现代汉语短语本位语法体系在机器翻译中的应用及其问题[①]

詹卫东
（北京大学中文系　北京 100871）
常宝宝　俞士汶
（北京大学计算语言学研究所 100871）

摘 要：本文全面介绍了在一个汉英机器翻译系统中，以现代汉语短语本位语法体系为背景构建形式化词典和句法分析规则的情况，同时也指出了面向机器翻译，汉语短语本位语法体系有待进一步研究的问题。

Applications and Problems of Chinese Phrase-Based Grammar in Machine Translation

Zhan Weidong
（Dept. of Chinese ,Peking University　100871）
Chang Baobao　Yu Shiwen
（Institute of Computational Linguistics,Peking University　100871）

Absract: This paper introduces how the Chinese Phrase-Based Grammar is formalized in a Chinese-English Machine Translation system. The formalized grammar mainly includes two parts:a dictionary and a set of rules. In the dictionary we record what functions words have; in the set of rules we describe what types phrases can composed and what constraint conditions each acceptable composed-phrase needs. Furthermore, some problems about the grammar are also discussed.

一、引　言

现代汉语语法理论依据在句子观和析句方法上存在的根本不同，可区分为句本位和短语本位（也可称作词组本位）两大体系[1]。关于这两大体系在理论方法上以及教学实践上的优劣，语言学界已有过广泛深入的讨论，而将这两个不同的语法体系应用到自然语言处理研究方面[2]，特别是机器翻译领域，其各具特色之处，以及是否存在高下之分，却还未见有详细比较。此外，有关以句本位语法体系为基础构造汉语形式语法系统，已有专著论文介绍[3]。而以短语本位语法体系为基础构造汉语形式语法系统的研究文献则相对较少[4]。

本文尝试对现代汉语短语本位语法体系在一个书面语汉英机器翻译系统[5]中的应用做初步的介绍。此外，面向计算机自然语言处理对语言研究提出的新要求，也指出这种语法体系有待进一步改进的地方。

[①] 本文的研究工作受国家“863”项目（编号 863-306-03-06-2）和国家自然科学基金项目（编号 69483003）支持。

二、基于短语本位语法体系的形式语法系统的基本框架

2.1 基本原则

（1） 功能分类思想

短语本位语法体系以语法功能为主要标准对词进行分类[6]，对短语的分类，则既有从结构角度分的，如述宾短语、述补短语、偏正短语等等；也有从功能角度分的，如名词性短语、动词性短语等等[7]。前者在语法研究与教学中，用得相对比较多和比较成熟一些。但在构造形式语法系统时，则需全面贯彻功能分类思想。所谓功能，实际上是对一个语言成分能够出现的位置的高度概括，也即是对一个语言成分跟其他语言成分组合能力的抽象描述。功能分类的结果可以直接说明一个语言成分怎样向外组合。所谓结构，则是对一个语言成分内部组成成分及其相互关系的描述。结构分类的结果可以直接说明一个语言成分内部构造的情况，但对一个语言成分怎样向外组合却是间接影响。而语法系统的直接目的同时也是最终目的，就是要说明一个语言成分如何跟另一个语言成分组合构成一个更大的语言成分。显然，从词到短语都宜采用功能类标记，用以组织形式语法系统。

（2） 功能实现思想

短语本位语法体系认为在语言成分的各级单位中，从词到短语是组成关系（composition），从短语到句子是实现关系（realization）[8]。在构造形式语法规则系统时，我们有意识地在一定程度上模糊词和短语的差别，认为从词到短语，除组成关系外某些时候同样也可以有实现关系[9]。句子作为目前形式语法系统处理对象中的最大单位，从整体结构上看是由短语加上其后标点符号（只能是三类点号“。？！”中的一种）组成的；仅就其中心成分短语而言，句子跟短语是同构实现关系。

2.2 基本标记符号

（1）为处理现代汉语中的语素字，形式语法系统中包括一个语素标记：g

（2）语素不再分类。现代汉语词分为 18 类。词类标记[10]如下：

名词 n 代词 r 连词 c 处所词 s 时间词 t 方位词 f
数词 m 量词 q 助词 u 区别词 b 状态词 z 形容词 a
动词 v 副词 d 介词 p 叹词 e 语气词 y 拟声词 x

（3）我们按功能标准将现代汉语短语分为 11 类。短语标记[11]如下：

名词短语 np 数量短语 mp 时间词短语 tp 处所词短语 sp
数词短语 mcp 动词短语 vp 介词短语 pp 形容词短语 ap
副词短语 dp 单句型短语 dj 复句型短语 fj

（4）整句是目前语法系统处理的最大单位。标记为：zj

2.3 基本组成

形式语法系统包括两大部分[12]：词典和规则。词典对词语的语法语义信息作详细记录；规则对汉语短语组合类型和组合条件加以说明。以下我们对词典和规则的内容展开论述。

三、词典：句法语义属性特征描述

机器词典中以复杂特征集[13]（complex feature set）的方式对词语的句法语义属性进行描述。目前我们词典中的句法语义信息可以大致分为三类，在不同程度上体现了功能思想。

（一） 基本信息：一个词所属词类、语义类。这是对该词的功能作最一般性的概括。

（二） 搭配信息：一个词跟其他成分的组合能力。这包括句法和语义两方面。

（三） 位置信息：一个词充当句法成分的能力。

下面试以词典中对名词“白杨”的描述为例说明[14]：

$$ 白杨 ** {n} n $=[名词子类:na,个体量词:棵|株,前名:是,前动:否,后名:是,名状语:否,
名主语:是,名宾语:是,名谓语:否,临时量词:否,语义类:植物]

词目“白杨”的复杂特征集合中即包含了上述三类信息。（1）基本信息：词类、名词子类、临时量词、语义类等；（2）搭配信息：个体量词、前名、前动、后名等；（3）位置信息：名状语、名主语、名宾语、名谓语等。这些属性中，“名词子类”取值为“na”，这个代码表示该名词是可数名词，可受数量短语修饰；“临时量词”为“否”表示该名词不能像“脸”之类的名词那样临时充当量词使用（如“一脸泥”）；“个体量词”用以标记能修饰该名词的量词；“前名”、“前动”、“后名”分别用来标记该名词跟其前后名词、动词搭配的能力；“名状语”表示该名词是否能作状语，其余“名主语”等类推。在句法信息之外，还设置了“语义类”项来标明其语义性质。下面我们再以动词的一项属性为例，来进一步说明这样标记词的功能对机器翻译具有的重要意义。

动词大多可受介词“被”的修饰，表示被动意义。但并非所有的动词都有此用法。在词典中，我们以“被”属性项来标记动词的这一功能特征[15]。如：“邀请、发现、逮捕”可受“被”的修饰，标记为“是”；“参加”则不能，标记为“否”等等。这一功能差异对分析和翻译将产生影响。请看对比例句：

a. 刘教授被一家公司邀请参加了大会

b. 小偷被警察发现逮捕了

这两例划线谓语部分的排列格式是同形的，都是 pp+vp+vp 格式。但内部组合方式却不一样。例 a 中“参加”不能受“被”修饰，谓语部分的层次结构因而是：

被一家公司 邀请 参加了大会
| vp | | vp |
| pp | | vp |

“参加了大会”不跟“被”直接发生结构关系，而是跟“被一家公司邀请”组合后作谓语。“参加”是“刘教授”主动参加，而非被动的。“邀请”是被动的。

例 b 中“发现”和“逮捕”都可受“被”修饰，谓语部分的层次结构因而是：

被警察 发现 逮捕 了
| pp | | vp |
| vp | | vp |

“发现”跟“逮捕”都是被动的，二者先结合在一起，再共同受“被”的修饰。结构层次跟例 a 不同。计算机根据词典中动词的“被”属性取值不同，对上两例中的结构差异可作出准确判断。从而得到正确的译文[16]。

a ` Professor Liu is invited by a company to have attended meeting.

b ` Thief has been found and arrested by police.

由此可见，词典中对词语的功能特征也即用法作准确描述，对机译系统进行源语句法分析有显著的帮助。限于篇幅，以上仅简单举例说明。下面再介绍规则组织的情况。

四、规则：组合模式与组合条件

我们的形式语法系统以上下文无关文法（Context Free Grammar）的产生式（rewriting

rule），结合合一等式（Unification ）构建句法分析规则集。上下文无关文法产生式用以描述短语组合的模式类型；合一等式用以说明短语组合的条件限制。

形式语法系统中的产生式规则不仅描写词语的组合类型，还包含词直接上升为短语的规则。汉语中词跟短语之间并无天然的形式差异。甚至还存在如“来信”这样的，既可理解为短语，又可理解为词的歧义现象[17]。另外汉语中有所谓的独词句，有的单个词加上句调即可成句（如：“好！”）。再从构造形式规则的简洁性要求考虑，规则必需是递归的。由于这些因素，我们允许有的词直接上升为短语。

此外，形式语法系统中的产生式规则大多是二叉组合的，既基本遵循层次分析中层层二分的做法。但鉴于汉语句法结构中尚有许多复杂结构，如中间嵌“得、不”的述补 vp，对其性质的认识还不是非常清晰，同时为了降低处理的复杂度，我们允许多叉组合短语[18]。在短语内部也可以包含逗号、顿号等标点符号。

规则的合一等式部分主要是充分利用词典信息以及短语属性信息，来描述短语组合的条件限制。我们这里以上节 PP 加 VP 组合为例简要说明。请看规则[19]：

```
&& {vpzz1} vp->pp !vp :: $.内部结构=状中,$.状语=%pp,$.中心语=%vp,…
                         IF %pp.yx=被 THEN $.被动=是 ENDIF,
                         IF  %pp.yx=被,%vp.被=否 FALSE,…
```

上面规则中“IF … FALSE”语句，就是用来描述“被”字形成的介词短语跟 VP 组合时，对 VP 的属性要求。如果 VP 是不能受“被”修饰的动词（%vp.被=否），组合失败。限于篇幅，文中规则的合一等式没有全部例出，仅用省略号表示还有其他合一约束。规则中“内部结构”是短语的属性描述，用于标示汉语短语的结构性质。目前设置的其他短语属性还包括像“状语”这样指明短语内部成分的项目，以及“被动、否定”等一些。无论是词典信息还是短语属性信息，都是在短语本位语法体系框架内进行组织的。实践证明，这些信息用于说明汉语短语组合的约束条件是颇有成效的。

综合上两节的内容，可以看到，词典跟规则组成了一个有机系统。其机制允许整个形式语法系统可在内部调整语言知识的组织方式。为以形式化方式检验并改进短语本位语法体系提供了良好的基础。

五　有待进一步研究的问题

本节我们集中在理论层面，讨论短语本位语法体系用于机器翻译时面临的问题。可以从两方面谈。

5.1　对短语分析的深度要求

短语分析是短语本位语法体系的核心内容。就机器翻译而言，要求供计算机使用的词典和规则，要描述得更为明确严格。什么样的短语在什么条件下以什么方式组合，都应该是毫不含糊的。这其实也就对研究语言规律的人提出了更高的要求。举例来说，对有多项定语修饰的定中 np，人容易作出正确分析，机器如果没有明确的规则约束条件帮助，是很难分析得到正确结果的。如：对“他是（我们学校最好的学生）”中括弧内的短语，机器分析得到两个结果：

a　我们学校最好的学生
　　|　np　||　　np　　|
　　　　|　ap　||　np　|

b　我们学校最好的学生
　　|　　dj　　|　|np|
　　|　　np　　||ap|

在人看来，显然 a 是正确结果，b 是错的。但道理何在，却不一定能说出多少来。就这两种分析各个局部阶段而言，都没有什么不对。但总体上看，却有对错之分。按 a 分析，

“学生”前有两个短语（np 和 ap），在不同层次上先后跟其发生定中修饰关系。 按 b 分析，“学生”前的两个短语组合起来在同一层次上对其进行修饰。两种分析得到两种译文：

a ` the best student of our school （正确译文）

b ` student who our school is the best （错误译文）

像上面这样的问题在计算机处理中是大量存在的，譬如汉语中的连谓式 vp、无标记联合式 np、vp、ap 等等，内部组合条件目前都尚未概括清楚。这些最终可归结为，对汉语中的各类短语组合类型，我们都必须问一句，在什么条件下可以这样组合？毫无疑问，面向机器翻译，短语分析的深度是短语本位语法体系必须继续加强研究的努力方向。有关从词到短语产生了哪些功能变异，短语本身应该设置哪些属性来描述其功能特征，都是迫切需要进一步研究的问题。

5.2 对句子分析的广度要求

如果说前一个问题是因为自然语言处理的特殊性，对短语本位语法体系只是在原有框架内提出了更高的研究目标，那么现在要讨论的问题，就带有对短语本位语法体系基础框架重新进行评价的性质。

包括机器翻译在内的自然语言处理研究，都需面临真实语料。这就容易发现我们对句子和短语的处理，在理论上和实践上都还存在不完善之处。突出的表现在两个方面：一是现有短语包括单句型短语 dj 和复句型短语 fj。这两个类较之 np、vp 等显然更为复杂，通常都包含一些不易作为短语成分处理的东西（如语气词、连词、各种标点等），或者内部并不是简单的主谓构造，而是所谓的存在句、流水句等等，但理论框架限制了 dj 和 fj 都不能作为句子对待；二是句子目前都严格规定为由一个短语加上标点构成，但对短语的成句条件，即对什么样的短语能够实现为一个句子却并不十分明确[20]。我们可以简单举两例说明：

（1）这样看来，有两种不完全的知识，一种是现成书本上的知识，一种是偏于感性和局部的知识，这两者都有片面性。

（2）下雪了，房顶上、树上、地上一片白。

例（1）内部结构复杂，其中“这样看来”还带有插入语性质，整句作为短语不易分析。例（2）翻成英语恐怕得两个整句，但汉语是一个很自然的流水句。分析和转换都比较困难。对其成句条件，现在可以说也是所知甚少。

上述两方面统括起来，就是短语本位语法体系面临的对句子分析的广度要求问题。即基于短语本位语法体系构造的形式化句法分析系统，是否能为所有的语言形式贴上合适的功能标签；是否能为所有的句子找到短语归宿。显然，对这两个问题的回答需要短语本位语法体系在拓宽分析的适用范围上做出努力。

总的看来，以短语本位语法体系为基础构建的形式句法分析规则，覆盖了大部分的短语组合类型。但对句子的实际复杂性仍有照顾不到的地方。对覆盖范围内的短语构造，需往深处探索，尽力挖掘短语组合的条件限制；对分析句子鞭长莫及之处，需拓宽广度，在研究跟短语异构的句子方面下番工夫。

六 结语与致谢

本文对以短语本位语法体系为背景构造面向机器翻译的汉语形式语法分析系统，做了一些介绍。以往对汉语语法体系的研究主要是面向人的。我们希望，通过本文的讨论分析，可以为在计算机自然语言处理领域进一步比较不同语法体系的得失提供一个基础。目前本系统的词典规模还在扩充之中；规则经过调试现基本稳定在 300 条左右，对汉语句法的覆

盖面和分析深度都达到一定水准。

中科院计算所二室刘群先生是本系统形式语法描述规范的设计者。我们借此机会特别向他表示感谢。课题组成员刘颖、王斌同学对文中内容都提出过宝贵意见，也一并致谢。

附注：

[1] 语言学史上也有所谓词本位的语法体系，但其影响远不及句本位和短语本位两个体系。可参见 朱德熙（1985）《语法答问》，商务印书馆；陆俭明（1993）《八十年代中国语法研究》，商务印书馆。

[2] 目前在自然语言处理领域采用的语法体系除句本位和短语本位两种体系之外，还有依存语法体系影响较大，分析方法跟前二者都有很大不同。可参阅黄昌宁等（1992）《语料库、知识获取和句法分析》，载《中文信息学报》1992 年第 3 期；刘伟权等（1996）《建立现代汉语依存关系的层次体系》，载《中文信息学报》1996 年第 2 期。其他还有基于范畴语法的汉语形式语法体系，可参阅翟成祥等（1991）《汉语组合类型语法》，载《中文信息学报》1991 年第 3 期。

[3] 参见吴蔚天 罗建林（1994）《汉语计算语言学》，电子工业出版社。

[4] 参见俞士汶（1996）《自然语言理解与语法研究》，载《计算语言学文集》，内部资料。

[5] 自 1994 年开始，中科院计算所二室与北大计算语言所合作开发汉英机器翻译系统，笔者一直参加该系统语言知识库的工作。目前系统还在调试扩充阶段。

[6] 参见俞士汶等（1996）《现代汉语语法信息词典规格说明书》，载《中文信息学报》1996 年第 2 期。

[7] 参见吕叔湘（1979）《汉语语法分析问题》，商务印书馆；俞士汶（1993）《关于计算语言学的若干研究》，载《语言文字应用》1993 年第 3 期。后者可以看作是本文对短语结构语法体系进行形式化描述的思想渊源。

[8] 参见朱德熙（1985）。

[9] 郭锐（1996）《汉语语法单位及其相互关系》，载《汉语学习》1996 年第 1 期。在语法理论层面讨论了从词到短语贯彻功能实现思想。可以参阅。

[10] 同注 6。目前我们的机译系统对拟声词尚未作处理。

[11] 参见周强 俞士汶（1996）《汉语短语标注标记集的确定》，载《中文信息学报》1996 年第 4 期。

[12] 目前北大计算语言学研究所正在进行国家自然科学基金项目“面向理解的汉语短语信息库的构造”方面的研究。其目标跟词典建设是一致的。最终成果也可以结合到形式语法系统中，服务于机器翻译应用。

[13] 有关复杂特征集和下文合一运算的细节内容，可参看冯志伟（1990）《汉语句子描述中的复杂特征》，载《中文信息学报》1990 年第 3 期；冯志伟（1991）《Martin Key 的功能合一语法》，载《国外语言学》1991 年第 2 期。

[14] 机器词典中为节省空间，很多“属性：值”信息是采用默认值方式描述的，这里为了说明词典的一般情况，把词典中名词的属性全列了出来。描述中用到“$$、**”等一些标识符，就不详细说明了。

[15] 参见詹卫东（1996）《现代汉语 VP 的结构层次和结构关系判定》，北京大学硕士学位论文。

[16] 目前所谓正确的译文，主要是考虑对汉语句法结构的正确分析在译文中能体现出来。对译文中一些时态、词形变化等细节问题，尚未进行很好的处理。这些是下一步工作的重点。

[17] 参见陆俭明（1988）《名词性“来信”是词还是词组》，载《中国语文》1988 年第 5 期。

[18] 同注 11。

[19] 规则中“{ }”内字母数字表示规则序号，“::”用来将产生式跟合一等式分隔开，“$”表示产生式左部根结点，多个合一约束之间以“，”分隔开，“%”表示产生式右部结点的顺序位置。如“%vp”表示右部的第一个 vp 结点。依次类推。

[20] 参见朱德熙（1985）。