现代汉语树库（TreeBank）加工规范（Version 1.0）

詹卫东
北京大学中文系 北京大学中国语言学研究中心
最近更新：2009-07-30

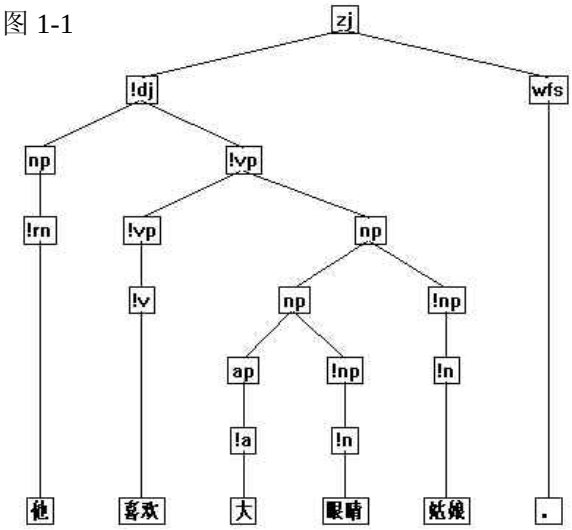
目 录

本规范文件的配套文档是“现代汉语树库标注常见问题举例”

一 引言

对自然语言句子的结构进行全自动的分析，是计算机进行自然语言信息处理的核心环节。这个环节的任务可以概括地描述为：将一维的线性字符串（句子）转换为二维的句法树结构的过程。例如，给计算机一个输入：“他喜欢大眼睛姑娘。”，如果计算机能够对这个句子进行正确的结构分析，它就可以输出如图 1-1 所示的树结构（有关标记的含义可见下文及附录三的说明）。

图 1-1



计算机要自动完成这个任务，并不是一件容易的事情，对于“他喜欢大眼睛姑娘”这个输入来说，计算机也有可能分析为下面这样的树结构：

图 1-2

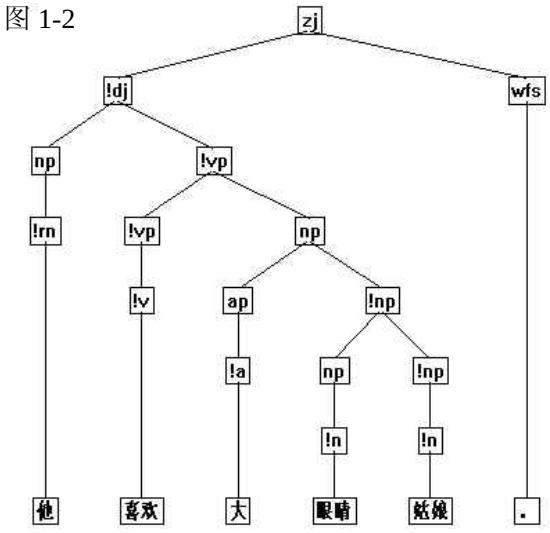


图 1-2 与图 1-1 中关于“大眼睛姑娘”的分析结果是不一样的。显然，图 1-2 是错误的分析。为了帮助计算机完成上面这样的任务（从某种程度上说，这也就是“理解”了句子的意义），需要我们人自身先对自然语言句子的结构有全面系统的把握。为此，就需要我们对大量的实际句子进行句法结构标注，为提高计算机自动句法分析的正确率（包括基于规则的分析方法

和基于统计的分析方法）提供一个坚实的基础。

像上面图 1-1 所示的句法结构树，在计算机中一般采用加括号（bracket）的一维线性形式来表示，例如：

```
[ zj [ !dj [ np [ !rn [ 他 ] ] !vp [ !vp [ !v [ 喜欢 ] ] np [ np [ ap [ !a [ 大 ] ] !np [ !n [ 眼睛 ] ] ] !np [ !n [ 姑娘 ] ] ] ] ] wfs [ 。 ] ] ]
```

其中 zj，np，n 等是语言成分的功能标记，左右括号 [ ] 则用来确定成分的边界。人们通常把带有这样的句法结构信息标记的句子组成的语料库称为“**树库**”（**TreeBank**）。

在树库基础上进一步获取句法分析规则，或者提取用于概率句法分析的参数，可以帮助计算机更好地进行自动句法分析。此外，在建成大规模树库后，对于进行定量的句型研究和语言教学都有极高的价值。

汉语语料库一般的加工流程是：

➔ 分词和词性标注

➔ 短语结构层次划分

➔ 短语功能类别标注

➔ 短语结构关系标注

➔ 语义关系标注

➔ 篇章信息标注

理想情况下，语料库中人工标注的信息越多，能够为计算机自动分析和语言研究与应用提供的帮助就越大。但在实践中，限于时间和人力等客观条件的限制，目前树库一般还难以按照上述流程全面地标注各个层极的信息，实际中标注下面两种信息的情况较多：

（1）结构边界信息，通常用[ ]或( )等符号来标记，表示语言成分的**结构层次**。

（2）功能范畴信息，比如 np，vp，…等等，用来表示一个语言成分的**句法功能**。

本加工规范将规定如何对汉语句子内部成分进行结构边界划分，以及如何来确定一个语言成分的功能范畴。对于实际的工作环节，包括从原始语料到制作成树库，请参见附录一的说明。规范正文的主体内容是对“短语层次划分”与“短语功能类别标注”两部分的描述（同时也涉及到一定的短语结构关系的问题）。在树库加工的过程中，也还会碰到分词和词性标注的一些问题，但限于篇幅，有关分词和词性标注的规范及相关问题这里从略（可以参看参考文献中列出的北京大学计算语言学研究所的相关规范文档）。本规范基本不涉及到语义关系标注、篇章信息标注。

二 语法单位的分层与分类

分层和分类是人们认识事物的基本方式。树库实际上就是用一种具体的形式来反映人们对语言成分分层和分类的认识。这包括分哪些层，哪些类；以及如何分。

2.1 语法单位的分层

本规范所描述的树库加工过程中涉及到的语法单位可以分为三个层次（三级）：

一级单位是树库加工中所面对的最大的加工单位，即“整句”。树库加工中暂不涉及“句段”“语篇”等更大的单位。对整句，目前没有进一步分类[1]；

二级单位是中层单位，即短语结构（或者说是词组），对这一层单位，根据功能差异，可以区分为复句性短语、单句性短语、名词性短语、动词性短语，形容词性短语、副词性短语、……等等类型[2]（详见下文第四节的具体说明）；

三级单位是基层单位，即词和语素。

下面表 2-1 对三级语言单位的性质作了进一步的说明。

表 2-1：

一级单位（整句）	不能被任何其他单位包含	这级单位不能作构成成分，所以不再进行功能分类，可依据结构分类
二级单位（短语）	可以被一级和二级单位包含，也可以包含二级和三级单位	这级单位可以相互嵌套；这级单位的分类应该同时考虑其功能和内部结构两个方面
三级单位（词）	只能被二级单位包含，不能包含其他单位	这级单位只能作构成成分；这级单位的分类只依据功能差异

上面三级单位，每一级单位之间既有功能上质的差异，也有长度上量的差异。一般来说，一级单位比二级单位的长度长，二级单位比三级单位长度长（也有的情况下，二级单位跟三级单位等长，详见下文 2.3 小节的说明）。对此可简单图示如下：

一级单位 > 二级单位 ≥ 三级单位

最大的单位只能从结构构成类型的角度进行分类。最小的单位只能从成分功能类型的角度进行分类。中间的单位则既可从功能的角度进行分类也可从结构的角度进行分类。因为从功能角度分出的类，更直接地表明了一个语言成分所具有的结合能力，对句子的结构分析起到更强的判别作用，所以树库在标记语言成分的性质时，二级单位和三级单位都统一采用功能分类的标记。在功能分类的基础上，对二级单位还可以进一步采用结构分类，比如同是动

[1] 一般可以根据整句的句末标点，大致将整句区分为“陈述”（句号结尾）“疑问”（问号结尾）“祈使/感叹”（叹号结尾）等不同的表达（功能）类型。需要注意的是，这种区分不是从句法功能角度出发分出的类。

[2] 事实上，从短语类的名称上也大致可以看出，“复句性短语”“单句性短语”跟“名词性短语”“动词性短语”等不同，后者是从词的功能分类沿用到短语层的功能分类，前者并不是真正意义上的功能分类，但对于“单句”“复句”这样的介于最大的单位（整句）和中级单位（短语）之间的单位，我们暂时选择将它们跟中级单位列在一个层级上，并沿用语言学上传统的“结构”名称“复句”“单句”来称呼。

词性短语（vp），再从结构不同上区分出更多的小类，实际上又能进一步在相当程度上显示不同小类 vp 的功能差异（详见下文 4.1 小节的说明）。

2.2 语法单位的分类

这里先将从功能角度区分出的基本的短语类型和词类及其标记列出来，有关短语、词和语素的更详细的分类及其标记可参见附录三及参考文献[2]。有关各类短语的功能描述可参见下文 4.1 的说明。

表 2-2：基本的短语标记（一级语法单位 + 二级语法单位）

短语标记	含义	短语标记	含义
zj	整句	np	名词性短语
dj	单句	pp	介词性短语
fj	复句	qp	数量短语
ap	形容词性短语	sp	处所性短语
dp	副词性短语	tp	时间性短语
mp	数词性短语	vp	动词性短语

表 2-3：基本的词类标记

词类标记	含义	词类标记	含义
a	形容词	n	名词
b	区别词	o	拟声词
c	连词	p	介词
d	副词	q	量词
e	叹词	r	代词
f	方位词	s	处所词
g	语素	t	时间词
h	前缀	u	助词
i	成语	v	动词
j	缩略语	w	标点
k	后缀	x	非语素字
l	习用语	y	语气词
m	数词	z	状态词

2.3 从词到短语

2.3.1 词性标记是否直接上升为短语标记的判别依据

从理论上说，一个词可以是以“词语”身份参与组合，也可以是以“短语（词组）”身份参与组合。比如“是”和“阿 Q”组合成“是阿 Q”这个词组，要对这个词组的结构进行分析和标注，有四种可能性，分别对应如下的组合形式（箭头 → 表示“由……组成”）：

vp → v n

vp → vp np
vp → v np
vp → vp n

这些组合形式可分别图示如下：

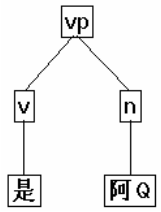


图 2-1A

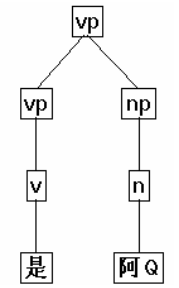


图 2-1B

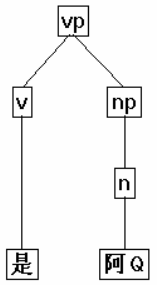


图 2-1C

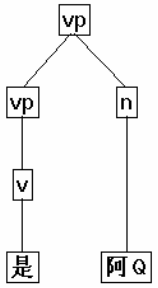
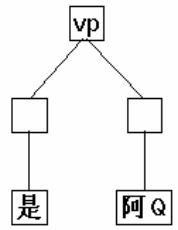


图 2-1D

那么，对“是阿 Q”这样的组合，到底如何来分析和标注比较好呢？

我们的认识是，应该根据一个结构位置允许什么样的语法单位进入该位置，来决定具体如何分析和标注。所谓“结构位置”，指的是类似于下面图 2-2 中空白方框所示的“树节点”位置，也就是说，像“是阿 Q”这样的 vp，它所对应的抽象结构模式为 vp → □ □ ，而这两个结构位置（由“□ □”表示）分别是由词来填充呢？还是可以由词组来填充？

图 2-2



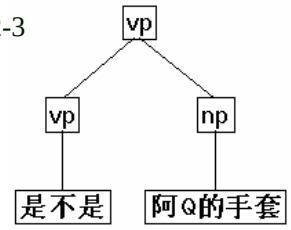
这通常包含两种情况：

（甲）一个结构位置既能由词来填充，也能由同功能类的词组来填充[3]。

（乙）一个结构位置只能由词语填充，而不能由词组来填充。

很显然，在图 2-2 所示的两个结构位置上，都是既可以由词来填充（如图 2-1A 所示），也可以由词组来填充，比如，下面图 2-3 所示就是由词组（vp，np）来填充结构位置的情形。

图 2-3



那么，**像甲这样的结构情况，我们的处理原则是：应该先将词性标记上升为短语标记，然后再跟其他语言成分组合。**也就是说，对于像“是阿 Q”这样的组合，按照图 2-1B 来进行分析和标注，是比较好的方式。其他三种方式（图 2-1A，2-1C，2-1D 所示的方式）尽管表现形式不同，但实质上处理原则是一致的，都是不可取的。

为什么按照图 2-1B 处理更好呢？因为这种处理方式会使得规则数量更少，更经济。

[3] 从理论上说，也应该包含这种情况，即一个结构位置只能由词组填充，不能由词语填充。不过，在实际语言中，这种情况比较少见。只有在一些特定的结构中，会有所表现，比如“把”形成的介词短语（pp），它所修饰的成分一般是动词短语（即由动词词组填充pp后的结构位置），而不是单个的动词（可对比： 把酒喝了 —— * 把酒喝）。

下面我们通过一个例子来对此加以说明。假设要用一套规则集来分析下面的词组：

（1）是阿 Q　　（2）是不是阿 Q 的手套

（3）是不是阿 Q 的朋友的旧手套　　（4）阿 Q 是

（5）阿 Q 的皮手套是朋友的旧皮手套　　（6）是皮手套公司

如果按照图 2-1B 所对应的的分析原则进行分析，需要的规则如下表中右边的 B 栏所示，而如果不按照这种分析原则进行分析，需要的规则就如下表中左边的 A 栏所示。

表 2-4：

A	词组实例	B
vp → v　n	是 阿 Q	vp → vp　np
vp → vp　np	是不是　阿 Q 的手套	
vp → v　np	是 皮手套公司	
vp → v 不 v	是 不 是	vp → vp 不 vp
dj → n　v	阿 Q 是	dj → np　vp
dj → np　vp	阿 Q 的皮手套　是朋友的旧皮手套	
np → n 的 n	阿 Q 的 朋友	np → np 的 np
np → n 的 np	阿 Q 的 皮手套	
np → np 的 np	阿 Q 的朋友 的 旧手套	
np → a　n	旧 手套	np → ap　np
np → a　np	旧 皮手套	
np → n　n	皮 手套	np → np　np
np → np　n	皮手套 公司	
		vp → v np → n ap → a
13 条规则		9 条规则

不难看出，A 栏的规则数量多于 B 栏，而且可以预见，随着实例的增多，A 栏的规则数量还会以比 B 栏更快的速度增长。A，B 两栏的规则数量不同，实际上正是不同的结构标注原则的反映。A 栏的处理原则是，让词性标记直接参与组合。B 栏的处理原则是，将词性标记先直接上升为同功能类的短语标记，然后再参与组合；如果按照 A 处理方式，在一个结构位置允许词组也允许词填充时，就必然需要两条规则来刻画；而如果按照 B 处理方式，则只需要一条规则来刻画，因为词组涵盖了同功能的词类标记（通项公式为：xp → x）。进一步说，如果有大量的结构位置都具有既允许词组填充该位置，也允许词填充该位置的特点，那么，按照 A 处理方式所需要的规则数量，就会明显地多于按照 B 处理方式所需要的规则数量。正是因为这个原因，我们确立了按照 B 这种方式进行结构标注的原则。当然，这一原则适用于上文提到的情况甲：一个结构位置既可以由词组填充，也可以由词填充。

下面再来看情况乙的例子：

图 2-3A

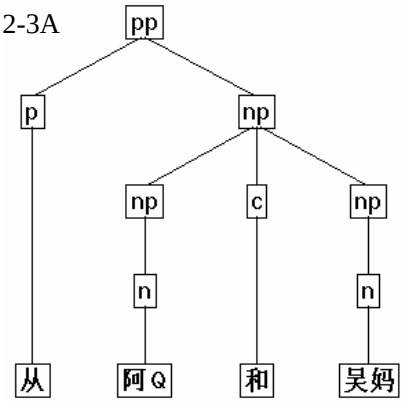
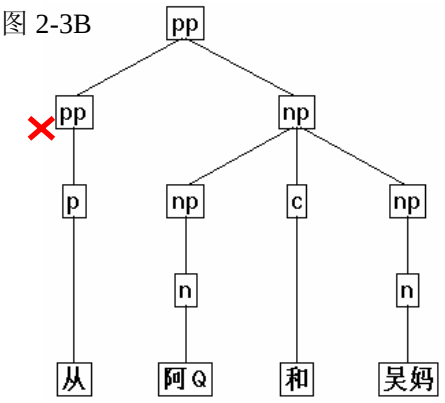
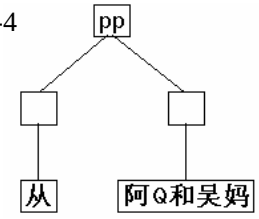


图 2-3B



抽象地看，上例对应的组合模式可以图示如下：

图 2-4



在“pp → □ □”这个组合格式中， 第一个□所在的结构位置上只能出现词语（介词），而不能出现介词词组（pp）[4]，因此，图 2-3A是比较好的标注形式，图 2-3B相比较之下则是不大好的标注形式。换句话说，汉语中没有下面这两种组合格式：

pp → p

pp → pp np

由此，我们可以提出第二条处理原则。**对于像乙这样的情况，不应将词性标记上升为短语标记，而应该由词性标记直接参与组合。**

把上文所说的这两条原则归纳在一起，就是：**当一个结构位置上只能出现词，而不能自由地出现短语时，就由词语所对应的词性标记直接参与组合；当一个结构位置上可以自由地出现短语（词组）时，就应该由短语参与组合，即词性标记应先上升为短语标记，然后再参与组合。**

这里不妨再举一些例子对上述原则做进一步阐述。

2.3.2 词性标记不上升为短语标记的情况举例

例 1： 按照上述原则，对于动词加“着”的组合，动词不应上升为vp，而应该直接由“v+着”进行组合，因为在“v + 着”的v位置上，不能出现vp[5]。规则组合模式为：vp→!v uzhe。与此相对，对于动词加“了”“过”的组合，则应该将v上升为vp再与“了”“过”组合，因为在“了”“过”前面的位置上，可以出现vp。比如“摔碎了（过）一个花瓶”。规则

[4] 第二个□所在的结构位置可以出现词组（np），也可以出现词（n），这个结构位置属于上文所说的情况甲。

[5] 参加课题工作的赵欣同学指出，也有像“（一路）小跑着前进”这样的例子。这种情况下是“小跑”加“着”，而不是“小”+“跑着”或“小”+“跑着前进”，而“小跑”属于状中型的vp。这也就意味着“着”前面的位置上可以出现vp，而不仅仅是v。不过，除“小跑，高举”这类很少数的状中型vp可以加“着”外，大多数的vp是不能加“着”的。如果因为有“小跑着”这样的用例，就设置“vp+着”组合规则，会带来明显的过度概括问题。权衡利弊之后，可以将“小跑着”“高举着”这类vp组合处理为三分的模式：vp → a !v uzhe(着)。这样既可以反映这类结构的特点，也可避免设置“vp→ vp uzhe”组合规则可能带来的问题。另一种处理策略是将“小跑”、“高举”看作是单个动词，而不是状中式的vp。从替换性角度和扩展性角度讲，“小跑”“高举”看作是单个动词也有一定的合理性。将“小跑”“高举”看作单个动词，而不是vp，同样避免了“vp+着”的组合模式。

组合模式为：vp→!vp ule　，　vp → !vp uguo。

例 2：汉语中粘合述补结构（如“摔破、喝醉”等）也是直接由词类组合形成短语结构的，而不是由短语组合形成短语结构的。在粘合述补结构的述语位置和补语位置，都是直接由词类标记来填充，而不是由词组（短语）标记来填充。规则组合模式为：vp → !v a。

例 3：汉语中副词一般是作状语修饰谓词性成分，但也有一些副词可以直接放在名词性成分的前面。这些副词主要是表范围数量的副词。比如“仅两人”“一共三本书”等等。像这样的“副词+np”的组合模式，副词不上升为 dp，而是直接以副词的身份参与组合。规则组合模式为：np→ d !np。程度副词修饰名词性成分的时候，整个结构是形容词性的。比如“很中国”。副词修饰的名词通常也是单个名词，不是名词词组。对这类组合，处理为副词和名词直接参与组合的形式，即 ap → d !n。除此之外，“副词 + np”组合还有一类情况，比如“好好一个人”。对于这类由单音节形容词重叠形成的副词性成分，有两种处理方式，一种是作为单词收入词库，一种是作为结构，分析为 dp → !a a。我们采用后一种处理方式，这样，这类“副 + np”组合的模式即为 np → dp !np。 其中 dp 为单音节形容词重叠而来。

2.3.3 词性标记应上升为短语标记的情况举例

在“把我的姓名写在这条线的上边还是下边”这个句子中，“这条线的上边还是下边”可能被分析成 sp -> !sp c f 这样的结构，如下面图 2-5 所示。但后面这个 f 所处的结构位置上也可以出现 sp，因此，f 应该上升为 sp 再参与组合，这样就形成 sp-> !sp c sp 结构，如下面图 2-6 所示。

图 2-5

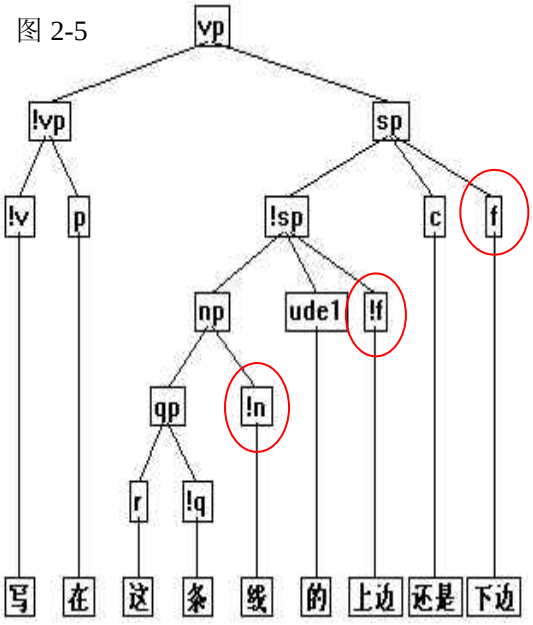


图 2-6

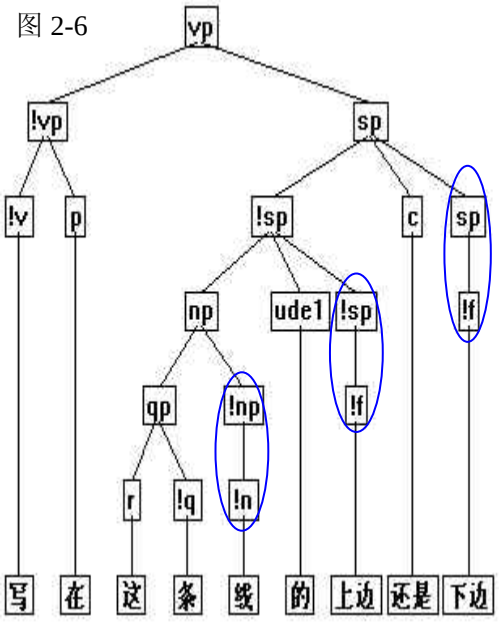


图 2-5 是不符合规范的标注，图 2-6 才是符合规范的标注形式。

有了对语言单位的以上认识，当我们面对一个句子时，就好像庖丁解牛一样，首先要将句子分解成大大小小的语言单位，然后再按照一定的类别标准给这些单位贴上事先分好的类别标签。需要特别注意的是，人在理解句子时可能这两部分工作是交织在一起进行的，很难说孰先孰后，但为了行文的需要，在下面的章节中，我们还是“人为地”把分解句子的过程按照先划分出单位，再贴标签的步骤来加以说明。

三 结构层次标注

把句子分割成大大小小相互间有嵌套关系的语言单位，就是对句子进行结构层次划分。结构层次划分反映的是语言单位之间的结合顺序关系。如果只有两个成分结合，因其是对称的，就无所谓先后顺序，而当有三个或三个以上的成分发生结合关系时，就必然会发生先后顺序的问题。下面以最基本的三项成分组合来看结构层次划分（结合的先后顺序）的问题（超过三项以上成分组合的结构层次划分显然可以参照三项成分组合时的情形加以判别）。

3.1 结构层次划分的一般原则

对于 A + B + C 形成的短语，这里记做{A　B　C}，内部构造层次会有三种可能性：

（P1） [[A　B]　C]（"A　B" 先结合为一个单位，再跟 C 结合，比如：大眼睛姑娘）

（P2） [A　[B　C]]（"B　C"先结合为一个单位，再跟 A 结合，比如：新操作系统）

（P3） [A　B　C] （"A　B　C"分不出结合的先后顺序，比如：张三李四王五）

实际上，在对{A　B　C}具体进行结构层次划分时，总是要考虑下面两个问题：

（1）二分还是三分（三分其实也就是内部不分层次）。

（2）如果是二分，是按照 P1 模式分，还是按照 P2 模式分。

要回答上面这两个问题，在具体操作时，可以把语言成分之间的结合分成下面这样几种情况，根据自己的语感来进行判别：

（i）如果“[A　B]” 和“[C]”分别都能单说，而[B　C]不能单说，则层次分析为 P1 模式，即 B 跟 A 结合紧密，而 B 不能跟 C 结合，比如：

羊皮领子大衣　　——　[[羊皮　领子]　大衣]

走遍中国　　——　[[走　遍]　中国]

（ii）如果“[A]”和“[B　C]”分别都能单说，而[A　B]不能单说，则层次分析为 P2 模式，即 B 跟 C 结合紧密，而 B 不能跟 A 结合，比如：

很有才华　——　[很 [有　才华]]

买大电视　——　[买 [大　电视]]

从理论上　——　[从 [理论　上]]

（iii）如果“[A　B]”、“[C]”和 “[A]”、“[B　C]”分别都能单说，但[B　C]单说时的含义跟它在 {A　B　C}中的含义不同，则层次分析为 P1 模式。比如：

三本书　　——　[[三　本]　书]　　（P1 模式）

相应地，如果[A　B]单说时的含义跟它在 {A　B　C}中的含义不同，则层次分析为 P2 模式。比如：

喜欢舞蹈老师　——　[喜欢 [舞蹈　老师]]　（P2 模式）

当 C 是由一些后置虚词充当（比如“了”）的时候，会发生向前黏附到哪一个成分上的问题，这时也应该根据“了”黏附位置不同对整个短语意义理解的影响来判别（参见 3.3.1 小节的说明）。比如：

允许他抽烟了　——　[[允许　他抽烟]　了]　（P1 模式）

以为他抽烟了　——　[以为 [他抽烟　了]]　（P2 模式）

（iv）如果“[A　B]”“[C]”和“[A]”“[B　C]”都能单说，并且都跟在{ A　B　C }中的含义相容，则层次可以分析为 P1 模式，也可以分析为 P2 模式，比如：

努力学习文化 —— [[努力 学习] 文化] （P1 模式）
[努力 [学习 文化]] （P2 模式）
明代诗歌研究 —— [[明代 诗歌] 研究] （P1 模式）
[明代 [诗歌 研究]] （P2 模式）

上面两例中，前一例两种分析意思没有不同，是**多切分结构**；后一例两种分析意思不同，是**歧义切分结构**。

对于多切分结构，可以综合考虑更多因素后，规定一种分析模式。比如对于“努力学习文化”这样的例子，综合考虑汉语的短语结构系统，一般取 P2 分析模式。P2 分析模式的含义是：汉语中的动词跟宾语的结合关系比跟状语的结合关系更紧密。支持这一看法的语言事实是，汉语中存在一定数量的所谓“粘宾动词”，比如“觉得，企图，加以、予以、装作、处于、……”等等。这些动词都要求跟宾语形成述宾结构后才能参与更大的组合。而相应地，汉语中似乎不存在“粘状动词”。由此，当状语（A）和宾语（C）竞争居于二者中间的动词（B）时，更具系统性的分析方式应该是 [A [B C]]，即按照 P2 模式分析。这种分析模式反映了汉语中宾语和动词的结合优先于状语的事实。

对于歧义分析结构，应根据实际情况选择能够反映语义内容的切分方式。比如对于“明代诗歌研究”这个歧义例子，如果在具体的语境中，指的是“对明代诗歌所做的研究”，就应该按照 P1 模式分析；如果在具体的语境中，指的是“明代的时候所做的有关诗歌的研究”，就应该按照 P2 模式分析。

（v）如果“[A B]”“[B C]”都不能单说，则又可以区分出下面两种情况：

情况甲： “[A B]”“[B C]”确实都不能单说，比如：“买了+ 张 +电影票”，其中“买了张”和“张电影票”一般都不单说，因此应该按照P3 模式三分[6]。

情况乙： “[A B]” 和“[B C]”中有一个能单说，但单说时的意义跟在{ A B C }中的含义不相容，层次也分析为 P3。比如“（我）喝 的 牛奶”，其中“喝 的”虽然能单说，但在作为“你早晨喝什么了？”的答句时，其中“喝的”的含义与单说的“喝的”显然不同（单说“喝的”是转指人或食品），整个短语的内部层次就应该分析为 P3 模式（参见下文图 3-5，3-6）；

对具体的句子（短语）结构的内部层次进行分析，一般都可以分为上述五种情况。总的操作原则是：**如果能够二分，应尽量二分**；如果无法二分，或者二分会带来负面的影响，则做多分的处理。

下面给出一些常见的可能有争议的结构的层次分析处理办法，作为标注时的参照。首先看一些多分支结构。

[6] 当然，从一般语言学分析的角度说，可以把“买了张电影票”分析为一般的二分模式：买了 + 张电影票。并认为这里“张”的前面省略了数词“一”。这是汉语普通话中非常普遍的现象，即量词前面的数词“一”往往可以省略。但是，对于计算机分析来说，处理为二分结构会带来一个明显的问题，就是“张电影票”将成为一个短语结构，其组合模式为 np → q !np 。这样，当计算机分析 “两 张 电影票”的时候，就会出现两种结构分析的可能性： A.[[两 张] 电影票] B. [两 [张 电影票]]。后一种结构分析正是因为在处理“买了张电影票”的时候，将“张电影票”看作一个结构引起的。为避免这种情况，我们不按照一般语言学分析的方式将“买了张电影票”看作一般二分结构的省略形式，而是直接看作三分结构。换句话说，这种处理方式是用一种“特别”的形式来标记由于省略造成的“特殊”结构。

3.2 多分支结构举例

3.2.1 双宾语构造

双宾语构造都处理为三分结构。例如：

送她一束花 → vp [[送] [她] [一束花]]

偷了他三本书 → vp [[偷了] [他] [三本书]]

双宾语构造的组合模式为： vp → !vp np np

其中述语位置可以出现单个动词，也可以出现“v 了”“v 过”等形式，因此，述语位置的动词先上升为 vp，再参与组合。

3.2.2 兼语构造

兼语结构都处理为三分结构 （以区别于一般的连谓结构 vp+vp）。例如：

派人去 → vp [派 人 去]

帮助他复习 → vp [帮助 他 复习]

责备他不听话 → vp [责备 他 不听话]

没有人不喜欢花 → vp [没有 人 不喜欢花]

兼语结构的组合模式为： vp → !vp np vp

其中第一个 vp 位置可以出现单个动词，也可以出现“v 了”“v 过”等形式。因此该位置上的动词需上升为 vp，再参与组合。

值得注意的是，对于一些“特殊”的动词，如“使、让，有、没有、是”等，在兼语构造中出现时，为突出这些动词的特殊性，不上升为 vp，直接以 v 的形式参与组合。组合模式为： vp → !v np vp

需要特别说明的是，传统的语言结构分析一般会将双宾语构造处理为述宾结构再带宾语的二分构造，将兼语结构看作是连谓结构的一种形式，从而也跟一般的连谓结构一样做二分处理。但从计算机分析的角度来看，这样的处理方式更容易造成双宾动词结构混迹与普通单宾动词结构中，同样地，也容易造成兼语结构和一般的普通连谓结构混为同类。实际上，**一般语言学分析更多的是“求同”，结构分析要尽可能体现“同类结构”的“共性”**，比如双宾语结构跟一般的述宾结构确实是有共性的，无论是双宾构造，还是单宾构造，将述语与宾语二分，确实可以体现这种共性。但是，**为了计算机分析短语结构的目的，取舍时则更倾向于“求异”，即结构的分析最好能体现出“同类结构”内部的差异，凸现出一个结构的“个性”**。从这个原则出发，对于双宾语构造，兼语构造，我们都选择做三分的处理。主要目的就是将这些“特殊”的结构从它们的所谓的“同类”中区分出来。这也是本规范在进行结构层次分析和短语功能类标注时的一个基本精神。了解了这一点，有助于理解本规范中其他具体例子的处理方式（比如对下面 3.2.3，3.2.5，3.2.6，…等等例子的处理）。

3.2.3 “v + 给 + np”构造

处理为三分结构。例如：

递给他 → vp[[递] [给] [他]] vp → !v v np

放在桌子上 → vp[[放] [在] [桌子上]] vp → !v p sp

扔到垃圾堆里 → vp[[扔] [到] [垃圾堆里]] vp → !v v sp

类似的，“v + 在 + sp”“v + 到 + sp”等构造也分析为三分支结构。

需要注意的是，如果在“给”“在”“到”等和后面的 np、sp 成分之间有“了”，则整个结构处理为四分支结构。例如：

放在了桌子上 → vp [[放] [在] [了] [桌子上]] vp → !v p ule sp

“v p sp”格式中 v 的位置上也可能出现紧凑的动词性结构，比如：

绊倒 在 泥潭里

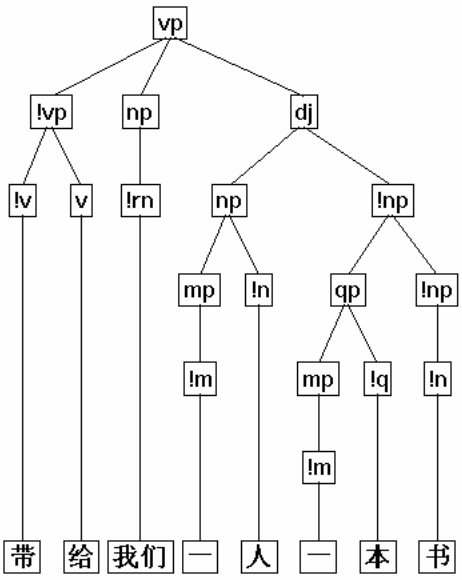
飘打 在 我的脸上

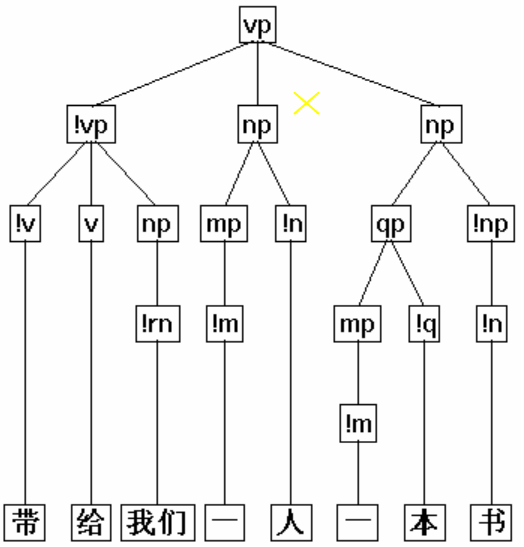
溺死 在 海里

3.2.4 多分支动词结构的套合

多分支动词结构可能共现，出现复杂的动词结构套合的现象。这时应根据语言学分析的一般方法来寻找更为合适的层次分析方式。比如对下面这个例子，可能会在左图和右图之间进行选择。

“带给我们一人一本书”





!vp(带给) np(我们) dj(一人一本书) !vp(带给我们) np(一人) np(一本书)

上面这个例子是一个双宾结构跟“v 给 xp”结构的套合。除此之外，这个例子中还包含了一个汉语中典型的数量分配句式，即“一人一本书”。

按照左图方式分析，突出了双宾结构的性质，但没有按照处理一般“v 给 xp”三分支结构的方式处理；

按照右图方式分析，在第一个 vp 的层次分析中体现了一般“v 给 xp”三分支结构的处理方式，同时也在上层的 vp 的层次分析中体现了双宾结构的特点。

从表面上看，右图的分析更好一些。但是，右图的分析有一个比较严重的问题，就是没有照顾到数量分配句式的分析，把数量分配句式的两部分“一人”和“一本书”割裂成两个部分，没有很好地体现二者的相关性。而按照左图的分析，“一人”和“一本书”在一个 dj

之中，这跟数量分配句式单用时，或者用于其他结构位置（比如主语，“一人一本书是不可能的”）时的分析是一致的。换句话说，按照左图的分析，对数量分配句式的处理更好一些。而按照右图的分析，表面上体现了双宾结构的模式，但实际上是错误地将数量分配句式分析为一个假双宾结构模式。这一点，通过下面的例子可以更清楚地反映出来。

“孩子出生，你精心呵护，疼爱备至，他就会用他的健康成长及甜甜的微笑带给你一天一个惊喜”

在上面的例子中，“带给你一天一个惊喜”按照左图的模式分析更好。如果按照右图的模式分析，就造成“带给… + 一天 + 一个惊喜”，近宾语是“一天”，远宾语是“一个惊喜”，这跟原句的语义不吻合。在“带给我们一人一本书”中，近宾语是“一人”，远宾语是“一本书”。因为近宾语也是指人的，所以从语义分析上，右图的分析模式也可以接受。但是，如果从更广泛的例子来看，按照右图的模式分析这类复杂动词结构套合现象，有可能造成句法分析和语义解释之间的不一致问题。

通过上述分析，可以得出结论，尽管按照左图分析会造成对“v 给 xp”处理上的不一致（有时候“v 给”二分，有时候“v 给 xp”三分），但整体而言，句法分析和语义解释之间的对应性更好，因此，对这类动词结构套合，应该按照左图的模式进行层次分析。

3.2.5“v + 有 + np”构造

处理为三分结构。例如：

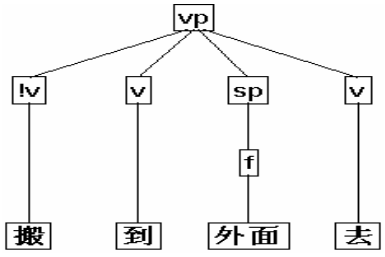
设有三个办事处 → vp[[设] [有] [三个办事处]]

（这是大自然）带有几分野性美 → vp[[带] [有] [几分野性美]]

这类结构的构造模式为：vp → !v v np

3.2.6“v + 到 + sp + 去”构造

处理为四分结构。例如：



类似的，“v + 到 + sp + 来”也处理为四分支结构。规则模式为：vp→!v v sp v

3.2.7“v + 趋向动词 1 + np + 趋向动词 2”构造

处理为四分结构。例如：

搬进一张桌子来 → vp[[搬] [进] [一张桌子] [来]]

走出两个人来 → vp[[走] [出] [两个人] [来]]

这类含复杂趋向动词的结构的构造模式为：vp → !v v np v

如果在“趋向动词 1”后面出现“了”，则处理为五分结构。例如：

闯进了两个人来 → vp[[闯] [进] [了] [两个人] [来]]

构造模式为：vp → !v　v　ule　np　v

3.2.8“v + 得 + 补语”构造

处理为三分结构。例如：

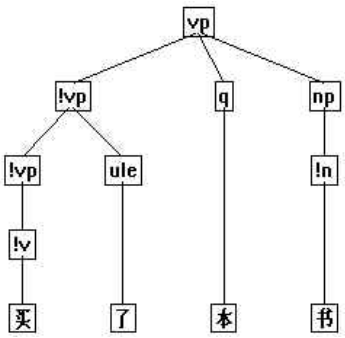
看得不清楚　→　vp[[看] [得] [不清楚]]　　vp → !v ude3 ap

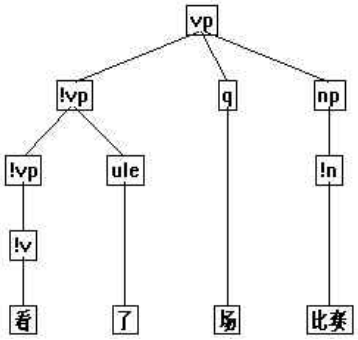
类似的，“v + 不 + 补语”也分析三分支结构。例如：

看不懂　　→　vp[[看] [不] [懂]]　　vp → !v d v

3.2.9“vp + q + np”构造

处理为三分结构。例如：



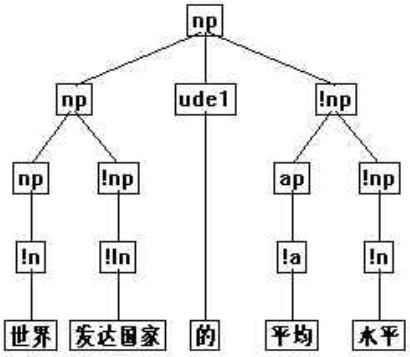


量词夹在 vp 和 np 之间（或者说量词前缺少数词）时，量词既不靠前，也不靠后。

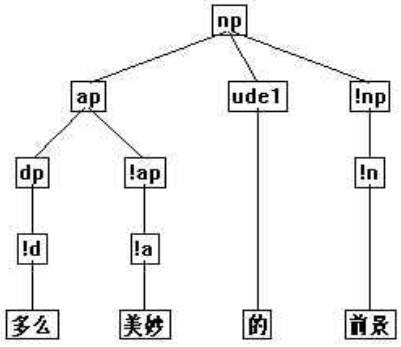
3.2.10 “x 的 y”构造

这类结构一般都处理为三分结构。例如：

例 1：世界发达国家的平均水平

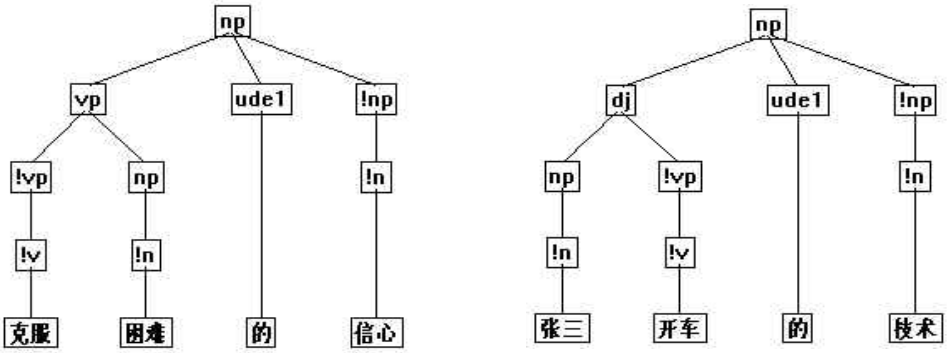


例 2：多么美妙的前景

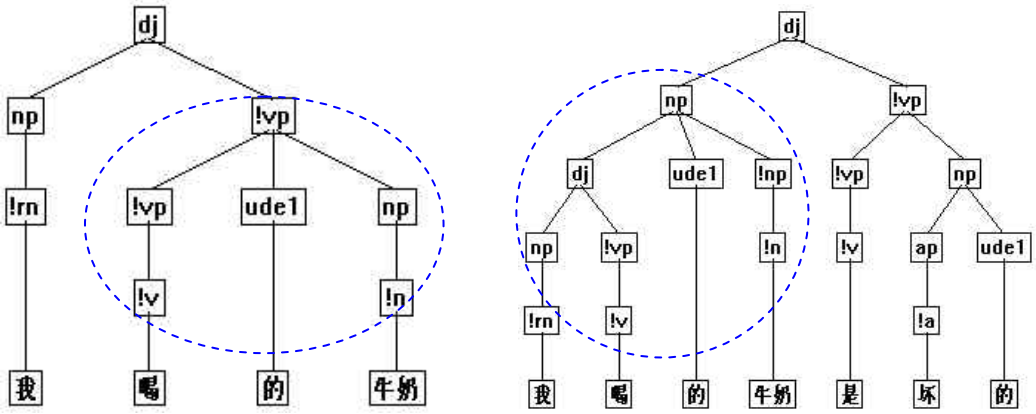


例 3：克服困难的信心

例 4：张三开车的技术

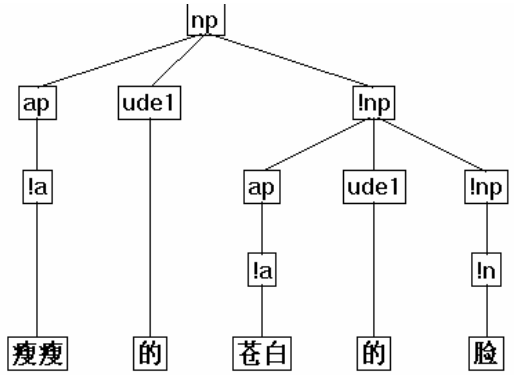


需要说明的是，以上四例的“x 的”处于定中结构的定语位置上，也可以考虑做这样的处理：如果“x 的”任何时候都不能用于独立指称，则分析为三分结构（比如“多么美妙的”，就不能独立指称）。否则，就分析为二分结构，即按照P1 模式分析为：[[x 的] y]。这时，如果“x的”具有指称性、领属性等，则将“x 的”标为np；如果“x的”主要是修饰性用法，则将“x 的”标为ap。但是，在工程实践中发现，“x 的”的功能性质不容易作出一致判定。为此，我们不得不寻找一个折中的方案：以整体功能性质的判定作为首要考虑因素，淡化对“x 的 y”内部构造的分析，统一将“x 的 y”处理为三分结构。“x 的 y”的整体功能性质判别一般不会有大的争议。上面所举的例子都是np类型的“x 的 y”，此外，也有一些vp类型的例子。比如上文已经提到的“我喝的牛奶”的例子。这个例子实际上是有歧义的，既可以理解为陈述形式（回答“你早晨喝什么了？”），这时候整个结构的功能是vp，也可以理解为指称形式，指具体的牛奶，这时候整个结构的功能是np。两种理解下具体的分析方式如下图所示[7]。



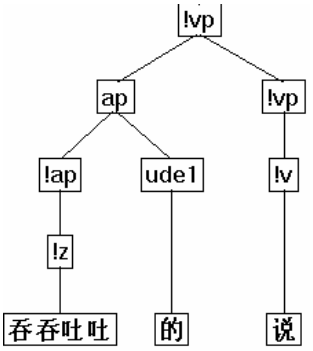
对于多个连续“X 的”作定语形成的结构，一般处理为三分结构嵌套三分结构。例如：“瘦瘦的苍白的脸”应按下图所示的方式进行分析。这里，我们不去纠缠是两个“的”字结构先并列再修饰中心 np，还是以先后叠加的方式来修饰中心 np。无论哪一种构造方式，事实上都不影响对这个 np 的语义解释。因此，我们规定，对于多个“X 的”修饰一个中心语的结构，均按照多层三分结构嵌套的方式进行结构分析。

[7] 作为陈述义理解的“我喝的牛奶”也可以处理为二分：dj[np[我喝的] np[牛奶]]，即名词谓语句（也可以认为两个 np 之间省略了“是”）。这时候，陈述义和指称义的对立，在句法结构关系上就体现为两个 np（“我喝的”和“牛奶”）之间是形成“主谓关系”还是“偏正关系”。我们不采取二分的分析方式，而是把这个结构处理为三分，是想“夸张”地突出这种结构的特性。如果处理为二分结构，容易“混迹”或者说“埋没”于其他二分结构中。



这样处理的好处是：两个“的”字结构得到的组合模式只有一个： np → ap ude1 !np

真实语料中“的”有时候的作用是状语标记（与“地”相同），这时候，“x 的”应该组合为一个单位，“x 的 y”应该做二分处理 [[x 的] y]，不应处理为三分支结构。这种情况下 y 一般都是谓词性成分 vp，而不是体词性成分。比如：



3.2.11 “是 x 的”构造

有些实例应该处理为二分结构，有些应该处理为三分结构。可以分为下面三种情况：

（i）“是 x”可独立形成结构，并且“是 x”形成的结构加“的”后能独立指称，这种情况下按照 P1 模式二分： [[是 x] 的]

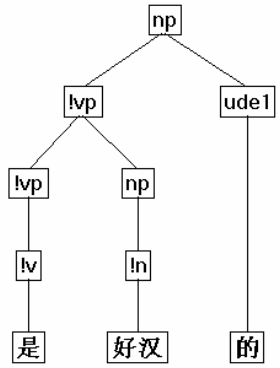
（ii）“x 的”能独立指称，这种情况下按照 P2 模式二分： [是 [x 的]]

（iii）除上述两种情况外，都按照 P3 模式三分

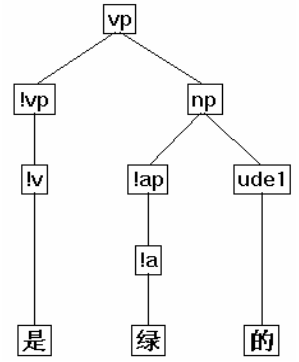
下面分别以具体例子来说明：

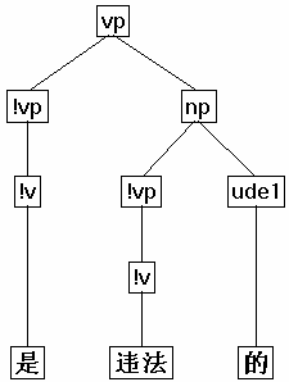
a）“是好汉的”。在“是好汉的站出来”这样的语境中，“是好汉的”指一类人，这类人具有“好汉”这种属性特征。这符合上面所说的情况i，因此应该按P1 模式二分[8]。

[8] “是好汉的”还可以出现在“这把大刀是好汉的”这样的语境中，这种情况下，“是好汉的”符合上文所说的情况ii，“好汉的”独立指称，应按P2 模式分析。

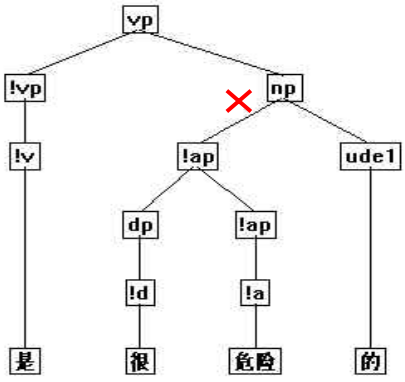


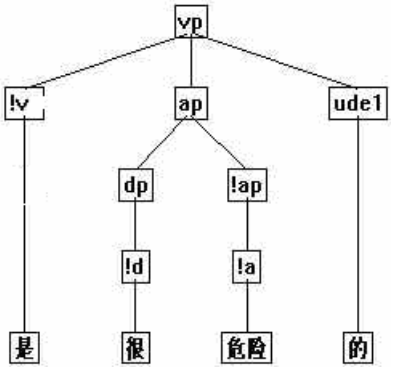
b）“是绿的”中“绿的”可以独立指称，应按 P2 模式分析；“是违法的”中的“违法的”也可独立指称（“违法的”既可指人——“违法者”，也可指事件——“违法事件”），也应按 P2 模式分析。





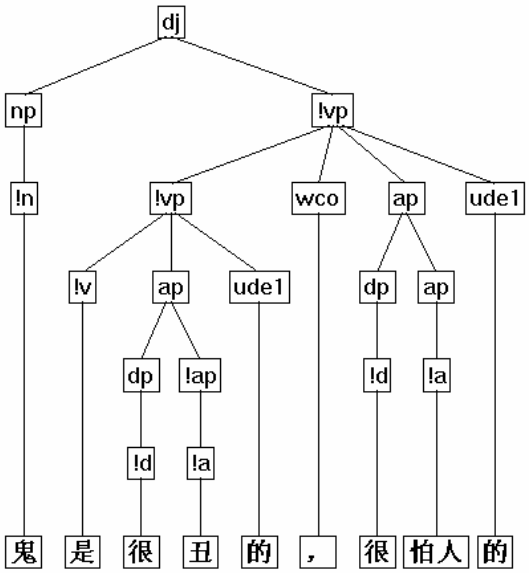
c） “是很危险的”中，“是很危险”可独立形成结构，但形成结构之后，再加“的”，不能指称（不符合情况 i），而“很危险的”不能独立指称（不符合情况 ii），因此，属于上文所说的第 iii 种情况。应该按照右图所示的分析，而不应该按左图分析。





需要特别注意的是，上面 i，ii 两种情况下，“是”是作一般的述宾结构中的述语，这个位置可以出现 vp，因此“是”要从 v 上升为 vp，再参与组合；在情况 iii 中，“是”和“的”形成特殊的框式结构（参见 3.2.13），“是”作为 v 直接参与组合，不应上升为 vp（参见 2.3）。

语料中还有“是 x 的，y 的”这样的骈合结构，比如“鬼是很丑的，很怕人的”。像这样的情况，我们处理为多分结构嵌套的形式。如下图所示：

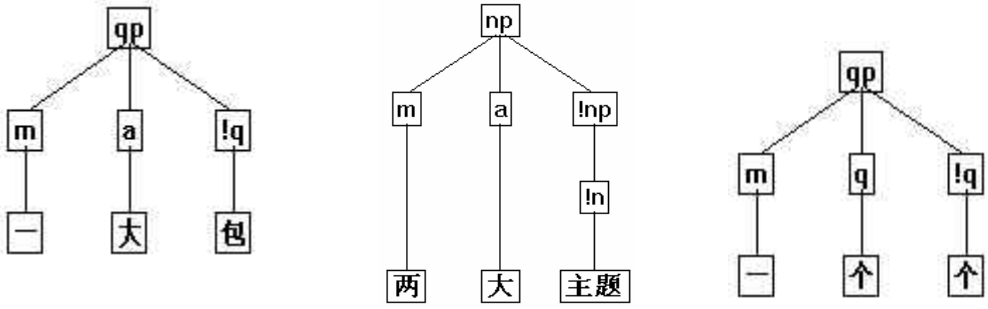


这样处理的指导原则跟上面处理多个“的”字定语修饰一个中心语的结构是一致的。

3.2.12 m + a + q 构造

类似的还有 m + a + n　　m + q + q 等，均处理为三分结构，例如：

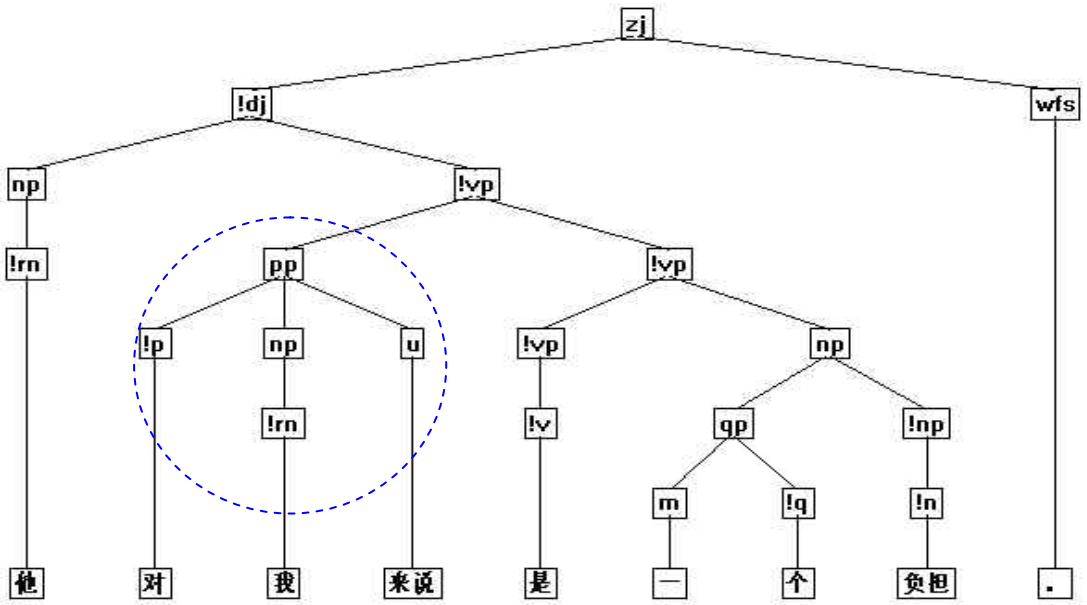
一大包（书）　　一大群（人）　　两大主题　　一个个　　一片片



3.2.13 框式结构

框式结构指的是这样的结构类型：至少有两个成分（不妨记作 X，Y）非连续共现（discontinuous co-occurrence），其他成分（不妨记作 m，n，…等）在 X 和 Y 的周围出现，形成形如“X m Y”，“X m Y n”，“m X n Y”，……等等结构，这类结构（1）不能按一般的层次分析方式，将邻近的成分组成为一个单位（比如以“X m Y”为例，“X m” “m Y”单独都不能形成结构）；或者（2）X、Y 跟它们周围的成分具有同等程度的结合力，使得无法按照一般的二分层次组合方式来进行结构分析。在框式结构中，X，Y 构成一个“框”（frame），它们周围是等待填充的“槽”（slot）。“框”是结构中的常量，“槽”则是结构中的变量。下面是一些具体的框式结构的例子。

a）比较常见的是介词形成的框式结构，例如：



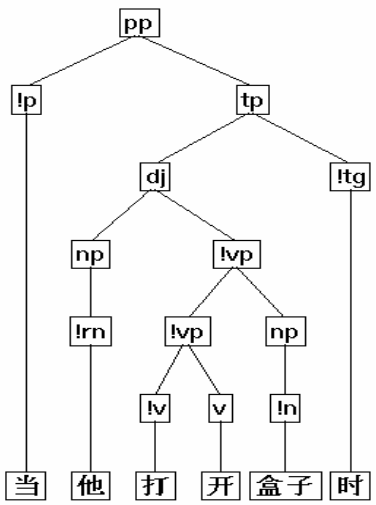
上例中“对……来说”就形成所谓的框式结构。类似的例子还有“在…下”[9]、“直到……为止”、“就……而言”、“除……以外”、“自从……以来”、“到……为止”等等。

这类由介词跟方位词、助词等搭配形成的框式结构，一般整体的功能类标注为 pp，由介词充当中心成分。

需要注意的是，并不是所有的介词形成的结构都处理为多分支的框式结构。事实上，大多数介词形成的介词结构，应该按照一般二分的模式进行层次分析，不需要处理为多分的框式结构。比如，“在 + sp” “从 + sp” “把 + np” “被 + np”等等，都应该按照一般的二分结构模式处理。此外，像“当 … 时（时候）”这样的例子，表面上看，也似乎有一个框架，即“当 + x + 时”，但这一个结构也可以按照一般的二分结构模式处理，因为“x 时（时候）”[10]都可以单独成立。比如“当他打开盒子时”，可以分析为：当 + 他打开盒子时。介词“当”的宾语“他打开盒子时”分析为tp。整个结构分析如下图所示：

[9] 指“在他的带领下”这样的例子，而不是“在桌子下”这样的例子，后者应按普通的二分结构分析，因为“桌子下”是可以单独成立的结构。而“他的带领下”是不能单独成立的结构，只能跟“在”一起形成三分支结构。

[10] 在“当…时（时候）”格式中，“时”标记为tg，“时候”标记为时间词t。

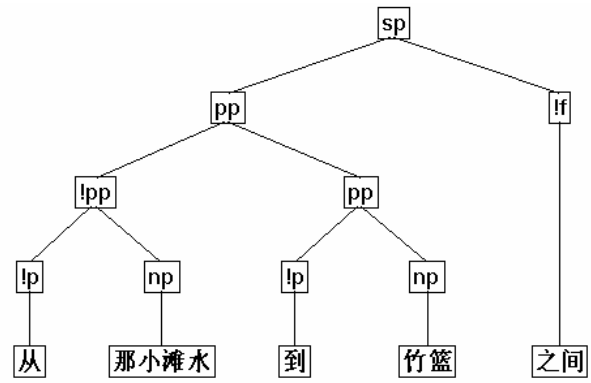


再比如，对于“从…到…”这样的结构形式，也不需要处理为框式结构。“从…到…”有以下一些具体用法情况：（1）作状语性成分。例如：“他**从八点到十点**一直在看球赛”。在状语位置上的“从…到…”结构分析为类似于动词复谓结构的组合， pp → !pp(从八点) pp (到十点) ， 中心词标在前一个 pp 上。（2）作定语性成分。例如：“在**从八点到十点**的两个小时里”。其中的“从八点到十点”处理方式同上。（3）作谓语性成分。例如：“**从量变到质变**了”。这时候“到”分析为 v，“从量变到质变”分析为 vp → pp (从量变) !vp(到质变) ，中心词标在后面的 vp 上。 这种用法情况下，“到”的前面还可以插入其他状语性成分，例如“从上午九点**一直**到下午五点”。（4）作主语性成分。例如：“**从慢到快**，是必然的一个过程”。其中的“从慢到快”也处理为 vp，结构层次分析方式同（3）。（5）在“从 … 到 …”中，“到”之前还可以出现动词，形成“从…v 到…”结构，例如：“**从早唱到晚**”。v 前还可以嵌入副词成分，例如：“从早**一直**唱到晚”。这类结构的处理方式也跟上面（3）相同。 vp → pp(从早) vp(一直唱到晚) 其中“一直唱到晚”分析为状中结构 vp → dp(一直) vp(唱到晚) “唱到晚”分析为三分支结构 vp→ v v a 。

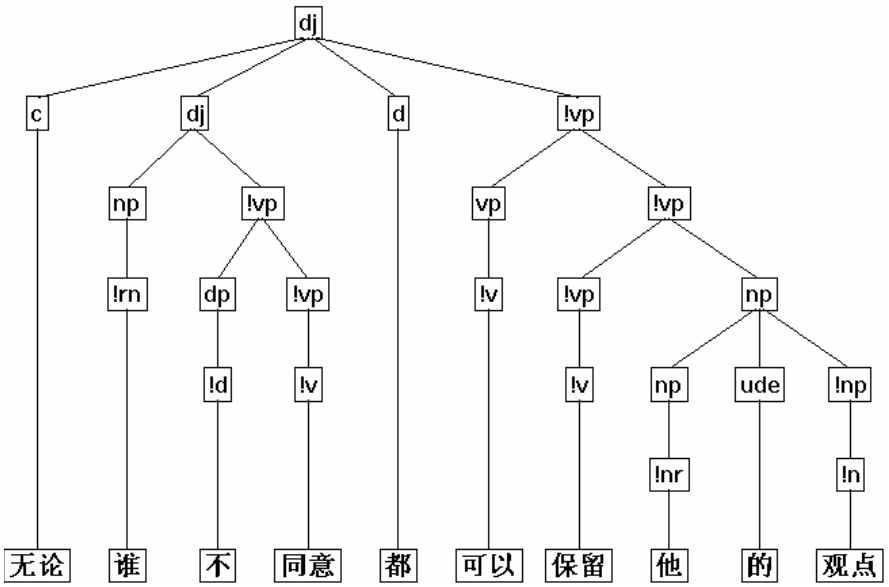
“从…到…”结构还可以出现在方位词（f）“之间”的前面，比如：

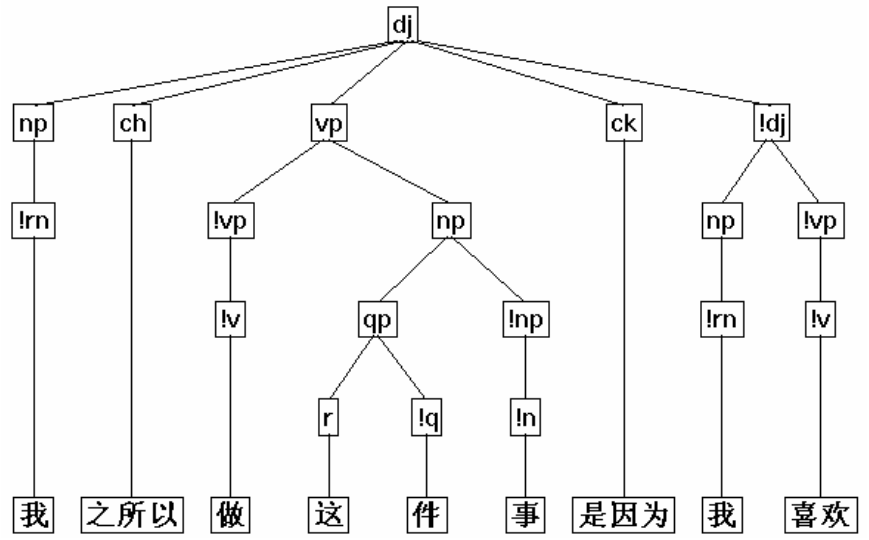
走近一看，我发现从那小滩水到竹篮之间的水泥地上有它的足迹。

这种情况下，结构分析为：

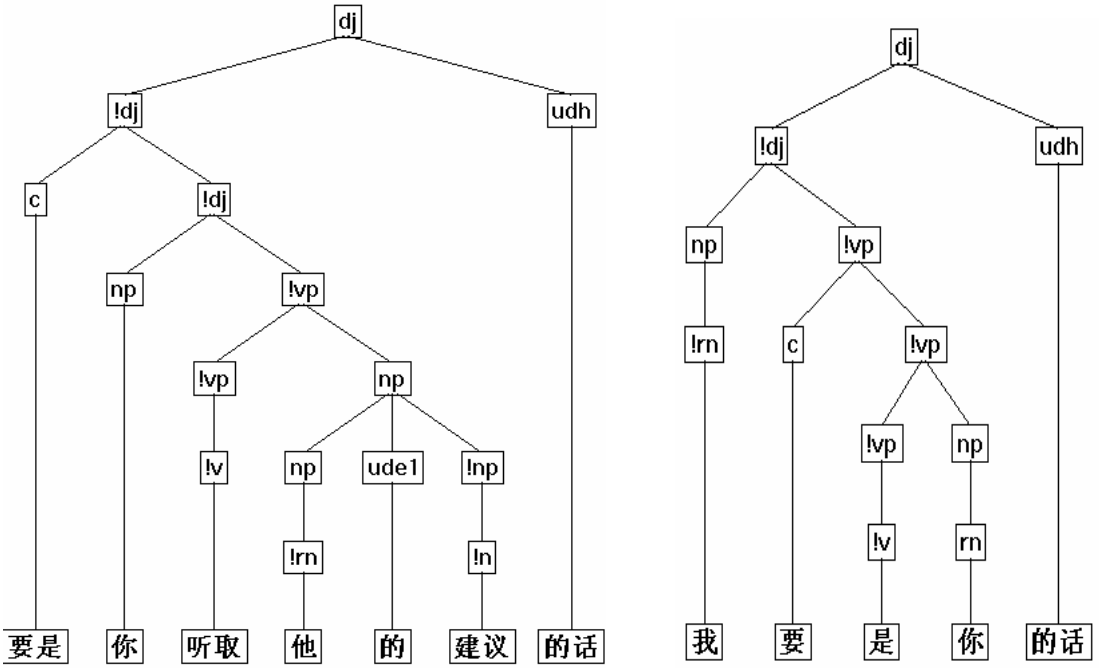


b）一些关联成分形成的连锁结构，应处理为多分支的框式结构。比如“…一…就…”，“…连…也|都…”，“…越…越…”，“无论…都…”、“…之所以…是因为…”，“为…而…”、“因…而…”，等等。



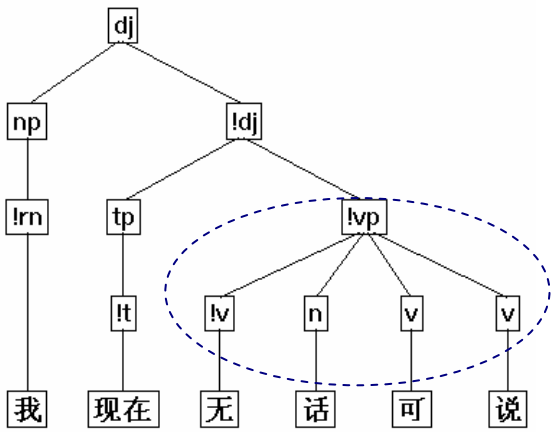


需要注意的是，有些关联成分形成的结构，尽管在使用中也经常以前后配对的方式出现，但整个结构可以分析为二分支结构，这时候就不分析为多分支的框式结构。比如“要是 … 的话”，后面的助词“的话”可以省略不用，前面的“要是”有时候也可省略不用，分别形成“要是 X”和“X 的话”结构。无论是哪种情况，都说明“要是 … 的话”可以分析为二分结构，因此不需要分析为框式结构。下面是两个例子。



“要是”可能处于主语前，也可能处于主语后的位置。如果在主语前，“要是”分析为连词 c；如果在主语后，“要是”可以分析为一个连词，也可以拆开来，将“要”分析为连词，将“是”分析为动词，这要看谓语的情况来定。在上面的例子“我要是你的话”中，应该把“要是”分开为两个词，这样以“是”为谓语动词，整个结构就比较好处理。如果是另外的例子，比如“我要是没有打好这场球……”，“没有打好这场球”可以作谓语。这样，“要是”就可以作为一个整体，分析为连词。

c）一些动词性结构，也应处理为多分支的框式结构。比如“无…可…”、“包括……在内”、“为…所…”、“非…不可”等。

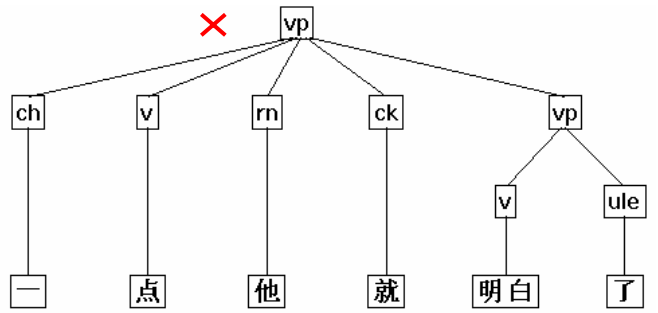


需要特别注意的是，下面这些情况，虽然从“形式”上看，也符合本规范对框式结构的要求，但在做具体的层次分析和标注时，不宜按照多分结构形式（框式结构形式）处理。比如

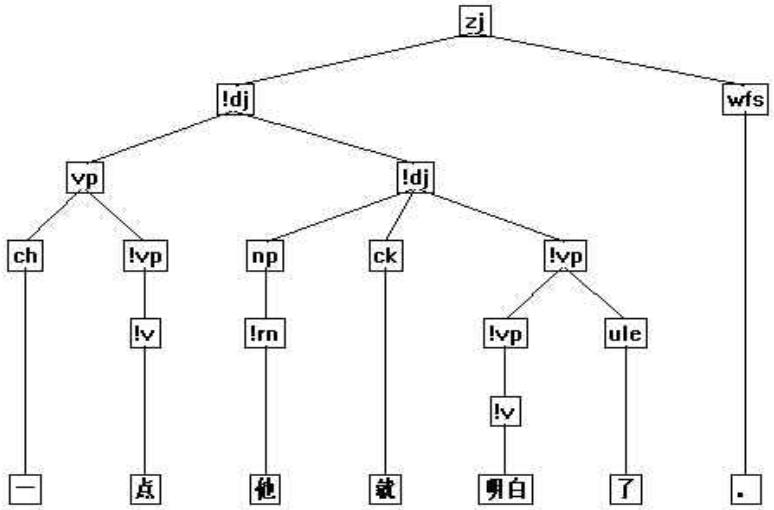
I. 像“一点他就明白了”这样的例子[11]，如果处理为五分结构，在“一”和“就”之间

[11] 这实际上是一个歧义结构。比如类似的例子“一看到他就明白了”。既可能是“他明白了”，也可能“某

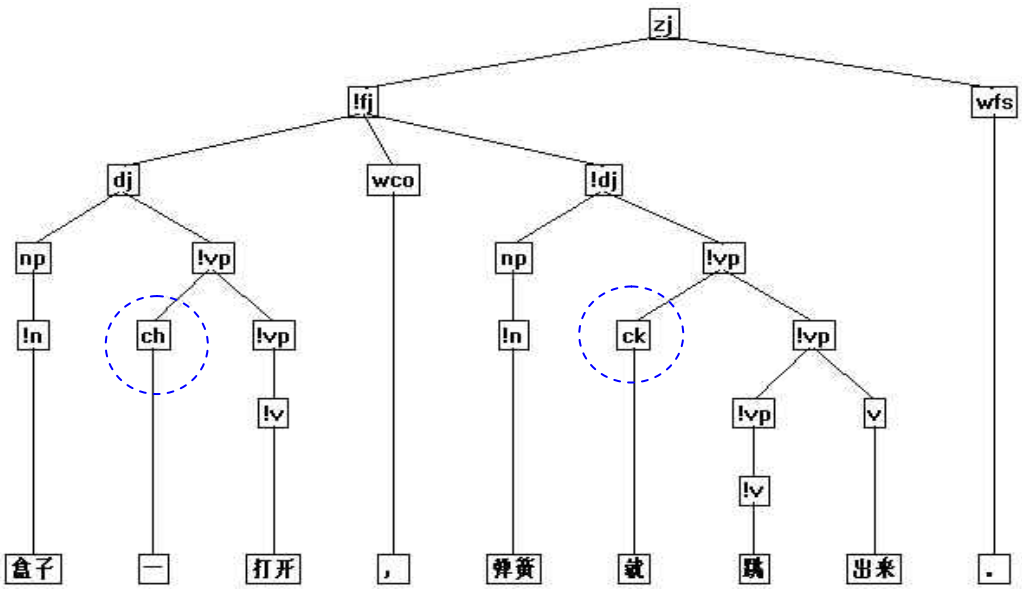
的“点 他” 就有可能被误解为一个单位（述宾式vp）。但实际上这个例子中，“点”和“他”之间并没有直接的结构关系。因此下面是错误的分析方式。



这个例子中应该按下图所示的方式进行分析（通过 ch，ck 的照应，也可以体现这个结构的“框式特点” —— 一般语法书上通常把这类结构称为“连锁结构”）：

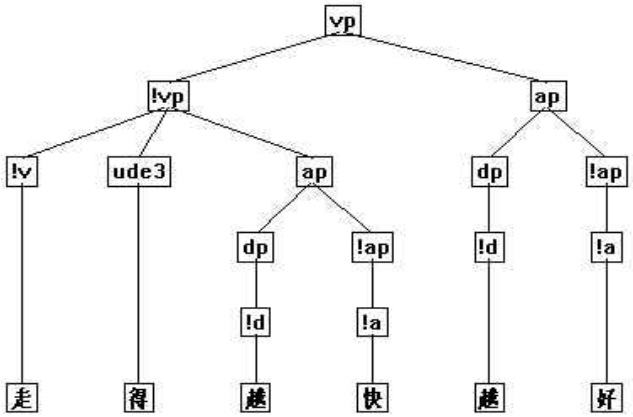


此外，如果“一…就…”这种框式结构中的“一”，“就”分属两个小句，中间有标点隔开，也不宜处理为多分结构，比如：

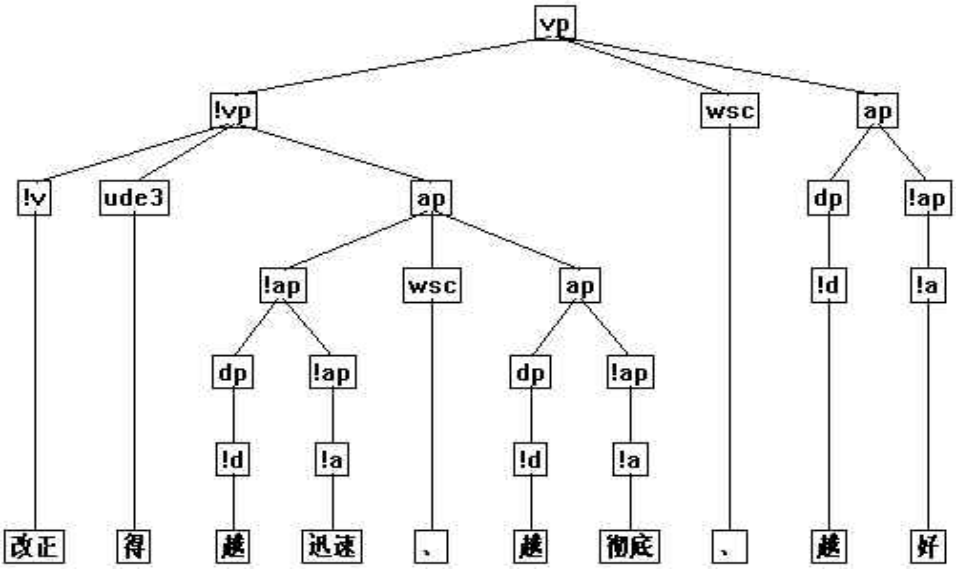


人（一看到他），某人就明白了”。

II. 像“走得越快越好”这样的例子，“越…越…”在这里并不直接构成框式结构。或者说，这是框式结构跟述补结构“杂揉”形成的假框式结构。应该按照下图所示的方式分析。



类似的例子如“改正得越迅速、越彻底、越好”，应该按照下图所示的方式分析。



3.2.14 并列结构

上文已经提及，并列结构的内部成分之间分不出组合的先后顺序，因而分析为多分支结构。这里再举一些例子。这些并列结构中都涉及到成分省略。

（1） “洗没洗衣服” “洗没洗过衣服”
其中的“洗没洗” 分析为 vp → !v d(没) v

（2） “喝没喝醉” 分析为 vp → !v d(没) vp
其中“喝醉”分析为 vp→! v v
“喝不喝得醉” 分析为 vp → !v d(不) vp
其中“喝得醉”分析为 vp → !v ude3(得) v
“吃不吃饭” 其中 “吃不吃”分析为 vp → !v d(不) v
然后是“吃不吃” 带宾语“饭”（np） vp → !vp np

（3） “喜不喜欢” 分析为 vp → !vg d(不) v

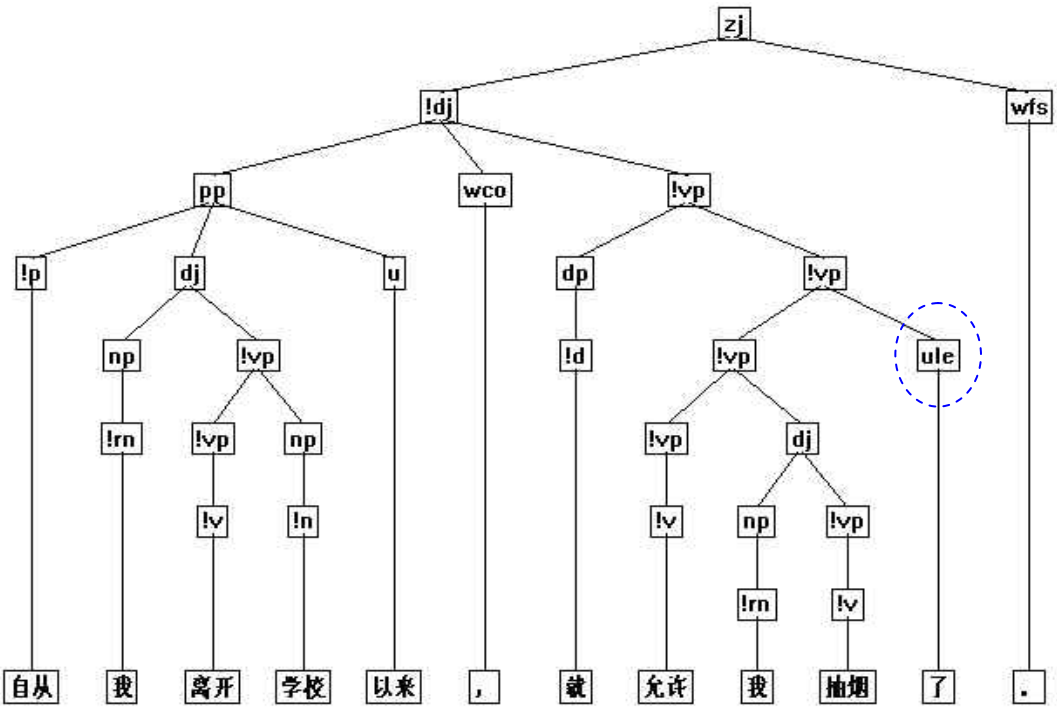
“游不游泳” 分析为 vp → !v d(不) v
“洗没洗过澡” 分析为 vp → !v d(没) vp
其中“洗过澡” 分析为 vp → !v uguo ng

此外，“复指成分”造成的所谓“同位结构”，也可以广义地看作是一种并列结构（有的语法书上把这类结构归入“偏正结构”）。比如“苦柚，那一袋苦柚，将永远留在我的记忆里。”其中“苦柚，那一袋苦柚”是“同位结构”，整个结构的组合模式可分析为：

np → np(苦柚) wco(，) !np(那一袋苦柚)

3.3 二分支结构的层次分析问题举例

3.3.1 “了”附着在前面哪一个成分上



在上例中，句末的“了”不是直接附着在离它最近的“抽烟”上，而是附着在“允许我抽烟”这个更大的单位上。“了”附着在前面哪一个单位上，反映了对“了”的语义的理解。“了”一般表示“变化”，就这个例子来说，变化的是从“不允许”到“允许”，而不是从“不抽烟”到“抽烟”，因此“了”应该附着在以“允许”为中心成分的 vp 上。

3.3.2 数量短语向前还是向后组合

在“指示代词 + qp + 名词”构造中，其中的“数量”（qp）部分应该先跟后面的名词组合，再一起跟前面的“指示代词|数词”组合。例如：

这两所学校 → np[r[这] np[qp[mp[m[两]] q[所]] np[学校]]]
每三个病人 → np[m[每] np[qp[mp[m[三]] q[个]] np[n[病人]]]]

这两例的句法结构分析可图示如下：

图 3-26

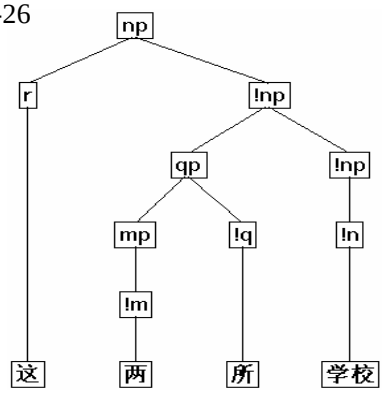
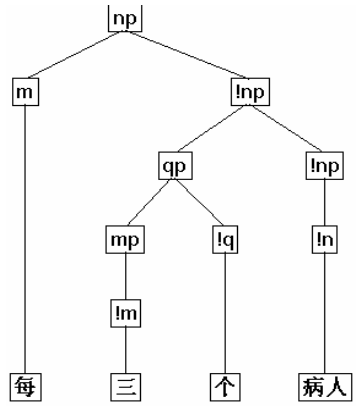


图 3-27



对这类构造的层次，可以利用范继淹先生的“并立扩展法”来加以判别。例如：

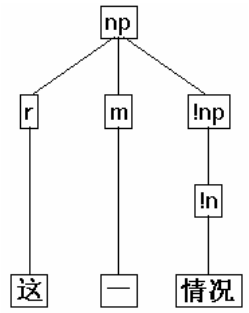
这两本书 → 这两本书，三个笔记本，四支笔

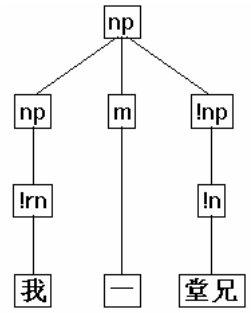
→ * 这两本，那两本 书

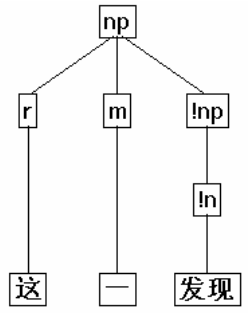
从扩展来看，组合层次应为 P2 模式。

类似的，“另 + 一个 + 人”也应分析为： 另 + 一个人。

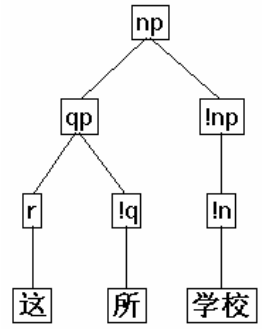
有时候这种结构中的量词也可以省略，比如“这 一 情况”、“我 一 堂兄”、“这一发现”等，这时处理为三分结构。

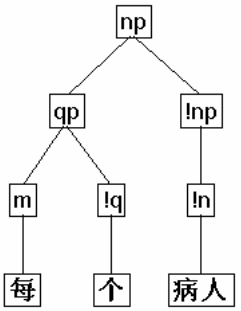






如果这种结构中的数词为“一”，则数词可以省略，形成“这 所 学校”、“每 个 病人”这样的组合，则分析为二分结构。量词先跟“这”“每”组合，再跟后面的名词组合。





3.4 标点符号在结构中的位置

3.4.1 逗号不应出现在多分支结构的末尾

图 3-28

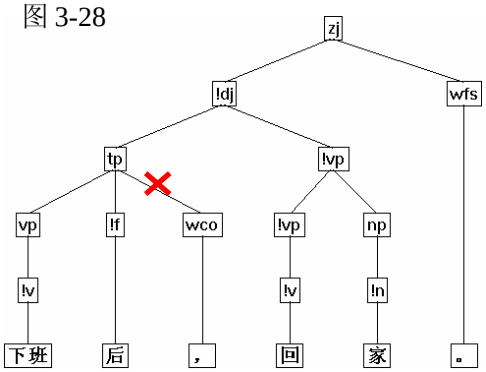
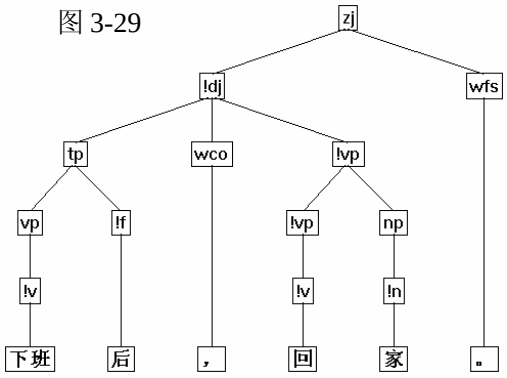


图 3-29



一般来说，逗号（wco）要么在二分结构的右子节点位置，要么在一个三分结构（或多分结构）的中间位置（如上面图 3-29 所示），而不应该在一个三分结构的最右子节点位置（如图 3-28 所示），当然，更不能出现在结构的最左子节点位置。

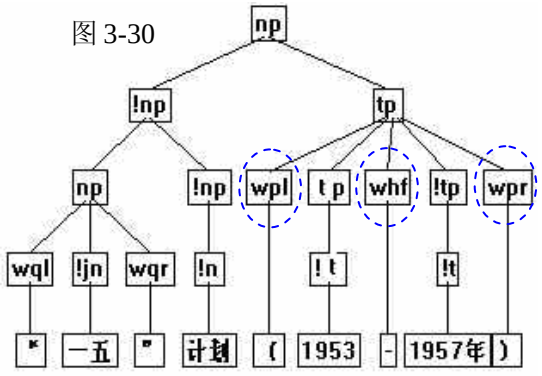
3.4.2 成对出现的标号

成对出现的标号（如括号，引号等），前后呼应，应该跟处于它们中间的成分一起，分析为多分支结构。

3.4.3 破折号、连字符、省略号

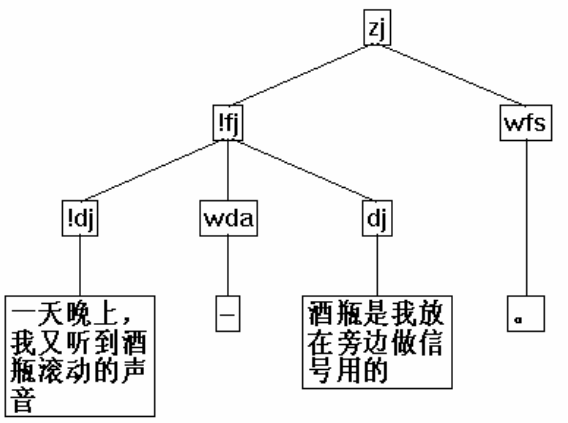
像连字符“—”这样的中置型标点，起分隔作用，也应该出现在多分支结构中，且只能出现在中间的节点位置。如图 3-30 所示，wpl，wpr 是成对的标点，whf 是中置的分隔符号。这些标点在结构层次分析中，都应出现在多分支结构中。

图 3-30



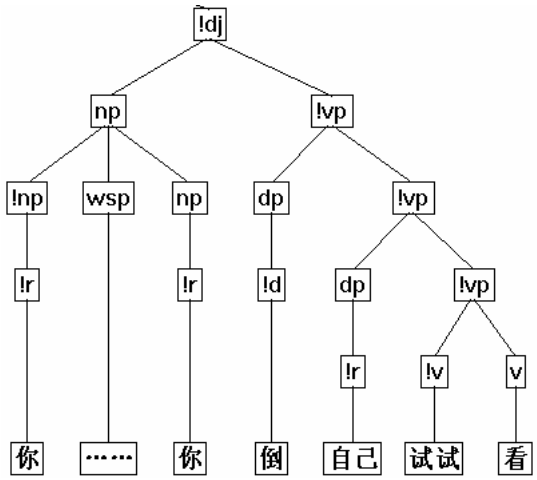
破折号“——”多数情况下起分隔作用，也属于中置型标点，一般出现在多分支结构的中间节点位置。例如：

一天晚上，我又听到酒瓶滚动的声音——酒瓶是我放在旁边做信号用的。



省略号有时候用在句末，功能相当于完句标点（见附录 2.1.2），有时候则起停顿的作用，这时候可以处理为多分支结构的中间节点。例如：

你……你倒自己试试看。



四 语法功能标注

分出大小语言单位后，进一步就需要给这些单位确定适当的功能标记。对于有系统的形态变化的语言来说，语法单位的功能差异往往通过形态标记体现出来。汉语缺乏系统的形态变化，因而语法单位的功能差异主要通过一个语法单位所能占据的结构位置来显现。

4.1 短语功能与结构位置的对应关系

下面表 4-1 给出了汉语的基础句法结构关系，也即句法结构位置的一个描述；表 4-2 给出了一个依据该结构位置体系初步确立的短语功能分类体系。

表 4-1：汉语短语的结构类型

序号	结构关系	句法结构位置		实　例
1	主谓结构	主语	谓语	树叶黄了;小明喜欢看电视;感冒传染
2	述宾结构	述语 1	宾语	喝了三杯酒;学了三年;企图逃跑;送他香烟
3	述补结构	述语 2	补语	洗干净;做得非常好;好得很;吃得完;拿出来
4	定中结构	定语	中心语 1	一斤白菜;老师的眼泪;大红灯笼;削梨的刀
5	状中结构	状语	中心语 2	快跑;认真地学习;把饭吃完;明天见;屋里坐
6	连谓结构	前项	后项	开着窗户睡觉;打电话请医生;派助手办理;请他来
7	联合结构	前项	后项	小说和戏剧;又高兴又难过;批评教育
8	附加结构	中心语 3	附加语	红着;吃了;砍光了;努力奋斗过
9	的字结构	中心语 4	附加语	买菜的;老师表扬了的;冰凉的;慢性的
10	所字结构	附加语	中心语5	所知道;所了解

表 4-2：汉语短语的功能分类[12]

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**	单句型短语		+		+										+
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	**qp**	数量短语							+							+
10	**mp**	数词短语							+							+

[12] 表中没有列出复句fj、引句yj、整句zj等的功能特征，对于这些‘大尺寸’的语言单位来说，传统上一般都是从结构上来界定，而非从功能角度来认识的。根据树库标注的需要，我们设定了这些标记，但实际上很难从功能角度界定这些单位。有关fj，yj，zj在标注中涉及的问题，详见下文 4.2.8，4.2.9，4.2.10 以及“现代汉语树库标注常见问题举例”中的具体示例说明。

a:作主语；b:作谓语；c:作述语 1；d:作宾语；e:作述语 2；f:作补语；g:作定语；h:作中心语1；
i:作状语；j:作中心语2；k:作连谓结构前后项[13]；l:作联合结构前后项；m:作中心语3；n:作中心语4；

需要注意的是，(1) 短语的结构关系跟短语的功能类型有明显的对应关系，但又并非简单的一一对应关系。比如“述宾关系”对应着 vp，ap 等，定中关系对应着 np,sp,qp 等。“主谓结构”对应着 dj，但反过来并不意味着 dj 一定是主谓结构，因为 dj 中也可以是由“状中结构”构成的（比如“以往他都在家里吃早饭”）；(2) 有不少短语的功能类型有对应的词类，比如名词性短语对应着名词，这也意味着二者的句法功能基本一致。但并非所有的短语类型都如此，比如短语的功能类型中 dj，pp，就都找不到对应的词类。

说到底，目前短语的功能类还缺乏非常严格的界定标准，人们更多的可能是通过类比来把握一个短语属于哪一类。在这种情况下，对实例的分析就非常重要。下面表 4-3 列出了各类短语的一些实例。

表 4-3：汉语短语功能类型示例（加底色的例子的类型归属需加以注意，同类例子应类推）

短语	实　　例
dj	爱夸张事实的孩子往往喜欢喜剧；三点钟全体集合；今天星期一；他二十来岁；**长两米；重三斤；**
np	粒子碰撞噪声检测仪；计算机在国外应用的现状；世界名牌服装；新问题；自己的；桌椅门窗；理想与现实；支持总统的群众；给孩子们的；服装设计；两国之间的合作；几十年的努力；他们两位；**录象带两百盘**；**最善良的一个**；**三斤重；两米宽；**
vp	把杂志放进抽屉里；进行多方面的经济结构的调整；从暴风雪中救出了一群羊；来了；请客人吃饭；去外婆家玩；烧毁证物并袭击警察；跑得我累死了；
ap	很不高兴；冷得发抖；比他们房间冷得多；干干净净的；通红通红的；亮了；干净不了三天；不礼貌而且不诚实；**长三米**；**小两岁**；
dp	飞快地；轻松而愉快地；波浪式地；
pp	关于专家系统；从桌子上；被我们；在后面；比这里；从北京到那里；除他之外；
sp	报纸上；我前面；我们班里；
tp	一个秋天的早晨；下星期一；吃饭前；
qp	两百张；三十岁；三场；多少斤；
mp	六七百；三万两千零五十；四又二分之一；五点三二；大多数；不少；几；

有关 np，vp，ap，dj 等的内部组成情况的详细描述（短语功能类型与结构类型之间的对应关系），请参看詹卫东（2000）。

作为原则，对短语功能类型的判别，应该依据**短语的整体功能性质**（占据句法结构位置的能力）来决定，而在实际操作中，比较常用的方式则是看**短语的中心成分的性质**来加以判定，短语功能类与其中心词的功能类（词类）多数情况下是一致的。比如 np 的中心成分一般也是名词，vp 的中心成分一般也是 v，等等（详见下文第五节）。

[13] 这里为表格的简单起见，把连谓结构前项位置和后项位置并作一个区别特征加以看待了。实际上这两个位置是有差异的，比如pp和ap短语都允许出现在连谓结构后项位置，但不能出现在前项位置。

4.2 短语功能定类需注意的一些问题

4.2.1 词和短语的功能偏移现象

在汉语的实际表达中，也常常可见汉语语法范畴的模糊性，比如“一切缴获要归公”这个例子，其中“缴获”实际上是指代“缴获的东西”，这种情况下，“缴获”做为动词范畴，起了本该由名词性范畴才起的作用（承担的功能），于是就出现了 r + v 组成 np 这样的句法结构。类似的例子还有“以压倒多数通过”。其中“压倒多数”是“v+m”构造，但从表达的实际意思来说，“压倒”可理解为是“压倒性”的省略，“多数”可理解为是“多数票”的省略，或者说，“压倒多数”可以理解为“压倒多数的票数”。这里介词“以”后面需要一个 np 短语，但因为省略，这个 np 只好由 v+m 来组成了。这两例的结构分析如下面图示。

图 4-1

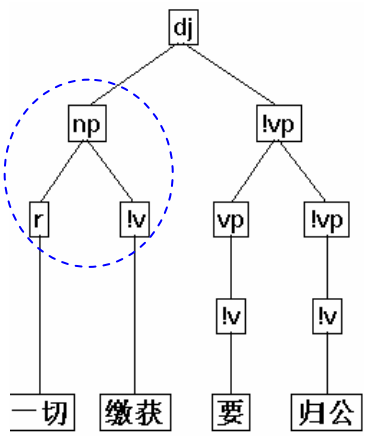
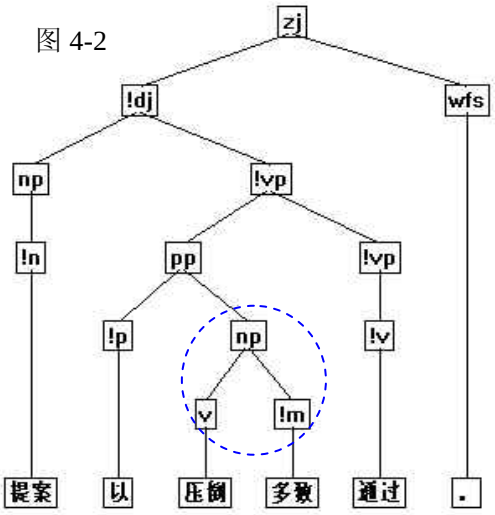
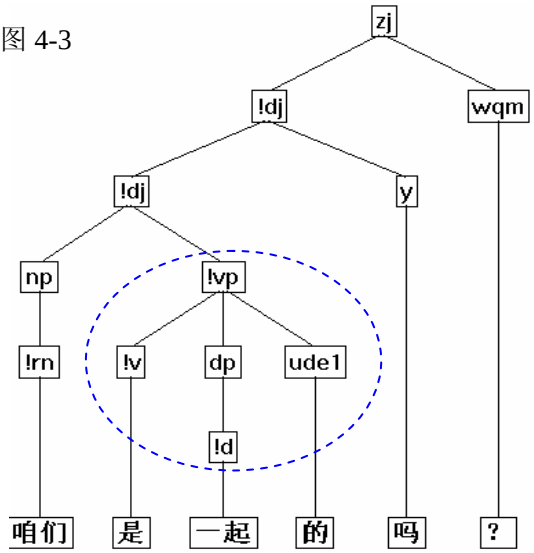


图 4-2



再比如，在“咱们是一起的吗”这个例子中，“一起”本属于副词，但因为它所修饰的中心动词省略（比如原式可能是“咱们是一起来的吗”），造成“一起”超越了副词的功能限制，在“是…的”构造中占据了一个句法位置，整个短语是 vp。

图 4-3



“一起”还可占据更多的句法位置，从而更大范围地超越副词的功能限制。比如“他们

一直在一起”，“他们终于走到了一起”。在这些例子中，如果把“一起”标为副词，句法结构就不好分析。这种情况下，我们把“一起”标为处所词 s。“在一起”“走到了一起”等结构分析为 vp → !vp sp。

以上问题可以说是在实际使用中词的语法功能偏移(之所以偏移则是由于省略了成分使然）造成的功能定性困难。在语义层面上，省略成分造成语言成分的转指（在语法层面即造成语言成分功能的偏移)。比如“两位打算怎么做？”其中“两位”是数量结构，后面省略了名词性中心成分（比如“两位老师”“两位先生”等等)，这样，就使得 qp“两位”转指原本由 np 负担的语义指称功能。

除上述由于省略造成的语法功能偏移现象之外，还有其他的功能定性困难的情况，下面再举一些例子说明。

4.2.2 “x 的”结构的功能类别

（1）“x 的”如果在主语、宾语、定中结构的中心语位置上，则标为 np。比如：

吃荤的比吃素的多　　三个男的都走了

上面例中“吃荤的”“吃素的”“男的”等都标为 np。

（2）“x 的”如果在状语、补语等谓词性位置上，一律标为 ap。

遗憾的通知您　　洗得干干净净的

上面例中“遗憾的”“干干净净的”等都标为 ap。

（3）“x 的”如果起陈述功能，单独成为一个小句，则标为 dj，此时“的”应标为语气词 yde，而不应标为结构助词 ude1。

例如：　…… 大“寿”字，陈抟老祖写的，……

上面例中“陈抟老祖写的”就构成一个陈述性的小句。因此应标为 dj。组合模式为 dj → !dj　yde

值得注意的是，上面（1）（2）（3）都是同形的“x 的”形式，这也就意味着“x 的”是有歧义的，在具体的语境中其功能类别应如何标注，应根据上面所提到的条件分别对待。比如（3）中的“陈抟老祖写的”如果在主宾语位置上，则应标为 np，而不是 dj。

4.2.3 “x 的 y”结构的功能类别

“x 的 y”结构比较常见的是体词性（指称功能）用法。当 y 为 np、qp、ap、vp 时，“x 的 y”的功能类型一般就是 np，例如：

这本书的出版 → np [np[这本书]　ude1[的]　vp[出版]]

我借的那本　→ np[dj[我借]　ude1[的]　!qp[那本]]

（中心温度是）表面温度的三千倍　→ np[np[表面温度] ude1[的] !qp[三千倍]]

而如果 y 是由体词性短语成分 tp，sp，mp 等充当时，“x 的 y”的功能类一般按照 y 的类别来标注，例如：

总收入的 20%　→ mp [np[总收入] ude1[的]　!mp[20%]]

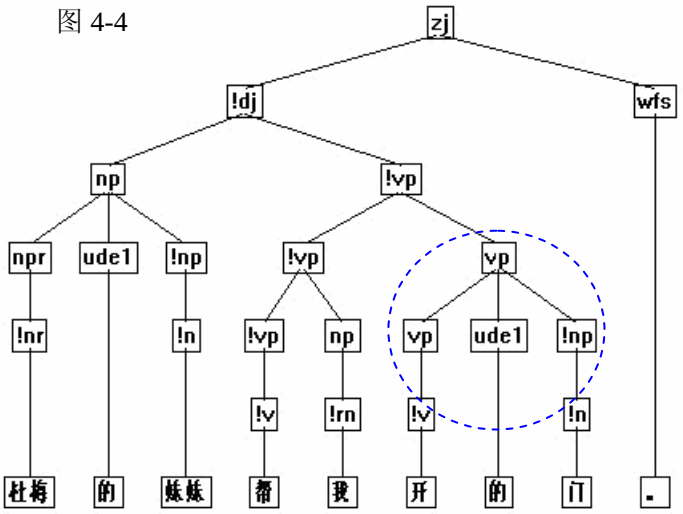
中心成分的功能类是 mp，“x 的　y”的整体功能分析为 mp。

除上述体词性的“x　的　y”结构外，上文第三节已经举过“喝　的 牛奶”这样的谓词性（陈述功能）的例子。下面再举一个 vp 类型的“x　的　y”结构的例子。

比如：杜梅的妹妹帮我开的门

其中“开 的 门”显然是在陈述，而非指称。因此应分析为 vp 短语。如下图所示：

图 4-4



4.2.4 联合式结构的功能类别

联合式短语的功能类一般跟构成成分的功能类一致，比如：

唱歌跳舞 → vp[vp[唱歌] vp[跳舞]]

但有时候如果联合式短语整体的功能跟组成成分的功能相差比较大，则应该以整体所能承担的功能为准，来判别其功能类。比如：“东”、“西”、“南”、“北”是方位词，但四个词组合形成的联合式短语“东西南北”却是一个 np。

我们在判断联合式短语的功能类别的时候，总的原则仍然是，要根据整个短语所能承当的功能去判断（即看它在句子中占据什么样的句法位置），而不能仅仅依据组成成分的功能来判别。

4.2.5 “np + qp”结构的功能类别

有可能是主谓结构（dj），也有可能数量定语后置的定中结构（np），分述如下：

（1）如果 np 和 qp 的语义关系构成“话题 — 评述”关系，则标为 dj。比如：

小王 七十五公斤 → dj [np[小王] qp[七十五公斤]]

张三 五千元 → dj [np[张三] qp[五千元]]

这本书 两毛 → dj [np[这本书] qp[两毛]]

对于上面这种结构类型，np 和 qp 不能颠倒顺序形成 qp 修饰 np 的结构，比如：

小王 七十五公斤 → * 七十五公斤 小王

张三 五千元 → * 五千元 张三

这本书 两毛 → * 两毛 这本书

事实上，这种结构一般都是单独成句。

（2）如果 np 和 qp 的语义关系构成“数量限定 — 对象”关系，则标为 np。比如：

（建成）创新基地 十五个 → np [np[创新基地] qp[十五个]]

（招收）服务员 十名 → np [np[服务员] qp[十名]]

这种情况下，np 和 qp 可以颠倒顺序形成 qp 修饰 np 的结构。

创新基地 十五个 → 十五个 创新基地

服务员 十名 → 十名 服务员。

这种结构一般都出现在宾语位置，而不是单独成句。

4.2.6 “qp + qp”结构的功能类别

“qp + qp”可能形成主谓结构（dj），也可能形成并列结构（qp），比如：

（1） dj → qp !qp

例如： 每秒 三百米

应分析为主谓结构，标为 dj。

（2） qp → !qp qp

例如：一斤 三两

应分析为并列结构，标为 qp

从理论上说，“一斤 500 克”可能有歧义，既可以理解为主谓结构（dj），也可以理解为并列结构（qp）。

4.2.7 “tp + vp”结构的功能类别

vp 前面的时间词性成分（tp）和处所词性成分（sp）是作状语还是作主语？语法学界一直都有争论。

在树库标注中，如果将 vp 前面的 tp（或 sp）分析为状语，则组合模式为：vp → tp !vp。如果将 vp 前面的 tp（或 sp）分析为主语，则组合模式为：dj → tp !vp

本规范暂时约定，vp 前面的 tp（sp）一般情况下分析为主语，即按照 dj→tp !vp 组合模式来分析这类结构。

对于像“明天见、屋里坐”这样的组合，应分析为 vp → tp !v ， 即 tp 作状语，此时动词不上升为 vp，以 v 的身份直接参与组合。

对于 tp + dj， tp + fj 这样的组合形式，一般也按照中心成分的短语类来标记整个短语的功能类，即处理为：dj → tp !dj ， fj → tp !fj

4.2.8 语篇成分（yp）的标注

句子中的某个语言片段在结构分析时如果很难被包含在它所处的上下文中，成为该上下文所表达的句法结构中的一个自然的组成成分，那么，该语言片断尽管身在句子线性排列中，但它所起的功能已经超出了句子的范围，是篇章层次上的构成成分。这样的语言片段在目前的树库标注中一般就判定为“语篇成分”。句子中的语篇成分不能跟其他一般成分一样参与句法结构的分析。我们的处理方式是：

（1）跟一般成分一样，对这种成分进行内部结构层次分析；

（2）对整个成分不进行功能分类，而是笼统地标记为语篇标记。

语篇标记具体分为下面两种情况：

（I）一般的插入语成分，标为 ypc，如下面图 4-5、4-6 的示例。

（II）呼语成分，标为 yph，如下面图 4-7、4-8 的示例。

图 4-5

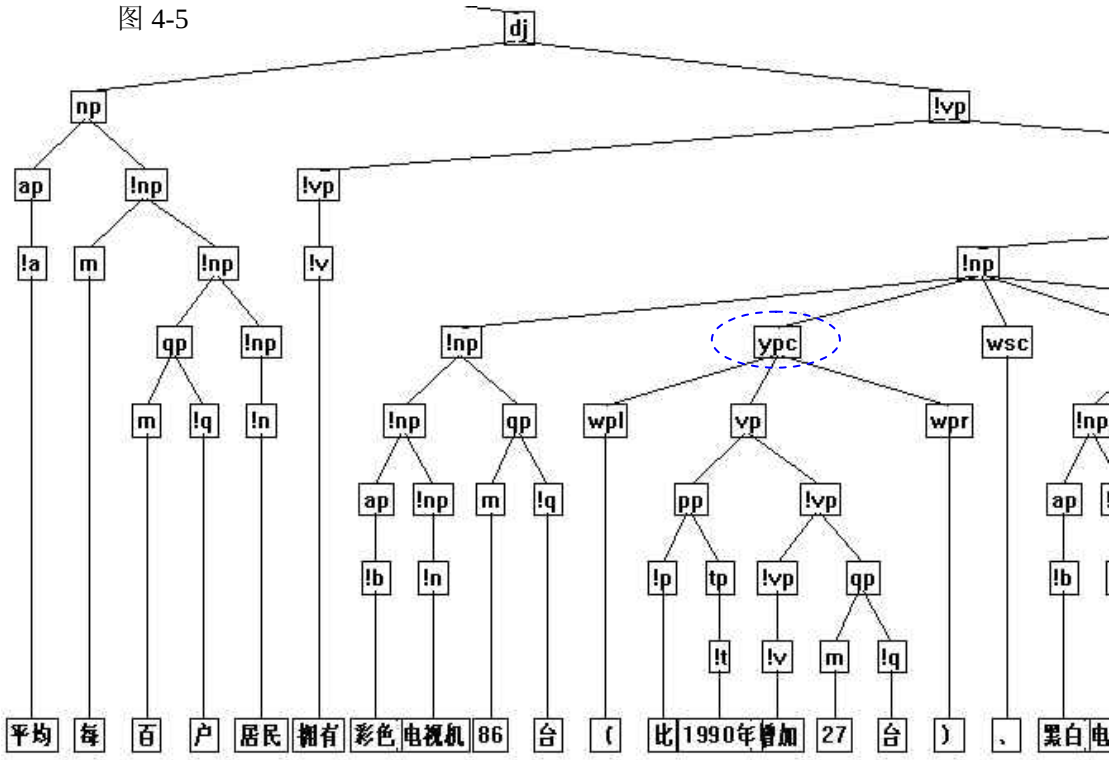

dj
np
!vp
ap
!np
!vp
!a
m
!np
!v
!np
qp
!np
!np
ypc
wsc
m
!q
!n
!np
qp
wpl
vp
wpr
!np
ap
!np
m
!q
pp
!vp
ap
!b
!n
!p
tp
!vp
qp
!b
!t
!v
m
!q
平均
每
百
户
居民
拥有
彩色
电视机
86
台
(
比
1990年
增加
27
台
)
、
黑白
电

图 4-6

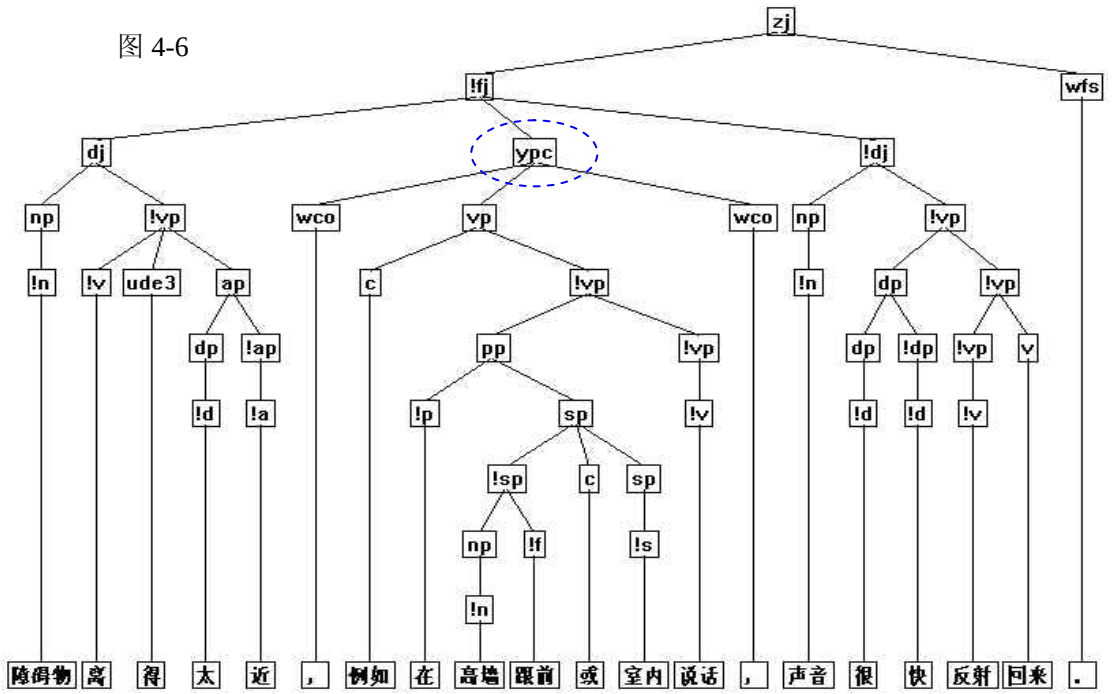

zj
!fj
wfs
dj
ypc
!dj
np
!vp
wco
vp
wco
np
!vp
!n
!v
ude3
ap
c
!vp
!n
dp
!vp
dp
!ap
pp
!vp
dp
!dp
!vp
v
!d
!a
!p
sp
!v
!d
!d
!v
!sp
c
sp
np
!f
!s
!n
陈得物
离
得
太
近
，
例如
在
高墙
跟前
或
室内
说话
，
声音
很
快
反射
回来
。

图 4-7

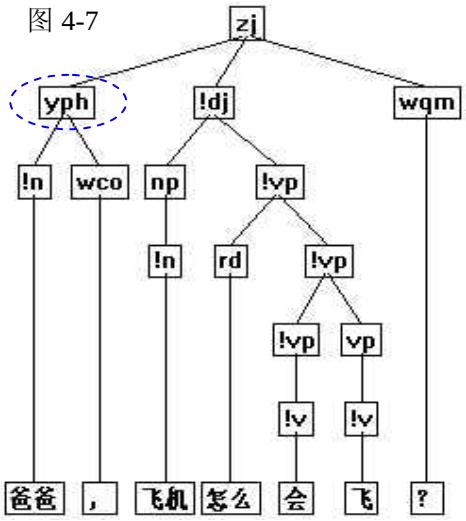
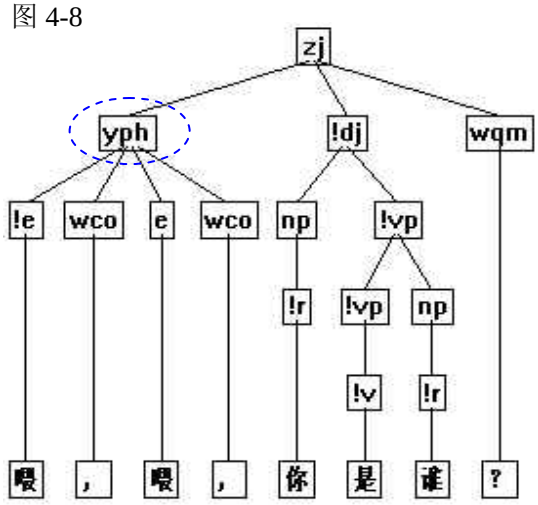
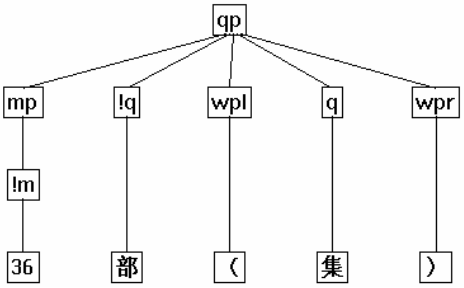


图 4-8



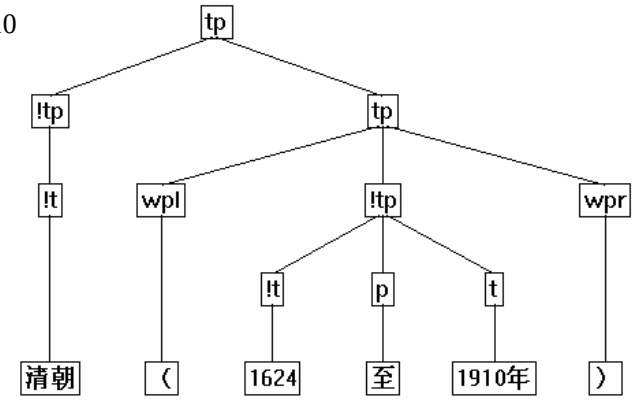
一般来说，插入语成分和呼语成分在句子中会有括号、逗号、破折号等分隔标记出来，可作为判别时的参考。但总的来说，判别时的主要依据还是要看该成分在句中的功能地位。原则上，句法功能明确、不影响整句句法结构分析的成分，一般不宜处理为插入语。比如：“完成电视剧 36 部（集）”，其中“36 部（集）”整体的句法功能就是一个 qp，处理为如下图所示的多分结构：

图 4-9



类似的例子还有“清朝（1624 至 1910 年）”，其整体功能是 tp，“清朝”是 tp，括号内的“1624 至 1910 年”也是 tp，处理为如下图所示的多分结构：

图 4-10



像上面这些括号中的成分，从功能（指分布意义上的句法结构功能，而不是语义功能、语篇功能）上说，是能够合并到前面的成分中的，都不应做插入语处理。

呼语成分一般出现在对话中，通常在句子的前面，一般后面有一个停顿（书面上用标点符号如逗号、感叹号把它们与句子的其他成分隔开），相对比较容易识别。但有的成分一般不作为“称呼”使用，从表达上看，有明显的传递语气信息的作用，这类成分不宜处理为呼语。比如下面图 4-11 的处理就欠妥，而应该按照图 4-12 的方式进行分析：

图 4-11

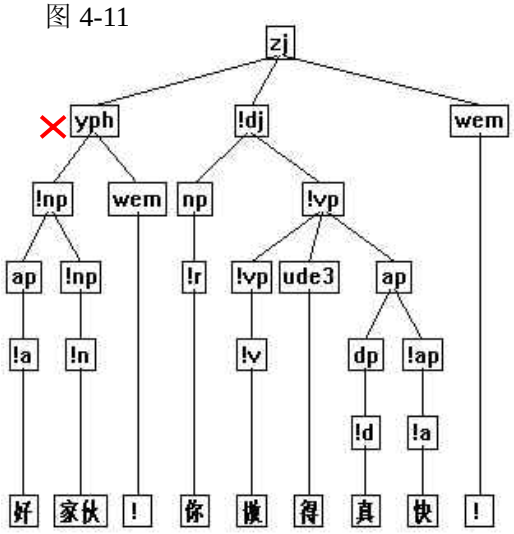
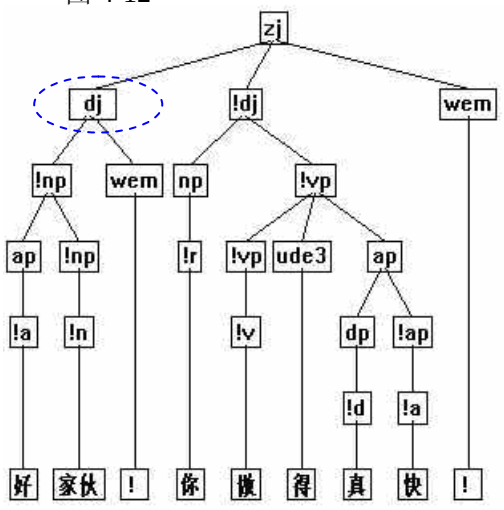
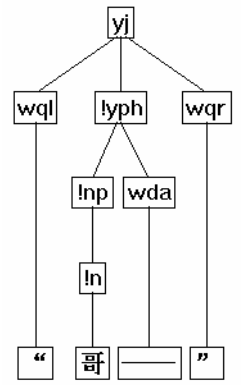


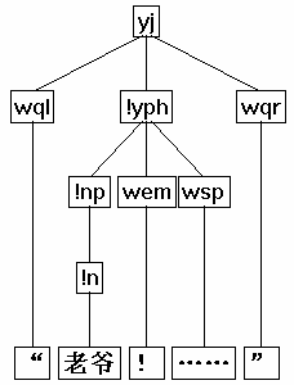
图 4-12



在上面的例子中，“好家伙”并不是作为（称）呼语来使用的，而是传达了说话人的一种情绪。可以分析为名词性成分形成的感叹句，以 dj 标记（目前树库标记中尚未特别为感叹句设立一个标记）。

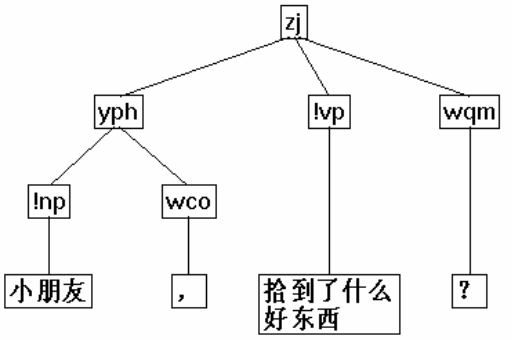
有些呼语成分在断句的时候作为一个句子处理了，如：“哥——”，“罗市长——”，“老爷！……”这样的例子。这些呼语在行文中处于引号之中，独立成句，后面没有跟其他的语言成分。这种情况下 yph 作 yj 的中心成分。如下图所示：





有的时候，还会出现连续多个呼语紧挨着出现的情况。比如“妈妈，我的好妈妈，……”。这时候处理为呼语并列的结构，组合模式为： yph → !yph yph。

一个成分如果被标记为 yp 成分，则意味着该成分不再参与句法结构的组合分析。有时候语篇成分跟句法结构成分之间的界限似乎不是很明晰，比如“小朋友，拾到了什么好东西？”其中“小朋友”是呼语 yph，但似乎又可以分析为后面 vp“拾到了什么好东西”的主语。对于这种情况，我们的原则仍然是 yp 成分不参与句法结构组合。在上例中“小朋友”分析为 yph，不再跟后面的 vp 形成 dj。这个句子应视作省略了主语，直接由 vp 加问号形成的 zj。结构分析如下：



这样分析，实际上体现了作呼语的成分与句法上作主语的成分之间的差异。试比较：

（1） 小朋友，拾到了什么好东西？

（2） 小朋友拾到了什么好东西？

上面例 1 是说话人和听话人之间的对白，“拾到…”前面可以补出省略的主语“你”。例 2 则很可能用在描述性的语境中，比如说话人指着一幅图画对听话人说例 2 这个句子，“小朋友”在这种语境下不能分析为呼语成分，只能分析为主语，后面还可以补出复指成分“他”。在口语中，呼语和主语在停顿的长短上有区别，在书面上，呼语后面一般有逗号隔开，也因为如此，我们把逗号作为 yph 的一个组成成分，跟 np 结合在一起形成 yph。

需要说明的是，暂时将这些语篇插入成分独立出来，一方面是简化树库标注的工作，另一方面也便于今后进一步集中来研究它们的分布特点。上述处理方式实际上体现了我们对目前树库加工工作性质的认识：(1) 分析的对象是句子；(2) 分析的对象是结构上相对比较“规矩”的句子。因此，凡超出句子范围之外的成分，或者虽然身在句中，但性质不属于句子层面的成分，都可以暂时“被忽略”（目前想处理也处理不好，不如暂不处理）。

4.2.9 复句（fj）的标注

如前所述，复句实际上不是短语的功能分类结果。树库标注实践也表明，对于复句的判别，有一定的主观性，目前也很难做具有严格客观标准的硬性规定。这里只对有标记的 fj 做如下规定：对于有成对关联词出现的句子，一般标为 fj，同时要保证前置关联词 ch 和后置关联词 ck 处于树结构的同一层级上。例如：

图 4-13

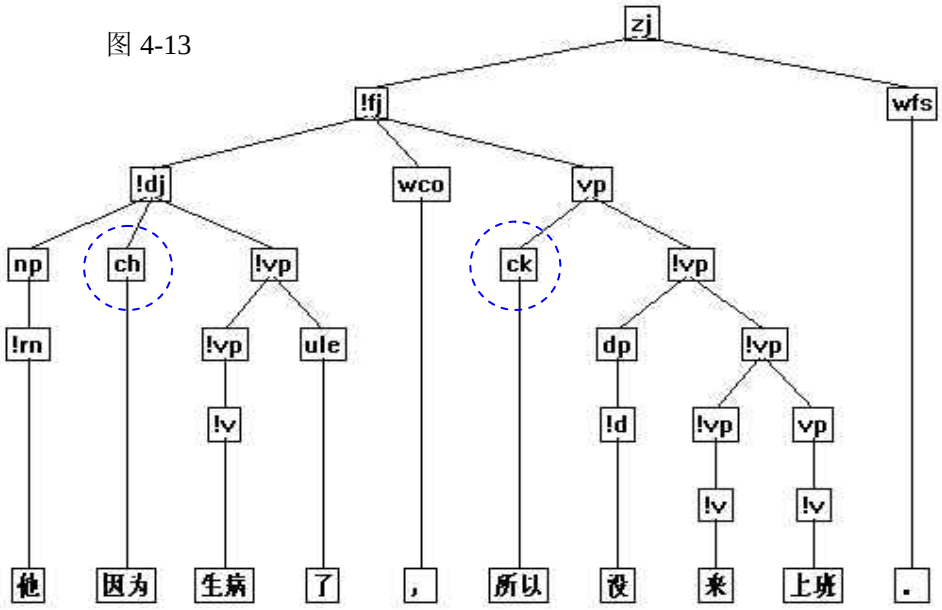
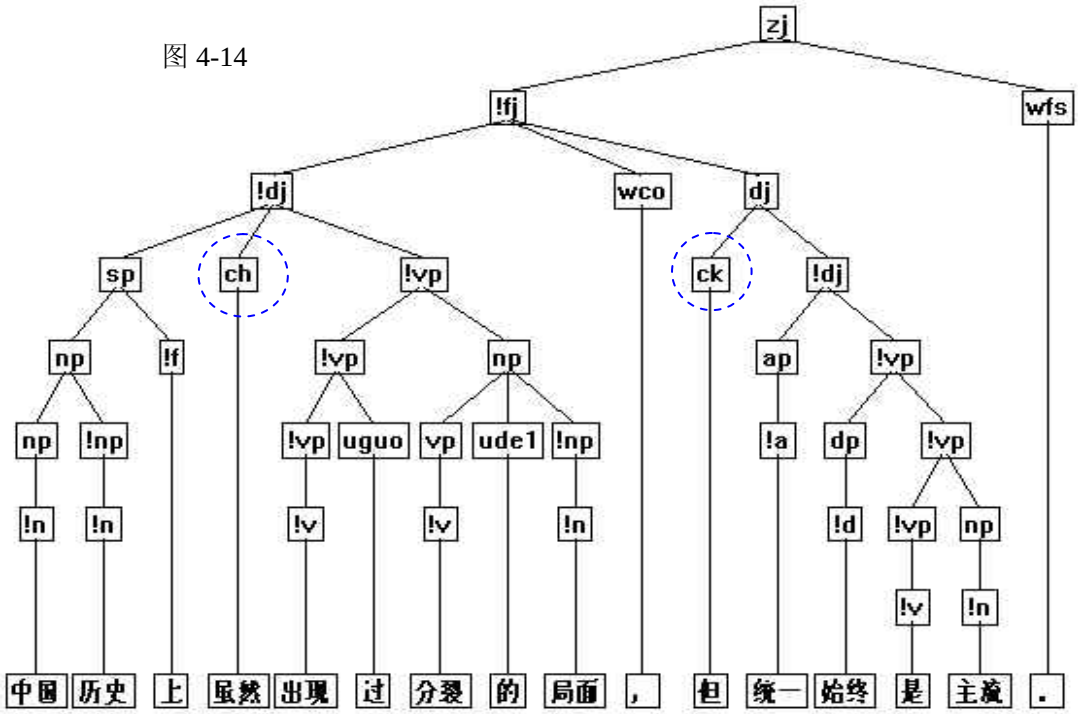
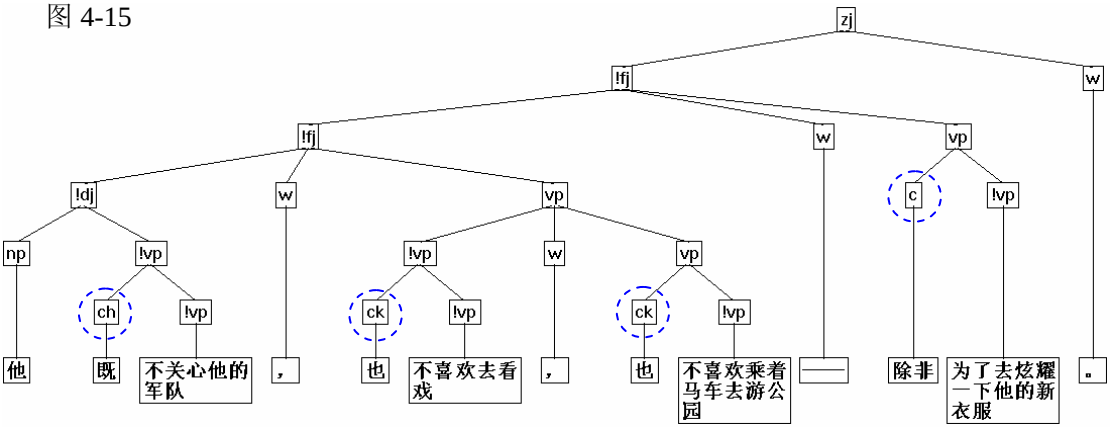


图 4-14



在实际语料中，ch，ck 并不总是一一配对出现，可能出现一个 ch，后面多个 ck 的情况，也可能出现多个 ch 跟一个 ck 配合的情况。对于这类有多个关联词的复杂的 fj，在标注结构层次时，应尽可能根据小句之间的关系紧密程度，分出层次来（一般不宜笼统地作多项并列处理）。下面图示的例子包含了四个关联词。

图 4-15



他**既**不关系他的军队，**也**不喜欢去看戏，**也**不喜欢乘着马车去游公园，**除非**为了去炫耀一下他的新衣服。

一般来说，fj 内部都包含有标点。但反过来，包含标点，不一定分析为 fj，也可能需要分析为 dj。dj 内部结构关系通常有“主谓”、“状中”两种，后者一般是出现在句首的时间、处所等状语性成分，比如“在南方，夏天是很潮热的”，句首的介词结构“在南方”应分析为 dj 中的状语性成分。如果一个句子内部不能分析为这种两种结构关系，而是“因果、假设、条件、转折、并列、递进、……”等等常见的复句关联关系，且内部包含标点，则该句子应标注为 fj。

4.2.10 引句（yj）的标注

yj 用于标记由引号“ ”引起来的句子，一般是某人说的一句或若干句完整的话。显然这是从结构特征角度对语言单位的一个定性，而非功能分类的结果。

语料中常见的 yj 有下面三种情况，均在对话的语境中出现：

（1） 独立 yj： 例如： “我买的东西，为什么让给你？”

（2） a. yj 在后： 例如： 姑娘说：“老大爷，那是令箭荷花。”

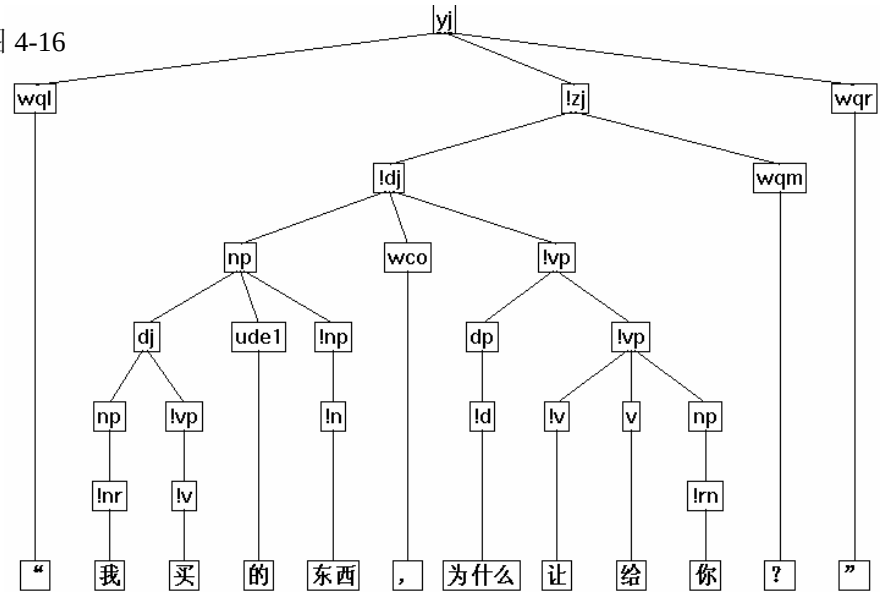
b. yj 在前： 例如： “老大爷，那是令箭荷花”姑娘说。

（3） yj 分列两端： 例如： “是吗？”魏王信不过自己的耳朵，问道，“你有这样的本事？”

其中（1）是由左引号“开头，右引号 ”结尾的完全引句；（2）a 是引句在言说类动词后的典型用法；（2）b 在文本中也并不少见。2b 跟 2a 顺序刚好颠倒。可采取类同的分析方式；（3）在书面文本中也并不少见，两个有关联的引句被某个语言成分隔开成了两段。

对于上面三种情况，分析方式如下面图示：

图 4-16

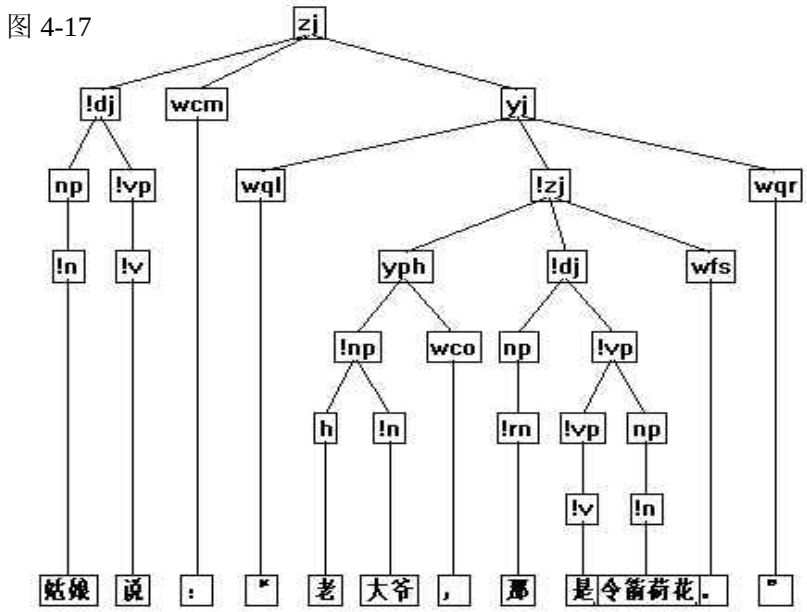


按照这种分析方式，yj 的第一种用法，可以形成的组合模式为：

yj → wql !xp wqr

其中 xp 可以是 zj，dj，vp，…… 等等成分

图 4-17



按照这种分析方式，上面 yj 的第二种用法（2a），就形成如下的组合模式：

zj → !dj wcm yj

类似的，顺序刚好跟 2a 颠倒的 2b 形成的组合模式为（图示从略）：

zj → yj !dj wfs

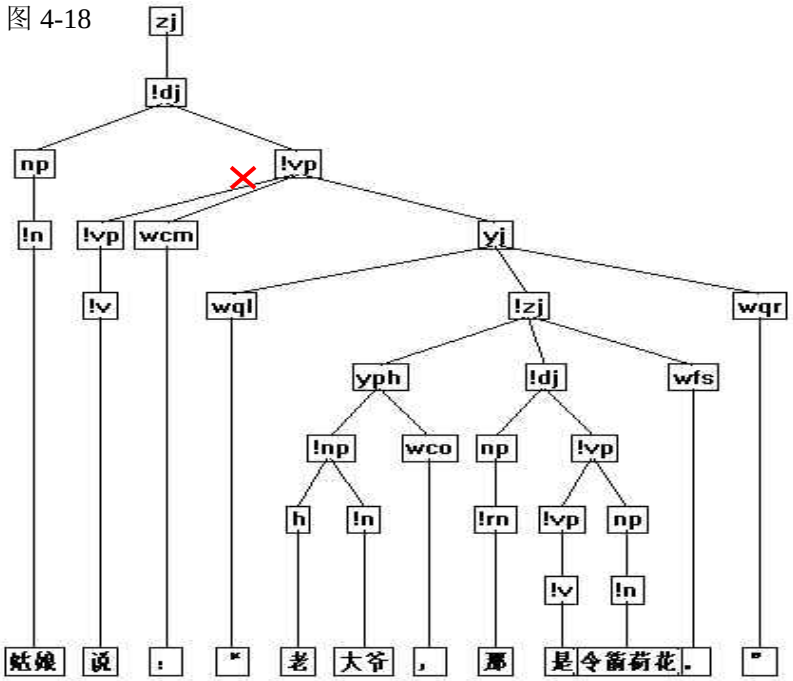
像例 2b 这样的例子，有时候可能在“姑娘说”前面出现逗号：

（2b’） “老大爷，那是令箭荷花”，姑娘说。

这时候组合模式应该参照 2b 的处理方式。具体规则为：zj → yj wco !dj wfs。

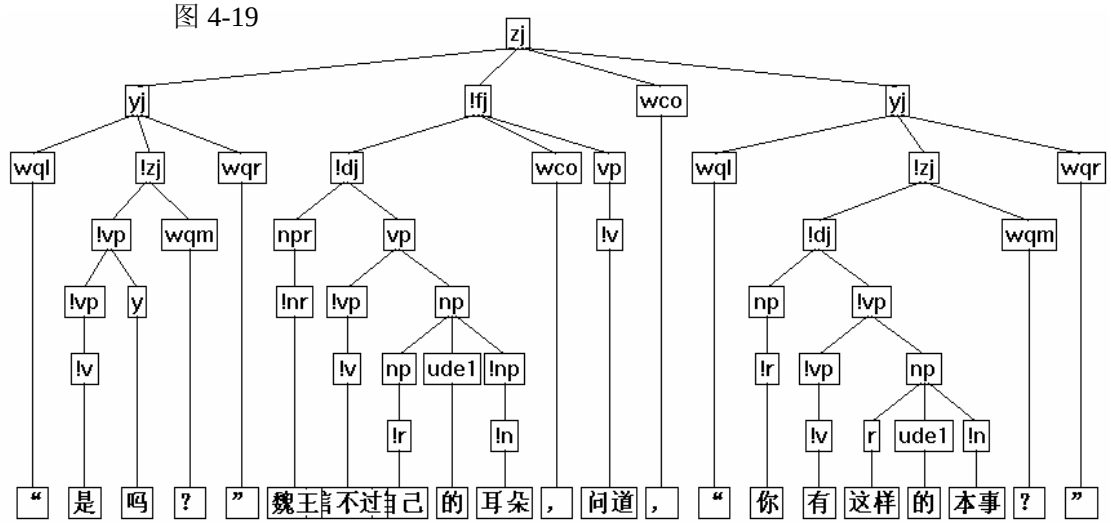
需要注意的是，例 2a 不宜分析为：

图 4-18



yj 一般应该跟 dj，fj 等在同层进行组合，并且居于比较靠近 zj 的层级。

图 4-19



按照这种分析方式，上面 yj 的第三种用法就形成如下的组合模式：

zj → yj　!xp　wco　yj

这一组合模式体现了 yj 被分隔为两段，分列两侧的结构特点。

以上例子中 yj 中都只包含了一个 zj。在语料中也存在 yj 中包含多个 zj 的情况。如果一个 yj 中包含的 zj 过多（超过 5 个），应将该 yj 断为多个句子（分为多个文本行）进行处理（参见“附录二：现代汉语文本断句的操作标准”）；如果一个 yj 中包含 5 个以内的 zj，则将这些 zj 并列，和首尾的引号一起构成 yj。下面例子中，我们不取图 4-20A 的标注方式，而是采用图 4-20B 的标注方式。

图 4-20A

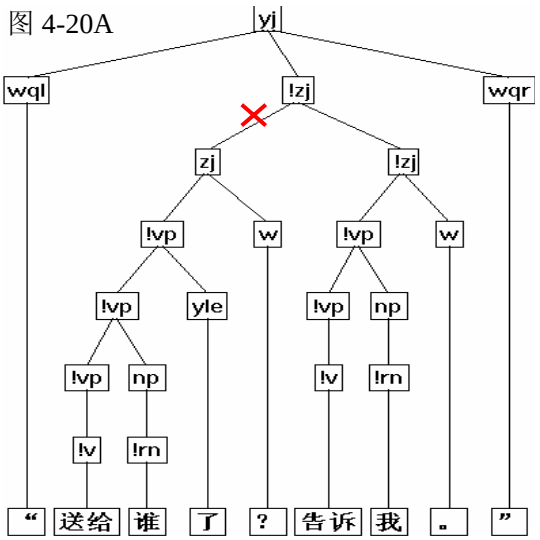
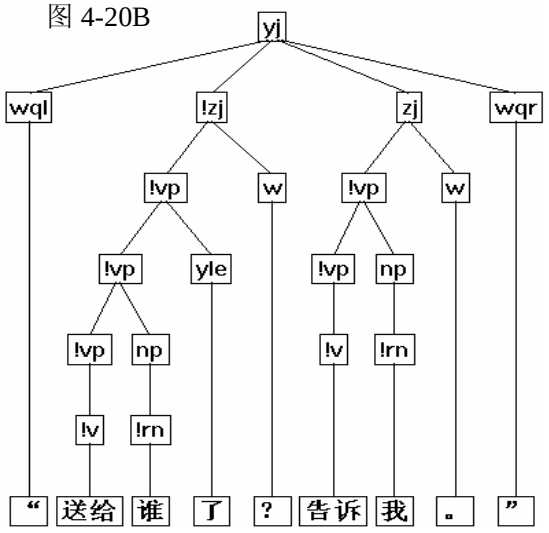


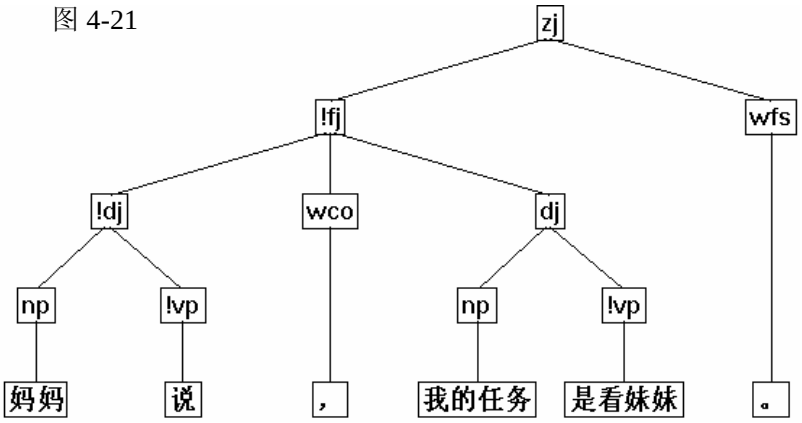
图 4-20B



总的来说，yj本身的形式标志是很明显的，凡是由 “ + **xp** + ” 组成的语言单位[14]，其整体起到句子的表达功能，就都可以标为yj。从yj的内部组成来说，yj的组合模式均可以表示为：yj → wql !xp wqr 其中xp以zj最为常见。如果wqr前面不是以断句符号结束，那么xp可能是dj，vp等短语成分。此外，yj内部也允许包含多个zj（不超过 5 个），zj之间不需要再标注层次组合关系。从yj的外部组合关系来说，yj一般处于多分支结构中，yj本身一般不是中心成分。yj参与组合时的模式以上面例 2a，2b，例 3 所示为常见类型，其余变体形式可参照处理。yj前后一般是“某某说”这样的形式，除此之外，还有“某某想”“某某道”“某某心中暗想”等形式，对于这些形式，处理方式均与“某某说”相同。

需要注意的是，对于真实文本语料中“间接引语”的例子（比如：妈妈说，我的任务是看妹妹），因为句中没有引号，不应分析为 yj。这种情况应按照一般的 fj 处理。例如：

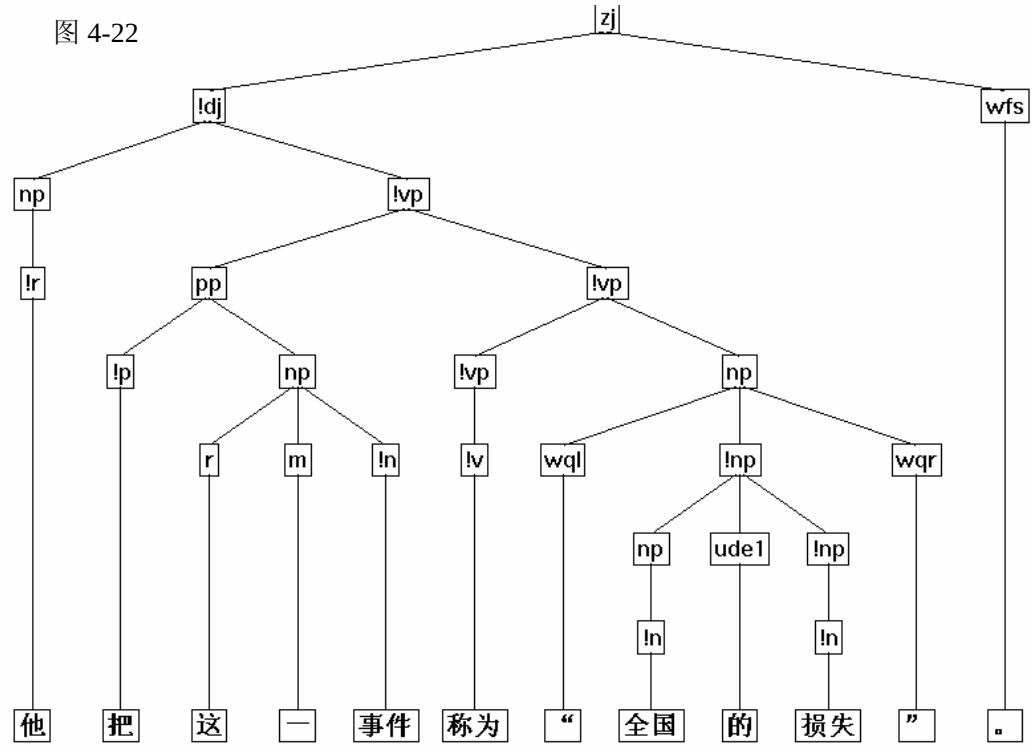
图 4-21



此外，yj 的首尾由引号“ ”标记，但由引号标记的语言片断并不都是 yj。比如下例中的“全国的损失”，功能是一个 np，该 np 前后以引号标记，形成的语言单位功能仍然是 np，不需要标为 yj。

[14] 中文真实文本中引号不仅仅是“ ”，还可能包括其他的形式，比如：『喝下去』，他命令说。

图 4-22



这个带引号的 np，其组合模式为　np → wql　!np　wqr。

再比如：这里却是一个“草木茂盛，多禽兽”的地方。

其中 〔 “草木茂盛，多禽兽” 〕 不应标为 yj。而应该标为 fj。

其组合模式为： fj → wql　!fj　wqr。

4.2.11 整句（zj）的标注

前文 2.1 已经对树库标注中的语法单位的层级性质做了整体说明，整句（zj）是一级语法单位。按照一般语法学的认识，zj是抽象的短语结构加上语气（句调）后形成的相对完整的表达单位。在书面形式上，zj是短语加上完句标点（一般常见的为。！？）[15]构成的语法单位。因此，zj一般的组合模式为：

zj → !xp　wfs　　其中 xp 一般可以是各种二级语法单位（即短语）

zj → !xp　wem

zj → !xp　wqm

也就是说，一般 zj 都是二分的，以完句标点 wfs（。） wem（！） wqm（？）结尾。

文本中也可能出现两个标点符号连用结句的情况，比如：

祥林嫂竟肯依？.....

这种情况下，不分析为 zj 嵌套的组合模式，而是采用多分支处理方式：

zj → !xp wqm wsp

也就是说，如果有多个标点符号连用结句，则这些标点符号以并列的身份作为 zj 的构成成分。

除上述一般构成模式外，zj 也可以包含 yj，以 yj 结尾，这时候可以是多分的组合模式。例如上文图 4-19 所显示的结构。因为 yj 一般都是以完句标点加右引号结尾的，因此，包含

[15] 省略号也可以起到完句标点的作用，不过，需要注意的是，省略号的作用不限于作句末标点。比如在剧本语料中有这样的例子：“吴：……现在的三块钱，值什么？”。这里省略号表示说话之前的停顿。

yj 的 zj 仍然应视为是以完句标点结尾的。这样处理体现了 zj 是抽象的句法结构带上句调（书面上体现为完句标点）所得到的语法单位。

上文在谈到 yj 的标注时已经涉及到 zj 的标注问题。这里再重复一下。关于 zj 的标注，要注意三点：

（1） zj 是一级语法单位，除了可以被 yj 包含外，一般不能被其他短语标记包含；

（2） zj 不能包含 zj，如果在一个 yj 中有多个 zj（不多于 5 个），应处理为多个 zj 并列的组合形式。

（3） zj 应直接（或间接）以完句标点（或右引号）结尾。反之，如果不是以完句标点（或右引号）结尾的单位，一般不应标为 zj。

4.2.12 独词句

独词句指语料中单个词（或简单词组）形成的句子。有多种情况：

（1）称呼。　“老张！”

（2）惊叹。　“红灯！”

（3）应答。　“是的。”

（4）省略。　“可是……”

（5）疑问。　“为什么？”

除称呼标为 yph 外，其他的独词句都分析为词语加标点符号上升为 zj 的模式。比如：

“是的。” 分析为 zj → vp(是的) wfs(。) “是的”分析为 vp → !vp(是) yde(的)

“可是……”就分析为 zj → !c wsp 即直接由连词加省略号形成整句。

“为什么？”分析为 zj→!vp wqm “为什么”标记为 rv，上升为 vp，再跟问号 wqm 组合为 zj。

有的时候，在句中像“可是”这样的连词会接上语气词。比如“可是呢，他一直没有等到”。对于“可是呢，xp”这样的形式，组合模式分析为：xp -> c y wco !xp，即“可是”“呢”“逗号”多项平列参与组合，中心成分标记在 xp 上。整个结构的功能类也从 xp 继承得到。

4.3 语言成分的自指（self-referential）用法

语料中绝大多数表达形式中符号都是以转指用法出现，但也有的时候符号是以自指义出现。比如，在介绍标点符号用法的文章中，会出现下面这样的一个小标题：

“一、句号（。）”

其中“。”在这里不是像一般的句号那样起结句作用，而是指它自身。对于这种符号的自指用法，标注时一律标记为 x，上升短语类时标注为 np，即 np→!x 。

五 中心成分标注

5.1 短语结构类型与短语中心成分的对应关系

一个短语结构由多个成分组成，一般来说，其中总有一个成分显得特别突出，该成分的性质在很大程度上决定了整个短语结构的性质，这个成分就是整个短语的中心成分。树库中通过在一个短语（或词）标记前加“!”号的形式来标记中心成分。下面表 5-1 反映了一般情况下各种短语结构的中心成分的类型。

表 5-1：汉语短语结构与中心成分对照表（黄颜色标记的成分为结构的中心成分）

标记	功能类名称	结构关系	结构成分		中心语	中心语常见类型
zj	整句		fj,dj,vp,...	完句标点	fj, dj, vp, ...	dj, fj
fj	复句型短语	主从结构	从句（主句）	主句（从句）	主句	dj, fj, vp
dj	单句型短语	主谓结构	主语	谓语	谓语	vp, ap, np, dj
vp	动词性短语	述宾结构	述语 1	宾语	述语 1	vp
		述补结构	述语 2	补语	述语 2	vp
		状中结构	状语	中心语 2	中心语 2	vp
ap	形容词性短语	状中结构	状语	中心语 2	中心语 2	ap
		述补结构	述语 2	补语	述语 2	ap
		的字结构	中心语 4	附加语	中心语 4	ap
np	名词性短语	定中结构	定语	中心语 1	中心语 1	np, vp
		的字结构	中心语 4	附加语	中心语 4	np, vp, ap, dj
		所字结构	附加语	中心语$_5$	中心语$_5$	vp
dp	副词性短语	地字结构	中心语 6	附加语	中心语 6	dp, ap
pp	介词性短语	介宾结构	介词	宾语	介词	p
sp	处所词性短语	定中结构	定语	中心语 1	中心语 1	sp, f, n
tp	时间词性短语	定中结构	定语	中心语 1	中心语 1	tp, f, n
qp	数量短语	定中结构	定语	中心语 1	中心语 1	q
mp	数词短语	定中结构	定语	中心语 1	中心语 1	m

从表中可以看出，汉语短语结构和中心成分之间的对应关系多数情况下比较清楚。这是因为大多数结构是所谓的向心结构（endocentric construction）。dj 一般认为不是向心结构，但因为 dj 是谓词性结构，所以其中心成分定在谓语上应该也没有太大争议。

5.2 助动词不作中心成分

“能愿动词 + vp”形成的结构是述宾结构，但跟一般的述宾结构中心词的处理不同，其中心成分标在助动词后面的 vp 上，而不是作述语的“能愿动词”。将“能愿动词+vp”结构中的中心成分定在 vp 上，主要考虑是：在这个结构参与更大的组合时，动词与名词之间的配价关系主要是由 vp 决定的，而不是由能愿动词决定的，为了将来在树库句法结构标注的基础上进一步作语义分析方便，我们选择将中心成分标在能愿动词带的 vp 宾语上。

5.3 倒装结构的中心成分

有些定中 np 的定语后置了，这时中心词应该标记在前置的中心语上。比如“录像带两百盘”中的中心词为“录像带”。

再比如主谓结构倒装的情况，中心词也应标在前面的谓语成分上。比如“烫手啰，热白果”。组合规则为：dj->!vp wco np　中心词标记在前置的谓语“烫手”上。

对于倒装结构，标记中心词的原则是标在倒装之前结构的中心成分上。

5.4 连谓结构和联合结构的中心成分

上表中没有涉及到连谓结构和联合结构。这两种结构都是由同类成分组合形成的非向心结构，而且经常是两项以上的成分参与组合，成分之间很难分出地位的高下。因此它们的中心成分不容易确定，实际上要给联合和连谓结构确定中心成分，也是出于统一的考虑，即假定了所有的结构都应该有一个“中心成分”。有了这样一个假定，即便是以硬性指派的方式，也要给联合和连谓结构确定一个中心成分。为简单起见，联合结构的中心词一般可确定为联合结构中的第一个成分[16]。连谓结构则视连谓各项的语义重要性来确定。语义重心所在就是中心成分，如果难以确定语义中心，就将第一项成分确定为中心成分。

5.5 多分支结构的中心成分

对于多分结构来说，其中心成分的确定一般参照对应的二分结构的中心成分。比如双宾结构是三分的结构，其中心成分根据相应的二分结构（即普通述宾结构），可以确定为述语动词。再比如“v + q + n”结构（处理为三分结构），中心成分同样仿照一般述宾结构确定为 v。对于像“这本书的出版”这样的“x 的 y”三分型定中结构，中心成分标记也跟一般定中结构一样，一律标在中心语 y（“出版”）上。

5.6 一个短语有且只能有一个中心成分

除叶子节点外，一般的子树结构中，应该有且仅有一个中心成分前面标记了“!”。在具体标注工作中，应避免兄弟节点中出现两个或多个成分前有标记“ ！ ”的情况出现。

[16] 一般来说联合结构中各组成成分的功能类是一样的，这时中心词定在任何一个组成成分上，都没有差别。不过也可能出现联合结构中组成成分本身的功能类不同的情况，这时，应遵循联合结构整体功能类别跟中心词功能类别保持一致的原则，先确定联合结构的整体功能类别是什么，然后再确定中心词标在哪一个组成成分上。

六 树库标注中需要注意的其他问题

除语言学方面的问题外，工程上的一些技术处理问题和一致性问题，也要特别注意，为此，特别提出如下两条约定：

6.1 应避免将一个标记直接上升为同级标记

这包括两类情况：一是不允许将一个标记上升为跟其自身一样的同级标记，即不允许出现“自身嵌套”A [A[...]] 。如下面图 6-1 所示，将 np 直接上升为 np 是不允许的；二是除一些特殊规定的情况外（比如一些篇章插入成分的处理），也不允许将一个标记上升为其他的同级标记，即不允许出现 A[B[...]]这样的结构。如下面图 6-2 所示，将 pp 直接上升为 vp 是不允许的。

图 6-1

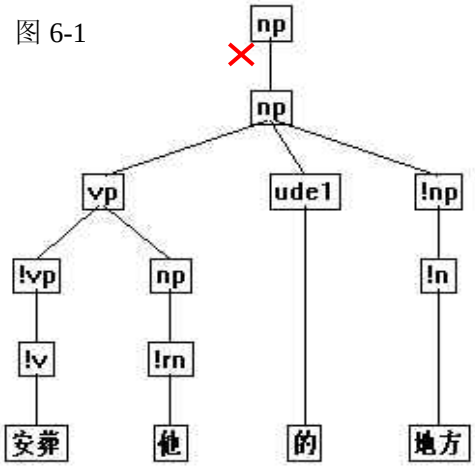
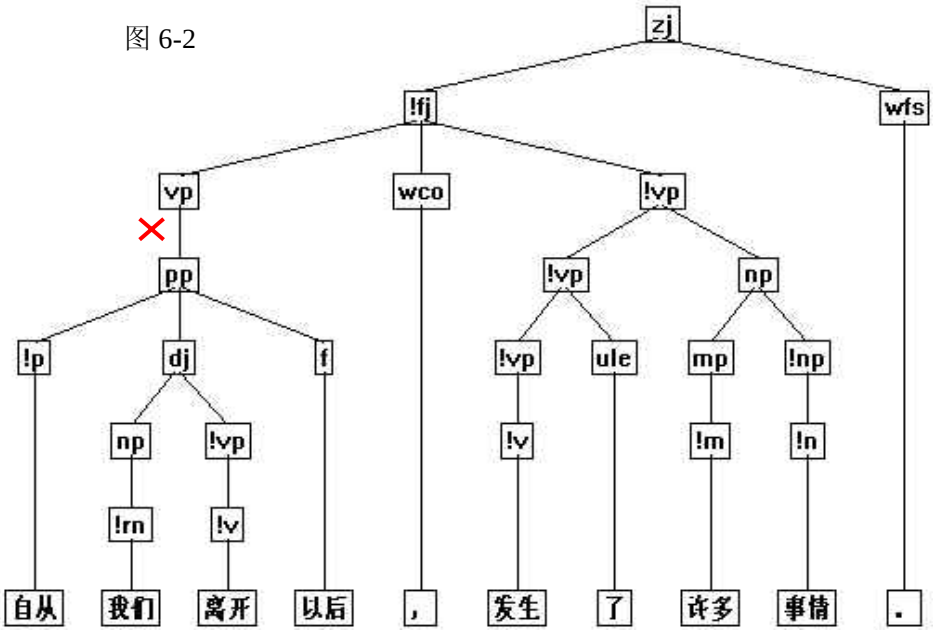


图 6-2



6.2 同类现象应做同样标注

这是为了追求标注的一致性和系统性。

比如：如果“v 一 v”处理为三分结构，那么“v 了 v”，“v 了 一 v”则相应地也应该处理为三分、四分等多分支结构（同为“扁平结构”）[17]；

再比如“称赞他勇敢”处理为兼语结构（三分：称赞 + 他 + 勇敢），相应地，“责备他懦弱”也应该处理为兼语结构（三分：责备 + 他 + 懦弱）。

树库只有一致性得到保证，当我们从树库中抽取语法规则和语言规律时才能得到准确的信息。

需要注意的是，有时候为了追求表面的一致性，可能却忽视了客观存在的差异，造成错误。比如前文说过，对“v＋到＋np”结构，应作三分处理，但同时应注意仔细甄别，比如“坐船到那儿”就不能分析为：vp[v[坐船] v[到] np[r[那儿]]]

图 6-3

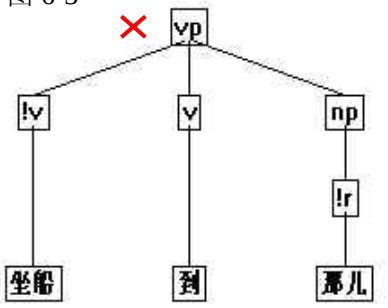
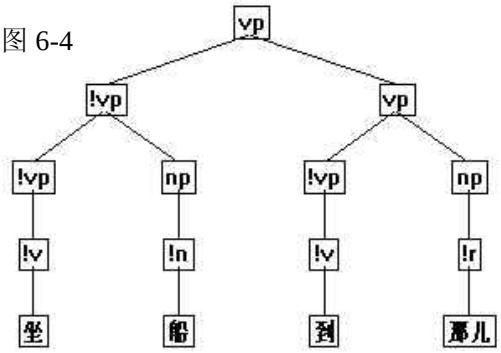


图 6-4



“v＋到＋np”这个格式中的“v”必须是不含宾语成分的单个动词，“坐船”实际上应该分析为述宾式 vp，而不是一个动词。“坐船到那儿”应该分析为连谓式 vp（如图 6-4 所示）。

[17] 语言学分析中一般把“v 一 v”看作是跟“v 一 下”同类的结构，即动词带动量或时量成分形成的结构（有的语法书把“v 一 下”这类结构看作述宾结构，有的则看作述补结构）。而对“v 一 下”采用的是二分处理：v + 一下。我们不做二分处理，而是做多分的处理，目的是强调这类结构跟普通的动词带动量/时量成分形成的结构之间的差异（参见上文 3.2.2 的说明）。

参考文献

[1]北京语言学院句型研究小组编,“现代汉语基本句型”,连载于《世界汉语教学》1989(1,3,4), 1990(1), 1991(1)

[2]北京大学计算语言学研究所（1999）“现代汉语语料库加工规范——词语切分与词性标注”，http://icl.pku.edu.cn/icl_groups/corpus/coprus-annotation.htm。

俞士汶、段慧明、朱学锋、孙斌（2002）北京大学现代汉语语料库基本加工规范，《中文信息学报》，2002年，第16卷第5期，P49-64；第6期，P58-65。

俞士汶、段慧明、朱学峰、孙斌、常宝宝（2003）北大语料库加工规范：切分·词性标注·注音，《汉语语言与计算学报》（新加坡），2003 年 6 月，第 13 卷 2 期。

[3]吴竞存, 梁伯枢（1992）《现代汉语句法结构与分析》， 语文出版社。

[4]李子云（1991）《汉语句法规则》，安徽教育出版社。

[5]范继淹（1986）《范继淹语言学论文集》，语文出版社。

[6]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。

[7]周强（2004）汉语句法树库标注体系，《中文信息学报》2004 年第 4 期。

[7] http://turing.iis.sinica.edu.tw/treesearch/ （台湾中研院树库）

[8] http://www.cis.upenn.edu/~treebank/ （宾州大学树库）

本规范文件的配套文档：“现代汉语树库标注常见问题举例”

致谢

本规范的制订，得到许多人的帮助。参加树库加工的工作人员对规范的修订贡献良多。在此深表感谢。下面是先后参加过这项工作的人的名单（大致按照参加时间先后为序）：

常宝宝、吴拥华、应晨锦、李恩京、叶娜、张洁、沈薇、杨灵叶、王秋萍，郭青剑、张则峰、杨丙涛、赵欣、运红娜、张娟、廖娟、曲丹、曾石铭、丁伟伟、姜巍、邓高、胡曼妮、孙薇薇、王楠、王祖明、金智英、王展、夏军、成方、陆烁、徐鹏波、王媛、杨霁楚、陈锡华、丘彦斌、杨帆、刘佳媛、崔延燕、白静茹、裴雨来、魏红华、蒋静忠

北大中文系袁毓林老师在本规范制定过程中提出过宝贵意见，在此也深表感谢。

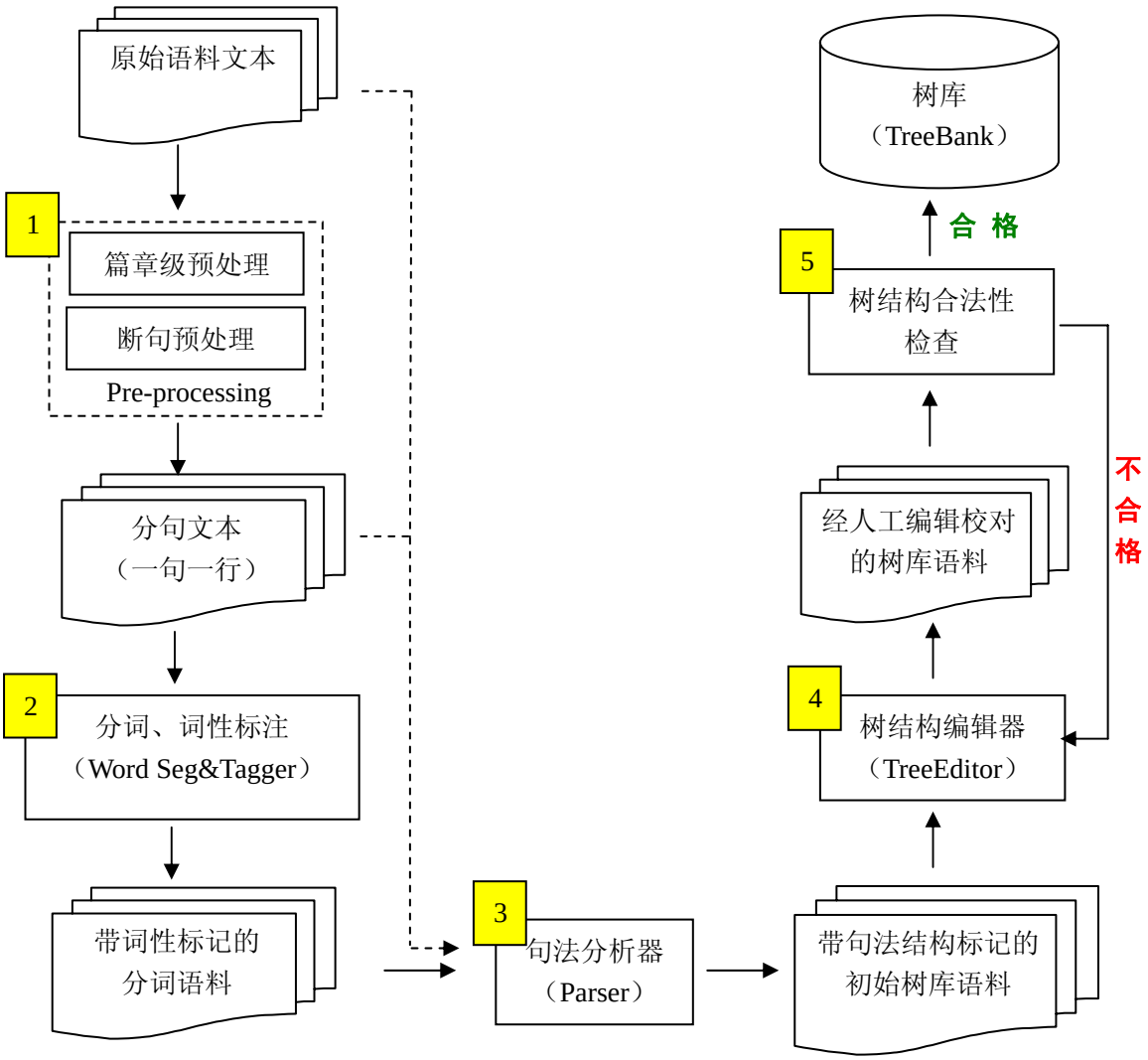
后记

通常技术规范在行文时多数是僵硬的条条框框，而较少对这些条条框框进行摆事实讲道理的解释。但我们考虑到，语言学中的问题本来就错综复杂，不少问题目前也还没有清晰一致的结论，如果只是生硬地拿出一个规定，对于参与工程的工作人员如何处理具体问题，帮助还是有局限的。事实上，语言工程的质量好坏，很大程度上是要依赖工作人员对于错综复杂的语言问题的认识水平，要保证一致性，尽量减少内部矛盾，就需要对规范所规定的内容尽可能准确地去理解，既要知其然，更要知其所以然。因此，在制订规范的时候，我们比较注意通过实例来阐明我们对于树库加工各个环节上的处理方式的认识（请参看“现代汉语树库标注常见问题举例”文件）。除了做具体的规定外，一般都对规定背后的依据（或者我们主要是基于什么样的考虑）做了解释。一方面这有助于工作人员更好地把握规范的精神，在具体操作时能举一反三，触类旁通，另一方面也有助于工作人员在具体实践中发现现有规范可能存在的问题，而不拘泥于现有的可能还不成熟的规范。

限于水平，本规范对汉语句法结构的处理还难免有不少缺漏或者不当之处，希望在实践中不断修正改善，也希望读者和参与树库加工的工作人员能够提出宝贵意见和建议。

附录一：现代汉语树库加工流程

1.1 现代汉语树库加工流程示意图



步骤说明：

1 对原始文本进行格式处理，使得一个自然篇章为一个文件（.txt，纯文本文件）。按照标点符号分界，将篇章中的句子分割出来，使得一句占一个文本行。关于断句的标准及操作时的注意事项，请看附录二的说明。
步骤一由自动断句程序，或利用 UltraEditor 等编辑工具辅助人工完成。

2 对原始句子进行分词和词性标注处理，得到带词性标记的分词语料。交给人工校对，然后将分词和词性标注基本正确的语料交给句法分析器进行分析。（也可以不进行人工校对，直接将机器处理得到的分词语料交给句法分析器进行分析）

3 句法分析器输出带有句法结构标记（短语类，层次标记，中心词标记）的初始树库(T_0)。

4 在 TreeEditor 编辑环境下，由人工对初始树库进行校对，修改其中的错误。得到经过人工校对的树库语料(T_1)。

5 利用TreeEditor程序中包含的树结构查错功能以及规则抽取功能等，对T_1进行检查。达到合格要求，就形成最终的树库，如果不合要求，就返回第 4 步，继续对T_1进行编辑修改，产生T_2，T_3，…… 等等，直至合格为止。（详见下文 1.3 的说明）

1.2 树库加工中用到的计算机辅助软件

程序列表及功能说明

序号	功能	名称	说明
1	断句程序	GetSentences.exe	对原始 txt 文件进行自动断句处理
2	分词和词性标注程序	ICTCLAS.exe	对断句文件进行分词和词性标注处理
3	检查分词格式程序	CheckSegTag.pl	对分词和词性标注文件进行格式检查
4	句法结构分析程序	GFRGParser.exe	进行句法结构自动标注
5	句法结构编辑、校对、核查程序	TreeEditor.exe（含 TreeCheck 功能）	辅助人工对自动产生的句法结构进行校改
6	句法结构后编辑程序	SubStringReplace.pl	对人工校改后的树库文件进行批量替换处理，对某些标记做调整
7	汉字频度统计	HZ_Freq.exe	对文件中的汉字字数、频度进行统计

上述 1，5，7 三个程序中包含了对树库语料文件的字、词、句、短语分布等进行统计的功能，如果想从宏观上了解一个树库语料的总体情况，或者想在校对好的树库语料中研究某类短语的分布情况，可以利用这些功能获得统计结果。

1.3 树库校对工作注意事项

对树库文件进行初次校对，一般是采取逐句顺序校对的方式。在顺次校对过一个文件中的每个句子，以及顺次校对完多个文件后，应该进一步从总体上对树库的句法结构标注质量进行检查，看是否存在不符合规范或者标注不合理的情况。具体来说，可以通过 TreeEditor 程序提供的功能，发现两类问题并予以纠正：（1）在逐句校对过程中，可能因为手误，造成错误的标注；（2）逐句校对时处理对象是每个孤立的句子，这样有可能出现前后处理上不一致的地方，通过对树库文件的整体检查，有可能发现这类问题。

因此，在完成了顺次逐句校对之后，还需经过下面三个步骤的检查，并且是循环进行这三个步骤的检查。

（1） 利用 TreeEditor 工具中的“查错”功能，对全部树库文件进行查错，得到 errorlist_2.txt。这时，应以 0_errorlist_2.txt 为文件名，保留第一次查错时的这个错误报告文件。

根据 0_errorlist_2.txt 文件中的提示，对树库标注内容进行修改。

（2） 在经过上面第一步检查后，按照规则频度序，抽取规则，对低频规则（频度为 1 或 2）进行逐一检查。

具体操作时，用 excel 打开 rule.txt 文件，将文件另存为 excel 表格文件，文件名为“文件夹名_rule_freq.xls”。比如校对者如果负责的是 13 号文件夹，则规则文件为 13_rule_freq.xls。

在这个 excel 表格文件中，对怀疑有误的规则组合模式，要用 ？ 加以标记。并利用 TreeEditor 工具中的查找功能，找到该规则模式所在的句子，进行核实。如确有错误，应进行修改。

（3） 在经过上面第二步检查后，按照规则右部节点排序（RHS 序），抽取规则，对 RHS 相同，但组合规则却不同的情况进行检查，修改处理上不一致的错误。然后，抽取兼类词表，看是否存在因为处理上前后不一致造成的“假兼类词”。

具体操作时，用 excel 打开 rule.txt 文件，将文件另存为 excel 表格文件，文件名为“文件夹名_rule_rhs.xls”。比如校对者如果负责的是 13 号文件夹，则规则文件为 13_rule_rhs.xls。

在这个 excel 表格文件中，对怀疑有误的规则组合模式，要用 ？ 加以标记。并利用 TreeEditor 工具中的查找功能，找到该规则模式所在的句子，进行核实。如确有错误，应进行修改。

上述三步检查，应循环进行。也就是说，在经过第三步检查后，要回到第一步，再次用查错功能进行查错。需要注意的是，在经过上述第二步、第三步纠错后，有可能造成新的错误，这样，再次进行查错，可以发现后来产生的新的错误。

经过循环检查，直至最终产生的错误报告文件中不再有真正的错误，低频规则中也不再

包含错误规则为止，即可提交校对结果。

在提交校对结果时，除树库文件外，还应包括以下文件：

（1） 0_errorlist_2.txt；　　（第一次查错的报告文件）

（2） errorlist_2.txt；　　（最后一次查错的报告文件）

（3） *_rule_freq.xls；　　（第一次抽取的规则文件，规则按频度排序）

（4） *_rule_rhs.xls；　　（第二次抽取的规则文件，规则按右部子节点排序）

（5） rule.txt；　　（提交正式校对结果时抽取的规则文件，按频度排序）

（6） MultiTagWordList.txt （提交正式校对结果时抽取的兼类词表，按词语排序）

上面两个.xls 文件中 * 代表树库文件夹编号

提交校对结果时请特别注意：

（I） 以上所有文件均应和树库文件一起，压缩为 *.zip 或 *.rar 形式提交。

（II） 校对前后，应保持树库文件名不变。

附录二：现代汉语文本断句的操作标准

对文本进行断句时，有两个主要因素可以考虑：（1）标点；（2）句子长度。

前一个因素是书面文本中可用来作为断句依据的形式标志；后一个标准是在考虑计算机处理时是否方便的依据。

上面这两个因素比较具体、客观，适合作为操作标准。此外，对文本进行断句，还有一个因素，相对抽象一些，但也是应该考虑的。那就是断句之后的结果应该适合作为一个完整的分析单位。换句话说，那些对句法结构分析造成干扰的语言片段，在断句之后，不应该再“混迹”于“正常”的结构单位之中，而应该单列出来（详见下面 2.4 中的示例）。

2.1 根据标点进行断句

2.1.1 结句标点：句号、问号、感叹号

一般情况下，句号、问号、感叹号为断句的标志性标点。如：

a. 中国人民有着热爱和平的光荣传统。

b. 你知道我在等你吗？

c. 中华人民共和国万岁！

逗号、分号、破折号等一般看作是句内的分隔符号，不是断句的标志性标点。

2.1.2 省略号

省略号的功能比较多，比如下面句中的省略号用于表示说话时较长的停顿。这种情况下省略号不是断句标志性标点。

你这种人啊……是出了名的！

有的时候，省略号也用在一个完整句子的句尾，表示后面还有一些类似的情况，略去不说，或者意犹未尽，但又不想多说了。这时候省略号可以当作完句标点看待。比如：

这些年我跑过好些地方，到处都听到人们在倾诉，水的污染使鱼虾减少了，乱砍滥伐使山林面积缩小了，肆意行猎使禽兽锐减了……

省略号还可能用在一般完句标点（。？！）等的后面。比如：

呵呀，这样的婆婆！……

吓，你看，这多么好打算！……

祥林嫂竟肯依？……

像上面这样的情形，应该以省略号作为断句的标记。

2.1.3 左右匹配型标点

左右匹配的标点，比如“ ”（ ）‘ ’[ ] 等等，我们称为前后封闭型标点。封闭型标点所包含的区域内，有时候是一个语言片断，而非完整的句子，但也有的时候是一个完整的

句子。如果是后一种情况，则封闭型标点连同其所包含的内容，作为一个断句单位处理。如：

a. “她吃过东西吗？这锅里是什么？”

b. “离敌人越近，越觉着打得过瘾，越觉着打得解恨！”

c. （新华社长江前线 22 日 22 时电）

上面这些例子都应该作为一个句子处理。

如果在封闭型标点的封闭区域内包含了一个以上的断句标志性标点，而封闭区域内的语言片断很长，这时需要根据句子长度来决定断句方式（详见 2.3 节的说明）。

2.2 无标点结尾的“句子”

无断句结尾标点的标题或作者等篇章单位看作一句，占一行。如：

a. 从百草园到三味书屋

b. 谁是最可爱的人

c. 鲁迅

2.3 断句时考虑句长因素

2.3.1 跟引句相关的长句

如果引号内的语言片断超过 30 个词的长度，可以将封闭区域内的句子按照一般断句方式断为多个句子。这时封闭型标点独立占一行，即形成类似于下面这样的断句形式：

“

xxxxxxxx 。

yyyyyyyy 。

zzzzzzzzzzzz ？

”

例如，原文为：

“请问你们闯的目的是什么？你们乘船的目的又是什么？绝不会是想送死吧？我们给你们‘帆船’是要你们体验风浪，去探求道路，而通过‘帆船’的行驶去认识海洋，去闯更艰难的道路，胜利到达彼岸。而不是送你们进漩涡，让你们去淹死。我们给你们‘摇篮’，是让你们快快长大，使你们具有做人的权利和义务，让你们在社会允许的范围之内去闯、去炼，告诫你们‘闯’不能出格，难道去闯小流氓之路、阿飞之路、杀人犯之路？任何事物都有一个范畴嘛！”

断句结果应为：

1. “
2. 请问你们闯的目的是什么？
3. 你们乘船的目的又是什么？
4. 绝不会是想送死吧？

5. 我们给你们‘帆船’是要你们体验风浪，去探求道路，而通过‘帆船’的行驶去认识海洋，去闯更艰难的道路，胜利到达彼岸。
6. 而不是送你们进漩涡，让你们去淹死。
7. 我们给你们‘摇篮’，是让你们快快长大，使你们具有做人的权利和义务，让你们在社会允许的范围之内去闯、去炼，告诫你们‘闯’不能出格，难道去闯小流氓之路、阿飞之路、杀人犯之路？
8. 任何事物都有一个范畴嘛！
9. ”

2.3.2 以分号为断句标点的长句

如果句子中包含分号（;），而且分号左边的句子长度很长（超过 30 个词），可以以分号作为断句标点，将分号左边的字符串处理为一个句子。

例如，原文为：

做工的人，傍午傍晚散了工，每每花四文铜钱，买一碗酒，这是二十多年前的事，现在每碗要涨到十文，——靠柜外站着，热热的喝了休息；倘肯多花一文，便可以买一碟盐煮笋，或者茴香豆，做下酒物了，如果出到十几文，那就能买一样荤菜，但这些顾客，多是短衣帮，大抵没有这样阔绰。

断句结果应为：

1. 做工的人，傍午傍晚散了工，每每花四文铜钱，买一碗酒，这是二十多年前的事，现在每碗要涨到十文，——靠柜外站着，热热的喝了休息;
2. 倘肯多花一文，便可以买一碟盐煮笋，或者茴香豆，做下酒物了，如果出到十几文，那就能买一样荤菜，但这些顾客，多是短衣帮，大抵没有这样阔绰。

原文：

“这是官批本，”鲁迅先生认真地说，“你就另外去印你自己的别集。快了！一个政权到了对外屈服，对内束手，只知道杀人、放火、禁书、掳钱的时候，离末日也就不远了。他们分明的感到：天下已经没有自己的份，现在是在毁别人的、烧别人的、杀别人的、抢别人的。越是凶，越是暴露了他们卑怯和失败的心理！”

断句结果应为[18]：

1. “这是官批本，”鲁迅先生认真地说，
2. “
3. 你就另外去印你自己的别集。
4. 快了！
5. 一个政权到了对外屈服，对内束手，只知道杀人、放火、禁书、掳钱的时候，离末日也就不远了。
6. 他们分明的感到：天下已经没有自己的份，现在是在毁别人的、烧别人的、

[18] 断句后第一句是以逗号结尾的。为了避免句子过长，只好造成一个以逗号结尾的“句子”。

杀别人的、抢别人的。

7. 越是凶，越是暴露了他们卑怯和失败的心理！

8. ”

2.3.3 “一逗到底”的长句

长句（包括含引句的长句）如果在结构上有可能断成简单的、短一些的句子，一般就倾向做断开处理（如上面对引句的处理）；如果结构上前后连贯紧密，不容易从中断开，则维持长句原貌。

下面是两个不容易“断开”的长句的例子。

例 1：

原文：

毛泽东的这个号召，很快地在中国共产党内和党外引起了怎样以从实际出发的观点而不是以教条主义的观点来对待马克思列宁主义原理，怎样使马克思列宁主义的基本原理和中国革命的实际相结合，以及怎样对待1931年初至1934年底这段时期党内两条路线的斗争这样一些重大问题的大讨论，巩固了马克思列宁主义思想在党内外的阵地，使广大干部在思想上大大地提高了一步，使中国共产党达到了空前的团结。

断句结果应为：

1. 毛泽东的这个号召，很快地在中国共产党内和党外引起了怎样以从实际出发的观点而不是以教条主义的观点来对待马克思列宁主义原理，怎样使马克思列宁主义的基本原理和中国革命的实际相结合，以及怎样对待 1931 年初至 1934 年底这段时期党内两条路线的斗争这样一些重大问题的大讨论，巩固了马克思列宁主义思想在党内外的阵地，使广大干部在思想上大大地提高了一步，使中国共产党达到了空前的团结。

例 2：

原文：

他的衣着过分随便，走路的姿态也不慎重，走上五六十米路便选定一处地方，一只脚踏在石凳上或土埂上或树墩上，解下腰间的酒瓶，解酒瓶的当儿迷起眼睛把一百八十度视角内的景物细细看一遭，然后以迅雷不及掩耳之势倒一大口酒入肚，把酒瓶摇一摇再挂向腰间，平心静气地想一会什么，便走下一个五六十米去。

断句结果应为：

1. 他的衣着过分随便，走路的姿态也不慎重，走上五六十米路便选定一处地方，一只脚踏在石凳上或土埂上或树墩上，解下腰间的酒瓶，解酒瓶的当儿迷起眼睛把一百八十度视角内的景物细细看一遭，然后以迅雷不及掩耳之势倒一大口酒入肚，把酒瓶摇一摇再挂向腰间，平心静气地想一会什么，便走下一个五六十米去。

需要特别说明的是，以句子长度（词数）来作为断句的参考标准，其中词数应该理解为是个概数，上面以平均句长（30 词）作为参考标准，但在实际操作中并不需要拘泥于此。

断句时一方面要考虑句子的完整性，另一方面要考虑结构标注的方便。一般以后者为主要的考虑因素（即如何断句对标注方便有利，就选择那种断句方式）。句子长度是一个辅助的考虑因素。

2.4 剧本类文本的断句

以上是对普通文本进行断句时采用的操作标准。对于一些比较特殊的文本，比如剧本，在断句时有一些特殊性，下面是针对剧本的断句操作规定：

一般剧本内容中常见的主要有四种类型。一是对场景的描写，属于背景性的文字；二是说对白的人物名；三是对人物动作的说明；四是对白的内容。其中一、四这两种类型，可以跟普通文本中的断句方式一样进行断句处理。二、三这两种类型是剧本中的特殊篇章成分，一律独立成行，然后进行相应的结构层次分析和功能标注。其中对白人物名一般都是简单的专名 np；对人物动作的说明则可能是各类短语结构（以 vp 居多）。

下面是剧本断句的例子：

断句前原文文本：

常四爷:（凑过来，要对马五爷发牢骚）这位爷，您圣明，您给评评理！

马五爷:（立起来）我还有事，再见!（走出去）

常四爷:（对王利发）邪！这倒是个怪人！

王利发: 您不知道这是马五爷呀！怪不得你也得罪了他！

常四爷: 我也得罪了他？我今天出门没挑好日子！

王利发:（低声地）刚才您说洋人怎样，他就是吃洋饭的。信洋教，说洋话，有事情可以一直地找宛平县的县太爷去，要不怎么连官面上都不惹他呢！

常四爷:（往原处走）哼，我就不佩服吃洋饭的！

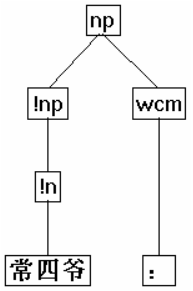
王利发:（向宋恩子、吴祥子那边稍一歪头，低声地）说话请留点神!（大声地）李三，再给这儿沏一碗来!（拾起地上的碎瓷片）

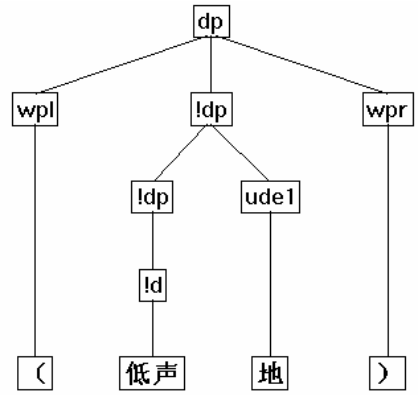
断句后的结果文本：

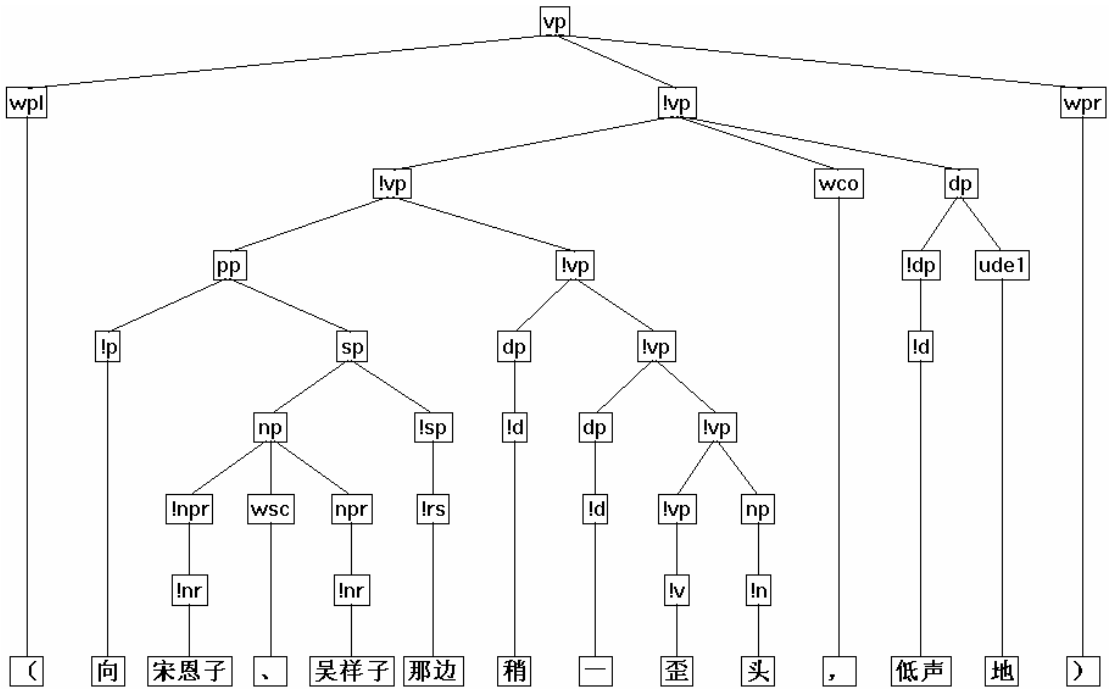
1. 常四爷:
2. （凑过来，要对马五爷发牢骚）
3. 这位爷，您圣明，您给评评理！
4. 马五爷:
5. （立起来）
6. 我还有事，再见!
7. （走出去）
8. 常四爷:
9. （对王利发）
10. 邪！
11. 这倒是个怪人！
12. 王利发:
13. 您不知道这是马五爷呀！
14. 怪不得你也得罪了他！
15. 常四爷:
16. 我也得罪了他？
17. 我今天出门没挑好日子！

18. 王利发:
19. （低声地）
20. 刚才您说洋人怎样，他就是吃洋饭的。
21. 信洋教，说洋话，有事情可以一直地找宛平县的县太爷去，要不怎么连官面上都不惹他呢！
22. 常四爷:
23. （往原处走）
24. 哼，我就不佩服吃洋饭的！
25. 王利发:
26. （向宋恩子、吴祥子那边稍一歪头，低声地）
27. 说话请留点神！
28. （大声地）
29. 李三，再给这儿沏一碗来！
30. （拾起地上的碎瓷片）

下面是对上例中一些行进行句法结构标注的结果示例：







有时候剧本中介绍场景的片断比较长，这时候可以参照长引句的处理办法。下面是一个示例。

断句前的原始文本：

〔 杨就此下台，回到象棋的战场，继续未完的棋局。吴太太也继续回到她未完的家事。少停，外边先传进一阵敲门的声音，接着走进一男一女，男的一望而知是一个警察，女的一手提了一个小包袱，从她的可怜神情，也不难猜出，她就是闯了祸的李嫂。〕

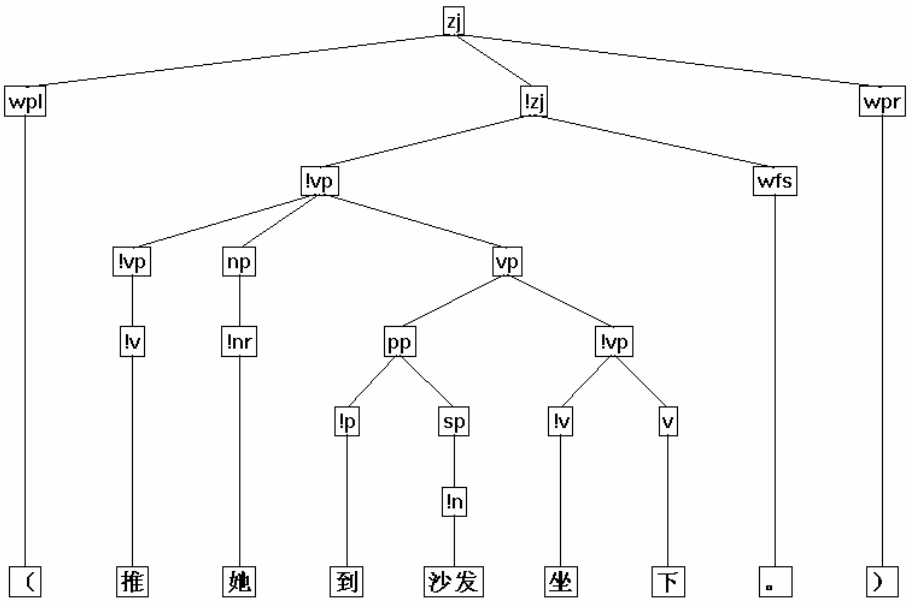
断句后的结果文本：

1. 〔
2. 杨就此下台，回到象棋的战场，继续未完的棋局。
3. 吴太太也继续回到她未完的家事。
4. 少停，外边先传进一阵敲门的声音，接着走进一男一女，男的一望而知是一个警察，女的一手提了一个小包袱，从她的可怜神情，也不难猜出，她就是闯了祸的李嫂。
5. 〕

剧本中也有的描述动作的文字以句号结尾，同时处于括号之中，对这种情况，可以将括号中的内容分析为 zj，然后再加上前后括号，上升为 zj。比如：

（推她到沙发坐下。）

结构标注如下图所示：



组合模式为：zj → wpl !zj wpr

2.5 小结：断句处理总的指导原则

树库加工目前阶段的主要任务是进行句法结构分析。分析对象是相对完整的“句子”。一个“句子”在计算机文件中占据一个文本行。反过来，一个文本行中的字符串并不一定是一个“句子”。这些并非“句子”的单行字符串，是为了切割出合适的分析单位（句子）而“不得已”造成的结果。

由此不难看出，在进行断句操作时，总的原则是尽量让一行内的字符串构成一个相对完整的、适合进行句法结构分析的语言单位。这样的分析单位不能太长 —— 太长的话，其结构关系就很难分析了（这时候需要进行篇章分析，而不是句法结构分析）；这样的分析单位也不能太短 —— 太短的话（比如只有一个词），就没有进行结构分析的必要了。事实上，上面的示例中已经有不少很短的单行字符串，比如文章的作者、剧本中的人物名，长引句的引号，等等，都是单独占一行的，这也就意味着，在树库加工中，对这些语言成分，我们已经放弃了对它进行结构分析，只是为了保持文章的完整性，将这些成分保留在树库中而已。这些成分中，有的不作特别的标记（比如文章作者），有的则增加了特殊的标记来标明其特殊性（比如文章的标题以 hl 标记）。

断句原则归结为一句话就是：断句后所得的语言片断应该适合进行句法结构分析。

这个原则进一步推衍可得：凡对句法结构分析造成干扰的语言片断，应该单列出来[19]。

[19] 像上面剧本示例中对人物动作的描述文字，比如“（大声地）”，单列一行，就体现了这一原则。本规范前面提到了语篇插入成分（ypc），这些实际上也是对句法结构分析造成干扰的片断，但因为嵌在“正常”的适合进行句法结构分析的片断中间，而不是在首尾的位置，很难单列出来，只好让这些ypc片断跟一般的句子成分混在一起进行句法结构分析，但对这些片断，需标以特殊的ypc标记。

附录三：现代汉语树库标记一览表

```
//////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////
//  共计 120 个标记
//////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////
hl // headline  文章标题（篇章标记，不是句法结构标记 — 超级单位）
//////////////////////////////////////////////////////////
zj //  整句（一级单位）  标记数（仅上位）：  1
//////////////////////////////////////////////////////////
//
//  以下是二级单位: 标记数（仅上位）：14，标记数（含下位）：21
qj //  句群（篇章单位）
yj //  用来标记引号“”及引号所包括的句子单位（篇章单位）
dj //  小句，主谓结构短语
fj //  复句
ap //  形容词性短语
dp //  副词性短语
mp //  数词性短语
np //  名词性短语
npr //  指人专名短语，如：混世魔王程咬金
nps //  指处所专名短语，如：北京海淀中关村
npt //  指机构专名短语，如：北京大学中文系
npx //  用来标记非中文字符串（词组），如：good bye
npz //  其他专名短语，如：“发现”号航天飞机
pp //  介词性短语
qp //  数量词性短语
sp //  处所词性短语
tp //  时间词性短语
vp //  动词性短语
yp //  语篇成分（篇章标记）
ypc //  语篇成分-插入成分
yph //  语篇成分-呼语成分
//  以上是二级单位
//
//////////////////////////////////////////////////////////
//  以下是三级单位，标记数（仅上位）：26，标记数（含下位）：97
a    //  形容词
ad //  形容词用作状语
an //  形容词用作名词
b    //  区别词
c    //  连词
ch //  前置关联词，比如“一”
```

ck // 后置关联词，比如“就”
d // 副词
e // 叹词
f // 方位词
g // 语素
ng // 名语素
vg // 动语素
ag // 形语素
dg // 副语素
bg // 区别语素
tg // 时间语素
sg // 处所语素
fg // 方位语素
h // 前缀
i // 成语
in // 名词性成语 如：稗官野史
iv // 动词性成语 如：暗箭伤人
ia // 形容词性成语
id // 副词性成语
j // 缩略语
jn // 名词性缩略语 如：妇救会
jv // 动词性缩略语 如：打砸抢
ja // 形容词性缩略语
k // 后缀
l // 习用语
ln // 名词性习用语 如：鹅毛大雪
lv // 动词性习用语 如：摆臭架子
la // 形容词性习用语
m // 数词
n // 名词
nr // 指人专名，如：张三、李四、王同志
ns // 指处所专名，如：中国，中关村
nt // 指机构专名，如：北京大学
nx // 用来标记非中文词，如：Ade，あなた
nz // 其他专名，如：京九铁路
o // 拟声词
p // 介词
pba // 介词“把”
pbei // 介词“被”
q // 量词
r // 代词
rn // 具有名词功能的代词（注意：标记是 r n，不是 m。要避免字母的字形混淆）
rs // 具有处所词功能的代词
rt // 具有时间词功能的代词

rm // 具有数词功能的代词
rd // 具有副词功能的代词
rv // 具有动词功能的代词
s // 处所词
t // 时间词
u // 助词
ude1 //“的”
ude2 //“地”
ude3 //“得”
usuo //“所”
uetc //“等” “等等”
uzhe // “着”
ule // “了”
uguo // “过”
udh // “的话”
usd // “似的”
v // 动词
vd // 动作作状语
vn // 动词用作名词
w // 标点 下面是具体的标点，除此之外的标点都笼统地标 w，
比如·作为外国人名用分隔符，以及像●这样的符号
wqm // 问号？question mark
wem // 感叹号！excalmatory mark
wcm // 冒号：colon
wfs // 句号。full stop
wsc // 顿号、sign of coordination
wco // 逗号，comma
wsm // 分号；semicolon
wsp // 省略号 …… suspension points
wda // 破折号 —— dash
whf // 连字符 -
wql // 左双引号“ quotation mark left
wqr // 右双引号” quotation mark right
wal // 『
war // 』
wbl // 左书名号《 book mark left
wbr // 右书名号》 book mark right
wcl // 左尖括号〈 左方括号〔
wcr // 右尖括号〉 右方括号 〕
wdl // 左单引号‘
wdr // 右单引号’
wpl // 左圆括号（parentheses Left
wpr // 右圆括号）parentheses Right
x // 中文非语素字，中文符号的自指用法一律标为 x

y　// 语气词
yle // 语气词“了”
yde // 语气词“的”，“他一定会成功的”“他跑起来很快的”
z　// 状态词

说明：

（1）如果代词的功能类别暂时不好确定，就仍标为 r，比如“这”“那”；如果代词的功能类别容易确定，就应标为 r 的下位标记，比如“我”应该标为 rn；“这么”应该标为 rd。

（2）从句法功能的角度看，拟声词 o 的功能不是很明确，在进行短语结构标注的时候应该注意将 o 上升为合适的短语功能类，然后再参与组合。

附录四：现代汉语树库样例

原始文件

中国人权事业的进展

前　言

1991 年 11 月，中国政府发表了《中国的人权状况》，向国际社会阐述了中国在人权问题上的基本立场和实践。四年来，中国的人权事业又取得了新的进展。

1991 年以来的四年，是中国全面实施国民经济和社会发展第八个五年计划的重要历史时期。中国的国民经济和社会发展突飞猛进，原定到 2000 年国民生产总值比 1980 年翻两番的计划已于 1995 年提前完成，人民生活显著改善，正在向小康目标前进。当前，中国政治稳定、经济发展、社会进步、民族团结，人民安居乐业、生活水平不断提高，人权状况呈现全面改善的良好态势。实践证明，将人民的生存权、发展权摆在首位，在改革、发展、稳定的条件下全面改进人权状况，是符合中国国情和全体人民的根本利益的，也是举世公认的。

四年来，中国积极维护《联合国宪章》促进人权和基本自由的宗旨与原则，坚决反对一些国家对其他国家特别是发展中国家在人权方面采取双重标准，或者以自己的模式强加于人，借口“人权问题”干涉他国内政的霸权主义行径。中国在人权问题上的原则立场，得到了世界许多国家的支持，为维护世界和平，促进国际人权事业的健康发展，作出了自己的努力。

虽然这几年中国在促进人权方面取得了显著成绩，但是，受历史和发展水平的限制，中国的人权状况还存在着一些不如人意的地方。继续维护和促进人权，不断地提高全体人民享受人权的水平，仍然是中国人民和政府的一项长期的任务。

经断句整理后的文件

1. 《中国人权事业的进展》
2. 前言
3. 1991 年 11 月，中国政府发表了《中国的人权状况》，向国际社会阐述了中国在人权问题上的基本立场和实践。
4. 四年来，中国的人权事业又取得了新的进展。
5. 1991 年以来的四年，是中国全面实施国民经济和社会发展第八个五年计划的重要历史时期。
6. 中国的国民经济和社会发展突飞猛进，原定到 2000 年国民生产总值比 1980 年翻两番的计划已于 1995 年提前完成，人民生活显著改善，正在向小康目标前进。
7. 当前，中国政治稳定、经济发展、社会进步、民族团结，人民安居乐业、生活水平

不断提高，人权状况呈现全面改善的良好态势。

8. 实践证明，将人民的生存权、发展权摆在首位，在改革、发展、稳定的条件下全面改进人权状况，是符合中国国情和全体人民的根本利益的，也是举世公认的。
9. 四年来，中国积极维护《联合国宪章》促进人权和基本自由的宗旨与原则，坚决反对一些国家对其他国家特别是发展中国家在人权方面采取双重标准，或者以自己的模式强加于人，借口“人权问题”干涉他国内政的霸权主义行径。
10. 中国在人权问题上的原则立场，得到了世界许多国家的支持，为维护世界和平，促进国际人权事业的健康发展，做出了自己的努力。
11. 虽然这几年中国在促进人权方面取得了显著成绩，但是，受历史和发展水平的限制，中国的人权状况还存在着一些不如人意的地方。
12. 继续维护和促进人权，不断地提高全体人民享受人权的水平，仍然是中国人民和政府的一项长期的任务。

分词和词性标注文件

1. 《/w 中国/ns 人权/n 事业/n 的/u 进展/v 》/w
2. 前言/n
3. 1991 年/t 11 月/t ，/w 中国/ns 政府/n 发表/v 了/v 《/w 中国/ns 的/u 人权/n 状况/n 》/w ，/w 向/p 国际/n 社会/n 阐述/v 了/u 中国/ns 在/p 人权/n 问题/n 上/f 的/u 基本/a 立场/n 和/v 实践/v 。/w
4. 四/m 年/q 来/f ，/w 中国/ns 的/u 人权/n 事业/n 又/d 取得/v 了/u 新/a 的/u 进展/v 。/w
5. 1991 年/t 以来/f 的/u 四/m 年/q ，/w 是/v 中国/ns 全面/a 实施/v 国民经济/n 和/c 社会/n 发展/vn 第八/m 个/q 五年计划/n 的/u 重要/a 历史/n 时期/n 。/w
6. 中国/ns 的/u 国民经济/n 和/c 社会/n 发展/vn 突飞猛进/i ，/w 原定/v 到/v 2000 年/t 国民/n 生产/vn 总值/n 比/p 1980 年/t 翻/v 两/m 番/q 的/u 计划/n 已/d 于/p 1995 年/t 提前/v 完成/v ，/w 人民/n 生活/vn 显著/a 改善/v ，/w 正在/d 向/p 小康/n 目标/n 前进/v 。/w
7. 当前/t ，/w 中国/ns 政治/n 稳定/vn 、/w 经济/n 发展/vn 、/w 社会/n 进步/vn 、/w 民族/n 团结/vn ，/w 人民/n 安居乐业/i 、/w 生活/vn 水平/n 不断/d 提高/v ，/w 人权/n 状况/n 呈现/v 全面/a 改善/v 的/u 良好/a 态势/n 。/w
8. 实践/v 证明/v ，/w 将/v 人民/n 的/u 生存权/n 、/w 发展/v 权/n 摆/v 在/p 首/m 位/q ，/w 在/d 改革/v 、/w 发展/v 、/w 稳定/v 的/u 条件/n 下/f 全面/a 改进/v 人权/n 状况/n ，/w 是/r 符合/v 中国/ns 国情/n 和/c 全体/n 人民/n 的/u 根本/a 利益/n 的/u ，/w 也/d 是/v 举世/n 公认/v 的/u 。/w
9. 四/m 年/q 来/f ，/w 中国/ns 积极/a 维护/v 《/w 联合国/nt 宪章/n 》/w 促进/v 人权/n 和/c 基本/a 自由/a 的/u 宗旨/n 与/p 原则/n ，/w 坚决/a 反对/v 一些/m 国家/n 对/p 其他/r 国家/n 特别/d 是/v 发展中国家/l 在/p 人权/n 方面/n 采取/v 双重/b 标准/n ，/w 或者/c 以/p 自己/r 的/u 模式/n 强加于人/l ，/w 借口/n “/w 人权/n 问题/n ”/w 干涉/v 他国/r 内政/n 的/u 霸权主义/n 行径/n 。/w
10. 中国/ns 在/p 人权/n 问题/n 上/f 的/u 原则/n 立场/n ，/w 得到/v 了/u 世界/n 许多/m 国家/n 的/u 支持/v ，/w 为/p 维护/v 世界/n 和平/a ，/w 促进/v 国际/n 人

权/n 事业/n 的/u 健康/a 发展/v ，/w 做出/v 了/u 自己/r 的/u 努力/a 。/w

11. 虽然/c 这/r 几/m 年/q 中国/ns 在/d 促进/v 人权/n 方面/n 取得/v 了/u 显著/a 成绩/n ，/w 但是/c ，/w 受/v 历史/n 和/c 发展/vn 水平/n 的/u 限制/v ，/w 中国/ns 的/u 人权/n 状况/n 还/d 存在/v 着/u 一些/m 不如/v 人意/n 的/u 地方/n 。/w
12. 继续/v 维护/v 和/c 促进/v 人权/n ，/w 不断/d 地/u 提高/v 全体/n 人民/n 享受/v 人权/n 的/u 水平/n ，/w 仍然/d 是/v 中国/ns 人民/n 和/c 政府/n 的/u 一/m 项/q 长期/b 的/u 任务/n 。/w

树库校对结果文件

1. (hl (wbl (《) !np (np (nps (!ns (中国)) !np (np (!n (人权)) !np (!n (事业)))) ude1 (的) !vp (!v (进展))) wbr (》)))
2. (np (!n (前言)))
3. (zj (!fj (tp (tp (!t (1991 年)) !tp (!t (11 月))) wco (，) !fj (!dj (np (nps (!ns (中国)) !np (!n (政府))) !vp (!vp (!vp (!v (发表)) ule (了)) np (wbl (《) nps (!ns (中国)) ude1 (的) !np (np (!n (人权)) !np (!n (状况))) wbr (》)))) wco (，) dj (pp (!p (向) np (ap (!b (国际)) !np (!n (社会)))) !vp (!vp (!vp (!v (阐述)) ule (了)) np (nps (!ns (中国)) !np (pp (!p (在) np (np (!n (人权)) !np (!n (问题))) f (上)) ude1 (的) !np (!np (np (!n (基本)) !np (!n (立场))) c (和) np (!n (实践))))))))) wfs (。)))
4. (zj (!dj (tp (!qp (m (四) !q (年)) m (来)) wco (，) !dj (np (nps (!ns (中国)) ude1 (的) !np (np (!n (人权)) !np (!n (事业)))) !vp (dp (!d (又)) !vp (!vp (!vp (!v (取得)) ule (了)) np (ap (!a (新)) ude1 (的) !vp (!v (进展))))))) wfs (。)))
5. (zj (!dj (qp (tp (tp (!t (1991 年)) !u (以来)) ude1 (的) !qp (m (四) !q (年))) wco (，) !vp (!vp (!v (是)) np (dj (nps (!ns (中国)) !vp (!vp (dp (!d (全面)) !vp (!v (实施))) np (np (!np (!n (国民经济)) c (和) np (n (社会) !v (发展))) !np (qp (mp (!m (第) mp (!m (八))) !q (个)) !np (qp (m (五) !q (年)) !np (!n (计划))))))) ude1 (的) !np (ap (!a (重要)) !np (np (!n (历史)) !np (!n (时期))))))) wfs (。)))
6. (zj (!fj (!fj (!dj (np (nps (!ns (中国)) ude1 (的) !np (!np (!n (国民经济)) c (和) np (n (社会) !v (发展)))) !vp (!iv (突飞猛进))) wco (，) dj (np (vp (!v (原定)) !np (dj (tp (!v (到) tp (!t (2000 年))) !dj (npz (!nz (国民生产总值)) !vp (pp (!p (比) tp (!t (1980 年))) !vp (!vp (!v (翻)) qp (m (两) !q (番)))))) ude1 (的) !np (!n (计划)))) !vp (dp (!d (已)) !vp (pp (!p (于) tp (!t (1995 年))) !vp (vp (!v (提前)) !vp (!v (完成))))))) wco (，) fj (!dj (np (np (!n (人民)) !np (!n (生活))) !vp (ap (!a (显著)) !vp (!v (改善)))) wco (，) dj (dp (!d (正在)) !vp (pp (!p (向) np (np (!n (小康)) !np (!n (目标)))) !vp (!v (前进)))))) wfs (。)))
7. (zj (!fj (tp (!t (当前)) wco (，) !fj (fj (!dj (nps (!ns (中国)) !fj (!fj (!fj (!dj (np (!n (政治)) !vp (!v (稳定))) wsc (、) dj (np (!n (经济)) !vp (!v (发

展)))) wsc (、) dj (np (!n (社会)) !vp (!v (进步)))) wsc (、) dj (np (!n (民族)) !vp (!v (团结))))) wco (，) fj (!dj (np (!n (人民)) !vp (!iv (安居乐业))) wsc (、) dj (np (n (生活) !n (水平)) !vp (dp (!d (不断)) !vp (!v (提高)))))) wco (，) !dj (np (np (!n (人权)) !np (!n (状况))) !vp (!vp (!v (呈现)) np (vp (ap (!a (全面)) !vp (!v (改善))) ude1 (的) !np (ap (!a (良好)) !np (!n (态势)))))))) wfs (。)))

8. (zj (!dj (np (!n (实践)) !vp (!vp (!v (证明)) wco (，) dj (vp (vp (pp (!p (将) np (np (!n (人民)) ude1 (的) !np (!np (v (生存) !ng (权)) wsc (、) np (v (发展) !ng (权)))))) !vp (!v (摆) p (在) n (首位))) wco (，) !vp (pp (!p (在) np (vp (!vp (!vp (!v (改革)) wsc (、) vp (!v (发展))) wsc (、) ap (!a (稳定))) ude1 (的) !np (!n (条件))) f (下)) !vp (ap (!a (全面)) !vp (!vp (!v (改进)) np (np (!n (人权)) !np (!n (状况)))))))) wco (，) !vp (!vp (!v (是) vp (!vp (!v (符合)) np (!np (nps (!ns (中国)) !np (!n (国情))) c (和) np (np (ap (!b (全体)) !np (!n (人民))) ude1 (的) !np (np (!n (根本)) !np (!n (利益)))))) ude1 (的)) wco (，) vp (dp (!d (也)) !vp (!v (是) vp (!iv (举世公认)) ude1 (的))))))) wfs (。)))
9. (zj (!dj (tp (!qp (m (四) !q (年)) u (来)) wco (，) !dj (nps (!ns (中国)) !vp (!vp (ap (!a (积极)) !vp (!vp (!v (维护)) np (np (wbl (《) npt (!nt (联合国)) !np (!n (宪章)) wbr (》)) !np (vp (!vp (!v (促进)) np (!np (!n (人权)) c (和) np (a (基本) !a (自由)))) ude1 (的) !np (!np (!n (宗旨)) c (与) np (!n (原则))))))) wco (，) vp (ap (!a (坚决)) !vp (!vp (!v (反对)) np (dj (np (qp (m (一) !q (些)) !np (!n (国家))) !vp (pp (!p (对) np (!np (rn (其他) !np (!n (国家))) c (特别是) np (!ln (发展中国家)))) !vp (pp (!p (在) np (np (!n (人权)) !np (!n (方面)))) !vp (!vp (!vp (!v (采取)) np (ap (!b (双重)) !np (!n (标准)))) wco (，) c (或者) vp (!vp (pp (!p (以) np (np (!rn (自己)) ude1 (的) !np (!n (模式)))) !vp (!v (强加) p (于) n (人))) wco (，) vp (vp (!vp (!v (借口)) np (wql (“) np (!n (人权)) !np (!n (问题)) wqr (”))) !vp (!vp (!v (干涉)) np (np (r (他) !ng (国)) !np (!n (内政))))))))) ude1 (的) !np (np (!n (霸权主义)) !np (!n (行径))))))))) wfs (。)))
10. (zj (!fj (!dj (np (nps (!ns (中国)) !np (pp (!p (在) np (np (!n (人权)) !np (!n (问题))) f (上)) ude1 (的) !np (np (!n (原则)) !np (!n (立场))))) wco (，) !vp (!vp (!vp (!v (得到)) ule (了)) np (np (np (!n (世界)) !np (ap (!a (许多)) !np (!n (国家)))) ude1 (的) !vp (!v (支持))))) wco (，) dj (pp (!p (为) vp (!vp (!vp (!v (维护)) np (np (!n (世界)) !np (!n (和平))))) wco (，) vp (!vp (!v (促进)) np (np (ap (!b (国际)) !np (np (!n (人权)) !np (!n (事业)))) ude1 (的) !vp (ap (!a (健康)) !vp (!v (发展))))))) wco (，) !vp (!vp (!vp (!v (做出)) ule (了)) np (np (!rn (自己)) ude1 (的) !ap (!a (努力)))))) wfs (。)))
11. (zj (!fj (!dj (ch (虽然) !dj (qp (r (这) !qp (m (几) !q (年))) !dj (nps (!ns (中国)) !vp (pp (!p (在) np (vp (!vp (!v (促进)) np (!n (人权))) !n (方面))) !vp (!vp (!vp (!v (取得)) ule (了)) np (ap (!a (显著)) !np (!n (成绩)))))))) wco (，) dj (ck (但是) wco (，) !dj (vp (!vp (!v (受)) np (np (!np (!n (历史)) c (和) np (vp (!v (发展)) !n (水平))) ude1 (的) !vp (!v

(限制)))) wco (，) !dj (np (nps (!ns (中国)) ude1 (的) !np (np (!n (人权)) !np (!n (状况)))) !vp (dp (!d (还)) !vp (!vp (!vp (!v (存在)) uzhe (着)) np (qp (!m (一些)) !np (vp (!iv (不如人意)) ude1 (的) !np (!n (地方)))))))))) wfs (。)))

12. (zj (!dj (vp (!vp (!vp (!v (继续)) vp (!vp (!vp (!v (维护)) c (和) vp (!v (促进))) np (!n (人权)))) wco (，) vp (dp (!dp (!d (不断)) ude2 (地)) !vp (!vp (!v (提高)) np (dj (np (ap (!b (全体)) !np (!n (人民))) !vp (!vp (!v (享受)) np (!n (人权)))) ude1 (的) !np (!n (水平)))))) wco (，) !vp (dp (!d (仍然)) !vp (!vp (!v (是)) np (np (nps (!ns (中国)) !np (!np (!n (人民)) c (和) np (!n (政府)))) ude1 (的) !np (qp (m (一) !q (项)) !np (ap (!b (长期)) ude1 (的) !np (!n (任务)))))))) wfs (。)))

附录五　北大中文树库与宾州大学树库标注体系对比

一 概述

从整个体系上，宾州树库（UPENN）跟北大中文树库（BDTB）相比有两个大的差别：宾州树库（UPENN）总体的理论指导是生成语法的管约（GB）理论，考虑了句子的深层结构假设。北大树库（BDTB）则只针对句子的表层结构进行标注。UPENN 有一套空语类标记（7 个）。BDTB 中没有这类标记。这些标记对反映动词的论元结构比较有帮助。但空语类标记中有的存在不易判别的问题，比如用于标记关系化结构中的空成分的“*OP*”标记，就比较费解。此外，空语类标记用来标示介词短语的空位结构时也十分的繁琐和费解，甚至感觉有些累赘。

另一个是大的差异是 Upenn 除句法功能标记外，还设计了语义功能标记，如：APP(同位),BNF(受益),TMP(时间),CND(条件),DIR(趋向),EXT(范围,程度,频率),FOC(焦点),LGS(逻辑主语),LOC(方位),MNR(方式),PRP(目的或原因),TPC(话题)等等。BDTB 中没有这些标记。

此外，在一些具体语言单位的标注处理上，两个系统也存在一些差异。比如：

（一） UPENN 没有区分单句\复句，采用的是 IP 跟 CP 的对立划分方式。

（二） UPENN 对 VP 的分类比较细。大致上有：

（1）VA,VC(是),VE(有),VV 四个部分，将"是"和"有"作特殊的处理。

（2）将"继续,正在,已经"这类副词归为 ASPECT VERB（体动词）

（3）提出五条标准严格区分复合动词和动词短语。这里用来标记复合动词的标记共 6 个：VCD,VCP,VNN,VPT,VRD,VSB。将象"建立起,下降到,表达出,看作是"这样的结构处理为词一级的单位，甚至可进一步将“A 一 A”，“A 不 A”，“A 得 A”这种“三分”结构处理为词一级的单位。

（4）将“SUBJECT CONTROL VERB”（给，还，送，欠，罚，骗，教，问）和 “OBJECT CONTROL VERB”（劝，逼，使，引诱，原谅）区别开，同时又将“SUBJECT CONTROL”，“OBJECT CONTROL”动词和心理动词也用五条标准区分开，虽然他们可以进入一样的结构（NP+VP+IP），但由于“vp”分属于“sub ctrl”，“obj ctrl”和心理动词使得其内部语法语义关系发生变化。

（5）UPENN 将把字句，被字句中的“把”“被”处理为动词。UPENN 对被字句进一步细化，以“被“后有无 NP 分为长被字句和短被字句.

（6）UPENN 将“长，宽，重，值，达“处理为量度动词。

（7）UPENN 将“是否“处理为助动词。

（8）UPENN 有一类动词叫“RAISING VERB”。比如出现在句首的情态动词，“好像，看起来，似乎”等等。与之相关的是空语类的“*RNR*”(right node raising)。

（9）UPENN 的 VA 和用来做修饰语的 JJ 合起来构成形容词类，也区分了作谓语的形容词，区别词和状态形容词。

（三）UPENN 的并类结构有 UCP 的标记，用来标注由不同词类构成的并列结构。

二 UPENN树库的基本框架

1 词性标记

标记	含义	标记	含义
AD	adverbs		
AS	aspect marker	M	measure word (including classifiers)
BA	in ba-const	MSP	some particles
CC	coordinating conj	NN	common nouns
CD	cardinal numbers	NR	proper nouns
CS	subordinating conj	NT	temporal nouns
DEC	的 for relative-clause etc.	OD	ordinal numbers
DEG	associative 的	ON	onomatopoeia
DER	得 in V-de const. and V-de-R	P	prepositions (excluding and)
DEV	地 as the head of DVP	PN	pronouns
DT	determiner	PU	punctuation
ETC	tags for and in coordination phrases	SB	in short bei-construction
FW	foreign words	SP	sentence-final particle
IJ	interjection	VA	predicative adjective
JJ	noun-modifier other than nouns	VC	copula
LB	in long bei-construction	VE	as the main verb
LC	localizer	VV	other verbs

特点：

（1） 长“被” – 短“被”的区别

（2） FW 标记

（3） VC，VE 分得比较细

（4） ON（onomatopoeia 拟声词）

（5） DEG 和 DEC 分开

（6） BA 跟 P 没有形式上的联系。标记未体现层级性。

2 短语句法标记

标签	含义	标签	含义
ADJP	adjective phrase		
ADVP	adverbial phrase headed by AD (adverb)	PP	preposition phrase
CLP	classifier phrase	PRN	parenthetical
CP	clause headed by C (complementizer)	QP	quantifier phrase
DNP	phrase formed by “XP + DEG”	UCP	unidentical coordination phrase
DP	determiner phrase	VP	verb phrase
DVP	phrase formed by “XP + DEV”	VCD	coordinated verb compound
FRAG	fragment	VCP	verb compounds formed by VV + VC

IP	simple clause headed by I (INFL)	VNV	verb compounds formed by A-not-A or A-one-A
LCP	phrase formed by ‘‘XP + LC’’	VPT	potential form V-de-R or V-bu-R
LST	list marker	VRD	verb resultative compound
NP	noun phrase	VSB	verb compounds formed by a modifier + a head

特点：

（1） Verb compound

（2） FRAG

（3） LST （破折号开头的句子、 可以提示篇章连贯/排比句|多项并列句）

（4） CP 和 IP 对应一般的句子

（5） PRN、UCP

3 短语功能标记

标签	含义	标签	含义
ADV	adverbial		
APP	appositive	PRP	purpose or reason
BNF	beneficiary	Q	question
CND	condition	SBJ	subject
DIR	direction	SHORT	short form
EXT	extent	TMP	temporal
FOC	focus	TPC	topic
HLN	headline	TTL	title
IJ	interjective	WH	wh-phrase
IMP	imperative	VOC	vocative
IO	indirect object	*OP*	operator
LGS	logic subject	*pro*	dropped argument
LOC	locative	*PRO*	used in control structures
MNR	manner	*RNR*	right node raising
OBJ	direct object	*T*	trace of A’-movement
PN	proper names	*	trace of A-movement
PRD	predicate	*?*	other unknown empty categories

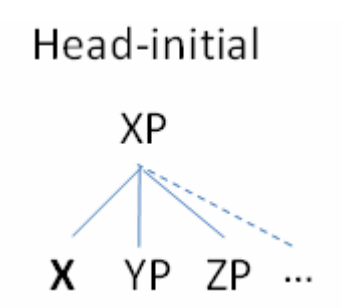
特点：

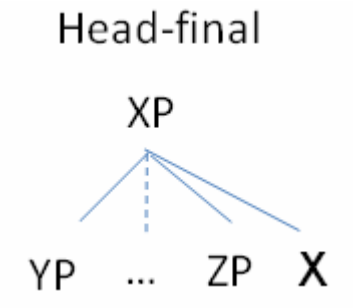
（1） 设置了 7 个空范畴标记

（2） 语义范畴、功能标记详细。

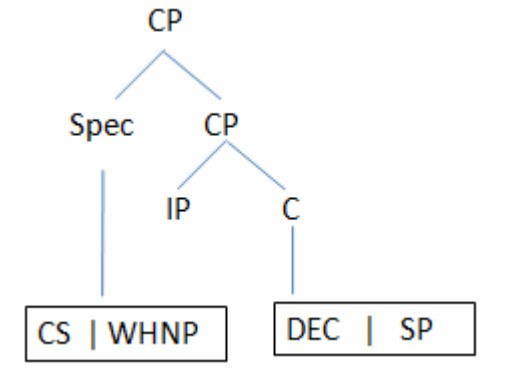
4 句子结构模式

1）补足语结构

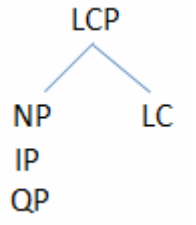




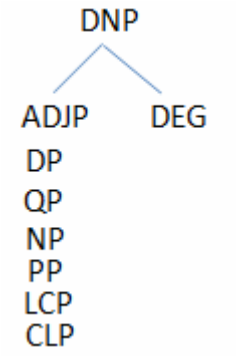
中心词后置的具体短语类型：（1）-（4）　　中心词前置的具体短语类型：（5）-（6）



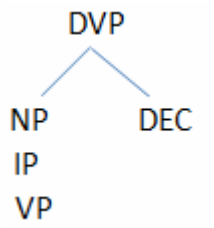
（1）



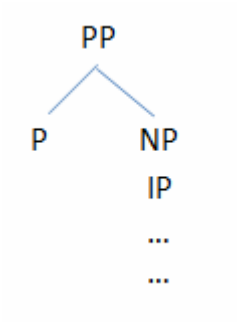
（2）



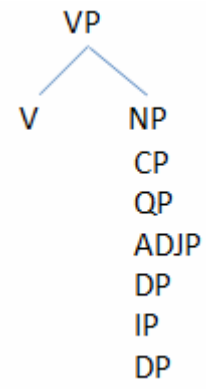
（3）



（4）

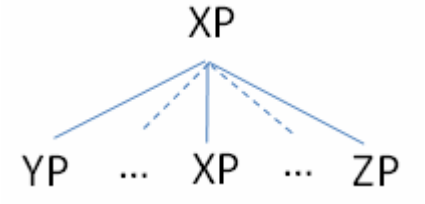


（5）

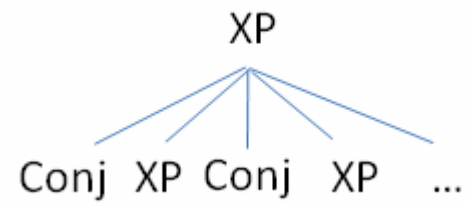


（6）

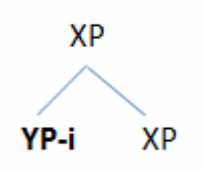
2）附接语结构

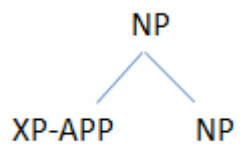


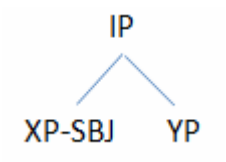
3）并列结构



4）修饰语结构　　5）同位语结构　　6）陈述结构

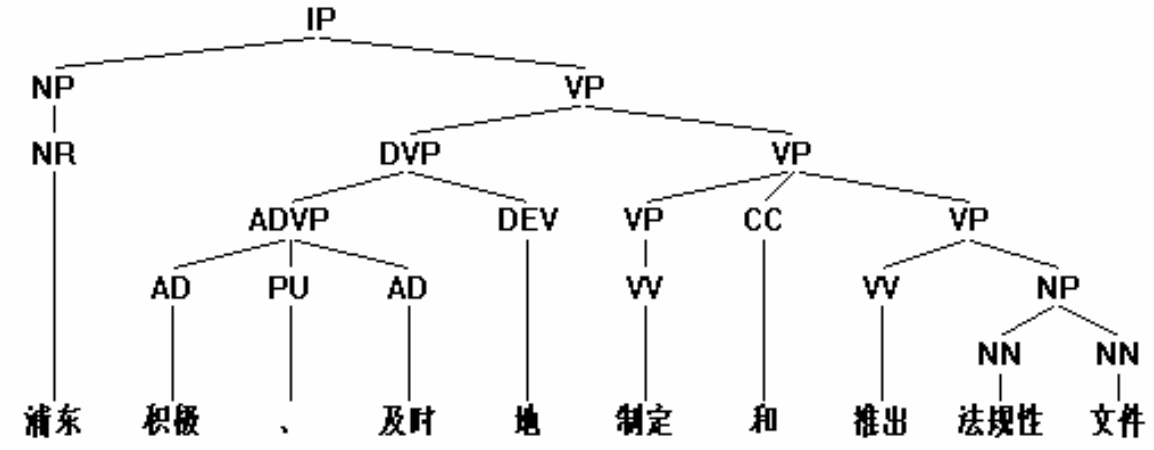


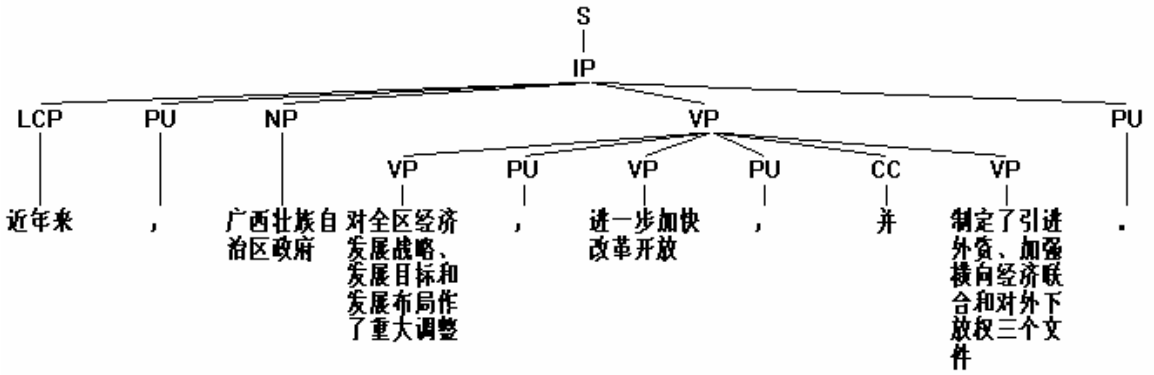




i ∈{TPC,TMP,LOC,DIR,...}

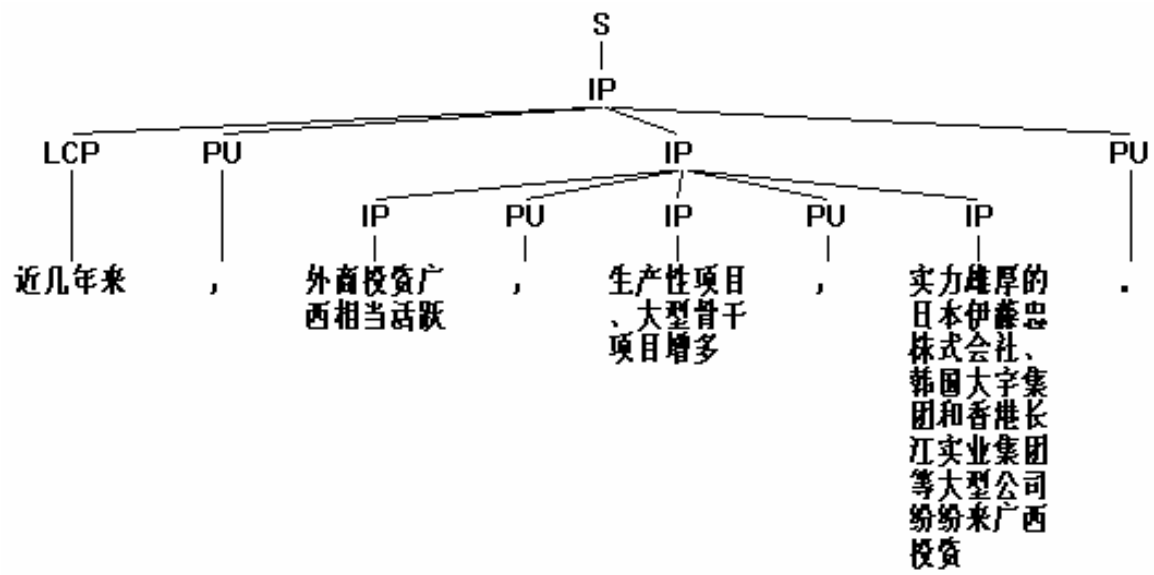
5 UPENN 树结构示例





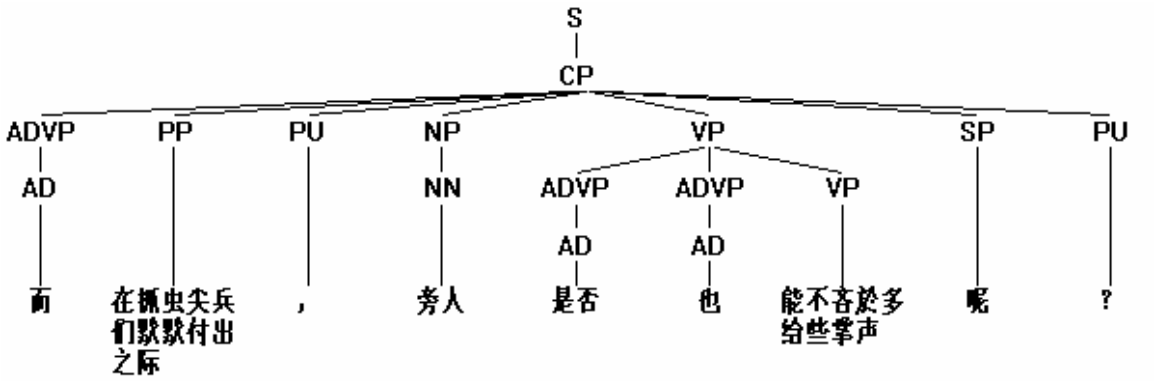
对应 CFG 规则：　　IP → LCP PU NP VP PU

VP → VP PU VP PU CC VP



对应 CFG 规则： IP → LCP PU IP PU

IP → IP PU IP PU IP



对应 CFG 规则： CP → ADVP PP PU NP VP SP PU

VP → ADVP ADVP VP

三 UPENN 与 BDTB 标记体系详细对照

<table>
<tr><td colspan="4">BDTB 标记集</td><td colspan="4">Upenn 标记集</td></tr>
<tr><td>标记</td><td>语法功能范畴</td><td>说明</td><td>示例</td><td>标记</td><td>语法功能范畴</td><td>说明</td><td>示例</td></tr>
<tr><td colspan="8">一级标记：不能被任何其他标记包含</td></tr>
<tr><td>zj</td><td>整句</td><td>末尾带有标点的完整表达单位</td><td>他喜欢大眼睛姑娘。</td><td colspan="4">条件：每行开头的第一个 IP/CP，100% 对应
if (IP 的父节点为空 or CP 的父节点为空)
then IP=zj; CP＝zj;
else if（IP 的子节点的左结点＝NP & 子节点的右结点＝VP）
then IP＝zj;
zj 的子节点＝dj 和其他层;
dj 的子节点的左结点＝NP;
dj 的子节点的右节点＝VP;</td></tr>
<tr><td colspan="8">二级标记：可被其他标记包含，也可包含其他标记</td></tr>
<tr><td>fj</td><td>复句</td><td>抽象的句法结构单位</td><td>他喜欢大眼睛姑娘，我不喜欢。</td><td colspan="4">IP → CP/IP+PU+(ADV)+IP+PU＋……</td></tr>
<tr><td>dj</td><td>单句</td><td>抽象的句法结构单位</td><td>他喜欢大眼睛姑娘。</td><td colspan="4">IP→ (PP)＋IP+(SP)</td></tr>
<tr><td colspan="4">dj
vp
ap</td><td>IP</td><td colspan="3">IP－>NP+VP（除去行首的 IP）
IP－>VP+ *
IP(它的子节点中有 a)</td></tr>
<tr><td colspan="4"></td><td>CP</td><td colspan="3">CP 下层一般都为 IP＋空语类，所以只需处理空语类即可，剩下的工作就由 IP 的对应转换完成</td></tr>
<tr><td>np</td><td>名词短语</td><td></td><td></td><td>NP</td><td>名词短语</td><td colspan="2"></td></tr>
<tr><td rowspan="2">ap</td><td rowspan="2">形容词短语</td><td rowspan="2"></td><td rowspan="2"></td><td>ADJP</td><td>形容词短语</td><td colspan="2"></td></tr>
<tr><td>VP</td><td>动词短语</td><td>VP－>VA</td><td></td></tr>
<tr><td>vp</td><td>动词短语</td><td></td><td></td><td>VP</td><td>动词短语</td><td colspan="2"></td></tr>
</table>

dp	副词短语			ADVP	副词短语		
		dp－>xp+u		DVP	状语	标示由“XP+地”构成的短语	内地经济长期稳定地增长。
pp	介词短语			PP	介词短语		
tp	时间短语			TMP	时间短语		
sp	处所短语			LOC	处所短语	标示隐含了时间发生地点的短语	在上海
				LCP	方位短语	标示由“定位词＋补语化成分（如上，中，下）”构成的短语	皖南事变中
qp	数量短语	数词后有量词		QP	数量短语		三个
				CLP	量词短语		他们家三口人。
mp	数词短语	数词后无量词		QP	数量短语		一百零一
				LST	列举短语	标示用来列举条款项目的短语	他为群众办事，一不怕吃亏，二不怕吃苦。
yp	语篇成分		总而言之，他喜欢大眼睛姑娘。	??			
yph	呼语			VOA			**老叶**，这车子怎么老往下走呢？
ypc	插入		据说，他喜欢大眼睛姑娘。	PRN	插入成分附加成分	*两插入成分意义不同，鉴于文本中出现的可能性小，不做考虑*	“我们相信”，**张三说**：“你们会赢的。”

以上为较整齐的对应，以下是将 UPENN 的标记逐个拆分为 BDTB 的标记对应形式

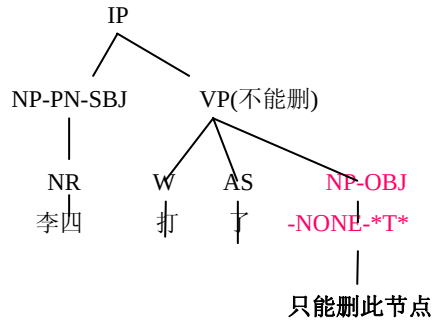
BDTB	UPENN
“xp＋u（的）”	“XP＋dec（的）”即 DNP

把 UPENN 中结构里的 XP 具体标记映射到 BDTB 的标记集中，即 NP — np

np+“的”（二分）	DNP → NP+DEC			
pp+“的”（二分）	DNP → PP+DEC			
sp+“的”（二分）	DNP → LCP+DEC			
mp+“的”（二分）	DNP → QP+DEC			
ap+“的”（二分）	DNP → ADJP+DEC			
np → r	DP	指称短语	标示由“指示代词+数量/名词短语”构成的短语 NP —>DP+NP	**任何**人、 **这五个**学生
没此层次 删除即可	UCP	非同类成分并列结构	标示内部各成员不属于同一语类的并列结构	养老、医疗保险
np_foc vp_foc	NP_FOC VP_FOC（后者居多）	焦点	标示提前到动词前，主语后位置上的原宾语	张三**作业**作完了。 ？甘肃省**利用外资**蓬勃发展 和代表们**合影留念** **进展顺利**
np_sbj np_pn_sbj dj_sbj sp_sbj mp_sbj	NP_SBJ NP_PN_SBJ IP_SBJ LCP_SBJ QP_SBJ	主语	标示表层结构主语	**这座楼**十八层。 对此，**浦东**不是简单的采取… …中心…，***pro***运转至今，成交… **世界上最大的二百二十五家国际承包商中**，有十几家已经… 如果**两队**以 7：7 战平
pp_lgs	PP_LGS	逻辑主语	标示隐含逻辑主语的状语	《新中东》已**由新华出版社**出版发行
np_tpc np_pn_tpc vp_tpc （大多数）	NP_TPC NP_PN_TPC IP_TPC	话题	（如有*pro*则有空位出现，删去此层）	**李四**张三打了。 但今天**中国队**发挥较好，所以获得了胜利。 **接到通知书**，我真想跳起来。 **水果**我喜欢苹果

np qp pp dj vp	40％ NP(PN)_PRD 50％QP_PRD PP_PRD IP_PRD VP_PRD DP_PRD	断言	70%在“是\为\近”后面出现	建筑是**开发浦东的一项主要经济活动**。 去年为**七十六点八亿美元**。 中外合作合资的企业近**两千家**。 运量**在万吨上**的企业就有两千多家。 推动经济发展的主要因素是，**亚洲经济发展依然强劲有力**。 候家驹指出，台湾当局说，“戒急用忍”是**限制大型企业赴祖国大陆投资**，对中小企业并无影响，实则非然。 全年出口值**前十位**的省市是，…。
np dj qp	NP_OBJ IP_OBJ QP_OBJ	宾语	标示动词的直宾，不含间宾和范围宾语	张三参加了**会议**。 他认为，**这次比赛我国参赛的新选手多**。 暂居**第三位**的津巴布韦选手
np	NP_(PN)_IO	间接宾语		留给**你**一定的难度
np → dj+np np → dj+的+np np	IP_APP CP_APP NP_APP	同位短语		在**难民完全自愿前提**下 在**全国尚有40多个贫困县接受财政补助的情况**下 **美国总统尼克松**
pp	PP_BNF	受益短语		西门子**为地铁二号线**提供设备和资金。
dj → c + dj	CP_CND （ADV + IP）	条件短语	标示表示必要或充分条件的条件的短语	如果那天我登上了那个险峰
pp	PP_DIR	方向短语	标示可以回答“从那儿来”“到那儿去”问题的短语	**向其他省市**输送部分商品气。
qp	LCP_EXT QP_EXT NP_EXT	范围短语	动词后描述范围、频率、数量的短语	**两成半以上** 跑了**两次** 出口总值一**百亿美元**

dp		NP_ADV	状语		**年**产汽车五万辆
pp		PP_MNR	方式短语	表示动作发生的方式、工具、凭借	**以百分之三十的速度**增长
np		NP_PN	专有名词短语	标示表示专有名词（人名、地名、机构等）的短语	**中国建设银行**
zj		IP_Q CP_Q	疑问语气	标示含有疑问语气的短语或子句	**你做了什么？** **这不是天津吗？**
zj		IP_IMP(文本无例子)	祈使语气	标示含有祈使语气的短语或子句	**弟兄们，开门！**
ypc		NP_IJ NP_PN_IJ	感叹语气	标示含有感叹语气的短语或子句	我仰头看去，**好家伙**，至少还有一千米高。
“被”字句（在删除空语类之前先做）	pp－>p＋np（np 为“被”的兄弟结点往下找到的 np-sbj）	LB－>被			
vp+vp		VCD	并列的复合动词		**开发建设**无人区 经济发展**强劲有力**
vp+vp		VCP		由“VV+VC”构成的复合动词	**估计是**、**看成是**
vp+dp+vp		VNV		由“v+one+v”“v+bu+v”构成的复合动词	**能不能**
vp+dp+vp		VPT	可能式复合动词	由“v+得+v”“v+不+v”构成的复合动词	**分不开** **进得了**
vp+vp		VRD	趋向复合动词	“V1+V2”V2 指示 V1 的方向	**建立起**、 **提取出**、 **发展成为**
vp+vp		VSB	修饰性复合动词	“vi+v”vi 修饰 v,且之间不能有体动词和附加成分	**驱车行程** 800 多公里 他**介绍说** **加以推进**

查找到空语类（*） 且空语类的兄弟结点只有一个，则 ①删除空语类自身 ②删除父节点 ③兄弟节点上升到父节点位置 若空语类的兄弟结点>=2,则 不执行②③ 只执行① 譬如：张三被 李四 打了 。  	*OP*	空操作者	标示关系从句的空操作者 NP, PP	
	pro	省略的论元	标示省略的主宾语的空位，该空位可以用显性成分补出	
	PRO	隐含的论元	标示控制结构中隐含的空语类，且该空语类不能用显形成分不出	
	RNR	标示右节点上升的成分		
	T	A’的语迹	标示如话题或宾语前置产生的空位	
	*	A的语迹	标示上升和被动结构的空位	
	?		未知的空语类	