

汉语块依存语法与树库构建

钱青青

北京语言大学/ 北京海淀学院路15号

qianqingqing19961@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

本研究依据以谓词为核心的块依存语法构建块依存树库，在句内和句间寻找谓词所支配的组块，利用汉语中组块和组块间的依存关系补全缺省部分，明确谓词支配关系。目前共标注2199篇文本，涵盖百科、新闻两个领域，共约187万字语料。本文简述了块依存语法的原则，并对组块及其依存关系进行了定义。将详细介绍标注流程、标注一致率、数据分布等情况。基于现有的树库，本研究发现汉语中有约25%的小句是非自足的，约有88%的核心谓词可支配1~3个从属成分。

关键词： 组块；块依存语法；树库

Chinese Chunk-Based Dependency Grammar and Treebank construction

Qian Qingqing

BLCU / 15th Xueyuan Road, Beijing

qianqingqing19961@foxmail.com

Wang Chengwen

BLCU / 15th Xueyuan Road, Beijing

chengwen_wang15@126.com

Abstract

This study create a treebank according to Chunk-Based Dependency Grammar ,which take predicate as the core and chunk as the object. With this grammar, predicate-dominated chunks can be found within and between sentences, default parts of sentences can be completed by the relations between chunks, and the dominant relations can also be clarifies .We have currently labeled 2199 texts, covering encyclopedia and news based on the Chunk-Based Dependency Grammar system, totaling about 1870,000 words.This paper briefly describes the principles of Chunk-Based Dependency Grammar and defines the chunks and relations.The labeling process, labeling consistency, data distribution, and so on are described in detail.Based on the existing tree bank, this study found that about 25% of clauses in Chinese are not self-sufficient, and about 88% of core predicates can control 1-3 subordinate components.

Keywords: chunk , Chunk-Based Dependency Grammar , treebank

1 引言

依存句法是自然语言处理领域的热门研究也是基础研究，其目的是将输入句子从序列形式变为依存树状结构，通过判断句内的词语之间是否存在依存关系以及存在何种依存关系，能够

适应灵活的语序特征，将句子分析为更加扁平的结构，从而降低了分析、标注、储存的难度，在问答系统、知识图谱、信息抽取等任务上发挥着重要作用。

值得注意的是，传统依存句法分析大多以词作为最小单元，在汉语中的应用也存在不适应的地方。汉语实际语篇中，词的词性、词义较为灵活，存在大量的活用、增加语境义等现象，传统依存句法分析以词作为分析节点的处理方式较难适应该特性；汉语具有典型意合特征，同样的语义内容可由语序不同的语言单元表达，关注其中的“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，也会因此影响到分析器的适应面和鲁棒性。此外，一些传统句法分析难以解决的问题在依存句法分析中也存在。句法分析一般以标点作为边界，而汉语中多流水句，主语、宾语、状语等的省略现象层出不穷，为分析结果的实际应用带来了困难。

为了解决以上问题，本研究提出了块依存语法，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，利用汉语中的组块和组块间的依存关系，既能够适应汉语灵活的语序特征¹，又能够将小句间成分缺省的问题转化为句间组块缺省成分补全的问题。同时以谓词为核心进行块依存关系构建能清晰呈现出句子的骨干结构，为后续任务提供准确的分析单元，关于块依存理论的详细说明请见另文讨论。

基于块依存理论，本研究对汉语组块理论、依存树库构建进行深入研究，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，标注百科文本、新闻文本，构建了汉语块依存树库。

2 相关研究

在传统的句法分析中，首先对句子进行分词和词性标注，再进行后续的句法语义分析工作。分词和词性标注的错误会带来较大的错误级联问题。与此同时，汉语有许多形式和语义上比较凝固的块成分，尤其是一些构式性成分，整体表示一定的语义，反而不适宜进行分词和词性标注基础上的句法语义分析。

组块分析理论由Abney在20世纪90年代初提出，CoNLL2000会议将组块分析作为Share Task提出(Erik & Buchholz 2000)，使该理论得到推广应用。国内学者也开展了大量的块研究工作。其中，刘芳、赵铁军等(2000)将块(Chunk)定义为一种包含一层或二层的符合一定句法功能和反映组成意义的短语结构，并将其分为八种类型；周强(2001,2007)从功能的角度对汉语中的语块进行了研究，定义了主语语块、述语语块、宾语语块、兼语语块、状语语块、补语语块、独立语块、语气块8类语块，形成了一套基于拓扑结构的汉语语块描述体系；其后，陈亿、周强等人(2008)设计了多层次功能块分析体系，进一步分析长功能块的内部结构；李素建(2002)将组块定义为符合一定句法功能的非递归短语，在划分组块时遵循非递归、无重叠、全覆盖的原则。

在依存句法树库方面，哈工大的汉语依存结构句法树库发表于2012年，以句法关系为主，语义信息知识作为补充，标注了人民日报约111万词的汉语语料。北大汉语依存结构句法树库发表于2015年，以依存句法为核心，并形成多种视图的标注体系，标注了新闻、专利及医药等约140万词的汉语语料。苏州大学面向多领域多来源文本构建了3万句左右的汉语依存句法树库。

在将组块理论与依存分析结合方面，Zhou(2000)较早地提出了一种基于块的依存分析器，分析块之间的依存关系，在非限制性的中文文本翻译中取得了较好的效果。但闻媛等人(2018)也指出由于中文中的模态词提升、话题化、成分分离等，在中文中存在较多的非投影结构，遵循依存语法的四条准则，为分析中文也带来了一定的难度。

此外，为解决汉语中多缺省的现象，宋柔(2013)归纳了广义话题结构遵从的堆栈模型和拓展后的流水模型，并将汉语的句子大致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(宋柔, 2017)，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或主语，大量句中或句尾的宾语、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，与句子更深层次的结构分析之间缺少衔接，大多还是停留在拆分复杂结构，生成“能说”的自足句层面。苏州

¹汉语语序灵活，但组块内成分具有相对稳定性

大学的多领域文本依存句法树库中也设置了表示谓语之间共同主语的依存关系，但并未全面地有针对性地解决缺省的问题。

3 汉语块依存树与组块关系

相对于细粒度的词来说，组块内部的句法、语义结构更加稳定，更符合语言的认知规律，是一种整存整取的单位。以组块为研究对象，能够避免纠结于“词-词”之间的依存关系，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，也能够达到减少分词碎片、加强鲁棒性的目的，因此本文的依存关系构建以“组块”为最小单元。

本文将组块定义为：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。在句法结构层面的组块按照功能可分为谓词块、主语块、宾语块、状语块、补语块，其中主语块和宾语块按照其性质可继续下分为谓词性主语块、谓词性宾语块、体词性主语块和体词性宾语块；除此之外，组块还包括篇章层面的衔接组块和辅助组块。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。

我们认为核心谓词组块是句子的核心，各类短组块均受核心谓词组块的支配并依存于核心谓词组块之上，在块依存关系分析中以谓词块作为句子的核心，寻找谓词所支配的各类组块。若某短语块和核心谓词组块之间存在依存关系，则称该短语块为核心谓词组块的从属成分，核心谓词组块为该短语块的依存对象。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，短语块至少依存于一个核心谓词组块之上。

核心谓词组块作为句内各组块的依存对象，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。

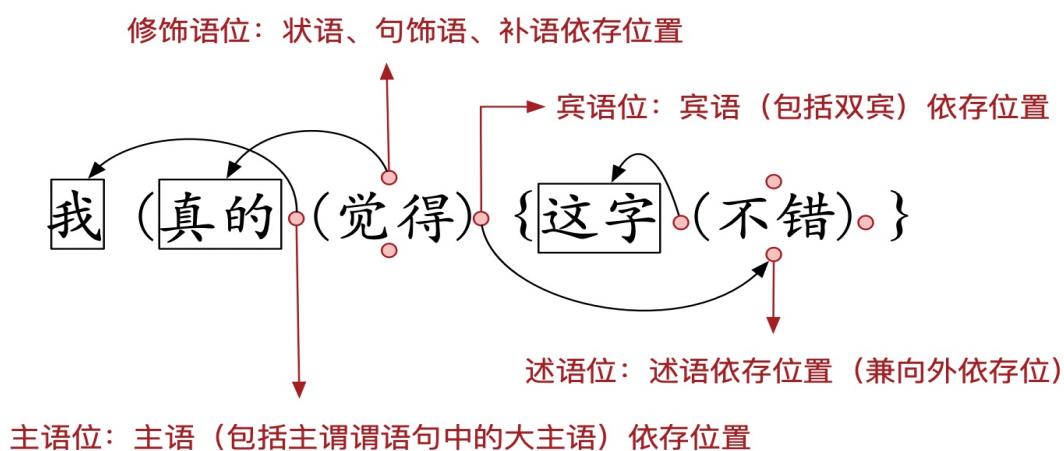


图 1: 块依存标注图示

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；在后续分析中，我们将谓词块与1号位上的体词性成分之间的关系定义为NP-SBJ，与谓词性成分之间关系定义为VP-SBJ。

状语、补语依存于谓词组块的2号位；在后续分析中，我们将谓词块与2号位上的成分之间的关系定义为NULL-MOD。

宾语，包括双宾语中的远近宾语依存于谓词组块的3号位；在后续分析中，我们将谓词块与3号位上的体词性成分之间的关系定义为NP-OBJ，与谓词性成分之间关系定义为VP-OBJ。

谓词性关联块与核心述语的4号位置连接，当某谓词组块依存于其他谓词组块时从4号位向外依存。在后续分析中，我们将谓词块与4号位上的成分之间的关系定义为VP-EMP。

因此，我们可以将谓词与其依存块之间的关系初步区分为以下6种：

标识	关系说明	举例
NP-SBJ	表示该谓词的体词性主语块	我说了一件事儿
VP-SBJ	表示该谓词的谓词性主语块	学习对我来说总是快乐的
NP-OBJ	表示该谓词的体词性宾语块	吃苹果
VP-OBJ	表示该谓词的谓词性宾语块	进行调查、研究
NULL-MOD	表示该谓词的修饰语块	上周她出门了。 跑得很快。
VP-EMP	表示该谓词的谓词性关联块， 即与谓词块相同的空述语。例 句中的空述语“()”是“有”的谓 词性关联块	有书三本，()笔三支。

表 1: 块依存标签说明

以谓词为抓手使得分析更具有灵活性，经过块依存语法分析的句子，能够展现为块依存图的形式。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。在句间关系分析中，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。

在进行块依存分析时，不限于小句或句子内部，而是补全因上下文而缺省的单元，能够将句子还原为更完整的形式，也为后续分析提供更完整的单元。例1中包含两个句子，其中第二个句子的两个核心谓词“苦”“蕴含”的主语是缺失的。通过观察上下文，我们可以找到对应主语应为前一句子主语修饰语“哥伦比亚咖啡”。相同地，例2中第二句缺失的状语“这些年”也可通过相似的方法在上下文中找回。

1) 哥伦比亚咖啡的风味是丰富多样的。不仅不苦，还蕴含着水果、坚果、谷物等不同气息。

2) 这些年，他通过努力进步了不少，是我们学习的榜样。而很多人却没有珍惜时间，仍在原地踏步。

3) 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。

块依存方法能够在补全缺省成分的同时明确句中成分的指向、句子的结构。在例3中，话头为“吕先生和许多严肃的学者一样，不会”，其中既有体词性成分，也有修饰性成分，修饰性成分中存在框式结构。利用块依存方法进行补全更具有理据性——话头部分能够成为另一个小句的一部分是因为它内部的两个组块都受到其中核心谓词的支配。

4 块依存树库建设

本文所标注的语料，均来自于“基于篇章的汉语句法结构树库”（下文称句法结构树库）。该树库目前已人工构建约1000 万字符集规模，包括1 万语篇文本、26.5 万单复句、64.4 万小句；以新浪及新华社新闻、百度百科、专利申请书、小学生作文、法律案件判决书等应用性文本为标注语料；树库中人工标注数据一致性校验Kappa 值均大于0.8。

在句子成分标注中已经明确了句、小句、组块的边界，并运用标注符号标识组块功能，本文所进行的标注是在句法结构标注基础上所进行的块依存标注。句法结构树库将句子成分分为主语、宾语、述语、句饰语、衔接语、辅助语，并分别以不同符号进行标注，如例4、5所示，“<>”表示衔接语，“<<>>”表示辅助语，“[]”表示句饰语（即与述语分离的状语或补语），“()”表示述语，述语内部又可利用“()”区分出状语、核心述语、补语，“{}”表示谓词性的主语或者宾语，体词性主语或宾语则无需用标注符号标注。

4) <但是>，[到今天为止]我(还是(放心)不下)你<<啊>>。

5) {(大力(发展))经济}(是)我们目前的工作重心。

在句号、分号、叹号等8个标句点号的基础上对篇章进行划分，通过语篇句子成分标注对句

子边界进行校准，能够明确句、小句的边界。当原本应属同一小句的主宾语、定状补语等向核成分被标点切分开的时候，使用标注符号将标点括在内部，达到取消标点切分的分句功能的目的：

6) [抗日战争胜利前夕，]党中央和毛主席(发出)号召和命令。

7) 他(是)河北辛集马兰村的一名普通农民。

8) “独自飞行1840.018公里的北京山茗网络科技有限公司创始人”\\，(是)彭少仪名片上的唯一头衔。

9) 我(吃)饭| 他(睡觉)。

10) [五个多月来，]中国海监船编队(一直(坚守)在祖国的钓鱼岛海域)(巡航)，他们(克服)重重困难，(涌现了)无数可歌可泣的先进事迹，他们“特别能吃苦、特别能战斗、特别能奉献”的“海监精神”，(也极大地(鼓舞了))全国人民。

例句6~10均为一个“句”，也是一个小句单元，而例9~10是通过分析校正后确定的句子，分别以两个和五个小句构成。

为了提高标注效率和质量，本研究实施在线标注，以下为标注界面：



图 2: 标注页面示例

本研究利用kappa值计算加权的一致率，计算公式为：

$$Similarity = \sum_{i=1}^n k_i p_i \quad (1)$$

其中k为该依存关系的kappa一致率，p为该依存关系在全篇任务总依存关系数量中的占比。一般认为若kappa值在0.8以上，则表明二者的一致率较高，本树库为保证高水平的一致率，将阈值定为0.9。数据标注一致率控制流程如图4所示，在早期的标注任务中，采用双人标注的模式，只有当基于kappa计算的一致率不低于0.9时，标注任务才算通过，若一致率低于0.9，则需要通过讨论修改、专家介入等方式提高一致率。质量比对通过后选取当期标注任务中平均kappa值较高人员文本入库。

以上策略开展标注实践，截止2020年8月，共标注2199篇任务，涵盖百科、新闻两个领域；共计约187万字语料，包含超过4万个句子，10万多小句，最终平均kappa值为0.945。

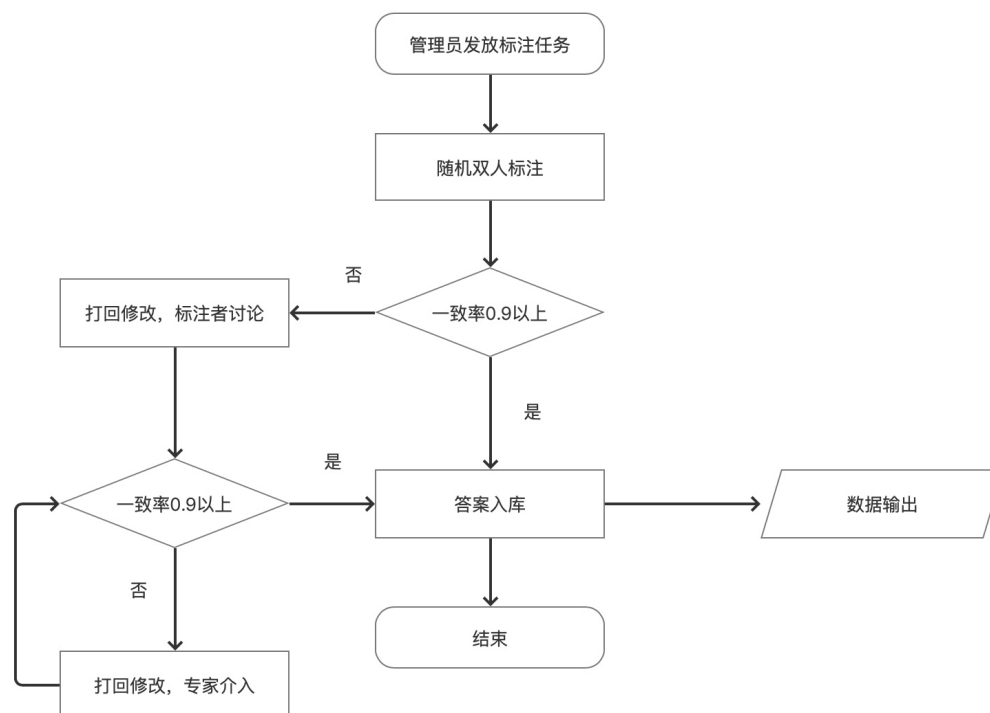


图 3: 数据标注流程

5 汉语组块关系

5.1 自足句与非自足句

下面以树库中的708个任务作数据分析, 内含554521字, 14181个复句, 30724个小句。

此处我们定义两类句子, 即“自足句”和“非自足句”。自足句指句中不缺省成分的句子, 句中核心谓词的所有依存块均位于该句内部。从此定义出发, 可知有一类较为特殊的自足句是独词句或仅包含篇章成分的句子, 例如“哗啦哗啦”“因为”等, 在后续分析中会将此类独词句独立分析; 此外, 某些无法补全缺省成分的句子如“(下)雨<<了>>”, 也认定为自足句。“非自足句”则指句中缺省成分, 句中核心谓词有位于其他句的依存块。相应的, 可以将小句和句子分为“自足小句”“非自足小句”“自足句”“非自足句”。

11) 国家发改委(联合)相关单位(连续(出台))中长期发展规划和场地建设规划,

12) (督促)地方政府(加快(制定))实施细则;

13) 督查涉及部门(很(多)), 内容(非常(具体))<和>(细化), <应该说>(是)一次无缝的、立体的和全方位的督查。

14) ()研究空气和燃气与发动机各零部件相对运动及其相互作用的学科, (是)流体力学的一个分支。

在例11中, 谓词“联合”的主、宾语块分别为“国家发改委”、“相关单位”, 谓词“出台”的主语块为“国家发改委”、“相关单位”, 宾语块为“中长期发展规划和场地建设规划”, 状语块为“连续”。例11内部所有谓词块的从属成分均位于该小句内部, 因此我们称这样小句为“自足小句”。

例12中, “制定”的各个从属成分均位于该小句内部, 但“督促”的主语块位于其他小句中, 在这里应该是“国家发改委”, 那么我们称某些小句内部存在谓词块的从属性成分在句外的是“非自足小句”, 需要通过块依存标注来补全结构。

例13是一个复句, 内部的几个核心谓词“多”、“具体”、“细化”、“是”的各个从属成分均位于该复句内部, 虽然“是”的主语块是跨了一个小句的大主语“督查”, 但我们依然认为在复句层面, 这个句子是自足的。当然, 割裂来看, 在小句层面, 第三个小句是不自足的, 经过分析并补全之后, 该句可以形成三个内部自足的小句: “①督查{涉及部门(很(多))}, ②内容(非常(具体))<和>(细化), ③<应该说>督查(是)一次无缝的、立体的和全方位的督查。”。

与之相对应的是例14，其中包含的一个空述语和谓词“是”在该复句内部均没有相对应的主语，我们需要在上文中找出“某某学科”作为它们的主语。因此我们认为这一类在内部存在无法找到所有从属成分的句子是“非自足”的。

分析数据中的小句和句子中，自足和非自足的分布结果如下图所示：

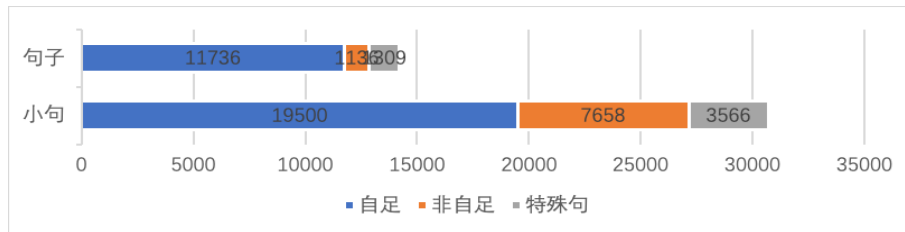


图 4: 句子/小句内自足与非自足分布图

通过统计，发现其中有23066个小句和13045个句子是自足的，分别占总数的75%和92%。若排除上文提到的特殊句，则分别占63%和83%。另外有7658个小句和1136个复句是非自足的，分别占总数的25%和8%。

上述结果在经过结构标注的句边界校准之后得到，由此可知的是：在汉语中，成分缺省是普遍存在的事实。若简单以标点符号分割后的文本来作为分析单元，则将会有25%以上的小句存在内部成分缺省，会造成指代、时间、地点等情态信息不明等问题。而本研究通过块依存的方式，补全了缺省的小句和句子，使这25%的句子变为“自足”的句子，能够极大程度地填补句子缺省信息，便于后续基于自足句子的分析。

5.2 谓词及其依存块

基于已有的树库，分别从核心谓词及和核心谓词支配块数量等角度做了统计。在标注数据中共包含5479项核心谓词，其中有4676项动词存在含有体词性主语块的实例，4089项动词存在含有修饰语块的实例，2943项动词存在含有体词性宾语块的实例，477项动词存在含有谓词性宾语块的实例，229项动词存在含有谓词性主语块的实例，6项动词存在含有谓词性关联块的实例，还有340项动词存在没有任何从属成分的实例。

标识	谓词项	占谓词总比	谓词实例数	依存块数	平均可支配组块数
NP-SBJ	4676	0.853	25744	26673	1.036
VP-SBJ	229	0.041	477	617	1.2935
NP-OBJ	2943	0.537	20786	21093	1.0148
VP-OBJ	477	0.087	2873	3529	1.2283
NULL-MOD	4089	0.746	19351	27278	1.4096
VP-EMP	6	0.001	7	7	1
NoDep	340	0.062	494	0	0

表 2: 各类依存块依存情况统计

如图5所示，在30487个谓词实例中，约有88%支配1~3个从属成分。值得注意的是，从属成分数量在2~3个的谓词数量均多于从属成分为1个的谓词，这表明多从属成分的谓词在汉语中是更加普遍存在的，谓词能够支配的句法成分不单一。从属成分最高为10个，为例15中的“是”，它的从属成分包含1个体词性主语块、7个体词性宾语块和2个谓词性宾语块。在从属成分数量大于6的31个实例中（图7），绝大多数的谓词并非实义动词，“是”“有”“包含”“分为”“进行”等含义抽象的系动词或是形式动词等，且这些动词大多能够支配多个主语块或宾语块；而部分实义动词，如“列入”“拓展”等，更多的是支配了多个修饰语块；对于一些具有认知、言说义的动词，如“要求”、“意味”等，则相对可能性更多些，它们存在支配多个宾语块的能力，同

时也有较为明确的意义，也能够支配较多的修饰语块。在依存块方面，不同于宾语块，体词性主语块和谓词性主语块的差别较为明显，体词性主语块的数量多维持在一个较为稳定的水平，一般不超过2个，而谓词性主语块的数量则起伏较大，一般无谓词性主语块但也存在谓词性主语块数量大于5的情况。

在340项没有任何从属成分的动词中，有58个动词（约17%）为古诗词或熟语，如“谋长远计”“智者善谋”“功夫不负有心人”等。在结构标注中，为了保持其内部语义的完整性，并未将其切分为多个部分，也因此使其内部一般不缺省成分，其他则多是在上下文语境中难以补全其缺省成分的，如例15中的“学习”。

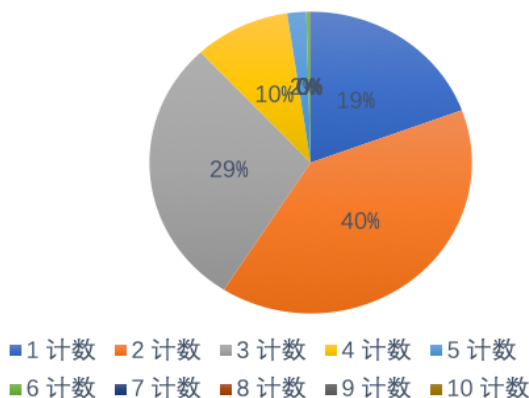


图 5: 谓词块从属成分统计

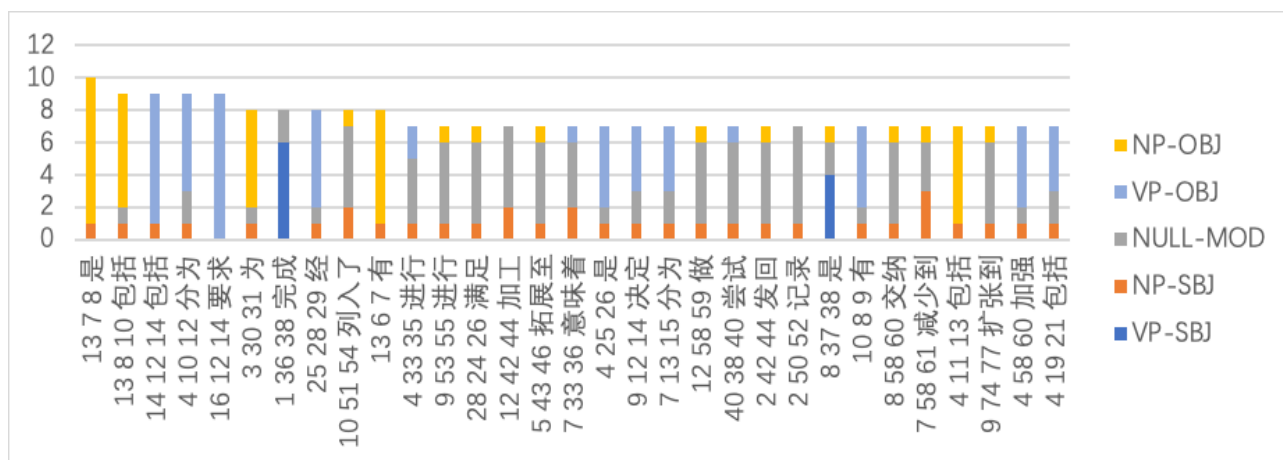


图 6: 从属成分数量最多谓词的从属块类别统计

17) {(学习)、(看书)}(是)让人很愉快的事情。



图 7: 块依存标注结果

从依存块的角度进行分析，该语料中共包含了79197个语块。其中体词性主宾语、修饰语占了绝大部分，均约占30%。与表2结合，我们可以发现，体词性主语和修饰语的分布较为平均，语料中绝大多数的谓词均能够支配主语修饰语组块，仅有约一半的谓词项能够支配宾语

块，这与不及物动词的存在有着密切的关联。而谓词性主语、谓词性宾语则都少于0.1，最少的是VP-EMP组块，仅占0.001，表明该类关系在语言中较为特殊。从可支配组块数量看，每个谓词实例平均可支配体词性主语块为1.04个，体词性宾语块也在1.01左右，谓词性的主、宾语块数量略高，约1.2左右，而修饰语组块最高，为1.41个，表明在实际语料中，往往有较多的修饰语修饰谓词，能够表达较丰富的情态、时态等信息，若简单地线性分析，则会丢失大量的信息。

6 总结

本文介绍了块依存树库建设的相关工作，截至2020年8月为止，本树库共标注了约187万字的高质量语料数据，包含超过4万个句子，10万多小句。本树库主要面向的是汉语中多流水句、缺省普遍的现象，通过依存标注的方法，以谓词为核心，从上下文中寻找支配的组块，从而补全缺省部分、明确句内支配与被支配的关系，也为接下来的语义分析奠定坚实的基础。基于此树库，我们发现汉语中有约25%的小句、8%的复句内部存在缺省现象；汉语中体词性主宾语块、修饰语块占绝大部分，约有90%，谓词块支配体词性主语、修饰语的能力更强，支配多组块的可能性更高。

基于块依存的树库建设充分遵循汉语语序灵活但块内顺序相对稳固的特点，同时将分析单元上升到块，可以有效抓住句子的骨架进行分析，避免了“词-词”依存分析所导致的句子依存结构繁琐问题。另一方面，谓词为中心的依存表征体系也为语义分析提供了结构支撑。通过树库的建设验证了组块依存分析理论的可行性，同时该数据资源能够为自然语言处理发展提供资源支撑。

在接下来的工作中，我们会进一步扩大标注，增加树库的规模和覆盖范围，增加不同领域的文本；探索各类关系内部依据特点的细分，例如修饰块内部可按照其语义关系细分为时间、地点、情态、数量等等。

参考文献

- Steven P. Abney. 1991. *Parsing By Chunks. Principle-Based Parsing*,257-278. Springer Netherlands.
- 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.
- 郭丽娟,彭雪,李正华,张民. 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.
- 郭丽娟,李正华,彭雪,张民. 适应多领域多来源文本的汉语依存句法数据标注规范[J]. 中文信息学报.2018. 32-39+56.
- 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).
- 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.
- 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.
- 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. 汉语句法树库标注体系[J]. 中文信息学报, 2004, 18(4):2-9.
- 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.