基于抽象语义表示的汉语构式的标注与分析*

黄彤[1] 李斌[1,2] 闫培艺[1] 戴玉玲[1] 曲维光[3]

（1. 南京师范大学 文学院，江苏 南京 210097；
2. 哈佛大学 计量社会科学研究所，剑桥 美国 02138；
3. 南京师范大学 计算机科学与技术学院，江苏 南京 210023）

摘要：构式作为组成成分与实际意义不能完全对应的结构，与常规句子差异较大，对句法和语义分析器的影响较大，构式的自动分析则更是困难。因此，需要研究构式的内部结构标注与语料构建。由于构式的语义结构与句法结构有较大差异，我们使用中文抽象语义表示（CAMR）来直接标注构式的语义结构。目前收录最全的构式库是北京大学现代汉语构式知识库，通过对该构式库共 1057 条构式进行人工标注并统计后，发现 CAMR 可以表示出 61.2%的基本符合组合原则的构式；而 38.8%不符合组合原则的构式需要修改或添加概念，存在缺少概念、组成成分难以拆分、修辞意义难以表示等情况。该文给出的策略是将其整体作为一个谓词标注或只标注其表层义。汉语构式库的标注可以为构式语义的自动分析提供理论与数据基础。

关键词：抽象语义表示；构式；形式化表示；构式语料库；中文信息处理

中图分类号：TP391　　**文献标识码：**A

Chinese Constructions Annotation and Analysis Based on the Abstract Meaning Representation

HUANG Tong[1], LI Bin[1], YAN Peiyi[1,2], DAI Yuling[1], QU Weiguang[3]

(1. School of Chinese Language and Literature, Nanjing Normal University, Nanjing, Jiangsu 210097, China;
2. Institute for Quantitative Social Science, Harvard University, Cambridge, Massachusetts, 02138, USA;
3. School of Computer Science and Technology, Nanjing Normal University, Nanjing, Jiangsu 210023, China)

Abstract: As a structure whose meaning can't correspond to the literal meaning, construction is quite different from the regular sentences. Construction has a great influence on the accuracy of parser, and automatic analysis of construction is even more difficult. Therefore, it is necessary to study the internal structure and corpus construction of constructions. In this paper, Abstract Meaning Representation (AMR) is used to annotate the semantic structure of constructions, because the semantic structure is quite different from regular sentences. According to the annotation of the 1,057 Construction corpus, it is statistically found that CAMR can represent 61.2% of constructions that basically follow the principle of compositionality. However, 38.8% of the constructions that do not follow the principle of compositionality need to modify or add concepts, and there are some problems such as lack of concepts, difficult to separate components, and difficult to express rhetorical

收稿日期：2019-07-31　　**定稿日期：**2019-08-15

基金项目：国家社科基金项目（18BYY127）；国家自然科学基金（61772278）。

作者简介：黄彤（1996—），女，硕士研究生，主要研究方向为计算语言学；李斌（1981-），男，副教授，主要研究方向为计算语言学；闫培艺（1995-），女，硕士研究生，主要研究方向为计算语言学；戴玉玲（1996—），女，硕士研究生，主要研究方向为计算语言学；曲维光（1964），男，主要研究方向为自然语言处理。

meaning. Therefore, the more feasible strategy is to label the whole structure as word or only annotate its surface meaning. The annotation of Chinese construction can provide theoretical and data basis for automatic analysis of the meaning of construction.

Key words: abstract meaning representation; constructions; formal representation; construction corpus; Chinese information processing

0 引言

在目前的自然语言处理中，句法分析大体还停留在基于书面的常规文本上，对于非常规的不符合组合原则的结构的处理则力有不逮，如“就差没跪下”实际语义为“就差跪下了”，存在冗余的会影响语义分析的成分“没”。这些非常规的结构一般被称为构式，构式是日常口语对话中的重要部分，随着对话机器人、智能问答等系统的发展，构式的自动分析也逐渐受到关注，这就需要对构式进行合理的形式表示，将构式表示成计算机可读的形式，为自动句法分析提供基础。但总体而言，目前系统探讨构式表示的研究较少，且缺少面向计算的形式化表示研究，不利于构式的自动分析。

本文拟通过一种新的语义表示方法——抽象语义表示（Abstract Meaning Representation，AMR）来对汉语构式进行描写，并统计出汉语构式的具体类型和比例，介绍其标注方法。中文抽象语义表示（Chinese AMR，CAMR）是在 AMR 的基础上针对汉语特点改进而成，能够更合理地表示汉语语义。本文旨在探索 CAMR 对汉语构式的表示能力，对构式进行标注，完善 CAMR 标注体系，并对构式的类型进行统计分析，以期对构式自动分析和构式理论研究提供帮助。

全文结构如下：第 1 节梳理了构式的理论研究脉络以及目前构式的形式表示相关研究。第 2 节介绍了数据来源和标注原则。第 3 节介绍了构式的具体标注方式。第 4 节是探讨了典型构式的类型和处理方式。第 5 节是结论和未来工作。

1 相关工作

1.1 构式的理论研究

在语法研究中，习语等不规则的现象一直受到忽视，20 世纪出现的结构主义语言学重视描写语言现象，转换生成语法则重视语言的生成机制，如何用有限的规则生成无限句子。虽然习语、熟语等不规则结构始终存在，但只是被认为“和语词或词语一样，具有一定意义和功能的形式”，是语法中的“边缘现象”。直到 80 年代，随着认知语法的发展构式开始受到语言学界的重视。

初期的构式研究集中在对某些具体习语的探讨上。Lakoff 分析了“there”构式的构造和用法[1]。Fillmore 和 Kay 对习语“let alone”进行分析，认为“let alone”的意义不能从“let”或“alone”中推出，有其自身的句法语义特征，因而只能作为一个整体，并且探讨了这些习语类构式的规则性[2]。这些研究对构式理论的产生起到了奠基性作用。Goldberg 在此基础上形成了系统的构式语法理论，将构式界定为“C 是一个构式当且仅当 C 是一个形式——意义的配对<Fi，Si>，且其形式（Fi）或意义（Si）的某些特征不能从 C 的组成成分或其他已存在的构式中得到严格的预测”[3]，真正全面开启了构式语法研究。

此后，构式语法理论不断发展，成果斐然。如 Croft 创建的激进构式语法（Radical Construction Grammar）把构式的外延扩大至语素、词汇、短语甚至是语义规则，“激进”地认为构式是句法表征最基本的单位[4]。Goldberg 出版的专著对论元结构构式进行了概括分析，并在其中调整了对构式的定义：任意语言构型只要其形式或功能的一些方面并不能从其组成成分或已存在的其他构式中被严格地预测，或者只要他们出现的频率足够多就被认为是构式[5]。此时，Goldberg 也扩大了构式讨论的范围。Bergen 和 Chang 创建的体验构式语法（Embodied Construction Grammar）同样认为构式包括了所有的单位，重视构式的

语言理解过程[6]。可见，国外的构式理论研究趋势是意图将构式语法应用到语素、词、短语、句子等各个语言单位中，将构式作为语言研究的基本单位，扩大构式理论的解释力，使其能够解释各种语言现象。

国内较早基于构式理论进行研究的是张伯江，他引入了 Goldberg 的构式理论，将其翻译为“句式语法”，用以分析汉语双宾语构式[7]。此后，不断有学者对汉语特定结构进行研究。如郑娟曼认为“还 NP 呢”存在表达反预期信息的构式和表达反期望信息的构式两种标记构式[8]；吴为善分析了“A 不到哪里去”构式的语义特点、话语功能并探究其成因[9]；夏雪、詹卫东将“X 什么”一类构式区分为“言语行为否定”和“命题真值否定”两种否定义，分别分析两种否定的基本要素和要素间的关系和使用情况[10]。对于汉语具体构式的研究较为丰富，在这里无法一一罗列。同时也出现了一系列总结性的著作，牛保义[6]、王寅[11]等系统梳理、总结了国内外构式语法理论的产生背景、流派和研究方法，并将之与汉语构式进行结合。

在汉语学界，对于汉语构式的内涵和外延的争论贯穿始终。石毓智认为构式语法的局限正是在于对构式概念的“不合理扩大”[12]。陆俭明认为不能将构式仅认为是“形式和意义的配对或者匹配”，构式必须是一个结构体，同时也指出了具有可预测性的句式没有必要用构式语法来解释[13]。也有一部分学者认同构式的范围包括不同层面的语法单位，如陈满华认为单词的意义同样不能从其构成成分中推出，语素仍然可以有更小的语言单位构成并且不能从这些语言单位里推出其意义，因此语素和单词作为构式是合理的[14]。

通过对国内外构式综述的梳理，可以看出国外构式理论研究趋势是将构式的范围不断扩大，旨在将构式语法应用到语法研究的方方面面，扩大构式语法的解释力；国内构式理论研究更注重对传统语法很难解释清楚的语法现象进行分析，这就影响了国内外学界对构式的定义。国外更倾向于将构式定义为形式和意义的结合体，这就使得构式包括语素、词、短语等语言单位，乃至语言规则也囊括进构式内涵。国内的构式定义则更强调结构的不可预测性，即结构的意义不能从组成成分中得到。

1.2 构式的形式化表示研究

随着构式理论的不断发展，国内外开始有学者尝试对构式进行表示，并衍生出面向计算的构式语法理论，其中最具代表性的构式语法理论是 90 年代兴起的针对英语提出的流体构式语法（Fluid Construction Grammar，FCG）和基于符号的构式语法（Sign-based Construction Grammar，SBCG）。

流体构式语法是面向计算语言学提出的，它是一个可操作的计算平台[15]。FCG 对所有构式都有一个表示形式：意义和形式之间的双向映射，使用标记词类和输出的机制来完成对构式的计算表示，其特点在于强调构式的动态性，因为语言使用者常常变动和更新自己的言语习惯。FCG 将语言信息表示为语义级（semantic pole）和句法级（syntactic pole）之间的映射，一个单元（unit）可以被认为是一个封装特征值对的标记框。以人名 Kim 为例，FCG 将 Kim 表示为一个特征值对[16]。

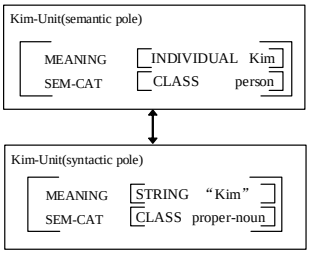

图 1 “Kim”FCG 表示方式示意

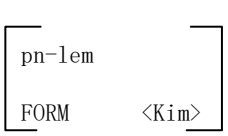

图 2 “Kim”SBCG 表示方式示意

图 1 中语义级（semantic pole）描述了 Kim 的语义范畴特征是“人（person）”，语法级（semantic pole）描述了 Kim 的形式是字符串“kim”，范畴意义则是“专有名词

（proper-noun）”。

基于符号的构式语法（SBCG）是一个包含有复杂符号的库[17]，库里包括形式、意义和各种限制的语言信息，构式就是简单的符号组合成复杂的方式，SBCG 没有对构式的单一表示，而是区分了词汇项（lexical items）和组合结构。同样以 Kim 为例说明 SBCG 的表示方式。

Kim 的词汇类别（lem）被表示为 pn（proper-noun，专有名词）类，形式是“Kim”。其他词同样可以用这些符号表示，互相组合形成一个更大的符号。

框架语义学被视为构式语法姐妹理论，在框架语义学基础上构建的词库工程框架网（FrameNet）对构式进行了专门的标注。FrameNet 从大量的语料实例中仿照词汇的标注对构式的框架进行描述，标注出构式的组成成分，例如对于典型构式“let alone”，标出其第一个并列项（conjunct）和第二个并列项。同时对于某些特别的构式还会标注出其语法功能和短语类型，最终 FrameNet 的表示包括对构式的描述和例句两部分。

国内对于构式的计算研究还处于起步阶段，大多数研究都只针对某一类构式的形式表示进行探讨，一部分学者认为构式的研究需要和语块的研究结合起来，认为一个构式由一个或几个语块组合起来。苏丹洁认为现代汉语的存在句可以用“存在处所”、“存在物”和“两者链接”三个语块表示[18]。2017 年詹卫东等人系统构建了现代汉语构式知识库，收录了 1057 条构式，是目前收录最全的构式库[19]。

随着自然语言处理的发展，国内外学者越来越重视构式的形式表示。在国外学界，构式的形式表示范围包括了语素、词、短语等各级语法单位。国内则更关注某些特殊结构的表示，但对其的整体研究还较少，且较为零散，这既不利于构式的计算机自动分析，也不利于构式的系统研究。因此本文将基于 CAMR 标注构式，总结其不同类别并进行统计分析，希望能对构式的研究和自动语义分析起到一定作用。

1.3 抽象语义表示研究

抽象语义表示（AMR）是一种新兴的语义表示方法。AMR 将一个句子的语义抽象为一个单根有向无环图，其中句子中的实词抽象为概念节点，实词之间的关系抽象为带有语义关系标签的有向弧，且忽略虚词和由形态变化体现的较虚的语义（如冠词、单复数、时态），允许增加、删除或修改概念[20]。AMR 区分了核心语义关系和非核心语义关系，共有核心语义关系标签 5 个、非核心语义关系标签 40 个。Bonial 等人运用 AMR 标注了 5 类英语构式，通过使用原有的关系标签和新增表示程度的概念标签 Have-Degree-91、表示数量的概念标签 Have-quant-91 等解决了一部分构式的表示[21]。AMR 新增的 Have-Degree-91 概念框架如图 3 所示。

图 3　AMR 表示程度的 Have-Degree-91 概念框

Have-Degree-91
arg1: domain, entity characterized by attribute
arg2: attribute
arg3: degree itself
arg4: compared-to
arg5: superlative: reference to superset
arg6: consequence, result of degree

图 4 “她平时不化妆不出门的”CAMR 表示文本

她[1] 平时[2] 不[3] 化妆[4] 不[5] 出门[6] 的[7]
x8/condition
:arg1 x4/化妆-01
:polarity x3/-
:arg0 x1/她
:arg2 x6/出门-01
:polarity x5/-
:arg0 x1/他
:time x2/平时
:smood x7/的

由于 AMR 对汉语特有的语言现象缺少处理规范，Li 等人将 AMR 引入，进行了概念与关系对齐，根据汉语的特点进行兼容扩充，形成了中文抽象语义表示方法（CAMR）[22]。CAMR 规定了汉语中一些特别的语言现象的标注方法：为量词、时、体新增了语义关系标签；对重叠式进行还原，如“试试”还原为“试”；组合离合式，如“洗了一晚上澡”，把洗和澡合成一个概念“洗澡”；为复句关系增加了 10 个关系概念标签，比如“condition（条件）”等等。为了实现概念与原句

中的词的对齐，李斌等（2017）又对 CAMR 标注体系进行了进一步改良，提出了融合概念对齐的一体化标注方案，并设计构建了适合于对齐版 CAMR 的人工标注平台。下面以“她平时不化妆不出门的。”为例，对改良后的 CAMR 标注方法进行介绍。

从图 4 中可以看出，CAMR 在进行句义标注前，会先对句子进行分词，并根据句子序列为每个词编号，概念与句中的词的对齐则是根据每个概念对应的词或词内的字的编号来实现的。本文主要基于 CAMR 对构式进行标注。

一个结构或句子根据其特征分为符合组合原则的结构和不符合组合原则的结构，符合组合原则的结构指结构的实际语义可以从其组成成分的意义得知，而不符合组合原则的结构则是指结构的实际语义难以（完全）从其组成成分的意义推知。目前 CAMR 已经发布了 11711 句标注语料，标注的句子多为常规的符合组合原则的句子，还未涉及大多数非常规的结构。

2 数据标注

2.1 数据来源

本论文中所出现的构式语料均来自詹卫东（2017）的北京大学现代汉语构式知识库（以下简称构式知识库），该知识库共收录构式 1057 条，删除了完全相同的重复构式后共 1038 条，其收录的标准包括四条：内部成分有组合性；构式内部成分无递归性；“形式——意义”超常规搭配；具有独特交际价值。构式知识库包含 5 个与构式标注相关的基本字段（表 1）：

表 1 “现代汉语构式知识库”基本字段

构式形式	a+ing
构式类型	短语型
构式特征	语法错配
释义模板	处于+a+的+状态+中
实例	开心 ing|抓狂 ing|伤心 ing

（1）构式形式。由常项和变项构成。如“a+爆；”

（2）构式类型。构式类型有凝固型、半凝固型、短语型、复句型四种。我们将凝固型构式作为一个谓词收录 CAMR 的词典。

（3）构式特征，特征分为语法形式特征、语义特征和其他特征，语法形式特征包括复现、省略、冗余、异序、语法错配、含否定成分、含疑问成分；语义特征包括论元异常、否定义、负面评价、周遍、主观大量、主观小量、语义错配，其他特征包括修辞、网络用语。本文在用 CAMR 对构式进行标注时参考了构式知识库中的特征类型。

（4）释义模板，即构式的直接释义，我们在标注构式时参考了释义模板的解释。

（5）实例。除了凝固型构式不需要填写外，其他类型构式均至少收录三个实例。除了凝固型构式外，我们只对构式实例进行标注，而不是直接标注构式形式。

同时构式知识库还包括了其他更详细的信息：（a）构式变体，即构式的变异形式，一般是该构式的某个常项存在差异，如“a 爆”的构式变体填值为“a+到+爆”、“a+爆+了”、“a+到+爆+了”。（b）义项，只有一个义项的构式该字段值为 0，有多个义项的构式分别用 1、2、3 表示有 1、2、3 个义项；音节数，即构式的音节数；变项数量和常项数量；近义构式、反义构式、上位构式、下位构式；备注，对构式信息的补充说明。

2.2 数据标注

本文将采用 CAMR 标注体系标注 1038 条构式。对于表层义和深层义一样或相近的符合组合原则的构式尽可能根据其组成成分运用语义关系标签、增删修改概念方式表示出构式的实际语义，对于表层义和深层义存在差异的不符合组合原则的构式则标注两遍，一遍

根据表层的字面义标注，一遍根据构式的实际的深层意义标注。

构式库的 CAMR 标注共有 2 名语言学学生参与，同时标注了 1038 条构式语料，标注一致性达 81%，对不一致的标注进行探讨并调整标注。在 1038 条构式语料标注完成后，我们发现符合组合原则的构式共 635 个，所占比例为 61.2%，不符合组合原则构式共 322 个，比例为 38.8%。由组合原则构成的构式是指虽然与常规句子不同，但其意义基本上可由其组成成分推知。

3 组合原则的构式的标注与分析

本节将探讨符合组合原则的构式的标注，通过调整语序、删除冗余成分、使用已有的关系标签和谓词框架、补充复句关系概念和新增 Have-Degree-91 概念标签能够标注 61.2%的构式。表 2 给出了构式标注结果的详细类型，比例之和超过 1 是因为一个构式可能同时使用两种标注方法。

3.1 语序

知识库中存在一部分异序特征类型的构式，所谓异序，在于假定构式有一个对应的常规语序的句子。CAMR 将句子结构抽象为单根有向无环图，异序特征的构式与常规句子的标注一致，因此语序对标注结果的影响较小。

图 5 异序构式的 CAMR 表示文本

```
狗熊1 一-2 个3
x1/狗熊
    :quant   x2/1
    :cunit x3/个
```

图 6 互文异序构式的 CAMR 表示文本

```
头1 昏2 脑3 胀4
x2_x4/昏涨-01
    :arg0   x1_x3/头脑
```

图 5 例子所对应的常规语序应该是“一个狗熊”，CAMR 对两种形式的标注是相同的，会将“狗熊”作为根（root），“一”和“个”分别是狗熊的 quant 和 cunit。具有异序特征的构式的标注与常规句子无异。

表 2 构式类别统计

序号	构式类别		数量	比例（%）
1	组合原则	异序	51	4.9
2		删除冗余成分	111	10.6
3		使用已有关系标签、谓词框架	188	18.1
4		补充复句关系概念	225	21.7
5		新增程度概念	83	8.0
6		小计	658	63.4
7	非组合原则	缺少概念	63	6.1
8		成分难以拆分（凝固型）	164	15.8
9		范围难以界定	24	2.3
10		需要语境推导言外之意/修辞	110	10.6
11		其他	42	4.0
12		小计	403	38.8

此外，在标注使用互文修辞手法的四字习语时需要调整语序，互文指在结构相同或相似的上下文中，上文里隐含着下文里出现的词语，下文里隐含着上文里出现的词语，参互成文，合而见义，以“头昏脑涨”为例，其实际意义是“头脑昏胀”，在标注时会重新组合他们的成分（图 6）。

3.2 删除冗余成分

构式中常常含有多余成分，少了这些成分不会对构式的意义产生影响。构式库冗余构式可分为有两种，一是构式中的成分存在重复，如“万一的万一”；二是构式中多了一些从字面上对构式整体意义没有影响的成分，如“就差没跪下”。针对这种情况，CAMR 在表示句子时会对其中意义较虚或冗余的词语进行删除的操作，在处理时不予表示多余成分。

图 7 冗余构式的 CAMR 文本表示

```
就[1] 差[2] 没[3] 跪[4] 下[5]
x2/差-03
    :arg1   x4/跪-01
    :polarity x3/-
    :direction x5/下
    :mod x1/就
```

图 8 使用非核心语义关系标签标注

```
小[1] 姑娘[2] 哭[3] 得[4] 我[5] 闹心[6]
x3/哭
    :arg0   x2/姑娘
      :arg0-of   x1/小-01
      :cause-of(x4/得)   x6/闹心
      :arg0   x5/我
```

图 7 例子实际语义为“就差跪下”，在标注时根据语义省略了“没”这个冗余成分。

有的构式在复现时其实蕴含了其他意义，复现的成分实际上触发了不同的复句关系，我们在标注时会将这种隐含语义补充出来。如“好是好”，这一结构无法单独成句，必须要有一个带有转折意义的后续句，因此在标注时不仅删除了句中冗余的成分（“是好”），同时也根据其实际语义增加了复句概念。

3.3 利用已有关系和概念标签表示语义

CAMR 有一套概念和关系标签。概念标签主要从实词中抽象出来，大部分概念都是由句中的词语充当。关系标签可分为核心语义角色关系标签（argx, x 取值为 1、2、3、4）和非核心语义关系标签，如 cause（起因）。

大部分由词语抽象得到的概念都有其论元结构，但在标注过程中会出现词典中概念的框架无法表示意义的具体构式，在具有语法错配、语义错配、论元异常特征的构式中体现得尤为明显，这一类构式的组成成分在传统句子中无法进行搭配，对于该类构式一般采用两种方式：一是使用谓词自身框架包含的语义角色标签和非核心语义关系标签进行标注。二是为谓词另外添加义项，重新设计概念的框架。

首先，可使用已有的语义角色标签对构式中的构成成分进行标注，尤其是借助 cause（起因）、result（结果）等非核心语义角色关系标签。如图 8，在这个句子里，“哭-01”是句子的核心，“姑娘”是哭的施事 arg0，“闹心”是“哭”的结果，“我”则是“闹心”的主体，这个句子使用谓词“哭”的事件框架和非核心角色关系标签 cause-of(cause 的反关系)可以将其语义表示出来。

具有错配特征的构式一部分是词类活用，其中以形容词作为动词使用最为常见，如“幸福着她的幸福”，类似的用法已经较为普遍，大多数表示心理情绪的形容词都可以出现这种变化，所以我们的做法是为这一类形容词都添加一个动词用法的词条，同一般动词类似，将 arg0 作为施事，arg1 作为受事。

3.4 标注复句关系概念

表示复句关系的构式所占的比例为 21.7%，是组合原则构式里比例最高的一种类型。复句关系是 CAMR 新增的语义关系标签，CAMR 增加了 10 个复句概念，当语料由多个独立句子构成时，可以用这 10 个概念来表示不同分句之间的关系。在标注时，语义较虚的关联词会标注在语义关系标签后面的括号里，出现在 CAMR 图中则表示为概念之间关系的弧上。

复句类构式可分为两种：一是省略了表示复句关系的关联词，二是使用了通常不作为关联词使用的常用语作为关联词。省略了关联词的构式在该类构式中占比为 29.3%，参见图 4，在标注时补充条件复句关系作为根。

使用了非常规的关联词的构式占比 71.7%，如图 9 例子是一个递进（progression）复句，我们通常不把“不说”作为表示递进关系的关联词。

图 9 特殊关联词复句构式的 CAMR 文本表示

```
低效[1] 不说[2] ，[3] 还[4] 拒绝[5] 合作[6]
x8/progression
    :arg1(x2/不说)   x10/低效-01
    :arg2(x4/还)   x5/拒绝-01
        :arg1   x6/合作
```

图 10 Have-Degree-91 概念框架

```
Have-Degree-91
arg0:domain,entity characterized by
attribute
arg1:attribute
arg2:degree itself
arg3:consequence,result of degree
```

3.5 表示程度

构式库中有一部分构式表示某些特殊的意义，主要是表示否定和程度的意义，表示否定的构式标注时可通过增删、修改概念表示否定义，表程度义的构式有其特殊性并且数量可观，因此单独进行探讨。

AMR 有一批特殊的概念，使用后缀“-91”作为标记，由于含有程度意义的构式数量在构式库中较为可观，共有 91 个构式用来表示程度，并且对于程度的语义表示难以直接用现有的关系标签 degree 概括，因此我们仿照 AMR 对英语构式的标注新增了一个概念“Have-Degree-91”来表示程度，并根据汉语构式的特点设置了“Have-Degree-91”的谓词框架（图 10）。

构式中在形容某物的程度时，常用一个夸张的结果来表达某物的程度之深，因此在 have-degree-91 框架中加入了 arg3 表结果。如图 11 所示，用“帅”导致的结果“没有朋友”来表示“帅”的程度。

图 11 Have-Degree-91 概念框架示例

图 12 缺少概念构式的 CAMR 表示文本

```
帅[1] 到[2] 没有[3] 朋友[4]
x5/have-degree-91
    :arg1 x1/帅
    :arg3(x2到) x4/有–01
        :polarity x3/-
        :arg1 x5/朋友
```

表示程度的成分会出现语法、语义超常错配的情况，如“差得可以”，或是使用原本不能表示程度的修饰词进行修饰，例如“运气好得不要不要的”，我们在标注时会直接将承载程度意义的成分作为 arg2（degree itself）。

4 非组合原则的构式的标注与分析

意义不能从其组成成分推知是较为公认的构式定义的重要部分，也就是说，实际上构式是不符合组合原则的，这一类构式是典型的构式，特点是由组成成分构成的表层义和真正想表达的深层义不同。根据统计，构式知识库中不符合组合原则的构式的比例为 38.8%，这一类型构式 CAMR 只能标注其表层义，而不能或者很难准确地表示出其深层义。据分析，这一类构式存在以下几种情况：缺少概念、组成成分难以拆分、范围难以界定、需要语境推导言外之意及修辞和一部分难以归类的构式，对于典型构式我们倾向于作为一个整体进行标注或只标注其表层义。

```
你[1] 真[2] 是[3] 的[4]
x7/好
  :polarity   x8/-
  :domain(x3/是) x1/你
    :mod   x2/真
    :smood   x4/的
```

4.1 缺少概念

缺少概念的构式是指缺少了与其构式义相关的成分，有两种情况：一是缺少表示概念的成分，二是使用与构式义不直接相关的成分来表示概念。如果要准确地表示语义，就必须对句中的实词概念进行修改或者添加。虽然 CAMR 本身有一套增删修改概念的机制，但

其只允许补充句子中省略或隐含的概念。以“卖菜的走了”为例，标注时需添加概念“person”，但这个新增的概念是没有争议的。而这类构式相较常规句子省去了一些与构式语义直接相关的成分，在标注时无法确定地补出其实际意义，甚至不同的语境下会有不同的意义，因此难以表示，如“你真是不好”（图 12）。

根据理解，例句“你真是的”的实际意思是“你真是不好”，在其表层结构中省略了“（不）好”这个谓词，在标注时为了还原语义就需要补出这个概念。修改概念也存在同样的情况，当遇到构式里部分成分的字面意义难以看出其实际意义时，CAMR 可以将构式的成分修改为符合实际语义的概念，但具体修改的概念可能会存在争议。如“曹雪芹长曹雪芹短”，根据语义可将“……长……短”抽象成概念“talk”，在修改概念时为了避免歧义将概念表示为英语单词（图 13）。

图 14 “东一榔头西一棒槌”框架

```
曹雪芹[1] 长[2] 曹雪芹[3] 短[4]
x2_x4 / talk
   :arg1 x6 / person
        :name x1_x3 /  曹雪芹
```

在标注时，根据标注人员对构式的理解和出现的语境不同，所添加和修改的概念也会存在不同，难以达到标注一致性，因此目前我们倾向于标注其表层义。

4.2 组成成分难以拆分

构式知识库中凝固型构式的特点是没有变项，且长度不变，构式的表层义和深层义相去甚远，作用与词一致。主要包括习语、成语、网络新词等。以习语“东一榔头西一棒槌”为例，这一类构式难以根据语义拆分，如果拆分反而不能准确表达该构式的意义，因此目前我们只能将这类构式作为一个谓词收录进谓词词典中，在标注时和其他谓词一样，设置该谓词的框架（图 14）。

图 14 “东一榔头西一棒槌”框架

图 15 “点赞”框架

```
东一榔头西一棒槌
arg0: entity that is messy
```

此外，还有一系列应用较为广泛且使用稳定的网络新词，他们无法拆分或变换语序，并且意义固定，因此将这类构式作为一个谓词收入到词典中，如图 15 所示。

```
点赞
arg0: people described
arg1: entity arg0 gives a like
```

对于凝固型的成语、新词、习语等类型构式，CAMR 的处理方式是将其收录词典中，并设计其框架。

4.3 范围难以界定

这一类构式常常使用“对举”修辞手法，句中实际所指的位置范围或时间范围不同于字面意义所指的范围。如“东逛逛西逛逛”，实际意义是指到处逛，而不是仅仅逛字面上的“东”和“西”，类似的还有“今天吃一种药，明天吃一种药”指天天都吃不同的药，而不只是“今天”和“明天”。这类构式所指的范围常常难以确定，同样以“东逛逛西逛逛”为例，如果 location 标为“everywhere”显然与实际意义也有所不同，因此很难准确地表示出这类构式的语义。此类构式只能根据其字面意义进行标注，以“东逛逛西逛逛”为例，将其标注为“逛逛东边逛逛西边”。

4.4 需要语境和语用推导

有些只能在语境中进行理解，在不同的语境中可能会产生不同的意义。例如“这不是钱不钱的问题”，根据不同的语境会出现不同的语义。在“这不是钱不钱的问题，只是我也需要一点安慰。”这个语境下，这一结构表示“这是钱的问题”。在“这不是钱不钱的问题，

你要认识到自己的错误。”语境中则表示“这不是钱的问题”，这类结构的意义需要放在具体的上下文语境中才能确定。

此外，有一部分构式使用了诸如比喻的修辞手法，如“友谊的小船说翻就翻”，将“友谊”比作小船，修辞的运用属于语用范畴。目前 CAMR 还无法对这些语用推导的句子进行处理，这类构式只能标注其字面意义。

4.5 小结

通过对 1038 个构式的标注，我们可以看出中文抽象语义表示能够清晰地表示汉语大部分构式。在传统的表示方法中，仅凭结构入手很难表示出构式的实际意义。通过 CAMR 的标注结果，我们发现 61.2%的构式可以从其组成成分的意义推知构式的整体意义，38.8%的典型构式意义不能从其组成成分推知，这一部分构式目前只能作为一个谓词收录词典中或标注出表层意义。

结合国内外对构式的界定，我们认为构式强调的是形式和意义的结合，但是其整体含义不能完全从构成成分的意义或形式中得知，因此以下两种结构不是构式：

首先，省略了关联词的复句不是构式。知识库中收录了这类特征为省略的构式，但我们认为这一类结构并不是真正的省略，汉语复句中各分句的关系可以用关联词语来连接表示，有时也可以不用或不能用关联词语表示，省略了关联词语的复句/紧缩句较为常见，这类构式是可以从其他已存在的结构中得到预测的，因此这类结构并非构式。

其次，实际意义只能从语境中得到的意义不是构式。这一部分结构只能存在于句子或者有上下文语境，在不同的语境中甚至会产生变义，因此这类结构并不满足构式是“形式和意义的配对”，我们认为这一类短语不是构式。

5 结论及未来工作

近年来，随着构式理论的不断发展和自然语言处理的需求，构式的形式表示也越来越受到重视，但汉语构式表示还未得到充分的关注。本文梳理了国内外构式的形式表示，基于 CAMR 体系对北京大学现代汉语构式知识库 1038 条构式进行标注，通过统计分析得到符合组合原则的构式比例为 61.2%，不符合组合原则的构式比例为 38.8%。总结出符合组合原则的构式可通过调整语序、删除冗余成分、使用已有关系标签、谓词框架、补充复句关系概念以及补充谓词 Have-Degree-91 等进行合理的标注，不符合组合原则的构式则出现了缺少概念，组成成分难以拆分、范围难以界定、需要语境推导四种情况，目前对这一类构式只能根据组成成分标注表层义，尚难合理地表示其深层义。

虽然使用 CAMR 可以表示出大部分构式的语义，但在标注不符合组合原则的构式时我们仍然遇到了许多问题。因此在未来的工作中，第一，我们将加强对构式理论的研究，进一步分析构式的形式与意义之间的关系，以提高对构式的表示能力；第二，继续完善 CAMR 的标注体系，探寻深层意义的语义表示方法，以期能够更合理准确的表示构式语义，对计算机自动分析构式语义提供帮助，同时也进一步扩大 CAMR 可以表示的范围。最后，我们希望通过对构式的标注语料库进行机器学习，提高中文 AMR 分析器的效果。

参考文献

[1] George Lakoff. 女人，火和危险事物[M]. 李葆嘉等，译. 北京：世界图书出版公司北京公司，2011.

[2] Charles J. Fillmore, Paul Kay, Mary Catherine O’Conn. Regularity and Idomaticity in Grammatical Constructions: The Case of Let Alone[J]. Language, 1988, 64(3): 501-538.

[3] Adele E Goldberg Constructions:A Construction Grammar Approach to Argument Structure[M]. Chicago: University of Chicago Press, 1995.

[4] Croft W. Radical Construction Grammar：Syntactic Theory in Typological Perspective[M].

Oxford&NY: OUP, 2001.
[5] Adele E. Goldberg. Construction at Work: The Nature of Generalization in Language[M]. Oxford: Oxford University Press, 2006.
[6] 牛保义.构式语法理论研究[M]. 上海:上海外语教育出版社, 2011.
[7] 张伯江. 现代汉语的双及物结构式[J].中国语文，1999，3: 175-184.
[8] 郑娟曼. “还 NP 呢”构式分析[J]. 语言教学与研究，2009，2: 9-15.
[9] 吴为善，夏芳芳. “A 不到哪里去 ”的构式解析、话语功能及其成因[J]. 中国语文，2011，4：326-333.
[10] 夏雪，詹卫东.“X 什么”类否定义构式探析[J]. 中文信息学报，2015，29(5): 1-8.
[11] 王寅. 构式语法研究[M]. 上海: 上海外语教育出版社，2011.
[12] 石毓智. 构造语法理论关于 construction 定义问题研究[J]. 重庆大学学报，2007，13(1): 108-111.
[13] 陆俭明. 构式语法理论研究中需要澄清的一些问题[J]. 外语研究，2018，168(2): 1-15.
[14] 陈满华. 关于构式语法理论的几个问题[J]. 外语教学与研究，2009，41(5): 337-400.
[15] Steels, L. Design Patterns in Fluid Construction Grammar[J]. Computational Linguistics: 2011, 39(2): 447-453.
[16] Luc Steels, Joachim de Beule L. A (very) Brief Introduction to Fluid Construction Grammar[C]//Proceedings of the 3rd Workshop on Scalable Natural Language Understanding, New York, 2006, 73-80.
[17] Charles Fillmore, Paul Kay, Laura Michaelis. Sign-Based Construction Grammar[M]. Stanford: Center for The Study of Language And Information, 2013.
[18] 苏丹洁，陆俭明.“构式—语块”句法分析法和教学法[J]. 世界汉语教学，2010，24(4): 557-567.
[19] 詹卫东. 从短语到构式：构式知识库建设的若干理论问题探析[J]. 中文信息学报， 2017，31(1): 230-238.
[20] Laura Banarescu, Claire Bonial, Shu Cai, et al. Abstract meaning representation for sembanking[C]//Proceedings of Linguistic Annotation Workshop and Interoperability with Discourse, Sofia, 2013: 178-186.
[21] Claire Bonial, Bianca Badarau, Kira Griffitt, et al. Abstract meaning representation of constructions:the more we include,the better the representation[C]//Proceedings of language resources and evaluation, Miyazaki, 2018, 1667-1684.
[22] 李斌，闻媛，宋丽等. 融合概念对齐信息的中文 AMR 语料库的构建[J]. 中文信息学报，2017，31(6): 93-102.

黄彤（1996—），硕士生，主要研究领域为计算语言学。
E-mail: iwanttardis@163.com


李斌（1981—），博士，副教授，主要研究领域为计算语言学。
E-mail: libin.njnu@gmail.com


闫培艺（1995—），硕士生，主要研究领域为计算语言学。
E-mail: ypyheta@gmail.com

作者联系方式：黄彤　地址：江苏省南京市鼓楼区宁海路 122 号南京师范大学随园校区文学院　邮编：210097　电话：18159592988 电子邮箱：iwanttardis@163.com