第三章 现代汉语 np、ap、vp、dj 的句法语义规则

§3.1 短语规则的形式表达

本文讨论的短语规则每一条都涉及两部分内容[1]。一部分是短语的内部构成情况；另一部分是对这条短语规则所做的详细说明。请看一个简单的例子。

“一件衣服”是一个具体的名词短语。如果用自然语言来描述的话，这个具体的名词短语对应的抽象规则至少包括这样两部分：（1）汉语中一个名词性成分（np）可以由数量成分（mp）加上名词性成分（np）组合而成。（2）要完成这个组合，需要满足的条件至少包括：其中的 np 可以受量词修饰；并且 mp 中的量词要跟 np 对量词的选择要求一致。

上述规则的作用是明显的。它可以接受“一件衣服”这样合法的汉语 np，而排斥“* 一件书”、“* 两个眼光”这样非法的 np。当计算机碰到 a“一件书上没有提到的衣服”、b“两个眼光很不错的男人”这样的形式时，这条规则能够帮助计算机做出正确的判断。对于 a，“一件”是修饰“衣服”的，而不是就近修饰名词“书”的；对于 b，“两个”是修饰“男人”的，而不是就近修饰名词“眼光”的。因为根据词典中已经标明的语言知识，“件”能跟“衣服”搭配，不能跟“书”搭配，“眼光”根本就不能跟个体量词搭配，所以对于 a 和 b，量词都是修饰距离远的名词，而不是修饰与它距离最近的名词。

这个简单的规则例子是用自然语言进行表述的。人容易理解，但计算机却很难看懂。要让计算机掌握这样一条规则。最好是将规则以一定的形式语言来进行表述。

自从乔姆斯基提出形式语言理论以来，已经发展出不少用来表述自然语言语言知识的形式模型。对这些形式模型，本文不作讨论，可参见翁富良等（1998）[2]。本文短语规则所用的形式化表达方法主要有两部分内容，分别对应着规则的上述两个组成部分。一部分是上下文无关文法**产生式（Rewrite Rule）**，用来描述短语的内部组成模式；另一部分是**合一等式（Unification Equation）**，用来对一个短语进行详细的说明。对本文规则的形式语言的详细定义可参见附录一。这里先对规则的一般情况作一些扼要的说明。

上下文无关文法产生式，又叫作转写规则，它的一般形式[3]为： A -> α

其中 A 是非终结符（Non-terminal Symbol），比如上文提到的短语功能类标记 np、vp、ap、dj、sp、……等等。α 可以是非终结符、终结符（Terminal Symbol）、或由二者组成的字符串，比如“vp”、“v”、“学习”、“mp np”、“v np”、“vp u<的>”、“。”……等等。

还是拿上面的例子来说明，它用到的产生式规则是：np->mp np。从组成的角度讲，这个规则描述了 np 可以由 mp 加 np 组成；从转写的角度讲，这个规则是说，一个 np 可以转写成（生成）mp 加上 np 的形式。用这样的有限的转写规则去描述无限的自然语言句子，是乔姆斯基创立转换生成语法的最高目标。尽管这个理论目标是否合适，以及是否能够达到，目前尚无定论，但从方法论的角度讲，上下文无关文法规则作为一种刻画自然语言句法结构的表达手段，至今仍然是形式语法以及自然语言处理实践中的主要模式，很多理论研究和实际工作都是在这个基础上做一些补充或者调整，参见方立（1993）[4]，姚天顺等（1995）[5]。这里需要说明的是，我们对产生式规则右部非终结符和终结符的个数没有限制，即一棵句法树可以是二叉的，也可以是多叉的，根据实际需要决定。而且右部终结符还可以有标点符号，即包含形如“np->np w<″、″> np”这样的规则（这是为了描述像“电影、音乐、文学”这样的联合式 np）。限于篇幅，本文对包含标点符号的短语组合暂不加讨论[6]。

“合一”是基于复杂特征集（Complex Feature Set）的运算。有关复杂特征集与合一运算的形式定义，可以参看冯志伟（1995）[7]、沙新时等（1993）[8]。这里仍以简单的例子来

做一些概要说明。

所谓复杂特征集，是指一些“属性名：属性值”对的集合。比如“zhuyu：是”就是一个“属性名：属性值”对，表示某个语言成分可以作主语（如“眼光很不错”）。显然，我们在第二章中确立的关于短语的句法和语义知识的范畴都可以作为“属性名”来理解。值得指出的是，属性的取值，除了简单的“是”或“否”这样的情况，还可以是复杂的形式。比如动词的语义知识范畴中有“主体”、“客体”等属性。这些属性的取值就不是“是”或“否”这样的简单情况，而是用另一个“属性名：属性值”对来作为它们的取值。下面给出一个词的复杂特征集的全部内容作为示例。

高兴 a $=[形容词子类:ab,谓词性主语:可,准宾语:可,形定语:的,形谓语:可,形补语:组,带补:粘|得,形趋:可,形状语:地,准谓宾:否,有的宾语:否,不:可,很:可,前名:否,重叠词性:z,语义类:境况,配价数:1]{主体:[语义类:人]}

“高兴”词条后面的“a”表示“高兴”的词性是形容词。从“$=”开始就是“高兴”这个词的复杂特征集。有关句法属性范畴“形容词子类”、“谓词性主语”、“准宾语”等等的说明可参见俞士汶等（1998）[9]。方括号“[……]”里面包括了一系列“属性名：属性值”对，互相之间以逗号隔开，都是简单特征。花括号“{ …… }”里面“主体”属性的取值是一个“属性名：属性值”对 ——“[语义类：人]”，表示“高兴”这个形容词的“主体”配价成分的“语义类”要求是“人”。上述复杂特征集内的元素是顺序无关的。对所有的词语，都可以用这种复杂特征集的表达形式来描述其句法语义属性。

再来看合一等式。合一是用来对短语规则进行详细说明的。这包括两部分[10]：一是着眼于一个短语的整体性质，对其整体结构情况以及内部组成成分进行定性描述，同时说明它作为一个整体向外组合的功能性质。二是向内看一个短语，给出这个短语的内部组成成分需要满足的限制条件。下面是用于分析 NP“一件衣服”的规则（部分）。

np->mp !np :: $.内部结构=定中,$.定语=%mp,$.中心语=%np,$.dingyu=否,…,
%np.数量名=是,IF %mp.量词子类=个体 THEN %np.个体量词=%mp.原形 ENDIF,…

其中 np 前的“!”号表示它所标记的 np 是这个短语的中心词（head）[11]。“::”是分隔符，后面开始是合一等式。“$”表示产生式的左部根节点 np。“.”号可以理解为汉语中的助词“的”。“%”表示一个短语在结构中的顺序[12]（比如“%mp”是规则右部第一个 mp，下文我们也称之为“前项 mp”）。“…”不是规则中用到的形式语言符号，表示省略（下同）。

从“$.内部结构=定中”开始是对这条规则做整体说明。这可以直观地理解为赋值操作，即这个 np 的“内部结构”属性（句法知识范畴）被赋值为“定中”。“$.定语=%mp”，则标示这个 np 的“定语”是其组成成分中第一个“mp”。“$.中心语=%np”表示这个 np 的“中心语”是规则右部第一个 np。“$.dingyu=否”表示这个 np 不能作定中结构的定语[13]。在上文§2.3.2 中我们已经说明过，像“dingyu”这类短语句法功能范畴，默认值都为“是”，因此当具体到某个短语其“dingyu”属性取值为“否”时，应在规则中显性标明。但也并不是对每一个取值为“否”的属性，都要说明，比如上面这条规则中，就没有说明“$.buyu=否”，尽管这个 np 的确不能作补语。下文讨论具体规则时，对一个短语取值为“否”的那些范畴，一般是有针对性地选择一些来加以说明，而不是逐一罗列的。

从“%np.数量名=是”开始是向内看一个短语的组成成分，给出约束条件[14]。这个合一等式要求右部第一个中心语 np 必须是那些能在前面加数量成分的名词[15]。我们把这类约束称为“**绝对条件**”，即一个语言成分不需要参照它的搭配成分的情况，自身必须满足的条件。与这种“绝对条件”不同的是“**相对条件**”。比如“IF %mp.量词子类=个体量词 THEN %np.个体量词=%mp.原形 ENDIF”表达的就是相对条件（因为要参照组成成分 np 和 mp 双方的情况进行判断，所以称之为相对条件）。其中“IF…THEN”之间是测试条件，如果满足测试条件，就进行“THEN”后的合一判断。整个表达式的具体含义是：如果 mp 的“量词子类”属性取值是“个体量词”，

那么 mp 中量词的形式必须跟 np 的“个体量词”[16]属性取值吻合。如果不满足测试条件，当然就不进行后面的合一判断。

下面再看跟上述 np 规则相关的词的例子：“件”跟“衣服”的复杂特征集表示。

件 q $=[量词子类:个体,表数:数]

衣服 n $=[名词子类:na,数量名:是,个体量词:件|套|身,前名:否,前动:否,后名:是,名状语:否,临时量词:否,语义类:服饰]

显然，“衣服”的“数量名”属性取值为“是”，符合绝对条件的要求。而“件”的“量词子类”属性取值是“个体”，满足条件测试要求，于是进行合一判断。“衣服”的“个体量词”属性取值为“件|套|身”（其中“|”表示逻辑“或”的关系）。这样词典中静态的值“件”跟输入短语中的“件”词形吻合，满足规则的要求，“一件衣服”被系统接受。以同样的方式，系统也可以判断“*一件书”、“*两个眼光”不能被接受。此外，碰到“一次会议”这样的np，由于“一次”中的“次”不是“个体量词”，不满足规则中条件测试的要求，因此也就不需要去符合这条规则的合一约束“%np.个体量词=%mp.原形”的要求。这样，“一次会议”也能被这条规则接受[17]。

以上简要介绍了本文所用短语规则的一般模式及其形式表示方式。下面章节有关具体的短语规则的讨论都将以此为基础展开。这里有两点需要特别补充说明。

一是像“mp np”这样的规则，本文称为**全局规则**（global rule），在全局规则库中描述（一条全局规则通常不是针对具体的特定词的，而是针对语类之间的组合的）。此外还有描述“乐观主义者，同学们，红了，所见，大师所说，……”这样一些短语实例的规则模式。这些短语形式紧凑，同时内部包含一些“特别”的成分，如“者、们、了、所”等等。从处理方便的角度考虑，本文把描述这些短语的规则放在一个专门的库中作为**局部规则**（local rule）处理（通常一条局部规则跟一个或多个特定的词语发生关联）。下面看几个局部规则的例子[18]，这些规则分别跟词典中“者”、“所”、“了”等条目发生关联。

np->np !g<者> :: $.内部结构=单词,…,$.语义类=人,$.dingyu=否,…

np->np g<所> !vp :: $.内部结构=所字,…,%vp.配价数=2|3,%vp.主体=%np,$.dingyu=否,…

ap->!ap u<了> ::$.内部结构=附加,$.中心语=%ap,$.附加语=%u,$.zhxyu3=否,%ap.zhxyu3=是,…

区分全局规则和局部规则的做法对更好地描写汉语的短语组合是有利的，因为很明显，跟全局规则相比，局部规则的针对性强，可以把一个短语的整体性质及其组合成分的条件刻画得尽可能详细准确。在实际的中文信息处理系统中，这样的局部规则甚至可以说是多多益善。从技术角度讲，所谓全局规则和局部规则，本质上是规则的使用顺序问题，即局部规则优先于全局规则。直观而言，当系统分析一个输入时，如果用局部规则能够分析出正确结果，就不必调用全局规则。

二是在上一章归纳的 10 类短语中，除 dj 外，其他短语都可以由词直接实现得到[19]，即存在“np->!n”，“np->!r”[20]、“vp->!v”，“sp->!s”，“sp->!r”[21]，dp->!d,……”这样的直接上升**实现式规则**。这些规则的语言学意义在 2.1.2 中说明短语的句法功能分类框架时已经提到了。而像“vp->!vp np”这样的规则，本文称之为**组合式规则**。从表达的简洁性角度讲，引入实现式规则，可以使整个规则体系的组织更加紧凑。比如“vp->!vp np”就能涵盖“vp->!vp n”、“vp->!v n”等规则的组合情况。而由单词直接实现得到的短语，可以靠“内部结构”属性来标示它跟其他同一功能类短语的区别。比如“np->!n”这条规则，其“内部结构”属性将被赋值为“单词”。

本文从把握现代汉语的短语规则的全貌着眼，下面的讨论主要集中在全局规则和组合式短语规则上。对每条具体的短语规则，文中都列出了我们目前已经归纳出来的关于该短语的整体性质说明及该短语对内部组成成分的条件约束。介绍的总体原则是有话则多，无话则短（有的组合模式没有展开讨论，一般就放在各节的“概述”部分作了一些必要的交代），尽

可能阐明我们已有的关于汉语具体短语结构的语言知识，同时也摆出难以归纳条件的情况，期待引发更多的研究。

附 注：

[1] 短语规则的表达模式可以有不同的选择。本文采用产生式加合一约束的方式。对其他可能的方式不加以讨论。事实上，最为简单的情况是，仅有产生式就可以构成规则了，即典型的上下文无关文法规则。

[2] 参见翁富良、王野翊（1998）《计算语言学导论》，中国社会科学出版社 1998 年版。第三、四章。方立（1993）《美国理论语言学研究》，北京语言学院出版社 1993 年版，P26-76。

[3] 参见翁富良、王野翊（1998），P36。姚天顺等（1995）《自然语言理解 —— 一种让机器懂得人类语言的研究》，清华大学出版社、广西科学技术出版社 1995 年版，P18。

[4] 方立（1993）《美国理论语言学研究》，北京语言学院出版社 1993 年版。

[5] 姚天顺等（1995）《自然语言理解 —— 一种让机器懂得人类语言的研究》，清华大学出版社，广西科学技术出版社 1995 年版。

[6] 在实际分析中，内部含标点的短语组合情况比较多样，比如“勤劳、勇敢、善良的中国人民”是联合式 AP 中包含顿号。而“大家把为大桥出的一点力，流的一点汗，当作光荣和幸福”中，是联合式 NP 以及状中式 VP 中包含逗号。本文不讨论各种标点的使用对短语规则的影响。

[7] 冯志伟（1995）《自然语言机器翻译新论》，语文出版社 1995 年版。P122-142。

[8] 沙新时等（1993）《基于合一语法的通用句法分析器：设计与实施》，载《中文信息学报》1993 年第 2 期。

[9] 俞士汶等（1998）《现代汉语语法信息词典详解》，清华大学出版社 1998 年版。

[10] 实际上用到的合一表达式还有别的情况，涉及到技术上的一些处理细节，跟本文讨论的内容关系不大，因此没有全部谈到。可以参见刘群（1994）《汉英机译系统数据说明》（“中科院计算所与北京大学计算语言学研究所汉英机器翻译课题组”内部工作手册）。关于对一个短语可以向外看，同时也要向内看的思想，受到朱德熙（1985）《语法答问》，商务印书馆 1985 年版，P43 内容的启发。

[11] 中心词是个技术概念。其作用在于属性传递：规则左部根节点的属性，默认情况下是从中心词节点的属性继承得到的。比如在“np->mp !np”这条规则中。如果不特别说明，左部根节点 np 的属性就从右部中心词 np 的属性继承。比如一个词的词性属性（用“ccat”代码表示）就是继承的。“脸”是名词（ccat=n），它作为中心语参与形成的 mp“一脸”的“ccat”属性仍然是“n”。而不是“q”，这样就可以跟“一张”区别开，即同样是 mp，“一脸”的“ccat”属性为“n”，而“一张”的“ccat”属性为“q”。它们的“ccat”属性都是从中心词那里继承的（相关规则是：mp->m !q 和 mp->m !n）。有关中心词与属性继承，还有不少技术细节问题，比如哪些属性继承，哪些属性不继承等等，都要在语言模型中进行定义。本文对此不进行详细讨论，可以参见刘群（1994），出处见上一条附注。一般情况下，词的大部分句法语义属性都向上传递（除配价信息外），短语的句法语义属性都不继承。此外还有一点需要注意的是，规则里的中心词跟语法上讲的中心成分（比如定中结构的中心语成分）并不完全对应，尽管在大多数情况下是对应的。比如上面这条 np 规则，中心词就是这个定中式 np 的中心语。但也有这样的情况，比如联合式 np 的规则“np->!np np”，从语法上讲，两个 np 联合形成一个大的 np，无所谓哪一个是中心成分，但出于属性传递的目的，可以“硬性”规定第一个 np 作为整个联合式 np 的中心词。

[12] 这个规则只有一个“%”的情况，体会不到顺序问题。碰到“np->np !np”这样的规则，就需要两个“%”来区分规则右部第二个 np（%%np）跟第一个 np（%np）了。

[13] 值得强调指出的是，类似“$.dingyu=否”这样的赋值合一等式，是本文在说明一个短语的整体功能性质时最主要的手段。当我们说一个短语某项功能属性的取值为“否”时，并不是指它绝对地丧失了该功能。对此要作相对理解。即当这个短语参与形成更大的组合时，如果在后续分析规则中出现了“dingyu=是”这样的约束条件，该短语的这项功能属性值才有意义，否则这个属性值是“是”还是“否”，并没有什么实际的效用。可以看一个简单的例子。比如“找他”是个述宾式 vp，汉语中一般述宾式 vp 不能再带宾语，即这样的 vp 不能出现在“述语1”位置上（“$.shuyu1=否”）。但是，语言事实中，有“找他半天”这样的短语，我们把其中的“半天”分析为“找他”的宾语，这跟“找他”的“shuyu1”属性取值为“否”不是矛盾了吗？对此，正确的理解是，虽然“找他”通常不能再带宾语了，但也不排除它带一些“特殊”宾语的可能性。本文的处理策略正是，把“找他”的“shuyu1”属性取值为“否”，同时，又在“vp->!vp mp”这条述宾式 vp 规则中放宽限制，当 mp 是“半天”这样的时量成分时，不要求前面的 vp 的“shuyu1”属性值为“是”详细分析可参见下文 3.4.3 有关规则的讨论。

[14] 约束条件是分层级对一个短语组合进行限制的。可以从句法范畴考虑给出约束条件，也可以考虑从语义范畴考虑给出约束条件，条件也可根据实际情况调整松紧度。关于分层级对组合规则进行约束，还可参见 Johannes Heinecke and Juregen Kunze.1998. Eliminative Parsing with Graded Constraints,In Coling'98 P526-530.

[15] 关于名词的“数量名”属性，参见俞士汶等（1998）《现代汉语语法信息词典详解》，P67。

[16] 关于“量词子类”及“个体量词”等属性，参见《现代汉语语法信息词典详解》，P67，P72。

[17] 当然，这条规则也会把“一次桌子”当作“合法”的形式接受的。这说明这条规则的条件约束还不够完善。必须坦率承认的是，本文中的许多规则都还不够完善。

[18] 从语法上讲，“乐观主义者”可以视为一个单词，一般不能直接作定语，因此“dingyu”属性值为“否”。“np g<所>!vp”组合的内部结构属性值是“所字”，以跟上面提到的其他结构相区别。参见陆俭明（1989）《关于“他所写的文章”的切分》（载《陆俭明自选集》，河南教育出版社 1993 年版，P220-230）。对这个短语组合的内部成分，我们限制其中 vp 的配价数为 2 或 3，并且要求 np 应满足 vp 的主体的选择限制要求(%vp.主体=%np)。此外这种结构是黏着的，不能直接作定语，需要加“的”后才能作定语，因此“dingyu”属性取值也为“否”。而像“红了”这样的附加式 ap，类似的情况还有附加式 vp（如“吃了，看过了”），对此本文都采用局部规则的方式进行描述。下文分析时还将提及。

[19] 介词短语 pp 一般得由一个介词加上宾语构成述宾结构后才能参与组合，但汉语中表被动义的介词“被”、“给”等，也可以直接以 pp 身份，出现在状语位置。如“杯子被（给）打破了”。所以我们也允许这样的介词直接上升为 pp，同时在规则中特别说明只能是“被、给”等介词。

[20] 根据朱德熙先生在《语法讲义》里对代词的分类，像“我、你、这、谁、哪儿……”等一大部分人称、指示和疑问代词，是体词性的，从功能上讲，属于 np。因此这些代词可以直接实现为 np。其他的一些谓词性的代词（如“怎么，这么，这样……”等），不能直接上升为 np。对此，规则 np->!r 的条件判断部分都可以做出详细说明，这里从略。

[21] 代词中有“这里，那里”等指代处所的。它们可以直接上升为 sp，即有规则“**sp->!r :: …,%r.指代处所=是**”。“指代处所”是《现代汉语语法信息词典》中为代词设置的一个属性，用以标记一个代词是否能指代处所性成分。如“我、你”等“指代处所”属性值就是“否”。