面向计算的构式研究：现状、问题与展望

北京大学　詹卫东　王佳骏*

［提　要］　本文首先概要介绍面向计算的构式相关研究工作，包括面向计算的形式化构式语法理论模型的构建，以及面向计算的构式语言数据资源建设工作。在此基础上，结合我们所做的现代汉语构式知识库建设和构式语料标注工作，本文提出将构式的句法语义分析与传统的“词库＋短语规则”语言知识系统融合的思路，并指出当前面向自然语言处理的构式研究的重点任务是构式数据资源的建设。

［关键词］　构式知识库；语料库标注；形式化表征；语言数据资源建设；自然语言处理

❶ 引言

关于“构式”（construction）的含义，可以有两个层次的理解，一个是作为一种语法观的“构式”，一个是作为特定语法实体的“构式”。

理解构式语法观的形成，离不开生成语法理论这个大背景。生成语法的理论主张之一是语法系统的模块化，让形式的归形式，语义的归语义，然后再构建模块之间的“界面”（interface），把句法形式系统和语义解释系统联系起来。此外，生成语法学者还主张解释语言的无限能产性是语法系统的使命，为此就需要采取具有递归表达能力的短语组合规则来表征句法系统。与此相对立，构式语法观突出的特点是主张语言具有“构式性”，其实质也就是索绪尔所谓的语言的“符号性”，即语言系统是“形式－意义”同步建构的，在组织语言知识的表征系统时，应把语言符

* 作者简介：詹卫东，北京大学中文系教授、博士。研究方向：现代汉语形式语法、中文信息处理、汉语语言知识工程。Email: zwd@pku.edu.cn。通信地址：100871　北京大学中文系。王佳骏，北京大学中文系博士在读。研究方向：形式语法理论、语言知识工程、统计学习方法。Email: wangjiajun_1991@pku.edu.cn。通信地址：100871　北京大学中文系。

本文相关研究得到国家科技创新2030“新一代人工智能”重大项目（2020AAA0106701）和教育部人文社科基地2015年度重大项目（15JJD740002）的支持。

号的形式和意义同等看待，同时进行形式和意义的配对描写，而不是把形式和意义分开来处理，也不需要以规则推导的方式来生成句子结构。

作为语法实体的构式，指的是“形式–意义”配对具有特殊性的某些自然语言实体，比如“let alone”（更不必说）、“kick the bucket”（去世）、“What’s X doing Y”（X 做 Y 这件事不合情理）等。这些语言单位的特殊性在于，其构成成分的字面意义，不足以说明整体的意义。为了准确说明整体的意义，就有必要在构成成分（词语）之外，增加一个“抽象的成分”，即“构式”。[①] 从这个角度说，常规的语言实体，其整体和组成部分的关系，好比“1+1=2”。而构式的整体和组成部分的关系，好比“1+1=3”。为了解释“1+1”为何会等于 3，就需要将这个式子改为“1+1+‘1’=3”。其中的“1”就是除了词语之外的那个（隐性）语言实体——构式。

上述两个层次的构式概念，前一个可以说是“构式性”，后一个则是“构式体”。而这两种解读方式，在发展过程中出现了一种融合为一的趋势，即构式是普遍存在的，是语言符号系统的基本性质，所有语言实体皆为构式（Croft，2001）。

从计算机自然语言处理（Natural Language Processing，NLP）的应用角度来说，这样的语法观是否会得到认同，又如何在计算系统中落实呢？本文对从计算视角开展的构式相关研究工作进行初步的考察调研（第 2 节）；然后介绍我们在现代汉语构式知识库建设方面做的工作，提出将“构式”视作特殊短语，对构式的句法语义分析，应融入传统的“词汇 + 短语结构规则”框架的处理思路（第 3 节）；最后在结语部分展望面向计算的构式研究的发展前景，指出重点任务是构建和积累与构式相关的语言学知识资源（第 4 节）。

❷ 面向计算的构式相关研究工作

早期构式研究以固定习用语为代表，跟普通短语结构相比，可以说是一种边缘语法现象。在语法理论研究中，构式长期以来都不是主流分析对象。同样地，在 NLP 中，构式现象受到的关注也不多。20 世纪 90 年代至今，跟计算相关的有代表性的构式研究工作，大体可分为两类，[②] 一是基于构式语法观的语法理论模型研究；二是以狭义的实体构式为对象，构建构式知识库和标注构式语料的工作。

2.1 基于构式语法观的形式化语法模型

根据我们目前所见到的文献，最早以“构式”名义搭建计算机用的形式化语法框架的工作，是加州大学伯克利分校的 Daniel Jurafsky（1991）提出的名为 CIG 的语法模型，即基于构式的解释型语法（Construction-based Interpretive Grammar，CIG）。CIG 把不同层级的语言单位统称为“语法构式”（grammatical constructions），对

① “构式”在口语中没有语音形式，听不见；在书面上也没有符号形式，看不见。

② NLP 中的构式相关研究工作并不仅限于本文介绍的这两类，还有其他的一些类型，比如从语料库中自动提取构式的算法研究（Dunn，2017）、语料中构式的自动识别和标注研究（黄海斌等，2020）。但总的来说，跟计算有关的构式研究类型不多。限于篇幅，本文聚焦在面向计算的语法模型研究和资源构建这两类。应该说，这也是跟计算有关的构式研究的主要大类。

每条构式的语音、词汇、句法、语义、语用信息都采用统一的模式进行表征。运用CIG对句子进行分析和语义解释的系统名为Sal。[③] 该系统由3部分组成：（1）工作区（working store）：存放对当前句子做并行分析过程中的各个构式；（2）知识库（long-term store）：存放一个语言的全部语法构式的全部知识；（3）解释函数（interpretation function）：用于解释句子语义的一组函数，是Sal系统算法的核心，具体包括：（a）检索函数（access function）——通过不同知识源以交互作用方式来搜索能够匹配当前待分析句子的正确构式，得到并行的构式候选集；（b）整合函数（integration function）——通过合一运算（unification）整合构式组成成分的信息，得到构式的语义解释（以框架表示）；（c）选优函数（selection function）——通过一致性评价以及评分剪枝等方式从候选的句义解释中淘汰低分构式，选择最优结果。

从Sal的系统架构和实现方式来看，Jurafsky设计的CIG的使用方法与一般的基于合一的短语结构文法（詹卫东，2000）实质是一样的，其分析算法的核心框架是自底向上的移进 — 归约算法，分析结果也是用层级句法树和语义特征集来表示，对于分析过程中产生的中间候选结果，也按照通常的剪枝策略进行了干预，以提高分析的时间效率和空间效率。限于当时的条件，CIG+Sal系统只在Common Lisp编程环境中测试了包含50个构式的一个小型语法。Jurafsky（1992）展示系统用的例句是“How can I create disk space”（我应当如何创建磁盘空间），这不是典型的狭义实体构式。从计算视角来说，把它看作是构式还是普通短语，并无太大区别。对于作为语法实体的构式，CIG尚未做比较深入且系统的探究。

除早期的CIG模型外，还有若干以“构式语法”冠名的形式化语法模型的研究，Hoffmann和Trousdale（2013）的第7章到第10章有集中的介绍，包括伯克利构式语法（Berkeley Construction Grammar，BCG）、[④] 基于语符的构式语法（Sign-based Construction Grammar，SBCG）、体验构式语法（Embodied Construction Grammar，ECG）、流变构式语法（Fluid Construction Grammar，FCG）[⑤] 等。值得注意的是，这些理论的来源背景和设计目标存在比较大的差异：BCG和SBCG的语法理论本体研究色彩更浓厚；而ECG和FCG的交叉学科视野特点更鲜明，更关注从计算过程和计算机模拟人的语言能力角度去构建语法模型。

据Sag和Hans（2012）的介绍，SBCG是中心语驱动的短语结构语法（Head-driven Phrase Structure Grammar，HPSG）的形式化表征框架和BCG构式语法观融合的产物。HPSG的描写对象是普通的短语结构，SBCG则拓宽到关注构式语法现象。BCG虽长期致力于英语构式的发掘和分析，但缺少像HPSG这样的统一的形式化表征工具。SBCG正是在这样的背景下逐步发展起来的。从Pollard和Sag（1994）到Sag和Wasow（1999）以及Sag等（2003），再到Boas和Sag（2012）以及Michaelis（2013），

③ Sal是一头骡子的名字，取自一首著名的美国民谣歌曲*Erie Canal*（伊利运河）。
④ 关于BCG的介绍，可参考Fillmore（2013），高波、石敏（2010），王寅（2011）。
⑤ 从影响范围和参与学者的数量角度看，面向计算的形式化构式语法研究在语法学界和NLP领域都应算是相对小众的。最主要的研究力量来自所谓的美国西海岸语法学圈（如斯坦福大学和伯克利加州大学），CIG、BCG、HPSG、SBCG等形式语法体系的关系相对更密切，都出自这两所大学；ECG也是由伯克利加州大学的学者提出。FCG则由法国巴黎的索尼计算机实验室的研究人员提出。

以 Sag 等为代表的一批构式语法学者，在语法体系中不断提升“构式语法观”（也即“构式性”）的地位，最终发展到用“基于语符的构式语法”来概括他们的主张。从中不难体会 SBCG 的“心路历程”：从短语结构语法观发展到构式语法观。构式语法观的核心是强调语言的符号性，即“形式 – 意义”配对的“一次加工成型”⑥的性质。而短语结构语法的核心则是语言单位组合的推导性，即语言单位的形式层面和意义层面，在从小单位到大单位组合过程中，都可以用规则推导来描写。以 CFG（上下文无关文法）为代表的短语结构文法，就是表达这个推导过程最基本的工具。在 HPSG 的框架下，语法知识的组织方式是分为句法和语义两个模块来描写的，而在 SBCG 里，形式和意义成为构式（语符）不可分离的一体两面，构式语法的知识库就是一个构式清单，每条构式的形式和意义特征，均采用“特征结构”（feature structure）的描写形式来记录，而构式之间的关系，则通过类型化特征结构的层级体系来表示。

从语言知识的形式化表征工具角度来说，SBCG、ECG、FCG 等都普遍采用特征结构的形式，都主张以类型层级来表示构式的上下位关系，从而在知识组织方面，可以利用继承性提高表征的效率（下位构式可以承继上位构式的信息）。跟 SBCG 更关注具体的构式现象、⑦比较凸显理论语法学本体研究意识不同，ECG 和 FCG 直接从计算需求切入，注重计算机模拟人的语言行为，前者强调具身认知，即人体运动神经系统在语言理解中的作用；⑧后者强调语言系统的动态性，主张在言语主体的动态交互中维护和更新语法知识库。⑨ECG 和 FCG 的研究文献中提及的语言现象大多是常规语言现象，并不是狭义的实体构式。从这点上说，ECG 和 FCG 更多的是借鉴“构式”的语法观念来搭建形式化的语法模型，为计算机生成句子和理解句义服务。有关 ECG 和 FCG 的 NLP 算法和程序实践及其最新进展，均可访问官网⑩了解。近年来，随着深度学习在 NLP 领域大行其道，也有学者开始尝试在构式语法建模中引入分布式语义表示，如 Rambelli 等（2019）提议的分布式构式语法（Distributional Construction Grammar，DisCxG）框架。限于篇幅，这里不再展开。

⑥ Goldberg（1995）对于构式的定义强调构式意义不能从其构成成分的意义严格预测（not strictly predictable），实际上就是要求形式和意义的“直接”绑定关系，而不是将语言单位的形式（生成）和意义（理解）诉诸一个“可推导的”（derivational）过程。

⑦ Sag（2012）提到，尽管 HPSG 在英语资源语法（English Resource Grammar，ERG）建设中得以应用，有长期的实践积累，但 SBCG 的工作还主要在语言学理论分析层面，尚未大规模转化为语言工程。有关 HPSG 的语言工程成果，可以访问 http: //lingo.stanford.edu 和 http: //www.delph-in.net 查询。

⑧ 代表性的研究如 Feldman 等（2010）在语言神经理论的基础上进行 ECG 建模。值得一提是，ECG 研究文献中举自然语言句子的实例展开分析时，一般都是跟运动有关的句子，如“The horse raced past the barn.”（那匹马跑过了谷仓）“The cat jumped down”（那只猫跳了下来），引自 Bergen 和 Chang（2013）。

⑨ FCG 的官方网站（https: //www.fcg-net.org/）收录了 47 个语法构式。FCG 语法工程建设过程中还跟 FrameNet 知识库结合，借助框架网库的知识资源来描写构式。数据资源下载地址：https: //www.fcg-net.org/demos/FrameNet/Downloads.html。关于 FCG 的介绍，可参考 Steels（2013）。

⑩ ECG 的官方网站：https: //github.com/icsi-berkeley/ecg_homepage。英语 ECG 语法分析器可参考 Bryant（2004），汉语 ECG 语法分析器，有张灿（2010）的研究。有关 ECG 研究的最新进展，可参考 Feldman（2020）。关于 ECG、FCG，也可参考牛保义（2011）第 6、7 章，郑开春、刘正光（2010）。

2.2 面向计算的构式语言资源建设

要让计算机真正能处理语言中的构式现象，除 2.1 节中讨论的形式化语法模型外，更重要的还是建设计算机可用的构式语言资源，对狭义实体构式的构成成分、形式和意义特点、使用条件等诸多语言知识要素进行分项细致描述。这方面最具代表性的研究工作当属加州大学伯克利分校的 Fillmore 等学者主导的基于 FrameNet 项目[11]开展的 FrameNet 构式库（Constructicon）的建设。FrameNet 网站上公布的构式库中描写了 73 个英语构式的信息，标注了 1 481 个例句（平均每条构式 20 多个例句）。构式库中的具体信息包括 7 个部分（参见 Fillmore et al.，2012）：（1）构式名称；（2）构式形式表征（母亲节点符号 + 女儿节点符号）；（3）构成成分的语义范畴；（4）构成成分的句法范畴；（5）构式实例；（6）构式整体的句法范畴；（7）构式的语义解释（类似传统词典释义的方式，并非形式化的释义）。

英语 FrameNet 构式库还影响到其他语种构式库的构建工作，包括瑞典语、巴西葡萄牙语、日语、俄语、德语构式库等（参见 Lyngfelt al.，2018）。不过跟英语构式库类似，这些构式库的规模都还不大，比如德语 GCon 构式库描写了 39 个构式，瑞典语 SweCcn 构式库描写了 400 条构式，巴西葡萄牙语 FN-Br 构式库描写了 289 个构式。这些构式库在数据库架构的设计上都或多或少地参照了伯克利英语 FrameNet 构式库的建设经验，并结合语言自身的特点进行了改造。

近年来，抽象语义表示（Abstract Meaning Representation，AMR）[12]在各项语料标注任务和语义分析评测中逐渐受到较多的关注。AMR 尝试绕开句法，用单根有向无环图直接对句子的语义进行标注：图中节点代表概念，节点间的连线上加注角色标签，代表概念间的关系。AMR 可以看作在早期宾州命题角色标注语料库基础上的扩展体系，即从句子中核心谓词的语义角色标注，扩展到句中更多类型的词语成分所代表的概念间的二元关系的标注。在实践中 AMR 的标注人员发现，仅以词间关系来表示句义，会碰到难以跨越的障碍，即语言中的构式，无法仅通过词间关系的描述得到整句的确切语义。例如英语中表示比较义的构式实例“The girl is taller than the boy”（这个女孩比这个男孩高）就需要借助新增构式，即虚拟一个“比较构式”概念来表达。Bonial 等（2018）用“Have-Degree-91”作为比较构式的概念节点，设置了 6 个构式成分（角色）来表征比较的语义：[Arg1：比较主体；Arg2：比较项；Arg3：程度；Arg4：比较客体；Arg5：最高级；Arg6：结果]。[13]由此，上面例句的 AMR 表示为：(have-degree-91: Arg1(girl): Arg2(tall): Arg3(more): Arg4(boy))。值得一提的是，Bonial 等（2018）文章的副标题是“The more we include，the better the representation”［（构式条目）我们囊括得越多，（语言知识）表示得越好 ］，[14]这

⑪ FrameNet 工程的理论基础是 Fillmore（1982）提出的框架语义学。有关 FrameNet 项目，可参考 Baker 和 Fillmore（1998），或访问 FrameNet 官网（https: //framenet.icsi.berkeley.edu/fndrupal/）。

⑫ 关于 AMR 的体系，可参看 Banarescu 等（2013），或到 AMR 官网 https: //amr.isi.edu/ 了解详情。中文 AMR 的研究近年来也开始关注构式的处理，可参看黄彤等（2020）的研究工作。

⑬ 在 AMR 的命名规范中，“XXX-91”中的 91 为非词汇概念的常用编号；Arg（论元）是角色简称，后接编号 1、2……等用于区分不同的类型。这种编号方式借鉴自命题库（Propbank）中的角色命名方式。

⑭ 可以大致意译为：（构式）收录得越多，（句义）表达得就越好。

本身也是一个英语中的典型构式，文中记作“The X-er，The Y-er”构式，给出了这个构式的 AMR 表征框架，把 X 和 Y 处理为该构式的两个论元 Arg1 和 Arg2。从这个副标题的寓意不难看出，AMR 力图将更多的构式当作跟词一样的语法单位来表征其语义，为构式的字面意义的表征提供一个语义解释方案。

NLP 中有关多词单位（Multi-Word Expression，MWE）的研究，也可归属于构式数据资源建设相关的工作。Constant 和 Eryiğit（2017）在长达 57 页的 MWE 研究综述中，详尽示例说明了 MWE 所具有的特点，包括：共现性、非连续性、语义非组合性、歧义性和可变性。MWE 可以看作是一类图式化程度低，实例化程度高的习语性构式，例如“ad hoc”（特别的）、“part of speech”（词类）、“hand out”（分发）等。不过，有关 MWE 的工作有很强的语言工程色彩，虽针对的语言现象跟构式明显有很大交集，但似乎缺少理论构建的兴趣，游离于主流的构式语法研究之外。

❸ 现代汉语构式知识库的构建及构式的语义计算

汉语学界自觉借鉴国外构式语法理论对现代汉语构式展开系统的研究已有较长时间并取得了很多成果（张伯江，2008，2018），但将构式语法的理论研究成果用于中文信息处理的研究却长期处于空白状况（张娟，2013）。自 2013 年以来，我们在北京大学现代汉语语法信息词典以及现代汉语句法结构树库的语言工程建设经验基础上，开始了现代汉语构式知识库的构建工作，尝试为中文信息处理补充现代汉语构式的语言学知识资源，同时也可以对汉语构式语法理论的研究提供更多语料，推动理论研究。[⑮] 将构式的理论分析成果转化为数据，首先就需要界定构式的收录标准，制定构式的形式和意义表示规范（可参看詹卫东，2017）。从跟传统的语法单位（词和短语）的性质对比角度，我们可以确定作为语法实体的构式的性质并相应地制定可操作的构式判定标准和收录标准，如下面表 1 和表 2 所示：

表 1　从句法结构四要素对比短语和构式

结构要素 / 语言单位	关系	中心	范畴	层次
短语	+	+	+	+
构式	-	-	-	-

表 2　语法单位的组合性与递归性

语法单位	组合性	递归性
词	-	-
短语	+	+
构式	+	-

语法是组词造句的规则。词与词组合形成更大的结构时，必然存在一定的关系，结构中的内部成分一般会有主次之分，主要成分就是结构的中心，结构内部成分因替换性和扩展性的能力差异自然聚合成类，形成不同的语法范畴，语言成分的组合由小到大，就形成结构层次。表 1 中的结构四要素，是表示一个语法结构必然涉及的四个方面。一个语法系统，对普通的词组（即短语结构）在这四个方面均需做详

⑮ 北京大学现代汉语构式知识库（暂记为 CCL-CxnBank）可访问 http: //ccl.pku.edu.cn/ccgd 查询详情。

细描述，通过不同取值区分出不同的短语类型（如主谓结构短语、述宾结构短语、定中结构短语等等）；跟短语相对，构式在这四个要素上的表现均出现退化或弱化，构式不凸显内部成分之间的关系，构式内部没有地位突出的中心成分，因缺乏递归能力，构式内部成分的范畴感不强，也难以形成复杂的层次构造（树结构）。因此，构式可以简单分析为词语的线性序列组合。在表 2 中，构式的性质进一步概括为有组合性（构式是词的线性组合）而无递归性，因此既不同于词，也不同于短语。在构式知识库中，不同构式需要像词一样，逐条列举；同时又要像短语一样，给出组成成分的模式。例如“n 中的 n（天才中的天才）”“东 v 西 v（东想西想）”“v 就 v 吧（去就去吧）”⑯ 等，这些构式都是有限词语的组合，不像一般短语那样可以自由递归扩展，仅用词类标记（如 n、v 等）作为变项，跟常项成分（如“东”“西”等具体词语）一起，就可以描写构式的线性组合模式，进而给出构式的“形 — 义”配对或者说是语义解释。概括来说，我们对构式知识的描写策略，是采用短语结构语法已有的范畴体系（如词类、短语功能类），用简单线性序列模式来表征构式的形式，将来随着构式知识库建设的深入，再将构式（视作一类特殊的短语）融入短语结构文法体系中。

语法分析的最终目的是给出语言形式的语义解释。对构式知识的描写，也同样如此。根据我们目前对汉语构式的分析，面向计算的构式语义解释可以分为三种难度不同的情况：

（1）构式的语义相对简单，可以用普通短语结构的形式来对构式进行释义。例如“a+ 得 + 不行（厉害得不行）”“a+ 到 + 没有 + 朋友（帅到没有朋友）”“a+ 到 + 不 + 能 + 再 +a（细到不能再细）”“a+ 了 + 去 + 了（大了去了）”“a+ 出 + 新 + 境界（美出新境界）”等。这些构式的释义，均可以简单地用“非常 +a”作为释义模板，即把构式形式转写为普通短语结构的形式，然后融入短语结构的释义系统，完成整句语义的计算。

（2）构式的语义相对复杂，除字面义外，往往会激活相关的推论义且带有主观性，需要跟语境、常识等外在知识结合来表述其所在句子的完整意义。这种情况无法仅用释义模板给出构式的语义，需要使用释义框架，即复杂特征结构来列举构式的语义。例如：

[1] a 这么新潮的衣服，别说穿过，连见都没见过。
b 这么新潮的衣服，别说没穿过，连见都没见过。

例 [1] 中的构式可记作“别说 +X，连 +Y+ 都 +Z”。除字面义外，这个构式还

⑯ 这里列举的构式都是汉语中的“同形复现”构式，在北大现代汉语构式知识库目前收录的 1 074 条构式中，标记了“同形复现”特征的构式有 391 条（占 36.41%）。可以对比的是，FrameNet 英语构式库收录的 73 条构式中，具有同形复现（Tautology）特征的构式只有 1 个（因构式的语义有差异，FrameNet 构式库实际是分了 3 条记录来描写这个构式），形式为“noun+be+noun”（两个 noun 同形），如“a campaign is a campaign”（竞选就是竞选）、“the past is the past”（过去的就让它过去）、“men were men”（男人就是男人）等。

表达了一定的推理语义，即“这个衣服新潮的程度非常高，说话人从未见过这么新的衣服，因此，就更不可能穿过这样的衣服了”。例 [1a] 和 [1b] 的构式实例中 X 的形式刚好相反，a 为肯定式“穿过”，b 为否定式“没穿过”，但例 [1a] 和例 [1b] 基本同义，因此，需要在构式的语义表征中说明：要先分析构式变项 X 的内部形式，如果 X 为肯定形式，则构式语义成分中应解释为 [~ X]（即 X 的否定义）；如果 X 为否定形式，则构式语义成分中相应的解释为 [X]。

例 [1] 的正反格式同义现象可以从“命题否定”和“元语否定”的差异角度加以解释，例 [1a] 中的“别说穿过”是命题否定，即在命题层面否定“穿过”；例 [1b] 中的“别说没穿过”则是元语否定，即在语用层面否定，表达“说没穿过”的语用条件不成立，这样说不合时宜。对此，可以通过例 [1] 的变换形式来佐证上述解释：

[1’] a 这么新潮的衣服，连见都没见过，更别说穿过了。
b ? 这么新潮的衣服，连见都没见过，更别说没穿过了。

例 [1’a] 成立，而例 [1’b] 不大能接受。原因就是其中的“别说没穿过”是“元语否定”，“没穿过”是引语性成分（quotation），即引用上文已经说过的话，这样的成分处于“连见都没见过”之后，跟其指代的被引成分距离比较远，会导致认知加工负担加重，就不大容易接受。而例 [1’a] 中的“别说穿过”是命题层的否定，既可以位于“连见都没见过”前面，也可以移到后面，不受顺序的影响。

（3）构式在使用中并不是固定的形式，而可能有较多的形式变体，难以用简单的线性序列模式来表征。这种情况下，需要先用一般短语结构文法的分析模式来分析整句，在得到句子句法结构的基础上再识别其中的“构式”片段，给出构式和整句的语义表示。例如下面例句展示的有“远距相关”特点的构式：

[2] a. 他们种的优质小麦成为市场走俏的产品，“**种多少卖多少**”。
b. 每天**想吃多少吃多少**，想喝什么喝什么，自然大是高兴。
c. 有**多少**根发梢便会传递**多少**缕柔情蜜意。
[3] a. 乱弹琴！只能有**多少报多少**，不能虚报也不能少报，怎么能这样胡来！
b. 让游客选食，**吃多少收多少钱**。
c. 大哥每个月给老九**多少**钱，也就给我们**多少**钱。

例 [2] 和例 [3] 中都包含了“X 多少 Y 多少”[17] 这一形式的构式，表示 X 事件涉及的量跟 Y 事件涉及的量存在倚变关系（参见陆俭明、王黎，2006）。这个构式中的常项“多少”的位置并不确定，一般在不同结构中处于语法性质相同的位置，如例 [2a]、[2b] 中“多少”处于宾语位置，例 [2c] 中两个“多少”分别跟量词“根”和“缕”组合为数量短语。但因为两个“多少”相隔较远，我们很难再按照线性组合的方式

[17] 这里仅仅是用“X 多少 Y 多少”来称呼这一构式，并不代表构式实例都能写成这样的线性序列。

去匹配整个构式片段，比较好的策略是在短语结构树分析的基础上，识别其中的“多少”及整个构式片段。

构式“X 多少 Y 多少”除表达倚变义之外，还有明显的主观量义。例 [2] 和例 [3] 表达的语义并不完全相同，例 [2] 有“事件 X 的量可以任意增大，且 Y 的量也随之增大”的意思，即 X 的量不受限制，Y 的量也不受限制，相当于“无论 X 多少，都可以（有）Y 多少”；例 [3] 没有这个意思，而是相反，即事件 X 的量受到限制，Y 的量也相应受到限制的意思，相当于“只要 X 的量定下来了，Y 的量也得定下来，Y 的量受限于 X 的量”。因此也可以说这个构式是多义的，其语义描写可分为两个层次，其中基本义为：事件 Y 涉及的量跟事件 X 涉及的量存在倚变关系；[18] 交际义则有两个义项：（1）事件 X 的量不受限制，因而 Y 的量也不受限制，Y 的量有不限量义（引申为主观大量义）；（2）事件 X 的量受限，因而 Y 的量也相应受限，Y 的量不能过大，Y 的量有限量义。根据 X 和 Y 之间的事理关系，Y 的量有时跟 X 的量（限量）等值，比如例 [3c]。

实际语料中跟构式“X 多少 Y 多少”同形的例句，可能并无上述构式义中的数量倚变关系，如例 [4]。还有一些用例，仅仅是句子中存在两个“多少”共现的情况，但无法按照“X 多少 Y 多少”构式的模式进行成分匹配，当然也更不会有语义上的数量倚变关联，如例 [5]。

[4] a. 掌柜的每个月**卖出多少买进多少**，都有一本账。
b. 去年全镇有多少缺吃少穿的？**出生多少死亡多少**？多少是老死病故的？多少是投河上吊的？

[5] a. 邱建伟是何等人，那**是多少人请多少次**都请不到的。
b. 每一个数字的累进，赤峰人民都要**付出多少艰辛多少**汗水。

例 [4a] 和 [4b] 中的“X 多少”和“Y 多少”是单纯的并列关系，二者没有数量倚变关系，应解读为独立的两个事件数量；例 [5a] 中的“是多少人请多少次”和 [5b] 中的“付出多少艰辛多少”都不是合法的直接成分，两个“多少”对应的数量当然也不存在倚变关系。

以上例 [1] 到例 [5] 所展示的构式形式和语义表征问题，对计算机自动分析构成了很大的挑战。比如要区分例 [2] 和例 [3] 的语义差异，要识别例 [4] 和例 [5] 并非构式用例，都非易事。在语言的实际使用中，实体构式并不孤立存在，而是跟普通短语结构共同组句成篇，因此，分析构式的语义，也必然要结合传统的词汇和句法结构分析。就工程而言，构式知识库的构建可以独立于词库和短语结构规则库，但要在 NLP 系统中真正发挥作用，仍需考虑不同知识源的融合。

⑱ 倚变方式由事件 X 和事件 Y 之间的事理关系决定。具体如何倚变，由常识或世界知识决定，这里从略。

❹ 结语

构式语法观强调的是从整体视角来认识语言单位，但这并不意味着对具体构式的分析要抛弃还原主义哲学的指引。从 NLP 的实践角度来看，对构式语法现象的分析仍应遵循跟短语结构分析一样的原则、走还原主义的分析道路。从来源角度说，我们认为，构式是脱胎于普通短语的，即“构式从短语中来”（詹卫东，2017）；而从使用角度说，构式跟短语是“你中有我、我中有你”的关系，构式的计算分析，最终应跟短语结构分析融为一体，即“构式应回到短语中去”。

从 NLP 中的构式研究现状来看，有关构式知识的表征理论和分析框架固然仍值得探讨，但建设更大规模的构式数据资源则显得更为重要。包括 FrameNet 构式库、CCL-CxnBank 构式库等在内，目前已有的构式数据库规模都不大，在 NLP 的实际应用中会有很大局限。在现有资源基础上，我们要尽快探索利用计算机辅助工具，构建大规模构式实例标注资源，进而滚动迭代，挖掘更多的可形式化的构式知识，特别是对构式内部成分的句法语义条件和所处外部环境的语用条件做出更精准的描述。

参考文献

- Baker, C. F., Fillmore, C. J. & Lowe, J. B. 1998. The Berkeley FrameNet project. In *The 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*. 86–90.
- Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M. , & Schneider, N. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*. 178–186.
- Bergen, B.& Chang, N. 2013. Embodied construction grammar. In T. Hoffman & G. Trousdale (eds.), *The Oxford Handbook of Construction Grammar*. New York: Oxford University Press. 133–148.
- Boas, H. C. & Sag, I. A. 2012. *Sign–Based Construction Grammar*. Stanford: CSLI Publications.
- Bonial, C., Badarau, B., Griffitt, K., Hermjakob, U., Knight, K., O’Gorman, T., Palmer, M., & Schneider, N. 2018. Abstract meaning representation of constructions: The more we include, the better the representation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. 1667–1684.
- Bryant, J. 2004. Scalable construction-based parsing and semantic analysis. *Proceedings of the 2nd International Workshop on Scalable Natural Language Understanding*

(ScaNaLU 2004). 33–40. .

❏ Constant, M., Eryiğit, G., Monti, J., van der Plas, L., Ramisch, C., Rosner, M. & Todirascuk, A. 2017. Multiword expression processing: a survey, *Computational Linguistics* 43(4): 837–892.

❏ Croft, W. 2001. *Radical Construction Grammar: Syntactic Theory in Typological Perspective*. New York: Oxford University Press.

❏ Dunn, J. 2017. Computational learning of construction grammars. *Language and Cognition* 9: 254–292.

❏ Feldman, J. A. 2020. Advances in embodied construction grammar. *Constructions and Frames* 12(1): 149–169.

❏ Feldman, J., Dodge, E. & Bryant J. 2010. Embodied Construction Grammar. In Heine, B. & Narrog, H. (eds.), Oxford Handbook of Linguistic Analysis. New York: Oxford University Press. 111–138.

❏ Fillmore, C. J. 1982. Frame semantics. In The Linguistics Society of Korea (ed.), *Linguistics in the Morning Calm*. Seoul: Hanshin Publishing Co..111–137.

❏ Fillmore, C.J. 2013. Berkeley construction grammar. In T. Hoffman & G. Trousdale (eds.), *The Oxford Handbook of Construction Grammar*. New York: Oxford University Press. 92–108.

❏ Fillmore, C. J., Lee-Goldman, R. R. & Rhomieux, R. 2012. The FrameNet constructicon. In H. C. Boas. & I. A. Sag (eds.), *Sign-Based Construction Grammar*. California: Stanford CSLI Publications. 309–372.

❏ Goldberg, A. E. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: The University of Chicago Press.

❏ Hoffmann, T. & Trousdale, G. (eds.), 2013. *The Oxford Handbook of Construction Grammar*. New York: Oxford University Press.

❏ Jurafsky, D. 1991. An on-line computational model of human sentence interpretation: A theory of the representation and use of linguistic knowledge. Doctoral dissertation, Department of Electrical Engineering and Computer Sciences, University of California, Berkeley.

❏ Jurafsky, D. 1992. An on-line computational model of human sentence interpretation. In *The Tenth National Conference on Artificial Intelligence*. 302–308.

❏ Lyngfelt, B., Borin, L., Ohara, K. & Torrent, T. T. (eds.), 2018. *Constructicography: Constructicon Development across Languages*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

❏ Michaelis, L. A., 2013. Sign-based construction grammar. In T. Hoffmann & G. Trousdale (eds.), *The Oxford Handbook of Construction Grammar*. New York: Oxford University Press. 109–122.

- Pollard, C. & Sag, I. A. 1994. *Head-Driven Phrase Structure Grammar*. Chicago: The University of Chicago Press.
- Rambelli, G., Chersoni, E., Blache, P., Huang, C. R. & Lenci, A. 2019. Distributional semantics meets construction grammar: Towards a unified usage-based model of grammar and meaning. In *First International Workshop on Designing Meaning Representations (DMR 2019)*. 110–120.
- Sag, I. A. & Wasow, T. 1999. *Syntactic Theory: A Formal Introduction*. California: Center for the Study of Language and Information.
- Sag, I. A., Wasow, T. & Bender, E. M., 2003. *Syntactic Theory: A Formal Introduction (2nd edition)*. California: Center for the Study of Language and Information.
- Sag, I. A. & Hans, C. 2012. Introducing sign-based construction grammar, In H. C. Boas & I. A. Sag (eds.), *Sign–Based Construction Grammar*. California: Center for the Study of Language and Information. 1–30.
- Sag, I. A. 2012. Sign-based construction grammar: an informal synopsis. In H. C. Boas & I. A. Sag (eds.), *Sign-Based Construction Grammar*. Stanford: CSLI Publications. 69–202.
- Steels, L. 2013. Fluid construction grammar. In T.Hoffmann & G.Trousdale (eds.), *The Oxford Handbook of Construction Grammar*. New York: Oxford University Press. 123–132.
- 高波、石敏，2010，构式语法家族概览。《外语学刊》(1): 57–61。
- 黄海斌、常宝宝、詹卫东，2020，基于高斯混合模型的现代汉语构式自动标注方法。《中文信息学报》(9): 1–8。
- 黄彤、李斌、闫培艺、戴玉玲、曲维光，2020，基于抽象语义表示的汉语构式标注与分析。《中文信息学报》(10): 1–9。
- 陆俭明、王黎，2006，开展面向对外汉语教学的词汇语法研究。《语言教学与研究》(2): 7–13。
- 牛保义，2011，《构式语法理论研究》。上海：上海外语教育出版社。
- 王寅，2011，框盒图：构式语法的形式化方案。《外国语文》(3): 42–47。
- 詹卫东，2017，从短语到构式：构式知识库建设的若干理论问题探析。《中文信息学报》(1): 230–238。
- 詹卫东，2000，《面向中文信息处理的现代汉语短语结构规则研究》。北京：清华大学出版社。
- 张伯江，2008，句式语法理论与汉语句式研究。载沈阳、冯胜利主编，《当代语言学理论和汉语研究》。北京：商务印书馆。497–507。
- 张伯江，2018，构式语法应用于汉语研究的若干思考。《语言教学与研究》(4): 2–11。
- 张灿，2010，《面向计算机的汉语体验构式语法试验研究》。北京大学硕士论文。

❑ 张娟，2013，国内汉语构式语法研究十年。《汉语学习》(2): 65–77。
❑ 郑开春、刘正光，2010，体验构式语法：认知语言学的形式化模型。《湖南大学学报（社会科学版）》(1): 57–62。

Computational Models and Resources of Construction Grammar: State of the Art and Future Prospects

Abstract: A construction is a conventionalized "form-meaning" pairing of words. For a long time, both constituency model and dependency model of parsing in NLP did not take constructions as the main subject of analysis. This paper investigates construction-oriented and construction-based studies for NLP in the past three decades, including the design of formal systems of construction grammars and the development of linguistic resources of constructions across languages. The authors claim that the knowledge representation of constructions and the knowledge representation of traditional grammar rules of common phrase structures should be integrated into one system, so as to provide more valuable linguistic support for NLP. The scale of the current resources of construction knowledge is still too small to meet the needs of NLP systems, so further efforts are needed to enlarge the scale of constructions and annotated corpus.

Key words: construction; corpus annotation; formal representation; linguistic resources; natural language processing

（责任编辑：胡旭辉、王思雨）