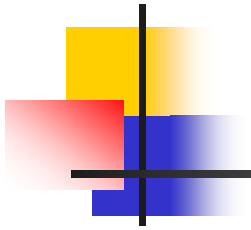
第四讲 汉语的句法结构分析（II）

詹卫东

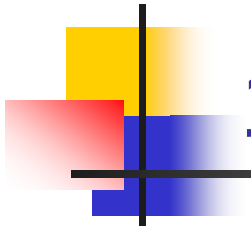
http://ccl.pku.edu.cn/alcourse/nlp



提纲

1. 句法结构分析中的歧义问题
   - 自然语言歧义现象举例
   - 句法结构歧义的不同类型
2. 增加合一约束的句法结构分析
   - 特征结构与合一运算
   - 增加合一约束的Earley算法
3. 小结

附录：计算机句法结构分析歧义的程度



1 句法结构分析中的歧义问题

英语结构分析中常见的三类结构歧义

- Attachment ambiguity
- Coordination ambiguity
- Noun-phrase bracketing ambiguity

Jurafsky & Martin(2000) Speech and Language Processing, Prentice-Hall, Inc. Chapter 10.3

歧义示例

▪ pp-attachment ▪ gerundive-vp attachment ▪ np-attachment	1. I shot an elephant in my pajamas. 2. We saw the Eiffel Tower flying to Paris. 3. Can you book TWA flights?
▪ Coordination ambiguity	1. old men and women 2. John or Tom and Dick
▪ Noun-phrase bracketing ambiguity	1. complete peace plan 2. dead poets' society

不同语言层面的歧义

- 结构层次歧义（bracketing ambiguity）

- 结构关系歧义（syntactic relation ambiguity）

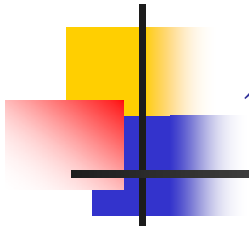
  出租汽车　　牛奶面包

- 语义关系歧义（semantic relation ambiguity）

  张三 谁 都 不 认识　　张三 的 笑话 说 不 完

- 语用歧义（pragmatic ambiguity）

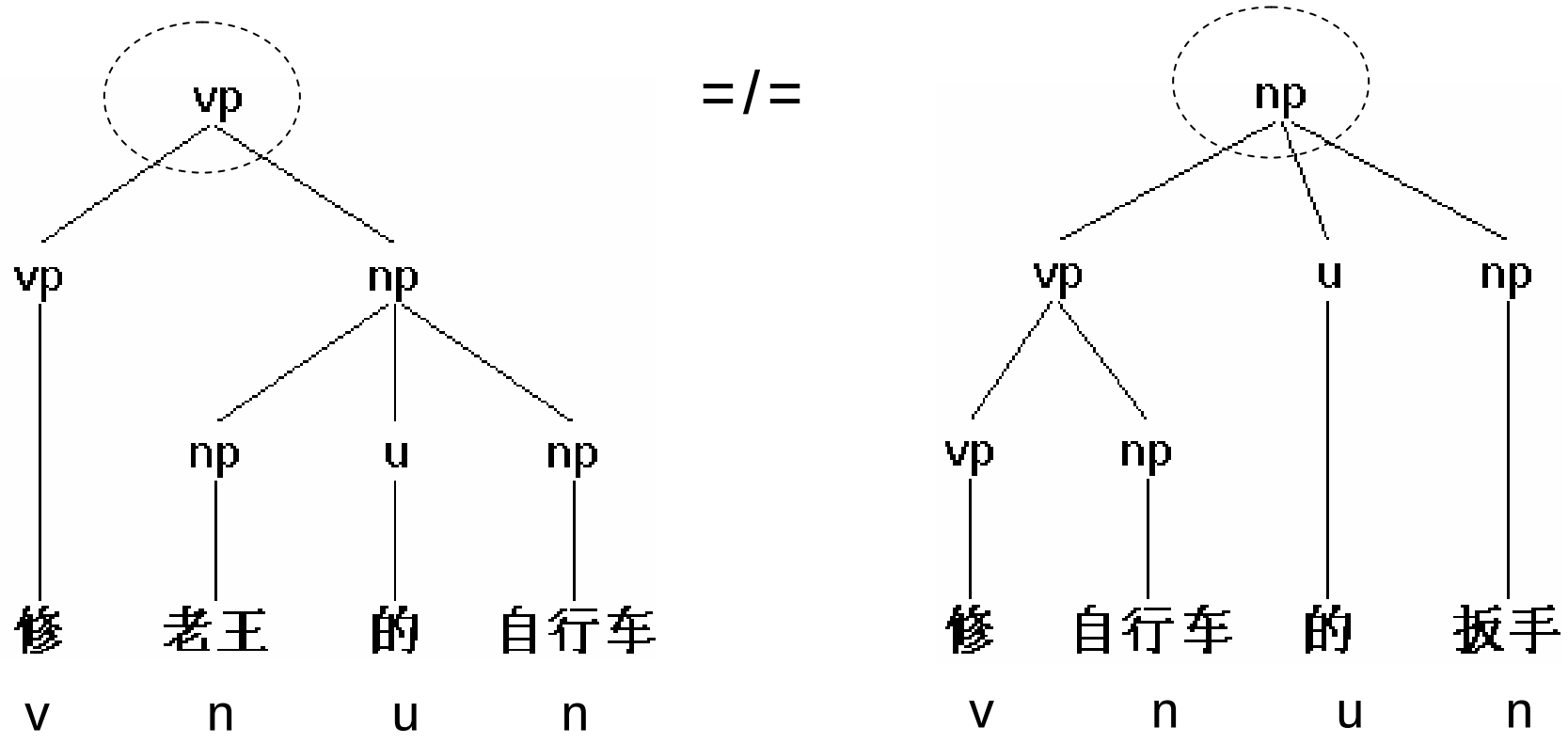
  张三 跟 李四 真 是 没 话 说



句法结构歧义的不同类型

- 显性歧义与隐性歧义
  - 歧义格式对环境敏感 vs. 歧义格式对环境不敏感
  - 句子层（终端）歧义 vs. 结构层（模式）歧义

外显型歧义



外显型歧义（续1）

咬死了　猎人　的　狗

发现了　敌人　的　哨兵

怀疑　张三　的　老师

骑了　三年　的　自行车

没有　买票　的　人

支持　罢课　的　学生

擦洗　干净　的　桌子

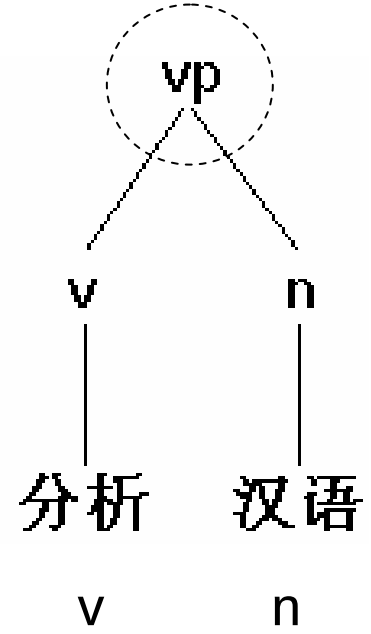
……

v　*　u　n

vp | np

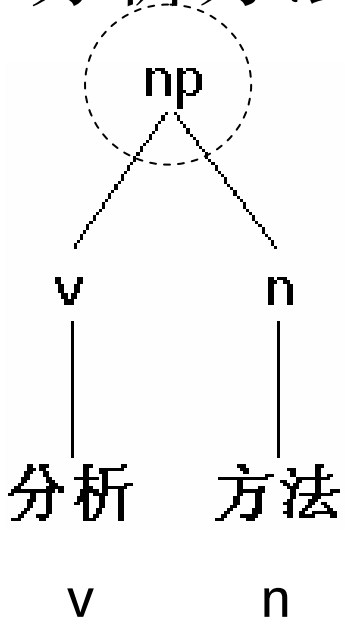
外显型歧义（续2）

分析汉语



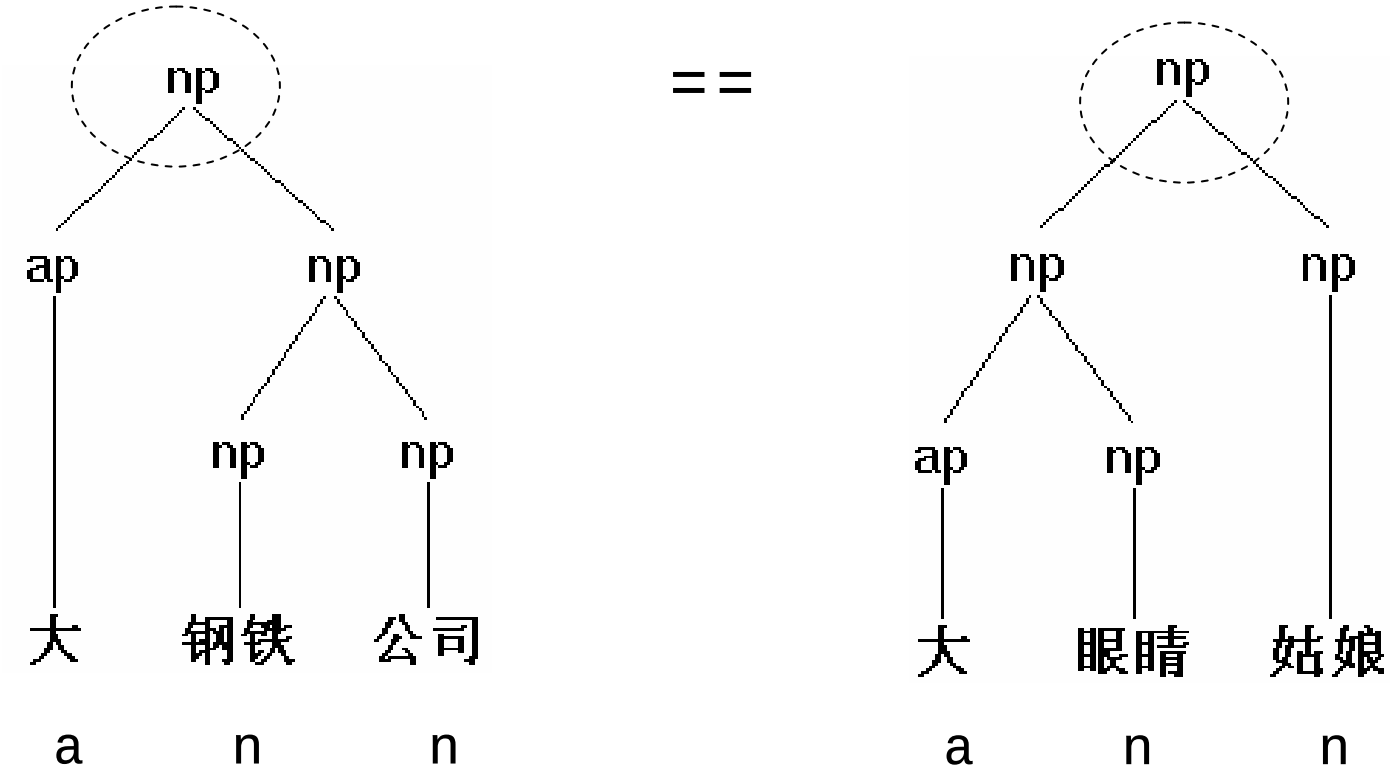
=/=

分析方法



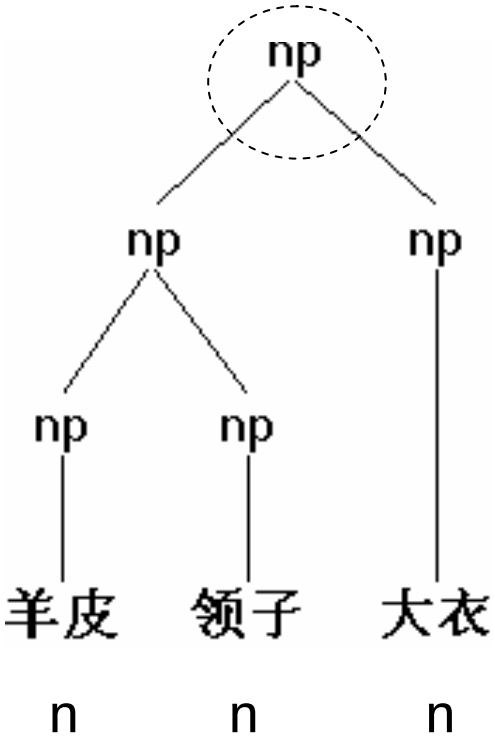
出租汽车　np | vp

内含型歧义



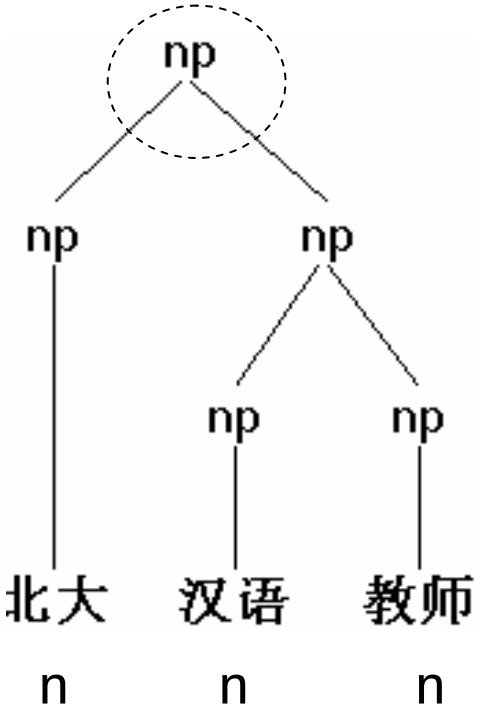
内含型歧义（续1）

羊皮领子大衣



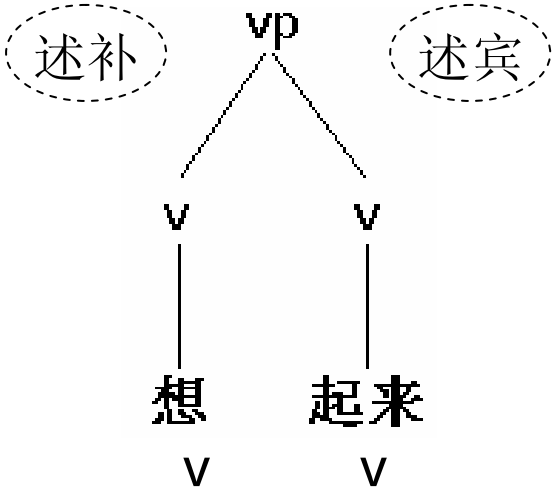
==

北大汉语教师



内含型歧义（续2）

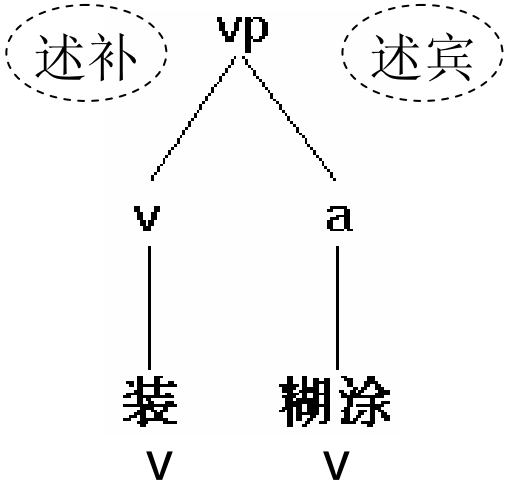
想起来



我终于想起来那天发生的事情了

奶奶躺了一整天，现在想起来了。

装糊涂

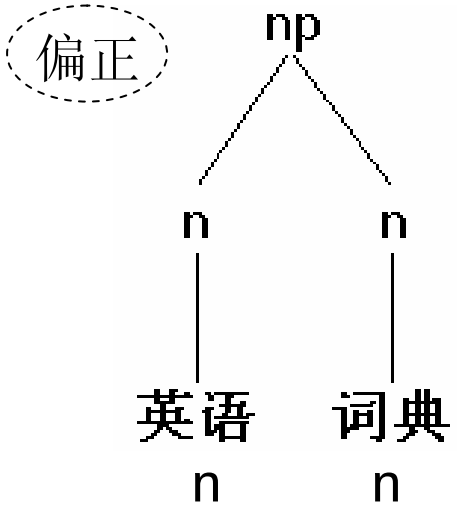


他就会装糊涂，其实他心理比谁都清楚

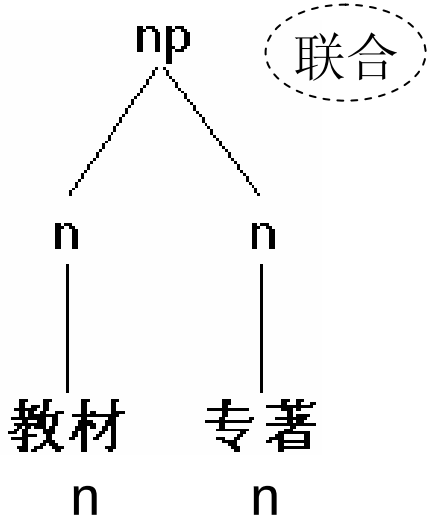
装了一上午家具，我都装糊涂了

内含型歧义（续3）

英语词典

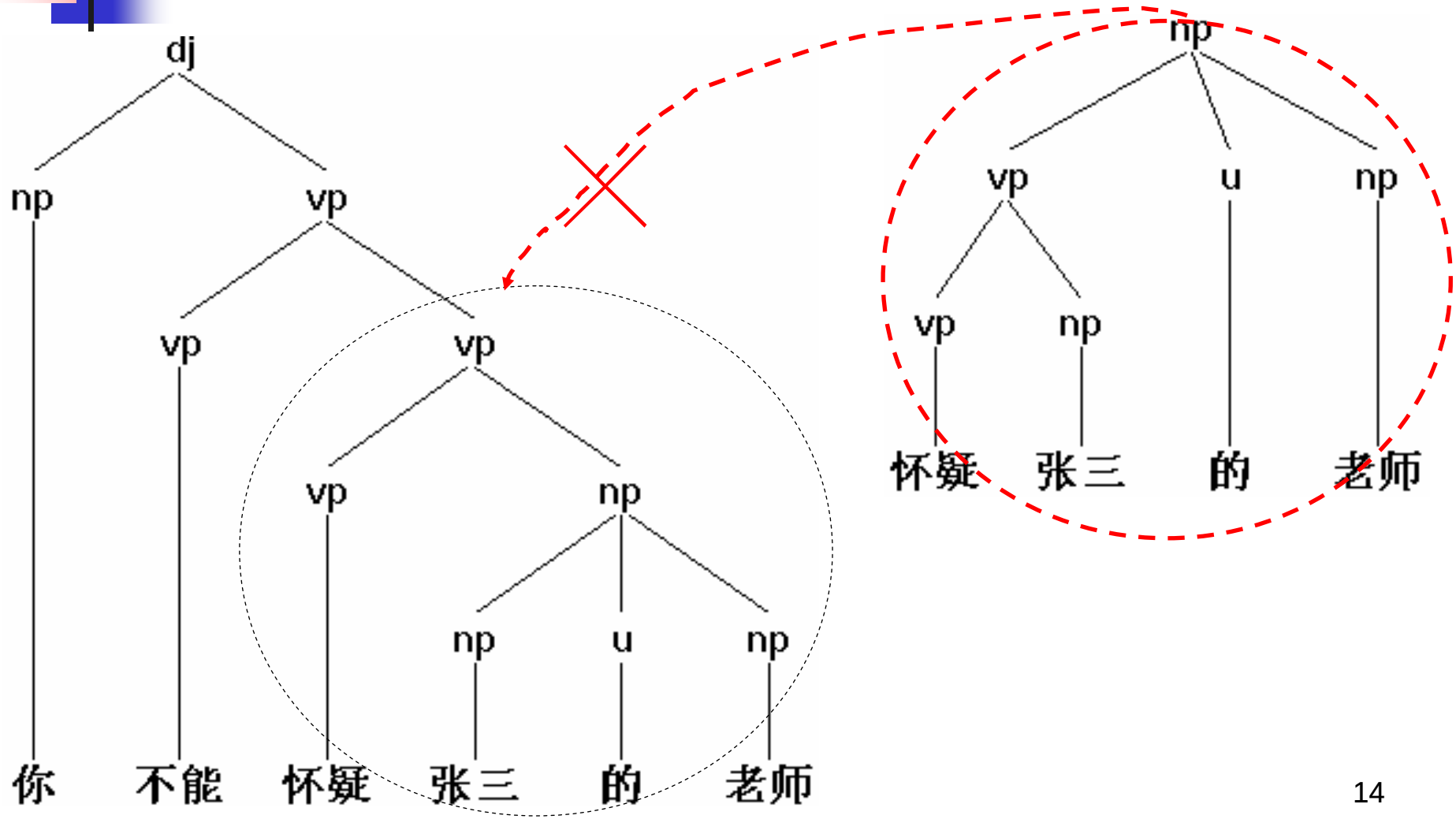


教材专著

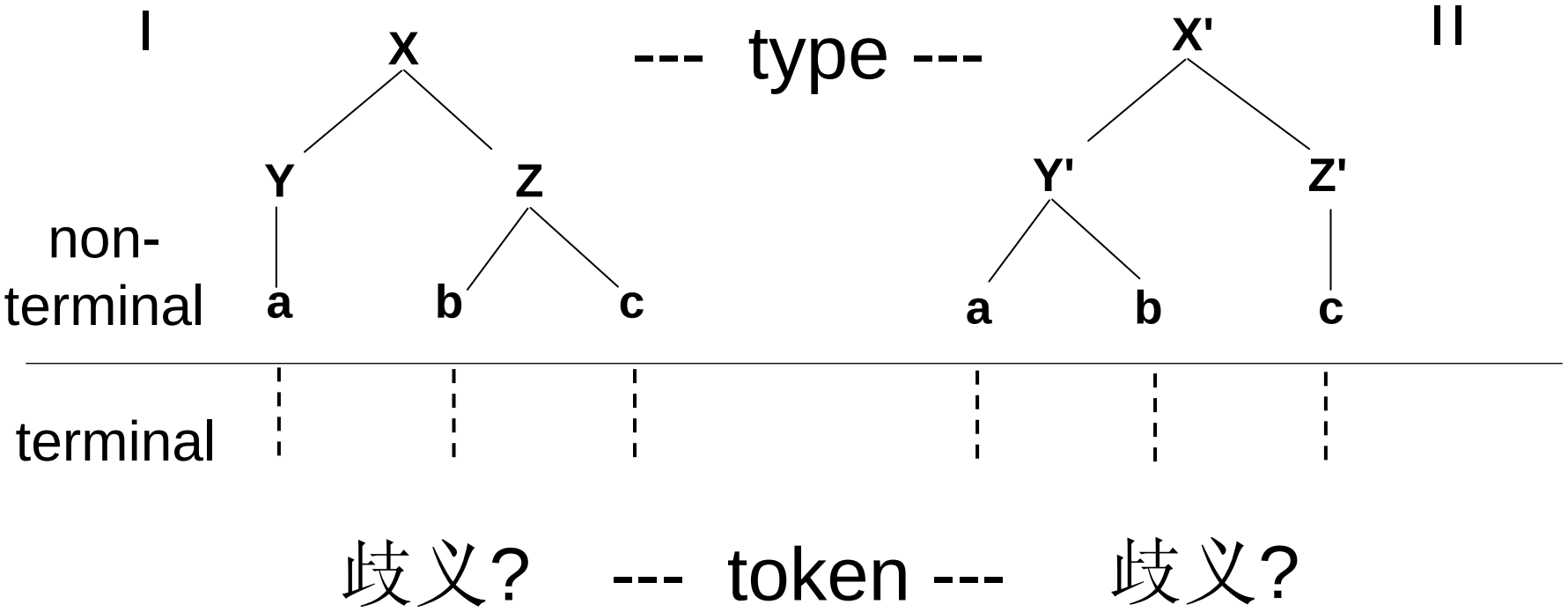


偏正？ 牛奶饼干 联合？

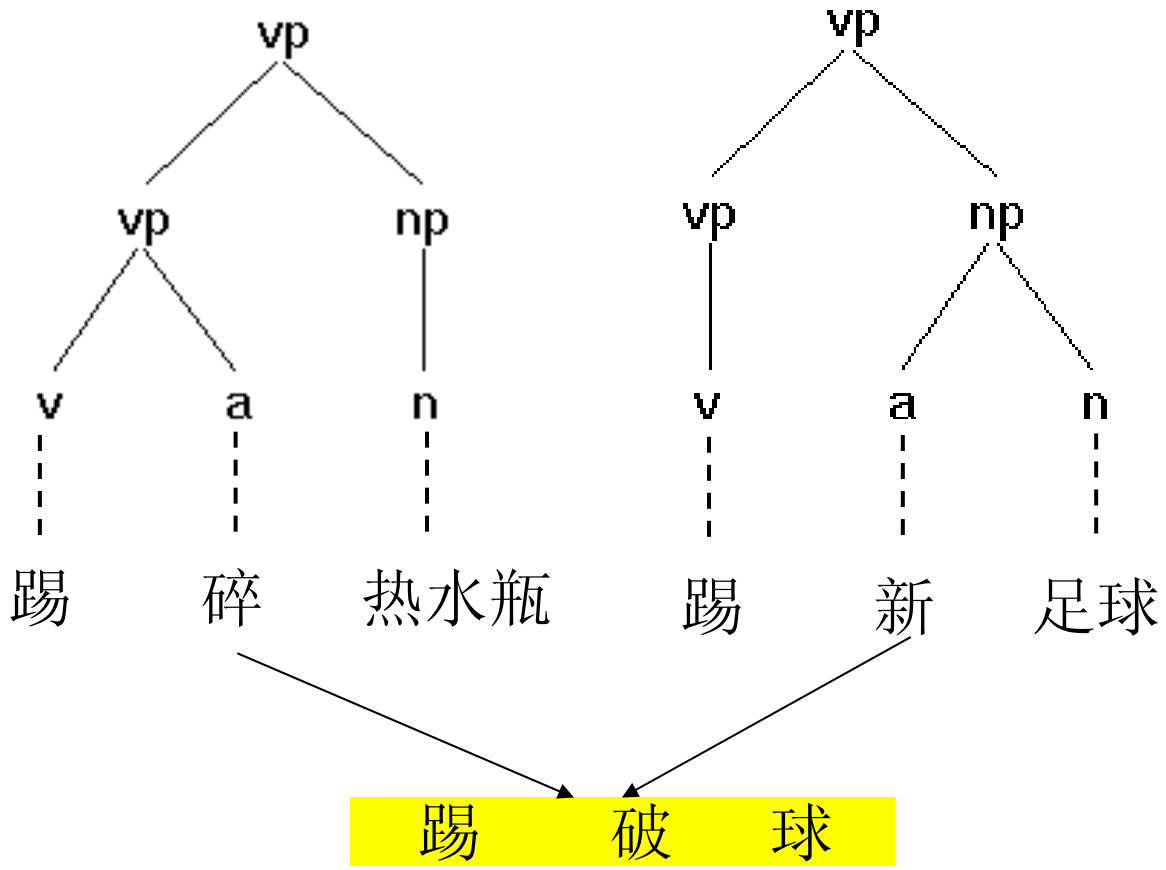
区分“外显”与“内含”的作用



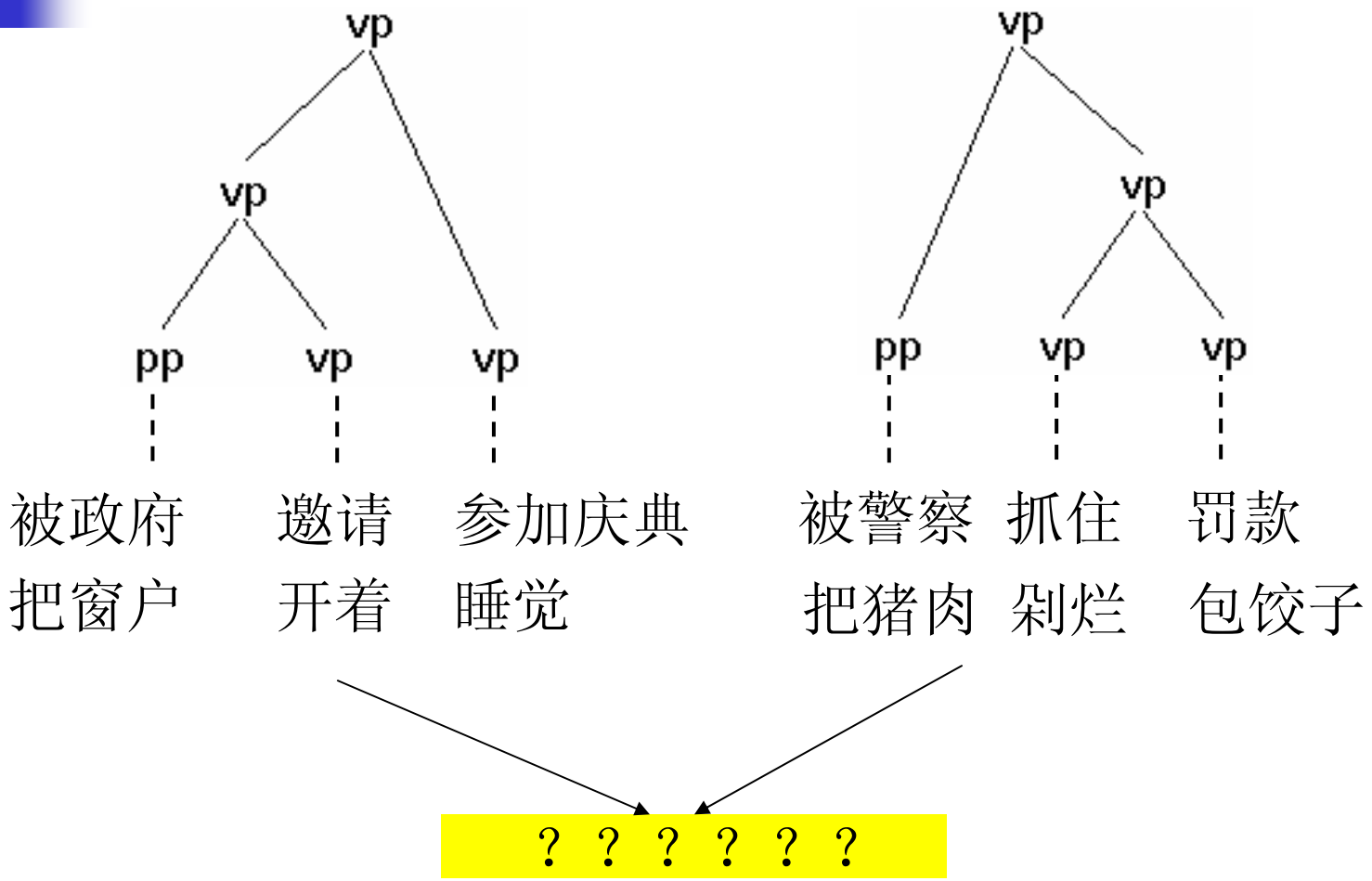
真歧义 准歧义 伪歧义



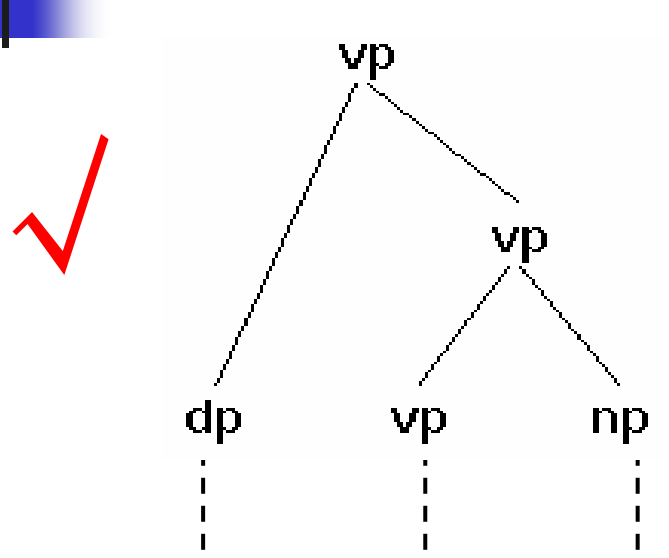
真歧义

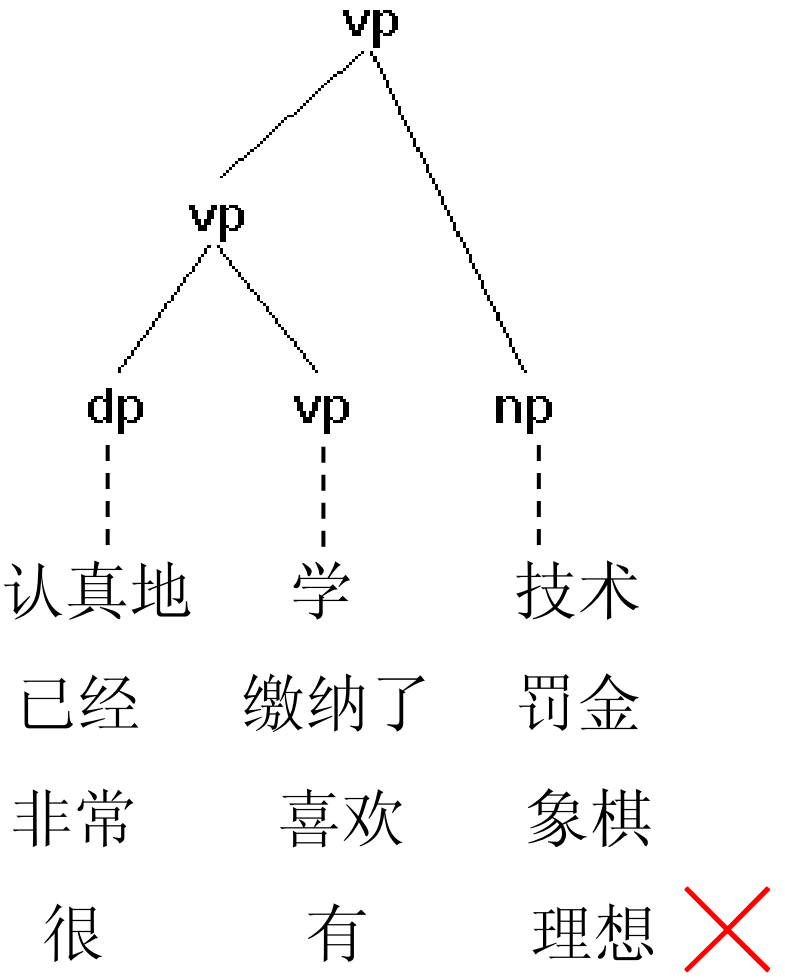


准歧义


vp
vp
pp
vp
vp
被政府
把窗户
邀请
开着
参加庆典
睡觉
vp
vp
pp
vp
vp
被警察
把猪肉
抓住
剁烂
罚款
包饺子
？ ？ ？ ？ ？ ？

伪歧义


vp
vp
dp
vp
np
认真地
已经
非常
很
学
缴纳了
喜欢
有
技术
罚金
象棋
理想


vp
vp
dp
vp
np
认真地
已经
非常
很
学
缴纳了
喜欢
有
技术
罚金
象棋
理想

区分“真/准/伪”歧义的作用

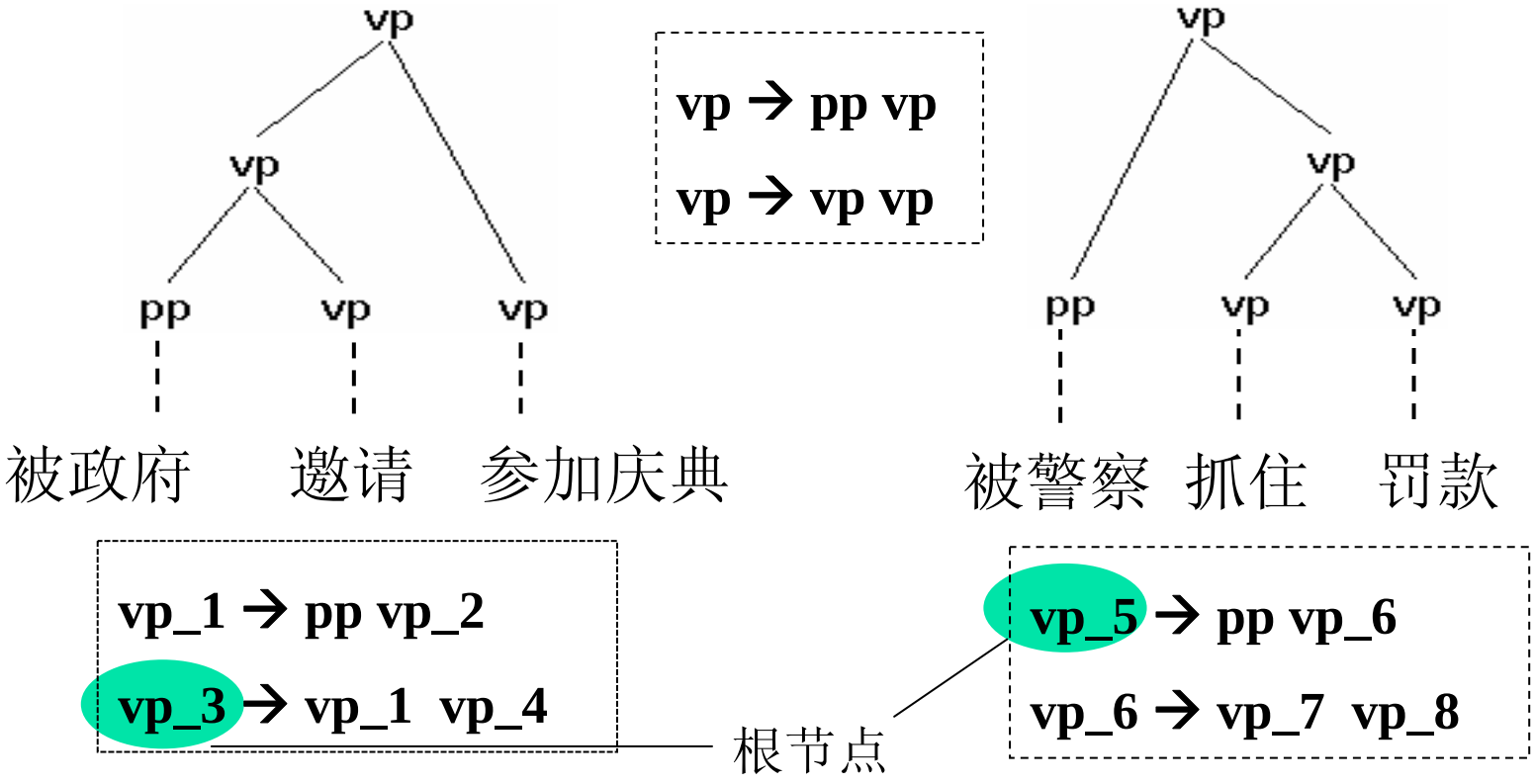
- 计算机针对不同类型的短语结构歧义，可用不同的策略

伪歧义	可通过安排规则的使用顺序来消岐
准歧义	可通过区分**单个语言成分**的不同特征消岐
真歧义	需要描述**两个语言成分**之间的相互约束关系

- 有助于提高人们对“准歧义”格式的关注度，在以往针对人的歧义研究中，“准歧义” 格式不大会引起人们的注意。

2 增加合一约束的句法结构分析

CFG 文法要更精确地描述句法成分的组合，就要增加非终结符



基于简单范畴的CFG文法的不足

- 范畴划分有不同的颗粒度（granularity）

  np vp ap 等各类短语都可以区分更小的子类

- 范畴划分有不同的角度（perspective）

为了增强对句法组合的约束，CFG文法中的非终结符需要不断细化。

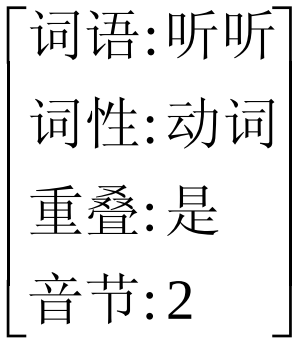
但增加非终结符的方式不可取，因此引入“**特征结构**”的表达形式。

特征结构（Feature Structure）

- 特征结构又名复杂特征集（Complex Feature Set）
- 特征结构定义为“特征”的集合
- 所谓“特征”，是一个由“属性”和“值”组成的二元组，“属性”也称为“特征名”，“值”也称为“特征值”
- 在特征结构中，要求所有的“特征”的“属性”互不相同
- 空特征结构：不含任何特征的特征结构 记作：[ ]

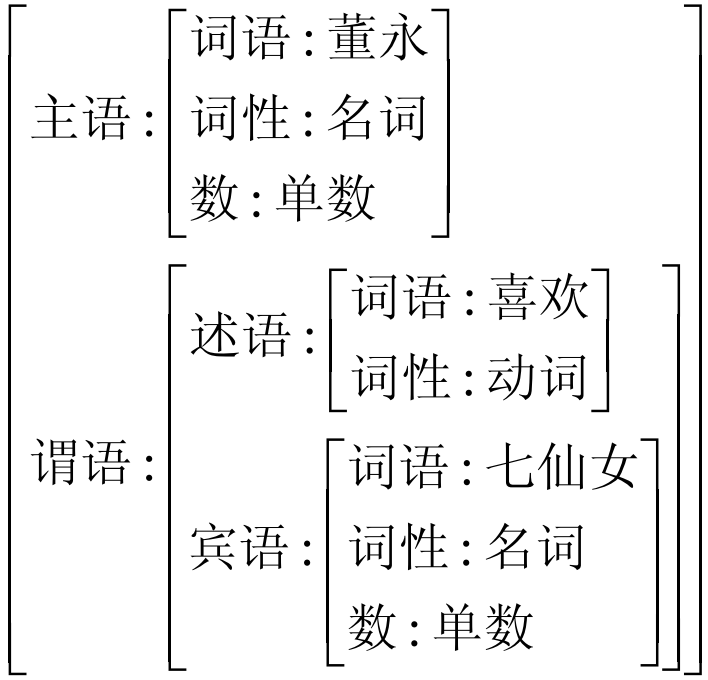
$$\begin{pmatrix} attribute_1=value_1 \\ attribute_2=value_2 \\ \cdots \\ \cdots \\ \cdots \\ attribute_n=value_n \end{pmatrix}$$

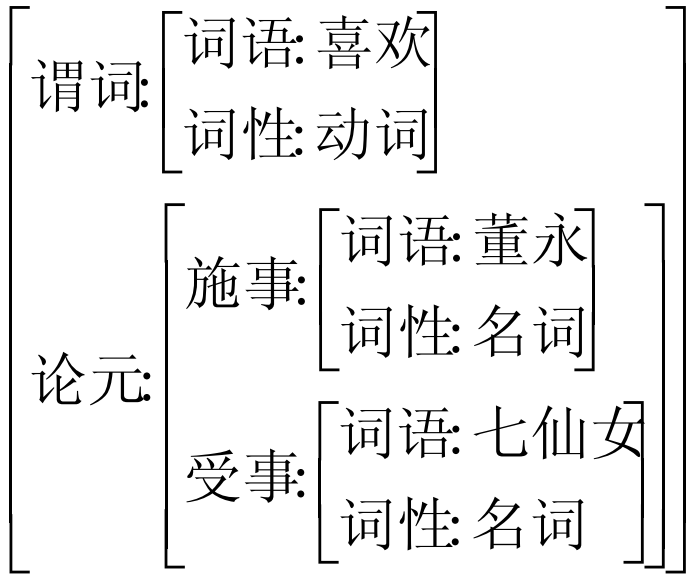
特征结构的嵌套



a. 简单特征结构

特征值是“原子”





b. 复杂特征结构（嵌套）

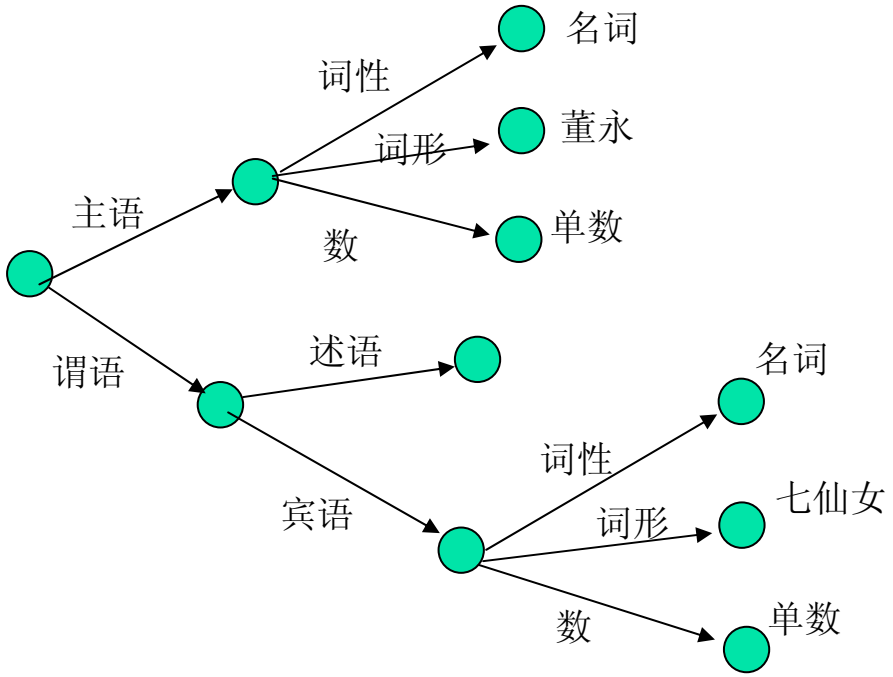
特征值是“特征结构”

特征结构的其他表示法

表（list）表示法

((主语: (词语:董永)
(词性:名词)
(数:单数))
(谓语: (述语:(词语:喜欢)
(词性:动词))
(宾语:(词语:七仙女)
(词性:名词)
(数:单数))))

图表示法
（**D**irected **A**cyclic **G**raph）



边（edge）表示特征

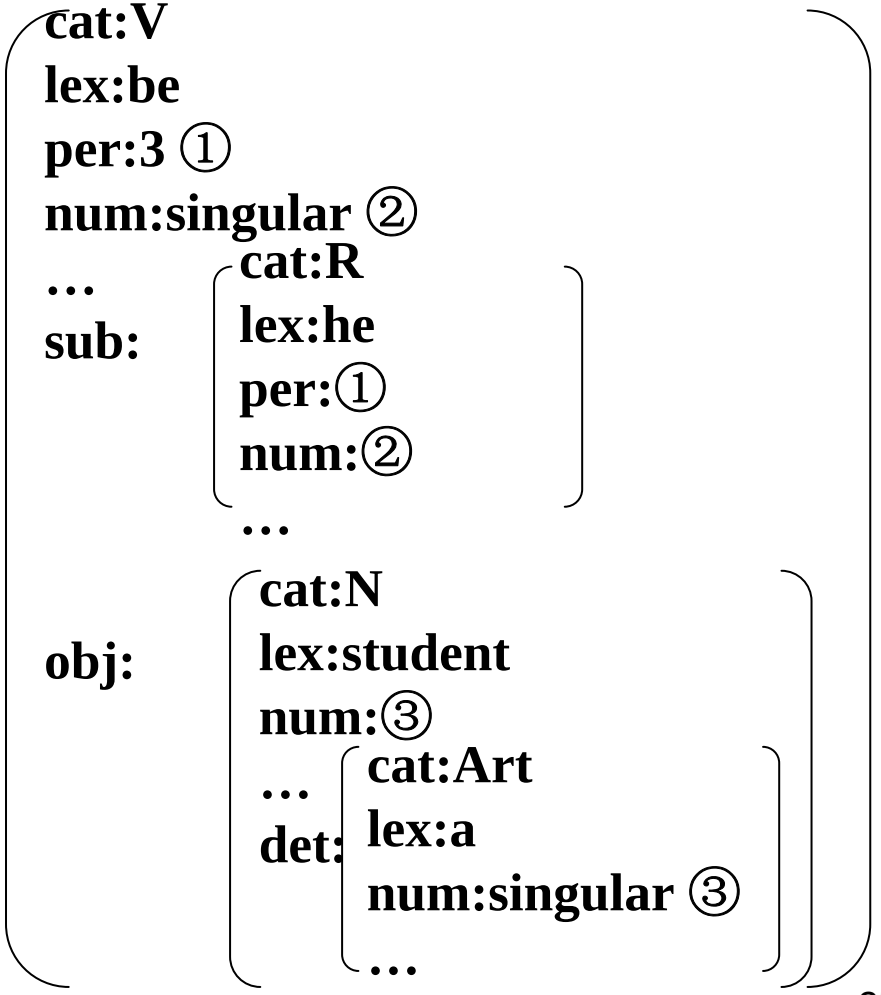
节点（node）表示特征值

特征结构的值共享

- 两个特征可以共享一个值，这是所谓的特征值的“共享”（也称为“重入”）。
- 在特征结构表示中，一般用数字表示重入的特征结构。
- 在重入的多个特征结构中，只需在一处说明其特征值。

- 例子：

  He is a student.



特征结构的值共享示例

- “同意”和 “答应”的区别

1 a 他同意去

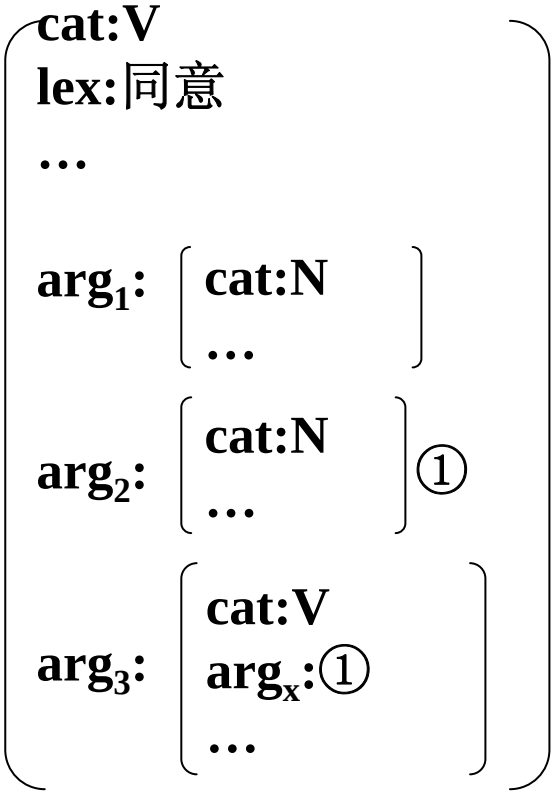
b 他答应去

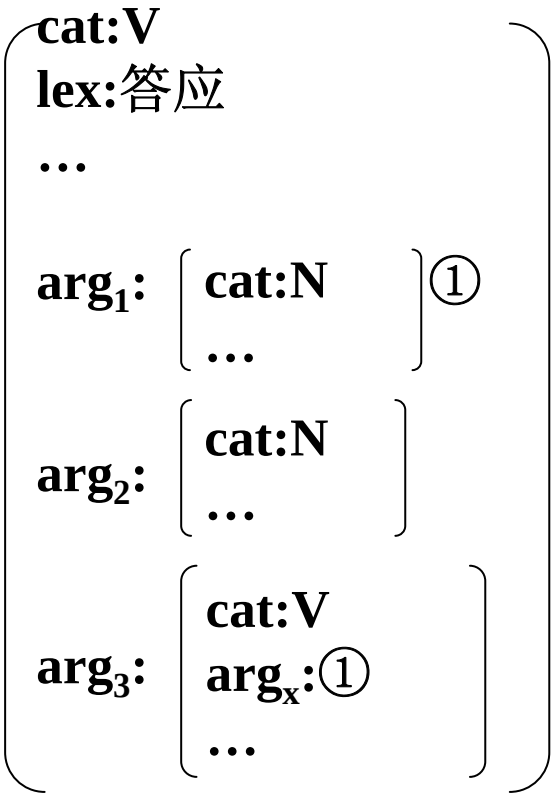
2 a 他同意明天去

b 他答应明天去

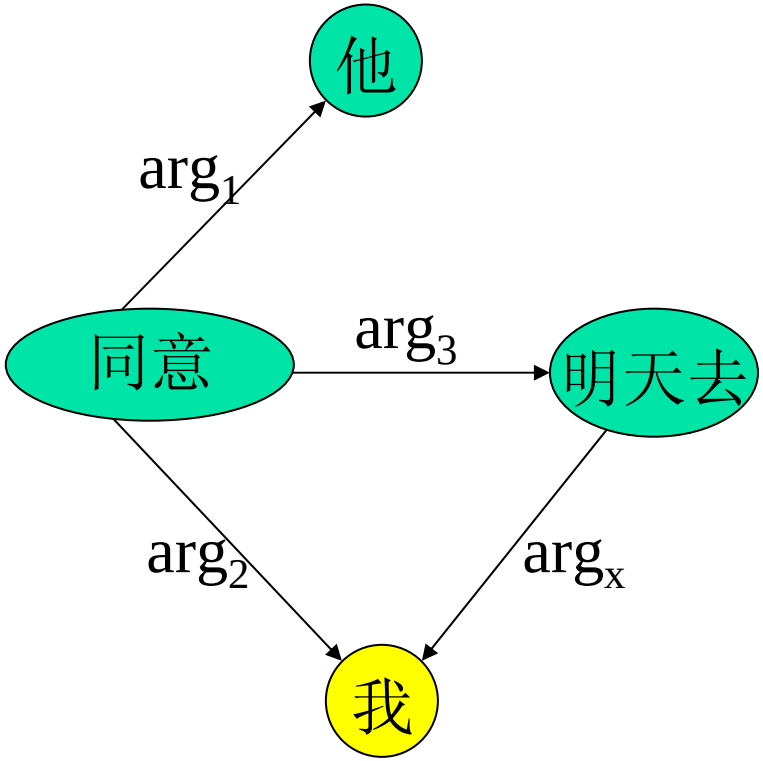
3 a 他同意我明天去

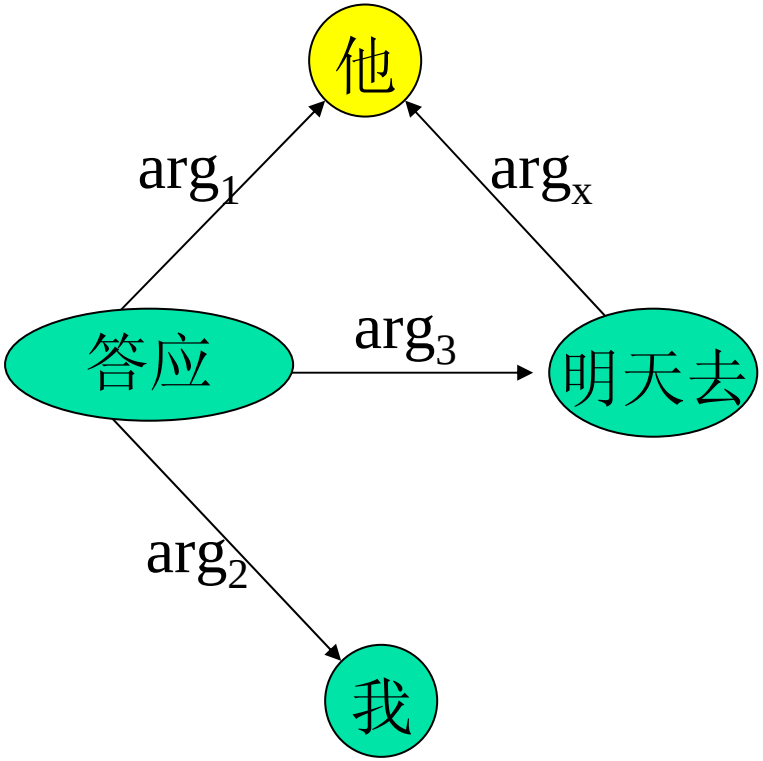
b 他答应我明天去





特征结构的值共享示例（续）





他同意我明天去　　vs.　　他答应我明天去

特征结构间的包孕关系subsumption

- 特征结构F1包孕F2，记作 $F1 \dot{\subseteq} F2$ ，当且仅当

  （1）若特征 $f \in F1$，则 $f \in F2$ ，并且

  （2）若f的值是特征结构，则 $value_{F1}(f) \dot{\subseteq} value_{F2}(f)$

  （3）若f的值是简单原子，则 $value_{F1}(f) = value_{F2}(f)$

- 空特征结构包孕任何特征结构

递归定义

特征结构包孕关系举例

$$\begin{bmatrix} Number & SG \end{bmatrix} \dot{\subseteq} \begin{bmatrix} Number & SG \\ PERSON & 3 \end{bmatrix}$$

$$\begin{bmatrix} Agree & \begin{bmatrix} Number & SG \end{bmatrix} \end{bmatrix} \dot{\subseteq} \begin{bmatrix} CAT & NP \\ Agree & \begin{bmatrix} Number & SG \\ PERSON & 3 \end{bmatrix} \end{bmatrix}$$

$$[\,] \dot{\subseteq} \begin{bmatrix} Number & SG \\ PERSON & 3 \end{bmatrix}$$

$$\begin{bmatrix} Number & SG \end{bmatrix} \dot{\not\subseteq} \begin{bmatrix} Number & PL \end{bmatrix}$$

特征结构的合一运算

- 合一运算（Unification）：将两个独立的特征结构F1，F2组合为一个新的特征结构F3，满足条件：$F1 \dot{\subseteq} F3$ 并且 $F2 \dot{\subseteq} F3$
- 合一的含义是：对两个特征结构进行类似于集合求**并**的一种运算，从而可以在“小”的特征结构基础上形成“大”的特征结构，这种运算非常适于刻划 “小”的语言单位发展为“大”的语言单位的过程中的信息增加，即F3中包含了F1，F2所包含的信息。

合一运算实例（一）

$$A = \begin{bmatrix} 结构:述宾 \\ 功能:述语 \\ 词性:动词 \\ 及物:是 \end{bmatrix} \qquad B = \begin{bmatrix} 词语:咳嗽 \\ 词性:动词 \\ 及物:否 \end{bmatrix}$$

$$A \overline{\cup} B = \phi$$

合一失败

“合一”的作用：检查两个特征结构所包含的信息是否相容

合一运算实例（二）

令 A = $\begin{bmatrix} 施事:C \\ 谓词:知道 \end{bmatrix}$ B= $\begin{bmatrix} 词语:董永 \\ 语义类:人 \end{bmatrix}$，其中 C = [语义类:人]

则将 C 和 B 合一后，特征结构A变为：

$$\begin{bmatrix} 施事:\begin{bmatrix} 语义类:人 \\ 词语:董永 \end{bmatrix} \\ 谓词:知道 \end{bmatrix}$$

合一成功

“合一”的作用：合一成功，特征结构的信息量增加

合一运算实例（三）

$$E = \begin{bmatrix} \text{Agree:} & \begin{bmatrix} \text{Number: Singular} \end{bmatrix} & ① \\ \text{Subject:} & \begin{bmatrix} \text{Agree:} & ① \end{bmatrix} & \end{bmatrix}$$

$$F = \begin{bmatrix} \text{Subject:} & \begin{bmatrix} \text{Agree:} & \begin{bmatrix} \text{Person: 3} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

$$E \overline{\cup} F = \begin{bmatrix} \text{Agree:} & \begin{bmatrix} \text{Number: Singular} \\ \text{Person: 3} \end{bmatrix} & ① \\ \text{Subject:} & \begin{bmatrix} \text{Agree:} & ① \end{bmatrix} & \end{bmatrix}$$

合一成功

合一运算的性质

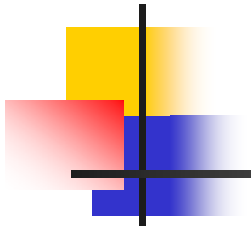
- 交换律：$A \bar{\cup} B = B \bar{\cup} A$
- 结合律：$A \bar{\cup} (B \bar{\cup} C) = (A \bar{\cup} B) \bar{\cup} C$

说明：

•合一运算的结果与执行顺序无关（order independent）；

•合一运算使得特征结构真正成为一种“描述性”知识表示方法，而不是“过程性”的表示方法；

•“描述性”知识表示方法的含义在于，对于一个变量的约束和赋值是等同的，我们可以在对一个变量赋值之前就给出对它的约束，而不必等到对这个变量赋值之后才对它进行约束。比如，我们可以在词典中指出，汉语动词“同意”的arg_3的arg_1必须和“同意”的arg_2合一，虽然这时我们并不知道在具体的句子中“同意”的各个arg是什么。因此，特征结构的“描述性”特点有利于在词典中给出词语的个性化描述。



为CFG文法增加特征结构合一描述

- 产生式规则：

  描述一个语言的基本范畴及其组合模式；

- 基于特征结构的合一约束：

  ①描述基本范畴之间发生组合关系的条件；

  ②描述组合后整体的功能特征；

规则描述内容示例

规则描述内容 / 实例	**内部构成**	**外部功能**
一件衣服	qp + np， 定中结构， qp是定语， np是中心语；	√ 名词性短语（np）， √ 主谓结构的主语， √ 述宾结构的宾语， √ 定中结构的中心语； × 述补结构的补语， × 定中结构的定语， × 状中结构的状语和中心语；

“一件衣服”相关的规则示例

&& {R1} np -> qp !np

① $.内部结构=定中,$.定语=%qp,$.中心语=%np,$.zuodingyu=否,…,

② %np.数量名=是,…,

③ IF %qp.量词子类=个体 THEN %np.个体量词=%qp.原形 ENDIF,…

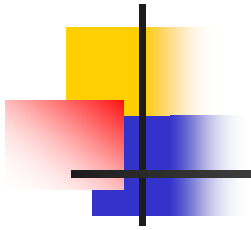
&& {R2} qp -> mp !q

$.内部结构=数量,$.定语=%mp,$.中心语=%q,$.zuodingyu=是,…,

&& {R3} np -> !n

$.内部结构=单词

① 说明np整体的特征，包括内部结构，定语，中心语，整体的功能分布特征等；
② 说明对中心语np的约束条件（独立条件）
③ 说明对定语mp与中心语np之间的相互约束条件（相关条件）

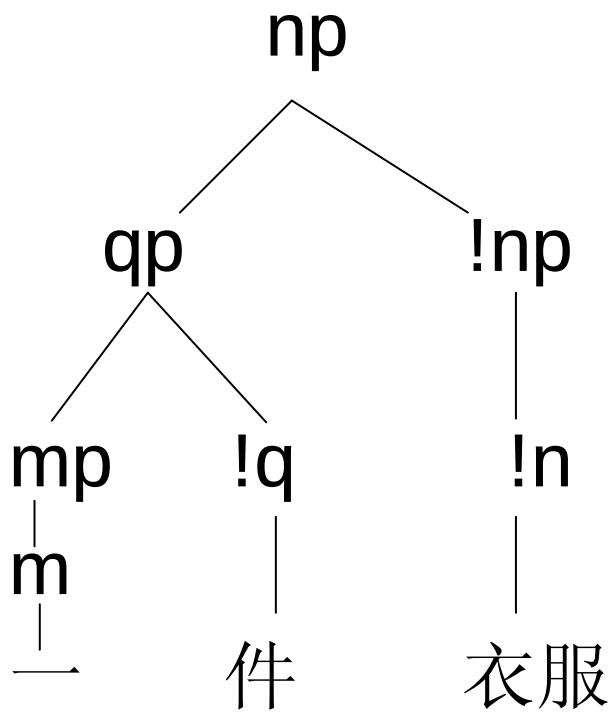


在词典中描述关于词的规则

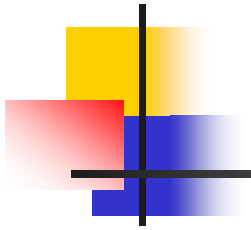
词语	特征结构
……	……
一	[词性:m,数词子类:基数]
件	[词性:q,量词子类:个体,表数:数]
衣服	[词性:n, 数量名:是,个体量词:件|套,…]
心胸	[词性:n, 数量名:否,…]
……	……

俞士汶等，《现代汉语语法信息词典详解》（第2版），北京：清华大学出版社，2003年2月

“一件衣服”的分析结果



$$\begin{bmatrix} \text{短语类型} = np \\ \text{内部结构} = \text{定中} \\ \text{定语} = \begin{bmatrix} \text{短语类型} = qp \\ \text{内部结构} = \text{定中} \\ \text{定语} = \begin{bmatrix} \text{短语类型} = mp \\ \text{内部结构} = \text{单词} \\ \text{词语} = \text{"一"} \end{bmatrix} \\ \text{中心语} = \begin{bmatrix} \text{词类} = q \\ \text{内部结构} = \text{单词} \\ \text{词语} = \text{"件"} \end{bmatrix} \end{bmatrix} \\ \text{中心语} = \begin{bmatrix} \text{短语类型} = np \\ \text{内部结构} = \text{单词} \\ \text{词语} = \text{"衣服"} \end{bmatrix} \end{bmatrix}$$



规则的应用示例

1. 她买了一件衣服
2. 董永拿走了七仙女的一件衣服
3. * 一个心胸
4. * 一个衣服
5. * np[[一件衣服] [领子]]
6. * np[[一件] [衣服领子]]

规则的应用示例：状中式ap的内部差异

1. 很 好
2. 更 好
3. 不 好
4. 更 不 好

ap → dp ap :: $.内部结构=状中,...

IF %dp.yx=很|不 THEN $.comp=No ENDIF

ap → pp ap :: $.内部结构=状中, ...

%ap.comp=Yes

1. （张三）比李四 好
2. 很好 ×
3. 更好
4. 不好 ×
5. 更不好

“很好、更好、不好、更不好”都是 dp + ap 形成的状中式ap组合模式，但是在“比 np __ ”环境中，“很好、不好”不能进入。这个限制可以通过特征结构的合一约束表达。

规则的应用示例：述宾式vp的内部差异

1a. 相信 上帝

→ 1b. 曾经 相信上帝

→ 1c. 相信上帝 的 人

2a. 相信 不 相信 上帝

→ 2b. 曾经 相信 不 相信 上帝

→ 2c. 相信 不 相信 上帝 的 人

3a. 曾经 相信 不 相信 上帝 的 人

vp → vp np :: $.内部结构=述宾,...

IF %vp.内部结构=联合,%vp.语气=疑问 THEN $.Modified_by_dp = No, $.BeModifier=No ENDIF, ...

vp → dp vp :: $.内部结构=状中,...

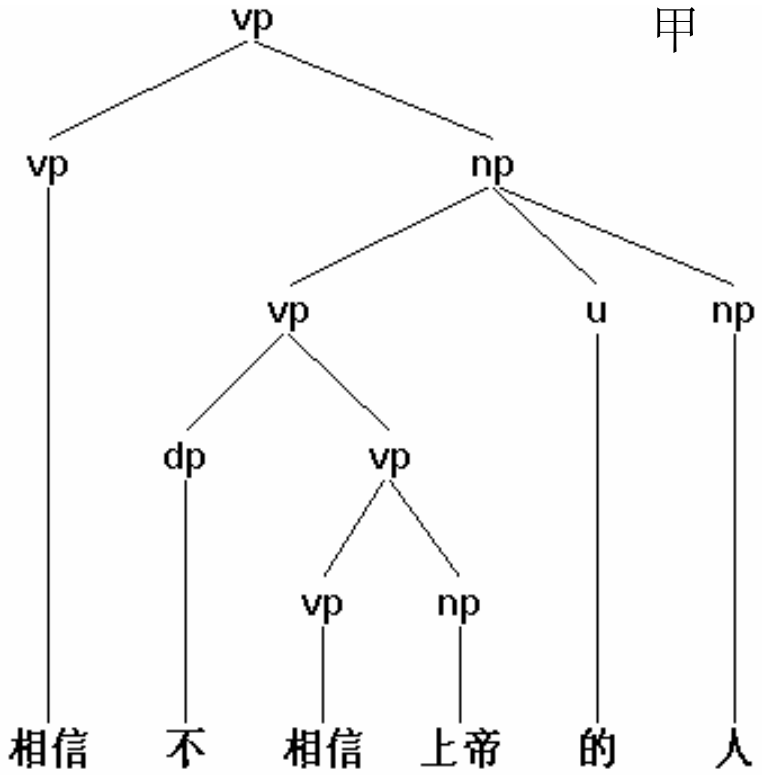
%vp.Modified_by_dp = Yes, ...

np → vp u<的> np :: $.内部结构=定中,...

%vp.BeModifier = Yes, ...

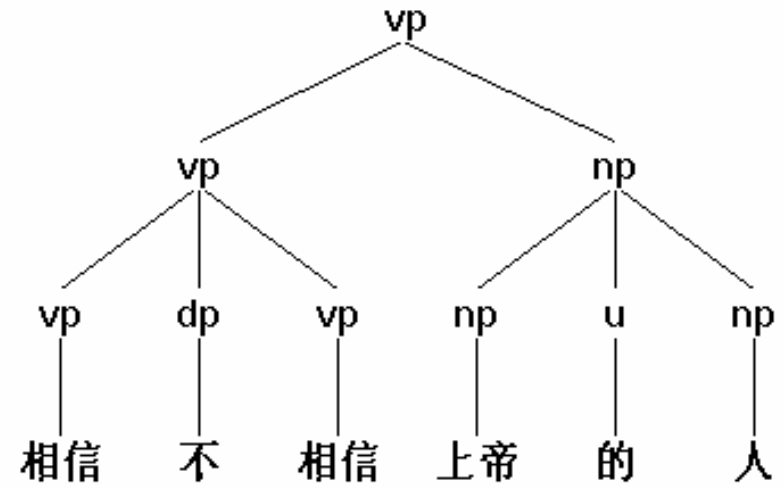
规则的应用示例：述宾式vp的内部差异

甲



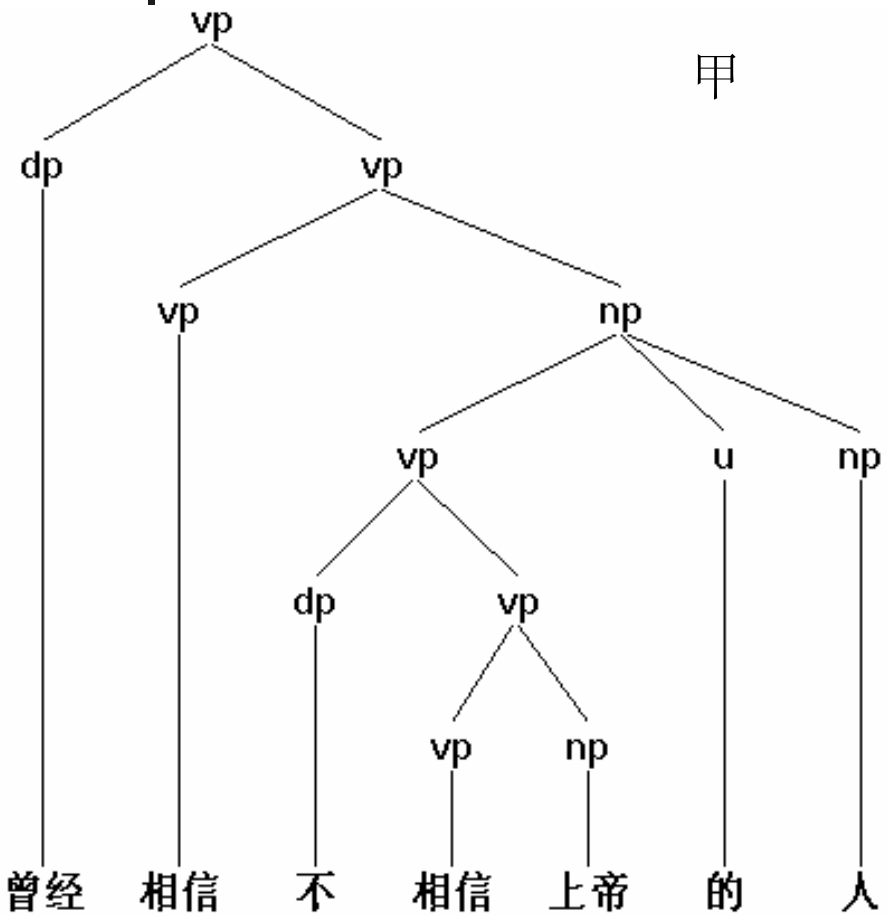
2c. 相信 不 相信 上帝 的 人

乙



规则的应用示例：述宾式vp的内部差异

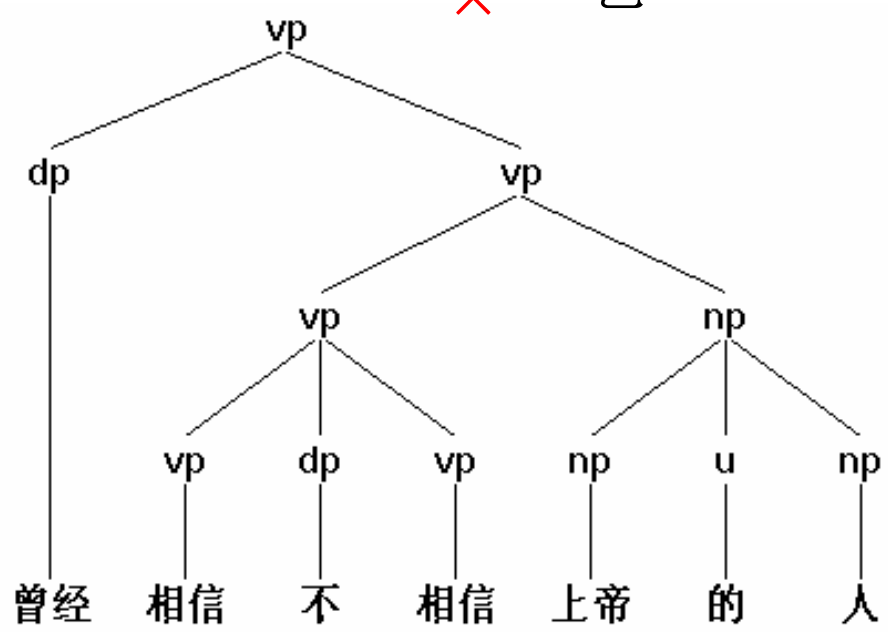
甲



3a. 曾经 相信 不 相信 上帝 的 人

×

乙

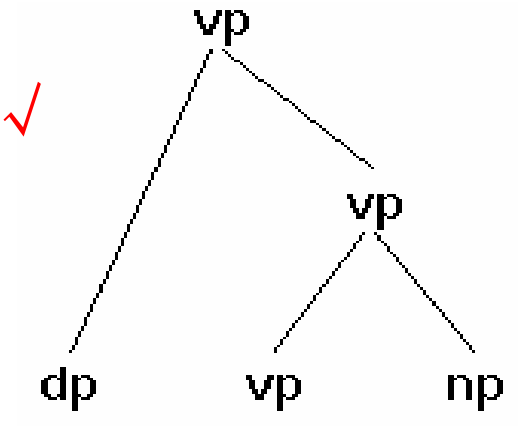


规则的应用示例：以歧义消解为例来说明

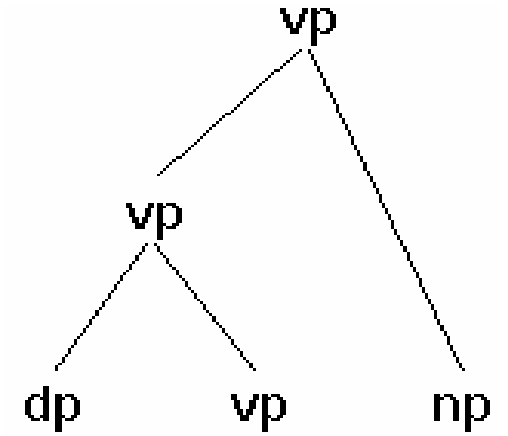
1. 伪歧义格式的处理举例 “dp vp np”格式的分析
2. 准歧义格式的处理举例 “qp qp 的 np”格式的分析
3. 真歧义格式的处理举例 “v a n”格式的分析

“dp vp np”格式的分析

√



X



vp_zz → dp vp_sb

vp_sb → vp np

vp_zz：状中式vp

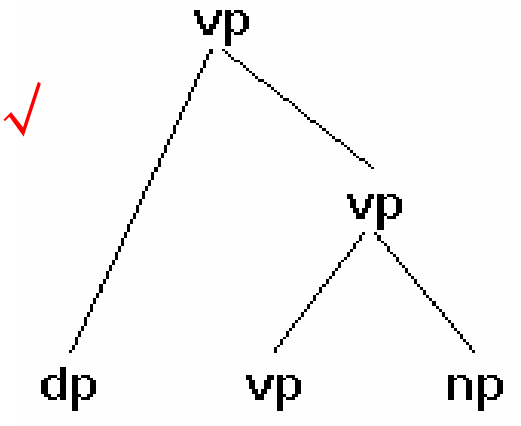
vp_sb：述宾式vp

vp：非状中、述宾式vp

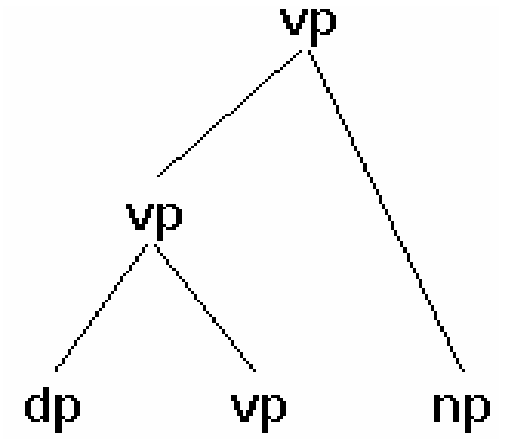
方案I

“dp vp np”格式的分析

√



X



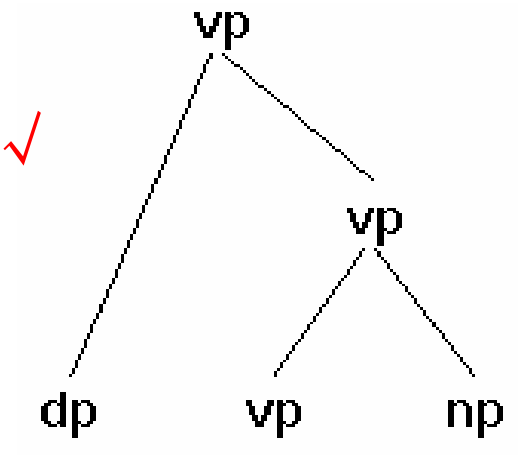
vp → dp vp :: $.内部结构=状中

vp → vp np :: $.内部结构=述宾,%vp.内部结构=~状中

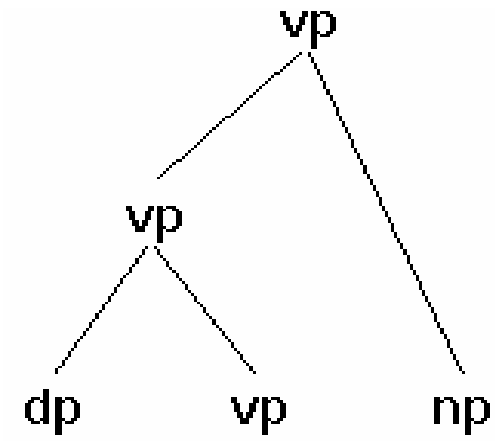
方案II　　根据“内部结构”特征值来进行约束

“dp vp np”格式的分析

√



X



vp → dp vp :: $.内部结构=状中,$.daibinyu=否

vp → vp np :: $.内部结构=述宾,%vp.daibinyu=是

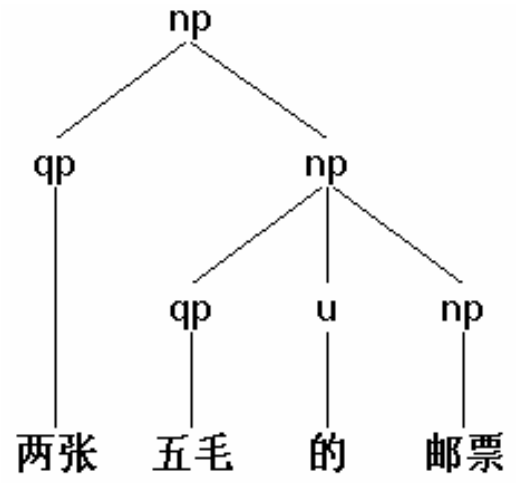
方案III

根据功能特征“daibinyu”（描述一个语言单位能否带宾语）来进行约束

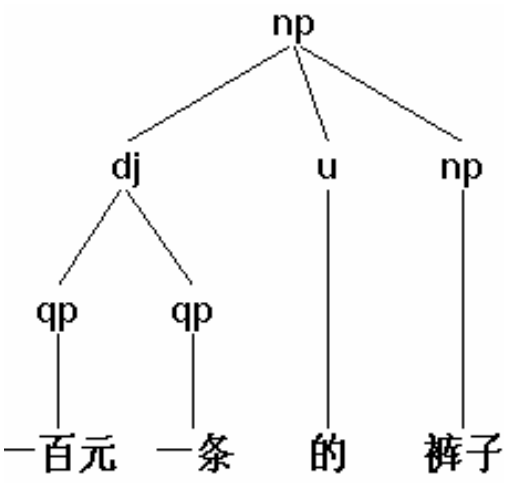
方案III更合理

“qp qp 的 np”格式的分析

两张 五毛 的 邮票



一百元一条的裤子



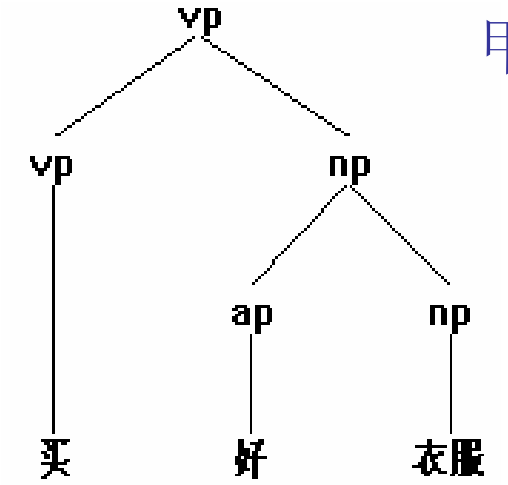
“qp qp 的 np”格式的分析

1. np → qp 的 np
   五毛 的 邮票
   :: $.内部结构=定中, %qp.zuodingyu=是,
   %qp.个体量词=否, …

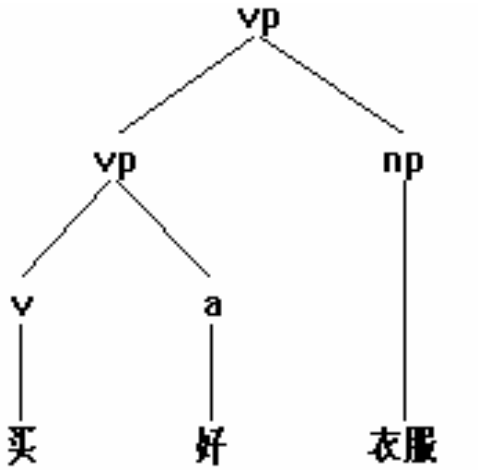
2. np → qp np
   两张 邮票
   :: $.内部结构=定中, %qp.zuodingyu=是,
   IF %qp.个体量词=是 THEN %np.个体量词=%qp, ENDIF…

3. dj → qp qp
   五十元 一斤
   :: $.内部结构=主谓,
   IF %qp.量词子类=%%qp.量词子类 FALSE,
   …

“v a n”格式的分析



甲



乙

如何给出区分甲和乙的判别条件？

“v a n”格式的分析

1. np → ap np :: $.内部结构=定中, %ap.zuodingyu=是, …

   新 球

2. vp → v a :: $.内部结构=述补, %ap.zuobuyu=是, …

   踢 破

3. vp → vp np :: $.内部结构=述补, %vp.daibinyu=是, …

   踢 球

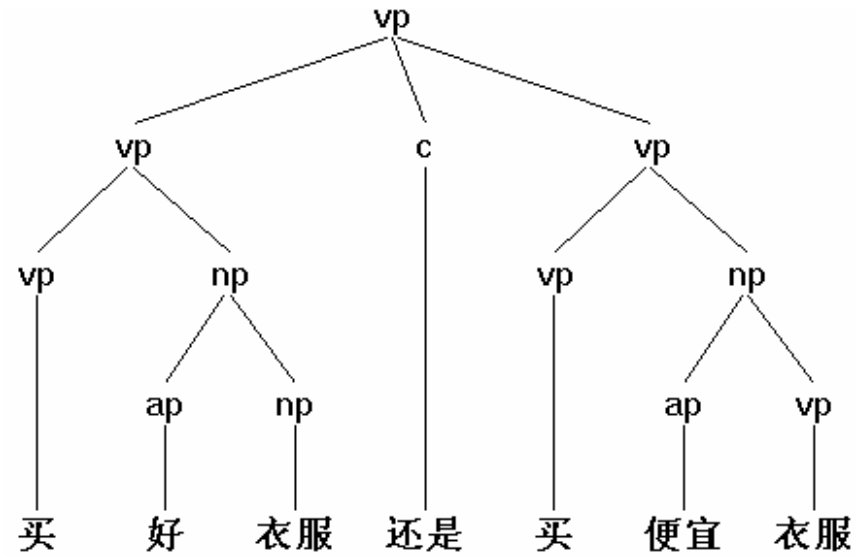
买 好 衣服

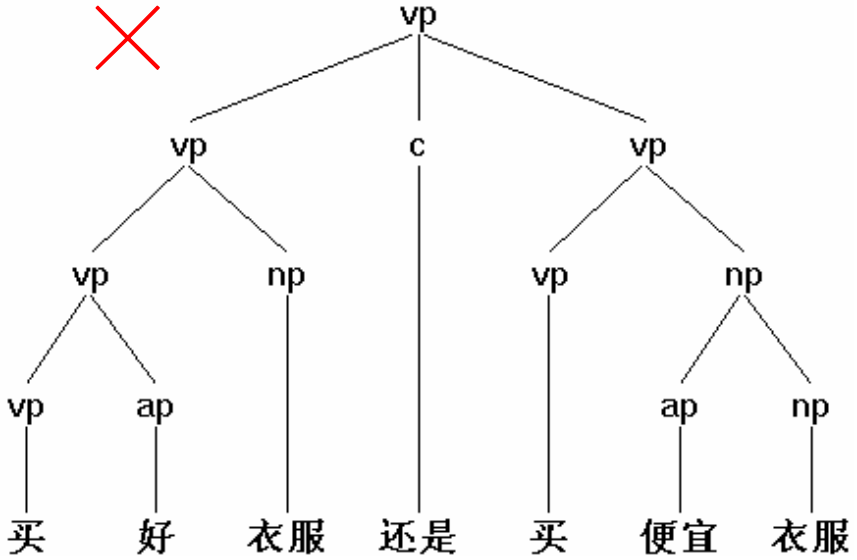
“好”同时满足规则1跟规则2

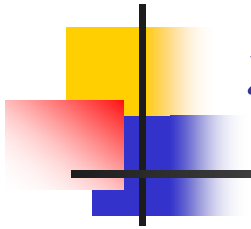
“v a n”格式的分析

你打算买好衣服还是买便宜衣服

vp → vp c vp :: ... %vp=%%vp @内部结构, %vp.述语.内部结构 =%%vp.述语.内部结构,...





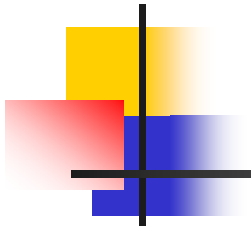


带合一约束的Earley算法

1. 合一算法（unification）

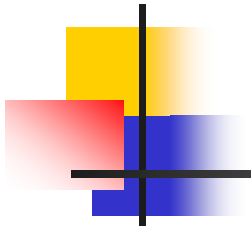
2. 将合一运算融入到Earley算法中

Daniel Jurafsky & James H. Martin, 2000, Speech and Language Processing, Pearson Education, Inc., Prentice Hall. chapter 11.



3 小结

- 自然语言的句法结构蕴含了大量的歧义。要消解歧义，就需要更准确地描述语言成分间的组合条件。
- 从 CFG 到 特征结构（FS），语言模型的表达灵活性大大提高了，人们可以更便利地描述语言成分间的组合条件。
- CFG产生式规则辅之以基于特征结构的合一约束，可以把语言知识中句法、语义等不同层次的知识纳入到一个统一的形式表达框架里加以描述。



进一步阅读文献

- 冯志伟等译（2005）《自然语言处理综论》第11章。
  Daniel Jurafsky & James H. Martin, 2000, Speech and Language Processing, Pearson Education, Inc., Prentice Hall. Chapter 11.

- 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。