



第三讲 汉语的句法规则系统

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



提纲

一 汉语形式语法系统的构建

- 1 现代汉语的基本语法范畴
- 2 理论语言学的描述方式：X-bar结构理论
- 3 计算语言学的描述方式：上下文无关文法与特征结构合一描述的结合

二 汉语句法结构歧义类型及歧义消解举例

- 1 从人的角度看句法结构歧义
- 2 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 句法结构歧义的统计分析
- 3 歧义消解举例



1 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



2 从计算机的角度看结构歧义

- 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 2.1
- 句子层（终端）歧义 vs. 结构层（模式）歧义 2.2

- 歧义程度的量化分析

- 计算机分析歧义格式可能产生的结构树的数量 2.3



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]]

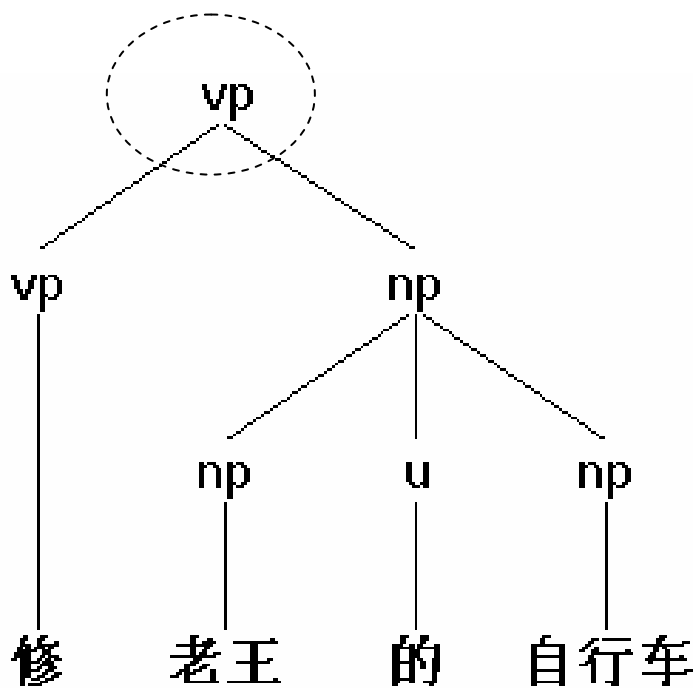
1b [[修 自行车 的] 扳手]

2 ap np np

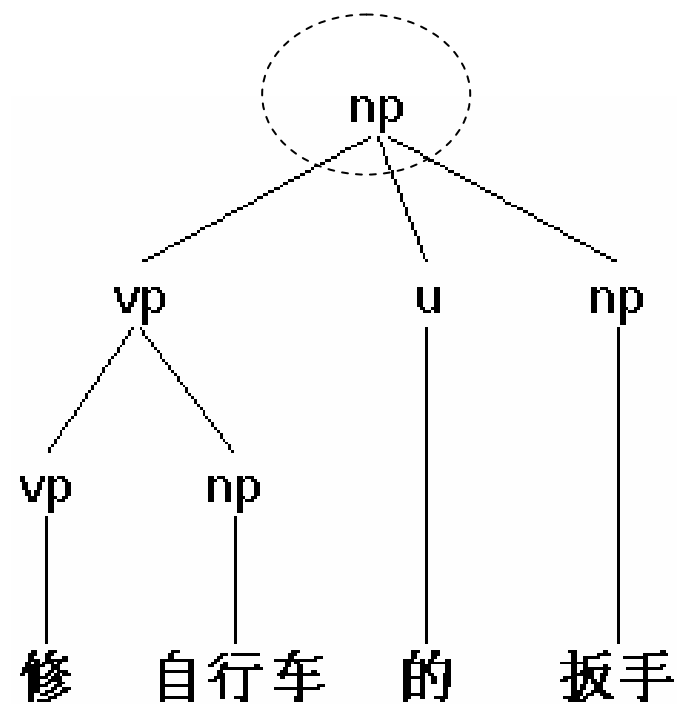
2a [大 [钢铁 公司]]

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

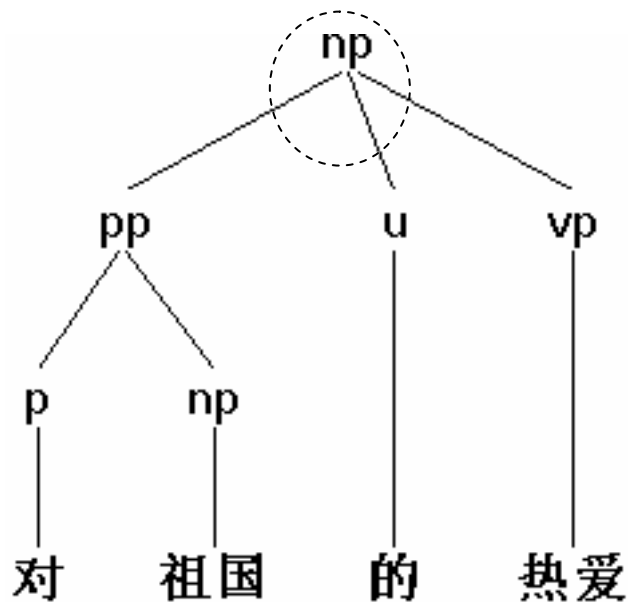
.....

vp

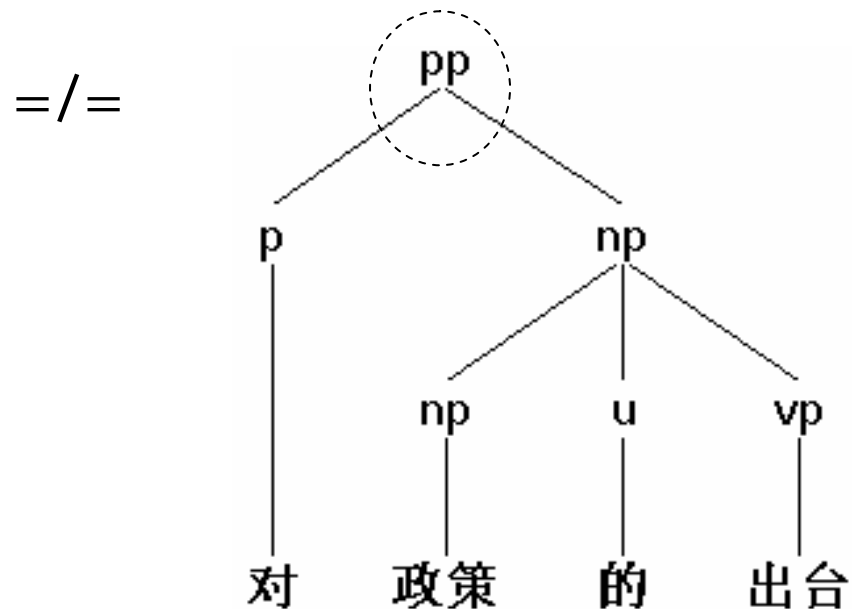
np

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

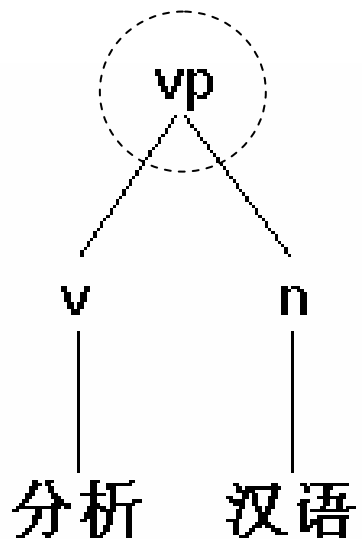
他对校长的批评很尖锐

他对校长的意见持否定态度

他对校长的批评充耳不闻

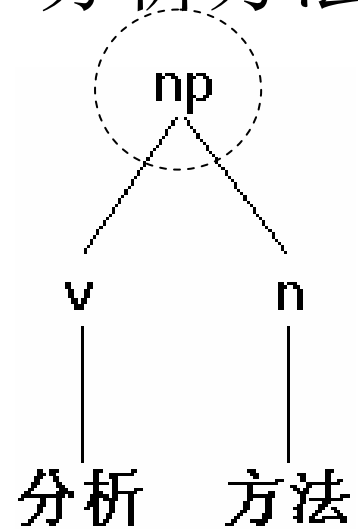
外显型歧义 (续)

分析汉语



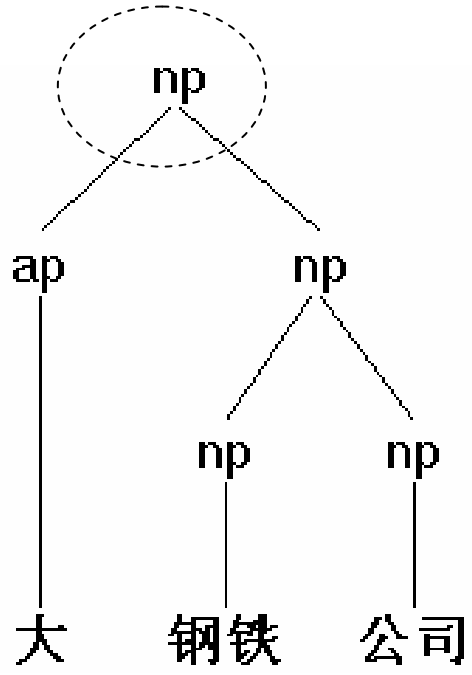
=/=

分析方法

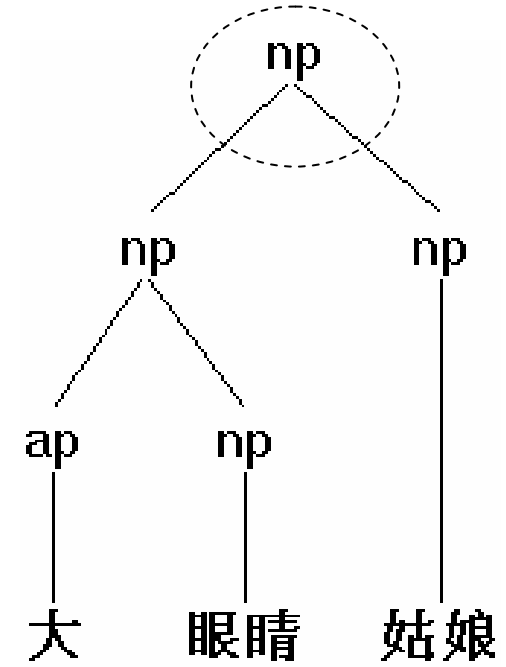


出租汽车 np or vp

内含型歧义

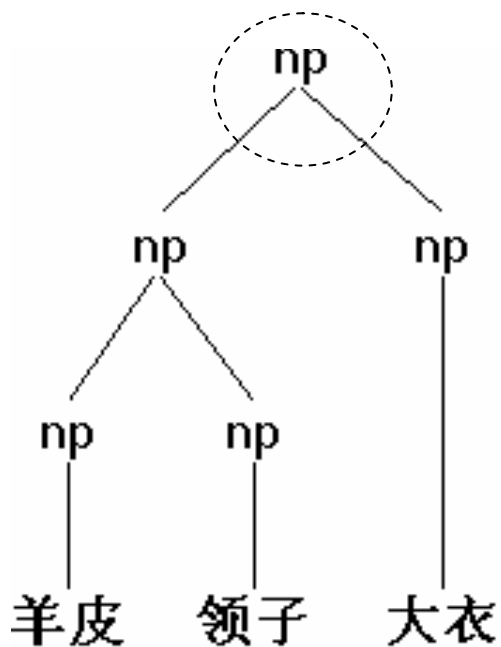


==



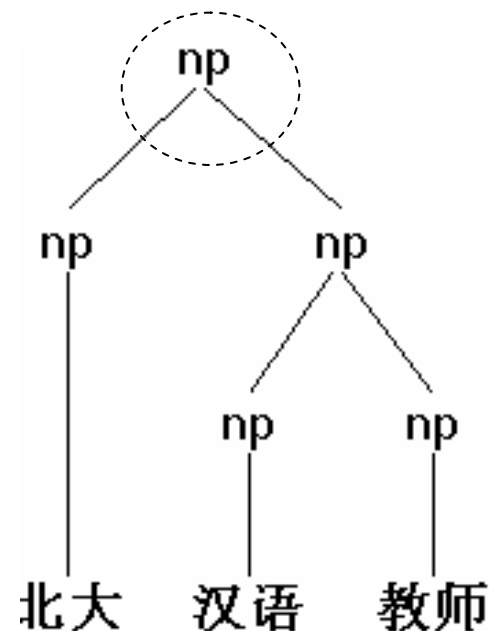
内含型歧义 (续1)

羊皮领子大衣



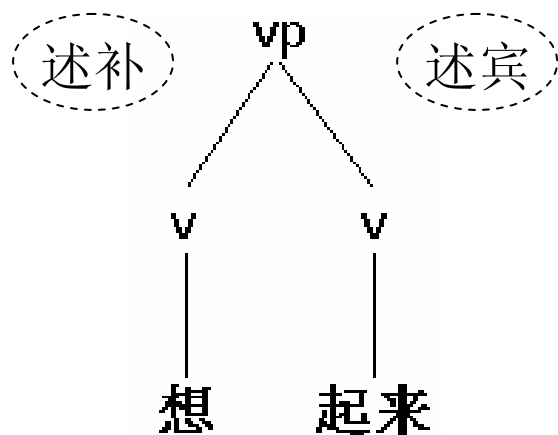
==

北大汉语教师



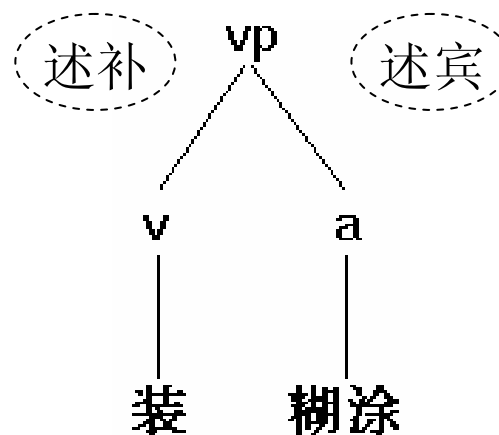
内含型歧义（续2）

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。

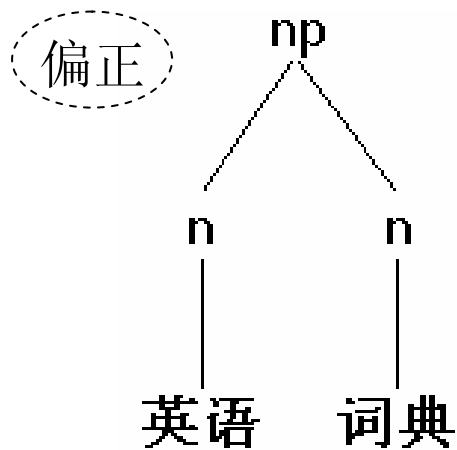
装糊涂



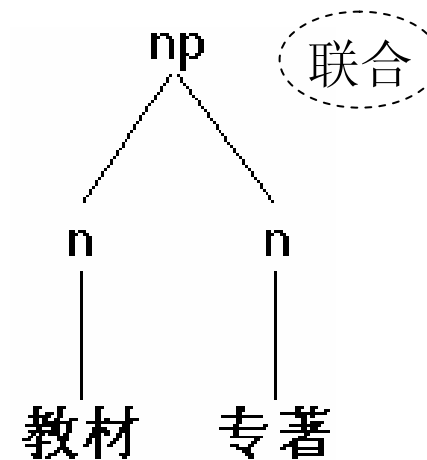
他就会装糊涂，其实他心理比谁都清楚
装了一上午家具，我都装糊涂了

内含型歧义 (续3)

英语词典

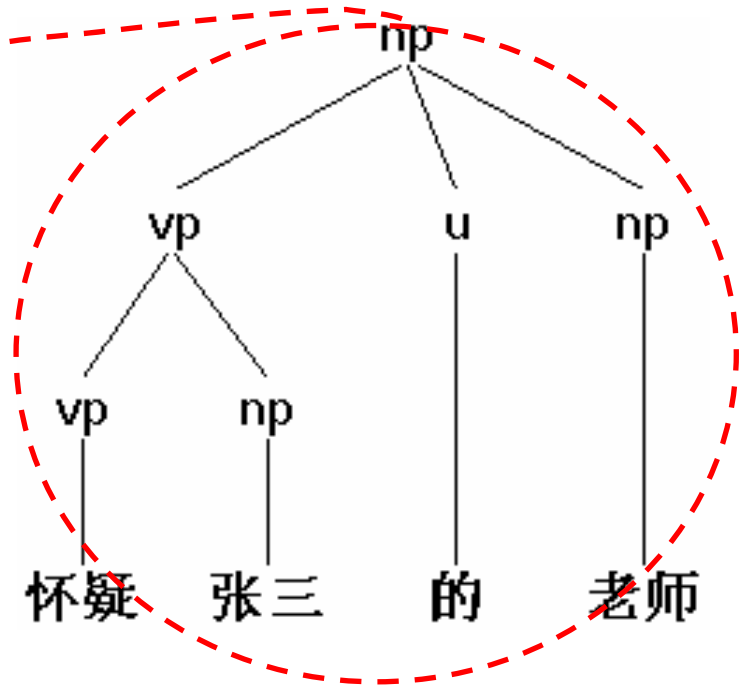
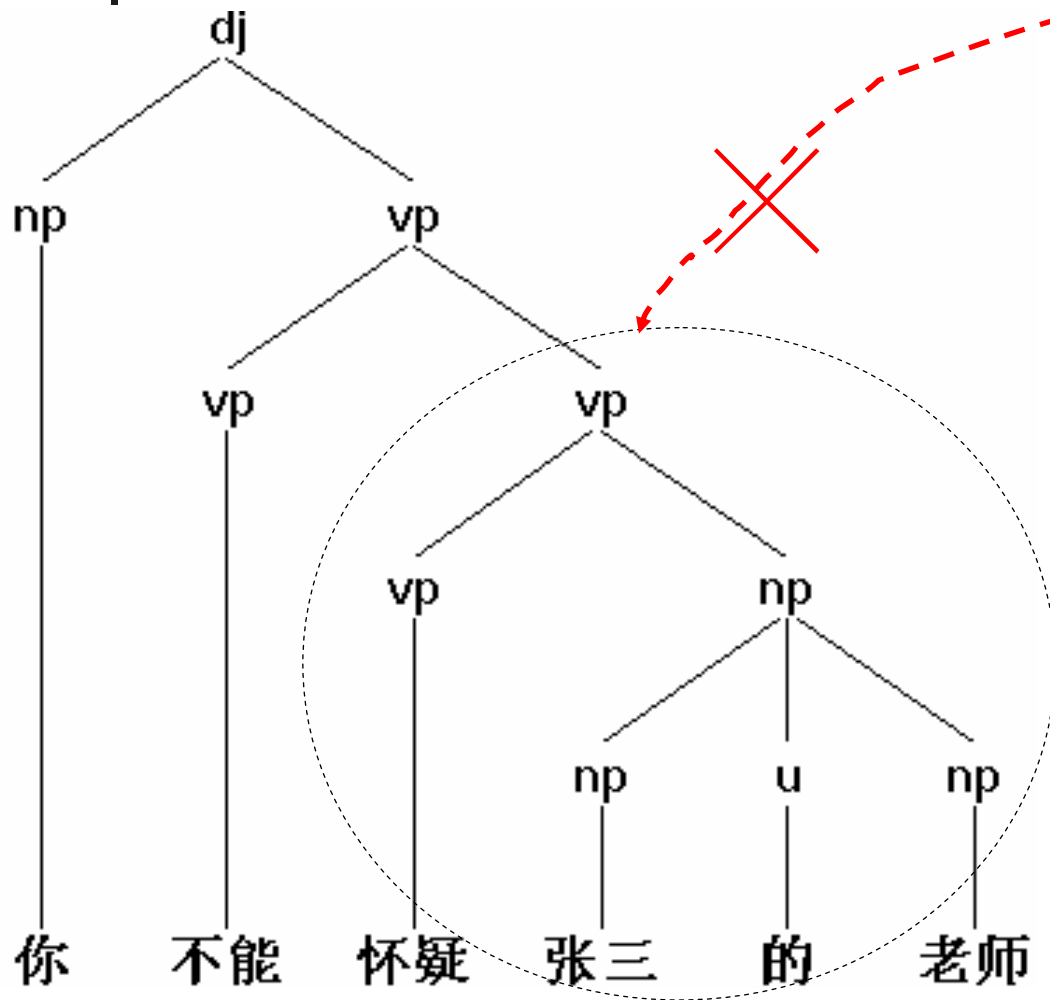


教材专著

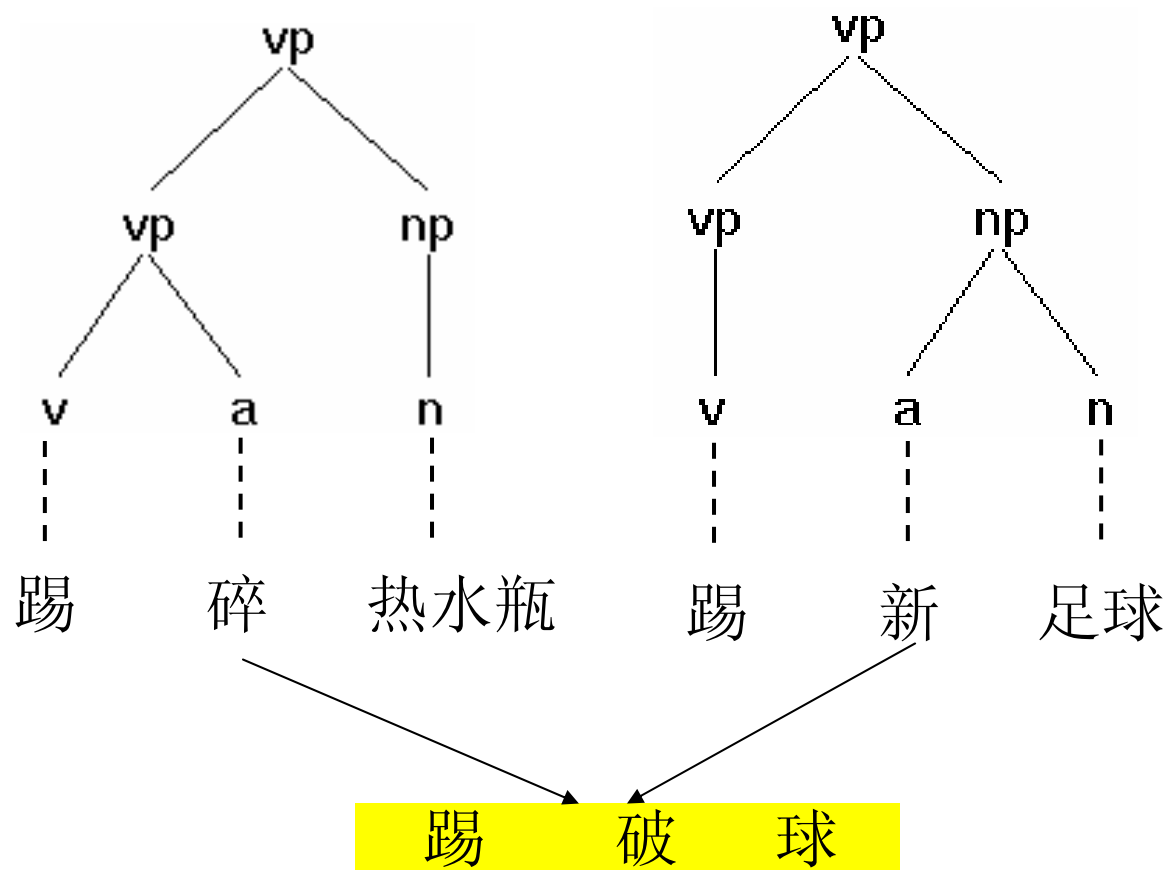


偏正? 牛奶饼干 联合?

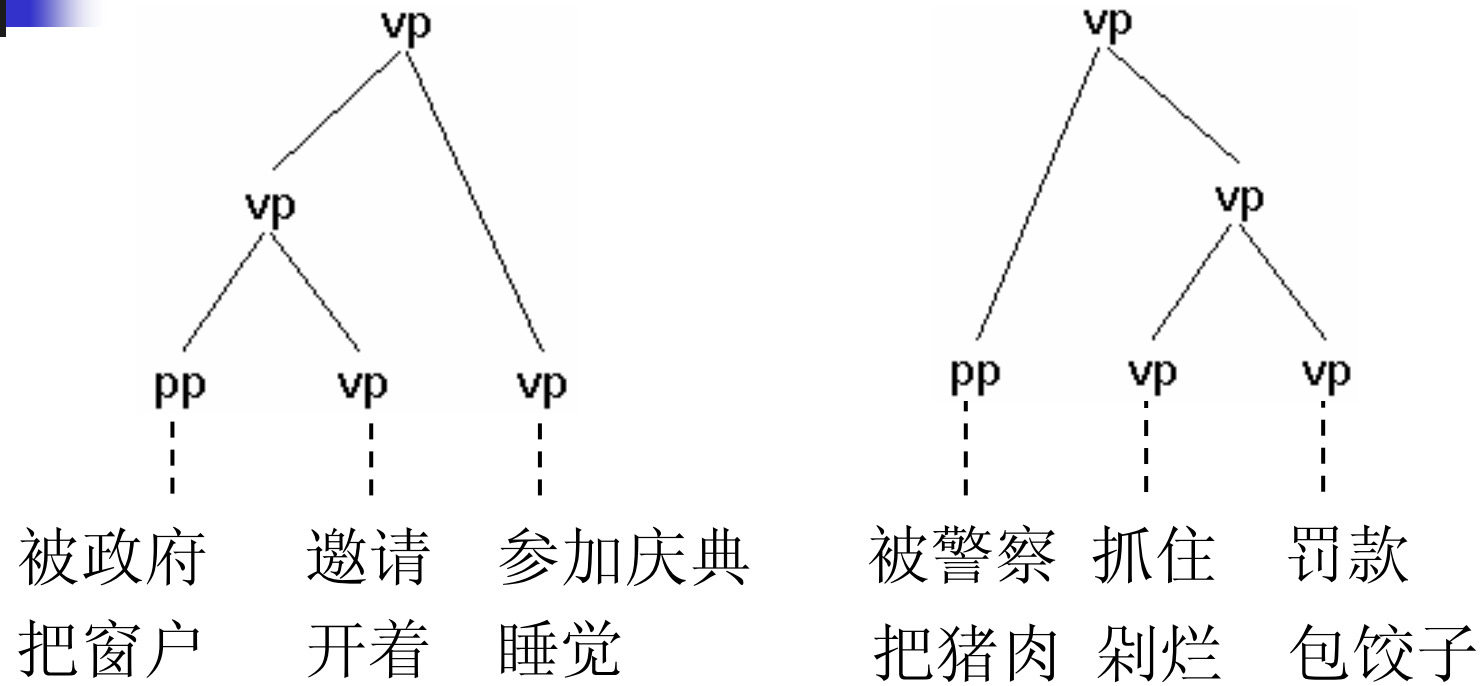
区分“外显”与“内含”的作用



真歧义

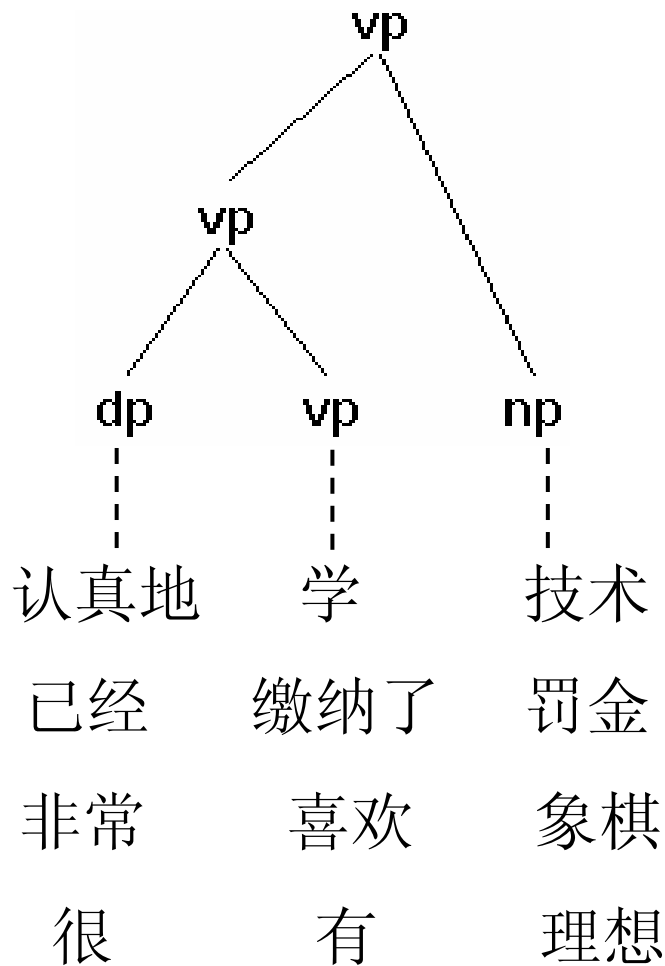
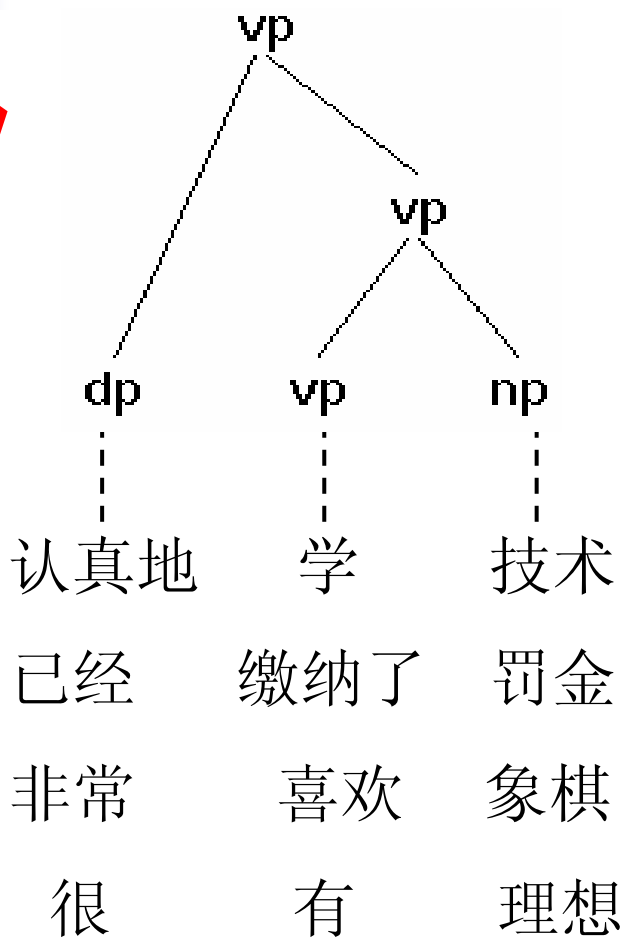


准歧义



??????

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。



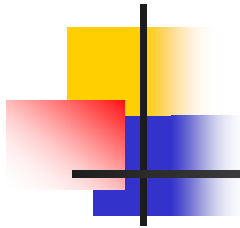
2.3 短语结构歧义的统计分析

- 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

考察对象： n^m 种排列格式

n 是形式语法系统中的非终结符个数；

m 是一个具体格式中包含的符号个数；



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

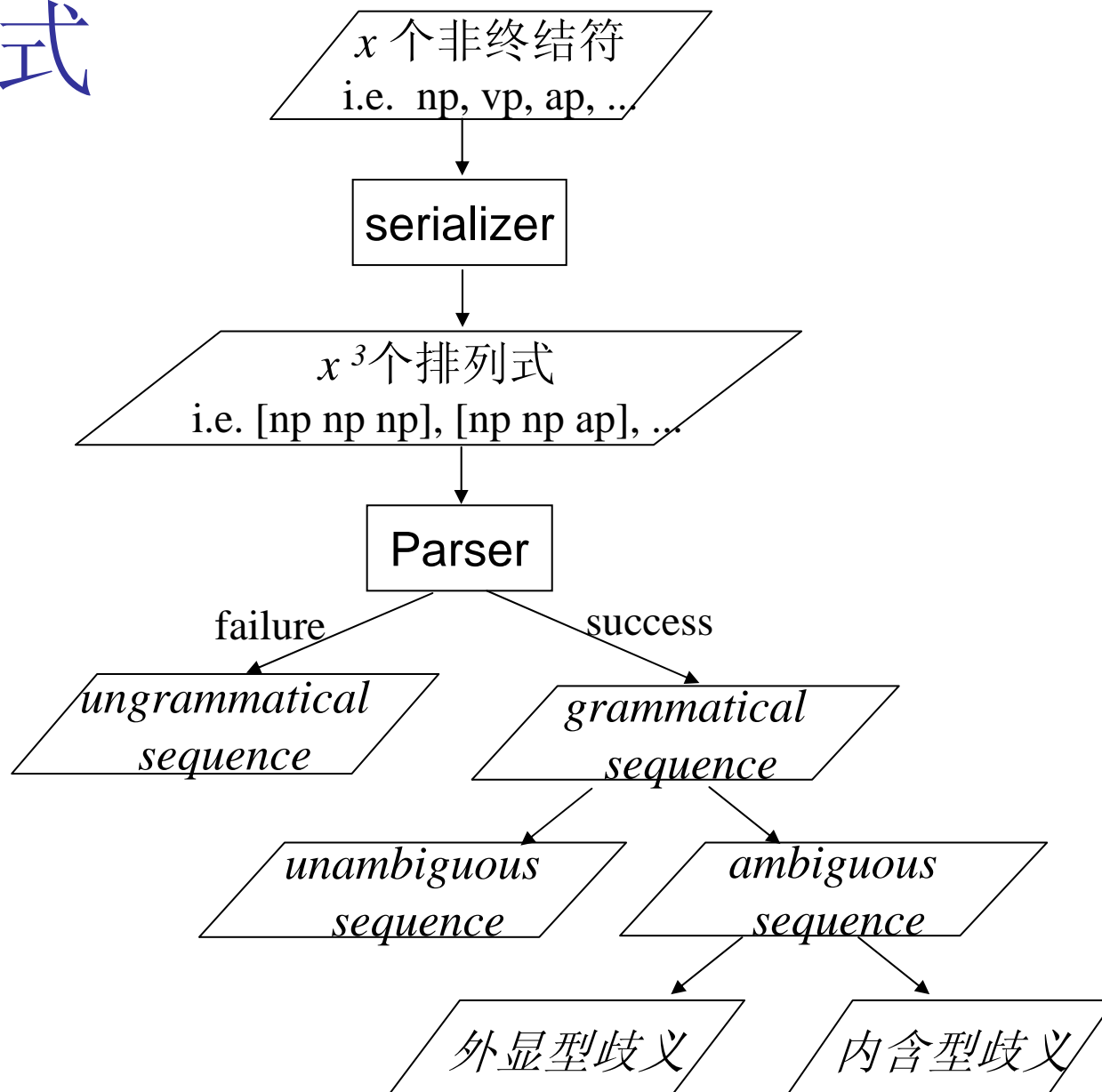
vp np np

.....

- 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

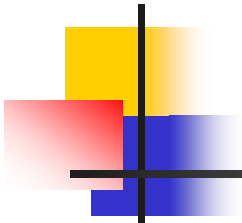
统计方式



$$9^3 = 729$$

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
- 3 从歧义的角度认识语言系统对相关句法格式的选择差异

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？

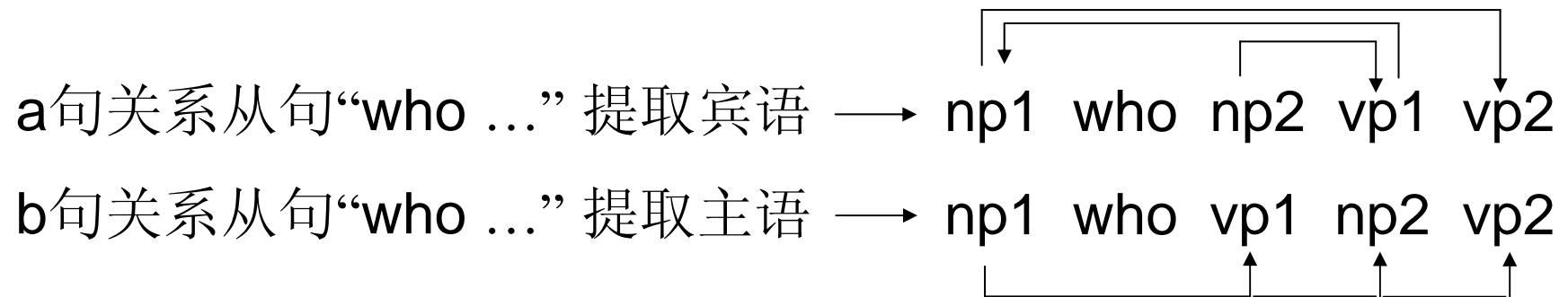
格式	可能的结构树数量
■ np vp np u<的> np	40
■ np np vp u<的> np	79

- 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.





小结

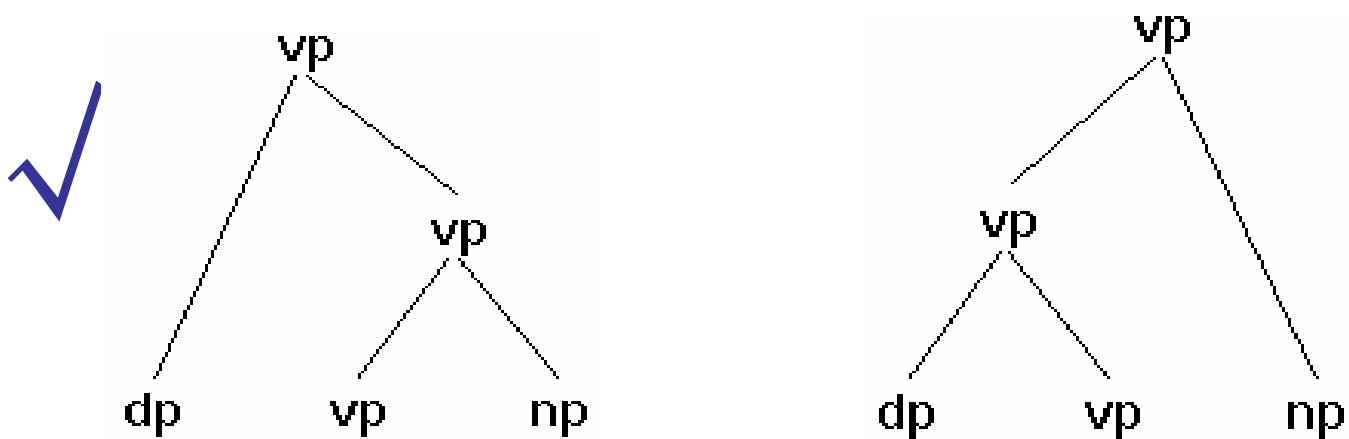
- 1 计算机“眼”里的歧义远远多于人眼里的歧义
- 2 应该区分计算机所面临的歧义的不同类型，有针对性地寻求消解歧义的方法
- 3 对于外显型歧义格式，有可能在短语结构规则层面消解歧义；对于内含型歧义格式，如果是准歧义，有可能在短语结构规则层面消解歧义；如果是真歧义，不可能在短语结构规则层面消解歧义。
- 4 区分不同的歧义类型，也有助于面向人的语言教学



3 歧义消解举例

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

3.1 伪歧义格式的消歧



vp_zz → dp vp_sb

vp_sb → vp np

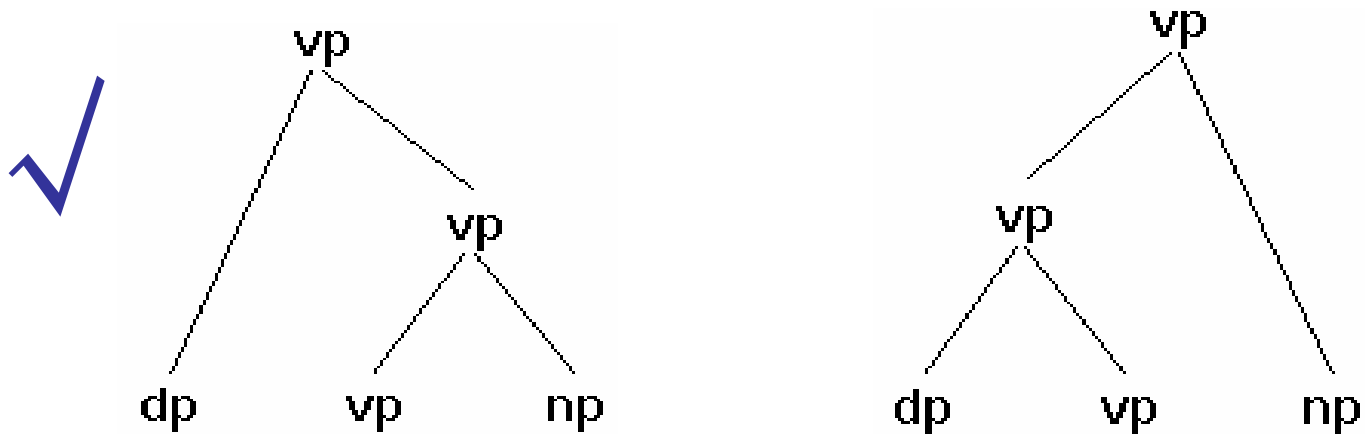
vp_zz: 状中式vp

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

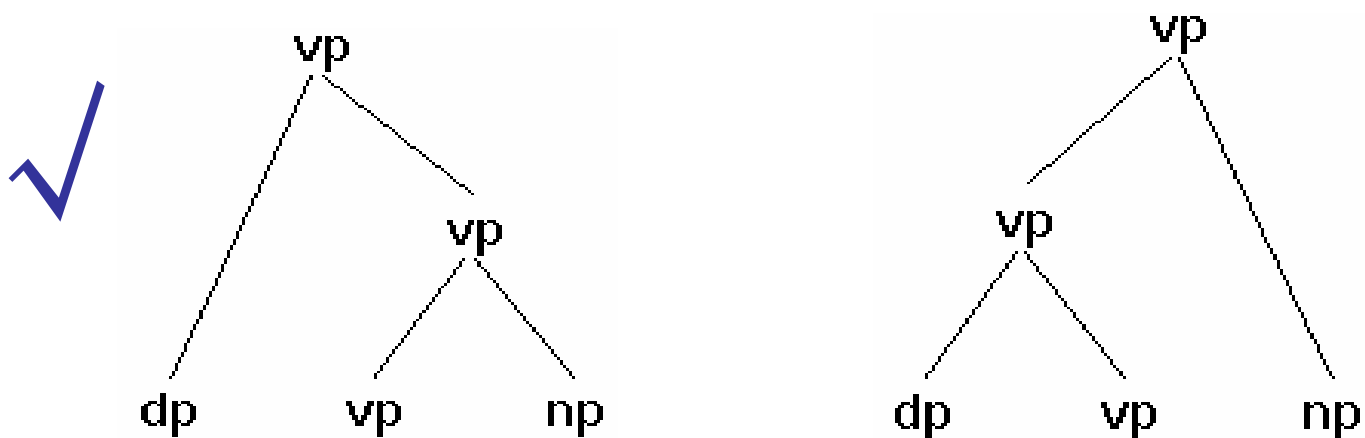


`vp` \rightarrow `dp vp` :: \$.内部结构=状中

`vp` \rightarrow `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



vp → dp vp :: \$.内部结构=状中,\$.daibinyu=否

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

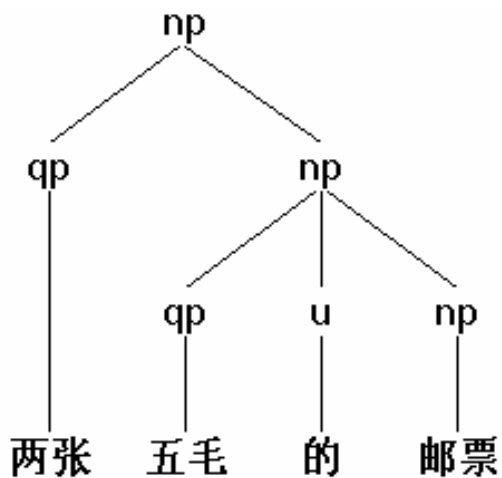
方案III

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

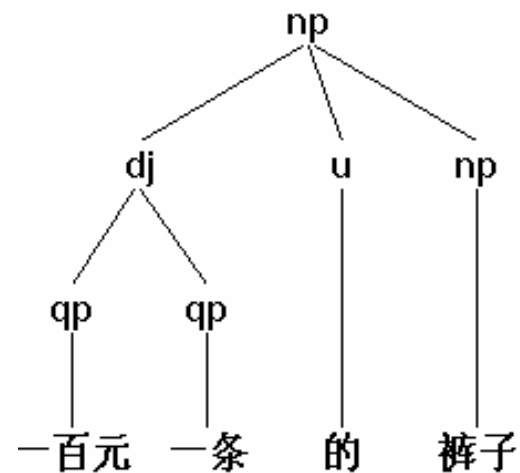
方案III更合理

3.2 准歧义格式的消歧

两张 五毛 的 邮票



一百元一条的裤子



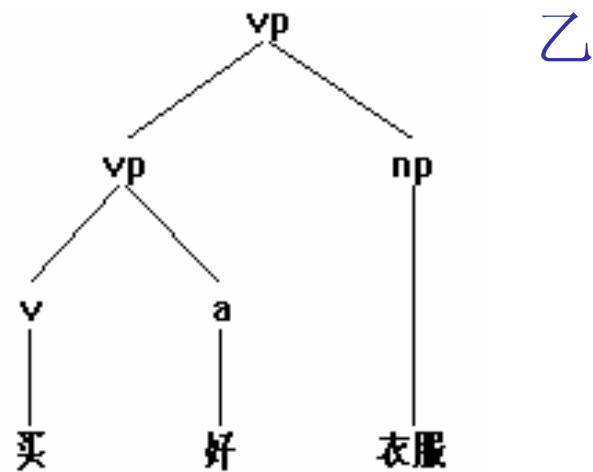
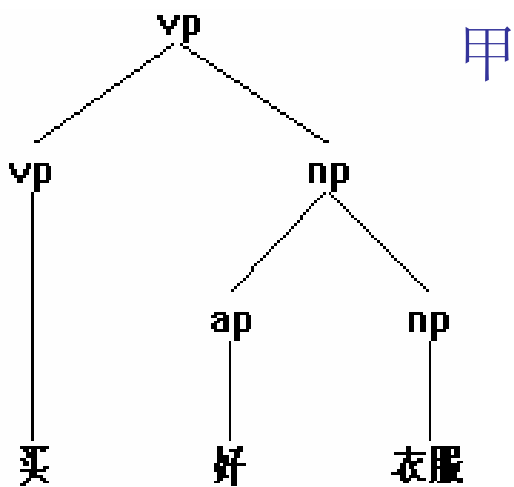
五十元一米的电缆



qp qp 的 np

1. **np → qp 的 np** :: \$.内部结构=定中, %qp.zuodingyu=是,
五毛 的 邮票 %qp.个体量词=否, ...
2. **np → qp np** :: \$.内部结构=定中, %qp.zuodingyu=是,
两张 邮票 IF %qp.个体量词=是 THEN %np.个体量词=%qp, ENDIF...
3. **dj → qp qp** :: \$.内部结构=主谓,
五十元 一斤 IF %qp.量词子类=%qp.量词子类 FALSE,
...

3.3 真歧义格式的消歧



如何给出区分甲和乙的判别条件？



v a n 的相关规则

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

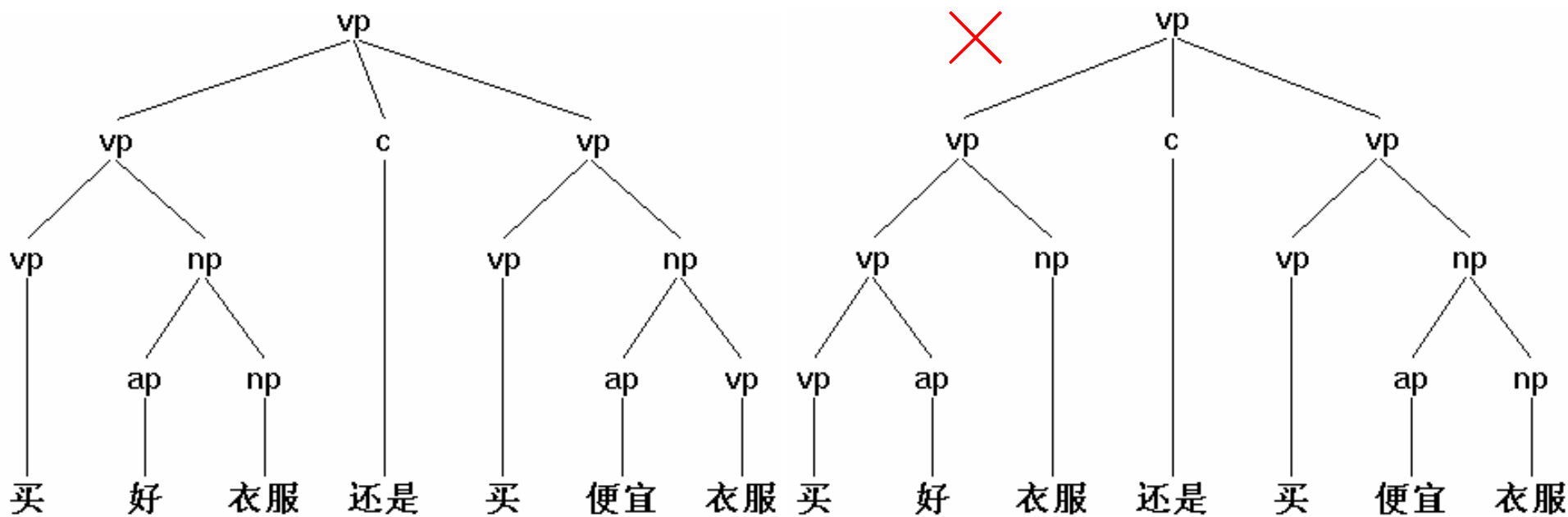
3. vp → vp np :: \$.内部结构=述补, %vp.daibinyu=是, ...
踢 球

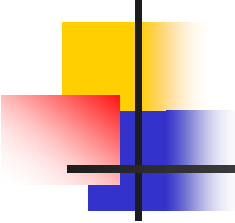
买 好 衣服

“好”同时满足规则1跟规则2

“v a n”出现在并列结构中

- 你打算买好衣服还是买便宜衣服





小结

- 句法分析就是消歧的过程。
- 消歧中可以依据的句法知识就是各级语言单位（词和短语）的功能特征，包括：
 - （1）记录在词典中的每个词语占据句法结构位置的能力；
 - （2）通过规则的合一约束描述的短语的功能特征；
 - （3）从词到短语，从小的短语到大的短语，功能特征的具体取值处于动态变化之中



思考题

1. 从课程网页上下载歧义格式分析程序，对一个你感兴趣的可能有歧义的序列（至少含3个以上非终结符）进行分析，得出各种可能的结构分析方式，并找出消歧的条件。



进一步阅读文献

- 冯志伟等译（2005）《自然语言处理综论》第8.1，8.2，第9章。
- 詹卫东（2000），《面向中文信息处理的现代汉语短语结构规则研究》，清华大学出版社，广西科学技术出版社。第1，2，3，4章。
- Robert D. Borsley, 1996, *Modern Phrase Structure Grammar*, No. 11 in Blackwell textbooks in Linguistics, Blackwell Publishers Inc..
- Sag, Ivan A. & Thomas Wasow, 1999, *Syntactic Theory: A Formal Introduction*, CSLI Publications, Stanford, California.
- 陆致极（1986），《关于广义短语结构语法》，载《国外语言学》1986年第4期。
- 姚天顺 等（1995），《自然语言理解》，清华大学出版社，广西科学技术出版社。
- 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
- Jurafsky & Martin(2000) *Speech and Language Processing*, Prentice-Hall, Inc. Chapter 10.3
- 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.



附录：基于简单CFG语法分析句子结构，可能产生的歧义结构的数量：Catalan number

■ Catalan number的计算公式

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

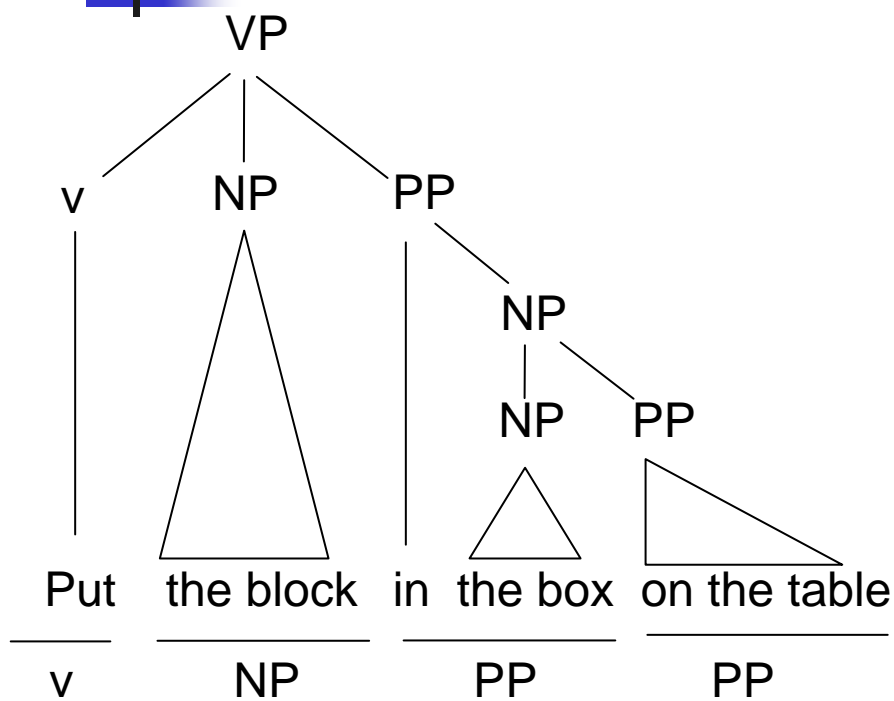
实例：设含有pp-attachment的英语句子中存在 n (n 为自然数)个介词短语，则计算机基于简单CFG语法进行句法分析，得到结果的数目（记作 C_N ）是一个Catalan数。



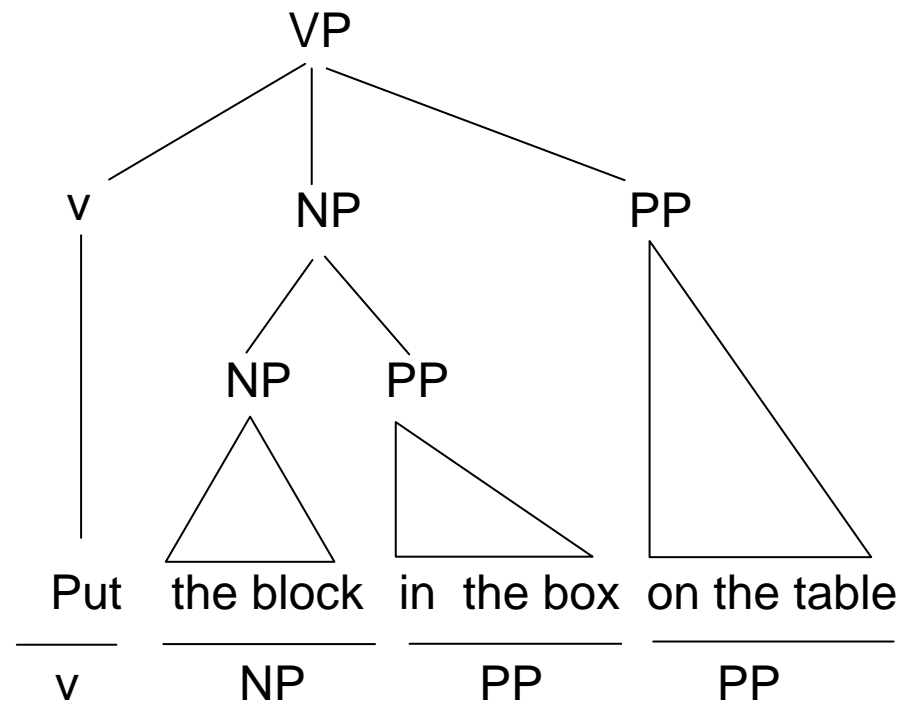
Catalan number

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

pp-attachment歧义实例



I



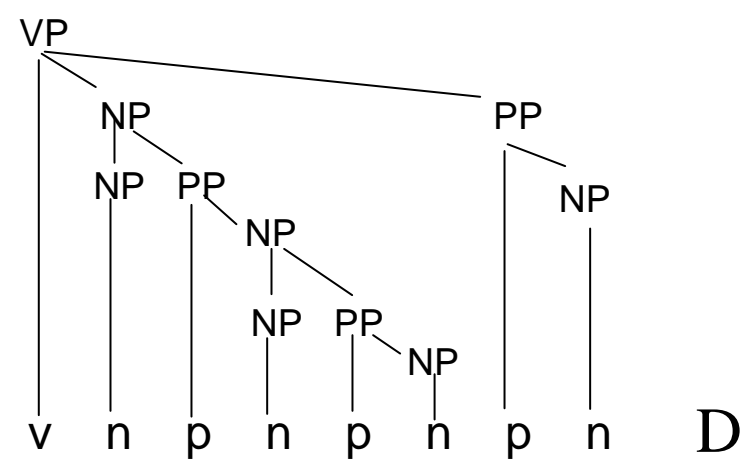
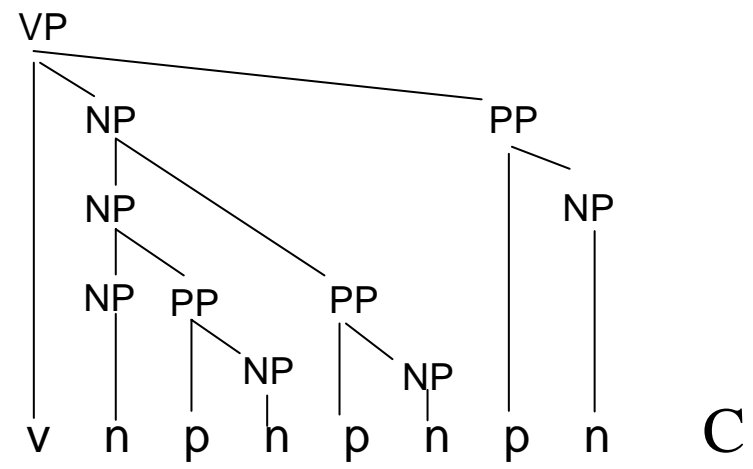
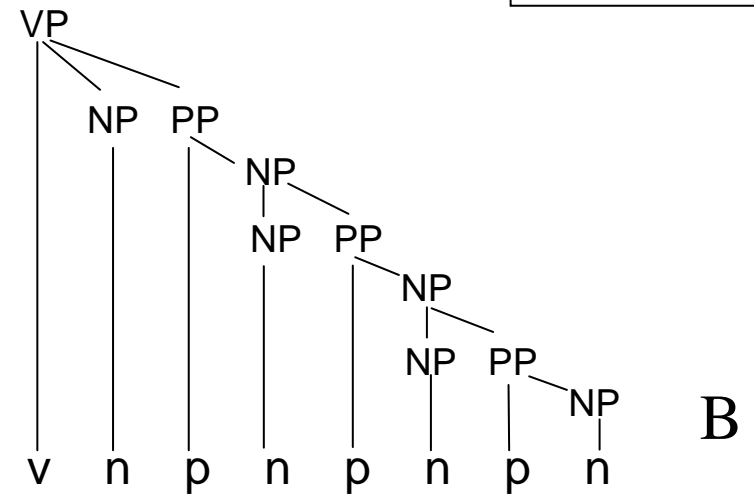
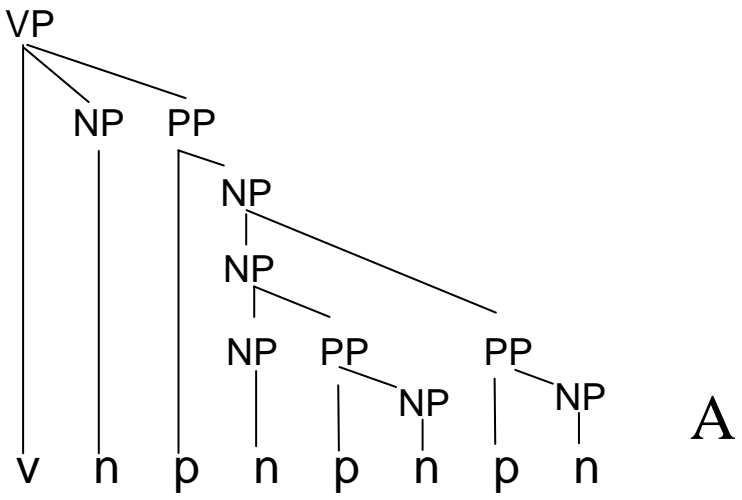
II

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

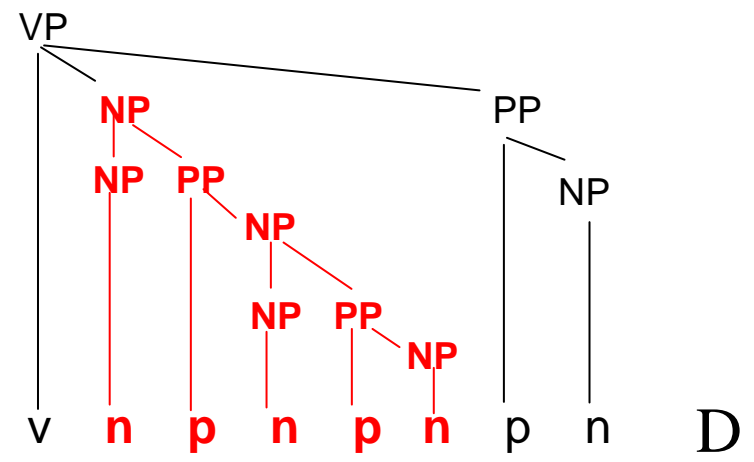
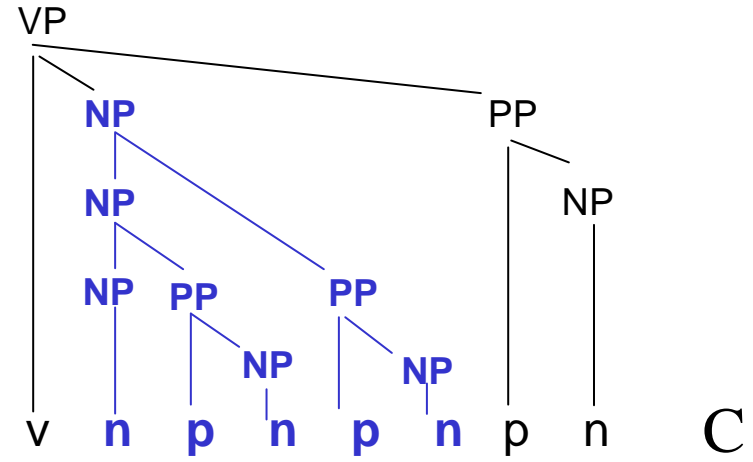
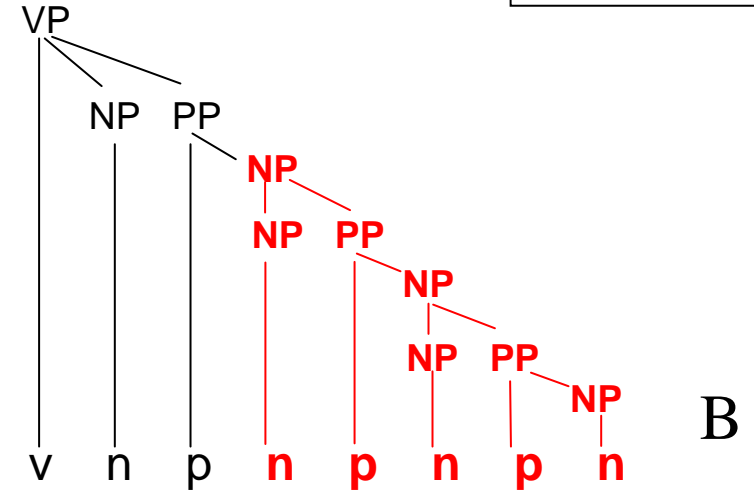
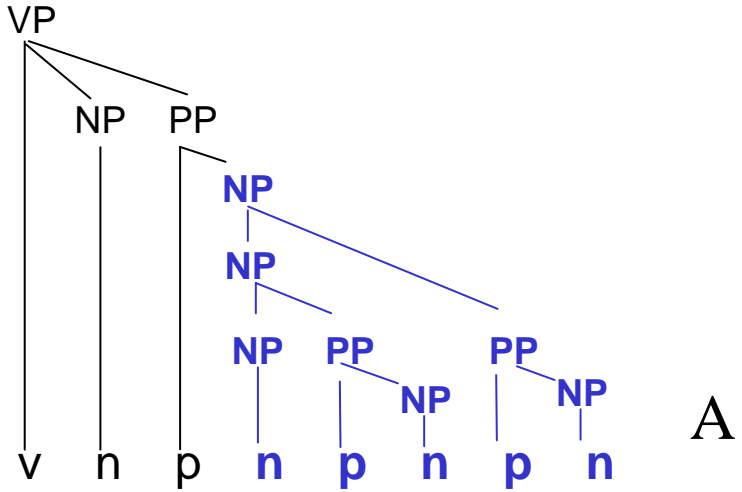


pp 个数为3

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

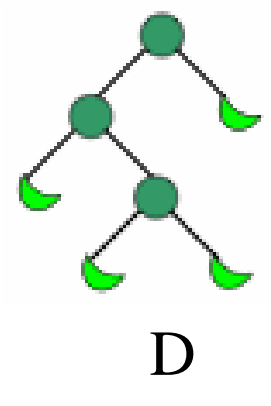
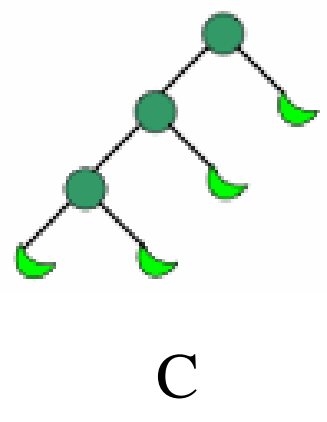
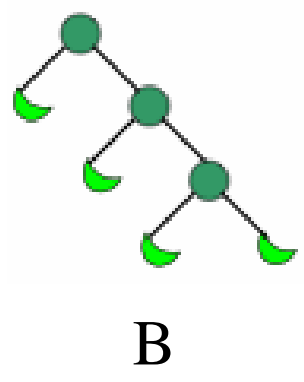
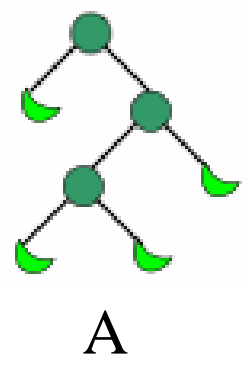
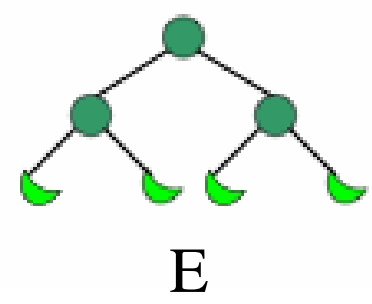
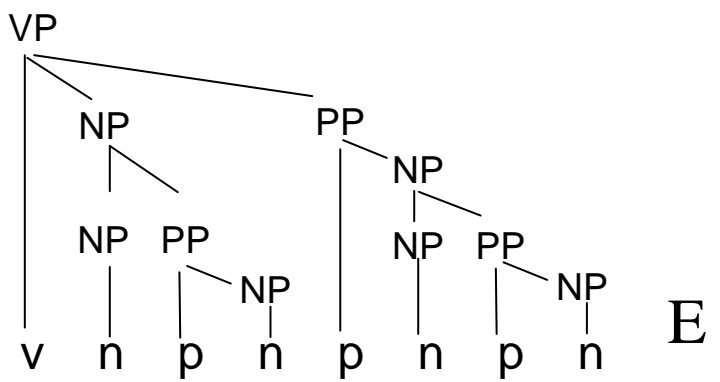
NP -> NP PP
 NP -> n
 PP -> p NP
 VP -> v NP PP



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数计算公式的说明

- 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

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

Donald E.Knuth著 苏运霖译, 《计算机程序设计艺术》(第三版) 第一卷。508页。国防工业出版社。