基于组块分析的汉语块依存语法

钱青青
北京语言大学/ 北京海淀学院路15号
qianqingqing1996l@foxmail.com

王诚文
北京语言大学/ 北京海淀学院路15号
chengwen_wang15@126.com

摘要

基于词单元的经典依存语法在面向中文的句子分析中遇到诸多汉语特性引起的困难。为此，本文提出汉语的块依存语法，以组块为研究对象，以谓词为核心，在句内和句间寻找谓词所支配的组块，构建句群级别的句法分析框架。这一操作不仅仅是提升叶子节点的语言单位，而且还针对汉语语义特点进行了分析方式和分析规则上的创新，能够较好地解决微观层次的逻辑结构知识，并为中观论元知识和宏观篇章知识打好铺垫。本文主要介绍了块依存语法理念、表示、分析方法及特点，并简要介绍了块依存树库的构建情况。截至目前为止，树库规模为187万字符（超过4万复句、10万小句），其中包含67%新闻文本和32%百科文本。

关键词： 块依存语法；依存语法；组块；谓词

Chinese Chunk-Based Dependency Grammar

Qian Qingqing
BLCU / 15th Xueyuan Road, Beijing
qianqingqing1996l@foxmail.com

Wang Chengwen
BLCU / 15th Xueyuan Road, Beijing
chengwen_wang15@126.com

Abstract

Classical dependency grammar based on word unit encounters many difficulties in Chinese oriented sentence analysis. Therefore, this paper proposes a Chinese Chunk-Based Dependency Grammar, which takes predicates as the core and chunks as the research object. It seeks the chunks controlled by predicates within and between sentences, and constructs a syntactic analysis framework at the level of sentence group. This operation not only improves the language unit of leaf nodes, but also innovates the analysis methods and rules according to the semantic characteristics of Chinese. It can solve the logical structure knowledge at the micro level and lay a good foundation for the meso argument knowledge and macro textual knowledge. This paper mainly introduces the concept, representation, analysis method and characteristics of Chinese Chunk-Based Dependency Grammar, and briefly introduces the construction of Tree-Bank. Up to now, the size of the tree database is 1.87 million characters (over 40,000 complex sentences and over 100,000 single sentences), including 67% news texts and 32% encyclopedia texts.

Keywords: Chunk-Based Dependency Grammar , Dependency Grammar , chunk , predicate

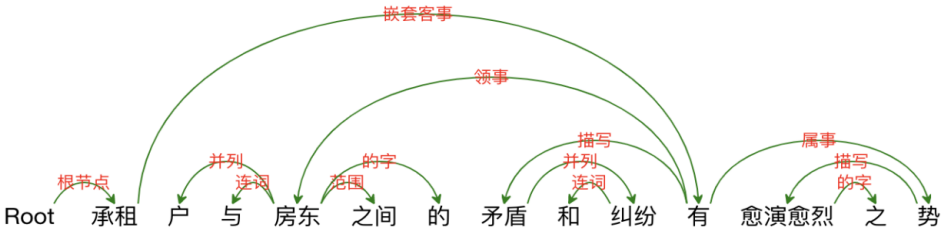
第十九届中国计算语言学大会论文集，第118页-第128页，海南，中国，2020年10月30日至11月1日。

1 引言

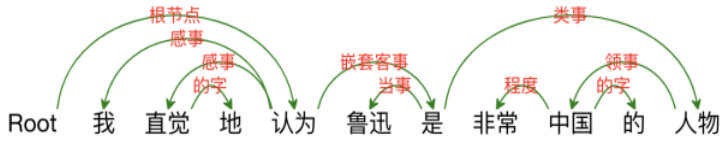
句法分析是自然语言处理领域中重要的基础研究问题之一，依据句法结构的差异性，可分为短语结构和依存结构。其中依存句法以能够适应汉语灵活语序特征且将句子分析为更加扁平的结构以降低分析、标注、储存难度的优势，近年来获得了更为广泛的应用，在问答系统、知识图谱、信息抽取等任务上发挥着重要作用。

在句法分析中，明确分析的单元是最基础、最根本的要求。传统依存句法分析大多以词作为最小单元，但分词及词性标注可能带来错误级联；汉语实际语篇中，词的词性、词义较为灵活，存在大量的活用、增加语境义的现象，传统依存句法分析较难适应该特性；汉语具有意合特征，同样的语义内容可由语序不同单元表达，过于关注“词-词”关系，使句子依存结构更为繁琐；词与词之间的关系复杂、多变，依存关系类划分的太细，降低了标注的可操作性，带来数据稀疏问题，也会影响到分析器的适应面和鲁棒性。

1)承租户与房东之间的矛盾和纠纷有愈演愈烈之势。



2)我直觉地认为鲁迅是非常中国的人物。



在例1中，主语相对复杂，此处就将主语内部词“承租户”切分开，把“承租”当成了全句的核心，从而也导致了整句依存结构的错误，依存分析时容易陷入复杂“词-词”关系分析的困境而产生错误；在语序方面，若交换“承租户”“房东”或“矛盾”和“纠纷”的语序，甚至将整个主语倒装，变为“矛盾和纠纷，在承租户与房东之间的”，句子的语义都不会产生较大的变化，但分析结构却会因此改变，这是不必要的。而例2中“中国”意为“具有中国品质的”，但此处在分析时仍然将“中国”和“人物”定义为“领事”关系，认为“中国”是一个实体，这是由于无法识别其中活用的信息而导致的。

除了基于词的依存句法分析本身存在的问题，汉语的特殊性也为句法分析带来了困难。

中文多小句、流水句，经过分析，我们发现汉语中至少有25%的小句存在成分缺失的现象[1]。而当前的中文树库中大多利用逗号、句号等标点划分分析边界，容易导致分析单位缺少成分、信息丢失，当流水句中后续小句的主语缺失时，还可能产生歧义：空主语既可能跟先行小句的主语(A)同指照应，又可能跟先行小句的宾语(B)等其他成分同指照应。修饰词（如否定词等）的辖域问题也会导致歧义的产生。

3)她不像她母亲,认为做家务的男人都是没有出息的。

4)他有票，我没有。

5)１９９１年，女足世界杯首次举行，有12支队伍参赛。

在例3中，句子呈现为两个小句，“她不像她母亲”和“认为做家务的男人都是没有出息的”。这个句子形成的图结构是分离的，后一小句的主语既可能是前一小句的主语“她”，也可能是前一小句的宾语“他母亲”，显然主语的不同会导致语义的差别，若割裂地看这个句子，会产生歧义。除了主语缺失之外，例4、5分别为宾语缺失、修饰语缺失。主宾语缺失的问题，已有学者从“篇章回指”“指代消解”等角度进行分析，如陈平（1984）、徐赳赳（1992）等，但仅限于实体之间的指代关系，忽视了提供大量情态信息的修饰语的缺失问题。宋柔（2017）关注到了除

[1]具体分析请见《汉语块依存语法与树库构建

实体之外缺省补全的重要性，他将汉语的句子界定为自足的广义话题结构，把小句界定为基于广义话题结构的话题自足句，利用流水模型生成这两类汉语篇章结构单位，为自然语言处理篇章分析单位提出了新的角度，从汉语篇章微观话题结构的角度为流水句提供了佐证和启示。但汉语中标点句并非只缺省句首的话题成分，句中或句尾的状语、宾语、补语等的缺省也值得关注；按照广义话题结构所生成的句子仅仅提示其话题-说明结构，与句子更深层次的句法语义分析之间缺少衔接，大多还是停留在拆分复杂结构，生成“能说”的自足句层面。

6)他把衣服抖了抖，然后穿上。

7)没有人民民主专政，就不可能保卫和建设社会主义。

话头理论的目的是寻找缺省的话头并生成话头自足句，但生成的话头自足句可能由于句法不通、语义不明等导致“不成句”。如例6中的第二个小句，生成自足句应当为“他把衣服然后穿上”，这是由于话头结构是线性分析的，强调“话头”和“说明”的语序，遇到语言中一些比较灵活的现象时，就会产生不成句的问题；此外，“话头-说明”的关系情况多样，可能是句法上的主谓关系，也可能是语义上的衔接关系，就使得在标注时存在两可情况，也可能与篇章级别的分析产生混杂，如例7中的“话头-说明”关系一般认为是复句中的条件关系。

指向不明确也会使句子分析不准确，下面这几个例子结构相似，但句子中名词性短语、修饰短语受哪些动词的支配却不尽相同。

8)老师让小张来办公室一趟。

9)我们洗衣服挺累。

10)我劝他手术好几天了。

针对以上的问题，我们提出汉语的块依存语法，以组块为研究对象，以谓词为核心，在句内和句间寻找谓词所支配的组块。分析时，利用汉语中的组块和组块间的依存关系，将成分缺失和指向不明的问题转化为小句内组块依存问题和小句间的组块缺省问题。补全缺失的成分，为后续任务提供准确的分析单元，消除由于指向不明确而导致的歧义。

2 组块及其类别

由于汉语句法的特殊性，“块”具有很好的现实意义。“块（Chunk）”概念最早由Abney（1991）提出，他认为句法分析可以分为三个阶段来进行，以达到简化句法分析任务的目的。即：对块进行识别、分析块的内部结构、分析块之间的关系。本阶段的工作，主要为第三步。

我们将组块定义为：由连续词语或语素整合而成的序列，表现为同一句子层级中充当句法成分的各个连续单元，例如下面这个句子被分为4个组块。

11)这句话| 只| 是| 一个例子。

组块按照其功能，可分为句法结构层面和非句法结构层面两部分（图1），其中句法结构层面的组块指在句子内部与谓词存在向心关系的组块，按照与谓词的关系，又可层层下分，而非句法结构层面的组块指通常在篇章层面用于衔接或表示语气作用的组块。

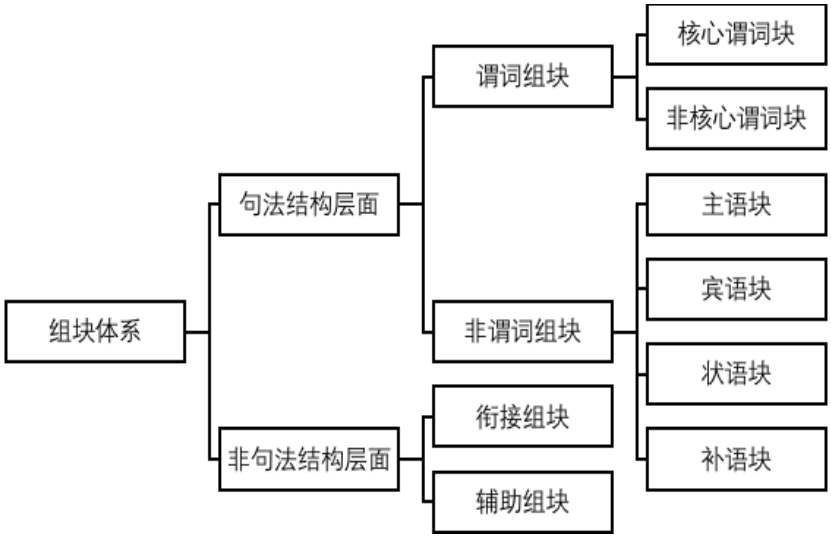


图 1: 组块体系

(1)谓词组块 谓词组块即由核心述语构成的组块，能够支配句中的非谓词块，是所在句子层级的核心，由最内部的小括号“()”表示。谓词组块主要由动词性、形容词性的词或短语[2]来充当，在一些特殊句中也会有空谓词组块的存在。句子中最顶层的谓词组块（即整个句子的核心）是核心谓词组块，出现在修饰语[3]、谓词性主宾语中的谓词组块为非核心谓词组块。

12)他(狼吞虎咽地(**吃完了**))饭。

13)这个人**()**黄头发。

14)我(现在(**承认**))｛你((**做**)得比我好)｝。

以上划线部分均为核心谓词组块，其中例13由补充的空述语充当。例14的核心谓词组块“承认”是整个句子的核心，而非核心谓词组块“做”是宾语“你做得比我好”中的核心。

(2)非谓词组块

非谓词组块指在结构上依存于谓词组块的组块，主要有主语块、宾语块、状语块、补语块几类。

a)主语块

主语块即结构中的主语，包括主谓谓语句中的大小主语。按照其内部是否还嵌套有谓词组块可将其分为体词性主语块和谓词性主语块。主语块在结构上依存于谓词组块。以下几例中的黑体部分为主语块:

15)**他**((说话)很快)。

16)**电脑**｛**我**(可(是))门外汉｝。

17)｛**(很(丰富))(却不(精细))**｝(也不(是))我们说的优秀。

b)宾语块

宾语块即结构中的宾语，按照其内部是否还嵌套有谓词组块可将其分为体词性宾语块和谓词性宾语块。宾语块在结构上依存于谓词组块，谓词性宾语用“｛ ｝”表示，双宾之间用“||”隔开。以下几例中的黑体部分为宾语块:

18)[在他壮年时，]他(爬上过)**珠峰**。

19)我(现在(承认))｛**你((做)得比我好**)｝。

20)(感谢)你(告诉)**我||这个好消息**。

c)状语块

状语块指述语中位于谓词组块前部与其紧邻和被其他成分或标点隔离的组块，对核心语块起到修饰作用，受谓词组块支配。以下几例中的黑体部分为状语块:

21)(**一年内**(新增))培育科技型企业||3465家。

22)[**别把孩子的教育，**](全(寄))希望[于教育机构上]。

[2]一般由V+着了过、V+单音节补语、两个连续的单音节V组成，字典中收录成语、常用俗语等也作为谓词组块。

[3]出现在修饰语中的非核心谓词组块将在下一步工作中进行处理

d)补语块

补语块指在句中充当补语的组块，一般位于谓词组块后部，可与谓词组块紧邻或被其他成分或标点隔离，对谓词组块起到修饰作用，受谓词组块支配。以下几例中的黑体部分为补语块：

23)她(哭着)((跑)**出来**)。

24)[别把孩子的教育，](全(寄))希望[**于教育机构上**]。

(3)衔接组块

衔接语块由连词、话语标记、插入语等组成，在句中主要发挥衔接功能，属于篇章成分。用尖括号“<>”表示，以下黑体部分为衔接语块：

25)她(非常不想(去))，<**因为**>(今天(下))雨。

(4)辅助组块

辅助组块由辅助语构成，句法上与句中其它各个成分之间没有结构上的关系，在句中主要承载表达语气的功能，用“<<>>”表示。以下各例中黑体部分为辅助语块。

26)他(走了)<<**吗**>>?

27)<<**嗯**>>，<<**好的**>>，我(知道了)。

3 块依存语法

3.1 块依存语法的表示

块依存语法主要分析非篇章成分的组块，即基于句法结构层面的6类组块，通过分析对象的选择，可将构建自足小句的过程与篇章关系的界限划分清楚。衔接组块用于表示句间的衔接关系，辅助组块则承载了表达语气的功能，均不应与句内的成分混杂。在分析句子内部成分时，我们认为核心谓词组块是句子的核心，各类非谓词块均受核心谓词组块的支配并依存于核心谓词组块之上，若某非谓词块和谓词组块之间存在依存关系，则称该非谓词块为谓词组块的从属成分，谓词组块为该非谓词块的依存对象。除了一些特殊的独词句，一般认为句子中都存在一个或多个核心，非谓词块至少依存于一个谓词组块之上。谓词组块作为句内各语块的依存对象，其左右上下各有四个点位，分别表示其主语位（1号位）、修饰语位（2号位）、宾语位（3号位）、述语位（4号位），各非谓词块按照其类别分别依存于谓词组块的四个节点上，依存线条从谓词组块的四个节点指向其从属成分。

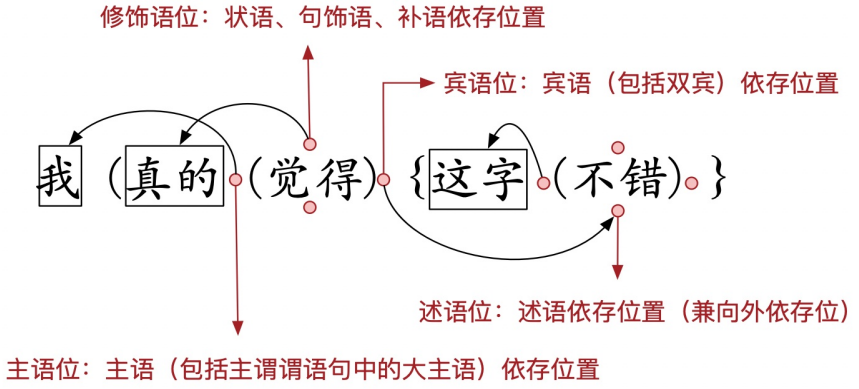


图 2: 块依存标注图示

主语，包括主谓谓语句中的大小主语依存于谓词组块的1号位；

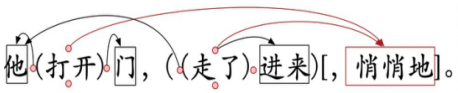
状语、补语依存于谓词组块的2号位；

宾语，包括双宾语中的远近宾语依存于谓词组块的3号位；

述语省略时从4号位置与相关述语连接，当某谓词组块依存于其他谓词组块时从4号位向外依存。

不同于Robinson（1970）所提出的四条依存分析方法的公理，块依存语法分析中，允许非谓词块、非核心谓词组块有一个或多个依存对象，允许谓词组块有多个从属成分，且允许线条交叉、跨句。中文中存在较多的非投影结构（闻媛，2018），允许线条交叉、组块多依存对象，能够使分析结果更清晰、准确。

28)他(打开)门，((走了)进来)[，悄悄地]。



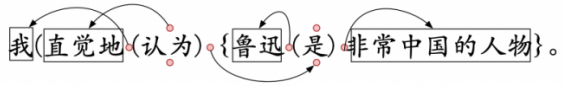
此例中，前一小句缺少了修饰成分“悄悄地”，后小句缺少了主语“他”，必然导致分析不完整。在块依存语法中，允许线条跨句、交叉，找到两个小句中核心谓词的所有从属成分，即可将两个小句补充完整。

3.2 块依存语法的分析方法

在下述两例中，“承租户与房东之间的矛盾和纠纷”“非常中国的人物”均为一个组块，语义具有相对的稳定性，更符合语言的认知规律。以组块为研究对象，能够减少分词碎片，降低活用、语境义等带来的分析错误；同时，避免纠结于“词-词”之间的关系，使得依存关系得到了精简，更关注于句子的整体结构，进一步降低存储和分析的复杂性，加强鲁棒性。在此基础上进行分析，能够在保证浅层结构正确的情况下为更深层次的分析打下基础。

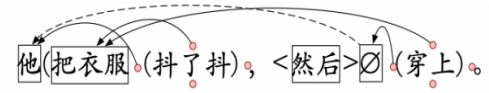
29)承租户与房东之间的矛盾和纠纷有愈演愈烈之势。

30)我直觉地认为鲁迅是非常中国的人物。



其次，通过跨句找回依存块，能够补全句子成分。组块缺省指在线性的结构标注中由于承前蒙后省略或小句分割等情况导致核心谓词组块在该小句内缺省了从属成分，在这样的情况下需要将句子在上下文中进行分析并在其四个节点处补全缺省的从属成分。

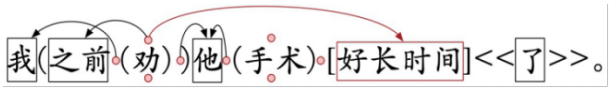
31)他(把衣服(抖了抖)，<然后>(穿上)。



在这个例子中，小句“穿上”缺省主语块和状语块，我们在这里以缺省的主语块为例，将其用“”表示，因此依存于“穿上”的主语块是“”，而“”是前一小句“他”的省略。因而为了寻回缺省的组块，使后一个小句成分完整，我们认为前一个小句的主语块“他”除了依存于所属小句的核心谓词组块“抖了抖，也依存于后一个小句的核心谓词组块“穿上”。在补全了缺省的组块之后，我们还可以将前后两个小句拆分为：“他(把衣服(抖了抖)”和“他(把衣服(穿上))”，这样，二者在这一个简单的上下文中，就没有缺省的从属成分了。篇章层面的组块“然后”并没有依存的对象，也就不进入自足句构建的过程，仅用于表示两个小句之间的顺承关系。以上的补全过程，是在排除了篇章层面的组块之后、以结构为指导的、句法层面的补全，能够与下阶段分析句间关系相衔接，且更具有理据性——能够成为另一个小句的一部分是因为它受到其中动词的支配。

针对依存对象不明确的问题，则通过寻找谓词的依存块，更好地明确句意。我们看以下这个例子：

32)我(之前(劝))他(手术)[好长时间]<<了>>。

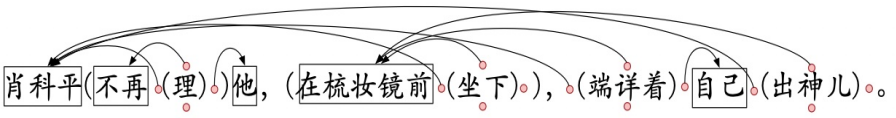


对于这样的句子，一般的处理原则是“默认左归”或者“默认右归”，采取“左归”方法时，认为“他”是“劝”的宾语，但和“手术”之间没有关系，“好长时间”是“手术”的修饰语。如果按照这样分析，这个句子的意思可能就变成了：我之前劝他，我手术好长时间了。但显然，这句话并非这个意思。因此我们判断其依存对象，认为“他”既是劝的从属对象，也是“手术”的从属对象，而“好长时间”则是“劝”的从属对象。这样，能够对这一类句子达到更好的分析效果。对兼语句、连谓句等特殊句式，也能做到很好的区分和分析。

按照缺省的组块类型，我们将组块缺省分为非谓词块缺省和谓词组块缺省。以下各举几例。**（1）主语块缺省**

主语块缺省即句子或小句中的谓词成分因省略或标点等原因缺少从属的主语块。事实上，有相当一部分的主语块缺省是由于语音上的停顿、语篇成分的插入造成的，在书面上表现为标点、衔接语、辅助语等。当忽略这些成分时，我们可以发现这类小句可与前后带有主语块的小句形成复谓或并列结构，从而找回主语块。

33)肖科平(不再(理))他，(在梳妆镜前(坐下))，(端详着)自己(出神儿)。



此句中，“坐下”“ 端详着”“ 出神儿”缺省了主语，“端详着”“出神儿”还缺省了状语，找回后，我们可以将其补充为完整的三个小句：

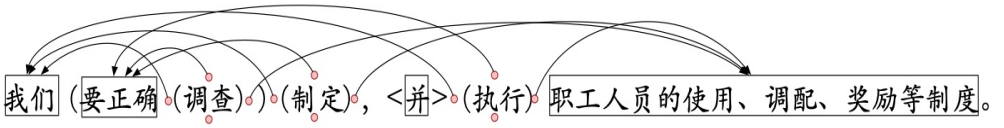
34)肖科平(不再(理))他，

35)肖科平(在梳妆镜前(坐下))，

36)肖科平(在梳妆镜前(端详着))自己(出神儿)。

（2）宾语块缺省

37)我们(要正确(调查))(制定)，<并>(执行)职工人员的使用、调配、奖励等制度。



宾语块缺省即句子或小句中的谓词成分因省略或标点等原因缺少从属的宾语块。在这个例子中，两个小句都缺省了一些成分，其中前一小句中的两个核心谓词缺省了宾语块，后一个小句的核心谓词组块“执行”缺省了主语、状语。进行分析后，我们可将两个小句补全为：

38)我们(要正确(调查))(制定)职工人员的使用、调配、奖励等制度，

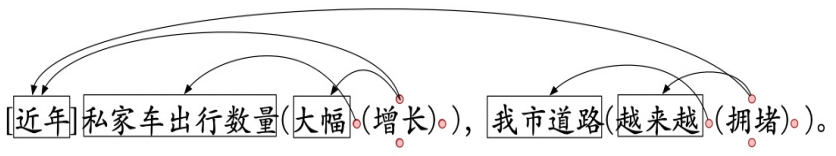
39)我们(要正确(执行))职工人员的使用、调配、奖励等制度。

此句中的“并”属于衔接组块，用于提示篇章中上下文的衔接关系，是我们下一步工作所需要关注的对象。

（3）状语块缺省

状语块中承载了大量的时地信息、情态信息，然而位于句首的状语在分句的时候，易随第一个小句进行切分，而第二个小句就因此缺少了这个状语。如下例中，我们可以将“近年”重新依存至“拥堵”，将后一小句的时间信息补充完整。

40)[近年]私家车出行数量(大幅(增长))，我市道路(越来越(拥堵))。



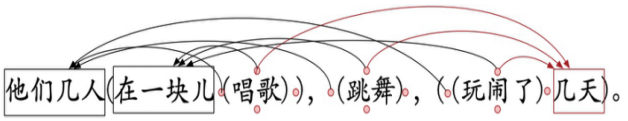
拆分后的完整小句为：

41)[近年]私家车出行数量(大幅(增长))，

42)[近年]我市道路(越来越(拥堵))。

（4）补语块缺省

43)他们几人(在一块儿(唱歌))，(跳舞)，((玩闹了)几天)。

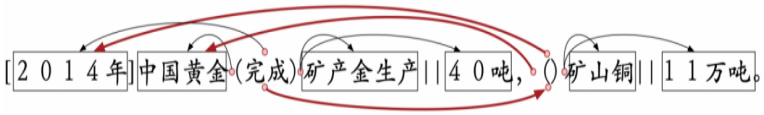


状语块缺省即句子或小句中的谓词成分因省略或标点等原因缺少从属的状语块。在上例中，补全“几天”作为“唱歌”“跳舞”的补语之后，为其增加了时间信息，句意更完整了。

（5）谓词组块缺省

谓词组块缺省是我们认为的一类特殊缺省情况。指由于省略前文中已出现过相同的核心谓词组块而造成的缺省。在这样的情况下，需要将缺省的核心谓词组块依存到原有核心谓词组块上。通过这种方法，我们可以补全原本缺省的谓词，使得句意更加清晰。

45)[2014年]中国黄金(完成)矿产金生产||40吨，()矿山铜||11万吨。



经过分析之后，生成的完整小句为：

46)[2014年]中国黄金(完成)矿产金生产||40吨，

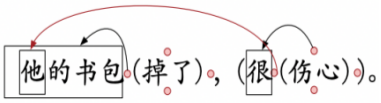
47)[2014年]中国黄金(完成)矿山铜||11万吨。

3.3 组块分割与小块依存

一般进行块依存分析时，非谓词块以整体的形式充当谓词组块的从属成分，但在某些特殊情况下，存在小块依存的现象。小块依存指在一个组块内部划分更小组块，进行依存关系构建。在小块依存中，谓词组块的从属成分并非是一个完整的组块，而是某个组块的一部分。小块依存现象在体词性的主宾语组块以及状语、补语组块中较为多见。

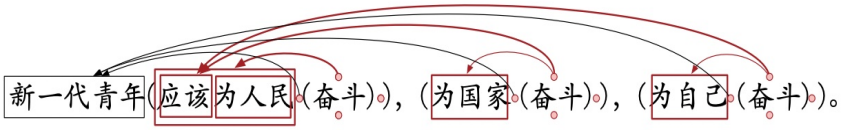
体词性主宾语块的小块依存多出现在定语和中心语之间存在从属或整体部分关系的情况下。下例中第二个小句通过块依存方法可找回主语并补全。

48)他的书包(掉了)，(很(伤心))。



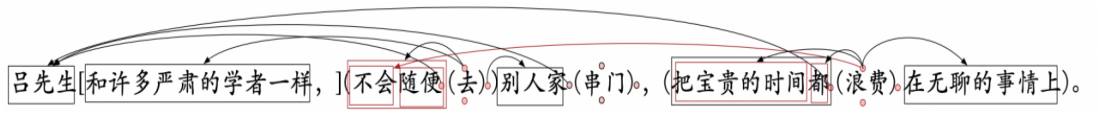
有部分状语或补语组块内部不同的部分从属于不同的一个谓词组块，此时也存在小块依存的现象。如下例中，若不分割组块，则“应该为人民”仅修饰第一个“奋斗”，将状语分割并重新分析其依存关系之后，能够更明确三个核心谓词“奋斗”的状语，在补全主语之后，即可形成3个完整的小句。

49) 新一代青年(应该为人民(奋斗))，(为国家(奋斗))，(为自己(奋斗))。



50)吕先生[和许多严肃的学者一样，](不会随便(去))别人家(串门)，(把宝贵的时间都(浪费)在无聊的事情上)。

否定性词语对于确定文本中的事件到底发生与否和是非评价有决定性的影响，尤其是否定词的辖域到底管到哪儿也决定了信息抽取的准确性。上例中，若无小块分割及跨小句的依存，则后一小句的语义与正确语义截然相反。而正确的语义应为：



51)吕先生和许多严肃的学者一样，不会随便去别人家串门，
52)吕先生和许多严肃的学者一样，不会把宝贵的时间都浪费在无聊的事情上。

4 块依存树库构建

目前，我们正在展开基于块依存语法的树库构建，经标注实践验证，该理论体系及表示方法能够覆盖绝大部分的语言现象，详细构建方法、过程及数据分析请见另文讨论[4]，以下简要进行介绍。基于块依存理论，以数据标注规范作为指导，以两两对比标注的模式，在基于浏览器的在线标注系统中，我们标注了百科和新闻领域文本，构建了汉语块依存树库。截至目前为止，树库规模为187万字符其中包含67%新闻文本和32%百科文本（仍在扩展中）。其中，新闻文本来源于新浪2006年新闻、新华社2012-2018 间新闻，百科文本来源于百度百科，分属自动化控制系统、电子学与计算机、轻工、大气与海洋及水文科学、航空航天、经济学等领域。各类别文本信息见表1。

	文件数	字符数	复句数（ROOT）	单句数（IP）
新闻	1461	1266466	29471	73973
百科	738	608389	16702	32289
总计	2199	1874855	46173	106262

表 1: 不同来源文本标注统计

当前树库中共包含299763个谓词词符，13425个谓词词形。其中约有1877个谓词（token）无从属成分，其余谓词均至少支配一个从属成分，依据目前定义的6类谓词和组块之间的依存关系，统计结果见表2。

依存关系	核心谓词（type）	核心谓词（token）	依存块数量（token）	谓词平均可支配组块数
NP-SBJ	11705	96199	101877	1.059
NP-OBJ	6757	72938	73850	1.013
VP-OBJ	1151	11901	14357	1.206
NULL-MOD	10267	73309	104980	1.432
VP-SBJ	765	2297	2805	1.221
VP-EMP	12	13	17	1.308

表 2: 各类依存块依存情况统计

从统计结果上看，在出现的一万三千多个谓词中，进行缺省补全后约有87%的谓词可支配名词性主语块，其次为修饰语块，为76%左右。这表明汉语中谓词支配主语和修饰语的普遍性，在明确动词具备支配该类组块能力的情况下，进行缺省补全是有必要的。另外，谓词支配修饰组块的能力最强，树库中平均一个谓词可支配1.432个修饰语块或小块。修饰语块中携带了大量的情态、时间、地点等各类语义信息，但内部结构相对复杂，存在框式结构、介宾短语等内部成分，因此进行小块切分，能够便于后续的语义角色分析、情态结构分析等工作。单个谓词支配名词性主语和宾语的组块数量相对较少，但仍略大于1，则表明语言中主谓谓语句及双宾

[4]具体分析请见《汉语块依存语法与树库构建》

语的现象占少数，后续工作中分析单主语和单宾语与谓词间的语义关系应作为重点，而相对于双宾句，主谓谓语句优先级更高。

5 块依存语法的特点

块依存语法是一种结合了组块分析、依存语法的语言分析方法。按照块依存语法所生成的句子，与宋柔所提到的“自足句”有相似之处，但也更进一步关注句子内部。

块依存语法以组块为研究对象，能够避免纠结于“词-词”之间的依存关系，关注句子的整体结构，进一步降低存储和分析的复杂性，也能够达到减少分词碎片、加强鲁棒性的目的；关注句法结构层面的组块，能够厘清句内-句间的界限，为篇章关系分析打下基础；以谓词为核心，在上下文中找到其支配对象，能够在句子层面补全缺省成分的同时明确内部成分的指向、句子结构。此外，块依存语法不仅关注常出现在句首的主语、状语成分，也关注经常出现在句中或句末的宾语、补语等，使生成的句子更加完整。

我们还注意到，以谓词为分析对象使得句法分析根据灵活。块依存语法分析能够以块依存图的形式对句子进行展现。整个句子以空节点为根，指向句中的核心谓词，核心谓词又有各个线条指向其支配成分。在篇章关系分析中，无论是寻找句间关系还是直接分析谓词间关系，都能够以更准确的分析单元为着力点。

袁毓林（2002）曾将信息抽取所需的语义知识分为三类，分别为宏观篇章知识、中观论元结构知识、微观层次的逻辑结构知识。块依存语法能够解决微观层次的逻辑结构知识，并为中观论元知识和宏观篇章知识打好铺垫。事实上，核心谓词的支配成分除了特殊的空述语之外，均可与汉语中的论元结构相挂钩，其余的状语成分、补语成分也可提示情态信息，此时的谓词论元、情态成分等均已齐全，仅需进行分类即可。在宏观层面，已明确的篇章分析单位，结合其余的辅助组块、衔接组块，为分析篇章的逻辑语义关系带来便利。

6 结论

本文创造性地提出了汉语块依存语法，并介绍了其标注体系和目前构建的树库规模。块依存语法在句内和句间寻找缺省的组块，补全缺省成分，以此为基础，也能够更深入地进行篇章层面的“小句-¿句间-¿篇章”关系探索。块依存语法与具体的语境、语用环境相结合，能够较好地解决当前中文自然语言处理中存在的分析对象不明确、依存对象不清晰、成分缺失等问题，更好地服务于事理图谱、知识图谱、问答系统、信息抽取各项任务。

参考文献

Steven P. Abney. 1991. *Parsing By Chunks. Principle-Based Parsing.*,257-278. Springer Netherlands.

Robinson, J.J. 1970. *Dependency Structures and Transformation Rules.*,1970,46(2) [J]Language.

Zhou Ming. 2000. *A Block-Based Robust Dependency Parser for Unrestricted Chinese Text.* The second Chinese Language processing workshop attached to ACL 2000, HongKong.

陈平. 汉语零形回指的话语分析[J]. 中国语文, 1987,(5):363-378.

陈亿,周强,宇航. 分层次的汉语功能块描述库构建分析[J]. 中文信息学报.2008(03):24-31+43.

郭艳华,周昌乐.. 面向多领域多来源文本的汉语依存句法树库构建[J]. 中文信息学报.2019. 38-46.

李素建. 汉语组块计算的若干研究[D]. 中国科学院研究生院（计算技术研究所, 2002.

刘伟权,王明会,钟义信. 建立现代汉语依存关系的层次体系[J]. 中文信息学报,1996(02):32-46.

卢露,矫红岩,李梦,荀恩东. 基于篇章的汉语句法结构树库构建[J/OL].自动化学报:1-12[2020-08-18].http://kns.cnki.net/kcms/detail/11.2109.TP.20200521.1558.007.html.

宋柔. 汉语篇章广义话题结构的流水模型[J]. 中国语文, 2013(06):483-494+575.

宋柔,葛诗利,尚英,卢达威. 面向文本信息处理的汉语句子和小句[J]. 中文信息学报, 2017,31(02):18-24+35.

徐赳赳. 现代汉语篇章回指研究[M]. 中国社会科学出版社,北京,1992.

闻媛,宋丽,吴泰中,李斌,周俊生,曲维光. 基于中文AMR语料库的非投影结构研究[J]. 中文信息学报, 2018,32(12):31-40.

尹鹏. 基于SVM的中文组块间依存关系分析[D].大连理工大学,2006.

袁毓林. 流水句中否定的辖域及其警示标志[J].世界汉语教学,2000(03):22-33.

袁毓林. 信息抽取的语义知识资源研究[15][J].中文信息学报,2002(05):8-14.

周明,黄昌宁. 面向语料库标注的汉语依存体系的探讨[J]. 中文信息学报, 1994(03):35-52.

周强. 汉语基本块描述体系[J]. 中文信息学报,2007(03):21-27.

周强,孙茂松,黄昌宁. 汉语句子的组块分析体系[J]. 计算机学报,1999(11):1158-1165.