# 自然语言理解

第七章 语法理论

#### 宗成庆

中科院自动化研究所 模式识别国家重点实验室

cqzong@nlpr.ia.ac.cn

http://www.nlpr.ia.ac.cn/English/cip/cqzong.htm



Beijing 100080, China





### 第七章 语法理论

- □语法理论的几个 Landmarks
  - ❖ 20 世纪 50s Chomsky 提出了4类形式语法
  - ◆ 20 世纪 80s Chomsky 提出管辖与约束理论(GB)
  - ❖ 随后,研究者们提出了-词汇功能语法(LFG)
    - 广义结构语法(GPSG)
    - 树连接语法(TAG)
    - 链语法 (Link Grammar)等
- □ 出发点: (1)描写自然语言的结构、语义;
  - (2)揭示人类语言的共性和人类思维的奥秘。

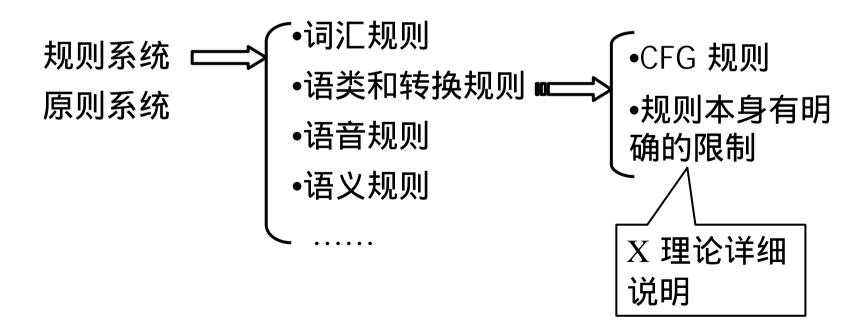






#### □ 语法构成

管辖约束理论 (Government and Binding theory, GB) 认为语法由两大系统构成:







□ 语法构成

GB 理论认为语法由两大系统构成:

规则系统 原则系统 **二二**>

- X 理论
- 格理论
- •管辖理论
- ?理论
- 约束理论
- 控制理论
- 界限理论





★ X 理论

以 CFG 为基本骨架。该理论认为:无论哪种词组,其结构均以中心语加上补语,或者中心语加上标志语方式构成。

形式上为:

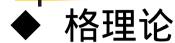
 $X^n \to \text{SPEC} \ X^{n-1}$ 

 $X^{n-1} \rightarrow X^{n-2}$  COMP

其中, SPEC 为标志语; COMP 为补语;  $X^i$  ( $i \in \{n, n-1, n-2\}$ )为中心语。







属于句法结构范围,它表示句中各成分之间的关系,以及 这种关系实现的形式或手段,确定在怎样的结构条件下和在句 子生成的哪个阶段,一些词可以赋予格,哪些词必须取得格。 格理论研究的是一种抽象的格。





格理论对词组移位后留下的语迹是否取得格有相应的规定。

如: (1)[那本书],张三看完了。

(2) 张三看完了[那本书]。

都合法且语义相同

相同的结构

(3)[张三[看完了[那本书]]]。

词组移位

(4) [[那本书], [张三看完了]]。





◆ 管辖理论

从结构上定义哪些成分可以管辖,哪些成分可以赋格给另外一些成分。

可以赋格的成分称之为主管成分,如:动词,介词等;

受其管辖的成分称之为<u>受管成分</u>,如:名词,形容词等;

这种关系称之为管辖关系。

管辖理论研究成为主管、受管成分的条件;研究代词、空语类以及它们与先行词之间的管辖关系怎样。





# NLPR

#### 7.1 GB 理论



#### - 理论

属于语义范畴,研究词语所处的句法位置与它们在句子描写的动作中所起的作用关系。它试图揭示句中陈述的行为、各个行为参与者的性质和相互关系,以及这些参与者在句子位置中应受何种约束。

参与者的角色包括:施事者、受事者或主题、行为的地点、 主题的来源、目的地,等。<u>这些角色通称为题元</u>( -role)。





约束理论

研究名词词组之间的指称关系,在怎样的结构范围(辖域) 内,代词(如:你、我、他、它等)、反身代词(如:自己)、 指称语、空语类等,可以受到先行词的约束,与先行词同指。

如:水边草地上,一头生啃着嫩草,放牛娃背对着它在吹

笛子。





◆ 控制理论

研究的对象是空语类 PRO,研究它怎样受到先行词的控制。PRO具有代词的性质,但必须有一个先行词。如:

老张叫儿子去买点酒来。

老张叫儿子 [去买点酒来]。

老张叫<u>儿子</u> [PRO 去买点酒来]。







◆ 界限理论

研究在一些位移情况下,如何用界限结点(如 S 结点、NP 结点)来对位移转换加以限制,使得语迹与迁走成分之间间隔不超过一定的界限。





- ◆ 管辖理论、格理论讨论管辖范围、管辖对象;
- ◆ 格理论、 理论则在结构和语义两个层次上讨论同一个 对象;
- ◆ 约束理论和控制理论都是研究名词性词组之间的同指问题。
- ◆ 几种理论相互补充、相互交互、相互限制,以解释各种语言现象。

详细内容请参阅 徐烈炯著《生成语法理论》, 上海外语教育出版社,1988。





功能合一文法(Function Unification Grammar, FUG) 是1985年 Martin Kay (Stanford University)提出的。

#### □ 提出起因

- ◆ Chomsky 短语结构语法生成能力太强,产生许多不符合语法或有歧义的句子;
- ◆标记十分简单,分析能力有限,难以反映自然语言的复杂特性。





- □ FUG 对短语结构语法的改进
  - ◆ 采用复杂特征集来描述词、句法规则、语义信息,以 及句子的结构功能。

试图以单一形式的结构模式来描述特征组合、功能分配、词条和组成成分的顺序,以达到对句子的完全功能描述。

◆ 采用合一运算对复杂特征集进行运算





#### □复杂特征集

◆ 复杂特征集功能描述的定义

设 α 为一个功能描述 FD (Functional Description), 当且 仅当 α 可以表示为:

$$\begin{pmatrix}
f_1 = v_1 \\
f_2 = v_2 \\
\dots \\
f_n = v_n
\end{pmatrix}$$
 $n \ge 1$ 

其中, $f_i$ 表示特征名, $v_i$ 表示特征值,且满足以下条件:





(1)特征名 $f_i$ 为原子,特征值 $v_i$ 为原子或另一个功能描述;

(2) 
$$\alpha(f_i) = v_i$$
  $(i = 1, ..., n)$  , 读作:复杂特征集  $\alpha$  中,特征  $f_i$  的值等于 $v_i$ 。

◆ 如何用复杂特征集描述词汇

在词典中单词的特征可以包括词类、形态、句法和语义等

多方面的信息,如:

Cat = Verb

Tense = Past

Transitivity = Mental-Process

Root = see

Lex = saw





◆ 如何用复杂特征集描述规则

$$S \rightarrow NP + Verb$$

$$Cat = S$$

$$Subject = [Cat = NP]$$

$$Cat = Verb$$

$$Predicator = \begin{cases} Cat = Verb \\ Number =  \\ Person =  \\ \end{bmatrix}$$





如何用复杂特征集描述句子: We helped her.

$$Cat = S$$

$$Subject = \begin{cases} Cat = Pron \\ Number = Plur \\ Person = Third \\ Lex = we \end{cases}$$

$$Objective = \begin{cases} Cat = Pron \\ Number = Sing \\ Person = Third \\ Lex = her \end{cases}$$

$$Predicator = \begin{cases} Cat = Verb \\ Lex = help \end{cases}$$

$$Tense = Past$$

$$Voice = Active$$







- ◆ 复杂特征集的特点
- (1)允许利用多个语言特征描述同一个语言单位;
- (2)从结构上看,复杂特征集是一种嵌套结构,可以有效地表示复杂词组或句子结构;
- (3)特征名的定义及其相互关系具有明显的层次性,而所有自然语言的结构都是层次性的,复杂特征集的这一特点显然对语言的层次分析有益;
- (4)复杂特征集便于运算,两个复杂特征集通过合一运算可以产生另一个复杂特征集,这与句法分析中词组和句子的产生是一致的。





- □ 合一运算
- ◆ 复杂特征集相容的定义

若 $\alpha$ 、 $\beta$ 均为复杂特征集,则 $\alpha$ 、 $\beta$ 是相容的,当且仅当:

- (1)如果  $\alpha(f) = a$ 、 $\beta(f) = b$ ,且a、b都是原子,那么 $\alpha$ 、 β是相容的,当且仅当a = b;
- (2)如果 $\alpha(f)$ 、 $\beta(f)$ 均为复杂特征集, $\alpha$ 、β是相容的, 当且仅当 $\alpha(f)$ 、 $\beta(f)$ 相容。





- ◆合一运算的递归定义
- (1) 在 a、b都是原子的情况下,如果 a = b,那么  $a \cup b = a$ . 否则  $a \cup b = \emptyset$ ;
  - (2) 如果 $\alpha$ 、 $\beta$  均为复杂特征集,则
- (a) 若  $\alpha(f) = v$ , 但  $\beta(f)$  的值未经定义,则 f = v 属于  $\alpha \cup \beta$ ;
  - (b) 若  $\beta(f) = v$ , 但  $\alpha(f)$  的值未经定义,则 f = v 属于
- $\alpha \underline{U} \beta$ ;
- (c) 若  $\alpha(f) = v_1$ , 但  $\beta(f) = v_2$ , 且 $v_1$ 与  $v_2$ 相容(不相抵触),则  $f = (v_1 \cup v_2)$ 属于  $\alpha \cup \beta$ , 否则  $\alpha \cup \beta = \emptyset$ 。





- ◆合一运算的作用
  - (1) 合并原有的特征信息,构造新的特征结构;
- (2)检查特征的相容性和规则执行的前提条件是否满足,如果参与合一的特征相冲突,就立即宣布合一失败。





$$\alpha$$
:  $Cat = N$ 

$$Lex = 1/3 \pm$$

$$Person = Third$$

$$\alpha \underline{U} \beta = ?$$

$$\alpha \underline{\mathbf{U}} \beta = ?$$

$$\beta$$
 Cat = N

Number = Sing

Semantic = Agent

Cat = N
$$Lex = 小王$$

$$\alpha \, \underline{U} \, \beta =$$
Number = Sing
$$Semantic = Agent$$

$$Person = Third$$



#### ◆ 例 2:

$$\alpha$$
:  $Cat = N$ 

$$Lex = 1/3$$

$$Semantic = Agent$$

$$\alpha \underline{U} \beta = ?$$

$$\beta$$
 Cat = N

Number = Sing

Semantic = Patient





如果把自然语言看作是一个传递和负载信息的系统,并且 承认自然语言中的句法成分都可以由较小的单位合并成较 大的成分,那么,合一作为句法和语义分析的基本运算便 是非常合理的了,因为:

- (1) 一个语言单位(如句子或词组等)所负载的信息可以分布在各个成分中,每个成分所负载的可以只是部分的信息;
- (2) 通过合一运算,在小的成分组成大的成分的过程中,小成分所负载的信息也可以同时被传递或累加为大的成分所负载的信息,在合一运算过程中,信息只会增加而不会减少;





- (3) 句法和语义分析都以合一作为基本运算,不仅句子的合法性可以通过语义的手段来判断,而且,还可以把句子的句法结构和语义表示用合一运算这种方式更加自然地衔接起来;
- (4) 对不同的复杂特征集进行合一运算,其结果同运算所进行的先后次序无关,合一的结果都是相同的。合一运算的这种无序性非常便于并行处理。

复杂特征集和合一运算的方法,是现代计算语言学研究的主流方法之一。



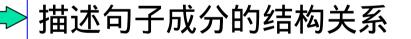


词汇功能语法 (Lexical Functional Grammar, LFG) 于上个世纪70年代末期由 R. Kaplan and J. Bresnan 在美国 MIT 提出。

#### □基本观点

句子由两个相对独立的层次来描述:

- 1) 成分结构层次
- 2) 功能结构层次



描述句子主语、谓语、宾语等 之间的关系





#### 口 要点

1)突出词汇的作用,体现"语法结构可以由某些词的意义预示出来"

LFG 认为,动词、形容词和一部分名词在句子中的语法结构作用,相当于数理逻辑中的谓词逻辑(predicate),它们的词义决定它们拥有哪些论元(argument)。即如果知道了谓词的意义,那么就可以知道以该谓词为中心的句子还会有哪类词会出现,它们的语法语义功能是什么。

- (a) 可以准确地解释语言现象:谓词的管辖范围 + 谓词对论元的预示→ 确定语法结构和语义解释
  - (b) 可以减轻语法规则的作用





#### 口 要点

2)把功能结构的描述作为语言描述中一个基本的独立层次 LFG 中用到的功能主要指语法功能(如主语、宾语、补语、修饰语,与传统的主、宾语概念一致);时态、数、人称、格等语法特征;谓词功能。

以功能为基础,定义句子的合格条件作为对成分结构的制约。

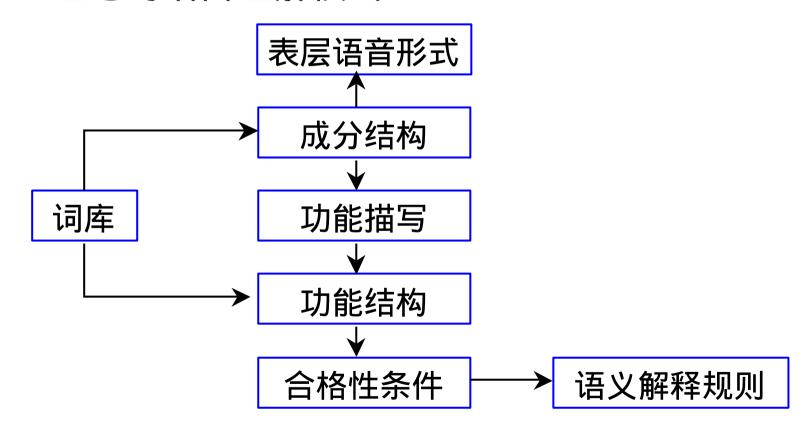
有成分结构的句子不一定是合乎语法的句子,只有存在合法功能的句子,才是合乎语法的句子。

LFG本质上是一种以功能为基点的文法。





□ LFG 理论的语言理解模式







#### □ LFG 的两个语法层次结构

◆ 成分结构 (Constitute Structure, c-结构)

用上下无关文法表示;树上的结点带有句子中词或短语 所预示的功能信息,这些信息由语法规则右部的符号所带 的功能注释表示,例如:

(1) S 
$$\rightarrow$$
 NP VP  
 $(\uparrow SUBJ) = \downarrow$   $\uparrow = \downarrow$ 

" $\uparrow$ "和" $\downarrow$ " 称为直接支配元变量(immediate domination meta-variable)。" $\uparrow$ "表示规则的左部符号;" $\uparrow$ SUBJ"表示 S 的主语;" $\downarrow$ "表示带有该注释的符号本身。





规则(1)表示:

句子 S 由 NP 和 VP 组成,其中,NP 所带的全部功能就是 S 的主语功能信息;VP 所带的全部功能信息就是 S 的功能信息。

(2) NP  $\rightarrow$  DET N

表示 NP 由限定词和名词组成。





(3) VP 
$$\rightarrow$$
 V  $(\uparrow OBJ2) = \downarrow$   $(\uparrow OBJ) = \downarrow$ 

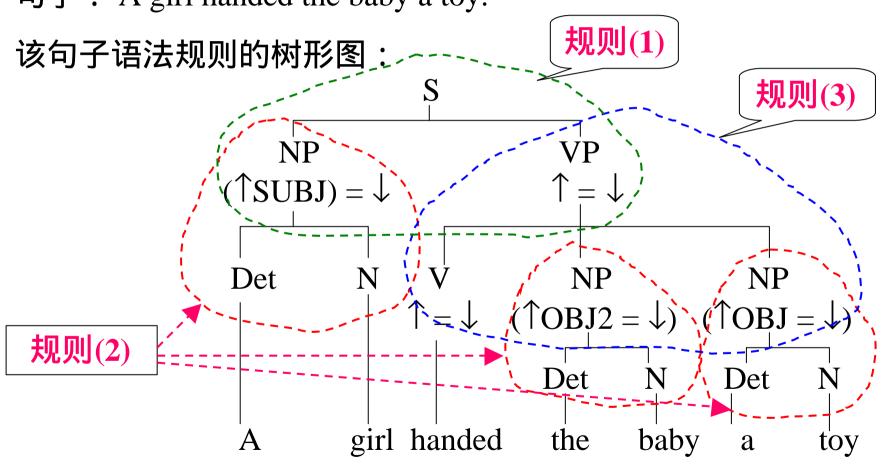
NP 外的括号表示 NP 是可选的。

该规则表示动词所带的全部功能信息就是 VP 的功能信息; VP 可由一个动词(不及物动词)组成;或者由一个动词和一个 NP(及物动词带单宾语),该 NP 的全部功能信息是 VP 的宾语的功能信息;或者有另外一个 NP 参加(及物动词带双宾语),该 NP 的全部功能信息是 VP 的第二宾语的功能信息。





句子: A girl handed the baby a toy.







以上为 LFG 的第一套规则,即<u>句法规则</u>。

<u>词法规则</u>由词典信息提供。词法规则在 LFG 中占有重要地位,它带有语法功能结构的预示信息。例如:

a: DET, 
$$(\uparrow SPEC) = A$$

$$(\uparrow NUM) = SG$$

girl: N, 
$$(\uparrow NUM) = SG$$

$$(\uparrow LEX) = 'GIRL'$$

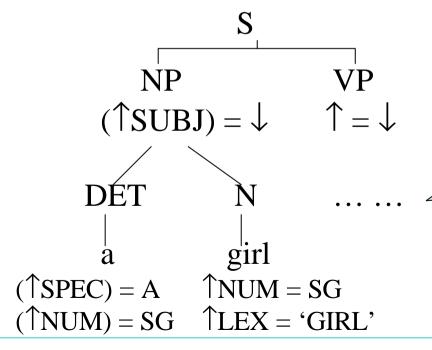
"(<sup>↑</sup>NUM) = SG"表示"我的父结点具有的功能NUM(数)其值为SG(单数)"。





LFG 把词汇按词的不同意义立项,词汇项所含的信息具有 语法范畴和功能注释。功能注释的形式与语法规则的功能 注释完全一致。如: A girl handed the baby a toy.

其c-结构图为:



参阅:[1]姚天顺等, 自然语言理解,清华 出版社。P.191.

[2] 刘颖,计算语言学 清华出版社. P.73.





◆ 功能结构 (Functional Structure, f-结构)

用以表示句子的功能关系。LFG 认为,功能结构是语言学描述上的一个独立层次,功能结构中包含语法信息,也包含语义信息,但它又不同于成分结构或语义解释。

f-结构是成分结构和语义结构的中介面,既表示语法范畴之间的功能控制关系,又是建立语义结构的基础。

上面例句的f-结构为:





```
SUBJ
          Spec
          Number
                         Sing
                         girl
          Lex
TENSE
         Past
         hand (\uparrowSUBJ) (\uparrowOBJ2) (\uparrowOBJ)
PRED
OBJ2
          Spec
                         the
          Number
                         Sing
                         baby
          Lex
OBJ
          Spec
                         a
          Number
                         Sing
          Lex
                         toy
```





- ◆ f-结构的形式定义
- (1) f-结构是有序对的集合,每个有序对含有一个属性和该属性的值;
  - (2) 语法功能的名字或特征名字是属性;
  - (3)f-结构中的属性值有以下4类:
    - (a)简单符号,如:girl, past, Sing 等;
    - (b) 语义形式,如:hand(↑SUBJ)(↑OBJ2)(↑OBJ)
    - (c)子f-结构
    - (d)上述三类属性的集合。





◆ 由c-结构构造f-结构

LFG利用c-结构提供的短语结构信息和功能信息,按照严格的程序和明确的合格条件构造 f-结构。

Step-1: 从c-结构求出功能描述式 (functional descriptions, 简称f-描述);

Step-2: 从f-描述构造f-结构。







◆ f-描述

f-描述是一系列等式的集合,每个等式的形式与c-结构中各结点所带的功能注释类似,区别仅在于把直接支配元变量<sup>↑</sup>和↓换成相应的结点代号(实际变量) - 代真(instantiation)。实际上是将c-结构中短语结构信息和功能信息合而为一。

代真前的注释只能表示树中各结点的孤立的功能信息,代真后的f-描述则表示了结点之间的功能关系。





◆ 代真过程

Step-1. 把词汇项的功能注释移交给父结点;

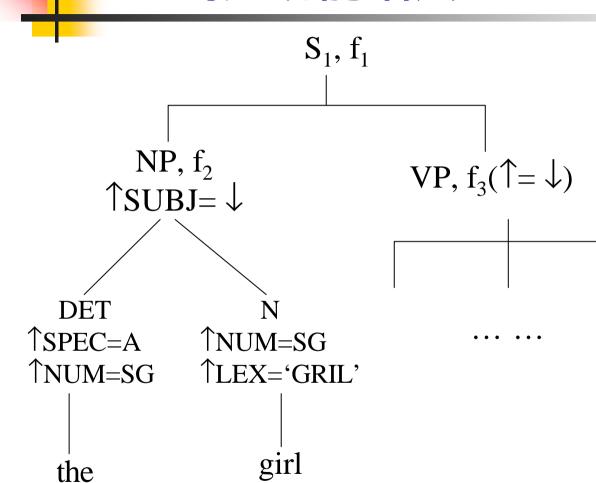
Step-2. 把实际变量  $f_n$  标注在树的非终结结点;

Step-3. 把 ↑ 和  $\downarrow$  换成实际变量  $f_n$  。

例如,前面的c-结构图经过代真后成为:







f-描述:

- (1)  $(f_1 SUBJ) = f_2$
- $(2) f_1 = f_3$
- $(3) (f_2 SPEC) = A$

• • • • • •





◆ 代真过程的合法性

代真过程的合法性由 LFG 中最重要的惟一性条件 (uniqueness)保证,这个条件同时也是从 f-描述构造f-结构的 主要合格条件。

惟一性条件:在一个给定的f-结构中,一个属性最多只能有一个值。





◆ 从f-描述构造f-结构

算法输入:f-描述中的所有等式,等式顺序没有限制;

算法输出:f-结构

操作:定位(locate)

合并 (merge)

一致性检验:如不符合f-结构定义,定位失败;

如合并时不兼容,合并操作失败。





□ 功能合格条件

由惟一性条件得到的f-结构的句子还不一定合法,LFG还要进一步用一系列功能合格条件(functional well-formedness conditions)检验这个f-结构的合法性,即给定的c-结构中,句子成分的功能必须独一无二。







#### 另外两个条件:

- 完全性条件(completeness):一个f-结构是局部完全的,当且仅当它包含了它的谓词所应支配的全部语法功能;一个f-结构是完全的,当且仅当它所有的子f-结构都是局部完全的。
- 关联性条件(coherence):一个f-结构是局部关联的,当且仅当它所包含的全部被支配的语法功能都能被一个局部谓词支配;一个f-结构是关联的,当且仅当它所有的子f-结构都是局部关联的。







## 口 词汇功能语法的特点

- ◆ 采用复杂特征集表达功能结构
- ◆ 以合一运算作为句法-语义分析过程的基本方式
- ◆ 语法信息主要来源于词典中对词汇信息的标注
- ◆ 功能结构是无序的





#### □ 概要

- ●广义的短语结构语法(Generalized Phrase Structure Grammar, GPSG)是一种语境自由语法,创始于20世纪70年代。
- ●代表任务:英国语言学家 Gerald Gazdar, Ivan Sag, Ewan Klein 和美国语言学家 Geoffry Pullum 等。
- ●代表著作: Gerald Gazdar 发表于1982年的《短语结构语法》







#### □ 基本主张

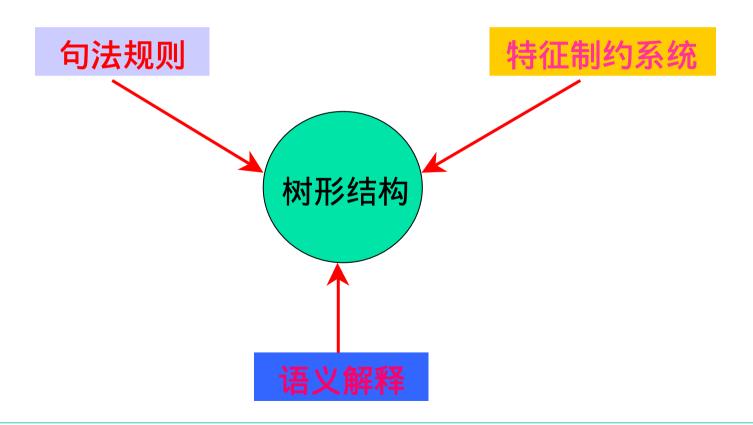
GPSG 的规则系统要经过一系列的合格性条件检查,才能跟句子的表层结构联系起来,每一条规则只产生一个候选的局部树形结构,这个树形结构是否被接受,要经过一系列合格性条件检验。

将普通的短语结构语法中的单纯推导过程变成了一步一步 的检验过程,通过检验把不合法的句法结构排除出去。





□ GPSG 的三个组成部分

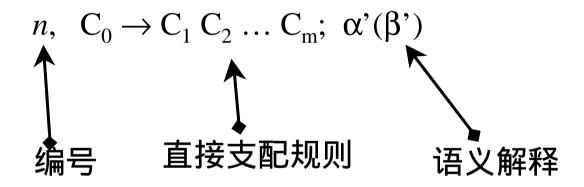






#### □ 句法规则

#### 一般形式:



#### 直接支配规则例子:

- 1) VP → H, NP 对应类似短语:close the window
- 2) NP → NP[+poss], H 对应类似短语: John's new book

7



□ 特征制约系统

句法特征:对句法规则得出的树形结构加以限制

- (1) 主特征(head feature):主语、助动词、谓语等
- (2)次特征(foot feature):描写疑问代词和关系代词的特征、反身代词和相互代词的特征等
  - (3)一般特征(normal feature):动名词、否定等
  - □ 语义解释系统

采用内涵逻辑的方法。





#### □ 概要

树连接语法(Tree Adjoining Grammar, TAG)1975年由美国学者 Joshi 提出。

提出的动机是试图弥补上下无关文法过于简单、不能限制 一些不合法语言现象的弱点。

TAG 是介于上下文无关和上下文有关文法之间的一种语法表示形式。





56

## 7.5 树连接语法

#### TAG 的形式化定义

```
G = \langle V_N, V_T, S, T_0, T_a \rangle
```

V<sub>N</sub>: 非终结符集合;

 $V_T$ : 终结符集合;

S: 起始符;

T<sub>0</sub>: 初始树集 (initial trees);

Ta: 附属树集 (auxiliary trees)。

基本树 ( elementary tree)的有限集

2005-4-19 NLPR, CASIA

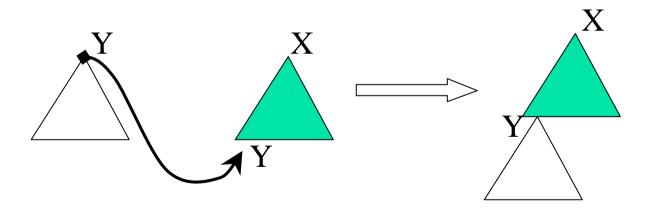




#### □ TAG 中的两种操作

◆ 替换 (substitution)

把一个树的根结点与另一个初始树中标为替换的非终结符叶结点合并,生成一棵新的树。



根结点和替换结点必须具有相同的标识。

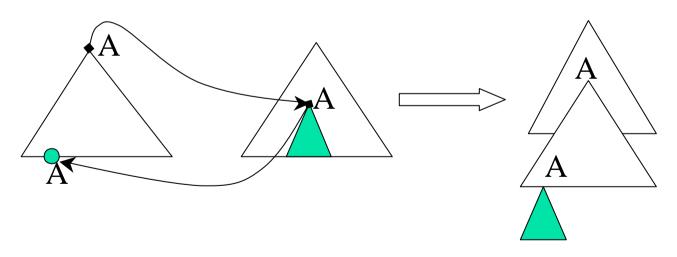




#### □ TAG 中的两种操作

◆ 附加 (adjunction)

把一个附属树嫁接(graft)到一个初始树的任意非终结结点上。









#### □ TAG 与短语结构语法的区别

- ◆ 短语结构语法试图定义什么是合法的成分,用合法的成分组装合法的句子的结构树。而TAG则直接定义合法的树, 句子的结构树是合法的树组装出来的。
- ◆ TAG一开始就是"立体的",最后形成句子的时候才"展平"。而短语结构语法树一开始是"扁平的",推导过程隐含了一个立体的树结构。





## 本章小结

- □ GB 理论:2个组成系统、7个基本理论
- □ 功能合一语法 (FUG)
  - > 对短语结构语法的改进
  - > 复杂特征集
  - > 合一运算
- □ 词汇功能语法 (LFG)
  - ▶ 基本观点、特点
  - ➤ 两个语法层次结构 ( c-结构、f-结构 )
  - ➤ 由c-结构经f-描述构造f-结构
  - > 功能合格条件





## 本章小结

- 口广义的短语结构语法 (GPSG)
  - > 三个组成部分:句法规则、特征制约系统、语义解释
- □ 树连接语法 (TAG)
  - ➤ TAG 的形式化定义
  - ➤ TAG 的操作
  - ➤ TAG 与短语结构语法的区别







- 1、阅读有关 GB 理论、FUG、LFG、GPSG、TAG的资料,了解有关这些语法理论的详细内容。
- 2、阅读有关转换生成语法、中心驱动的短语结构语法、依存文法和链语法的资料,了解这些语法的主要思想。





# Thanks 谢谢!

