应用 HPSG 理论实现汉语 NP 的自动分析*

王璐璐 陈芯莹 李森 林晨
中国传媒大学应用语言学研究所 北京 100024
E-mail: {hetty_luluwang, cici13306, lovelyls }@hotmail.com, heshang110@cuc.edu.cn

摘要： 在汉语 NP 自动分析中，名词和量词的搭配问题是难点。本文在 HPSG 理论框架下，构建了汉语 NP 的形式化模型。并且重点分析了该模型在 LKB 系统中的实现情况。文章也总结了用 LKB 实现 HPSG 的优势与不足之处。

关键词： NP、LKB、HPSG

Automatic Analysis on Chinese NPs applying HPSG

Wang Lulu, Chen Xinying, Li Sen, Lin Chen
Applied Linguistics Institute, Communication University of China, Beijing, 100024
E-mail: {hetty_luluwang, cici13306, lovelyls }@hotmail.com, heshang110@cuc.edu.cn

Abstract: The noun-classifier matching is one of the key issues in the automatic analysis of Chinese NPs. In this paper, we propose a formalized model of Chinese NPs in the framework of HPSG, and especially focus on the implementation of the model in the system of LKB. We also figure out the advantages and disadvantages of the LKB system implementing HPSG.

Key words: NP、LKB、HPSG

0 引言

现代汉语中，名词短语分析中的一个主要难点是名词和量词的相互选择，因为“个体量词和个体名词之间的搭配是约定俗成的”（朱德熙，1982）。由此，对汉语 NP 结构，仅有浅层句法分析是远远不够的，还需要更精细的理论来描述。HPSG 理论是一种基于约束的形式化句法理论。该理论不仅能够描写多种语言，还被应用于许多语言工程的具体实践中。并且，基于该理论框架构建了以 LKB 和 TRALE 为代表的语法开发平台（Melnik，2005）。这样就可以将手工建立起来的语法形式化模型在计算机上进行实现，以实现语言的自动分析。

本文采用 HPSG 理论框架，以汉语 NP 为研究对象，考察 HPSG 理论在 LKB 应用平台上实现汉语的情况。文章第一节简要介绍 HPSG 理论以及 LKB 系统的功能。第二节提出汉语 NP 的形式化模型，并分别从句法，词汇和语义三方面进行分析。第三节重点介绍了所构建的语法文件，并给出在 LKB 系统里分析汉语 NP 的实验结果，包括树图、规则图和语义关系的展示和说明。第四节就在应用过程中发现的问题做一总结。

1 HPSG 理论和 LKB 系统

* 衷心感谢中国传媒大学应用语言学研究所的刘海涛教授对本文的悉心指导与建议。

HPSG（Head-Driven Phrase Structure Grammar）理论，即“中心语驱动短语结构语法理论”，由俄亥俄州立大学的 Carl Pollard 和斯坦福大学的 Ivan A. Sag 于 1987 年首次提出相关概念。后经过不断完善，于 1994 年形成了比较系统的理论框架。此过程中借鉴了许多相关理论，例如广义短语结构语法（GPSG）、词汇功能语法（LFG）、范畴语法（CG）、情景语义学和话语表达理论（DRT）等。HPSG 理论具有较好的普适性，能够应用于多种语言的语法分析之中；同时，HPSG 理论还具有强大的弹性，能够对具体的语言知识进行精确的描述。目前，英语、法语、德语、日语、韩语等多种语言都有基于 HPSG 理论的较成熟的语法框架。HPSG 理论主要有三个主张：表层导向，基于约束和词汇主义（吴云芳，2003：231）。简单来说，HPSG 理论把语言的句法、语义和语用信息利用复杂特征表现在词汇中，并通过语法规则来约束这些关系，从而达到描写一种语言的目的。

LKB（Linguistic Knowledge Builder）系统是一个用类特征结构语法来开发语法和词典的平台（Copestake, 2002）。它不仅是为了分析大规模语法，还涉及基于合一运算的自然语言处理的分析与生成（化柏林，2004）。其理论框架是基于 PATR 的，并被最广泛地用于测试基于 HPSG 理论的大规模语法。LKB 系统自 1991 年问世以来，一直在网络上开放资源，有许多人参与到研究中来（如 Carroll、Malouf 和 Oepen 等），研究成果显著。如斯坦福大学开发的英语语法项目，即 LinGO English Resource Grammar（ERG）。

LKB 系统软件的主要功能包括编译类特征结构的语法，剖析句子是否合乎语法，以及句子的生成情况。而且，该系统还提供了多种可视化手段，如层级体系结构可以看出整个语法的框架，各节点之间的关系，树形图则提供了最直观的句法树信息，还可以查看句子的特征结构图和规则等等。另外，系统软件还提供调制和跟踪功能。简单地说，LKB 系统使得人工构建的语法得以在计算机上实现。这样，语法开发人员只需要写出基于类特征结构的语法文件，然后利用该系统软件查看句子的剖析和生成情况，进而对构建的语法有更深入的理解。所以，本文采用 LKB 系统来分析 HPSG 理论在处理汉语 NP 时的实现情况。

2 汉语 NP 的形式化模型

HPSG 理论旨在通过表层的句法现象来揭示人类的语言能力。这种表层的句法现象就是所谓合乎语法的句子。HPSG 理论要对这种合乎语法的句子给与描述，通常通过大量的词汇信息和少量的规则来构成对合法性的约束。由此，本节中我们先对汉语 NP 结构进行描述，继而用 HPSG 理论从句法规则、词汇特征和语义约束这三方面内容来进行分析。

现代汉语 NP 结构的主要特点是：修饰语在前，中心词名词在后。如：一本汉语书；如果几个定语都不带“的”，修饰语的顺序一般是：（1）领属性定语，（2）数量词，（3）形容词，（4）名词。如：他那件羊皮大衣。（朱德熙，1982）并且，NP 内部成分间的主要特点是名词和量词存在搭配关系。根据 Wang/Liu（2007）对《人民日报》2000 年数据中 NP 的统计，得出 NP 基本结构的频率分布（详见表一）。

序号	NP 结构类型	频率	例子
1.	Dem + CL + N	158	这本书
2.	Num + CL + N	93	一本书
3.	Dem +Num+ CL + N	19	这两本书

表一 NP 基本结构的频率分布

本文选取最简单的 NP 结构，即不考虑带"的"等复杂定语的情况，只针对上表中 "Dem + CL + N"、"Num + CL + N"和"Dem +Num+ CL + N"这三个基本机构来进行分析.下面我们就用 HPSG 理论来分析各成分间的关系，以及对量词名词的搭配给予约束。

首先从句法的角度出发，我们知道现代汉语的 NP 结构由指示词、数词、量词和名词等成分组成。Wang/Liu（2007）认为，在 NP 结构中，名词是中心语（Head），指示词和数量结构分别作名词的限定语（SPR）。由此，我们需要构建一个新的规则，即双限定语规则，如图一所示。

$$X\left[\text{SPR}\ \boxed{1}\right] \rightarrow \boxed{2}\left[\text{SPEC}\ \boxed{3}\right]\ .\ \boxed{4}X\left[\text{SPR}\ \boxed{1} + \left\langle \boxed{2} \right\rangle\right]$$

图一 双限定语规则

其次，在 HPSG 理论中，词典本身可以被看作是一个类型层级体系。（Sag/Wasow，2003）词汇间通过缺省传承将上级词汇的类特征传递给下级词汇，即下级词汇一定具备上级词汇的类特征。如我们定义 "ng-lxm" 代表个体名词，它的类特征从属于名词（noun-lxm）的类特征。

最后，对于名词量词的搭配问题，我们试图从语义的角度给与约束。HPSG 理论的语义部分来源于情景语义学，但是它不考虑语用方面（如场景）或者词汇语义学的内容的知识，而注重短语结构中各成分的语义关系。比如短语的语义特征是各成分的语义汇总，而上下级节点则可以承袭语义特征。由此，词项有这样几个基本概念，即 INDEX 指代某一情景，RESTR 中加入语义限制的特征。为了解决上文中汉语名词量词的搭配问题，Wang/Liu（2007）提出 CLS 特征来表示搭配的约束特性（如图二所示）。

$$\left\langle \text{sbn},\ \left[\begin{array}{l} \textit{ng-lxm} \\ \text{SEM}\left[\begin{array}{ll} \text{INDEX} & i \\ \text{RESTR} & \left[\begin{array}{ll} \text{RELN} & \text{book} \\ \text{CLS} & \text{bound} \\ \text{INSTANCE} & i \end{array}\right] \end{array}\right] \end{array}\right] \right\rangle$$

图二 "书"的词项

3 如何用 LKB 系统处理汉语 NP 结构

LKB 系统所应用的类特征结构语法由两方面因素所决定，即类特征结构的数据和合一的运算（Copestake，2002）。而不管是规则，还是词汇，每一部分都是一个类特征结构。在 HPSG 理论中，这些特征由属性值矩阵（Attribute Value Matrix,简称 AVM）来表示（如图二）。而在 LKB 的语法中，这些特征由特征描述语言（Type Description Language,简称 TDL）来表达。由此，我们需要将这些矩阵图转换为类特征描述语法。下面的内容将详细介绍语法中构建主要类特征结构语法，包括词汇、规则和类特征层级体系。

另外一点值得注意的是，这种语法的构建，不同于编程语言，而是以人们的语言行为作为依据。这些语言行为就是合乎语法的、合格的句子。由此，我们将剖析（parse）一组汉语名词短语，并通过这些剖析结果，判断哪些是合乎语法的，哪些是不合乎语法的。最后，我们将给出主要的剖析结果，包括树图、规则图、语义图等。

3.1 LKB 系统里构建的语法

一套基本完整的 LKB 程序包括 11 个文件。其中，最主要的部分有类型系统（type.tdl）、词汇列表（lexicon.tdl）和语法规则（rules.tdl）。下面，我们分别就这三个文件来具体说明我们是如何构建语法的。

第一，类型系统是 LKB 语法的基础，它定义了整个语法的框架，以语法知识的层级体系为表现形式。由此，所有的语法单位都要在一个层级体系里体现，并且以一个特征结构来表达，这样才能达到语法单位特征的承袭以及系统的概括。例如，我们在类型系统中定义“*top*”统领所有语法知识，feat-struc 从属于这一属性，表现为“feat-struc := *top*.”。而每一个具体的语言单位则都要从属于 feat-struc，如下面的代码所示。

expression := feat-struc & [ORTH *dlist*, HEAD pos, OPT *bool*, SPR *list*, COMPS *list*, SEM semantics, ARGS *list*].

这里 expression 的特征语法定义了各子类型，包括拼写（ORTH），中心语（HEAD），限定语（SPR），补足语（COMPS），语义信息（SEM）和论元结构（ARGS）。所以，在描述每一个语法单位时，都要有这几个类型。其中，拼写（ORTH）在词汇列表中给出具体的拼写形式，其他都在类型系统中给出定义。

在类型系统中，语法知识都被嵌入词汇的层级体系中。因此，词汇类型是语法的基础部分，我们以名词的词汇类型定义来说明。如下面的代码所示，名词（noun-lxm）通过层层传承，从属于 expression。再者，在名词的句法和语义特征定义中，可以看出其中心语(HEAD)为名词(noun)，限定语（SPR）的中心语为量词（clf）。而且，限定语的中心语量词的语义属性（INDEX）要与中心语名词的相一致。

lex-item := expression.

lexeme := lex-item.

contentful-lxm := lexeme & [SEM [INDEX #index, KEY #key & [ARG0 #index], RELS <! #key !>]].

noun-lxm := contentful-lxm & [HEAD noun, SPR < phrase & [HEAD clf, SPR <>, COMPS optional-list, SEM.INDEX #1] >, COMPS optional-list, SEM.INDEX object & #1].

可以看出，通过对名词类特征结构的定义，我们可以在语法中表达 NP 的句法关系，即在名词短语中，名词是中心语，而量词是名词的限定语。另一方面，对于量词名词的搭配限制，我们在语义类型中给与约束，下面就是语义类型的定义。这里的 CLS 特征就是为了解决量词名词搭配而定义的新变量。

semantics := feat-struc & [INDEX individual, KEY relation, RELS *dlist*].

relation := feat-struc & [PRED *string*, ARG0 index].

individual := *top* & [CLS *string*].

第二，词汇项给出了词汇的线性序列和语法知识的描述。如下面所示“书”（shu）的词汇项，它包括三条基本信息：（1）句法信息，如 noun-lxm 表示名词；（2）拼写信息，即输入的词语，如 ORTH <! "shu" !>，表示“书”；（3）语义信息，SEM.KEY.PRED "shu_rel"表示该词基本的语义特征, SEM.INDEX.CLS "bound" 表示该词与其他词汇发生关系的语义属性，即名量搭配限制。相应地，量词“本”(ben)和“台”(tai)分别具有不同的语义属性，即“bound”和“machine”。

shu := noun-lxm & [ORTH <! "shu" !>, SEM.KEY.PRED "shu_rel", SEM.INDEX.CLS "bound"].

ben := clf-lxm & [ORTH <! "ben" !>, SEM.KEY.PRED "ben_rel", SEM.INDEX.CLS "bound"].

tai := clf-lxm & [ORTH <! "tai" !>, SEM.KEY.PRED "tai_rel", SEM.INDEX.CLS "machine"].

第三，语法规则定义了如何将词汇项和短语相结合，从而构成新的短语。也就是说，它实现了词汇项到短语的构建。下面就是根据双限定语中心语规则改写成的 SPR 规则 1，其中 ARGS 是 argument structure 的缩写，即论元结构。这里指包括其限定语和补足语的表，表中元素用逗号隔开。如上面的#1 即表示逗号后短语的限定语。而且，规则指名#1 与短语的限定语属性一致，都为#1。我们用#rest 属性来找到短语中的每一个限定语。

specifier-head-rule-1 := binary-head-final & [SPR #rest, COMPS #comps, ARGS < #1, [SPR [FIRST #1, REST #rest] , COMPS #comps & optional-list] >].

3.2 剖析实验及其结果

上面就是我们在 LKB 系统里构建的汉语 NP 的语法。接下来，我们来做一个自动剖析的实验。在 LKB 系统的菜单里，可以选择批量剖析句子，即在文件中写好待剖析的句子，从而可以一次得到多个剖析结果。我们在测试文件（all.item）中给出如下 5 个名词短语。这里需要说明的是，目前 LKB 系统的软件还不能处理汉字，所以我们在做剖析实验时，用拼音来代替汉字。其中，“*”表示不合乎语法。

1 yi ben shu

2 yi tao shu

3 *yi tai shu

4 zhe yi ben shu

5 zhe ben shu

在运行语法后，我们在测试结果文件（all.resluts）中得到剖析结果，如下所示。

1 yi ben shu 1 21

2 yi tao shu 1 21

3 *yi tai shu 0 20

4 zhe yi ben shu 2 37

5 zhe ben shu 1 18

;;; Total CPU time: 15 msecs

;;; Mean edges: 23.40

;;; Mean parses: 1.00

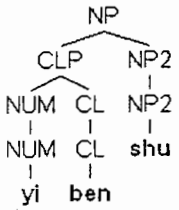
我们注意到第三条短语“一台书”（yi tai shu）的起始数字是 0，这表示它未能成功地被剖析。如果我们单独剖析这一短语，则会得到“No parses found”。可见，目前我们的语法是成功的，它成功地排除了不合乎语法的短语。并且，如果我们在 LKB 系统中选择查看语义信息，即 MRS。则会得到如下结果，也就是说“本”（ben）和“书”（shu）的语义特征一致，都被定义为“x1[x CLS：bound] ”(如下图所示)。

```
[
 INDEX: x1 [ x CLS: bound ]
 RELS: <
     [ "yi_rel"
       ARG0: u2 [ u CLS: *STRING* ] ]
     [ "ben_rel"
       ARG0: x1 ]
     [ "shu_rel"
       ARG0: x1 ] > ]
```

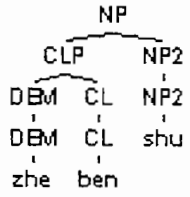
图三 “一本书”的语义信息

4 存在的问题

通过上面的结果显示，我们构建的语法成功地处理了量词名词的搭配问题。但是，如果进一步分析上述的几条名词短语，我们发现目前的语法还是有一些问题的。具体来说，以“一本书”和“这本书”为例，二者目前都可以被成功地剖析。不过，我们在深入分析其句法树，甚至是规则图时，可以发现，目前 LKB 的分析并不能严格按照 HPSG 理论来分析问题。

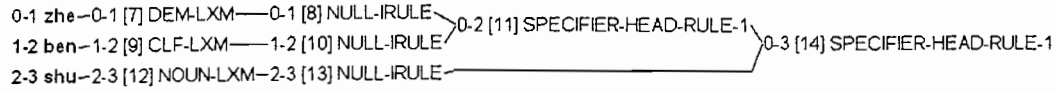


图四 “一本书”的句法树图



图五 “这本书”的句法树图

如上面图四和图五所示，“一本书”和“这本书”的句法分析是一样的。但是实际上，“这”应该和名词构成限定关系，而不是和量词。虽然我们在类型系统里定义了限定词的补足语为名词，但是在实际分析中，它却和量词发生关系。据此，我们继续考察“这本书”（zhe ben shu）的规则图（如图六所示），发现“这”（zhe）和“本”（ben）根据 SPR 规则 1 而联系到一起。由此可见，虽然 SPR 规则 1 可以分析包括“一本书”、“这本书”和“这一本书”的结构，但是并不十分理想，有描述过度的倾向。



图六 “这本书”的规则图

而且，对于名词量词的语义限制，CLS 特征应该加在 INDEX 还是 RESTR 中也是一个有争议的问题。因为按照 MRS（Minimal Recursive Grammar）的理论，INDEX 的约束太严格，对于我们处理大规模的语法存在不利的因素。

5 结论

本文在 HPSG 理论框架内提出了汉语 NP 的形式化模型，分析了其句法和语义的关系。并且

将这些语言知识转化为 LKB 系统的形式化语法来对汉语简单 NP 结构进行自动分析。该分析在一定程度上解决了量词、名词的搭配问题。并且我们发现 HPSG 的复杂特征结构集中在词汇中，这样就大大减少了语法规则的数量，这点对于我们语法工程的开发有较大的参考价值。而与此同时，我们注意到，目前的自动分析还存在若干问题。例如在现代汉语中，名词量词的搭配不仅仅限于一对一，更多的是一对多的关系。而且在真实语料中，还有许多复杂的 NP 结构和省略成分的现象。这些问题都有待于我们进一步地研究。

参考文献

1. Copestake, A. *Implementing Typed Feature Structure Grammars.* Center for the Study of Language and Information, Stanford, CA: CSLI Publications. 2002.
2. Melnik, Nurit. *From "Hand-written" to Computationally Implemented HPSG Theories.* In proceedings of HPSG 2005. Center for the Study of Language and Information, Stanford, CA: CSLI Publications. 2005. 311-321.
3. Pollard Carl and Sag, I. A. *Head-Driven Phrase Structure Grammar.* CSLI Publications and University of Chicago Press. 1994.
4. Sag, I. A. and Wasow, T. *Syntactic Theory: A Formal Introduction.* Center for the Study of Language and Information, Stanford, CA: CSLI Publications. 1999.
5. Wang Lulu and Liu Haitao. *A Description of Chinese NPs using HPSG.* In Proceedings of HPSG 2007. Center for the Study of Language and Information, Stanford, CA: CSLI Publications. 2007. 287-305.
6. 化柏林. 语法开发平台，我们落后了. 中国计算机用户. 2004 (3).
7. 吴云芳. HPSG 理论简介. 当代语言学. 2003 (3)：231-242.
8. 朱德熙. 语法讲义. 北京：商务印书馆. 1982.