书”方式，主要有象形、指事、形声、会意四种造字方式。其中，形声字占现代汉字总数的80-90%。但是，形声字的声旁，由于本身不是音素符号，再加上语音演变的影响，其有效表音率是很低的；形声字的形旁，由于词义的复杂性，再加上词义的不断发展，其表义功能极为宽泛、模糊而又十分有限。另一类是经过简化后不能再归入象形、指事、形声、会意四种构造方式的字。它们的构造方式有轮廓字(如“齐、变”)、符号字(如“办、邓”)、省略字(如“声、际”)、草书楷化(如“专、长”)等。传统和简化这两大类所包含的构造方式的总和，就是现代汉字全部的构造类型，也就是它的全部造字方式。

现代汉字的形体可以分为三个层次：

汉字 → 部件 → 笔画

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

汉字是最高层次，部件是中间层次，笔画是最低层次。如“湖”字的三个层次如下图所示：



层次越高，表示一个汉字所用的符号越少，表示全部汉字所用的符号的总数越多。如最高一层，表示一个符号只需用一个符号，如果有五万个汉字，就得用五万个符号。层次越低，表示一个字所用的符号越多，而表示全部汉字所用的符号的总数越少。如最低一层，表示一个汉字最多要用几十个笔画符号(笔画最多的汉字有64画)，而笔画符号的总数可减少到横、竖、撇、点、折等有限的几种。部件处于中间层次，它是组成现代汉字的能够相对独立的结构单位，它比笔画完整，又比汉字本身简单、灵活，所需符号数目适中。

如何把汉字输入输出计算机，成为汉字信息处理的关键性问题。电子计算机是西方发明的，它是建立在西方文化的基础之上的。它使用西文打字机键盘，会用西文打字机的人来操作电子计算机，马上就可以驾轻就熟，而如果要用电子计算机来处理汉字，就会遇到巨大的困难。计算机是一种文化的载体，而汉字则是汉字文化圈的文化基础。在汉字文化圈内，计算机如果不和汉字相结合，如果不能处理汉字，就不可能得到生存和发展，而要使计算机能处理汉字，就要解决汉字的数字化、信息化、智能化以及汉字输入输出计算机的问题。

我国在六十年代末期就开始对汉字信息处理进行探索和实践，1968年研制成汉字电报译码机，七十年代中期明确提出“汉字信息处理系统”的研究课题(即“七四八”工程)。自1978年以来，我国开始广泛应用大规模集成电路存储器和成套的微处理机芯片，为汉字输入计算机提供了物质条件，研制成了一些新型的汉字输入输出设备，并配制成各种应用系统。汉字信息处理的研制成果已经在中国的现代化建设中发挥着重要作用。

目前我国的汉字输入方法大致可以分为六类：编码输入法、整字输入法、拼音-汉字转换法、印刷体光学输入法、手写输入法、声音输入法等。这里着重谈谈编码输入法。

所谓编码输入法，就是给汉字规定一种便于计算机识别的代码，使每一个汉字对应于一个数字串或符号串，从而把汉字输入计算机。学者们提出的汉字编码方案已有近千个，其中上机通过实验和已被采用的编码方案已达数十种之多。这些汉字编码方案大致可以分为四种：

形码：根据汉字的字形来进行的编码。如笔形编码法和五笔字形编码法。笔形编码法在笔画层进行编码。这种方法把汉字的笔画分为一(横)、丨(竖)、丿(撇)、丶(点)、乛(折)、乚(弯)、×(叉)、口(方)八类，分别用1、2、3、4、5、6、7、0等数字来代表，横、竖、撇、点为单笔，折、弯、叉、方为复笔。汉字代码是不等长码，最大码长为九码。五笔字形编码法在部件层进行编码。这种方法是把汉字分解为部件，把汉字的部件归并为664个，并进行部件的优选，合理安排部件在键盘上的布局。平均码长为四码，使用高频字简码和词汇码后，平均码长为2.8码。

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

音码：根据汉字的读音来进行编码。音码一般以汉语拼音方案为根据，汉语拼音方案已有三十多年的历史，1982年成为国际标准，标准号是ISO 7098。由于汉语拼音方案是以国际通行的拉丁字母字符集以及与它们相近的发音为基础制定的，有利于国际交流。采用音码最大的困难是区分同音字的问题。汉字的音节不计声调共408个，而汉字的数目成千上万，这就必然导致大量的同音字的出现。现有的音码方案都把区分同音字作为主要的研究目标。例如智能ABC输入法采用以词定字的方法，在计算机中存储双音词和多音词数万个，按词输入，以词来定字，从而减少了重码。有一种叫做“自通”(autoway)的输入法，按照马尔可夫(Markov)信源模型来进行汉字序列处理的工程设计，用户可使用拼音直接向计算机输入汉语语流，系统可对用户输入的拼音形式的语流进行自动分词、自动筛选和自动频度统计，从而把输入的拼音语流自动转换为汉字文本。

形音码：这种编码法基本上立足于字形分解，把字分解为部件或笔画，部件和笔画统称为字元(element)，各个字元又通过它们的读音来帮助记忆。

音形码：这是一种以音为主，以形为辅的编码，利用字形来区分同音字。

从1978年至今，汉字编码的研究出现了百家争鸣的局面，各种方案如雨后春笋，源源不绝，数不胜数。为了促进汉字编码的研究更加健康地向前发展，有必要对已有的汉字编码方案进行评测，以便由“百家争鸣”逐渐地发展到“百家归一”，优选出最佳的汉字编码方案。

印刷体光学输入法(optical character recognition，简称OCR)：首先把在纸面上的汉字信息转换成离散的电子信号，然后再由计算机来识别这些离散的电子信号，从而把汉字输入计算机。目前，OCR汉字识别系统能识别GB 2312-80《信息交换用汉字编码字符集－基本集》中的6763个汉字，识别正确率达99.6%。

声音输入法的目标是直接把汉语的语音转换为汉字。“四达863A”系统可识别汉语的398个基本音节，识别正确率为93%，响应时间小于0.1秒，输入速度为每分钟80个汉字。汉语的音节数比较少(不计声调只有420个音节，计声调也只有1300个音节)，而英语的音节有4030个，俄语的音节有2960个，因而有人估计，汉语的语音识别会比英语和俄语容易一些。

汉字输入计算机之后，就形成了机器可读的汉语书面文本，这就为建立汉语语料库提供了条件。从1979年到1992年，在国内建立的主要的语料库有：

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)，2千万字，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。

现代汉语语料库(1983年)，180万字，北京语言学院。

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。

北大汉语语料库(1992年)：500万词，北京大学。

1992年以后，大量的语料库在我国研究中文信息处理的单位建立起来，语料库已经成为研究

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

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。

此外，国家语言文字工作委员会语言文字应用研究所还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》(Republic)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。

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，其规模计划将达7000万汉字，这将成为世界上最大的汉语语料库。这个语料库的语料要经过精心的选材，语料应受到如下因素的限制：

时间的限制：选取从1919年到当代的语料，以1977年以后的语料为主。

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。

使用领域的限制：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，尚待进一步的加工，还是“生语料库”，因而还不能供社会使用。

2. 语料库中汉语书面文本的自动切词

语料库输入和校对之后是生语料库，还需要进行深加工，使语料库由“生”变“熟”，这样才能从熟语料库中获取蕴藏在语言中的各种知识。为了进行语料库的深加工，首先就要实现书面文本的自动切词。

书面汉语的句子，是连续的汉字流，词与词之间没有空白，除了标点符号之外，单词之间的界限无明显的标志。而中文的自动句法分析和语义分析，都是以单词为基本单元的，这样，书面汉语的自动切词，就成了中文信息处理的一个基本问题。

为了自动地找出隐藏在汉语文本中的单词，我们一般的做法是把文本中的汉字符号串与中文词典中的单词条目相匹配。主要的匹配方法有：

最大匹配法(maximum matching method, MM法)：选取包含6-8个汉字的符号串作为最大符号串，把最大符号串与词典中的单词条目相匹配，如果不能匹配，就削掉一个汉字继续匹配，直到在词典中找到相应的单词为止。匹配的方向是从右向左。

逆向最大匹配法(reverse maximum method, RMM法)：匹配方向与MM法相反，是从左向右。实验表明：对于汉语来说，逆向最大匹配法比最大匹配法更有效。

双向匹配法(bi-direction matching method, BM法)：比较MM法与RMM法的切分结果，从而决定正确的切分。

最佳匹配法(optimum matching method, OM法)：将词典中的单词按它们在文本中出现频度的大小排列，高频度的单词排在前，频度低的单词排在后，从而提高匹配的速度。

联想-回溯法(association-backtracking method, AB法)：采用联想和回溯机制来进行匹配。

尽管采用这些方法，某些切分有歧义的符号串(ambiguous segmentation strings, ASSs)和词典中的未登录词(unregistered words, URWs)仍然严重地影响着切词的准确性，这些问题在自动切词中必须解决。

ASSs有两种类型：

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

交集型歧义切分字段：例如，“太平淡”可能切为“太平”或“平淡”，“平”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。

URWs 主要是专有名词，即人名、地名、机构名，它们一般在词典中没有登录。例如：“冯志伟”是一个不见经传的普通人，在词典中决不会登录；“蒂豪尼”(Tihany)是匈牙利的一个小城市，词典一般也不登录。这样的未登录词，在自动切分时将无法匹配，造成切分的困难。

为了解决这些问题，可以利用各种知识，特别是词类的知识。因此，如果把词类的自动标注与自动切词结合起来，将可以提高切词的精确度。

1992 年，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。

3. 汉语语料库的自动词类标注

自动词类标注的方法有两种：基于统计的方法；基于规则的方法。

采用基于统计的方法，词类自动标注过程可按如下步骤进行：

(1)从语料库中选出一定数量的文本，作为训练集(training set)。手工分析这个训练集，采用二元语法(digram grammar)，从中归纳出统计数据。(2)根据对训练集的语料分析得出的统计数据，构造统计模型。(3)根据统计模型去标注语料库中新的文本。(4)标注时所用的标记都记录在词典中的单词上。

清华大学计算机系黄昌宁等采用统计方法建立了一个自动词性标注系统，标注正确率达 96.8%，自动标注的速度为每秒 175 个汉字。

对于基于规则的方法来说，最为严重的问题是兼类词。在汉语中，兼类词主要集中在动词、名词、形容词等常用词上。各种兼类现象的比例如下：

动词－名词兼类：37.6%

动词－形容词兼类：24.3%

名词－形容词兼类：10.4%

形容词－副词兼类：4.55%

动词－介词兼类：4.04%

动词－副词兼类：2.27%

名词－动词－形容词兼类：2.27%

名词－副词兼类：2.02%

其他兼类现象：12.55%

基于规则的方法主要根据句法、语义、上下文等语言学规则来消解兼类歧义。

事实上，基于统计的方法是一种经验主义的方法，而基于规则的方法则是一种理性主义的方法，我们应该把经验主义的方法与理性主义的方法很好地结合起来，并且在词性自动标注中吸收

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

不同方法的长处。北京大学计算语言学研究所就采用这样的策略，实验结果如下：切词正确率：97.68%(封闭语料)，词性标注正确率：96.06%(封闭语料)，95.72%(开放语料)。

4. 语料库中汉语书面文本的自动短语定界和句法标注

在对汉语书面文本进行自动切词和自动词性标注之后，我们应该认真地检查实验的结果。如果我们确认这些结果都是正确无误和无懈可击的，那么，就可以开始自动短语定界和自动句法标注的工作。这些工作可按如下步骤进行：

根据单词的信息、词类类别和句法特征，确定哪一个单词是短语的左边界，哪一个单词是短语的右边界，哪些单词是短语的中间部分。

短语定界的格式如下：

[w w...w w]

其中，[w 是开括号，它是短语的头，w] 是闭括号，它是短语的尾。

自动短语定界的步骤是：(1)根据上下文信息，把开括号与相应的闭括号对应起来。(2)根据歧义消解规则和统计信息，消解短语定界的歧义。(3)生成表示句子结构的成分结构树。

现在，北京大学计算语言学研究所正在开发一个汉语语料库的多级加工系统(Chinese corpus multilevel processing, CCMP)。这个 CCMP 系统包括两个子系统和一些辅助工具。

子系统是自动切词和词性标注子系统、自动短语定界和句法标注子系统。

辅助工具有查询工具、样本采取工具、统计工具、语料库管理界面等。

实验结果如下：交叉括号的百分比为 13.98%；错误短语标记的百分比为 8.65%。

从实验结果来看，汉语语料库的自动标注和多级加工处理，还有相当多的问题等待我们解决。

下面是一篇短文前 6 句的短语定界和句法标注结果，每句前面都标有序号。标注时采用北京大学计算语言学研究所的标注符号。

1 [zj 纱笼 /n 。/w]

2 [zj [fj [dj 纱笼 /n [vp 是 /v [np [np 马来 /n 民族 /n] 的 /u [np 传统 /n 服装 /n]]]] ，/w [vp [vbar 富 /a 有 /v] [np 浓厚 /a 的 /u [np 热带 /n 情调 /n]]]] 。/w]

3 [zj [fj [dj [np 纱笼 /n 的 /u 用途 /n] [ap 很 /d 广 /a]] ，/w [dj [pp 除了 /p [vp [tp 出外 /v 时 /n] 穿 /v]] ，/w [vp 也 /d [vbar 被 /p [vp 当做 /v [np 浴衣 /n 、/w 睡衣 /n 和 /c [np 婴孩 /n 的 /u 摇篮 /n]]]]]] 。/w]

4 [zj [fj [np 纱笼 /n 的 /u 图案 /n] ，/w [fj [dj 不但 /c [dj 设计 /n 别出心裁 /a]] ，/w [dj 而且 /c [dj 色彩 /n 鲜艳夺目 /a]]]] 。/w]

5 [fj [dj [np 它 /r 的 /u [np 制作 /v 方法 /n]] [vp 有 /v [mp 三 /m 种 /q]]] ，/w [vp 即 /v: /w [zj [np [np [vp [pp 用 /p 模型 /n] 印制 /v] 的 /u] 、/w [np [np [np 人工 /b 纺织 /n] 的 /u] 和 /c [np 蜡染 /n 的 /u]]] 。/w]]]

6 [zj [dj 其中 /r [vp [pp 以 /p [np 蜡染 /n 的 /u]] [vp 最 /d [vp 受 /v 欢迎 /v]]]] 。/w]

最近由国家科委基础研究高技术司、国家高技术计划智能计算机系统主题 863 专家组、全国信标委非键盘输入分技术委员会组织了自然语言处理系统的评测。这个评测简称 863 评测。

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

在汉语文本的自动分词和自动标注方面，参加评测的有北京大学计算语言学研究所、北京工业大学计算机学院、北京邮电大学提交的三个系统。评测时，有自动分词和自动标注同时进行的一体化的综合测试，也有分词和标注同时进行的一体化的单项测试，也有只测试分词而不测试标注的单项测试。评测结果如下：

对北京大学计算语言学研究所10万语料的一体化综合测试结果，分词正确率为87.42%，词性标注准确率为74.51%。

对120个测试点的动名兼类一体化测试，词性标注准确率为51.67%；对30个测试点的形名兼类一体化测试，词性标注准确率为33.33%；对50个测试点的形动兼类一体化测试，词性标注准确率为42.00%。可以看出，兼类词判别的水平还有待提高。

北京工业大学计算机学院提交的系统主要进行自动分词的单项测试。对229个测试点的交集型歧义切分字段定点测试结果，准确率为68.56%；对20个测试点的多义组合型歧义切分字段定点测试结果，准确率为40.00%；对741个测试点的我国人名定点测试结果，准确率为91.28%；对855个测试点的中国地名定点测试结果，准确率为69.12%；对456个测试点的外国人名地名的汉语译名定点测试结果，准确率为82.83%。可以看出，多义组合型歧义切分字段的切分结果，不如交集型歧义切分字段的切分结果，地名的切分结果不如人名的切分结果。今后还须加强多义组合型歧义和地名切分的研究。

北京邮电大学提交的系统也是主要进行自动分词的单项测试。对229个测试点的交集型歧义切分字段定点测试结果，准确率为36.68%；而对20个测试点的组合型歧义切分字段定点测试结果，准确率反而比交集型歧义切分字段为高，这可能是因为组合型歧义切分字段的测试点只有20个，不足以真实地反映真实语料的面貌。

5. 机器词典的建造

机器词典是最重要的语言资源。北京大学计算语言学研究所俞士汶、朱学锋等开发的现代汉语语法知识库，就是一种机器词典。这项研究与北京大学中文系密切合作进行，.在现代汉语语法知识库的基础上，他们又编写了《现代汉语语法信息词典》。这部机器词典以语法－义项相结合的原则以及词典编纂的普遍原则，选取了5万多个词语，又根据语法功能分布的原则，建立了面向语言信息处理的现代汉语词语分类体系，完成了这5万多个词的归类，确定了每个词的词性。由于属于同一类的各个词语的语法属性仍然有很多差别，采用了关系数据库文件格式来描述每一个词语及其语法属性的二维关系。机器词典中共有32个数据文件，其中包含全部词语的总库1个，各类词库23个。总库设21个属性字段，各类词库又分设若干属性字段，例如，名词库设27个属性字段，动词库设46个属性字段，等等。除此之外，某些类词库下面又设分库。例如，动词库下面设6个分库，代词库下面设2个分库，分别描述每一个子类的更细微的语法属性。所有的库都可以根据主关键字段(词语+词类+同形)进行连接。这样一来，32个数据库文件构成了有上位下位继承关系的“树”，在这样的树中，子结点可以继承父结点的全部信息，将父结点与子结点连接起来就可以得到关于每个词的更加全面的信息。如果把每个库所包含的词语数同该库的属性字段数的乘积定义为该库的“信息量”，那么，现在总库的信息量约为60万，32个库的信息量达250万。这些信息量所需的存储空间约为16兆字节。

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

这部语法信息词典已经为国内外不少计算语言学研究单位所采用，作为重要的语言资源。他们建立了一个比较完善的现代汉语词语的语法功能分类体系，把现代汉语的基本词类分为18类（括号内的英文字母是其代码）：

名词(n)：　例如，牛、书、水、教授、国家、心胸、北京
时间词(t)：　例如，明天、元旦、唐朝、现在、春天
处所词(s)：　例如，空中、低处、郊外
方位词(f)：　例如，上、下、前、后、东、西、南、北、里面、外头、中间
数词(m)：　例如，一、第一、千、零、许多、百万
量词(q)：　例如，个、群、克、杯、片、种、些
区别词(b)：　例如，男、女、公共、微型、初级
代词(r)：　例如，你、我们、这、哪儿、谁
动词(v)：　例如，走、休息、同意、能够、出去、是、调查
形容词(a)：　例如，好、红、大、温柔、美丽、突然
状态词(z)：　例如，雪白、金黄、泪汪汪、满满当当、灰不溜秋
副词(d)：　例如，不、很、都、刚刚、难道、忽然
介词(p)：　例如，把、被、对于、关于、以、按照
连词(c)：　例如，和、与、或、虽然、但是、不但、而且
助词(u)：　例如，了、着、过、的、得、所、似的
语气词(y)：　例如，吗、呢、吧、嘛、啦
拟声词(o)：　例如，呜、啪、丁零当啷、哗啦
叹词(e)：　例如，哎、喔、哦、啊

这些基本词类可以合并成为较大的词类。名词、时间词、处所词、方位词、数词、量词统称体词，动词、形容词、状态词统称谓词。代词一部分属于体词，一部分属于谓词。体词、谓词、区别词、副词又合称实词。介词、连词、助词、语气词合称虚词。实词和虚词是汉语的两个最大的词类。此外，还有拟声词和叹词，它们被列在这两大词类之外。当然，这18个基本词类还可以再划分小类，这里不再细说。

语法信息词典中登录的基本是词，但是，在实际的文本中，常常会出现一些小于词的语言成分，如前接成分、后接成分、语素、非语素字等，又会出现一些大于词的成分，如成语、习用语、简称略语等，因此，词典中还增加了如下的词语类别：

前接成分(h)：　例如，阿、老、非、超、单
后接成分(k)：　例如，儿、子、性、员、器
语素(g)：　例如，民、衣、失、遥、郝
非语素字(x)：　例如，鸳、枇、蚣
成语(l)：　例如，按部就班、八拜之交
习用语(i)：　例如，木头疙瘩、光杆司令、跑龙套
简称略语(j)：　例如，三好、政协

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

再加上文本中常出现的标点符号(代码为w)，形成了覆盖英文26个字母的现代汉语词语标记集。

这部《现代汉语语法信息词典》有电子版本，也有书面版本，叫做《现代汉语语法信息词典详解》。

中国人民大学语言文字研究所林杏光等研制了《现代汉语动词大词典》，这部词典的编制目的是“人机两用”，因此，它对于汉语的信息处理有很大的参考价值。作者根据汉语的格语关系，把22个格组成的格系统分成三个层次：第一层由“角色”和“情景”组成；“角色”下面包括“主体”、“客体”、“邻体”、“系体”四个要素，“情景”下面包括“凭借”、“环境”、“根由”三个要素，这七个要素构成了格系统的第二层，它们以述语动词为核心，完整地表达了一个句子的意义；第三层是22个具体的格，它们分别属于上一层的七个要素。层次格局很清楚。本词典还根据格关系把汉语动词分为他动词、自动词、外动词、内动词、领属动词、系属动词六个次类，并从两条视线来考察一个动词：从动词往后看其客体，将动词分为两类，带客体的叫及物动词，不带客体的叫不及物动词；从动词向前看其主体，也可以把动词分为两类，连接施事主体的叫自主动词，连接当事主体的叫非自主动词。据此，自主而及物的动词是他动词（如“踢、吃、研究”等），自主而不及物的动词是自动词(如“跑、蹲、飞”等)，非自主而及物的动词是外动词(如“听见、看见”等)，非自主而不及物的动词是内动词(如“病、死”等)。再从主体和动词的领属关系、系属关系划分出领属动词(如“具有、属于”等)和系属动词(如“是、等于”等)。这样的分类把动词的语法属性同格关系联系起来，有助于计算机进行自动句法语义分析。

董振东建立的知网(how-net)是一种特殊类型的电子词典。它实际上是一个词典知识描述系统，描述的词汇包括汉语和英语两种语言，这两种语言是相对独立的，它们在词语之间的对应是建立在相同的属性描述的基础之上的。目前，知网有汉语词汇33069条(41791个概念)、英语词汇38774条(48834个概念)。

知网对概念作了形式化描述，把概念和它们的属性组织在一个完整的知识系统中，它对于自然语言的计算机处理是很有价值的。最近，董振东把他设计的知网在Internet上公布，免费提供非商业性的使用，成为网络上的一种有用的语言资源。

6. 术语数据库的建造

术语数据库是在专业领域内的机器词典，它的研究与机器词典密切相关。

科技术语是人类的科学技术知识在自然语言中的结晶，是一种非常重要的语言资源。

1990年，我国成立了计算机辅助术语工作技术委员会(简称SC3)，挂靠在国家语言文字工作委员会语用所，由冯志伟担任秘书长。SC3与ISO国际标准化组织的TC37/SC3对口，积极展开国际交流活动。

SC3成立之后，制定了一系列的有关术语数据库的国家标准：

GB/T 13725-92　《建立术语数据库的一般原则和方法》，1992年。

GB/T 13726-92　《术语与辞书条目的记录用磁带交换格式》，1992年。

GB/T 15387.1-94　《术语数据库开发指南》，1994年。

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

GB/T15387.2-94　《术语数据库开发文件编制指南》，1994年。

GB/T15625-95　《术语数据库技术评价指南》，1995年。

GB/T16785-97　《术语工作　概念和术语的协调》，1997年。

GB/T15786-97　《术语工作　计算机应用　数据类目》，1997年。

GB/T17532-98　《术语工作　计算机应用　词汇》，1998年。

最近正在编制《术语工作 计算机应用 机器可读术语交换格式(MARTIF)-协商交换》，该标准采用国际标准SGML语言来描述术语数据库。

1988年以来，我国的术语数据库如雨后春笋一样地建立起来：

GLOT-C：英汉数据处理术语数据库，这是中国科学院软件研究所与德国大琅禾费研究院(FhG)的合作研究课题，由中国科学院派冯志伟代表软件所在德国斯图加特FhG于1988年完成。这项国际合作项目开了我国术语数据库研究的先河。

TAL：英汉应用语言学术语数据库，含1万条术语。这是国家社会科学基金资助项目，由国家语言文字工作委员会语言文字应用研究所冯志伟等于1990年完成。

COL：英汉计算语言学术语数据库，含1万条术语。国家语委语用所与联邦德国特里尔大学合作于1993年完成。

多语言计算语言学术语数据库：汉-英-日-德四种语言对照，含5000多条术语，由北京大学计算语言学研究所于1994年完成，英汉对照的书面本《英汉对照计算语言学词语汇编》于1996年由北京大学出版社出版。

多语言机械术语数据库：汉-英-日-德-法-俄六种语言对照，含25万条术语，由机械部信息研究院术语中心研制，第一期工程已完成，现正在开发中。

农业科学叙词表：汉英对照，含2万5千条术语，由中国农业科学院于1991年完成。

化工叙词表：汉英对照，含2万5千条术语，由中国化工信息中心于1989年开始开发。现已完成，有两种版本：书面出版物版本和电子出版物(软盘)版本。所有的术语可以通过网络检索。

大百科全书术语数据库：汉英对照，每条术语都有定义，18万条，由新闻出版署拨款给中国大百科全书出版社开发，1995年开始，正在开发中。

标准化术语数据库：汉英对照，这是中国标准化与信息分类编码研究所(China Standardization and Information-Classification-and-Coding Institute，简称CSICCI)与奥地利标准化协会的合作研究项目，正在开发中。

汉英科技术语数据库：含5万条术语，由中国科学技术信息研究所开发，1995年开始，正在开发中。

上述术语数据库的开发，是汉语词汇研究的重要内容，是词汇学和词典编纂现代化不可分割的部分，可惜这方面的研究常常被语言学界忽略，这是令人遗憾的。

7. 机器翻译

我国是继美国、苏联、英国之后，第四个开展机器翻译研究工作的国家。早在1956年，国家便把机器翻译研究列入了我国科学工作的发展规划，成为其中的一个课题，课题的名称是：

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

“机器翻译、自然语言翻译规则的建立和自然语言的数学理论”。从此，机器翻译研究便在我国开展起来。国内主要的机器翻译系统如下：

(1)俄汉机译系统：1957年，中国科学院语言研究所刘涌泉等与计算技术研究所合作，开展俄汉机器翻译的研究。1959年，他们在我国制造的104大型通用电子计算机上，进行了俄汉机器翻译试验，翻译了9个不同类型的、较为复杂的句子。这是我国最早研制的机器翻译系统。

(2)英汉题录机译系统：1975年11月，在中国科学技术情报研究所成立了一个由情报所、语言所和计算所等单位的工作人员组成的机器翻译协作研究组，以冶金题录5000条为试验材料，制定英汉机器翻译方案并上机试验。1978年5月，在计算所111机上进行抽样试验，抽样20条，达到了预期的效果。

(3)汉-法/英/日/俄/德多语言机器翻译系统：1981年，冯志伟根据依存语法和配价语法的理论，采用模块化程序设计的方法，研究汉语到外语的机器翻译，在法国格勒诺布尔理科医科大学的IBM-4341计算机上通过试验，首次将20多篇中文科技文章用计算机翻译成法语、英语、日语、德语和俄语等5种外语，译文可读性强。研究成果在COLING'82国际计算语言学会议上得到好评。

(4)“译星”英汉系统：中国人民解放军军事科学院董振东等研制成功“科译1号”实用型全文与题录兼容的英汉机器翻译系统，于1987年在北京通过了技术鉴定。“科译1号”系统的基本原理是：由原语的线性结构出发，经过多层次、多次数的扫描，按规则的顺序匹配，形成以动词为根结点，以逻辑语义项为主结点的多结点、多标记的树形图，最后，从根结点逐层展开，形成译语的线性结构，得到相应的译文。该系统还采用了自行设计的专用的形式描述语言来书写自然语言的处理规则，实现了语言规则与计算机程序的彼此独立。此外，该系统还具有如下的翻译支援手段：1)词典与规则库的增添和修改手段；2)翻译过程的追踪和监测手段；3)为用户提供批量专业术语的增添手段；4)人用词典编制手段；5)英语词汇动态分析统计程序。该系统于1988年由中国计算机软件与技术服务总公司实现了商品化，命名为“译星1号”。“译星1号”在商品化过程中，发展成现在的“译星”系统，在语言词典和规则方面作了进一步的改善，在软件硬件的开发环境方面作了进一步的优化，还建立了专业词典，主要专业领域有计算机、经济、通讯、陶瓷、火力发电、印刷机械、汽车拖拉机、石油物探、地质、化工等。

(5)“高立”英汉系统和日汉系统：“高立”英汉系统由北京市高立电脑公司与中国社会科学院语言研究所刘倬等合作开发。高立英汉系统以具有普遍意义的语言学公理理论和原则作为语言分析器的理论基础，以智能化的机器词典代替传统的信息参数词典，使句法规则与词的个性相结合，使词义与词的参数和规则相结合，整个机器翻译系统实质上是一个词专家系统。这个机器翻译系统还建立了背景知识库，把语义分析与句法分析有效地结合起来，在抽象的形式分析中，充分地利用语义信息。由于机器词典与系统的运行程序彼此独立，用户可以通过追踪信息和词典维护程序来修改机器词典的内容，这样，用户就有可能在自己的使用过程中不断地修改机器词典，不断地提高机器翻译的译文质量。该系统具有良好的可扩充性和可移植性，系统的程序采用模块化的方法来设计与实现。整个机器翻译系统由翻译子系统、语言知识管理子系统、支援子系统三个部分组成。系统的基本词库收词60000条，语法规则库收规则800条，背景知识库

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

收规则150条，译准率达80%以上。

此外，高立电脑公司还开发了一个“日汉自动翻译系统”，该系统采用了高立英汉系统的理论方法与开发策略。

(6)863-IMT/EC英汉系统：该系统由中国科学院计算技术研究所陈肇雄等开发，从1986年开始研究，经历了理论探索(1986年-1988年)、模型系统试验(1989年-1990年)和实用系统开发等三个阶段，现已实现商品化。机器翻译系统研制的内容，包括语言学工程、翻译处理软件环境和知识处理环境三个部分。语言学工程研究如何把语言学知识和用于机器翻译的非语言学常识进行归纳和形式化描述，以适合于计算机处理。其中，语言学知识包括机器翻译过程中需要用到的词法、语法、语义以及语用知识，而非语言学常识包括机器翻译过程中常常涉及的学科分类、背景文化知识以及专业知识。翻译处理软件环境研究如何应用形式化的语言学知识和非语言学常识实现从原语输入到译语输出的转化，这一过程包括词法分析算法、结构分析算法、上下文相关处理、译语生成等分析和推理机制的实现技术。知识处理环境研究如何提供一套有效的软件工具环境，帮助语言学家归纳语言学知识和简单的非语言学常识，实现这些知识的形式化描述。

(7)Matrix英汉系统：该系统是国防科技大学1994年由史晓东研制成功的，已经开始商品化。该系统翻译速度在IBM PC386-DX33计算机上，每分钟能译5000-10000个英语单词，比国内外大多数机器翻译系统的速度高出1-2个数量级。按照日本电气工业促进协会JIEDA发布的关于1992年国际自然语言处理现状的报告中提出的标准，Matrix系统的翻译速度是当今世界上最快的。

(8)通译英汉-汉英系统：该系统由天津大通通译计算机软件研究所陈光火等研制，有英汉全文翻译、汉英全文翻译、Internet在线翻译三个系列产品，专业词典丰富，涉及机械、电信、化学、冶金、医学、建筑、广播、石油、环境保护、能源、汽车、电力、造纸、船舶、农林牧、纺织、航空、计算机、水利、航海、经贸等专业。

(9)雅信英汉系统：该系统由北京雅信诚软件技术有限公司开发，翻译方式有联机、自动和交互三种，可适应不同水平用户的需要。词库和语法库都向用户开放，用户不仅可以修改词库中的单词或词组，而且可以修改已有的语法模式或自己定义语法模式，突出人在翻译中的主导作用。

(10)LIGHT英汉系统：该系统由深圳桑夏(Sunshine)科技发展有限公司史晓东等研制，翻译速度与Matrix相当，基本上可以满足实时翻译的需要。这个系统又叫做“桑夏译王”。近来他们在Internet上开发了自动翻译网站“看世界”(readworld)，可以将网上的英文自动地翻译为中文，有力地帮助了网上英文信息的获取。这是一个非常有应用前景的翻译网站(网址：www.readworld.com)。

(11)Sino Trans 汉英-汉日机译系统：该系统由中国计算机软件与技术服务总公司吴蔚天等开发，于1993年9月通过了电子工业部的部级鉴定。包括汉英和汉日两个商品化的机器翻译系统。其中汉英系统的三个用户已翻译了数十万字的科技资料，节省了50%的工作量。Sino Trans还是一个多功能的中文信息处理系统，具备汉语自动切词、当前词的词性自动确定、词组生成、汉语语法树生成、汉语外语转换及外语生成等功能。由于其中的每一个模块都可以单独使用，所以Sino Trans还能为自然语言理解研究、基于语词的语言学研究提供条件，为汉语教学提供帮助。

(12)E-to-J英日机器翻译系统：由北京日电华公司冯志伟、董亦农等开发，英语分析采用短

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

语结构语法，日语生成采用依存语法，现已经商品化，在日本市场上销售。

此外，哈尔滨工业大学计算机系的汉英机器翻译系统 CEMT，东北工学院计算机科学与工程系的汉英机器翻译系统 CETRANS 也正在向实用化的方向努力。

对机器翻译系统也进行了863评测。在英汉机器翻译方面，参加评测的共有三个系统，其中，深圳桑夏公司提交了两个系统，哈尔滨工业大学提交了一个系统。测试平台为奔腾II/266的微型计算机和Windows 95操作系统。桑夏公司的两个系统成绩分别为73.31分和72.94分，哈尔滨工业大学的成绩为64.08分。

在汉英机器翻译方面，参加评测的也是三个系统，哈尔滨工业大学、中国科学院计算所和微电子中心各提交一个系统，成绩分别为69.45分、72.24分和72.84分。

从评测结果看来，我国的汉英机器翻译系统有了很大的发展，研究水平也提高了。这与我国从90年代以来汉语计算语言学的基础研究是分不开的。

8. 计算机辅助书面文本校对

1992年台湾工研院施得胜等在ICCIP'92上发表了《基于统计的中文错字侦测法》，首次提出利用计算机进行汉语文本校对的问题。此后，北京工业大学计算机学院、哈尔滨工业大学计算机系、清华大学中文系、山西大学计算机系等单位先后开展了这方面的研究。90年代中期开始，先后上市的校对产品有“黑马校对系统”、“工智校对通”、“方正金山校对系统”、“三欧校对系统”、“文捷校对系统”、“WORDPRO中文校对”等。由于种种原因，其中有一些系统已经停止开发。微软公司购买“工智校对通”的技术后开发的WORD中文校对系统已经发布测试版。所有这些系统都是以词语查错为主的，是人工校对的辅助工具。

其中，北京工业大学计算机学院宋柔等在国家和北京市的科研基金的支持下，开发出计算机辅助校对系统《工智校对通》。这个软件查错速度特别快，每秒钟达到数万字，每小时可达上亿字。查错和提供修改建议比较准确，能自动标示中西人名地名供人核查，还有多项辅助的知识检索核查功能，受到用户的欢迎。

9. 情报检索系统

我国从1963年开始进行机械情报检索的研究工作。1965年进行了机械情报检索试验。70年代以来开始研究计算机情报检索。1975年进行了首次计算机情报检索试验。1977年进行了计算机联机检索试验。1983年在中国科学技术信息研究所建立了连接美国和欧洲主要国家的数据库联机检索系统，这个系统通过意大利的ITALCABLE分组交换中心，连接到欧洲空间组织的ESA-IRS系统，并由数据交换网转接美国的DIALOG、ORBIT系统，这样，我国就可以在北京利用通信卫星检索到欧美200多个数据库的几十万篇文献。目前，不少单位在建立各种中文文献库，有的单位在研究自动标引和自动做文摘的问题。全国的科技情报部门已配备大中小型计算机120台以上，已建立各种科技文献数据库、事实数据库、数值数据库400多个，其中，中文科技文献数据库累计记录量约为150万条。我国的计算机情报检索已经取得了令人瞩目的进步。

我国从70年代末期开始探讨汉语文献的自动标引问题，“七五”期间先后建立了一批试验性的自动标引系统，如上海交通大学王永成等研制的基于汉字部件词典的中文篇名自动标引系统，

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

北京大学图书馆情报学系研制的基于规则和词典的中文文献自动标引系统，中国软件技术服务总公司吴蔚天等研制的基于非用字后缀表法的中文文献自动切词标引系统（“非用字”是指那些不能做标引词的字，如“其、起、且、首”等。而“用字”是指那些可以做标引词的字，抽词时，如果为用字则取，如果为非用字则舍）。

在自动文摘方面，上海交通大学计算中心在IBM-5550微机上初步开发出一个自动编制中文科技文献文摘的试验性系统。该系统根据“大多数反映文献主要内容的句子往往出现在段首或段尾”以及“文献的篇名基本上能反映其主题内容”的统计性结论，把包含预置关键词与标题关键词的句子从文献的某些重要部分中选出，作为文摘的句子，然后再适当地把这些句子组织成文献的文摘。

我国的全文检索研究开始于八十年代中期。1986年，武汉大学开始接受国家教委文科博士点科研项目“湖北省地方志全文检索系统”，建立了“湖北省地方志大事记”和“中国人民解放军大事记”两个全文数据库。接着，北京文献服务处（BDS）研制了“基于自然语言处理的中文情报检索和处理系统CIRPON”，用于BDS的文献自动标引和文摘自动处理，文献标引的查全率和查准率大体上相当于手工标引的质量。1990年初，北京信息工程学院与《人民日报》社合作开发了全文检索系统Biti FTRS（full text retrieval system的简称），在人民日报开始使用，并已实现了商品化。山西大学计算机科学系刘开瑛等使用自动切词、自动分类、自动词性标注等自然语言处理技术，于1991年研制了“中文全文检索软件系统”，现已被南京金陵石化总公司精细石化文献检索系统和山西省政府办公厅和太原市政府办公厅信息处理系统所采用。电子部计算机与微电子技术发展研究中心（CCID）中文信息处理开放实验室（CIPOL）张潮生等研制了中文全文检索系统TIR，该系统可以对各种文本型资料和某些数据库的文件进行操作，避免了传统检索系统只能检索主题词，而对主题词之外的信息无能为力的局限。该系统现在能够检索一切输入文本，对原始文献里的字符无特别限制，可以处理各种通用的字符。此外，上海交通大学建立了“法律条目全文数据库”，陕西省中医研究院建立了中医经典古籍《素问》、《灵枢》、《甲乙》、《难经》的全文数据库，江苏省中医研究所建立了《伤寒论》、《金匮要略》、《脾胃论》等20余本中医古籍的全文数据库，深圳大学建立了古典文学名著《红楼梦》的全文数据库。所有这些全文数据库都为用户提供了有效的检索服务，也为汉字全文检索系统的进一步发展奠定了基础。

全文文本检索是西文情报检索软件普遍实现的基本功能。瑞典的PROLOG公司研制的TRIP全文检索软件具有全面的全文文本检索功能。1988年，中国科技信息研究所与该公司合作，实现了TRIP系统的汉化。汉化TRIP系统的特点是：以每个汉字单字切分（最简单的汉语书面语自动切分）实现全文检索功能，可按字段（作者、标题、分类、日期、标引词等）检索，可用命令方式和菜单方式检索，可在主题词控制下进行检索。这一系统的缺点是空间开销偏高，不能自动抽出关键词。目前这一系统只能在VAX/VMS计算机上运行，有一定的局限性。该系统已在中国科技信息研究所用于建立“中国学术会议论文数据库”和“中文科技期刊联合目录系统”，又被北方交通大学用来为《经济日报》建立了“《经济日报》新闻资料检索系统”。汉化TRIP全文检索系统的开发和应用，为中文全文文本的检索提供了可行的技术途径和有益的实践经验。如果以汉化TRIP全文文本检索系统为基础，在系统的存贮部分适当地增加关键词自动抽词功能，在系统的检索部

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

分适当增加后控主题词表的管理和检索功能，将大大地提高这一软件对中文全文检索的适应能力。

10. 汉语语音识别系统

我国在离散单词、简单口令的语音识别方面已经取得不少进展。中国科学院声学研究所于50年代后期就研制出汉语单元音识别装置。60年代对汉语的清晰度进行过系统的实验，取得了基本数据。70年代末、80年代初，采用模式匹配的方法，事先存入发话人的语音作成标准模式，计算机可识别该特定说话者的几十条口令，内容包括数字、算术四则运算符号及一些操作指令。1980年，清华大学计算机系采用模式匹配法研制成我国30个大城市的地名识别系统，只要进行口呼地名输入，计算机就可以显示汉字。他们还于1984年建成“800台电话声控查号系统”，用于清华大学校内电话查号，已经投入实用。用户查询电话时，需由话务员复述单位名称，并由话务员通过自己的语音把单位名称报给计算机，计算机屏幕上就显示出该单位的电话号码，并可通过语音合成装置将号码自动地报给用户。1986年，清华大学计算机系在长城0520C-H国产微型机的汉字编码输入的基础上，增加了汉字语音输入方式，他们研制的汉字语音输入系统具有约1000个汉字的字表，在这个字表内的字以及由这些字组成的词，都可以通过语音输入到计算机中去，操作者无须经过专门训练，只要预先念一遍字词，让计算机熟悉其口音就行了，语音识别的正确率为90%，字表的内容还可以根据使用领域任意确定。中国科学院声学研究所研制出“汉语孤立字全音节实时识别系统”，该系统可识别1300个汉语全音节，分为四声识别、辅音粗识别和音节细识别三个层次。四声识别的正确率达到99.4%。辅音粗识别主要用来提取辅音强频区的分布、清辅音的长度、声母与韵母的时长比等辅音的音征，根据音征从全部辅音中选出候选声母，起到粗分类的作用。在粗分类之后进行音节识别，只限定识别包含上述6个候选声母的那些音节。这样做既可以节约匹配时间，又可提高识别的正确率。该系统在1988年西欧高技术展览会(TEC-88)上获得国际大奖，在此基础上，已制成语音打字机。清华大学研制了“大词汇量汉语语音识别系统”，该系统采用分段矢量量化和分段概率模型，没有专门分割声母和韵母的步骤，但在建立矢量码本时以及在识别策略上，都考虑了二者的区别。该系统采取了两级匹配的策略，先是计算音节匹配的概率，继而计算词组匹配的概率，系统中建有单音节字表、双音节至四音节词表，可以直接口呼词进行识别，识别精度高，响应速度快。中国科学院自动化研究所研制了“汉语大词汇量语音识别与口呼文本输入系统”，以声韵调为基元来进行语音识别，识别时采用了隐马尔可夫模型及人工神经网络方法。

我国在非特定说话者语音识别方面也取得了进展。清华大学研制成功非特定说话者中词汇量语音识别系统。非特定说话者的语音识别难度很高，识别时要强调众多说话者的语音共同参数，采用类聚和模糊处理使其具有一般性，并要解决语音多变性和语流速度变异问题，采用更为有效的时间规正技术。采用这样的语音识别系统，使用者不必经过训练，在400多个词汇的范围内，有很高的识别率。另外，清华大学还研制成基于神经网络方法的非特定说话者小词汇量语音识别系统，以30个军事用语作试验，使用者不必经过训练，识别正确率接近100%。北京四达技术开发中心和哈尔滨工业大学合作，研制了汉语语音识别系统“四达863A”。该系统以单音节作为语音识别的基本单元，选择398个无声调单音节作为语音识别的基本内容，这398个单音节包含了国家标准一、二级汉字库中所有汉字的语音。用户在初次使用该系统时需要作短暂的训练，因此，该系

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

统是认人的。该系统还把语音识别技术与拼音汉字简单转换技术结合起来，使用者只需朗读所要输入的汉字，属于同一音节的若干个汉字由拼音－汉字转换程序来确定是哪一个汉字。“四达863A”系统的一次识别正确率超过93%，系统的响应时间小于0.1秒，四个声调的识别正确率为99%，每分钟可口呼输入80个汉字。

IBM公司最近推出的Via Voice汉语语音识别系统，已经达到实用水平。

11. 汉语语音合成系统

中国科学院声学研究所与瑞典皇家工学院语言通信和音乐声学系合作，于1983年研制成“汉语文语转换系统”，采用规则合成方式来合成汉语语音。该系统首先分析了汉语的语音频谱和音位规则，建立了合成规则。可以通过键盘或光电阅读装置输入用汉语拼音拼写的文章，让计算机根据合成规则，读出合成后的语音。该系统还可以根据句型调整语调，根据句子中某些单词上标出的着重点进行重读，它合成语音的词汇量是无限的，已经可以用计算机来朗读故事。

中国社会科学院语言研究所近年来从声学语音学和发声语音学两方面入手，研究汉语语音特征，以提高合成语音的自然程度，在单元音和复合元音的研究方面已取得一定成绩，建立了汉语普通话规则合成系统。

清华大学计算机系在文语转换系统的研制中，采用了以词为单位的合成策略，这个系统不但能够合成单字的语音，而且还能够根据对文章的理解进行自动切词，并根据语言的上下文和音变规则确定正确的发音，将书面的文本按单词的自然停顿实时地读出来，保持了自然语言的韵律，提高了文语转换的可懂度和自然度。

12. 汉字识别系统

我国自70年代开始汉字自动识别的研究，自1986年以来取得了很大的成绩。联机手写体汉字识别已经商品化，有些产品的性能达到了国际水平，识别的汉字字数为6763-12000个，初次使用的识别正确率为80%左右，经常使用可达95%以上，识别速度基本上能跟上人的书写速度。清华文通信息技术公司研制的“文通笔”，可以用来直接书写汉字输入计算机，用户用不着学习任何汉字输入法，只要会写汉字，就可以在书写板上把汉字输入到计算机中。

印刷体汉字识别也开始实用化。有十多个单位推出了实用化系统，可识别国家标准的1级和2级简体汉字3755到6763个，繁体汉字5401个；可识别的汉字字体，简体有宋、仿宋、报宋、黑、楷以及多体混排，繁体有明、楷、仿、黑等，也可以识别多体英文混排；识别速度用286微机时为每秒9-14个汉字，用386微机时为每秒20个汉字；识别正确率，对低等质量的印刷品为95%以下，对中等质量的印刷品为98%-99%，对高等质量的印刷品则达到99%以上；输入设备大多采用普及型图形扫描器或传真机，能识别印刷体的字号为3号到5号。这些系统配备了方便的用户界面，能够进行版面分析、文本识别、识别结果的后处理、自动纠错、编辑、输出等。

脱机手写印刷体汉字和无书写限制的脱机手写体汉字的识别近几年也进行了许多研究，建成了一些试验系统。现已有近于实用的交互式自学脱机手写体汉字识别系统，可识别国标一级汉字3755个，如果加上专用特征库就可识别不加任何书写限制的汉字，识别速度用386微机时为每秒1个汉字。

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

由于我国的汉字识别系统几乎都是在汉字操作系统下工作的，识别结果为汉字内码，因而可以把识别出的汉字直接在计算机上显示或打印出来。

汉字识别如果不是仅仅局限于一个字一个字地孤立地进行模式匹配，而且还能利用词以及上下文关系的信息，那么将会显著地提高识别的正确率。例如，在汉字识别系统中，可利用汉字单词和词组的信息来进行自动纠错，利用语言知识修改部分误识字，利用词的联想来修改误识字和拒识字，在这些方面都获得了很好的识别效果。因此，把自然语言计算机处理的技术应用到汉字的自动识别中，将会使汉字自动识别系统如虎添翼。

对于汉字自动识别系统也进行了863评测。评测结果如下：

印刷体汉字文本识别可以分为简体汉字文本识别和繁体汉字文本识别两类。在简体汉字文本识别方面，参加评测的有清华大学电子工程系和中自汉王公司两个单位。对于质量较差的文本，清华大学电子工程系的识别率为95.61%，中自汉王公司的识别率为95.23%。对于质量较好的文本，清华大学电子工程系的识别率为98.46%，中自汉王公司的识别率为98.26%。使用奔腾II/266微型计算机和Windows 95操作系统，清华大学电子工程系的识别速度为95字/秒，中自汉王公司的识别速度为68字/秒。在繁体汉字文本识别方面，参加评测的单位只有清华大学电子工程系，识别率为97.27%，识别速度为71字/秒。这些测试结果表明，我国的印刷体汉字识别的技术已经相当成熟。

如果在汉字文本中加入表格，识别的难度就会增加，因此，我国还进行了印刷体表格识别的评测，参加评测的有清华大学电子工程系、智能计算机研究开发中心、北京信息工程学院三个单位。表格分析正确率都比较高，清华大学电子工程系的表格分析正确率为84.89%，智能计算机研究开发中心为92.38%，北京信息工程学院为84.48%。但是，如果在汉字中加入表格，尽管单独的表格分析正确率比较高，但是，汉字的识别率却大大下降。清华大学电子工程系的识别率为64.37%，智能计算机研究开发中心的识别率为77.36%，北京信息工程学院的识别率为72.63%。看来，对于带表格的汉字自动识别还有待加强。

手写体数字识别的评测也有较好的成绩。参加评测的清华大学电子工程系、中国科学院自动化所、清华大学计算机系，识别正确率都在95%以上。

脱机手写体汉字识别难度较大。在脱机手写体汉字单字识别方面，参加评测的单位有六个：中自汉王公司的识别率为88.87%，中国科学院自动化所的识别率为75.85%，北京邮电大学信息系的识别率为75.81%，清华大学计算机系的识别率为69.60%，武汉工业大学的识别率为67.37%，清华大学电子工程系的识别率为64.86%。在脱机手写体文本识别方面，参加评测的单位有四个：中自汉王公司的识别率为95.44%，清华大学计算机系的识别率为80.62%，武汉工业大学的识别率为73.49%，清华大学电子工程系的识别率为64.86%。在脱机手写体汉字人民币大写字符识别方面，参加评测的单位有六个：中自汉王公司的识别率为99.90%，北京邮电大学信息系的识别率为99.39%，中国科学院自动化所的识别率为99.15%，武汉工业大学的识别率为98.90%，清华大学计算机系的识别率为98.74%，清华大学电子工程系的识别率为95.86%。中自汉王公司在脱机手写体汉字识别的各项评测中都处于领先地位。

联机手写体汉字识别的测试样本分为工整手写体样本和自由手写体样本两类。参加评测的单

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

位有六个，其中一个单位在测试中出现故障，无法继续进行现场测试，其他各单位的评测结果如下：

在工整手写体联机汉字识别方面，中自汉王公司的识别率为94.40%，清华大学电子工程系的识别率为94.63%，台湾蒙恬公司的识别率为89.79%，北京大学计算所的识别率为89.05%，美国摩托罗拉公司的识别率为88.19%。

在自由手写体联机汉字识别方面，中自汉王公司的识别率为91.04%，清华大学电子工程系的识别率为87.50%，台湾蒙恬公司的识别率为82.60%，美国摩托罗拉公司的识别率为76.49%，北京大学计算所的识别率为74.88%。

中自汉王公司在联机手写体汉字识别的各项评测中，也处于领先地位。

计算机中文信息处理技术的发展任重道远，我们要为在中国实现初步的信息化，为在中国不同行业、不同背景下的人广泛应用计算机信息技术建立起中文信息处理技术的计算机文化基础，我们还要发展中文信息处理的基础技术，优化键盘输入技术，实现手写和语音汉字输入技术，突破汉字、词和句的信息处理技术，发展人工智能技术，逐步实现图象、汉字、声音宽带高速传输技术，实现互联网上的多文种机器翻译技术，克服信息网络时代的语言障碍，为迎接本世纪高度的信息化社会奠定坚实的基础。在全世界范围内，本世纪必将是计算机文化的世纪，计算机将真正成为一种文化载体，普及到各行各业，进入到千家万户。在汉字文化圈内，一个高度信息化的计算机汉字文化的新时代必将到来。

我国在中文信息处理方面的研究是与我国的语言学研究密切相关的。希望我国的语言学工作者能够注意这方面的研究，迎接这个计算机文化的新时代。

参考文献

陈肇雄主编，1992，《机器翻译研究进展》北京：电子工业出版社。

董亦农等，1995，MMT(ODA)项目中基于中间语言的分析和生成的机制。《中文信息学报》第4期。

董振东，1998，语义关系的表达和知识系统的建造。《语言文字应用》第3期。

冯志伟，1979，形式语言理论。《计算机科学》第1期（创刊号）。

冯志伟，1984，汉－法／英／日／俄／德多语言自动翻译试验。《中国的机器翻译》上海：知识出版社。

冯志伟，1985，机器翻译的困难性和它的工程化。《情报学报》第4期。

冯志伟，1989，《现代汉字和计算机》北京：北京大学出版社。

冯志伟，1991，Martin Key的功能合一语法。《国外语言学》第2期。

冯志伟，1995，《自然语言机器翻译新论》北京：语文出版社。

冯志伟，1997，《现代术语学引论》北京：语文出版社。

冯志伟，1999，《应用语言学综论》广州：广东教育出版社。

冯志伟，1999，中国情报检索的历史和现状。《资讯传播与图书馆学》（台湾）第5卷，第4期。

冯志伟，2000，《术语浅说》北京：语文出版社。

国家标准GB13715，1992，《信息处理用现代汉语分词规范》北京：中国标准出版社。

林杏光等，1994，《现代汉语动词大词典》北京：北京语言学院出版社。

刘开瑛，1993，中文全文检索技术研究。《计算语言学研究与应用》北京：北京语言学院出版社。

刘开瑛，2000，《中文文本自动分词和标注》北京：商务印书馆。

刘涌泉、刘倬、高祖舜，1962，俄汉机器翻译规则系统新旧方案比较。《中国语文》第10期。

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

刘志杰、刘倬，1997，基本词典与专业词典的关系。《语言工程》北京：清华大学出版社。
钱文浩，1956，文字与通讯。《科学通报》10月号。
宋柔等，1993，基于语料库和规则库的人名识别法。《计算语言学研究与应用》北京：北京语言学院出版社。
孙茂松、黄昌宁等，1997，利用汉字二元语法关系解决汉语自动分词中的交集型歧义。《计算机研究与发展》第5期。
王永成，1992，《中文信息处理技术及其基础》上海：上海交通大学出版社。
吴蔚天等，1994，《汉语计算语言学—汉语形式语法和形式分析》北京：电子工业出版社。
吴文虎，1992，汉语语音识别的现状与展望。《语文建设》第6期。
杨顺安，1992，语音合成与语音学研究。《语文建设》第8期。
俞士汶、朱学锋、E.Kaske、冯志伟，1996，《英汉对照计算语言学词语汇编》北京：北京大学出版社。
俞士汶、朱学锋等，1998，《现代汉语语法信息词典详解》北京：清华大学出版社。
张潮生等，1995，中文信息全文检索系统。《中文信息处理应用平台工程》北京：电子工业出版社。
张忻中，1992，计算机汉字识别技术。《语文建设》第10期。

作者通讯地址：100010 北京朝内南小街51号 国家语言文字工作委员会语言文字应用研究所

英国学术院院士 Geoffrey N. Leech 教授来华讲学

英国著名语言学家英国学术院院士 Geoffrey N. Leech 教授应中国社会科学院语言研究所和北京外国语大学的邀请，定于2001年5月18-26日来北京讲学，内容包括(1)“语料库语言学与计算语言学 —研究与实践研讨会”主题演讲；(2)“中国外语教学研讨会”主题演讲；(3)中国社会科学院语言研究所学术报告。

美国著名语言学家 William Labov 等将来华讲学

北京语言文化大学出版社将于2001年5月14-16日举办“海外著名语言学家讲习所”第一期，邀请美国著名语言学家威廉·拉波夫(William Labov)教授、安东尼·克洛克(Anthony Kroch)教授来华讲学(用英语讲，配翻译)。希望有意聆听者于2001年3月31日前与接待单位取得联系，以便妥善安排。

联系人：王飙 / 侯明　　电话：(010) 82303647　　邮编：100083

E-mail: wangbiao@blcu.edu.cn　houming@blcu.edu.cn

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

《当代语言学》第3卷2001年第1期77-78页，北京

Abstracts of Articles

Feng, Zhiwei, The computer processing of Chinese characters and Chinese language

The author describes the computer processing of Chinese characters and Chinese language. The following aspects are discussed in detail: the various input approaches of Chinese characters and Chinese corpus, the automatic segmentation, the automatic POS tagging, the automatic phrase bracketing and syntactic annotation of Chinese written text, electronic dictionary, terminological databank, machine translation, computer-aided proofreading, automatic information retrieval, automatic Chinese speech recognition and synthesis, and automatic Chinese characters recognition.

Sun, Maosong, and Zou Jiayan, A critical appraisal of the research on Chinese word segmentation

This paper firstly discusses the practical significance and feasibility of word segmentation system for unrestricted Chinese texts, then focuses on three basic issues in the field, that is, segmentation ambiguity disambiguation, unknown word processing and language resource construction. The history of the researches and the important methods developed in the past are reviewed and assessed. Suggestions for future study are proposed.

Dong, Zhendong, and Dong Qiang, Construction of a knowledge system and its impact on Chinese research

The paper presents a full account of how-net, viz. a knowledge system for natural language processing. How-net has been recently released in the Internet. The knowledge dictionary of how-net contains 50,000 Chinese word forms, 62,000 Chinese concepts and 55,000 English equivalents with 70,000 English concepts. The paper also covers some important issues on the building of relational semantic net. The authors make a detailed discussion about the impact and enlightenment of how-net on the Chinese language, among which the key issue is how to construct Chinese semantic syntax.

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net