# 中文信息处理六十年[1]

宗成庆　曹右琦　俞士汶

**摘要**　中文信息处理是自然语言处理领域的一枝奇葩，近 60 年来不断放射出奇光异彩，吸引着众多语言学家、计算语言学家和从事自然语言处理技术研究、开发的工程人员为之奋斗，取得了丰硕成果。本文首先简要回顾中国语文现代化走过的历程、取得的重要成果及其对中文信息处理的影响，然后对汉字信息处理和汉语信息处理的其他工作予以归纳阐述，并对这一领域的学术活动与国际交流情况做简要介绍，最后对中文信息处理所面临的挑战和未来发展的目标给予粗略的展望。作者希望本文提供的信息和阐述的观点能够对未来中文信息处理的发展起到一定的借鉴和促进作用。

**关键词**　中文信息处理，自然语言处理，自然语言理解，计算语言学

# Sixty Years of Chinese Information Processing

ZONG Chengqing, CAO Youqi, YU Shiwen

**ABSTRACT**　The Chinese Information Processing (CIP) is a charming flower in the garden of natural language processing (NLP). In the recent sixty years, she has been emitting fantastic colors and always attracts so many linguists, computational linguists, and engineers working with NLP to devote themselves to working for her. Fruitful results have been made in the past decades. This paper first briefly reviews the process of Chinese language modernization in China, the important results, and the affects on CIP, and then summarizes the related work on Chinese character information processing and Chinese language information processing as well. The academic activities and international exchanges in this area are also introduced. Finally, the challenges that CIP faces to are roughly analyzed and the future development is simply prospected. The authors sincerely hope the information and the viewpoints given in this paper may promote the study of CIP and could be well used for reference.

**KEYWORDS**　Chinese information processing, Natural language processing, Natural language understanding, Computational linguistics

[1] 本文的相关研究工作得到国家自然科学基金项目（60736014）、“863”计划项目（2006AA010108－4）、国家支撑计划项目（2006BAH03B02）和国家 973 课题（2004CB318102）的支持。

## 1. 引　言

顾名思义，“中文”就是中国的语言文字。从广义上理解，她可以是中国各民族使用的所有语言文字的总称。但是，由于汉族在人口数量和地域分布上都占有绝对优势，而且长期以来，中国境外（如新加坡、马来西亚等）华人使用的汉语文字被称为华文或中文，因此，在不引起混淆的情况下，我们认为“中文”与“汉语”指同一概念。根据国家标准 GB12200.1－90“汉语信息处理词汇 01 部分：基本术语”的解释，“**中文（Chinese）**”特指**汉语**。本文不涉及民族语言文字信息处理的内容，拟另请其他专家撰文介绍。

中文信息处理是自然语言处理领域的一枝奇葩，几十年来不断放射出奇光异彩，吸引着众多语言学家、计算语言学家和从事自然语言处理技术研究、开发的工程人员为之奋斗，取得了累累硕果。尤其近 20 年来，随着计算机网络和手机等现代通讯技术的迅速发展与普及，自然语言处理成为计算机科学与语言学交叉领域研究的热点。伴随我国经济实力和国力的不断增强，汉语在世界范围内逐渐成为一种继英语之后的强势语言，世界华人和中国市场对自然语言处理技术的巨大需求，吸引着众多科学家和企业界的目光。因此，不管是发达国家还是落后国家，没有人敢忽视或藐视汉语。中文信息处理技术已经不再是中国人自己关注的问题，而成为整个国际自然语言处理领域共同关注的焦点。

本文其余部分首先简要回顾中国语文现代化所走过的主要历程、取得的重要成果及其对中文信息处理的影响，然后重点对汉字信息处理和汉语信息处理的方方面面予以归纳阐述，并对这一领域的学术活动和国际交流情况做简要介绍，最后对中文信息处理目前所面临的挑战和未来发展的目标给予粗略的展望。

## 2. 早期语文现代化工作回顾

语言文字的信息化或者说语言文字信息处理技术的发展水平是关乎国家现代化、社会信息化的大事。中国语文现代化的早期工作及其成果对后来汉字信息处理技术的发展起到了奠基性的作用[1]。

中国语文现代化的开始可以追溯到中华人民共和国建国前后。1949 年 8 月 7 日吴玉章等发起组织了中国文字改革协进会，同年 10 月 10 日新中国的第一个全国性文字改革组织－中国文字改革协会宣告成立。1952 年 2 月 5 日新中国第一个主管文字改革工作的国家机构－中国文字改革研究委员会成立。整理和简化汉字成为中国文字改革研究委员会的既定工作任务之一。1956 年 1 月国务院通过了《关于公布⟨汉字简化方案⟩的决议》，并正式公布首次《汉字简化方案》。1986 年 10 月经国务院批准决定，国家语委重新发表了《简化字总表》，共收 2235 字，对原《简化字总表》中的个别字做了调整。1988 年国家语委、国家教育委员会分布《现代汉语常用字表》，1997 年国家语委、新闻出版署发布《现代汉语通用

字笔顺规范》[2]。

如果说汉字简化与规范化对汉字信息处理有重要影响，其重要意义更多地体现在有利于汉字教学和应用、提高国民文化水平，那么，与其并称为文字改革三大核心任务的汉字拼音化和推广普通话则是汉字信息化进程中不可或缺的关键环节。从 1958 年 2 月 11 日全国人大一届五次会议通过决议，正式批准《汉语拼音方案》，到 1982 年汉语拼音被国际标准化组织（ISO）接纳，成为拼写汉语的国际标准，以及 1984 年 10 月中国文字改革委员会发表《汉语拼音正词法基本规则（试用稿）》和 1996 年正式颁布国家标准（GB/T 16159-1996）《汉语拼音正词法基本规则》。汉语拼音的推广使用对于普及汉字和汉语教学、促进国际交流起到了非常重要的作用，对中国社会生活的各个方面都产生了极其深远的影响[3][4]。尤其值得指出的是，汉语拼音对于计算机汉字输入和中文电脑普及起到了至关重要的作用。50 多年的实践证明，汉语拼音方案是既能体现拉丁化优点，又符合汉语汉字本身特点的最优方案[5]。

纵观几十年来中国语文现代化的历史，老一代专家学者高瞻远瞩的战略思想和一系列英明举措对汉字信息处理技术的发展起到了重要的奠基作用。勿庸置疑，语文现代化与社会信息化、知识经济化有着密切关系。或许可以说，语文现代化是对我国工业现代化、农业现代化、国防现代化和科学技术现代化的重要补充，其历史意义和现实意义不容低估。

## 3．汉字信息处理概要

我们知道，语言和文字既是信息、知识、文化的载体，也是文化的组成部分[6]。汉字作为中华民族璀璨文化中独具特色的一项发明在数千年一脉相传的源源历史中，为记载、继承和传播中华文化建立了不朽的功勋。然而，当 20 世纪 40 年代电子计算机问世，并迅速引发席卷全球的信息技术革命，如何对汉字进行编码、存储、输入和输出等一系列关于汉字处理的难题，曾一度成为电脑在中国普及和推广的“拦路虎”。因此，从上个世纪 70 年代中期到 80 年代末期，汉字信息处理技术成为当时的研究热潮。

汉字信息处理主要指以汉字为处理对象的相关技术，包括汉字字符集的确定、编码、字形描述与生成、存储、输入、输出、编辑、排版以及字频统计和汉字属性库构造等等[6]。一般而言，汉字信息处理关注的是文字（一种特殊的图形）本身，而不是其承载的语义或相互之间的语言学关系，因此，本文将其分离出来单独介绍，而后面将要重点介绍的“汉语信息处理”部分则是指对传递信息、表达概念和知识的词、短语、句子、篇章乃至语料库和网页等各类语言单位及其不同表达形式的处理技术。

在汉字信息处理中，有两个问题最引人注目，一是汉字的输入问题，二是

汉字的排版、印刷问题。其中，汉字输入问题又分为键盘输入和非键盘输入两种。所谓键盘输入是指通过对汉字进行“编码”，即利用普通计算机键盘上的英语字母键之间的组合，建立与汉字之间的对应关系，并将这种对应关系以编码对照表的形式存储在计算机内部，最终利用转换软件将键入的字符串转换为对应的汉字。最早的计算机汉字编码输入始于上个世纪 50 年代的俄汉机器翻译研究，当时只能用电报码和四角号码做汉字编码。60 年代完成了“见字识码”的方案设计和码本。1978 年 5 月上海推出了一台汉字信息处理实验样机。20 世纪 80 年代，在联想汉卡、四通中文电脑打字机之后，中国的汉字编码出现了“万马奔腾”的局面，从“五笔字型”，到自然码、郑码、拼音输入法、智能 ABC、智能狂拼等，较规范、易学易用的输入法层出不穷。国家七五、八五重点科技攻关项目“PJS 普及型中文输入系统”、“规范码汉字输入系统”和“认知码”等都对汉字编码输入方法进行了深入研究，并取得了一批研究成果[7]。尤其值得提及的是，速记专家唐亚伟先生发明的亚伟中文速录机，实现了由手写速记跨越到机械速记的历史性突破，这一成果被迅速推广应用，催生出了速录行业和速记师职业。2005 年 92 岁高龄的唐亚伟获得我国中文信息处理领域的最高科学技术奖—钱伟长中文信息处理科学技术奖一等奖。

非键盘输入是指不借助键盘直接将汉字或数字等字符输入计算机的技术，常用的方法包括文字识别、语音识别等。汉王文字识别技术是一个成功的代表。

以北京大学王选院士为代表的从事汉字照排和印刷技术研究的老一代专家在解决巨量汉字字形信息存储和输出等问题中做出了卓越贡献。1981 年第一台汉字激光照排系统“原理性样机”通过鉴定，1985 年激光照排系统在新华社正式运行。1987 年《经济日报》采用激光照排系统出版了世界上第一张采用计算机屏幕组版、整版输出的中文报纸，成为国内第一家全部废除铅字排版的报纸。此后，国产激光照排系统迅速推广应用，在中国掀起了“告别铅与火，迎来光与电”的印刷技术革命[7]。

另外，上个世纪 80 年代完成的《汉字频度表》、《现代汉语频度词典》、GB2313－80、6763 汉字属性信息库等一系列基础性工作，都为后来的汉语信息处理研究奠定了很好的基础。

## 4．汉语信息处理技术成果与应用

本部分重点关注在汉语词、短语、句子、篇章乃至语料库等各类语言单位处理方面所取得的研究成果及应用情况。为了便于描述，我们将其分为“基础资源建设”、“理论方法研究”和“应用技术开发”三个方面。

### 4.1 基础资源建设

语言资源库（包括语料库、词汇知识库、语法语义词典等）在不同层面构成

了自然语言处理各种方法赖以实现的基础，有时甚至是建立或改进一个自然语言处理系统的“瓶颈”。因此，世界各国对语言资源库的开发建设都投入了极大的关注。自 1979 年以来，中国开始进行机读语料库建设，并先后建成汉语现代文学作品语料库（1979 年，武汉大学，527 万字）、现代汉语语料库（1983 年，北京航空航天大学，2000 万字）、中学语文教材语料库（1983 年，北京师范大学，106 万字）和现代汉语词频统计语料库（1983 年，北京语言学院，182 万字）[8]。近 20 多年来，北京大学、清华大学、教育部语言文字应用研究所、山西大学、哈尔滨工业大学、北京语言大学、东北大学、中科院自动化所、科技部中信所、中国传媒大学、台湾中央研究院和香港城市大学等相当一批大学和研究机构都对汉语资源库建设做了大量工作。其中，北京大学计算语言学研究所开发的“综合型语言知识库”、董振东等开发的“知网”（HowNet）是两项有代表性的成果，而中文语言资源联盟（Chinese Language Data Consortium, 缩写：Chinese LDC[2]）则是为推动我国语言资源共享所建立的第一个联盟性学术组织。

**（1）综合型语言知识库**

北京大学计算语言学研究所的语言资源建设工作始于 1986 年，从研制《现代汉语语法信息词典》[9]起步。该词典曾获 1998 年度教育部科技进步奖二等奖。在此成果的基础上，于 1995 年提出建立“综合型语言知识库”的规划，经过十多年的努力，“综合型语言知识库”取得了阶段性成果，并于 2007 年获教育部科技进步奖一等奖。

申报奖励之前，“综合型语言知识库”通过了教育部组织的技术鉴定：“其规模、深度、质量和应用效果在我国语言工程实践中是前所未有的。该成果是以汉语为核心的多语言知识库建设中最全面、最重要的研究成果，总体上达到了国际领先水平”。该项成果为推动以汉语为核心的多语言信息处理技术的发展做出了重要的贡献，并取得了显著的经济效益。作为单项技术成果，在北京大学创下了转让次数最多的纪录。

“综合型语言知识库”[10]在汉语计算语言学理论、汉语语言知识形式化描述、语言知识库构建技术以及多语言知识融合技术等方面都有所创新。目前它包含的语言资源包括现代汉语语法信息词典、现代汉语语义词典、中英文概念词典、汉语短语结构知识库、现代汉语大规模基本标注语料库、汉英双语对齐语料库以及多个专业领域的术语库。

“综合型语言知识库”仍在继续发展。研制中的“综合型语言知识库系统”不仅把现有的语言知识资源集成为一个有机的整体，各个成员知识库可以相互参照，互相印证，而且进一步挖掘深层的语言知识，发展概率型汉语词汇知识库，让语言知识库建设更上一层楼，同时，将有新成员不断加入“综合型语言知识库”

[2] www.chineseldc.org

的大家庭。

**（2）知网[3]**

知网（HowNet）是董振东教授提出并创建的语言知识库，是一个以汉语和英语的词语所代表的概念为描述对象，以揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库。

知网作为一个构思严密的知识系统，是一个名副其实的意义网络，它着力要反映的是概念的共性和个性。在知网中，义原是一个很重要的概念，他是指最基本的、不易于再分割的意义的最小单位。知网体系的基本设想是，所有的概念都可以分解成各种各样的义原，同时，也存在一个有限的义原集合，其中的义原组合成一个无限的概念集合。董振东教授认为，中文中的字（包括单纯词）是有限的，并且它可以被用来表达各种各样的单纯的或复杂的概念，以及表达概念与概念之间、概念的属性与属性之间的关系。因此，知网从大约六千个汉字中提取出了这个有限的义原集合[11][12]。知网的规模主要取决于双语知识词典数据文件的大小。由于它是在线的，修改和增删都很方便，因此，它的规模是动态的。目前知网已作为中文信息处理技术研究和系统开发重要的基础资源，被广泛地应用于词汇语义相似性计算、词义消歧、名词实体识别和文本分类等许多方面。

**（3）中文语言资源联盟**

在国家重点基础研究发展规划项目（973 项目）“图象、语音、自然语言理解与知识挖掘”（资助号：G19980305）的支持下，由中科院自动化所、清华大学、教育部语用所和中科院计算所发起，于 2003 年成立了中文语言数据联盟。该联盟挂靠在中国中文信息学会，其目标是建成具有国际水平的具有完整性、系统性、规范性和权威性的通用中文语言资源库以及中文信息处理的评测体制，为汉语语言信息处理的基础研究和应用开发提供支持，促进汉语语言信息处理技术的不断进步[13]。目前该联盟已拥有会员单位 70 多个、各类语言资源 80 余种，包括 8~10 万词的《汉语通用词表》、25000~30000 词的《汉语语法信息词典（高频词）》、500 万字的《分词词性标注语料库》、100 万字的《汉语语法树库》、20 万句对的《中英双语语料库》，等等。其中 30%数据资源对会员免费，从而在全世界范围内实现中文语言数据资源的共享。Chinese LDC 于 2006 年运营以来，平均每天都有数十人次的网站访问和电话咨询。到目前为止，该组织已共享资源 200 多套，授权评测单位使用 40 多个，包括美国、加拿大、德国、日本、澳大利亚等国内著名科研机构和公司若干单位已经通过该平台获取了中文信息处理科研工作所需的基础资源[14][15]。

[3] http://www.keenage.com/html/c_index.html

## 4.2 理论方法研究

我国最早利用计算机进行自然语言处理研究的项目是机器翻译。1956 年国家把机器翻译研究列入科学工作发展规划并设立课题，1957 年中科院语言所和计算所合作开展了俄汉机器翻译研究。机器翻译是一个高度综合性的研究课题，涉及词法分析、句法分析、语义分析和语言生成等各个层面，因此，伴随机器翻译研究，中文信息处理相关的各种理论方法研究随之展开。在过去 50 多年的曲折历程中，中文信息处理理论研究的脚步从来都没有停止过。

1958 年刘涌泉、刘倬等提出的“中介成分理论”曾在早期的中国机器翻译研究中发挥了重要的作用。70 年代末期冯志伟最先开展了对汉字信息熵的研究，经过几年的语料收集和手工统计，在当时艰苦的条件下测定了汉字的信息熵为 9.65 比特（bit），这与 80 年代末期北京航空学院刘源等通过计算机对大规模语料统计得到了汉字信息熵为 9.71 比特的结论相当接近。

进入上个世纪 80 年代以后，汉语分词与词性标注方法研究得到了快速发展。全切分分词方法、最短路径分词方法、N-最短路径分词方法、基于隐马尔可夫模型（HMM）或 *n* 元语法（*n*-gram）的分词方法等一系列分词方法相继提出。1992 年《信息处理用现代汉语分词规范》被国家技术监督局批准（GB13715），并于 1993 年 5 月 1 日在全国正式实行[16]。

1990 年代，面向机器翻译提出的 SC 文法[17]，从某种意义上拓展了复杂特征集理论和合一文法，而《现代汉语语法信息词典》和《知网》是我国学者结合汉语特点和规律对词汇主义思想的进一步发展和应用。

另外，概念层次网络理论的提出也是中文信息处理研究中一个有益的探索。

## 4.3 应用技术开发

相对于理论方法研究而言，中文信息处理应用技术开发和产业化进程中的成果可谓琳琅满目。除了前面提到的汉字存储、显示、输入、激光照排等实用技术以外，机器翻译、搜索引擎、文语转换等应用系统也如雨后春笋不断涌现。

上个世纪 80 年代中期到 90 年代初期，我国的机器翻译研究开始走向繁荣。军事科学院研制的“KY-1”英汉机器翻译系统获得了国家科技进步二等奖，后来发展为“译星”，成为中国第一个商品化机器翻译系统。中科院计算所研制的“IMT/EC863”英汉机器翻译系统于 1995 年荣获国家科技进步一等奖，获得了可观的经济效益。

进入 21 世纪以后，基于大规模语料库的统计方法在自然语言处理中得到快速发展，以语料库为研究对象和基础的语料库语言学（corpus linguistics）迅速崛起，并反过来进一步推动了自然语言处理相关技术的快速发展，统计机器翻译逐渐成为国际机器翻译研究的主流。中科院计算所、自动化所、哈尔滨工业大学、厦门大学和中科院软件所等在统计机器翻译研究中进行了富有成效的探索和实

践。中科院自动化所还在语音翻译研究方面做了大量开创性的工作，先后实现了基于个人电脑、PDA 和普通手机的汉英、汉日双向语音翻译系统。

近几年来，以机器翻译技术为支柱发展起来的中科院华建集团公司和沈阳格微软件有限公司在机器翻译应用方面取得了十分可喜的成就。

与此同时，在语音识别、语音合成和人机对话系统等方面，中科院自动化所、声学所、中国科大、清华大学、北京交通大学、哈尔滨工业大学等都做了大量研究和开发工作。语音识别、语音合成系统已在实际应用中取得了丰硕的成果。

近 10 年来，随着国际互联网技术的迅速发展和普及，国内一批面向计算机网络的信息搜索系统脱颖而出，TRS[4]、百度[5]和中搜[6]等一批优秀企业成为当前信息领域十分耀眼的明星。

值得提及的是，由国家语言文字工作委员会组织编纂发布的《中国语言生活绿皮书》[7]是当代中国语言规划的一项重要举措，体现着新世纪国家语言文字工作的一些新理念，体现着中国语言研究的一些新进展。编辑出版《中国语言生活绿皮书》的目的，是为国家语言方针政策的决策提供参考，为语言文字研究者、语言文字产品研发者和社会其他人士提供语言服务，引领社会语言生活走向和谐[18]。勿庸置疑，《中国语言生活绿皮书》既是中文信息处理研究成果的具体体现，也是中文信息处理研究的重要参考。

## 5．学术活动与国际交流

随着中文信息处理研究的逐步深入和人才队伍的迅速壮大，由钱伟长、甄健民、安其春等老一代专家发起的中国中文信息学会（Chinese Information Processing Society of China）[8]于 1981 年 6 月宣告成立，成为具有独立社团法人资格的国家一级学会。在学会的引导和支持下，中文信息处理学术活动与交流蓬勃兴起。

每两年一次的全国计算语言学学术会议（CNCCL）（2007 年前的名称为“全国计算语言学联合学术会议（JSCL）”）到 2009 年为止已经举办了十届，是中国中文信息处理领域最具影响力的全国性学术会议。自 2002 年开始的全国学生计算语言学研讨会（SWCL）到 2008 年为止举办了四届，是面向中文信息处理领域学生的全国性学术会议，整个会议由学生组织，深受同学们的喜爱。中日自然语言专家研讨会（CJNLP）自 2001 年起每年召开一次，奇数年在日本召开，偶数年在中国召开，旨在推动中日两国自然语言处理研究的学术交流与合作。自

[4] http://www.trs.com.cn/
[5] http://www.baidu.com.cn/
[6] http://www.zhongsou.com/
[7] 第一部《中国语言生活绿皮书》—《中国语言生活状况报告（2005）》于2006年9月18日正式出版。此后每年发布一次，至今已经连续发布了三年。
[8] http://www.cipsc.org.cn/

2004 年起每年召开一次的自然语言处理青年学者研讨会则着眼于促进青年学者之间的学术交流，加强与国际学术界和企业界的联系。另外，中国中文信息学会下属各专业委员会的学术活动也呈百花齐放之势：每两年一次的人机语音通讯学术会议到 2009 年已经举办了十届；每两年一次的中国少数民族语言文字信息处理学术研讨会 2009 年为第 12 届；每年一次的全国机器翻译研讨会和全国信息检索学术会议到 2009 年均已举办了五届。

值得一提的是，评测对于促进中文信息处理技术的发展起到了非常重要的作用。早在上个世纪 90 年代初期，我国“863”计划中文与接口技术评测组就多次组织汉语分词与词性标注、机器翻译等技术评测。基于测试集与测试点的机器翻译评测系统 MTE 最早实现了译文质量的自动评测[19]。进入本世纪以来，汉语自动分词、词性标注、句法分析、机器翻译、信息检索、文本分类、语音识别、语音合成等针对不同技术和系统的评测如雨后春笋般迅速成长。2003 年国际计算语言学学会汉语兴趣小组（SIGHAN）[9]举办了首届汉语分词技术国际评测（Chinese word segmentation bakeoff），至今已经举办了四次。勿庸置疑，这些评测对于促进同行专家之间的互相交流、互相学习，共同提高，起到了不可替代的作用，同时，评测技术本身也在不断研究和实践中得到了改进和提高。

令人鼓舞的是第 23 届国际计算语言学大会（COLING）[10]将于 2010 年 8 月在北京举办。COLING 是由国际计算语言学学会（ICCL）[11]直接领导组织的学术大会，是国际计算语言学领域参加人数最多、涉及学科范围最广、历史最悠久的国际盛会之一，在 40 多年的风风雨雨中经久不衰，能够获得 COLING 大会的主办权是各国计算语言学专家追求的梦想。几十年来，我国几代计算语言学专家为了实现这一梦想付出了不懈的努力。我们相信，这一盛会在北京的成功举办必将为推动中文信息处理研究的发展产生积极而深远的影响。

伴随中国改革开放的步法，中文信息处理国际交流与合作活动日益增多。早在上个世纪 80 年代，中国参加了由日本发起，印度尼西亚、泰国和马来西亚共同参与的五国多语言机器翻译合作项目，为当时中国机器翻译研究的人才培养、技术传播和资源积累等都产生了重要影响。进入 90 年代以后，尤其是进入 21 世纪以来，包括 IBM、微软、Google、Yahoo、Sohu、富士通、东芝、Nokia、法国电信等在内的一大批国际著名企业纷纷落户中国，在中国设立研究机构，其研究兴趣无不包含中文信息处理，这从另一个侧面为中国大学和研究院所直接与国际企业合作打开了方便之门。

在语音翻译研究中，中科院自动化所自 90 年代中期开始与美国 CMU、日本 ATR、法国 GETA 等开展国际合作，2001 年以核心成员的身份加盟国际语音翻

[9] http://www.sighan.org/
[10] http://www.coling-2010.org
[11] http://nlp.shef.ac.uk/iccl

译先进研究联盟（Consortium for Speech-to-speech Translation Advanced Research International, C-STAR），近 10 年来，参与发起、组织和实施了一系列有关口语翻译的国际学术活动和联合实验。

近年来，随着国际交流的全面展开，一方面一批国际著名的自然语言处理专家频繁来访中国，他们的学术讲座、报告为中国学者开阔了眼见；另一方面，每年都有一大批中国学者走出国门参加包括讲学在内的各种学术交流与合作。在这种互惠互利的国际交往中，中文信息处理技术得到了长足的进步。

## 6．挑战与未来

综上所述，中文信息处理 60 年的辉煌历史产生了一大批令人鼓舞的成果，这些成果概括起来可以归纳为如下几个方面：

（1）语文现代化取得丰硕成果，有关规范化汉字、汉语拼音和普通话的一系列的国家法规、标准及规范已经形成；

（2）汉字信息处理技术已达到实用化水平，并在实际应用中日趋成熟；

（3）已建设完成一批颇具影响的汉语信息处理用语言资源库，部分汉语信息处理技术已在实际应用中发挥作用；

（4）中文信息处理的国内外学术交流与合作环境已经建立，中文信息处理正在世界范围内迎来空前繁荣时期。

然而，在我们看到这些成果的同时，我们不能忘记中文信息处理毕竟是认知科学、语言学和计算机科学等多学科交叉的复杂问题，最终要达到汉语理解的目的，目前仍面临若干尚未解决的难题。首先，语义理解与计算问题成为当前中文信息处理面临的最大挑战。自然语言的语义如何表示？语义是否可计算？如何计算？这些问题仍没有答案。从目前情况来看，仅歧义消解这一个难题就已经让自然语言处理研究者左支右绌，力不从心，更何况人类运用语言还有多种多样的表现手法，诸如隐喻、幽默、夸张、双关、影射等，它们对自然语言理解研究都有深刻的影响。目前对有些问题刚开始研究，有的甚至尚未触及。显然，离自然语言理解这个目标尚有遥远的路要走。要实现机器对语言的理解，必须首先解开人类理解语言机制的秘密，这是有关人类认知机理、智能本质的科学难题[20]。

另外，随着计算机网络和各种通讯技术的迅速发展，许多新的应用需求对自然语言处理技术提出了更高的要求。例如，网络内容管理、信息监控、有害信息过滤和概念搜索等，这些任务不仅与自然语言处理技术有关，而且涉及图像理解、情感计算和网络技术等多种相关技术。而语音自动翻译则是涉及语音识别、机器翻译、语音合成、表情识别与理解以及通讯等多种技术的综合集成技术。面对这些新的任务，研究才刚刚开始，离问题解决的最终目标仍很遥远。

在语言资源库建设方面，至今仍缺乏基本的国家规范和标准，语料库和知识

库开发仍呈现“百家争鸣”的局面，许多成果难以共享和整合。而在理论模型和方法研究方面仍处于探索阶段，尽管许多理论模型和方法已经得到实际应用，如上下无关文法、HMM、噪声信道模型等，但是，许多重要的问题仍未得到彻底、有效的解决，包括汉语自动分词、命名实体识别等经典问题。综观整个自然语言处理领域，尚未建立起一套完整、系统的理论框架体系，许多理论研究甚至处于盲目的摸索阶段，如尝试一些新的机器学习方法或未曾使用的数学模型，这些尝试和实验带有很强的主观性和盲目性。在技术实现上，许多改进往往仅限于对一些边角问题的修修补补，或者只是针对特定条件下一些具体问题的处理，未能从根本上建立一套广泛适用的、鲁棒的处理策略。尤其如何针对汉语自身的特点和规律，建立真正适合中文信息处理的一整套理论体系和实现方法，将是中文信息处理研究者长期面临的严峻挑战。

无论如何，我们相信中文信息处理像其他学科一样，需要经过众多学者长久的、坚持不懈的探索和实践。我们期待着语言学（包括计算语言学）、脑科学、认知科学、智能科学、哲学、数学等各个领域的专家密切合作，在中文信息处理中实现“规则与统计共舞，语言随计算齐飞”。

**参考文献**


[1] 俞士汶、朱学锋，语文现代化与汉语信息处理技术[M]，见苏培成主编《语文现代化论丛（第六辑）》，第 176-189 页。北京：语文出版社，2006 年 9 月

[2] GF 3002-1999《GB13000.1 字符集汉字笔顺规范》[S] 第 345－423 页

[3] 周有光，21 世纪的华语与华文[J]，《中国语文现代化学会通讯》，2001 年 9 月，第 27 期，第 1-3 页

[4] 俞士汶、苏祺、胡景贺，汉语拼音与汉语信息处理技术[C]，见苏培成主编《信息网络时代的汉语拼音（汉语拼音方案公布 45 周年纪念文集）》，北京：语文出版社，2003 年 10 月，第 7-20 页

[5] 王均，再论汉语拼音方案是最佳方案[J]，《语言文字应用》，2003 年第 2 期，第 1－9 页

[6] 俞士汶， 民族特点的文化要求——汉字汉语民族语言进入信息系统[M]，见罗沛霖主编《信息电子技术知识全书》第 15 章，第 298-311 页，北京：北京理工大学出版社，2006 年 5 月

[7] 中国工程院编，常平主编:《20 世纪我国重大工程技术成就》[M]，暨南大学出版社,2002 年 5 月，第 29－41 页

[8] 冯志伟，中国语料库的历史与现状－语料库研究回顾与问题[C]，见《中文计算国际会议论文集（ICCC）》，2001 年 11 月 27－29 日，新加坡，第 1－15 页

[9] 俞士汶、朱学锋 等，《现代汉语语法信息词典详解（第二版）》[M]，北京：清华大学出版社，2003 年 2 月

[10] 俞士汶，建设综合型语言知识库的理念与成果的价值[J]，《中文信息学报》，2007 年第 6 期第 3-12 页

[11] 董振东，董强，知网，见网页：http://www.keenage.com，1999 年

[12] Dong Zhendong Dong Qiang. HowNet and the Computation of Meaning [M]. Singapore, World Scientific Publishing Company, 2006

[13] 赵军，徐波，孙茂松，靳光瑾，中文语言资源联盟的建设和发展[C]，见《中文信息处理若干重要问题》，北京：科学出版社，2003 年 11 月，第 218-225 页

[14] 宗成庆，高庆狮，中国语言技术进展[J]，《中国计算机学会通讯》，2008 年 8 月，第 4 卷第 8 期，第 39-48 页

[15] Chengqing Zong, Qingshi Gao. Chinese R&D in Natural Language Technology [J]. *IEEE Intelligent Systems*, Vol.23, No. 6, 2008. Pages 42-48

[16] 刘源，谭强，沈旭昆，信息处理用现代汉语分词规范及自动分词方法，北京：清华大学出版社、广西科学技术出版社，1994 年 6 月

[17] 陈肇雄，高庆狮，SC 文法功能体系，《计算机学报》，1992 年第 11 期，第 801 －808 页

[18] 李宇明，关于《中国语言生活绿皮书》[J]，《语言文字应用》，2007年第1期，第12-19 页

[19] Yu Shiwen, Automatic Evaluation of Output Quality for Machine Translation Systems[J], *Machine Translation*, 1993，V8, 117-126, Kluwer Academic publisher, Netherlands

[20] 俞士汶，语言随计算齐飞，《当代语言学》，2009 年第 2 期，第 97-99 页