# 树库中的歧义组合考察*

李艳娇，杨尔弘
北京语言大学 应用语言学研究所，北京 100083
E-mail: yanjiao8119@126.com

**摘　要：**汉语树库是汉语信息处理的宝贵资源，其中包含了丰富的句子结构及成分组合信息，对树库中的词性串组合进行考察，是有效利用树库信息的基础工作。本文对汉语树库中歧义组合进行考察，发现汉语中的歧义组合、歧义结构很大程度上要靠词语的语义关系来化解，单纯依靠外在的句法信息是无法实现的。

**关键词：**歧义组合；树库

# The Study of Ambiguous Combinations in Treebank

Li Yanjiao, Yang Erhong
Applied Linguistic Institution at BLCU, Beijing 100083
E-mail: yanjiao8119@126.com

**Abstract:** Chinese Treebank, a kind of valuable resource in Chinese Information Processing, includes rich information on sentences structure and composition combination. The study on the combination of POS string is the basic work for the effective use of treebank information. This paper makes a study on the ambiguous combination in Chinese Treebank, and finds that it largely depends on semantic relations to resolve the ambiguous combination and structure in Chinese, and sole external syntactic information can not make it.

**Keywords:** ambiguity combinations; treebank

## 1　引言

树库作为包含语言结构信息的资源，其价值与作用得到了人们的肯定。“首先，它可为基于统计的自动句法分析器提供必要的训练数据和统一的测评平台；其次，它能为汉语句法学研究提供真实文本标注素材，便于语言学家从中总结语言规则和规律；第三，它是进一步进行句子内部的词语义项和语义关系标注的基础。”（王跃龙、姬东鸿，2009）[1]

歧义是指一个句法结构可以对应多种组合方式，即对应多棵树。对于计算机而言，要在多个结构中选择一个合适的句法结构，需要各种知识，通常统计的训练模型主要利用的是一种结构在特定环境中的概率分布知识。本文通过对树库语料中三元词性序列组合方式的考察，发现汉语中的歧义组合很大程度上要靠词语内部的语义关系来化解，上下文环境的句法信息作用甚小。

## 2　基于树库的考察

### 2.1　语料说明

清华大学树库（Tsinghua Chinese Tree-bank，TCT）是国内第一个大规模汉语树库，也是一个标注信息最丰富的短语结构树库（详见标注规范[2]）。1998 年到 2002 年间完成了 100 万词的建设。本文所使用的语料是 TCT 中经过人工校对的 150 个文件，共 7063 个句子。

### 2.2　考察对象

名词、动词、形容词是汉语的三大词类，清华树库中名词、动词、形容词共有 14 种不同的标

*本文受“中央高校基本科研业务费专项资金”资助。
在本文的写作中，董振东教授提出了很多宝贵的意见，在此表示诚挚的感谢！

记符号（包括小类）。本文主要考察名词、动词、形容词（包括小类）在连续线性序列上的组合情况，共27（$3^3$）种。具体方法是，将连续出现的三个词串（名词n、动词v、形容词a的任意组合）提取出来，然后统计27种模式的实例数量，将排在前十位的模式作为本为的考察对象（下文一些具体的标注符号请参见相关规范[3][4][5]）。

## 2.3 考察结果与分析

### 2.3.1 总体情况

通过数据统计，可以得到十种模式的不同组合情况，如表1所示。

表1 十种模式的组合情况

| 模式 | 实例总数 | 不同的组合方式 | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | [[AB]C] | | [A[BC]] | | [ABC] | | 其他 | |
| | | 数量 | 百分比 | 数量 | 百分比 | 数量 | 百分比 | 数量 | 百分比 |
| n+v+n | 1905 | 235 | 12.3% | 295 | 15.5% | 0 | 0 | 1375 | 72.2% |
| v+n+n | 1802 | 153 | 8.5% | 632 | 35.1% | 1 | 0.09% | 1016 | 56.4% |
| n+n+v | 1796 | 229 | 12.8% | 38 | 2.1% | 0 | 0 | 1529 | 85.1% |
| n+n+n | 1563 | 596 | 38.1% | 345 | 22.1% | 30 | 1.9% | 592 | 37.9% |
| v+v+n | 1374 | 120 | 8.7% | 342 | 24.9% | 0 | 0 | 912 | 66.4% |
| v+n+v | 1205 | 139 | 11.5% | 156 | 12.9% | 1 | 0.08% | 909 | 75.4% |
| n+v+v | 1152 | 6 | 0.5% | 82 | 7.1% | 0 | 0 | 1064 | 92.4% |
| v+v+v | 618 | 13 | 2.1% | 119 | 19.3% | 10 | 1.6% | 476 | 77.0% |
| v+a+n | 454 | 8 | 1.8% | 208 | 45.8% | 0 | 0 | 229 | 50.4% |
| a+n+n | 370 | 123 | 33.2% | 57 | 15.4% | 0 | 0 | 190 | 51.4% |

在a[A[BC]]、b[[AB]C]两种组合方式中，十种模式分布相异，有的倾向a式组合，如“v+a+n”模式，有的倾向b式组合，如“a+n+n”模式。相比起来，“n+n+n”模式a、b两种组合所占的比例都很高，“n+v+v”模式，两种组合方式所占的比例都很低（二者之和不足10%）。

“v+n+n”“n+n+n”“v+n+v”“v+v+v”四种模式分别有三种不同的组合方式，比其他模式多了一种[ABC]的组合。三个词内部不再分层次、直接组合成一个整体，在“n+n+n”和“v+v+v”模式中尤为明显，说明三个词连续出现，若词性相同，则直接组合成一个整体的可能性要大于词性不同的情况，从另一个角度看，词性相同的词连续出现，内部的组合情况更为复杂。

其他组合是十种模式都存在的组合情况，并且除“n+n+n”外，其余九种模式的其他组合所占比例都在50%以上。这里的其他情况是指三者之间或者互不相干，如，收回/v 澳门/nS，/，…… ，/，是/vC [np-DZ [np-DZ [np-DZ 我/rN 国/n ] 人民/n ] **[np-DZ 长期/n** 的/u [np-DZ 共同/a 愿望/n ] ] ] ] ] …….。/。（n+n+n模式，“国”没有直接和“人民”结合，而是与前面的代词“我”组合，“长期”则是“共同愿望”的定语，所以这里连续的三个n没有直接联系）或者是局部组合，即连续的两个词语构成短语，另一个词（称其为“无关词”）与句中的其他成分相关，三者之间不能直接构成一个结构整体。

局部组合中，根据“无关词”的位置，上述每种模式还会有两个小类，一个是前无关，一个是后无关，以“a+n+n”模式为例，

例1 ……农产品/n 的/u 深加工/n 和/c 饲养业/n 的/u 进一步/d 发展/v ，/，必须/vM [vp-PO [vp-SB 解决/v **好/a ] [np-DZ 市场/n** 问题/n ] ] ] ] …… 。/。

例2 科学家/n 们/k 回顾/v 我/rN 国/n 建造/v 对撞机/n 取得/v 的/u **[np-DZ 巨大/a**

**成就/n ] ] ] ] 时/n ]** ，/，总/d 忘/v 不/u 了/vB 小平/nP 同志/n 。/。

例 1 中，“市场”与“问题”组合成一个整体，“好”与“市场问题”没有直接关系，这种情况称为“前无关”。例 2 中，“巨大”和“成就”先局部组合成定中结构，后面表示时间的“时”与“巨大成就”也没有直接关系，这种情况称为“后无关”。在实际的语料中，无论哪种模式，前无关和后无关的例子都是大量存在的，说明三个词虽然线性出现，依次排列，但在语义上并不一定直接相关联，也正是由于语义的这种不相关造成了“前/后无关”的现象。

广义上看，无论是局部组合，还是三者互不相干，也是一种组合方式，并且这种组合方式在每种模式中所占的比例都很大，除 n+n+n 模式外，其他九种模式的其他组合都超过 50%，说明汉语的线性序列中，三个连续出现的词串（名词、动词、形容词中的任意组合）不倾向于内部组合形成一个整体，要么局部组合，要么互不相干。或者说，它们在位置上是“邻居”，但在句法组合中可以没有直接或任何“血缘关系”。这表明，汉语短语的组合与词语之间的语义信息密切相关，而与线性序列上所处的位置无关。

前面提到，每种模式都有 a [ A [ BC] ]、b [ [ AB] C ]两种不同的组合，有些情况下，同种模式的不同组合根据上下文环境或者本身的词性序列是可以化解掉的，如“n+v+v”模式中若 A 位置上是人名（nP）或地名（nS），B 和 C 位置上同时为普通动词（v），B 位置上是能愿动词（vM），C 位置上的动词是趋向动词（vB）四种情况中的一种时，倾向于 a 式组合，如：

例 3 …… “/“ 你/rN 做/v 得/u 对/a ，/，**[dj-ZW 车/n [vp-SB 留/v 下来/vB ] ]** ，/，我们/rN 走/v 进去/vB 就/d 行/v 了/u 。/。”/”

例 4 医生/n 说/v ，/，[dj-ZW 骨头/n [vp-ZZ 可能/vM 坏死/v ] ] …… 。/。

例 5 [dj-ZW 主人/n [vp-LW 介绍/v 说/v ] ] ……。/。

但很多情况下，这种歧义组合在词性标记符号的基础上是消解不了的，以下分别说明十种模式的歧义组合。

### 2.3.2 十种模式的歧义组合

1) n+v+n 模式

a1 ……成立/v 了/u 600/m 多/m 个/qN **[dj-ZW 民兵/n [vp-PO 送/v 温暖/n ] ]** 小组/n ……

b1 ……先后/d 分/v 片/qN 举办/v 了/u X 30/m 多/m 期/qN **[np-DZ [dj-ZW 商品/n 交易/v ] 知识/n ]** 培训班/n ]，/，……。/。

上面的两个例子分别属于 a 式（[ A [BC] ]）和 b 式（[ [AB] C ]）组合。“民兵/n 送/v 温暖/n”与“商品/n 交易/v 知识/n”两个短语的词性序列相同，说明两个短语中对应的具体词属于相同的词类范畴，但它们内部的组合方式却完全不同：a1 是动词“送”与后面的名词“温暖”先结合形成动宾结构，然后与前面的名词“民兵”结合，最外层形成主谓结构；而 b1 是动词“送”与前面的名词先结合形成主谓结构，主谓结构作定语然后修饰后面的名词，形成定中结构。说明决定两种组合方式的是由其内在的语义关系决定的，无论哪种组合，动词总是与其动作对象先结合（“送”的对象是“温暖”，“交易”的对象是“商品”），不论动作对象的位置在前还是在后，而这种歧义结构靠词性序列是很难化解的。

2) v+n+n 模式

a2 ……以往/t 那/rB 种/qN 认为/v **[vp-PO 干/v [np-DZ 人武/n 工作/n ] ]** 是/vC “/“ 敲边鼓/iV ”/” 的/u 思想/n 打掉/v 了/u 。/。

b2 …… **[np-DZ [vp-PO 种/v 菜/n ] 开支/n ]** 增加/v ，/，自然/d 影响/v 价格/n 。/。

在上面两个例子中，“干/v 人武/n 工作/n”与“种/v 菜/n 开支/n”两个短语不仅词性序列相同，更严格的看，两个短语句法位置也都是一样的，在句子中作主语，但它们的组合方式依然不

同。“种/v 菜/n 开支/n”是动词“种”与“菜”先组合成动词性的述宾结构，述宾结构作定语修饰“开支”，最外层形成名词性的定中结构；而“干/v 人武/n 工作/n”是“人武”与“工作”先组合，然后与前面的“干”在最外层形成动宾结构。这种组合的差异主要体现在内部语义关系的不同：“种/v 菜/n 开支/n”中“菜”是“种”的动作对象，“种菜”作为一个整体限定说明“开支”的用途；“干/v 人武/n 工作/n”中“人武”不是“干”的直接对象，而是限制说明“工作”的性质，“人武”要与“工作”组合成一个整体来作为“干”的对象。可见，这种不同的组合是由短语内部的语义信息决定的，与句法信息无关。

3) n+n+v 模式

a3 …… [dj-ZW 法乌斯蒂诺/nP [dj-ZW 头部/n 中弹/v ]] 当场/d 死亡/v …… 。/。

b3 [dj-ZW [np-DZ 张鸣岐/nP 同志/n ] 遇难/v ]以后/f ，/当地/s 老百姓/n 说/v ，/， ……

这两个句子中的短语，第一个名词都是人名（nP），属于名词的小类，可以看成是一种较粗的语义标注，但它们的组合方式还是有歧义，这说明类似较粗的语义标注达不到化解歧义的要求，内部需要更细微的语义信息：a3 中“中弹”的直接部位是“头部”，“头部中弹”的对象是“法乌斯蒂诺”，所以它们是 a 式组合；而“张鸣岐/nP 同志/n 遇难/v”中“遇难”的对象是“张鸣岐同志”，所以“张鸣岐”与“同志”先组合，然后跟“遇难”发生联系。

4) n+n+n 模式

a4 自从/p1939/m 年/qT 5/m 月/qT 14/m 日/qT **[np-DZ 博卡/nO [np-DZ 青年/n 队/n ] ]**与/p 拉努斯队/nO 比赛/v 时/n ，/，…… 。/。

b4 **[np-DZ [np-DZ 鞍钢/nO 炼铁厂/n ] 值班室/n ]** 56/m 岁/qN 的/u 老/a 工人/n 陆森/nP ，/， …… ，/，听到/v 收音机/n 里/f 传出/v 的/u 白雪洁/nP 英雄/n 事迹/n ，/，……

具体来说，上面两例中的模式是“nO+n+n”（nO 表示组织机构名，属于名词的小类）。在前面的“n+n+v”模式中，我们将名词小类（人名 nP）的设立看成一种粗粒度的带有语义信息的标注，但没有起到化解歧义组合的作用，在“n+n+n”的模式中同样证明了这一点。“博卡”与“青年队”是整体和部分的关系，“青年”与“队”之间是限定、修饰的关系，所以“青年队”要作为一个整体，然后与“博卡”组合；而“鞍钢”与“炼铁厂”是整体和部分的关系，“鞍钢炼铁厂”与“值班室”也是整体和部分的关系，所以“鞍钢/nO 炼铁厂/n 值班室/n”采取 b 式组合方式体现了由整体到部分的层层递进。由此可见，组合方式与内部的语义信息有直接关系，上下文的句法信息，甚至是名词下面的小类信息，是很难达到排除歧义组合的要求。

5) v+v+n 模式

a5 有/vJY 两/m 位/qN 医生累瘫/v 在/p 手术台/n ，/， 稍事休息/v ，/，又/d 重新/d 上场/v **[vp-PO 继续/v [vp-PO 作/v 手术/n ] ]**。/。

b5 要/vM 增强/v 责任/n 意识/n 、/、全局/n 意识/n 、/、[np-DZ [vp-LH 改革/v 开放/v ] 意识/n ] 、/、[np-DZ [vp-LH 调查/v 研究/v ] 意识/n ] ] ] ] ，/， …… 。/。

上面两个例子中的短语词性序列相同，意味着两个短语中对应的具体词属于相同的词类范畴，它们充当句法成分的能力、与其他词语的组合能力有某些共性。但在具体的实例中，“v+v+n”模式的组合情况却不相同。a5 中，“继续”的对象是“作手术”这一行为，“手术”是“作”的直接对象，所以“作手术”要先结合；而 b5 中，“改革”与“意识”、“开放”与“意识”都不是动作与对象的关系，而是一种限定关系（一种什么意识），“改革”与“开放”需要先组合成一个整体，然后限定后面的名词。所以，短语内部语义关系的差别决定了 a5、b5 组合方式的差异。

上面五种模式的分析说明了同一个问题，即相同的词性序列可能产生不同的组合方式，也就是所说的组合歧义，而这种不同的组合由短语内部不同的语义关系决定的，与外在的制约因素无关。由于篇幅限制，后面几种模式仅列出具体实例，不再进行逐一分析。

6) v+n+v 模式
a6 [vp-PO 造成/v [dj-ZW 人员/n 伤亡/v ]]
b6 [vp-LW [vp-PO 去/v 养猪场/n ] 参观/v]
7) n+v+v 模式
a7 [np-DZ 价格/n [np-LH 监督/vN 检查/vN ]]
b7 [np-DZ [np-DZ 经济/n 发展/vN ] 要求/vN]
8) v+v+v 模式
a8 [dj-ZW 供/v [vp-PO 大于/v 求/v]]
b8 [dj-ZW [vp-ZZ 对外/v 联络/v ] 中断/v]
9) v+a+n 模式
a9 [vp-PO 成为/v [np-DZ 热门/a 话题/n ]]
b9 [vp-PO [vp-SB 洗/v 干净/a ] 手/n]
10) a+n+n 模式
a10 [np-DZ 基本/a [np-DZ 伦理/n 观念/n ]]
b10 [np-DZ [np-DZ 寻常/a 百姓/n ] 家/n]

## 3 讨论

纵观汉语的这些类型的歧义，不难发现它们歧义的排除绝大多数是靠自身的意义，而不受外部结构各种元素的制约，如“成为/v 热门/a 话题/n”，“洗/v 干净/a 手/n”的歧义靠自身的意义排列就可以排除，与线性序列上的词性标记、句法信息等无关。

前面提到，当前统计的方法得到越来越多的认可，人们构建树库，在很大程度上是“可为基于统计的自动句法分析器提供必要地训练数据和统一的测评平台”，在现有树库的基础上进行机器学习，让计算机获得尽可能多的句法知识，实现更大规模的标注，提高正确率，节省人力、物力。而大量歧义组合的存在必定影响机器学习的效果，对正确的组合方式产生干扰，不利于句法正确率的提高。

构建树库不是要增加歧义，而是要消除歧义，尽量使一个句子对应一棵树（因为人的理解是没有歧义的）。目前的考察不得不令人怀疑，依靠句法信息、在词性标注的基础上构建汉语树库能否达到消歧的目的。或许，我们可以尝试采取一些新的策略和方法，在词类标记小类更加细化的基础上，对歧义短语进行集中标注，或许情况会有所改善。

## 4 结论

汉语大规模树库是一项重要的资源，它为汉语语言研究和信息处理作出重要的贡献，如何有效地利用其中的数据，需要通过对数据的考察、分析、研究。

连续出现的三个词（名、动、形）大量存在局部组合或互不相干的情况，说明位置上是“邻居”，句法组合中可以没有关联，因为决定组合方式的是词语之间的语义信息，而非线性序列上所处的位置。

通过十种模式的考察与分析发现，从词性序列的角度看，汉语中的歧义组合是普遍存在的。一般来说，不同的组合方式与词语内部的语义信息有密切的关系，而并不受外部句法信息的制约，甚至名词小类如人名（nP）、组织机构名（nO）等这种粗粒度的语义标注，某些情况下也达不到化解歧义组合的要求。这说明在分词、词性标注的基础上，对汉语进行句法标注会产生大量的歧义组合。汉语没有明显的形态变化，词类和句法功能不能一一对应，同一位置上可能出现不同功能的词类（或短语），因此在句法层面上歧义组合很难化解。

歧义组合的存在会影响计算学习的效果，这需要引起我们的思考和重视，尝试采取一些新的策略和方法，更好的解决树库中存在的歧义组合。

## 参 考 文 献


[1] 王跃龙，姬东鸿. 汉语树库综述［J］. 当代语言学，2009.

[2] 汉语句子的句法树标注规范 V2．0[R]．清华大学计算机系智能技术与系统国家重点实验室，技术资料，2002.

[3] 周强，俞士汶．汉语短语标注标记集的确定［J］．中文信息学报，1996(4).

[4] 周强，张伟，俞士汶．汉语树库的构建［J］．中文信息学报，1997(4).

[5] 周强．汉语句法树库标注体系［J］．中文信息学报，2004(4).