国家社科基金
GUOJIA SHEKE JIJIN HOUQI ZIZHU XIANGMU
后期资助项目

面向语言信息处理的现代汉语并列结构研究

Research on Chinese Coordinate Structure for Natural Language Processing

吴云芳 著

北京师范大学出版集团
BEIJING NORMAL UNIVERSITY PUBLISHING GROUP
北京师范大学出版社

国家社科基金
GUOJIA SHEKE JIJIN HOUQI ZIZHU XIANGMU
后期资助项目

面向语言信息处理的现代汉语并列结构研究

Research on Chinese Coordinate Structure for Natural Language Processing

吴云芳 著

北京师范大学出版集团
BEIJING NORMAL UNIVERSITY PUBLISHING GROUP
北京师范大学出版社

图书在版编目（CIP）数据

面向语言信息处理的现代汉语并列结构研究／吴云芳著. —北京：北京师范大学出版社，2013.7
（国家社科基金后期资助项目）
ISBN 978-7-303-16267-3

Ⅰ.①面… Ⅱ.①吴… Ⅲ.①现代汉语—并列（语法）—研究 ②现代汉语—中文信息处理—自然语言处理—研究
Ⅳ.①H146.3 ②TP391.1

中国版本图书馆CIP数据核字（2013）第094207号

营销中心电话 010-58802181 58805532
北师大出版社高等教育分社网 http://gaojiao.bnup.com
电子信箱 gaojiao@bnupg.com

出版发行：北京师范大学出版社 www.bnup.com
北京新街口外大街19号
邮政编码：100875
印　　刷：北京市易丰印刷有限责任公司
经　　销：全国新华书店
开　　本：165 mm × 238 mm
印　　张：18
字　　数：310千字
版　　次：2013年7月第1版
印　　次：2013年7月第1次印刷
定　　价：49.00元

策划编辑：杨　帆　　责任编辑：杨　帆
美术编辑：毛　佳　　装帧设计：毛　淳　毛　佳
责任校对：李　菡　　责任印制：孙文凯

国家社科基金后期资助项目

出版说明

后期资助项目是国家社科基金设立的一类重要项目，旨在鼓励广大社科研究者潜心治学，支持基础研究多出优秀成果。它是经过严格评审，从接近完成的科研成果中遴选立项的。为扩大后期资助项目的影响，更好地推动学术发展，促进成果转化，全国哲学社会科学规划办公室按照“统一设计、统一标识、统一版式、形成系列”的总体要求，组织出版国家社科基金后期资助项目成果。

全国哲学社会科学规划办公室

目 录

下　编　面向真实文本的并列结构信息处理研究

引 论

一、研究的意义

并列结构(coordinate structure)的自动识别和标注长期以来一直是语言信息处理中的难点，严重阻碍了自动句法分析器整体性能的提升。本书的研究目标是面向语言信息处理，系统地研究现代汉语中的并列结构，发掘、归纳、整理并列结构形成的语言规律，并进一步依据这些语言规律，对真实文本中的并列结构进行计算机自动识别和标注。

并列结构的自动识别和标注是用计算机程序自动识别并标注出并列结构的边界范围。如识别下面句子中的并列结构(用记号“[ ]”来标注文本中的并列结构)：

[1]输入：推动/v 两岸/n 经济/n 文化/n 交流/vn 和/c 人员/n 往来/vn ，/w 促进/v 两岸/n 直接/ad 通邮/v 、/w 通航/v 、/w 通商/v 的/u 早日/d 实现/v

输出：推动/v 两岸/n [[经济/n 文化/n]交流/vn 和/c 人员/n 往来/vn] ，/w 促进/v 两岸/n 直接/ad [通邮/v 、/w 通航/v 、/w 通商/v] 的/u 早日/d 实现/v

现代汉语并列结构在文本中的分布是非常广泛的。我们统计了《人民日报》1998 年 1 月 1 日－10 日的语料，数据结果如下表所示。

现代汉语并列结构在文本中的分布

语料规模	句子总数	含并列结构句子数	含并列结构句子所占比例
56 万字	13947	4486	32%

孙宏林(2001)考察了 20 万词的语料，其中 37%的句子中包含有并列连词。苗艳军等(2009)对宾州树库进行了考察，25.9%的句子中含有并列结构。语料中有近 1/3 的句子中包含有并列结构，可见并列结构分布之广泛。有些句子中包含两个或多个并列结构，例如下面这个并不算很长的句子中就包含 2 个有标记并列结构和 1 个无标记并列结构，整个句子

几乎就是由并列结构堆砌起来的。

[2][监督、检查][本条例及根据本条例制定的规章]的[贯彻执行]。

如果能正确识别出句子中的并列结构，那整个句子的句法结构有时就可“豁然开朗”地呈现在我们的眼前。

跟并列结构相关的一个问题是长句的分析处理。语言信息处理面向大规模真实文本时，一个必须攻克的难关就是长句的处理。随着句子长度的增加，自然语言的不确定性和歧义性成指数级增长，而并列结构是句子长度增加的一个主要方式。例如下面这两个长句就是由并列结构生成的：

[3] a　建设[一支面向21世纪的[革命化、现代化、正规化]的常备军和一支[数量充足、质量较高、动员快速、机制完善]的强大后备力量]。

b　《简明版》顺应时代需要，增设了一大批新条目。如[社会主义初级阶段、公务员制度、个人所得税、黑市、灰市、同性恋、希望工程、小康、盲流、扫黄、走穴、失业率、软科学、美容、收藏]等。

并列结构的正确识别，即是把并列结构规约为句法树上的一个节点，将使后续的处理难度大大降低。Roh(2001)对英语长句分割中，包含大量的并列结构识别处理。黄河燕、陈肇雄(2002)对复杂长句的翻译处理中，包含对并列分句的分析处理，应用了“断句拼合”的方法。

从语言理论研究的角度考虑，并列结构是多中心的扁平结构，明显不同于其他短语结构类型，具有特殊的理论研究价值。Radford(1988)指出，一个观察充分的英语语法应该能够提供原则性的回答：英语中什么样的成分跟什么样的成分可以并列以及什么样的成分跟什么样的成分不能并列。对一个观察充分的汉语语法而言，当然也应该提供原则性的回答：汉语中什么样的成分跟什么样的成分可以并列以及什么样的成分跟什么样的成分不能并列。本书将从计算的角度来系统考察现代汉语并列结构，来努力回答这一难题。

然而，并列结构自动识别和标注长期以来一直徘徊在较低的水平，

是语言信息处理中的普遍性难题。Agarwal and Boggess(1992)设计了一个英语并列结构识别算法，平均准确率为 81.6%。Okumura and Muraki (1994)利用并列成分之间的平行特性(parallelism)来识别英语技术报告和手册中的并列结构，这种模型集成到 PIVOT 英日机器翻译系统中，得到了 75% 的准确率。Kurohashi and Nagao(1994) 提出了一个日语并列结构的句法分析方法，平均识别准确率为 88%。Resnik(1999)采用基于词语相似度的方法来辨识英语名词性并列结构的结构歧义，在宾州树库上实验结果的 F 值为 70%。Chantree et al. (2005)采用分类方法来处理并列结构歧义，实验结果的 F 值为 47.4%。Kawahara and Kurohashi (2007)采用概率统计方法来识别日语的并列结构，并将其集成在一个依存句法分析器中。最近的研究中，Bergsma et al. (2011)致力于解决 n1+n2+n3 的并列结构歧义问题，可以生成两种结构[n1 + n2]+ n3(如[rocket and mortal] attacks)和 n1+[n2+n3](如 asbestos and [polyvinyl chloride])，借助于大规模的单语语料库 Google N-grams 语料和大规模的双语语料库，识别准确率从 79%提升至 96%，但这种方法需要大规模的语料，计算复杂度大，而且只是处理了 n1+n2+n3 的歧解结构。

汉语方面，周强(1996)借鉴 Kurohashi and Nagao (1994)的方法设计了汉语并列结构的识别算法，结果不是很理想，文章指出："并列结构的分析是汉语句法分析的一个难点。目前的算法在这方面的处理错误很严重。"詹卫东(2000)指出，对联合式内部组成成分的限制条件，"目前还难以做到准确描述"。孙宏林(2001)利用并列成分之间的对称性，通过一个简单的概率模型来识别并列结构边界，使实语块识别的准确率提高了 3 个百分点。据周强(2008)的研究报告，在所有类型的短语结构中，并列结构自动识别的准确率最低，只有 59.7%。赵然和晋耀红(2009)抽取出含有若干个 and 的句子进行翻译测试，基于统计的 Google 翻译系统在处理并列结构时效果很差，正确率只有 58.97%。我们利用 MST Parser 的开源软件，基于哈尔滨工业大学的汉语依存树库语料，利用其中的 8000 条数据作为训练语料开发了一个汉语依存句法分析器，对另外 1000 条数据进行了测试。下表列出了不同依存关系的正确率，其中 UAS 表示依存链路的正确率，LAS 表示依存链路和依存关系都正确的正确率，并列结构关系(COO)LAS 的正确率只有 56.34%，严重影响了汉语句法分析器的性能。

汉语依存句法分析器主要依存关系的正确率

关系类型	结点数量	UAS	LAS	错误结点数
QUN(数量关系)	618	0.9692	0.9644	19
MT(语态结构)	308	0.90584	0.9058	29
ATT(定中关系)	4755	0.8691	0.8658	622
ADV(状中关系)	2704	0.8380	0.8176	438
HED(核心)	1000	0.8	0.8	200
VOB(动宾关系)	2358	0.7960	0.7595	481
DE(的字结构)	1081	0.7955	0.7946	221
SBV(主谓关系)	1613	0.7699	0.7482	371
POB(介宾关系)	817	0.8433	0.84332	128
VV(连谓关系)	730	0.6493	0.4315	256
COO(并列关系)	701	0.6362	0.5634	255
IC(独立分句)	334	0.5269	0.2125	158

综上所述，并列结构的自动识别和标注对汉语句法分析而言是举足轻重的，现有的并列结构自动识别技术远远不能达到实用的水平。上述并列结构自动识别研究所暴露出来的共同问题是：缺乏充足的关于并列结构的语言知识，或者说是缺乏充足的直接提供给语言信息处理用的关于并列结构的语言知识。Agarwal and Boggess(1992)指出，对异词类形成的混合并列结构(mixed coordination)算法无能为力。Okumura and Muraki(1994)指出，对短语特征、词语特征和形态特征三种不同并列特征的权重是手工调试得来的，需要进一步优化。Kurohashi and Nagao(1994)指出，提高并列结构识别准确率的一个有效途径，就是设计更为精确的相似度测量方法。至于汉语方面，语言知识的不足更是制约汉语并列结构识别准确率提高的瓶颈。本书将基于大规模语料库，全面系统地研究现代汉语中并列结构形成的语言规律，为真实文本中并列结构的自动计算准备语言知识，进而基于这些语言知识实现真实文本中并列结构的自动识别和标注。

二、研究的方法

(一)研究范围

对于并列结构，汉语语法学界有不同的称说，如“并列结构”(丁声树等，1961；赵元任，1968)；“联合结构”(朱德熙，1982；吴竞存、梁伯

枢，1992)；“联合词组”(胡裕树，1987)；“联合关系”(陆俭明，2000)。不同的称说在所指的内涵和外延上并没有太大的区别，本书使用“并列结构(coordination)”这一名称。并列结构由并列成分(conjuncts)组成，并列成分有时也称之为并列项。

给并列结构一个准确、科学的定义很难。丁声树等(1961)“偏正结构的成分有偏有正，……并列结构的成分是平等的”，这是一种描述性的说明。赵元任(1968)“并列结构是一种有两个或更多中心的内中心结构，每个中心都有大致跟整个结构相同的功能”，这是从结构上来定义。朱德熙(1982)“联合结构是由两个或更多的并列成分组成的”，用“并列成分”定义“联合结构”，有循环定义之嫌。吴竞存，梁伯枢(1992)“联合结构指直接成分并列的结构”，用“并列”定义“联合结构”，有循环定义之嫌。并列结构在人们的思维中几乎是“自证(self-proved)”的，但对其内涵特征人们却还是模糊不清的。

现代汉语中，由语素到词，由词到短语，由短语到单句，由单句到复句，由复句到句群，每一个层面上由小单位组合成大单位时，都存在并列的组合形式。本书的并列结构研究限定在词和短语的层面上。下面两种并列形式不属于本书的研究范围：(1) 语素到词的并列组合，如[1]a；(2)单句到复句、复句到句群的并列组合，如[1]b、c：

[1] a 灯火　奇怪　摩擦　风雨　无依无靠　三天两头

b 成绩能够鼓励人，同时会使人骄傲。

c 成绩能够鼓励人，同时会使人骄傲；错误使人倒霉，使人着急，是个敌人，同时也是我们很好的教员。

吴振国(2007)将类似[1]a 的并列结构称为“黏合式联合结构”，一般由单音节词或语素构成，结构紧凑，中间没有连接标记，并指出黏合式联合结构的凝固性与词相同，而与一般短语不同，但其意义的非融合性又与一般词不同，而与一般短语相同。由于大部分黏合式联合结构已收入词典，因此不纳入本书的研究范围。

宾州树库(Penn Chinese Treebank)将并列结构划分成了 3 个层次：词层(word-level)，如[2]a；短语层(phrase-level)，如[2]b；子句层(clause-level)，如[2]c：

[2] a 热战、冷战、动荡、冲突和剧变

b 大型工程和小型项目

c 不但张三来了，而且李四也来了。

宾州树库对这三个不同的并列层次有自己严格的定义，本书的研究涵盖其词层和短语层，而不包括其子句层。

适应本书的研究需要，我们对词层(word-level)并列和短语层(phrase-level)并列加以简单的界定：并列成分都是单个词语(single-word)时是词层并列，或称之为词语并列；有并列成分是短语时为短语层并列，或称之为短语并列。词语并列如“凄美而沉毅”，“合作与交流”；短语并列如“[诚挚的问候和良好的祝愿]”，“[明智而有创造性]的选择”，“防止[重复建设和浪费]”。我们将词和短语的并列界定为短语层并列，这是因为在句法上词可以直接上升为短语，汉语中词和短语可以直接形成并列而没有特别的限制，如[3]所示：

[3] a 我只怕你不信，就给你带了点[花生、小枣、芝麻和你爱吃的绿豆杂面]。

b 地球表面有四个圈层，即[气圈、水圈、土壤—岩石圈以及在这三个圈交会处适宜于生物生存的生物圈]。

从句法理论上讲，并列都是短语意义上的：Kayne(1994)证明并列成分应该是一个最大的句法投射(maximal projection)；Johannessen(1998)进一步指出，即使并列项只包含中心词，也应看作是一个完整的短语(maximal phrase)。本书区分词层并列和短语层并列只是为了操作和描述上的方便，例如，“与”连接动词性成分经常在词层上形成并列，“而”经常连接形容词性成分在词层上形成并列，汉语由于缺乏形态变化而产生的异词类并列一般在词语层面上实现。

但要严格判定是不是短语并列，有时并不是一件很容易的事，如[4]a中的例子。本书把研究范围限定在由逗号隔开的一个小句范围内，这样可能会丢失一些本该研究的对象，如[4]b，不过这样限定带来的一个明显好处是，可以使我们更集中精力去研究短语并列结构的更本质的问题，而免于纠缠“这是或不是你研究的范围”。

[4] a 泽仁桑珠说，[有党的领导，有全国人民作坚强后盾，有优越的社会主义制度]，我们有信心战胜雪灾。

b 把[增产增效显著，易于掌握]的技术送给农民。

我们不研究并列式疑问结构“VP 不”“VP 不 VP”“VP 没有”“VP 没 VP”，如[5]中的那些例子：

[5] a 吃饭不？
b 吃饭不吃饭？
c 写完没有？
d 写完没写完？

这些并列式疑问结构有自己特殊的构成方式，和一般的肯定形式的并列结构有着显著差异。

汉语并列结构某种程度上是修辞的产物。普通美学上有所谓“对称美”，并列结构是展示“对称美”最好的语言形式。修辞格中的“反复”“对偶”“排比”无不跟并列结构相关。修辞对并列结构的形成可能会有一些影响，但本书却无力顾及这个因素。另外，我们不研究因为修辞而形成的“奇怪”的并列结构，因为那远远超出了我们的观察视野，同时在实际的《人民日报》语料中基本不会出现。例如，并列式标题常带有修辞的意味，诸如“未名湖与我的青春岁月”“鸡腿与情诗”“篱笆·女人·狗”，这些并列成分间的语义距离很大，作者正是借助于这种语义张力来达到一种特殊的修辞效果。尹世超(2006)指出，标题中非典型的异性并列短语已经超越了句法层面，是篇章层面的语言现象。

我们以《人民日报》语料作为研究的底本，这样就不可避免地会带上一些新闻语体的色彩，这种语体色彩在并列结构的句法语义约束上也会有一些反映。但我们相信这些并列结构还是能够大体上反映现代汉语并列结构的整体风貌。

(二)研究方法

从方法论的角度讲，语言信息处理中的研究可分为基于统计的方法和基于规则的方法，两种方法经常为孰优孰劣展开激烈的争论。

毋庸置疑的是，语料库无论是在语言本体研究还是在语言信息处理研究中都扮演着越来越重要的角色，无论是在统计方法还是在规则方法中都发挥着越来越重要的作用。人的语言知识从哪里来？个人的语言直觉(intuition)和历史记忆在某些方面比不上庞大的、真实存在的语料库。语料库对语言信息处理研究的意义在于：它是知识的源泉。统计的方法

是基于语料库，用统计来发现、用概率数字来表现语言规律；规则的方法是基于语料库，通过语言学者的观察从中总结和归纳语言规律。两种方法都离不开语料库的支撑。

本书总体上采用规则方法，采用基于语料库的定量考察方法，并辅之以定性的语言理论分析和描述。我们的语言规律和语法规则是通过观察语料，受语料中语言事实的启发而得来的，因此它首先是基于语料库的；对观察来的语言事实，语言学者综合自己的语言知识对其进行甄别和补充，因此它又是高于语料库的。我们充分利用了基于语料库的定量研究方法，同时力求对统计数字背后的语言本质给予可能的解释。

本书在并列结构的自动识别研究中也采用了机器学习方法，但在机器学习方法中充分利用和融合了语言知识。

面向语言信息处理，我们着重研究的是现代汉语并列结构形成的句法语义约束条件。前贤著作中(如储泽祥等，2002；马清华，2005)多是从语言理论的角度来审视并列结构，这与本书的研究旨意多有不同。不同角度的并列结构研究可以互相借鉴，互相补充，共同形成一个完整的现代汉语并列结构描述体系。

(三)研究资源

本书研究用到了四部分语言资源。

其一是《人民日报》1998 年 1 月份的语料。这是由北京大学计算语言学研究所研制开发的，经过了词语切分和词性标注，词性标记有 39 个(参见附录 1)。关于语料库开发的详细情况，请参见段慧明等(2000)。选择《人民日报》作为研究的语料，这一方面是适应“走向非受限大规模真实文本”这一语言信息处理发展的大趋势(黄昌宁，2002；1993)；另一方面，考虑到并列结构更多地出现在书面文体中，《人民日报》是书面文体的典范，行文流畅，语言规范，可以是很好的试验场地(testbed)；第三个原因是这个语料已在北京大学计算语言学研究所的网站上公布，是免费的开放资源，很多研究者基于此语料进行了相关研究，研究者同人尽可以在这个语料上来验证、评测本书的研究结果。

本书的例句绝大部分取自于《人民日报》1998 年 1 月份的语料，少数例句是作者自造的或摘自别处。作者手工标注了《人民日报》1998 年 1 月份 1～10 日语料中的所有词和短语层面的有标记形式的并列结构(样例请参见附录 2)，语料共计约 56 万字，标注出来的并列结构有 6217 个，这 56 万字、6217 个并列结构即是本书主要的考察对象。当数据不充分时，将适当把考察范围扩充至 1998 年 1 月份整个月的语料。这 10 天的并列

结构标注语料已经在北京大学计算语言学研究所的网站上公布，欢迎下载研究使用。

其二是《现代汉语语法信息词典》，下文简称《语法词典》。关于该词典的详细情况，请参见俞士汶等(2003)。《语法词典》是机器可读词典，它结合从语料库来的知识和语言学家的内省知识，对现代汉语不同词类(共18个基本词类)的词语进行了详尽的语法描述。《语法词典》的著作者敏锐地觉察到了语言知识库在中文信息处理中的基石作用，感觉到了词汇信息在句法构建和理解中的驱动作用。俞士汶等(2003)指出："词典中为每个词项所附加的信息同语法规则相结合，可以实现由词项驱动规则"，这与本书的主张是一致的。

《人民日报》语料建设和《语法词典》编制都是以朱德熙先生所倡导的"词组本位语法体系"为理论支撑(朱德熙，1982)，这也是本书对现代汉语并列结构进行研究的理论基点。

其三是中文概念词典(Chinese Concept Dictionary，CCD)。CCD是北京大学计算语言学研究所开发的、与WordNet兼容的汉语语义词典，关于其详细情况，请参见于江生、俞士汶(2001)。WordNet是由普林斯顿大学研制开发的、在计算语言学界具有广泛影响的英语概念词典(请参见Fellbaum，1999)，其词典构建的理念和框架被词汇语义学界和计算词典学界所公认。CCD的构建思想主要体现在三个方面：(1)传达的是概念关系，概念的承载者是词语，词义在概念中体现。(2)用同义词词集(Synset)表示概念。同义词词集用可替换性原则来确定，当两个词语在某个语境中可以相互替换而不改变语义时，它们属于同一个同义词词集。例如，{手段，方法}可以在下面的语境中替换："要采用合适的手段｜方法来解决这个问题"，它们因此属于同一同义词词集。(3)上下位关系是概念之间的主要关系，即是主要的语义关系。CCD(沿袭WordNet)设定了25个名词初始概念，15个动词初始概念(参见附录3)。

CCD主要为我们提供语义知识。研究中还参照了北京大学的现代汉语语义词典(下文简称《语义词典》)(王惠等，2003)，以及董振东先生的知网(HowNet)。

其四是国际语言资源联盟LDC的Chinese Gigaword语料。在基于并列结构进行同义词集的自动获取研究中，为了获得更多并列结构实例，选用了更大规模的Chinese Gigaword语料。选取其中的新华社语料加以研究，该语料库收录了1990—2004年15年的新华社全部文本，共约471110千字。对语料进行了前期处理工作：(1)从Unicode到GB编码的

转换；(2)利用中科院计算所的分词软件 ICTCLAS 对全部文本进行了自动词语切分和词性标注。

《人民日报》语料是我们考察的内容和对象，《语法词典》和 CCD 是我们描述现代汉语并列结构语言规律的知识体系，其中《语法词典》提供句法知识，CCD 提供语义知识，它们共同在“基于约束的文法”中发挥作用。

三、全书的组织

引论。交代选题的依据和背景，着重从语言信息处理的角度探讨了现代汉语并列结构研究的意义和价值；交代研究范围、研究方法和所用资源。

上编：基于语料库的并列结构语言研究。

第一章：并列结构研究的理论基础。说明本书并列结构研究的有关理论主张——扁平结构、句子并列；阐明本书对并列结构进行描述的理论体系——基于约束的文法。从语言信息处理的角度考虑，本书对有标记和无标记并列结构分别对待处理。

第二章：无标记并列结构研究。深入细致地考察了 n＋n、v＋v、a＋a这三大类歧义格式形成并列结构的句法语义约束条件。对 n＋n，主要是从语义角度探求其形成并列结构的条件；对 v＋v，主要是从句法角度探求其形成并列结构的条件，这过程中采取了逐步淘汰、逐渐逼近的方法；对 a＋a，主要采取了“反证”的方法。

第三章：有标记并列结构研究。又分为三个部分。一是同词类形成的有标记并列结构，从中心语相似、结构平行两个方面深入细致地考察了体词性并列结构、动词性并列结构、形容词性并列结构的句法语义约束条件；二是异词类形成的有标记并列结构，探讨了由于汉语缺乏形态变化而造成的异词类并列现象，并提出了可能的解决方案；三是有标记并列结构的外部句法特征，通过简单的统计方法找出了并列结构的前边界特征词和后边界特征词。

第四章：并列标记研究。对“与”“及”“并”“而”“或”几个标记形式进行了个案考察，指出了它们的“个性”之处及其对并列结构句法语义约束的影响；还从并列标记系统的角度，探讨了不同标记形式之间的句法语义差异，指明了它们各自在系统中的存在价值。

下编：面向真实文本的并列结构信息处理研究。

第五章：规则方法的并列结构自动识别。综合利用本书所发掘出的