

北京大学

博士学位论文

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

姓名: 詹卫东
学号: 19620822
系别: 中国语言文学系
专业: 现代汉语
研究方向: 计算语言学
导师: 陆俭明教授 俞士汶教授

一九九九年五月二十五日

摘 要

本文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,并探讨解决短语结构歧义问题的途径。全文共七章。

第一章对中文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,确定了本课题研究所针对的对象为短语结构,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。

第二章贯彻词组本位语法理论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,本文吸收了汉语配价理论、动词格框架等的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”,以及一个简化的语义分类体系。

第三章在上述句法语义属性范畴基础上,对四类主要的现代汉语短语结构:np、ap、vp、dj 的组合规则进行了系统而具体的形式化描写。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,加上作者本人的研究和实践,组织成了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一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四类短语的规则共 89 条。

第四章细致分析了计算机处理汉语短语结构时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。在已有短语结构规则的基础上,利用一个简单的分析程序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系统的认识基础上,通过对三个典型的短语歧义格式进行分析,进一步探讨了排歧策略,并对难以在短语结构规则层面解决的歧义分析问题,指出困难所在,为将来的排歧研究打下了基础。

第六章以计算机分析实例的结果展示了本文研究所得到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。

第七章对全文的研究工作进行了总结,包括具体的研究成果,对中文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。

本文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,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,而且也使得语法研究中的许多问题能够在形式系统的框架中得到更明确、更规范的表述。

关键词: 短语结构语法, 句法范畴, 语义范畴, 规则, 合一, 广义配价模式, 中文信息处理

Abstract

This thesis, which is oriented towards Chinese Information Processing, or 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 paper 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. Fortunately, such an electronic lexicon has been developed by the Institute of Computational Linguistics of Peking University. To some extent, the main research presented in this thesis can be regarded as a natural extend of the research of the lexicon.

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 of a word or a phrase. Furthermore, a simple semantic taxonomy of Chinese content words is also built up. Apparently, all of those syntactic and semantic categories set up in this chapter are the basis of constructing phrase structure rules of Modern Chinese that will be discussed in detail in the Chapter three.

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, which generally consist of syntactic constraints and semantic constraints. As a result, there are 89 rules induced in this chapter for most types of phrases in Modern Chinese, 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 true-ambiguous phrase, the quasi-ambiguous phrase, and the pseudo-ambiguous phrase. Depending on the above analysis and the set of rules proposed in the 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. Furthermore, I also explained the reasons for why some results are not good.

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

This research done in my thesis is devoted to two fields: grammar of Modern Chinese and CIP. On the one hand, the result of this research can be used directly, or as a significant reference for support 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 thesis.

Keywords: phrase structure grammar, syntactic category, semantic category, rule, unification, generalized valence mode, Chinese Information Processing

关于本文所用符号代码的说明

一 语素标记、词类标记、短语类标记、标点标记等

代码	含 义	代码	含 义	代码	含 义
a	形容词	b	区别词	c	连词
d	副词	f	方位词	g	语素
m	数词	n	名词	p	介词
q	量词	r	代词	s	处所词
t	时间词	u	助词	v	动词
w	标点	z	状态词	dj	主谓短语
ap	形容词性短语	dp	副词性短语	mp	数量短语
mcp	数词短语	np	名词性短语	pp	介词短语
sp	处所词性短语	tp	时间词性短语	vp	动词性短语

二 本文用到的词和短语功能范畴的符号标记

代码	含 义	代码	含 义
zhuyu	用于标记一个成分能否出现在主谓结构的主语位置	weiyu	用于标记一个成分能否出现在主谓结构的谓语位置
shuyu1	用于标记一个成分能否出现在述宾结构的述语位置	shuyu2	用于标记一个成分能否出现在述补结构的述语位置
binyu	用于标记一个成分能否出现在述宾结构的宾语位置	buyu	用于标记一个成分能否出现在述补结构的补语位置
dingyu	用于标记一个成分能否出现在定中结构的定语位置	zhxyu1	用于标记一个成分能否出现在定中结构的中心语位置
zhuangyu	用于标记一个成分能否出现在状中结构的状语位置	zhxyu2	用于标记一个成分能否出现在状中结构的中心语位置
zhxyu3	用于标记一个成分能否出现在附加结构的中心语位置	zhxyu4	用于标记一个成分能否出现在“的”字结构的中心语位置
lwqx	用于标记一个成分能否出现在连谓结构的前项位置	lwhx	用于标记一个成分能否出现在连谓结构的后项位置
lhqx	用于标记一个成分能否出现在联合结构的前项位置	lhhx	用于标记一个成分能否出现在联合结构的后项位置
ccat	用于标记一个词语的“词性”属性	cpcat	用于标记一个短语的“短语类”属性