基于结构检索的汉语介动搭配知识库构建

王诚文[1]，**饶高琦**[2]，**荀恩东**[1]
(1.北京语言大学 信息科学学院，北京市 100083;
2.北京语言大学 汉语国际教育研究院，北京市 100083)

摘要

以往的介词知识库构建重视介词语义和介宾的搭配研究，鲜有对介动搭配进行系统研究及知识获取的工作。而汉语介词发达及动词是句子中心的特征决定了介动搭配研究的重要性。本研究基于结构检索技术，充分借助短语结构属性和结构信息，从大规模语料中抽取介动搭配16033对。并提出了介动搭配紧密度的度量方法，初步分析证明其远优于依靠绝对频次进行搭配度量的方法。

关键词： 介动搭配；结构检索；介动紧密度

Construction of Preposition-verb Knowledge Base Based on Structure Retrieval

Wang Chengwen[1]，**Rao Gaoqi**[2]，**Xun Endong**[1]
(1.School of Information Science,Beijing Language and Culture University,Beijing,100083;
2.Research Institute of International Chinese Language Education,
Beijing Language and Culture University,Beijing,100083)

Abstract

The construction of preposition knowledge base in the past had given more prominence to the study of semantics of prepositions as well as collocations of prepositions and their complements, while there had been few work on systematic research and knowledge extraction of preposition-verb collocations. The characteristics of Chinese that prepositions are well-developed and that the heads of sentences being verbs determine that the study of preposition-verb collocations is crucial. Being based on a technology of structures retrieving and fully utilizing phrasal structure attributes as well as information of structures, we have extracted 16033 pairs of preposition-verb collocations from a large-scale corpus, and came up with a method for measuring the closeness of a preposition-verb collocation, which was preliminarily proved to be superior to the method of absolute frequency.

Keywords: Preposition-verb collocation , Structure retrieval , Closeness of preposition-verb collocation

第二十届中国计算语言学大会论文集，第513页-第524页，呼和浩特，中国，2021年8月13日至15日。

1 引言

汉语作为区别于印欧语系的语言有着自身独特的特点。从搭配角度来看，相较于跨语言通有的搭配类型，诸如动宾、状中和定中，汉语有着丰富的量词和虚词参与构建的搭配，因此有必要重视虚词在搭配中发挥的作用(胡韧奋，2019)。作为集词汇、语义和语用特征于一体的介词，在汉语中起着重要的句法语义作用，突出表现在介引宾语修饰相关谓词，同时起标明语义角色的作用。无论是语言本体研究还是自然语言处理领域，都较多关注介词与宾语的搭配、介词与方位词、助词和连词的搭配形成的框式结构。然而，从语言现实出发，可以看出介动搭配不仅具有频率上的高频稳态出现性，也有其独特的句法语义特征。试看例子1：

(1)

a 向老师问好

b 向远处投递

对于1a来说，“老师”充当的是邻体的语义角色，表示动词“问好”的对象。在例1b中，“远处”充当动词“投递”的方向语义角色。同样的介词“向”，介引出不同的语义角色，主要是在与不同动词结合时，凸显出来介引语义的不同。在与[+方向]义动词搭配时，便突出表示出来方向的含义。

通过上述例子分析，可以发现，能够介引不同语义角色的介词在与具体动词搭配后便能够凸显出来唯一的一种语义。介词作为虚词，是语法化的结果，语义往往比较虚化。正是通过与动词的搭配使用，凸显出来自身的语义。与此同时，汉语中一些动词，对于介词有着较强的选择性，突出表现在其充当谓语时，一般有介词与之共现。比如，对于动词“着想”来说，其做谓语时形成的表述为“为*着想”。

与此同时，囿于已有语料库检索技术的制约，从搭配获取方面来看，以往基本上是基于词语共现关系，并利用互信息等统计特征进行搭配知识获取。完全基于统计特征的搭配获取会得到许多相关词语。例如“医院-医生”。在语法或语义体系下界定的搭配，诸如主谓、动宾等的获取上，只能利用词性序列符号进行获取。词性符号序列是在线性的语言序列上进行符号匹配，往往会抽取出来许多伪搭配。例如“p*反应”会匹配到“跟人交流时难免不能快速反应”类似的噪音数据。本研究在介动知识获取阶段将会利用面向句法结构的检索技术来规避上述提到的问题。

本文将从理论上，关注介词与动词形成的二元搭配，分析形式特征，并在数据抽取基础上对搭配内部句法语义特征进行进一步研究；并从实践角度出发，利用面向句法结构的检索技术，结合介动搭配形式特征，制定介动知识获取规则，从大规模多语体语料中获取介动搭配知识，形成介动搭配知识库，以期为语言本体研究和自然语言处理领域提供语言知识资源。

2 相关研究

语言学本体领域关于介词的相关研究层出不穷，其理论上的研究工作为介词短语、搭配等语言知识库的构建提供理论指导作用。然而，目前，以汉语介动搭配为视角系统构建搭配资源的工作鲜有。

2.1 介词本体研究

在汉语研究中，关于介词的研究工作主要聚焦在介词分类体系、句法语义和历时演变等几个方面。从分类上来看，有的学者关注介词类别及数量的多少，界定出类别不一样的介词体系，例如，范晓(1990)从语义角度将介词短语分为了九类：处所、时间、受事、施事、与事、共事、凭借者、关涉者和条件。实际上，也是从介词短语划分的角度给介词做分类。张谊生（2000）则划分出了十五类介词。陈昌来(2002)则在7大类介词基础上，进一步总结汉语中介词的总数在150个左右；也有学者从介词的位置和形态出发进行分类。刘丹青(2002)则率先提出汉语中存在框式介词的现象。介词句法语义的研究方面，则重在与连词、动词等比较分析中明确介词的句法语义特征，而这种比较往往会从共时层面延展到历时层面，代表性的研究工作有(张旺熹，2004；何洪峰，2014；刘静敏，2015)。

2.2 介词相关知识库构建

围绕介动搭配进行系统语言知识库的构建工作还鲜有。其相关的知识构建工作有：介词组块的标注、介词词典编撰和介词知识库构建等。

邹宏梅(2007)在组块识别任务中，界定了一种由单一介词构成的块，并以宾州树库语料为基础，标注了用于实验的小规模组块数据。同样王莹莹（2006）、高红（2007）等以北京大学计算语言学研究所1998年的人民日报语料为标注对象，标注了单一介词及长度不超过3个词语的介词框架。上述工作在介词组块的界定上较为简单，一般只关注介词本身，没有注意到介词与动词之间的句法和语义上的密切相关特征。

方清明(2017)编撰的《现代汉语介词用法词典》主要从介词的语义和常规句式入手，围绕着149个介词进行了穷尽性的知识刻画。其中的介词框架中，框架的后置成分有些是由动词充当的，比如“与*相比”等，但是整个框架倾向于出现在句首位置做状语，交代后续句子的时间、地点和比较对象等信息，没能够系统注意到动词做谓语中心语时与介词的搭配情况。

作为现代汉语广义虚词知识库的一个子库(俞士汶，2003)，现代汉语介词知识库(彭爽，2009)包括机器词典、标注语料库和规则库三部分。其机器词典围绕介词刻画了构词语素、词族、语体色彩、体宾谓宾、否定、介词框架、介词短语充当定语、介词短语作主语前修饰语、单独成句、句法结构和格标记等12个属性字段。可以看出，其并没有对介词和动词的搭配进行细致刻画，还是传统上介词的一些句法功能和分布的描写。邢丹（2020）利用词性、词长和标点符号信息构造正则表达式从大规模语料库中获取介词结构搭配，主要是介词、介宾中心词和动词的三元搭配实例。囿于在搭配中考虑到了介词宾语的论元实例，因此在抽取数据上会有数据稀疏情况。与此同时，如引言所介绍，一定程度上，介词的格标凸显作用是在与动词结合后便能够明确，例如“向*问好”和“向*投掷”的介词介引论元的区别便可以直接明确下来。因此，介动搭配获取是更值得关注的视角。与本文聚焦的介动搭配研究工作最为接近的工作当属于胡韧奋。但其从本质上与本文的介动搭配还有明显区别。其从服务于二语教学的角度出发，构建搭配知识库，其中一种类型为介动搭配。只不过其只关注到“把、被、对”等引导动词施事或受事的介词与动词形成的句法格式，例如“把X 解决”、“把X 带过来”和“把X 买了回来”。一方面，其只关注有限的介词“把或被”；另一方面，其抽取的更多是一种句式，诸如“把X 带过来”等，不是系统化的介动之间的搭配知识。

3 结构检索

以往支撑语料库检索的语料主要是经过分词和词性标注的，因此决定了其后续的使用方式。从检索式上来看，只能由字符、词性和通配符等构成查询语句；从匹配方式上看，只能在线性的语句上进行模式匹配。而语言是具备层级化结构的，只有充分利用语言结构进行检索设计，才能够助力深度的语言知识的检索和获取。

结构化检索是指在结构化分析语料基础上，利用语言学特征设置支撑句法结构及短语属性检索的检索技术。为了便于后文介动知识获取的检索过程的介绍，在下面将会事先介绍结构化语料、结构化检索设计原则和结构化检索系统使用。

3.1 结构化语料

综合考虑语言结构表示与语言资源构建效率之间的关系，我们制定了一个以句法性质与功能为主的同时参考篇章功能、人际功能的组块体系，将句子标注为一个带有性质与功能信息的组块序列。目前人工和机器自动标注后达到的结构化语料规模在195G左右。

具体的人工标注组块的符号介绍如下，“()”表示述语块[0]，述语内部又可利用“()”区分出核心谓词块、状语块和补语块，“<>”表示衔接语块，“<<>>”表示辅助语块，“[]”表示句饰语块(即与述语分离的状语或补语)，“{}[1]”表示谓词性的主语块或者宾语块，体词性主语块或宾语块则无需用标注符号标注。试以下句子说明标注范式:

句子1：在那种情况下，首先，他应该尽可能自保，然后再去帮助别人吧！

标注1：[在那种情况下，]<首先>，他(应该尽可能(自保))，<然后>(再去(帮助))别人<<吧>>！

[0]述语块指以核心谓词为中心包含前后连续性状语和补语的，充当句子述语成分的组块。

[1]相当于对于句子中的表示命题义的语言单元可以进一步嵌套标注。

基于人工注入标注符号的组块序列，可以无歧义转换为带性质功能标签的组块状短语结构树的形式，上述句子的组块状短语结构树为图1。

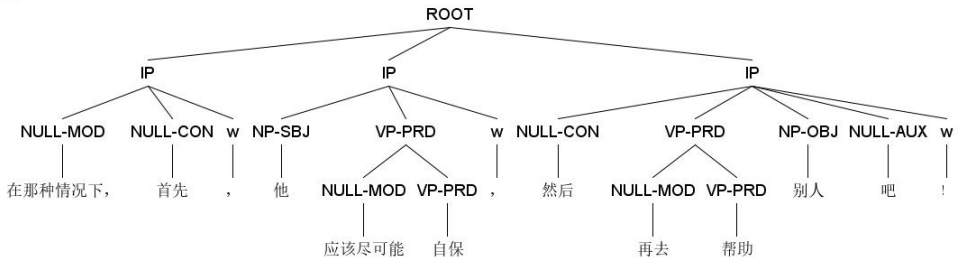

Figure 1: 组块短语结构树示例

具体的短语结构树成分标签见表1。

标识	标签说明	举例
NP-SBJ	表示体词性主语块	我说了一件事儿。
VP-SBJ	表示谓词性主语块	学习对我来说总是快乐的。
NP-OBJ	表示体词性宾语块	吃苹果。
VP-OBJ	表示谓词性宾语块	进行调查、研究。
NULL-MOD	表示修饰语块	上周她出门了。 跑得很快。
NULL-CON	表示衔接组块	首先，应该自保护。
NULL-AUX	表示辅助组块	啊，你竟然考过了。
VP-PRD	表示述语块	他的确十分努力

Table 1: 组块状短语结构树的成分标签

3.2 结构检索设计原则

结构化的语料主要是一种注入了性质和功能信息的组块状短语结构树，特别是对于句子的述语块来说，标注出了其中心词（核心谓词）。结合汉语的实际语言特征和语言知识抽取的需要，在设计检索系统时，考虑以下几个因素:

1 体词性短语中心突出原则

体词性短语是一种向心结构的短语，其主要语义落在右中心词语上。在检索系统设计时候，以体词块的后缀为索引对象建构索引，便于获取体词块的中心成分。

2 述语块结构的限制

结构化语料中，介宾和核心谓词是整体标注为述语块，对应一个VP-PRD的节点标签，同时标注出了核心谓词的左右边界。因此，介动抽取中，介词和动词形成的离合检索在匹配时候可以在整个述语块的内部进行匹配，这样会缓解许多噪音数据的影响。

3 支持短语属性检索

在设置检索系统时，也构建了以结构块的属性为对象的索引，支持利用短语结构块的属性进行查询。介词一般出现在状语中，所以利用状语块的标签来限制介词位置，进行搭配获取。

3.3 结构化检索系统

结构检索的查询表达式主要是由字、词、词性、通配符、短语块属性构成。其查询语句的形式形如“Query{Condition}Operation”。其中Query主要是查询单元；Condition可以对查询单

元中指定的部分进行长度、内容和输出的限制；Operation主要指统计查询单元的频次或者是上下文。具体的检索式解释见表2。

检索式	内容		举例
Query	字符	汉语中的字符集合	“我”、“你”
	词	汉语中的词	“打击”、“按照”
	词性	词性符号，这里指北大词性标注符号	v、a、n
	通配符	~代表一个词	吃~能够匹配“吃米饭”
		*表示小句内离合	“洗*澡”匹配“洗个澡”
		^表示小句间离合	“对^打击”匹配“对那种情况，应该猛烈打击”
	短语属性	|Tag*Tag|	|NP-SBJ*NP-SBJ|表示一个名词性主语块
	括号	()括在检索式内部构成单元外边	(p)*着想{len($1)=2}Freq中的括号捕获介词p，并用$1代替，在“{}”条件内设置长度为2。“Freq”表示输出介词频次
Condition	长度限制	len($1)=2	(p)*着想{len($1)=2}表示介词长度为2
	内容限制	$1!=[进行予以]	(v)起来{$1!=[进行予以]}限制后边能加趋向动词的动词不能为形式动词
Operation	Freq	输出查询单元频次	Item Freq
	Context	输出查询单元上下文	Contex

Table 2: 结构检索的检索式构成

以动词“着想”的高频搭配介词抽取为例，简要说明结构检索的使用方法。其检索式为“|NULL-MOD(p)*NULL-MOD|(|VP-PRD*着想VP-PRD|){print($1);len($2)=2}Freq”。检索式的Query部分为|NULL-MOD(p)*NULL-MOD|(|VP-PRD*着想VP-PRD|)。表示一个以介词开头的状语块(NULL-MOD)后边紧跟着一个述语块（VP-PRD）。小括号分别捕获介词和整个述语块分别对应$1和$2.Condition部分为print($1);len($2)=2,表示限制整个述语块就是动词“着想”,并最终输出介词的统计结果；Operation部分为Freq,表示统计与“着想”搭配的介词的频次。其检索结果见表3。

“着想”搭配介词	频次
为	16542

替	2015
从	1469

Table 3: “着想”搭配介词的分布

4 介动搭配抽取

4.1 加工对象

本研究以《现代汉语词典》第五版中的15891个双音节动词为研究对象，从语料中获取与其搭配的介词。之所以选择双音节动词主要是因为单音节动词的多义问题较为突出，以及与介词形成搭配时语义的不完备。例如“为*打”的语义完整性不如“为*打水”，类似的的例子有，“与*有”不及“与*有关系”的语义完备。而双音节动词在与介词形成搭配之后，形式上和语义上较于单音节动词来说具备更强的完备性。

介动搭配上只考虑形如“p*v”的搭配形式，即只考虑介词出现在动词前的形式。不考虑动介搭配的主要原因在于能够出现在动词后边与之紧邻的介词的种类是有限的，一般为“于、在、自、给”，相对于出现在动词前的介词类型来说不具备多样性。

4.2 加工流程

步骤1 动词文件读取

读取15891个动词，存储至文件VerbList中；

步骤2 动词检索式构造

遍历VerbList中的每个动词，利用代码自动生成介动搭配获取的检索式。具体来说，用遍历得到的具体动词来替换检索式“|NULL-MOD(p)*NULL-MOD|(|VP-PRD***Verb**VP-PRD|){print($1);len($2)=2}Freq”中的**Verb**来得到每个动词的搭配获取的检索式。对于动词“着想”、“请教”和“赋能”的对应检索式见表4所示。

动词	检索式
着想	|NULL-MOD(p)*NULL-MOD|(|VP-PRD*着想VP-PRD|){print($1);len($2)=2}Freq
请教	|NULL-MOD(p)*NULL-MOD|(|VP-PRD*请教VP-PRD|){print($1);len($2)=2}Freq
赋能	|NULL-MOD(p)*NULL-MOD|(|VP-PRD*赋能VP-PRD|){print($1);len($2)=2}Freq

Table 4: “着想”、“请教”和“赋能”的介动抽取检索式

步骤3 批量检索

利用结构检索提供的WebAPI进行批量的检索，对于每个动词的介词搭配按照频次自高到低进行排列，输出到Verbi文件中。将介词与动词进行拼接形成“介词_动词”的搭配形式，按照频次自高到低的统一输出到PVColl文件中，该文件存储了初步获取的所有介动搭配实例及其对应的出现频次。

步骤4 条件限制

对于PVColl中的所有搭配实例进行条件限制，以去除噪音数据。主要从以下三个方面进行考虑:首先，删除部分经常加谓词性宾语的介词，诸如“鉴于”、“由于”和“自从”等。这部分介词主要后边带一个句子或者谓词性成分，一般不跟某个动词形成类似“为*赋能”式的形义完备的搭配形式。其次，利用现代汉语语法信息词典(俞士汶，1996)来对一些虚化动词进行过滤，排

除助动词、趋向动词、形式动词、情态动词和经常做状语的动词参与构成的介动搭配实例。最后，根据频次的观察，将阈值定在10，删除频次低于10的搭配实例。

4.3 搭配度量

为了进一步度量介词与动词搭配的紧密程度，本研究借鉴Chen(2017)利用词频逆文档频(TF.IDF)进行事件抽取的做法，使用动词凸显度和介词相关度来度量介词与动词的紧密程度。其具体的计算公式如下：

$$\text{动词凸显度} = \frac{\text{某介词搭配的某动词实例数}}{\text{某介词搭配的最多动词实例数}}. \tag{1}$$

$$\text{介词相关度} = log(\frac{\text{介词的Type数}}{1 + \text{某动词搭配的介词Type数}}) \tag{2}$$

$$\text{介动紧密度} = \text{动词凸显度} * \text{介词相关度} \tag{3}$$

按照上述计算公式，对PVColl中的介动搭配实例计算了其相应的介动搭配紧密度，部分动词的搭配频次和搭配紧密度的值如表5所示。

介动搭配	频次	介动紧密度
朝_迈进	698	1.073
同_握手	5300	0.995
与_无关	152429	0.924

Table 5: 介动搭配的频次和紧密度

为了进一步衡量利用介动紧密度来衡量搭配的有效性。分别将搭配数据按照介动搭配的频次和介动紧密度两个标准从高到低进行排列，分别考察了介词“在”参与构成搭配在Top1000中所占的比例。之所以选择介词“在”的原因有二：其一，从介词搭配实例统计中发现，介词“在”能够搭配最多的动词；其二，介词“在”的语义比较多样，介引时间、地点、条件和范围等，与动词的搭配紧密度相对较差。我们期望抽取出来的介动搭配凸显出来的语义比较单一和明确，以帮助明确介引论元成分的语义，因此希望在质量高的高频搭配中，介词“在”参与的搭配出现占比相对较少。

根据频次和介动紧密度自高到低进行排序后，介词“在”在TOP1000中的累计出现次数见下图2：

Figure 2: “在”在不同指标下的Top1000的数量分布

从图2看出，相对于按照频次来进行搭配度量，按照介动紧密度的分布中，介词“在”在Top1000中的占比远低于按照频次来度量。这也达到了我们的预期期望。说明按照介动紧密度进行搭配度量是一种更为合适的方法。

4.4 搭配存储

根据齐普夫定律（Zipf's Law），对自高到低的按照介动紧密度排列的搭配数据，选择出现次数占到总搭配实例数85%的部分作为最终的介动搭配实例数据。

分别以介词和动词为索引，来进行数据的存储。其具体的数据存储形式为：

.PV_跟	
(	
[	
有关	0.747
无关	0.555
商量	0.376
见面	0.275
类似	0.272
说话	0.253
吵架	0.229
合作	0.222
求教	0.190
沟通	0.170
相处	0.169
作对	0.169
结婚	0.160
握手	0.157
分手	0.152

.PV_处理	
(	
[	
依照	0.117
依据	0.033
按照	0.017
按	0.013
经由	0.010
用	0.008
经过	0.007
以	0.007
由	0.007
通过	0.006
被	0.004
替	0.003
因	0.003
凭	0.003
在	0.003

Figure 3: 介动搭配的存储形式

5 搭配库分析

5.1 数据规模

介动搭配按照介动紧密度取高频的85%之后，共保留了16033条搭配。参与构成搭配的介词种类数为38类，动词种类数为6159。

搭配数	16033
介词(Type)	38
动词(Type)	6159

Table 6: 介动搭配的数据规模

对于38类介词，按照搭配动词的数量，自高到底进行排列，Top10的其分布见下图4所示：

Figure 4: 介词搭配的动词种类数分布（Top10）

5.2 类别分析

综合考虑介动搭配的强弱和介动凸显的论元角色因素，通过对16033条实例的抽样观察，可以将搭配根据上述两个因素分为以下几种类型，具体如表7所示。

动介搭配强度	核心/非核心论元	论元角色	例子
强	核心论元	主客体	被*称为
			把*当作
		邻体	为*着想
			为*赋能
			为*保守
			向*传播
			向*请教
	非核心论元	原因	因*闻名
		原因	因*闻名
		依据	以*见长
弱		处所/时间	在*打球

Table 7: 介动搭配的类型

首先，根据介动搭配的强度可以将搭配分为两种类型，一种为强搭配，另一种为弱搭配。这种强弱的划分一方面考虑到了频次和介动紧密度度量值的因素，更重要是从语义上，看介动介引论元角色来划分，通常来说能够介引动词主客体和邻体的介动搭配，其介动搭配强度更强。其次，对于非核心论元角色来说，其内部有巨大差异。例如，对于“因*闻名”来说，虽然介引原因论元角色，但是原因论元角色是动词“闻名”语义框架完整的不可或缺成分。同样的情况也适用于“以*见长”。然而，对于“在*开会”来说，其紧密程度就跟“因*闻名”和“以*见长”比起来，有一定的弱化。

能够介引邻体的介动搭配中，发现动词的价位存在区别。“着想”为1价动词，“请教”为3价动词，因此下一小节对介引邻体角色的介动搭配作专门分析。为叙述方便，因为“p*v”形式正如一个框式，所以将凸显邻体论元角色的搭配中的动词称作“邻体框式动词”。

5.3 “邻体框式动词”分析

通过分析，“邻体框式动词”根据配价数量的不同可以划分为以下几种类型:

(1)“准二价”型

这种动词主要指那些在谓词框架语义角度来看涉及到两个成分，但从语义向句法结构的投射过程中，其中的一个语义成分可以不用其他词进行介引直接出现在句子的表层句法结构中，而另外的一个语义成分，则往往通常借助介词才能够出现在句子的表层结构中。

A 我国90%以上有害气体排放都与煤的燃烧有关。

B 向被日本侵略的亚洲国家赔罪并进行战争赔偿。

(2)“二价”型

该类别动词主要为二价动词，即其语义涉及两个主要成分，其句法投射的时候有多种选择，其中有两条倾向性的选择，一种即分别充当谓语动词的主宾语成分，另外一种则是一个成分充当主语，另外一个成分靠介词介引出现在动词的状语位置上。

A 数字奥运将为信息化发展助力。

B 这同样会导致向普遍贫困复归。

(3)“准三价”型

该类动词和“二价”动词相比，从语义上能够关涉一个更多的成分。与“三价”动词比起来，在句法上却不如三价动词那样的灵活，集中表现在其一个成分只能出现在状语位置上，不能与三价动词一样，可以将两个非主体成分投射在动词后边做宾语。

A 还有一些性急的朋友向人们传播“中国很快会胜”的盲目乐观情绪。
B 我和丈夫想为家中老人分担一些节日前繁重的家务劳动。
(4)“三价”型
三价动词主要为语义上关联三个主要成分，在句法形式上主体外的两个成分可以出现在宾语位置的动词。一般情况下，其中的非直接宾语可以由介词介引出现在动词状语的位置上。
A 鲍托正在向新仓卫生院医护人员传授B超检查技术。
B 新委员们就向老委员们请教“开会经验”。
通过穷尽性考察，确定了2118个邻体框式动词并对其进行了价位的语法测试，标注了相应的价位[2]，其加工示例如表8所示：

动词	价位
叮嘱	3
传播	2.5
赋能	2
着眼	1.5

Table 8: 框式动词的价位分类

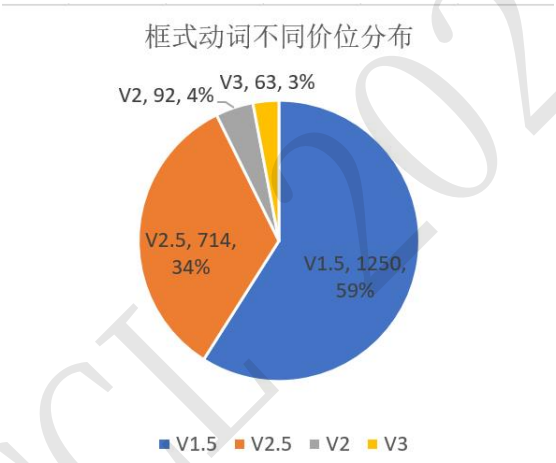

Figure 5: 框式动词不同价位分布

从上图可以看出，介引邻体的介动搭配中的动词主要为准二价和准三价的动词，这主要是因为其语义层面的某个成分投射到表层的句法结构时，只能够通过介词的介引出现在状语位置上，此时介词与核心谓词高频共现，形成稳定的前框式结构。

6 应用分析

语言研究有两个面向，一个是面向人，例如服务于语言教学；另外一方面，服务于机器，语言研究应该以助力智能语言理解为自己研究导向。相应而言，语言资源的构建也要以服务于人和机器为自己的目标，也是在服务于人和机器的过程中，衡量和评价语言资源的价值。
（1）介动搭配词典能够有效助力语言教学研究。
搭配教学是汉语教学中重要的一个环节。综合考虑汉语中虚词使用频次高，介词语义复杂的情况，介词参与的搭配研究应该引起足够的重视。与此同时，众多持动词中心说的学者（吕叔湘，1987；范晓，1987；吴为章，1994）认为动词是句子的中心和重心，把握句子动词对于理解语义起到提纲挈领的作用。因此，介动搭配的研究对汉语研究及教学来说都是重中之重，本介动搭配知识库能够为语言研究及教学提供足够的教学案例支持。

[2]为了方便进行动词价位的标注及统计，这里做了如下对应，1.5代表准二价，2代表2价，2.5代表准3价，3代表3价。詹卫东(2004)也指出，x元动词中x的取值不只是为0/1/2/3这样的整数，也能为1.5和2.5这样的小数。

（2）介动搭配词典能够为自然语言处理任务赋能。

介动搭配数据库能够为语义角色标注、信息抽取和句法语义分析提供数据支持。首先，汉语中的语义角色大部分是靠介词介引的，出现在介动之间，如何确定介引论元的语义角色是语义角色标注中的重要研究工作，通过我们的研究发现介动二元搭配对于自动凸显语义类型具有重要的作用。其次，事件抽取是自然语言处理中比较热门的研究，以往都是采用深度学习方法进行事件抽取。其对大规模高质量带标注语料依赖严重。介动搭配数据可以作为专家知识，利用远程监督方式自动构建带标注数据。最后，介动搭配的抽取也符合“大词库小规则”的句法语义分析研究范式。

7 结语

从研究侧面来看，本文关注到了以往介词研究中没有引起足够重视的介动搭配研究，并根据汉语介词发达和动词中心的特点，强调介动搭配研究的重要性，并进行介动搭配的知识获取。从研究方法上，相较于以往的基于词性字符的线性正则表达式匹配查询，本研究在句法结构树语料上，利用短语块属性和位置信息限制，进行介动搭配知识抽取，噪音少和准确率高。并借鉴Tf-Idf思想提出了介动搭配的度量方法，初步的实验证明该方法相较于绝对频次具备更好的表现。

本研究后续的改进方向有下：（1）将其应用到具体的自然语言处理任务中，例如事件抽取中，来进一步衡量验证目前知识库的价值。（2）探索单音节动词与介词的搭配研究及知识获取。

本研究的数据将会寻求合适方式与学界共享，推进相关研究工作。

参考文献

Y. Chen, L. Xu, K. Liu, D. Zeng, and J. Zhao, 2015. Alternation. *Event extraction via dynamic multi-pooling convolutional neural networks*, Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics 7th Int. Joint Conf. Natural Lang. Process., 2015, pp. 167–176.

胡韧奋, 肖航. 2019. 面向二语教学的汉语搭配知识库构建及其应用研究. 语言文字应用,2019(01):135-144.

范晓. 1990. 介宾短语·复指短语·固定短语. 北京：人民教育出版社.

张谊生. 2000. 现代汉语虚词. 上海：华东师范大学出版社.

陈昌来. 2002. 介词与介引功能. 合肥：安徽教育出版社.

刘丹青. 2002. 汉语中的框式介词. 当代语言学,2002(04):241-253+316.

张旺熹. 2004. 汉语介词衍生的语义机制. 汉语学习,2004(01):1-11.

何洪峰, 张文颖. 2016. 汉语动介并行现象. 2016,36(04):21-27.

刘静敏. 2015. 动词介词化的句法语义机制. 语文建设,2015(21):34-35.

俞士汶. 2015. 《现代汉语广义虚词知识库的建设》. 《汉语语言与计算机学报》2003年第1期.

邢丹, 饶高琦, 荀恩东, 王诚文. 2020. 基于大规模语料库的介词结构搭配库构建. 中文信息学报,2020,34(11):1-8.

彭爽,俞士汶. 2009. 现代汉语介词知识库的建设. 社会科学战线,2009(08):247-249.

方清明. 2017. 现代汉语介词用法词典. 北京：商务印书馆.

邹宏梅, 王挺. 2007. *SVM*和基于转换的错误驱动学习相结合的汉语组块识别. 计算机工程与科学,2007(04):91-94+123.

高红. 2007. 基于统计语言模型的汉语浅层分析研究. 大连理工大学.

王莹莹. 2006. 汉语组块识别的研究. 大连理工大学.

詹卫东. 2004. 论元结构与句式变换. 中国语文,2004(03):209-221+286.

俞士汶, 朱学锋, 王惠, 张芸芸. 1996. 现代汉语语法信息词典规格说明书. 中文信息学报,1996(02):1-22.

吕叔湘, 朱德熙. 1952. 语法修辞讲话. 北京:中国青年出版社.

范晓. 1987. 汉语动词概述. 上海：上海教育出版社.

吴为章. 1994. “动词中心”说及其深远影响——《中国文法要略》学习札记. 语言研究,1994(01):10-20.