体词性并列结构的结构平行*

吴云芳
北京大学计算语言学研究所 北京 100871
E_mail: wuyf@pku.edu.cn

摘 要：本文对现代汉语体词性并列结构的结构平行性进行了考察，论述了存在的两种平行分布：92%的并列结构在数量定语的分布上是平行的；91%的并列结构在“的”字定语的分布上是平行的。体词性并列结构的这种平行特性可帮助自动识别并列结构的边界，但对少量的不平行，我们还找不到有效的解决方案。

关键词：并列结构　结构平行　数量定语　“的”字定语

Structure Parallelism in Chinese Nominal Coordination

Wu Yunfang
Institute of Computational Linguistics, Peking University, Beijing, 100871
E_mail: wuyf@pku.edu.cn

Abstract: This paper demonstrates two kinds of structure parallelism in Chinese nominal coordination. As to the quantity modifier, the ratio of structure parallelism is up to 92%. As to the De modifier, the ratio is up to 91%. The characteristic of structure parallelism can help automatically identify the boundaries of most of the coordinate structures, but those with imparallelism still need good resolution.

Keywords: coordination　structure parallelism　quantity modifier　De modifier

1 引言

有标记并列结构各并列成分在结构形式上呈现出某种平行态势。设 x, y1,y2 都是语言成分，Conj 表示标记形式，y1+Conj+y2 形成一个合理的并列结构，在符合下面分布时，我们说成分 x 在并列结构中的分布是平行的：

[x+y1+Conj+x+y2]　or　x+[y1+Conj+y2]

* 本文研究工作得到国家 973 项目（G1998030507-4）和 863 项目（2001AA114040）的支持。
文章写作过程中，向穗志方博士多有请教，谨表示诚挚的感谢。

人们经常利用并列结构的平行特性来帮助识别其边界，如 Okumura and Muraki(1994)利用并列成分之间的平行特性来识别英语技术报告和手册中的并列结构；孙宏林（2001）利用并列成分的对称性来识别汉语并列结构。本文对现代汉语体词性并列结构的结构平行性进行了考察，论述了存在的两种平行分布：数量定语的平行分布和“的”字定语的平行分布。

本文主要应用了定量考察的方法。考察的对象是《人民日报》1998 年 1 月 1—10 日的语料，计约 56 万字，是北京大学计算语言学研究所研制开发的，本文的例句均取自于此。作者手工标注了语料中出现的所有的词和短语层面的有标记并列结构，共计 6217 个，本文的考察主要就是基于这 6217 个并列结构展开。现代汉语体词性并列结构还可以是 n+n 形式的无标记并列结构，如“语言文字”，“爸爸妈妈”，本文不考察无标记并列结构，因此标题及行文中的“体词性并列结构”应缺省地认为是“有标记体词性并列结构”。

2 数量定语的平行分布

这儿的数量定语是一个宽泛的概念，包括：1）“数词+量词”构成的数量短语，如“两亩、39 个”等；2）《现代汉语语法信息词典详解》（俞士汶等，1998）中称为“限定数词”的数词，如“一些、若干”等；3）“指示代词+量词”构成的指量短语，如“这份、那种”等；4）量词的重叠形式，如“种种、个个”等；5）光杆的数词，如“1400 万、18%”等。我们把这种宽泛的数量定语统一记作 A_mq。体词性并列结构各并列成分在数量定语的分布上表现出平行性。请看下面一组例句：

（1） a 再加[一瓶茅台、两盘卤菜]。

b **再加[一瓶茅台、卤菜]。

c **再加[茅台、两盘卤菜]。

d 再加[茅台、卤菜]。

a 若转换成 b、c 总觉别扭，转换成 d 好像还可以，这就是并列结构的平行性要求所使然。为了更详细地了解并列结构中数量定语的分布，我们对此进行了定量考察。考虑到并列结构的自动识别，我们主要关心以下 A、B、C 三类情况。下面的描述中，“NULL”表示没有数量定语；“+”表示语言成分之间的组合关系；“…+NULL…”表示不包含数量定语的任意字符串；“…+A_mq…”表示包含数量定语的任意字符串；“A_mq+[”表示数量定语与并列结构的左边界紧密相邻；用简省的两项并列结构代表所有的并列结构。

A 并列成分都有数量定语：

1）NULL+[…+A_mq…+Conj…+A_mq+…]

例：新开通了[4 条快速干线汽车邮路和 12 条普件汽车邮路以及两条自办航空邮路]。

B 并列成分都没有数量定语，在并列结构外部共享同一个数量定语：

2）A_mq+[…+NULL…+Conj…+NULL+…]

例：为灾区人民捐赠了大量[钱款、物资和食品]。

C 有的并列成分有数量定语，有的并列成分没有：

3）NULL+[…+NULL…+Conj…+A_mq+…]

例：他家中有[妻子和两个孩子]。

4）NULL+[…+A_mq…+Conj…+NULL+…]

例：与[万名首都各界群众和劳动模范代表]一起辞旧迎新。

上述不同情况构成一个总的事件 R。语言中不会出现整个并列结构有数量定语而且并列成分又同时有数量定语的情况，即不会出现 A_mq+[…+A_mq…+Conj…+A_mq+…]、A_mq+[…+NULL…+Conj…+A_mq+…]、A_mq+[…+A_mq…+Conj…+NULL+…] 的格式序列。基于《人民日报》1998 年 1 月 1—10 日的语料，A、B、C 三大类情况的分布如表 1 所示。

表 1　数量定语在并列结构中的分布

情况	A	B	C	R
频次	98	124	20	242

A 和 B 都是数量定语分布平行的表现，C 是数量定语分布不平行的表现。数量定语分布平行的频率为：

$$(2)\quad f=\frac{98+124}{242}=92\%$$

数量定语分布不平行的频率为：

$$(3)\quad f'=\frac{20}{242}=8\%$$

上面的计算告诉我们，数量定语在并列结构中分布平行的频率是很高的。

当并列成分在并列结构外部共享同一个数量定语时，如果数量定语为定量，则并列结构的数量定语不能分布到各并列成分上，否则所表数量会成倍增加：

（4）a　就拿回 23 个[金、银、铜及优秀]奖。

a'　** 就拿回 [23 个金、23 个银、23 个铜及 23 个优秀]奖。

a 总共只有 23 个奖，而 a' 的奖项达到了 23×4=92 个，比事实上的奖项增加了 3 倍。由此可得到一个推论：当数量定语为定量时，并列结构的数量定语所表数量等于各并列项的数量之和。但这个推论只适用于表合取（conjunction）的连词，而不适用于表析取（disjunction）的连词。在所考察的语料中，我们只发现了一例表析取连词连接的并列成分共享同一个数量定语的情况，见（5）a，这时并列结构的数量定语可以分布到所有的并列成分上，但所表数量是析取的关系，如（5）b。

（5）a　至少获得 1 项 [国家级奖励或省部级科技进步奖]。

b　至少获得[1 项 国家级奖励或 1 项省部级科技进步奖]。

当数量定语为不定量时，并列结构的数量定语可以平均分布到各并列成分上，而所表数量不会有变化：

（6）a　温州的许多[商店、饭庄、娱乐场所]，

a'　温州的[许多商店、许多饭庄、许多娱乐场所]，

这是因为“许多+许多=许多”，不定量之和仍然为不定量，上文的推论依然成立。

数量定语的平行分布可以帮助自动识别并列结构的边界。跟数量定语相联系的一个歧

义格式是 A_mq+np1+Conj+np2，它可以有两种解读方式：

（7） a A_mq+[np1+Conj+np2] （√）

b [A_mq+np1+Conj+np2]

根据上文的讨论，无疑 a 应该是缺省解读。

关于并列结构的数量定语，一个曾引起人们兴趣的话题是数量定语在并列结构中的辖域（scope）问题（Carpenter, 1995）。下面这个英语句子是有歧义的：

（8） a Every kid or adult just ran.

a1 [Every kid or adult] just ran.

a2 Every [kid or adult] just ran.

作 a1 解读时，连词“or”占广域（wide scope），量词“every”占窄域（narrow scope）；作 a2 解读时正好相反，量词“every”占广域，连词“or”占窄域。a1 解读历来受到批判和挑战。尽管人们可以设想一些特殊的场景使 a1 合法，例如在 a 句后附加一句“but I don't know which”，但 a1 的存在还是受到人们的质疑。而 a2 解读历来被认为是合情合理的。本文的探讨表明，汉语的类似歧义格式在绝大多数情况下应该作 a2 解读。换言之，现代汉语中当量词和并列连词共用时，一般是量词占广域，并列连词占窄域。

3 “的”字定语的平行分布

“的”字定语和数量定语不是同一个平面上的成分。数量定语是就定语的语义性质而言的，“的”字定语是就定语的结构形式而言的，前者和后者可以有交叉：

（9） 这种剪彩形式节省了大量的[人力、物力、财力]。

“大量的”是数量定语，同时也是“的”字定语，因此，“的”字定语的平行分布和上文探讨的数量定语的平行分布有时会有交叉，不过这并不影响我们对它们各自的探讨。“的”字定语是组合式定语的一种，可简单地记作“De”。并列结构在“的”字定语的分布上表现出普遍的平行性。早在 1956 年，肃父就已经观察到了并列结构对“的”字有无的一种制约作用。例如，“自然现象”本是熟语，一般情况下不说“自然的现象”，但在（10）的并列结构中，就需要加上“的”字，这是为了“从修辞方面来斟酌语气的整齐”：

（10） 人的认识，主要地依赖于物质的生产活动，逐渐地了解自然的现象、自然的性质、自然的规律性、人和自然的关系。 （毛泽东《实践论》，转引自肃父，1956）

为了更详细地了解“的”字定语在并列结构中的分布，我们对此进行了定量考察。下面的描述中，“NULL”表示没有“的”；“…+NULL…”表示不包含“的”的任意字符串；“…+De…”表示包含“的”的任意字符串；“De+[”表示“的”与并列结构的左边界紧密相邻；用简省的两项并列结构代表所有的并列结构。

A 并列成分都有“的”字定语，不管并列结构外部是否有“的”字定语：

1）De+[…+De…+Conj…+De+…]

例：他们要把她永远定格在漳州的[永恒的历史和美好的生活]中。

2）NULL+[…+De…+Conj…+De+…]

例：不能盲目追求[资产规模的膨胀和新兴业务的拓展]。

B　并列的成分都没有“的”字定语，在并列结构外部共享同一个“的”字定语：

3）De+[…+NULL…+Conj…+NULL+…]

例：人的[境况和发展]。

C　有的并列成分有“的”字定语，有的并列成分没有：

4）De+[…+NULL…+Conj…+De+…]

例：图片展概览了本溪的[风光、名胜、古迹及本溪解放50年来的辉煌成就]。

5）De+[…+De…+Conj…+NULL+…]

例：车主租用的[公安部门招待所的停车场和部队仓库]。

6）NULL+[…+NULL…+Conj…+De+…]

例：[童志成和他的同事们] 高兴得无以言状。

7）NULL+[…+De…+Conj…+NULL+…]

例：这笔钱包括[赔偿费的利息和诉讼费用]。

以上不同情况构成一个总的事件R。基于《人民日报》1998年1月1—10日的语料，A、B、C三大类情况的分布如表2所示。

表2　“的”字定语在并列结构中的分布

情况	A	B	C	R
频次	254	946	120	1320

A和B都是“的”字定语分布平行的表现，C是“的”字定语分布不平行的表现。“的”字定语分布平行的频率为：

（11）　$$f=\frac{254+946}{1320}=91\%$$

“的”字定语分布不平行的频率为：

（12）　$$f'=\frac{120}{1320}=9\%$$

上面的计算告诉我们，“的”字定语在并列结构中分布平行的频率是很高的。

当并列成分都有“的”字定语时，有时并列成分的定语和中心语都不相似，甚至相差很远，但因为它们具有共同的“的”字结构而依然可形成并列：

（13）a　他们[思想的形式及学术生涯的起始]与四川有着千丝万缕的联系。

b　音乐会仍将以[众多著名音乐家的参与、低廉的票价和边演出边讲解的方式]面对广大学生。

a中两个并列成分的中心语“形式”和“起始”的词性分别是n和v，语义距离很大，第一个并列成分的定语是光杆名词，第二个的定语是定中结构。b中三个并列成分的中心语“参与”、“票价”、“方式”的词性分别是v、n、n，三者的语义距离很大，第一个并列成分的定语是定中结构，第二个的定语是形容词，第三个的定语是并列结构的动词短语。这告诉我们：“同为‘的’字结构”是体词性并列结构形成的一个很强的条件，这个条件可以不在乎定语和中心语是否相似。从并列结构反观“的”字，可以推知“的”不仅具有表义（语法义）的功能，而且还具有协调语气、调整结构的功能。那么，对“的”字有无的研

究（例如，张敏，1998），就不应当忽略结构对“的”字隐现的影响。

“的”字定语的平行分布可以帮助自动识别并列结构的边界。跟“的”字定语相联系的一个歧义格式是 De+np1+Conj+np2，它可以有两种解读方式：

（14） a De+[np1+Conj+np2] （√）

b [De+np1+Conj+np2]

其中 a 应该是缺省解读。冯志伟（1995）指出，“n1+的+n2+和+n3”是一个歧义格式，可以有两种解读方式：

（15） a n1+的+[n2+和+n3]

b [n1+的+n2+和+n3]

对 b 解读的两个例子是：

（16） a （衣服的袖子）和口袋

b （衣服的袖子）和拐杖

作 a 理解的几率是非常小的，除非人们绞尽脑汁去设想一个合适的语境；b 在语料中事实上是很少出现的，因为有很少的场景让人们可以把“衣服的袖子”和“拐杖”联系起来。这两个例子因此是不合适的。从中也可推知，（14）a、（15）a 应该是最通常的解读方式。

4 结束语

汉语体词性并列结构各并列成分在结构上显现出某种平行态势，除了本文所探讨的数量定语的平行分布、“的”字定语的平行分布之外，还有人名并列时称谓定语的平行分布等，我们拟另文讨论。体词性并列结构的这种结构平行特性可帮助计算机自动识别有标记并列结构的边界。不过，所谓的“平行”是频率意义上的：92%的并列结构在数量定语的分布上是平行的；91%的并列结构在“的”字定语的分布上是平行的。对于少数的不平行我们目前依然应付乏力。

参考文献

[1] Carpenter, B. : *Lectures on Type-Logical Semantics*. MIT Press, Cambridge, 1995.

[2] Johannessen, J. B. : *Coordination*.. Oxford University Press, Oxford, 1998.

[3] Okumura, A. and Muraki, K. : Symmetric pattern matching analysis for English coordinate structures. In *Proceedings of the 4th Conference on Applied Natural Language Processing*, 1994.

[4] 冯志伟：“论歧义结构的潜在性”，《中文信息学报》，1995 年第 4 期。

[5] 肃父：“名词词组中‘的’字的作用”，《中国语文》，1956 年 3 月号。

[6] 孙宏林：《现代汉语非受限文本的实语块分析》，北京大学计算机系博士论文，2001 年。

[7] 俞士汶等：《现代汉语语法信息词典详解》，清华大学出版社，1998 年。

[8] 张敏：《认知语言学与汉语名词短语》，中国社会科学出版社，1998 年。