现代汉语短语结构标记规范的制定原则及相关问题

詹卫东 应晨锦
北京大学中文系
北京大学汉语语言学研究中心
{zwd; yingchenjin}@pku.edu.cn

本研究工作得到国家语言文字应用研究“十五”科研项目资助

信息处理用现代汉语短语及句型标记规范

(项目编号：YB105-49)

提 纲

1) 国内有关汉语句型系统的研究
2) 国内外的树库（Treebank）研究
3) 我们对汉语短语结构标注的认识
4) 短语结构分析中的若干问题
5) 介绍一个正在建设中的汉语树库

一 国内有关汉语句型的系统研究

- 面向语言教学
- “扁平”线性序列的描述模式
- 以句子成分（主、谓、宾等）做为主要的句型构造成分，同时也包含词类成分（比如“动词”）和一些特殊词语（比如“是”“有”等）

具体的句型系统：

- 各家语法著作：《汉语知识》《中学教学语法系统提要）《现代汉语八百词》《现代汉语》（黄廖本、胡本） 80年代以前
- 北京语言学院句型研究小组（219个句型） 80年代末-90年代
- 清华大学汉语句型自动分析和分布统计研究（209个句型） 90年代

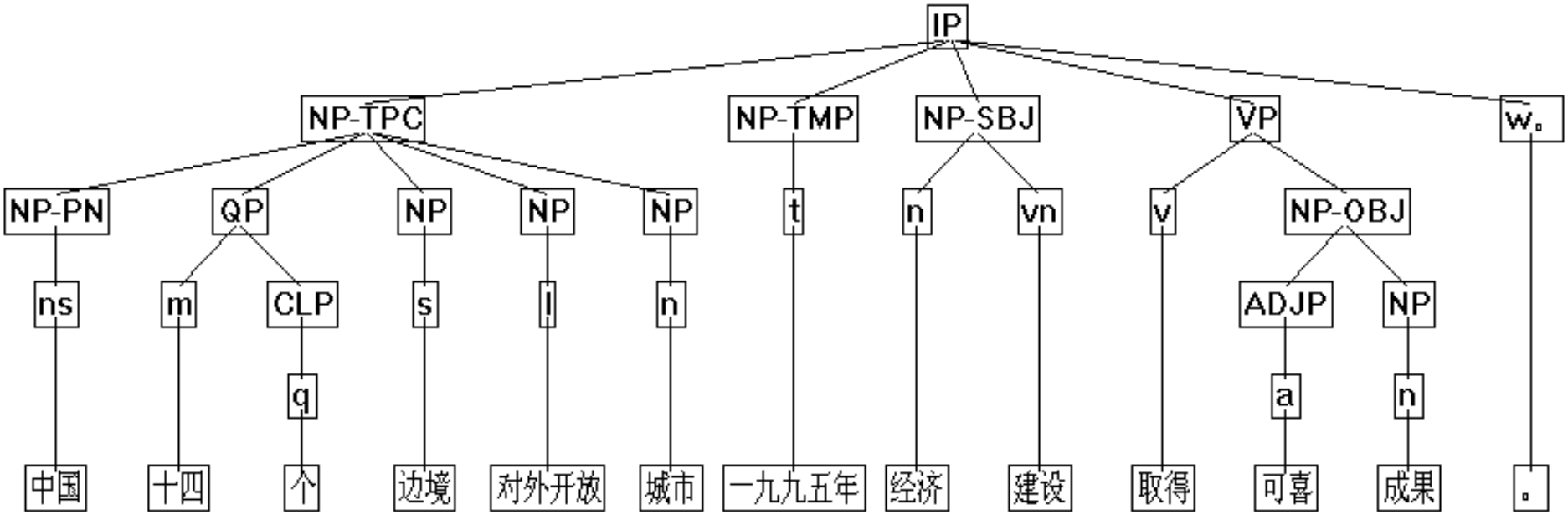
二 国内外的树库（Treebank）研究

- 面向自然语言处理
- 句法结构的层级树状描述模式
- 以词和短语功能类为主要句法标记，来刻画句子的组成类型，同时也包含一些特殊的词语标记（比如“的”）

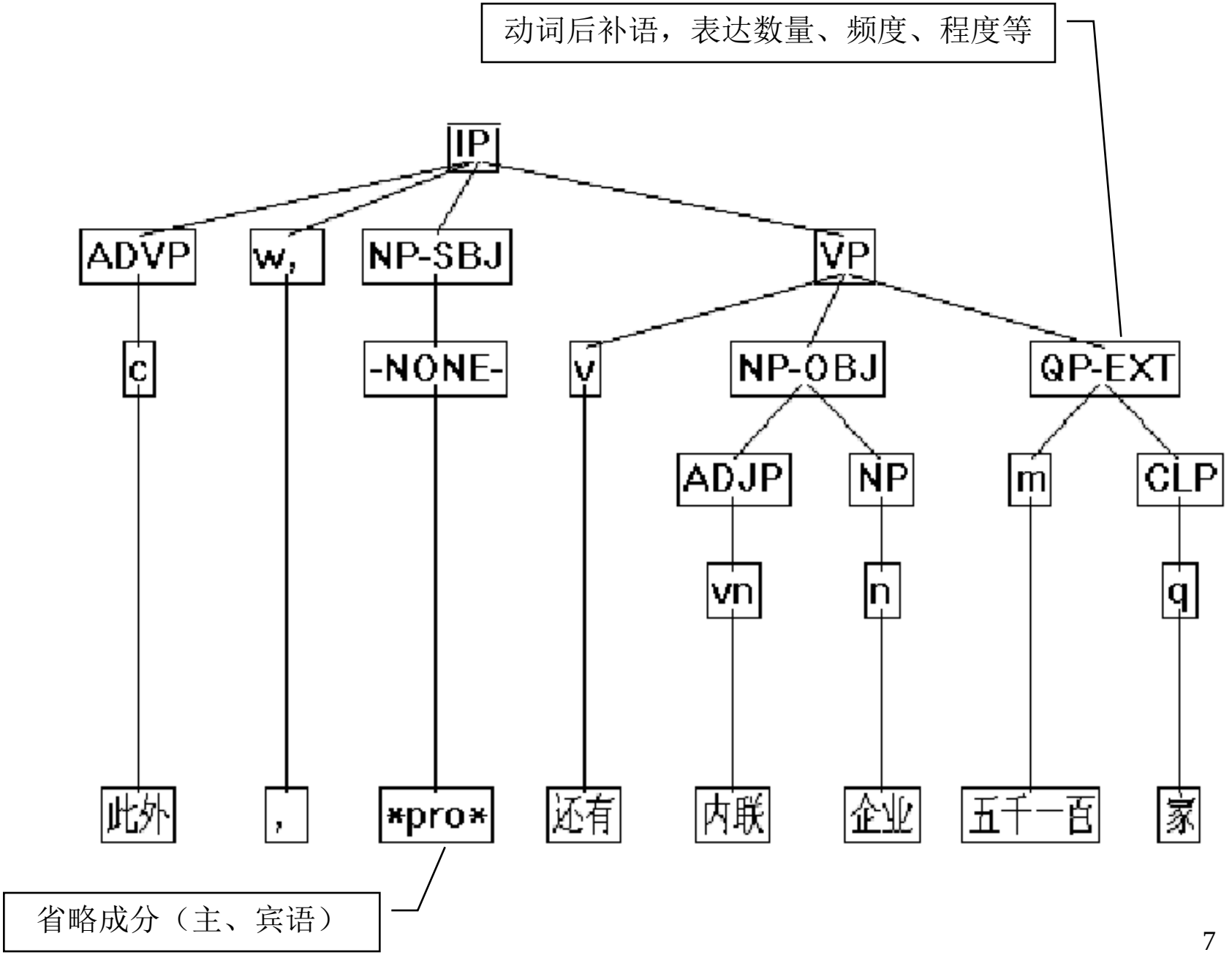
具体的树库项目：

- 清华大学周强的树库研究，2003年 （100万字）
- 美国宾州大学中文树库的研究，2000年 （17万字）

UPenn Chinese Treebank示例



(IP (NP-TPC (NP-PN (ns<中国>) QP (m<十四> CLP (q<个>)) NP (s<边境>) NP (l<对外开放>) NP (n<城市>)) NP-TMP (t<一九九五年>) NP-SBJ (n<经济> vn<建设>) VP (v<取得> NP-OBJ (ADJP (a<可喜>) NP (n<成果>))) w。<。>))


动词后补语，表达数量、频度、程度等
IP
ADVP
w,
NP-SBJ
VP
c
-NONE-
v
NP-OBJ
QP-EXT
ADJP
NP
m
CLP
vn
n
q
此外
,
pro
还有
内联
企业
五千一百
家
省略成分（主、宾语）

三 我们对汉语短语结构标注的认识

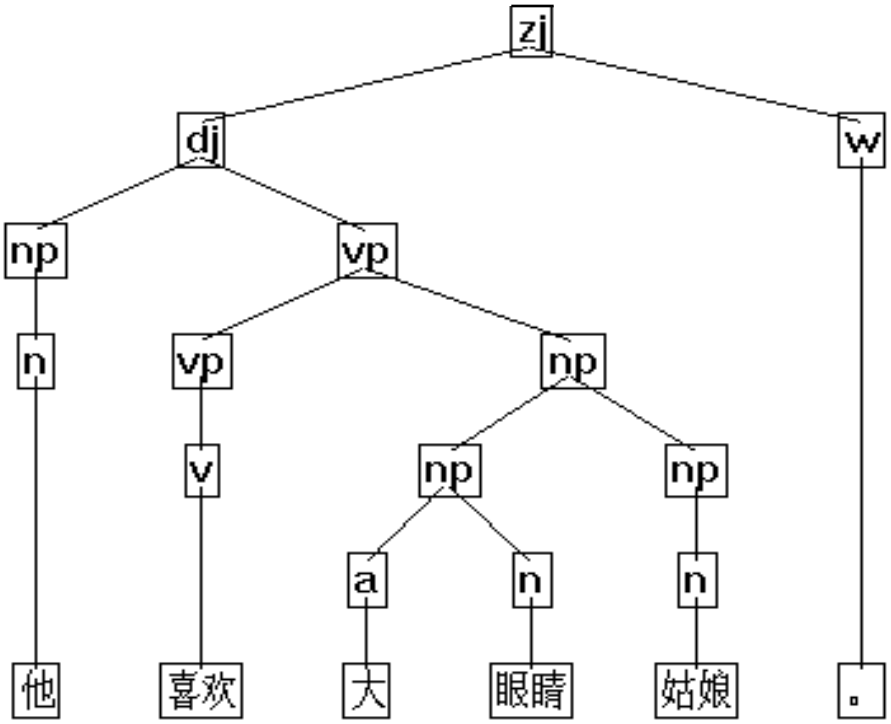
- 标注内容：

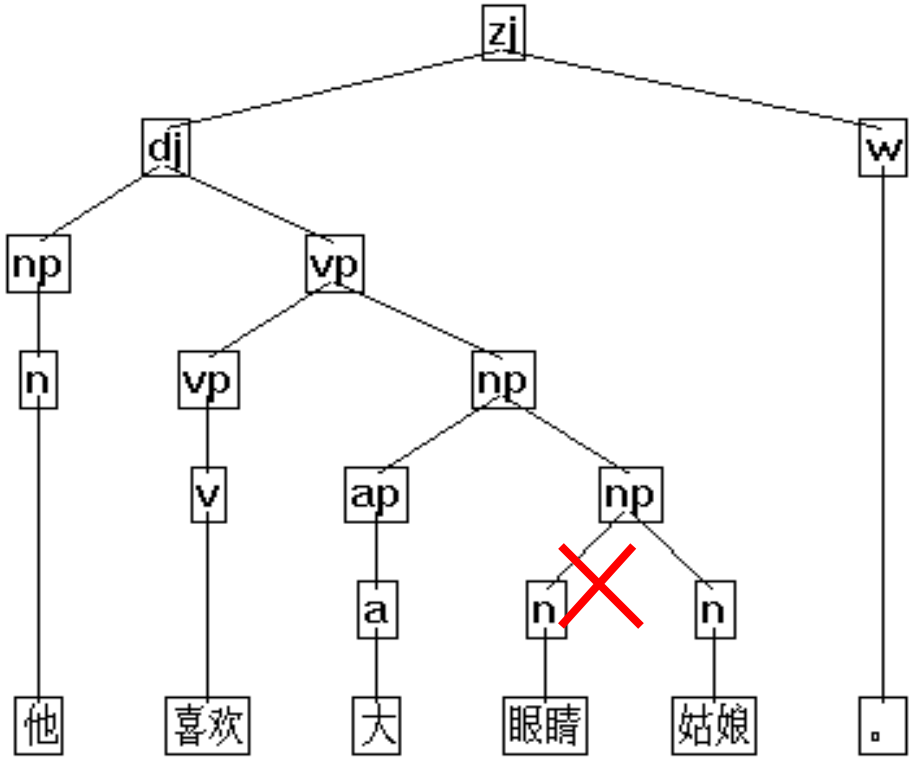
结构层次 —— 定界

短语类型 —— 定性

- 结构类型标记
- 句法功能标记
- 语义功能标记
- 篇章功能标记

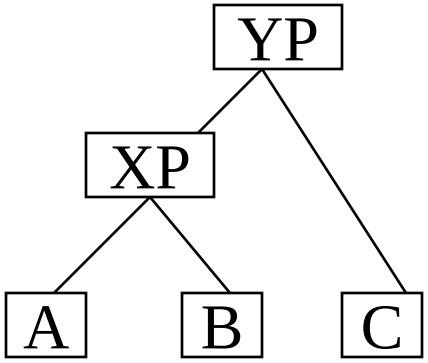
定界问题


zj
dj
w
np
vp
n
vp
np
v
np
np
a
n
n
他
喜欢
大
眼睛
姑娘
。

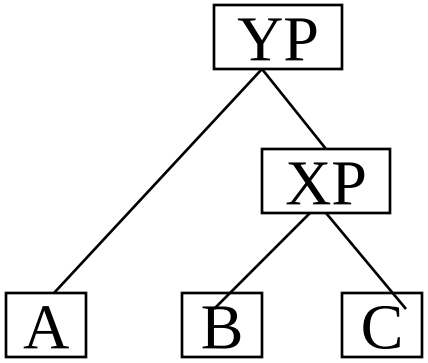

zj
dj
w
np
vp
n
vp
np
v
ap
np
a
n
n
他
喜欢
大
眼睛
姑娘
。

n v a n n

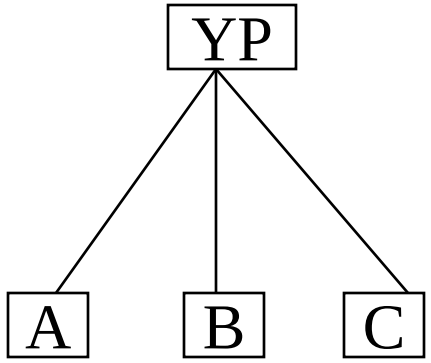
短语结构定界的原则



P1



P2



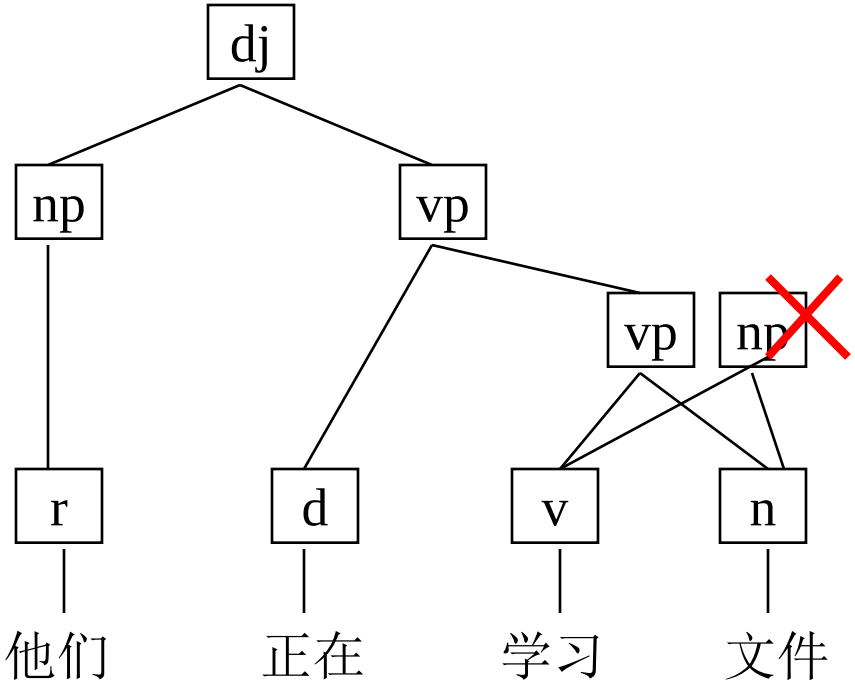
P3

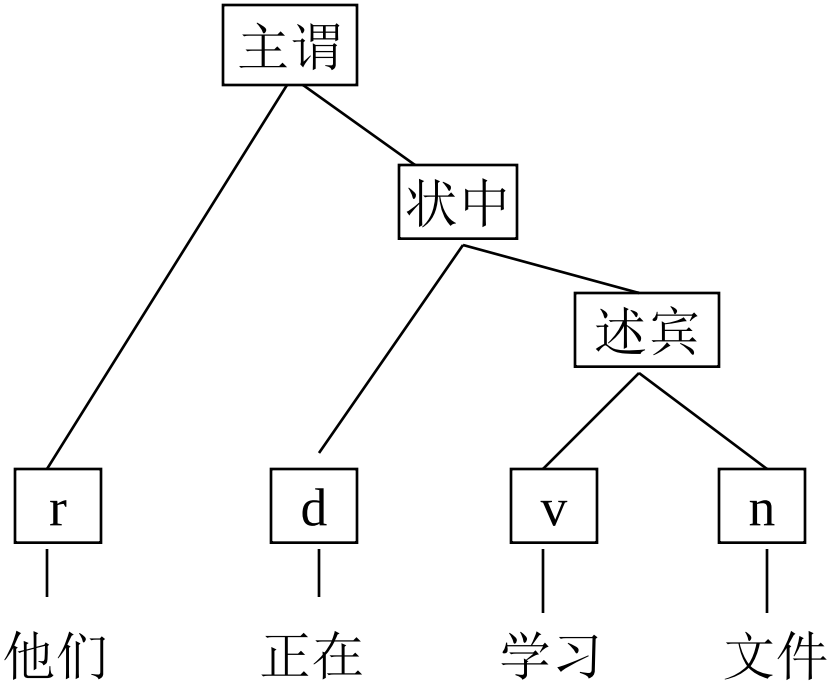
（1）不能分析为P1和P2时，才分析为P3；

（2）对于多切分结构，统一规定一种切分方式；

（3）对于歧义结构，要按照具体语境中的意思来进行切分；

定性问题





短语结构定性的原则

(1) 结构标记是“向内”看的结果，功能标记是“向外”看的结果。

短语结构标注的目标是描述短语的组合能力，所以应该优先使用功能标记；

(2) 一个短语结构单位的属性是有层次性的，可以根据需要分级描述，因此短语结构标记应体现层次性；

(3) 确定一个短语结构单位(xp)的功能类，应综合考虑三个因素：①xp的句法位置; ②xp的中心成分; ③xp的内部结构关系；

(4) 对于兼类短语（歧义短语），应根据具体语境中的意思进行归类。

对定界和定性原则的解释

定界：

- 二分结构使产生式规则可重用度高，规则总数少
- 多分结构使结构内部成分之间的相互约束易于描述

定性：

- 功能标记比结构标记更适宜组织规则
- 分级标记使得树库的扩充性和兼容性比较好

标注树库所用的功能标记

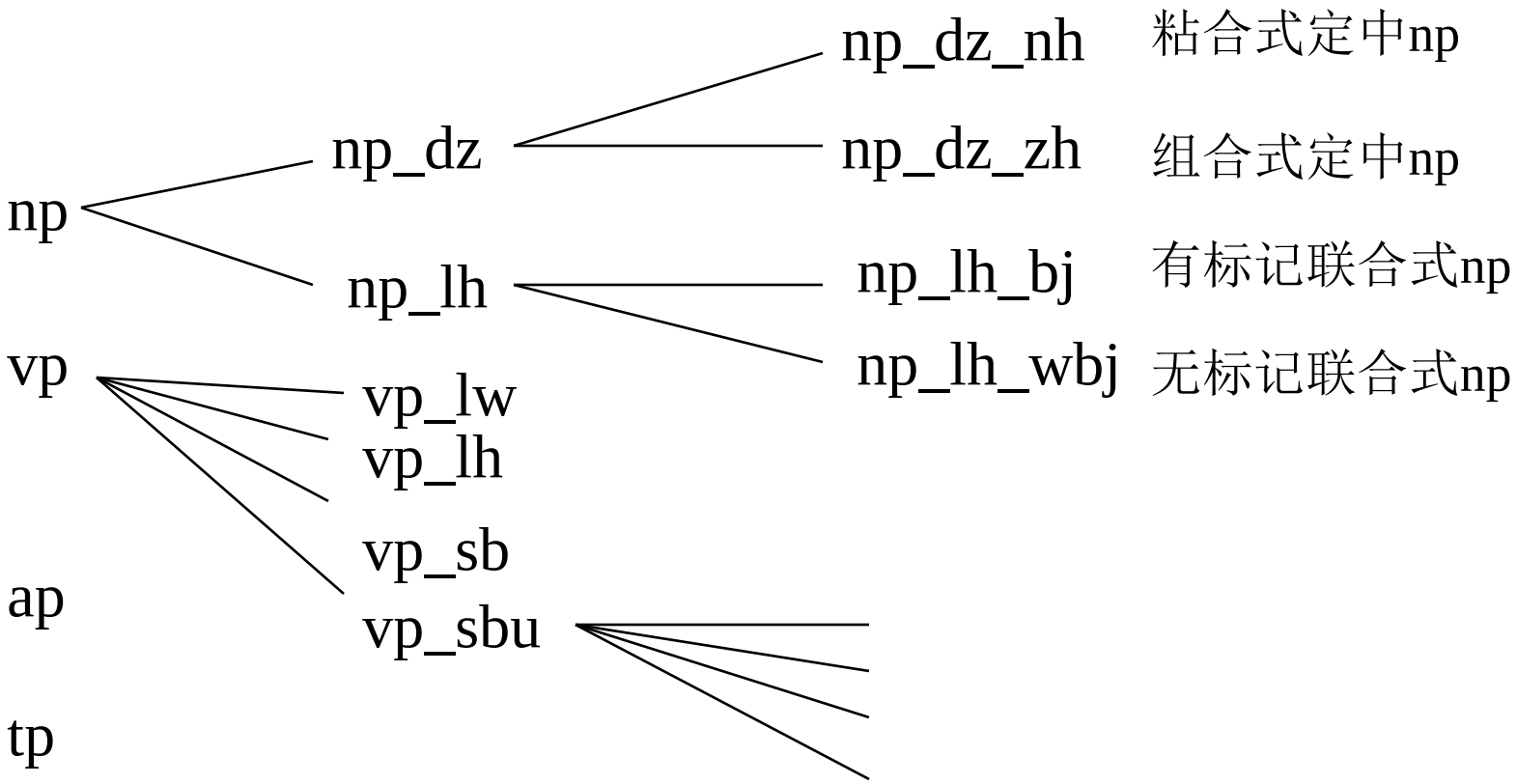
第一级	zj	整句（形式标记是句号等）	不被其他任何标记包含，只能包含第二级标记。
第二级	dj	小句	既可以包含第二级或第三级标记，又可以被第二级或第一级标记包含。
	fj	复句	
	np	名词短语	
	vp	动词短语	
	……	……	
第三级	n	名词	不包含其他任何标记，只能被第二级标记包含。
	v	动词	
	a	形容词	
	……	……	

结构与功能

序号	标记	功能类名称	典型功能												
			a	b	c	d	e	f	g	h	i	j	k	l	m	n
1	dj	小句型短语		+		+										+
2	np	名词性短语	+			+			+	+				+		+
3	vp	动词性短语		+	+		+					+	+	+	+	+
4	ap	形容词性短语		+			+	+	+		+	+		+	+	+
5	dp	副词性短语									+					
6	pp	介词性短语									+		+			
7	sp	处所词性短语				+										+
8	tp	时间词性短语				+										+
9	mp	数量短语							+							+
10	mcp	数词短语							+							+

a:作主语；b:作谓语；c:作述宾结构的述语；d:作宾语；e:作述补结构的述语；f:作补语；g:作定语；h:作定中结构的中心语；i:作状语；j:作状中结构的中心语；k:作连谓结构前后项；l:作联合结构前后项；m:带“着了过”等时体标记；n:“的”前位置

短语标记的层级体系



短语内部结构分类是对短语功能分类的进一步细分类

结构与功能

功能标记	数量 + _____	vp + _____
np_dz	+ （三双皮凉鞋）	+ 看语法书
np_lh	- （*三双皮鞋和凉鞋）	- * 看书和音乐 喜欢书和音乐

功能标记	_____ + np	ap + _____
np_dz_nh	+ （语法理论基础）	+ （高档 红木家具）
np_dz_zh	- （* 他的理论 基础）	- （* 高档 红木的家具）

短语类型-结构关系-中心词之间的对应

功能类	结构关系	中心词
vp	SB SBU LW ZZ LH	v
np	DZ DE LH TW	n v a
dj	ZW ZZ FJ	v n q m z a ~p ~b
ap	ZZ LH SB SBU	a
mp	DZ LH	m q
pp	JB LH	p
sp	DZ LH	s f n
tp	DZ LH	t f n
dp	FJ LH	d a v n

其他标记

- fj：复句
  1) 功能与dj类似
  2) 结构组成比dj复杂

- yp：语篇成分 — 不参与整句的结构分析
  1)插入成分　　据说，他喜欢大眼睛姑娘
  2)呼语　　张三，你喜欢大眼睛姑娘吗？
  3)引语　　他问我，“张三，你喜欢大眼睛姑娘吗？”

  yp_cr　　yp_hy　　yp_yy

四 短语结构分析中的若干问题

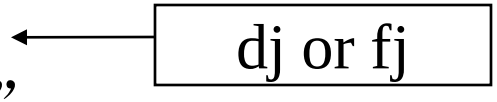
- “现在上课”： dj 还是 vp
- “看了三天”： vp_sb 还是 vp_bu

- “坐的飞机”： “的”？
- “骑了三年的破自行车”
- “太大了”： “太 + 大了” 还是 “太大 + 了”

- “在市长的带动下”
- “从这些小事上”
- “从三点到八点”
- “张三偷了李四一百块钱”
- “拿出一本书来”

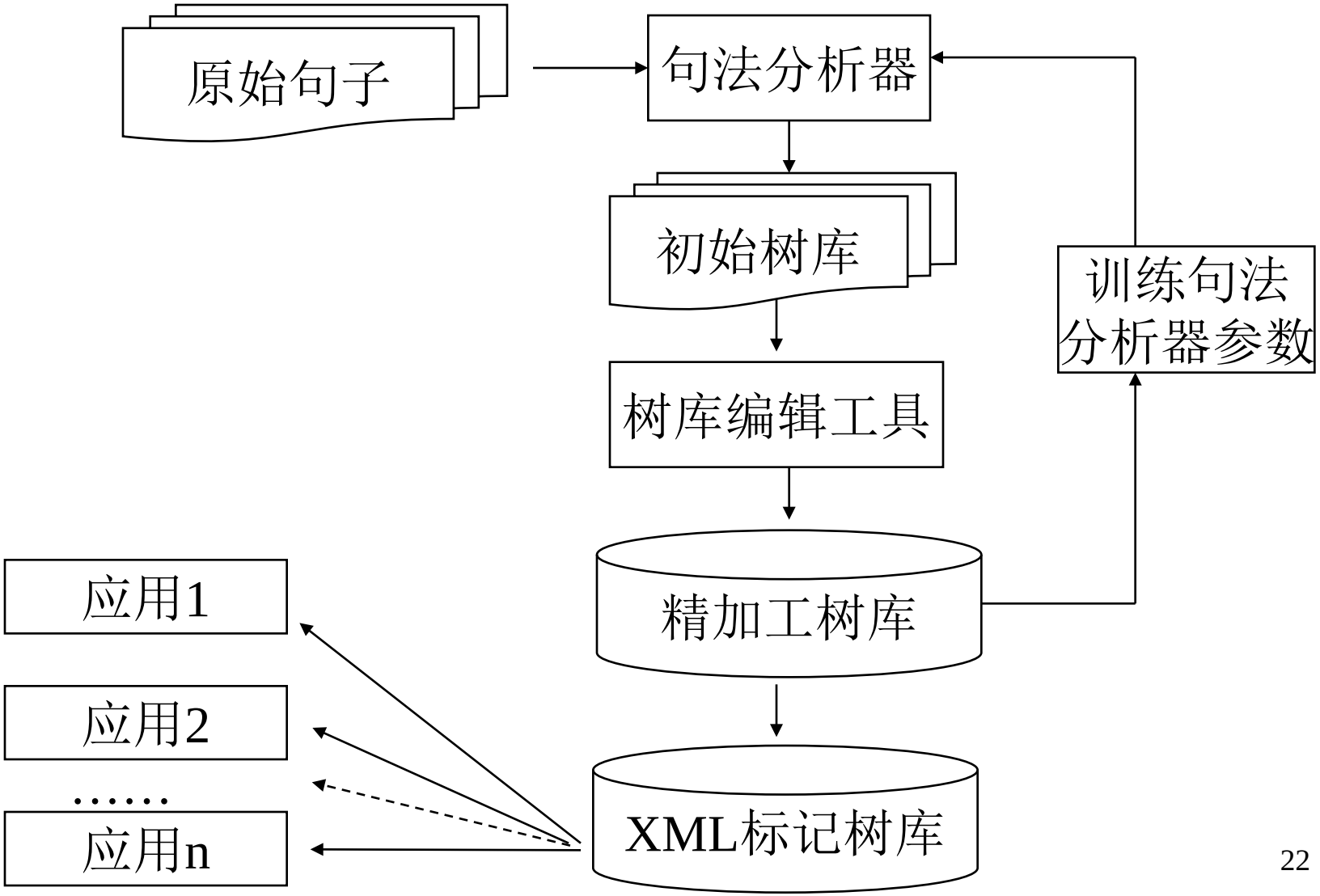
短语结构分析中的若干问题（续）

- “她之所以这么做是想取得更好的成绩”
- “无论谁不同意都可以保留他的观点。”
- “无论怎样，我都会赢。”



- “这不是什么大不了的成就。”
- “再有那么五六元钱就好了。”
- “戏法人人会变，各有巧妙不同。”
- “习惯成自然”

五 一个正在建设中的树库



树库示例

zj[dj[tp[m[10],q[年]],vp[vp[v[是]],np[mp[m[一],q[段]],np[ap[dp[d[很]],ap[a[长]]],u[的],np[n[时间]]]]]],w[。]]

zj[fj[tp[t[1991年]],w[，],fj[dj[np[ap[vp[vp[v[感染]],np[n[霍乱]]],u[的]],np[n[病人]]],vp[vp[d[多],v[达]],np[mp[m[3026]],np[n[人]]]]],w[，],dj[np[np[r[其中]],np[mcp[m[440]],np[n[人]]]],vp[v[死亡]]]]],w[。]]

zj[dj[e[啊],w[，],dj[r[这],vp[vp[v[是]],np[ap[dp[d[多么]],a[美妙]],u[的],np[n[前景]]]]]],w[！]]

zj[fj[vp[l[爱国一家]],w[，],dj[vp[v[爱],n[国]],vp[dp[d[不]],vp[v[分],t[先后]]]]],w[。]]

zj[dj[np[ap[vp[v[安葬],r[他]],u[的]],np[n[地方]]],ap[dp[d[很]],ap[a[美]]]],w[。]]

zj[dj[np[ap[vp[vp[v[安装]],np[n[灯]]],u[的]],np[n[人]]],vp[vp[v[是]],np[ap[r[我],u[的]],np[n[同学]]]]],w[。]]

zj[dj[np[n[安]],dj[np[r[自己]],vp[vp[v[打开],u[了]],np[n[门]]]]],w[。]]

zj[dj[dp[d[八成]],dj[dj[np[r[他]],vp[dp[d[不]],vp[v[来]]]],y[了]]],w[。]]

……

树库的应用

- 面向语言研究和教学
  抽取规则（基于实际语料发现结构模式）
  查询带结构的语法形式
  ……

- 面向自然语言处理
  自动句法分析（常宝宝，2003）
  机器翻译
  人机对话
  高级信息检索
  ……

句法分析实验结果

open test		close test	
Number of sentence	= 245	Number of sentence	= 119
Number of Error sentence	= 0	Number of Error sentence	= 0
Number of Skip sentence	= 0	Number of Skip sentence	= 0
Number of Valid sentence	= 245	Number of Valid sentence	= 119
Bracketing Recall	= 71.67	Bracketing Recall	= 90.84
Bracketing Precision	= 75.24	Bracketing Precision	= 95.18
Complete match	= 26.53	Complete match	= 45.38
Average crossing	= 2.30	Average crossing	= 0.36
No crossing	= 47.76	No crossing	= 84.87
2 or less crossing	= 66.12	2 or less crossing	= 95.80

常宝宝（2003）在宾州树库上的试验

树库工具演示

- 树库编辑、校对工具
- 树库规则抽取工具
- 树库结构模式查询工具

感谢北大计算语言所常宝宝博士和硕士研究生吴拥华提供的软件支持和帮助

参考文献

[1] 邵敬敏,1985《汉语句型研究述评》，载《语文导报》1985年第4期。

[2] 赵淑华,1991《谈80年代和90年代的汉语句型研究》，载《语言教学与研究》1991年第4期。

[3] 北京语言学院句型研究小组,1989 《现代汉语基本句型》，载《世界汉语教学》1989年第1期- 1991年第1期（分5次刊出）

[4] 范继淹,1986《范继淹语言学论文集》，语文出版社。

[5] 李临定,1986《现代汉语句型》，商务印书馆1986年版。

[6] 吴蔚天、罗建林,1994《汉语计算语言学》，电子工业出版社1994年版。

[7] 罗振声、郑碧霞,1994《汉语句型自动分析与分布统计算法与策略的研究》，载《中文信息学报》1994年第2期。

[8] 鲁川,2001《汉语语法的意合网络》，商务印书馆2001年版。

[9] 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2000年版。

[10] 周强、詹卫东、任海波,2001《构建大规模的汉语语块库》，载黄昌宁、张普主编（2001）《自然语言理解与机器翻译》，清华大学出版社2001年版，pp102-107。

[11] 周强、俞士汶，1996，《汉语短语标注标记集的确定》，《中文信息学报》1996年第4期。

[12]周强，1997，《汉语短语的自动划分和标注》，《中文信息学报》1997年第1期。

[13]周强，1997，《汉语树库的构建》，《中文信息学报》1997年第4期。

[14] Mitchell P. Marcus, et al, 1993, Building a large annotated corpus of English: the Penn Treebank, Computational Linguistics, Vol.19.

[15] http://www.cis.upenn.edu/~treebank/，（美国宾州大学树库网址）

[16] Xue,Nianwen & Xia, Fei, 2000, The Bracketing Guidelinges for the Penn Chinese Treebank (3.0)

谢谢大家

请多多批评指正