第四章　现代汉语短语结构歧义类型分析及分布统计

§4.1 从计算机处理的角度看现代汉语短语结构歧义

在上一章对短语结构组合规则的具体分析中，为说明一些规则约束条件的用意，我们已经举了不少计算机分析汉语短语结构时碰到的歧义例子。这样的歧义问题都是计算机分析汉语的结构必须面对的困难，有效的解决歧义问题无疑对中文信息处理有重要的理论和实际意义。要解决问题，对问题本身的性质、造成问题的原因、以及问题的难度到底有多大，事先有个清晰的认识，显然是必要的。尤其值得注意的是，从计算机处理的角度考虑歧义问题，跟从人的角度考虑歧义问题[1]，有很大不同。本章就在已有的短语结构规则基础上，以计算机处理为背景，对汉语短语结构歧义做更为全面系统地分析整理。

目前我们主要是从定界歧义和结构关系歧义两方面来看短语结构歧义[2]。所谓定界歧义，也就是短语结构的层次切分歧义。层次切分歧义通常会伴随着结构关系歧义。而所谓结构关系歧义，则是指两个成分发生组合，能以不同的关系形成一个组合体。实际中发生的短语结构歧义几乎总是同时包含这两个方面。

要发生短语结构的定界歧义，一定是发生在三个以上的成分之间。考虑最简单的情形，我们以三个符号形成的线性序列为考察对象，分析可能造成短语结构定界歧义的排列格式。假定 A、B、C 为三个任意的符号标记，“ABC”即是一个由这三个标记排列形成的格式。就组合情况来讲，这个格式存在以下六种可能性：

（1）AB+C；　如：[[缺 水] 地区]　　[[办 完] 手续]
（2）A+BC；　如：[许多 [职业 军人]]　　[打击 [走私 活动]]
（3）ABC；　如：[美女 和 野兽]　　[激动 得 流泪]
（4）AB C；　如：[监狱 看守] 和　　[香港 特区] 最
（5）A BC；　如：的 [幸福 家庭]　　了 [计算机 科学]
（6）A B C；　如：在 了 屋顶　　完 饭 看

前三种是 A、B、C 三者之间可能发生结构组合关系构成一个结构整体的情况。后三种是 A、B、C 三者之间或者只能局部构成结构体，或者干脆互不相干，总之是不能构成一个结构整体的情况。广义而言，上述六种情况都可以称为对于“ABC”这个序列的解释。换句话说，也就是 ABC 这个抽象的形式有六种可能的解释方式（意思）。而对一个具体的“ABC”序列，通常是以其中的一种或几种方式来进行解释，如果只能以一种方式进行解释，则该序列是无歧义的，如果可以有一种以上的方式来解释，则该序列是有歧义的。

需要特别说明的是，所谓一个具体的“ABC”序列，可以指上面像“缺水地区”这样的三个词排列形成的具体的短语，也可以指像“ap np np”这样的由抽象的短语功能类排列形成的格式（只不过比起“ABC”来说，“np np vp”显得更“具体”一些），还可以指像“vp np 的”这样的既包含抽象的短语功能类标记，又包含具体的汉语词形成的格式。不难看出，一个格式抽象还是具体，是相对而言的。在面对人的歧义研究中，比较关注由具体的词语符号排列造成的歧义，如“咬死猎人的狗”这样的歧义例子。而面对计算机的歧义研究，则除了注意这种具体的歧义例子外，更重视像“vp np 的 np”这样抽象的歧义格式的研究。通过对抽象的歧义格式的研究，可以对所有具体的歧义例子进行全面系统地归类整理。从计算机分析自然语言的方式来讲，也是以对抽象的规则进行操作为中介，来控制对具体的语词符号进行分析的。显然，在比具体的歧义例子更抽象的模式歧义层面分析短语结构的歧义格式，

对计算机而言，具有更重要的意义。此外还有一点需要强调，就是面对人分析歧义，往往是指出有歧义就达到目的了，因为指出歧义后，人可以通过诸多知识来判断实际使用中应该如何来准确地理解或者表达。但计算机却不一样，指出歧义只是解决实际问题的起点而不是终点。必须找到切实的可以用来排除歧义的因素，并形成一定的范畴，以一定的规则表达成形式化知识，计算机才能以它作为判断依据，来解决一个具体的例子是否有歧义的问题。

短语组合的结构关系歧义的模式很简单，即任意两个成分如果能形成结构，它们之间可以选择多种结构关系，就存在关系歧义，如果只能选择一种结构关系，就没有关系歧义。比如 vp 跟 np 发生组合，可能形成述宾、定中两种结构关系。这两个抽象的短语类之间就存在结构关系歧义。当然，一个具体的 vp 加上一个具体的 np 可能没有歧义，比如“参加了大会”只能是述宾结构。也可能真的有歧义，如“复印资料”。而 dp 跟 vp 发生组合，只能形成“状中”结构关系，这两个抽象的短语类之间就没有结构关系歧义，即任何一个具体的 dp 加上任何一个具体的 vp，永远都不会有歧义（参见上一章有关规则）。

在对计算机分析短语结构时面临的两方面歧义问题有了概括的认识后，下面我们对现代汉语短语结构的定界歧义和结构关系判定歧义做系统的考察。考察对象主要是抽象的歧义格式，同时也兼顾具体的歧义实例。对任何一个抽象的歧义格式，都举例来说明歧义的性质。我们选择了这样三个考察角度。

（I）考察歧义格式中组成成分有何特征；

（II）考察不同的结构定界方式造成的对外影响；

（III）考察抽象的格式歧义和具体的实例歧义的对应关系；

下面分别讨论从这三个角度出发区分出的不同歧义类型。

§4.2 包含终结符的歧义格式与不包含终结符的歧义格式

从歧义格式中组成成分的特征看，歧义格式可以简单地区分为两种，即包含终结符的歧义格式和不包含终结符的歧义格式。以下分别说明。

（一）包含终结符的歧义格式

看两个歧义格式的例子。（1）mp np u<的> np；（2）vp u<的> np c<和> np

这两个排列式的组成成分中都含有终结符，如“的”、“和”（本文所说的终结符指汉语中的词），同时这两个格式的结构边界都是有歧义的，即（1）、（2）都可以有两种组合方式：

1a. [mp [np u<的> np]]； 2a. [vp 的 [np 和 np]]

1b. [[mp np u<的>] np]； 2b. [[vp 的 np] 和 np]

而且，这两种组合方式在汉语中都能找到语言实例来体现。如：

1A. [一张 [电影院 的 海报]]； 2A. [捐赠 的 [时间 和 地点]

1B. [[一家 电影院 的] 经理]； 2B. [[倒塌 的 房子] 和 难民]

（二）不包含终结符的歧义格式

看两个歧义格式的例子。（3）np np np；（4）np vp np

跟（1）、（2）不同，（3）、（4）中都不含终结符。不过这两个排列格式也都是有边界歧义的。它们都至少有下面这两种组合方式：

3a. [np [np np]]； 4a. [np [vp np]]

3b. [[np np] np]； 4b. [[np vp] np]

上述不同的组合方式，也都可以在汉语中找到语言实例来体现。如：

3A. [公司 [项目 经理]]； 4A. [老师 [辅导 学生]]

3B. [[羊皮 领子] 大衣]； 4B. [[电器 修理] 教材]

包含终结符还是不含终结符，只是在考察有结构边界歧义的排列格式的组成成分特征

时，得到的一种区分结果。从这个角度考察歧义格式，也可考虑其他的区分标准，比如以排列式中包含 np 还是不含 np 来作区分。这也可以作为二级分类标准进一步把上面两类歧义格式区分出更多的小类来。本文以是否包含终结符作为首要区分标准，主要有两方面的考虑，一是认为形式上非常明显。跟汉语的“的”、“和”等特定虚词相关的结构边界歧义问题一向很突出（常有人跟英语的 pp-attachment 歧义相提并论），特别强调一下也不为过。至少可以促使对短语结构定界歧义的研究目标相对更集中一些。二是一般有“的”、“和”这样的终结符参与造成的定界歧义，通常都要针对三项以上的排列格式（比如上面例 1、2，歧义格式内部分别包含了四项和五项成分），才容易显出歧义来。而仅由非终结符参与形成的歧义，三项以内就可以清楚地显示出定界歧义问题了（见下文例子）。

§4.3 外显型歧义格式与内含型歧义格式

从一个格式不同的结构定界方式对外部环境的影响或受外部环境的制约这个角度，可以区分出外显型歧义格式和内含型歧义格式两类不同的歧义类型[3]。下面分别说明。

（一）外显型歧义格式

看一个歧义格式的例子。（5）vp np u<的> np

这个歧义格式的两种组合方式：5a.[vp [np u<的> np]； 5b.[[vp np u<的>] np]

分别都可以找到具体的实例：5A. [修 [老王 的 自行车]；

5B. [[修 自行车 的] 扳手]

对这个歧义格式，两种不同的定界方式造成的后果是有显著差异的。所谓显著差异，是指不同的定界形成的结构整体功能类不同。按 5a 切分，结构整体功能类是 vp，内部结构关系是述宾结构；按 5b 切分，结构整体功能类是 np，内部结构关系是定中结构。这一差异可以显著地在结构整体参与组合时体现出来。比如“修老王的自行车”可以作谓语（“他修老王的自行车”），可以受状语成分的修饰（“正在修老王的自行车”），不能受数量结构的修饰等等；而“修自行车的扳手”，可以受数量结构的修饰（“两把修自行车的扳手”），不能受状语成分的修饰，不能作谓语等等。

（二）内含型歧义格式

看一个歧义格式的例子。（6） ap np np

这个歧义格式的两种组合方式： 6a. [ap [np np]]; 6b. [[ap np] np]

可以分别对应实例： 6A. [大 [钢铁 公司]]； 6B. [[大 眼睛] 姑娘]

对这个歧义格式，两种不同的定界方式并不造成结构整体有显著的功能差异。在这两种定界方式下，结构整体功能类都是 np。这也就意味着，就句法条件而言，这两种不同的定界方式对应的结构整体对外差不多有相同的组合能力。比如都可以作主语（“大钢铁公司容易造成垄断”，“大眼睛姑娘最讨人喜欢”），都可以受数量成分修饰（“一家大钢铁公司”、“一个大眼睛姑娘”）等等。

不难看出，从上面的角度对歧义格式作区分，直接的意义是有助于正确地考虑排歧的策略。像（5）那样的歧义格式的实例，在一定的上下文环境中不同的结构定界会受到显著的制约。比如对经典的歧义例子“咬死了猎人的狗”，如果这个表达式出现在“那只狼咬死了猎人的狗”中，毫无疑问，它得按 5a 的方式进行组合。因为这时在它所处的位置上，需要的是 vp，而不是 np。这就排除了以 5b 的方式组合的可能性。换句话说，对这样的外显型歧义格式，我们更需要关注它的外部限制条件。这样对寻找排除歧义的规则可能会更有效一些。而对（6）那样的歧义格式，不同的结构定界通常并不显著地受外部环境的制约。或者更准确的说，是受外部环境制约的条件更不确定一些，而常常是由内在的组成成分之间的制约关系来决定整个排列式该以何种方式进行组合。比如，“大”不大能跟“钢铁”组合，但可以

跟“公司”组合。所以“大钢铁公司”得按 6a 的方式组合。因此，对这类内含型歧义格式，我们就更需要关注它的内部组成成分之间的组合限制。

当然，我们只是说在考虑这些歧义格式的消歧策略时可以有所侧重，并不是提倡偏废一方，对任何一个有边界歧义的排列式，向外考察其可能的上下文环境制约，向内探求其组成成分之间的搭配约束关系，都是不可或缺的。

此外还有一点值得一提，那就是这两种歧义情况对整句分析的影响程度也明显不同，如果是（5）那样的歧义情况，局部的歧义分析出错对整句的分析会造成很大的影响。而如果是（6）那样的歧义情况，局部的歧义分析出错对整句分析则影响较小，错误基本会局限在歧义语段内部，不大会因为局部歧义的分析错误造成对整句格局的破坏。就这点而言，做句法分析系统时，应该首先在（5）这类歧义格式多下些工夫。

从逻辑上讲，外显型歧义格式内部可以有局部内含型歧义的情况[4]。像例（5）这个歧义格式，内部没有内含型歧义的情况，可以称为**简单外显型歧义格式**。而像下文统计中提到的“vp vp vp”这个外显型歧义排列格式，内部还有内含型歧义的情况，可以称为**复杂外显型歧义格式**。

§4.4 真歧义格式、准歧义格式、伪歧义格式

从抽象的格式歧义和具体的实例歧义之间的对应关系这个角度，可以把歧义格式区分成真歧义格式、准歧义格式和伪歧义格式三种情况。以下分别说明。

（一）真歧义格式

看一个歧义格式的例子。（7） vp ap np

这个歧义格式的两种组合方式：7a. [vp [ap np]] ； 7b. [[vp ap] np]

分别对应着实例：7A. [踢 [新 足球]]； 7B. [[踢 碎] 热水瓶]

对人理解而言，上面这两个实例本身是没有歧义的。只是计算机在分析这两个实例时，要判断各自分别该按 7a 定界还是该按 7b 定界。

对这个格式，我们另外还可以找到这样的实例：7C. “踢破球”。这个实例，人理解起来也是有歧义的。即可以理解为踢的本来就是破球（按 7a 的方式组合），也可以理解为把一个球（本来可能是好的）给踢破了（按 7b 的方式组合）。

像（7）这样的格式，在抽象的句法结构层面有定界歧义，同时也很容易找到歧义实例，即一个具体的表达式有两种理解的可能性。这样的歧义格式我们称之为真歧义格式。

（二）准歧义格式

看一个歧义格式的例子。（8） pp vp vp

这个歧义格式的两种组合方式是：8a. [pp [vp vp]]； 8b. [[pp vp] vp]

对应的实例：8A. [被警察 [抓住 罚款]]； 8B. [[被政府 邀请] 参加庆典]

对人理解而言，8A 跟 8B 本身都不造成理解上的歧义。 只是计算机在碰到这样的实例时，要判断到底该按哪一种方式进行结构定界，即对计算机而言，8A 可能被分析为 8a，也可能被分析为 8b，从而造成计算机分析时的歧义问题。但是不像上面的（7）格式那样，对于（8）这个格式，在汉语中不大容易找到一个具体的实例可以有两种理解。我们把（8）这样的情况，即在抽象的句法结构层面有定界歧义，但语言中对应的具体表达式都只有一种理解的可能性，称为准歧义格式。

（三）伪歧义格式

汉语中还有这样的组合格式。（9） dp vp np

这个格式也可以有两种组合方式： 9a. [dp [vp np]]； 9b. [[dp vp] np]

不过不像上面的各种格式的歧义情况，这两种看上去不同的组合方式，并没有真正的歧

义实例来与之分别对应，从另一个角度讲，就是符合这个格式的实例，都是既可以按 9a 方式定界，也可以按 9b 方式定界，同时又基本不影响对意义的理解，即没有歧义。比如：

9X. ［认真地［学 英语]] —— 9a 或 [[认真地 学] 英语] —— 9b

9Y. ［已经 ［吃完 饭]] —— 9a 或 [[已经 吃完] 饭] —— 9b

对这样的格式，我们可以系统地考察汉语的句法结构格局后，确定按某一种方式进行结构定界就可以了。比如对（9）这样的情况，我们就可以规定它按 9a 方式组合。像这样的歧义格式，我们称之为伪歧义格式。

从抽象的格式歧义和具体的实例歧义的对应关系这个角度对歧义格式加以分类，是我们在冯志伟（1995）有关潜在歧义格式[5]的讨论基础上进一步深入分析得到的结果。这个结果有助于对消解不同格式歧义的难易程度有个大致的估计。真歧义格式在现实的语言表达中容易找到显性的歧义实例，排歧涉及的因素相对多一些，也就困难一些。可以考虑暂时把这类歧义排除出计算机能够解决的问题范围。准歧义格式在现实语言表达中只有单义实例，但具体到不同的实例，又会有不同的定界方式。对准歧义格式，刻画排歧条件相对真歧义格式要容易一些。因而也就应该列入计算机可以解决的问题范围之内。而伪歧义格式，纯粹是来干扰分析的，可以根据实际情况做出规定，而不必费心去理会。有了这些认识，我们在考虑具体的歧义格式的消歧策略时，就能更有针对性。此外，真歧义格式在实际语料中容易碰到具体的歧义实例，因而也容易引起人们的注意（特别是语言学家的注意）。而准歧义格式因为仅仅是抽象的格式歧义，不是实际语料中的具体的实例歧义，因而不大容易引起注意。因此以往汉语学界关注和研究较多的是真歧义格式，对准歧义格式的研究相对就很少。从中文信息处理的角度讲，准歧义格式的的研究尤其应该引起重视。

以上三小节，我们对汉语短语结构定界歧义从三个角度进行了类型分析。从这三个角度观察得到的不同歧义类型是有重叠的（比如“vp ap np”歧义格式就既是内含型歧义又是真歧义）。这也很正常。因为我们并不强调这三个角度要有内在的系统性和层次性，事实上这三个观察角度差不多是相对独立的。进一步说，我们的目的并不是为了给歧义格式分类而分类。而是想通过以上对歧义格式所做的类型分析，进一步考虑相应的排歧策略，因此也就并不在意是否能给出一个严密的歧义格式分类系统。而是注重歧义类型的分析结果对认识歧义的成因和性质是否有帮助。

§4.5 现代汉语短语结构歧义格式分布统计

按照上述对歧义格式类型的认识，我们在现有的 147 条汉语短语结构分析规则基础上，对汉语中可能造成短语结构定界歧义的排列式进行了初步统计[6]。

（一） 统计方式

区分含终结符的歧义格式和不含终结符的歧义格式，应该是最为直观的歧义格式划分了。我们的统计是把这两种情况分开来进行的。

先来看不含终结符的歧义格式，这只要考虑三个非终结符标记的排列情况就可以了。目前我们考虑了 9 个短语功能类标记，包括：np,tp,sp,mp,ap,dp,pp,vp,dj（没有考察含 mcp 的情况）。每条规则由上下文无关文法产生式和产生式左部根结点的内部结构信息描述两部分组成（如：vp->!vp np :: $.内部结构=述宾），不包含其他条件约束。

对于任意选取对象标记集中的三个符号排列形成的格式（如“np np np”，共有 9^3=729 个这样的排列式），仅靠一套上下文无关文法产生式规则（包含短语内部结构信息），不考虑其他任何约束条件，通过程序[7]自动分析，可以得到这些短语标记发生组合关系的各种可能情况。包括哪些三项排列可以形成合法组合，哪些不行。形成合法组合的三项排列式中，哪些是有歧义的，哪些是无歧义的。有歧义的三项排列式，再统计哪些是外显型的，哪些是内

含型的，对外显型歧义格式和内含型歧义格式，再按每个歧义的组合可能性（不妨暂命名为“歧义指数”）从大到小排序，并计算了平均歧义程度。至于一个歧义格式跟具体歧义实例的对应关系，即该歧义格式是真歧义类型、准歧义类型、还是伪歧义类型，由于要跟实际语料中的情况相印证，需要大规模树库的支持，本文暂时没有做统计。

对含终结符的歧义格式的统计，方式跟上面不含终结符的三项排列式的情况基本一样。不同之处在于，对含终结符的歧义格式，我们考察的对象是四项和五项的排列式，即在上面产生的三项终结符全排列的基础上，插入助词“的”跟连词“和”，共考虑了 8 种插入的方式，如对“np np np”三项排列式，进行插入操作后，有下面这 8 种排列形式：

4 项排列：np 的 np np　　np 和 np np　　np np 的 np　　np np 和 np

5 项排列：np 的 np 和 np　　np 和 np 的 np　　np 的 和 np np　　np np 的 和 np

这样即得到全部排列式为 729×8=5832 个。对这 5832 个排列式，我们也通过程序自动分析，得到这些排列式内部各项成分之间发生组合关系的各种可能的情况。

（二） 统计结果

（1）不含终结符的排列式（共 729 个）的情况

<table>
<tr><td colspan="3">可能形成合法结构的排列:420 个（57.6%）</td><td>不可能形成合法结构的排列:309 个（42.4%）</td></tr>
<tr><td colspan="3">np np np
np np mp
np np tp
np np sp
……</td><td>np np dp
np np pp
np mp sp
np mp dp
……</td></tr>
<tr><td colspan="2">有歧义的排列式:265 个（36.4%）</td><td>无歧义的排列式:155 个（21.3%）</td><td rowspan="3">dj mp mp
dj mp tp
dj mp sp
dj mp dp
……
pp tp sp
pp tp dp
pp tp pp</td></tr>
<tr><td>外显型歧义格式:141 个（19.3%）</td><td>内含型歧义格式:124 个(17%)</td><td rowspan="2">np mp np
np mp tp
np mp dj
np ap dj
……</td></tr>
<tr><td>np np np
np np ap
np np vp
np vp vp
……</td><td>np np mp
np np tp
np np sp
np np dj
……</td></tr>
</table>

（表 4-1：汉语短语标记三项排列的统计结果）

外显型歧义格式（共 141 个）	歧义指数	内含型歧义格式（共 124 个）	歧义指数
[1] vp vp vp	43	[1] mp vp vp	6
[2] vp vp ap	34	[2] tp vp vp	6
[3] vp ap ap	25	[3] sp vp vp	6
……		……	
[141] vp pp np	2	[124] pp pp pp	2
平均歧义数	6.24	平均歧义数	2.52

（表 4-2：外显型歧义和内含型歧义格式歧义程度排序结果）

说明：表 4-2 大致反映了三项排列式中歧义格式的歧义程度。就具体格式的绝对歧义指数来讲，外显型歧义格式中歧义最多的达到 43 种分析结果，而内含型歧义格式中最高的也不过 6 种分析结果。就平均歧义数来讲，外显型歧义是内含型歧义平均指数的两倍多。这两方面的情况说明，在多数情况下，歧义格式通过结构之间的外部环境制约而得以自动消除歧义的发生几率还是比较高的。因此，我们在考虑不同歧义类型的消歧策略时，应该自觉地注意到各自不同的性质而加以针对性的研究。

（2）含终结符的排列式（共 5832 个）的情况

<table>
<tr><td colspan="3">可能形成合法结构的排列:971 个（16.6%）</td><td>不可能形成合法结构的排列：4861 个（83.4%）</td></tr>
<tr><td colspan="3">np u<的> np np
np u<的> np mp
np u<的> np tp
np u<的> np sp
……</td><td>np u<的> np dp
np u<的> np pp
np u<的> mp dp
np u<的> mp pp
……</td></tr>
<tr><td colspan="2">有歧义的排列式:795 个（13.6%）</td><td>无歧义的排列式:176 个（3.0%）</td><td rowspan="3">sp c<和> np np
sp c<和> np mp
sp c<和> np tp
sp c<和> np ap
……
ap c<和> np u<的> dp
……
dp u<的> c<和> tp tp
……</td></tr>
<tr><td>外显型歧义格式:507 个（8.7%）</td><td>内含型歧义格式:288 个（4.9%）</td><td rowspan="2">np u<的> mp dj
pp tp u<的> tp
np c<和> np tp
tp np c<和> tp
……</td></tr>
<tr><td>np u<的> np np
sp pp u<的> ap
mp c<和> mp u<的> vp
sp c<和> sp u<的> ap
……</td><td>np u<的> np mp
vp u<的> np tp
ap tp c<和> tp
np u<的> np c<和> np
……</td></tr>
</table>

（表 4-3：含终结符的四项及五项排列式的统计结果）

外显型歧义格式（共 507 个）	歧义指数	内含型歧义格式（共 288 个）	歧义指数
[1] vp ap u<的> vp	38	[1] ap ap u<的> dj	11
[2] ap u<的> vp vp	34	[2] ap ap u<的> tp	10
[3] ap ap u<的> ap	30	[3] ap ap u<的> sp	10
[4] ap ap u<的> vp	30	[4] np ap u<的> dj	9
[5] vp ap u<的> ap	30	[5] ap u<的> ap dj	7
……		……	
[506] sp c<和> sp u<的> ap	2	[287] sp c<和> sp u<的> sp	2
[507] sp c<和> sp u<的> vp	2	[288] sp c<和> ap u<的> sp	2
平均歧义数	7.60	平均歧义数	3.01

（表 4-4：外显型歧义和内含型歧义格式歧义程度排序结果）

（3）以上是对抽象的语类符号序列进行分析得到的歧义格式结果。不含终结符的歧义格式中最高可以分析出 43 种结果，含终结符的歧义格式中，最高可以分析出 38 种结果。下面我们给出外显型歧义和内含型歧义中各自歧义指数最高的格式的分析结果示意。

vp vp vp	mp vp vp

[1](dj:主谓(vp,dj:主谓(vp,vp))) [2](vp:述宾(vp,dj:主谓(vp,vp))) [3](dj:主谓(vp,vp:述宾(vp,vp))) [4](vp:述宾(vp,vp:述宾(vp,vp))) [5](vp:述补(vp,vp:述宾(vp,vp))) …… [43](vp:联合(vp:联合(vp,vp),vp))	[1](dj:主谓(mp,dj:主谓(vp,vp))) [2](dj:主谓(mp,vp:述宾(vp,vp))) [3](dj:主谓(mp,vp:述补(vp,vp))) [4](dj:主谓(mp,vp:连谓(vp,vp))) [5](dj:主谓(mp,vp:联合(vp,vp))) [6](dj:主谓(dj:主谓(mp,vp),vp))
vp ap u<的> vp	ap ap u<的> dj
[1](dj:主谓(vp,dj:主谓(ap:的字(ap,u<的>),vp))) [2](vp:述宾(vp,dj:主谓(ap:的字(ap,u<的>),vp))) [3](dj:主谓(vp,vp:状中(ap:的字(ap,u<的>),vp))) [4](vp:述宾(vp,vp:状中(ap:的字(ap,u<的>),vp))) [5](vp:述补(vp,vp:状中(ap:的字(ap,u<的>),vp))) …… [38](np:定中(np:的字(vp:连谓(vp,ap),u<的>),vp))	[1](dj:主谓(ap,dj:主谓(ap:的字(ap,u<的>),dj))) [2](dj:联合(dj:主谓(ap,ap:的字(ap,u<的>)),dj)) [3](dj:主谓(ap:述补(ap,ap:的字(ap,u<的>)),dj)) [4](dj:主谓(ap:联合(ap,ap:的字(ap,u<的>)),dj)) [5](dj:主谓(ap,dj:主谓(np:的字(ap,u<的>),dj))) …… [11](dj:主谓(np:的字(ap:联合(ap,ap),u<的>),dj))

（表 4-5：外显型歧义和内含型歧义程度最高的歧义格式分析结果示意）

这里有必要对分析结果做些补充说明。上表中“(dj:主谓(vp,dj:主谓(vp,vp)))”这个分析结果的含义是：“vp vp vp”三项排列式可以有一种组合方式，即后面两项 vp 先形成主谓式 dj，然后再跟前面第一项 vp 形成一个更大的主谓式 dj。分析结果中标记了组合体的内部结构关系，如“主谓、述宾、连谓、定中、的字……”等等。很显然，这样的分析结果并不一定能找到实际的例子来对应，换句话说，人几乎是不会把一个三项 vp 连续排列的格式进行这样的分析的，但计算机按照一定的短语结构规则就能把“vp vp vp”分析出这么多结果来。这是计算机分析短语结构会碰到麻烦的主要症结。

§4.6 小 结

本章从歧义格式的内部组成成分特征、歧义造成的外部影响、抽象的模式歧义和具体的实例歧义的对应关系三个角度，考察了现代汉语短语结构歧义的整体情况。我们的认识是，针对不同的歧义类型，在考虑排歧策略时应有不同的侧重。但不管怎样，彻底解决短语结构分析涉及到的这些歧义问题，必须建立在对汉语各个短语类之间的组合条件有相当准确清晰的知识基础上，比如对两项短语成分组合 np+vp，假使对汉语中任意的 np 加 vp 的组合情况都能准确描述，当计算机碰到“np vp”连续排列时就一定能得到正确的分析结果。相应地，对包含这两个短语类的三项排列式，也不会被歧义难倒了。

与上述看法相关，还有一点值得一提，即这里谈到的汉语短语结构歧义现象，仅仅在观察视角和表述上是以我们所用的短语结构规则描述体系为背景的，而歧义问题本身，跟具体采用什么语法体系以及基于何种短语标记体系来描述是无关的。不管基于何种语法理论，在中文信息处理的一定阶段，都必然要面对这些汉语短语结构歧义现象。

附注：

[1] 参见朱德熙（1980）《汉语句法中的歧义现象》，载《中国语文》1980 年第 2 期；赵元任（1959）《汉语中的歧义现象》，载袁毓林编《中国现代语言学的开拓和发展》，清华大学出版社 1992 年版；吕叔湘（1984）《歧义类例》，载《中国语文》1984 年第 5 期；黄国营（1985）《现代汉语歧义短语》，载《语

言研究》1985 年第 1 期；邵敬敏（1994）《歧义分化方法探讨》，载《九十年代的语法思考》，北京语言学院出版社 1994 年版。

[2] 语言学家在面向人的歧义研究中早已归纳了不同层次上的多种歧义现象，如多义词歧义、结构成分间的语义关系歧义、跟上下文环境相关的语用歧义等等，就目前计算机处理的水平来讲，暂时只考虑结构定界和结构关系歧义两种情况，是比较适宜的。

[3] 关于外显型歧义格式和内含型歧义格式，已经有学者在研究中做过区分，并用了“他囿性”和“自囿性”或“他囿型”和“自囿型”等概念名称来称述。参见孙茂松 黄昌宁《汉语中的兼类词、同形词类组及其处理策略》，载《中文信息学报》1989 年第 4 期；罗振声 郑碧霞《汉语句型自动分析和分布统计算法与策略的研究》，载《中文信息学报》1994 年第 2 期。值得一提的是，本文在上述概念的基础上，还进一步指出了外显型歧义格式中“简单”与“复杂”的差别。

[4] 我们把外显型歧义格式进一步区分成“简单外显型歧义格式”和“复杂外显型歧义格式”，实际上内含型歧义格式相应地可以改称为“简单内含型歧义格式”，只不过没有“复杂内含型歧义格式”与之相对罢了。

[5] 参见冯志伟（1995）《论歧义结构的潜在性》,载《中文信息学报》1995 年第 4 期。

[6] 需要说明的是，规则是可以调整的，规则的条数也并不是固定的，比如我们在第一章就曾提到，目前在汉英机器翻译系统中的全局规则有近 300 条。用不同的规则集来进行统计，结果会有不同。但就本章试图说明的问题而言，这种影响是无关紧要的。

[7] 统计程序由北京大学计算机系常宝宝博士设计提供，特此致谢。