一种短语结构制导的范畴表达式演算

白　硕
国家智能计算机研究开发中心
1998年12月

背景与动机

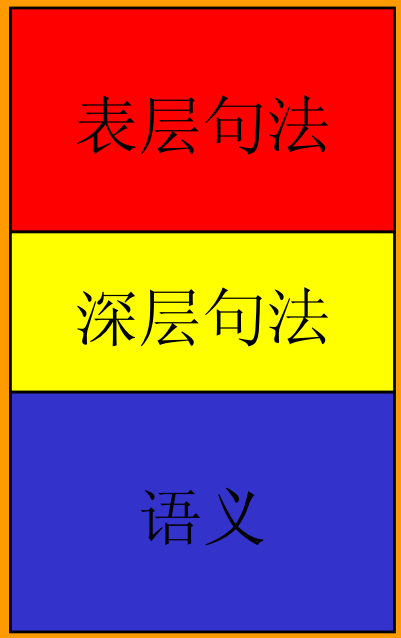
- 自然语言理解、信息提取
- 计算语言学研究的现状：句法理论与语义不匹配，汉语尤甚
- 汉语中的著名例句：
  - 鸡不吃了
  - 台上坐着主席团
  - 王冕六岁上死了父亲
  - 圆圆的画了一个圈

问题的所在

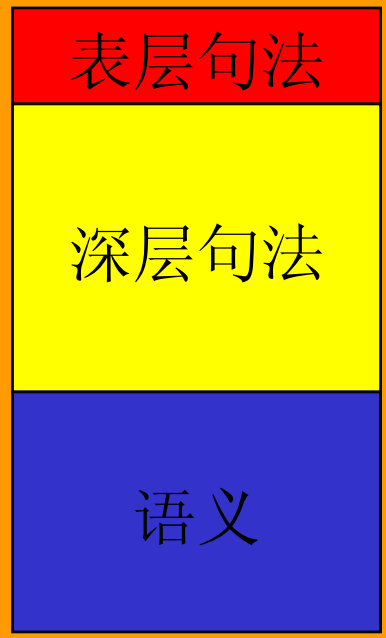
- “超距相关”，对表层句法结构的语感和语义成分的相关关系不吻合
- 这种不吻合不依赖于所采用的表层句法结构理论
- 这种不吻合说明存在着一种“深层的”句法结构，它既是句法的，又是与语义结构相一致的

汉语与英语的对比

英语



汉语



“鸡不吃了”

- “鸡”是主语还是宾语？
- 为什么“鸡”的位置上既可以是“吃”的施事，也可以是“吃”的受事？
- 为什么“鸡”换成“张三”，“张三”就只能是施事？
- 为什么“鸡”换成“菜”，“菜”就只能是受事？
- 什么因素在影响对“谁吃”“吃谁”的理解？

“台上坐着主席团”

（1）主－谓－宾

（2）状－谓－宾

（3）状－谓－主

（4）宾－谓－主

- 语序之外，还有什么决定着成分之间的语义结合？
- 主谓宾等句法范畴对汉语有无意义？

“王冕六岁上死了父亲”

- 凭什么死的是“父亲”，不是“王冕”？
- 不及物动词如何可以带“宾语”？
- 为什么“父亲”是“王冕”的父亲？
- “王冕”在这个句子里充当的是什么角色？仅仅说明“父亲”是谁的父亲吗？这样的角色为什么可以放在我们通常叫做“主语”的位置上？

“圆圆的画了一个圈”

- “圆圆的”是定语还是状语？
- “圆圆的”到底修饰谁？
- 跨越成分的修饰结构为什么是可能的？
- 既然表层结构里没有直接结合的依据，是什么决定了“圆圆的”和“圈”之间的深层语义联系？

问题是与表层句法理论无关的

- 即使不承认主谓宾等句法概念，只要承认有“分析树”的结构，前面的问题依旧存在
- “分析树”结构无法刻划上面例子里说明的那些位置上出现那样的成分结合关系的原因

问题的实质

- 成分有“稳定感”上的差别
- “不稳定”的成分填入适当的其他成分就变得相对“稳定”
- 对要求填入的成分有选择性
- 对语言的理解过程就是从不稳定走向稳定的过程，不稳定是语言理解要排除的障碍，也是语言理解的动力

三种做法

- 基于“合一”的复杂特征集演算
- 基于配价的填充演算
- 基于继承和约分的范畴表达式演算
- 深层句法与语义的界限：
  - 只确定成分间的结合（匹配）关系
  - 不进行超出形式上的结合（匹配）关系以外的语义分析和表示

话说“范畴”

- 范畴**Category**
- 范畴语法**Categorial Grammar**
- 句法结构的构造↔范畴表达式的演算
- **concatnation** ↔范畴表达式的乘法
- **combination** ↔范畴表达式的约分

范畴表达式

- 基本范畴：一个集合**A**，其中至少有“*”
- 复合范畴：**a/b1,b2,…,bn**
- 范畴表达式：
  - 基本范畴是范畴表达式
  - 如果**a,b1,b2,…,bn**是范畴表达式，则复合范畴**a/b1,b2,…,bn**也是范畴表达式
  - 如果**a**是不含**#**的范畴表达式，则**#ia,a#i**也是范畴表达式，其中**i**是自然数或者是**‘max’**

太抽象了，来点具体的

- 比如，**A包括Human, Event**
- 动词“打”对应的范畴表达式可以写为

  Event/Human, Human

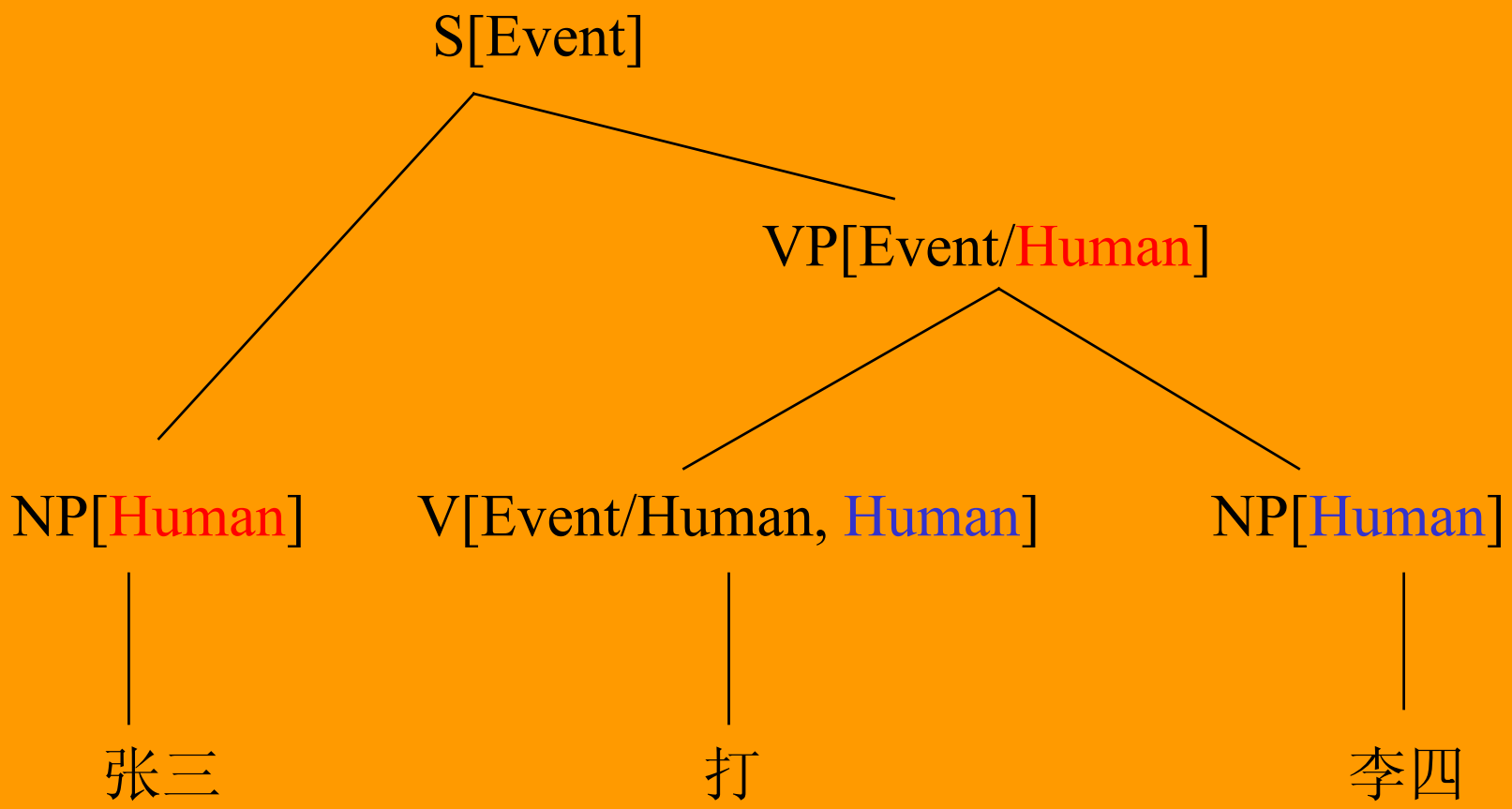
  “张三”、“李四”对应的范畴表达式可以写为**Human**
- “张三打李四”对应的范畴表达式是什么？怎么得来的？

范畴表达式：语言成分的量纲

张三	打	李四
~~Human~~	Event/~~Human~~, ~~Human~~	~~Human~~

约分关系就是成分结合关系！

短语结构如何“制导”


S[Event]
VP[Event/Human]
NP[Human]
V[Event/Human, Human]
NP[Human]
张三
打
李四

建构范畴体系的一般手续

- 确定基本范畴
- 确定继承关系，注意它有传递性
- 确定每个词对应的范畴表达式和每个短语结构（产生式规则）对应的范畴约束，它们都可能不是唯一的，若干个可能的范畴表达式和范畴约束用||号隔开，表示“或”逻辑关系

继承与约分

- 两个范畴表达式可以“约分”，如果一个的分子与另一个的分母中的一项之间有“继承”关系
- 比如，如果**Human→Animate**，则范畴表达式**Event/Human**和**Animate**之间就可以进行带继承的约分，其结果为**Event**，留下的痕迹为**Human**而不是**Animate**。

面向汉语的尝试

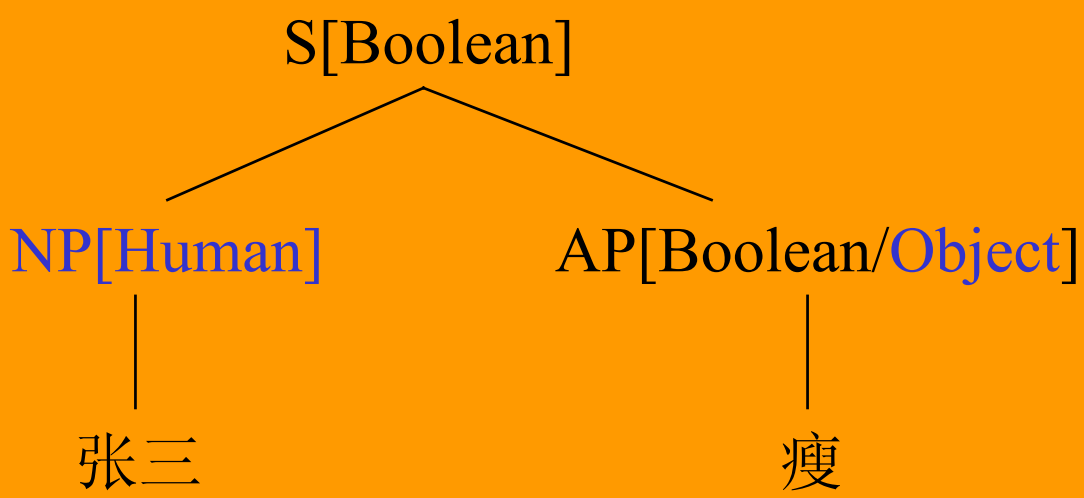
- 主谓结构
- 定中结构
- 述宾结构
- 述补结构
- 状中结构
- 连谓结构
- 的字结构
- 时体结构
- 把字结构
- 被字结构
- 介宾结构
- 数量结构

主谓结构

- 普通主谓结构
  - 规则：**S → NP VP|NP AP|NP S|NP NP**
  - 原则：
    - 如果**VP**的范畴表达式为**x/...,y,...,** **NP**的范畴表达式为**z**，而继承关系**y →z**或者**z →y**成立，则可以进行约分，约分的结果为**x/...,...**
    - 如果**VP**的范畴表达式的分母上有多的项与**z**有正向或反向的继承关系，则最左方的项最优先参加约分

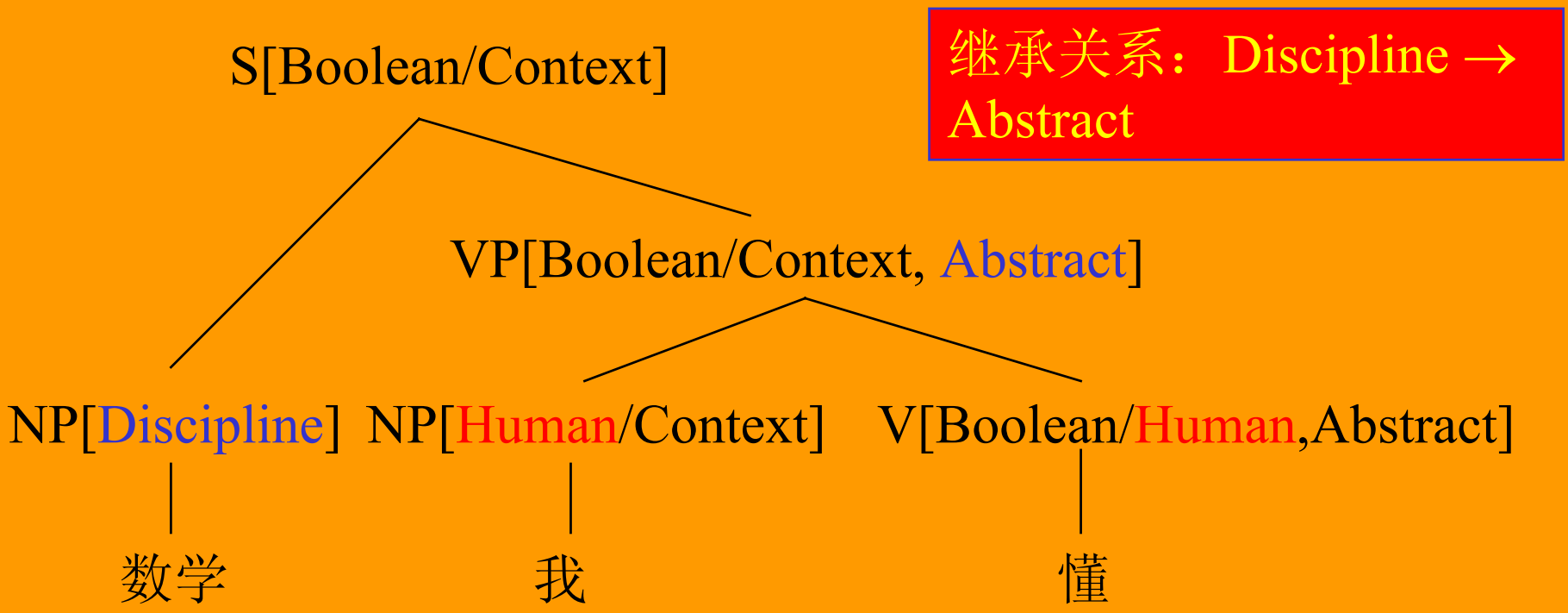
主谓结构

- “张三瘦”：普通主谓结构
  - 继承关系：**Human → Object**



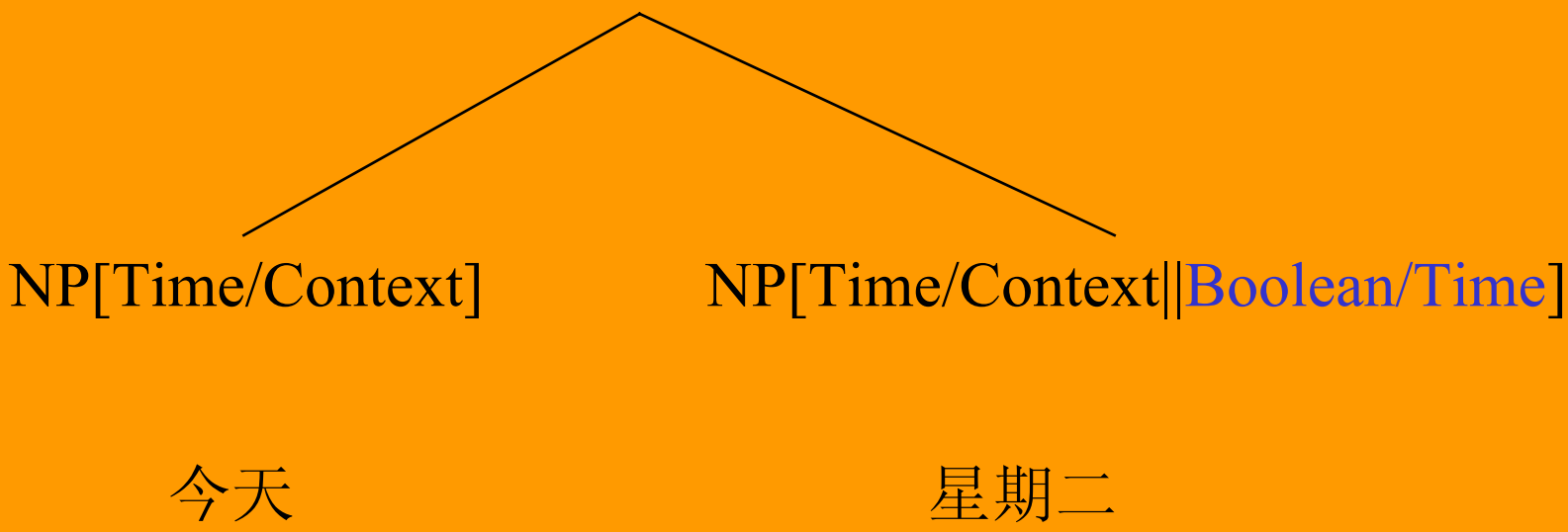
主谓结构

- “数学我懂”：主谓谓语型主谓结构



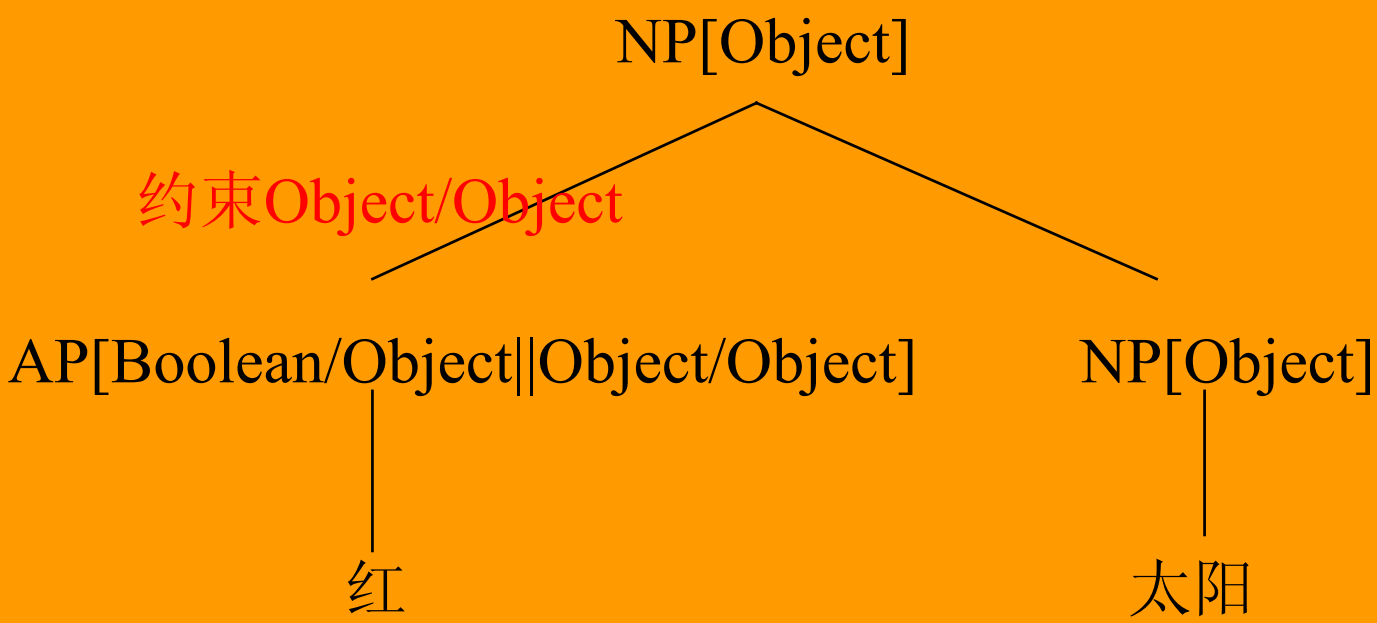
主谓结构

- “今天星期二”：名谓型主谓结构



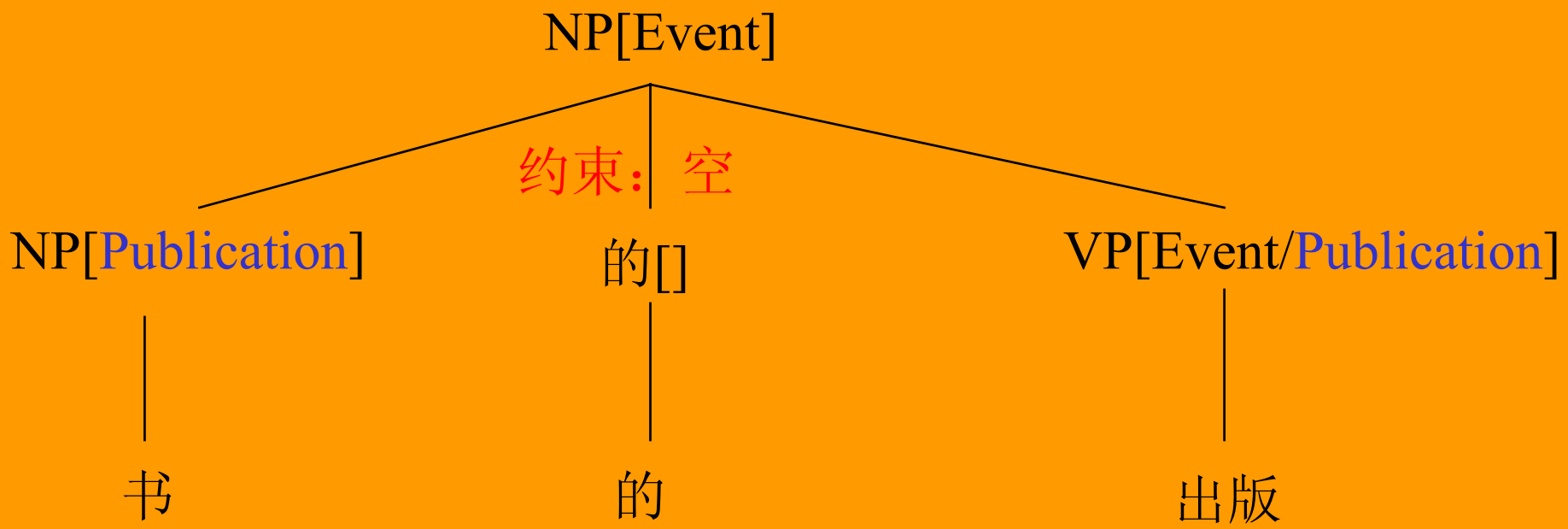
定中结构

- “红太阳”：一般定中结构 **NP →AP NP**



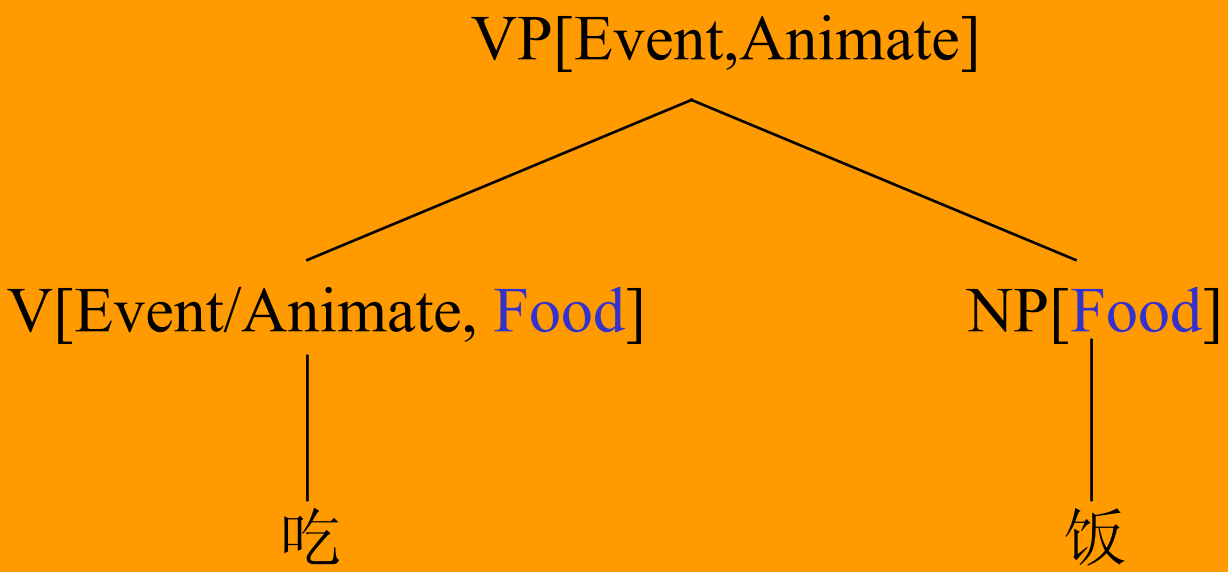
定中结构

- “书的出版”：名物化定中结构**NP→ NP的VP**



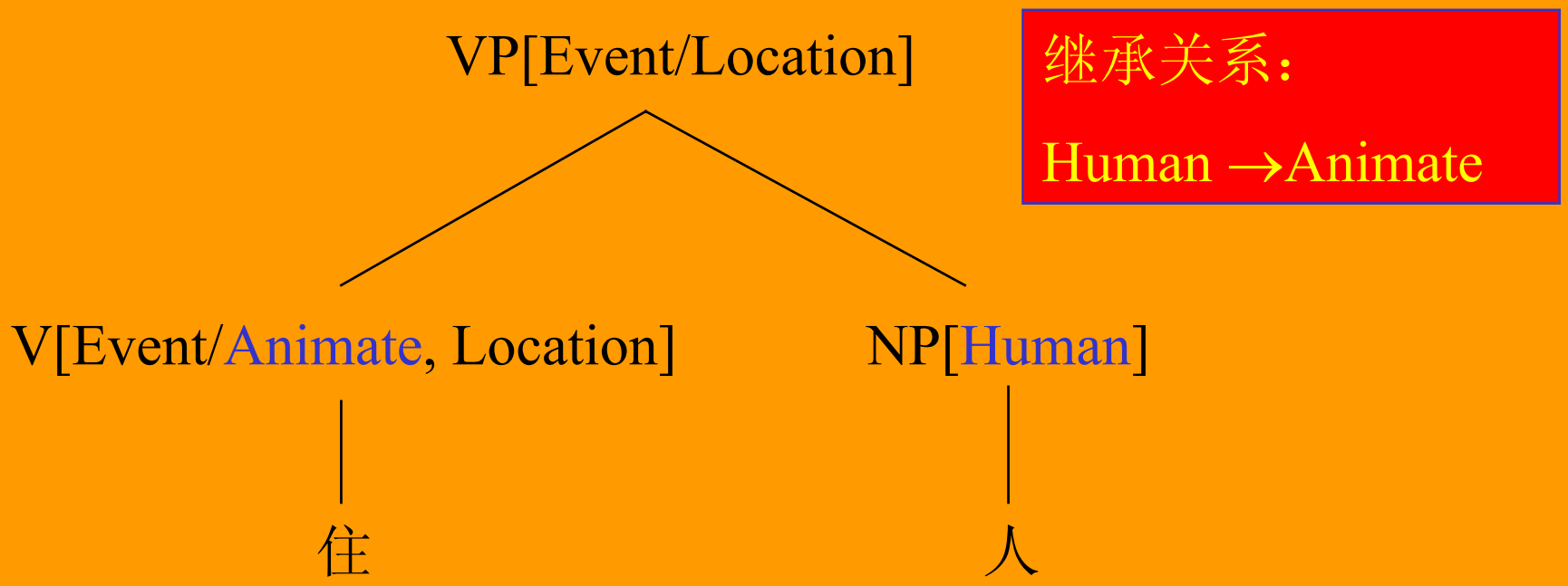
述宾结构

- “吃饭”：一般述宾结构**VP →V NP**



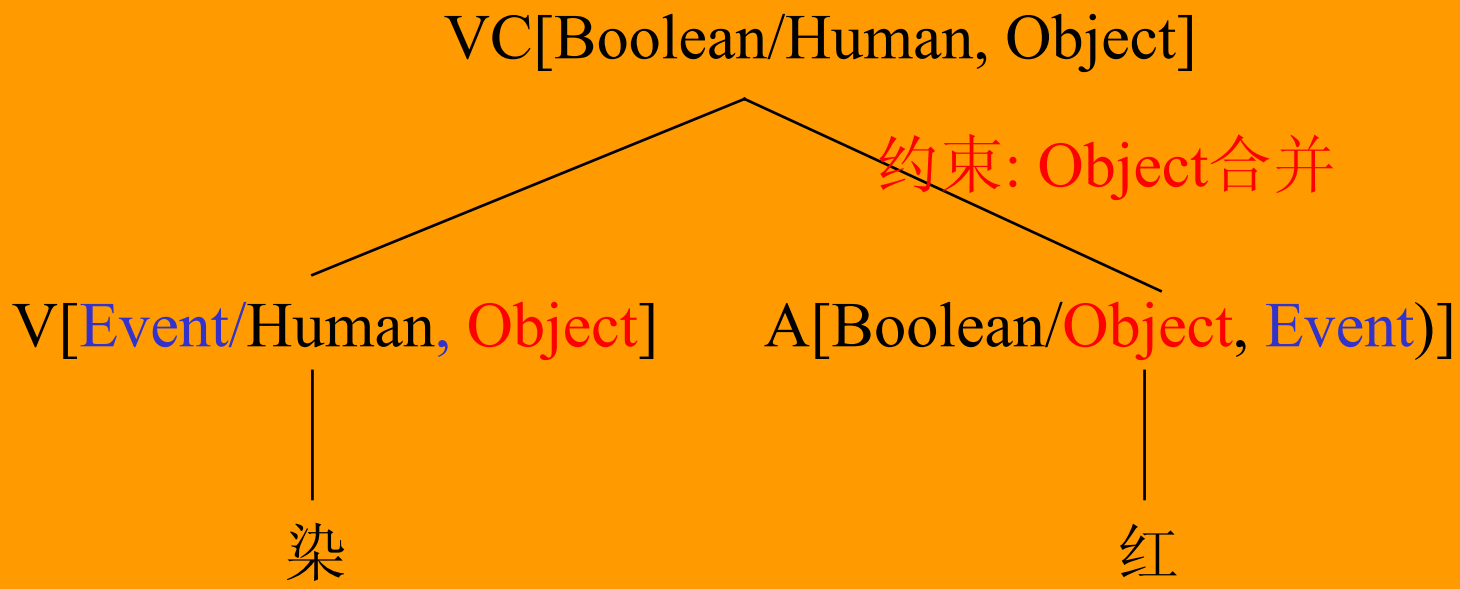
述宾结构

- “住人”：单向述宾结构**VP →V NP**



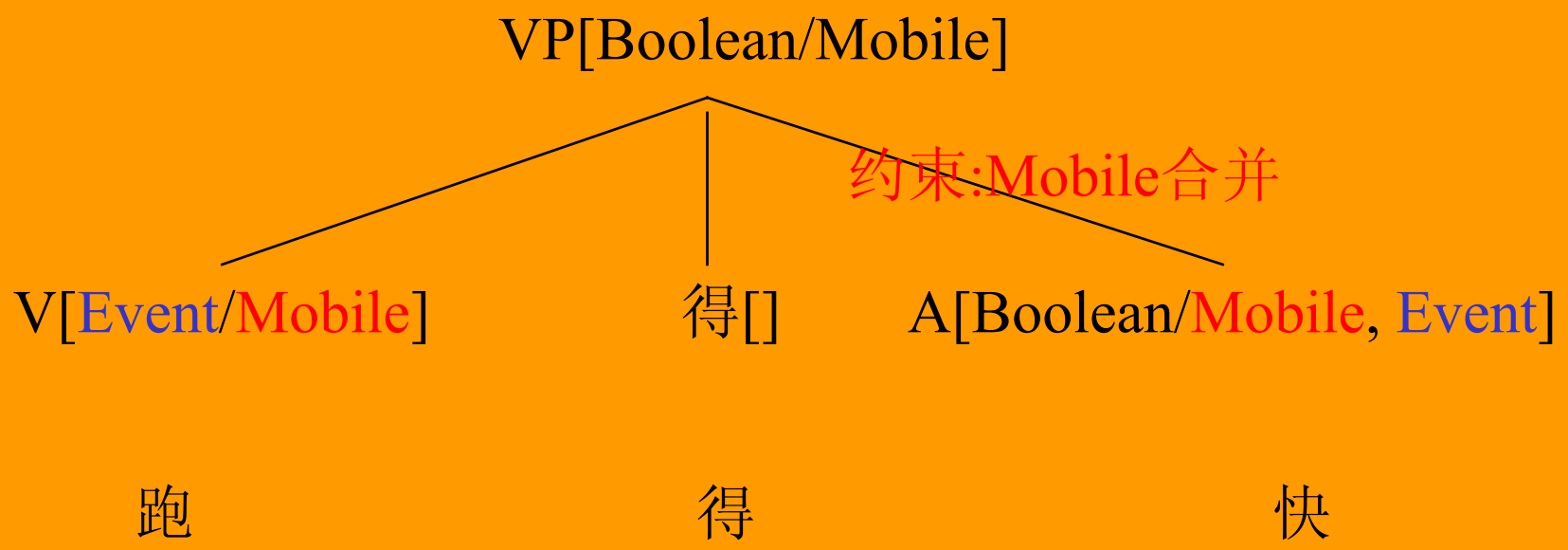
述补结构

- “染红”：动结(趋)式**VC →V V| V A**



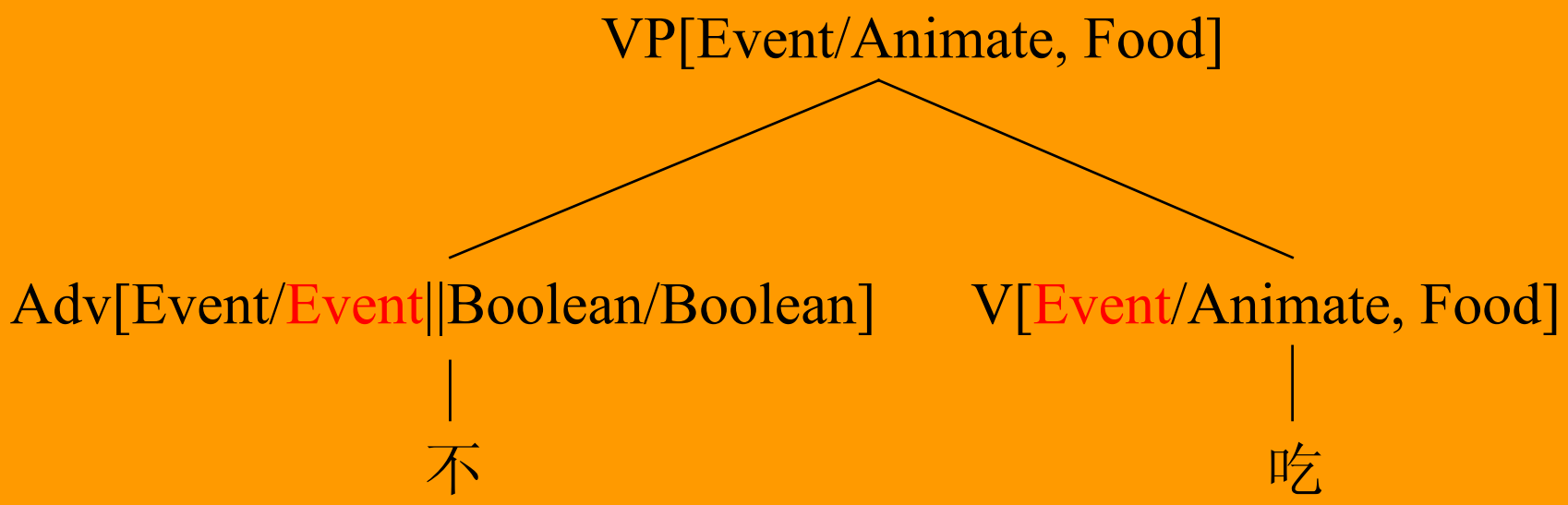
述补结构

- “跑得快”：得字结构**VP →V 得 A**



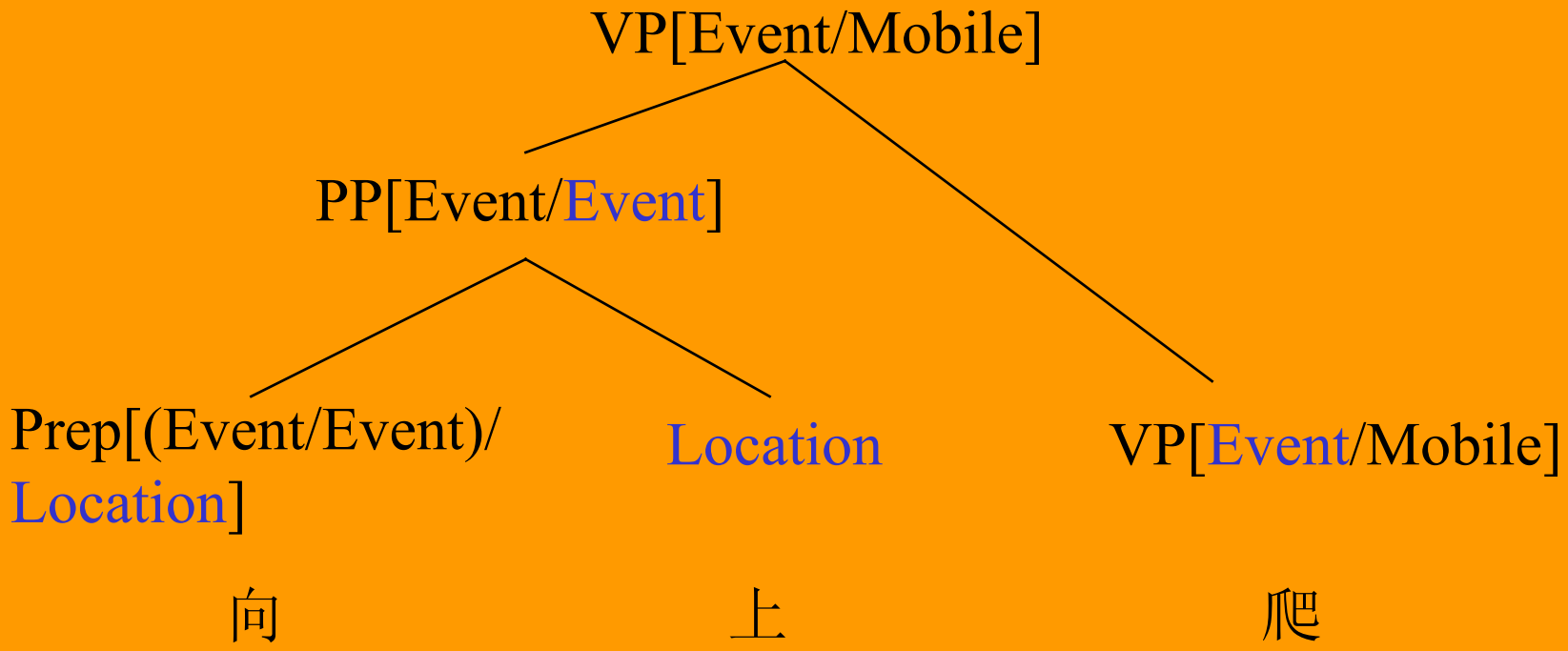
状中结构

- “不吃”：紧状中结构**VP→Adv V**



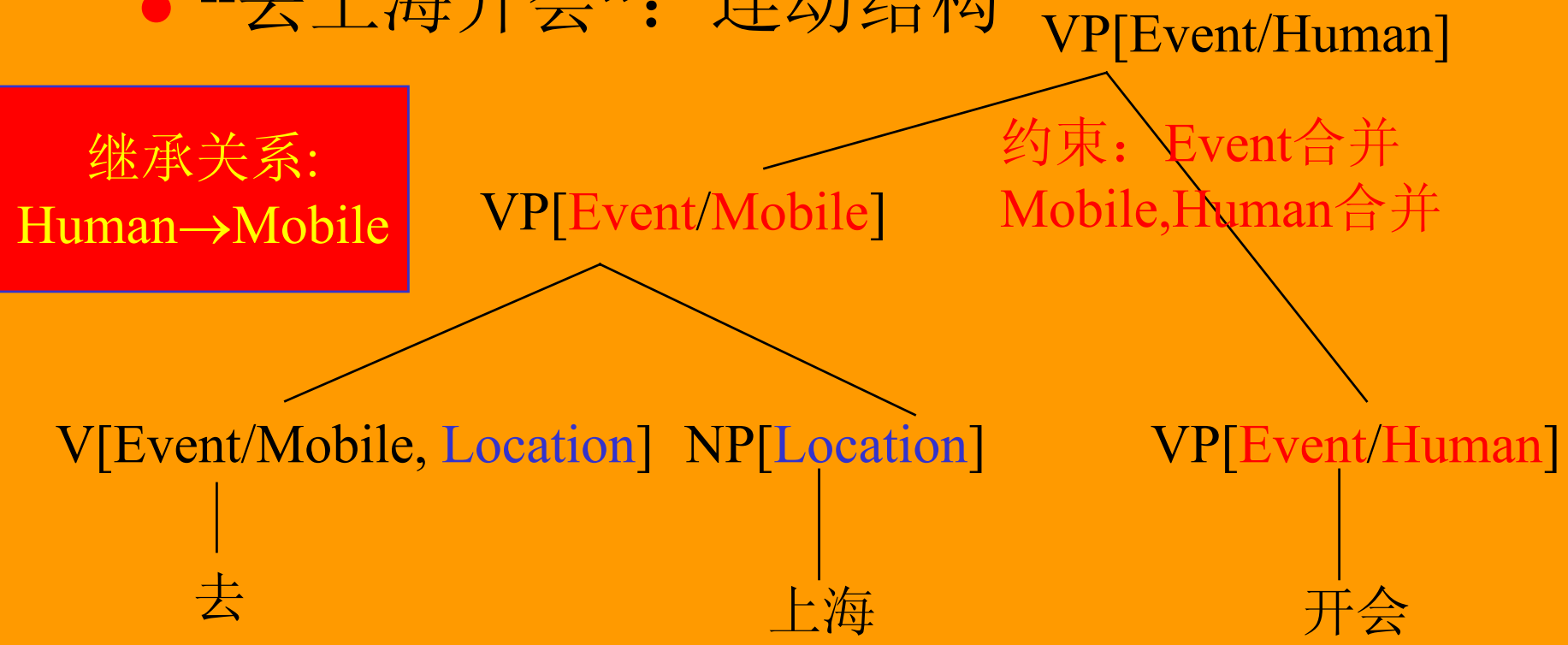
状中结构

- “向上爬”：松状中结构**VP→PP VP**



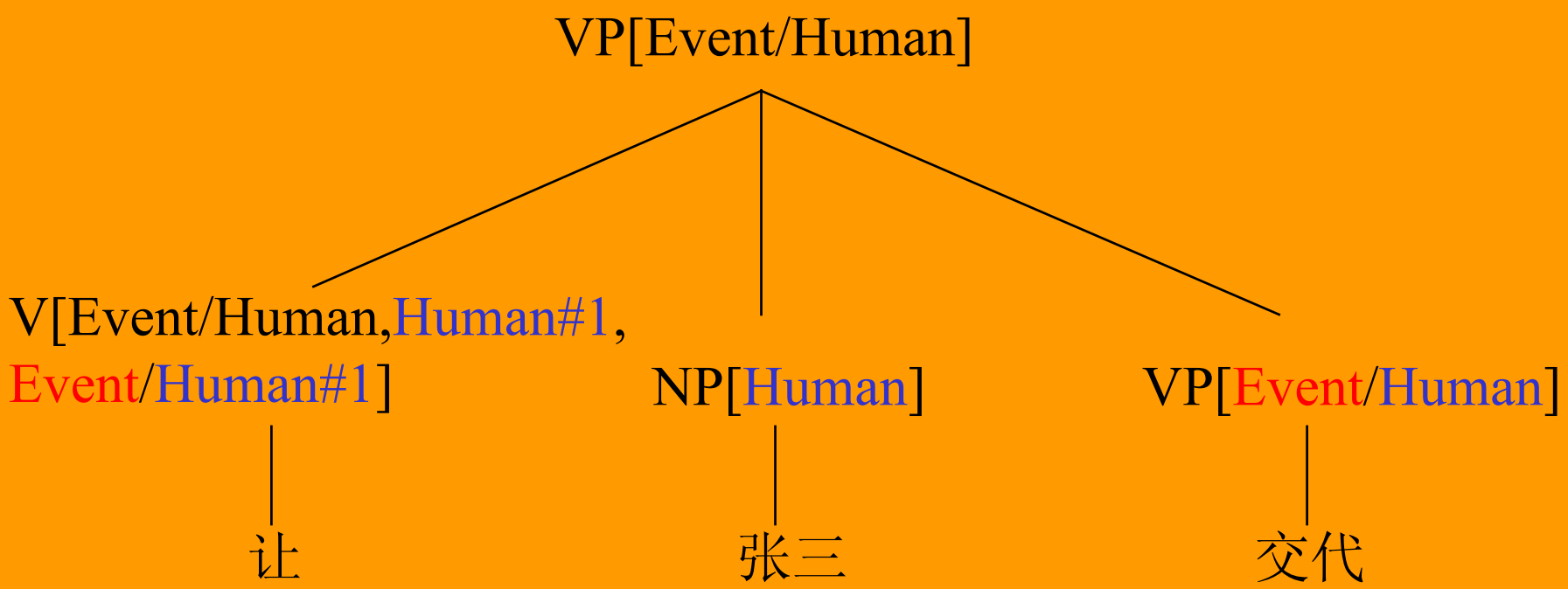
连谓结构

- “去上海开会”：连动结构



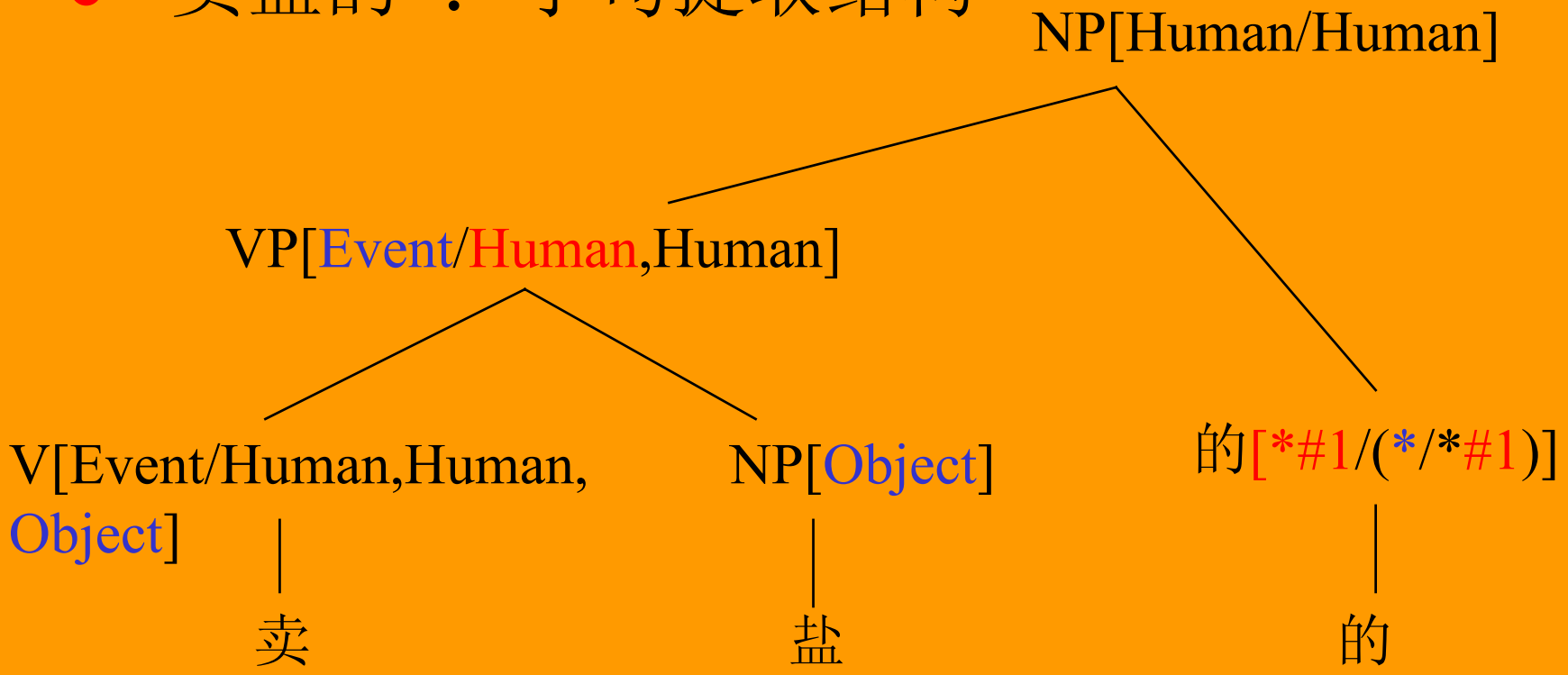
连谓结构

- “让张三交代”：递系（兼语）结构



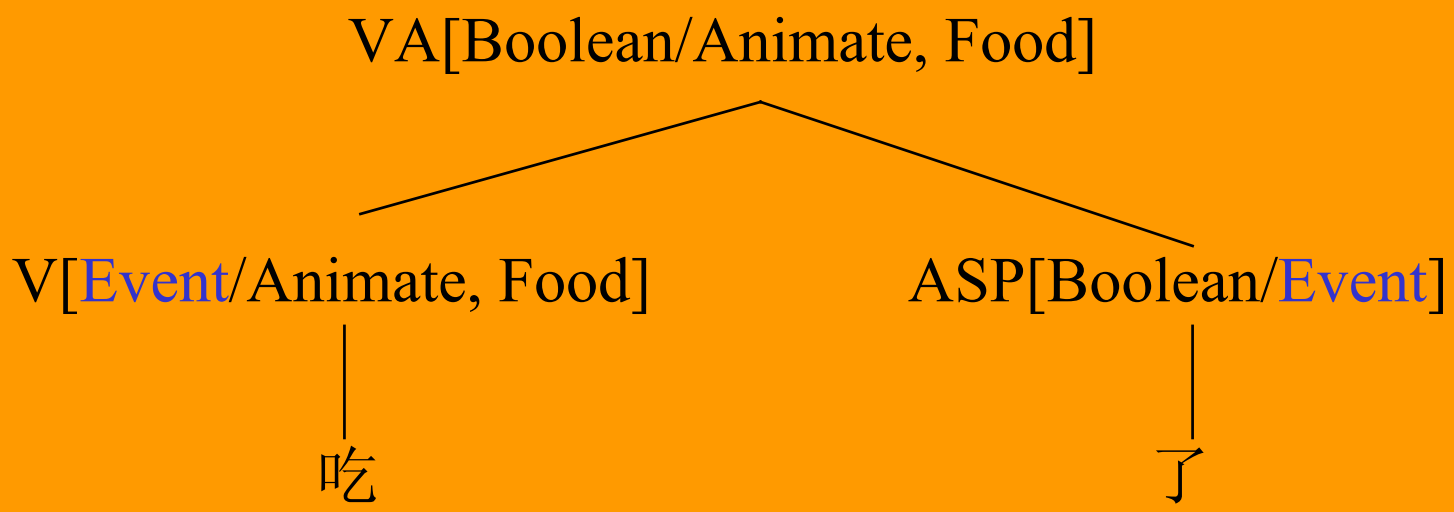
的字结构

- “卖盐的”：子句提取结构



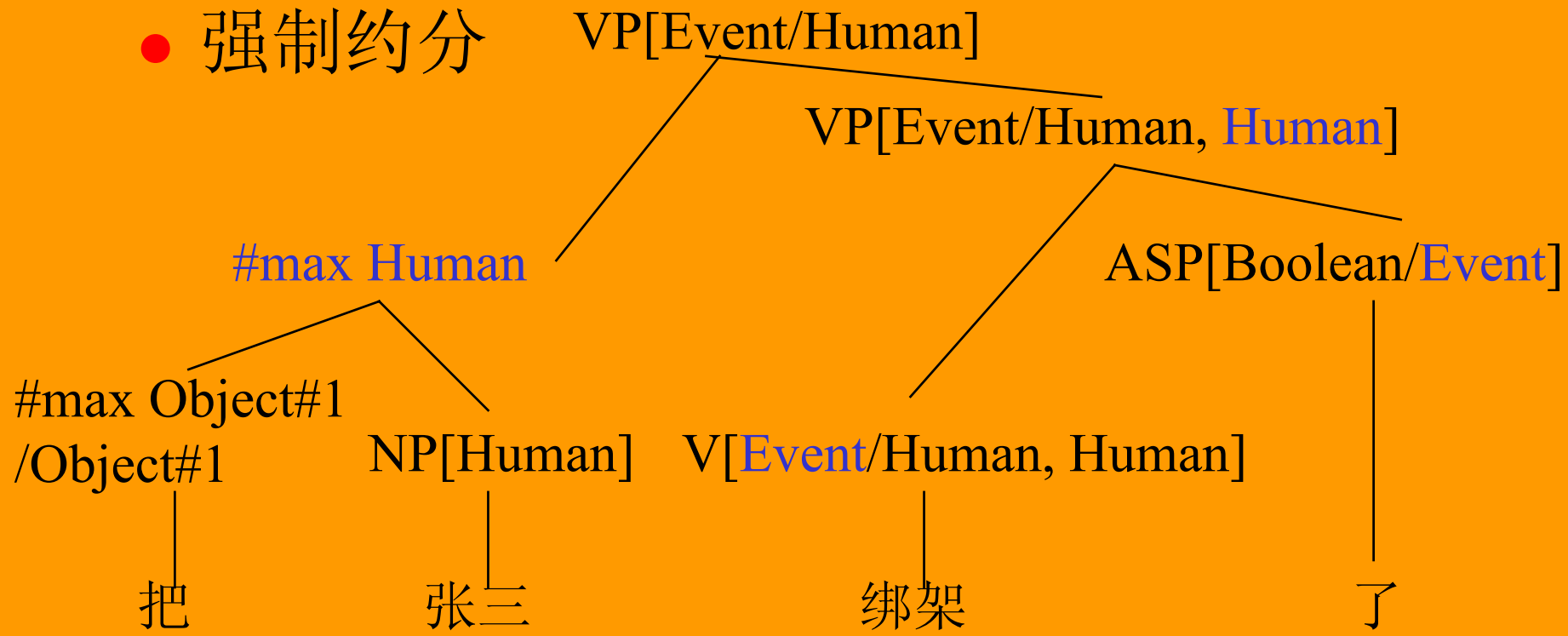
时体结构

- **“吃了”：VA →V ASP|VC ASP**
- **(ASP →了|着|过)**



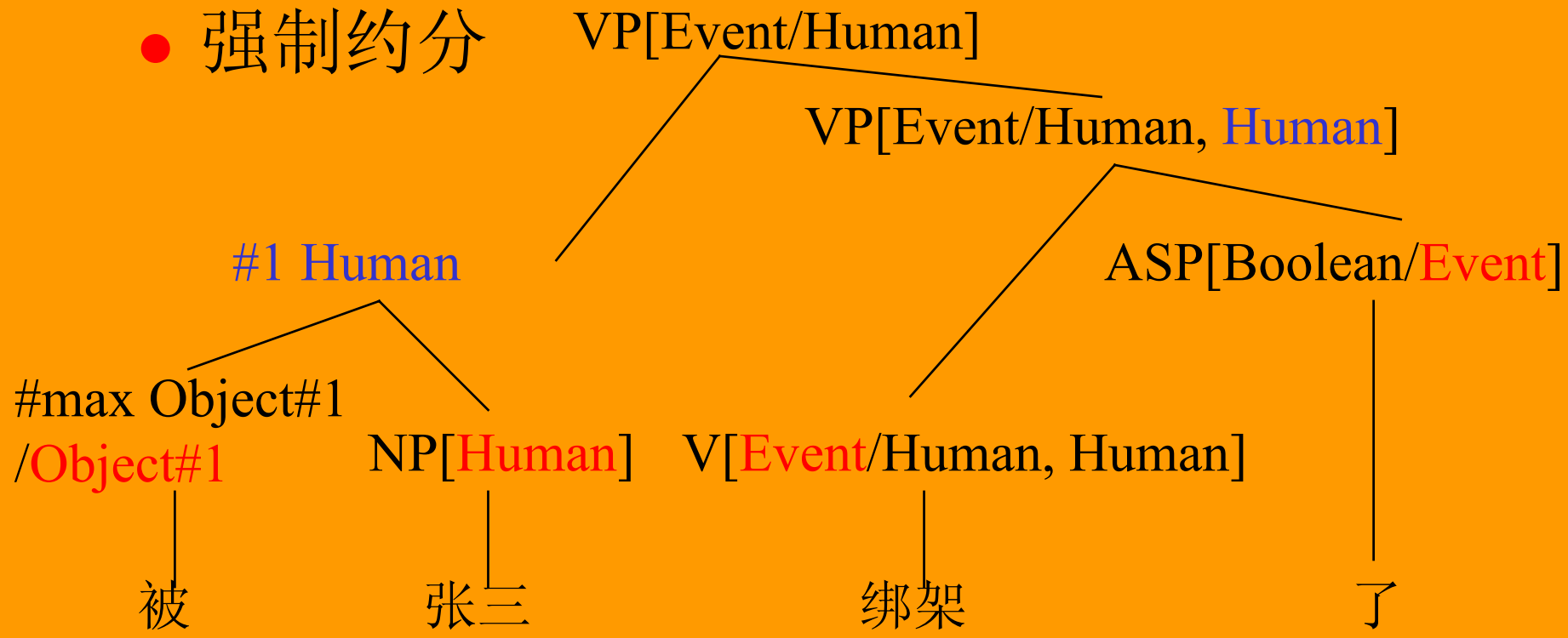
把字结构

- “把张三绑架了”：**VP →把 NP VP**
- 强制约分



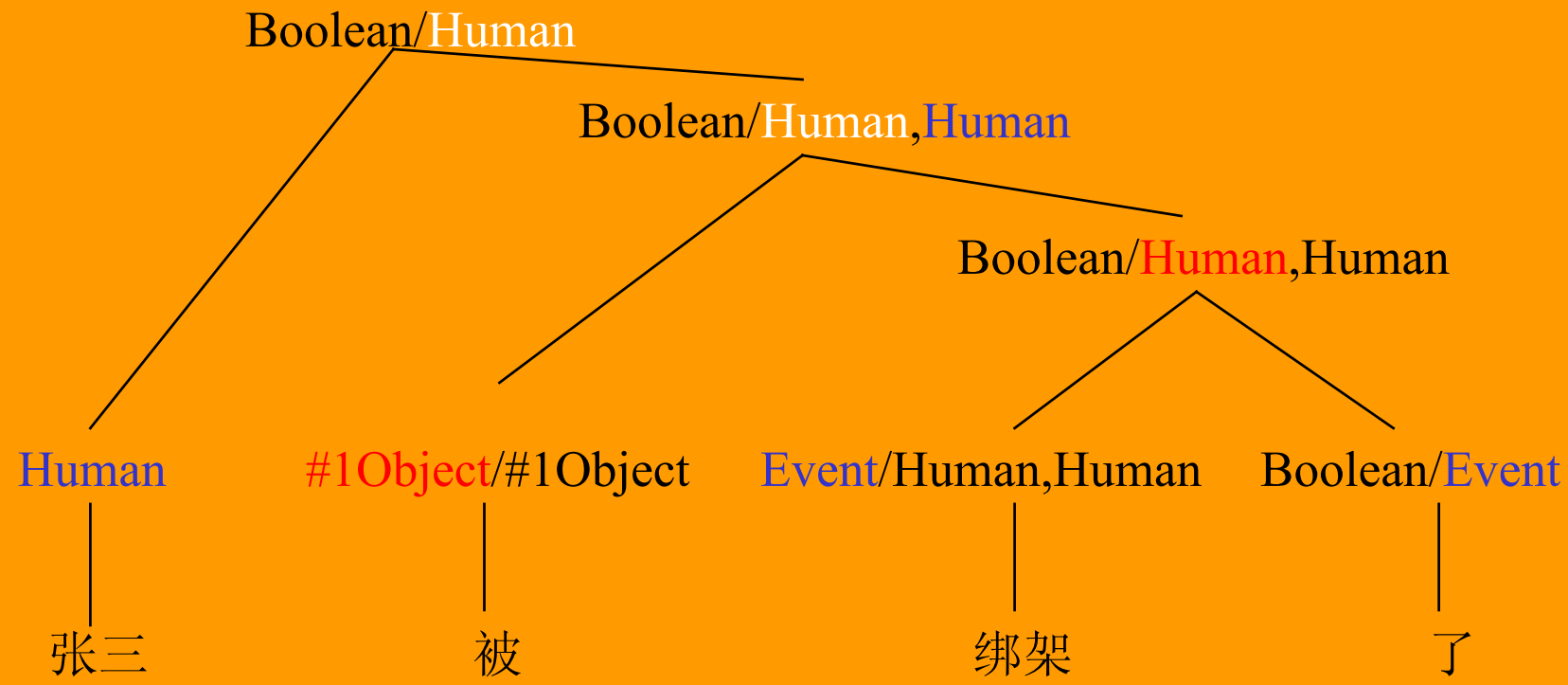
被字结构

- “被张三绑架了”：**VP →被 NP VP**
- 强制约分



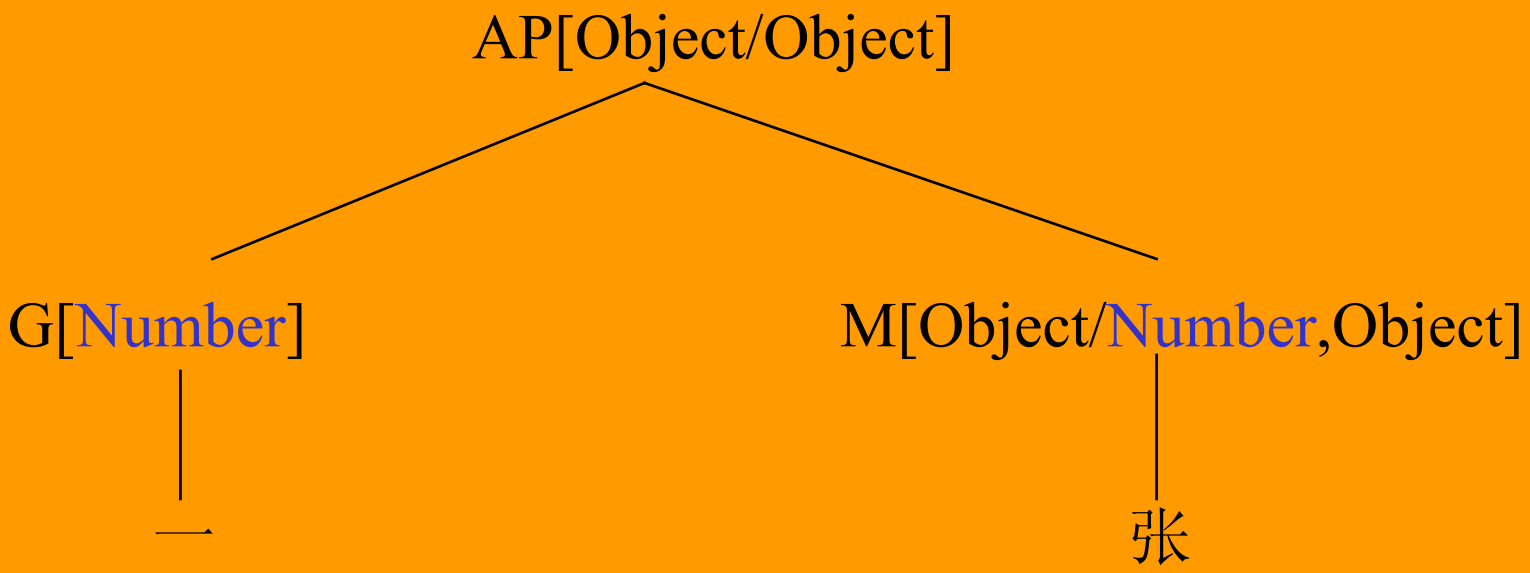
被字结构

- “张三被绑架了”：带缺省的被字结构

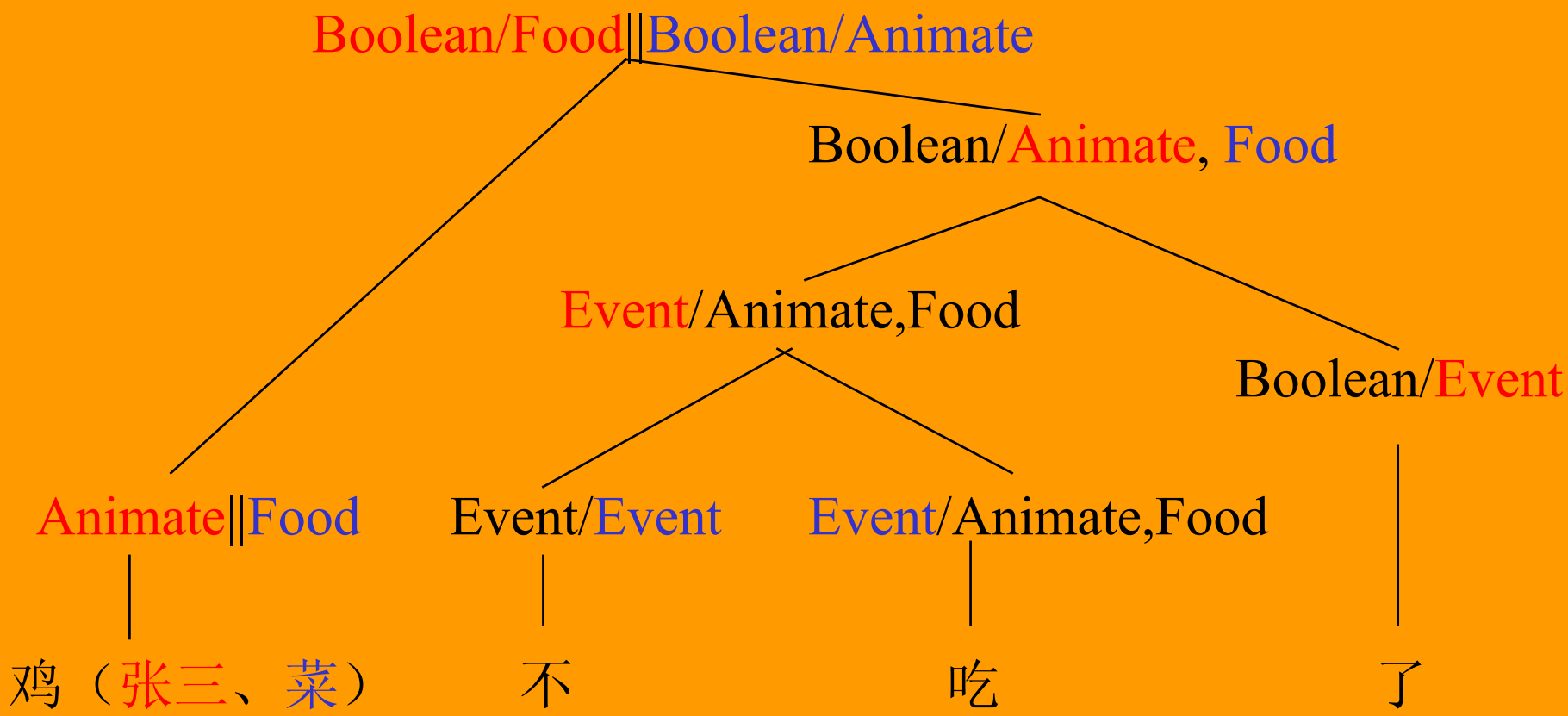


数量结构

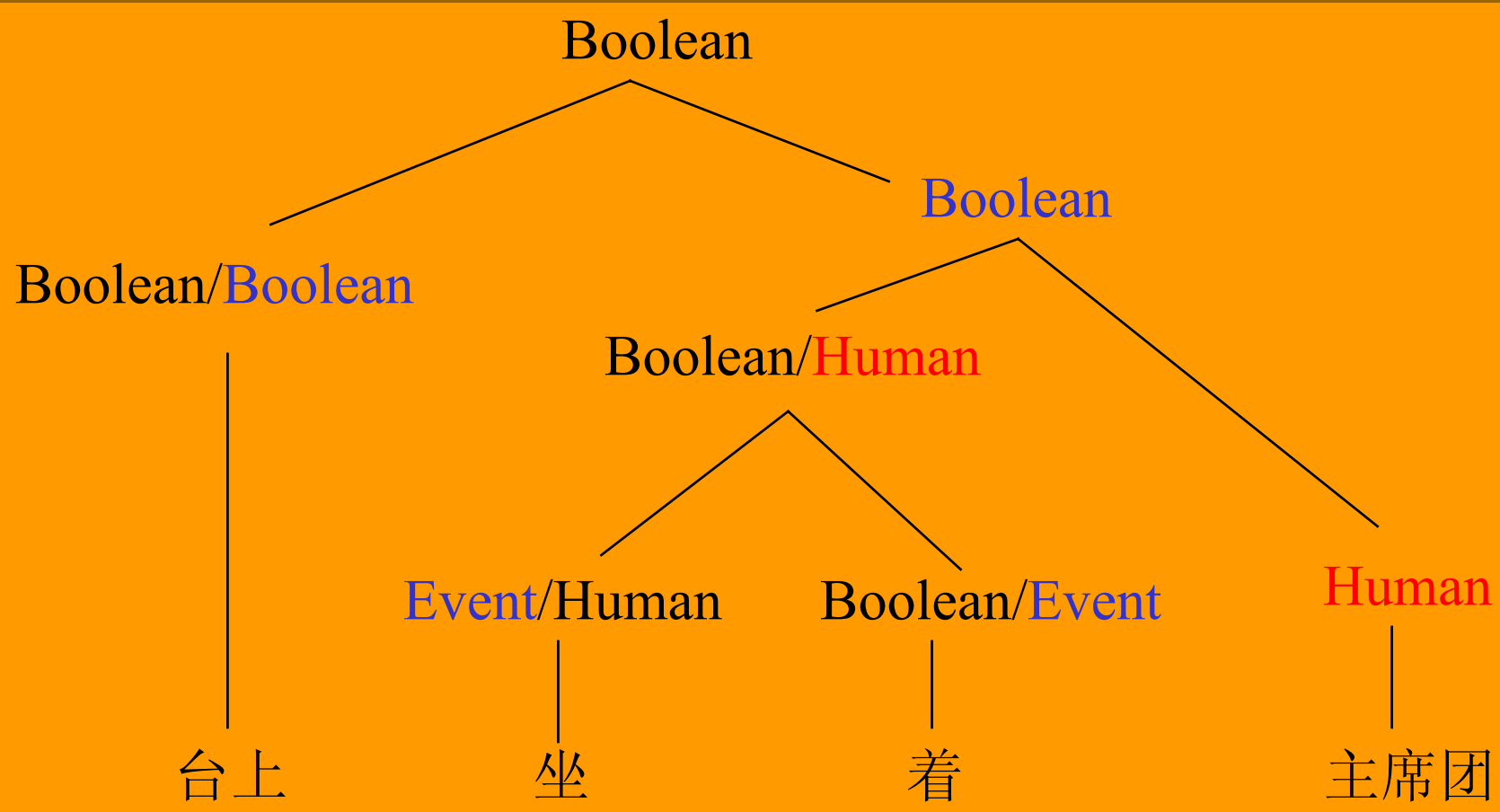
- “一张”:



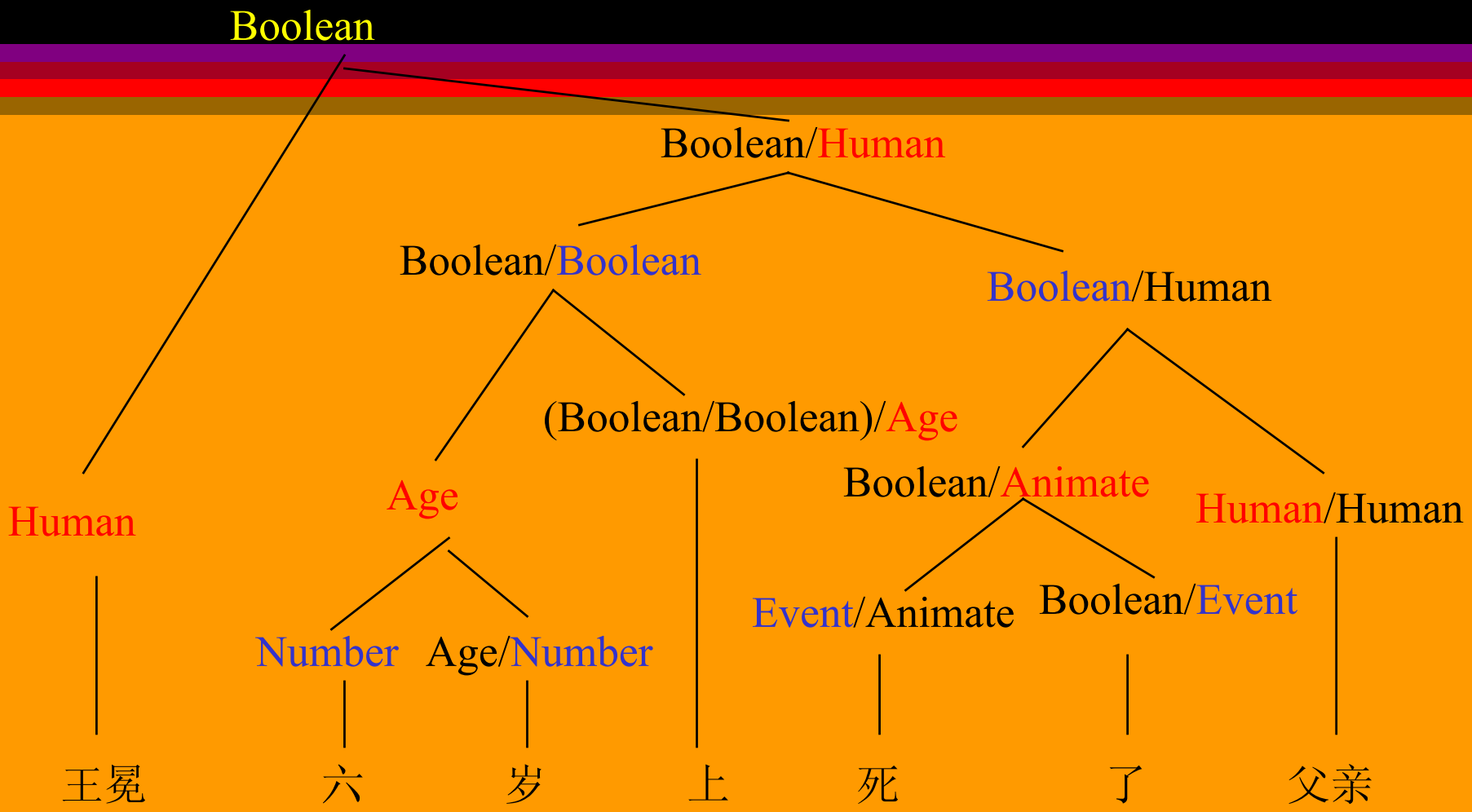
“鸡不吃了”


Boolean/Food||Boolean/Animate
Boolean/Animate, Food
Event/Animate,Food
Boolean/Event
Animate||Food
Event/Event
Event/Animate,Food
鸡（张三、菜）
不
吃
了

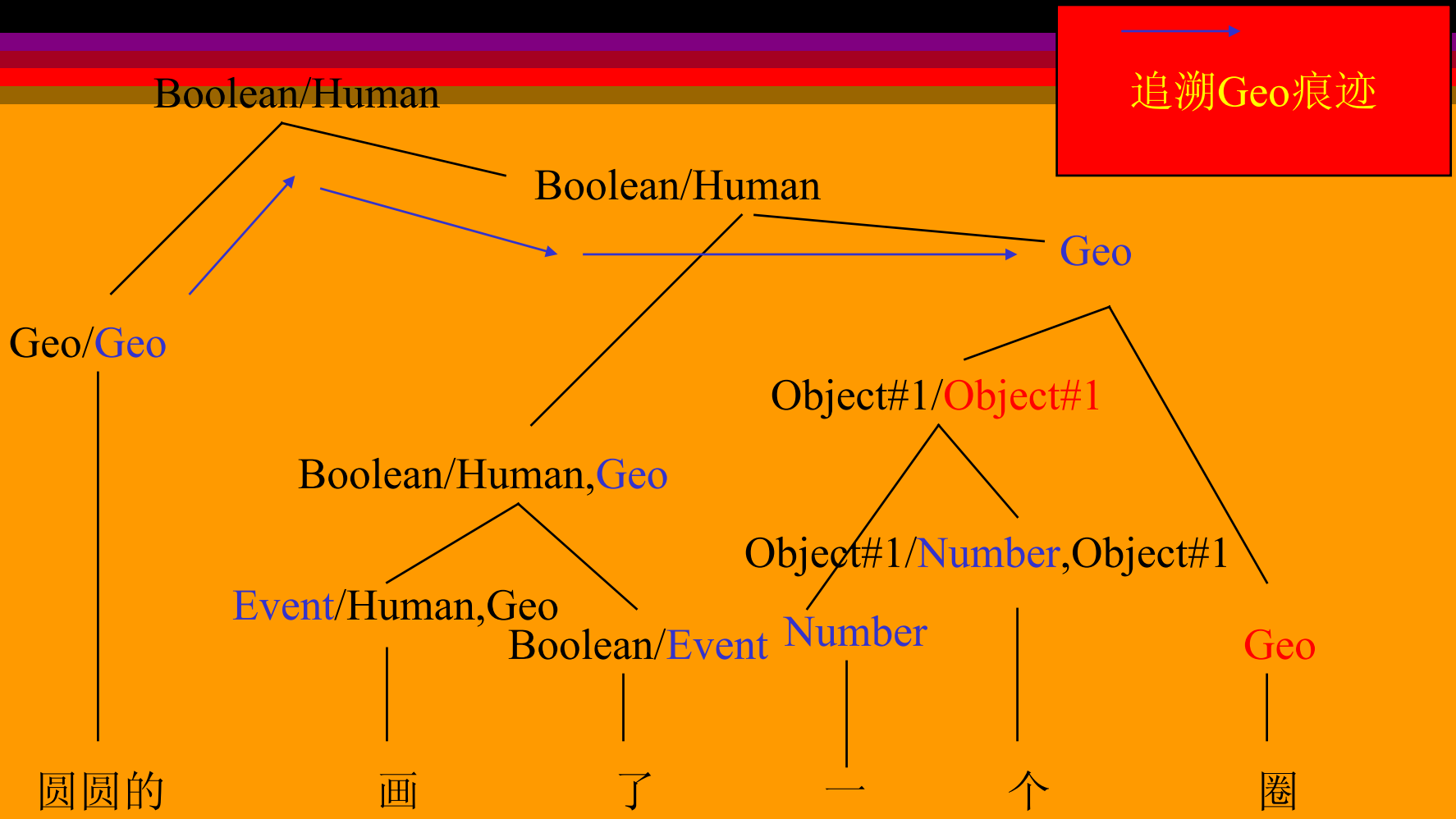
“台上坐着主席团”


Boolean
Boolean
Boolean/Boolean
Boolean/Human
Event/Human
Boolean/Event
Human
台上
坐
着
主席团

“王冕六岁上死了父亲”


Boolean
Boolean/Human
Boolean/Boolean
Boolean/Human
(Boolean/Boolean)/Age
Human
Age
Boolean/Animate
Human/Human
Event/Animate
Boolean/Event
Number
Age/Number
王冕
六
岁
上
死
了
父亲

“圆圆的画了一个圈”


追溯Geo痕迹
Boolean/Human
Boolean/Human
Geo
Geo/Geo
Object#1/Object#1
Boolean/Human,Geo
Object#1/Number,Object#1
Event/Human,Geo
Boolean/Event
Number
Geo
圆圆的
画
了
一
个
圈

结论

- 深层表达精确，表达力强
- 短语结构驱动，约分合理
- 向语义结构映射变得容易
- 帮助表层分析消歧
- 尤其适合汉语
- 范畴继承体系和范畴表达式词典是庞大的语言系统工程

谢谢大家