“现代汉语语法信息词典”简介

詹卫东
zwd@pku.edu.cn

提 纲

1. 词典概要
2. 知识表达范式：词类 + 属性描述
3. 词典结构与规模

- 北大计算语言所 + 北大中文系联合研制
- 俞士汶 等（1998，2003）《现代汉语语法信息词典详解》（第1，2版），清华大学出版社、广西科学技术出版社1998年版。

1 概要：面向语言工程的词语观

（1）类型：不拘一格

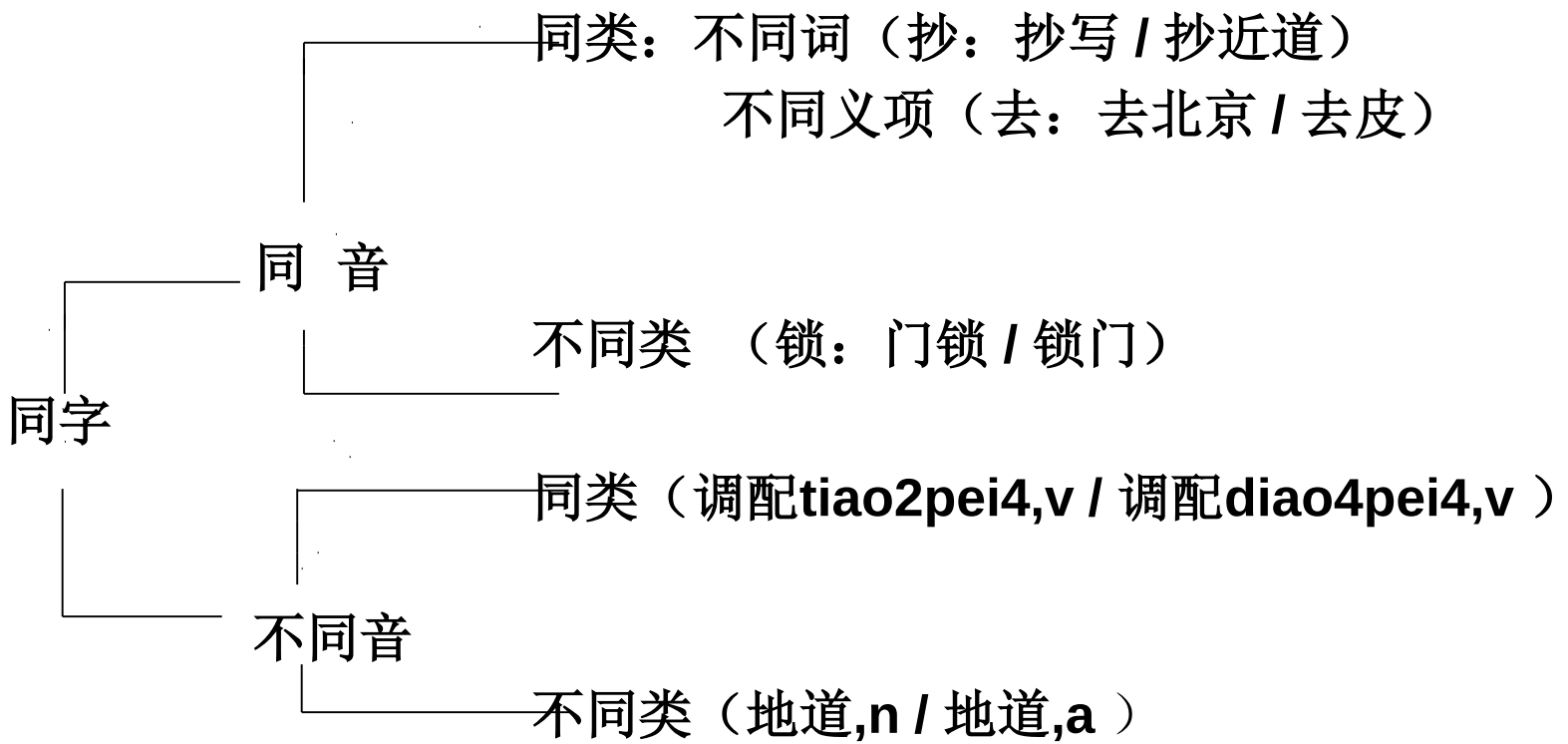
词/词组（短语），语素/单纯词，成语

（2）词条：严格筛选

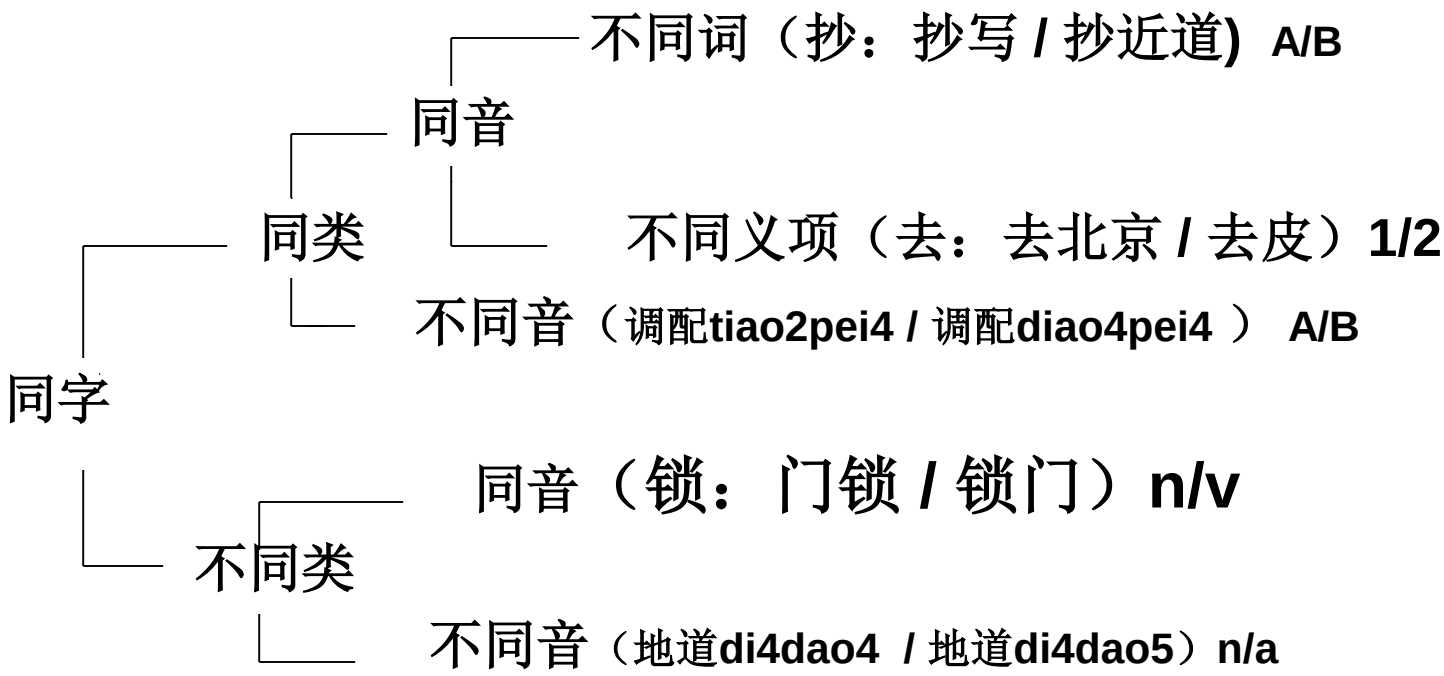
“义项”与“语法功能”相结合的原则

- 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）

2 知识表达范式：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。

知识表达范式：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（$m \geq 1$），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（$M \geq 2$）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）。

知识表达范式：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

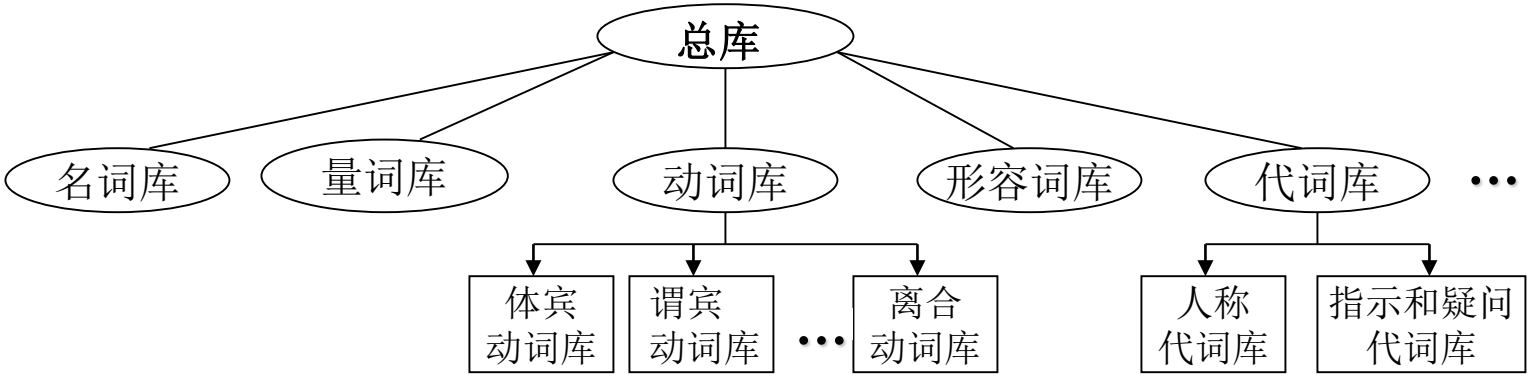
3 数据库的总体结构与规模

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。

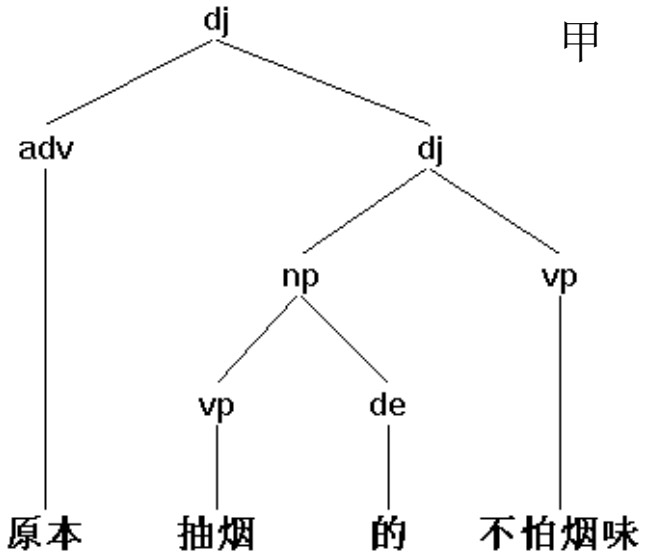
最后粗略地计算一下这项工程所需要的人力。假设一个人确定一个属性信息只需要2分钟，填写3,369,828个信息需要6,739,656分钟，约等于112327.6小时。如果每天工作8小时，则需要14040.95天。按现在每年的工作日约为250天计算，折合的结果为56年，即完成这项工程至少需要投入56个人年的工作量**(16年，每年4个人**）。

词典信息还需要扩充

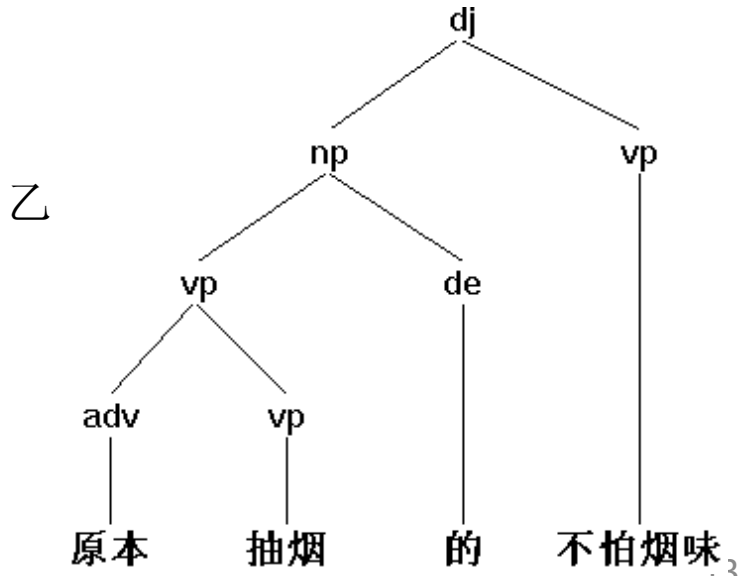
例1　　adv　　vp　　的　　vp

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



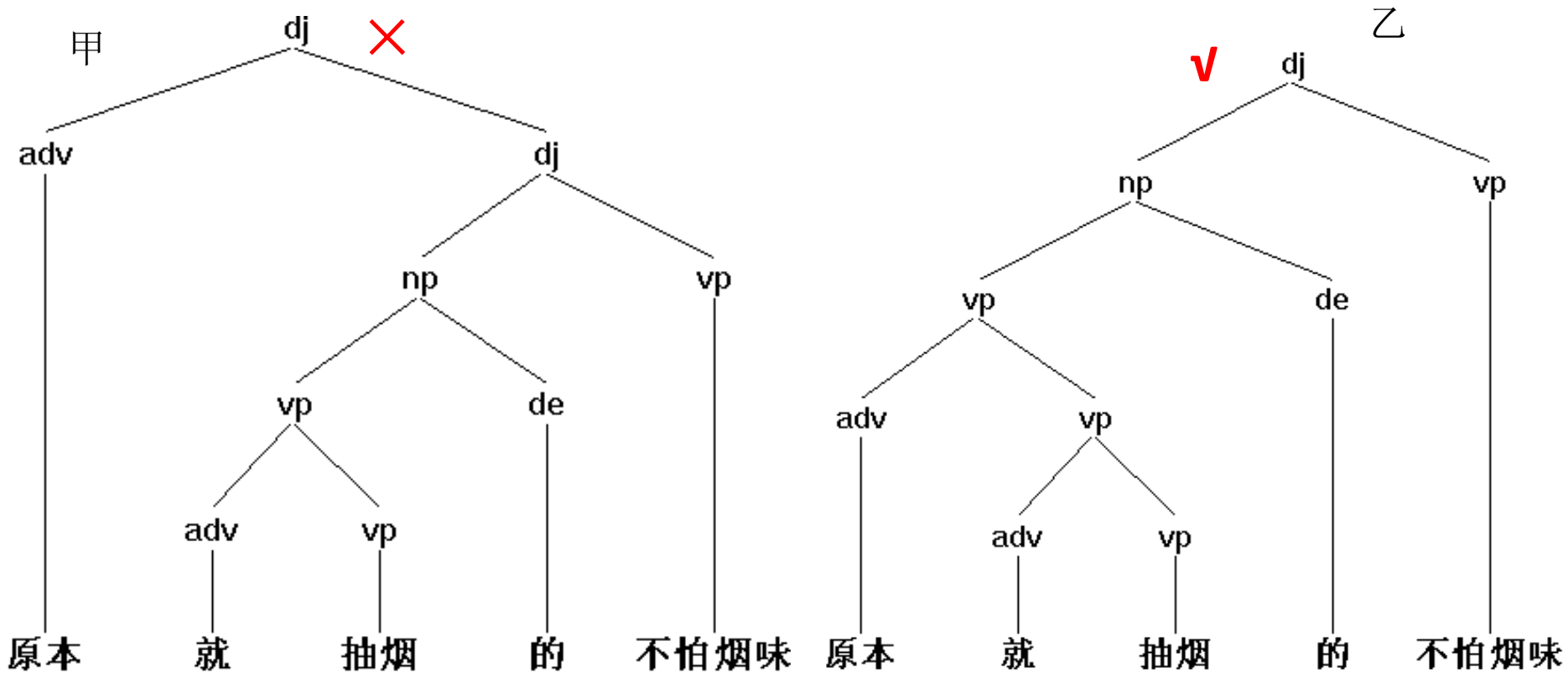
乙



为什么多了一个副词就没歧义了呢？

例1’　adv　adv　vp　　的　　vp

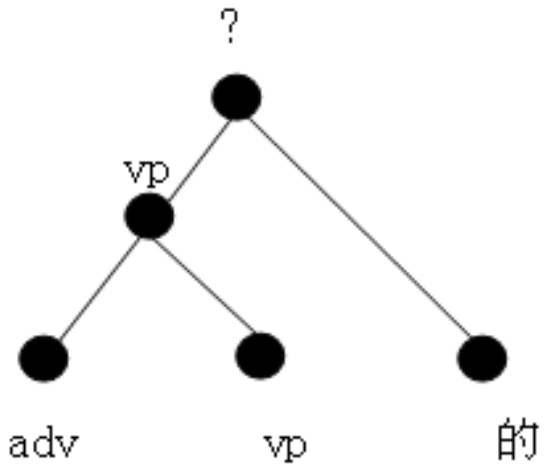
4　原本　就　抽烟　的　不怕烟味



现在的知识粒度够细吗？

（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

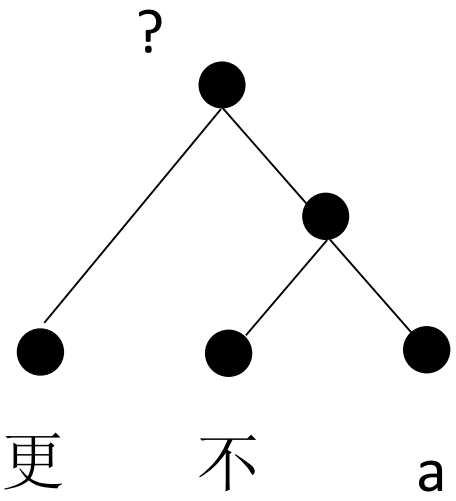
（2）《现代汉语语法信息词典》中没有“副词 + V”后能不能再加“的”的特征描述



例2：形容词“好”跟“高”分布不同

不 好 —— 不 高

更 不 好 —— * 更 不 高



张三 比 李四 更 不 好

* 张三 比 李四 更 不 高

（1）《现代汉语语法信息词典》中**有**形容词能否受“不”修饰的描述
（2）《现代汉语语法信息词典》中**没有**“不+形容词”后能否再受“更”修饰的描述