文章编号：1003-0077(2018)10-0019-09

《现代汉语动词语义知识词典》的开发与应用

孙道功[1]，亢世勇[2]

（1. 南京师范大学 文学院，江苏 南京 210097；2. 鲁东大学 文学院，山东 烟台 264025）

摘　要：该文吸收已有动词研究的相关成果，提出了动词语义词典开发的相关原则和研制思路，界定并描写了词典中所涉及的相关属性信息，并对词典的总体文件结构及其各个库的信息进行了描写和说明。最终开发了融合词汇语义和句法语义，涵盖词形、词性、释义、义类、义场、句法范畴信息、语义范畴信息、语义句模等多种信息参数的开放性的动词语义知识词典。该词典可以在歧义分化、词义关系考察、句法—语义接口、句模抽取等方面提供支持。

关键词：动词语义知识库；语义词类；句法语义范畴

中图分类号：TP391　　**文献标识码**：A

Development and Application of *Modern Chinese Verb Semantic Knowledge Dictionary*

SUN Daogong[1], KANG Shiyong[2]

(1. School of Chinese Language and Literature, Nanjing Normal University, Nanjing, Jiangsu 210097, China;
2. School of Chinese Language and Literature, Ludong University, Yantai, Shandong 264025, China)

Abstract: Based on the existing researches on verbs, this paper puts forward principles and ideas of developing verb semantic dictionary, defines explains the attribute information involved, and explains the overall file structure and each library. An open verb semantic knowledge dictionary is finally constructed, covering both lexical meaning and syntactic meaning, including morphologies, word classes, paraphrases, word meanings, semantic fields, syntactic category information, semantic category, semantic pattern. The dictionary provide support for ambiguity interpretation, lexical relation research, syntax-semantics interface, semantic pattern extraction, etc.

Keywords: verb semantic knowledge base; classification of lexical meaning; syntactic-semantic category

0　引言

自然语言处理的迅速发展，不仅召唤语言研究向技术化层面延伸，同时也进一步凸显了语义分析的重要性和迫切性。众所周知，语义知识是语言信息处理的难点。如何解决语义问题，如何为计算机的理解生成提供可形式化、可计算化的语义网络，成为语义研究的核心[1]。从 20 世纪 60 年代欧美语言学实现从语形研究到语义研究的历史性转向后，越来越多的学派和学者开始关注语义问题。尤其从 20 世纪 80 年代中期开始，为了克服语言处理中普遍存在的“语义障碍”(semantic barrier)，越来越多的国家开始开发语义词典。

1　语义词典研究评述

语义词典作为自然语言处理系统的重要组成部分，为语言处理提供语义资源，目前比较有影响的语义知识词典，国外如 WordNet、MindNet、FrameNet 等；国内如《同义词词林》、知网(HowNet)、《现代汉语语义词典》(SKCC)、汉语框架语义知识库

收稿日期：2018-01-04　**定稿日期**：2018-05-18

基金项目：国家社会科学基金(12CYY052)；教育部人文社科青年基金(14YJC740077)；山东省语言资源开发与应用重点实验室项目(1311023)

(CFN)等。

国内外语义词典研究把“语义关系”作为描写重点。作为20世纪80年代后国外语义词典的重要代表,WordNet的特色表现为根据词义关系而不是单纯词形标记组织词汇信息。具体言之,首先基于词义关系,把名词、动词和形容词聚类为代表某一基本词汇概念的同义词集合,然后在这些同义词集合之间建立语义关系。目前,WordNet已对95 600个不同的词形(51 500个简单词和44 100个搭配词)进行了分析,形成了70 100个词义集合(或者说同义词聚类)[2]。基于语义分类构建聚类系统以及语义关系构建关联体系这一做法,成为国外语义知识词典建构的重要方法。FrameNet是以框架语义学理论为基础,以英语真实语料为依据,涵盖1 007个语义框架11 797个词的在线语义词典,目前已经对近7 000个词、9 000多个核心框架元素和30多个外围元素进行了注释和描写。作为一个在线的词典编纂工程,在对语义框架、框架元素、句子语义标注体系处理方面富有特色,尤其是其研制思路和理念,对国内语义知识库构建产生了很大影响[3]。MindNet的特色表现为完全采用自动的方式来获取语言知识,其理论基础仍然是语义关系,库中共定义了24种不同的语义关系标记,分析了近16万个词。在技术层面,仍然是基于规则的方法,采用广域句法分析器(Broadcoverage Parser)获取语义信息,其根本目的是建立一个范围广泛的自然语言理解系统[4]。

国内语义词典编纂历史悠久,秦汉时期的《尔雅》是世界上最早的义类词典。20世纪80年代之后,以《同义词词林》为先导,国内出现了多种不同的语义知识词典。《同义词词林》作为国内最具影响力的语义词典之一,通过对50 000多词语约67 000义项进行整理分析,分为三个层级,其中大类12个,中类94个,小类1 428个。并在小类下列出所对应的词语(或义位)[5]。《同义词词林》所建构的语义分类体系被国内语义词典编撰者参考或模仿。

董振东等人开发的知网,其实质是基于英汉词语所代表概念的描写,揭示概念之间的相互联系的在线语义知识库。根据词语的语义特征,在概念分类和关系描写的基础上,使语义信息形成了相互关联的知识网络系统。概念系统涵盖万物、时间、空间、属性、属性值、事件、部件七大类。但在语义分类上仍然兼顾了对应的词性信息,大致对应情况是:实体、属性、单位对应名词;事件对应动词和部分形容词;属性值对应形容词和副词。分别称为N范畴、V范畴、A范畴。通过概念、属性等纵横交错关系最终形成一个网状知识系统[6]。《现代汉语语义词典》(SKCC)在大类的划分上采用词性标准,小类上采用了语义标准,共收实词66 539条,并以数据库的形式进行呈现,包括12个数据库:1个总库,11个子库,分别是:名词、时间词、处所词、方位词、代词、动词、形容词、区别词、状态词、副词、数词;但是词典中没有设立虚词库。总库中包括词语、拼音、同形、义项、语义类、词类、子类、兼类八个字段。每类词的特有属性填在各类词库中,如名词库设15个属性字段,动词库设16个属性字段。每个库文件都详细刻画了词语及其语义属性的二维关系,最终目的是为计算机语义自动分析、词义消歧等提供支持[7]。该词典中的语义分类主要参考了WordNet的分类体系,在大类划分上仍然基于词性角度。另外,刘开瑛等人以框架语义学理论为基础,以FrameNet为参照,构建了汉语框架语义知识库(CFN)。CFN数据库由框架库、句子库和词元库三部分组成。目前已构建了130个框架,涉及动词词元1 428个、形容词词元140个、事件名词(即有配价的名词)词元192个,句子8 200多条[8]。其特色表现为结合汉语特点,把词元库和句子库结合起来,并不是对FrameNet的简单汉化。

综上所述,目前语义词典尤其是国内语义词典编撰存在的问题,主要表现为四个方面:①多数仍停留在词语的语义分类层面,且分类依据一般是哲学或逻辑,通常以词性标准为纲、词义分类为辅,并不是完全意义上的语义分类。②在语义分类后仅列出符合某一义类的词语,缺乏对内部成员的分析描写,尤其缺乏对不同义类成员的语义关系和语义差异的深度刻画。③有些词典虽然增设了词汇语义关系的分析说明,但尚未对所收录词语的语义进行多维度刻画,尤其是缺乏句法语义信息的深度描写。④大都着眼于传统的词汇语义视角,尚未对批量词汇进行词汇语义和句法语义的一体化描写,也未揭示其内在关联性。本文基于受限原则,先以少量词汇为典型样本,构建语义词典,解决上述存在的问题。

2 收词原则和研制思路

《现代汉语动词语义知识词典》(简称“词典”)实

质上是一个在线语义知识库,与常规词典的不同之处主要表现为:通过对批量动词的词汇语义和句法语义的标注,揭示了两者的内在关联,从而实现了对词汇语义和句法语义一体化的分析和描写,为语义形式化研究和语言信息处理提供语言资源。动词语义词典研制的首要任务就是要选取具有代表性的常用动词作为典型分析对象,在此基础上进行相关信息的标注和描写。

2.1　收词原则

结合本词典的相关特点,确定了以下收词原则[9]。

第一,典型原则。典型原则指词典所收录词汇应该具有代表性和权威性,使用频度和熟知度高,是目前大部分动词类辞书收录词汇的交集部分。基于《现代汉语词典》(商务印书馆第 7 版,2016)、《现代汉语频率词典》(北语语言教研所,1986)、《现代汉语动词大词典》(林杏光等,1994)、《现代汉语动词分类词典》(郭大方,1994)等筛选出交集部分的动词。

第二,广布原则。广布原则指词典所收录词汇应该分布范围广,通行于各个领域,不应该仅适用于某一特定领域或特定人群。这与典型原则有一定相似之处,但又有差异。典型原则强调使用率,即使用频度高;广布原则侧重分布率,即使用领域广。针对某些词表在语域方面的局限性,选词时会多方面兼顾,把多个语域中广泛使用的词语吸收进来,提高覆盖率。

第三,单义原则。单义原则指词语选择和词义描写时,以词元为单位。词元是按照一形一音一义对应原则对词语进行分化的结果,一个词元在语义上仅对应一个能够独立使用的义项。故包含多个独立运用的义项的词语,可以分化为多个词元,分别用 A1、A2、A3……表示。之所以使用词元对词语进行分化,一方面,同一词形对应的多个词元,其使用率和分布率并不相同,以词元为单位可以使词义描写更加精细化;另一方面,同一词语分化形成的多个词元,在语义搭配、语义句模、“句法—语义”接口等方面的表现也大相径庭。

基于以上原则,进行筛选并确定词典的收录对象。到目前为止,共选取 6 000 个词元作为词典分析对象。

2.2　研制思路

具体研制思路如图 1 所示。

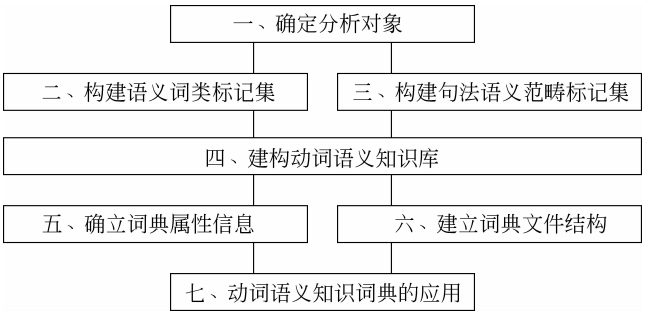

图 1　语义词典研制路线图

3　属性信息

为了满足语义形式化和中文信息处理的需要,对所涉及范畴标注时尽量采用字母标记形式。与以往的动词语义词典相比,本《词典》设置的属性信息更为丰富,不仅涉及词汇语义层面的常规信息,还涉及句法语义层面的信息,以及词汇语义与句法语义的内在关联信息。具体如下:

(1) 常规信息,指词典中所收录词元的拼音、声调等信息。其中四声分别用“1,2,3,4”表示,例如,“吃”是“chi1”。如果是轻声,则用“5”表示。

(2) 词类信息,指词元对应的词性信息。按照北京大学计算语言所的语法词典的词类标准和标记符号进行描写。因为本文构建的是动词语义知识词典,分析对象中的词类仅涉及动词一类,即动词(V)。

(3) 释义信息,指某词元在《现代汉语词典》中对应释义。虽然属性信息中包含了义类信息字段,但是二者并不完全一样。其内在关联主要表现为需要依据释义信息来确定词汇义类。

(4) 义类信息,指某词元所属的语义类,如动物、植物、人类等。与词元对应的释义信息不同,义类信息着眼于词元所属的上位语义范畴。由于本文所开发的动词语义词典,其目的之一是对词汇语义和句法语义进行一体化描写,涉及语义框架的描写,所以在句子标注中不能仅仅考察动词义类。实际标注中涉及名词、形容词等非动词的义类信息。该义类标记集包括 10 大类 32 小类,其中动词(陈述类)的义类信息共涉及 7 类。括号内为其语义类型和标记符号。如表 1 所示。

表 1 词语义类信息表

指称类(4)	陈述类(7)	描述类(6)	限定类(5)	询问类(1)	标记类(2)	关联类(1)	情态类(2)	语气类(1)	呼应类(3)
名称(mc) 称代(cd) 空间(kj) 时间(sj)	施动(sd) 祈使(qs) 参加(cj) 遭致(zz) 存有(cy) 施感(sg) 判断(pd)	性质(xz) 状态(zt) 声状(sz) 情状(qz) 方式(fs) 趋向(qx)	指别(zb) 数量(sl) 类别(lb) 计位(jw) 品类(pl)	询问(xw)	介引(jy) 助构(zg)	关联(gl)	时体(st) 评估(pg)	语气(yq)	呼应(hy) 寒暄(hx) 感叹(gt)

大规模的义类标注是规模浩大的语言工程。受时间、精力等多方面条件的制约,目前义类标注还停留在二级层面,共标注了 32 个小类。三级小类标注是下一步研究的重要任务。

(5) 义场层级信息,指从词汇语义层级的角度,从高层到低层分别列出某词元的上下位的语义关系图。义类分析和义场建构是互动的过程,基于所收录 6 000 个词元构建了 251 个义场。

(6) 语义范畴信息,指句子中语块所对应的语义信息,包括核心范畴、角色范畴、情态范畴、超句范畴,目前超句范畴暂不标注。

具体标注时,以语块为单位,标到语义体系的第二层级。为了便于统计和减少角色符号的重码率,标注中所涉及范畴也都采用了对应汉字拼音的首字母来表示。在同一大类中,如果首字母重合,会采用音节的第二个字母表示,如果依然重合,再采用第三个字母。语义范畴信息包括动核八类 19 种,基本角色九类 32 种,附加角色六类 26 种,共计 77 种。

动核包含的类型及标记符号如表 2 所示。

表 2 动核类型信息表

施动核(3)	祈使核(1)	参与核(2)	遭使核(3)	存有核(3)	感受核(1)	评判核(3)	性状核(3)
自动(HZD) 协动(HXD) 施言(HSY)	祈使(HQS)	加入(HJR) 担任(HDR)	遭遇(HZY) 致使(HZS) 致变(HZB)	存现(HCX) 变化(HBH) 领有(HLY)	感知(HGZ)	判断(HPP) 比喻(HBY) 评价(HPJ)	性质(HXZ) 状态(HZT) 显示(HXS)

基本角色范畴包含的类型及标记符号,如表 3 所示。

表 3 基本角色类型信息表

施动类(8)	祈使类(2)	参加类(3)	遭致类(4)	存有类(4)	感知类(2)	性状类(3)	评价类(2)	论断类(4)
施事(JS)、 共事(JGS) 与事(JYS)、 言事(JYH) 受事(JSS)、 成事(JCH) 所变(JSB)、 所言(JSY)	祈事(JQS) 所祈(JSQ)	任事(JRS) 所加(JSI) 所任(JSR)	遭事(JZS) 所遭(JSA) 致事(JZH) 所使(JSH)	变事(JBS) 变果(JBG) 领事(JLS) 所隶(JSL)	感事(JCH) 所感(JSG)	系事(JXS) 当事(JDS) 比事(JBH)	评事(JPS) 所评(JSP)	断事(JDH) 所断(JSD) 喻事(JUS) 所喻(JSU)

注:①表 3 中,因为施事和受事首字母重合,为了区分,施事使用了 JS,受事使用了 JSS。②表 3 和表 4 中,成事与处所、遭事与致事、所加与时间、受事与所使、共事与感事、变事与比事、当事与断事、涉者与所遭,首字母相同。其中处所、遭事、时间、受事、共事、变事、当事、涉者仍然采用音节首字母表示,而成事、致事、所加、所使、感事、比事、断事、所遭分别采用前音节首字母加后音节第二字母来表示,其中"J"表示角色。

附加角色范畴包含的类型及标记符号如表 4 所示。

情态范畴包括时体(TST)和评估(PPG)两类。时体表示事件中动作行为的开始、进行、持续或完成等。评估表示对事件中所发生的动作行为推测、估计、评价、强调等。目前暂时标注到时体、评估大类层面。

(7) 句法范畴信息,指动核及关联成分对应的句法成分信息。虽然所要建构的是语义词典,但是语义范畴信息的标注以语块为单位。同时句法范畴与语义范畴信息是密切关联的,开发本语义词典重要目的之一是为"句法—语义"接口的研究提供平台和语言资源,故在信息库中仍然保留了句法信息。包括主语、谓语、宾语、状语、补语。定语通常和后面的中心语作为一个语块承担某种句法成分或语义角色,所以不分开标注。

表 4　附加角色类型信息表

对象类(10)	场合类(3)	伴随类(4)	限定类(3)	因缘类(4)	凭借类(2)
感者(JGZ)、比者(JBZ) 归者(JGE)、对者(JDZ) 涉者(JSZ)、替者(JTZ) 向者(JXZ)、专者(JZZ) 除者(JCZ)、门类(JML)	方位(JFW) 处所(JCS) 时间(JSJ)	情状(JQZ) 方式(JFS) 结果(JJG) 趋向(JQX)	计量(JJL) 范围(JFV) 程度(JCD)	依据(JYJ) 条件(JTJ) 原因(JYY) 目的(JMD)	工具(JGJ) 材料(JCL)

注:因为归者采用两个音节首字母与感者重复,采用第二字母又会与感事重合,所以采用前音节首字母和后音节第三个字母的组合形式。

(8) 句模信息,即句子对应的语义结构信息。根据语义知识库中所标注的句法、语义范畴信息抽取某动词词元形成的句子语义模型,也是动词语义词典语义信息描写的重要组成部分。如 JS+HXD+JSS,指施事+协动核+受事。

(9) 义类与语义范畴对应关系信息,指某词元所属义类与语义范畴的内在关联。基于语义知识库提取动词词元关涉语义范畴所对应的词元信息,考察其义类,建立词元义类与语义范畴的对应关系模型。

4　文件结构

4.1　收词原则

《词典》采用关系数据库技术,在 Access 下实现。文件中信息都尽量地用汉字表示。根据研究需要共设置了三个库。其中总库一个,另外两个分别是:词汇义类信息库、句法和语义范畴信息库。这三个库通过"词汇、拼音"字段链接。其中总库中包含了其他两个库的义类、语义范畴和句法成分标注信息。该词典具有开放性,计划先收录 10 000 个词元,目前已经收录并分析 6 000 个。

4.2　库文件的结构及属性的描述

4.2.1　总库的文件及属性描述

总库的具体属性字段、字段宽度、属性值,以口部动作词"吃"为例,具体描述如表 5 所示。

4.2.2　词汇义类信息库文件结构及属性描述

该库包含四个部分:词类信息、释义信息、义类信息、义场层级信息。词类信息和释义信息如总库中结构信息表 5 中所述,不赘。义类信息相对简单,即某词元对应的《语义词类标记集》中的所属类型。义场层级信息比较复杂,对词典中所收录词元,库文件中会分层级列出所属的义场信息。同一义类动词的义场层级信息相似度高。如"动作"大类中的手部动作义场的四个词元"打$_2$(殴打)、拿、指、托"对应的义场层级信息,如图 2 所示。

表 5　总库文件结构信息表

属性字段	字段宽度	属性值
词元信息	30	填写词典中收录词元的拼音信息,"吃$_1$",拼音为 chi1
词类信息	2	填写词元的词性标记。依据北大语法信息词典的标注标记,即"v"。
释义信息	100	填写《现代汉语词典》(第 7 版)对应词元的释义信息。即"把食物等放到嘴里经过咀嚼咽下去"。
义类信息	10	填写词元所属的义类信息。即"施动(sd)"。
义场层级信息	50	填写词元所属的义场层级信息,从上到下分别列出所属的义场信息。即"动作—口部—牙齿"
语义范畴信息	10	填写词元在语句中所属的动核范畴信息及所关涉语义角色信息。如:我[JS]吃【HXD】饭[JSS]。为了统计方便,所有角色范畴用"[ ]",动核范畴用"【】"表示。
句法范畴信息	10	填写词元在语句中对应的句法范畴信息。包括五种类型:主语 S;谓语 V,宾语 O,状语 D;补语 P。"吃$_1$"的句法范畴信息标注为"V"。
句子标注信息	500	填写包含某动词词元并且标注了义类信息、句法语义范畴信息的句子。如{ S 你/rc} [JS]{ D 凭/jy 什么/sw} [JYJ]{V 吃/xd}【HXD】{O 白面/sw 馒头/sw} [JSS]? 其中"/"后面的是义类信息;"{}"后面的是句法成分信息;"[ ]"中的是语义角色信息;"【】"中的是动核信息。
句模信息	100	JS+JYJ+HXD+JSS,含义是施事+依据+协动核+受事,即上例,这是"吃$_1$"对应的句模形式之一。

第一层	第二层	第三层	第四层	第五层
打$_2$	动作	人类	上肢	手部	整手
拿	动作	人类	上肢	手部	手指
指	动作	人类	上肢	手部	手指
托	动作	人类	上肢	手部	手掌

图2 义场层级图示例

4.2.3 句法和语义范畴信息库的文件结构及属性描述

该库包含所收录的动词词元以及带有句法成分和语义范畴信息的句子实例。其中,句法成分包括S/V/O/D/P。语义范畴信息相对比较复杂,包括动核、角色和情态,具体信息如词典属性信息部分所述。

在此选取了现代汉语非常复杂的手部动作词“打”为例。“打”作为典型的动作动词,其义项多达24个,其中最高频义项是“打$_2$”(殴打)。该词元对应了43种句模,43种句法语义对应关系模式。其中原型句模是JS+HXD+JSS;原型句法结构是S+V+O。在句法和语义范畴信息库中提取的相关例句,具体如下:

1.{V打/xd}【HXD】{O他/cd}[JSS]啊/yq!

2.{D三/sl}[JJL]{V打/xd}【HXD】{O白骨精/mc}[JSS]。

3.{D棒/mc}[JGJ]{V打/xd}【HXD】{O鸳鸯/mc}[JSS]。

4.{D莫/pg}(PPG){V打/xd}【HXD】{O笑脸/mc人/mc}[JSS]!

5.{D按/jy军规/mc}[JYJ]{D要/pg}(PPG){V打/xd}【HXD】{O他/cd}[JSS]{O军棍/mc}[JJL]。

6.{D一/sl棒/mc[JGJ]{V打/xd}【HXD】{P死/zz}[JJG]了/st{O妖精/mc}[JSS]!

7.{D由于/jy不/pg小心/zt}[JYY]{V打/xd}【HXD】{P破/xz}[JJG]了/st(TST){O水银/mc温度计/mc}[JSS]。

8.{V打/xd}【HXD】{P死/zz}[JJG]{O侵略军/mc 400/sl多/sl人/mc}[JSS]。

9.{S他/cd}[JS]{D把/jy小三/mc}[JSS]{V打/xd}【HXD】了(TST)!

10.{S他/cd}[JS]{D把/jy人/mc}[JSS]{V给/jy打/xd}【HXD】{P死/zz}[JJG]{O一/sl个/jw}[JJL]?

11.{S凶残/xz的/zg敌人/mc}[JS]{D把/jy这个/zb青年/mc}[JSS]{V打/xd}【HXD】{P晕/zt}[JJG]了/st(TST)!

12.{S敌人/mc}[JS]{D把/jy他/mc}[JSS]{P往/jy死/zz里/kj}[JCD]{V打/xd}【HXD】。

13.{S你/cd}[JS]{V打/xd}【HXD】{O我/cd}[JSS]啊/yq!

14.{S林冲/mc}[JS]{D棒/mc}[JGJ]{V打/xd}【HXD】{O洪教头/mc}[JSS]。

15.{S外婆/mc}[JS]{D只/pg}[JFV]{V打/xd}【HXD】{O淘气/xz的/zg哥哥/mc}[JSS]!

16.{S你/cd}[JS]{D凭/jy什么/zb}[JYY]{V打/xd}【HXD】{O他/cd}[JSS]!

17.{S他/cd}[JS]{D为了/jy老婆/mc}[JMD]{V打/xd}【HXD】了(TST){O警察/mc}[JSS]。

18.{S那个/zb城管/mc}[JS]{D正在/sj}(TST){V打/xd}【HXD】{O人/mc}[JSS]呢/yq?

19.{S我/cd}[JS]{D一/sl拳/jw}[JGJ]{V打/xd}【HXD】{P烂/zt}[JJG]{你/cd的/zg狗头/mc}[JSS]。

20.{S列车长/mc}[JS]{D狠狠/xz地/zg}[JFS]{V打/xd}【HXD】了/st(TST){O他/cd}[JSS]{O一/sl巴掌/mc}[JGJ]!

21.{S我/cd}[JS]{V打/xd}【HXD】{P断/zt}[JJG]{O你/cd的/zg狗/mc腿/mc}[JSS]!

22.{S他们/cd}[JS]{D不敢/pg}(PPG){V打/xd}【HXD】{O你/cd}[JSS]!

23.{S武松/mc}[JS]{D酒/mc醉/zt后/sj}[JSJ]{D在/jy景阳冈/kj}[JCS]{D赤手空拳/fs}[JFS]{V打/xd}【HXD】{P死/zz}[JJG]{O老虎/mc}[JSS]。

24.{S泰森/mc}[JS]{D狠狠/xz地/zg}[JFS]{V打/xd}【HXD】{O他/cd}[JSS]{P一/sl拳/jw}[JGJ]。

25.{S你/cd家/mc孩子/mc}[JSS]{V被/jy打/xd}【HXD】了/st(TST)?

26.{S小贩/mc}[JSS]{V被/jy打/xd}【HXD】{P死/zz}[JJG]{P在/jy台阶/mc前/kj}[JCS]。

27.{D立即/sj}(TST){D把/jy那/zb只/jw疯狗/mc}[JSS]{V打/xd}【HXD】{死/zz}[JJG]!

28.{S他/cd的/zg右脸/mc}[JSS]{V被/jy打/xd}【HXD】{P肿/zt}[JJG]了/st(TST)!

29.{S妈妈/mc你/cd}[JS]{V打/xd}【HXD】啊/yq!

30.{S他/cd}[JS]{D很/qz重/xz地/zg}

[JFS]{V 打/xd}【HXD】{P 下来/qx}[JQX]!

31. {S 他/cd}[JS] {D 一/sl 棍子/mc}[JGJ] {D 狠狠/xz 地/zg}[JFS]{V 打/xd}【HXD】{P 过去/qx}[JQX]!

32. {S 我/cd}[JS] {D 没/pg}(PPG){V 打/xd}【HXD】啊/yq!

33. {D 敢/pg}(PPG){V 打/xd}【HXD】{P 一/sl 下/jw}[JJL]吗/yq?

34. {D 怎么/}(PPG){D 朝/jy 孩子/mc 脑瓜/mc 上/kj}[JCS]{V 打/xd}【HXD】呢/yq?

35. {V 打/xd}【HXD】{O 哪儿/kj}[JCS]呢/yq?

36. {V 打/xd}【HXD】{P 得/zg 哭/zd 爹/mc 喊/xd 娘/mc}[JJG]!

37. {D 一/sl 记/yw 重重/zt 的/zg 老/xz 拳/mc}[JGJ]{V 打/xd }【HXD】{P 得/zg 眼冒金星/zt}[JJG]。

38. {D 给/jy 我/cd}[JTZ]{V 打/xd}【HXD】!

39. {D 一/sl 电棍/mc}[JGJ]{V 打/xd}【HXD】{P 在/jy 他/cd 腰/mc 上/kj}[JCS]。

40. {D 无缘无故/pg}(PPG){V 被/jy 打/xd}【HXD】了/st(TST){P 一/sl 个/jw 多/sl 小时/sj}[JSJ]!

41. {S 老虎/mc}[JSS] {D 被/jy 武松/mc}[JS]{V 打/xd}【HXD】{P 死/zz}[JJG]了/st(TST)。

42. {S 他/cd}[JSS] {D 被/jy 一/sl 个/jw 花白/xz 胡子/mc 的/zg 人/mc}[JS]{D 用/jy 马鞭/mc}[JGJ]{V 打/xd}【HXD】{P 晕/zt}[JJG]了/st(TST)。

43. {S 嘎子/mc} [JS] {D 趁/jy 他/cd 不/pg 注意/sg} [JTJ]{V 打/xd}【HXD】{O 他/cd} [JSS]{P 一/sl 顿/jw} [JJL]。

动词词元在组合层面形成的句法结构和语义句模信息,都是基于该库中的句子实例的标注信息提取的。由于再大的语料库也无法涵盖所有的语言事实,随着语料库的扩大,手部动词“打$_2$”对应的模式类型和数量可能会有所增加,但都是基于原型模式通过添加附加角色或情态范畴递归形成的。该库为动词词元涉及的句法成分、语义范畴、句模形式的描写提供了语言资源。

4.3　总库文件样例

总库的具体词元样例,因篇幅所限,仅能部分列举分析,仍然以手部动作词“打”进行说明。“打”的24 个义项中,有些已经抽象化,属于手部动作的转义。在此仅分析与手部动作直接相关的七个具体义项,如表 6 所示。

表 6　总库文件样例信息表

编号	词元	词类	释义	义类	义类层级	语义范畴信息	句法范畴信息	实例	句模信息
1221	打 1	v	用手或器具撞击物体	sd	动作|上肢|整手	【HXD】	V	{S/孩子们 rc}[JS] {D 在现场}[JCS]{D 不停地}[JFS]{V 打/xd}【HXD】{O 鼓/sw}[JSS]。	JS+JCS+JFS+HXD+JSS
1222	打 2	v	殴打	sd	动作|上肢|整手	【HXD】	V	{S 他/rc}[JSS]{D 上/xd 学/sw 路上/kj}[JCS]{D 被/jy 人/rc}[JS]{V 打/xd}【HXD】了/zg(TST)。	JSS+JCS+JS+HXD+TST
1223	打 3	v	搅拌	sd	动作|上肢|整手	【HXD】	V	{S 她/rc}[JS]{D 亲自/fs}[JFS]{V 打/xd}【HXD】{P 好/xz}[JJG]了/zg(TST){O 鸡蛋/sw}[JCH]。	JS+JFS+HXD+JJG+TST+JCH
1224	打 4	v	捆	sd	动作|上肢|整手	【HXD】	V	{S 他/rc}[JS]{D 一/sl 大早/sj}[JSJ]{D 匆匆/fs}[JFS]{V 打/xd}【HXD】{P 好/xz}[JJG]了/zg(TST){O 铺盖卷/sw}[JCH]。	JS+JCJ+JFS+HXD+JJG+TST+JCH
1225	打 5	v	编织	sd	动作|上肢|整手	【HXD】	V	{S 你/rc}[JS]{D 用/jy 这/zb 毛线/sw}[JCL]{V 打/xd}【HXD】{O 件/lb 毛衣/sw}[JCH]吧/yq!	JS+JCL+HXD+JCH
1226	打 6	v	凿开	sd	动作|上肢|整手	【HXD】	V	{S 它们/sw}[JS]{D 春天/sj}[JSJ]{D 在/jy 树干/sw 上/fw}[JCS]{V 打/xd}【HXD】{O 洞/sw}[JCH]。	JS+JSJ+JCS+HXD+JCH

续表

编号	词元	词类	释义	义类	义类层级	语义范畴信息	句法范畴信息	实例	句模信息
1227	打7	v	举	sd	动作|上肢|整手	【HXD】	V	{S前排/kj 的/zg 小朋友/rc}[JS]{D都是/pg}[JFV]{D一/sl 人儿/rc}[JFS]{V打/xd}【HXD】{O一/sl 小伞儿/sw}[JSS]	JS+JFV+JFS+HXD+JSS

5 主要应用

与以往的语义词典相比,本词典的主要特点是对词汇语义和句法语义信息进行一体化描写,不仅标注了动词词元的义类信息,同时给出了在组合层面关涉的句法语义范畴以及形成的语义组合模式,为词汇与句法语义关系的描写,尤其是"句法—语义"接口研究提供了平台和语言资源。

首先,《词典》所标注词汇语义和句法语义信息,可以应用于词汇语义计算。词汇语义计算包括相关度计算和相似度计算两种类型。相似度着眼于词汇相互替换但不改变句法语义结构。相关度虽然涵盖了相似度的概念,但二者并不完全一致。目前学界对相关度的研究较少。基于《词典》中标注的义类知识和义场层级信息,可以计算同一义场词元的语义相似度,也可以计算不同义场词元的语义相关度。词汇语义计算的相关数据可以服务于信息检索、词义消歧、文本分类以及文本聚类等方面。

其次,《词典》中标注的句法成分、语义角色以及句模等信息,可以服务于语义关系的自动获取。目前获取方法主要有基于统计的机器学习方法或基于语言组合特征的关系获取算法等[10]。自然语言处理领域的语义关系有不带标记和带标记两种类型。前者通常基于同现统计的方法获得,只能表明词语之间存在关系,却不能体现是何种关系;后者能体现出词语存在关系以及何种关系。本《词典》标注的丰富的句法语义信息,尤其是组合中的语义范畴和语义关系类型,可以服务于语义关系的自动获取,从而呈现出带有标记的语义关系。

再次,《词典》为"词汇—句法语义"的接口(或链接/衔接)研究提供支持平台。汉语中大部分句子都是以动词为中心的,基于语料库构建的动词语义知识词典,对词汇语义和句法语义进行了一体化描写,为探讨"词汇—句法语义"的接口提供了基础。具体思路是基于动词语义词典中所标注的词汇语义和句法语义信息,考察词汇单位实现为语义范畴,尤其是语义角色的机制、语义角色的排序机制、语义角色句法实现机制以及语用制约机制。因为某一义类的词元类聚为同一义场,同一义场词元往往具有相同的句法表现。具体考察时以义场为单位,基于《词典》中的标注信息和统计数据,考察并得出义类与角色范畴的对应关系、角色范畴与句法成分的对应关系,以及角色范畴句法实现时与语用的制约关系。

最后,基于《词典》,开发了句法语义范畴标注工具。不仅可以对语料文本进行句法语义范畴的标注,还可以提取动词关涉的语义角色频度信息,以及所形成的语义结构信息。如基于《词典》提取的关于动词"打$_2$"(殴打)的部分语义结构信息,具体如图3所示。

图3 "打$_2$"的语义结构模式图

此外，《词典》还可以应用于：①某一词元的义类义场的提取和统计研究；②同一义场词元形成语义框架的对比研究等，不再赘述。

6　结论

综上，本文在对国内外语义词典评述的基础上，吸收动词研究的已有相关成果，提出了动词语义词典开发的相关原则和研制思路，界定并描写了词典所涉及的相关属性信息，并对词典的总体文件结构及其库的信息进行了描写和说明，并进一步指出了本词典的主要用途和应用前景。创新之处主要表现为：①词典中所确定的相关属性信息及描写方法为之后的动词语义词典开发提供了样例和参考模板；②对批量动词词元进行词汇语义和句法语义的一体化描写，为语义形式化和句法语义关系的获取提供了基础；③对常用动词词元从释义、义类、语义层级、语义关系到语义差异进行多层次深度刻画，为动词的语义分析和处理提供丰富的语义资源；④基于语义词典开发了相关的标注工具和软件，为大规模语料的句法语义标注提供了便利。

受字数等诸多因素的限制，文中仅对词典的整体框架进行展示，对于某些属性信息及关系缺乏更充分的描写和介绍。同时，动词语义知识词典的开发，需要根据研制目的，制定相应的标注规范和标注规模，其具体标注过程耗时费力，目前所开发的规模还比较小，希望在进一步的研究中扩大规模，完善标注信息，以期能够更好地服务于语义形式化和语言信息处理研究。

参考文献

[1] 孙道功. 词汇—句法语义的衔接研究[M]. 北京：世界图书出版公司，2011：1-7.

[2] Fellbaum, Christiane, et al. WordNet: An electronic lexical database[M]. MIT Press, 1998.

[3] C J Fillmore, C Wooters, C F Baker. Building a large lexical databank which provides deep semantics[C]// Proceedings of the Pacific Asian Conference on Language Information and Computation. Hong Kong, 2001:86-90.

[4] Richardson SD. Dolan WB, Vanderwende L. MindNet: Acquiring and structuring semantic information from text [C]//Proceedings of COLING-ACL'98, 1998: 1098-1102.

[5] 梅家驹. 同义词词林[M]. 上海：上海辞书出版社，1983.

[6] 董振东，董强. 面向信息处理的词汇语义研究中的若干问题[J]. 语言文字应用，2001(3)：27-32.

[7] 王惠，詹卫东，俞士汶.《现代汉语语义词典》的结构与应用[J]. 语言文字应用，2006(1)：134-141.

[8] 刘开瑛，由丽萍. 汉语框架语义知识库构建工程[C]. 中文信息处理前沿进展，北京：清华大学出版社，2006：64-71.

[9] 孙道功.《现代汉语析义元语言词典》的开发和应用[J]. 辞书研究，2011(5)：33-43.

[10] 刘兴林，陆建超，马千里. 基于互联网的词汇语义知识库构建框架研究[J]，计算机与现代化，2010(10)：8-11.

孙道功(1977－)，博士，副教授，主要研究领域为句法语义学和工程语言学。

E-mail：sundg9527@hotmail.com.

亢世勇(1964－)，博士，教授，主要研究领域为中文信息处理和汉语语法。

E-mail：kangsy64@163.com