中文信息处理丛书

詹卫东 著

A Study of Constructing Rules of Phrases in Contemporary Chinese for Chinese Information Processing

面向中文信息处理的现代汉语短语结构规则研究

清华大学出版社
广西科学技术出版社

中文信息处理丛书

面向中文信息处理的现代汉语短语结构规则研究

A Study of Constructing Rules of Phrases in Contemporary Chinese for Chinese Information Processing

詹卫东 著

清华大学出版社
广西科学技术出版社

(京)新登字 158 号
(桂)新登字 06 号

内 容 简 介

本书从语言学为中文信息处理提供支持的角度，对现代汉语的短语结构规则进行了全面和系统的研究。主要内容包括现代汉语短语的句法语义范畴体系，现代汉语短语结构规则的形式化描述，现代汉语短语结构歧义格式的类型分析及排歧策略研究等三个方面。

本书可供从事中文信息处理、现代汉语语法、计算语言学以及理论语言学研究的人士参考，也可以作为大学相关专业高年级本科生和研究生课程的参考书。

书　名：面向中文信息处理的现代汉语短语结构规则研究
作　者：詹卫东　著
出版者：清华大学出版社（北京清华大学学研楼，邮编 100084）
http：//www.tup.tsinghua.edu.cn
广西科学技术出版社
印刷者：北京昌平环球印刷厂
发行者：新华书店总店北京发行所
开　本：787×1092 1/16　**印张**：12.25　**字数**：281 千字
版　次：2000 年 7 月第 1 版　2000 年 7 月第 1 次印刷
书　号：ISBN 7-302-03940-2/TP・2303
印　数：0001～3000
定　价：19.00 元

清华大学出版社　广西科学技术出版社
计算机学术著作出版基金

评 审 委 员 会

主 任 委 员　张效祥

副主任委员　汪成为　唐泽圣

委　　员　（按姓氏笔画排列）

王鼎兴　杨芙清
李三立　施伯乐
徐家福　夏培肃
董韫美　黄　健
焦金生

出 版 说 明

近年来，随着微电子和计算机技术渗透到各个技术领域，人类正在步入一个技术迅猛发展的新时期。这个新时期的主要标志是计算机和信息处理的广泛应用。计算机在改造传统产业，实现管理自动化，促进新兴产业的发展等方面都起着重要作用，它在现代化建设中的战略地位愈来愈明显。计算机科学与其它学科的交叉又产生了许多新学科，推动着科学技术向更广阔的领域发展，正在对人类社会产生深远的影响。

科学技术是第一生产力。计算机科学技术是我国高科技领域的一个重要方面。为了推动我国计算机科学及产业的发展，促进学术交流，使科研成果尽快转化为生产力，清华大学出版社与广西科学技术出版社联合设立了“计算机学术著作基金”，旨在支持和鼓励科技人员，撰写高水平的学术著作，以反映和推广我国在这一领域的最新成果。

计算机学术著作出版基金资助出版的著作范围包括：有重要理论价值或重要应用价值的学术专著；计算机学科前沿探索的论著；推动计算机技术及产业发展的专著；与计算机有关的交叉学科的论著；有较大应用价值的工具书；世界名著的优秀翻译作品。凡经作者本人申请，计算机学术著作出版基金评审委员会评审通过的著作，将由该基金资助出版，出版社将努力做好出版工作。

基金还支持两社列选的国家高科技重点图书和国家教委重点图书规划中计算机学科领域的学术著作的出版。为了做好选题工作，出版社特邀请“中国计算机学会”、“中国中文信息学会”帮助做好组织有关学术著作丛书的列选工作。

热诚希望得到广大计算机界同仁的支持和帮助。

清华大学出版社
广西科学技术出版社　　计算机学术著作出版基金办公室

1992 年 4 月

中文信息处理丛书

序　言

中文信息处理技术在我国现代化及信息化建设中，越来越起着重要的作用，做为一个高新技术的重点，它已经列入国务院批准的“国家中长期科学技术发展纲领”。我国的中文信息处理事业正在不断向前推进，在技术研究、产品开发以及产业化发展等方面都取得了显著的成绩。现在有必要把这些方面的成果加以综合、提炼，以便更好地推广应用，并且作为一个起点，再上一个新台阶。中文信息处理在从汉字信息处理进入汉语信息处理之后，在从单机信息处理进入网络信息处理之后，已经面临着新的更大的挑战和机遇，需要我们重新对中文信息处理进行全面的审视与整合，这就是我们组织编写并出版这套中文信息处理丛书的目的。

在这套丛书出版之际，我愿向读者介绍以下几点：

第一，为什么我们要把中文信息处理技术作为高新技术的一个重点来发展呢？

语言文字是信息的首要载体。我们日常工作中的信息，绝大部分是以语言文字表达、记载、传播和交换的。因此随着计算机和因特网的推广应用，由数据处理、信息处理发展到知识处理，对语言文字处理要求的深度和广度越来越高，可以认为一个国家的语言文字的信息处理水平和处理量基本上代表了这个国家进入信息社会的程度，其语言文字信息的处理能力直接关系到它在网络社会和网络经济中的国际竞争能力。目前，网络社会和网络经济正以我们难以预料的速度在全世界发展，其阻碍发展的首要瓶颈问题就是自然语言的处理问题。网络社会也是人类社会，网络经济也是人类经济，需要以自然语言作为社会交际工具，一旦基于网络的自然语言处理问题得到突破，网络社会和网络经济将会突飞猛进。我们要在下一个世纪成为世界强国，就不能不把语言文字信息处理技术作为高新技术的一个重点来发展。在世界一流高新技术企业纷纷在中国设立“中国研究院”，争先把“中文信息处理”作为研究的重中之重的时候，我们当然要抢占中文信息处理这个高新技术发展的制高点。

第二，中文信息处理与印欧语系的语言信息处理的不同之处是什么？

计算机从诞生之日开始，就是以处理印欧语系的语言为基础的。换言之，计算机对于印欧语系的自然语言处理具有较好的支撑能力，计算机的推广应用在语言文字信息处理方面受到的阻力较小。我们的汉语却与印欧语系的语言差别很大，能够处理那些语言的计算机，面对汉语汉字，却显得无能为力。例如：

· 印欧语系为拼音文字，所使用的字符仅二十余个，而汉语是意音文字，常用的汉字就有六七千个，总数超过五万。这是一个根本性的问题。仅这一个差异就引起了处理汉语的计算机与处理印欧语言的计算机一系列的差异，需要我们自己去解决。包括键盘输入、汉字打印与显示、内部代码、汉字识别、程序语言的数据类型、数据库的检索和排序等等。

· 印欧语系的书写，词与词之间有空格，而书面汉语的词与词之间无空格，于是词的机器自动切分问题就成了计算机处理汉语的首要问题。

· 印欧语系的同音词较少，而汉语的同音词较多。例如，仅在《现代汉语词典》中 JI 音汉字就有一百多个，辨析同音词就成了汉语语音处理的关键。

· 印欧语系多有形态变化（例如：复数、单数，过去、现在，阴性、阳性等等），而汉语缺少形态变化。计算机对汉语的处理（例如机器翻译、人机接口等）无法利用形态变化，只能在句法、语义上找出路。

· 汉语的语法研究尚未形成规范化，而且人们习惯于约定俗成的语法。于是语义研究显得尤其重要。例如，“吃饭”“吃大碗”和“吃食堂”的理解只能靠语义来解决。

· 汉语的自动（计算机）处理是多学科和跨学科的研究工作，特别需要计算机科学与语言学、认知科学等学科的密切结合，而且要依靠长期积累的语言学的研究成果。但我国语言学界过去的研究多着重汉语教学，对象是人，而不是机器，因此对其丰硕的研究成果要经过改造、深化、量化、形式化，甚至要从头开始。要清醒地认识到面向机器的汉语研究的艰巨性，要持续不懈地抓下去。

以上只是几个突出的问题，还有一些其他问题，不再赘述。这些语言上的特点造成了计算机处理汉语的众多障碍，每前进一步都会遇到新问题，我们不得不花费比印欧语系的信息处理多得多的力量去解决。

再就计算机的发展趋势而言，计算机产业面临转型期，多媒体和笔记本式计算机成为热门产品，计算机从单机进入网络，网上的汉语信息处理正在成为强势和主流，并对语言文字的信息处理提出新的要求。这些产品的核心技术无不与中文信息处理技术有关。因此，加强中文信息处理的研究，取得网络化的自然语言处理的突破更为必要。

第三，中文信息处理技术包括哪些科目呢？

大体上包括下列一些科目：

· 词的切分和频率统计
· 汉语句型和短语的研究及频率统计
· 汉语语义的研究
· 键盘和非键盘汉字输入技术及处理系统
· 汉语语料库的开发及应用
· 汉字的机器代码，程序设计语言的数据类型
· 汉语开放系统的接口规范
· 语声输入与合成
· 汉字识别
· 字形生成
· 汉语分析及篇章理解
· 汉语生成
· 人机接口
· 汉外外汉机器翻译
· 信息检索

- 自动标引和抽词，自动文摘、文本自动分类与网站自动分类、信息自动提取与知识挖掘
- 全文检索
- 电子印刷出版系统
- 汉语辅助教学与现代远程教学
- 电子词典等

以上这些科目，有些是基础研究，有些是技术研究，也有些可以直接转化为产品。这些科目的分类并非学科分类，不过是按照编者本人日常接触的项目，把它们罗列出来而已。其分类的科学性、正确性和完整性尚待商榷。必须指出，有些基础性研究虽然看不到直接的经济效益，但它的研究成果则是其他研究工作所必需的，而且要先行。

到目前为止，在上述这些项目中，有些已经产业化，例如电子印刷出版和少数几个汉字输入系统；有些项目已经商品化，正向产业化迈进；很多项目已经实用化。但每个领域都有很多问题等待我们去解决。今后的工作只能加强，不能削弱，使我们中文信息处理的每个领域，每个项目都沿着实用化、商品化和产业化的道路奋勇前进。我相信我们这套丛书必将在促进中文信息处理技术的发展方面发挥它应有的作用。这套丛书将陆续出版。

最后，感谢“计算机学术著作出版基金评审委员会”把出版中文信息处理丛书列入了出版计划。感谢清华大学出版社和广西科学技术出版社给予出版基金的支持。

中国工程院院士　陈力为

1992 年 5 月于北京序

2000 年 4 月于北京修订

中文信息处理丛书编委会

A Study of Constructing Rules of Phrases in Contemporary Chinese for Chinese Information Processing

Abstract

This book, which is oriented towards Chinese Information Processing (CIP) by computer, proposes a set of formulized rules on Chinese phrase structures, and discusses the treatment of the phrase structure disambiguation. The full text consists of 7 chapters.

Chapter one: The status of development of CIP and the current level of study on Modern Chinese grammar are discussed preliminarily in a broad outline. Based on it, the system of Chinese phrase structures is chosen as the subject of this research, and the goal of the book is set to create a RuleBASE including a set of Chinese phrase structure rules with rich constraints. It is worth noting that such a RuleBASE must be supported by a lexicon. which contains vast amount of syntactic and semantic features related to every lexical entry.

Chapter two: A classification system of Chinese phrase is put forward firstly, and other syntactic categories for description of phrase structures are also defined. All of these syntactic attributes, which are based on the theory of Phrase-standard Grammar system proposed by Prof. Dexi Zhu, will be used for describing the functions of phrases. At the same time, a semantic expression framework, named as Generalized Valence Mode, is designed for describing the semantic features and functions of a word or a phrase. Furthermore, a simple semantic taxonomy of Chinese content words is also built up.

Chapter three: With the existent syntactic and semantic categories, a constraint-based Chinese phrase-structure rule system is constructed. Each rule includes two parts: a context free rewrite rule and a series of unification equations. The former is used for describing the construction of a compounded phrase, and the latter is used for describing the functions of the compounded phrase and the constraints of the constituents including syntactic constraints and semantic constraints. As a result, there are 89 rules induced in this chapter for most types of phrases, i.e. np,ap,vp,dj.

Chapter four: This chapter analyzes the ambiguity of determining boundaries and structural relations of Chinese phrases in automatic parsing by computer. Seen from different perspectives, all of the ambiguous phrases can be classified into different

types. In terms of components of ambiguous structures, ambiguous phrases can be classified into two categories: one including terminal symbols, the other not including terminal symbols but only non-terminal symbols. In terms of the influence of ambiguity, ambiguous phrases can also be classified into two categories: self-confined ambiguous phrases and non-self-confined ambiguous phrases. The influence of the former ambiguity is mainly inside the ambiguous phrases. The influence of the latter ambiguity is outside of the ambiguous phrases. As viewed from differentiated types of the relation between type and token, ambiguous phrases can be classified into three categories: the ture-ambiguous phrase, the quasi-ambiguous phrase, and the pseudo-ambiguous phrase. Depending on the above analysis and the set of rules proposed in Chapter three, I also survey all ambiguous phrases in Modern Chinese and their various types of ambiguity.

Chapter five: This chapter takes the further step to demonstrate how to solve ambiguities of phrases. And the reasons why some ambiguous phrases are not easy for disambiguation are also discussed and can be regarded as a reference for future research.

Chapter six: The results of parsing example sentences are presented to show the capability of the RuleBase that has been integrated into a parser.

Chapter seven: The last chapter concludes the main achievements and significance of this research. Planning of further research is also proposed.

The research done in my book is devoted to two fields: the grammar of modern Chinese and CIP. The result of this research can be used directly, or as a significant reference for supporting various CIP applications. On the other hand, CIP can provide a clearer background of application for study of Chinese grammar. As viewed from the computer, not human beings, some questions that were not observed before would be revealed more easily. In addition, these questions also can be expressed more definitely and normatively within the formula schema expatiated in this book.

Keywords: Phrase structure grammar, syntactic category, semantic category, rule, unification, generalized valence mode, Chinese Information Processing (CIP).

序

中文信息处理，我国从20世纪50年代就起步了，这是从俄汉机器翻译开始的。《中国语文》1959年11月号，报道了“俄汉机器翻译初步试验成功”的消息。该项研究是由当时隶属于中国科学院的语言研究所和计算技术研究所合作进行的。他们以俄文数学文献彼德罗夫斯基的《偏微分方程讲义》一书为主要材料，试图通过他们所研制的俄汉机器翻译系统将该书翻译成汉语。据报道这套俄汉机器翻译系统“对于翻译该书的大部分句子都是有效的”。这里，我们不想搞清楚当时的这套俄汉机器翻译系统是否真能如报道所说“对于翻译该书的大部分句子都是有效的”，但是有一点大概可以肯定，这套俄汉机器翻译系统并未真正解决汉语的字处理、词处理、句处理等问题。但是，这次成功的试验，对中文信息处理来说，毕竟是个可喜的、值得庆贺的开端。与此同时，当时的北京外国语学院、广州华南工学院、哈尔滨工业大学等高等院校，也分别成立了机器翻译研究组，开展俄汉或英汉机器翻译的研究试验。当时，在机器翻译方面，我国的研制水平可以说跟苏联和欧美不相上下。但由于众所周知的原因，从60年代中期至70年代中期，我国中文信息处理研究工作虽没有完全停顿，但进展缓慢。70年代末，80年代初，跟其他学科一样，中文信息处理迎来了发展的春天；特别是进入90年代之后，由于计算机学界与语言学界双方更紧密地结合，中文信息处理出现了大发展的情景。到目前为止，我们已基本上解决了“字处理”（汉字输入和显示）的问题，初步解决了“词处理”（中文自动分词、词性标注）的问题，但离信息科学发展的需要还有相当大的距离。

20—21世纪，可以说是人类社会又一个大的转折时期——从工业时代步入信息时代。进入21世纪之后，20世纪后期开始建立起来的“信息高速公路”将通遍全球，进入千家万户。正是这种时代发展趋势，促使自然语言信息处理成为目前全球关注的研究热点。据有关报导，许多国家，特别是美国、日本、欧共体等，从20世纪80年代开始就已投入大量人力、物力，有的作为政府行为，加速智能计算机的研制开发工作；在着手研制开发智能计算机的过程中，它们都不约而同地把语言信息处理放在非常重要的地位来考虑，并都希望自己在信息科学与工程领域内独占鳌头，能起操纵的作用。

就中文信息处理来说，眼下特别要集中精力解决好“句处理”问题。现在句处理有多种策略和途径——有基于句法规则的，有基于概念网络的，有基于语料库统计的，有基于语义计算的，等等，形成了一个竞相研究、竞相发展的局面。这不能不说是一个可喜的现象。但也不能不看到，在这种竞相研究、发展中，存在着各执一端、唯我独是的门户之见，这将严重影响我国信息科学的发展，削弱我国信息科学在国际上的群体竞争能力。其实，在目前，我们很难说哪一种策略和途径是唯一正确、唯一合理、唯一可取的。须知，无论使用哪一种策略与途径，都离不开我们对汉语的认识，离不开有关汉语的知识。而我们对汉语的认识，或者说我们应具有的有关汉语的知识，应该是一种涉及到语音、语义、语法、语用等

诸方面的综合的知识，因为人用语言向对方表达自己的思想、看法、情感，或者从对方的话语中准确理解对方的思想、看法、情感，都须经过一个复杂的编码或解码的过程，而在这个编码或解码的过程中事实上要调动各种各样的因素，单就语言这个角度说，起码也得调动语音、语义、语法、语用等各方面的因素（如果是通过书面语言进行交际，还得调动视觉方面，或者说图像方面的因素）。因此，各种策略和途径我们都需给以足够的重视，都应给以足够的支持；同时也都难免存在偏颇的缺陷。因此，各种策略和途径都应该继续深入研究下去，各种策略和途径可以而且应该各显神通；但同时一定要互相吸取，取长补短，通力协作，逐步形成在信息科学领域里能在国际上与他国抗衡的群体竞争力量。我们应该建立这样的共识：不是我自己或者我们单位自己，而应该是我们整个国家，在不太长的时间里，在中文信息处理，乃至自然语言理解和处理方面，从工程到理论，能达到世界先进水平，继而能居世界领先地位。

在这里我们不能不提醒大家注意这样一点：即使是中文信息处理我们也面临着严峻的国际挑战。我们需要清醒地看到，不要以为“中文信息处理中的句处理”我们一定是大拿，优势一定在我们中国人手里。就目前的形势看，我们只能说“中文信息处理中的句处理”的优势有可能在我们手里。我们需要了解这样一个事实：中文信息处理，国外早就注意并着手研究了。以往，他们是在国外或者将研究课题交给中国有关研究机构或高等院校来做，他们出钱；或者他们从中国雇人去他们那儿进行研究。这两年来，起了变化，他们陆续进驻中国，在北京、上海等地设立中文信息处理的研究机构或基地，以高薪雇佣中国研究人员（对他们来说，比在国内所花的费用还是低得多），与中国研究机构与高等院校争夺人才，争夺中文信息处理的“制高点”。因此，如果我们不觉醒，如果我们还是上面不重视、不积极支持，下面不团结、不合作，那么这中文信息处理的“制高点”不要几年就会被外国公司或研究机构所占领。这绝不是危言耸听，而是严酷的现实。

詹卫东同志的《面向中文信息处理的现代汉语短语结构规则研究》是属于基于规则的策略和途径方面的一项研究成果。汉语的短语结构是汉语句子的基础结构。这项研究工作的目的，是尝试以形式化的方式对现代汉语短语结构的组合规则进行全面的描写，并探讨解决短语结构歧义问题的途径，以便为计算机提供处理和理解汉语句子所必不可少的汉语知识。全书共七章，第一章“引论”，扼要地对中文信息处理技术的发展状况和目前现代汉语语法研究的水平进行宏观的评介，以此说明该项研究的基础和出发点；第二章“现代汉语短语句法语义属性范畴的确立”，主要是提出了一个综合运用句法语义属性的面向中文信息处理的分析、描写短语结构的理论框架，在这个理论框架中建立了汉语实词的分类系统和带有开创性的“广义配价模式”；第三章“现代汉语 np、ap、vp、dj 的句法语义规则”，对现代汉语里四类短语结构——名词性短语（np）、形容词性短语（ap）、动词性短语（vp）和主谓短语（dj）的组合规则进行了系统而具体的形式化描写，列出了 89 条关于这四类短语的句法语义规则，基本上呈现了现代汉语短语结构规则的主体面貌，从而把以往汉语学界从句法、语义两方面所作的面向人的有关现代汉语短语结构的研究成果跟作者自己在这一方面所作的面向计算机的研究成果结合起来，组织成了一个可以为计算机分析现代汉语短语结构提供直接支持的规则库；第四章“现代汉语短语结构歧义类型分析及分布统计”和第五章“现代汉语短语结构歧义的消解策略分析”，细致分析了计算机处理现代

汉语短语结构时所面临的“定界歧义”和“结构关系歧义”的问题，从不同角度对现代汉语短语结构歧义的不同类型进行了分析，而且通过统计获得了一份比较完整的、计算机分析现代汉语短语结构时可能碰到的种种歧义格式的清单，针对不同类型的短语结构歧义的特点，对相应的排歧策略做了探讨，并对一些典型的短语结构歧义格式，提出了虽是初步的但明显是有效的排歧办法。第六章“实验结果示例及难点分析”，向读者具体而如实地报告了作者运用上述种种规则对从调试规则所用的语料中抽取的100个例句让计算机进行自动分析的结果，结果显示用作者现有的短语结构规则分析短句，效果还是可以的，但有些歧义现象（如“我和我最好的朋友在这里堆雪人打雪仗”里的“和”既可以看作是连词，又可以看作是介词），在作者现在所提供的短语结构规则描写框架下还无有效的解决办法；第七章“结语”，一方面作者对自己所作的这项研究工作进行了较好的总结，指出了这项研究工作的意义与可能有的贡献，说明了作者自己通过这项研究工作所获得的有关认识，提出了今后在该项研究上的进一步的设想。

本书是作者在博士论文的基础上修改加工而成的。该书到底写得怎么样，应由广大读者，尤其是这方面的行家去加以评论，我作为作者的导师，不便在这里多说什么。不过，请允许我在这里介绍几位中文信息处理方面的专家在詹卫东同志博士论文答辩会上的一些评论意见。

冯志伟研究员（国家语言文字工作委员会）说：我是很挑剔的。我对詹卫东论文中的89条规则逐一进行了检查，想挑出些毛病或破绽，结果没有发现。他这些规则可直接用于语言信息处理。

张普教授（北京语言文化大学语言信息处理研究所所长）说：詹卫东同志的论文写得很朴实，毫无哗众取宠、故弄玄虚之处，不光写自己研究中获得成功的内容，也如实地摆出了问题与难处。另外，有很强的可读性。有的人把本来很好懂的道理讲得让人看不懂。中文信息处理中的许多理论规则对一般人来说不是很好懂的，而詹卫东这篇论文能把很难懂的理论规则说得深入浅出，通俗易懂，这是难能可贵的。

刘群副研究员（中国科学院计算技术研究所）说：我们正在搞汉英机器翻译，詹卫东同志所提出的短语结构规则和排歧策略在我们的系统中试用的效果是不错的。而且作为文科背景的研究人员，他能很好地跟计算机背景的研究人员合作，把语言知识尽可能合理地安排到一个形式化的框架中。

俞士汶教授（北京大学计算语言学研究所副所长）说：我对博士生论文的要求是很严的。过去也好，今年也好，好几篇博士论文曾有评委提出来是否可考虑评为优秀论文，并写进答辩委员会的决议中去，我都曾持否定意见。但詹卫东这篇博士论文我确实认为可以称得上优秀论文。（附注：俞士汶教授关于“詹卫东同志的博士论文可评为优秀论文”这个意见经答辩委员会一致同意后，写入了答辩委员会的决议中。）

黄国营教授（清华大学中文系）说：该文在理论与实际的结合上尽了最大努力，为中文信息处理提供了一个有价值的理论框架和许多可行的具体操作规则。

请原谅我在这里只介绍了他们几位对詹卫东论文的褒扬之词。当然他们也提出了许多宝贵的修改意见。论文答辩以后，詹卫东同志正是根据各位答辩委员和他的师兄弟们所提出的意见，对论文进行了认真的修改。

最后，我想用詹卫东同志自己书中的两段话来结束这篇序文。

一个研究课题总是针对一个或一些特定问题的。一方面，探索真理的路永远都没有尽头；另一方面，在一个具体的研究课题范围内，对现有问题的解决通常是有一定限度的。因此，在一个研究课题暂时告一段落，人们要思量下一步该如何去做的时候，也无非是在这两个方面做更多的努力，即：一面结合更多的实践，对现有的框架进行检验并向纵深挖掘；一面在现有的研究成果基础上，探索如何开辟更广阔的研究空间。

本研究工作可以看作是一个更为宏大的目标——“编写一部给计算机用的现代汉语语法”——的一部分。虽然距离语法大厦的最终建成还有许多路要走，但我们希望，已经迈出的这一步能够或多或少、或正面或反面地昭示未来的方向。如果本课题的研究工作能够成为将来真正完整意义上的“计算机用现代汉语语法”的一个组成部分，那么毫无疑问走这一步是值得的，因为它是通向成功的足迹中的一个；如果将来的“计算机用现代汉语语法”全然是另外一幅图景，那么这一步也是值得的，因为它虽然没有留下一个成功的印迹，但至少竖起了一个“此路不通”的标牌。

陆俭明

于北京大学中关园寓所

2000 年元旦

前　言

面向人写的汉语语法书已经非常多了，面向计算机写的汉语语法书则还很少见。众所周知，计算机处理自然语言困难重重，最常被人们提及的恐怕莫过于计算机不懂得人类所用的自然语言的语法这一问题。那么，如何让计算机懂得自然语言的语法呢？进一步地，要让计算机能够理解汉语，处理中文信息，需要汉语研究者为计算机准备一部什么样的汉语语法呢？

《面向中文信息处理的现代汉语短语结构规则研究》可以看做是在前人已经开始的许多研究工作的基础上，为回答上述问题所迈出的新的一步。

本书面向中文信息处理的实际需要，尝试以形式化的方式对现代汉语短语结构的组合规则进行全面的描写，并探讨解决计算机分析汉语短语结构碰到的各类歧义问题的途径。全书共分 7 章。

第一章对汉语信息处理技术的发展状况以及目前现代汉语语法研究的水平进行了宏观分析。以此为背景，确定了本书研究课题所针对的对象是短语结构，预期的目标是完成一个带有丰富约束条件的现代汉语短语结构规则库。特别值得指出的是，这样的短语结构规则库是以一部对现代汉语词语进行了全面句法语义属性描述的电子词典作为底层支撑的。有关电子词典的语法部分的详细介绍，请读者参阅本系列丛书中由俞士汶教授等人所著的《现代汉语语法信息词典详解》(简称《详解》)一书。本书跟《详解》一书选择同样的语法理论框架——词组本位语法体系——作为开展研究工作的基本立场。从某种意义上说，本书的研究内容是《详解》一书的自然延伸。

第二章贯彻词组本位语法体系以功能为原则建立句法范畴的精神，将以往对词的句法功能分类和属性特征的研究进一步全面拓展到短语结构上，得到了一个相对完整的短语结构功能分类体系，并初步确立了一套描述短语结构句法功能属性的范畴体系。同时吸收了汉语配价理论、动词格框架等具体研究成果并加以拓展，提出了一个面向中文信息处理的综合的语义信息描述框架 ——“广义配价模式”和一个简化的语义分类体系。这部分工作为进一步开发一个短语结构规则库打下了坚实的范畴基础。

第三章在上述句法语义属性范畴的基础上，对 4 类主要的现代汉语短语结构：np、ap、vp 和 dj 的组合规则进行了系统地形式化描写。这部分工作可以概括为，将以往面向人所做的有关汉语短语结构的句法语义研究的成果，加上作者的研究和实践，组织成一部可以为计算机分析汉语短语结构提供直接支持的规则库。从形式上讲，一条短语结构规则包括两部分，产生式规则和合一等式。产生式规则用于描述汉语短语结构的一种组合可能性，合一等式则进一步描述这个特定的组合模式的整体性质及组合条件。本章总结了有关上述 4 类短语的规则共 89 条。

第四章详细分析了计算机处理汉语短语结构时所面临的定界歧义和结构关系歧义问

题，从不同角度区分了抽象的歧义格式的不同类型：包含终结符的歧义格式与不含终结符的歧义格式；外显型歧义格式与内含型歧义格式；真歧义格式、准歧义格式、伪歧义格式等。在已有的短语结构规则的基础上，对现代汉语短语结构歧义格式（不含终结符的3项排列歧义格式和含终结符“的”跟“和”的4项和5项排列歧义格式）进行了统计，得到了计算机分析现代汉语短语时可能碰到的歧义格式的一个比较完整的清单。

第五章则在对汉语短语结构歧义有了全面认识的基础上，通过对3个典型短语歧义格式的分析，进一步探讨了排歧策略，并对难以或无法在短语结构规则层面解决的歧义问题，指出困难所在，以期为进一步的排歧研究提供参考。

第六章以计算机分析实例的结果展示了本书归纳的短语结构规则在一个具体的汉语句法分析器中使用的实际效果，同时对造成某些分析结果不佳的原因进行了解释。

第七章对本书涉及的研究工作进行了全面总结，包括具体的研究成果、对汉语信息处理研究所能提供的支持以及对汉语语法研究的意义等，最后对进一步的研究工作进行了规划。

本书的研究工作是跨现代汉语语法和中文信息处理两个领域进行的。一方面，研究的具体结果对推进中文信息处理技术的发展应该会有直接的应用和参考价值；另一方面，从中文信息处理的角度来审视现代汉语语法研究，也可以为研究工作提供一个清晰的实用背景。不仅可以注意到以往面向人的研究不容易注意到的一些问题，而且也使得在语法研究中的许多问题能够在一个形式系统的框架中得到更明确、更规范的表述。作者希望这本书对从事汉语信息处理实际应用开发工作的科研人员、在计算语言学这一交叉学科领域辛勤耕耘的研究人员，以及汉语语法研究工作者，都能起到一定的参考作用。

书中内容在得到许多专家学者的指导和宝贵意见后经过若干次调整修正，并经多次仔细校对，但错误疏漏之处，恐仍难免。在请读者包涵谅解的同时，也恳请专家同行多多批评指正。