汉语粘合式名词短语语义结构信息数据库

胡凤国　傅爱平

中国社会科学院语言研究所　北京　100732

E-mail: bushiwoshishui@yahoo.com.cn　fuap@linguistics.cass.net.cn

摘　要：本文选择汉语短语的一个小类——含有事件类词语的粘合式名词短语——作为切入点，在有限语料的范围内，相对穷尽地描写这一类 NP 的语义结构，着重考察语义关系，建立了一个小规模的语义结构信息数据库，并提供查询工具。希望通过相当数量的实例，验证汉语语法研究中某些定性分析的结果，或者发现新的语言现象和规律，并使其在自然语言信息处理中可以应用。

关键词：粘合式名词短语，语义关系，语义结构信息数据库

The Database of Semantic Structure Information of the Bound NP in Chinese

Hu Fengguo　Fu Aiping

Institute of Linguistics, Chinese Academy of Social Sciences, Beijing 100732

E-mail: bushiwoshishui@yahoo.com.cn　fuap@linguistics.cass.net.cn

Abstract: In this paper we have selected, in a limited amount of newspaper text, one type of the noun phrases in contemporary Chinese, which is the bound NP that contains words of events. And then we have described the structure of such noun phrases, especially their semantic structures. As a result of our survey, a small-scaled database of semantic structure information has been built together with its inquiry tool. The database can provide a quite large number of NP instances, so as to support the investigation of NP in Chinese Language and the development of Natural Language Processing.

Keywords: bound noun phrase, semantic relation, semantic structure information database

汉语中短语是连接词和句子的中间纽带，短语的识别和分析对于整个句子的理解十分重要。对计算机来说，短语处理更是句子处理的基础[1]。目前面向计算机处理的短语研究多从词类概念和句法功能概念[2]出发，较少涉及语义，这往往使得短语理解局限于某一层次而不能深

[1] 当前汉语的自动句法分析仍然是一个难点，现在已经倾向于将句法分析问题分解为短语处理问题。参见[许超，陈小荷 2002]。

[2] 参见 [詹卫东 2000 第 15 页]。

入。本文选择短语当中结构相对简单的一个小类——含有事件类词语的粘合式名词短语——作为研究的切入点。在有限语料的范围内，相对穷尽地描写包含事件类词语的 NP 内部的语义结构，着重考察语义关系，建立一个小规模的开放式语义结构信息数据库，并提供查询工具。希望通过相当数量的实例，验证汉语语法研究中某些定性分析的结果，或者发现新的语言现象和规律，并且为自然语言信息处理提供语言知识资源。

事件类词语统指表示动作、行为或事件的词语，在词类上主要表现为动词。我们把这一类 NP 称为“含有动词的粘合式 NP”，简记为 NP(v)。

1 语料、资源和 NP(v)的提取

1.1 NP(v)的范围

如果一个语法形式[3]能够同时满足以下三个条件，我们称它是 NP(v)：

1) 基本组成成分[4]不是介词、连词和助词“的”等虚词；
2) 直接组成成分[5]之一是事件类词语；
3) 外部功能是名词性的。

根据在考察语料范围内的初步统计，有多种符合上述条件的语法形式，仅基本组成成分数目小于等于三的，就有 15 种之多：

1) N+N+V	5) V+N	9) V+V+N	13) D+V+V
2) N+V	6) V+N+N	10) V+V+V	14) V+A+N
3) N+V+N	7) V+N+V	11) A+N+V	15) V+A+V
4) N+V+V	8) V+V	12) D+V+N	

其中后五种语法形式数量较少，实际上是前面 10 种的扩展，大都能化归为前面 10 种形式，我们对此暂不考虑。在前 10 种语法形式当中，8）和 10）的基本组成成分全由动词组成，比较特殊，也不作为研究对象，我们只考虑动词和名词全都具备的 NP(v)。实际上，经过上面的化归和特殊处理，NP(v)的基本组成成分已经全部是名词和动词了。而且，其基本组成成分和直接组成成分总是一致的，以后我们不加区分，称为组成成分。

1.2 语料和资源

我们从北京大学计算语言研究所的《人民日报》标注语料库(PFR)中提取 NP(v)。对 NP(v)进行语义结构信息标注所依据的语义知识体系是《知网》。从《知网》定义的概念、概念的属

[3] 关于“语法形式”的定义及细节问题，请参见[马真 陆俭明 1996]。
[4] 一个语法形式经某种分词系统切分后得到的一组词语，称为它的基本组成成分。
[5] 一个语法形式经过层次分析可以得到一系列更低一级的语法形式，其中直接组成该语法形式的那一组低级语法形式称为它的直接组成成分。

性及其相互关系出发，我们可以描写 NP(v)内部的语义结构，讨论其组成成分之间的组合规律，并使某些语义关系的识别具有可操作性。

1.3 NP(v)的提取

根据 NP(v)的定义和 PRF 的标注规定，我们从大约 2000 万字的 1998 年 1—5 月份的 PFR 当中提取所有标注为“/vn”和“/n”的连续词语串，共 8 类 50881 条，作为候选词语串初步取出，然后经过一系列的筛选确定 NP(v)。筛选时考虑的因素有如下几点：词语串的出现频率，是否语法形式，PFR 与《知网》在词目、词性等方面的互适性，直接组成成分是否含事件类词语，等等。我们用人工方式对这些候选词语串进行逐条审查。

经过人工筛选余留下 NP(v)当中，还有一部分需要简化，例如“党性党风教育”，这是“N+N+V”结构的 NP(v)，我们把它简化成“党性教育”和“党风教育”这两个“N+V”结构。经过简化之后，8 类 NP(v)共得到 5855 条。接下来要做的事情，就是考察这些 NP(v)内部的语义结构信息。

2. NP(v)的语义结构信息

2.1 NP(v)的语义结构

NP(v)的语义结构涉及其组成成分的语义类别序列和语义关系。前者取决于 NP(v)各组成成分的语义类别，后者是组成成分之间的组合关系。我们先说明这两个概念，然后，在此基础上讨论 NP(v)当中各个组成成分之间可能的语义关系类型。

语义类别序列：我们把《知网》对词语从概念角度所作的描述当作该词语的语义类别。例如“盐”在《知网》知识词典中描述为“DEF=material|材料,?food|食品,salty|咸”，“盐”的语义类别就是“材料,?|食品,|咸”。NP(v)的语义类别序列就是其组成成分的语义类别组成的一个有序序列。

语义关系：NP(v)的语义关系指其组成成分之间的意义组合关系。语义关系可以存在于动词和名词、动词和动词、名词和名词之间，我们着重考察第一种，第三种虽然存在但极少，暂不考察。

我们主要用论元关系描述动词和名词之间的语义关系，论元角色的设立参考了[袁毓林 2002]的论元角色[6]层级体系和《知网》的动态角色分类，结合 NP(v)的特点增删了一些论元角色。我们注意到，在 NP(v)当中，并非所有的动词和名词都构成论元关系，也有不是论元关系的情况。例如：整风运动、爱国肉、旅游部门等，因此也设立了若干非论元关系。

NP(v)中动词和动词之间的语义关系是一种事件与事件的关系。在 NP(v)中，这种关系相

[6] 关于“论元”、“论旨角色”、“论元角色”等概念，请参考[袁毓林 2002]。

对比较简单[7]，我们所遇到的事件关系有5种。

通过一定范围语料的试验，我们建立了一个分层次的NP(v)语义关系类别表（表1）。该表以动词和名词之间的语义关系为主，在动词和名词的语义关系当中，又以论元关系为主。

表1　NP(v)的语义关系类别表

			论元层级分类		
NP(v)中的语义关系	动词和名词的语义关系	论元关系	核心论元	主体论元	施事
					合作施事
					经验者
					关系主体
				客体论元	受事
					目标
					结果
					内容
					占有物
					关系客体
			外围论元	凭借论元	工具
					根据
					方式
				环境论元	处所
					源处所
					终处所
					时间
					原因
					目的
				属性论元	数量
					程度
					状态
		非论元关系			类属
					限定
	动词和动词的语义关系	事件关系	主从事件		结果事件
					内容事件
			补充事件		方式事件
					范围事件
					目的事件

[7] 按照我们的考察约定，含有并列事件的NP（v），已经被分解合并到其他类型的NP（v）里面，这就使得动词跟动词的语义关系大为简化。

2.2 NP(v)语义结构的判定

给定一个 NP(v)，首先要确定它的组成成分，然后确定组成成分的词类、语义类别，形成语义序列，再在语义序列的基础上判定语义关系。

为了保证判定的一致性，我们根据《人民日报》标注语料库的切分结果确定 NP(v)的组成成分及其词类。如果其中有值得商榷的，就暂时存疑，留待以后研究处理。

我们依据《知网》给 NP(v)的每个组成成分确定适当的语义类别。一个词语可能有多个义项，需要人工比较和选择。

判定语义关系时有两个原则：唯一性和顺序性。按照前者，在 NP(v)的某两个组成成分之间，如果有语义关系存在，就按照语义关系类别表的分类及相应的标准确定一个类别。后者是为保证人工判定语义关系的一致性而制定的操作优先顺序，这种优先顺序带有一定主观性，但只要能贯彻始终，就能够保证相似的词语和类似的结构得到一致的语义关系类别。

根据语义关系类别表（表 1）和判定原则，我们制定了语义关系判定流程。以类别表中的先后顺序作为操作优先顺序。

2.3 NP(v)语义结构信息的标注

我们逐条标注提取出来的 NP(v)，标注的项目有：音步、词类序列、语义序列和语义关系。有歧义的 NP(v)有多于一种的语义结构信息。这分为两种情况：一是词语有多个义项，如例 1 所示；另一种情况是确实存在两种可能的语义结构信息，如例 2 所示：

【例 1】	分配原则	2+2	V+N	<差遣>+<规矩>	(<1> → [根据]<2>)
	分配原则	2+2	V+N	<分发>+<规矩>	(<1> → [根据]<2>)
【例 2】	微机管理	2+2	N+V	<电脑>+<管理>	([工具]<1> ← <2>)
	微机管理	2+2	N+V	<电脑>+<管理>	([受事]<1> ← <2>)

对这样的 NP(v)，每一种语义结构信息都要标注一条记录。因此最终得到的标注记录数目比 NP(v)的数目略有增加：8 种类型共 5939 条标注记录。

3. NP(v)语义结构信息数据库

我们把这 5939 条标注记录做成了一个语义结构信息数据库，用 ACCESS 来存储。这个数据库的数据表共含有 20 个字段，表 2 给出了各个字段的说明。有必要指出的是，由于这些 NP(v)的组成成分个数不尽相同，我们在设计数据库结构的时候作了一些特殊处理，使得所有的标注记录能用统一的格式存放。我们假设所有的 NP(v)都有三个组成成分，如果某个 NP(v)只有两个组成成分，那么，跟第三个组成成分有关的字段赋值为零或者为空字符串。表中还专门设了一个字段“COUNT”标示 NP(v)组成成分的个数，因此不会造成混淆。

表 2　数据库的字段说明

字段名称	数据类型	功能介绍
WORDGROUP	文本	NP(v)实例
COUNT	数字	组成成分个数
F_C、S_C、T_C[8]	数字	音步
F_W、S_W、T_W	文本	组成成分
F_T、S_T、T_T	文本	组成成分的词类
F_M、S_M、T_M	文本	组成成分的语义类别
R12、R13、R21、R23、R31、R32	文本	各个组成成分之间的语义关系 Rij 表示从第 i 个组成成分指向第 j 个组成成分的语义关系（1≤i，j≤3）

数据库配有查询工具，根据用户输入的查询条件搜索记录。检索条件有以下几类，可以单独使用，也可以组合使用：

a）NP(v)含有/等于某个词语串；

b）NP(v)的组成成分个数；

c）NP(v)的某个组成成分含有/等于某个词语串；

d）NP(v)的某个组成成分含有的汉字个数；

e）NP(v)的某个组成成分的词类；

f）NP(v)的某个组成成分的语义类别；

g）NP(v)含有某个语义关系类别；

4．NP(v)语义结构信息数据库的初步应用

这个数据库收录的 NP(v)实例反映了汉语的实际使用情况，虽然目前规模还不够大，也已经可以初步用来验证汉语语法研究中某些定性分析的结果，发现新的语言现象，或者在自然语言处理系统中对名词短语的自动理解有所帮助。

比如，“N+V”这种 NP(v)，是一种表示事件的名词短语。[马真、陆俭明 1996]认为，其中 N 必须是 V 的配价成分，通常是 V 的受事，这一点在我们的数据库中得到了证实。与此同时我们也发现：N 还可以是 V 的主体论元，或者说是施事性成分。例如：大学生就业、职工互助、银行倒闭等等。这种结构数目不少，达到 292 条，占数据库中全部标注记录的 4.9%。

再如[李晋霞 2002]认为：泛指名词构成定中式“$V_{双}$+$N_{双}$”的能力强于非泛指名词，并举了释义基本一致[9]的泛指名词“群众”和非泛指名词“人民”为例子来支持其结论。这里说的

[8] “F”是“First”的缩写，在字段名当中表示第一个组成成分；“S”是“Second”的缩写，表示第二个组成成分；“T”是“Third”的缩写，表示第三个组成成分。下同。

[9] 该文的依据是《现代汉语词典》给出的释义。

定中式“$V_{双}$+$N_{双}$”，在我们的数据库中是音步为“2+2”、词类序列为“V+N”的 NP(v)。这类 NP(v)，与“群众”有关的有 7 条，跟.“人民”有关的有 1 条，这可以为李文的定性分析提供实例验证。

数据库收录的语义结构信息还可以对智能信息检索提供帮助。目前的信息检索系统多采用基于关键词机械匹配的检索方式，效果还不能令人满意。相同意思的关键词，检索出来的内容常常大相径庭。比如用“汽车修理”和“修理汽车”作关键词，分别在新浪、中文雅虎、263 在线等网站上面按照“网页”或者“中文网页”进行搜索，得到的结果在数量上的差异都是很明显的（例如：“汽车修理”：56100 条；“修理汽车”：3640 条[10]），而且检索结果大多互不相交，除非文档中同时含有这两个关键词语。

关键词通常是名词性短语，其中相当一部分是我们考察的 NP(v)。如果搜索引擎能确定关键词的组成成分的语义类别，正确分析他们之间的语义关系，就可以知道“汽车修理”和“修理汽车”是同义的关键词，通过义类关联还有可能把“车辆修理”、“汽车维修”、“车辆维修”、“修理车辆”、“维修汽车”、“维修车辆”也作为检索目标。语义结构信息数据库可以为这些分析提供语言知识资源。

本文所建立的语义结构信息数据库只收录了近六千条 NP(v)的实例，有待于进一步扩大规模。要想客观地反映这一类短语的结构规律，使数据库达到实用的程度，尚有许多工作要做。除了增加短语类型和数量以外，还应该进一步研究短语语义结构信息的描写手段，并且在《知网》等语义知识资源的支持下分析数据库提供的短语语义结构信息。

参考文献

[1] Fillmore, C.（1968）, The Case for Case, In E. Bach and E. Harms, eds. , Universals in linguistic Theory, Holt, Reinhart and Winston, New York..

[2] 董振东，董强：“知网”，http://www.keenage.com ，1999 年。

[3] 董振东，董强：“关于知网-中文信息结构库”，http://www.keenage.com ，2000 年。

[4] 李晋霞：“现代汉语名词性结构‘$V_{双}$+$N_{双}$’结构研究”，中国社会科学院研究生院博士论文，2002 年。

[5] 马真，陆俭明：“‘名词+动词’词语串浅析”，《中国语文》，1996 年第 3 期。

[6] 许超，陈小荷：“试评两种商用机译软件的汉语分析能力”，《机器翻译研究进展》，电子工业出版社，2002 年。

[7] 尹世超：“动词直接作定语与名词中心语的类”，《语文研究》，2002 年第 2 期。

[8] 俞士汶等：“北京大学现代汉语语料库基本加工规范”，《中文信息学报》，2002 年第 5、6 期。

[9] 袁毓林：“论元角色的层级关系和语义特征”，《世界汉语教学》，2002 年第 3 期。

[10] 詹卫东：《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，2000 年。

[10] 2003 年 4 月 28 日的搜索结果。