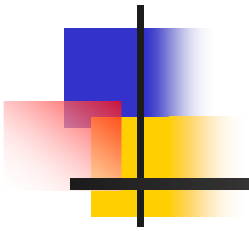


从形式文法看自然语言的歧义

詹卫东

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

提纲

一 外显型歧义与内含型歧义

二 真歧义、准歧义、伪歧义

三 含终结符的歧义格式

四 短语结构歧义的统计分析

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

一 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]

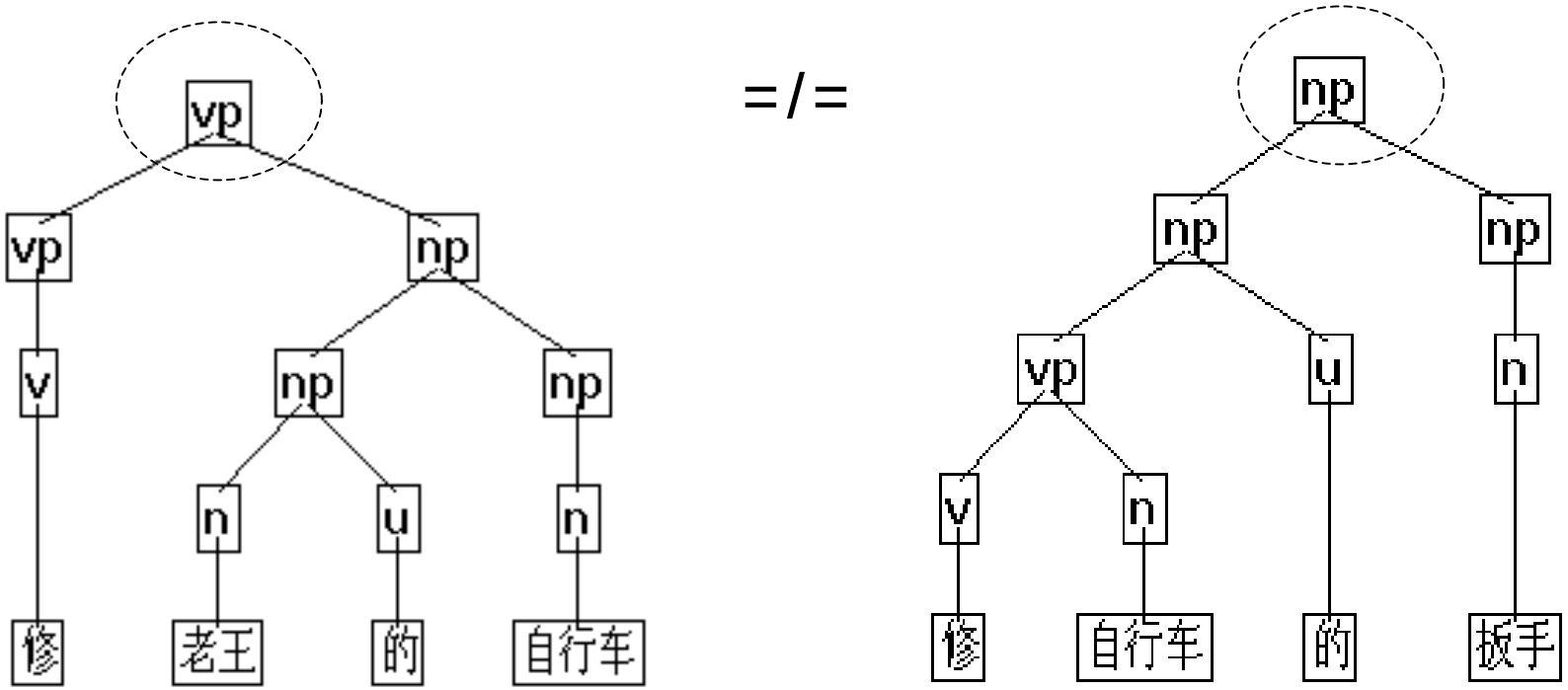
1b [[修 自行车 的] 扳手]

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　　罢工　的　学生

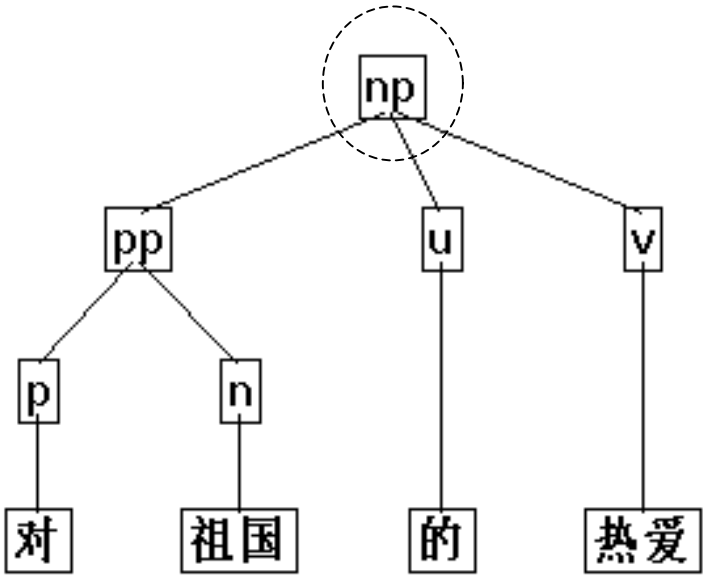
擦洗　　干净　的　桌子

……

vp　　np

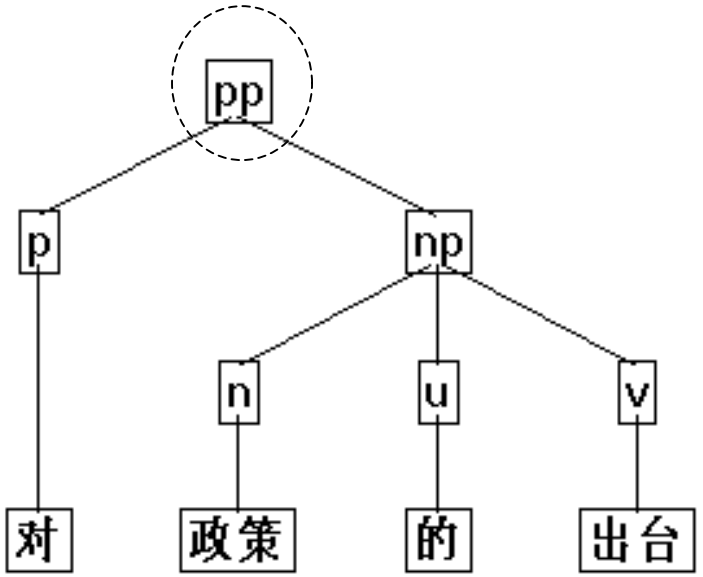
外显型歧义（续2）

对祖国的热爱



=/=

对政策的出台



他对校长的意见很大

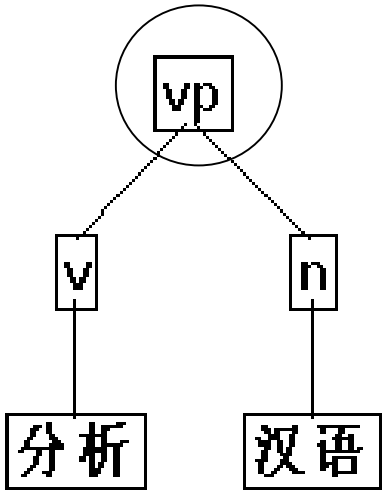
他对校长的批评很尖锐

他对校长的意见持否定态度

他对校长的批评充耳不闻

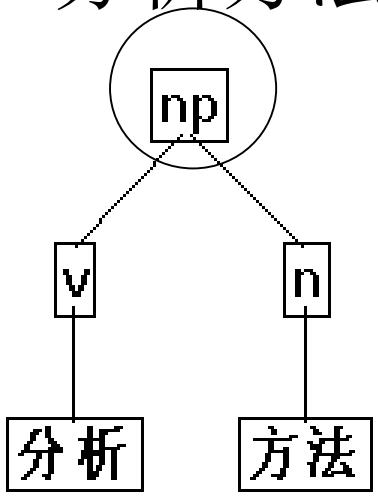
外显型歧义（续）

分析汉语



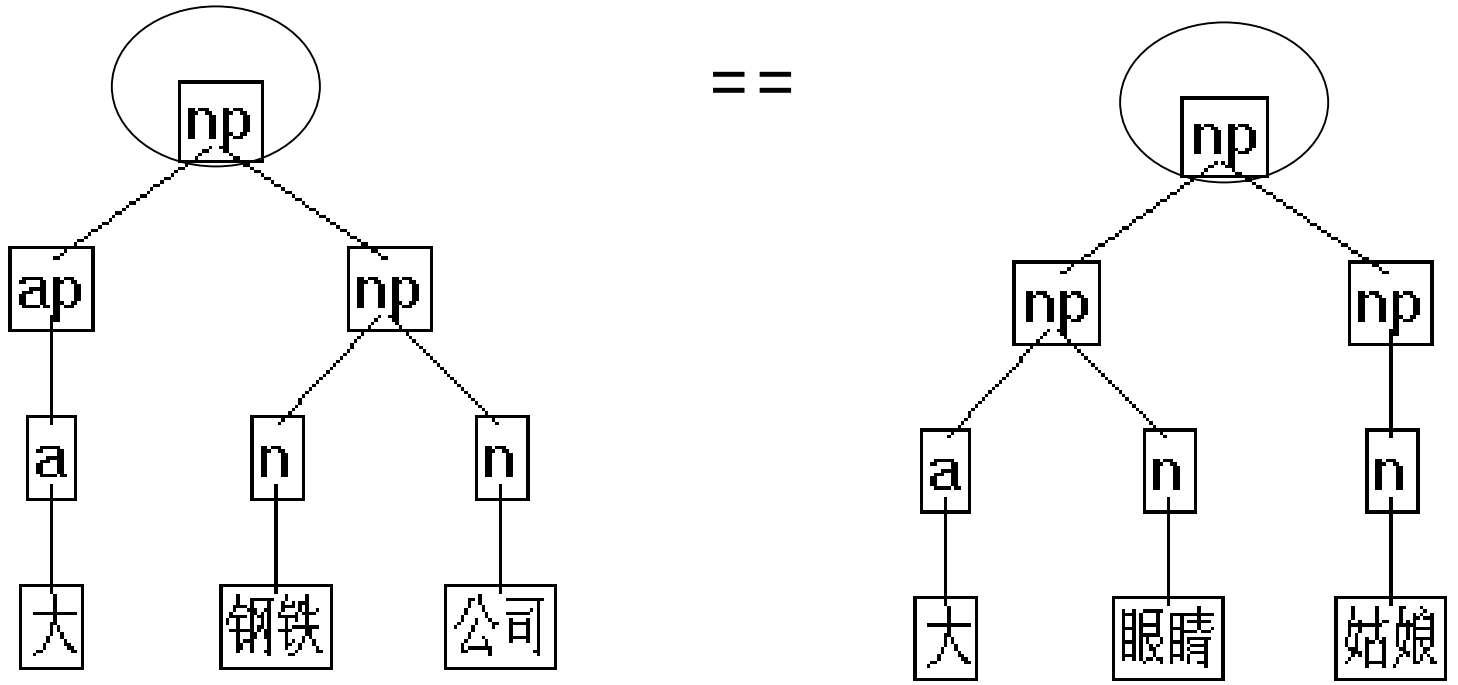
=/=

分析方法



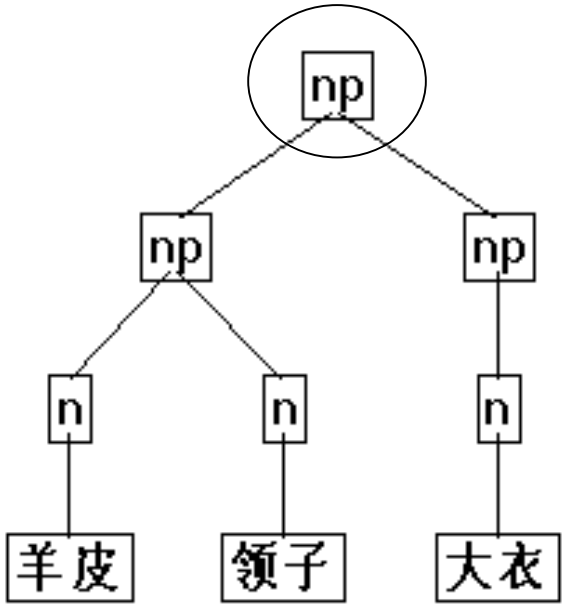
出租汽车　np or vp

内含型歧义



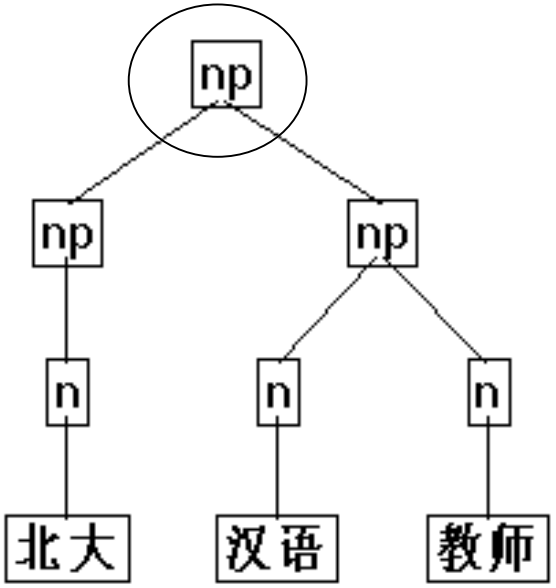
内含型歧义（续1）

羊皮领子大衣



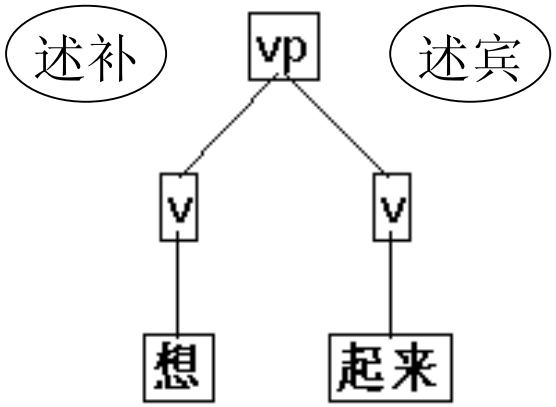
==

北大汉语教师



内含型歧义（续2）

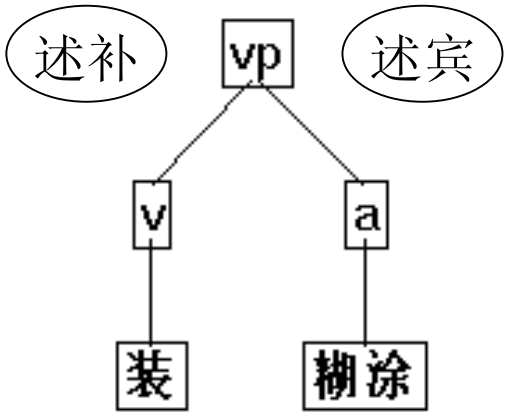
想起来



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

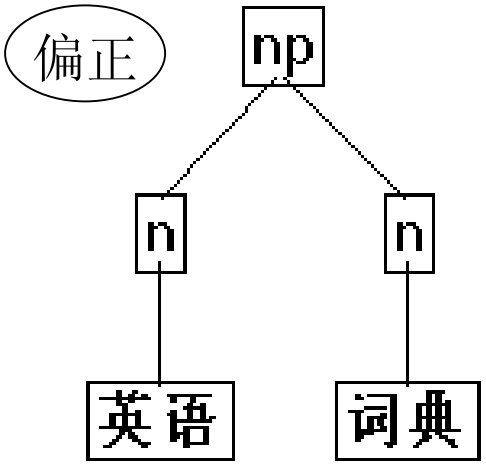


他就会装糊涂，其实他心理比谁都清楚

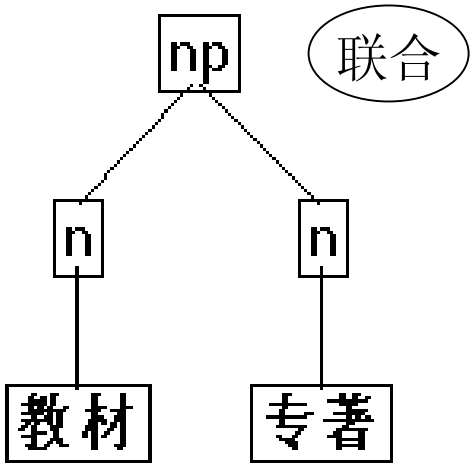
装了一上午家具，我都装糊涂了

内含型歧义（续3）

英语词典



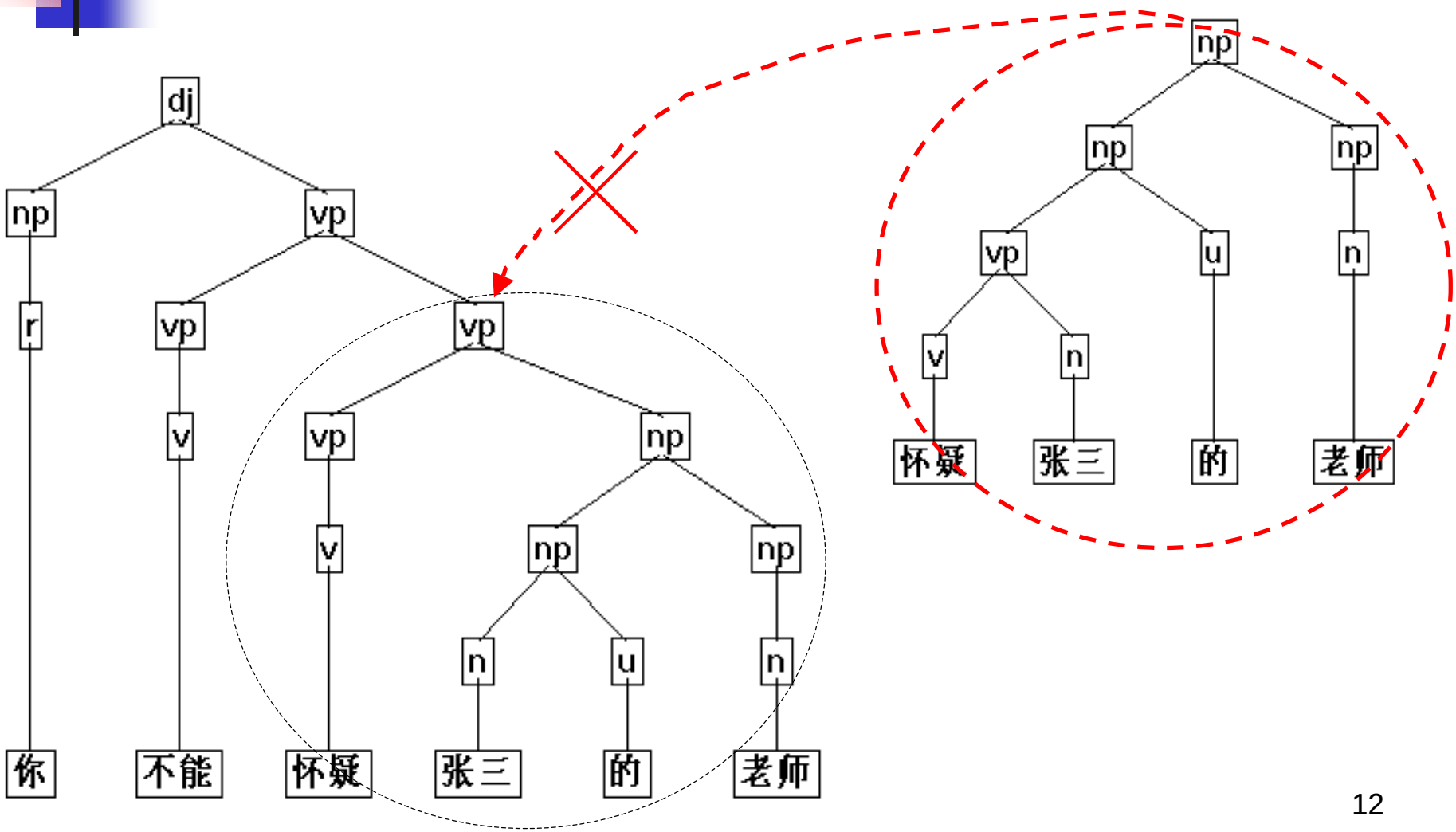
教材专著



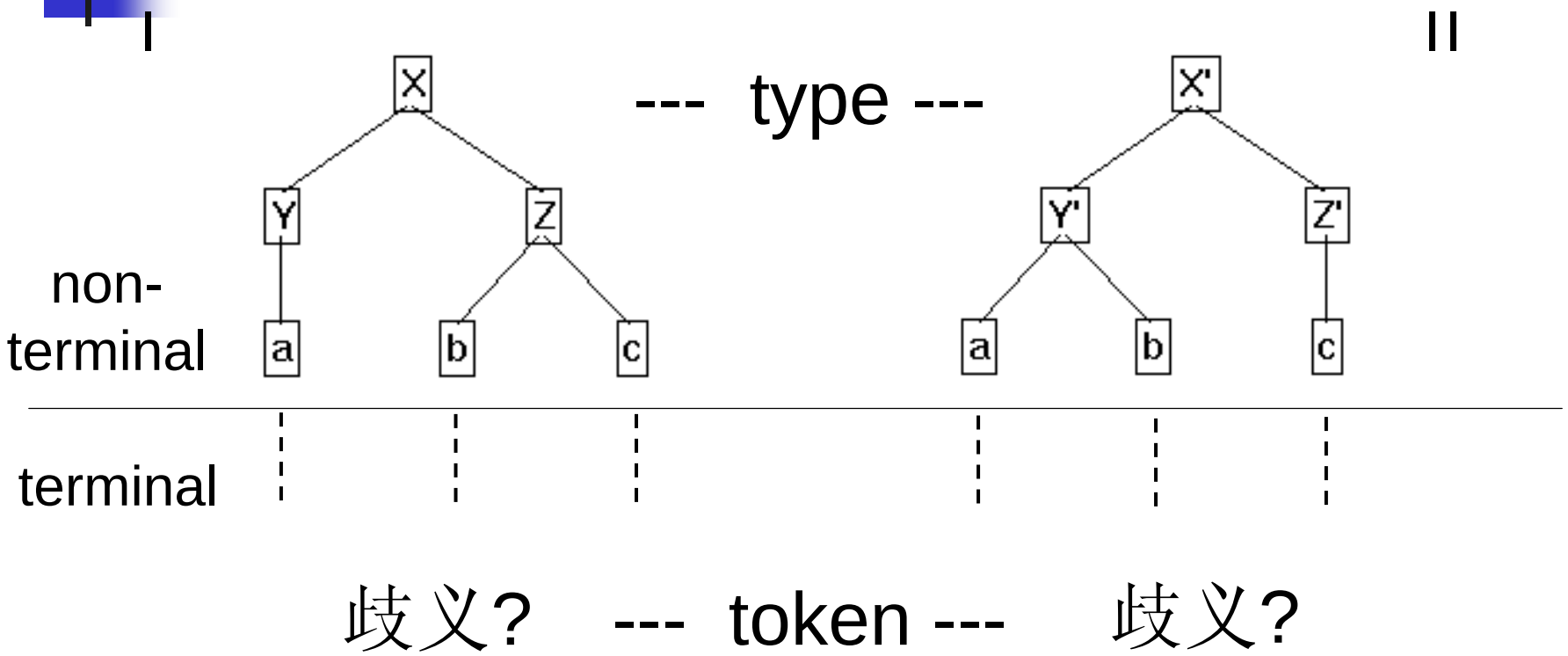
偏正？ 牛奶饼干 联合？

区分“外显”与“内含”的作用

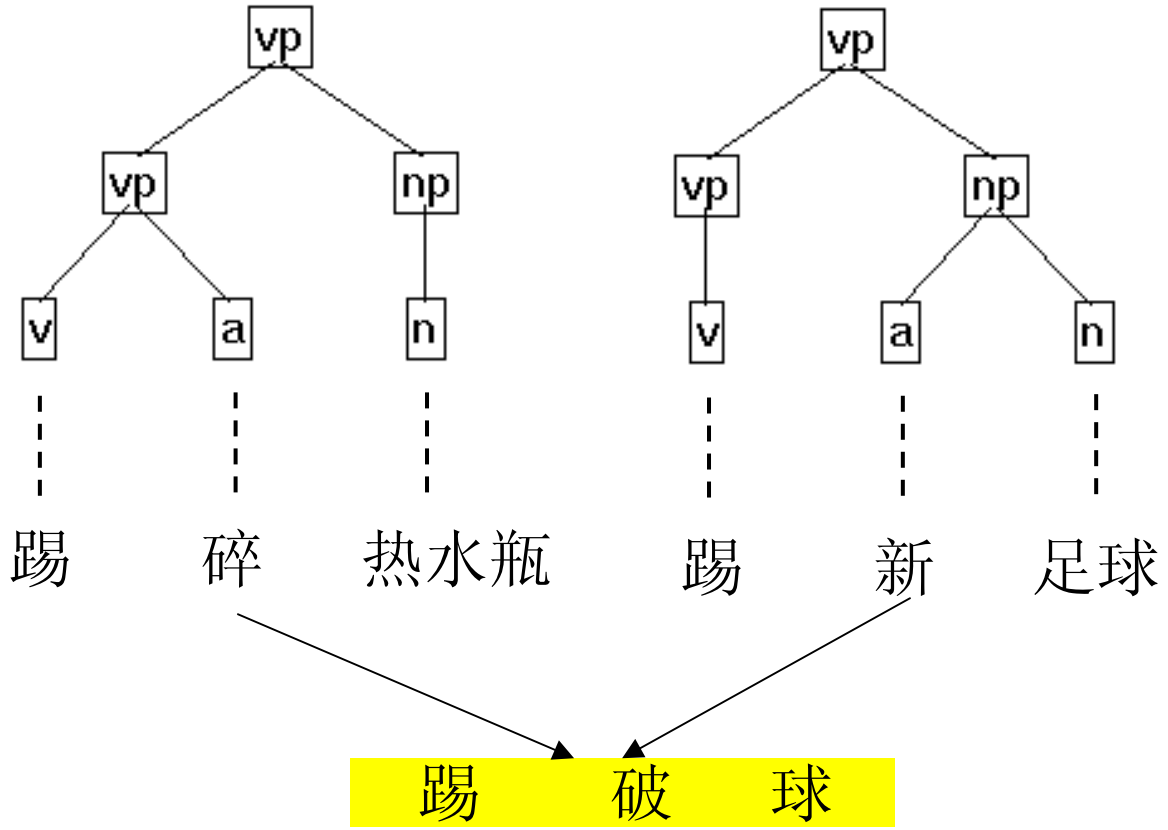
ex. 你不能怀疑张三的老师



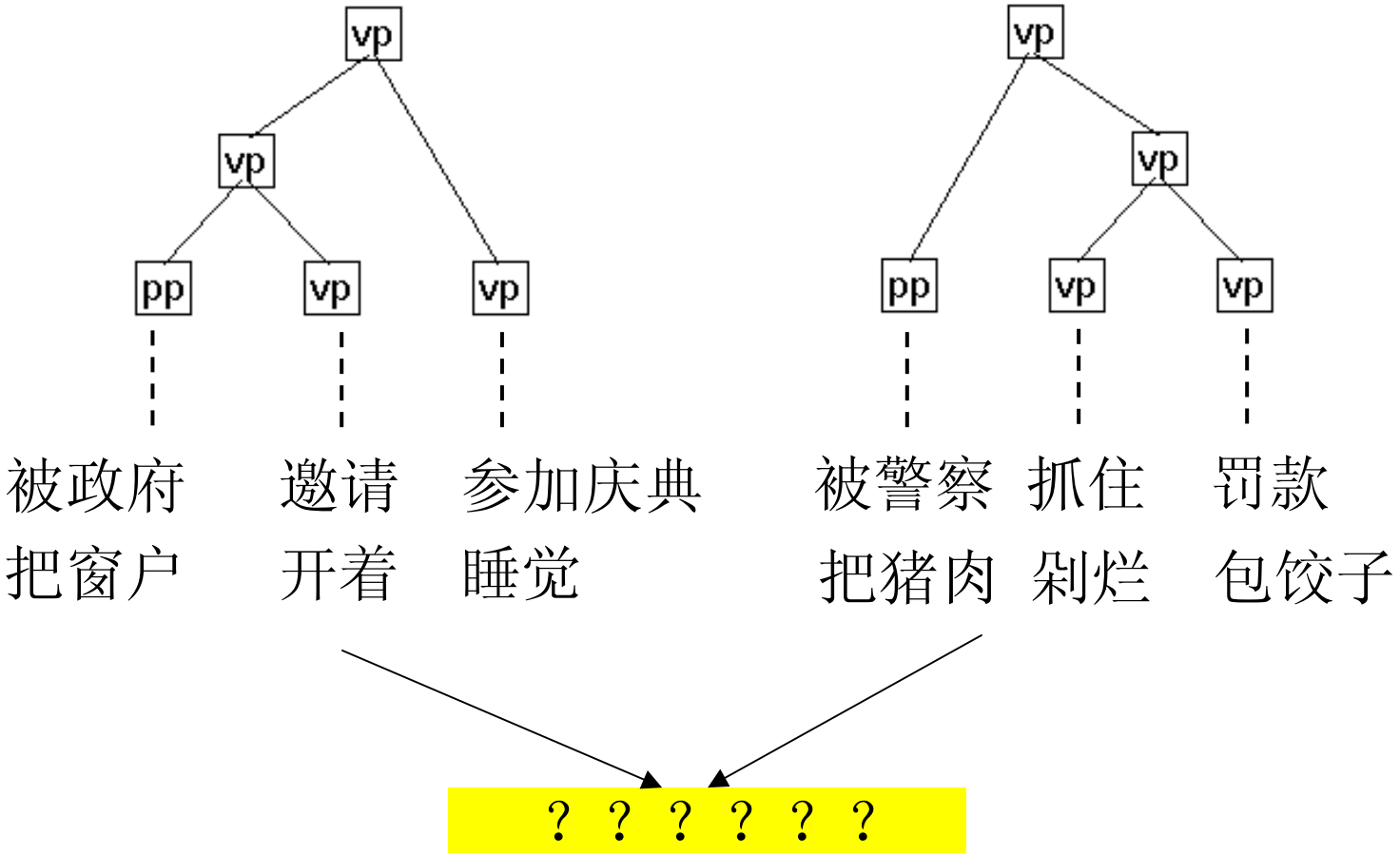
二 真歧义 准歧义 伪歧义



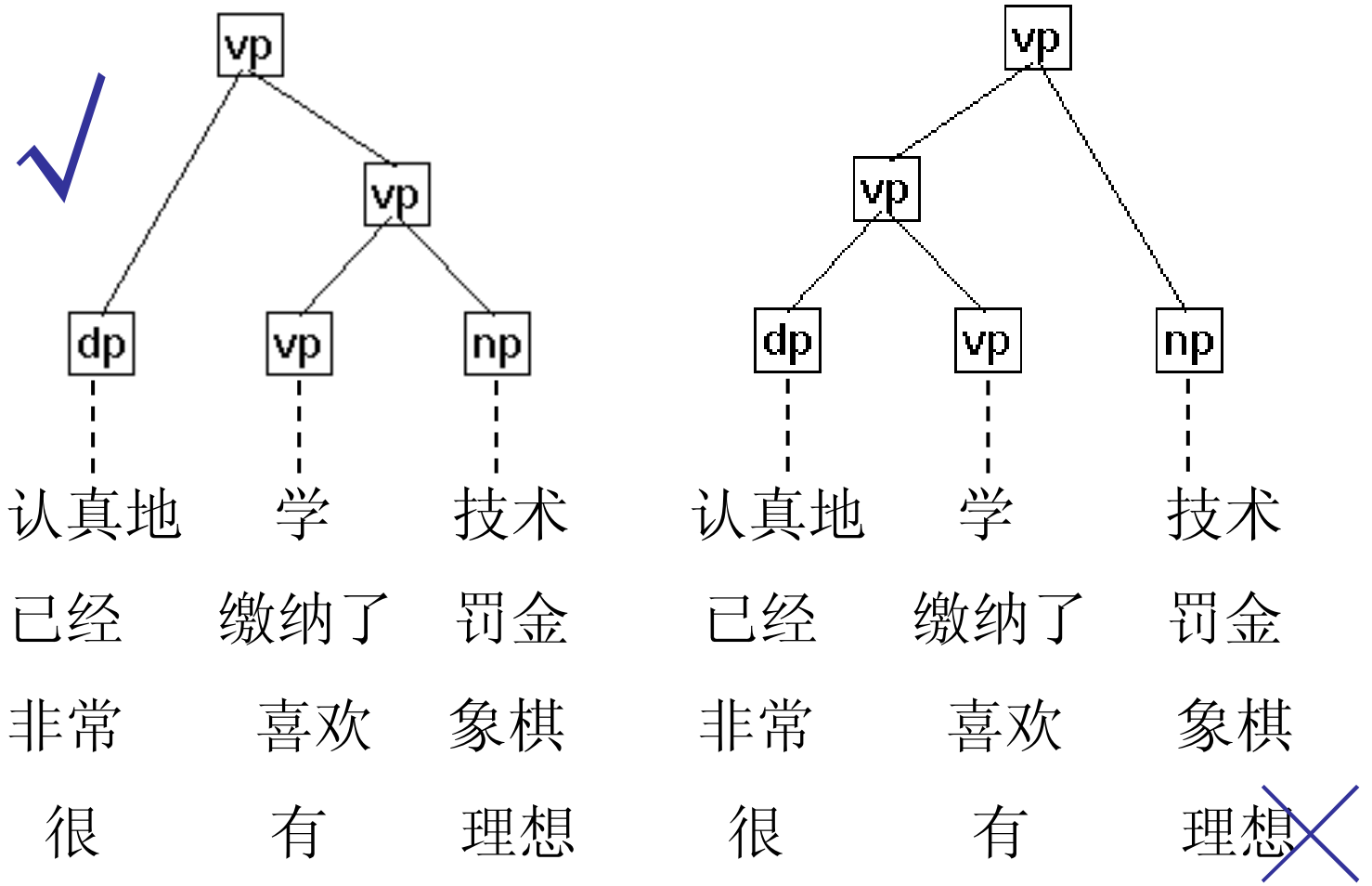
真歧义



淮歧义


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

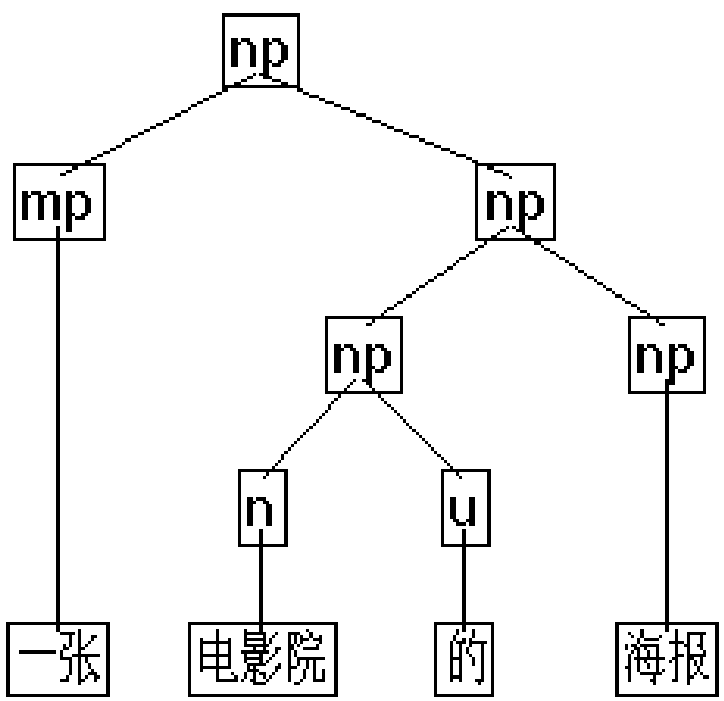

vp
vp
dp
vp
np
认真地
已经
非常
很
学
缴纳了
喜欢
有
技术
罚金
象棋
理想
vp
vp
dp
vp
np
认真地
已经
非常
很
学
缴纳了
喜欢
有
技术
罚金
象棋
理想

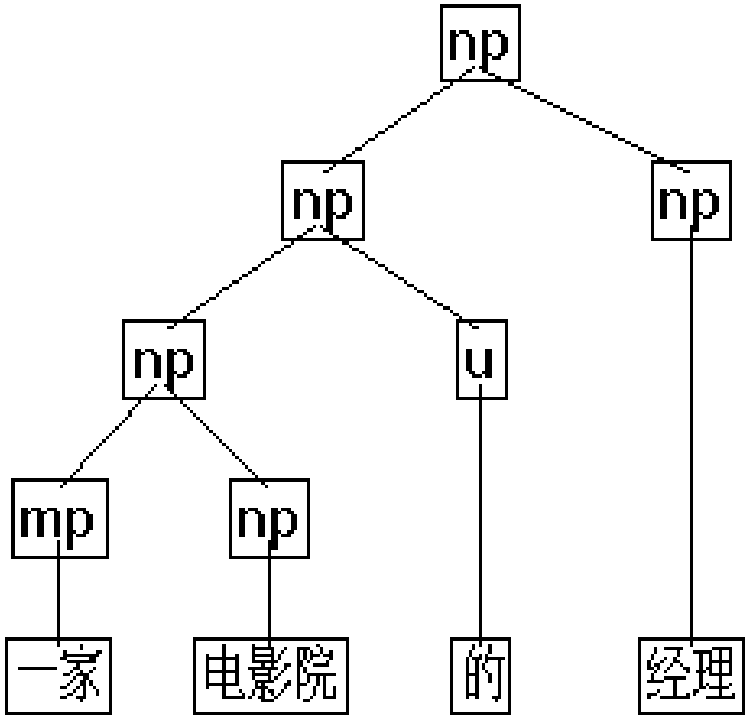
区分“真/准/伪”歧义的作用

- 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。

三 含终结符的歧义格式

mp np u<的> np





四　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

……

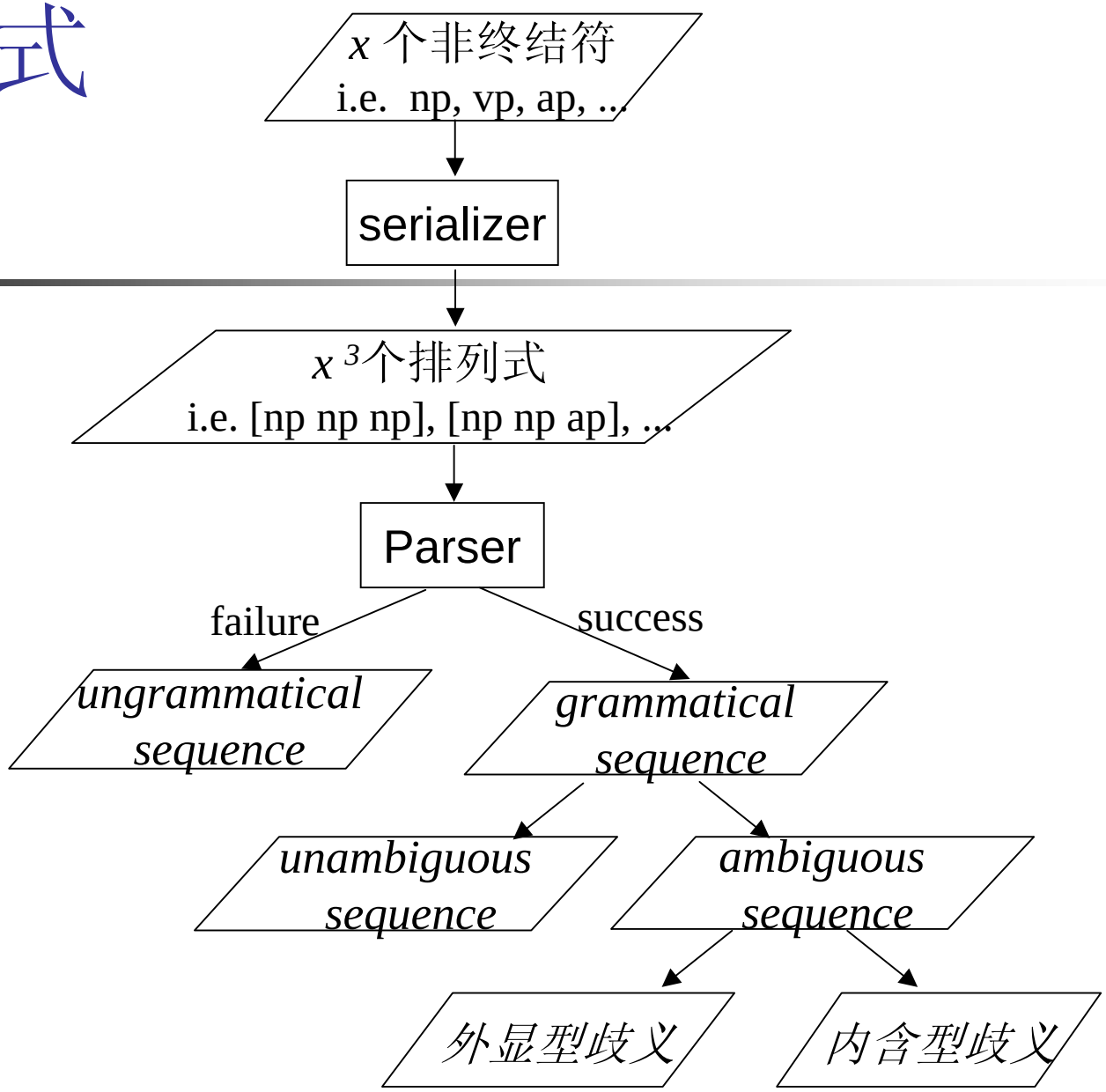
1　哪些格式有潜在歧义？

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


x 个非终结符
i.e. np, vp, ap, ...
serializer
x 3个排列式
i.e. [np np np], [np np ap], ...
Parser
failure
success
ungrammatical sequence
grammatical sequence
unambiguous sequence
ambiguous sequence
外显型歧义
内含型歧义

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

$9^3 = 729$

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	dj mp mp dj mp tp dj mp sp dj mp dp …… pp tp sp pp tp dp pp tp pp
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 ……	
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

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