汉语短语结构定界歧义类型分析及分布统计*

詹卫东　常宝宝*　俞士汶*

北京大学中文系　北京　100871

*北京大学计算语言学研究所　北京　100871

摘要　本文对汉语短语结构的定界歧义做了全面考察，从歧义格式的组成成分，歧义对外造成的影响，模式歧义和实例歧义的对应关系三方面考察了短语结构定界歧义的不同类型，并对汉语短语结构定界歧义的不同类型进行了初步统计。希望能将计算机处理汉语时碰到的短语结构边界歧义问题进一步清晰化，供理论研究者和应用系统开发人员参考。

关键词　短语　短语定界歧义　自然语言处理

一、引　言

计算机处理汉语会在不同层次上遭遇多种歧义问题。本文考察汉语短语结构的边界判定歧义，目的是使计算机处理汉语的短语结构定界问题进一步清晰化。全文分两大部分。

第一部分分析汉语短语结构定界歧义的不同类型。本文对短语结构定界歧义的考察从抽象的句法格式入手，而不是从语言中具体的歧义实例开始。对有定界歧义的排列式，我们选择了这样三个考察角度。

（1）考察其中组成成分有何特征；

（2）考察不同的定界方式造成的对外影响；

（3）考察抽象的格式歧义和具体的实例歧义的对应关系；

本文在阐述对汉语短语结构不同类型的歧义的认识时，理论基础是短语（词组）本位语法体系[1,2,14]，同时参考了前人有关汉语歧义现象的研究[3,4,5,6,7,8]，用到的短语功能分类体系及短语功能标记可参阅文献[11,12,14,15]。

这里必须强调的是，本文涉及的汉语短语结构定界歧义现象，仅仅在观察视角和表述上是以词组本位语法体系为理论背景的，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，都必然会遭遇到本文所谈到的这些汉语短语结构定界歧义现象。

本文第二部分在对具体歧义格式的理论分析基础上，结合开发汉英机器翻译系统的实践经验，进一步对汉语中可能造成结构定界歧义的三成分排列式进行了统计分析，给出了汉语短语结构定界歧义排列格式的分布。希望这部分内容能为研究人员全面把握汉语短语结构定界

* 本文研究工作得到国家 863 项目（编号 863－306－03－06－2）基金资助。
本文于 1999 年 1 月 26 日收到

歧义格式提供参考。

以下第二到第四节是第一部分内容，第五节是第二部分，最后是结语。

二、包含终结符的歧义格式与不包含终结符的歧义格式

2.1 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[①]

这两个排列式的组成成分中都含有终结符，如“的”、“和”(本文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 u〈的〉[np c〈和〉 np]]

1b. [[mp np u〈的〉] np]； 2b. [[vp u〈的〉 np] c〈和〉 np]

而且，这两种组合方式在汉语中都能找到实例。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]

2.2 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实例。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]

2.3 说明

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选区分标准，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)。

另外需要说明的是，要解决这些格式的定界歧义问题，无论包含终结符与否，最终都是要把短语结构之间的组合关系和条件研究清楚(通常是就两项短语成分组合如 np＋vp 等进行讨论)。我们当然也认识到，这个分类角度对解决这些格式的歧义问题并不直接有特别的助益，仅仅是为了把引起短语结构定界歧义的排列式的组成情况作一个大致的区分，以促使对歧义

① mp 表示数量短语，np 表示名词短语，vp 表示动词短语，u 表示助词，c 表示连词。

格式的研究定向性更强。

三、外显型歧义格式与内含型歧义格式

3.1 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的整体功能类不同。5a 的整体功能类是 vp,内部结构关系是述宾结构;5b 的整体功能类是 np,内部结构关系是偏正。这一差异可以显著地在结构整体参与形成更大的组合时体现出来。比如"修理老王的自行车"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,一般不能作谓语等等。

3.2 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","大眼睛姑娘很漂亮"),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。

3.3 说明

关于外显型歧义格式和内含型歧义格式,也有用"他囿性"和"自囿性"或"他囿型"和"自囿型"来称述的,可分别参见文献[9,10]。

从本小节所用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会得到更多的线索。

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,就更需要关注它的内部组成成分之间的组合限制。

当然,这里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。

最后，从理论上讲，外显型歧义格式通常内部可以有局部内含型歧义的情况，反之则不然。这就必然造成外显型歧义格式比内含型歧义格式的数量多(详见下文统计结果分析)。

四、真歧义格式、准歧义格式、伪歧义格式

4.1 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些实例时，要判断到底该按 7a 定界还是该按 7b 定界，造成计算机分析的歧义问题。

对这个格式，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。这样的歧义格式称之为真歧义格式。

4.2 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]

这两种组合方式分别也都有对应的实例：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 和 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，8B 也是如此。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。

4.3 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，并没有合适的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，在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。

4.4 说明

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进行这样的分类，是我们在冯志伟(1995)有关潜在歧义格式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，歧义涉及的因素相对多一些，排歧也就困难一些。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，刻画排歧条件相对于真歧义格式要容易一些。有了这些认识，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。但从事中文信息处理工作的研究者却应该给以充分的重视。

以上三节内容是对汉语短语结构定界歧义进行类型分析的结果。从这三个角度观察得到的不同歧义类型是有重叠的。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，本文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。关键在于有关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。

五、汉语短语结构定界歧义格式初步统计

按照上述对歧义格式类型的认识，我们在一个汉英机器翻译系统[13]所用的汉语分析规则基础上，对汉语中可能造成短语结构定界歧义的排列格式进行了一次初步统计。

5.1 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① 标记，包括：np，tp，sp，mp，ap，dp，pp，vp，dj。用到的汉语分析规则共 246 条。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)。这些规则是按照词组本位语法的理论体系组织的，基本覆盖了汉语短语结构的组合情况。在分析中之所以还考虑了短语内部结构信息，是由于短语结构的定界和内部结构关系有紧密的联系。受篇幅所限，上面三节仅对短语结构定界歧义进行了说明，而对与之有紧密联系的结构关系歧义则没有讨论。详细讨论见文献[15]。

① 实际上我们的短语结构规则用到的短语功能类标记有 10 个，除文中提到的 9 个外，还有数词短语 mcp。主要是描述阿拉伯数字，如 12、三百，等等。统计时我们没有把这类短语包括进来。

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”就是一个这样的排列式,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还是准歧义类型、伪歧义类型,由于要跟实际语料使用相印证,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“np 的 np np”是在三项 np 排列式的前两个 np 之间插入“的”,“np np 和 np”是在后两个 np 之间插入“和”,“np np 的 和 np”则是在后两个 np 之间连续插入“的”跟“和”。这样得到全部排列式为 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。

下面是统计的结果。

5.2 统计结果

1. 不含终结符的排列式(共 729 个)的情况

表 1 汉语短语标记三项排列的统计结果

<table>
<tr><td colspan="3">可能形成合法结构的排列: 369 个(50.6%)</td><td>不可能形成合法结构的排列: 360 个(49.4%)</td></tr>
<tr><td colspan="3">np np np
np np mp
np np tp
np np sp
……</td><td rowspan="4">np np dp
np np pp
np mp sp
np mp dp
……
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 colspan="2">有歧义的排列式: 285 个(39.1%)</td><td>无歧义的排列式: 84 个(11.5%)</td></tr>
<tr><td>外显型歧义格式: 194 个(26.6%)</td><td>内含型歧义格式: 91 个(12.5%)</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 2 外显型歧义和内含型歧义格式歧义程度排序结果

外显型歧义格式(共 194 个)	歧义指数	内含型歧义格式(共 91 个)	歧义指数
[1] vp vp vp	43	[1] vp ap np	5
[2] vp vp ap	34	[2] dj vp vp	5
[3] vp ap ap	25	[3] np sp dj	4

……		……	
[194] pp sp vp	2	[91] pp pp p	2
平均歧义数	6.55	平均歧义数	2.37

说明:表 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 5 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

表 3 含终结符的四项及五项排列式的统计结果

<table>
<tr><td colspan="3">可能形成合法结构的排列:1010 个(17.3%)</td><td>不可能形成合法结构的排列:4822 个(82.7%)</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 rowspan="4">np u〈的〉 np dp
np u〈的〉 np pp
np u〈的〉 mp dp
np u〈的〉 mp pp
……
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 colspan="2">有歧义的排列式:795 个(13.6%)</td><td>无歧义的排列式:215 个(3.7%)</td></tr>
<tr><td>外显型歧义格式:574 个(9.8%)</td><td>内含型歧义格式:211 个(3.6%)</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 外显型歧义和内含型歧义格式歧义程度排序结果

外显型歧义格式(共 574 个)	歧义指数	内含型歧义格式(共 211 个)	歧义指数
[1] vp vp u〈的〉 ap	45	[1] np np u〈的〉 tp	5
[2] vp ap u〈的〉 ap	41	[2] np np u〈的〉 sp	5
[3] vp vp u〈的〉 vp	39	[3] mp mp u〈的〉 tp	5
[4] ap ap u〈的〉 ap	36	[4] mp ap u〈的〉 tp	5
[5] vp ap u〈的〉 vp	36	[5] mp vp u〈的〉 tp	5
……		……	
[573] tp c〈和〉 tp u〈的〉 mp	2	[220] sp c〈和〉 sp u〈的〉 np	2
[574] sp c〈和〉 sp u〈的〉 mp	2	[221] sp c〈和〉 sp u〈的〉 sp	2
平均歧义数	9.02	平均歧义数	2.62

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 45 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。

这里有必要对分析结果做些补充说明。“(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、组合定中、粘合定中、的字……”等等,有关说明可参见[1]。限于篇幅这里就不说明了。表 5 中其他分析结果都可以依此类推。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。

表 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

vp vp vp	vp ap np
[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))) …… [43] (vp: 联合(vp: 联合(vp, vp), vp))	[1] (vp: 述宾(vp, np: 组合定中(ap, np))) [2] (vp: 述宾(vp, np: 粘合定中(ap, np))) [3] (vp: 述宾(vp: 述宾(vp, ap), np)) [4] (vp: 述宾(vp: 粘合述补(vp, ap), np)) [5] (vp: 述宾(vp: 连谓(vp, ap), np))
vp vp u〈的〉 ap	np np u〈的〉 tp
[1] (dj: 主谓(vp, dj: 主谓(ap: 的字(vp, u〈的〉), ap))) [2] (vp: 述宾(vp, dj: 主谓(ap: 的字(vp, u〈的〉), ap))) [3] (dj: 主谓(vp, vp: 述补(ap: 的字(vp, u〈的〉), ap))) [4] (vp: 述宾(vp, vp: 述补(ap: 的字(vp, u〈的〉), ap))) …… [45] (ap: 联合(ap: 的字(vp: 联合(vp, vp), u〈的〉), ap))	[1] (tp: 定中(np, tp: 定中(ap: 的字(np, u〈的〉), tp))) [2] (tp: 定中(np: 组合定中(np, ap: 的字(np, u〈的〉)), tp)) [3] (tp: 定中(ap: 的字(dj: 主谓(np, np), u〈的〉), tp)) [4] (tp: 定中(ap: 的字(np: 定中(np, np), u〈的〉), tp)) [5] (tp: 定中(ap: 的字(np: 联合(np, np), u〈的〉), tp))

六、结　语

本文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定界歧义的整体情况。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,彻底解决短语结构定界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,无论这样的知识是表现为规则还是表现为统计数据。我们已经将有可能造成歧解的排列式系统地整理出来,下一步就是有针对性地来研究排歧条件。希望本文的研究对从事中文信息处理的研究人员在处理汉语短语结构的歧义问题方面有参考价值,也请专家同行批评指正。

本文研究受到北京大学中文系陆俭明教授一次讲课“关于二项排列式歧义结构全分析”的启发,谨致谢意。作者在跟中科院计算所二室刘群副研究员的讨论中获益良多,特此致谢。

参　考　文　献

[1] 朱德熙. 语法讲义. 北京: 商务印书馆, 1982
[2] 朱德熙. 语法答问. 北京: 商务印书馆, 1985
[3] 朱德熙. 汉语句法中的歧义现象. 中国语文, 1980 年第 2 期
[4] 赵元任. 汉语中的歧义现象. 中国现代语言学的开拓和发展. 北京: 清华大学出版社, 1992
[5] 吕叔湘. 歧义类例. 中国语文, 1984 年第 5 期
[6] 黄国营. 现代汉语歧义短语. 语言研究, 1985 年第 1 期

[7] 邵敬敏. 歧义分化方法探讨. 九十年代的语法思考. 北京: 北京语言学院出版社, 1994
[8] 冯志伟. 论歧义结构的潜在性. 中文信息学报, 1995, 9(4)
[9] 孙茂松, 黄昌宁. 汉语中的兼类词、同形词类组及其处理策略. 中文信息学报, 1989, 3(4)
[10] 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. 中文信息学报, 1994, 8(2)
[11] 俞士汶. 关于计算语言学的若干研究. 语言文字应用, 1993 年第 3 期
[12] 周强, 俞士汶. 汉语短语标注标记集的确定. 中文信息学报, 1996, 10(4)
[13] 刘群等. 一个汉英机器翻译系统的计算模型和语言模型. 智能计算机接口与应用进展. 北京: 电子工业出版社, 1997
[14] 詹卫东等. 现代汉语短语本位语法体系在汉英机器翻译中的应用及其问题. 同上
[15] 詹卫东. 现代汉语 vp 的结构定界和结构关系判定[硕士学位论文]. 北京大学, 1996

Analysis on Types of Phrase Boundary Ambiguity in Contemporary Chinese

Zhan Weidong Chang Baobao * Yu Shiwen *
Dept. of Chinese Peking University Beijing 100871
* Institute of Computational Linguistics Peking University Beijing 100871

Abstract This paper analyses the ambiguity of determining boundaries of Chinese phrases in automatic parsing by computer. The type of ambiguity can be classified from three different perspectives. As viewed from component of ambiguous structures, ambiguous phrases can be classified into two kinds: one including terminal symbols, the other not including terminal symbols but only non—terminal symbols. As viewed from the influence of ambiguity, ambiguous phrases can also be classified into two kinds: self—confined ambiguous phrases and non—self—confined ambiguous phrases. The influence of the former ambiguity is mainly inside the ambiguous phrases. The influence of the latter ambiguity is outside of the ambiguous phrases. As viewed from differentiated types of relation between type and token, ambiguous phrases can be classified into three kinds: the true—ambiguity, the quasi—ambiguity, and the pseudo—ambiguity. Furthermore, the distribution of these types of ambiguous phrases in Modern Chinese is also surveyed depending on the above analysis and a set of rules used for a Chinese—English Machine Translation system. The authors hope that the analysis on various types of ambiguities mentioned above conduces to solve the problem of phrase structure ambiguities in Chinese.

Keywords phrase phrase boundary ambiguity nature language processing