基于合一的汉语短语结构规则*

Unification-based Chinese Phrase Structure Rules

詹卫东

北京大学中文系

摘要：本文介绍基于合一运算的汉语短语结构文法，分为三部分，五个小节。前三小节是第一部分，概要介绍基于合一的汉语短语结构规则这种形式化方法的整体情况，其中第一节对产生式规则与合一运算相结合的短语结构知识形式化表达模式进行了概括说明；第二节讨论汉语中产生式规则的不同类型；第三节讨论合一等式的不同性质；第四节，也即本文第二部分，以汉语中的典型结构“把”字结构为例，阐明了本文介绍的这种形式化方法的具体工作模式；最后一部分，即第五节是余论，扼要总结了以形式化模式来表述汉语短语结构规则的一些基本特点，同时对形式化表达方式的不足也做了一些概括说明。

关键词： 上下文无关文法　产生式规则　合一　　短语结构规则　复杂特征

§1 短语规则的形式化表达

语法研究的主要任务大致上就是要回答这样一个问题：语言中的小单位是如何组成大单位的？对一个具体的语言系统 L 而言，这个问题实际上可以分解为两个更具体一些的问题：

（1）L 中存在多少个组合模式 P_i（i = 1,2,3,...n）？

（2）对任意一个 P_i，其中的组成成分需要满足什么样的条件？

自 N.Chomsky 于 1957 年出版《句法结构》以来，一批批语言学家和计算语言学家在所谓形式主义的研究框架下所做的各种探索工作（参见文献[14]），差不多都可以看作是从不同的角度，尝试用不同的形式化方法，为回答上述问题所做的努力。尽管具体的策略有所不同，但目标是共同的，那就是力图以精确，清晰，并且更为经济的方式来寻求问题的答案。

本文介绍在此类研究工作的发展中，特别是在计算语言学的具体实践中，逐渐成为主流方法的一种语言知识形式化模式，即基于上下文无关文法（Context Free Grammar）的产生式规则（rewrite rule）与基于复杂特征（Complex Feature）的合一运算（Unification）相结合的形式化表达模式（参见文献[1，2，3]。其中上下文无关文法产生式用来给出上面问题（1）的答案；合一运算用来给出上面问题（2）的答案。下面我们就以一个简单的例子来对这种形式化模式扼要加以说明。

“一件衣服”是汉语中一个具体的表达式。如果用自然语言来描述这个具体的表达式对应的抽象规则，至少应包括这样三方面的内容：（1）作为一个整体，这个表达式是一个名词性范畴的语言成分；（2）这个表达式是由属于数量范畴的语言成分加上属于名词性范畴的语言成分组成的；（3）要实现这样一个组合，其中的名词性成分应该是能受数量成分修饰的那类名词，而且名词跟量词之间还需要满足一定的搭配关系。

上述规则的作用是很明显的。它可以解释“一件衣服”在汉语中为“合法的”表达式，而排斥“* 一件纸”、“* 两个眼光”这样“非法的”表达式。如果计算机掌握了这样的规则，

* 本文研究中所用“人民日报”语料及相关统计结果由北京大学计算语言所段慧明工程师提供，特此致谢。

当碰到 a“一件纸做的衣服”，b“两个眼光很不错的男人”这样的表达式时，就能够做出正确的判断：对于 a，“一件”是修饰“衣服”的，而不是就近修饰同样也属名词范畴的“纸”；对于 b，“两个”是修饰“男人”的，而不是就近修饰也属名词范畴的“眼光”。关于“件”能跟“衣服”搭配，不能跟“纸”搭配，“眼光”根本就不能跟个体量词搭配（不单单是不能跟“个”搭配），等等诸如此类的语言知识，是可以事先在一个人们称之为“词典”的地方一一加以记录的。也就是说，规则加上词典中的语言知识，可以构成计算机进行推理判断的已知条件（基础）。

在这个简单的例子中，规则以及跟规则配合使用的词典知识，都是用自然语言进行表述的。人容易理解，但计算机却不容易看懂。要让计算机掌握这样的规则和相关的词典知识，最好是将规则和词典知识以一定的形式语言来进行表述。下面就是一种形式化的方式（也就是把上面用自然语言表述的规则和词典知识“翻译”成用某种形式语言来表达）：

```
np -> mp !np :: $.内部结构=定中,$.定语=%mp,$.中心语=%np,$.dingyu=否,…,      ①
                %np.数量名=是,…                                              ②
                IF %mp.量词子类=个体 THEN %np.个体量词=%mp.原形 ENDIF,…      ③

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,前名:否,前动:否,后名:是,名状语:否,
          临时量词:否,语义类:服饰]
```

在“np -> mp !np”这条产生式规则中，箭头左部的 np 代表名词性范畴（比如“一件衣服”），右部的“mp !np”表示左部 np 范畴是由一个数量范畴（mp）加上一个名词性范畴（np）组成的。这里用“!”符号标记一个范畴在一个组合式中是中心成分（head）[1]。不难看出，基于上下文无关文法的产生式规则正是用来刻画一个语言中的组合模式（pattern）的。

“::”是分隔符，它后面是各种类型的合一等式，用来描述产生式规则中从左到右各个非终结符（non-terminal symbol）的特征（feature）。这里用“$”符号代表箭头左边的非终结符（即一条规则的根节点）；用“%”标记箭头右边符号的顺序[2]，“%mp”就表示箭头右边第一次出现的 mp；用“.”号表示对特征的引用，“$.内部结构”就表示箭头左部的 np 范畴的“内部结构”特征。在上面规则中，“$.内部结构=定中”就是一个最简单的合一等式，合一的结果是箭头左部 np 范畴的“内部结构”这一特征的值为“定中”（结构）。上面规则中条件合一式③还引进了程序设计语言中常用的条件控制句“IF ... THEN ... ENDIF”来表示特定条件下才使用的合一，意思是：如果 mp 是由个体量词形成的数量短语，就执行 THEN 后面的合一运算（即要求 mp 后面的 np 的“个体量词”特征跟 mp 的“原形”特征（即 mp 的量词本身）匹配），否则不执行 THEN 后面的合一运算。

规则下面就是在词典中以复杂特征集形式记录的关于词语的语言知识。所谓复杂特征集，是指一些“属性名：属性值”对的集合。比如“量词子类:个体”就是一个“属性名：属性值”对，表示某个量词是个体量词（比如“件”就是个体量词，而“斤”不是个体量词）。显然，这种形式很适合表达各级语言单位的各种属性特征。在上面的形式表达中，“件”后面的 q 表示量词范畴，符号“$”的含义跟规则中的相同，也代表最左的节点，即词典中的每个词条（比如“件”）。“[ ]”之间的内容就是复杂特征。不同的特征以逗号分隔开。同一个特征的析取型取值用“|”分隔开（比如“个体量词:件|套”，就表示“衣服”的“个体量词”特征取值既可以是“件”，也可以是“套”）。不难看出，词典中记录的关于词语的复杂特征描述，是规则中进行合一运算的基础。

以上只是就产生式规则和合一运算的基本面貌做了一个大致的勾勒，关于产生式规则和合一运算的技术细节的介绍，可以参看文献[1，2，3，5]。关于本文介绍的形式化模式的更

为详细的讨论，包括其中所使用的符号的含义，非终结符所代表的语言范畴所依据的理论体系，以及词典中所用到的各个属性特征及其取值的具体含义和规定，可以参看文献[4，9]。

下面两小节对产生式规则的不同类型以及不同性质的合一等式做一些概括的说明。

§2 不同类型的产生式规则

上文介绍的“np -> mp !np”这条规则是产生式规则中最典型的一种情况。从规则的形式出发，我们可以有三个角度来考察刻画汉语的全部组合模式所需要的产生式规则的类型。一是可以就规则中箭头左部的范畴来考察汉语中的短语有哪些功能类型；二是可以就规则中箭头右部范畴的组成情况来考察汉语中短语的不同的构成类型；三是可以根据规则中箭头右部的范畴的性质来看一个产生式规则的覆盖面。下面我们分别说明。

（一）产生式规则箭头左部的范畴实际上代表了一种语言中短语的功能类型，比如上面规则中 np 表示汉语有“名词性短语”这样一种短语类型，mp 表示“数量短语”，等等。那么，要刻画汉语的结构组合模式，需要多少这样的短语类型呢？就目前的研究来看，这个问题还难有确切的答案。这有两方面的原因，一是对短语进行功能分类是比较困难的一个问题，难度显然在词类之上。要得出一个得到广泛认可的分类系统，还有待许多基础研究的深入。比如“大规模种植”中的“大规模”该属什么短语类型，就值得进一步研究。因为分类实际上是对具体语言现象的一种概括，而具体语言现象的差别又是非常丰富的，这就造成了以有限的类型差异去刻画近乎无限的个体差异所必然面临的矛盾。二是分类本身是相对的，不见得一定有一个统一的标准答案。拿 np 来说，可以根据内部组成情况的不同进一步分成定中式的 np 和联合式的 np，定中式的 np 又可以根据组成情况分成所谓的组合式定中 np 和粘合式定中 np（参见文献[11]）。这些类型，也可以像 np 那样以一定的符号来表示（比如 nplh, npdz, npzhdz, npnhdz 等等），出现在产生式规则箭头的左部。不过，也可以只确立一个 np 类型，根据内部组成情况区分的那些 np 类型不通过箭头左部的范畴来表示，而是用“内部结构”这个特征的不同取值来加以说明（比如上面规则中的“$.内部结构=定中”）。可见，分类本身存在相对性，而且以分类形式承载的语言知识，也可以转换为以“属性特征：属性值”形式来承载。尽管分类是相对的，但对于一个具体的形式系统而言，要运转起来，还得有一个相对稳定的短语类型体系来搭起一个产生式规则集的架子。目前我们已经确立的基本短语类型有 np, ap,vp,tp,sp,dp,pp,mp,mcp,dj 这样十种。有关细节可参看文献[9]。

（二）就规则右部组成情况而言，可以根据非终结符数量的多少，分成：

（1）直接上升式规则（箭头右部只有一个非终结符）。比如：np-> !n ， 直观地说，这条产生式规则表示任何一个名词在参与句法组合时实际上就相当于一个名词短语（参见文献[12]）。类似地，还有 np->!r 这样的规则，其中 r 表示代词，这就意味着代词也可以作为一个名词短语参与句法组合；这样做的好处在于，使得整个规则系统的组织是基于短语的，而不是直接建立在词这一级单位之上。从而大大减少了规则的数目（参见文献[6]）[3]。此外，我们可以通过“内部结构=单词”这样的特征描述方式来区分这种直接上升式的规则跟其他组合规则，因而也不致引起混淆。

（2）二叉组合式规则（箭头右部有两个非终结符）。比如上文例子：np-> mp !np，这是典型的产生式规则。这里不再赘述。

（3）多叉组合式规则（箭头右部有三个以上非终结符）。比如：np-> np c np，其中 c 表示连词；np -> np w<、 > np，其中 w 表示标点符号。这两条规则箭头右部都有三个非终结符。理论上，所有的多叉组合式规则都可以通过在非终结符集合中增加非终结符的办法转化为二叉组合式规则（二叉产生式规则利于计算机操作），但就自然语言的具体情况来讲，使用多叉组合式规则在某些场合更符合人们的语言直观。

（三）根据产生式规则箭头右部范畴的性质，来看一条规则的覆盖面，这大致可以分成两种类型：

（1）全局规则：像上文“np->mp !np”这样的规则，我们称为全局规则（global rule）。这类规则箭头右部范畴都是由非终结符（如 mp,np）组成，即描述的是短语类之间的组合，而不是具体词语之间的组合，覆盖面是全局性的。

（2）局部规则：像“np->np !g<者>”，“vp->p<被|为> np g<所> !vp”这样的规则（其中 g 代表语素，p 代表介词，“|”表示逻辑或关系），我们称之为局部规则（local rule）。很明显，局部规则箭头右部的范畴包含终结符（terminal symbol），如“者”，“所”等。这些规则很显然是特别针对这些用法“特殊”的词语的，其覆盖面是局部性的。

区分全局规则和局部规则可以更好地描写汉语的短语结构，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。从技术角度讲，全局规则和局部规则的区分实际上涉及到规则的使用顺序问题，一般而言，局部规则优先于全局规则。当计算机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。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。只要认为某个结构组合足够“特殊”，就不妨特别针对该结构设立一条局部规则来加以描述，这样即使规则描述有偏误，影响也是局部的。

§3 不同性质的合一等式

为了对汉语的一类结构组合（对应一条产生式规则）进行详细的组合条件描述，合一等式需要有各种不同的运算类型（参见文献[3]）。这里我们仅在直观的层面，从表述语言知识的角度来区别合一等式的一些不同性质。

（一）从合一等式的效用看，可以分为描述结构整体情况的合一等式和对结构内部成分进行约束[4]的合一等式这样两类。上文举例时“np -> mp !np”这条规则后面附了三行合一等式（从形式上看也各有差异，本文对此不加讨论，可参见文献[9]）。其中①是对结构整体的情况进行描述的合一等式，这一行里的合一等式实际上刻画了这条产生式规则对应的结构的类型（定中），其中定语是左分支节点（mp），中心语是右分支节点（np）[5]；整个结构作为一个整体不能充任其他结构中的定语成分（$.dingyu=否）[6]，等等。直观地讲，用于描述结构整体情况的合一等式，可以理解为是对相关的属性特征进行赋值。再此之前，这些特征的值都为空。②③两行跟①不同，是对结构内部成分进行约束。

（二）从一个合一约束涉及到的组成成分的数量来看，对结构内部成分进行约束的合一等式可以分为绝对条件约束的合一等式（只涉及到箭头右部某一个成分）和相对条件约束的合一等式（同时涉及到箭头右部两个以上的成分）。例中②是绝对条件约束，因为这个合一等式只涉及到 np 范畴，其含义是要求参与组合的名词必须是能受数量词修饰的那些名词（这就排除了像“心胸”、“眼光”这样的名词参与这种组合）；例中③是相对条件约束，因为这个以条件语句组成的合一等式，同时涉及到 np 和 mp 范畴，其含义是要求参与组合的前后两个成分相互参照，在都满足约束条件的情况下才能组合成功（这就排除了像“纸”这样的名词跟量词“件”组合的情况）。

§4 “把”字结构的相关分析规则

本小节以对汉语中“把”字结构的分析为例，来简要说明要使上文介绍的形式化模式得到具体实现，一般的工作方式是怎样的。

首先凭语言直觉我们就可以把“把”字结构的组合模式表述为：（1）vp->pp !vp；

（1）式中 pp 部分对应不同的产生式规则，比如（2）pp->!p<把> np（整体结构即通常所说的介宾词组）[7]；（1）式中 pp 后的 vp 部分也对应有多种组合模式，比如（3）vp->!vp np（整体结构是述宾组合式的 vp，例如“把房子卖给了老王”）；（4）vp->!vp ap（整体结构是述补组合式的 vp，例如“把衣服洗干净”）；（5）vp->!v u<得> ap（其中 u 代表助词，整体结构也是述补 vp，例如“把香港建设得更美好”）；（6）vp->dp !vp（dp 代表副词性短语，整体结构是状中式 vp，例如“把张三狠狠地批评了一顿”）；（7）vp->!vp vp（连谓式或者联合式 vp，例如“把笔捡起来还给人家”）；……等等。（2）式中的 np 也可能有不同的组合模式，比如（8）np->ap !np（整体结构是定中式 np，例如“把脏衣服拿走”）；（9）np->!np c np（整体结构是联合式 np，例如“把照片和书寄来”）；……等等。

通过上面不完全的罗列就很容易看出，对“把”字结构的分析实际上可能涉及到汉语中几乎所有的组合结构。换言之，以往在讨论“把”字句时常用的“A+把+B+C”这样的公式（参见文献[7，10]），虽然表面形式并不复杂，但实际上其中的 A，B，C 三个位置上出现的成分，可能涉及到汉语所有的短语结构。这是自然语言结构套叠性的鲜明体现。同时这也意味着，一种语言的产生式规则集合中的每条规则之间都可能是相互联系，相互影响的，因而每考虑一条规则的情况，实际上都需要从系统整体的角度着眼。

比产生式规则更重要的是描述这些产生式规则的约束条件。在一篇文章的篇幅里当然不可能去讨论所有结构（产生式规则）的具体组合情况，下面我们集中在上文（1）式（其中 pp 介词为“把”），来讨论如何进行合一描述。实际上，要给出约束条件，通常采用的办法也无非就是内省加上从语料中归纳（半年“人民日报”语料中查找到九千多例“把”字句用于分析，参见附录一、二）两种方式。下面先把部分结果（给出全部约束条件实际上几乎是不可能的，参见下文第五节相关讨论）呈现出来。

```
vp -> pp !vp :: ...,
    IF %pp.原形=把 THEN %vp.把=是 ENDIF,                                    ①
    IF %pp.原形=把,%vp.配价数=3,%vp.内部结构=述宾,%vp.与事=%pp.宾语 FALSE,      ②
    IF %pp.原形=把,%vp.配价数=3,%vp.客体=%pp.宾语 THEN %vp.客体=%pp.宾语 ENDIF,  ③
    ......
```

合一约束①式的含义是要求跟“把”字结构搭配的 vp 的“把”属性取值为“是”。而这又是通过以下两种手段实现的：（1）在词典中需要对每个动词描述它是否能出现在“把”字之后跟“把”字结构形成合理组合。像“是、等于、濒临、搏击、姓、……”等等动词显然是不可能跟 “把”字结构组合的，这些“动词”在词典中的“把”属性值就为“否”；而像“打、洗、说服、作为、……”等等动词都有可能以某种形式跟“把”字结构组合，这些动词在词典中的“把”属性值就为“是”[8]。（2）在有关 vp 的规则中，需要根据具体情况对 vp 的“把”属性进行赋值。比如由“往往，曾经，已经”等副词作状语构成的状中式 vp，“把”属性值就为“否”，这样的 vp 不能跟“把”字结构组合（比如不说“*把台词往往记错了”、“*把拳王曾经打倒过”）。由否定副词“不”作状语构成的状中式 vp 一般也不能跟“把”字结构组合，其“把”属性也为“否”（比如不说“*把论文没写完”、“*把鸡蛋不放在一个篮子里”）[9]。此外，如果 vp 是由动词带谓词性宾语构成的述宾结构，其“把”属性值也为“否”（比如不说“*把篮球喜欢打了”，尽管“喜欢”以别种形式是有可能跟“把”字结构组合的，比如“把奶奶喜欢得了不得）。

合一约束②式的含义是，如果“把”后动词是三价动词（如“送给”、“教”等），并且动词是述宾结构（而不是其他形式，比如述补结构），则“把”后宾语不能充当动词的“与事”论元，即汉语中这样的句子是不合法的：“*张三把李四送给了两本书”，“*诸葛亮把姜维教了八卦阵”。

合一约束③式的含义是，如果“把”后宾语的复杂特征（通常是语义特征）符合“把”

后三价动词的“客体”[10]论元要求（合一成功），则“把”后宾语优先被指派（assign）给“把”后动词做它的“客体”论元。比如“王允把貂禅送给了董卓”，其中“貂禅”就应是“送给”的“客体”论元，而不能是其他语义角色，尽管“貂禅”的语义特征值也是[+人]，符合“送给”这个动词的“施事”论元要求，甚至也符合“与事”的论元要求。

上面这样的分析规则跟词典中有关词语的复杂特征描述相配合，就可以在排除短语结构组合歧义，以及多义动词义项确定等方面发挥实际效用。比如 a“把老板看不惯的人辞退掉”，b“把老板气得生了一场大病的人肯定会被辞退掉”这两例中划线部分的构造格式都是“把+np+vp+的+np”，但 a 应该分析为“[把 [np+vp+的+np]]”；b 则应该分析为“[[把+np+vp] 的 np]”。限于篇幅，这里就不展开讨论了（可参见文献[8，9]）。

§5 余论

通过以上四节从面到点的介绍，不难看出，以形式化的方式来组织汉语的短语结构规则系统，除了可以直接用于计算机分析外，还对汉语的短语结构研究本身有重要的促进作用，因为形式化的表述方式使得对有关汉语短语结构问题的描述变得更为清晰和明确，在这样的背景和基础上，就可以开列出一个系统的问题清单来。譬如：（1）描写汉语的短语结构，需要多少个非终结符？需要设立多少种内部结构类型？（2）一个 np 跟另一个 np 能发生什么组合关系？需要满足什么条件？……等等。而且这个清单上的问题是有层次的，不是笼统混排在一起的。有的问题是在产生式规则这个层次上的，有的问题是在合一约束这个层次上的。后者又可以区分出在句法、语义、语用等不同层次上寻求约束条件。

除了上述清晰明确的问题清单之外，更重要的是，在答案给出之前，我们关于答案的形式就已经有了理性的期待，即最终的答案必须能够表达成某种规则形式（比如本文介绍的“产生式规则+合一等式”这样的形式化模式），而具体的语言知识，也必然需要落实到具体的词语上，能够以“属性名：属性值”的形式在词典中加以记录。从另一个角度说就是，形式化的表述模式为我们思考语言学问题划定了求解的上限——在符号主义的理论框架里，问题只能解决到某个程度，超出这个限度的答案是不可计算的。

值得进一步指出的是，以上述形式化方式表述的短语结构规则都是可以调整的。这包括：（1）产生式规则可以调整。我们可以在非常概括的抽象语类层面上描述一个短语组合模式（全局规则），也可以在非常具体的词语层面上描述一个短语组合模式（局部规则）。比如“np->mp !np”这条规则描述的组合模式是高度概括的；“np->v !g<者>”这条规则描述的组合模式就具体一些，其中有一个特定的成分“者”；而像“vp->!v<吃> a<饱>”这条规则描述的组合模式，就已经具体到两个具体的词了。（2）合一等式可以调整。条件约束既可以放宽，也可以收紧，既可以用句法范畴来约束，也可以用语义范畴来约束。这取决于人对具体短语组合的性质的认识和把握。除此之外，为了更有效更精确地描述词语的用法，词典中的信息也是可以调整的，包括（1）属性项目可以增删归并；（2）具体的属性取值也可根据语料实际情况加以调整。无论上述何种调整，实际上都可看作是一个形式化系统在概括过分（over-generalization）和概括不足（under-generalization）之间寻求一个比较合适的平衡点，从而达到语言知识的概括度和精确度这对矛盾指标的某种平衡。

最后，还有必要简略地说说上述形式化表达模式的不足之处。理论上，本文所讨论的产生式规则加合一等式的形式化模型具有很强的描述能力，但人在理解语言时用到的知识或者太繁杂，或者难以范畴化，因而具体组织起来并不那么容易。比如，可以说“我认识很多名人”，但却不说“*我认识很少名人”，可以说“很少去看电影”，但却不说“*很多去看电影”。这实际上就要求我们区别“很多”跟“很少”。可以说“90 年代”，但不说“*96 年代”，因此又有必要去区别“90”跟“96”。这样的知识显然不在少数而且比较细微，但要想得到正

确的分析结果，这些知识又是必须具备的。再比如，人知道“走路去”、“散步去”、“跑步去”这三个短语的意义差别。“走路去”是“以走路的方式去”；“散步去”是“去散步”（散步是目的）；“跑步去”既可能是“以跑步的方式去”，又可能是“去跑步”（跑步是目的）。计算机却无法借助“方式”、“目的”这样的范畴来理解这三个短语的差别。因为这两个范畴不是静态的，而是在语言成分的使用中动态产生的。“走路”在别的场合可以是“目的”，“散步”在别的场合也可能是“方式”。这样，就难以在词典中区分“走路”跟“散步”实际存在的差别（即不大容易把上述知识以“属性名：属性值”的形式表述出来），相应地，短语结构规则也就难以做出有效地描述。

附录一：《人民日报》1999 年 1 到 6 月份的语料中“把”的分布情况

月份 / 频次 / 词性	一	二	三	四	五	六	合计
把/v	12	25	27	26	12	18	120
把/q	40	46	52	53	56	44	291
把/p	1486	1567	1730	1779	1643	1598	9803
中文字数	1839469	1877008	2056829	2104796	2034790	2051711	11,964,603

附录二：《人民日报》1、2 两月语料中跟“把”字结构发生组合关系的中心动词（V_{ba}）的情况

（1）一月份语料中 V_{ba} 共 468 个；二月份语料中 V_{ba} 共 478 个；两月语料中 V_{ba} 共 723 个；

（2）一月份语料中部分 V_{ba}（频次大于等于 10 的 V_{ba}，词后数字表示频次）：

作为/v 113 结合/v 73 放在/v 69 推向/v 55 送/v 44 办/v 21 提高/v 20 做/v 17

当做/v 15 摆/v 14 变成/v 14 列为/v 14 落/v 14 集中/v 13 建设/v 13 看做/v 12

引/v 12 带/v 11 纳入/v 11 搞/v 10 建成/v 10 交给/v 10 送给/v 10

（3）二月份语料中部分 V_{ba}（频次大于等于 10 的 V_{ba}，词后数字表示频次）：

作为/v 135 结合/v 79 推向/v 62 放在/v 56 送/v 33 建设/v 30 提高/v 24 带/v 19

办/v 17 做/v 17 摆/v 16 当做/v 15 搞/v 14 统一/v 14 变成/v 12 引/v 12

纳入/v 11 推/v 11 献给/v 11 置于/v 11 称为/v 10 带入/v 10 当/v 10 建成/v 10

附 注：

[1] “!”标示的中心成分是个技术概念。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。一般情况下，词的大部分句法语义属性都向上传递，短语的句法语义属性都不继承。

[2] 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。

[3] 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。

[4] 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。参见文献[11]。

[5] 这里的“中心语”概念跟上文提到的以“!”标示的规则里的中心成分概念含义是不同的（参见上面附注 1），尽管在大多数情况下二者是重合的。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 的中心成分，不过这个 np 就不是“中心语”了，而是联合结构的“前项”成分。

[6] 值得强调指出的是，类似“$.dingyu=否”这样的赋值合一等式，是我们在说明一个短语的整体功能性质时最主要的手段。当我们说一个短语某项功能属性的取值为“否”时，并不是指它绝对地丧失了该功能。对此要作相对理解。即当这个短语参与形成更大的组合时，如果在后续分析规则中出现了“dingyu=是”这样的约束条件，该短语的这项功能属性值才有意义，否则这个属性值是“是”还是“否”，并没有什么实际的效用。可以看一个简单的例子。比如“找他”是个述宾式 vp，汉语中一般述宾式 vp 不能再带宾语，即这样的 vp 不能出现在“述语”位置上（“$.shuyu=否”）。但是，汉语中有“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“shuyu”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。我们的处理策略正是，令“找他”的“shuyu”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 mp 前面的 vp 的“shuyu”属性值为“是”（即此种情况下“shuyu”这个属性不发生实际的效用）。

[7] 在实际语料中“把”后并不总是简单的 np，实际上还有可能是其他谓词性成分等，比如“把那些不痛快都忘了”，“天津国有大中型商场已经把不进假货、不卖假货变成了自觉的行动”，“把挑战化为机遇”。

[8] 《现代汉语语法信息词典》中收有动词 10283 个（1998 版），其中标记了能带体词性宾语的动词共 5933 个（“体谓准”属性值为“体”）。而这其中宾语语义类型是“受事”（属于我们的“客体”语义角色）的动词有 3267 个。词典中还标记了这些动词中有 1803 个动词的受事宾语可以提前做“把”的宾语。换言之，也就是在 3267 个可以带受事宾语的动词中，有 55%的动词能跟“把”字结构搭配。而 45%的动词没有标记能否跟“把”字结构组合。据我们抽样考察，实际上在没有标记能跟“把”字结构搭配的动词中，有一部分动词也是可以跟“把”字结构组合的，比如“颁发、更动、干洗……”等，即不能跟“把”字结构搭配组合的动词数量应该不到 45%，还要再少一些。此外，值得一提的是，从理论上讲，在实际语料中考察动词跟“把”字结构搭配的情况，只能得到能够跟“把”字结构搭配的动词清单，却无法得到不能跟“把”字结构搭配的动词清单。这大概也称得上是语料库语言学的一个先天不足吧。关于实际语料中“把”字结构跟动词搭配使用的情况，我们粗略地考察了人民日报 1999 年 1、2 两月的语料，结果见附录 2。

[9] 这是就一般情况说的，实际上汉语中也有“把人不当人”这样的用法，否定副词“不”出现在“把”之后，不过应该将这样的用法看作特例。我们查看了 1999 年人民日报 2 月份的全部“把”字句（1486 句），其中“把”后出现“不”的句子共 19 句（1.2%），但其中只有 1 句属于“不”修饰的 vp 跟“把”字结构组合的情况，“我国的个别企业把发售股票筹集到的资金不用于企业本身的发展，而用于高息放款。”仅有的这句也是因为“把”后面是由“不……而”关联起来的并列 vp 才如此组句的。

[10] 关于本文中所用“配价”、“客体”、“与事”等语义概念的说明，可参见文献[8]、[9]。

参考文献

[1]冯志伟（1991）《Martin Key 的功能合一语法》，载《国外语言学》1991 年第 2 期。

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

[3]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。

[4]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。

[5]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。

[6]徐烈炯（1988）《生成语法理论》，上海外语教育出版社 1988 年版。

[7]薛凤生（1994）《“把”字句和“被”字句的结构意义——真的表示“处置”和“被动”？》，载戴浩一、薛凤生编《功能主义与汉语语法》，北京语言学院出版社 1994 年版。

[8]詹卫东（2000a）《基于配价的汉语语义词典》，载《语言文字应用》2000 年第 1 期。

[9]詹卫东（2000b）《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 2000 年版。

[10]郑定欧（1999）《词汇语法理论与汉语句法研究》，北京语言文化大学出版社 1999 年版。

[11]朱德熙（1982）《语法讲义》，商务印书馆 1982 年版。

[12]朱德熙（1985）《语法答问》，商务印书馆 1985 年版。

[13]Heinecke,Johannes and Juregen Kunze.1998. Eliminative Parsing with Graded Constraints, In Proceedings of Coling'98, 526-530.

[14]Robert D. Borsley, 1996, Modern Phrase Structure Grammar, Blackwell Publishers Inc..