现代汉语树库的构建及其应用

詹卫东
zwd@pku.edu.cn

http://ccl.pku.edu.cn:8080/WebTreebank/

提纲

1. 树库（Treebank）概述
2. 树库的构建
   - 软件工具：分词/词性标注/句法分析器
   - 语言学理论：词类 | 短语类 | 层次分析
3. 树库的应用
   - 基于树结构的语言成分分布考察
   - 语言成分的功能变异分析

   为自动句法分析提供知识源

   ……

1 树库（Treebank）概述

- 历史发展简介

时间：1993 —— Marcus(1993)

语种：英语、德语、中文、阿拉伯语 ……

标注体系：生成语法 —— HPSG —— 依存语法

标注深度：树库 —— 命题库 —— 篇章库

Xue, Nianwen (2005)

http://en.wikipedia.org/wiki/Treebank

2 中文树库的构建

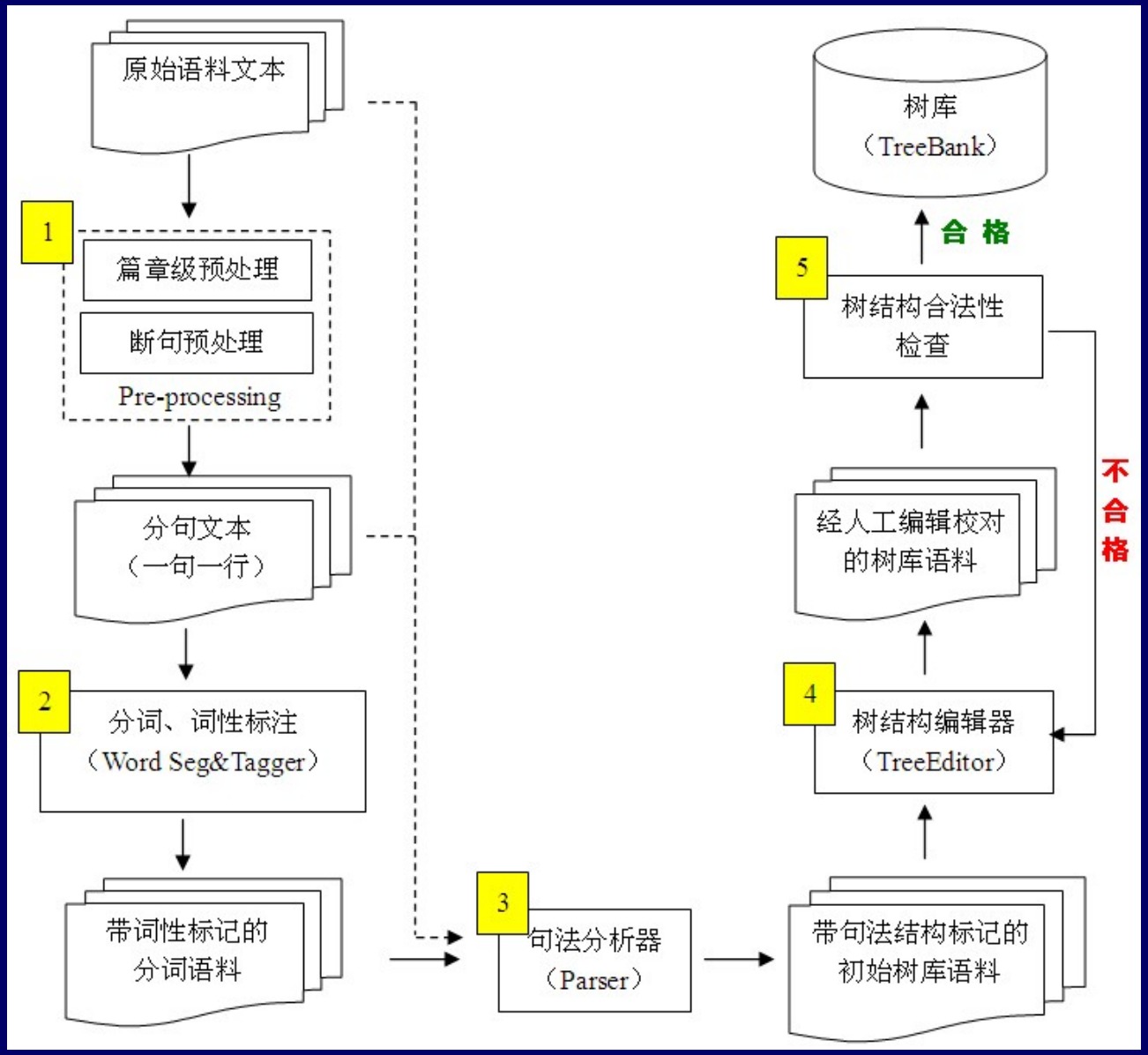
2.1 树库构建方法

2.2 树库加工中面临的语言学问题

2.2.1 短语层次分析问题

2.2.2 短语功能分类问题

树库加工流程（Workflow）



1，2，3，5：
程序自动完成

4：
人工校对

树库加工流程（示例）

原始文件

《中国人权事业的进展》

前　言

1991年11月，中国政府发表了《中国的人权状况》，向国际社会阐述了中国在人权问题上的基本立场和实践。四年来，中国的人权事业又取得了新的进展。

……

断句文件

1. 《中国人权事业的进展》
2. 前言
3. 1991年11月，中国政府发表了《中国的人权状况》，向国际社会阐述了中国在人权问题上的基本立场和实践。
4. 四年来，中国的人权事业又取得了新的进展。
5. ……

树库加工流程（示例）

断词文件

1. 《/w 中国/ns 人权/n 事业/n 的/u 进展/v 》/w
2. 前言/n
3. 1991年/t 11月/t ，/w 中国/ns 政府/n 发表/v 了/v 《/w 中国/ns 的/u 人权/n 状况/n 》/w ，/w 向/p 国际/n 社会/n 阐述/v 了/u 中国/ns 在/p 人权/n 问题/n 上/f 的/u 基本/a 立场/n 和/v 实践/v 。/w
4. 四/m 年/q 来/f ，/w 中国/ns 的/u 人权/n 事业/n 又/d 取得/v 了/u 新/a 的/u 进展/v 。/w
5. ……

树库文件

1. (hl (wbl (《) !np (np (nps (!ns (中国)) !np (np (!n (人权)) !np (!n (事业)))) ude1 (的) !vp (!v (进展))) wbr (》))))
2. (np (!n (前言)))
3. (zj (!fj (tp (tp (!t (1991年)) !tp (!t (11月))) wco (，) !fj (!dj (np (nps (!ns (中国)) !np (!n (政府))) !vp (!vp (!vp (!v (发表)) ule (了)) np (wbl (《) nps (!ns (中国)) ude1 (的) !np (np (!n (人权)) !np (!n (状况))) wbr (》)))) wco (，) dj (pp (!p (向) np (ap (!b (国际)) !np (!n (社会)))) !vp (!vp (!vp (!v (阐述)) ule (了)) np (nps (!ns (中国)) !np (pp (!p (在) np (np (!n (人权)) !np (!n (问题))) f (上)) ude1 (的) !np (!np (np (!n (基本)) !np (!n (立场))) c (和) np (!n (实践)))))))))) wfs (。)))
4. (zj (!dj (tp (!qp (m (四) !q (年)) m (来)) wco (，) !dj (np (nps (!ns (中国)) ude1 (的) !np (np (!n (人权)) !np (!n (事业)))) !vp (dp (!d (又)) !vp (!vp (!vp (!v (取得)) ule (了)) np (ap (!a (新)) ude1 (的) !vp (!v (进展))))))) wfs (。)))
5. ……

北大树库词类标记

词类标记	含义	词类标记	含义
a	形容词	n	名词
b	区别词	o	拟声词
c	连词	p	介词
d	副词	q	量词
e	叹词	r	代词
f	方位词	s	处所词
g	语素	t	时间词
h	前缀	u	助词
i	成语	v	动词
j	缩略语	w	标点
k	后缀	x	非语素字
l	习用语	y	语气词
m	数词	z	状态词

细化

词类标记	含义	词类标记	含义
rb	具有指示词功能的代词	rs	具有处所词功能的代词
rd	具有副词功能的代词	rt	具有时间词功能的代词
rm	具有数词功能的代词	rv	具有动词功能的代词
rn	具有名词功能的代词		

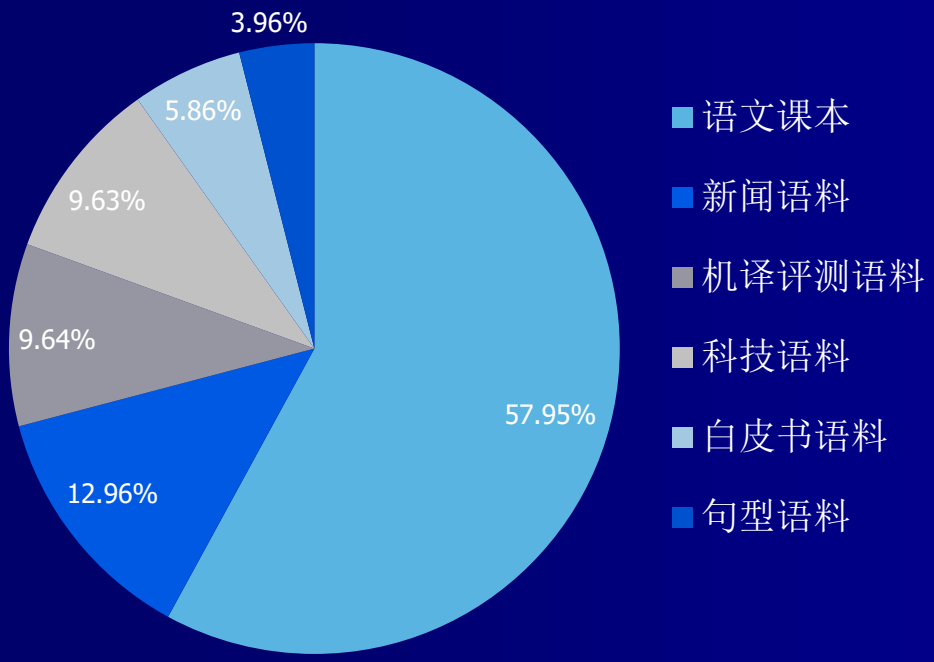
北大树库短语类标记

短语标记	含义	实例
zj	整句	三点钟全体集合。 你怎么了？ 快跑！ 他走了……
fj	复句	只要他在，你就过不去； 风也停了，雨也住了
dj	小句	爱夸张事实的孩子往往喜欢喜剧；三点钟全体集合；今天星期一；他二十来岁；**长两米；重三斤；**
np	名词性短语	粒子碰撞噪声检测仪；计算机在国外应用的现状；世界名牌服装；新问题；自己的； 桌椅门窗；理想与现实；支持总统的群众；给孩子们的；服装设计；两国之间的合作；几十年的努力；他们两位；**录象带两百盘；最善良的一个；三斤重；两米宽；**
vp	动词性短语	把杂志放进抽屉里；进行多方面的经济结构的调整；从暴风雪中救出了一群羊；来了；请客人吃饭；去外婆家玩；烧毁证物并袭击警察；跑得我累死了；
ap	形容词性短语	很不高兴；冷得发抖；比他们房间冷得多；干干净净的；通红通红的；亮了；干净不了三天；不礼貌而且不诚实；**长三米；小两岁；**
dp	副词性短语	飞快地；轻松而愉快地；波浪式地；
pp	介词性短语	关于专家系统；从桌子上；被我们；在后面；比这里；从北京到那里；除他之外；
sp	处所词性短语	报纸上；我前面；我们班里；
tp	时间词性短语	一个秋天的早晨；下星期一；吃饭前；
qp	数量短语	两百张；三十岁；三场；多少斤；三四十次
mp	数词性短语	六七百；三万两千零五十；四又二分之一；五点三二；大多数；不少；几；

结构类

仿照词类确定的短语**功能类**

北大中文树库规模及语料分布情况



句数： 55,161

词数： 882,326

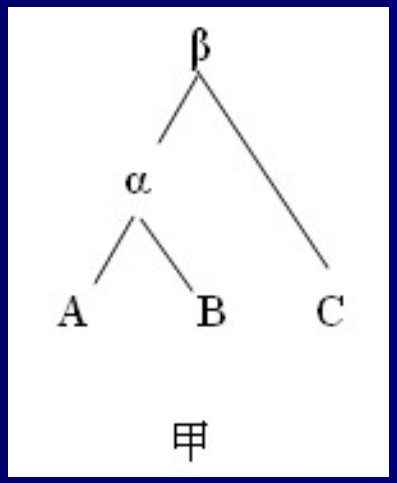
字数： 1,281,169

北大中文树库短语类和词类统计

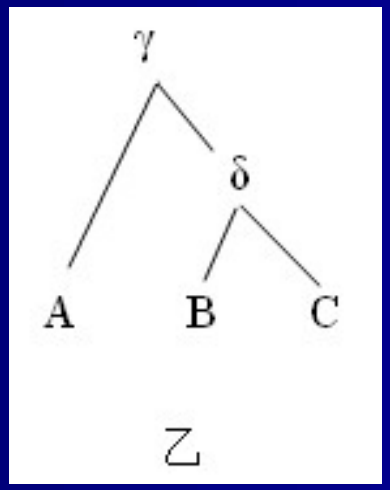
	短语	TYPE	TOKEN
	fj	372	43672
	np	364	261756
	dj	256	101198
	vp	254	211357
	tp	127	20446
	ap	117	54550
	zj	115	54637
	sp	87	26581
	pp	71	25522
	qp	71	33818
	mp	42	30835
	dp	39	63943
	START	19	55742
	yj	19	2823
	ypc	18	1693
	hl	13	428
	npr	10	831
	npz	9	98
	yph	2	747
	vn	1	1
合计	20	2006	990678

	词类	TYPE	TOKEN
	n	20423	159710
	v	11233	178341
	a	3098	34745
	m	2835	29325
	nr	2381	10391
	vn	1602	7355
	d	1494	58607
	t	1345	12287
	iv	985	1329
	b	858	4472
	nz	858	1944
	ns	850	6302
	z	759	1954
	ng	652	2418
	lv	502	668
	.....	.....	.....
	wfs	1	43013
	wsc	1	6979
	yde	1	582
	yle	1	3796
合计	95	56304	899373

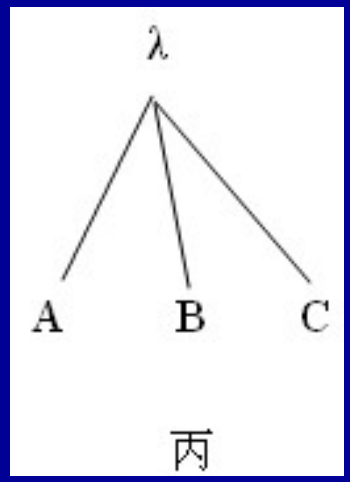
2.2.1 短语结构层次划分的问题



大　眼睛　姑娘



大　钢铁　公司

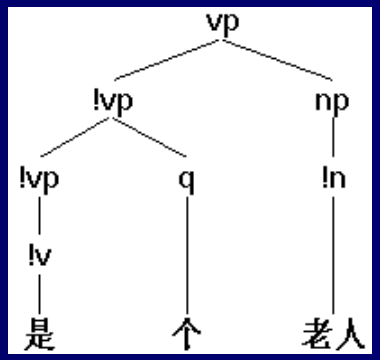


小王　和　小李

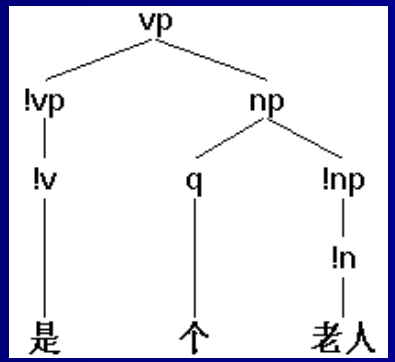
短语结构层次划分的问题

是　个　老人　　v　q　np

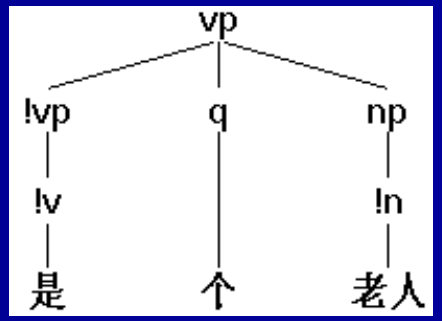
甲



乙



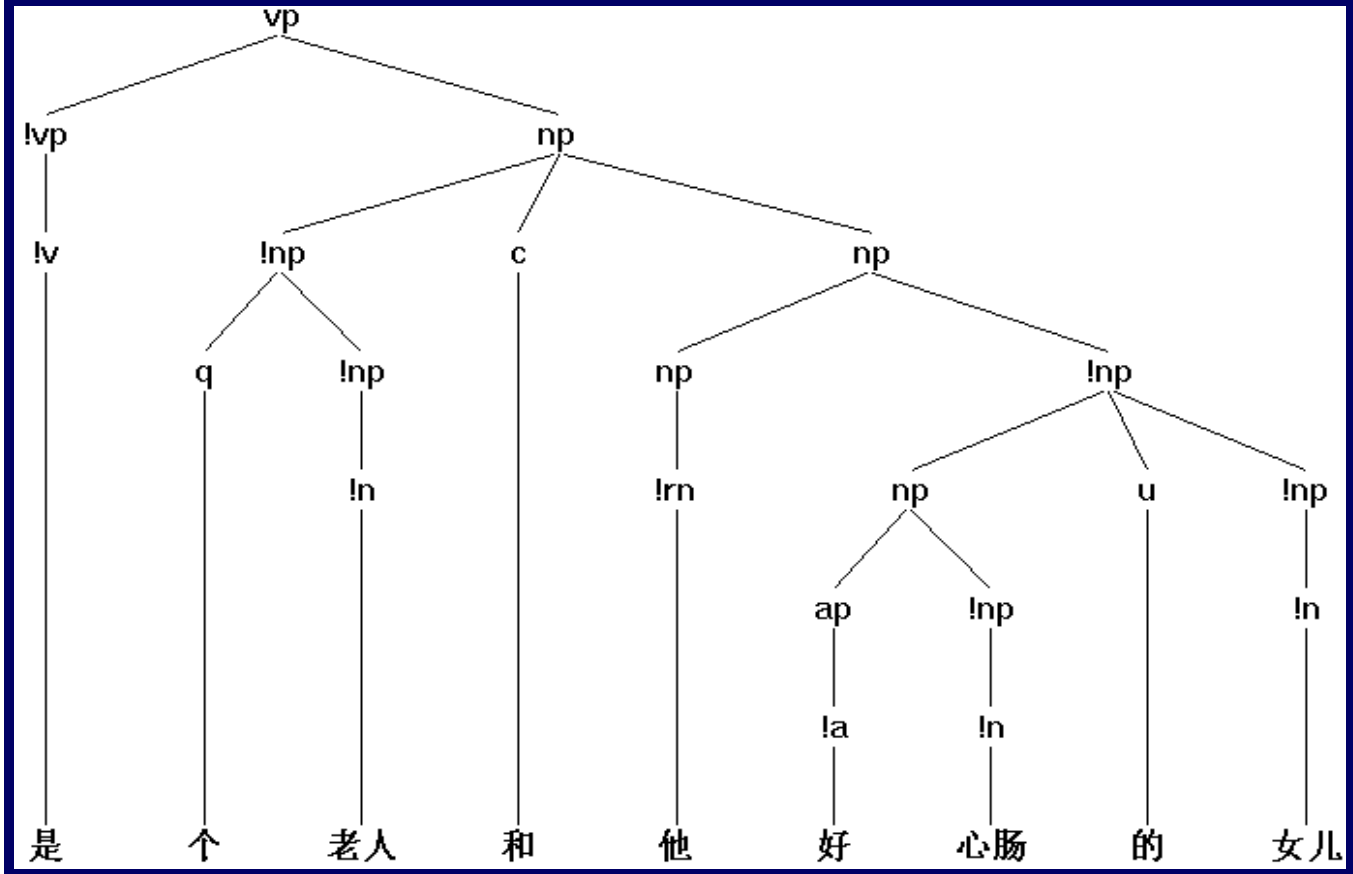
丙



买　本　瞧瞧

你 再 **坐　会儿**

短语结构层次划分的问题



选择：按 乙 方式分析

“q np” 的分布：

1）v 后宾语位置

2）“把、被”后宾语位置

3）联合结构前项位置

按 甲 方式 分析： ×

造成“个”后接复数结构

按 丙 方式 分析： ×

造成 vp 和 np 并列构造

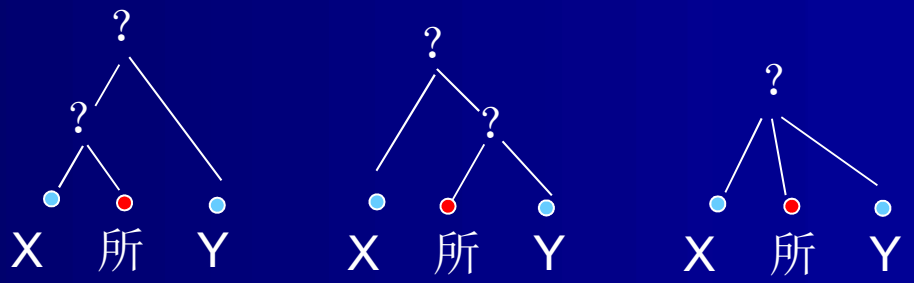
层次分析所得的单位应“分布最大化”

2.2.2 短语结构功能分类的问题

他 所 写 的 文章

今天 所 讲 的 内容

……



“X+ 所 + Y”的更多例子

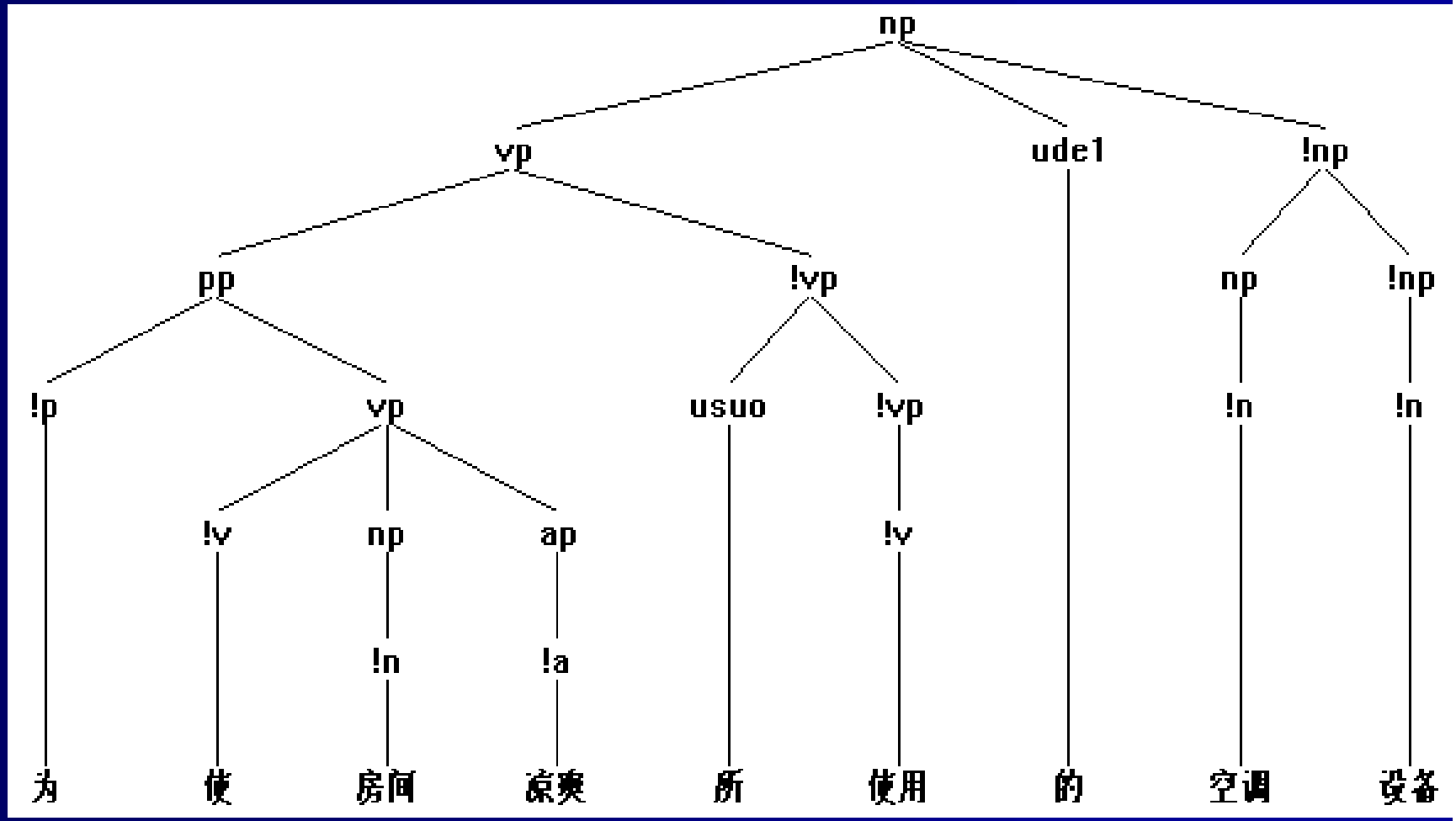
1. 所 **使用** 的 案例 还是 很早的
2. 所 **需** 建设费 平均 每瓦 为 二百五十日元
3. **所 生** 子女 属于 母亲 一 方
4. 全 靠了 他 **卖血** 所 **换得** 的 钱，才…
5. **为 使 房间 凉爽** 所 **使用** 的 空调 设备
6. 毛泽东 **在 这次会议 上** 所 **作** 的 报告
7. 即将 **由 这 次 停火** 所 **带 来** 的 新 形势
8. 前一次大老 **亲口** 所 **说** 的 话 ……

“所 + Y” 前面可以没有成分　例1-3

“所 + Y” 后面可以没有 “的”，直接修饰 np　例2-3

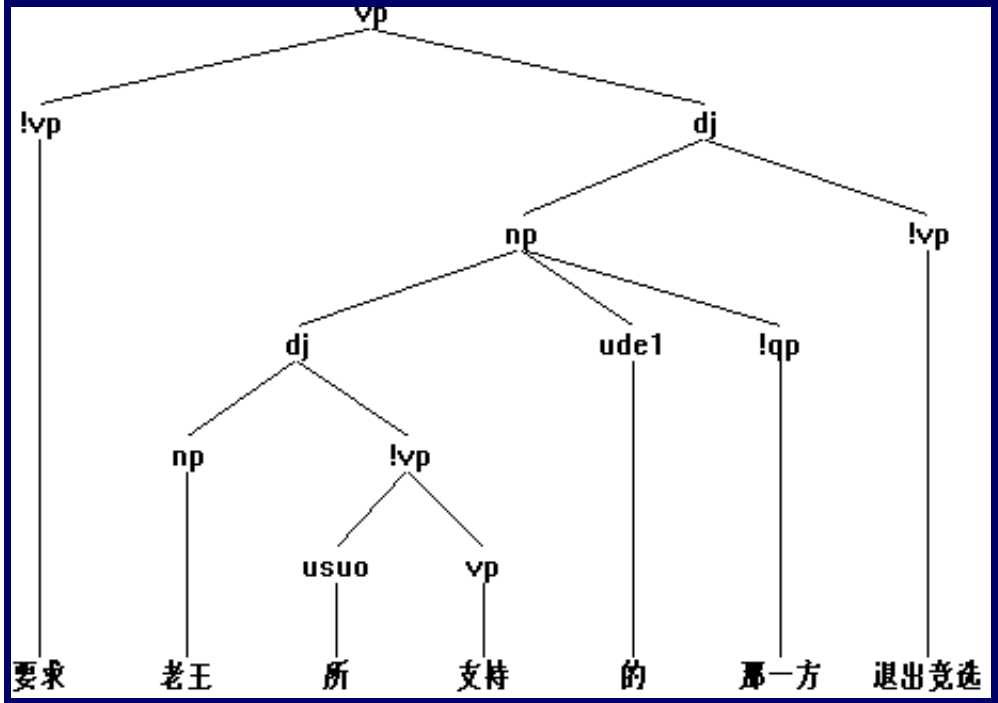
“所 + Y” 前面可以是vp，pp，dp等成分　例4-8

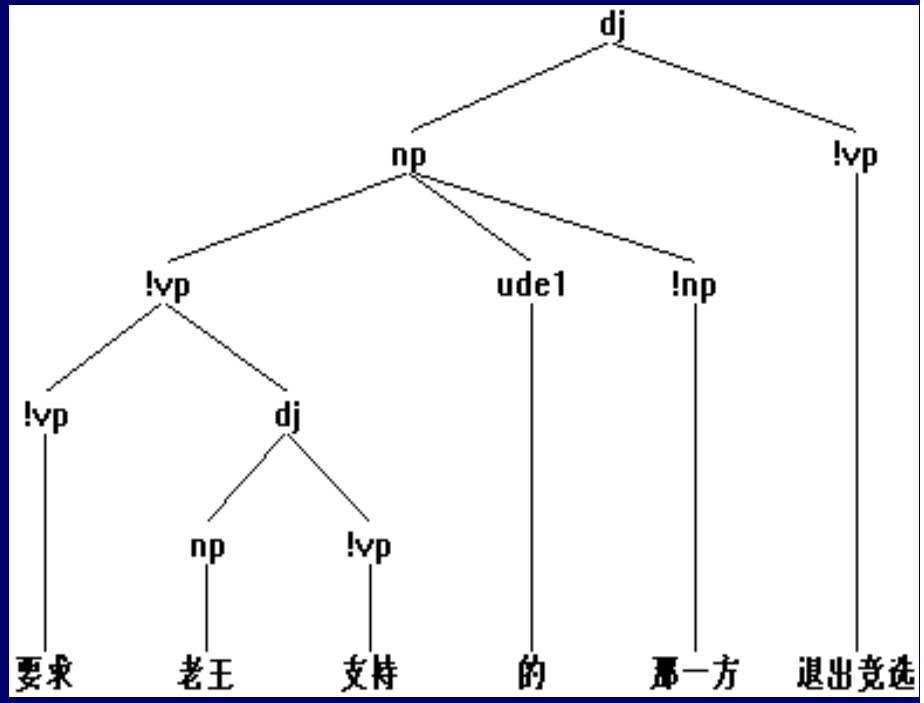
“X+ 所 + Y”的分析


np
vp
ude1
!np
pp
!vp
np
!np
!p
vp
usuo
!vp
!n
!n
!v
np
ap
!v
!n
!a
为
使
房间
凉爽
所
使用
的
空调
设备

“所 VP”是弱陈述性VP

a. 要求老王所支持的那一方退出竞选
b. 要求老王支持的那一方退出竞选





3 树库的应用

- 3.1 基于树结构的语言成分分布考察
- 3.2 语言成分的功能变异分析
- ……

3.1 从 基于线性串的分布 到 基于树结构的分布

- 分布分析是语言分析的主要手段。
- 以往的分布分析（面向人）主要是基于线性串的。
- 基于树结构的分布分析（面向计算机）可以获得粒度更细的语言知识。

两个例子：（1） 副词的内部差异

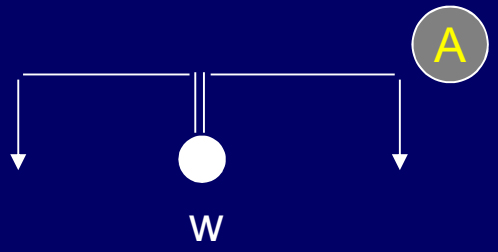
（2） “的”字结构的分布考察

案例1

关于语言单位的功能（分布）分类

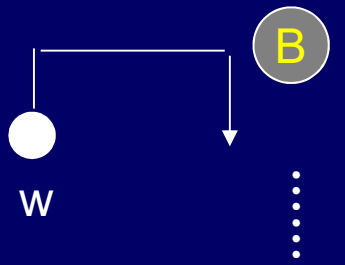
（1）一个语言单位（w）的组合方向：

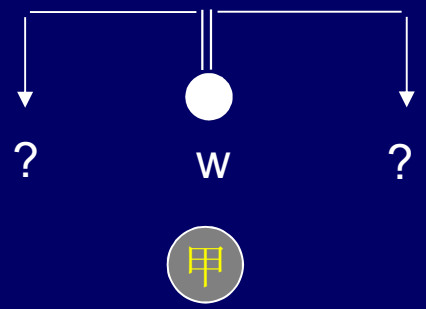
w在参与序列组合时朝哪个方向组合？

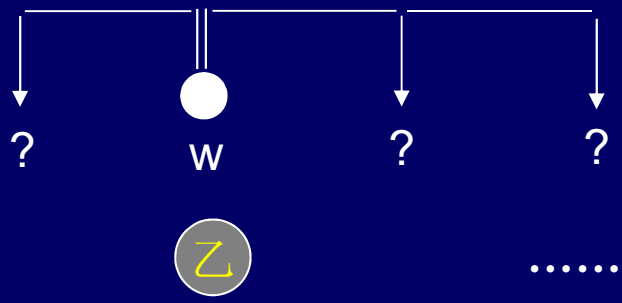


（2）一个语言单位（w）的组合对象：

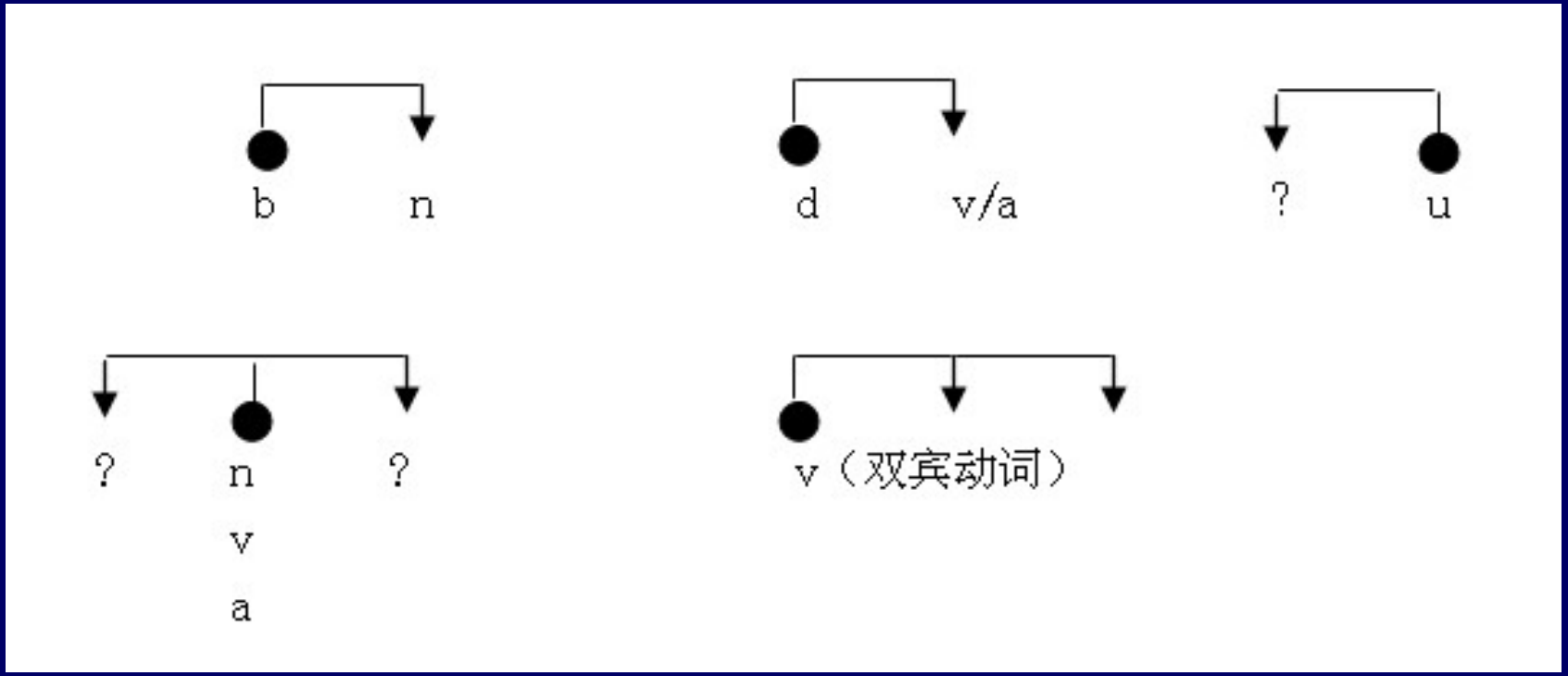
a. w要求跟几个成分组合？
b. w要求跟什么类型的语言成分组合？







“词类”（词的功能分类）示例

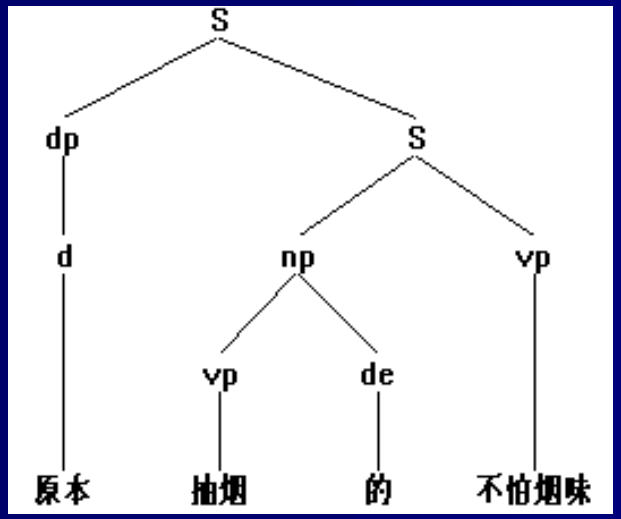


b: 区别词　d：副词　u：助词　v：动词　a：形容词　n：名词

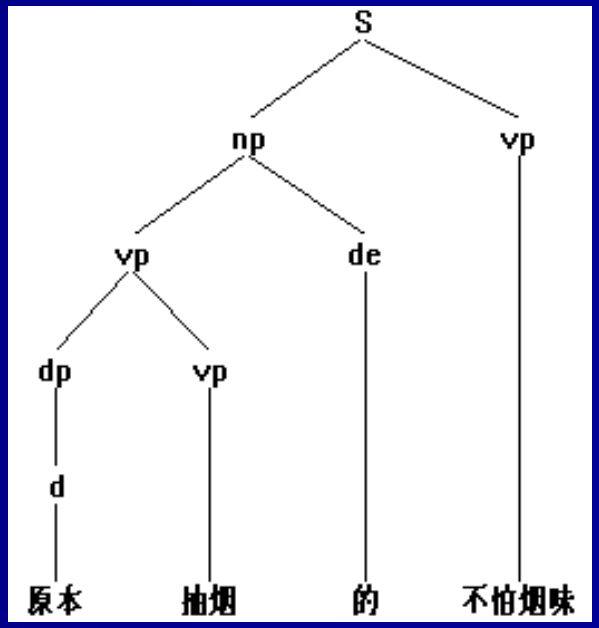
- b，d 是功能（分布）比较确定的词类；
- u 是组合方向相对确定，但组合对象不确定的词类；
- n，v，a等是组合方向和组合对象都不大确定的词类；

“dp vp 的 vp” 的结构歧义

1. 原本 抽烟　的　不怕烟味
2. 也许 抽烟　的　不怕烟味
3. 一直 抽烟　的　不怕烟味



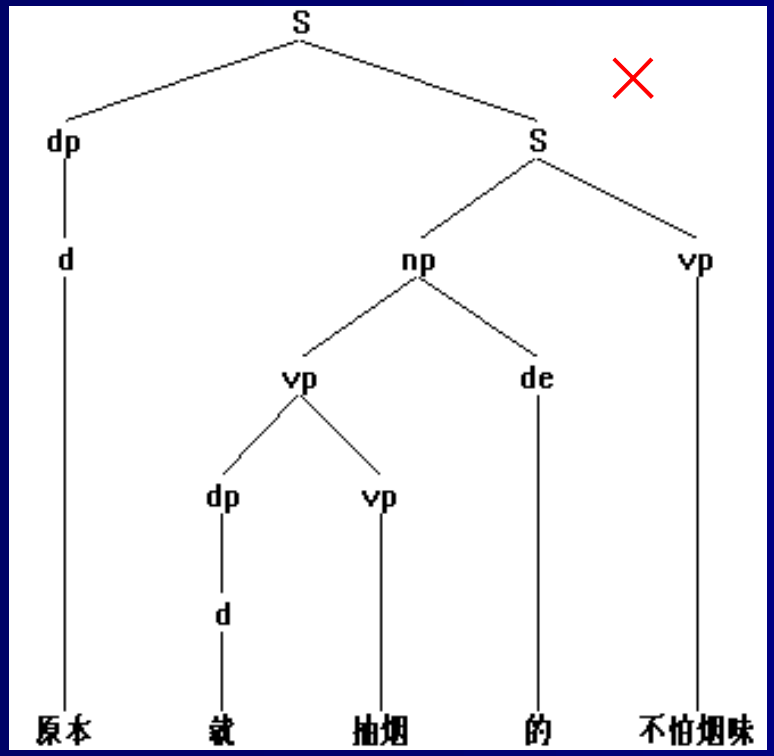
甲



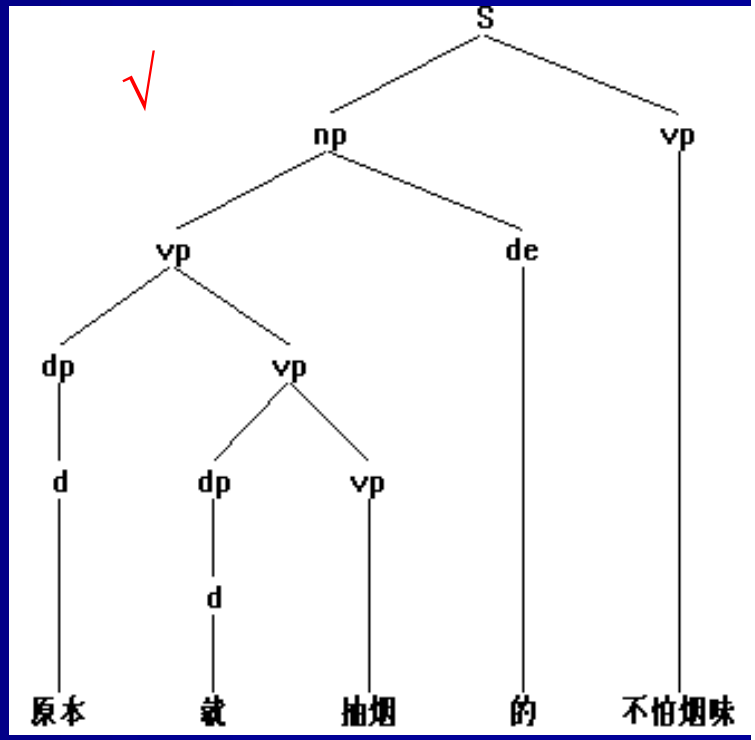
乙

增加一个副词，歧义消失

- 原本 就 抽烟 的 不怕烟味



甲

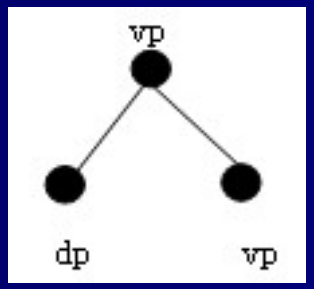


乙

副词的内部差异

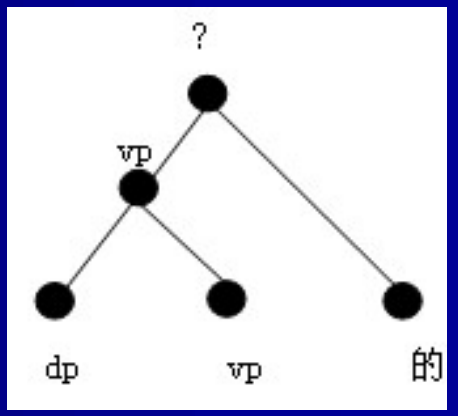
（1）《现代汉语语法信息词典》中副词有“主前后”的描述： 一个副词能否在“主语”前出现

（2）《现代汉语语法信息词典》中没有“副词 + V”后能不能再加“的”的特征描述



√ 就 抽烟

√ 原本 抽烟

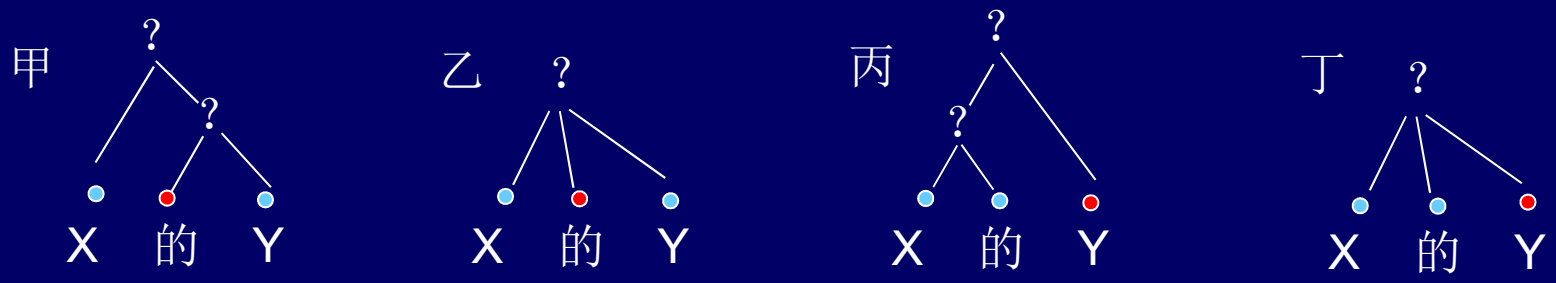


× 就 抽烟 的

√ 原本 抽烟 的

案例2

“的”字短语的功能类别与内部结构

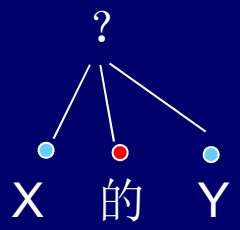


	二分/三分	中心成分	
甲	二分	的	司富珍（2004）熊仲儒（2005）
乙	三分	的	陆俭明（2003）仅针对“X 的 VP”
丙	二分	Y	李艳惠（2008）
丁	三分	Y	我们的处理方式

“的”在树库中的频次和分布

		的	地
句数：	55,161	25,726(46.64%)	2447(4.44%)
词数：	882,326	43,563(4.94%)	2644(0.30%)
字数：	1,281,169	(3.40%)	(0.21%)

甲



37758例

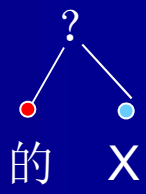
86.67%

乙

5801例

13.32%

丙



2 例

丁

“的”（di）

2 例

“的”在树库中的频次和分布（续）

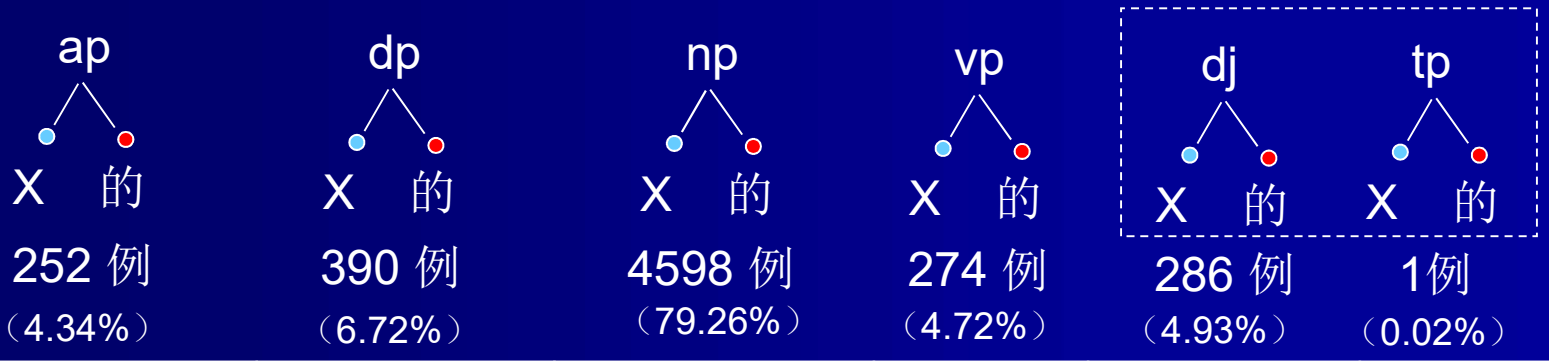
丙

1. 左邻右舍（的人），都捡了东西。
2. 两边的机关枪（的射击声）稍一停歇，大门外面的赤卫队……就冲进了公安局。

丁

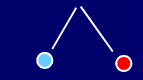
1. “有的放矢”中的“的”

“X 的”短语的功能与分布

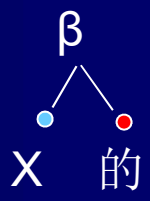


标点或连词、语气词之前	189 （75.00%）	9 （2.31%）	3774 （82.08%）	274	286	1
其他	63 （25.00%）	381 （97.69%）	824 （17.92%）	0	0	0
	慢腾腾的 晕头晕脑的 真够瞧的	不住的 又一次的 俨然的	红的 成套的 天蓝色的	是的 会着凉的 眨呀眨的	你一定喜欢的 他肺病死的 我报了名的	深更半夜的

“X 的”短语的功能与分布

	ap	dp	np	vp	dj	tp
						
	252 例 （4.34%）	390 例 （6.72%）	4598 例 （79.26%）	274 例 （4.72%）	286 例 （4.93%）	1例 （0.02%）
标点或连词、语气词之前	补语，谓语，并列项，分句（75.00%）	状语 （2.31%）	主语、宾语 （82.08%）	分句， 谓语 38例	分句， 谓语 2例	分句
其他	状语 （25.00%）	状语 （97.69%）	主语（98%） 宾语（2%）	0	0	0
	慢腾腾的 晕头晕脑的 真够瞧的	不住的 又一次的 俨然的	红的 成套的 天蓝色的	是的 会着凉的 眨呀眨的	你一定喜欢的 他肺病死的 我报了名的	深更半夜的

“X 的”短语 小结



X

vp	ap	dj	dp	fj	pp	np	qp	sp	tp	mp
2509	**1561**	**985**	**95**	**19**	**2**	556	29	24	18	3

8 : 1

β

np	dp	dj	vp	ap
4598	390	287	274	252

4 : 1

1203例非指称用法中，750例（62.34%）为陈述表达功能，且“的”位于句尾

“的”

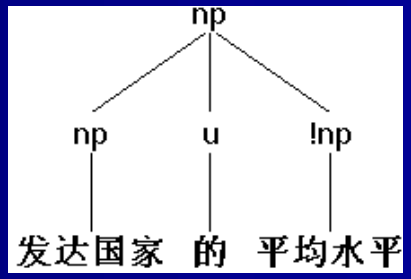
- “的”更多的是跟在“非指称性成分”后面
- “X 的”短语整体更多的是用作“指称性表达”
- 有些句尾“的”有明显语气词化倾向

“的”表“确认”语气用法的一些实例

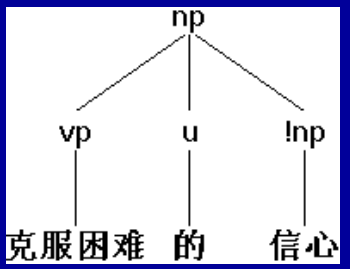
1. 所以他们才把这项工作委托给改良沙漠土壤方面具有丰富经验的林业部门的吧？
2. 我还听说施工人员以及车辆经过的路线也都列入了设计规划之中，不可以随意乱来的。
3. 历史上没有一个反对人民的势力不被人民毁灭的。
4. 酣眠固不可少，小睡也别有风味的。
5. 你 什么时候遇见他 的
6. 横竖 我 要去 的，不用 请 他 来。
7. 这些事情，是无论哪一个“友邦”也都有的，……
8. 懒洋洋地问道：“哪村来的？”
9. 您别又穷疯了，胡说乱道的。

“X 的 Y”短语的功能与分布

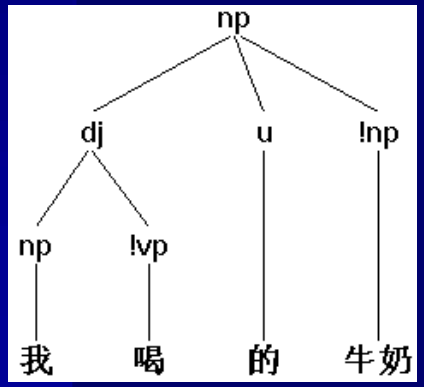
1. 发达国家的平均水平
2. 克服困难的信心
3. 张三开车的时候
4. 多么美妙的前景
5. 我喝的牛奶



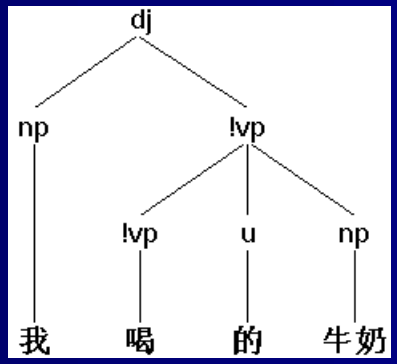
1



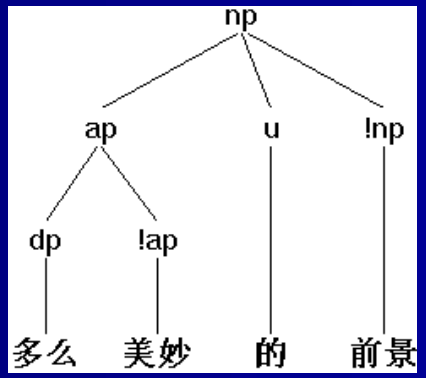
2



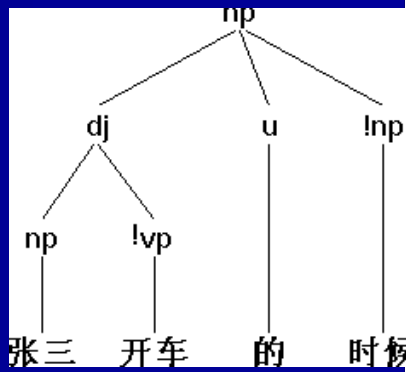
5b



5a



4

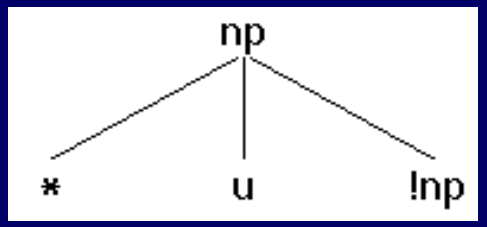


3

“X 的 Y”不同内部模式的频次

部分树库语料统计结果

16358例 94.29%

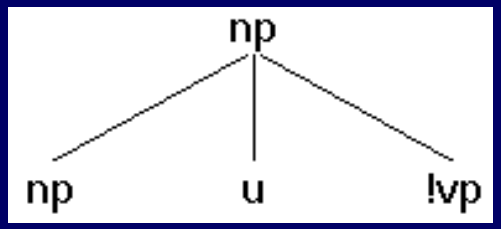


他 的 情绪

他 的 紧张情绪

紧张 的 情绪

667例 3.84%

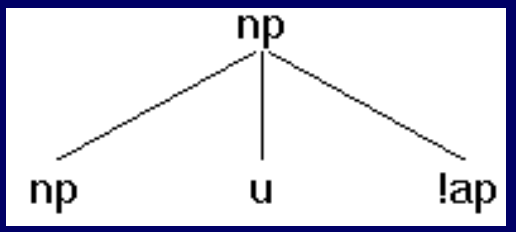


时间 的 推移

器官 的 生长发育

校长 的 尽力撮合

93例 0.54%

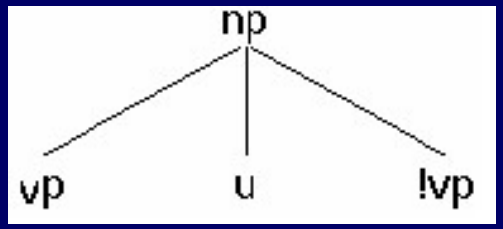


自己 的 莽撞

经济形势 的 逐步稳定

他 的 不诚实

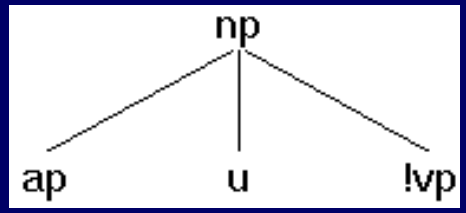
71例 0.41%



有组织 的 游说

可持续 的 增长

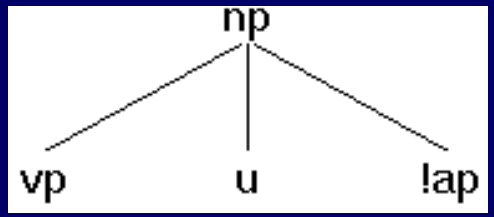
145例 0.84%



彻底 的 失败

越来越多 的 重视

15例 0.09%



说不出 的 兴奋愉快

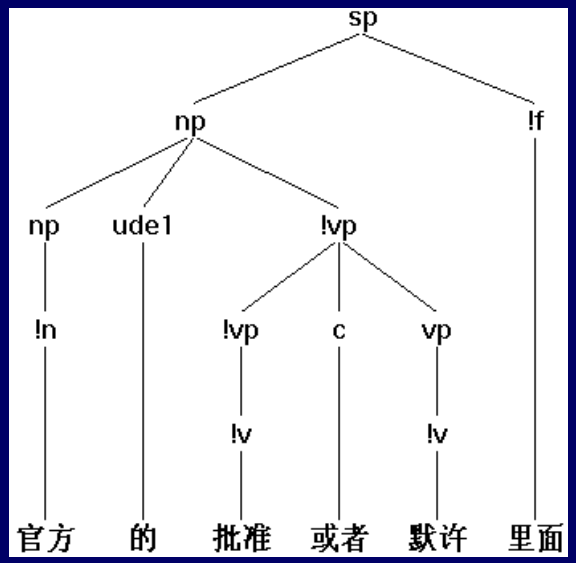
改革 的 深入

“X 的 Y”的分布环境

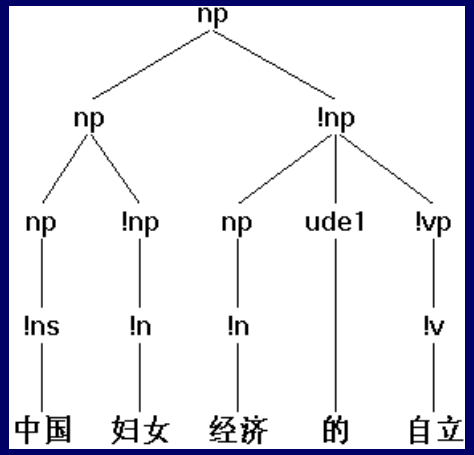
短语结构	分布种数	频次
np(np ude1 np)	165	6252
np(vp ude1 np)	130	3447
np(ap ude1 np)	120	3037
np(dj ude1 np)	97	1793
np(sp ude1 np)	61	763
np(np ude1 vp)	55	667
np(tp ude1 np)	38	281
np(pp ude1 np)	35	308
np(qp ude1 np)	33	219
np(fj ude1 np)	22	76
……		

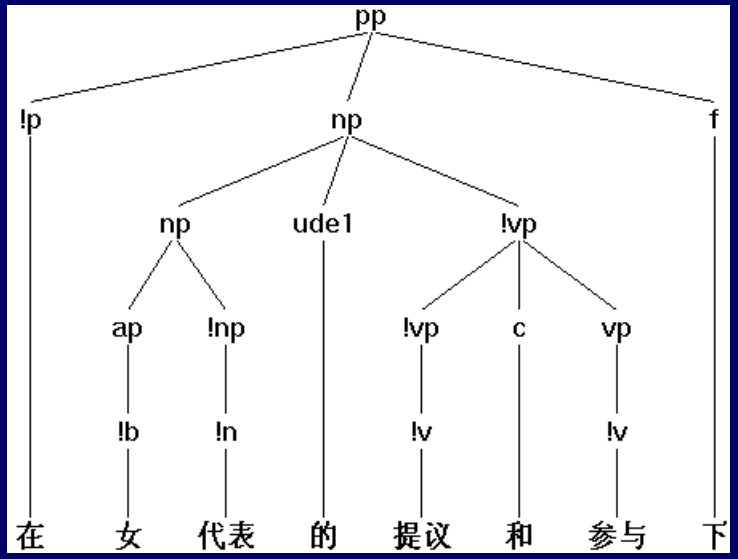
phrase	root	left	right	freq
np(np ude1 vp)	vp	vp	##	283
np(np ude1 vp)	pp	p	##	87
np(np ude1 vp)	dj	##	vp	68
np(np ude1 vp)	dj	##	wco vp	37
np(np ude1 vp)	np	##	c np	30
np(np ude1 vp)	np	np c	##	23
np(np ude1 vp)	sp	##	f	21
np(np ude1 vp)	np	np	##	12
np(np ude1 vp)	pp	p	f	11
……				

“np 的 vp” 高频分布示例



	phrase	root	left	right	freq
宾语	np(np ude1 vp)	vp	vp	##	283
	np(np ude1 vp)	pp	p	##	87
主语	np(np ude1 vp)	dj	##	vp	68
	np(np ude1 vp)	dj	##	wco vp	37
并列项	np(np ude1 vp)	np	##	c np	30
	np(np ude1 vp)	np	np c	##	23
	np(np ude1 vp)	sp	##	f	21
	np(np ude1 vp)	np	np	##	12
	np(np ude1 vp)	pp	p	f	11
	.....				





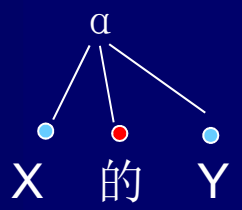
“np 的 vp” 与 “np 的 np”同分布的比例

序号	np的外部分布环境			np的内部结构	
	root	left	right	np的np	np的vp
1	vp	vp	##	2432	283
2	pp	p	##	374	87
3	dj	##	vp	1312	68
4	dj	##	wco vp	96	37
5	np	##	c np	63	30
6	np	np c	##	98	23
7	sp	##	f	291	21
8	np	np	##	81	12
9	pp	p	f	12	11
10	START	##	##	9	9
…	……	……	……	……	……
46	np	wq1	wqr	8	1
				5828/6252	658/667
				93.2%	[98.7%]

“np 的 vp” 中的vp的结构类型

	结构类型	一般vp	“np的vp”中的vp	
1	!v	75803	439	
2	!vp np	29892	0	×
3	dp !vp	23310	26	
4	!vp vp	11399	21	
5	pp !vp	9489	1	
6	!vp ule	6000	0	×
7	vp !vp	4542	14	
8	!v v	2548	1	
9	!vp dj	2452	0	×
10	!vp wco vp	2385	0	×
11	!v uzhe	2335	0	×
……				
	结构种数：	842	25 [2.97%]	
	结构例数：	203962	667 [3.30%]	

“X 的Y”短语 小结



X　所有短语类型均可。np占41.57%

Y　除 pp外其他短语类型均可。np占绝大多数（89.66%）。

α

np	**sp**	**tp**	vp	ap
36600	**604**	**509**	41	4

“的”　短语整体用作“指称性表达”占绝对多数；

有少数“的”用在vp后，np前，整体是“陈述性表达”

有极少量“的”相当于“得”。

“X 的 Y”短语整体为vp、ap的一些实例

1. 我是1964年上的大学。
2. 女人看出他笑的不像平常。
3. 您大概是想我想的梦里到过这儿
4. 你混的不错
5. 他去的匆匆，……
6. 要想住的安稳一些，……
7. 他说不出的新鲜而且高兴，……
8. 他老的不像样子了。
9. 我们先前——比你阔的多了。
10. 男社员当中，最数张老五挑剔的欢。

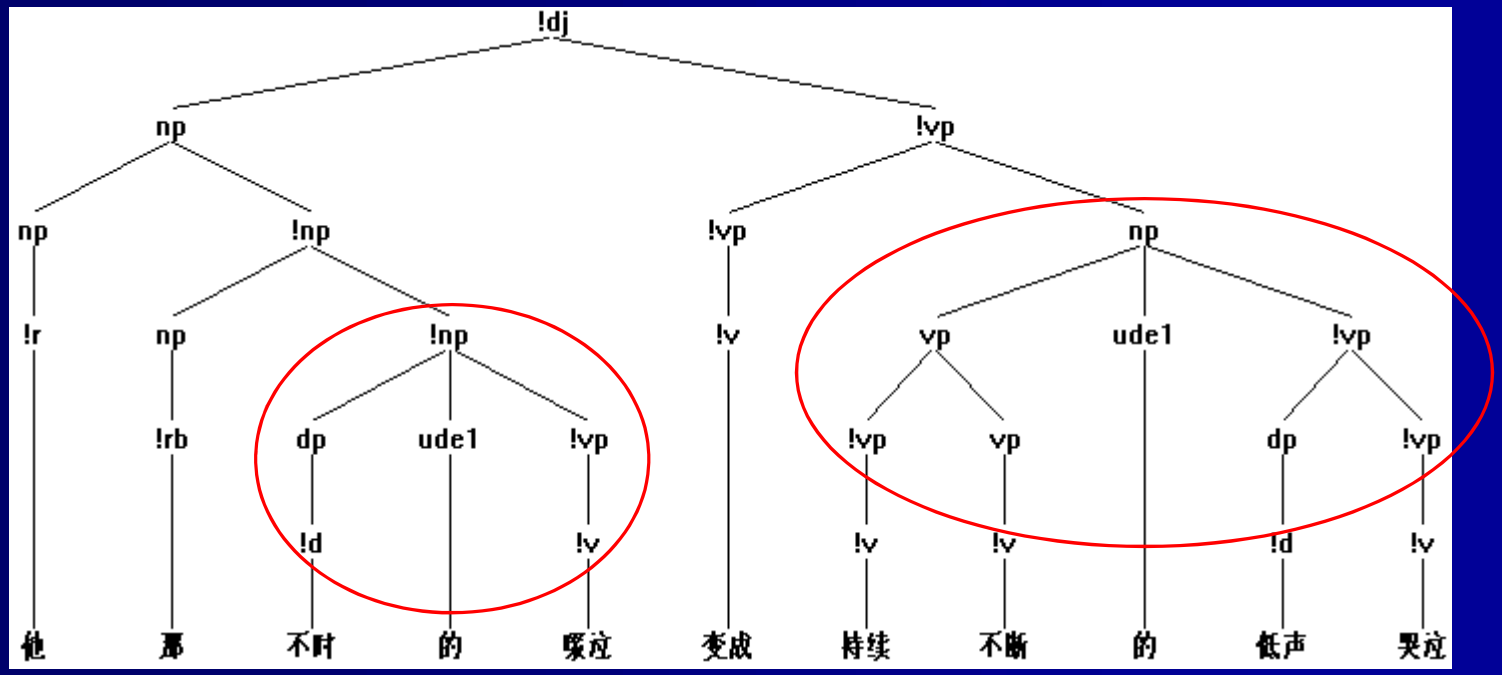
“X 的 （Y）” 短语 小结

基本格式： X 的 Y　　省略格式： X 的

	X 的 Y	X 的
内部（X、Y）的构成	X：指称（42%）陈述（58%） Y：指称（94%）陈述（6%）	X：指称（11%）陈述（89%）
整体的分布性质	与np同分布（98%） 其他（2%）	与np同分布（80%） 其他（20%）
整体的表述功能	指称（多） 陈述（少）	指称（多） 陈述（少）
“的”的性质	1. 把修饰成分和中心成分“间隔开” 2. 有一定的标记“指称性结构”的作用	1. 附着在修饰成分上。 2. 当修饰成分不再被解释为“修饰”成分时，“的”发展为语气词。

指称与陈述界限模糊的例子

- 他 那 **不时 的 啜泣** 变成 持续 不断 的 低声 哭泣



dp 的 vp

vp 的 vp

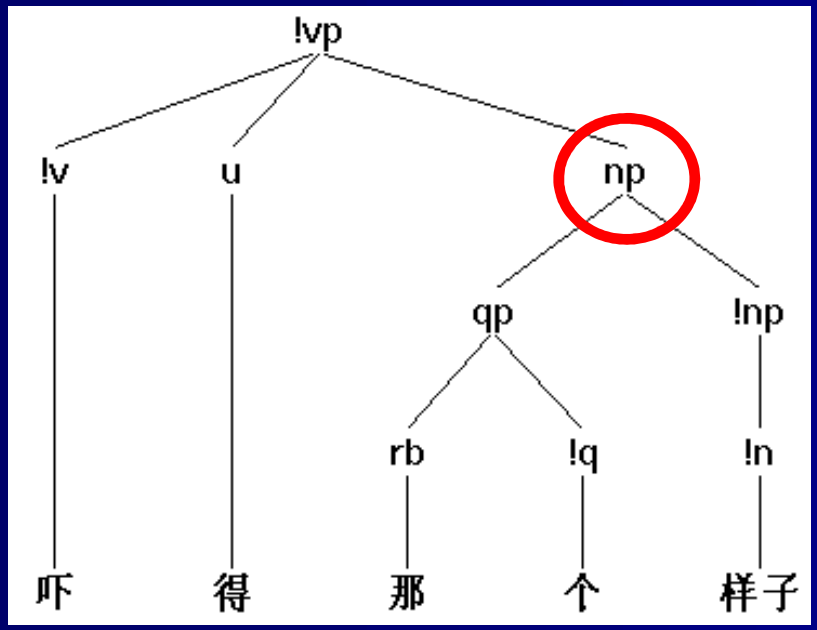
3.2 成分省略造成的功能变异分析

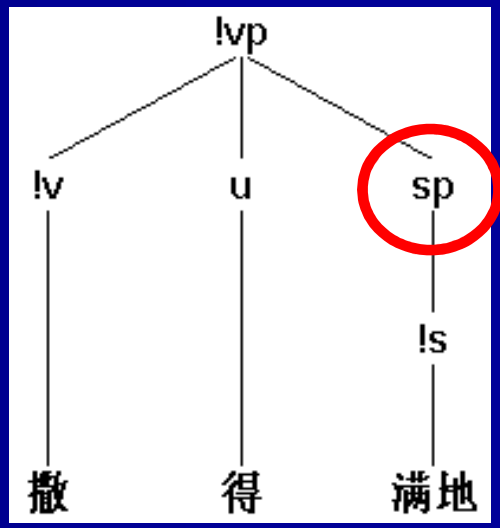
- 功能变异
  - 句法约束条件改变
    - 句法成分与句法位置
      - 体词性成分占据谓词性位置
      - 谓词性成分占据体词性位置
    - 中心成分与从属成分
      - 中心成分缺省
      - 从属成分缺省
  - 语义约束条件改变
    - 论元数发生变化
    - 论旨角色的约束条件发生变化

功能变异的后果

体词性成分占据谓词性位置

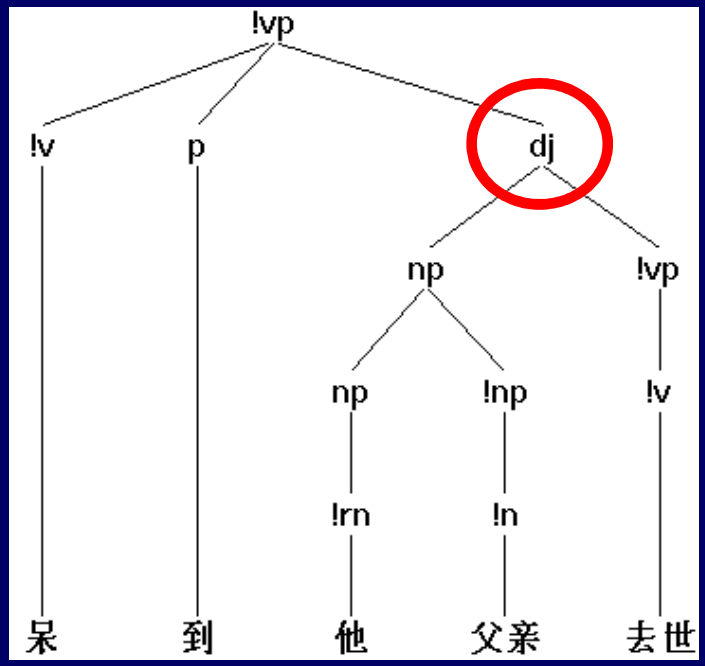
1. 看 你 把 闺女 **吓 得 那 个 样子**
2. 豆子 **撒 得 满地**

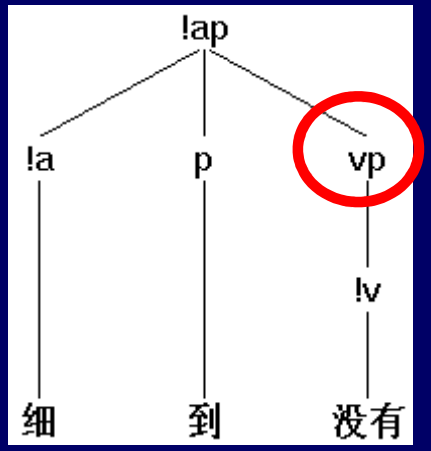


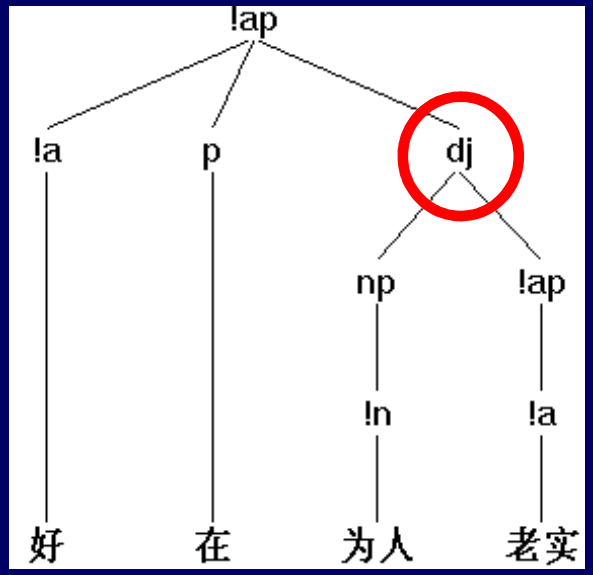


谓词性成分占据体词性位置

3. 一丝发抖的声音，在空气中愈颤愈细，**细 到 没有**，周围便都是死一般静。
4. 他在他父亲的公司里一直 **呆 到 他 父亲 去世**。
5. 他 好 就 **好 在 为人 老实**。







p+np	p+sp	p+tp	p+dj	p+vp	p+ap
6772	2513	992	190	524	24
93.3%			**6.7%**		

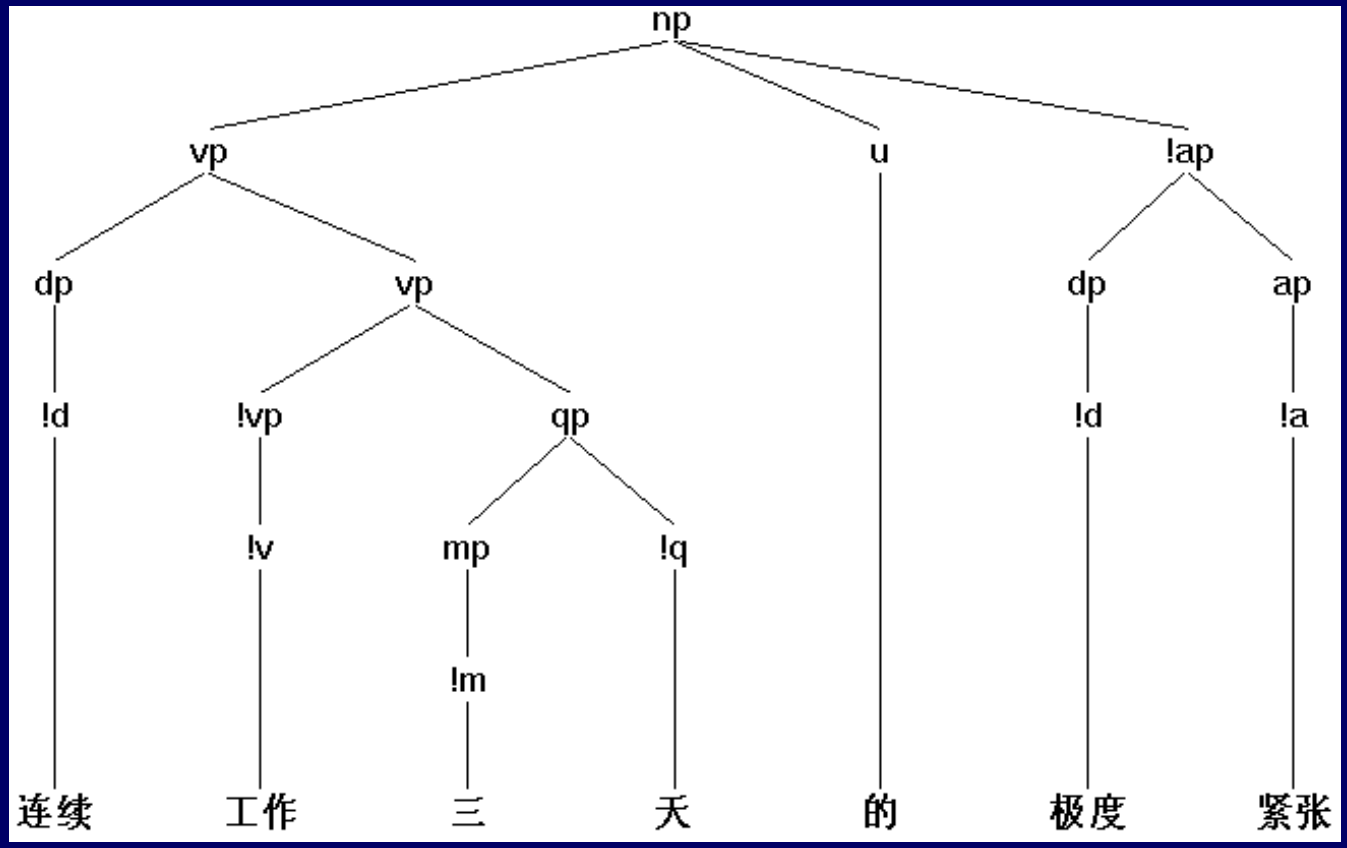
把+np	把+sp	把+tp	把+dj	把+vp	把+ap
1285	7	0	2	17	0
98.0%			**2.0%**		

被+np	被+sp	被+tp	被+dj	被+vp	被+ap
205	2	0	0	9	1
95.8%			**4.2%**		

在+np	在+sp	在+tp	在+dj	在+vp	在+ap
956	1644	392	0	4	0
99.9%			**0.1%**		

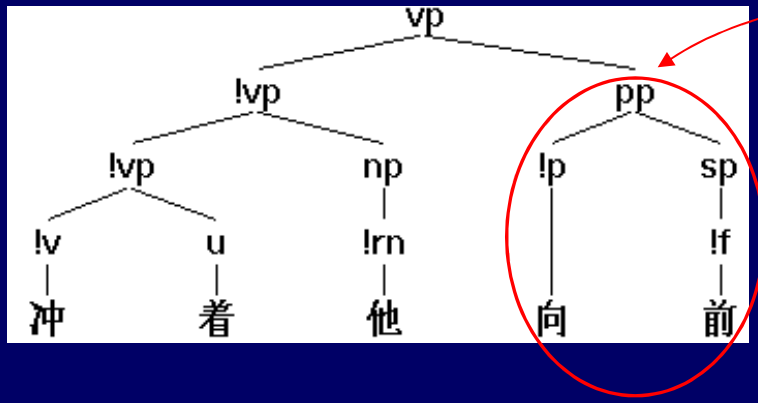
谓词性成分占据体词性位置

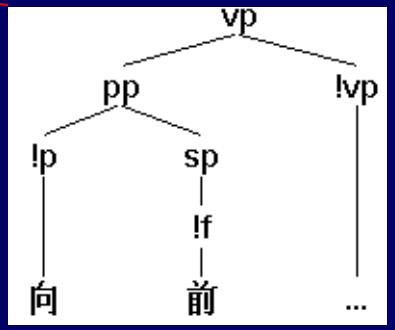
6. **连续 工作 三 天 的 极度 紧张** 使他几乎到了崩溃的边缘



中心成分缺省

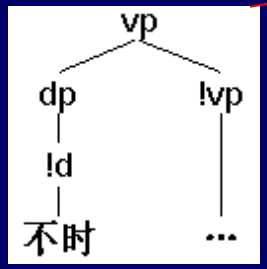
7. 让 河水 冲 着 他 **向 前**
8. 他 那 **不时 的 啜泣** 变成 持续 不断 的 低声 哭泣

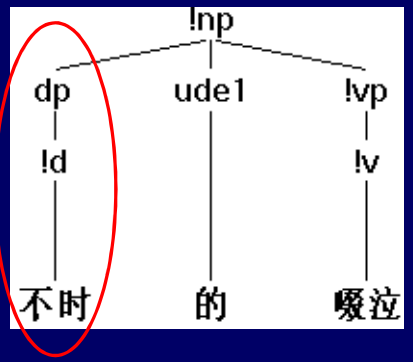




（静止）向前？

（运动）向前？

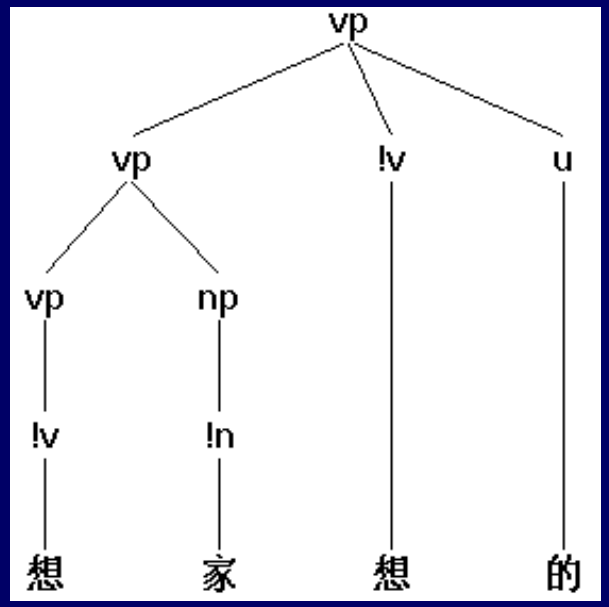


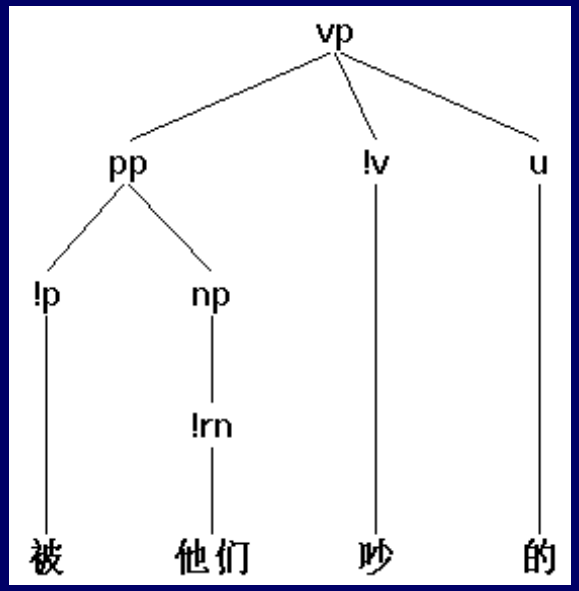


不时 发出/发生…

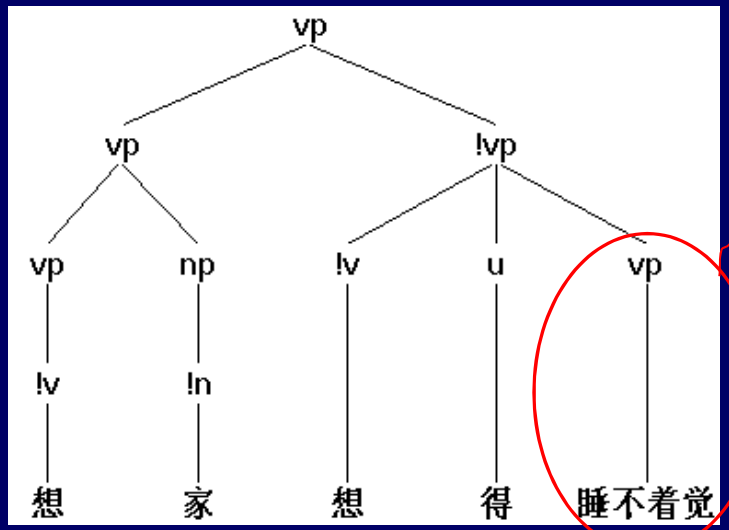
从属成分缺省

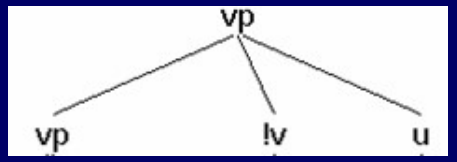
9. 他 是 **想 家 想 的**，晚上 总 睡 不 着 觉。
10. 你 为什么 睡 不 着 —— **被 他们 吵 的**



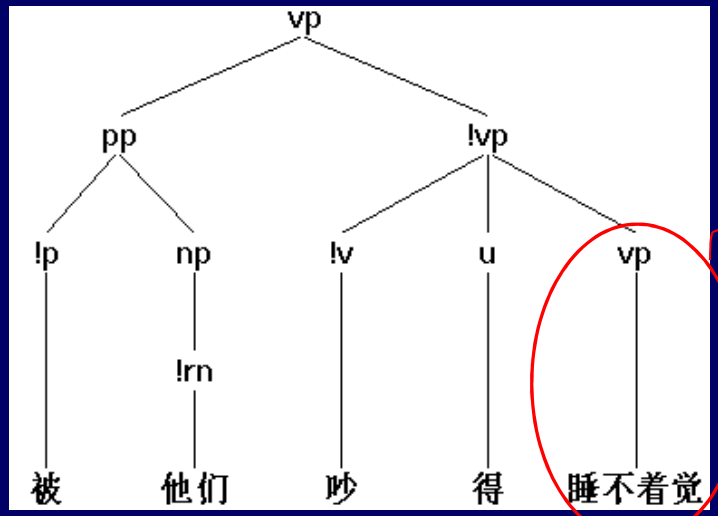


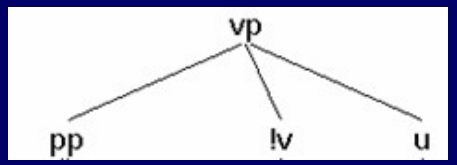
V X V de：的？ 得？





想家　想　得→的

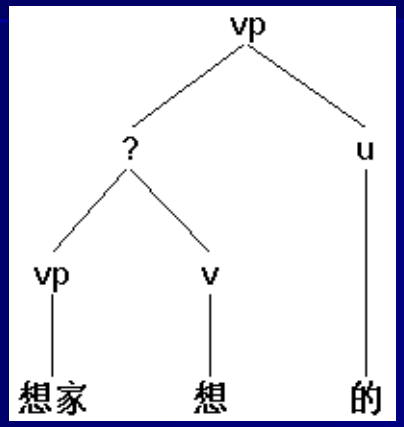


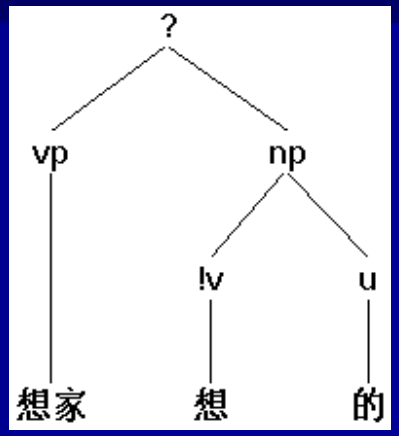


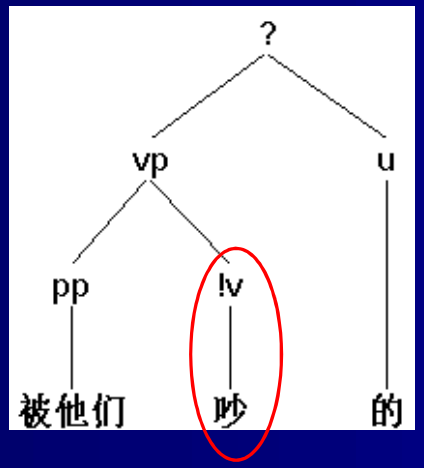
被他们　吵　得→的

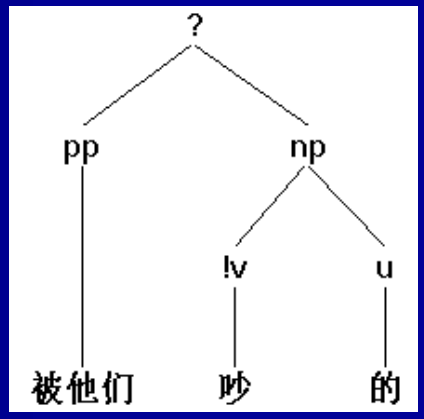
V X V de：的？ 得？

“的”
的困境





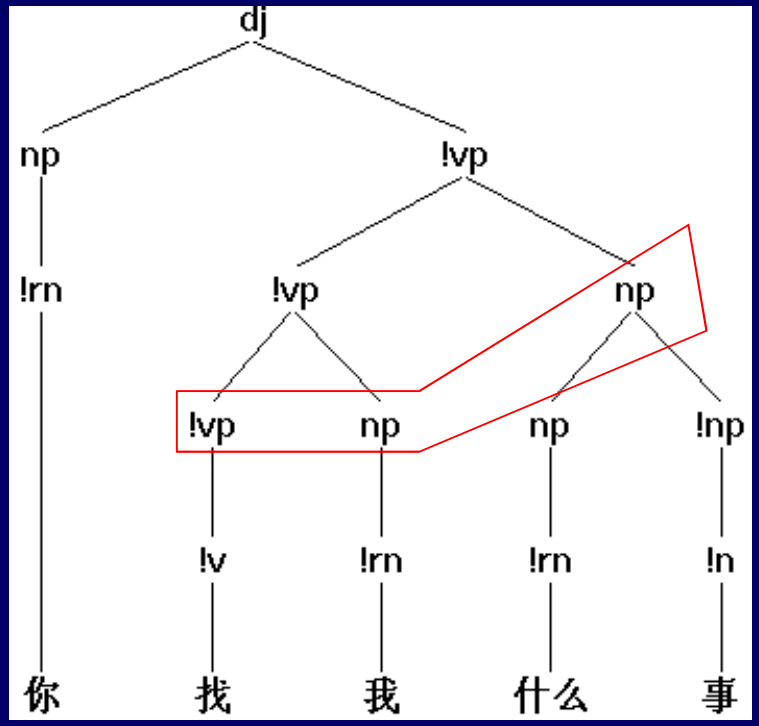




“把”“被”结构后面的vp不能是简单动词形式

论元数发生变化

11. 你 **找 我 什么 事**



找： 二价动词？

三价动词？

他 找 我 打 球

他 找 我 借 了 一 些 钱

你 找 我 干/做 什么事

? 他 找 你 三 件 事

* 他 找 你 几/多少/哪 件 事

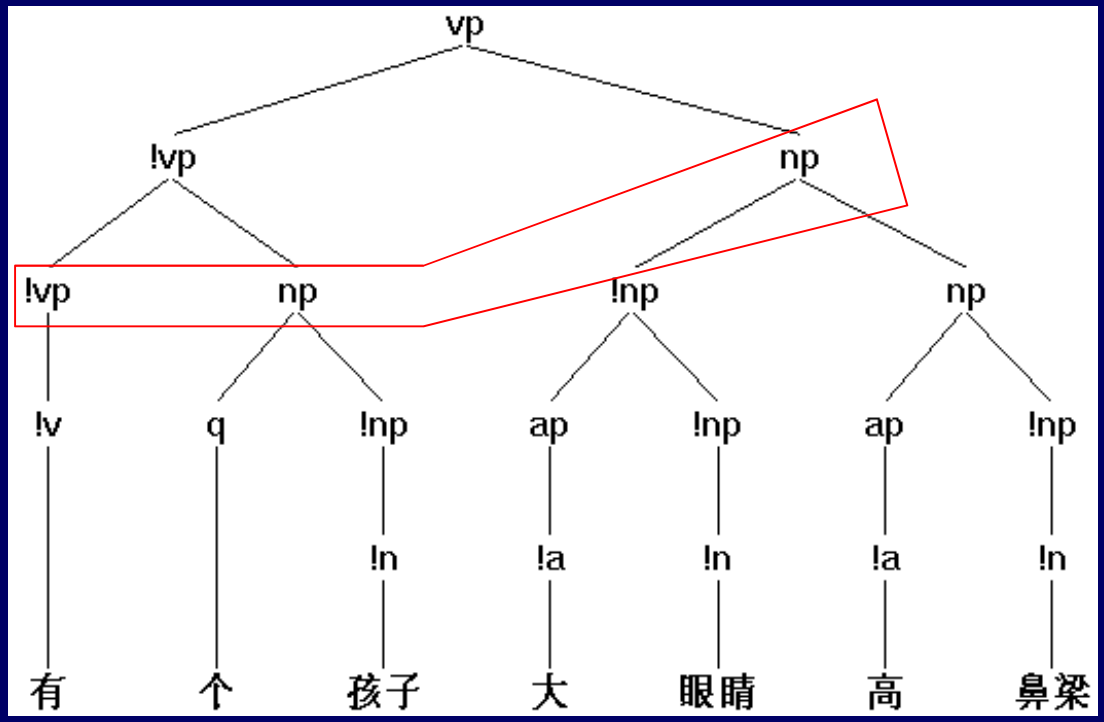
* 他 找 你 那 件 事

他 找 你 就 三 件 事

他 找 你 就 这 件 事

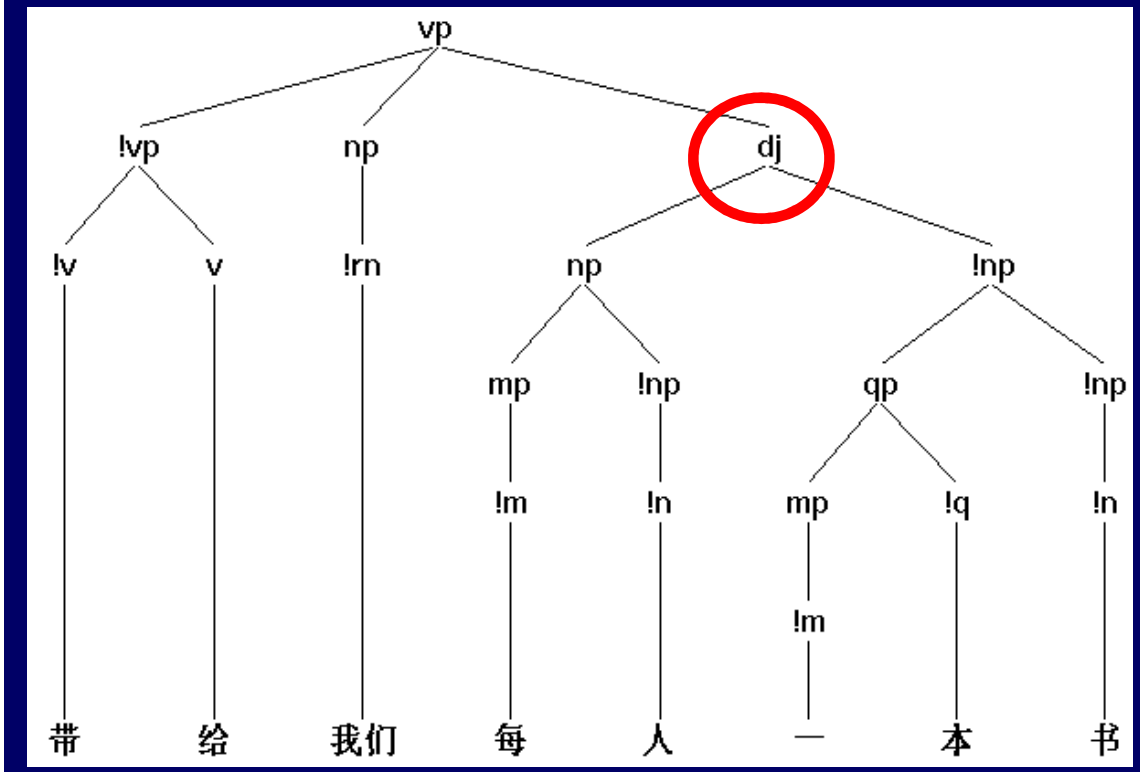
论元数发生变化

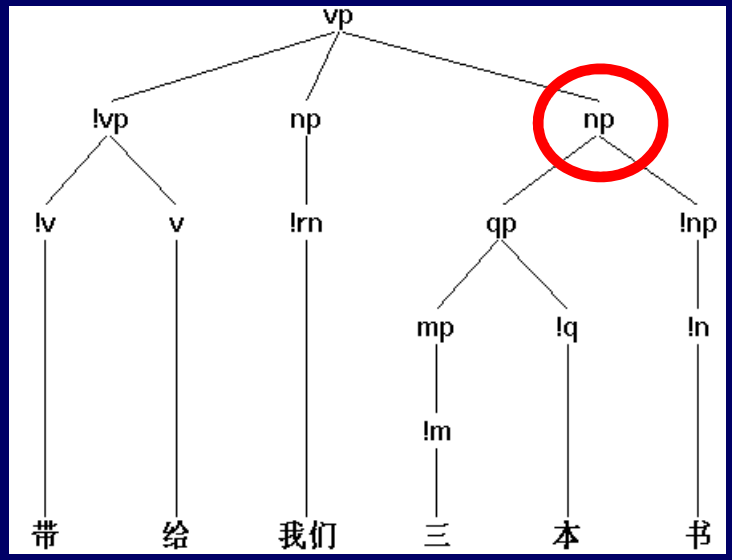
12. 有 个 孩子 大 眼睛 高鼻梁



论旨角色的约束条件发生变化

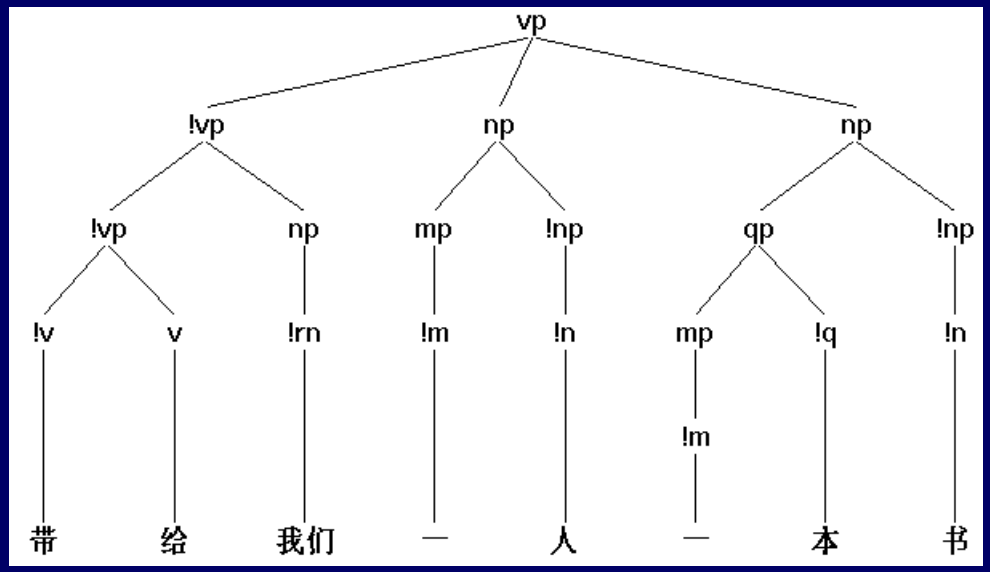
13. 老张 带 给 我们 每 人 一 本 书





带给1：___ np np

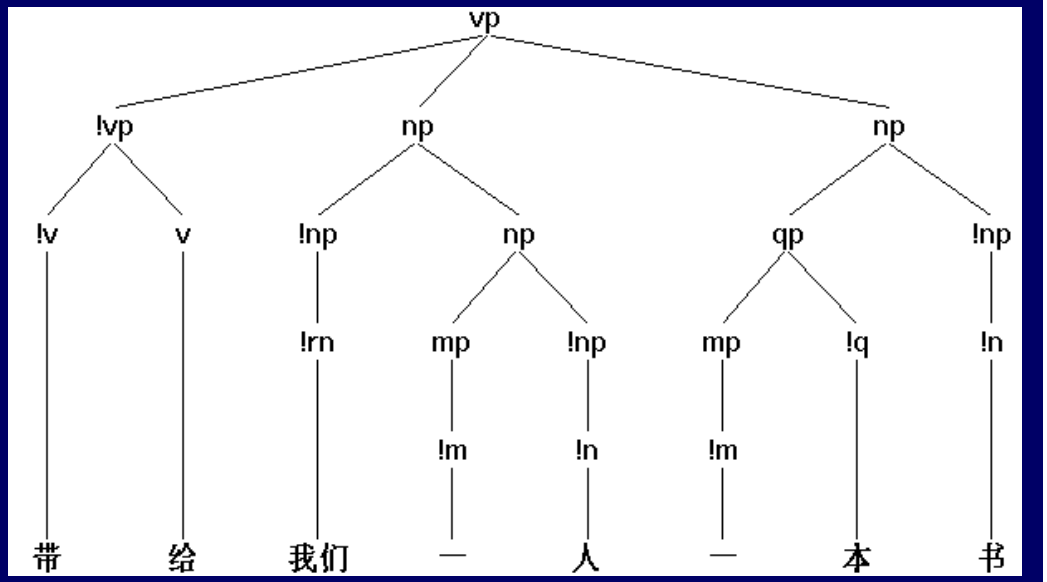
带给2：___ np dj



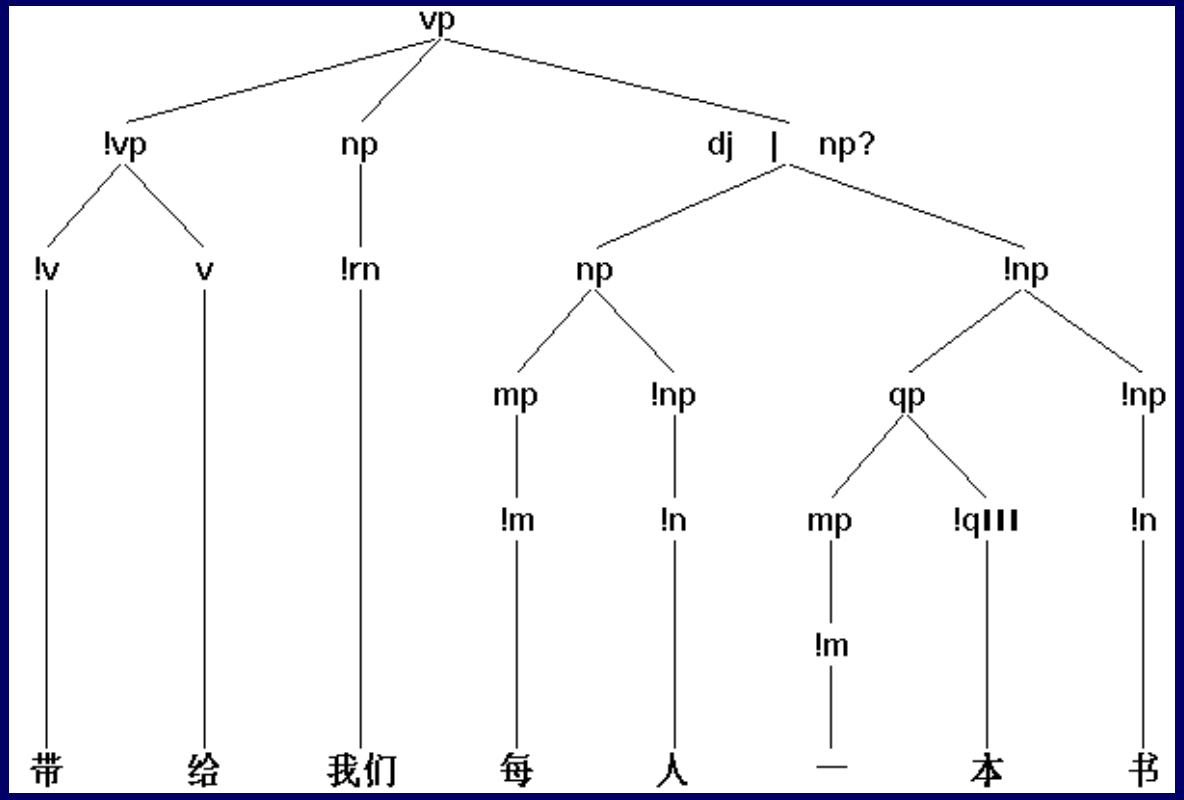
维持“带给”的论元结构不变？

⟶ 造成“带给”为4价动词

“我们一人”不能成立 ⟵



如何给“每人一本书”定性？



“每人一本书”是np，则维持了“带给”的论元结构，但这个np太特别！

“每人一本书”是dj，则造成dj能进入“带给”的论元位置！

“省略式”与“原式”的对比

- 每人 一 本书
- 他 八 岁

- 每人 分/发/买/... 一 本书
- 他 是/有/... 八 岁

- 带给 他们 每人 一本书
- 他 八 岁 那年

- * 带给 他们 每人 分/发/买 一本书
- * 他 是 八 岁 那年

省略（转指）造成的短语，即便归入“已有”的短语类，其功能跟“省略前”的构造也不可能完全相同。

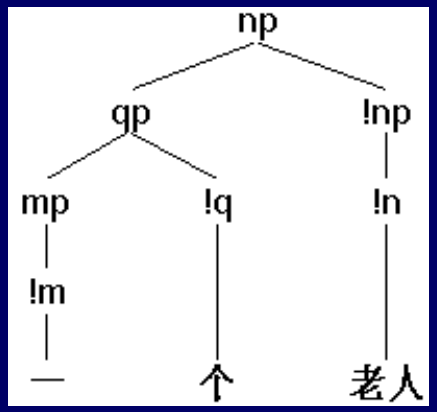
此外，我们认为，“省略式”理应比“原式”受到更多限制，因而分布功能较窄，除非“省略式”使用日久，不再被看作是“省略”，成为新的“常规格式”。

功能变异对句法分析的影响 —— 造成更多潜在歧义

1.	np → q !np	是 个 老人	m q n
2.	np → mp !np	一 老人 成功获救	m q n
3.	qp → mp !q	一 个 获救，一 个 遇难	m q n
4.	np → qp !np	一 个 老人 的 自述	m q n

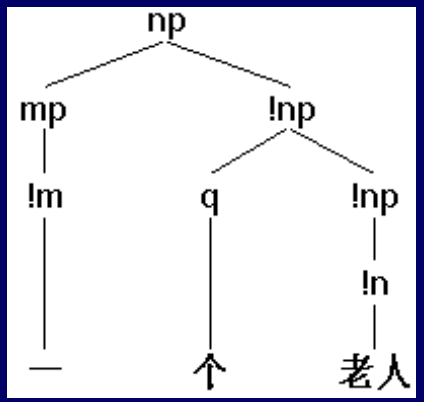
m q n 三个成分任何一个都可以省略

先规则4，再规则3



√

先规则2，再规则1



- 组合模式增加
- 组合条件改变

m q n 的各种形式

	m	q	n	示例
1	+	+	+	一 个 老人
2	-	+	+	个 老人
3	+	-	+	一 老人
4	+	+	-	一 个
5	+	-	-	一
6	-	+	-	个
7	-	-	+	老人
8	-	-	-	

潜在歧义

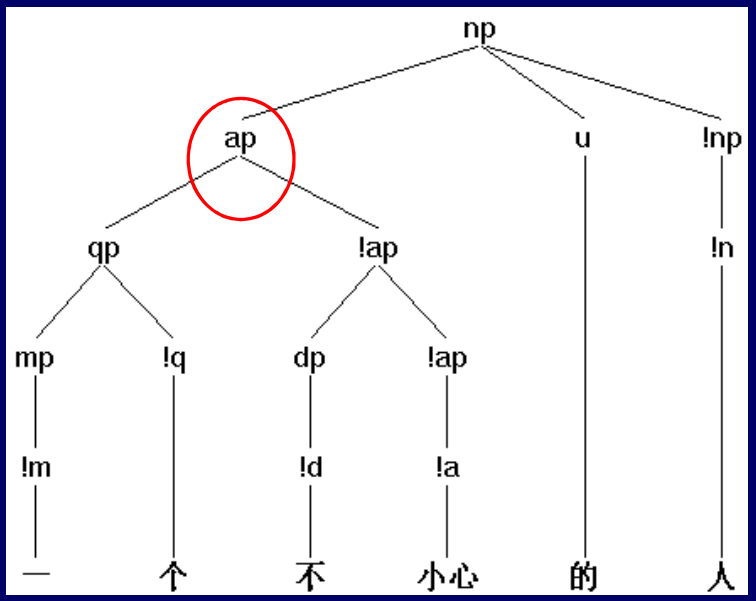
组合模式增加，潜在歧义增多

- 有 十倍 那么大
- （其中）一个 不小心
- 一 个 不 小心 的 人

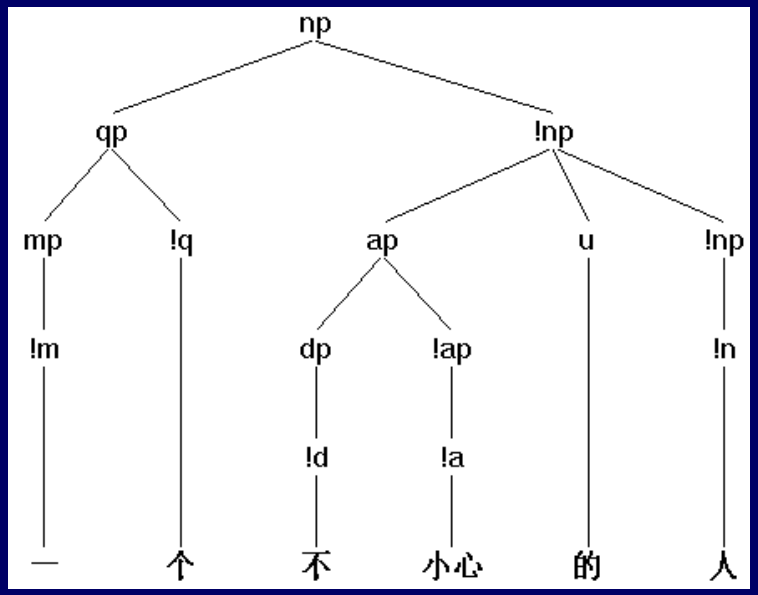
np → qp !np

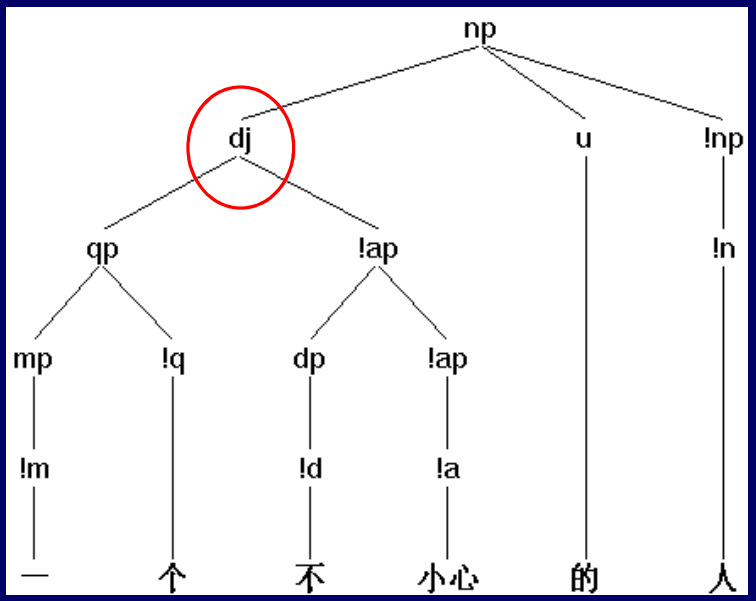
ap → qp !ap

dj → qp !ap



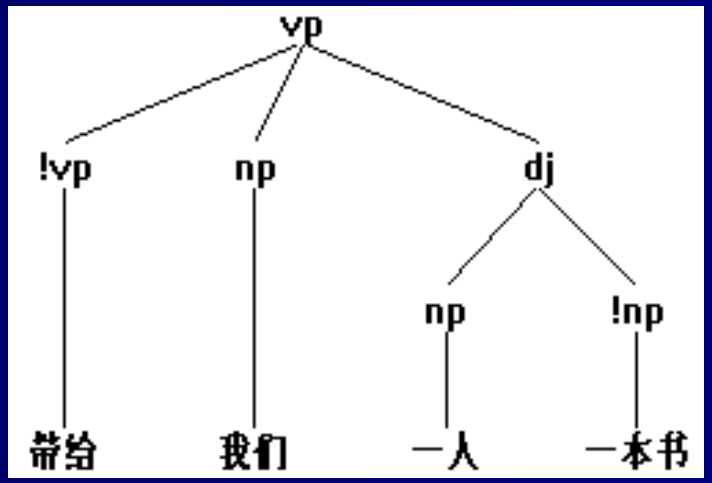
√

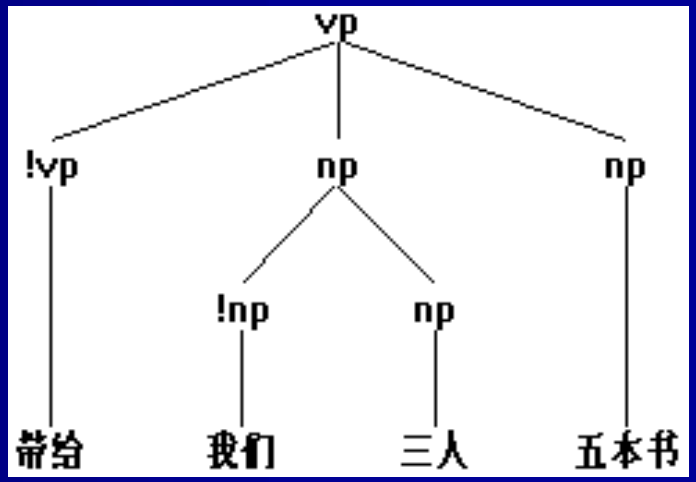




潜在歧义

1. （他）带给我们一人一本书
2. （他）带给我们三人五本书





结语

- 树库加工过程，可以看作是对语言学理论的一个检视过程，语法理论中建立的语法范畴覆盖面（适用性）如何，在加工过程中可以全面体现出来。
- 建好的树库，可以直接检索一个语言结构的分布情况；可以统计各种句法结构的频次。其中低频的分布（组合），可以为发现“非常规性”的语言现象（比如省略式）提供线索。
- 通过加工中文树库，我们体会到：汉语词语没有语法形态变化。词和短语所属的类别（范畴）主要是一种语义（表达功能）类，同一个语义类的成分在分布上自然会形成一定的同分布聚合，即语法类。通过树结构观察分布，有助于进一步发现同一类中成员的差异。

附录：树库标注的语言学问题示例

1. 我们曾家人都是**读书第一**。
2. 这些项目的建设时间，**最长三十个月**，短的只有十一个月。平均建设周期为十七点二个月。
3. **好家伙**！
4. 我应该**今天开始还是明天**？
5. **二五一十，五五二五**。
6. 下劣、凶残 到 这种地步
7. 经度的所以发生影响，是离海洋远近的关系。
8. 达 **34 座 之 多**
9. 这当然是**再保险不过**的了
10. **连夜三班**，急急忙忙，完成这一环节之后，还得等待旁的环节才能装配。

树库标注的语言学问题示例

11. 这时，原子核通常还会以光的形式释放出能量（称为**γ**射线）
12. 他**不肯也罢了**，连个回信也不给。
13. 全年 **国有 及 国有 控股 企业** 增加值 一点一七二六万亿元
14. 常常会出现 **皮肤潮红**、**出疹**、**头痛**、**恶心**等副作用
15. 这天，风雨又急又大，小乌鸦**一**早**就**飞出去找食物，为了不让妈妈担心，它们**一**找到食物，**就**飞回窝里去。
16. 早晨**七点差十分**到八点半左右
17. **这么一个破茶馆**竟然在市中心，是怎么回事？

参考文献

- Chu-Ren Huang, Feng-Yi Chen, Keh-Jiann Chen, Zhao-ming Gao & Kuang-Yu Chen.(2000). Sinica Treebank: Design Criteria, Annotation Guidelines, and On-line Interface, In *Proceedings of the Second Chinese Language Processing Workshop, HongKong*. pp.29-37.
- Nianwen Xue. 2005. Annotating discourse connectives in the Chinese Treebank, in Proceedings of the ACL Workshop in Frontiers in Annotation II: Pie in the Sky . Ann Arbor, Michigan.
- Nianwen Xue, Fei Xia, Fu-Dong Chiou & Marta Palmer (2005) The Penn Chinese Treebank: Phrase structure annotation of a large corpus, In *Natural Language Processing 11* (2): pp.207-238. Cambridge University Press.
- Mitchell P. Marcus, Beatrice Santoriniy, Mary Ann Marcinkiewicz, 1993, Building a large annotated corpus of English: the Penn Treebank, Computational Linguistics, Vol.19, No.2.
- 邓思颖（2006）以“的”为中心词的一些问题 《当代语言学》第3期， 205 - 12页。
- 郭锐（2000）表述功能的转化和“的”字的作用《当代语言学》2000年第1期，37-52页。
- 李艳惠（2008）短语结构与语类标记：“的”是中心詞？《当代语言学》2008年第2期，97-108页。
- 姬东鸿（2009）汉语树库综述，《当代语言学》2009年第1期。
- 陆丙甫（2006）不同学派的“核心”概念之比较 《当代语言学》第4期，289 - 310页。
- 陆俭明（2003） “对NP的+VP”结构的重新认识 《中国语文》第5期，378 - 391页。
- 陆俭明（1983）“的”字结构和“所”字结构。载中国语文杂志社編《语法研究和探索》（一）北京大学出版社。57 – 68页。
- 司富珍（2004）中心词理论和汉语的DeP 《当代语言学》第1期，26 - 34页。
- 司富珍（2006）中心语理论和“布龙菲尔德难题” 《当代语言学》第1期，60 - 70页。
- 熊仲儒（2005）以“的”为核心的DP结构《当代语言学》第2期，148 - 65页。
- 袁毓林（2003）从焦点理论看句尾“的”的句法语义功能《中国语文》2003年第1期。
- 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。
- 詹卫东（2000）语言成分的組合与功能传递，载陆俭明主编《面临新世纪挑战的现代汉语语法研究》，山东教育出版社。
- 周国光（2005）对“中心词理论和汉语的DeP”一文的质疑 《当代语言学》第2期，139 - 47页。
- 周国光（2006）括号悖论和“的X”的语感——“以‘的’为核心的DP结构”疑难求解 《当代语言学》第1期，71-75页。
- 周强（2004）汉语句法树库标注体系，《中文信息学报》2004年第4期，1-8页。
- 朱德熙（1961）说“的”，《中国语文》1961年12月号。
- 朱德熙（1978）“的”字结构和判断句，《中国语文》1978年第1、2期。