华中语学论库 第四辑

◎邢福义 主编

面向中文信息处理的汉语语法研究

姚双云 著

華中師範大學出版社

华中语学论库(第四辑)

邢福义　主编

本书为教育部人文社会科学重点研究基地
华中师范大学语言与语言教育研究中心成果

miàn xiàng zhōng wén xìn xī chù lǐ de

面向中文信息处理的

hàn yǔ yǔ fǎ yán jiū

汉语语法研究

姚双云　著

华中师范大学出版社
2012 年 · 武汉

新出图证(鄂)字 10 号

图书在版编目(CIP)数据

面向中文信息处理的汉语语法研究/姚双云 著. —武汉:华中师范大学出版社,2012.5

ISBN 978-7-5622-5478-2

Ⅰ.①面… Ⅱ.①姚… Ⅲ.①汉语—语法—汉字信息处理—研究

Ⅳ.①H14②H127

中国版本图书馆 CIP 数据核字(2012)第 083339 号

面向中文信息处理的汉语语法研究

作者:姚双云

责任编辑:肖和容　　责任校对:王　胜　　封面设计:罗明波

编辑室:高校教材编辑室　　电话:027—67863220

出版发行:华中师范大学出版社

社址:湖北省武汉市洪山区珞喻路 152 号

电话:027—67863426/3280(发行部)　　027—67861321(邮购)

传真:027—67863291

网址:http://www.ccnupress.com　　电子信箱:hscbs@public.wh.hb.cn

印刷:武汉理工大印刷厂　　督印:章光琼

字数:245 千字

开本:787mm×960mm　1/16　　印张:15.75

版次:2012 年 5 月第 1 版　　印次:2012 年 5 月第 1 次印刷

印数:1—1000 册　　定价:32.00 元

欢迎上网查询、购书

序

邢福义

随着历史的发展，社会的进步，科技的发达，语言学在整个世界范围内越来越展示出强大的活力和能量。中国语言学是世界语言学的重要组成部分。为了对中国语言学事业有所推动，我们组织撰写“华中语学论库”。作为专用名称，这里的“语学”主要指汉语语言学，近期的15年时间里以现代汉语语法专题研究为重点。“语学论库”，这是汉语语言学研究的一个系统工程，如果将来主客观条件具备，在研究范围上可以不断扩大，在研究时间上可以无限延展，在研究队伍上可以辈辈交接，代代传承。“华中”一词，既跟研究队伍的华中群体相关，又跟华中师范大学出版社的名称相关。

汉语语言学源远流长。千百年来，特别是《马氏文通》出版以来，尤其是20世纪70年代之后，由于一代代学者的不懈努力，汉语语言学沿着“创业——拓新——发展”的轨道不断推进。目前，汉语语言学所统括的汉语语法学、汉语语音学、汉语方言学、汉语词汇学、汉语语用学等学科，都已出现了初步繁荣的喜人局面。

但是，初步繁荣并不意味着已经成熟。对于语言学这样一门社会科学来说，成熟与不成熟的突出标志，应该是学派或流派是否已经形成。在这一点上，科学跟艺术情况相同。比方说，我国的京剧表演艺术已经达到了成熟的高峰，最基本的表现就是形成的这“派”那“派”，只要一提到“梅派”和“程派”，稍有京剧表演艺术知识的人就会知道这是两个具有各自特点的著名流派。又比方说，我国的书法艺术早已达到了成熟的高峰，最基本的表现就是形成了这“体”那“体”，只要一提到“颜体”，稍有书法艺术知识的人就会知道它是不同于“柳体”和“欧体”等的有独特风格的书写体，甚至还会知道颜真卿打破了“书贵瘦硬”的传统书风，开创了二王体系之外的新体。然而，汉语语言学的各门学科，即使是其中发展

速度最快的现代汉语语法学，仍然缺乏显示成熟的任何标志，距离真正成熟实际上还十分遥远。

当今的汉语语言学，面临的主要问题是“二求”：一求创建理论和方法，二求把事实弄清楚。这是互补互促而又互成因果的两个问题。没有理论和方法的成熟，一门学科不可能是成熟的。而理论和方法的创建，是学者们长期深入研究的成果，是有效地进行群体性思考、独立性思考和开拓性思考的结晶。因此，必然带有鲜明的个性，带有学派的印记，反映一派学者的思想体系、研究特点和总体成就。另一方面，没有对事实的清楚了解，理论和方法的创建便成为空中楼阁。从现代汉语语法研究来说，之所以至今尚未成熟，自成体系的理论和方法之所以尚未创立起来，最根本的原因还是对事实的了解基本上仍然处于朦胧的状态。真正适合于我国语言文字的理论和方法，最终只能产生在我国语言文字的沃土之上。因此，应该强调“研究植根于汉语泥土，理论生发于汉语事实”。不然，我国的汉语语言学在世界语言学中就可能永远处于附庸的地位，就永远不会有跟国外理论对等交流的时候。

学术派别的产生，起码应该具备三个条件：第一，有特定的学术领地，提示标志性的理论和主张；第二，有鲜明的治学特点，形成一套自己的研究方法；第三，有良好的学风，形成一支富有活力的队伍。近年来的研究状况表明，我国的学者们已经或多或少地显示了各自的风格特点，但是，顶多只能说其中孕育着某些派别意识，或者顶多只能说预示了某种派别意识的萌芽。汉语语言学的真正成熟，需要经历很长很长的历史阶段，有赖于众多的学者群策群力，更有赖于一辈一辈的学者发扬愚公移山的接力精神。我们华中研究群体人数很少，力量单薄，起点不高，功力不足，对于汉语语言学的发展起不了多大的作用，但是，我们愿意跟在前辈学者的后头，跟在全国各地学者的后头，尽心竭力地做点力所能及的工作。如果把建设富于特色的汉语语言学比作建筑一座大厦，那么，我们组织撰写“华中语学论库”，便是想为这座大厦的建筑献上几根钢筋几块石头。通过参加大厦的建筑，使我们这支小小的队伍受到训练，这是我们的最大愿望。各部著作在内容上具有独立性，但我们希望，在出版了以上二十部之后可以看到研究风格上的某些特色和理论方法上的某种网络。

“华中语学论库”的撰写和出版，得到华中师范大学出版社的大力支

持。年初，出版社社长朱峰先生和中文编辑室主任陈昌恒先生到我家，鼓励我牵头编写一套关于汉语语言学的丛书，要我拟订一个初步的计划。不久之后，新上任的总编辑王先霈先生了解了有关情况，立即审定计划，并且从内容到选题都提出了好些中肯的意见。他们为发展学术事业所作的决策，他们在出版事业上的决心、魄力和历史责任感，不管是对我个人还是对华中语言研究群体的所有成员，都是极为有力的鞭策。

千里之行，始于足下！

贵在努力，贵在坚持！

1996 年 5 月 4 日

目　录

第一章　中文信息处理的回顾、展望与理论探讨

第一节　中文信息处理30年的回顾与展望[①]

中文信息处理指的是利用计算机对汉语(包括书面语和口语)的音、形、义等信息进行转换、传输、存贮、分析等加工的科学。它是自然语言信息处理的一个分支,是一门与计算机科学、语言学、心理学、数学、信息学、自动化技术、控制论、声学等多种学科相关联的综合性交叉学科。其中,"中文"是指中国通用的所有语种,包括汉语与少数民族语言,但一般指的是汉语[②]。我们这里讨论的也是以汉语的信息处理为主,顺带提及了少数民族语言的信息处理问题。"信息"是指能通过视觉、听觉、嗅觉、味觉、触觉等器官或仪器获取具有交际功能的实体,主要指文字的形体与语音。

以1978年为基点,我国的中文信息处理走过了30多年的历程。30多年里,在计算机研究者和语言学研究者的艰辛合作与共同努力下,中文信息处理学科稳步发展,各个领域均取得了很大的成就,对我国的国民经济、社会发展以及语言生活等重要的领域均产生了深远的影响。下面我们通过九个方面,简明扼要地回顾中文信息处理在过去30年的发展概况,并对中文信息处理未来的发展趋势作简单的分析和展望。本节的介绍偏重于基础性研究成果,应用性研究成果只顺带涉及。

---

① 本节的主体内容发表在邢福义、汪国胜主编的《中国高校哲学社会科学发展报告(1978—2008)·语言学》上,因此,这里的30年指的是1978—2008年。

② 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。

1. 研究概览

1.1 概论性研究

从世界范围来看，我国的自然语言处理研究要晚于西方发达国家。这其中有政治、经济、科技、教育等多种原因。因此，在自然语言处理方面，我国要落后于发达国家。所以，我国中文信息处理的主流是学习和借鉴国外的理论方法和技术手段。学者们在引进和学习国外理论的同时，致力于结合汉语特点来研究汉语的计算机处理问题，取得了可喜的成就。

著作方面，20 世纪 90 年代初，国内有三本引论性质的计算语言学专著问世，分别是《计算语言学引论》(钱锋，1990)、《计算语言学导论》(陆致极，1990)、《自然语言处理》(刘开瑛、郭炳炎，1991)。三部著作全面介绍和总结了自然语言处理研究的理论和方法，基本上反映了国内外一个时期内研究的概貌。《中文信息处理技术基础》(王永成等，1991)、《汉语信息处理研究》(张普，1992)等也是这一时期的重要著作；90 年代中后期出版了《汉语计算语言学》(吴蔚天、罗建林，1994)、《自然语言理解》(姚天顺、朱靖波，1995)、《语言学知识的计算机辅助发现》(白硕，1995)、《自然语言的计算机处理》(冯志伟，1996)、《中国计算语言学》(姚亚平，1997)、《计算语言学导论》(翁富良、王野翊，1998)、《语言的认识研究和计算分析》(袁毓林，1998)、《中文信息处理》(傅永和，1999)以及《计算语言学与汉语自动分析》(侯敏，1999)等重要著作。这些著作及时吸收了国外研究的最新动态，对国内学者所做的系统研究和理论探索也多有涉及，是了解本学科理论方法的重要资料；进入 21 世纪，出版的重要著作有：《汉语语法的意合网络》(鲁川，2000)、《汉语自动分析——Visual C++实现》(陈小荷，2000)、《自然语言理解》(姚天顺、朱靖波，2002)、《计算语言学概论》(俞士汶，2003)、《计算机自然语言处理》(王晓龙、关毅，2005)、《自然语言处理》(江铭虎，2006)、《中文文本信息处理的原理与应用》(苗夺谦、卫志华，2007)等。

30 年来，出版了一些译著。如《形式语言及其句法分析》(阿霍、厄尔曼，1987)、《翻译算法》(G·米兰，1998)、《自然语言理解》(艾伦，2005)、《统计自然语言处理基础》(曼宁等，2005)、《计算语言学前瞻》(俞士汶、黄居仁，2005)、《现代信息检索》(巴伊赞·耶茨等，2005)等。这些译著

对国内学者了解国外本学科的研究动态具有重要的价值。

这一时期公开发表的论文，成果丰硕，主要刊登在《软件学报》、《计算机学报》、《中文信息处理》、《语言文字应用》、《当代语言学》、《计算机研究与发展》、《计算机科学》、《计算机工程与应用》、《计算机应用研究》、《计算机应用》以及一些大学学报等刊物上。《RJD-80 型汉语人机对话系统的语法分析》(范继淹、徐志敏，1982)、《国外自然语言理解系统简介》(冯志伟，1984)是较早介绍各种上机系统的论文。20 世纪 80 年代陆续发表了不少译介文章。如《国外机器翻译的新进展》(冯志伟，1980)、《自然语言理解的理论和方法》(范继淹、徐志敏，1980)、《自然语言的语义分析技术》(俞士汶，1998)等，对国内学者了解国外研究动态具有重要的意义。

理论方法和学科建设问题一直是学者关注的重点，不少学者对此做过一些有益的探索。宁春岩(1985)发表了《自然语言理解中的几个根本问题》一文，对自然语言理解中的有关重要问题做了深层次的思考。马希文(1989)的论文《以计算语言学为背景看语法问题》从计算语言学的角度探讨了汉语的语法问题，提出了独到的见解。《现状和设想——试论中文信息处理与现代汉语研究》(许嘉璐，2000)则阐述了中文信息处理技术的发展过程，以及朝向自然语言处理技术的必然趋势。《自然语言处理的计算模型》(张钹，2007)讨论了现有各种类型的语言计算模型的特点和局限性，并探讨了解决这些困难的途径。其他重要论文有：《关于中文信息处理的探讨》(黄典诚，1982)、《语言应用和现代化——中文信息处理研究》(刘涌泉，1983)、《中文信息技术和自然语言处理》(袁琦，1986)、《中文信息处理在中国的发展》(苏东庄、袁琦，1990)、《自然语言理解的语言学假设》(袁毓林，1993)、《中文信息处理研究的现状和前瞻》(曹右琦，1995)、《计算语言学应用中的模块化概念》(刘海涛，1995)、《人机并存，“质”“量”合一》(孙茂松、张磊，1997)、《语义的先决性·句法的强制性·语用的选定性》(鲁川，2000)、《自然语言处理技术的三个里程碑》(黄昌宁、张小凤，2002)、《计算符号学》(胡壮麟，2002)、《“句管控”与中文信息处理》(刘云、俞士汶，2004)、《自然语言理解的全信息方法论》(钟义信，2004)等。这些论文反映了学者结合汉语的实际情况思考计算语言学的理论方法问题，在中文信息处理的研究和实践中发挥了重要作用。

不少学者一直在致力于结合汉语的特点思考现有自然语言处理理

论在汉语中的应用问题。

有的学者主张计算机处理汉语不应该以图灵检验为标准，而应该以对语言的模糊的消解能力为第一标准，代表性理论是 HNC(概念层次网络)理论。HNC 理论是我国学者黄曾阳针对汉语提出来的关于自然语言理解处理的一个理论体系。其目标是建立自然语言的知识表述和处理模式，使计算机能够模拟人脑的语言感知功能。HNC 理论开拓了一条从语义分析入手的语句分析之路，走出了一条与传统的句法分析或语义分析不同的路子。黄曾阳的论文《HNC 理论概要》(1997)以及他的专著《HNC(概念层次网络)理论》(1998)是该理论的代表性成果。《HNC(概念层次网络)语言理解技术及其应用》(晋耀红，2006)、《面向机器翻译的汉英句类及句式转换》(张克亮，2007)等，则是对 HNC 理论的具体应用和进一步的发展。

有的学者则主张基于内涵模型的语义分析，代表人物是上海交通大学的陆汝占教授。其学术主张是在一个逻辑句义框架下来分析词汇及其分类，只要能明白表达句义，不必过于精细，也就是用逻辑框架处理词汇理论。基于这一考虑，将汉语表达式抽象成数学表达式，恰当地表示内涵和外延义，然后把这些语义表示在计算机内进行处理，亦即把汉语表达式与计算机数据结构之间直线联结，改变为汉语表达式—抽象数学表示—数据结构三者的间接联结，称之为基于形式方法——模型论的汉语语义计算理论。代表性论文有：《蒙太古语义学》(陆汝占、靳光瑾，1995)、《领属关系与逻辑语义解释》(陆汝占、靳光瑾，1996)、《从汉语句子中提取逻辑函子的一种方法》(靳光瑾、陆汝占，1998)、《现代汉语研究的新视角》(陆汝占、靳光瑾，2004)等。

30 年里，陆续出版了不少有影响的论文集，重要的有中国社会科学出版社 1982 年、1985 年、1986 年连续出版的《语言和计算机》第 1、2、3 期，这是我国当时唯一的机器翻译学术刊物。其他重要论文集有《计算语言学研究与应用》(陈力为主编，1993)、《中文信息处理应用平台》(陈力为主编，1995)、《计算语言学进展与应用》(陈力为、袁琦主编，1995)、《计算机时代的汉语和汉字研究》(罗振声、袁毓林主编，1996)、《计算语言学文集》(俞士汶、朱学锋主编，1996)、《语言信息处理专论》(黄昌宁、夏莹主编，1996)、《汉语言文字信息处理》(陈原主编，1997)、《语言工程》(陈力为、袁琦主编，1997)、《1998 中文信息处理国际会议论文集》(黄昌

宁主编，1998）、《汉语自动分词研究的若干最新进展》（孙茂松主编，2001）、《HNC与语言学研究》（张全、萧国政主编，2001）、《语言计算与基于内容的文本处理》（孙茂松、陈群秀主编，2003）、《中文信息处理若干重要问题》（徐波、孙茂松主编，2003）、《中文信息处理前沿进展——中国中文信息学会二十五周年学术会议》（曹右琦、孙茂松主编，2006）、《中文计算技术与语言问题研究——第七届中文信息处理国际会议论文集》（萧国政、姬东鸿主编，2007）、《语言·认知·信息处理》（李红等主编，2007）等，收录了不少高水平的论文，内容涉及中文信息处理的各个领域，反映了这一时期的整体研究水平。

1.2 文字字的分析与处理

文字的字信息处理主要是汉字的信息处理，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（俞士汶，2006）。汉字的自动分析与处理是中文信息处理的前提和基础，基本任务是解决汉字键盘输入技术、汉字的排版、印刷问题以及汉字的自动识别和汉语的语音识别。

我国于20世纪60年代末就开始对汉字的信息处理进行有益的探索和实践，1968年研制出汉字电报译码机。70年代中期明确提出研究“汉字信息处理系统”，1978年召开了全国汉字编码会议，成立了汉字编码研究会，汉字处理研究被提上重要议程。80年代汉字信息处理进入了飞速发展时期，出现了数百种汉字编码方案和多种输入方式，其中王永民的五笔字型汉字编码有着广泛影响。80年代初，北京大学王选院士主持研制了计算机汉字激光照排印刷系统，使汉字文献的印刷进入电子时代。我国在汉字手写体和印刷体自动识别等领域也取得了较大成就。

该领域的代表性著作有：《汉字属性字典》（傅永和主编，1989），本书收入《信息交换用汉字编码字符集基本集》的6763个汉字，对其读音、笔画数、笔顺、编号、部首、繁体、异体、汉字构成部件、使用频度及各种代码等五十种属性提供准确数据，是一部新型的规范性的多信息字典。《汉字编码方案汇编》（中国汉字编码研究会，1980）是我国第一部关于编码方案的专著，为汉字的信息处理打下了基础。其他重要著作有：《汉字信息处理技术》（郭平欣、张松芝主编，1985）、《汉字属性字典》（北京图书馆编，1988）、《汉字信息字典》（李公宜等，1988），《汉字识别技术》（张炘中，

1992)、《汉字识别——原理、方法与实现》(吴佑寿、丁晓青,1992)、《现代汉语用字信息分析》(陈原主编,1993)、《汉字键盘输入技术与理论基础》(陈一凡、胡宣华,1994)、《语言文字信息处理》(盛玉麟,2006)、《现代汉字特征分析与计算研究》(邢红兵,2007)等。少数民族语言方面出版了《蒙古文编码》(确精扎布,2000)等著作。

代表性论文有:《计算机汉字输入五笔字型编码方案简介》(王永民,1984),详细介绍了五笔字型编码的基本方法和取码规则等问题。《论汉字特征信息编码键盘输入》(陈一凡,1997),介绍了汉字键盘输入技术的发展、汉字的特征信息、汉字小键盘输入原理等五个方面的问题,对汉字的信息处理有重要的价值。《搭建中华字符集大平台》(李宇明,2003),从宏观上讨论了中华字符集的内容及需要解决的技术问题。其他重要论文有:《汉字的熵》(冯志伟,1984)、《谈谈汉字编码输入技术》(白水,1985)、《论汉语拼音、三拼、双拼、简拼的统一表达形式》(王晓龙,1988)、《现代汉语统计频度及其在电脑音轨输入系统中的应用》(陶沙,1988)、《计算机古籍字库的建立与汉字的理论研究》(王宁,1994)、《语句级汉字输入技术》(王晓龙、王幼龙,1996)、《基于模糊方向线素特征的手写体汉字识别》(马少平,1997)、《古文字字库建设的几个问题》(张再兴,2003)、《古维吾尔文(察合台文)及转写符号的智能输入法研究》(地里木拉提·吐尔逊等,2007)等。

1.3 词的分析与处理

词的分析与处理是自然语言处理的一项基础性工作,基本任务是从自然语言的字符串中分析出词串并作相应的处理。主要研究词语自动切分、词性自动标注、未登录词识别、人名与地名的识别等,与此相关的研究还有各种新兴词典的建设。我国的分词研究已有 20 多年的历史,在学者的努力下,无论是基础研究还是应用软件都取得了很大的进展。

20 世纪 80 年代初期,北京航空学院、中国人民大学等十几所院校、研究机构参加"现代汉语词频统计",这是国内首次使用计算机进行大规模语料词频统计研究的大型语言工程。80 年代末,第一个汉语自动分词系统——CDWS 研制成功。此后,许多高校和科研单位致力于分词的研究,开发了系列分词软件,取得了重大成就。我国在各类电子词典开发中也取得了很大的成就,突出成果有:北京大学计算语言所研制的《现代汉语语法信息词典》,清华大学和中国人民大学合作研制的《现代汉语述

语动词机器词典》等。最近，哈尔滨工业大学在《同义词林》(梅家驹等，1983)的基础上建成了《同义词林(扩展版)》，并实现了电子资源共享，成为自然语言处理中的重要资源。

《中文文本自动分词和标注》(刘开瑛，2000)是关于分词的一本重要专著，对汉语分词的基本理论和方法进行了全面系统深入的研究。相关的重要著作还有《汉语词汇的统计与分析》(北京语言学院，1985)、《现代汉语常用词词频词典》(刘源等，1990)、《信息处理用现代汉语分词规范及自动分词方法》(刘源等，1994))、《现代汉语动词大词典》(林杏光等，1994)、《汉字信息处理基础》(朱巧明，1997)、《汉字编码键盘输入文集》(张普，1997)、《现代汉语语法信息词典详解》(俞士汶等，1998)等。

本领域的论文主要集中在以下5个方面：

有的学者从宏观上探讨分词方法问题。代表性论文有：《论汉语自动分词方法》(揭春雨、刘源等，1989)，考察了目前中文信息处理领域中已有的几种主要的汉语自动分词方法，提出自动分词方法的结构模型，并讨论了各种分词方法的复杂度、速度和精度等问题。《从词性标注看小句的中枢地位》(温锁林，2004)，论述了小句在词性的辨别，特别是汉语兼类词的处理中的优势，认为小句是词类的最佳观测站。其他重要论文有：《谈谈词库问题》(刘涌泉，1986)、《基于短语结构文法的分词研究》(韩世欣、王开铸，1992)、《基于神经网络的分词方法》(徐秉铮、詹剑等，1993)、《中文信息处理中的分词问题》(黄昌宁，1997)、《有关汉语分词的几点意见》(进明，1997)、《一种规则与统计相结合的汉语分词方法》(赵伟、戴新宇等，2004)等。

有的学者探讨了分词系统的设计与实现。如《书面汉语自动分词系统—CDWS》(梁南元，1987)、《基于规则的汉语自动分词系统》(姚天顺等，1990)、《汉语自动分词实用系统CASS的设计和实现》(揭春雨，1991)、《串频统计和词形匹配相结合的汉语自动分词系统》(刘挺、吴岩等，1998)、《藏文自动分词系统的设计与实现》(陈玉忠、李保利等，2003)等。

有的学者讨论了分词标准相关的问题。代表性论文有：《关于分词规范和规范词表的若干意见》(袁毓林，1997)，针对分词的规范和规范词表的制定，发表了富有价值的见解。其他重要论文有：《字词频统计与汉语分词规范》(刘源，1992)、《关于分词规范的探讨》(宋柔，1997)、《关于

汉语分词问题之我见》(杨成凯,1997)、《谈谈制定信息处理用汉语词表的策略》(孙茂松、张磊,1997)、《浅谈汉语分词的标准》(孙宏林,1997)、《现代汉语五万词语归类的实践》(朱学锋、俞士汶等,1997)等。

有的学者研究了未登录词、人名与地名的识别问题。如《汉语姓名自动识别初探》(郑家恒、刘开瑛,1994),分析了汉语姓名在各种类型汉语文本中的分布情况、汉语姓名组成的复杂性和自动识别姓名的难点,提出了自动识别姓名的策略和规则。《新词语的监测与搜获——一个汉语本体研究者的思考》(邢福义,2007),从汉语本体研究的角度,就如何监测与搜获新词语的问题提出若干意见。其他重要论文有:《中文姓名的自动识别》(孙茂松、黄昌宁等,1995)、《中文机构名称的识别与分析》(张小衡、王玲玲,1997)、《基于统计的中文地名识别》(黄德根、岳广玲等,2003),《基于语料库的中文姓名识别方法研究》(郑家恒、李鑫等,2000)等。

有的学者探讨了分词中歧义字段的切分问题。重要论文有:《汉语自动分词及歧义组合结构的处理》(李国臣、刘开瑛等,1988)、《关于歧义字段切分的思考与实验》(刘挺、王开铸,1988)、《汉语自动分词中的歧义问题》(侯敏、孙建军,1996)、《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》(孙茂松、黄昌宁等,1997)、《中文文本歧义字段切分技术》(温锁林,2001)、《基于语料库的高频最大交集型歧义字段考察》(李斌、陈小荷等,2006)、《用基于词的二元模型消解交集型分词歧义》(陈小荷,2004)等。

1.4 句法分析

句法分析是中文信息处理领域一个重要的基础性课题,基本任务是通过设计某种算法,准确地分析自然语言中的具体句子的句法结构和语法特征。常见的方法有自顶向下分析法、自底向上分析法、左角分析法等。

由于词性的多样性和词义的复杂性使得自然语言中的句子结构多样,充满了歧义,且汉语缺少形态标记,在语法上具有趋简性、兼容性(邢福义,1997)和句位高效(储泽祥,2010)等显著特点,这些特点对意合型的汉语本身可能是优点,但对中文信息处理来说却是缺点。因此,汉语的句法自动分析面临诸多困难,成为计算语言学界公认的难题。当前最重要的任务是根据汉语语法体系分门别类地建立形式化的可为计算机

利用的语法规则。

与信息处理相关的句法理论在概论性著作中一般都会安排专门章节进行介绍。此外,《面向中文信息处理的现代汉语短语结构规则研究》(詹卫东,2000)值得关注,该书尝试用形式化的方式从句法和语义两个层面归纳现代汉语短语结构的组合规则,为自然语言处理提供了必要的语言知识。

代表性论文有:《汉语句型自动分析和分布统计算法与策略的研究》(罗振声、郑碧霞,1994),该文以结构特征为标准的句型系统,提出以谓语为中心的句型成分分析与句型匹配相结合的分析算法与策略,讨论了句型成分及其短语边界的识别与判定方法,给出了有关歧义结构的处理策略。其他重要论文有:《中文信息处理中的切词和句法分析》(刘倬,1985)、《统计与规则并举的汉语句法分析模型》(周明、黄昌宁等,1994)、《汉语单句谓语中心词识别知识的获取及应用》(穗志方、俞士汶,1998)、《汉语句法规则的自动构造方法研究》(周强、黄昌宁,1998)、《基于局部优先的汉语句法分析方法》(周强、黄昌宁,1999)、《浅层句法分析概述》(孙宏林、俞士汶,2000)、《基于短语结构语法的自动句法分析方法》(冯志伟,2000)、《句处理中排歧问题补议》(陆俭明、王黎,2003)、《一种基于句法语义特征的汉语句法分析器》(杨开城,2000)、《基于词性和语义知识的汉语句法规则学习》(苑春法、陈刚等,2001)、《四种基本统计句法分析模型在汉语句法分析中的性能比较》(孟遥、李生等,2003)、《基于句子对齐的汉语句法结构推导的计算模型》(王厚峰、王波,2003)等。

在我国中文信息处理领域,句子的研究比字、词的研究起步晚,研究也相对薄弱。目前对句子的研究重点是单句,但是随着句法研究的深入,复句研究的重要性日益明显。面向信息处理的复句研究也开始受到学界的重视。著作方面,《复句关系标记的搭配研究》(姚双云,2008),利用汉语复句语料库和大规模连续文本研究关系词的搭配问题,在复句的信息处理方面有一定的参考价值。代表性论文有:《汉语复句本体模型初探》(胡金柱、王琳等,2005)、《本体论在复句领域概念建模中的应用》(胡金柱、罗旋等,2006),两文依据邢福义的复句分类系统,引进本体知识建模方法,建立复句本体模型,反映了面向信息处理的复句研究的最新动态。其他重要论文有:《汉英机器翻译中描述型复句的关系识别和处理》(鲁松、宋柔,2001)、《汉语多重关系复句的关系层次分析》(鲁松、

白硕等,2001)、《汉语复句的结构分析》(张仕仁,1999)、《面向计算机的二重复句层次划分研究 》(李晋霞、刘云,2003)、《自然语言处理中句群划分及其判定规则研究》(吴晨、张全,2007)、《并列复句的自动识别初探》(周文翠、袁春风等,2008)等。

1.5 语义问题研究

语义问题也是中文信息处理领域的研究重点,这是比句法问题更困难的研究课题。但是,要真正实现计算机理解自然语言,语义是一个无法回避的问题。语义计算的主要任务是:解释自然语言句子或篇章各个部分(词、词组、句子、段落、篇章)的意义。语义研究目前遇到的困难主要体现在:(1)自然语言句子中存在大量的歧义格式(包括语音、词汇、句法、语用等引起的歧义),涉及指代、同义、多义、量词的辖域、隐喻等方面;(2)同样的句子对于不同的人来说可能有不同的理解,而导致不同理解的因素又是复杂多变的,如知识背景、文化程度、语言本身的主观性等;(3)语义计算的理论不少,但是缺少经得起检验的理论、方法与模型。特别是语义的形式化表示是一个大难题,因为人脑具有丰富的世界知识,可以迅速地激活语言符号所蕴含的意义。但是对缺少世界知识的计算机来说,显然是困难重重。

正因如此,语义处理的首要任务是利用有限的符号对各级语法单位进行语义的形式化处理和有效分析。语义形式化问题解决的好坏,将大大影响语义自动分析与语言加工的成效。鉴于语义研究的重要性,不少学者纷纷投入了这一领域的研究中。

语义研究的重要著作有:《汉语计算语义学——关系、关系义场和形式分析》(吴蔚天,1999),从语义学的角度来探讨汉语的形式分析问题,重点讨论了词语之间的语义关系,对名词、动词的语义进行分类,构建关系语义场等问题,处理方法具有较强的操作性。《词汇语义和计算语言学》(林杏光,1999),着力讨论自然语言处理中词汇和语义研究需要具备的理论知识以及对词义的分类,提出了一些新的理论和方法,也是计算语义研究的一项重要成果。其他重要著作有:《现代汉语动词语义计算理论》(靳光瑾,2001)、《句式语义的形式分析与计算》(吴平,2007)等。

代表性论文有:《一个面向工程的语义分析体系》(陈小荷,1998),介绍了语义分析的设计思想、基本结构、基本方法和应用范围。《基于配价的汉语语义词典》(詹卫东,2000),提出了一个“广义配价模式”表达语义