计算机时代的汉语研究
——从“面向计算”到“基于计算”

詹卫东

北京大学中文系 北京大学计算语言学研究所

zwd@pku.edu.cn

2006-08-20

提纲

- 从计算的角度看汉语研究
- 面向计算的汉语研究
- 基于计算的汉语研究
- 结语

一 从计算的角度看汉语研究

- 研究自然语言的目的

1 “正确地”说话 —— 知其然

不但我没见过槟榔，他**也**没见过槟榔 √

我**不但**没见过槟榔，他**也**没见过槟榔 ×

2 解释人何以能“正确地”说话 —— 知其所以然

人的语言能力——“正确地”说话

- 知道一个语言形式的对错

很多 学生 → 很多 长头发 学生

长头发 学生 → * 长头发 很多 学生

- 知道两个语言形式之间的关联

这件事 容易 办 → 办 这件事 容易

这件事 好 办 ⊗→ 办 这件事 好

- 知道“无形”的语言形式与有形语言形式间的关联

胡戈$_i$ 拍了 一部 新片$_j$， 大家 都 爱 看 [∅$_i$, ∅$_j$] ⊗

解释人的语言能力

- 规则制导（所谓的理性主义路线）

  实例 —> 范畴 —> 规则(模式)

- 经验制导（所谓的经验主义路线）

  实例 —> 频次 —> 参数(概率)

获取关于自然语言的知识

实例 → 范畴

a	b	c
三	本	书
一	群	人
五	个	苹果
两	斤	鱼
…	…	…

x	y	z
洗	不	干净
衣服	不	干净
很	不	干净
干净	不	干净
…	…	…

范畴 → 规则

三	本	书
一	群	人
五	个	苹果
两	斤	鱼
…	…	…
m	q	n
数词	量词	名词

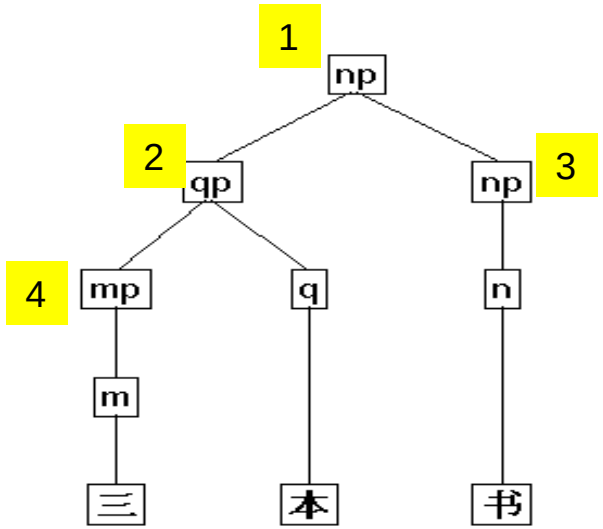
1 np → qp np

2 qp → mp q

3 np → n

4 mp → m

规则



结构模式

计算机模拟人的（部分）语言能力

- 让计算机掌握关于自然语言的知识

形式化的知识库

大规模的知识库

→ 系统地刻画自然语言形式与意义之间的关系

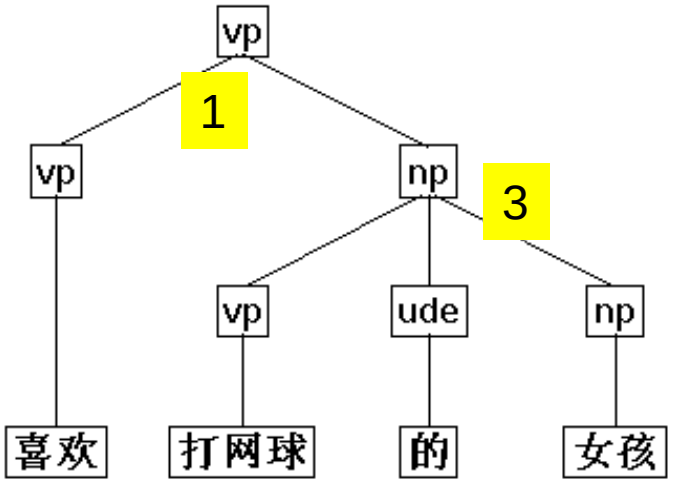
1 判断一个语言形式S的对错

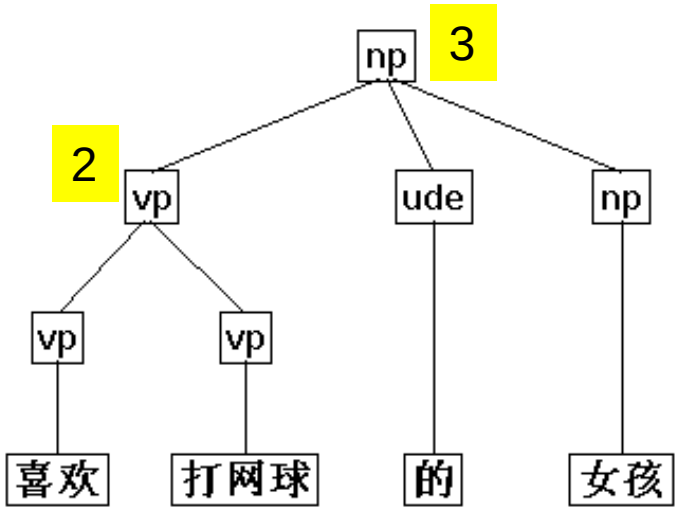
2 给出一个语言形式S的变换形式（S的意义）

当范畴和规则从少到多……

序号	规则	实例
1	vp → vp np	喜欢 网球； 喜欢 女孩
2	vp → vp vp	喜欢 打网球
3	np → vp ude np	打网球 的 女孩

喜欢 打网球 的 女孩





计算机需要处理歧义的语言知识

- 网球　　我买了一个网球　　网球他比我强
- 打网球　　老张在打网球　　单位里打网球老张总拿第一

- vp　vp　ude　np
  喜欢　打网球　的　女孩

  张三　特别　喜欢　打网球　的　女孩

  喜欢　打网球　的　女孩　一般 都 不 胖

vp	vp	ude	np
打	喜欢网球	的	女孩

计算机会如何分析？

语言知识在NLP中的应用

信息检索

- 季羡林老师　　季羡林的老师
- 李时珍医生　　李时珍的医生

- 北大教师　　北大的教师
- 红色围巾　　红色的围巾

二 面向计算的汉语研究

2.1 歧义格式类型分析

2.2 歧义格式消歧研究

2.1 歧义格式类型分析

（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]

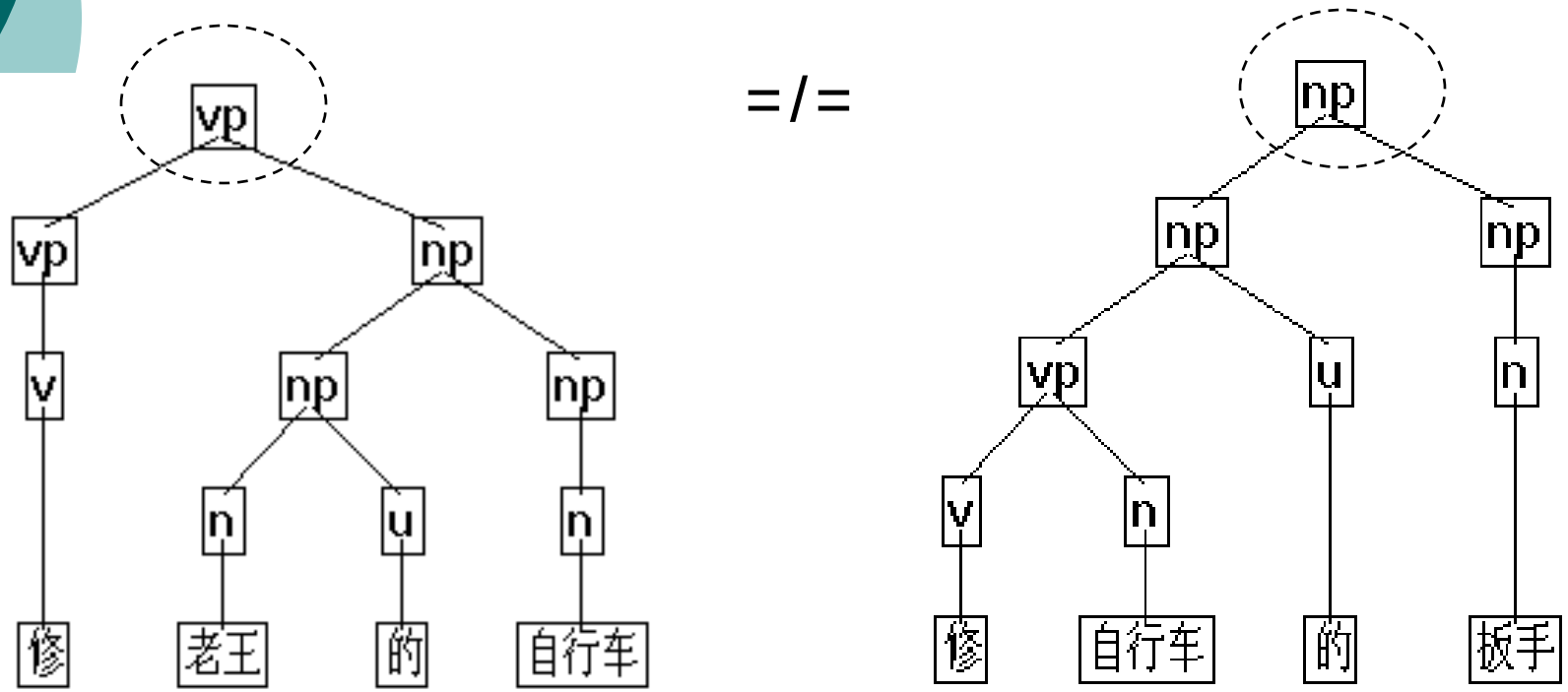
1b [[修 自行车 的] 扳手]

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

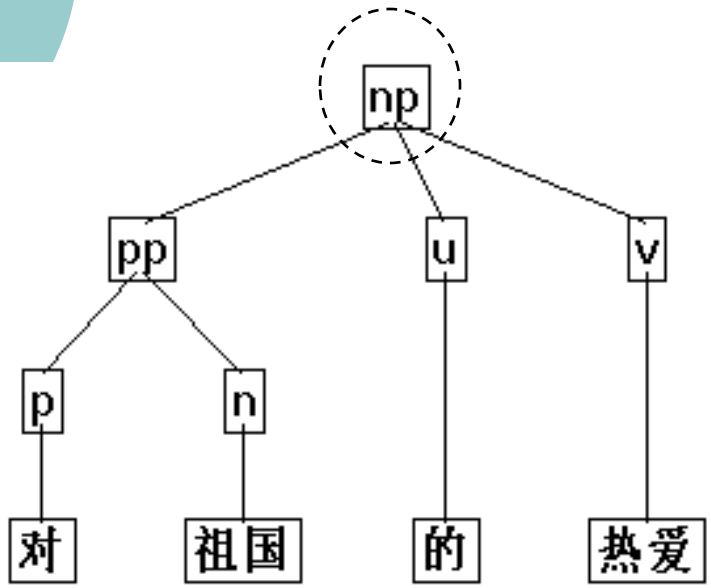
擦洗　　干净　的　桌子

……

vp　　np

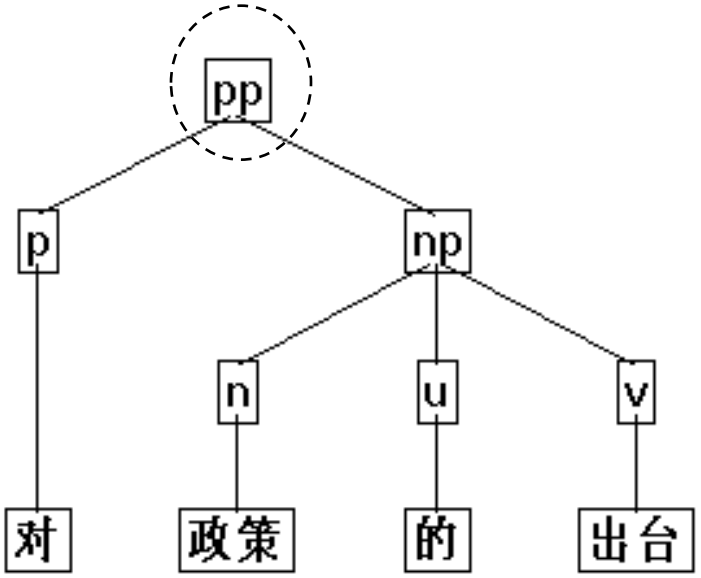
外显型歧义（续2）

对祖国的热爱



=/=

对政策的出台



他对校长的意见很大

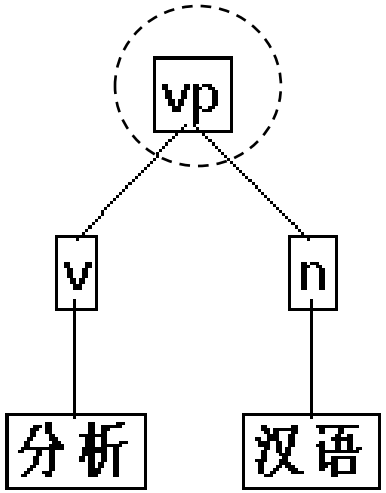
他对校长的批评很尖锐

他对校长的意见持否定态度

他对校长的批评充耳不闻

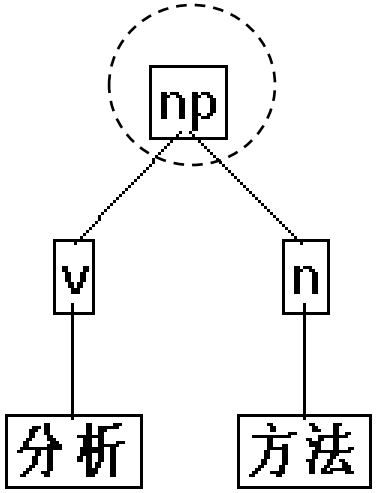
外显型歧义（续3）

分析汉语



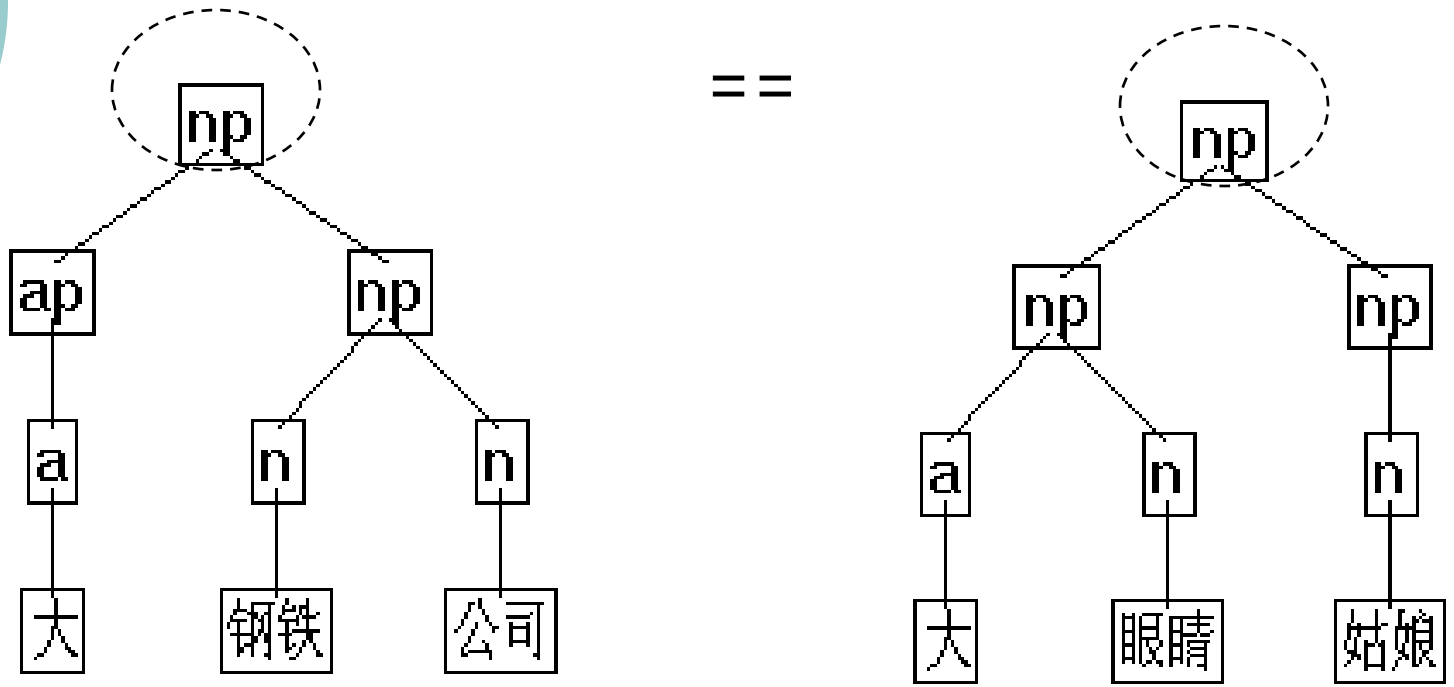
=/=

分析方法



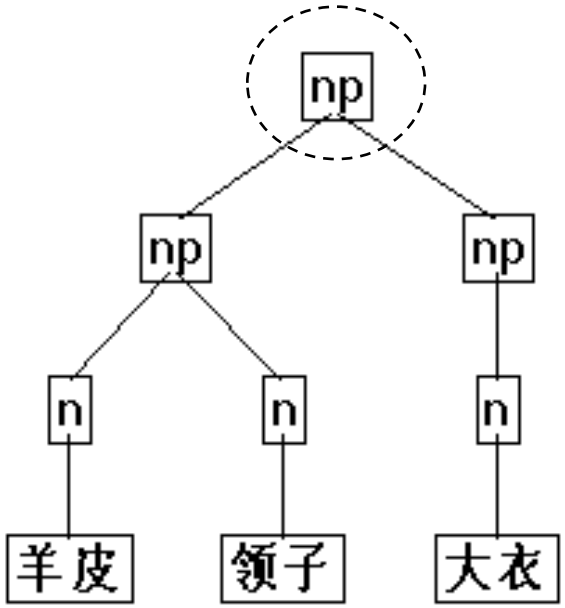
出租汽车　np or vp

内含型歧义



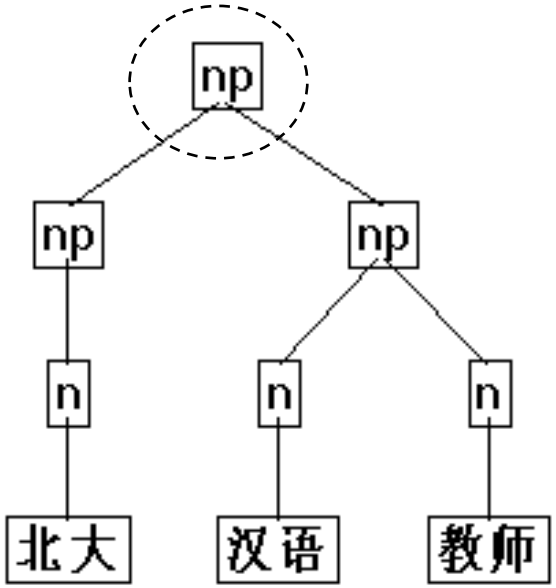
内含型歧义（续1）

羊皮领子大衣



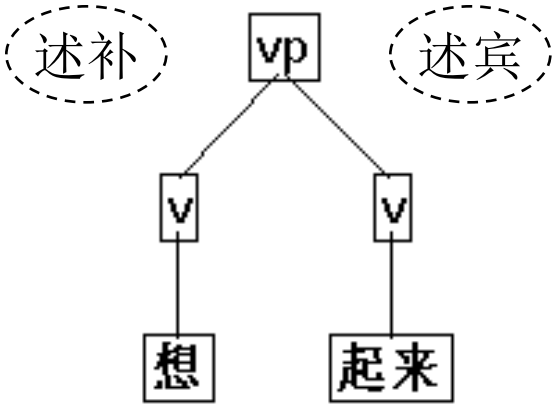
==

北大汉语教师



内含型歧义（续2）

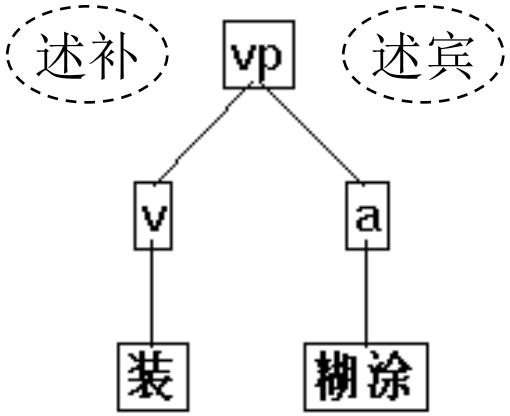
想起来



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

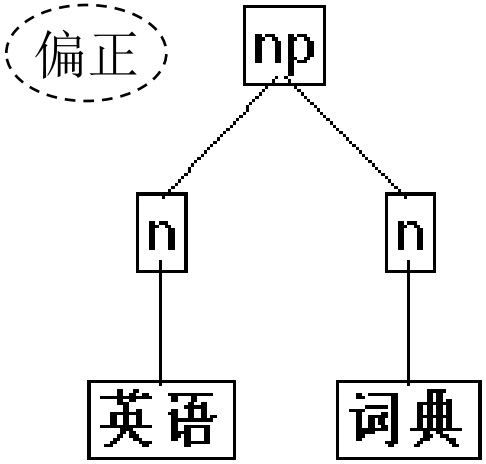


他就会装糊涂，其实他心理比谁都清楚

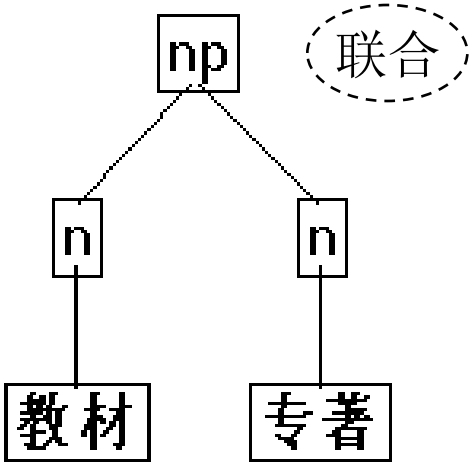
装了一上午机器，我都装糊涂了

内含型歧义（续3）

英语词典



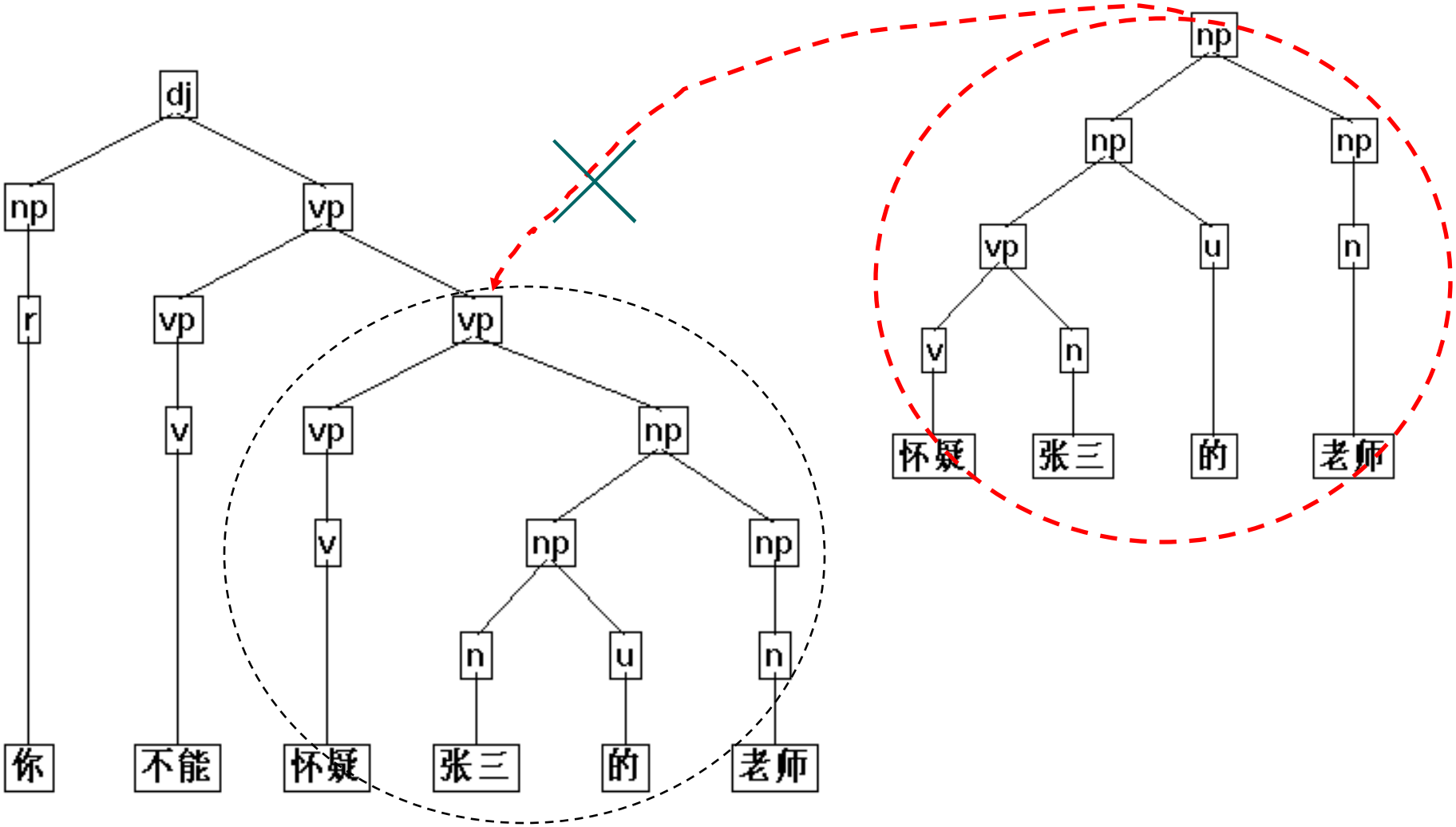
教材专著



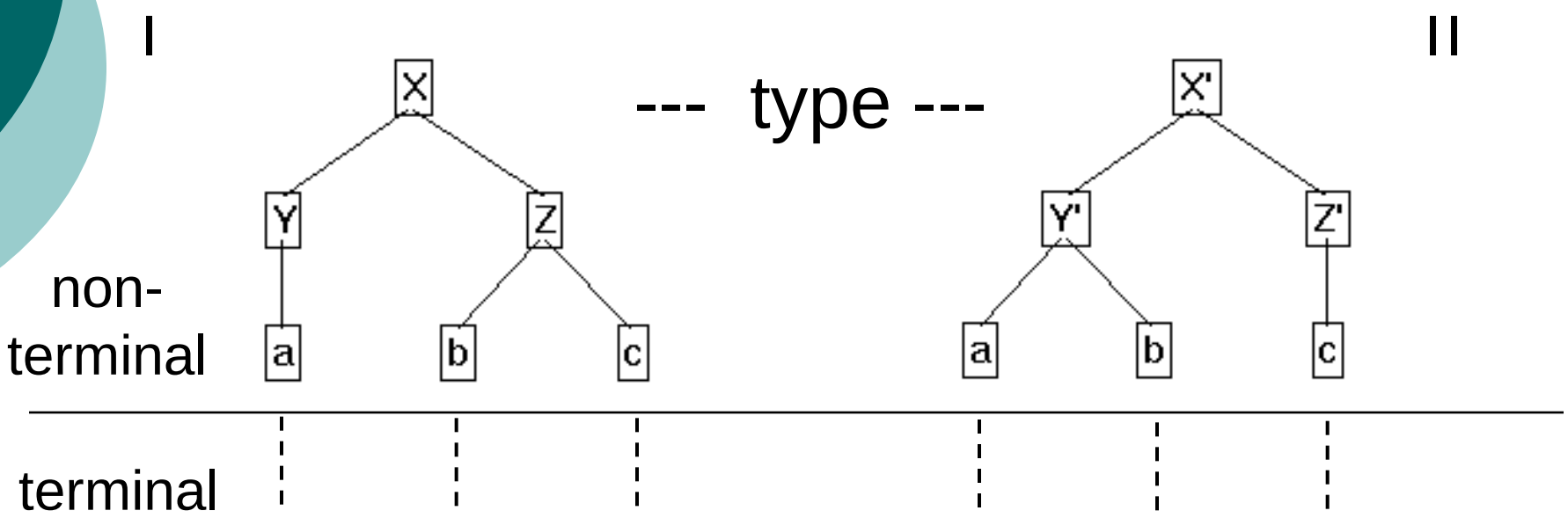
偏正？　牛奶饼干　联合？

区分“外显”与“内含”的作用

ex. 你不能怀疑张三的老师

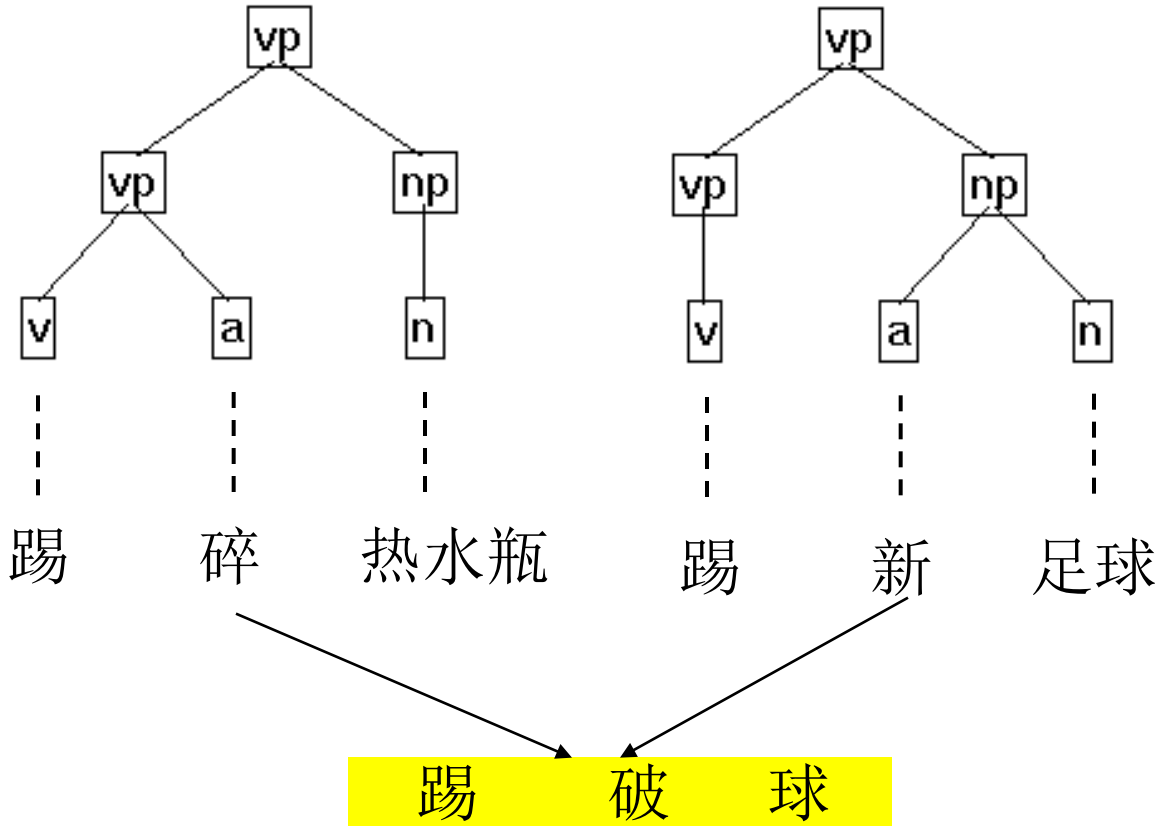


（二） 真歧义 准歧义 伪歧义

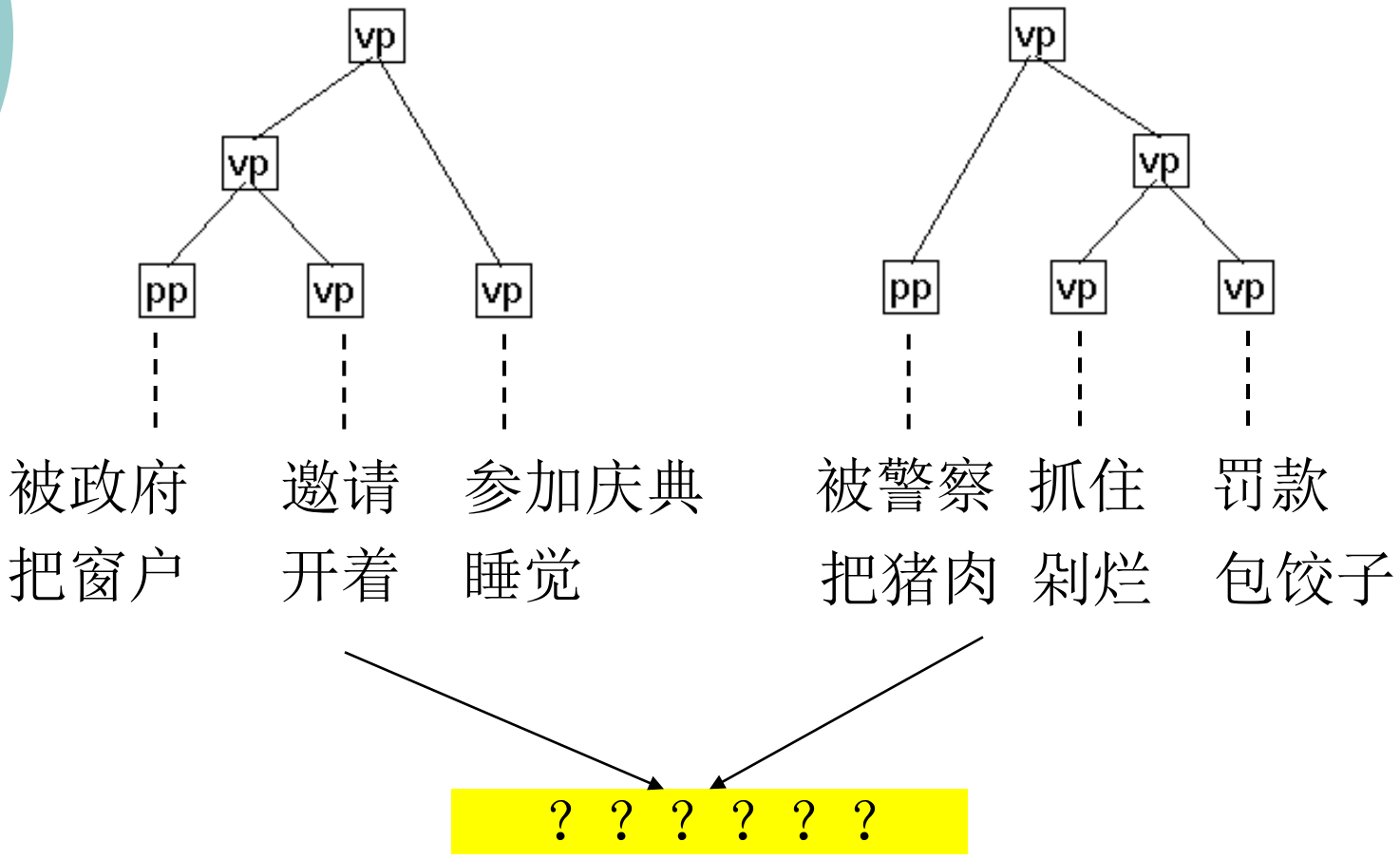


歧义? --- token --- 歧义?

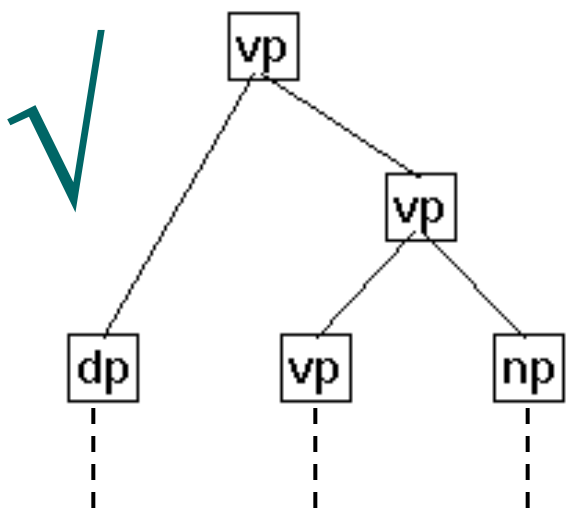
真歧义

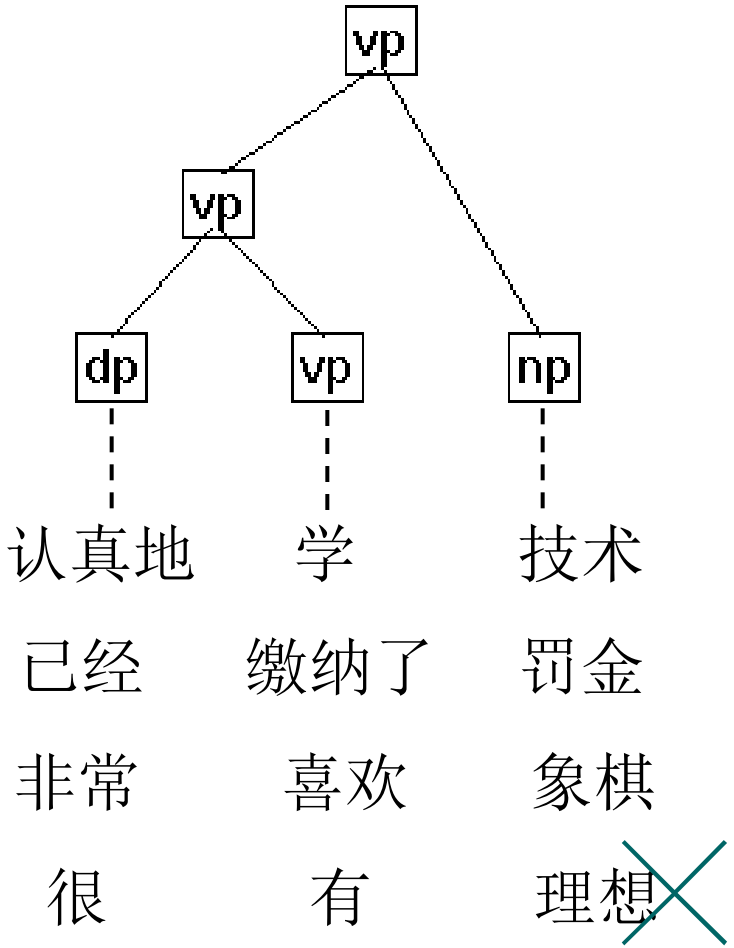


准歧义


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
学
技术
已经
缴纳了
罚金
非常
喜欢
象棋
很
有
理想


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

小结

- 计算机“眼”里的歧义远远多于人眼里的歧义
- 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
- 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。

2.2 歧义格式消歧研究

- 外显型准歧义举例

p np vp 的 np

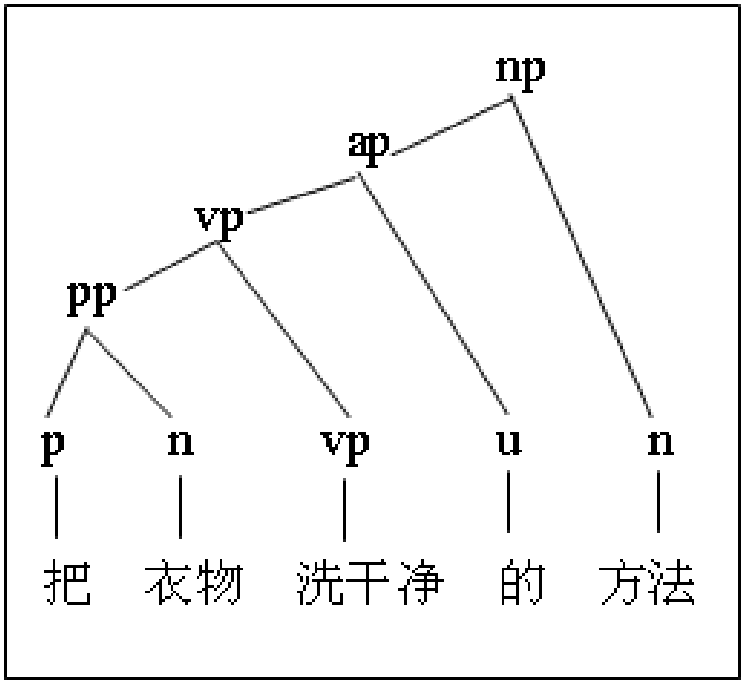
- 内含型真歧义举例

v a n

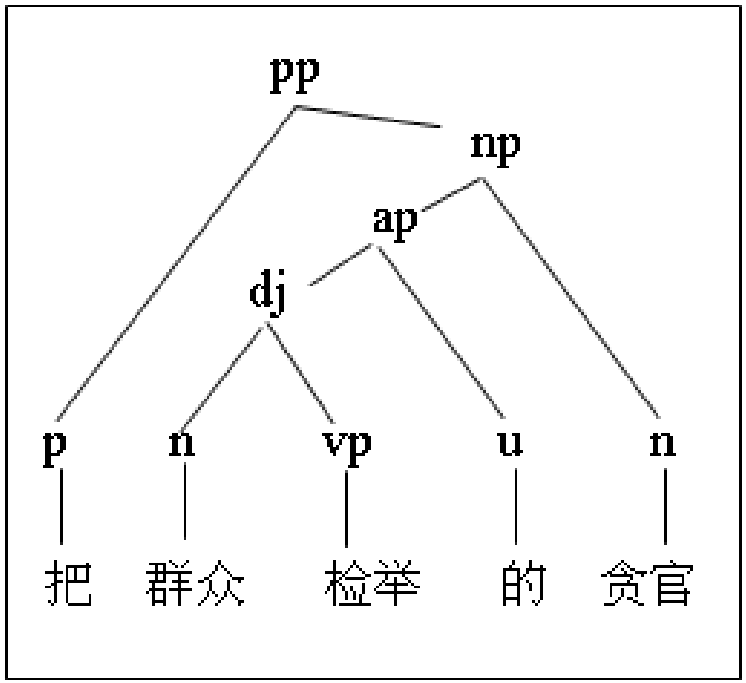
p np vp 的 np

A. 把衣物洗干净的方法

B. 把群众检举的贪官



(a)



(b)

“v a n”格式的歧义消解分析

- 例1 a. * 那张床张三买匀称了
  b. * 张三买贵了一张床

  } v a

- 例2 张三写错三个字

  } v a

- 例3 张三买好衣服 v a n

汉语动词与形容词的组配

- 哪些动词跟形容词可以组成述补结构？
- 如何描述VR结构作为一个整体的组配能力？

V A 组合能力考察：句法

动词V1	形容词A1	动词V2	形容词A2
借 浸 经营 精简 救 举 锯 聚 捐 卷 扛 考 搜刮 限制 笑 养 赢 ……	乏 肥 富 干 干净 高 鼓 乖 光 光滑 贵 好 黑 红 ……	拜见 报答 报 废 拚杀 变卦 驳回 查明 称 斗牛 告辞 公认 恭候 立足 遭到 ……	苍白 苍劲 诧异 孱弱 缠绵 长寿 长足 常见 常用 怅惘 畅达 超然 佳 险 ……

V1:有可能出现在VR结构中V位置的动词，3178/14479=21.9%
A1:有可能出现在VR结构中R位置的形容词，262/2856=9.2%

V2: 不能出现在VR结构中V位置的动词

A2: 不能出现在VR结构中R位置的形容词

数据来源: 北大计算语言所《现代汉语语法信息词典》1998版

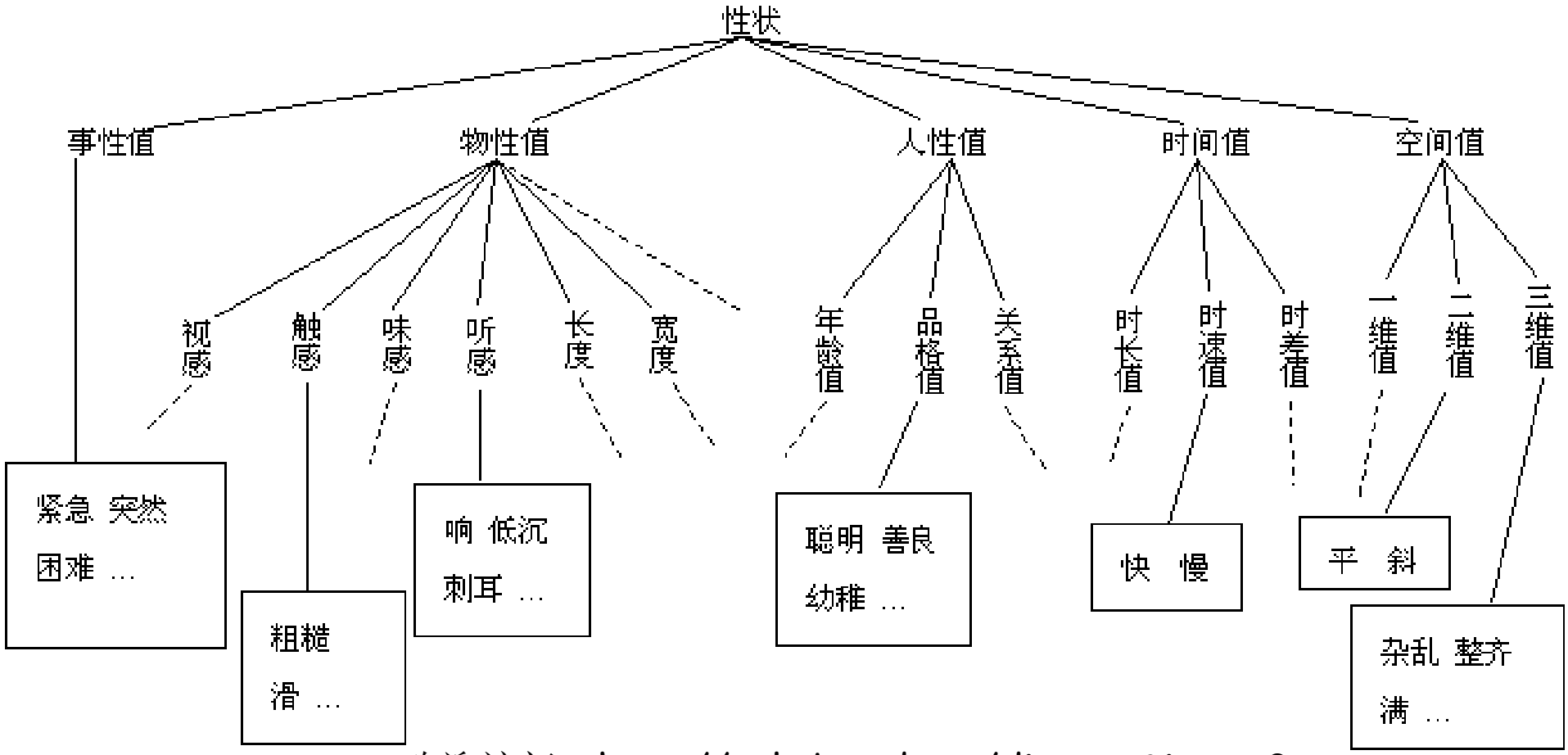
v a n 结构分析示例:句法约束

[v [a n]] 结构	[[v a] n] 结构
像/v 小/a 山/n	填/v 饱/a 肚子/n
走/v 和平/a 道路/n	找/v 准/a 突破口/n
持/v 不同/a 看法/n	结/v 满/a 伤疤/n
学/v 真/a 本事/n	擂/v 响/a 大鼓/n

动词带结果补语的能力 & 形容词作结果补语的能力

例子来源: <人民日报> 标注语料库

V A组合能力考察：语义



欢迎访问：http://ccl.pku.edu.cn/dict.asp?item=2

北大语义词典项目

动词的[论旨角色变化] 特征

$$\left[\begin{array}{l} \text{词语}:\text{洗} \\ \text{论元}:2 \\ \text{论旨角色}:\left[\begin{array}{l} \text{施事}:[\text{语义类}:\text{人}] \\ \text{受事}:[\text{语义类}:\text{衣物}] \end{array}\right] \\ \text{论旨角色变化}:\left[\begin{array}{l} \text{施事变化}:[\text{语义类}:\text{人性值}] \\ \text{受事变化}:[\text{语义类}:\text{触感}\,|\,\text{视感}\,|\,\text{色感}] \end{array}\right] \end{array}\right]$$

洗累了　洗干净了　洗白了　　*洗响了　*洗稳当了　*洗突然了

陆俭明. 1990.〈“VA了”述补结构的语义分析〉，载《汉语学习》1990年第1期。

马真、陆俭明. 1997.〈形容词作补语情况考察〉，载《汉语学习》1997年第1、4、6期。

动词的[论旨角色评价]特征

[词语：买
论元：2
论旨角色：[施事：[语义类：人]
受事：[语义类：商品]]
论旨角色变化：[施事变化：[语义类：人性值]]
论旨角色评价：[受事评价：[语义类：外形|价值]]]

买累了　买小了　买贵了　　*买深了　*买干净了　*买蓝了

v a n 结构分析示例:语义约束

[v [a n]]结构	[[v a] n]结构
穿/v 蓝/a 大褂/n	睁/v 大/a 眼睛/n
花/v 大/a 力气/n	调/v 低/a 利率/n
买/v 特别/a 国债/n	抹/v 干/a 眼泪/n
坐/v 早/a 船/n	摆/v 正/a 位置/n

v a 结合成一个整体后的组配能力？

v后a表示“论旨角色变化”（自然结果），则“v a”可带宾语

v后a表示“论旨角色评价”（非自然结果），则“v a”不能带宾语

v a构成述补式vp后组配性质的变化

* 你们盖晚了这座楼	——	你们盖晚了	（“结果”宾语被抑制）
* 你买早了房子	——	你买早了	（“受事”宾语被抑制）

他们$_i$搬 空$_j$了那间屋子$_j$	——	屋子搬空了	（“处所”角色）
他$_i$砍 钝$_j$了三把刀$_j$	——	刀砍钝了	（“工具”角色）

* 他$_i$砍 累$_i$了这把刀$_j$	——	他砍累了	（“工具”角色被抑制）
* 他$_i$学 聪明$_i$了这些办法$_j$	——	他学聪明了	（“受事”角色被抑制）
* 他$_i$睡 迷糊$_i$了这张床$_j$	——	他睡迷糊了	（“处所”角色被抑制）
* 他$_i$吃 饱$_i$了中餐$_j$	——	他吃饱了	（“方式”角色被抑制）

a 指向 v 的主体论元　还是　指向 v 的客体论元

→　[v a] 带宾语的能力

三 基于计算的汉语研究

3.1 歧义格式的分布统计

3.2 汉语句法规则知识库开发环境

3.3 大规模语料库建设

3.1 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

……

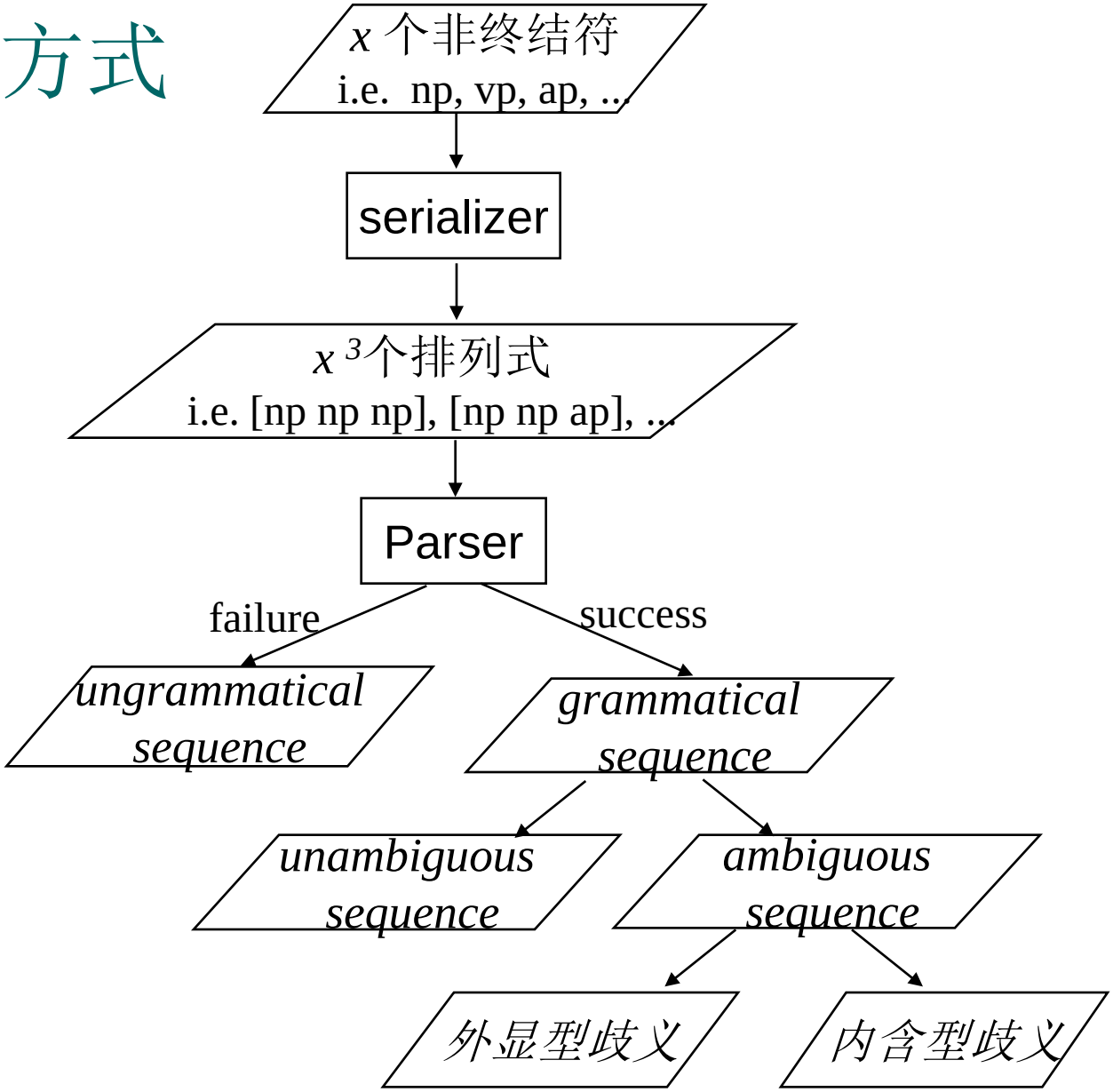
1　哪些格式有潜在歧义？

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

$9^3 = 729$

<table>
<tr><td colspan="3">可能形成合法结构的排列:369个</td><td rowspan="6">不可能形成合法结构的排列:360个

np np dp
np np pp
np mp sp
np mp dp
……

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

pp tp sp
pp tp dp
pp tp pp</td></tr>
<tr><td colspan="3">np np np
np np mp
np np tp
np np sp
……</td></tr>
<tr><td colspan="2">有歧义的排列式:285个</td><td>无歧义的排列式:84个</td></tr>
<tr><td>外显型歧义格式:194个</td><td>内含型歧义格式:91个</td><td rowspan="3">np mp np
np mp tp
np mp dj
np ap dj
……</td></tr>
<tr><td>np np np
np np ap
np np vp
np vp vp
……</td><td>np np mp
np np tp
np np sp
np np dj
……</td></tr>
</table>

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！

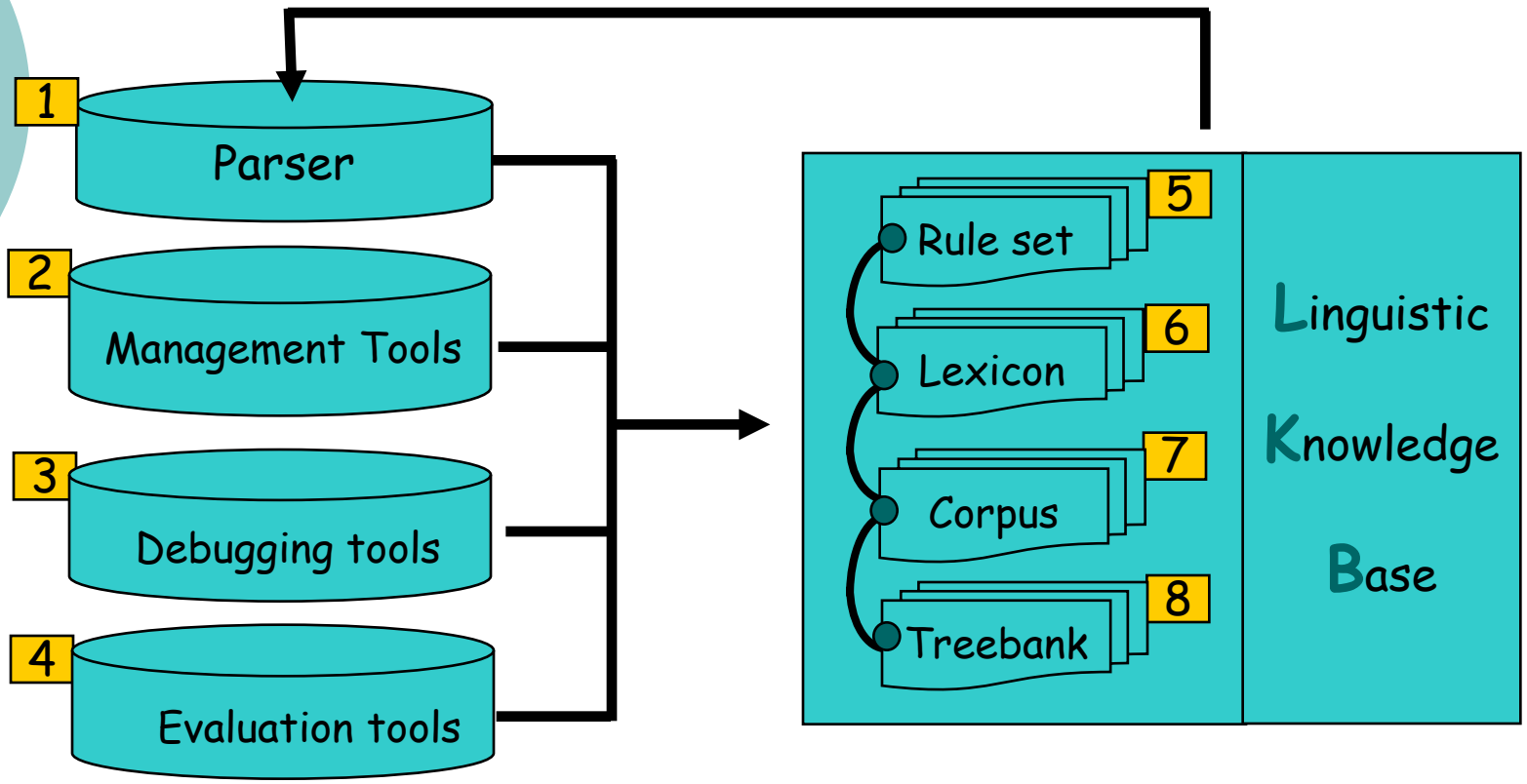
歧义格式统计研究的意义

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

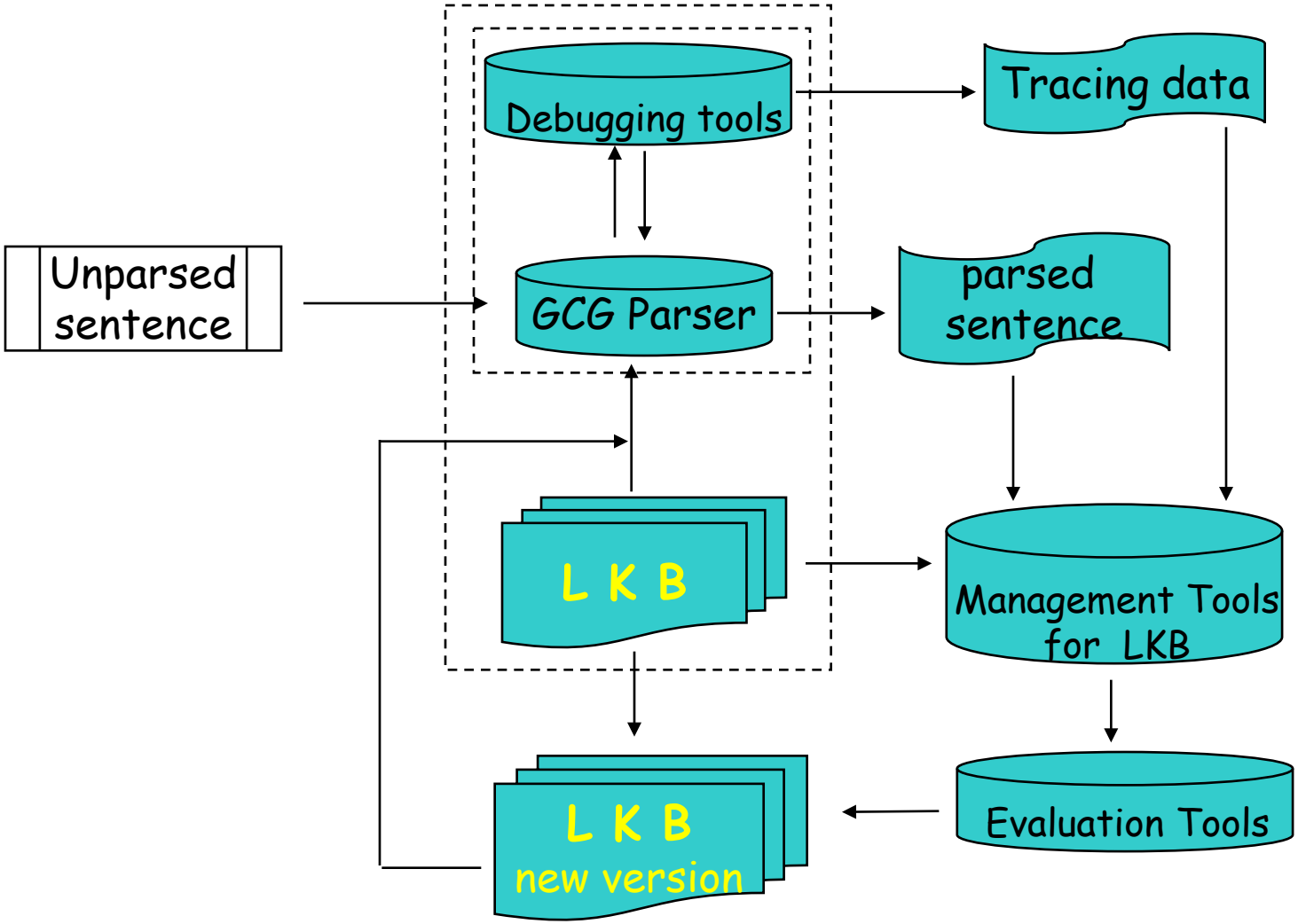
对“真歧义、准歧义”进行统计，需要树库数据作为基础

3.2 汉语句法规则知识库开发环境

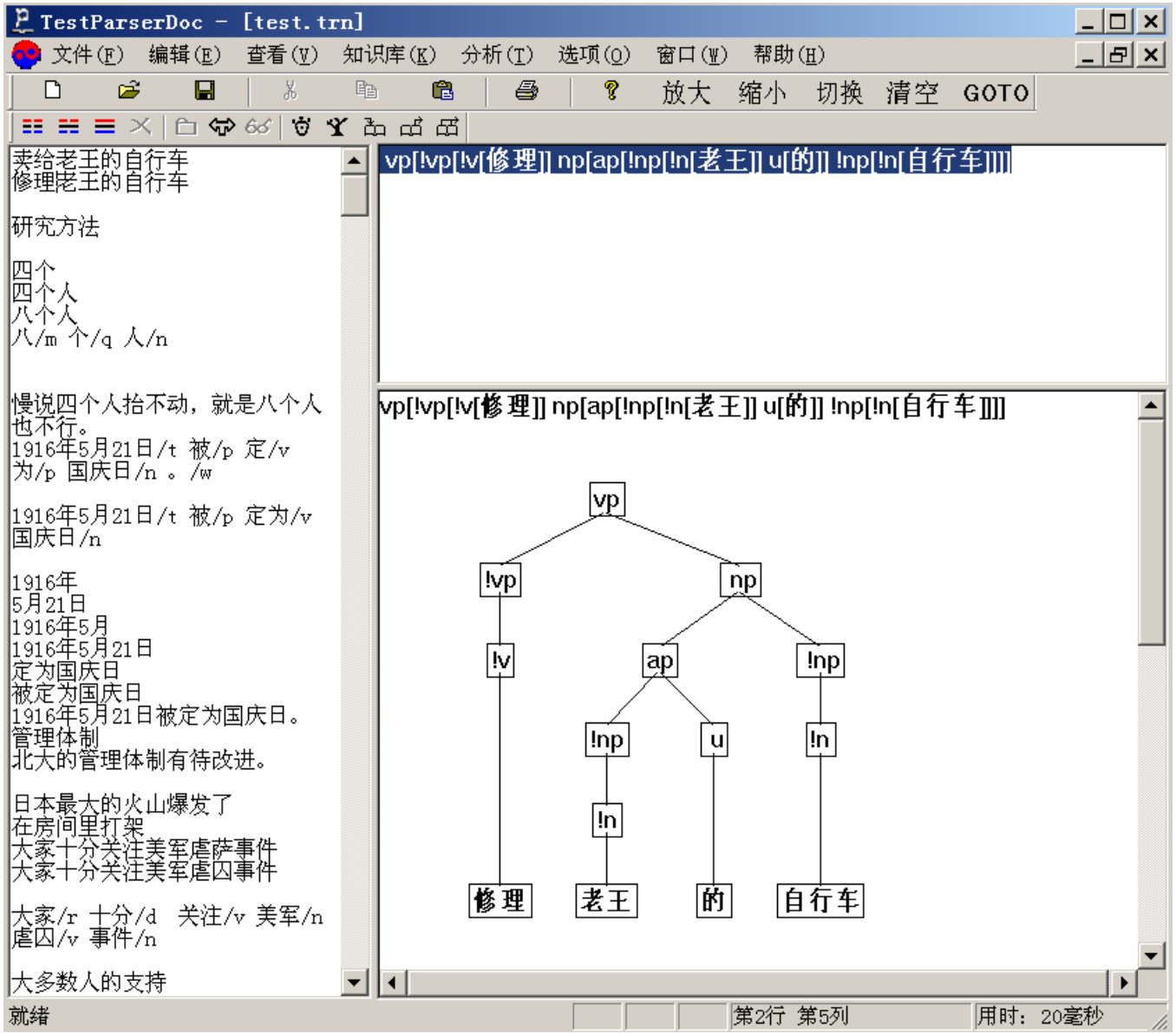


An Integrated Chinese Grammar Development Environment（ICGDE）

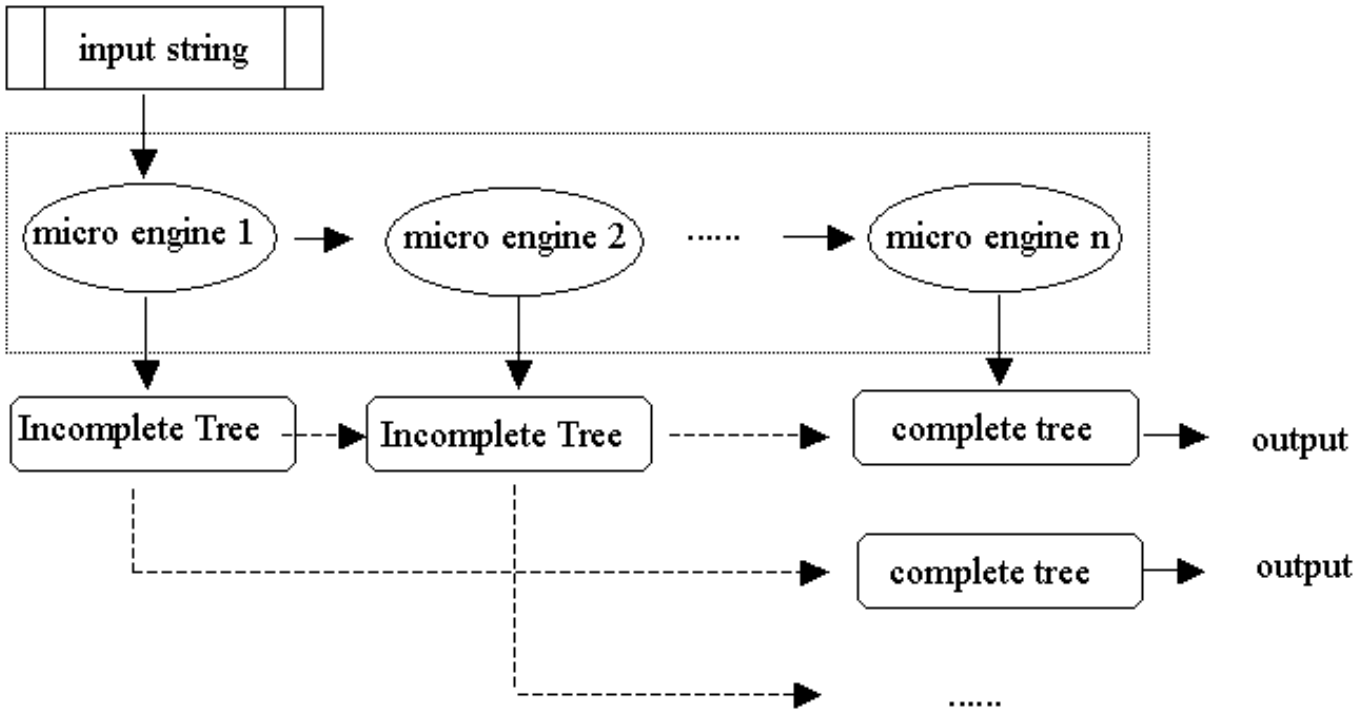
ICGDE的工作流程


Debugging tools
Tracing data
Unparsed sentence
GCG Parser
parsed sentence
L K B
Management Tools for LKB
L K B new version
Evaluation Tools

句法分析器界面

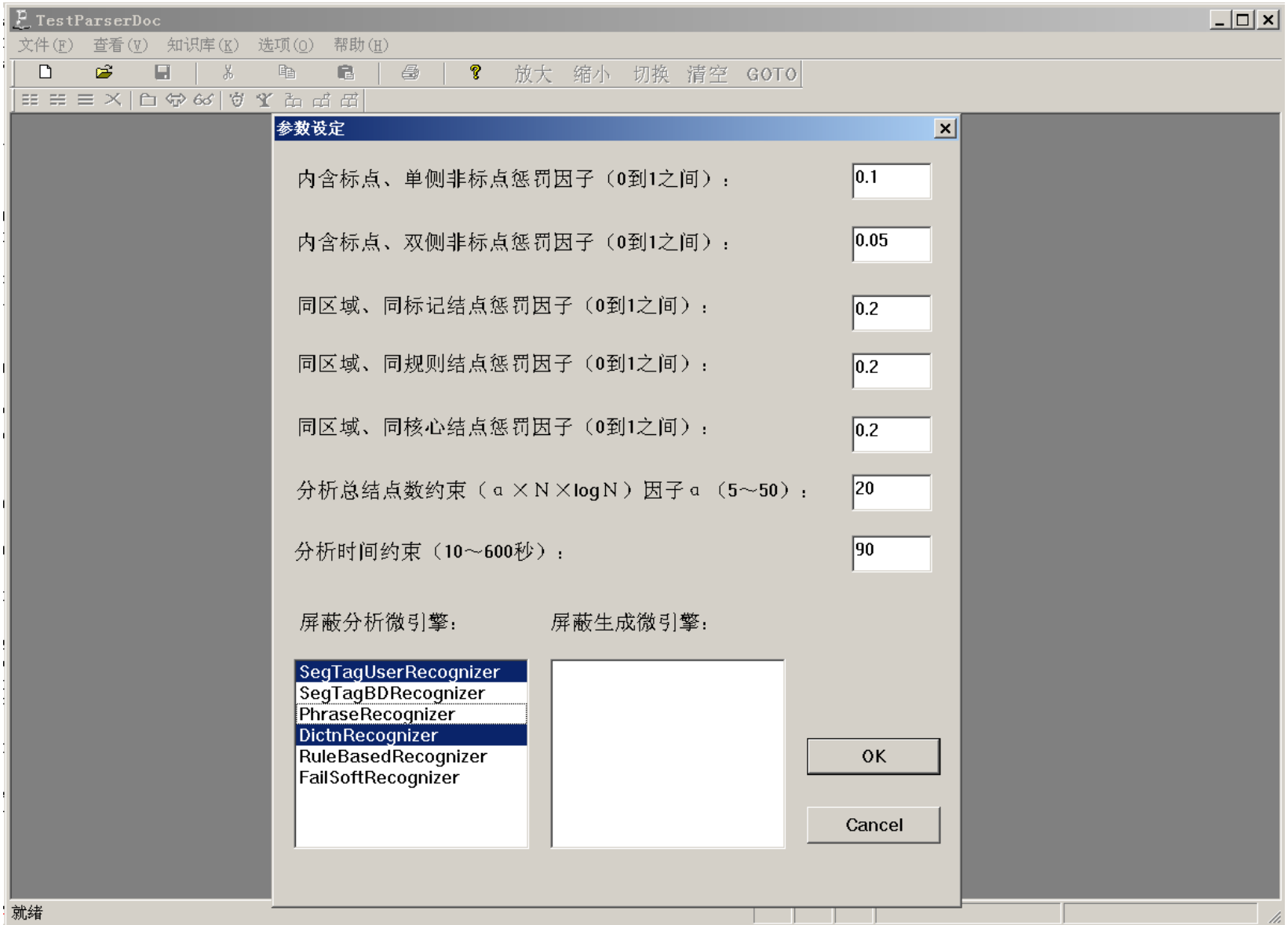

TestParserDoc - [test.trn]
文件(F) 编辑(E) 查看(V) 知识库(K) 分析(T) 选项(O) 窗口(W) 帮助(H)
放大 缩小 切换 清空 GOTO
卖给老王的自行车
修理老王的自行车
研究方法
四个
四个人
八个人
八/m 个/q 人/n
慢说四个人抬不动，就是八个人也不行。
1916年5月21日/t 被/p 定/v 为/p 国庆日/n 。/w
1916年5月21日/t 被/p 定为/v 国庆日/n
1916年
5月21日
1916年5月
1916年5月21日
定为国庆日
被定为国庆日
1916年5月21日被定为国庆日。
管理体制
北大的管理体制有待改进。
日本最大的火山爆发了
在房间里打架
大家十分关注美军虐萨事件
大家十分关注美军虐囚事件
大家/r 十分/d 关注/v 美军/n 虐囚/v 事件/n
大多数人的支持
vp[!vp[!v[修理]] np[ap[!np[!n[老王]] u[的]] !np[!n[自行车]]]]
vp[!vp[!v[修理]] np[ap[!np[!n[老王]] u[的]] !np[!n[自行车]]]]
vp
!vp
np
!v
ap
!np
!np
u
!n
!n
修理
老王
的
自行车
就绪
第2行 第5列
用时：20毫秒

多引擎结构



A pipelined multi-engine approach to Chinese parsing

参数配置

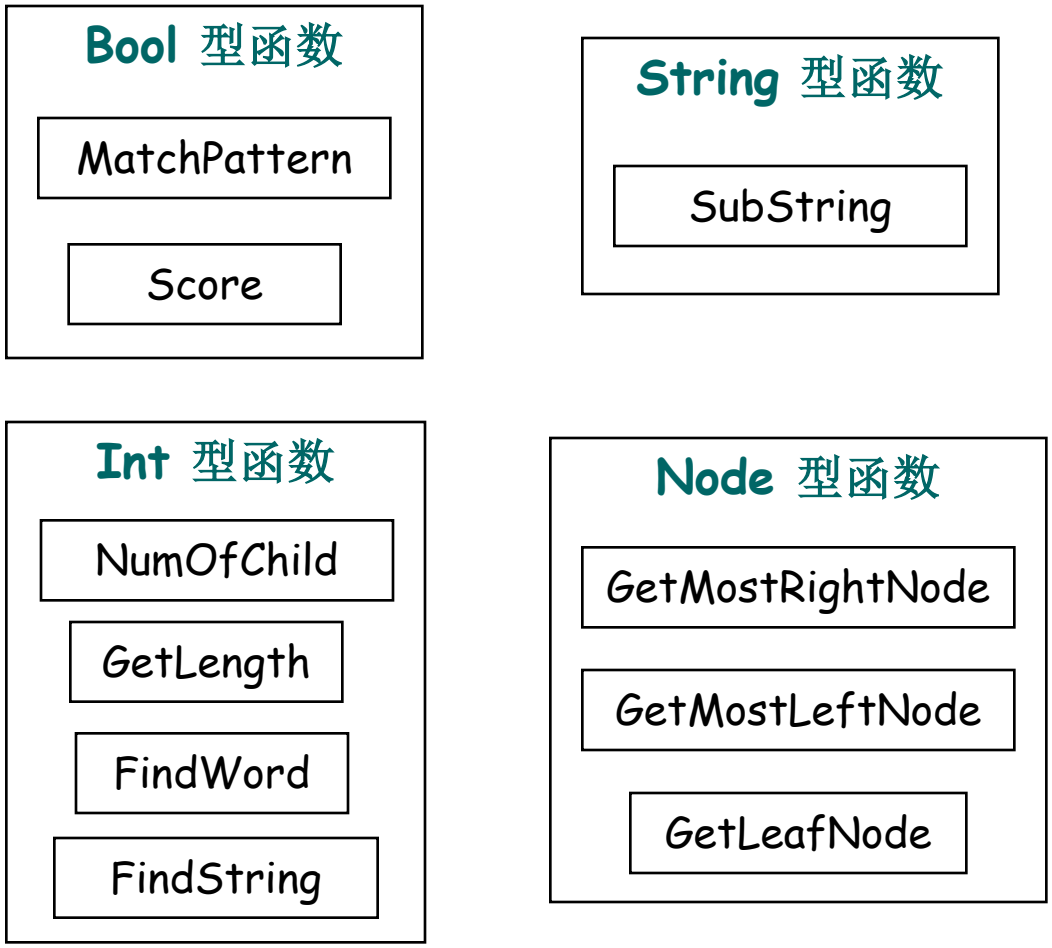

TestParserDoc
文件(F) 查看(V) 知识库(K) 选项(O) 帮助(H)
放大 缩小 切换 清空 GOTO
参数设定
内含标点、单侧非标点惩罚因子（0到1之间）：
0.1
内含标点、双侧非标点惩罚因子（0到1之间）：
0.05
同区域、同标记结点惩罚因子（0到1之间）：
0.2
同区域、同规则结点惩罚因子（0到1之间）：
0.2
同区域、同核心结点惩罚因子（0到1之间）：
0.2
分析总结点数约束（α×N×logN）因子α（5～50）：
20
分析时间约束（10～600秒）：
90
屏蔽分析微引擎：
屏蔽生成微引擎：
SegTagUserRecognizer
SegTagBDRecognizer
PhraseRecognizer
DictnRecognizer
RuleBasedRecognizer
FailSoftRecognizer
OK
Cancel
就绪

规则的表达能力

- CFG规则
- Unification
- 条件语句
- 函数与操作符

分级约束语法

规则中的函数和操作符

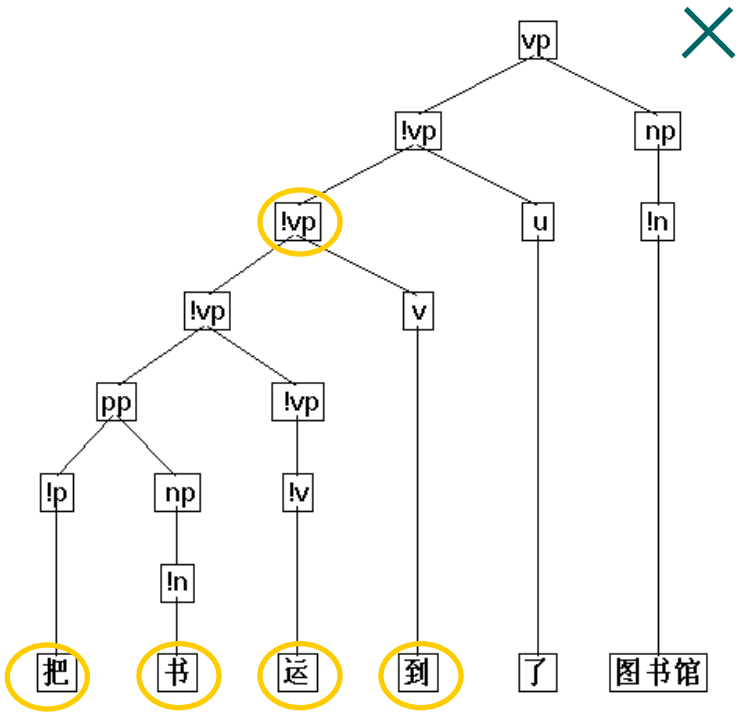


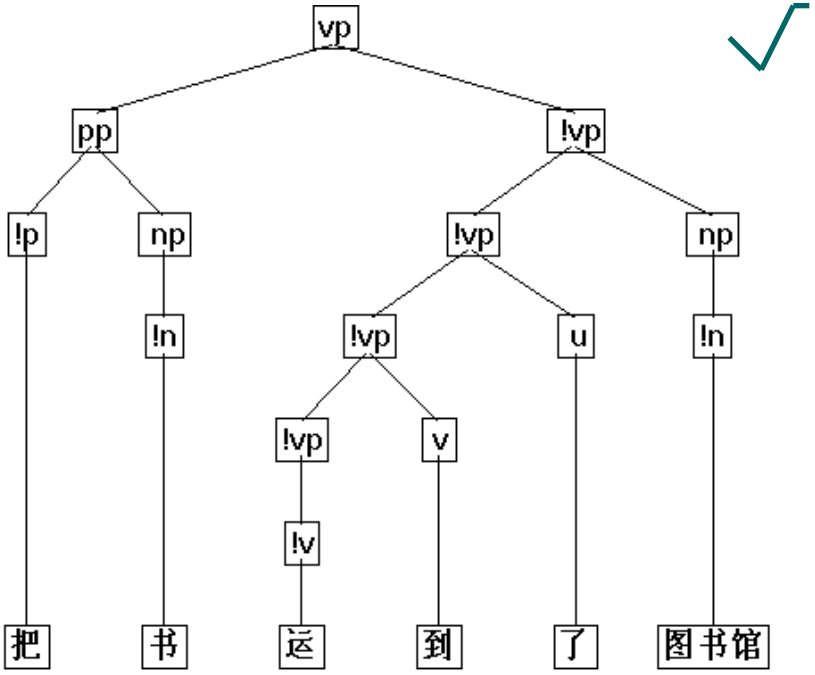
关系操作符： > < >= <= == !=

内置函数应用示例

ex.1 GetLength()

{vp1} vp -> !vp u :: ..., IF #GetLength(%vp) >=4 FALSE, ...

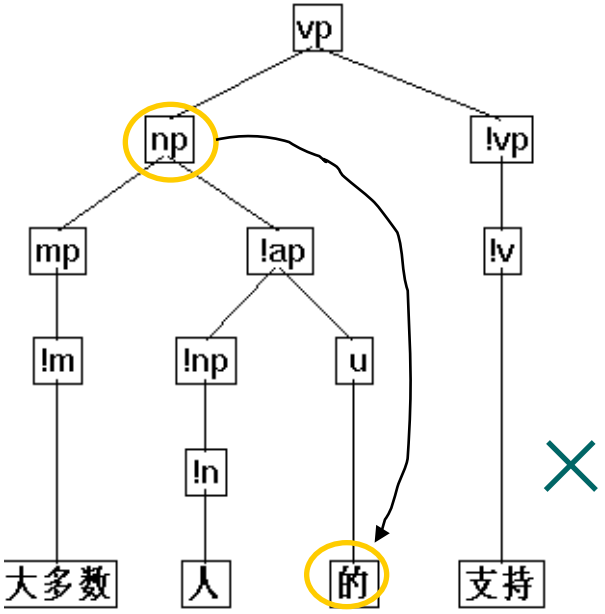


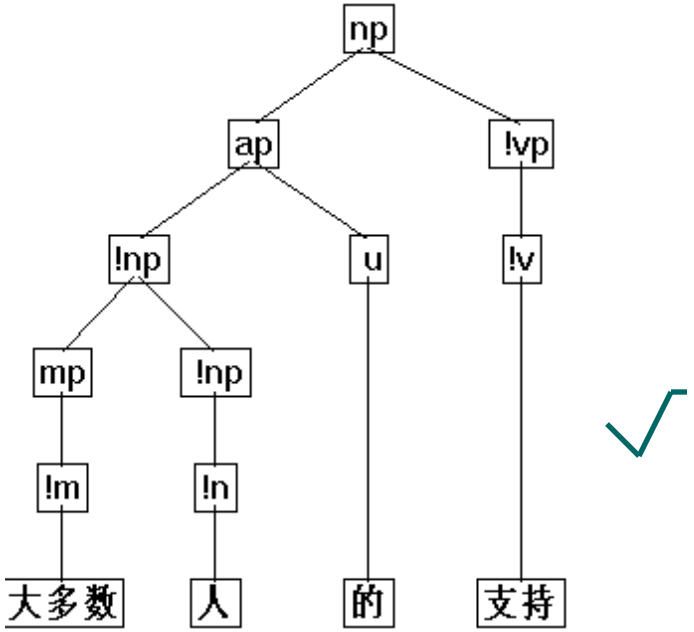


内置函数应用示例（续）

ex.2 GetLeafNode()

{vp2} vp -> np !vp :: ..., IF %GetLeafNode(%np,-1).原形=的 FALSE, ...





节点评分机制

- 评分的作用：知识颗粒度的细化

1) 不同知识源，可靠程度不同，设置不同的评分；
2) 相同知识源，可靠程度不同，设置不同的评分；
3) 相同知识源，主观感觉可靠程度不同，设置不同的评分；

ex.1 系统词典比较可靠，补充词典不可靠

ex.2 含终结符的规则（局部规则）比不含终结符的规则（全局规则）可靠

ex.3 一些不常见的组合模式（低频规则），感觉不可靠

节点评分示例

局部规则

&& {vpdao1} vp -> pp(!p<把> np) !vp(!vp vp(!vp<<到>> np))

```
vp -2.10694{vpdao1}
====pp -0.694124{pp3}
    ====p<把> 0{}
    ====np -0.00097704{np00}
        ====n<书> 0{}
====vp -1.40497{vpyundao}
    ====vp -0.00097704{vp00}
        ====v<运> 0{}
    ====vp -0.71085{vpsb1}
        ====vp -0.0167254{vp1}
            ====vp -0.00097704{vp00}
                ====v<到> 0{}
            ====u<了> 0{}
        ====np -0.00097704{np00}
            ====n<图书馆> 0{}
```

全局规则

&& {vpzz1} vp->pp !vp

```
vp -2.79127{vpzz1}
====pp -0.694124{pp3}
    ====p<把> 0{}
    ====np -0.00097704{np00}
        ====n<书> 0{}
====vp -1.404{vpsb1}
    ====vp -0.709873{vp1}
        ====vp -0.694124{vpsbu1}
            ====vp -0.00097704{vp00}
                ====v<运> 0{}
            ====v<到> 0{}
        ====u<了> 0{}
    ====np -0.00097704{np00}
        ====n<图书馆> 0{}
```

节点评分示例（续）

- 把衣物洗干净的方法
- 把群众检举的贪官

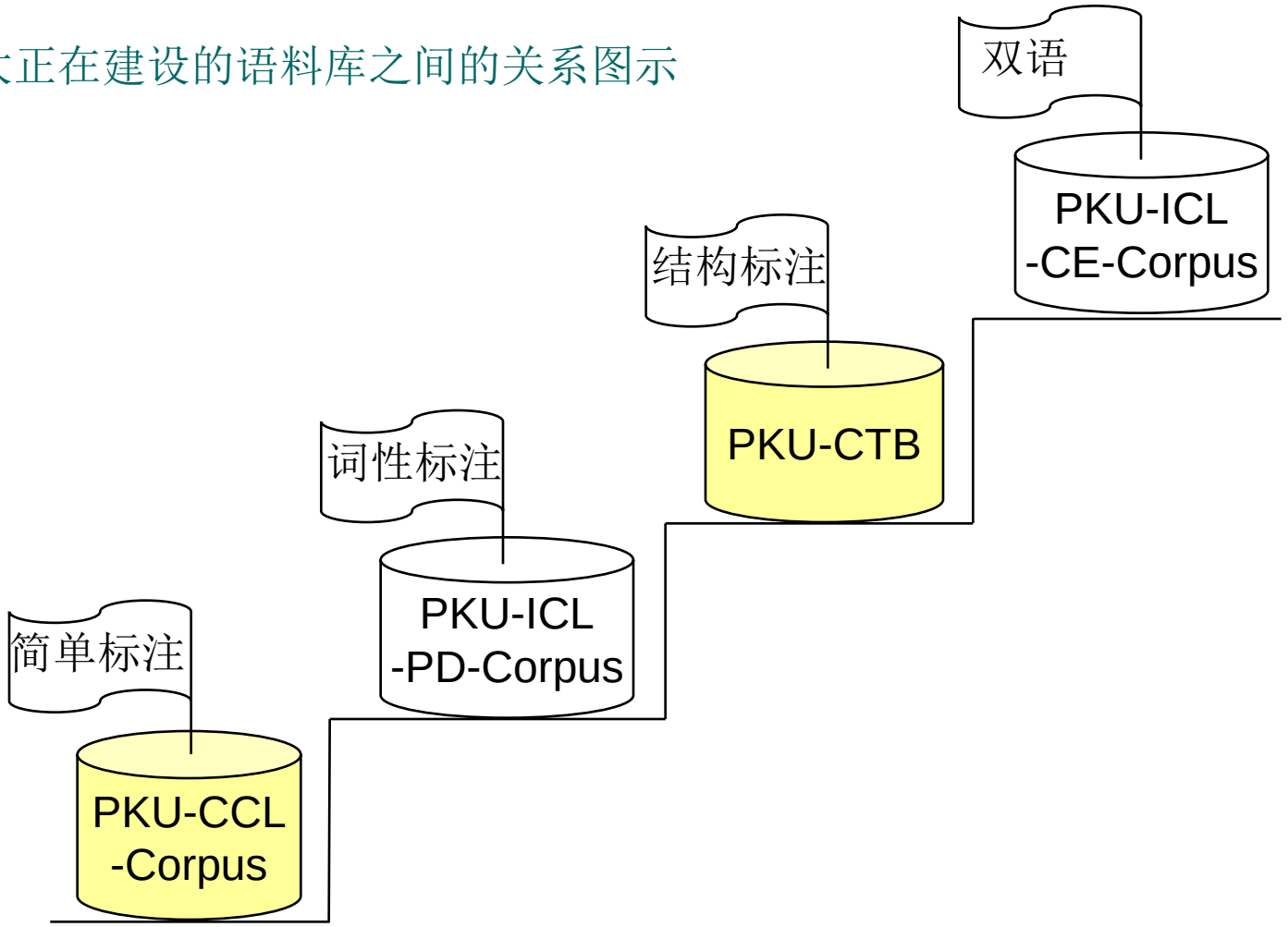
- 买好衣服还是买便宜衣服，……
- 老张昨晚喝酒还是喝多了，……

规则库调试工具

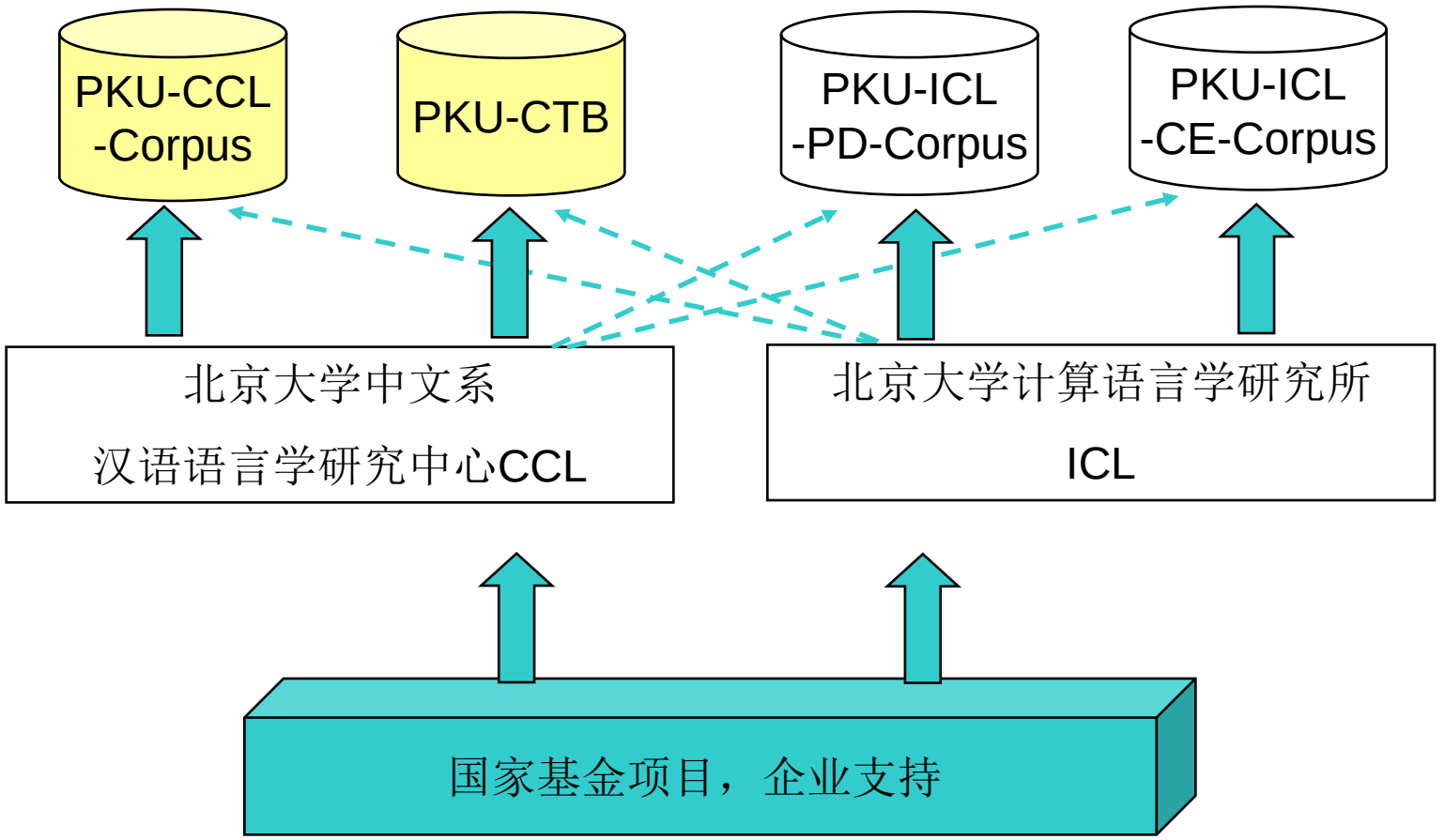
- 规则库管理（查询、编译）
- 跟踪分析过程
- 设置分析断点

3.3 语料库建设

北大正在建设的语料库之间的关系图示



背景简介

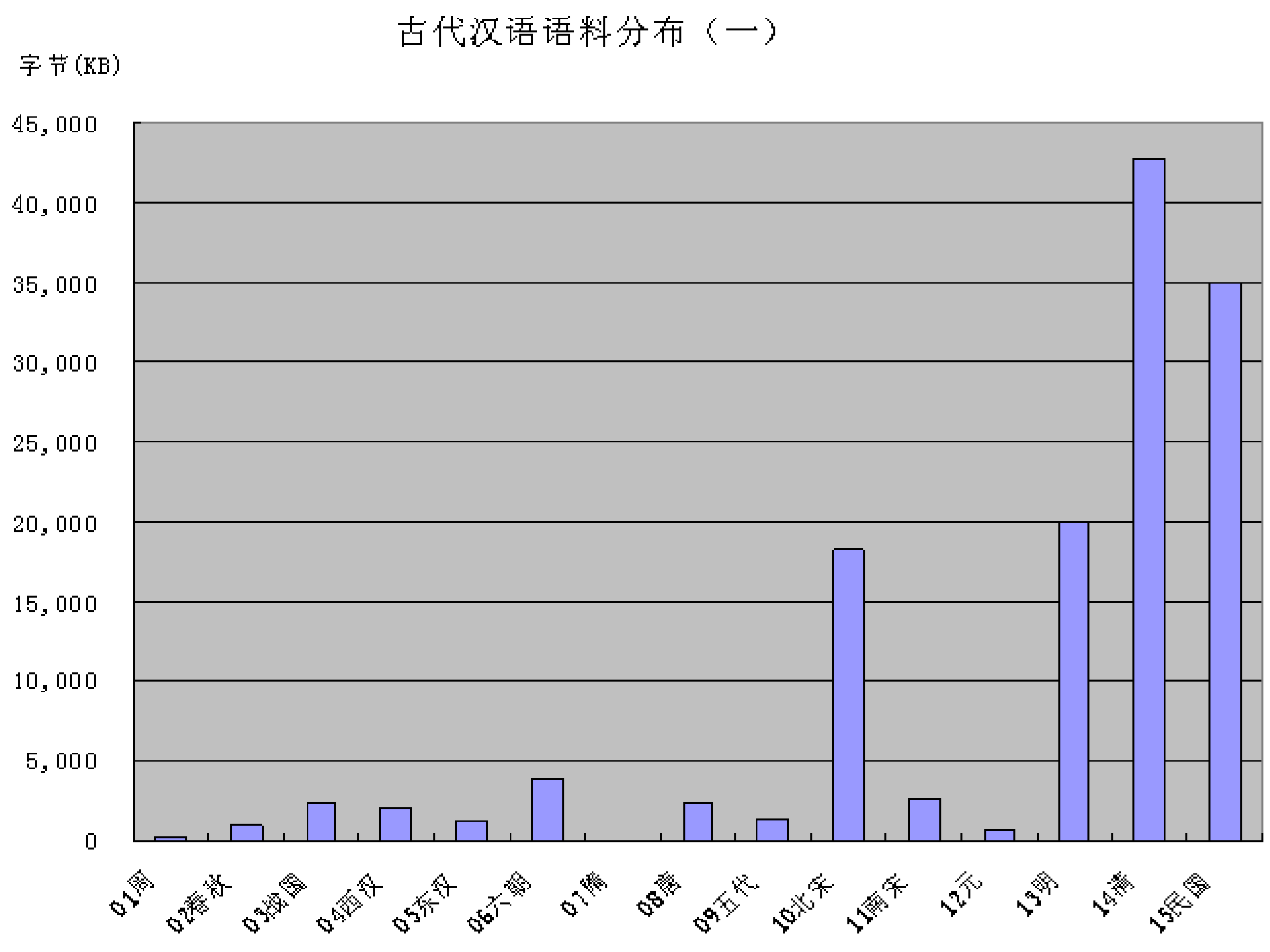

PKU-CCL
-Corpus
PKU-CTB
PKU-ICL
-PD-Corpus
PKU-ICL
-CE-Corpus
北京大学中文系
汉语语言学研究中心CCL
北京大学计算语言学研究所
ICL
国家基金项目，企业支持

3.3.1 大规模汉语在线语料库（PKU-CCL-Corpus）

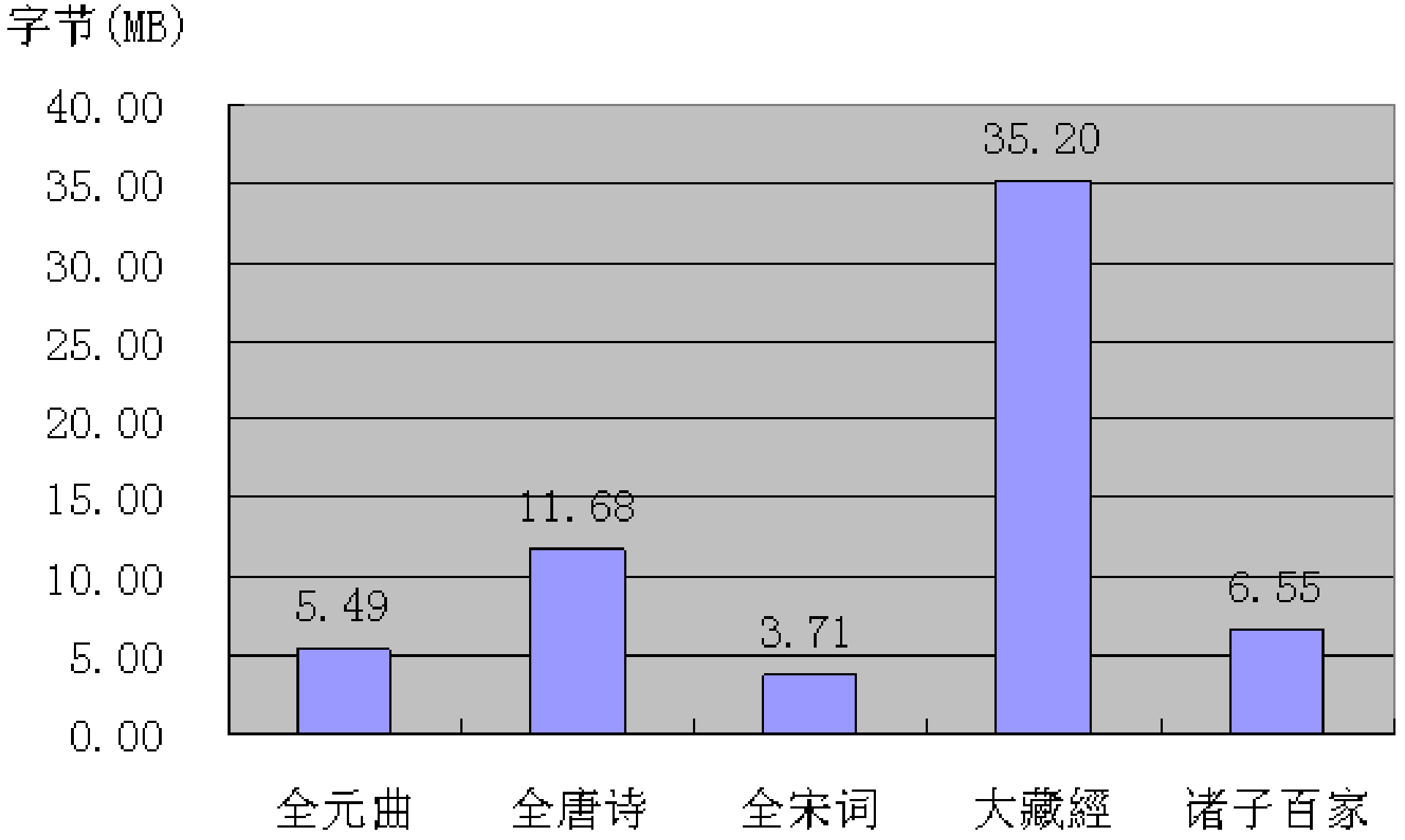
- 语料规模与分布
- 检索系统
- 检索示例
- 使用情况

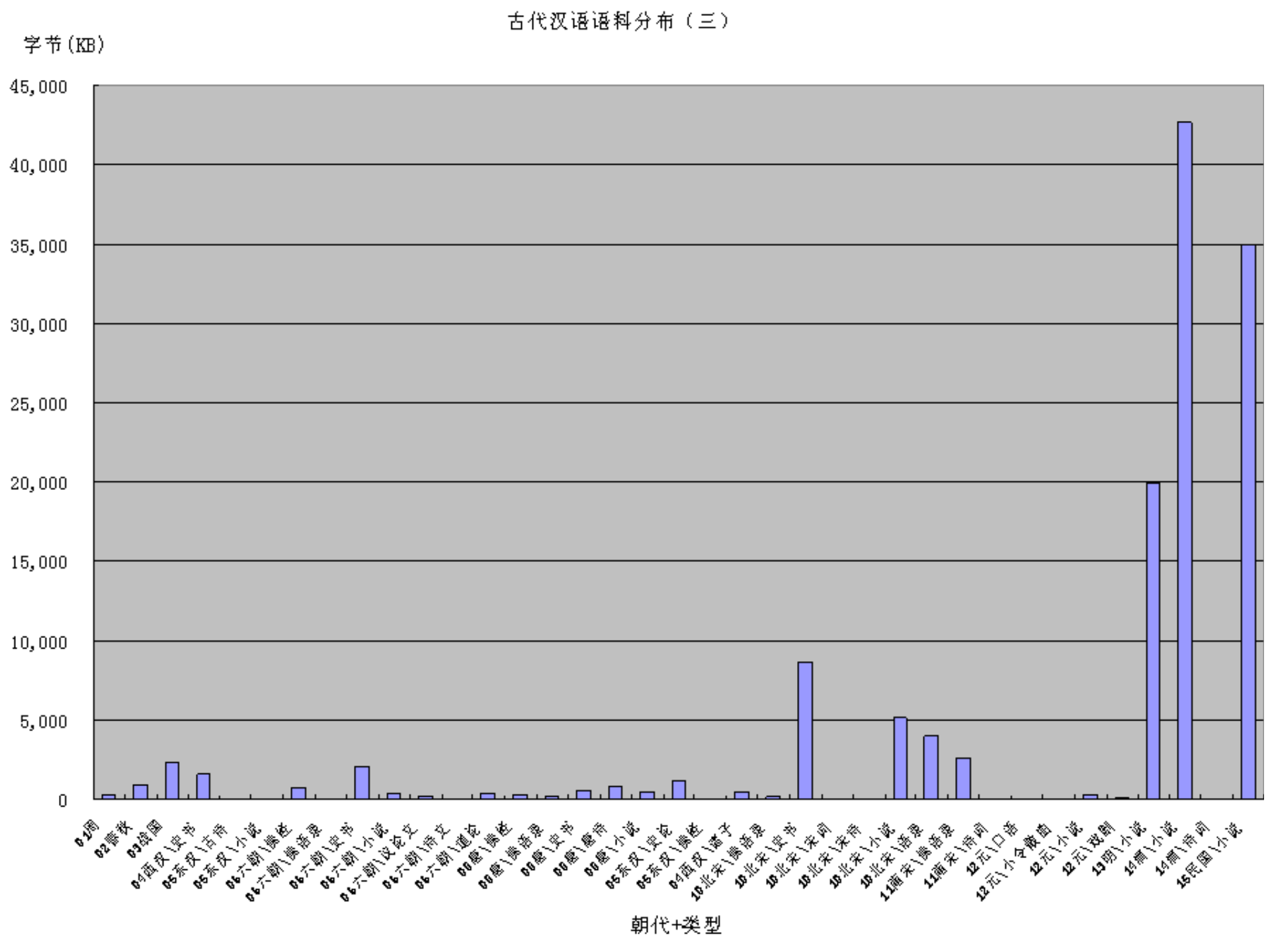
语料规模与分布

	古代汉语	现代汉语
文件夹	54	23
文件	486	157
字节	202,305,825	229,700,435
合计	432,006,260 （412 MB） （2.16 亿字）	

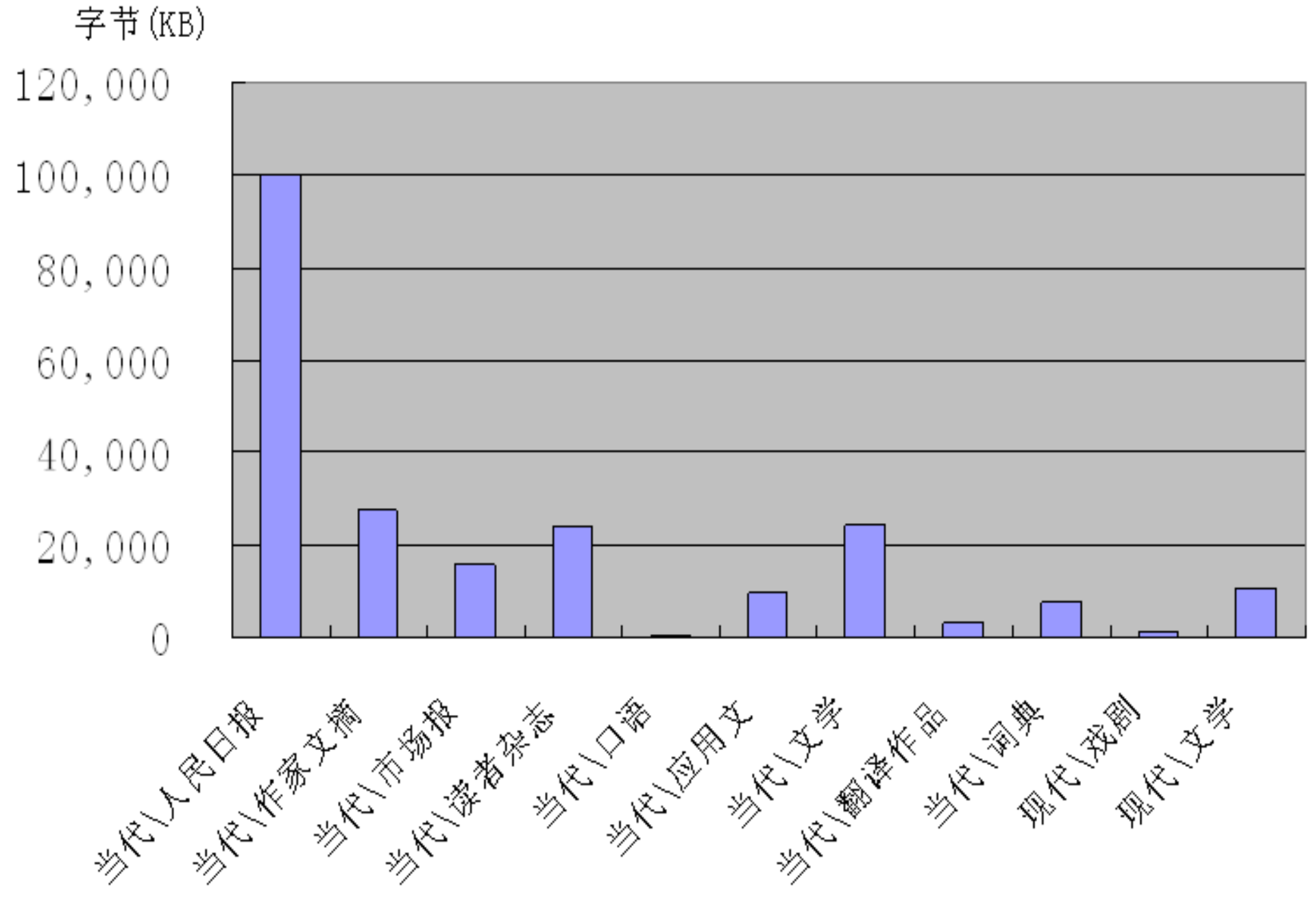

古代汉语语料分布（一）
字节(KB)
45,000
40,000
35,000
30,000
25,000
20,000
15,000
10,000
5,000
0
01周
02秦汉
03魏晋
04五代
05东汉
06六朝
07隋
08唐
09五代
10北宋
11南宋
12元
13明
14清
15民国

古代汉语语料分布（二）

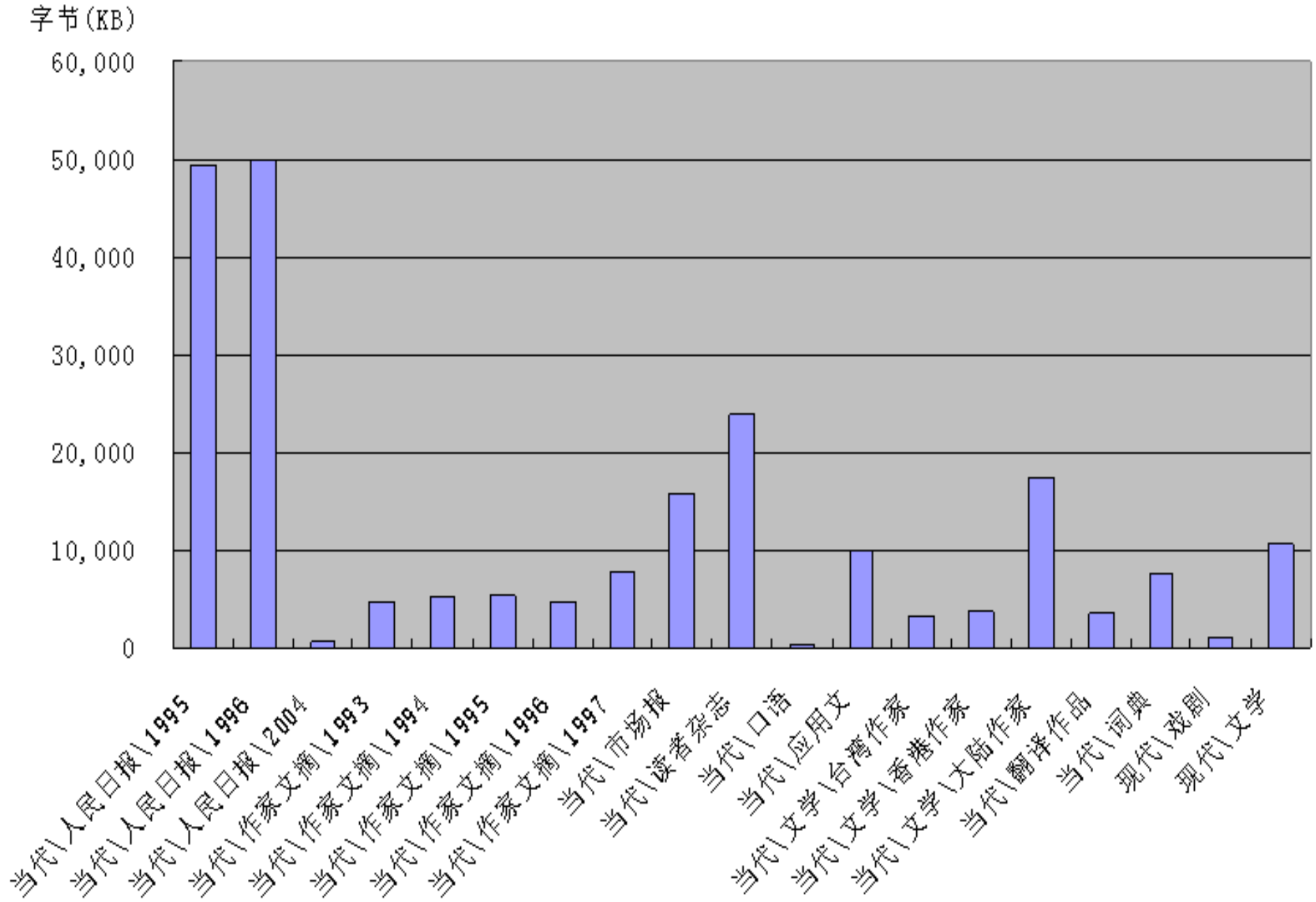

字节(MB)
40.00
35.00
30.00
25.00
20.00
15.00
10.00
5.00
0.00
5.49
11.68
3.71
35.20
6.55
全元曲
全唐诗
全宋词
大藏經
诸子百家


古代汉语语料分布（三）
字节(KB)
45,000
40,000
35,000
30,000
25,000
20,000
15,000
10,000
5,000
0
01周
02春秋
03战国
04西汉\史书
04西汉\古帝
05东汉\小说
05东汉\佛经
06六朝\佛经
06六朝\佛语录
06六朝\史书
06六朝\小说
06六朝\议论文
06六朝\诗文
06六朝\道论
08隋\佛经
08隋\佛语录
08隋\史书
08隋\帝
08隋\小说
05东汉\议论
05东汉\佛经
04西汉\诸子
10北宋\佛语录
10北宋\史书
10北宋\宋词
10北宋\宋诗
10北宋\小说
10北宋\语录
11南宋\佛语录
11南宋\诗词
12元\口语
12元\小令散曲
12元\小说
12元\戏剧
13明\小说
14清\小说
14清\诗词
15民国\小说
朝代+类型

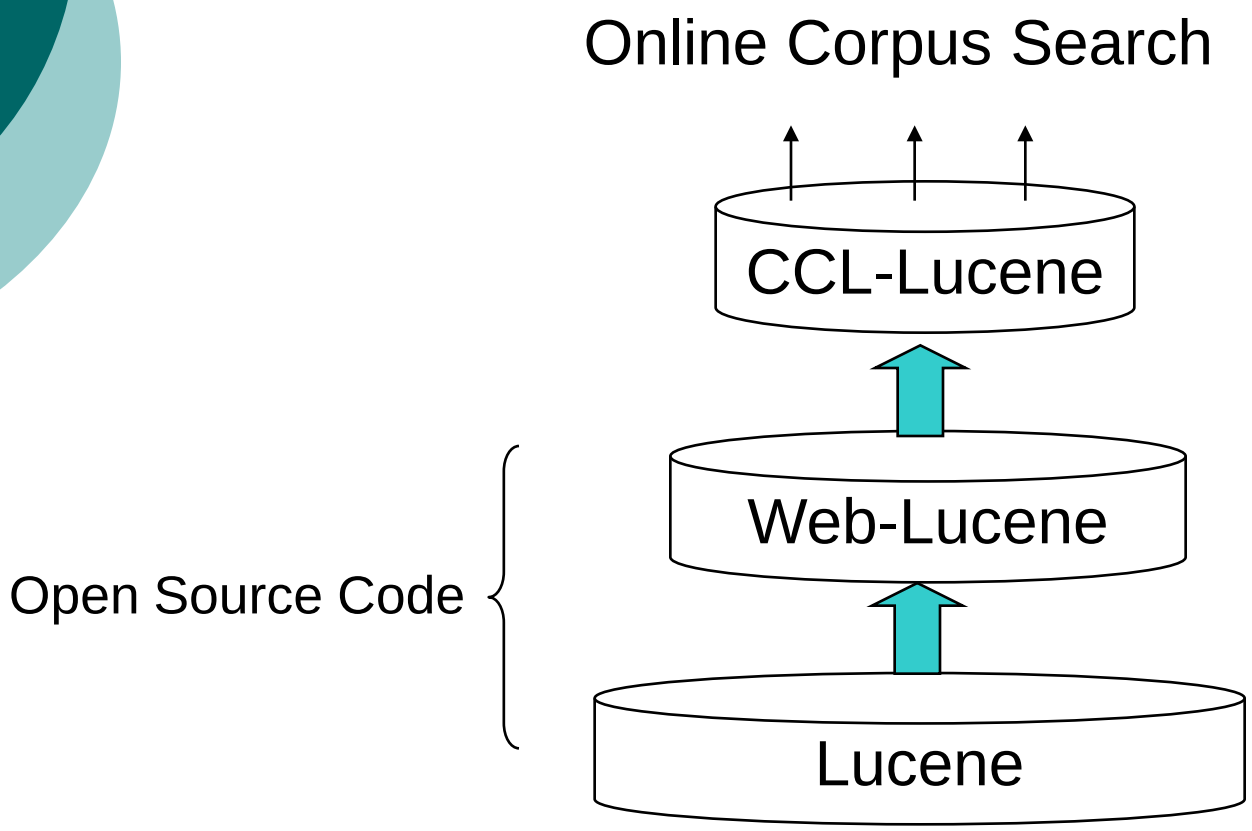
现代汉语语料分布（一）


字节(KB)
120,000
100,000
80,000
60,000
40,000
20,000
0
当代\人民日报
当代\作家文摘
当代\市场报
当代\读者杂志
当代\口语
当代\应用文
当代\文学
当代\翻译作品
当代\词典
现代\戏剧
现代\文学

现代汉语语料分布（二）


字节(KB)
60,000
50,000
40,000
30,000
20,000
10,000
0
当代\人民日报\1995
当代\人民日报\1996
当代\人民日报\2004
当代\作家文摘\1993
当代\作家文摘\1994
当代\作家文摘\1995
当代\作家文摘\1996
当代\作家文摘\1997
当代\市场报
当代\读者杂志
当代\口语
当代\应用文
当代\文学\台湾作家
当代\文学\香港作家
当代\文学\大陆作家
当代\翻译作品
当代\词典
现代\戏剧
现代\文学

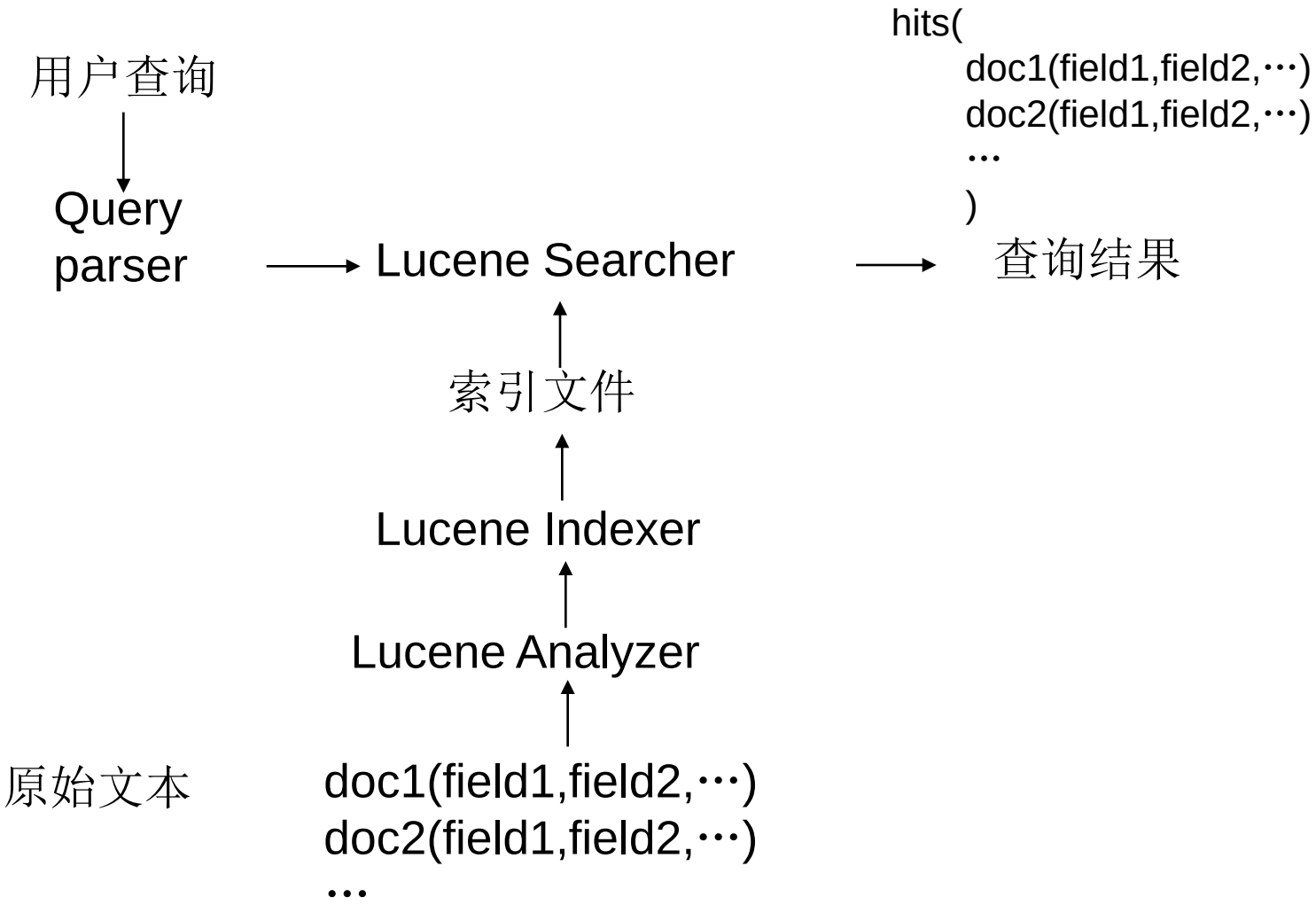
在线检索系统



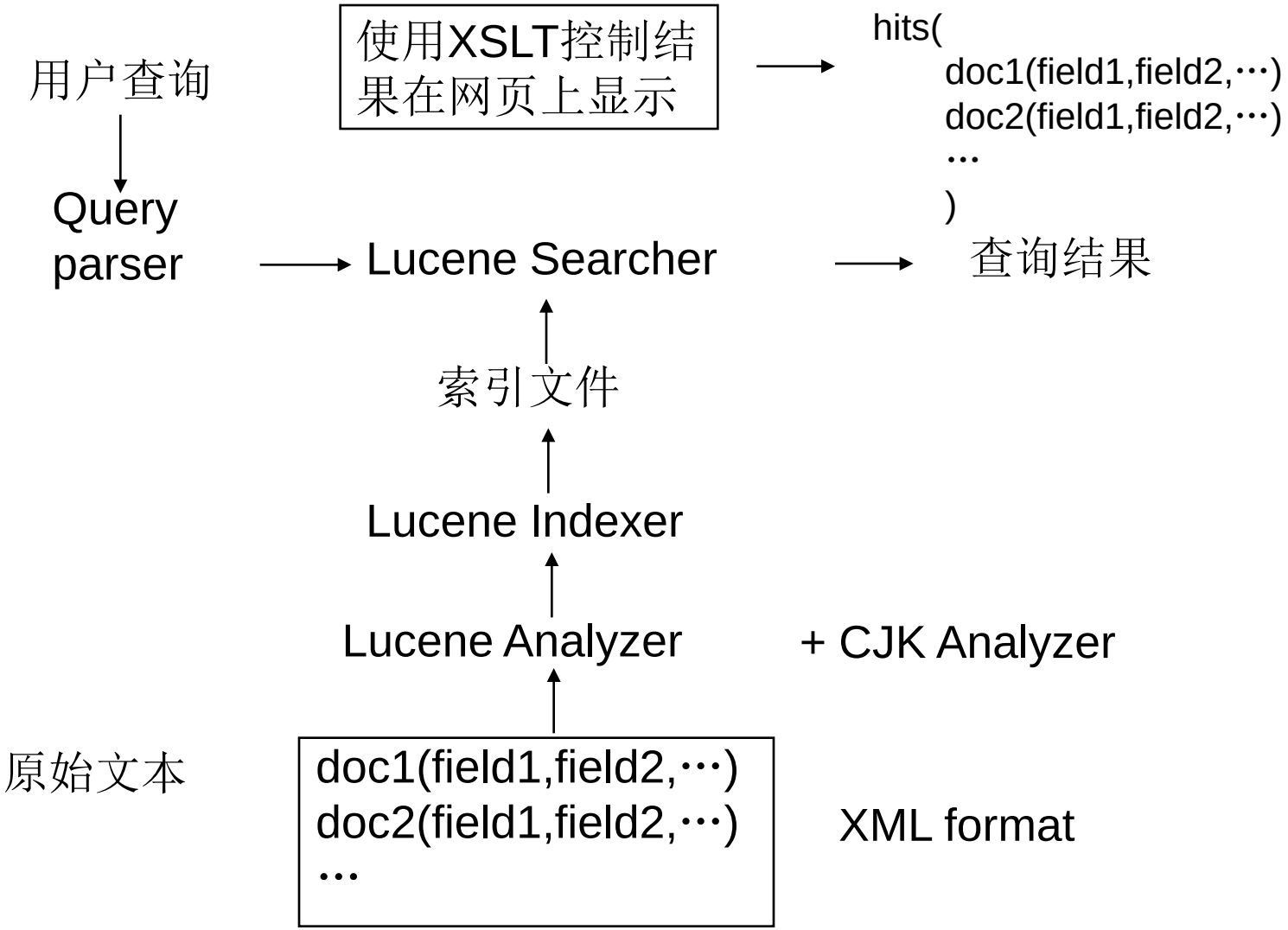
http://lucene.apache.org/java/docs/index.html

http://sourceforge.net/projects/weblucene/

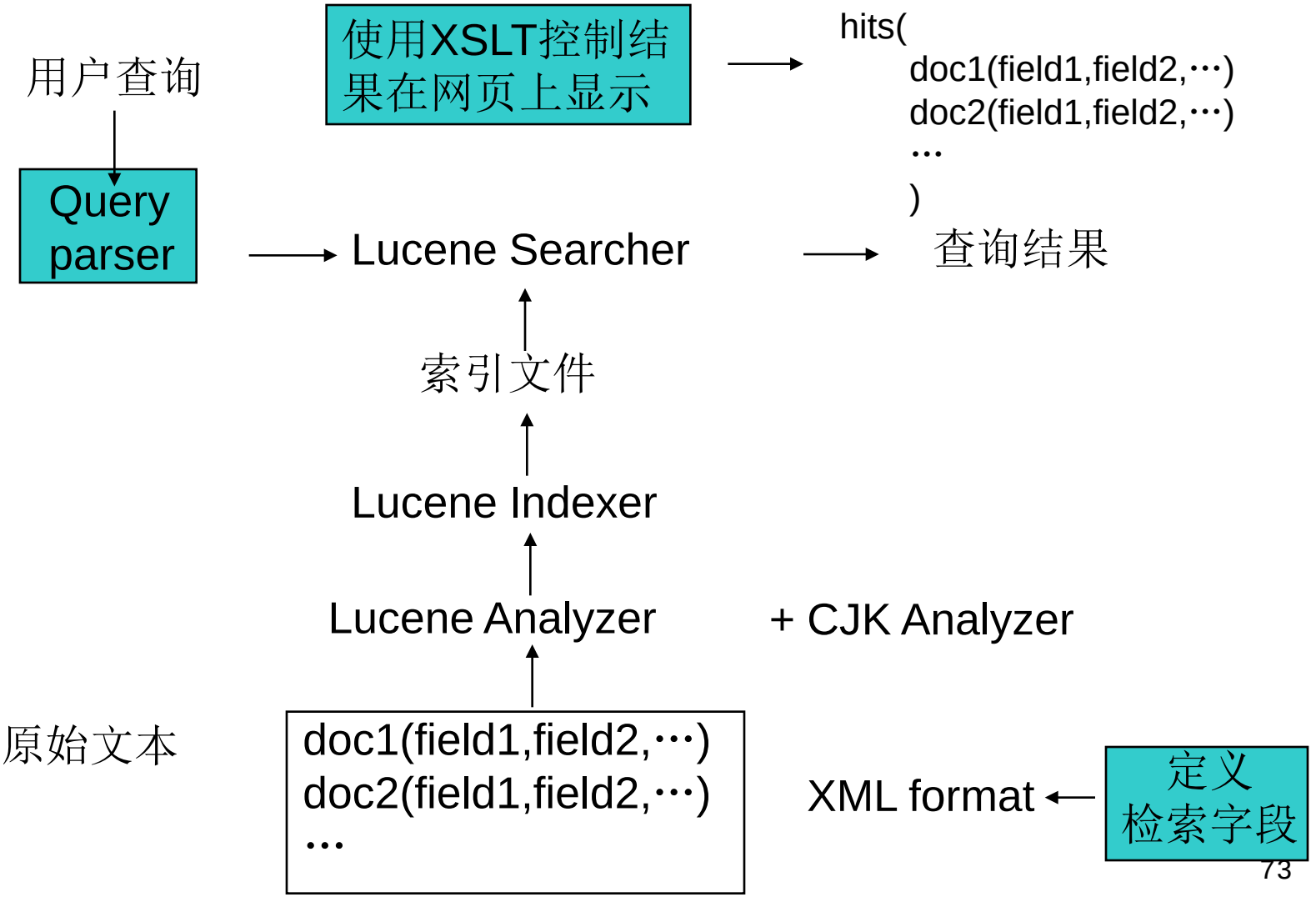
Lucene的基本架构(created by Doug Cutting)



Web-Lucene的基本架构(developed by 车东)



CCL-Lucene的定制



CCL-Lucene的定制

- 语料库文件的字段：作者，标题，…
- 查询表达式：重叠模式　定距查询非连续字符
- 结果显示：高亮 + 居中 + 左右字数 + 排序

检索应用示例

- 离合词用法
- 近义词比较

discontinuous compound：“洗澡”

查找	次数	查询表达式含义
洗	11,204	查出现“洗”的句子
澡	1,471	查出现“澡”的句子
洗 澡	1,185	同时出现“洗”“澡”的句子
洗澡	841	“洗澡”紧邻出现的句子
洗#100 澡 洗-0 澡	344	同时出现“洗”“澡”，并且两字相邻在 100 个字以内，同时不出现“洗澡”的句子
澡$10 洗 澡~0 洗	49	查‘澡’后出现‘洗’，并且两字相邻在 10 个字以内，但不出现‘洗澡’的句子

查询结果示例

query: 澡$10洗 澡~0洗

饭吃了一半、头剃了一半、	澡	洗了一半、梦做了一半
本来这个	澡	可以一直洗得十分惬意
我后来就想：人，只要有	澡	可洗，不是已经不错了？
跟随苦行僧去修道，一连奔波了六年，连	澡	也不洗，还是没有找到解决问题的办法。
还没吃上几口饭，已经瞌睡连连，连	澡	也不洗便倒头睡去。

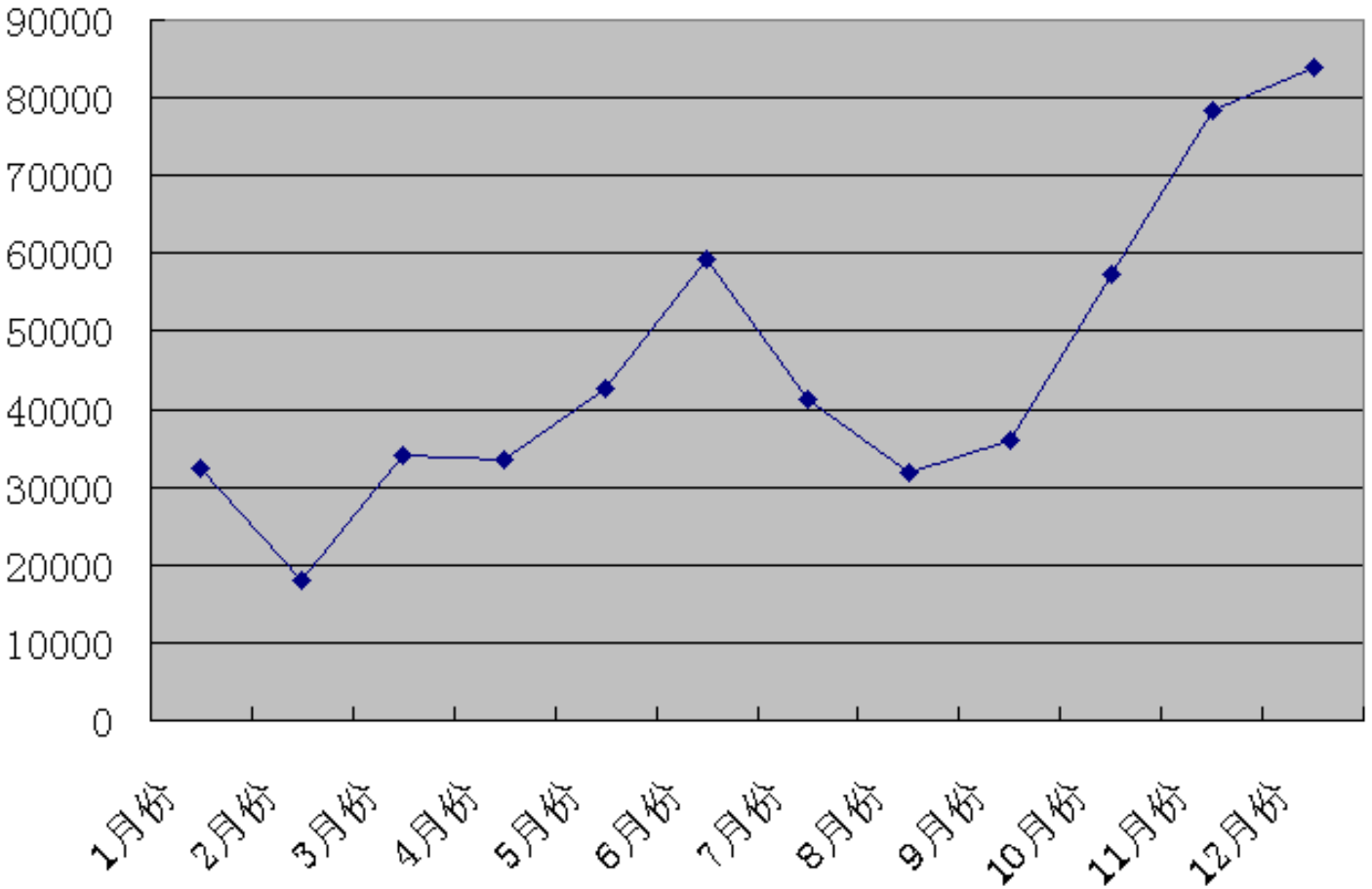
近义词比较：高兴 vs. 快乐

项目		高兴	快乐
总频次		10,065	3,077
AABB		329	25
ABAB		23	2
形容词用法	很~	1,368	101
	有点不~	40	1
	不大~	30	0
名词用法	许多~	0	12
	最大~	0	8
	得到~	0	11
用于标题或名称	《~》	1	55
搭配	很~看到	28	0
	由衷地~	92	0
	新春（圣诞）~	0	33

PKU-CCL-Corpus使用情况

2005年CCL在线语料库查询次数统计

平均月查询次数：45,711；日查询次数：1,562



3.3.2 北大中文树库

- 语料库基本信息
- 现代汉语树库的构建
- 树库应用举例

语料库基本信息

树库	字 种数	字 次数	词 种数	词 次数	句数	平均句长
T-I	1553	51295	4917	35480	1268	27.981
T-II	2415	151033	11763	93984	3553	26.452
T-III	1983	63499	5695	52202	4108	12.707
T-IV	2610	111494	9957	89794	10631	8.446
Total	3205	377321	22911	271460	19560	13.878

T-I: 政府白皮书；T-II:新闻语料；T-III：语文课本；T-IV：机译系统评测句子集

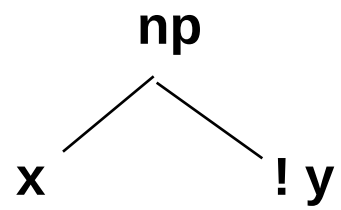
树库短语标记集（22个）

标记	含义	标记	含义
hl	文章标题	npt	机构名短语
zj	整句	npx	非中文字符串
yj	带引号的句子	npz	其他专名短语
dj	单句	pp	介词短语
fj	复句	qp	量词短语
		sp	处所短语
ap	形容词短语	tp	时间短语
dp	副词短语	vp	动词短语
mp	数词短语		
np	名词短语	yp	语篇成分
npr	人名短语	ypc	插入语
nps	地名短语	yph	呼语

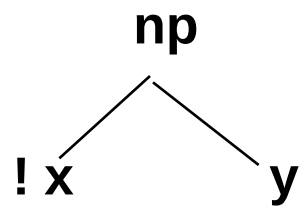
句法树结构标注示例

(zj (!fj (tp (tp (m (1992) !q (年)) !tp (m (3) !q (月))) wco (，) !fj (!dj (npt (!nt (联合国)) !vp (!vp (!v (决定)) dj (np (qp (m (第四) !q (次)) !np (np (np (!n (世界)) !np (!n (妇女))) !np (!n (大会)))) !vp (pp (!p (于) tp (m (1995) !q (年))) !vp (pp (!p (在) np (np (nps (!ns (中国)) !np (!n (首都))) !nps (!ns (北京)))) !vp (!v (召开))))))) wco (，) dj (rn (这) !vp (!vp (!v (使)) np (np (nps (!ns (中国)) !np (!n (妇女))) ude1 (的) !np (!n (状况))) vp (!vp (!v (倍受)) np (!n (世界)) vp (!v (关注))))))) wfs (。)))

(np (mp (!m (一)) wsc (、) !np (np (nps (!ns (中国)) !np (!n (妇女))) ude1 (的) !vp (np (!n (历史性)) !vp (!v (解放))))))

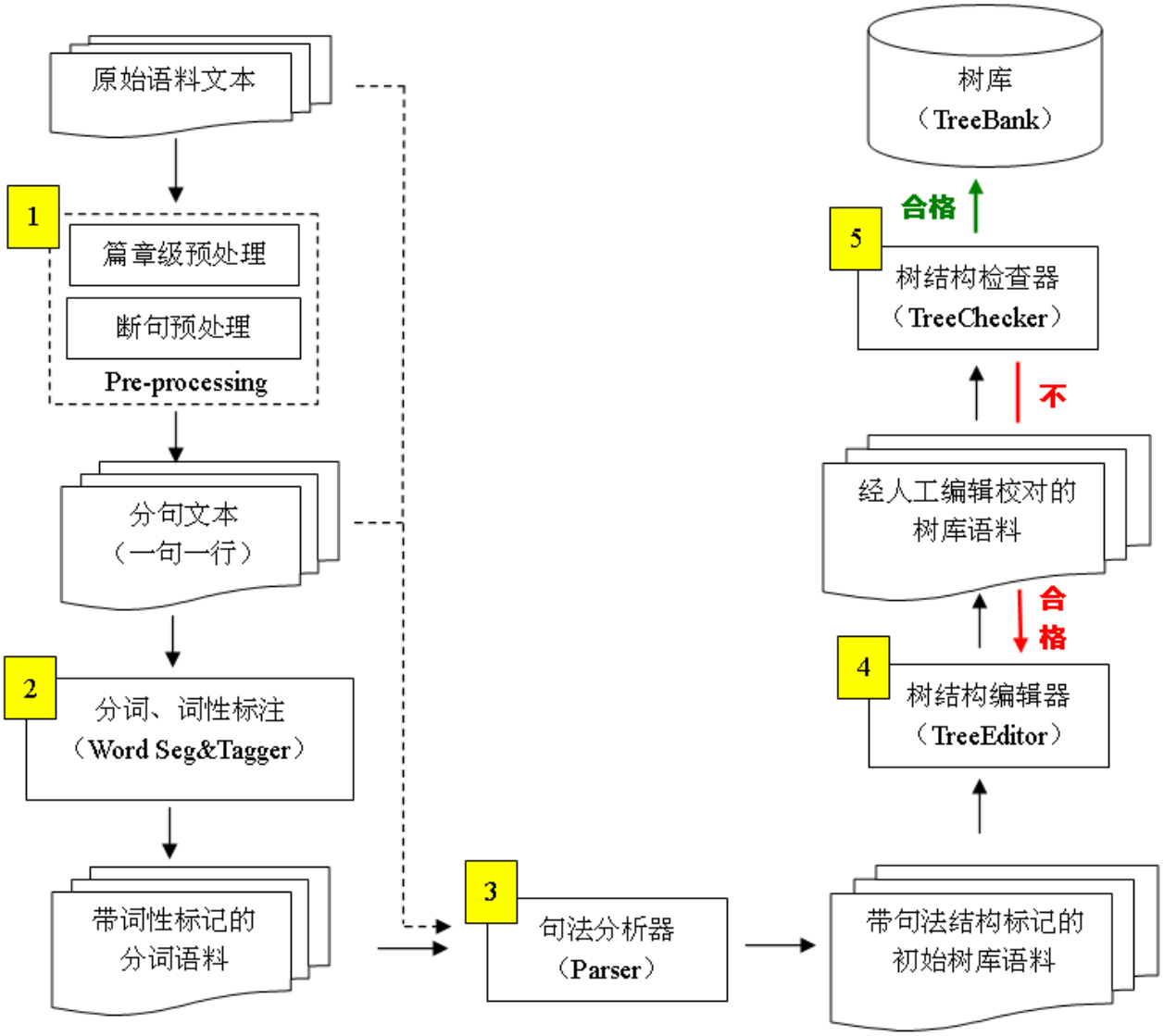


modifier – head structure



coordination structure

PKU-CTB的加工流程



基于规则的句法分析器的效果

样本来源	词数	句数	平均句长	召回率	精确率	F-Score
T-I	2270	100	22.70	67.24	71.97	69.53
T-II	3088	100	30.88	58.93	54.76	56.77
T-III	5587	254	22.00	68.44	75.71	71.89
T-IV	1031	120	8.59	81.56	73.09	77.09

-- All --

Number of sentence	=	100
Number of Error sentence	=	0
Number of Skip sentence	=	0
Number of Valid sentence	=	100
Bracketing Recall	=	67.24
Bracketing Precision	=	71.97
Bracketing Fscore	=	69.53
Complete match	=	3.00
Average crossing	=	5.73
No crossing	=	14.00
2 or less crossing	=	34.00

-- len<=40 --

Number of sentence	=	90
Number of Error sentence	=	0
Number of Skip sentence	=	0
Number of Valid sentence	=	90
Bracketing Recall	=	68.66
Bracketing Precision	=	73.01
Bracketing Fscore	=	70.77
Complete match	=	3.33
Average crossing	=	4.73
No crossing	=	15.56
2 or less crossing	=	37.78

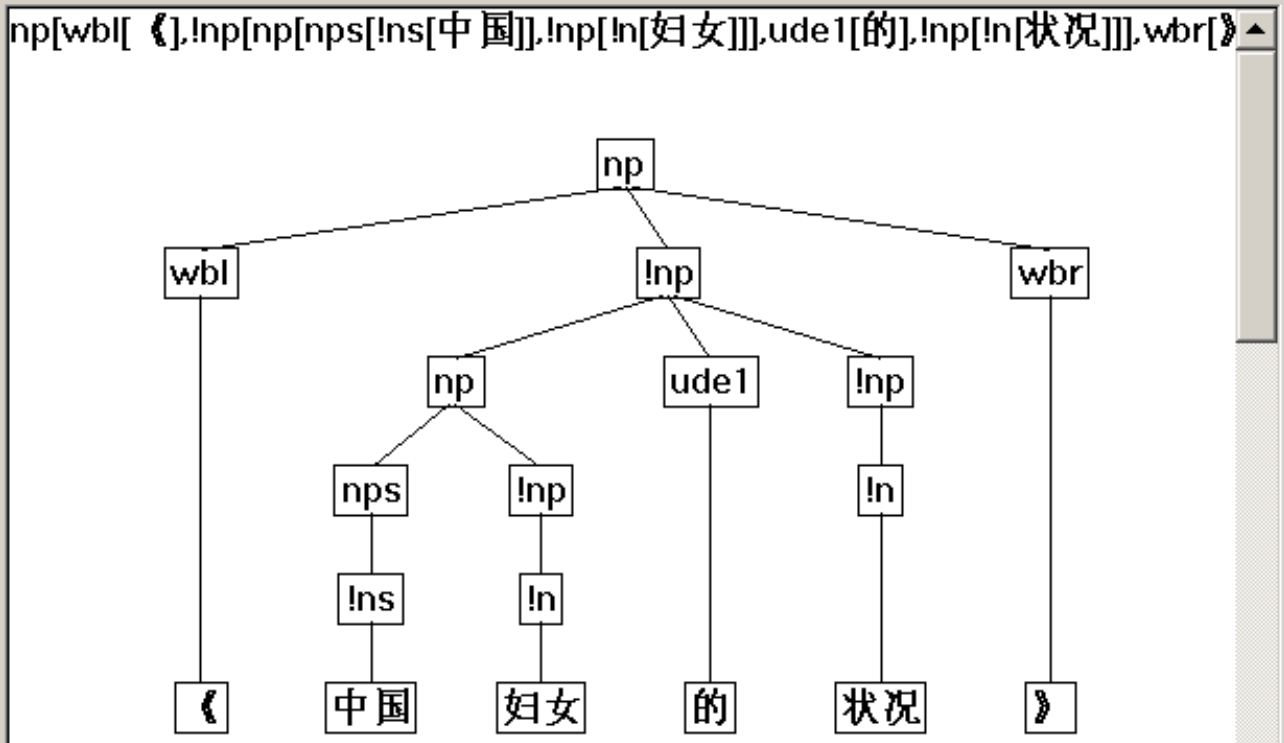
E:\TreeBank\Data\zhp_funvzhuangkuang_01_tree.txt

文件(F) 编辑(E) 核查(C) 工具(T) 文件管理(M) 句子管理(S) 帮助(H)

恢复 放大 缩小 切换 跳转 查错 查找

- np
  - wbl
    - 《
  - !np
    - np
      - nps
        - !ns
          - 中国
      - !np
        - !n
          - 妇女
    - ude1
      - 的
    - !np
      - !n
        - 状况
  - wbr
    - 》

np[wbl[《],!np[np[nps[!ns[中国]],!np[!n[妇女]]],ude1[的],!np[!n[状况]]],wbr[》]]
np[!n[前言]]
zj[!fj[tp[tp[m[1992],!q[年]],!tp[m[3],!q[月]]],wco[，],!fj[!dj[npt[!nt[联合国]],!vp[
np[mp[!m[一]],wsc[、],!np[np[nps[!ns[中国]],!np[!n[妇女]]],ude1[的],!vp[np[!



当前文件共 133 句 数字 当前选定第 1 句

树库应用示例

- 抽取规则
- 抽取短语实例
- 句法结构查询
- 短语分布统计

抽取规则

序号	短语	规则	频次
…	…	…	…
273	np	np → !n	9655
274	np	np → np !np	2905
275	np	np → ap !np	863
276	np	np → np ude1 !np	518
277	np	np → qp !np	495
278	np	np → !np c np	447
279	np	np → vp !np	415
280	np	np → vp ude1 !np	331
…	…	…	…

从T-I中抽取的部分规则示例

抽取短语实例（从T-I中抽取的短语实例）

短语	深度	宽度	实例
…	…	…	…
np	5	4	（np（np（np（!n（职业））!np（!n（技能））!np（vp（!v（开发））!n（工作））））
np	5	3	（np（vp（!v（就业））!np（vp（!v（培训））!n（中心））））
np	5	3	（np（vp（vp（!v（劳动））!vp（!v（服务））!n（企业）））
np	5	3	（!np（vp（vp（!v（就业））!vp（!v（服务））!n（体系）））
np	5	3	（!np（dj（np（!n（社会））!ap（!a（保险））!np（!n（事业））））
np	5	5	（np（!np（n（社会）!v（统筹））c（与）np（np（!n（个人））!np（!n（帐户））））
np	5	4	（!np（ap（!a（基本））!np（np（v（养老）!a（保险））!np（!n（模式））））
…	…	…	…

各类短语的平均深度和宽度

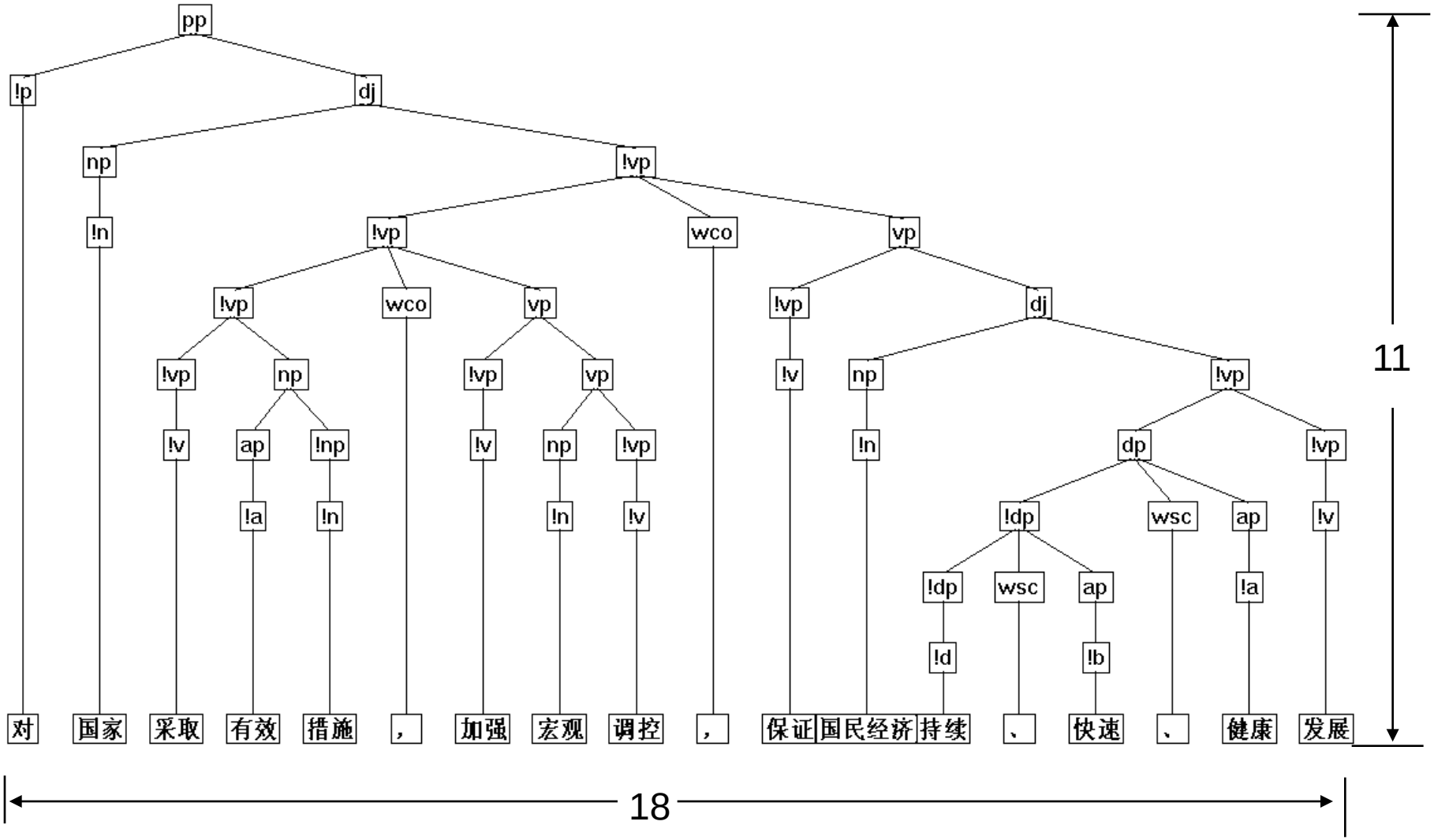
phrase	zj	fj	dj	pp	vp	np	sp	ap	dp	tp	mp	qp
average depth	11	10	8	6	6	5	5	4	4	4	3	3
average width	28	29	13	5	7	4	4	2	2	3	2	2

根据T-I语料库计算的短语深度和宽度平均值

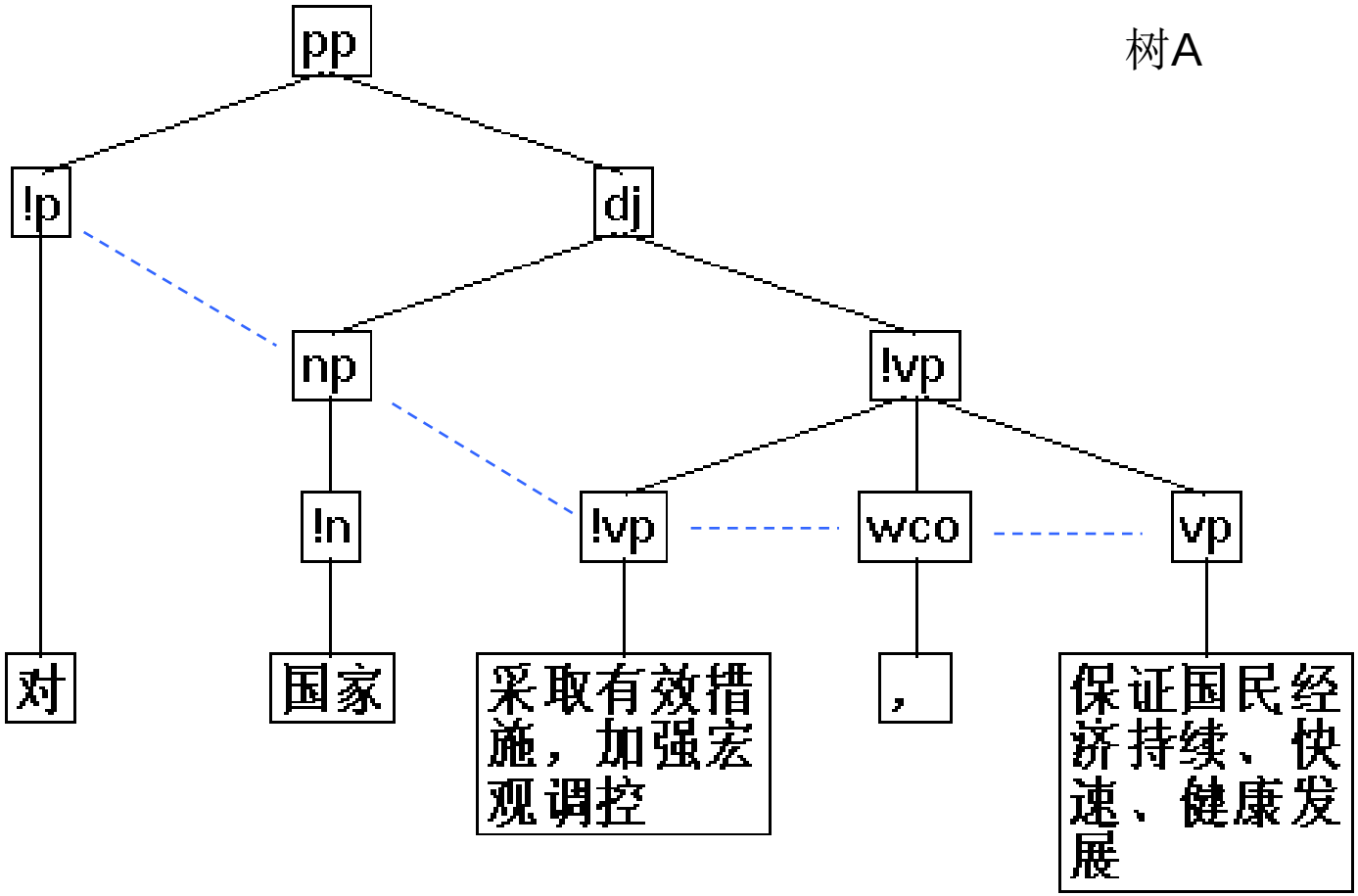
T-I树库中pp短语宽度分布

宽度	2	3	4	5	6	7	8	9
频次	231	189	135	102	88	76	54	37
宽度	10	11	12	13	14	15	…	45
频次	35	30	18	16	15	13	…	1

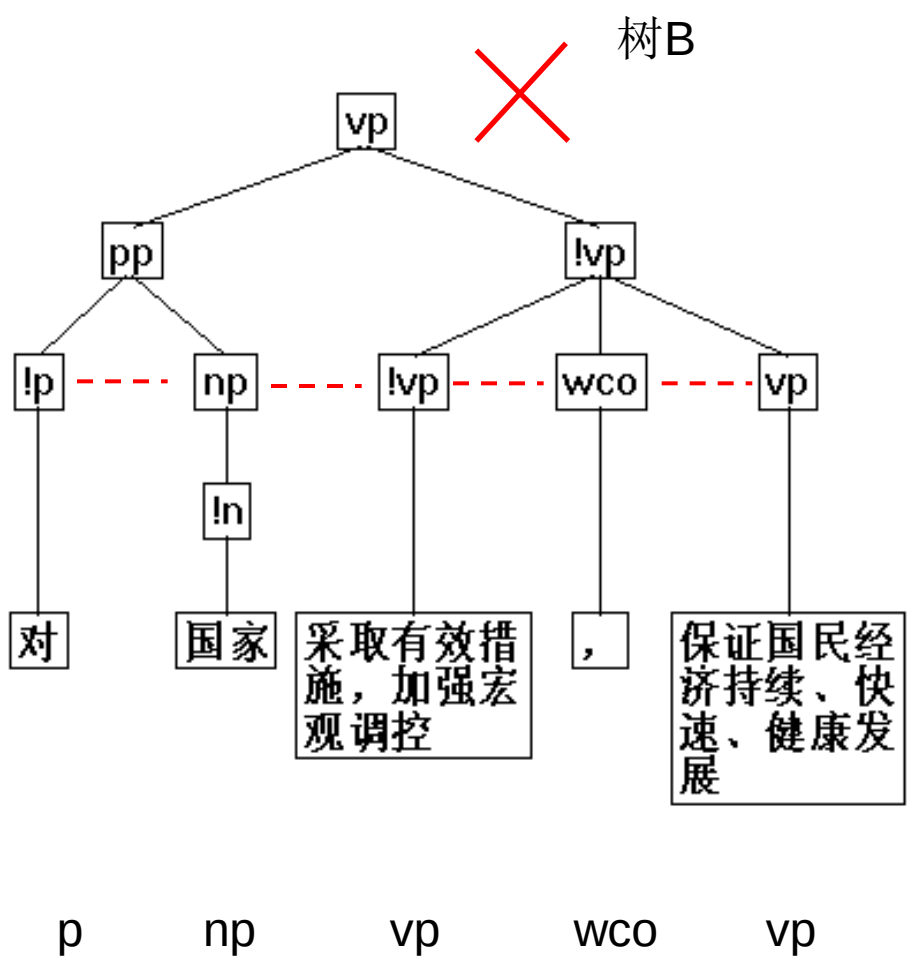
一个pp结构示例— 深度11；宽度18

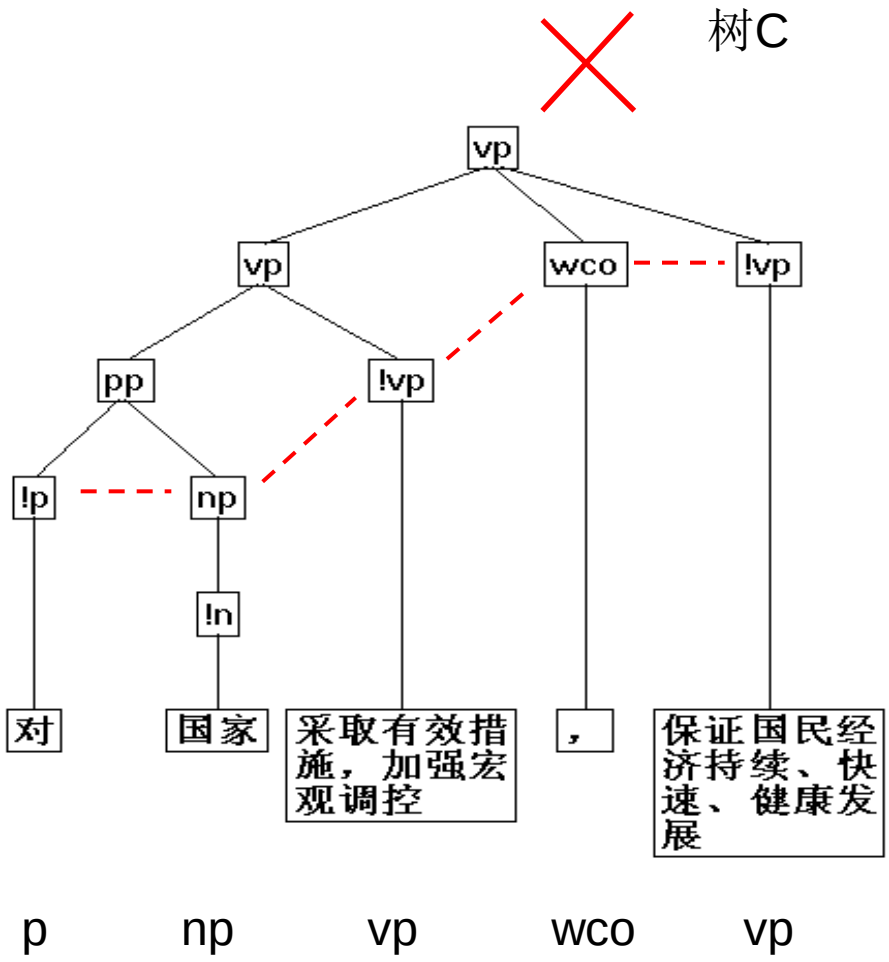


pp的结构骨架


pp
树A
lp
dj
np
lvp
ln
lvp
wco
vp
对
国家
采取有效措施，加强宏观调控
，
保证国民经济持续、快速、健康发展
p
np
vp
wco
vp

错误的结构标注方式

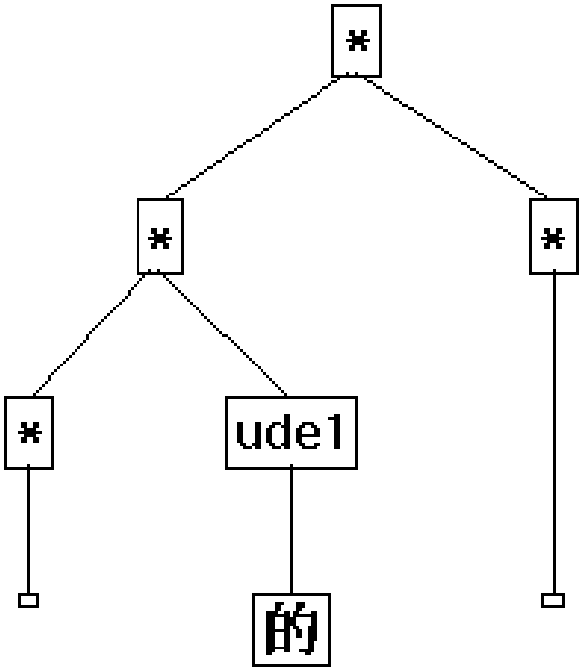

树B
vp
pp
!vp
!p
np
!vp
wco
vp
!n
对
国家
采取有效措施，加强宏观调控
，
保证国民经济持续、快速、健康发展
p
np
vp
wco
vp


树C
vp
vp
wco
!vp
pp
!vp
!p
np
!n
对
国家
采取有效措施，加强宏观调控
，
保证国民经济持续、快速、健康发展
p
np
vp
wco
vp

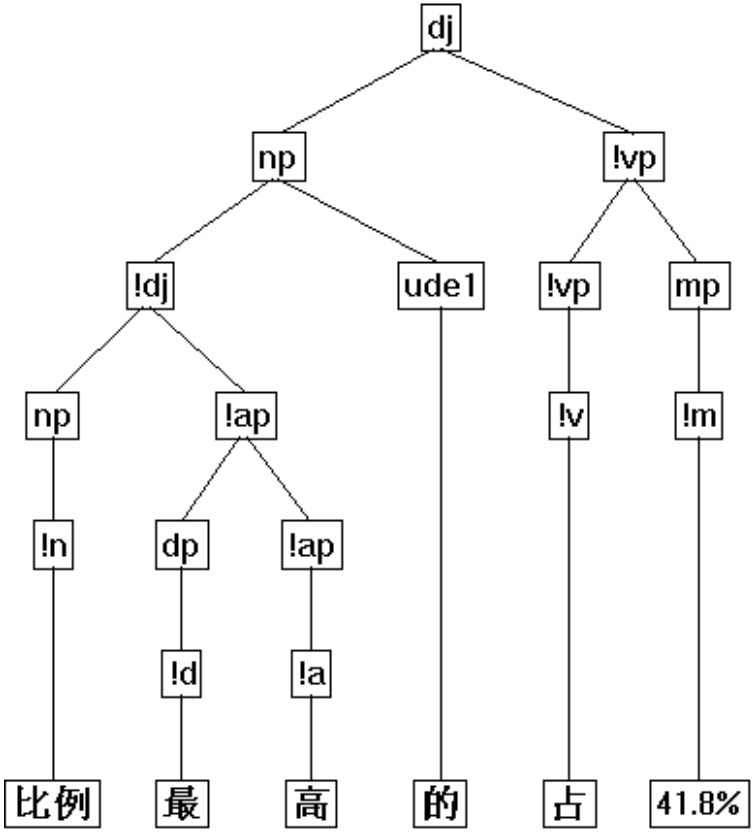
句法结构查询

查询： 根节点：*

子节点：*[*,ude1],*



结果示例：



短语分布统计

从T-I中抽取的pp短语分布情况

No.	root	left	**phrase**	right	Absolute freq.	Relative freq.	accumulative relative freq.
1	vp	##	**pp**	!vp	580	0.530650	0.530650
2	dj	##	**pp**	wco !dj	136	0.124428	0.655078
3	fj	##	**pp**	wco !fj	91	0.083257	0.738335
4	vp	##	**pp**	wco !vp	86	0.078683	0.817017
5	np	##	**pp**	ude1 !np	57	0.052150	0.869167
6	ap	##	**pp**	!ap	31	0.028362	0.897530
7	vp	!vp	**pp**	##	22	0.020128	0.917658
8	np	##	**pp**	ude1 !vp	17	0.015554	0.933211
9	dj	##	**pp**	!dj	9	0.008234	0.941446
10	dj	##	**pp**	wco !vp	7	0.006404	0.947850

四 结 语

计算机要具备处理自然语言的能力，就必须掌握语言成分之间的组合规则：

（1）什么成分之间可以组合？
（2）以什么关系组合？
（3）组合之后的整体具有什么样的性质（即如何跟其他成分再组合）？

为了解决上述问题，就要对语言成分（单位）进行分类。
有多少个词类、短语类？分多少类合适，如何分比较好？

困难：
（1） 一个语言成分（f）兼属多类 （多义）
（2） 两个语言成分（f1，f2）被归属同一类，结果个体差异被忽视了

结语（续）

- 自然语言成分组合的非传递性

  * 这个房间 比那个房间 不 干净

    这个房间 比那个房间 还 不 干净

  * 干净 一件 衣服

    这么 干净 一件 衣服

分类问题

这是一项非常基础性的工作

这样做是不诚实的

？这样做是不狡猾的

分类问题（续）

- 三米多　　—　　*三多米
- 十米多　　—　　十多米
- ？二十米多　—　二十多米
- 　二十米多一点　—　* 二十多一点米

- 1945年　　—　45 年　—　8年
- 四五年　三五年　二五年

参考文献

- 陆俭明. 1990.〈“VA了”述补结构的语义分析〉，载《汉语学习》1990年第1期。
- 马真、陆俭明. 1997.〈形容词作补语情况考察〉，载《汉语学习》1997年第1、4、6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，广西科学技术出版社。

- Ann, Copestake & Flickinger Dan (2000), An Open Source Grammar Development Environment And Broad-coverage English Grammar Using HPSG, In Proceedings of LREC 2000 (The 2nd International Conference on Language Resoruce & Evaluation), Zappeion Megaron, Greece, May 31 – June 2, 2000.
- Blevins, James(2003), Feature-based Grammar, In Borsley, R.D. & Borjars, K. eds., Non-transformational Syntax, Oxford: Blackwell, to be published in 2005.
- Blache, Philippe, Marie-Laure Guйnot, Tristan van Rullen (2003), A Corpus-based Technique for Grammar Development, In Proceedings of The Shallow Processing of Large Corpora Workshop (SProLaC 2003), Lancaster University (UK), 27 March, 2003.
- Borsley, Robert D., 1996, *Modern Phrase Structure Grammar*, No. 11 in Blackwell textbooks in Linguistics, Blackwell Publishers Inc..
- Chen,Feng-yi, et al. 1999, Sinica Treebank, Computational Linguistics and Chinese Language Processing,4(2):183-204
- Chen, Keh-Jiann & Yu-Ming Hsieh, 2002, Chinese Treebanks and Grammar Extraction, CJNLP’2002, Peking University, 2002.10.30-11.2

参考文献（续）

- Erbach, Gregor (1991), A Flexible Parser for a Linguistic Development Environment, In O. Herzog & C.-R. Rollinger eds., Text Understanding in LILOG, Springer, 1991, pp. 74-87
- Heinecke, Johannes, Jurgen Kunze, Wolfgang Menzel, and Ingo Schroder (1998), Eliminative parsing with graded constraints. In Proceedings of 17th International Conference on Computational Linguistics, 36th Annual Meeting of the ACL, Coling-ACL '98, Montreal, Canada, 1998.
- Knight,Kevin,1989, Unification: A Multidisciplinary Survey, ACM Computing Surverys, Vol.21, No.1.
- Sag, Ivan A. & Thomas Wasow, 1999, *Syntactic Theory: A Formal Introduction*, CSLI Publications, Stanford, California.
- Schmid, Helmut (1999), YAP: Parsing and Disambiguation With Feature-Based Grammar. PhD thesis, Institute of Maschinelle Sprachverarbeitung, University Stuttgart, Germany, 1999.
- Suzuki, Hisami (2002), A Development Environment for Large-scale Multi-lingual Parsing Systems, In Workshop on Grammar Engineering and Evaluation (Post-conference workshop in conjunction with COLING-2002, Taipei, Sept. 1, 2002).
- Uszkoreit, Hans(2002), New Chances for Deep Linguistic Processing, Coling2002,Taipei.
- Volk, Martin & Dirk Richarz (1997), Experiences with the GTU Grammar Development Environment, ACL workshop on Environments for Grammar Development, 1997, Madrid, Spain.
- Xu,Ruifeng, et al., 2004, The Construction of A Chinese Shallow Treebank, ACL2004 SIGHAN Workshop, July 21-16, 2004. Barcelona, Spain.

The End

谢　谢

zwd@pku.edu.cn