

# 第三讲 汉语的句法规则系统 (2)

从形式文法的角度看汉语句法结构歧义

詹卫东

<http://ccl.pku.edu.cn/doubtfire/>

# 提纲

- 一 从人的角度看句法结构歧义
- 二 从计算机的角度看句法结构歧义
  - 2.1 外显型歧义与内含型歧义
  - 2.2 真歧义、准歧义、伪歧义
  - 2.3 歧义格式的长度
- 三 句法结构歧义的统计分析

# 一 从人的角度看歧义

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

# attachment ambiguity

- pp-attachment

I shot an elephant in my pajamas.

- gerundive-vp attachment

We saw the Eiffel Tower flying to Paris.

- np-attachment

Can you book TWA flights?

# coordination ambiguity

- old men and women
- John or Tom and Dick

## noun-phrase bracketing ambiguity

- complete peace plan

完全 的 和平计划

完全和平 的 计划

- dead poets' society

逝去 的 诗社

过世诗人 的 社团

# 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

## 二 从计算机的角度看结构歧义

- 显性歧义与隐性歧义
  - 句子层（终端）歧义 vs. 结构层（模式）歧义
  - 歧义格式对环境敏感 vs. 歧义格式对环境不敏感
- 歧义程度的量化分析
  - 歧义格式自身的长度
  - 歧义格式的可能结构树的数量

## 2.1 外显型歧义与内含型歧义

1 v n u<的> n

1a [修 [老王 的 自行车]]

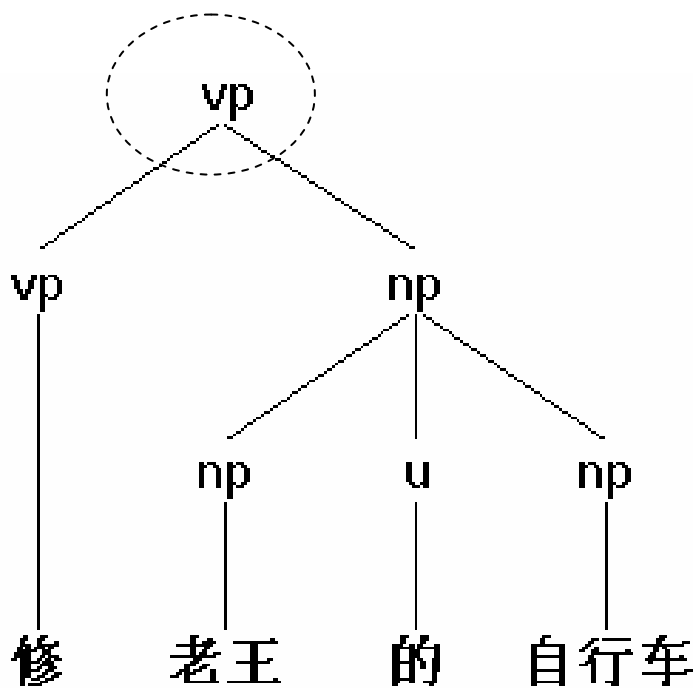
1b [[修 自行车 的] 扳手]

2 ap np np

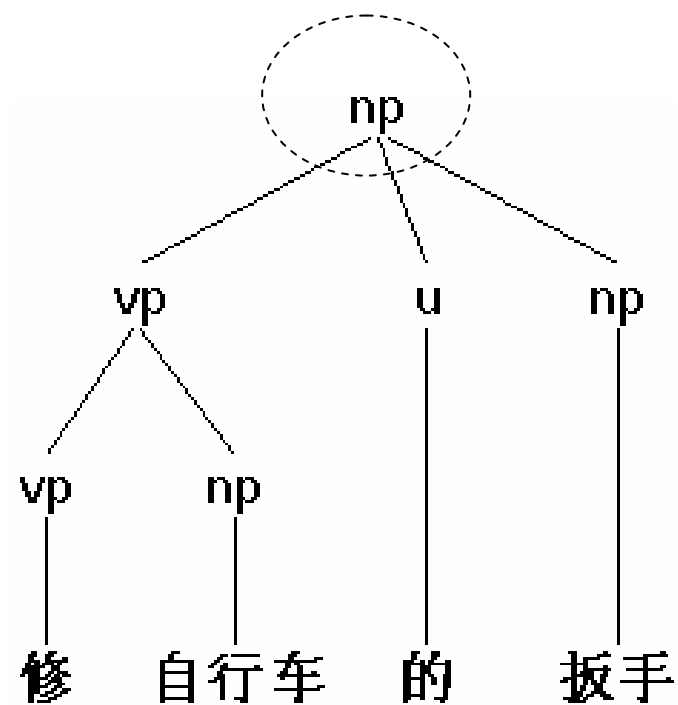
2a [大 [钢铁 公司]]

2b [[大 眼睛] 姑娘]

## 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

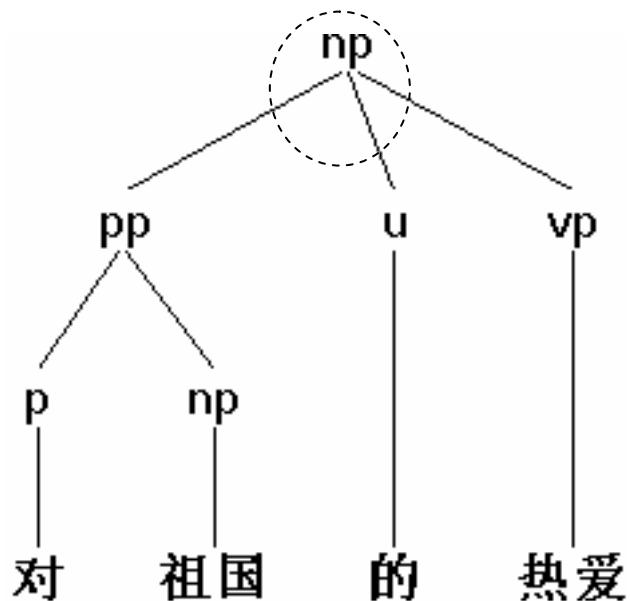
.....

vp

np

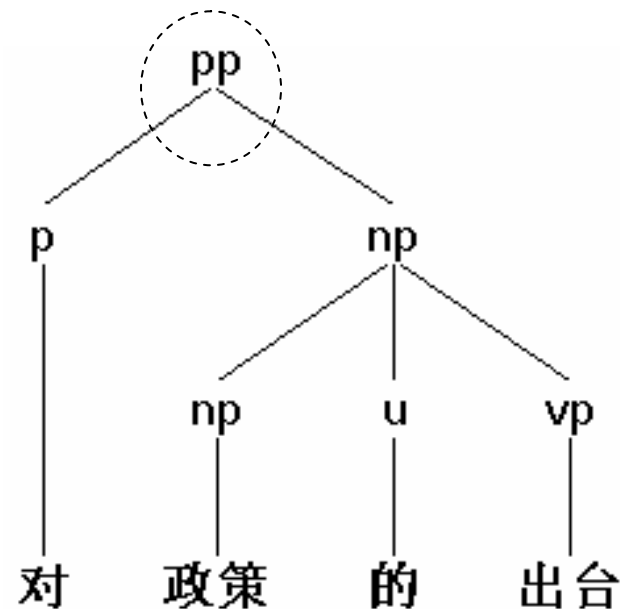
## 外显型歧义 (续2)

对祖国的热爱



=/=

对政策的出台



他对校长的意见很大

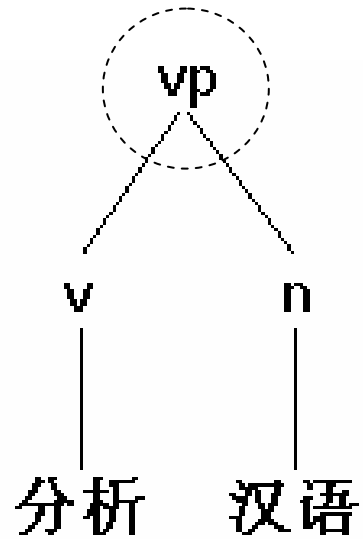
他对校长的批评很尖锐

他对校长的意见持否定态度

他对校长的批评充耳不闻

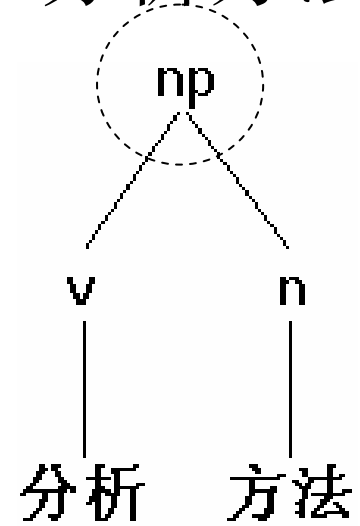
## 外显型歧义（续）

分析汉语



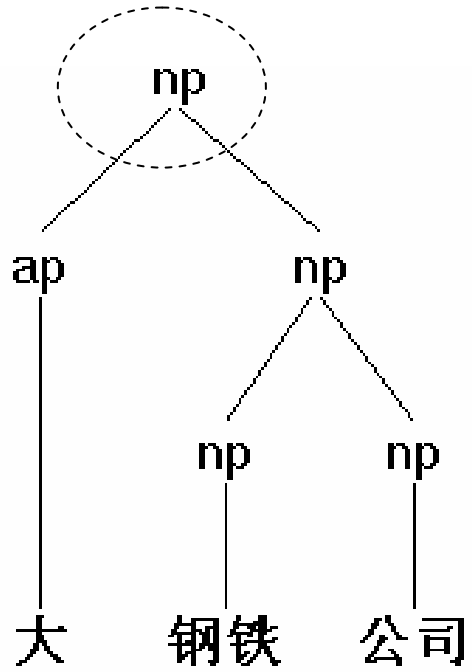
=/=

分析方法

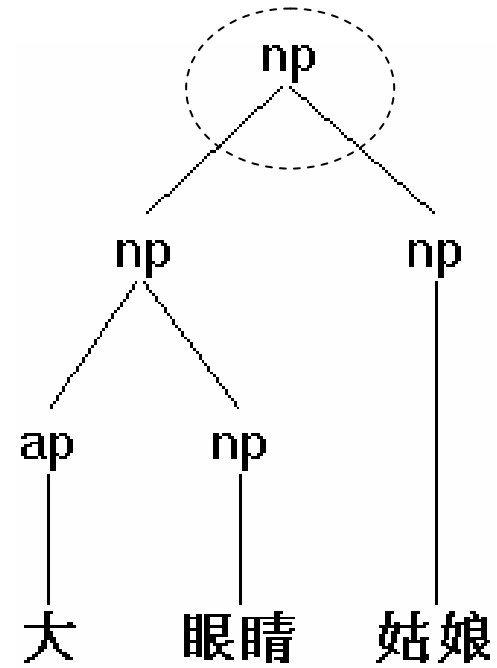


出租汽车 np or vp

# 内含型歧义

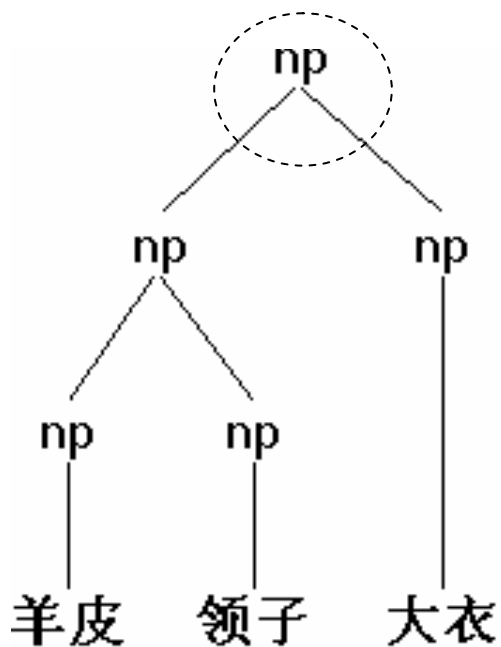


==



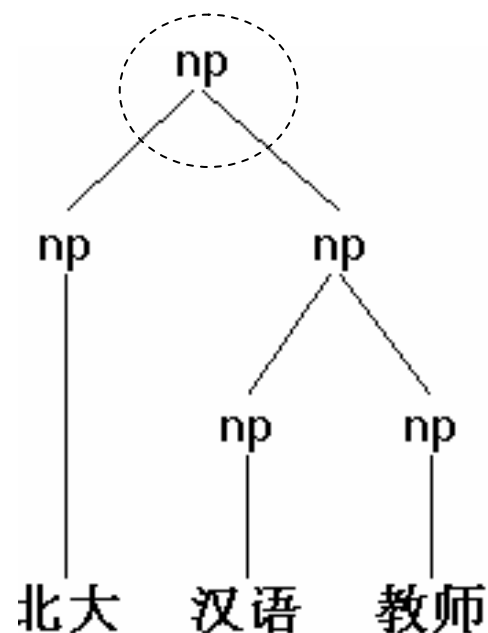
## 内含型歧义 (续1)

羊皮领子大衣



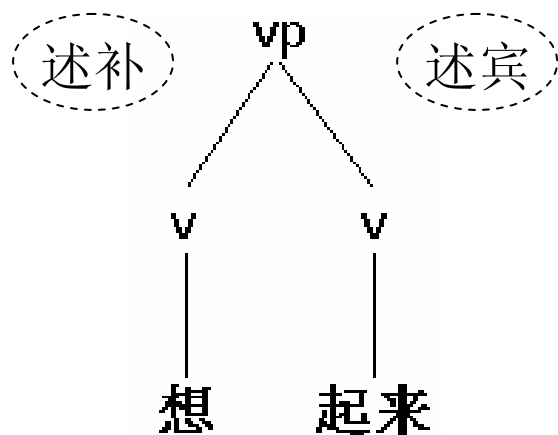
==

北大汉语教师



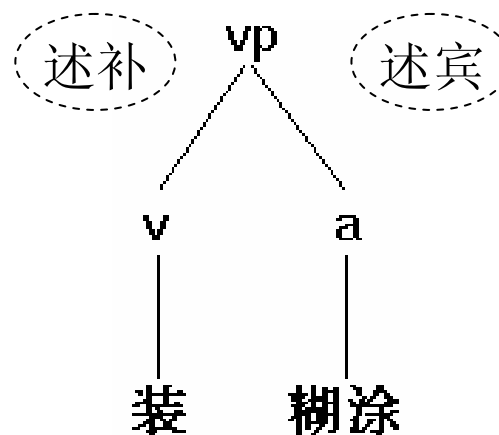
## 内含型歧义（续2）

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。

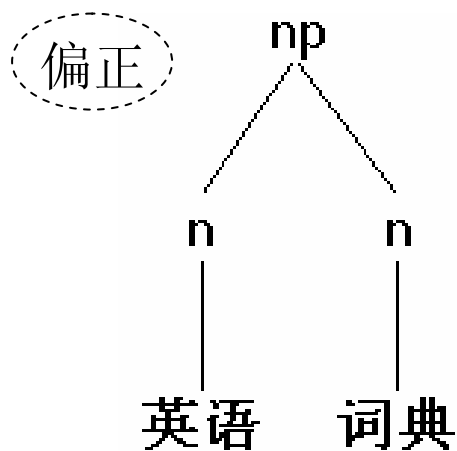
装糊涂



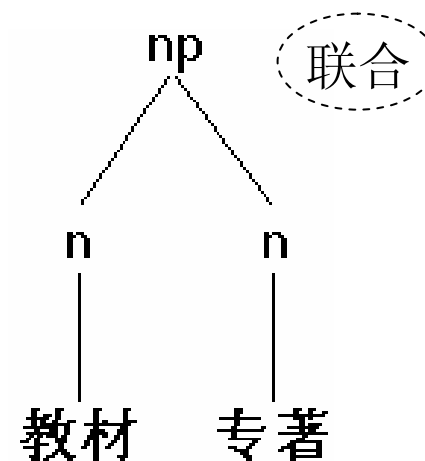
他就会装糊涂，其实他心理比谁都清楚  
装了一上午家具，我都装糊涂了

## 内含型歧义 (续3)

英语词典

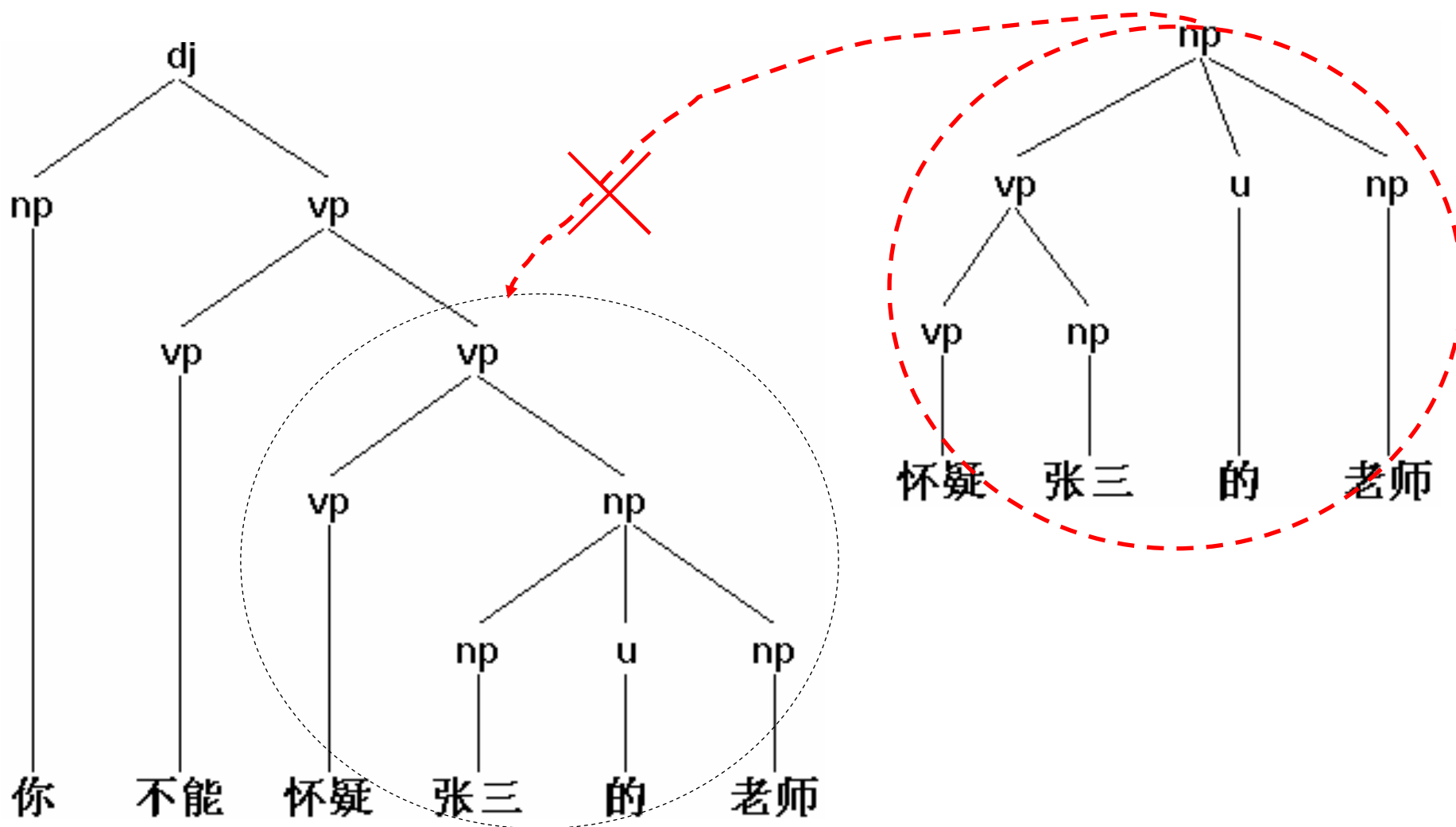


教材专著

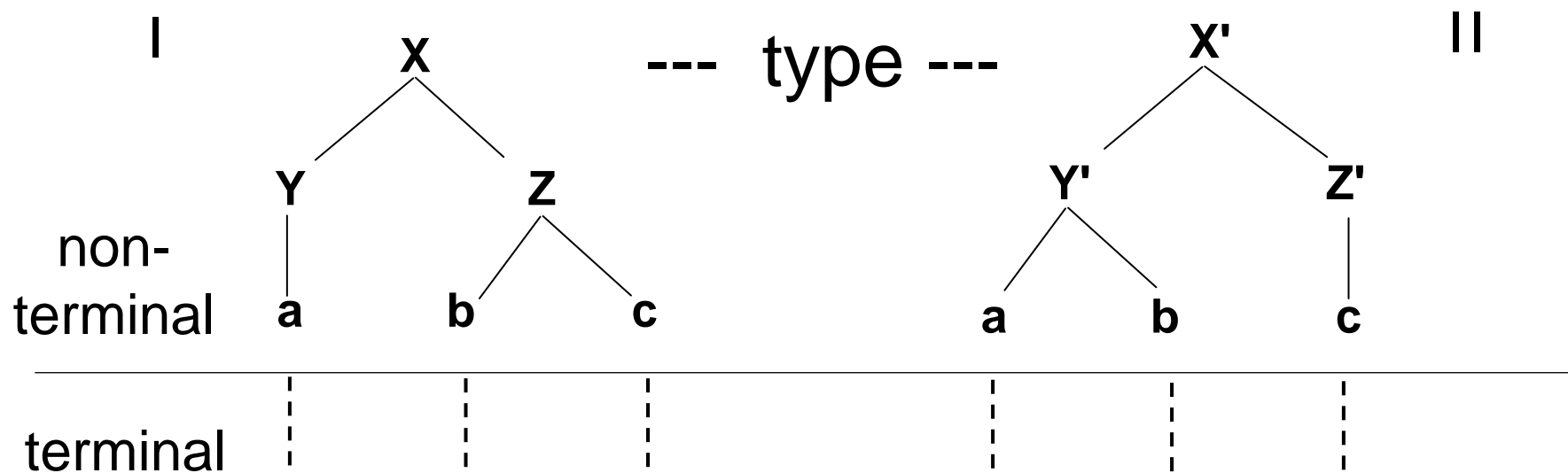


偏正? 牛奶饼干 联合?

## 区分“外显”与“内含”的作用

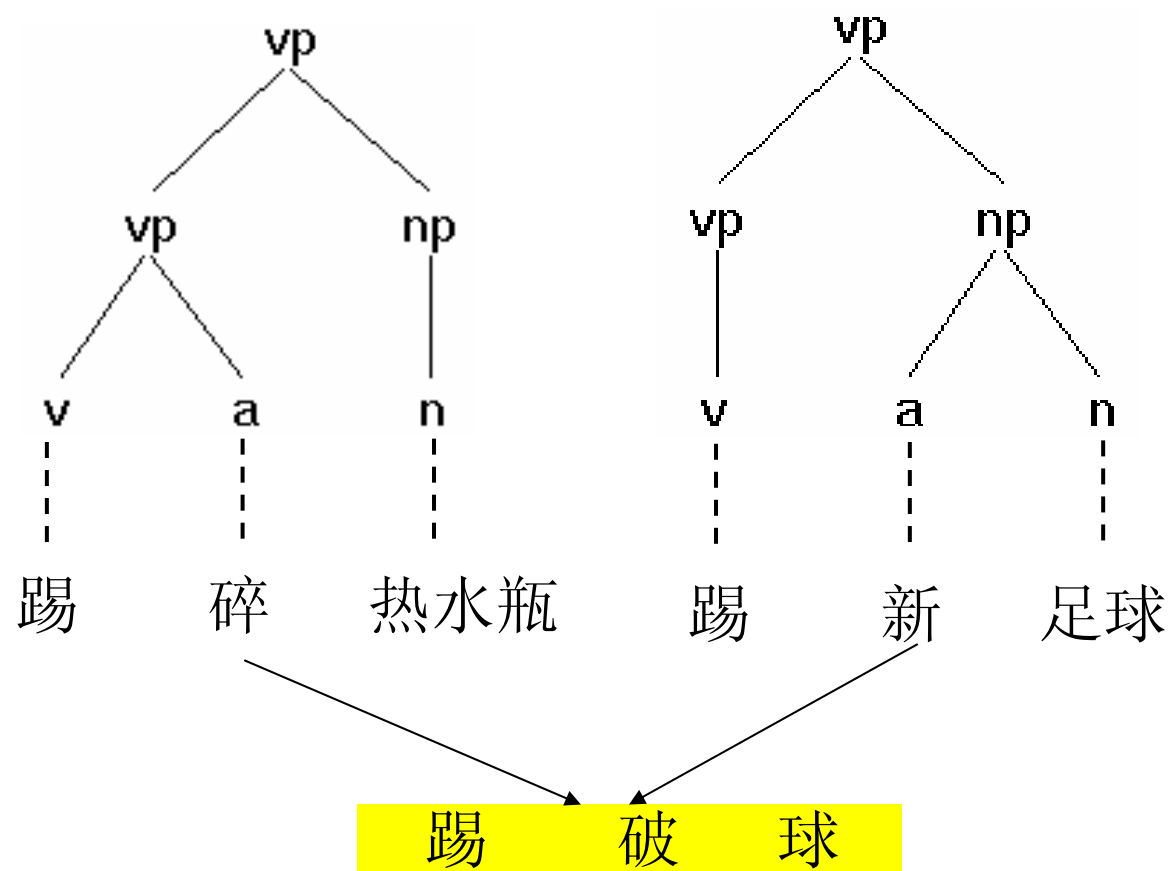


## 2.2 真歧义 准歧义 伪歧义

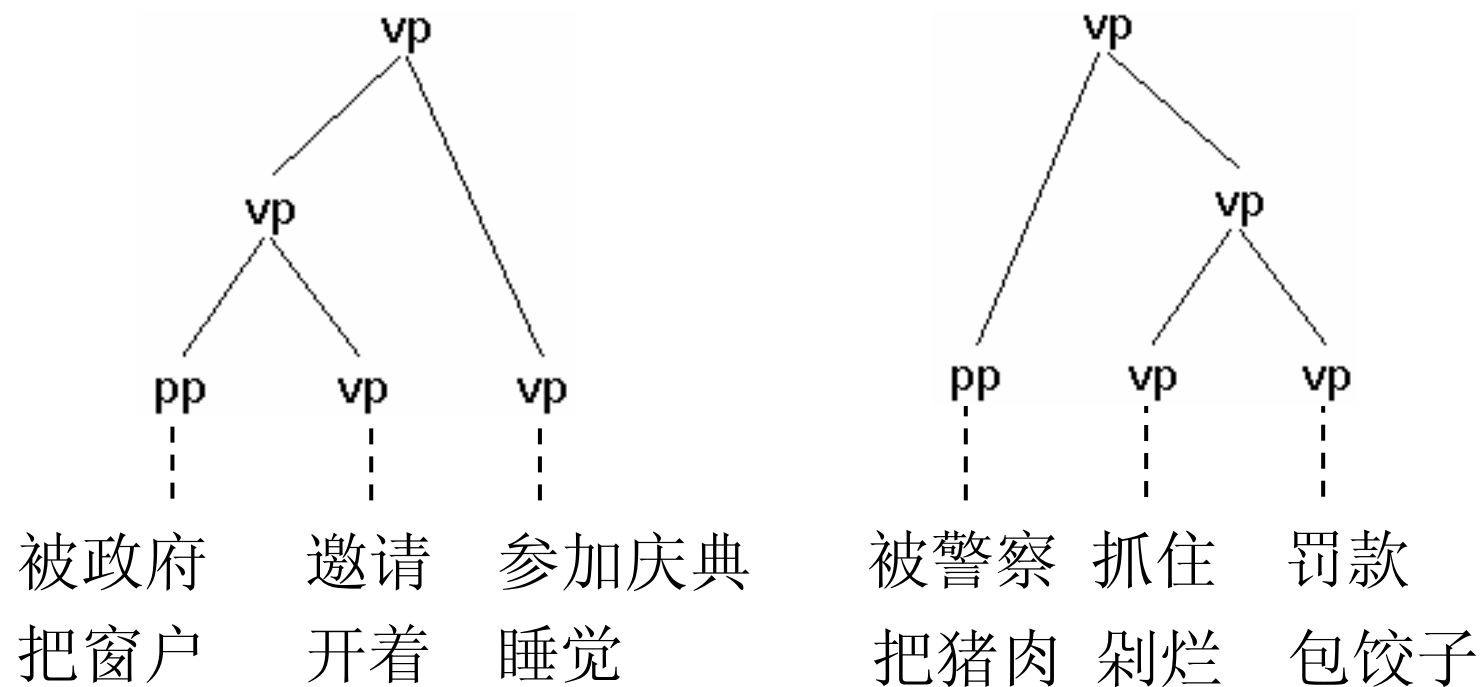


歧义? --- token --- 歧义?

# 真歧义



# 准歧义



把窗户

开着

睡觉

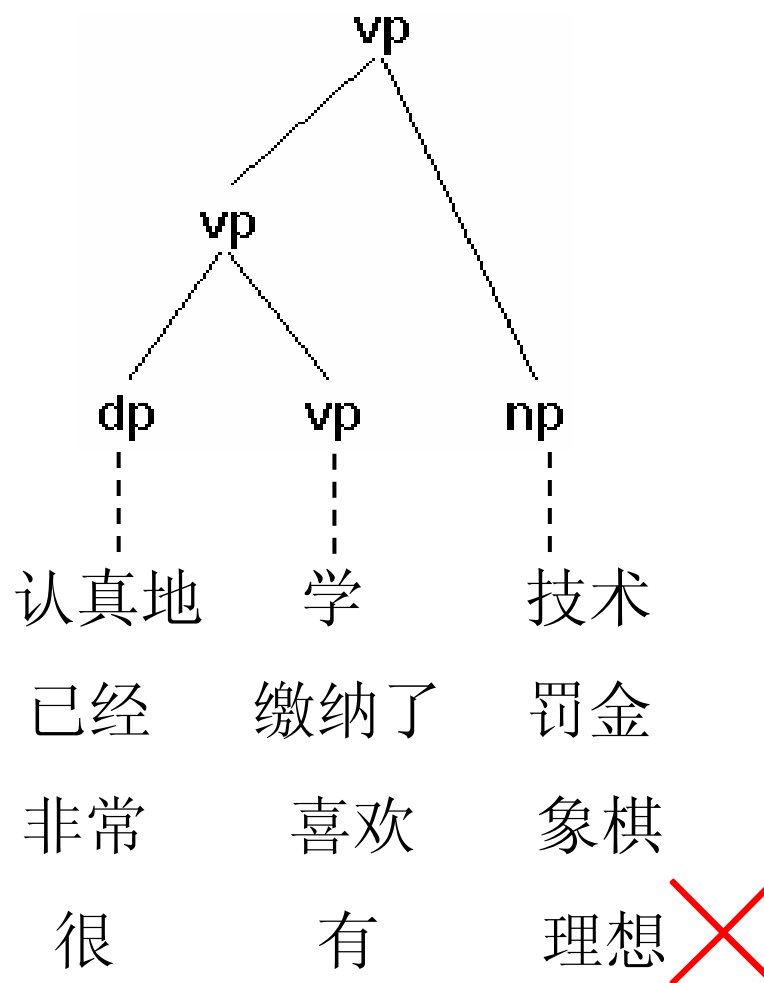
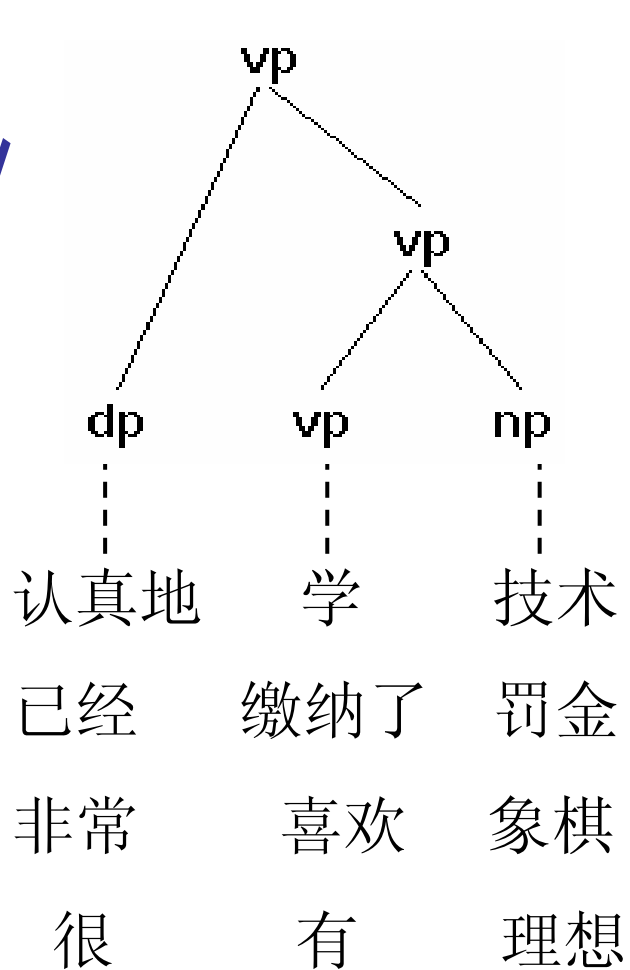
把猪肉

剁烂

包饺子

??????

# 伪歧义

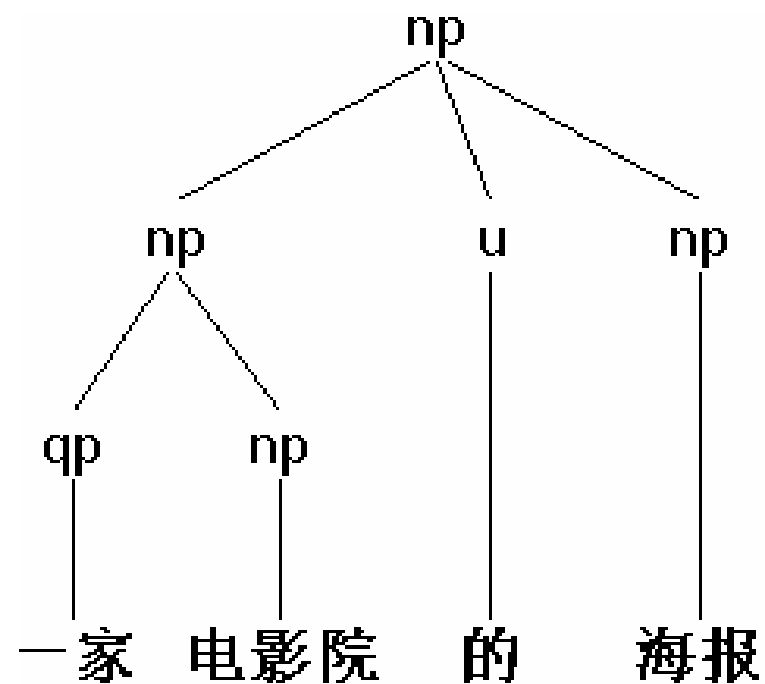
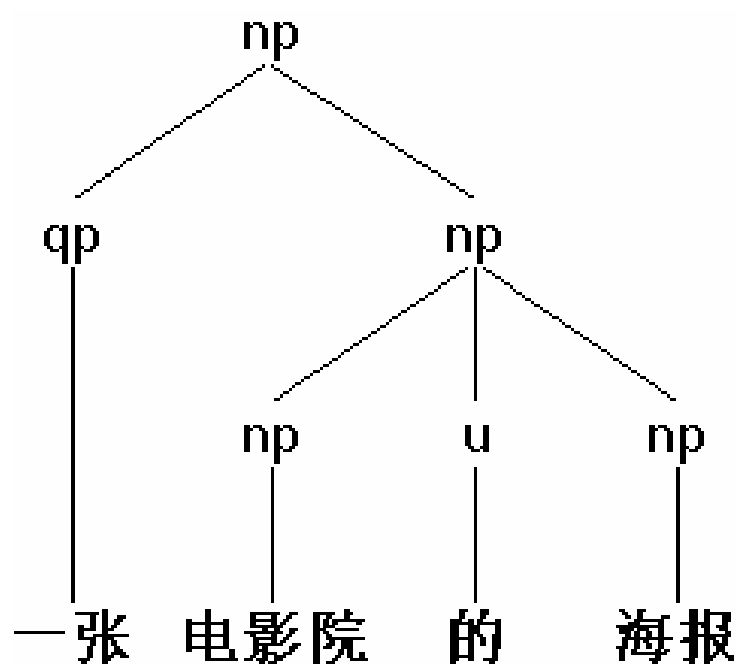


## 区分“真/准/伪”歧义的作用

- 有可能为计算机消解短语结构歧义制定不同的策略
- 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。

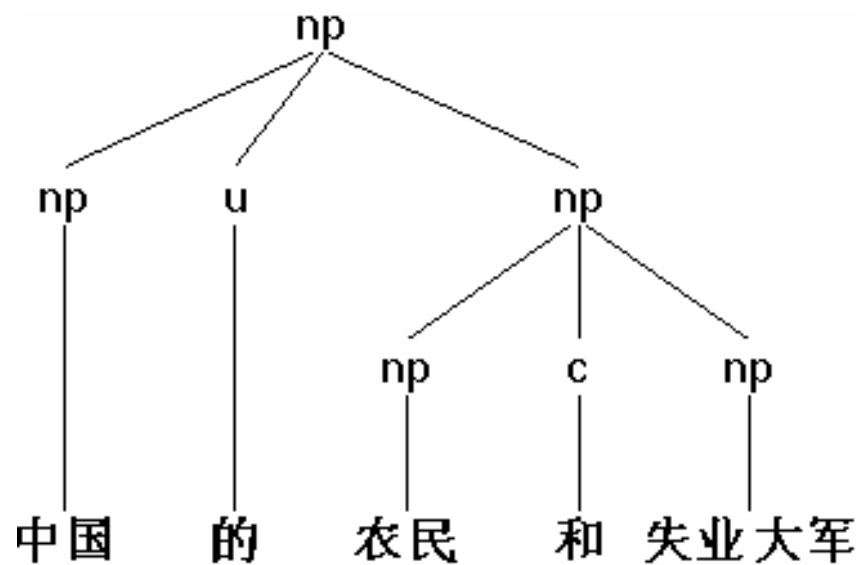
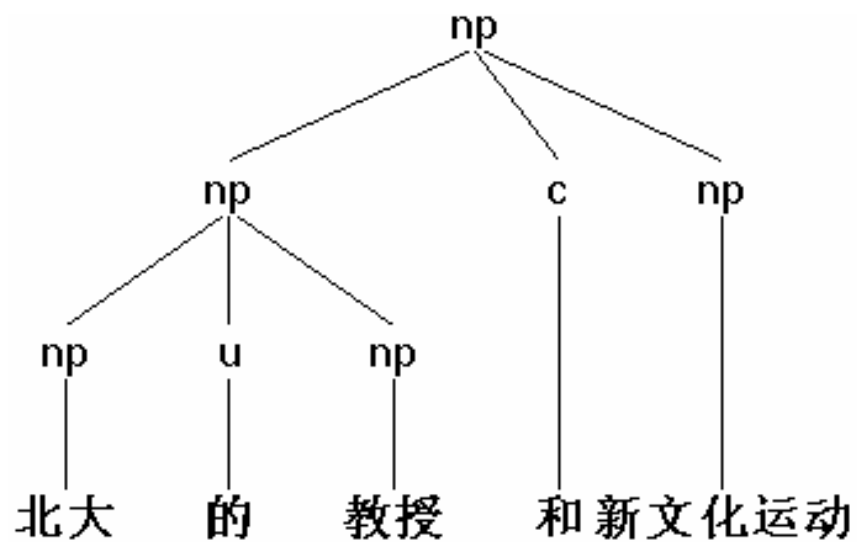
## 2.3 歧义格式的长度

qp np u<的> np



# 歧义格式的长度

np u<的> np c<和> np



# 歧义结构的数量分析

- Catalan number

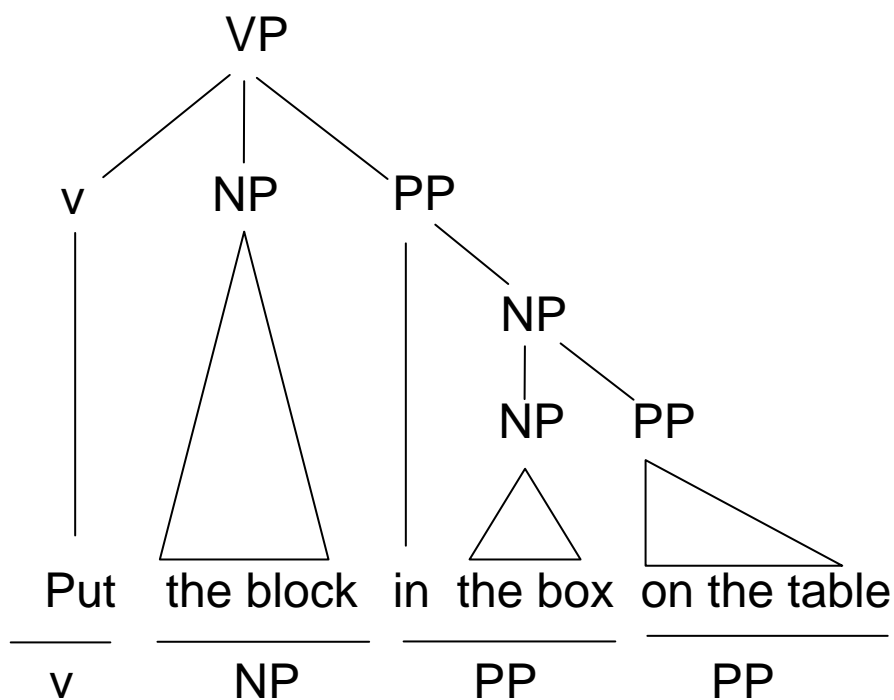
$$C_n = \frac{1}{n+1} \binom{2n}{n} = \frac{1}{n+1} \times \frac{(2n)!}{n!(2n-n)!} = \frac{(2n)!}{(n+1)(n!)^2} = \frac{(2n)!}{(n+1)!n!}$$

含有pp-attachment的句子中存在 $n$  ( $n$ 为自然数)个介词短语，则句法分析结果的数目（记作 $C_N$ ）是一个catalan数。

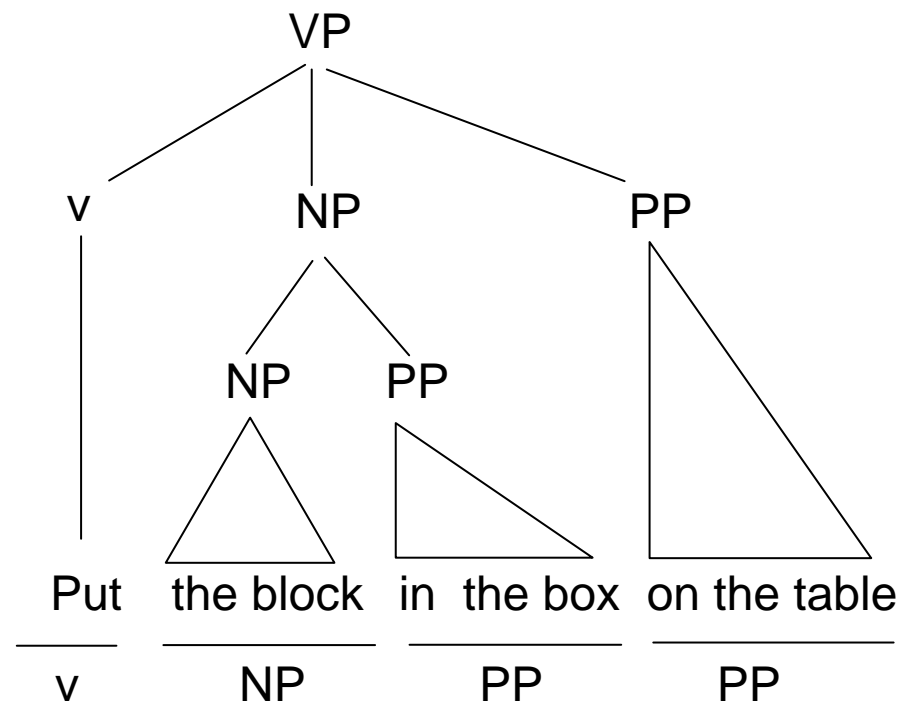
# Catalan number

- $n=2$        $(2 \cdot 2)! / (2+1)! \cdot 2! = 4! / 3! \cdot 2! = 2$
- $n=3$        $(2 \cdot 3)! / (3+1)! \cdot 3! = 6! / 4! \cdot 3! = 5$
- $n=4$        $(2 \cdot 4)! / (4+1)! \cdot 4! = 8! / 5! \cdot 4! = 14$
- $n=5$        $(2 \cdot 5)! / (5+1)! \cdot 5! = 10! / 6! \cdot 5! = 42$
- $n=6$        $(2 \cdot 6)! / (6+1)! \cdot 6! = 12! / 7! \cdot 6! = 132$
- $n=7$        $(2 \cdot 7)! / (7+1)! \cdot 7! = 14! / 8! \cdot 7! = 429$
- $n=8$        $(2 \cdot 8)! / (8+1)! \cdot 8! = 16! / 9! \cdot 8! = 1430$
- $n=9$        $(2 \cdot 9)! / (9+1)! \cdot 9! = 18! / 10! \cdot 9! = 4862$
- ...

# pp-attachment歧义实例



I



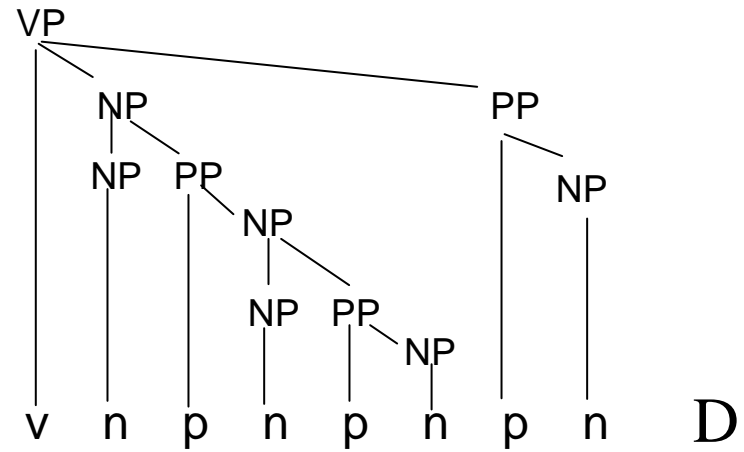
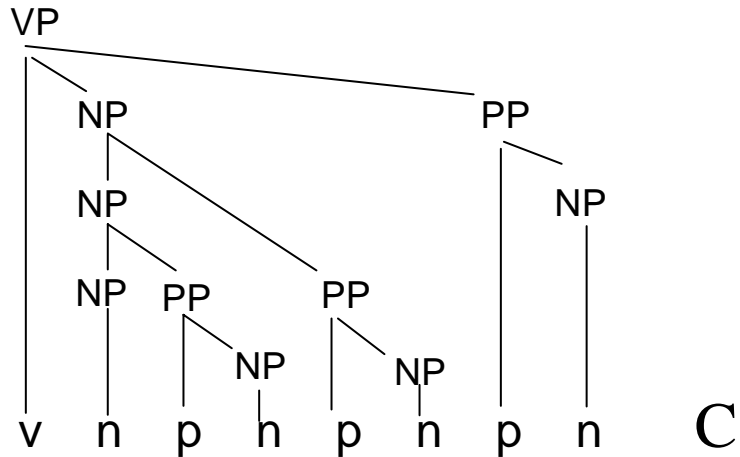
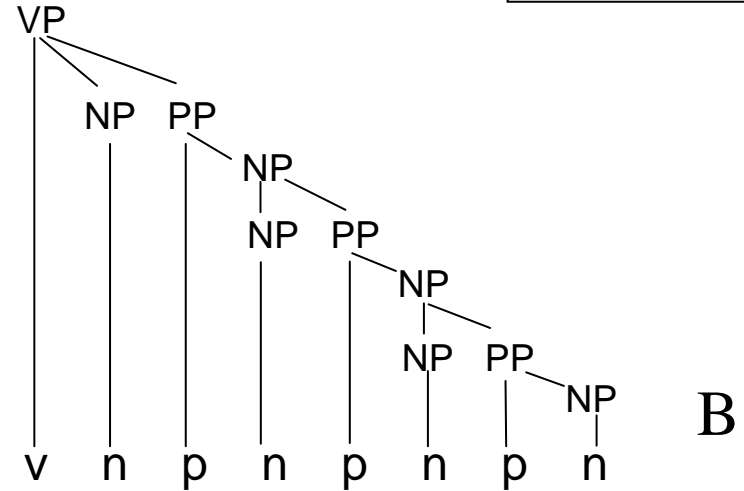
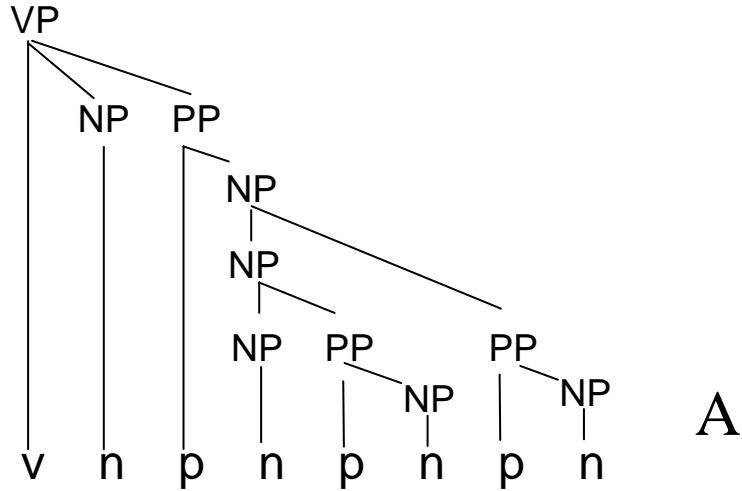
II

pp 个数为2

Put the block in the box on the table in the kitchen

v	np	pp	pp	pp
v	n	p n	p n	p n

NP -> NP PP  
 NP -> n  
 PP -> p NP  
 VP -> v NP PP



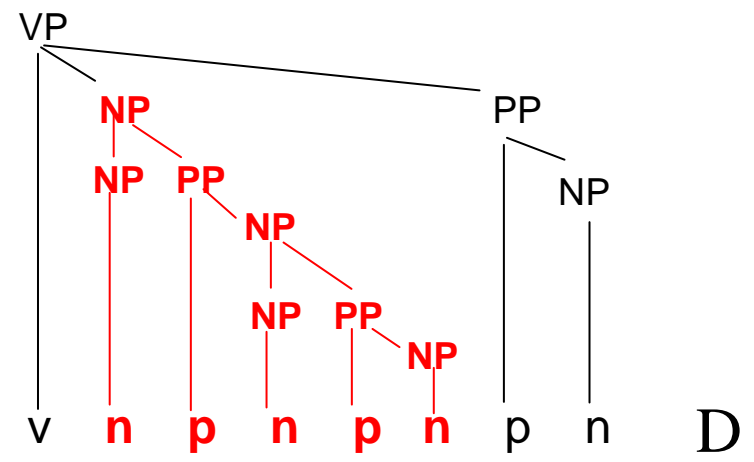
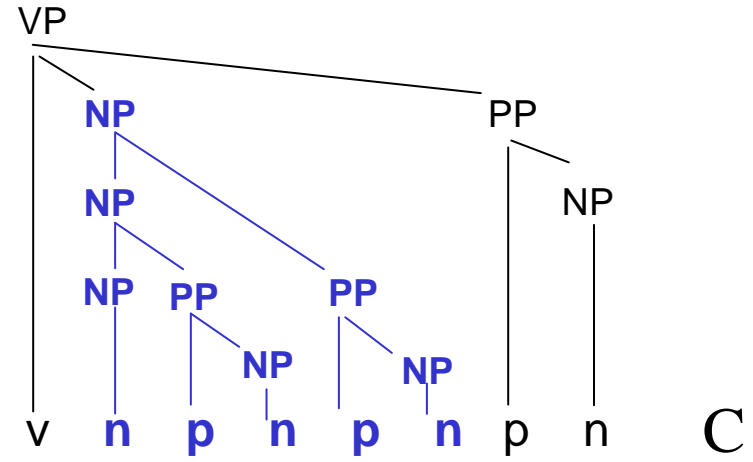
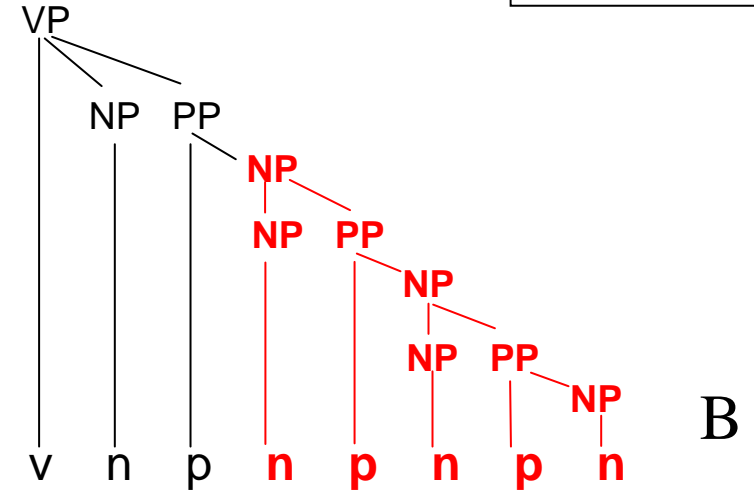
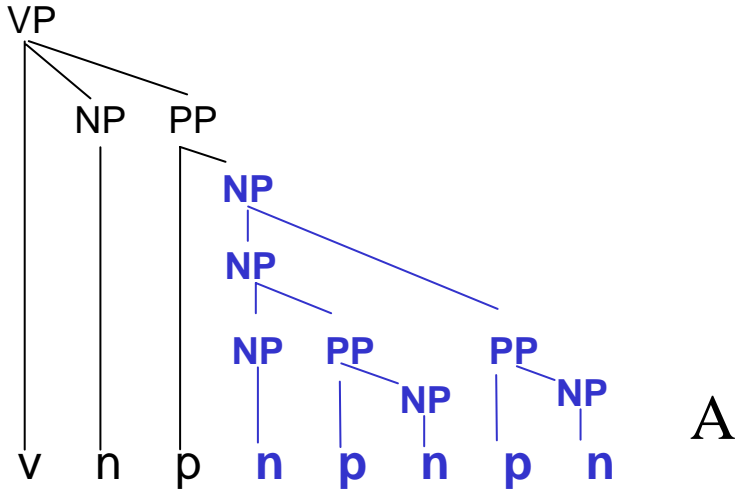
pp 个数为3

Put the block in the box on the table in the kitchen

---

v	np	pp	pp	pp
v	n	p n	p n	p n

NP -> NP PP  
 NP -> n  
 PP -> p NP  
 VP -> v NP PP



pp 个数为3

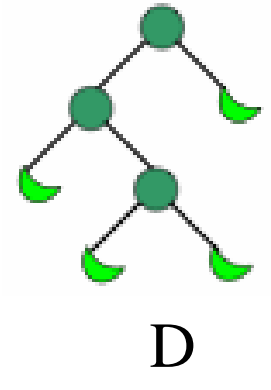
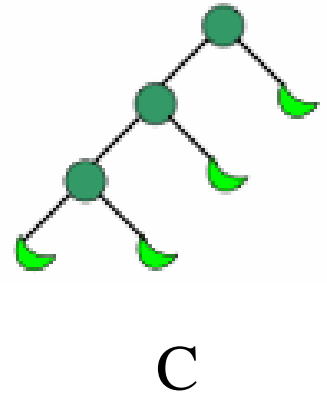
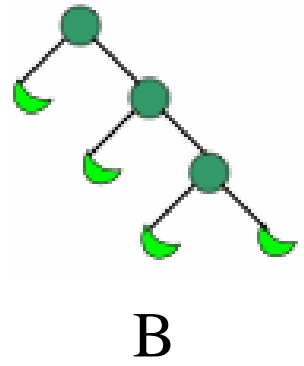
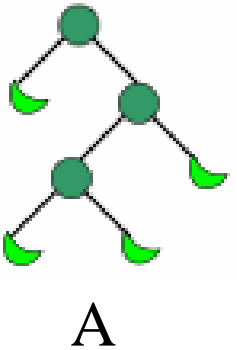
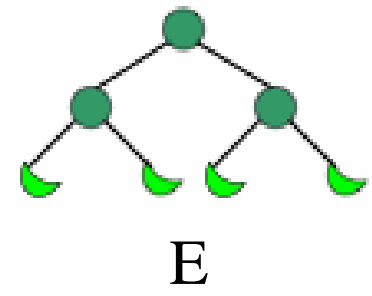
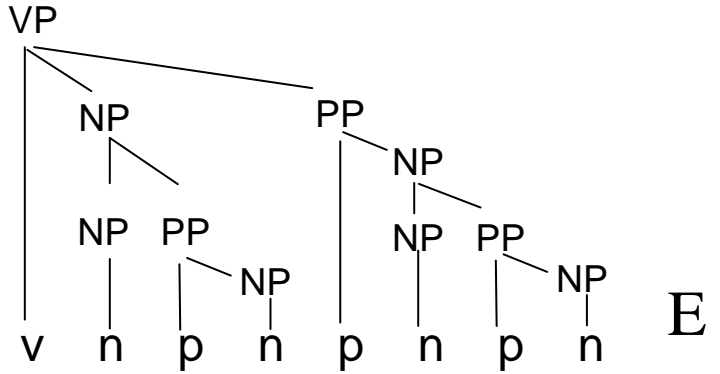
Put	the block	in the box	on the table	in the kitchen
v	np	pp	pp	pp
v	n	p n	p n	p n

NP -> NP PP

NP -> n

PP -> p NP

VP -> v NP PP



pp 个数为3

✓	1	(( ( ) ) )	11	) ( ( ( ) )
✓	2	( ( ) ( ) )	12	) ( ( ) ( )
✓	3	( ( ) ) ( )	13	) ( ( ) ) (
	4	( ( ) ) ) (	14	) ( ) ( ( )
✓	5	( ) ( ( ) )	15	) ( ) ( ) (
✓	6	( ) ( ) ( )	16	) ( ) ) ( (
	7	( ) ( ) ) (	17	) ) ( ( ( )
	8	( ) ) ( ( )	18	) ) ( ( ) (
	9	( ) ) ( ) (	19	) ) ( ) ( (
	10	( ) ) ) ( (	20	) ) ) ( ( (

✓	1	0 0 0 1 1 1	11	1 0 0 0 1 1
✓	2	0 0 1 0 1 1	12	1 0 0 1 0 1
✓	3	0 0 1 1 0 1	13	1 0 0 1 1 0
	4	0 0 1 1 1 0	14	1 0 1 0 0 1
✓	5	0 1 0 0 1 1	15	1 0 1 0 1 0
✓	6	0 1 0 1 0 1	16	1 0 1 1 0 0
	7	0 1 0 1 1 0	17	1 1 0 0 0 1
	8	0 1 1 0 0 1	18	1 1 0 0 1 0
	9	0 1 1 0 1 0	19	1 1 0 1 0 0
	10	0 1 1 1 0 0	20	1 1 1 0 0 0

## 关于Catalan number公式的说明

- 条件1: 左右括号数相等
- 条件2: 在任意位置, 左括号数不少于右括号数

$$\begin{aligned} \binom{2n}{n} - \binom{2n}{n-1} &= \frac{(2n)!}{n! \times n!} - \frac{(2n)!}{(n-1)!(n+1)!} \\ &= \frac{1}{n+1} \binom{2n}{n} \end{aligned}$$

### 三 短语结构歧义的统计分析

- 1 在格式层面（非终结符序列）考察歧义
- 2 歧义的量化考察
- 3 对一种语言的结构歧义情况的总体把握

$n^m$  种排列格式， $n$ 是非终结符个数， $m$ 是格式中包含的符号数

## 以 np, vp, ap 三个非终结符的排列为例

np np np

np np vp

np np ap

np vp np

np vp vp

np vp ap

np ap np

np ap vp

np ap ap

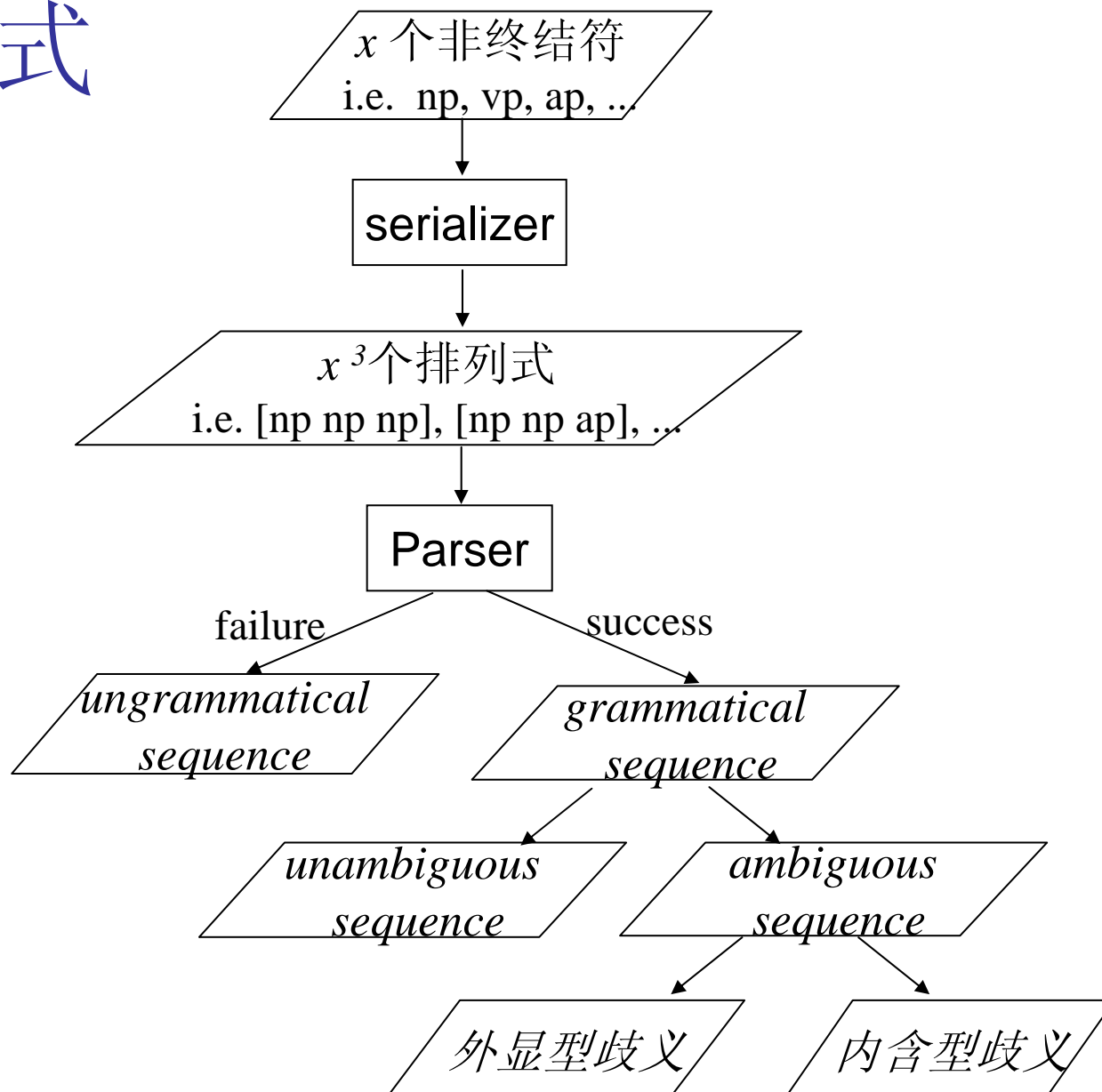
vp np np

.....

- 1 哪些格式有潜在歧义?
- 2 是外显型歧义还是内含型歧义?
- 3 一个有潜在歧义的格式歧义程度如何?

← 从“类”到“例”的观察视角 ↑

# 统计方式



$$9^3 = 729$$

np, tp, sp, mp, ap, dp, pp, vp, dj

可能形成合法结构的排列: 369 个		不可能形成合法结构的排列: 360 个	
np np np np np mp np np tp np np sp .....		np np dp np np pp np mp sp np mp dp .....	
有歧义的排列式: 285 个		无歧义的排列式: 84 个	
外显型歧义格式: 194 个	内含型歧义格式: 91 个	np mp np np mp tp np mp dj np ap dj .....	dj mp mp dj mp tp dj mp sp dj mp dp ..... pp tp sp pp tp dp pp tp pp
np np np np np ap np np vp np vp vp .....	np np mp np np tp np np sp np np dj .....		

外显型歧义格式 (共 194 个)	歧义指数	内含型歧义格式 (共 91 个)	歧义指数
[1] vp vp vp	43	[1] vp ap np	5
[2] vp vp ap	34	[2] dj vp vp	5
[3] vp ap ap	25	[3] np sp dj	4
.....		.....	
[194] pp sp vp	2	[91] pp pp pp	2
平均歧义数	6.55	平均歧义数	2.37

格式举例： np np np

[1](dj:主谓(np,dj:主谓(np,np)))

[2](dj:主谓(np,np:定中(np,np)))

[3](np:定中(np,np:定中(np,np)))

[4](np:联合(np,np:定中(np,np)))

[5](dj:主谓(np,np:联合(np,np)))

[6](np:定中(np,np:联合(np,np)))

[7](np:联合(np,np:联合(np,np)))

[8](dj:主谓(np:定中(np,np),np))

[9](np:定中(np:定中(np,np),np))

[10](np:联合(np:定中(np,np),np))

[11](dj:主谓(np:联合(np,np),np))

[12](np:定中(np:联合(np,np),np))

[13](np:联合(np:联合(np,np),np))

13种可能的分析结果！

# 歧义格式统计研究的意义

- 1 评估一个具体的歧义格式的歧义程度
- 2 评估非终结符的设置（分类）的合理性

对“真歧义、准歧义”进行统计，需要树库数据作为基础

# 话题句为何难以关系化？

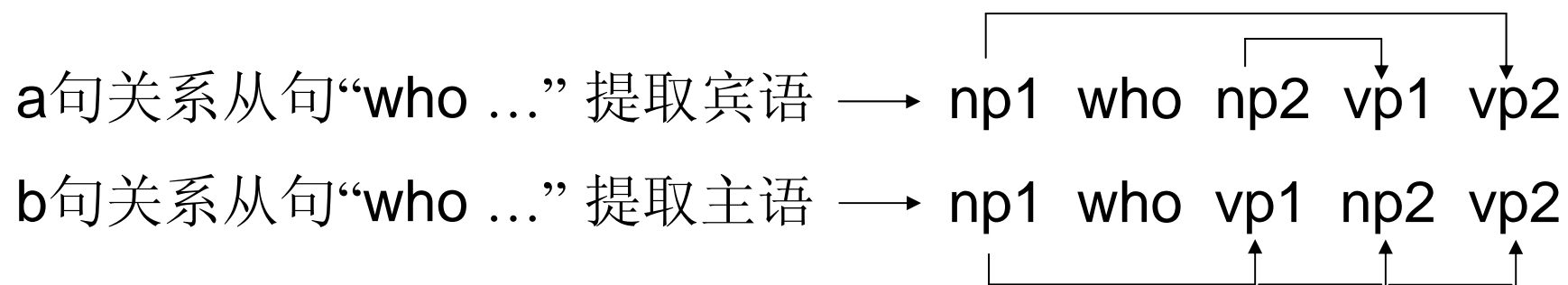
格式	结构数量
■ np vp np u<的> np	40
■ np np vp u<的> np	79

- 1 张三吃饺子      ———>      1' 张三吃饺子的碗
- 2 饺子张三吃      ———>      2' 饺子张三吃的碗      ✗

歧义数 与 句子的认知加工难度是否存在联系？

# 英语关系从句的加工难度

- a. The reporter who the senator attacked admitted the error.
- b. The reporter who attacked the senator admitted the error.



# 五 结 语

- 1 计算机“眼”里的歧义远远多于人眼里的歧义
- 2 应该区分计算机所面临的歧义的不同类型，有针对性地寻求消解歧义的方法
- 3 对于外显型歧义格式，有可能在短语结构规则层面消解歧义；对于内含型歧义格式，如果是准歧义，有可能在短语结构规则层面消解歧义；如果是真歧义，不可能在短语结构规则层面消解歧义。
- 4 区分不同的歧义类型，也有助于面向人的语言教学

一种语言语法系统里的错综复杂和精细奥妙之处往往在歧义现象里得到反映。因此分析歧义现象会给我们许多有益的启示，使我们对于语法现象的观察和分析更加深入。

—— 朱德熙

汉语句法里的歧义现象，《中国语文》1980.2

# 进一步阅读文献

1. 詹卫东、常宝宝、俞士汶，汉语短语结构定界歧义类型分析及分布统计，《中文信息学报》1999年第3期
2. Jurafsky & Martin(2000) *Speech and Language Processing*, Prentice-Hall, Inc. Chapter 10.3
3. Church, K.W. & Patil, R. 1982, Coping with syntactic ambiguity (or How to put the block in the box on the table), *American Journal of Computational Linguistics*,8(3-4),pps.139-149.