面向中文专利文献的有标记并列结构的统计分析

石翠，周俏丽，张桂平

（沈阳航空航天大学 知识工程中心，辽宁 沈阳 110136）

摘要：本文在中文专利语料的基础上，统计分析了中文专利文献中有标记并列结构的内部特征和外部特征。内部特征主要考察了中文专利文献中有标记并列结构的并列标记、并列结构内部分析和词性分布等。外部特征主要统计了可能的边界特征词，并分析了有标记并列结构在中文专利文献中出现的外部环境。

关键词：有标记并列结构;中文专利文献;内部特征;外部特征

Analyzing the Linguistics Features of Coordination with Overt Conjunctions Based on Chinese Patent Literature

Shi Cui, Zhou Qiaoli, Zhang Guiping

(Knowledge Engineering Research Center, Shenyang Aerospace University, Shenyang, Liaoning 11013, china)

Abstract: Based on the Chinese patent corpus, this paper counts and analyzes the internal and external features of Coordination with Overt Conjunctions (COC) in the Chinese patent literature. It mainly investigates the internal features including coordination tag, internal analysis of coordination structure and the distribution of Part-Of-Speech (POS). It's mainly counted the candidate boundary markers by the external features, and analyzes the contextual information of the coordinate structures in the Chinese patent literature.

Keywords: COC; Chinese patent literature; internal features; external features

1、引言

专利文献是一种非常重要的技术资料，它有较为固定的书写格式和表达方式[1]。与普通文献相比，专利文献的文本格式比较固定，用语较为规范，除含有一些高频词和未登录词之外，还存在着大量的并列结构。

并列结构[2]（coordinate structure），也称联合结构，它由两个或更多的并列成分组成，并列结构有时也称为联合短语。并列结构里的直接成分通常称为并列语（conjunct），并列语通常用连词、顿号或空的连接形式连接。

在中文专利文献里并列结构有下列的语言结构。

A. 该通信接口 1215 BL【发送和接收】BL【电、电磁、或光】信号。

B. 加压包装可包括合适的推进剂如 BL【二氯二氟甲烷、三氯氟甲烷、二氯四氟乙烷、二氧化碳或其它合适的气体】。

C. 任务 Z100 接收 BL【表征所述高频带部分的频谱包络的一组滤波器参数和表征所述高频带部分的时间包络的一组增益因数】。

D. 进给装置可以包括 BL【用于控制材料从第二部分 6 释放的缓冲系统或任何其它适合的系统】。

E. 计算机 802 通过 BL【BL【有线和/或无线】通信网络接口或适配器 856】连接至局域网 852。

A 为连续的两个并列结构；B 为包含多个并列语且并列语由多个并列标记连接的并列结构;C 和 D 为跨度较大的并列结构,但 C 中并列结构是规则的,而 D 中并列结构是不规则的；E 为嵌套的并列结构。

有标记并列结构是指并列语由连词或标点连接的并列结构，如 A、B、C、D、E 所列并列结构；无标记并列结构是指并列语由空(如：多输入单输出)连接的并列结构。本文主要研究有标记的并列结构，而不分析无标记的并列结构。

针对有标记并列结构，有关学者进行了多方面的考察与研究。吴云芳[3]利用现有的语言资源，从句法、语义两个层面详尽地考察了并列成分之间的约束关系，并对这些约束关系进行了形式化的描述，而后基于知识描述进行了并列结构的自动识别，基于并列词语进行了相似词语的自动聚类。王东波[4,5]在对并列结构进行自动识别前，详细统计和分析了有标记联合结构的内部语言学和外部语言学特征。苗艳军[6]，分析了宾州中文树库中并列结构的内部和外部的语言学特征。马清华[7]，立足于语言系统的自组织性这一理论基础，对并列结构的句法语义进行较为系统的动态研究。Daniel M. Bikel[8]，分析了英文中并列结构的句法特征。本文基于中文专利语料，考察了中文专利文献中有标记并列结构的并列标记和词性分布等内部语言学特征，并分析了有标记并列结构在中文专利文献中出现的外部环境，统计了可能的边界特征词等外部特征。这些关于中文专利文献的有标记并列结构的考察与分析，一方面为并列结构语料库构建提供了理论基础，另一方面为中文专利文献中的并列结构的自动识别提供了语言学知识。

本实验所使用的语料是由本实验室自己标注的，经自动分词、词性标注并人工校对的语料，且用 BL【】标记标注了语料中所有有标记的并列结构，语料的具体情况如下面表 1 和表 2 所示：

表 1 语料库的基本统计数据对比表

	文体	句子数	词数	字数	平均句长（词/句子）
本文	**专利**	**6133**	**190288**	**303082**	**31.10**
清华	综合	31970	739516	1000010	23.13
	学术	5589	158780	240289	28.4

表 2 语料库的句子长度分布数据

	句子数	词数	平均长度	所占比例
简单句子	1298	18427	14.20	21.16%
复杂句子	4835	171861	35.55	**78.84%**

王东波，谢靖[4]在《基于清华汉语树库的有标记联合结构统计分析》一文中关于清华汉语树库的基本统计数据显示清华汉语树库语料的平均句长为23.13，且其统计的语料中学术类的句子较复杂平均句长为28.4，而本文所统计的中文专利文献的平均句长为31.10，显然与非专利文献相比，专利文献的句子要复杂（表1给出了专利文献与非专利文献的对比数据）。本文对中文专利文献中的句子进行了划分，20词以下（含20）的句子为简单句子，20词以上的句子为复杂句子，则复杂句子占整个语料的80.89%。

2、 中文专利文献中并列结构基本情况统计

我们对标注的 6133 句中文专利语料进行了更细致的分类，从统计的结果更能看出并列结构在中文专利文献中是不容忽视的问题，具体分析情况如表 3、表 4 所示。

在中文专利文献中，不规则的并列结构占据很大的比重，而且不规则的并列结构有可能嵌套在规则的并列结构中，对于内层不规则并列结构的识别效果影响外层规则的并列结构的识别效果，所以只采用基于规则的方法识别中文专利文献中的并列结构是不够的，要借助于统计的方法进行识别。

在中文专利文献中，并列跨度大，即在整个句子中并列结构占较大比例的句子较多，这在非专利文献中也并不常见。如：任务/n Z100/ws 接收/v BL【表征/v 所/u 述/v 高/a 频带/n 部分/n 的/u 频谱/n 包络/n 的/u 一组/m 滤波器/n 参数/n 和/c 表征/v 所/u 述/v 高/a 频带/n 部分/n 的/u 时间/n 包络/n 的/u 一组/m 增益/n 因数/n】 。/wp

我们对中文专利文献中的并列结构进行了跨度统计，按并列结构内部包含的词语个数，将语料中的并列结构进行了划分，具体情况如表 5 所示，L 表示并列结构中包含的词语个数。

表 3 嵌套并列分布情况数据

	句子数	所占比例
包含单层并列	3097	82.91%
包含一层嵌套	602	16.12%
包含多层嵌套	36	0.96%

表 4 并列结构规律数据表

	句子数	所占比例
都是规则并列	2386	63.88%
包含不规则并列	1349	36.12%

5 并列结构跨度统计表

长度	个数	所占比例
L<=5	3144	50.15%
5<L<=10	1403	22.38%
10<L<=20	982	15.66%
L>20	740	11.80%

由表 5 可以看出，中文专利文献中的跨度大的并列结构占较大比重，这将对并列结构的识别效果产生一定的影响。

专利文献中的并列结构与非专利文献中的并列结构主要有下面几点差异：（1）包含嵌套并列结构多。（2）不规则并列结构分布广泛。（3）并列结构跨度大，甚至占据整个句子。

3、 中文专利文献并列结构内部特征

对于中文专利文献中并列结构的内部特征，我们主要从并列标记、内部并列分析和词性分布三方面考察。

3.1 并列标记

中文专利文献中并列标记主要有下面三种形式：（1）并列连词：连接并列结构的连词。例如：和、或、与、或者、及、及其、并、并且等。（2）标点符号：连接并列结构的标点符号。主要有：顿号（、）、斜杠（/）、分号（；），有时逗号（，）也起并列连词的作用。（3）复合标记：主要是并列连词与标点符号的复合。如：[，或者]、[；或者]、[；以及]、[和/或]等。

下面我们对中文专利文献中比较有特点的并列标记以及规律加以叙述。

3.1.1 并列标记斜杠“/”

在专利语料里，由“/”连接的并列结构都是包含两个并列成分的并列结构，并且这两个并列成分都是最理想、最严格的并列 ，即由词性相同、结构相同、语义类相同、音节相同的并列项组成。如例句 1 所示：

例句1：扩展/v 注入区/n 126S/ws 、/wp 126D/ws 、/wp 226S/ws 、/wp 226D/ws 与/c n-FET/ws 和/c p-FET/ws 的/u 主/b **源极/n //wp 漏极/n** 层/n (/wp 将/p 在/p 随后/d 形成/v)/wp 是/v 相同/a 导电/n 类型/n 的/u 杂质层/n 。/wp

由“/”连接的并列结构里有一种情况，使我们不得不重新考虑到底该如何分词。例如：形成/v 在/p 栅极/n 叠层/n 周围/s 的/u 受/v 压力/n 的/u 衬垫/n 、/wp 加高/v 的/u BL**【源/n //wp 漏区/n】**、/wp 掩埋/v 的/u 阱区/n 和/c //wp 或/c 掩埋/v 且/c 受/v 应力/n 的/u 包含/v Si:C/ws 和/c //wp 或/c SiGe/ws 的/u BL**【源/n //wp 漏区/n】** 都/d 可以/v 与/p 本/r 发明/n 一起/d 使用/v 。/wp

这里，显然要说的是源区和漏区，也就是说，应该是“源”和“漏”并列，那么分词为：

[源 / 漏 区]似乎更合理，但由于标注的专利语料里源区、漏区作为名词性的术语大量存在，且由“/”连接的并列结构较规则，我们可以将其作为一个整体即作为：**源/漏区/n**，我们将其切分为源/n //wp 漏区/n，便于根据并列标记斜杠“/”的特征将其进行整合。

3.1.2 并列连词 “与”

“与”有两个词性，连词（c）和介词（p），只有其作为连词时，才可作为并列标记。

“与”是双目的并列标记，即“与”只连接包含两个并列语的并列结构，而不连接包含多个并列语的并列结构。如：

错误标注：注意/v 到/vb 处于/v 简化/v 目的/n ，/wp 未/d 具体/v 示出/v BL【UE/ws 与/c 控制/n 功能性/n (/wp 例如/c S-CSCF/ws)/wp 以及/c 控制/v 功能性/n 与/c HSS/HLR/ws】之间/nd 的/u 所有/b 消息/n 。/wp

正确标注：注意/v 到/vb 处于/v 简化/v 目的/n ，/wp 未/d 具体/v 示出/v BL【BL【UE/ws 与/c 控制/n 功能性/n】(/wp 例如/c S-CSCF/ws)/wp 以及/c BL【控制/v 功能性/n 与/c HSS/HLR/ws】】之间/nd 的/u 所有/b 消息/n 。/wp

3.1.3 复合并列标记“和/或”

在专利语料里，由“/”连接的还有“和”与“或”，如例句 3 所示：

例句3：在/p 一些/m 实现/v 方案/n 中/nd ，/wp BL【监测/v 系统/n 130/m **和/c //wp 或/c** 管理/v 系统/n 160/m】 可以/v 是/v 在/p 计算机/n 165/m 上/nd 运行/v 的/u 虚拟/a 计算/v 系统/n 。/wp

此处，[和/c //wp 或/c]起并列连词的作用，所以我们把它作为复合标记使用，而不把它看作“和”与“或”的并列。

3.2 并列结构内部分析

3.2.1 包含多个并列语的并列结构的并列标记分析

包含多个并列语的并列结构，并列语通常由一种或两种并列标记连接，很少由三种及以上并列标记连接。如果包含多个并列语的并列结构是由两种并列标记连接的，那么只有最后一个并列标记不同于前面的并列标记。如：

错误标注：优选/v 地/u 在/p 用于/v BL【喷墨/v 装置/n 、/wp 直写/v 工具/n **或/c** 其它/r 类似/v 装置/n **或/c** 工具/n】 的/u 喷墨/v 墨水/n 中/nd 或/c 数字/n 墨水/ng 中/nd 。/wp

正确标注：优选/v 地/u 在/p 用于/v BL【喷墨/v 装置/n 、/wp 直写/v 工具/n **或/c** 其它/r 类似/v BL【装置/n **或/c** 工具/n】】 的/u 喷墨/v 墨水/n 中/nd 或/c 数字/n 墨水/ng 中/nd 。/wp

这里还需要说明的是在由两种并列标记连接的包含多个并列语的并列结构中，“、”（顿号）不作为最后一个并列标记。如：

错误标注：它们/n 或者/c 是/v BL【硬件/n 、/wp 硬件/n 和/c 软件/n 的/u 组合/n 、/wp 软件/n】。/wp

正确标注：它们/n 或者/c 是/v BL【硬件/n 、/wp BL【硬件/n 和/c 软件/n】 的/u 组合/n 、/wp 软件/n】 。/wp

在由两种并列标记连接的包含多个并列语的并列结构中，前一个并列标记大多情况下为“、”（顿号），有时也用“或”、“或者”连接，很少用其它并列标记连接，也就是说其它并列标记在包含多个并列语的并列结构中出现时，通常都是作为最后一个并列标记，其后面连接该并列结构的最后一个并列语。

3.2.2 相差一个前缀的并列结构分析

在我们考察的中文专利文献中，有 81 个（占并列总数的 1.3%）并列结构，并列语之间只差一个前缀词，如例句 4 所示。

例句 4：声道/n 缩减/v 混音/v 信号/n 103/m 可/v 被/p 分类/v 成/v BL【包括/v 头部/n 的/u 情形/n 和/c **不/d** 包括/v 头部/n 的/u 情形/n】 。/wp

我们对 81 个并列结构的前缀词进行了统计（括号中的数字表示个数）：不/d（15）、非

/d（4）、非/h（6）、未/d（13）、从/h（1）、毫微级/b（37）、半/m（2）、非常/d（1）、被/p（2），当“非”修饰动词时其词性为副词（d），当“非”修饰名词时其词性为前缀（h）。在这 81 个并列结构中，除了一个并列结构（如：例句 5 所示）包含 3 个并列语外，其余并列结构都是包含两个并列语的并列结构。

例句 5：但是/c ，/wp 近年/nt 来/v ，/wp 已经/d 开发/v 出/v 了/u BL【透射/v -/ws 、/wp 反射/v -/ws 和/c 半/m 透射/v -/ws 】液晶/n 显示器/n ，/wp 其中/r 倾角/n 不/d 总是/d 45/m °/ws ，/wp 因此/c ，/wp 优选/v 任意/d 地/u 调节/v 拉伸/v 方向/n 至/p 每/r 种/q LCD/ws 的/u 设计/n 。/wp

在例句 5 中，并列结构的第一个和最后一个并列语相差一个前缀词，所以我们也把该并列结构列为相差一个前缀的并列结构。

3.3 有标记并列结构的词性分布

为了能分析中文专利文献中并列结构的内部特征，我们对标注的 6262 个并列结构按照并列短语核心词的词性进行了细分类。中文专利文献的内部词性分布如下面表 6 所示。

表 6 有标记并列结构内部词性分布表

词性	频次	百分比	词性	频次	百分比
名词	3159	50.39%	形容词	57	0.91%
动词	1158	18.47%	句子并列	209	3.34%
数量词	581	9.27%	不同词性之间	466	7.43%
ws(英文字符)	371	5.92%	区别词	20	0.32%
“的”字并列	68	1.08%	其他	137	2.19%
动名词	43	0.69%			

根据中文专利文献自身的语言特点，下面几种词性的并列结构有其独特的特点和规律。

3.2.1 英文字符 ws

在中文专利文献中，有些词不属于纯正意义的外文词语，而是由英文字母和数字组成的，其没有真正的含义，经常表示一些设备号等，如：转移弧/n 1**02A/ws** 和/c **102B/ws**，这里我们也将它们的词性标注为 ws。由表 6 中的数据可以看出，由 ws 组成的并列在有标记的并列结构中占有 5.95%的比重，且这些并列结构是完全对称的并列结构。如：四/m 个/q 探测器/n BL【a/ws 、/wp b/ws 、/wp c/ws 和/c d/ws 】可以/v 位于/v 透明/a 屏幕/n 10/m 的/u 各个/r 角/n 上/nd 。/wp

ws 词性的词，除了与 ws 词性的词形成并列以外，只与名词性的或数词性的词语形成并列。例如下面的例句 6、7 所示。

例句 6：例如/c ，/wp BL【URLC8/ws 和/c 底物/n 】，/wp 例如/c 含有/v D-/ws 环/n 的/u ntRNA/ws 可/v 在/p 适合/v 于/p nt-RNA/ws 二氢尿苷/n 合成/v 的/u 测定/v 条件/n 下/nd 与/p 给氢体/n 孵育/v 。/wp

例句 7：将/p 该/r 替换/v 实施/v 方式/n 的/u BL【一个/m 或/c 多/m 个/q 】特征/n 与/p 附图/n BL【2A/ws 和/c 2/m 】中/nd 表示/v 的/u 代表性/n 薄膜/n 组合/v 。/wp

上面例句 7 中与 ws 词性的词“2A”并列的数词“2”起的也是标号的作用。事实上，在中文专利文献中，ws 词性的词与数词的并列，一种情况是数词起标号的作用，一种情况是 ws 词性的词充当数词的作用。

3.2.2 数量词

在中文专利文献中，数量词并列，除了上面例句 7 中数量词之间的并列和数词与英文字符之间的并列之外还有下面几种情况：例句 8 所示的数量词之间的并列，例句 9 所示的基数词之间的并列，例句 10 所示的数词与数词短语之间的并列，和例句 11 所示数词与形容词之间的并列等几种形式。

例句 8：在/p 本/r 实施/v 方式/n 中/nd ，/wp 磁场/n nd37/ws 对准/v 出口/n 装置/n 的/u BL【12/m

点钟/q 和/c 6/m 点钟/q】 位置/n 之间/nd 。/wp

例句 9：烤炉/n 在/p BL【第一/m 和/c 第二/m】 位置/n 之间/nd 的/u 旋转/v 运动/v 根据/p 需要/v 通过/p 过程/n 控制/v 重复/v 多/m 次/q 。/wp

例句 10：语音/n 模式/n 参数/n 具有/v BL【一个/m 或/c 一个/m 以上/nd】其它/r 状态/n 以/p 指示/v 例如/c 无声/n 或/c 背景/n 噪声/n 或/c 无声/n 与/c 浊/a 语音/n 之间/nd 的/u 转变/v 的/u 模式/n 。/wp

例句 11：纤维材料/n 片段/n 的/u 激光/n 切割/v 边缘/n 包括/v BL【两/m 个/q 或/c 更/d 多/a】纤维/n 熔合/v 在/p 一起/d 的/u 多/m 个/q 组/n G/ws 。/wp

3.2.3“的”字并列

在中文专利文献中，“的”字并列是指并列语的最后一个字是“的”的并列，如：在/p 使用/v 中/nd ，/wp 移动台/n 1401/m 的/u 用户/n 对/p 麦克风/n 1411/m 讲话/v ，/wp 并且/c BL【他/r 的/u 或/c 她/r 的/u】 话音/n 随同/v 任何/r 检测/v 到/v 的/u 背景/n 噪声/n 被/p 转换/v 为/v 模拟/v 电压/n 。/wp

上面表 6 所列的 68 个“的”字并列中，有 31 个是对称的并列结构，27 个并列结构中并列语包含相同个数的“的”字，10 处并列结构中并列语包含不同个数的“的”字，但这 10 处并列结构中有 5 个并列结构的并列语的倒数第二个词是相同的词，如例句 12 所示，1 个并列结构的并列语的第一个词相同。

例句 12：动作/n 模式/n 202/m 在/p 所/u 测量/v 的/u 信号/n 200/m 上/nd 沿着/p 时间轴/n 滑动/v ，/wp 并且/c 在/p 点/m 202/m 处/n ，/wp 观察/v 到/v 存储/v 在/p BL【动作/v 模式/n 202/m 中/nd 的/u 和/c 所/u 测量/v 的/u 信号/n 峰值/n 200B/ws 中/nd 的/u】 数据/n 足够/a 一致/a ，/wp 以/p 在/p 所/u 述/v 设备/n 中/nd 将/p 所/u 测量/v 的/u 信号/n 200/d 解释/v 为/p 表示/v 人/n 的/u 行走/n 。/wp

除此之外，在我们所考察的专利文献中，还有 5 个并列结构是“的”字并列与名词性、动词性和形容词性并列语之间的并列。

4、中文专利文献并列结构外部特征

吴云芳[9]对并列结构的外部句法特征进行了详尽的分析，下面我们将对中文专利文献中并列结构的外部句法特征进行分析，寻找有助于专利文献中有标记并列结构识别的语言学特征。

4.1 中文专利文献并列结构的左、右边界词分析

4.1.1 左边界词分析

并列结构的边界词属于并列结构的外部语言学特征，这里讲的边界词是指大多出现在并列结构外部，而不出现在并列结构内部的词语。我们把经常出现在并列结构左边界外部的词称为左边界词。根据这一语言学特征，我们把考察范围限定在一个句子的范围内，且专利文献的句子较长 ，我们进一步把考察的范围限定在子句的范围内，即由逗号分隔的句子。设 w 是句子内的任一个词，left 设定为并列结构的左边及并列结构内部的范围，f(w_left)表示词 w 在 left 范围内出现的频次，left_out 设定为并列结构左边的范围，f(w_left_out)表示词 w 在 left_out 范围内出现的频次，则词 w 作为并列结构左边界词的计算公式如下：

$$p(W)=\frac{f(w\_left\_out)}{f(w\_left)} \qquad (1)$$

通过下面两个例子对我们考察的并列结构的 left 和 left_out 范围加以解释，如例句 15 中，第一个并列结构的 left 范围是：**解映射/v 指令/n 的/u 执行/n 包括/v 把/p 复数/n 操作数/n 和/p 另外/b 的/u 复数/n 操作数/n**；left_out 范围是：**解映射/v 指令/n 的/u 执行/n 包括/v 把/p**。如果在子句范围内包含两个同级的并列结构，我们把彼此的边界作为考察的边界，如下面例句 16 中，第二个并列结构的 left 范围是：**网层/n 320/m 和/c 340/m**；left_out 范围是：**网层/n**。

例句 15：在/p 一个/m 实施例/n 中/nd ，/wp 解映射/v 指令/n 的/u 执行/n 包括/v 把/p BL【复数/n 操作数/n 和/p 另外/b 的/u 复数/n 操作数/n】 相乘/v ，/wp 然后/c ，/wp 将/v 该/r 结果/n 的/u BL【实分量/n 和/c //wp 或/c 虚分量/n】 跟/p 一个/m 边界值/n 进行/v 比较/v 。/wp

例句 16：薄膜/n 300/m 还/d 包括/v 将/p BL【第一/m 和/c 第二/m】网层/n BL【320/m 和/c 340/m】连接/v 在/p 一起/d 的/u 纵向/n 密封件段/n 352/m 。/wp

由公式(1)我们可以得到，当 p 值越大，w 作为并列结构左边界词的可能性就越大，这里我们将 p 的阈值设为 0.7，也就是说当 p 大于 0.7 时，我们将 w 作为并列结构的左边界词。在中文专利文献中，可以作为有标记并列结构左边界词的词如表 7 所示。

在专利文献中，我们将 p 的阈值设计为 0.7，主要是因为边界词出现在内部的几率很大，但是通过我们的分析发现，当边界词出现在并列结构中时，并列结构的并列语都包含该边界词且在并列语中的位置相同，即边界词与其自身形成并列。由于上述原因，在非专利文献中可以作为左边界词的词，如：在（0.63），例如（0.57），通过（0.69）等等，并未出现在左边界词的词表中，如下面例句 17 所示：

例句 17：就/d BL【**在/p** 详细/a 描述/n 中/nd 或者/c **在/p** 权利要求书/n 中/nd】 使用/v 的/u 术语/n “/wp 包括/v ”/wp 而言/u

表 7 有标记并列结构左边界词表

左边界词/词性	**频率值**	**例句**
包括/v	0.89	本/r 发明/n **包括/v** BL【药学/n 或/c 治疗性/n 组合物/n】
限于/v	0.88	用做/v 液体/n 收集/v 部件/n 41/m 的/u 柱体/n 在/p 横截面/n 可以/v 是/v 任何/r 形状/n ，/wp 不/d **限于/v** BL【圆/n 或/c 多边形/n】 例如/c 三角形/n 。/wp
涉及/v	0.85	本/r 发明/n 还/d **涉及/v** BL【由/p 所述/n 方法/n 所/u 制得/v 的/u 长生/n 红/a 桔梗/n 以及/c 包含/v 所/u 述/v 长生/n 红/a 桔梗/n 的/u 功能性/n 食品/n 材料/n】 。/wp
可以/v	0.85	电/n 连接器/n **可以/v** 位于/v BL【外侧/n 部分/n 15/m 上/nd 或/c 内侧/n 部分/n 14/m 上/nd】 。/wp
如/c	0.78	访问/v **如/c** BL【因特网/n 或/c 局域网/n (/wp LAN/ws)/wp】 等/u 网络/n 时/g 使用/v 的/u 计算机/n 可/v 读/v 电子/n 数据/n 。/wp
来/v	0.78	系统/n 1100/m 包括/v 能/v 用/v **来/v** 帮助/v BL【(/wp 诸/r)/wp 客户机/n 1110/m 与/c (/wp 诸/r)/wp 服务器/n 1130/m】 之间/nd 进行/n 通信/v 的/u 通信/v 框架/n 1150/m 。/wp
根据/p	0.75	这些/r 单个/m 粘合剂/n 器件/n 的/u 复杂性/n 可/v **根据/p** BL【位置/n 和/c 感测/n 性能/n】 从/p 较小/a 的/u 基本/a 传感器/n 系统/n 变化/v 到/p 更/d 复杂/a 的/u 系统/n 。/w
是/v	0.74	所/u 述/v 半透明/b 涂层/n 也/d 可以/v 可/v 选择/v 地/u **是/v** BL【导电性/n 的/u 和/c //ws 或/c 磁性/n 的/u】

4.1.2 右边界词分析

与左边界词相同，我们把经常出现在并列结构右边界外部的词称为右边界词。设 w 是句子内的任一个词，right 设定为并列结构的右边及并列结构内部的范围，f(w_right)表示词 w 在 right 范围内出现的频次，right_out 设定为并列结构右边的范围，f(w_right_out)表示词 w 在 right_out 范围内出现的频次，则词 w 作为并列结构右边界词的计算公式如下：

$$p(W)=\frac{f(w\_right\_out)}{f(w\_right)} \qquad (2)$$

由公式(2)我们可以得到，当 p 值越大，w 作为并列结构右边界词的可能性就越大，这

里我们将 p 的阈值设为 0.7，也就是说当 p 大于 0.7 时，我们将 w 作为并列结构的右边界词。在中文专利文献中，可以作为有标记并列结构右边界词的词如表 8 所示。与左边界词相同，在非专利文献中可以作为边界词的也（0.51）、中（0.51）等也未出现在右边界词的词表中。

表 8 有标记并列结构右边界词表

左边界词/词性	频率值	例句
等等/u	0.91	可/v 与/p 本/r 发明/n 蛋白/n 融合/v 的/u 蛋白/n 的/u 例子/n 包括/v BL【GST/ws (/wp 谷胱甘肽-S-转移酶/n)/wp 、/wp 流感凝集素/n (/wp HA/ws)/wp 、/wp 免疫球蛋白/n 恒定区/n 、/wp β -半乳糖苷酶/n】 **等等/u** 。/wp
都/d	0.86	BL【tRNA/ws 和/c 还原/v 的/u tRNA/ws】 中/nd 的/u BL【任何/r 一个/m 或/c 二者/r】 **都/d** 可/v 用/p 质谱分析/n 检测/v
来/v	0.85	所/u 述/v 频谱/n 翻转/v 操作/v 可/v 通过/p 将/p 信号/n 与/p BL【函数/n ejnπ /ws 或/c 序列/n (-1)n/ws】 相乘/v **来/v** 执行/v
分别/d	0.80	不/d 穿过/v 信号/n 变换/v 单元/n 而/c 直接/d 输出/v 的/u BL【X3/ws 和/c X4/ws】 **分别/d** 被/p 表示/v 为/v 不/d 分割/v 标识符/v 0/m 。/wp
等/u	0.78	背景/n 技术/n 诸如/c BL【晶体管/n 、/wp 电容器/n】 **等/u** 的/u 集成电路/n 元件/n 在/p 尺寸/n 上/nd 已经/d 显著/d 降低/v
一起/d	0.77	通过/p 将/p BL【RFID/ws 标签/n 130/m 与/c 主部/n 122/m】 **一起/d** 包覆/v 成型/v 而/c 形成/v 周缘/n 构件/n 120/m
之一/r	0.75	该/r 表面/n 活性剂/n 材料/n 优选/v BL【十二烷基硫酸铵/n 、/wp 线性/a 烷基苯磺酸/n 、/wp 十二烷基硫酸三乙醇胺/n 或/c 离子/n 表面/n 活性剂/n】 **之一/r** 。/wp
之间/nd	0.74	较高/a 的/u 每/r 分钟/q 转速/n (/wp RPM/ws)/wp 的/u 例子/n 在/p BL【大约/d 50/m RPM/ws 和/c 大约/d 1500/m RPM/ws】 **之间/nd** 。/wp

4.2 专利文献中有标记并列结构的依存关系分布

我们在依存树库的基础上，统计分析了中文专利文献中有标记并列结构的依存关系分布。从统计分析结果可以看出，专利文献中有标记并列结构主要出现在以下几种依存关系中：动宾关系（VOB）、定中关系（ATT）、介宾关系（POB）、“的”字结构（DE）、主谓关系（SBV），它们占据了整个并列结构的 66.47%。具体分析如下：

1、动宾关系（VOB）

做宾语的成分，与核心词之间的关系标注为动宾关系，一般位于核心词的后面。并列结构做动宾关系的句子如图 1 所示(其中，由方框框起来的是并列结构；椭圆中的是它们的依存关系)：

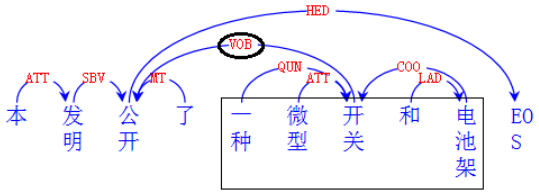

图 1 并列结构做动宾关系

2、定中关系（ATT）

定语和中心语之间的关系标注为定中关系。并列结构做定中关系的句子如图 2 所示：

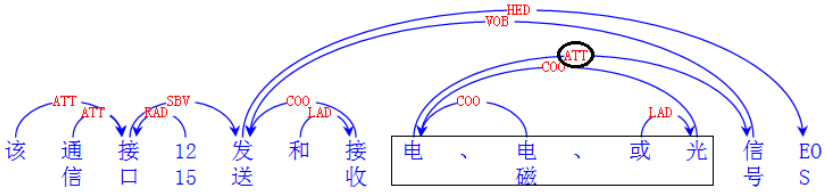

图 2 并列结构做定中关系

3、介宾关系（POB）

依存到介词的词语，则该词与依存词之间的关系标注为介宾关系。并列结构做介宾关系的句子如图 3 所示。

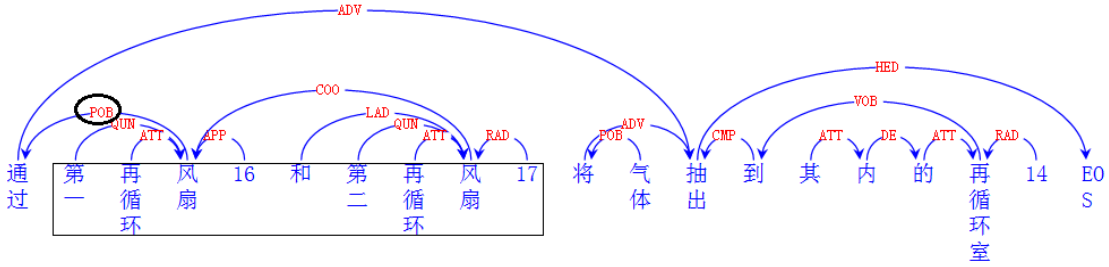

图 3 并列结构做介宾关系

4、“的”字结构（DE）

依存到“的”的词，该词与“的”之间的关系为“的”字结构。“的”字结构应该属于定语的一部分。并列结构做“的”字结构的句子如图 4 所示。

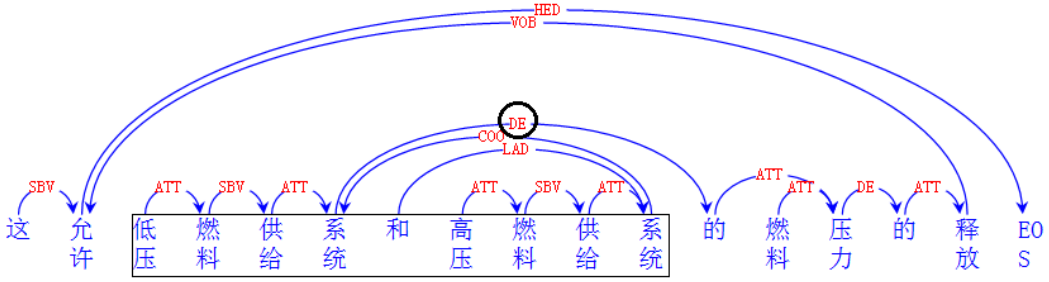

图 4 并列结构做“的”字结构

5、主谓关系（SBV）

做主语的成分，与核心词之间的关系标注为主谓关系，一般位于核心词的前面。并列结构做主谓关系的句子如图 5 所示。

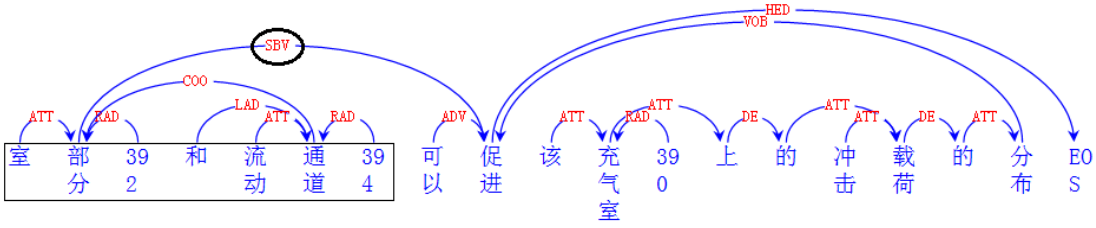

图 5 并列结构做主谓关系

4.3 专利文献中并列结构的其他外部规律

专利语料里，并列结构的左边界为介词“在”，右边界为方位名词的情况也较普遍，例如“在/p……之间/nd”、“在/p……中/nd”及“在/p……之外/nd”等。如下面例句 23 所示：

例句 23：在/p BL【所/u 附/v 权利要求书/n 及/c 其/r 等效物/n】 之外/nd 受到/v 限制/n 。/wp

在专利语料里，并列结构的左边界为介词，右边界为动词的情况也普遍存在，例如“由

/p……覆盖/v”、“从/p……去除/v”、“与/p……相关/v”等。如下面例句 24 所示：

例句 24：BL【通过/p 研磨/v 或/c 通过/p 化学/n 机械/n 抛光/v】 从/p BL【迹线/n 和/c //ws 或/c 通路/n 位置/n】 去除/v 多余/r 导电/v 材料/n 的/u 需要/n 。/wp

4、 结束语

本文通过对中文专利文献的考察，统计分析了有标记并列结构在专利文献中的内、外部语言学特征，省略了专利文献与非专利文献共有的一些语言学特征，这将为中文专利文献中有标记并列结构的自动识别提供语言学规则。但是，由于语料有限，仅依据这些规则进行有标记并列结构的识别显然是不够的。我们将扩大语料的考察范围，对中文专利文献中的有标记并列结构进行更全面的考察与分析。

参考文献

[1]任楚威.英文专利文献的汉译[J]. 湖南师范大学自然科学学报，2008，（9）:122-125.

[2]冯文贺，姬东鸿.并列结构的依存分析与连词的控制语地位[J].语言科学，2011，10(2)：168-181.

[3]吴云芳. 面向语言信息处理的现代汉语并列结构研究[D].北京:北京大学，2009.

[4]王东波.基于清华汉语树库的有标记联合结构统计分析[J].现代图书情报技术，2010，(4) : 12-17.

[5]王东波. 有标记联合结构的自动识别[D].南京：南京师范大学，2008.

[6]苗艳军. 汉语并列结构的自动识别[D].苏州：苏州大学，2009.

[7]马清华.并列结构的自组织研究[D].上海:华东师范大学，2004

[8] Daniel M. Bikel. 2005. Multilingual statistical pars-ing engine version 0.9.9c. http://www.cis.upenn.edu/～dbikel/software.html.

[9]吴云芳，并列结构的外部句法特征[C]//机器翻译研究进展-2002 年全国机器翻译研讨会论文集，2002：110-116