第三讲 汉语的句法规则系统（3）

汉语形式语法系统的建构

詹卫东

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

3.1 形式语法系统的组成

- 产生式规则：
  描述一个语言的基本范畴及其组合模式；

- 基于特征结构的合一约束：
  描述基本范畴之间发生组合关系的条件；

附录2：形式语法的BNF描述

一个简单的例子

- S=“一件衣服”
- S的内部组成情况；

  qp + np ，定中结构, qp是定语,np是中心语；“一件”跟“衣服”可以组合成定中结构；
- S的外部功能情况；

  S作为一个整体是一个名词短语（np），可以充任主谓结构的主语，述宾结构的宾语，定中结构的中心语；不能充任述补结构的补语，定中结构的定语，状中结构的状语和中心语；

规则

说明np的内部结构，定语，中心语，以及功能特征等

说明对中心语np的约束条件（独立条件）

&& {R1} np -> qp !np

$.内部结构=定中,$.定语=%qp,$.中心语=%np,$.zuodingyu=否,…,
%np.数量名=是,…,
IF %qp.量词子类=个体 THEN %np.个体量词=%qp.原形 ENDIF,…

①
②
③

&& {R2} qp -> mp !q

$.内部结构=数量,$.定语=%mp,$.中心语=%q,$.zuodingyu=是,…,

&& {R3} np -> !n

$.内部结构=单词

说明对定语mp与中心语np之间的相互约束条件（组配条件）

规则（续）

A

np → qp !np :: $.zuodingyu=否 /* “一件衣服”不能再作定语 */
np → np !np :: %np. zuodingyu=是, %%np. zuozhongxinyu=是

B

np → qp !np :: $.内部结构=组合定中
np → np !np :: IF %np.内部结构=组合定中,%%np.内部结构=单词 FALSE

可以用不同的约束条件，达到同样的约束效果

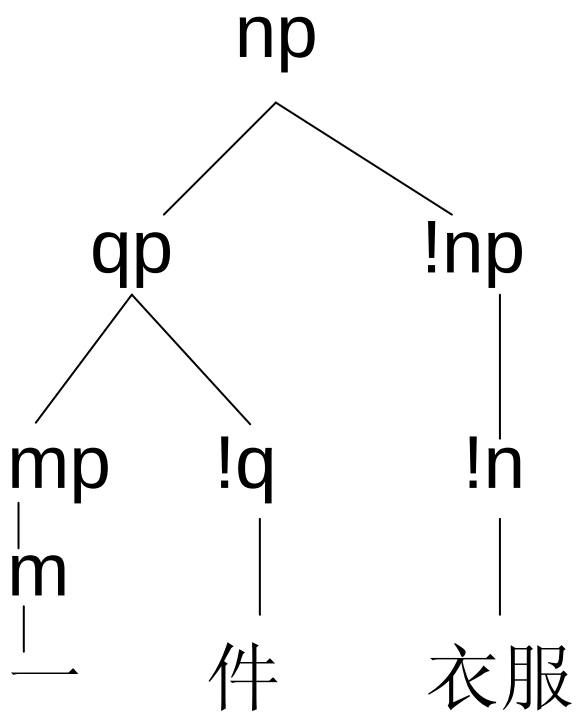
词典

词语	特征结构
……	……
一	[词性:m,数词子类:基数]
件	[词性:q,量词子类:个体,表数:数]
衣服	[词性:n, 数量名:是,个体量词:件|套,…]
心胸	[词性:n, 数量名:否,…]
……	……

俞士汶等，《现代汉语语法信息词典详解》（第2版），北京：清华大学出版社，2003年2月

附录1

“一件衣服”的分析结果



$$\begin{bmatrix} \text{短语类型} = np \\ \text{内部结构} = \text{定中} \\ \text{定语} = \begin{bmatrix} \text{短语类型} = qp \\ \text{内部结构} = \text{定中} \\ \text{定语} = \begin{bmatrix} \text{短语类型} = m\text{p} \\ \text{内部结构} = \text{单词} \\ \text{词语} = \text{"一"} \end{bmatrix} \\ \text{中心语} = \begin{bmatrix} \text{词类} = q \\ \text{内部结构} = \text{单词} \\ \text{词语} = \text{"件"} \end{bmatrix} \end{bmatrix} \\ \text{中心语} = \begin{bmatrix} \text{短语类型} = np \\ \text{内部结构} = \text{单词} \\ \text{词语} = \text{"衣服"} \end{bmatrix} \end{bmatrix}$$

应用

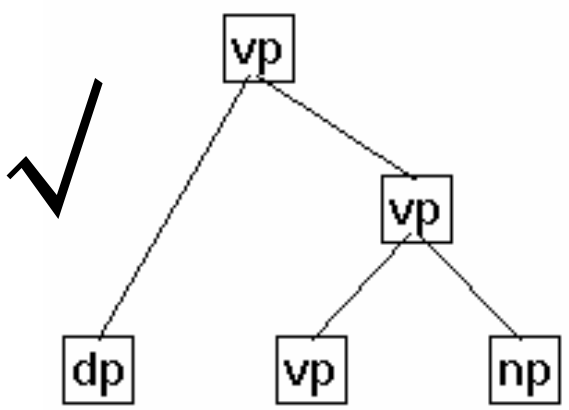
- 她买了一件衣服
- 董永拿走了七仙女的一件衣服
- * 一个心胸
- * 一个衣服
- * np[[一件衣服] [领子]]
- * np[[一件] [衣服领子]]

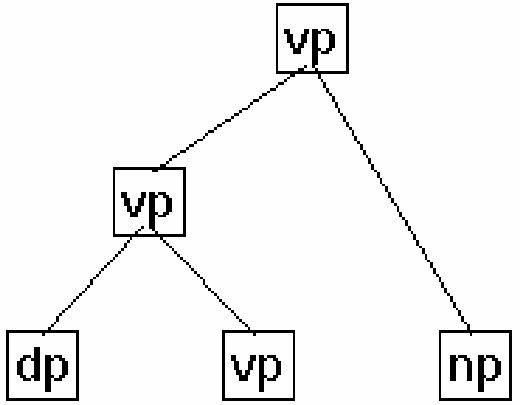
思考题

请为下面例子写出结构规则及合一约束条件

1. 一个心胸比较狭隘
2. 一个心胸宽广的人
3. 一件心胸宽广的人才会穿的衣服

3.2 伪歧义格式的消歧：“dp vp np”格式的分析（方案I）





vp_zz → dp vp_sb

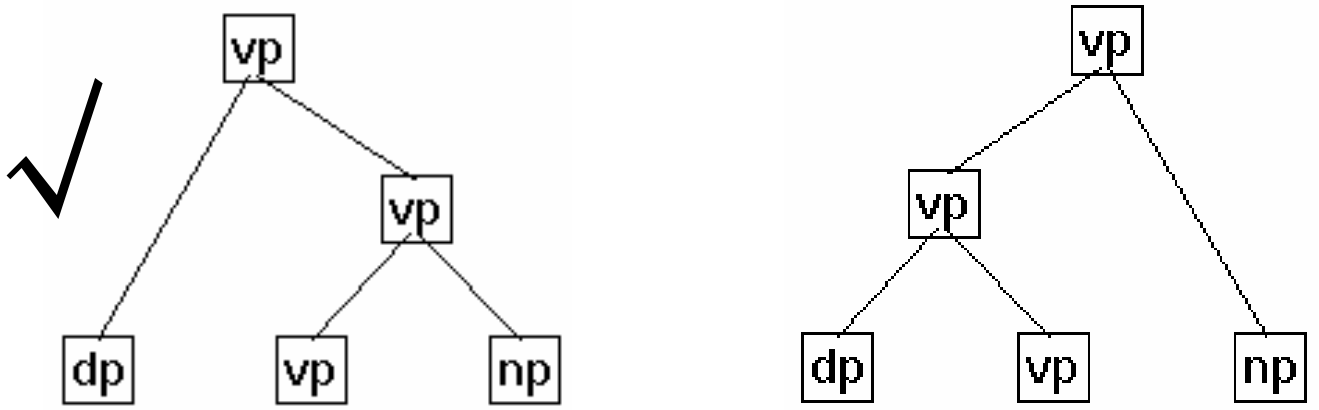
vp_sb → vp np

在vp内部分出两个非终结符（小类）：

vp_zz：状中式vp

vp_sb：述宾式vp

“dp vp np”格式的分析（方案II）

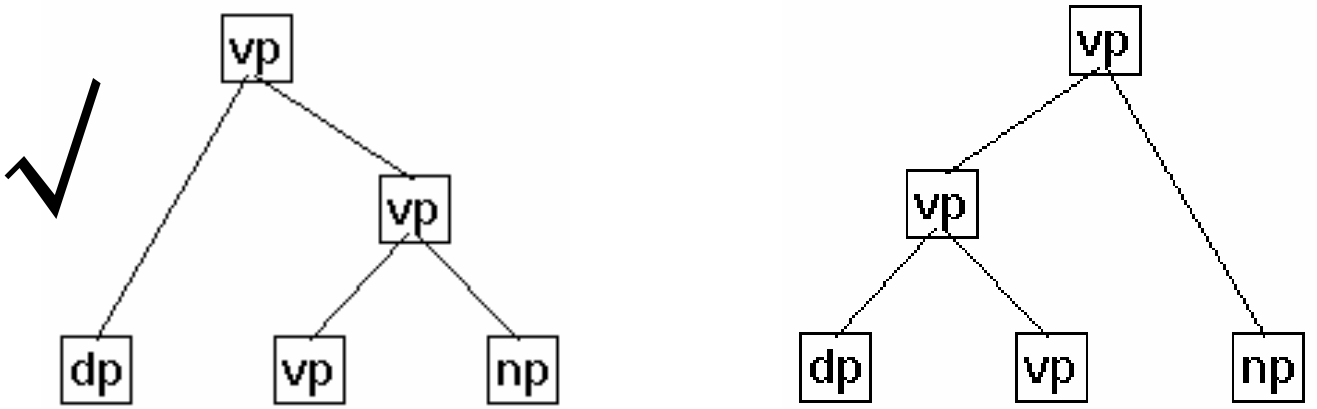


vp → dp vp :: $.内部结构=状中

vp → vp np :: $.内部结构=述宾,%vp.内部结构=~状中

根据“内部结构”特征值来进行约束

“dp vp np”格式的分析（方案III）



vp → dp vp :: $.内部结构=状中,$.daibinyu=否

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

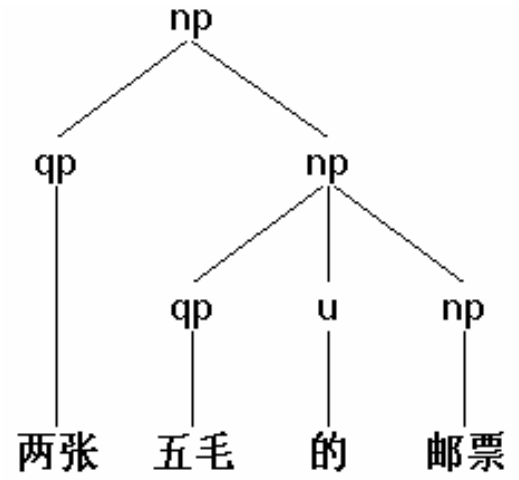
根据功能特征“daibinyu”（描述一个语言单位能否带宾语）来进行约束

这才是合理的描述方式！

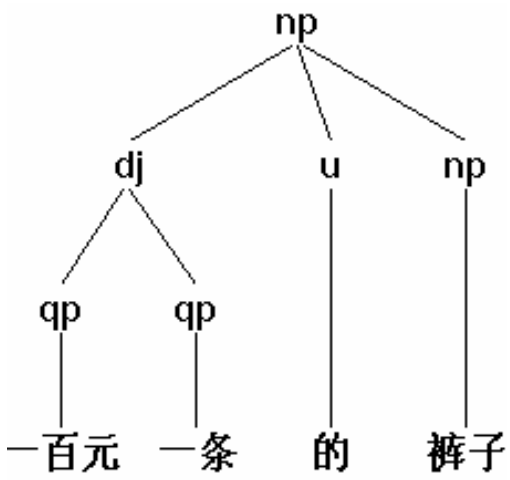
准歧义格式的消歧

- qp qp 的 np

两张 五毛 的 邮票



一百元一条的裤子

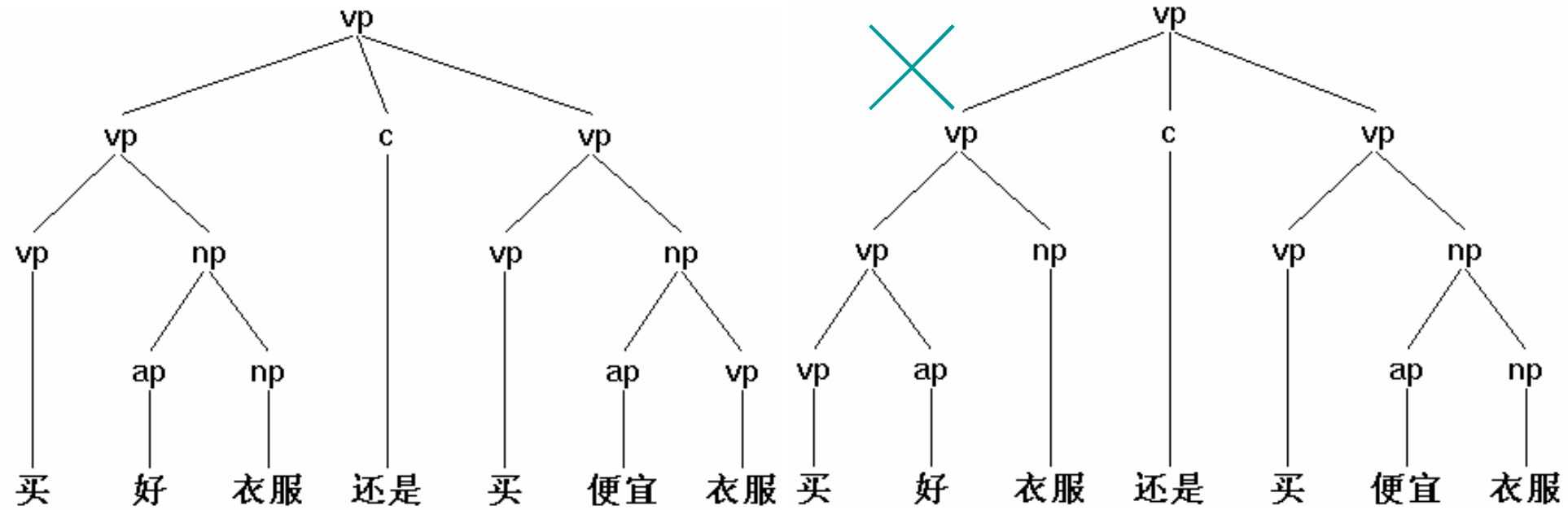


五十元一米的电缆

真歧义格式在并列结构中的消歧

v a n

- 你打算买好衣服还是买便宜衣服



3.3 内部结构特征与外部功能特征（示例I）

vp → dp vp :: $.内部结构=状中

述宾结构vp可以被dp状语修饰

是不是所有的述宾结构vp，都能被dp状语修饰？

1a. 相信 上帝

→ 1b. 曾经 相信上帝

2a. 相信 不 相信 上帝

→ 2b. 曾经 相信 不 相信 上帝 ？

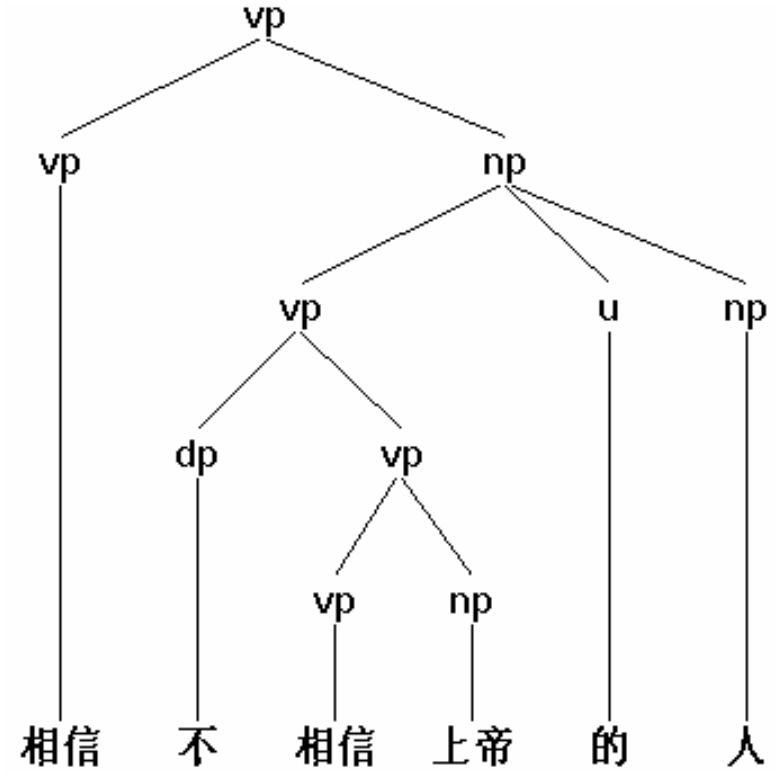
3a. 相信 不 相信 上帝 的 人

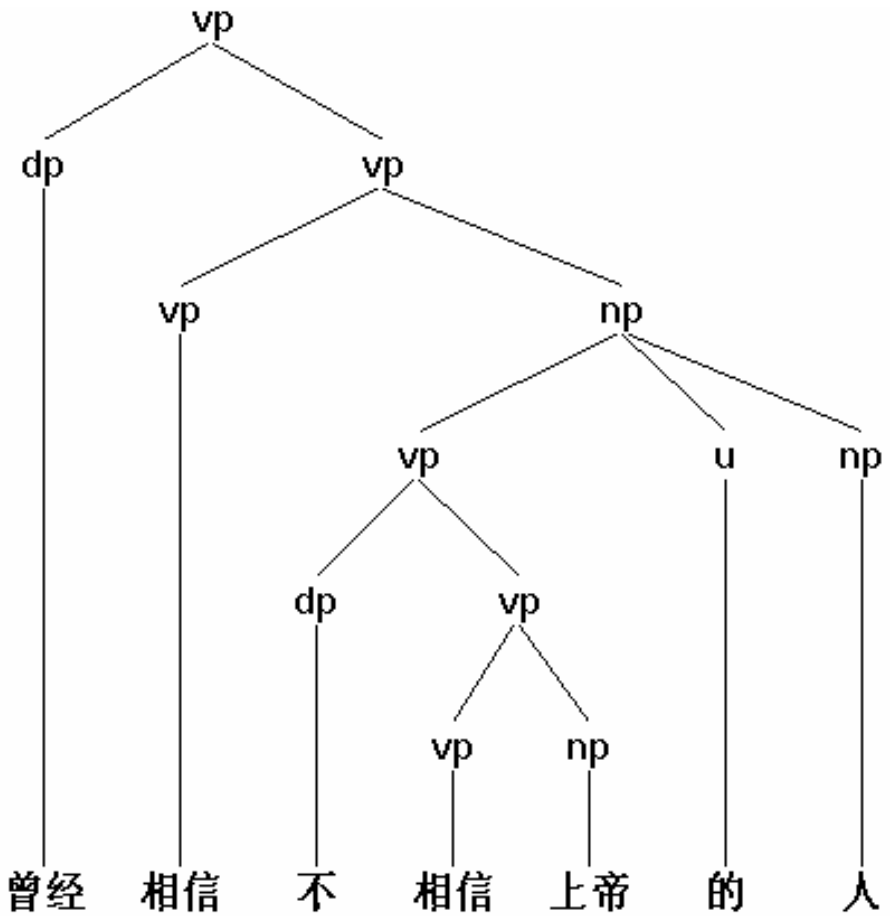
3b. 曾经 相信 不 相信 上帝 的 人

3a,3b能不能说？

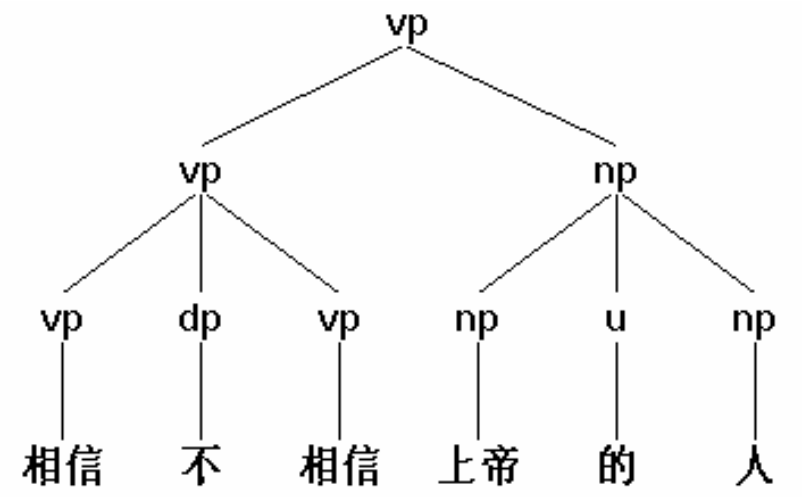
结构树怎么画？

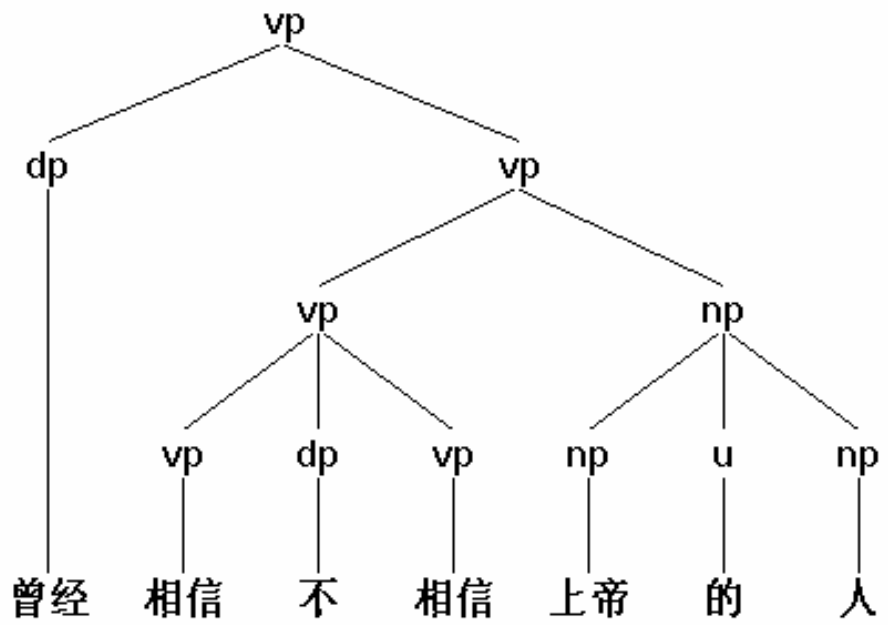
内部结构特征与外部功能特征（示例I）



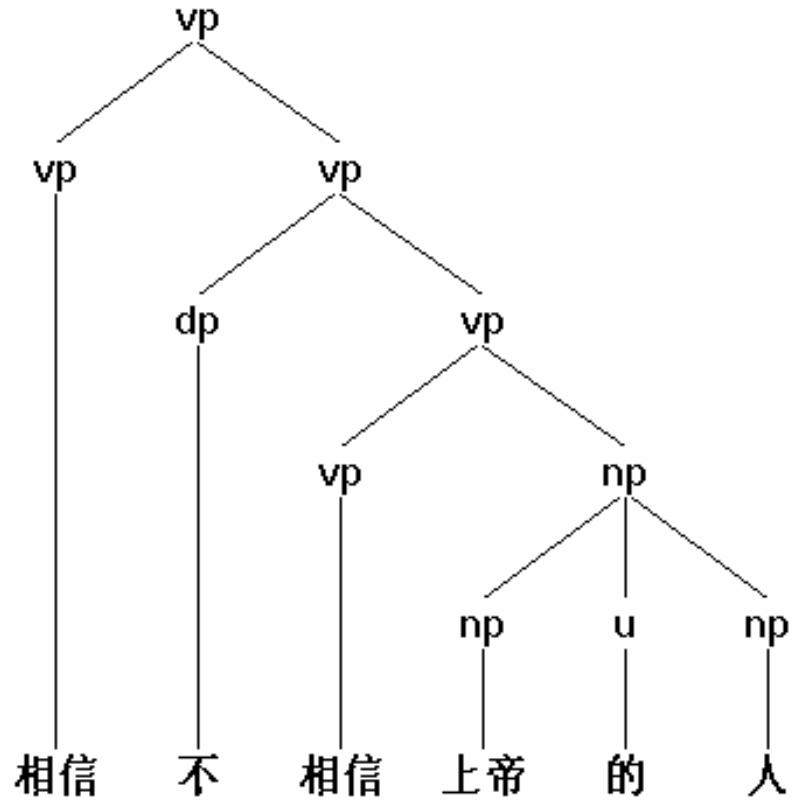


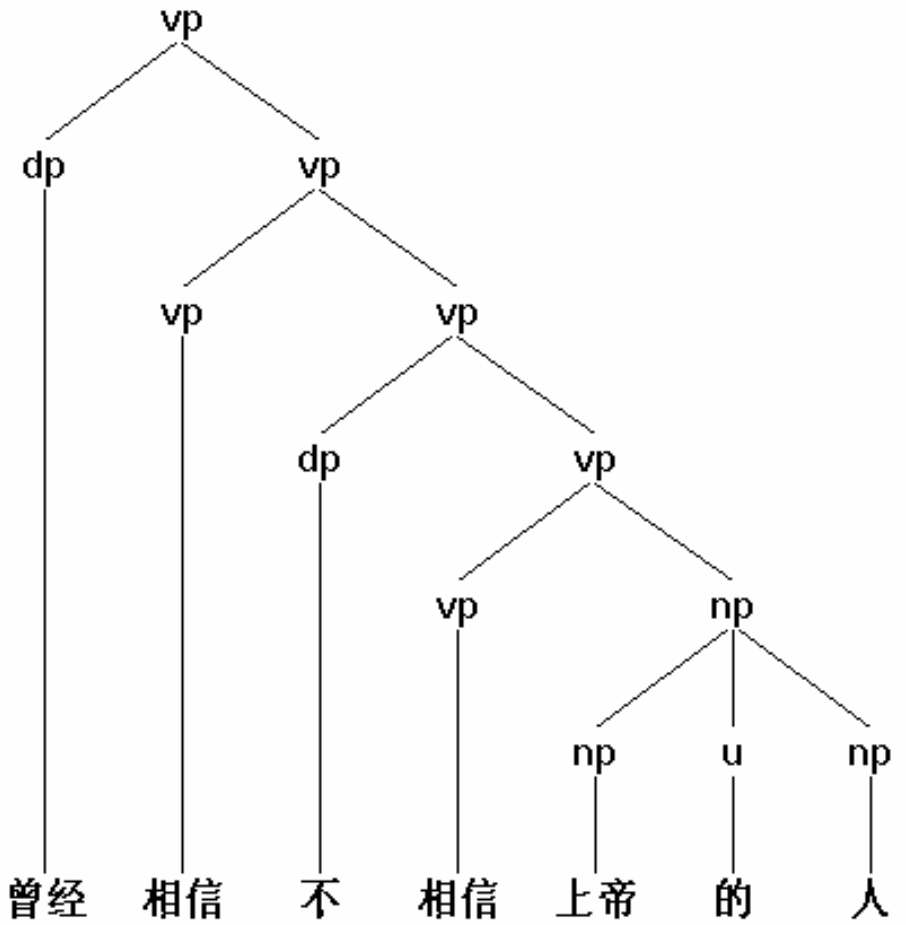
内部结构特征与外部功能特征（示例I）





内部结构特征与外部功能特征（示例I）





内部结构特征与外部功能特征（示例II）

状中结构ap

很 好
更 好
不 好
更 不 好

ap → dp ap :: $.内部结构=状中

这些状中结构的功能特征是否都一样？

（张三）比李四 好 ——— 比李四 很好 ×

——— 比李四 更好

——— 比李四 不好 ×

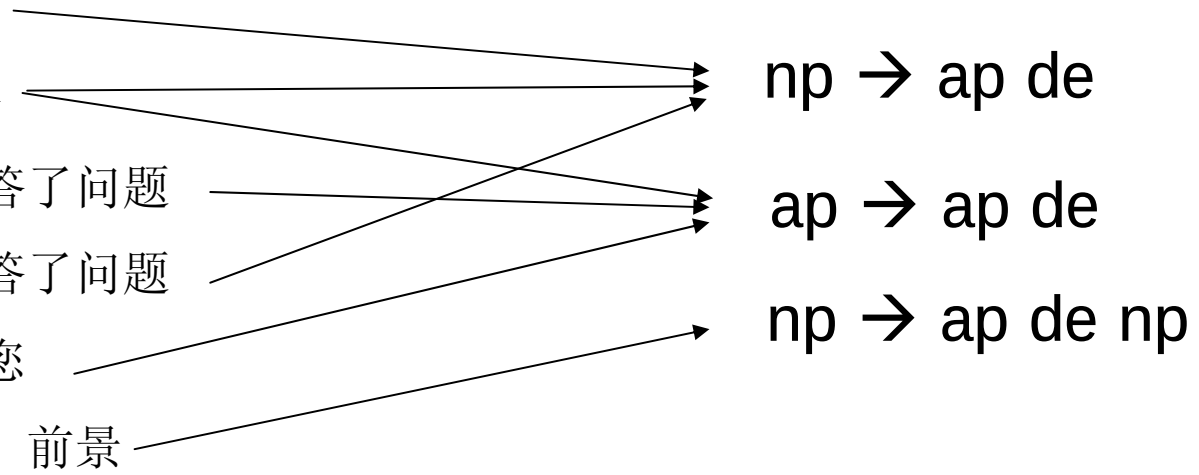
——— 比李四 更不好

在“比x”之后的位置，状中结构ap显示出分布差异

内部结构特征与外部功能特征（示例III）

“ap + 的”结构　　这些“ap+de”结构的外部功能是否相同 ？

1. 红de　给您
2. 认真 de　回答
3. 很认真 de 回答了问题
4. 最认真 de 回答了问题
5. 遗憾 de 通知您
6. 多么美妙　de　前景



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

小结

任何一个形式文法系统G，都必须回答：

1) 为了描述一个自然语言L，需要多少个基本范畴C_i？

2) 需要在$\{C_i\}$上定义多少种结构关系R_i？

3) 任给两个范畴C_m，C_n，它们之间是否有关系R_x？

4) C_m与C_n之间形成关系R_x，需要满足什么条件？

5) C_m与C_n组成的范畴$C'_{(m,n)}$，具有什么样的性质？具体要考虑：
哪些性质是从C_m，C_n继承来的？
哪些是C_m，C_n原本没有，而在$C'_{(m,n)}$中新增的功能属性？
哪些功能属性是C_m，C_n原本有，而在$C'_{(m,n)}$中反而没有的？

小结（续）

- 本来没有，后来新增的功能

* 很 有　　——　很 有理想
* 渠道 拓展市场 ——　多渠道 拓展市场

- 本来有，后来丧失的功能

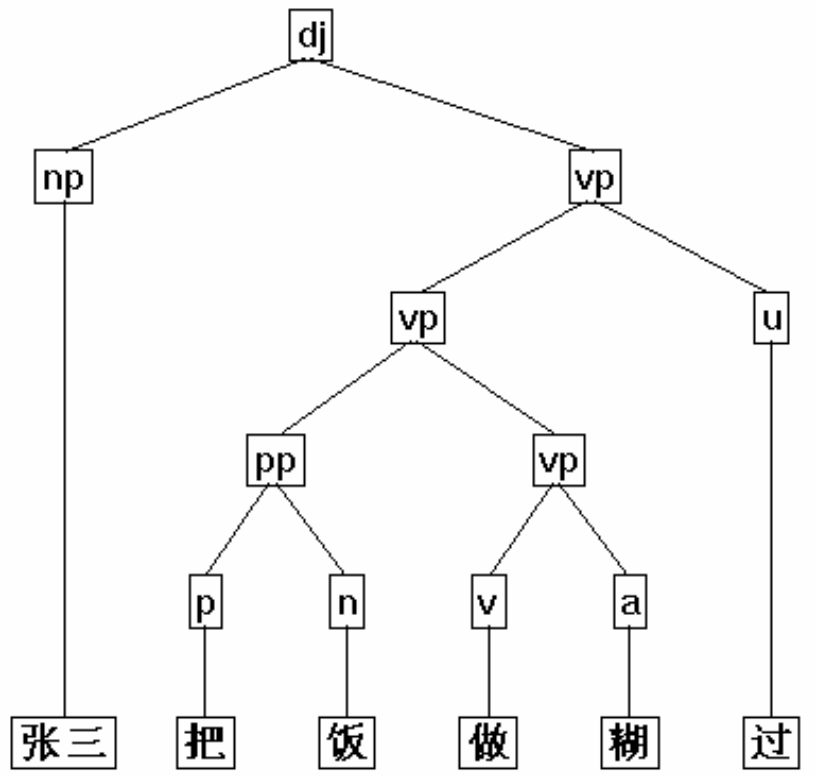
比李四 好　——　* 比李四 很好
比李四 高　——　* 比李四 不高

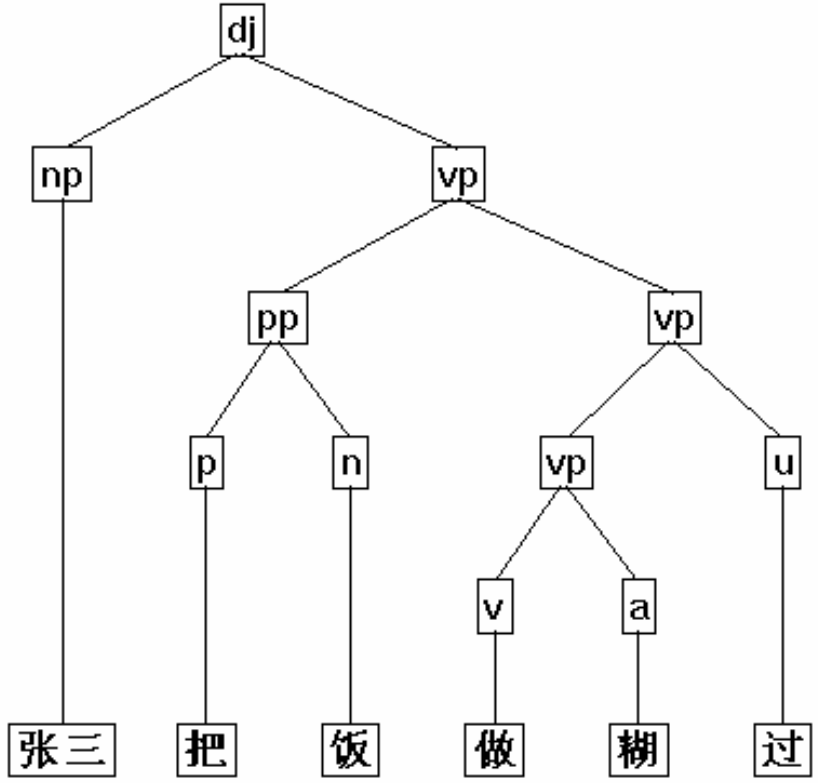
思考题

a. 张三 把 饭 做 了
b. 张三 把 饭 做 过
c. 张三 把 饭 做糊 过

（1）能说不能说？

（2）能说的句子，句法结构树怎么画？


dj
np
vp
vp
u
pp
vp
p
n
v
a
张三
把
饭
做
糊
过


dj
np
vp
pp
vp
p
n
vp
u
v
a
张三
把
饭
做
糊
过

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。

复习思考题

汉语的np短语包含哪些组合类型？请写出相应的规则及约束条件。

附录1：“现代汉语语法信息词典”简介

北大计算语言所 + 北大中文系联合研制

1. 词典概要
2. 知识表达范式：词类 + 属性描述
3. 词典规模

概要：面向语言工程的词语观

（1）类型：不拘一格

词/词组（短语），语素/单纯词，成语

（2）词条：严格筛选

“义项”与“语法功能”相结合的原则

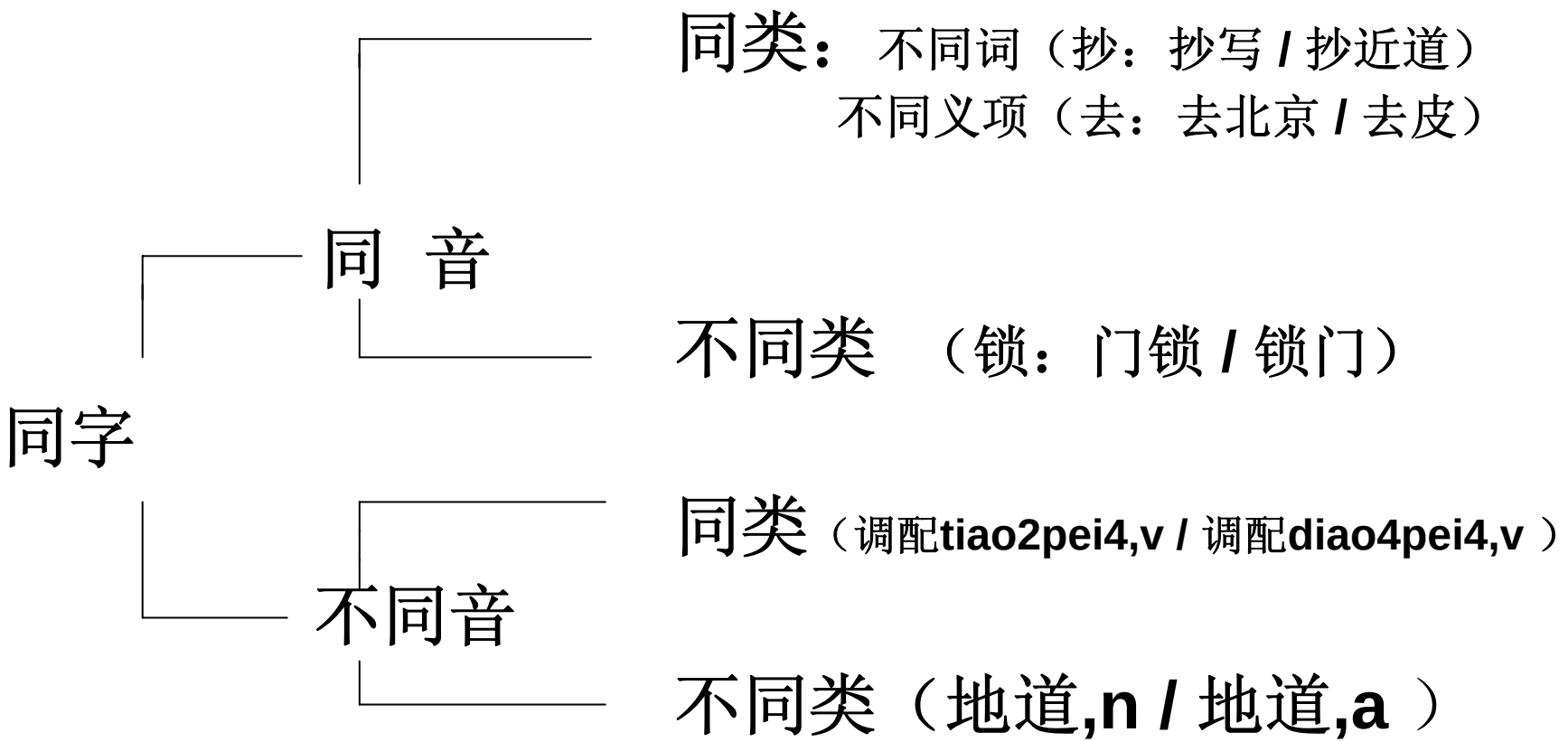
抄：不同词(抄写/抄近道）

去：不同义项（去北京 / 去皮）

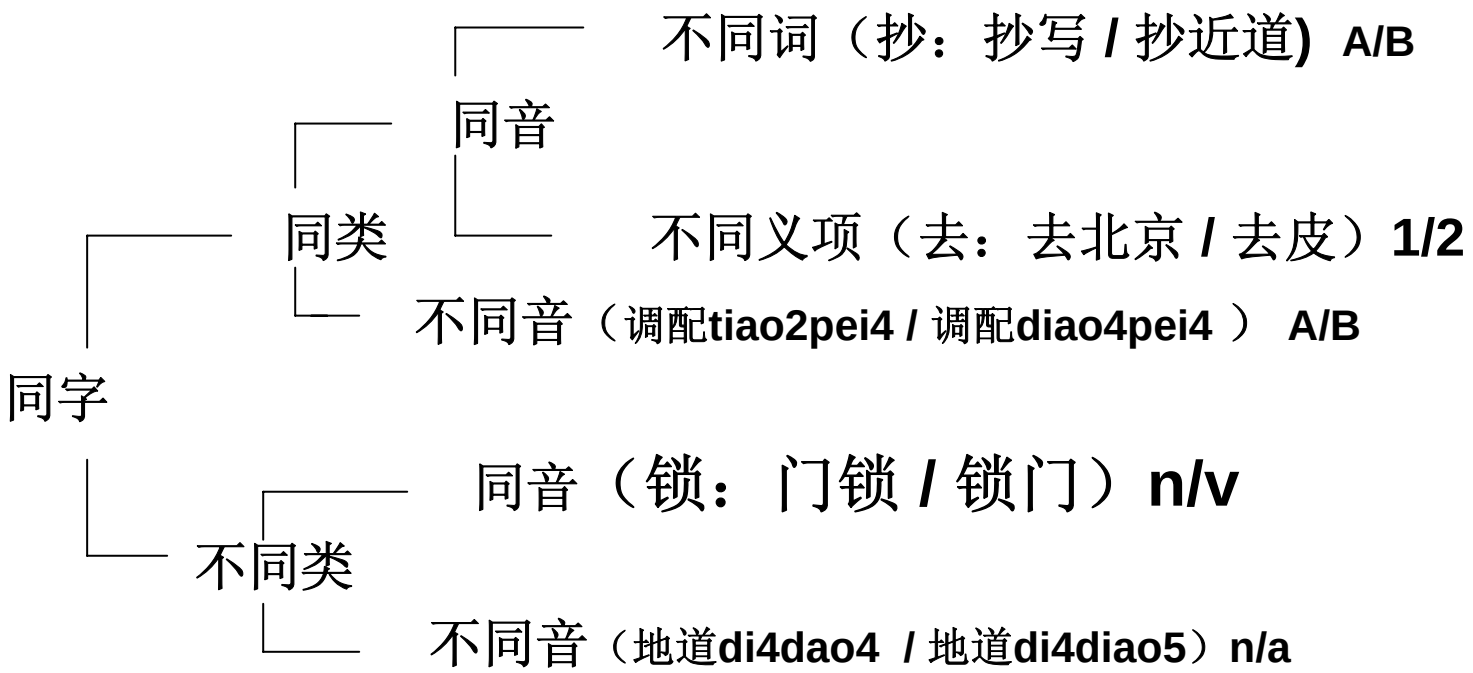
保管：两个义项，两个动词

材料：三个义项，一个词

同形词语的处理策略
（面向人的划分）



同形词语的处理策略
（面向机器的划分）



“词语”+“词类”+“同形”作为数据库的主关键字（Primary Key）

词语的归类

- 在本词典开发之前，众多的语法学家建立过各自的汉语词类体系，但都是以典型的例词说明各自体系的合理性、科学性。由于其研究是面向人的，又受限于当时的技术条件等原因，没有任何一位语言学家完成过数以万计的词语的归类。《现代汉语词典》只将数量很少的词（主要是虚词）归了类。
- 在汉语语法研究史上，本词典第一次完成了 **7** 万多词语的归类。到目前为止，还没有第二家。

词语的属性描述 (1)

从理论上来说，对于集合S上的一个分类P恒对应集合S上的一个属性表A，两者定义同样的等价关系。因此，分类与属性描述是可以相互转换的。

如果为所描述的对象确立m个二值属性（$m \geqslant 1$），则最多可将对象的集合划分为2^m个不相交的子集。反之，若将对象的集合划分为M（$M \geqslant 2$）个不相交子集，则至少要确立$[\log_2(M-1)+1]$个不同的二值属性（这里的方括号代表取整运算）。

词语的属性描述 (2)

由于划归同一类的词语仍有相互区别的属性特征，继续细分会造成分类体系庞杂，难以适应不断发现新的属性特征的研究过程，属性描述是恰当的策略。

《现代汉语语法信息词典》最重要的设计思想是在分类的基础上详细描述属于同一类的每个词语的详细的语法属性。

这个设计思想也成为北大计算语言所后来开发其他语言知识库建设的指导原则。

从动词库中抽取的部分语法属性字段

词语	同形	义项	系词	助动	趋向	体谓准	双宾	单作补	复数主	后名	很	着了过	重叠	离合	兼类
保存						体				可		着了过	ABAB		
成为			系			体									
得到						体准						了过			
告诉						体谓	双					了过			
协商						体谓			复	可		了			
加以						准									
冒险										可		过	VVO	离	a
去	A1	除掉				体						了过	VV		
去	A2	~上海			趋	体		可				了过	VV		
去	B	扮演				体						了过			
应	A	答应						可				了			
应	B	应该		助		谓									
支持	1	支撑				体						着了过			
支持	2	鼓励并帮助				体谓准					很	着了过	ABAB		
指挥						体谓				可		着了过	ABAB		n

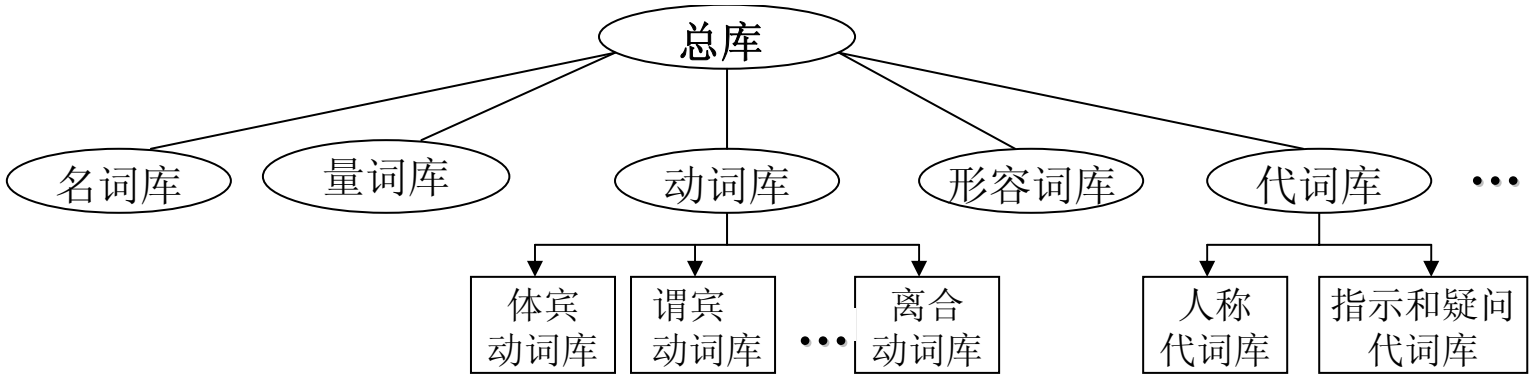
数据库的总体结构与规模

每一个数据库文件都刻画了词语及其属性的二维关系。

词典中共有34个数据库文件。总库1个，各类词库25个。其中，代词库下又设2个分库，动词库下设6个分库。

各类词的特有属性填在各类词库中。所有词的共同属性则容纳在总库中。总库中的属性包括读音、词类、拼音、虚实、体谓等，共计约20项。

所有的库都可以进行连结（JOIN），连结条件可以用“词语”+“词类”+“同形”主关键字表达。这样，34个库文件构成有上下位关系的“树”，子节点继承父节点的全部信息，或者说，将父节点与子节点连结起来就可以得到词语的更全面的信息。



库 名	记录数	属性 字段数
总库	73877	13
名词	35201	31
时间词	565	16
处所词	183	15
方位词	194	21
数词	165	26
量词	456	24
区别词	757	13
代词	205	19
人称代词分库	49	8
指示代词分库	157	15
动词	14496	47
体宾动词分库	7630	27
谓宾动词分库	1321	8
双宾动词分库	185	12
动结式分库	3178	10

库 名	记录数	属性 字段数
动趋式分库	6195	32
离合词分库	3420	8
形容词	2857	33
状态词	986	18
副词	1174	22
介词	108	28
连词	203	15
助词	38	12
语气词	53	13
前接成分	11	9
后接成分	43	9
成语	5264	15
简称略语	400	14
习用语	3031	15
语素	7223	14
标点符号	52	17
总计		**579**

工作量

73,874个词语的总信息量为3,369,828。这些信息所需存储空间为29,000,686字节。

最后粗略地计算一下这项工程所需要的人力。假设一个人确定一个属性信息只需要2分钟，填写3,369,828个信息需要6,739,656分钟，约等于112327.6小时。如果每天工作8小时，则需要14040.95天。按现在每年的工作日约为250天计算，折合的结果为56年，即完成这项工程至少需要投入56个人年的工作量**(16年，每年4个人）**。

附录2：一个汉语规则描述语言的Backus-Naur范式

<规则> ::= && <规则名> <产生式> [:: <约束>]

<规则名> ::= "{" <标识符> "}"

<产生式> ::= <短语类标记> → <短语结构>

<短语结构> ::= <结构成分> { <结构成分>}

<结构成分> ::= [<中心成分标识>] <语类标记>

| [<中心成分标识>]<语类标记> "<"<终结符>">"

<中心成分标识> ::= !

<语类标记> ::= <词类标记> | <短语类标记>

<词类标记> ::= n|v|a|t|s|……

<短语类标记> ::= np|vp|ap|tp|……

一个汉语规则描述语言的BNF（续）

<约束> ::= <约束式> {,<约束式>}
<约束式> ::= <合一等式>
| <条件测试句>
<条件测试句> ::= IF <合一等式> {,<合一等式>} TRUE
| IF <合一等式> {,<合一等式>} FALSE
| IF <合一等式> {,<合一等式>} THEN <合一等式> {,<合一等式>} ENDIF
| IF <合一等式> {,<合一等式>} ELSE <合一等式> {,<合一等式>} ENDIF
| IF <合一等式> {,<合一等式>} THEN <合一等式> {,<合一等式>} ELSE
<合一等式> {,<合一等式>} ENDIF
<合一等式> ::= <内部变量>=<原子>
| <内部变量>=<内部变量>
| <外部变量>=<外部变量>
| <外部变量>=<True>
| <外部变量>=<False>
| <外部变量>=<None>
| <外部变量>==<外部变量> <属性项目> {<属性项目>}
| <外部变量>!=<外部变量> <属性项目> {<属性项目>}
<外部变量> ::= <根结点标记>
| <位置标记><语类标记>
| <外部变量>.<外部属性>
……