中 文 信 息 学 报
第4卷 第3期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol.4 No3

中文信息处理在中国的发展

苏 东 庄

（北京信息工程学院）

袁 琦

（机电部计算机与信息发展研究中心）

我国在1958年研制成功第一台大型电子计算机。之后，从中俄机器翻译着手，开始了中文信息处理的研究。到1975年，围绕中文文字排版系统和汉字编码输入开始了中文信息处理的理论研究和实用开发。1978年以来，不论是在理论研究和实用技术、实用系统的研究开发方面都取得了巨大的进展，从各个方面推动了计算机在中国的应用和推广。

一、汉字输入技术

汉字输入方法可分为汉字键盘人工输入、汉字的自动识别和汉语的语音识别。

1. 汉字键盘输入

据统计，到目前为止，国内汉字键盘输入的设计方案已接近六百种，已在机器上实现的方案有五十几种，有一定知名度的约十几种。

依据不同的标准，汉字键盘输入可进行如下分类：

①按文稿类型分类可分为看打、听打和想打。

②按用户类型分类可分为普通用户和专业用户。

③按键盘类型分类可分为大键盘、中键盘和小键盘；也可分为普通键盘和专用键盘。

④按抽取汉字信息特征的不同可分为音码、形码和音形码。

目前，没有看到完全采用文字信息编码的方案，汉字字义的信息系统比音和形要复杂得多。

我国在汉字键盘输入技术方面的发展，可以归纳为如下几个方面：

①从汉字本身形、音、义的特征出发的编码方案研究发展到利用计算机软件支撑的汉字键盘输入技术研究。

②从人工设计编码方案发展到利用计算机辅助设计提高编码方案的质量。

③从汉字编码发展到汉语词语编码，进而朝着自然语言处理的方向发展。

④从单纯注重面向专职操作人员的编码方案设计，转向同时注重面向一般使用人员（非

本文1990年4月收到

* 本文根据1990年中文与东方语言计算机处理国际会议专题报告修改

专职输入人员）的编码方案研究。

国内已开始了几次大规模的汉字键盘输入方法评测，评测一般包括分类评测、综合评测、联合评测和达标评测等。

目前，我国汉字键盘输入技术的研究工作已经向系统化、机助化、智能化和系列化方向发展。

2. 汉字识别

①汉字识别的研究现状

自七十年代末期，汉字识别技术研究在我国开展以来，特别是从1986年后取得很大进展。

联机手写汉字识别已经商品化，总的性能已达到世界先进水平。识别字数为6763～12000。识别率初次使用时为80%左右，经常使用可达到95%以上。识别速度基本上能跟上人书写的速度。

印刷体汉字识别是我国汉字识别研究的主流。近几年，国内先后推出一批适合我国国情的实用系统。有的系统已商品化。以识别字体看，有单字体的、双字体和多字体（宋、仿宋、黑、楷）的。识别的字数一般为1～2级汉字（3755～6763个）。典型系统的识别速度：9～14个汉字／秒（用286微型机），20个汉字／秒（用386微型机），对样张识别率能达到99.9%的高指标，对中等印刷质量以上的印刷品可达到95%～98%。输入设备大多采用普及型图形扫描器（300dpi），能识别的印刷体的字号从3号到5号。有些系统有版面分析、文本识别、及识别结果的后处理、自动纠错、编辑和输出等功能，形成一个完整的识别输入系统。

脱机手写印刷体汉字识别，最近几年也进行了很多研究，并且建立了几个实验性系统。其中一个交互式自学习识别系统，可识别的字数为3755个，其前十位候选正确率为80～95%，使用386微型机识别速度为1字/秒。

②我国汉字识别方法的特色

我国的汉字识别研究工作虽然起步较晚，但发展迅速。近年来的进展使我国的汉字识别技术摆脱了对国外技术的模仿，在对汉字字形结构深入研究的基础上，提出了一些有特色的汉字特征选择和识别方法，如“汉字特征点法”、“脱壳透视分类和稳定框架法”和“可回溯式点跟综包含配选法”等等。它们共同的特点是：认为汉字是具有汉字结构特点的特殊图形，从如何区别几千或几万个汉字，来选择汉字结构中关键的、稳定的、富含信息的特征。在抽取特征的方位上，着重于文字的上、下、左、右四边或文字的四角。识别方法多为统计识别方法与结构识别方法相结合。

③今后的研究课题

重点在提高系统性能，加强实用化研究，改善人机综合环境、版面分析和识别结果的后处理等方面。

今后的汉字识别技术，应进一步利用词的上下文匹配和基本句法语义的上下文匹配来提高识别率，采用现场学习技术也是提高识别率的一个重要方面。

3.语音识别

①语音识别的研究现状

汉语语音识别主要沿着两条途径展开：孤立词的模式匹配识别和有限词汇的连续识别。中国科学院声学所1978年研制成功通用实时语音识别系统RTSRS(001)，1984年清华大学研制出3000个孤立词的语音识别系统。1988年清华大学利用矢量量化和稳式马尔可夫模型研制成功能识别30个城市名称的非特定人语音识别系统。次年又研制出能识别200多个汉语词汇的实时非特定人语音识别系统。中科院声学所具有2000孤立词的实时语音识别系统，在1988年西欧高技术展览会上，获国际大奖。

②今后的研究课题

今后将开展建立语音库和语音特征库的研究，完善适于汉语语音识别的新技术，继续探讨语音特征抽取和距离量度方法并进一步完善话者适应技术。加强非特定人、大词表、连续汉语语音识别系统的研究。另外，需要注意研究人工神经元网络(Artificial neural networks)在语音识别的应用。语音识别的最终目标是语音理解，应该结合汉语的特点，研究词法、句法和语义知识在语音处理系统中的应用。

二、编码字符集

信息编码是信息处理系统的基础，为了做到系统资源的共享，必须使信息编码标准化。到目前为止，我国共拟订了六个汉字编码字符集标准和四个少数民族文字编码字符集标准。

1. 信息交换用汉字编码字符集

我国于1980年颁布了第一个汉字编码国家标准GB2312-80《信息交换用汉字编码字符集—基本集》，奠定了中文信息处理技术的发展基础。

1984年全国计算机与信息处理标准化技术委员会“字符集和编码”分技术委员会经过研究，提出了汉字编码字符集标准的繁体字与简体字对应编码的原则，并做出了制订六个信息交换用汉字编码字符集的计划。这六个集分别命名为基本集、第一、第二、第三、第四和第五辅助集。其中，基本集和第二、四辅助集是简化字集；第一、三、五集辅助集是繁体字集。同时，基本集与辅一集、辅二集与辅三集、辅四集与辅五集中的汉字分别有简、繁体字的一一对应关系，也即第一、三、五辅助集分别是基本集、第二、四辅助集是繁体字射映集，并且简／繁体字在两个字符集中同码（个别简／繁关系为一对多的汉字除外）。

这六个集均采用双七位编码方式，每张代码表分为94个区和94个位，其中前15区作拼音文字及符号区或保留未用，16区～94区为汉字区。除基本集外，第一、二、三、四、五辅助集的标准文本中，还建议了它们在双八位编码环境中的使用方式。

这六个汉字编码字符集中，基本集（GB2312-80）已出版，第二辅助集（GB7589-87）和第四辅助集（GB7590-87）已正式发布，但尚未出版；第一辅助集已于1988年制订完成，但尚未正式发布；第三、五辅助集已基本制订完成，正在审批阶段。

2. 少数民族文字编码字符集

为适应少数民族语言文字处理技术的发展，已陆续制订了一些少数民族文字编码字符集。

GB8045-87《信息处理交换用蒙古文七位和八位编码图形字符集》是我国制订的第一个

少数民族文字编码字符集标准。国标《信息交换用朝鲜文字编码字符集》已制订完成，但还未经国家技术监督局正式批准发布。国标《信息交换用维吾尔文编码图形字符集》已于1988年制订完成，但尚未正式发布。国标《信息交换用彝文编码字符集》已于1989年制订完毕，也尚未正式发布。

三、汉字内码

在以西文8位单字节计算机系统为基础改造成为能够进行多字节汉字处理的中文信息处理系统中，随着国外各种机型的引进，软件工作者根据各自的体验，设计了各自的非标准的汉字内部码，致使国内外派生出十几种计算机汉字内部码制式，在计算机界出现了汉字内码的混乱。

内码的混乱潜伏在系统之间或机种之间的各个界面上。多种汉字内码共存的危害性在早期的单用户环境下，没有暴露出来。但是随着计算机应用，从单机到网络，从单用户到多用户的发展，以及各种信息系统的建立，汉字内码的不统一，给用户带来了资源共享的困难。

鉴于统一码的重要性，在87年10月成立了“中文内部码与数据类型标准化工作组”。这个工作组的内任务是制订汉字内部码规范，并以FORTRAN8×为试点制订中文数据类型规范。

1987年12月底起，该工作组开始对各公司中文内部码的现状进行调查。88年底工作组召开了汉字用量、有关G1集、ISO10646方案的专题研讨会。

国内外各公司、厂家对此非常重视，认识到交换码和内部码的标准化在信息系统发展中具有战略意义，为此中文内部码与数据类型标准化工作组又着手发起筹备组建“通用中文代码国际联合会”（简称ACCC）。于88年11月联合国内外十八家计算机公司在北京召开了通用中文代码国际联合会正式成立会。

在开展中文内部码与数据类型标准化的课题研究方面，国内组织了文字专家、计算机专家、标准化专家的跨部门协作。在深入研究文字、字符、代码、图符的关系的基础上，提出了HCC方案，有效地参与、影响了ISO的活动，深刻反映了中文信息处理的特点和需求，维护了我国的权益。我国的提案，已与港台同行取得共识，并得到ISO各成员和包括IBM公司在内的跨国公司的重视。

四、中文信息处理应用系统

1. 中文计算机情报检索

作为中文信息处理的应用领域之一，中文计算机情报检索在八十年代取得了长足进步，十年来，经历了由探索创建、基础建设到开始走向实用化的过程。

（1）中文数据库建设

数据库建设是开展计算机情报检索的基础和关键因素之一，据不完全统计，我国已建立各种数据库300个以上，并约有220个已在不同层次和范围使用。

（2）中文情报检索系统的研制

由于中文信息处理的特点和应用条件的不同，国外的情报检索软件无法直接应用于中文

情报资料的计算机检索，已经报现的上百个情报检索软件都是自行研制或在国外检索软件的基础上汉化和二次开发的。在汉化方面较著名的有MINISIS、TRIP、CDS/ISIS等，其中CDS/ISIS的微机版是目前国内使用最广泛的情报检索软件；属于二次开发的有BDSIRS系统。近年来，针对新闻资料检索系统开发了较大型的新闻资料检索系统。在研制和开发中文情报检索系统中，一些主要课题都取得了不同程度的进展：

①汉语主题表的编制　我国从七十年代后期开始编制“汉语主题词表”，目前已出现大中小型主题词表60多部，其中半数以上实现了词表的计算机管理和维护。

②汉语自动标引　八十年代以来，在汉语自动标引方面，已经提出的方法有陈培久的词典切分组配法、王永成的部件词典法、吴蔚天的后缀表法、北京大学的主题词典法、赵宗仁的组配抽词法、Andrew Choi的词串频度分析法和北京信息工程学院的新闻资料全文自动标引方法等，目前我国自动标引的研究正从以词频分析为主的单词标引向采用较复杂语言分析技术的短语标引过渡，人们试图建立自动标引的知识库，以实现智能标引。

③汉语全文检索　汉语全文检索系统的研制近年来受到重视，已开发出现试验系统。由于汉语自动分词的困难，有人提出并实现了以字为基本检索单位的汉语按字全文检索系统，包括汉化的TRIP系统。

④汉语智能接口　研究表明，用户要找到相关的资料，检索者需要大量的知识，它包括：不同检索系统的查询语言；特定数据库文档的主题覆盖面；如何构造正确的和有效的布尔检索式；正确地表达检索需求。汉语智能接口的主要研究课题包括：①汉语自然语言查询接口；②用户模型专家系统；③媒介专家系统；④浏览专家系统等。目前在汉语自然语言接口方面已取得一定进展。汉语的复杂性和检索者本身的特点决定了汉语智能检索接口的研制比西文更为重要，然而也更为困难。

在中文情报检索的其它方面（如联机检索服务、中文情报检索理论）也都有不同程度的进展。

由于对汉语情报资料的需求不断增长和计算机情报检索技术的进步，90年代将是中文情报检索事业大发展的时期，应重点研究的与中文信息处理相关的课题有：

①加速中文数据库的建设步伐

②深入研究汉语自动分类、自动分词和自动标引，实现情报处理的自动化

③综合性、多介质情报检索系统（全文文本、数据、事实、图象、声音等）的设计和研制

④中文智能接口的研究与实现

2.电子印刷排版系统

在中文信息处理应用领域里，中文印刷排版系统的研究开发卓有成效。到目前为止，成为商品化的系统不下十余种，象华光、科印这样的排版系统其用户已达数百乃至上千。

国内从事排版软件的研究开发可追溯到七十年代初第二代光机式照排机研制时期，这是最早的尝试和探索。1975年北京大学开始承担华光系统的研制任务。1981年推出的华光Ⅰ型系统上有一个功能较强，能自动成页，自动形成页码、书眉，能排有斜线表格的排版软件。1985年初，华光Ⅱ型系统投入生产性使用。

1985年夏天，中国印刷科学技术研究所研制成微机上的排版软件，适用于文科书籍的

排版。这是我国首次使用微型机完成排版出清样的工作。

1986年，福州大学等单位试制成我国第一套台式出版系统。该系统是国内第一套四种字体、多种字号，能在300DPI激光印字机上输出文科版面的台式出版系统。同一年，北京大学完成了国内第一个实用的高性能科技排版软件。

1987年夏天，中国印刷科学技术研究所又推出了称之为“科印”的微机科技排版软件。它是国内第一个微型机上的实用批处理科技排版软件。同年秋天，四通公司4S科技排版软件问世。它是国内第一个交互式书刊排版软件。

1988年春专用芯片支持的华光Ⅳ型系统投放市场。

1989年12月，我国对电子出版系统进行了全国性的评测，认为北大华光、4S、潍坊华光、前景系统、星汉系统这5个系统代表了汉字电子出版系统的最高水平。

在上述中文排版系统发展过程中，特别值得一提的是华光排版系统。该系统被评为1985年中国十大科技成就之一，荣获1985年中国发明协会发明奖，1986年日内瓦国际发明展览金牌奖，并多次到日本、新加坡、西德等国参加国际展览。

中文电子印刷排版系统在国内有巨大的潜在市场。报社排版计算机化发展极为迅速，书刊出版和办公室轻印刷采用计算机排版的趋势也将继续上升。市场的迫切需要和产品的竞争促使我国在这一领域的技术水平提高很快。

3. 机器翻译系统

(1) 现状

在我国中文信息处理应用研究方面，机器翻译是早在1956年就列入了我国科研工作的发展规划。在三十多年的发展历程中，走过了一个曲折的道路，直到七十年代末期，开始重建研究队伍，进入了我国机器翻译的再发展阶段。

目前，国内有三十多个单位从事机器翻译的基础理论或实验系统的研究，已建立的英汉，日汉和汉英等各种类型的实验系统有十几个，其中有些系统已商品化。国内已有四个系统通过了部级鉴定。这四个系统是：“英汉科技文献题录机器翻译系统”；“科译1号英汉机译系统”；“ISTIC—1型英汉冶金题录机译系统”；“人名翻译和题录翻译系统”。国内已推出的商品化系统有三个：“译星英汉机译系统”；“IECM英汉机译系统”和“英汉机器翻译系统Marcopolo”。

此外，中国社会科学院语言所研制的JFY—Ⅳ型全文机译系统在1987年完成定型设计，以其独特的设计思想区别于国内外其它系统。

近年来中国MT研究虽有很大发展，但与国外相比仍有不少差距。

(2) 难点和关键

①MT研究中首当其冲的是语言学问题。

②必须有良好的软件支撑环境。

③语言外的知识及知识表达问题是MT研究的关键。

(2) 今后的研究课题

①语言（尤其是汉语）结构的研究

②语义学的研究

③语言分析算法和分析器的研究

④机器翻译支援系统用人机界面技术的研究
⑤中性电子词典的开发
⑥机器翻译质量的评估技术
⑦机器翻译的利用及其受限语言的制定

4. 汉语理解和人机接口

我国自然语言理解方面的研究工作起步较晚，进入八十年代以后，随着自然语言理解成为人工智能研究中十分活跃的课题和新一代计算机的主要技术，自然语言理解在我国也开始得到重视。八十年代中，在国家支持下，将“自然语言理解和人机接口”列入重点研究课题。在新一代智能计算机的研制规划中，也把自然语言理解列为主要课题。

十余年来，我国在汉语理解方面已取得一些研究成果。例如借鉴国外自然语言理解的理论和模型，结合汉语的某些特点，提出了一些初步的汉语理解模型，并建立了一些汉语理解实验系统和汉语接口实验系统。近年来，国内也开始重视汉语理解的实用问题，自然语言理解已向广度和深度两个方向发展。

汉语理解最终要落实到篇章一级的理解上来。这不仅是因为句子的歧义（如省略和指代等现象）往往不能在其自身范围内消除，而是必须把它放到更大的语境即篇章中去才能理解。篇章生成和单句生成的关系也是如此。国内是三、四年前开始从事汉语篇章理解和生成的研究工作的。

我国自然语言理解的研究和国外有较大的差距，表现在：研究力量薄弱，低水平重复多；缺少深度，没有出现软件商品；未形成汉语理解的成熟模型，更未形成自己的流派。

5. 我国计算语言学的研究现状与面临的课题

(1) 我国计算语言学的研究现状

我国计算语言学的早期研究工作，可追溯到七十年代中期的汉字词频统计研究。1981年中国中文信息研究会成立以后，相继组建了基础理论专业委员会和自然语言处理专业委员会。1986年通过国家级鉴定的现代汉语词频统计工程，可以说是我国在计算语言学研究方面取得的一项重要成果。进入八十年代以后，随着中文信息处理从字处理逐步向词处理和句处理发展，计算机科学和语言研究愈加互相渗透，国内一批学者相继提出了汉语的多标记多叉树形图分析法、汉语格语法、汉语属性制约法，以及制定信息处理用规则汉语等等。为了进一步促进我国计算语言学的基础研究和学术交流工作，1986年6月中国中文信息学会吸收国内的计算机、人工智能、语言学和心理学专家组建了计算语言学专业委员会。

近年来通过应用开发和国际合作也推动了计算语言学的研究工作。我国参与了“日本及其邻国的机器翻译系统研究和开发项目”的实施计划。我们的研究重点是中文句子的分析和生成，以及中文基本词典的开发这三个与计算语言学关系极为密切的基础课题。这会促进国内中文信息处理的基础研究工作。

(2) 今后的研究课题

为了推动中文信息处理应用技术的发展，必须加强我国在汉语计算语言学方面的基础研究工作。它应包括语料库系统、电子词典系统和汉语语法库系统三个主课题。

语料库系统将为电子词典的研究提供词条，并且从中提取和验证词条的属性。同时，也

为语法库提供规则依据和测试、统计素材。电子词典系统是一切语言信息处理和某一特定应用系统的中性词典。语法库系统不仅应该描述汉语语法体系中句型、短语的类型，同时也应回答句型、短语构成的约束条件。

为了支撑上述三个主课题的研究，首先要在词语一级和句子一级上加强基础性研究。在词语一级上，应在汉语词频统计、汉语分词规范和建立通用词库的基础上，尽快制定信息处理用汉语词性划分及其二级分类规范。通过研究词的语义因子和词义分类体系，建立汉语词的语义解释。在句子一级上，应加强汉语的词组、句型规范和汉语句型统计研究，并且加强汉语语法体系和适用于汉语的句法分析算法的研究。与此同时，还应加强与句法分析密切相关的语义组合关系（即句子中概念之间的关系和属性）的研究。

以上我们从汉字输入技术、编码字符集、汉字内码和中文信息处理应用系统几个方面阐述了中文信息处理在我国的发展，其它方面如中文操作系统、中文高级语言等也有不同程度的进展。

中华文化有着数千年的悠久历史，全世界有五分之一以上的人使用中文，中文信息处理应该有也必定会有更为显著的发展。

致谢

本文引用和参考了中国中文信息学会各专业委员会的专家们的论文和资料，在此深表谢意。