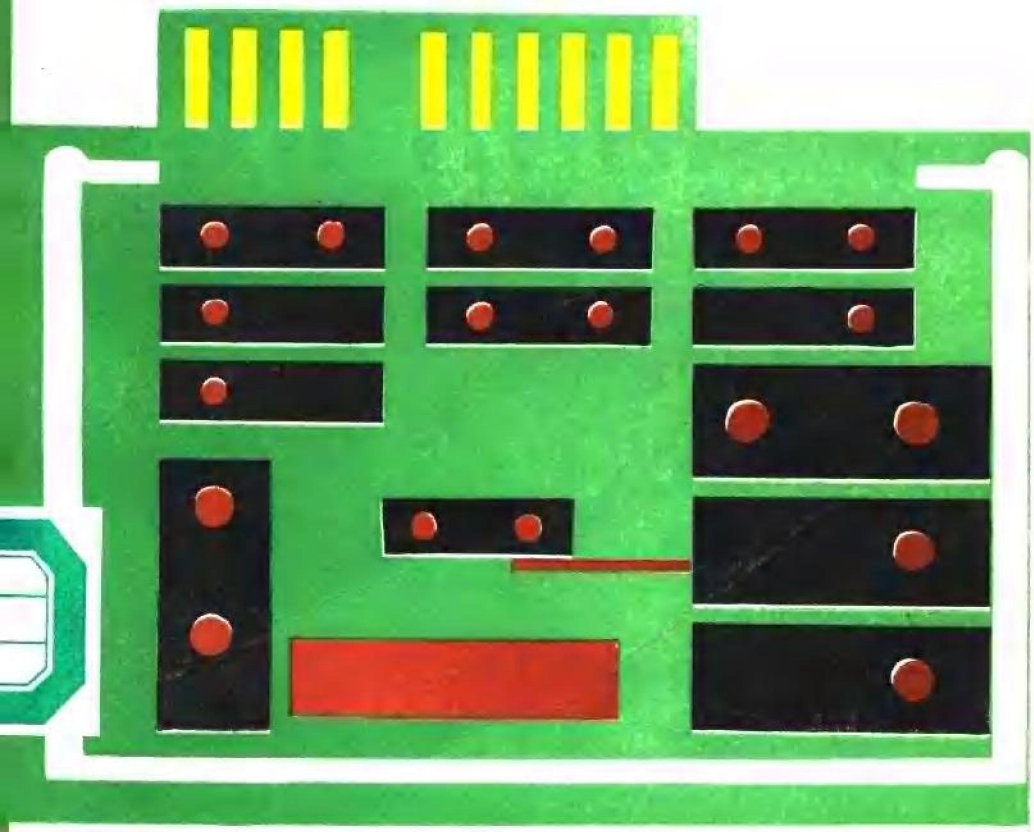
汉语信息处理研究

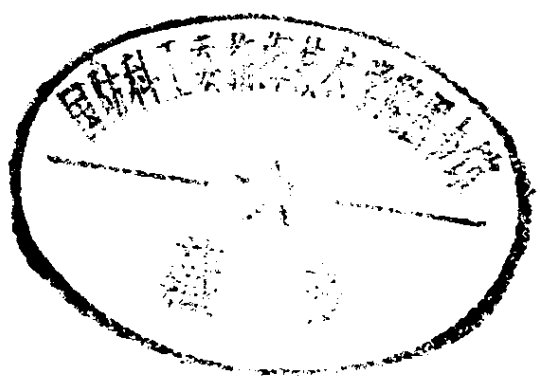
张 普 著



北京语言学院出版社

汉语信息处理研究

张普 著



北京语言学院出版社

（京）新登字157号

汉语信息处理研究

张　普　著

*

北京语言学院出版社 出版发行
（北京海淀区学院路15号　邮政编码100083）
新华书店北京发行所经销
北京门头沟胶印厂印刷

850×1168　1/32　9.25印张　223千字
1992年8月第1版　1992年10月第1次印刷
印数 1—2000
ISBN7—5619—0211—5/H·155　定价：4.20元

前　言

本书收入作者1986年以后撰写的有关汉语信息处理研究方面的论文19篇，附录一篇。全书分为总论、汉字键盘输入方法及其评测研究、汉语信息处理词汇（国家标准）研究、计算机与古籍整理研究、计算机与对外汉语教学研究、计算机与语义和语用研究及其他六个部分。作者1986年以前撰写的有关汉语信息处理研究方面的论文已收入《语言自动处理》一书，由武汉大学出版社1988年出版。如读者阅读本书（特别是第一部分）时感到有必要，可以参阅《语言自动处理》一书的有关内容。

本书收入的论文绝大多数是作者在承担国家七五重点科技攻关项目（普及型汉字输入方法研究、计算机辅助对外汉语教学系统及其所用之语言文字规范、汉语信息处理词汇标准研究）和国家八五重点科技攻关项目（信息处理用现代汉语语义分析词典）所撰写的有关论文，这些论文均已在国际、国内学术会议或相应刊物发表过，收入文集时未进行修订。许多论文在撰写时曾与共同承担国家任务的攻关组内的同事以及国内同行讨论过，有的论文正式发表前还经有关领域的专家传阅过，作者愿借论文集发表之机，向所有给予过帮助的专家和朋友再一次表示感谢，同时也期待着读者的批评。

汉语信息处理是关系我们国家和民族子孙万代的百年大计，是现代化的重要标志，是一代人的追求，是几代人的事业。汉语信息处理的重要意义和深远影响已越来越突出也越来越被

人们认识。作为一种事业，汉语信息处理还处于起步阶段，它的研究领域还要不断开拓，研究成果还要不断推广，研究队伍还要不断壮大，如果由于本书的出版，使一些年轻同志热爱并投身于汉语信息处理事业，将是作者最大的收获。

作者的论文受功力、精力所限，一定有许多不妥之处，但承安子介先生和北京语言学院出版社的大力支持，得以结集出版，谨向他们表示最诚挚的谢意。

张　普

1992. 4. 30

总论

中文信息处理研究与发展前瞻*

——中国语言研究面临的挑战与机遇

中文信息处理属于语言信息处理范畴。语言信息处理是指“用计算机对自然语言的音、形、义等信息进行处理，即对字、词、句、篇章的输入、输出、识别、分析、理解、生成等的操作与加工。”** 本文论述的中文信息处理的研究与发展主要是指汉语信息处理，所谈的语言方面的问题也以汉语为特定对象。

---

* 本文1989年12月19日发表于《计算机信息报》，该报现名为《中国计算机报》。

** 引自《汉语信息处理词汇·01部分：基本术语》4·1·1·1条“语言信息处理”。

一、汉语信息处理的研究内容

由于语言信息处理的手段是利用计算机，而处理的对象是自然语言，因此，当前其研究内容就包括计算机软硬件的开发和探讨自然语言的受限规则两大范畴。就汉语而言，是开发计算机的汉语言支撑能力（Chinese language support）* 和研究受限汉语（restricted Chinese language）的规则。

由于计算机的汉语言支撑能力的开发与受限汉语规则的研究都不是一件容易的事，不能一蹴而就，并且两者的研究又互相促进也互相制约，因此，汉语信息处理的研究与发展可能呈现四个阶段：字处理阶段、词处理阶段、句处理阶段、篇章处理阶段。这四个阶段主要是就汉语信息处理的发展水平而言的，而不是单纯时间上的划分。虽然各个阶段研究内容的侧重点不同，但在技术和成果方面，一些阶段之间是互相渗透的。比如：词处理阶段为句处理阶段打下了基础，句处理阶段还将发展词处理阶段的成果。总的看，汉语信息处理的四个发展阶段中，词处理是基础，句处理是中心。

很长一个时期以来，在我国学术界没有提“汉语信息处理”，甚至一开始也不叫“中文信息处理”，而是叫“汉字信息处理”(Chinese character information processing)。这是因为汉语信息处理的开端是由“字处理阶段”入手的，这个阶段是汉语的处理所特有的。西文无所谓“字”，二十六个拉丁字母（或其他字母）及必要的符号解决之后，就直接进入词处理阶段，同时，由于有较丰富的形态标志，词处理向句处理的过渡也有良好的

---

* 《汉语信息处理词汇·01部分：基本术语》4·1·1·7条收录“民族语言支撑能力”National Language Supprot 缩写为：NLS，定义为“使计算机具备能够处理民族语言的开发能力。例：中文化，汉字化。”

基础。书面汉语不实行分词连写，可以称得上形态的标志也很少，词处理的难度比西文大得多。但汉字是表意文字，含有意义信息，词由字组成，词与字之间有千丝万缕的联系，一个字就是一个音节，因而首先解决汉字的受限并开发计算机的汉字支撑能力相对是容易的，也是必需的。由“汉字编码”的研究开始，限制汉字数（6763）、制定汉字交换码、汉字点阵、汉字内码、推出汉字库、汉（字）卡等等构成了整个“字处理阶段”的研究。

就总体水平而言，我国的汉语信息处理水平正处于从字处理向词处理过渡的阶段。单纯解决字处理已经不能满足汉语信息处理的发展需要，从“六五”到“七五”期间，围绕词处理问题作了许多准备工作和实际的探索。例如：词频统计、分词规范、通用词表、自动分词以及词汇编码、词库设计等等。目前正向设计真正含有句法、语义、语用信息的词典推进（而不是单纯的词表），没有这样的机器词典或词库，“词处理”的水平就只能停留在初级阶段，不可能有彻底的解决。语言的三要素是语音、词汇和语法，文字只是记录语言的符号。说到底，汉语信息处理的解决必须向词处理和句处理的阶段发展。我们说当前正在解决词处理阶段的问题，并不是不能够或没有人探索句处理乃至篇章处理的问题，在自然语言理解、机器翻译、辅助教学甚至语音识别与合成、汉字自动识别、智能化输入系统、自动分词专家系统等课题中都已不同程度地涉及了句法、句型、语义、语用等方面的问题。这些探索都是有益的、有成效的。但是，词处理这个基础是不可超越的，在词处理没有较好地解决前，上述的超前探索或者是有限的，或者是探索者自己已经解决了有限的词处理手段。

二、中国语言研究面临的挑战

由于计算机的出现和信息处理事业的迅速发展，语言研究突然面临着一种挑战，这种挑战是全面的、深刻的、划时代的。它推动了语言研究的发展，给语言研究带来了一片生机。语言学界经历了短暂的迷茫、观望、沉思之后，已经开始自觉地适应和接受这种挑战。认识到它并不是宣告传统语言研究的“死亡”，而是为语言研究注入新的活力，语言研究必将也正在发生着一些变化，这些变化表现为：

·研究的目的：

从单纯面向人的研究到面向人与机两个方面。不仅要解决人与人的交际问题，还要解决机器理解、人机对话。

·研究的对象：

从单纯以语言为研究对象到以语言、人脑、电脑及其之间的关系为主要对象，并涉及到电子、生理、物理、数理、心理、社会、哲学等许多方面。由于信息处理具有国际性，以多种语言为研究对象的比较研究越来越突出。

·研究的意义：

语言是信息的载体，是符号系统，语言研究的水平已直接关系到语言信息处理的水平。而语言信息处理的水平和处理量已是当今衡量一个国家现代化水平的重要标志。

·研究的内容：

我们已经论述了汉语信息处理是从汉字编码入手，要经历字处理、词处理、句处理、篇章处理等发展阶段。因此，面向信息处理的语言研究已涉及到文字、语音、词汇、句法、语义、语用等各个方面，并要从新的角度（面向机器）来审视这些方面。

·研究的方法：

由于电脑不似人脑，人脑具有母语的背景知识，而电脑的起点是“零”。为取得教给电脑的语言知识，除传统的语言研究方法外，必须采用一些新的研究方法，例如：从举例法转到统计法或穷尽枚举法；从个别语言现象的深入研究到总体的系统工程的角度的研究；从单纯研究语言现象到涉及到多种学科交叉的综合研究。

·研究的手段：

从传统的手工研究、卡片方式到使用计算机辅助研究；从单纯利用计算机搜集资料到辅助标引、分析、统计、分类等等。面向计算机的语言研究必须由计算机来辅助，将欲取之，必先予之。

·研究的结果：

学术上的问题可以百花齐放、百家争鸣，可以存疑，可以有歧义，也可以简单地自圆其说，自成体系。工程化的语言研究结果更注重解决实际问题，计量化、精确化、模式化、标准化。

·研究的队伍：

当前语言研究队伍由于客观发展的需要，有扩大化的趋势、综合化的趋势、年轻化的趋势。特别要注意的是一批自然科学工作者已经加入到语言研究的队伍，他们从寻求、等待语言研究成果到自己动手为语言信息处理的目的研究语言。

结束语

由于信息处理的发展，世界语言学界面临着挑战。由于汉语汉字的一些不同于西文的性质，总的说，中国语言研究面临的挑战更加严峻。而开发计算机的汉语言支撑能力取决于面向计算机的汉语研究水平，从这个意义上说，这种挑战不单纯是

面向中国语言学界的，它从一开始就面向了整个中文信息处理学界。

有专家估计，就中文信息处理的研究水平而言，中国已居世界领先地位，并且领先台湾省两年时间。我们必须不断引进消化吸收，不断开发汉语言支撑能力，不断前瞻，才能保持研究方面的优势，才能真正在我国的现代化建设事业中普及计算机的应用。

目　　录

计算机与古籍整理研究

计算机与对外汉语教学

计算机与语义、语用研究

其他

附录：

一、汉字键盘输入方法及其评测研究

论汉字键盘输入技术*
——历史·现状·展望

1. 术语的界定

在展开本文论述之前，我们对下述术语进行界定，以利对本文的理解。

---

* 本文1992年4月发表于台北海峡两岸中文键盘输入研讨会，收入会议论文集。《计算机世界》（中美合刊）1992年5月27日刊出。
本文与华绍和、陈一凡、曹右琦合作，张普执笔。

1．1 汉字编码* **Chinese character coding**

按照一定的规则，对指定的汉字集中的元素编制相应的代码。

1．2 汉字编码方案* **Chinese character coding scheme**

汉字集元素映射到其他字符集元素的一组完整规则。

1．3 汉字编码（键盘）输入方法* **Chinese character coding (keyboard) input method**

运用某种编码方案、键盘设备及计算机资源，由操作者向计算机输入汉字的方法。

1．4 汉字编码（键盘）输入系统 Chinese character coding (keyboard) input system

由汉字编码（键盘）输入软件及其相关的造字软件、查字软件、词库生成软件、知识库管理软件、多文种兼容输入软件、非语言符号输入软件、繁简体汉字转换和软件等组成的配套的汉字输入软件。

1．5 汉字编码（键盘）输入技术 Chinese character coding (keyboard) input technology

指汉字编码（键盘）输入系统的研制涉及到的一切有关的技术。如：汉语语料库建造技术；字（词语）频度统计技术；语音频度分析统计技术；笔画、部件、结构分析统计技术；点阵（矢量）汉字库建造技术；字（词语）属性库建造技术；汉语知识库的建造技术；编码技术；计算机辅助编码技术；字（词语）信息压缩技术；屏幕引导技术；盘外造字技术；多文种兼容技术；非语言符号处理技术；繁简体汉字转换和兼容技术；字体、字号变换技术等。

---

* 凡加有*号的术语，均转引自《汉语信息处理词汇·01部分：基本术语》(GB12200．1－90)，下同。

1.6 汉字编码(键盘)输入方法评估 * **evaluation of Chinese character coding (keyboard) input method**

按照约定的或法定的规则和步骤，对汉字编码（键盘）输入方法的素质和特性进行定量的测试和定性的评价等。

1.7 汉卡 * **Chinese character card**

将汉字编码输入方法的码表和有关程序及汉字的字模数据固化在 ROM 器件中的一种逻辑电路插件。

1.8 语言信息处理 * **language information processing**

用计算机对自然语言的音、形、义等信息进行处理。即对字、词、句、篇章的输入、输出、识别、分析、理解，生成等的操作与加工。

1.9 汉语信息处理 * **Chinese information processing**

用计算机对汉语的音、形、义等信息进行处理，有时又称中文信息处理。

1.10 汉字信息处理 * **Chinese character information processing**

用计算机对汉字所表示的信息进行的操作和加工，如汉字的输入、输出、识别等。

1.11 汉字输入 * **Chinese character input**

利用汉字的形、音或相关信息通过各种方式把汉字输入到计算机中去的过程。

1.12 多文种信息处理 * **multilingual information processing**

在两种或两种以上语言文字字符集编码体系基础上，实现对多文种信息的兼容处理。

1.13 民族语言支撑能力 * **National language support NLS（缩写）**

使计算机具备能够处理民族语言的开发能力。

2. 汉字编码（键盘）输入技术

二十世纪中叶，一个巨大的难题历史地落在了当代中国人的肩上，同时引起了全世界高科技领域特别是信息处理界的专家们的极大兴趣和强烈关注。这就是世界上唯一仍在通行并且使用人口最多的最古老的文字——汉字和当今最现代的高科技的精灵——电脑所发生的划时代的激烈碰撞，这种强烈的时间反差已经决定了这一碰撞必然要引起的轰动效应。轰动效应首先瞄准了第一次碰撞的能量释放中心——汉字进入计算机。多少中国人废寝忘食、呕心沥血献身于“汉字编码事业”，一时间各种汉字编码方案此呼彼应，发明迭起，百花齐放，色彩纷呈。这种情况被比喻为历史上的“编码现象”，是“造福于子孙万代”，成功的编码方案被称为“重大突破”、“中国的第五大发明”，方案的作者被誉为攻克了“汉字的哥德巴赫猜想”，是“当今的苍颉”。这些都足以说明这场轰动是怎样地震惊新闻界、震惊国人，震惊世界了。

2.1 字处理阶段

2.1.1 早期的汉字编码研究

已知的最早的汉字编码始于本世纪五十年代的俄汉机器翻译，当时只能用电报码或四角号码充做汉字编码。但如何使数量庞大、笔画、结构复杂的汉字进入计算机仍时时困绕着少数先知先觉的科学家的头脑。

五十年代，钱文浩先生开始从信息论的角度研究汉字，并计算汉字熵值。

六十年代，戴着“反动学术权威”的帽子的支秉彝先生在“牛棚”的茶杯盖上完成了“见字识码”方案的设计和码本。

1978 年 7 月 19 日上海《文汇报》以几乎整版的篇幅在第一