构式的形式与意义表征*
——语言数据资源建设视野下的构式研究

詹卫东

摘要 构式知识需要转化为数据资源才能为自然语言处理应用系统提供支持。为此，北京大学中国语言学研究中心近年来开展了构式知识库建设和构式语料标注的语言知识工程实践，基于一个构式知识库网页编辑平台，从句法、语义、语用三个层面描述构式的特征，以及构式间的关系（包括近义、反义、上下位关系等），目前描写的构式条目超过 1000 条。本文结合构式工程实践，讨论了构式与传统语法单位（词、短语）的关系，提出构式形式的线性表征方案，以及构式意义的释义模板表征和语义框架表征相结合的策略。本文认为，构式知识资源应跟词库和短语规则库知识资源相融合，共同为中文信息处理提供语言知识服务。

关键词 构式　形式表征　语义表征　知识库　语料库标注

1　引言

过去 30 年来，有关构式的个案研究和理论研究成果已经相当丰富（Hoffmann & Trousdale 2013；Hoffmann 2017；张娟 2013；张伯江 2018）。而在自然语言处理领域，现有的知识库和语料标注，针对的对象主要还是普通的常规词组和句子，缺少从计算的角度对构式进行比较系统的研究（詹卫东 2017）。特别是对构式的形式表征和语义表征等基础问题，还需要做深入分析，这样才有助于对已有的构式研究成果进行系统地整理并转化为构式数据资源，为自然语言处理提供更直接的支持。国外在构式的形式化表征框架以及构式资源构建方面已经做了一些探索（Boas & Sag 2012；Bonial *et al.*

* 本文研究工作得到国家科技创新 2030“新一代人工智能”重大项目（2020AAA0106701）和教育部人文社科基地重大研究项目(15JJD740002)资助。

2018；Lyngfelt *et al.* 2018）。而在汉语构式知识资源的建设方面，近年来我们也已经开展了一些探索工作，收集了现代汉语中的常见构式1000余条，在构式知识库网页编辑平台上，从句法、语义、语用三个层面描述构式的特征，并尝试描写构式间的关系（包括近义、反义、上下位关系等）。此外，为了更全面地反映构式在实际使用中的情况，我们还开发了一个构式语料库在线标注系统，对真实语料中收集到的构式用例，通过网页界面标注其内部成分和主观态度义信息（下文以CCL-CxnBank称述这个构式库系统[①]）。

本文是在上述构式知识工程实践中的一些思考。第2节通过跟传统的“短语”对照，尝试更为系统地说明我们对“构式”性质的看法；第3节讨论构式的形式表征问题；第4节讨论构式的意义表征问题；第5节概要介绍我们在构式知识库和构式语料方面的工作情况；最后第6小节扼要阐述了我们对构式知识资源建设面临的挑战和未来发展方向的认识。

2　从语言结构理论视角和语言知识工程视角对比构式与短语

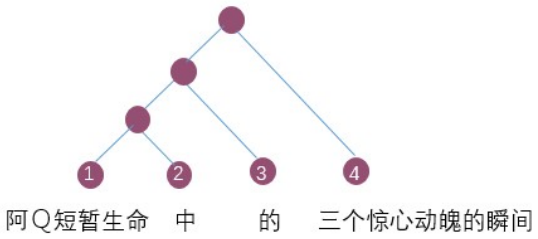
2.1 从句法结构四要素看构式与短语的对立

在语法学研究传统中，短语（词组）作为语法单位中的核心（朱德熙 1982，1985），对其结构描写，实质上包含了四个要素：关系、中心、范畴、层次。

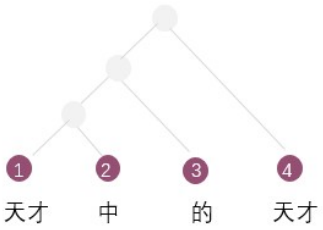
从形式文法（formal grammar）的角度说，语言的无限能产性，是通过短语的组合性与递归性来实现的。而在组织短语结构规则的形式化描写中，正是通过这四个要素的共同作用来刻画自然语言短语结构的组合性与递归性，即（1）词与词组合成短语，一个短语的内部成分之间就有一定的组合关系，不同的短语因组合关系的不同而有所区别；（2）在多数组合中，内部成分的地位有主有次，不同的短语因中心成分的不同而有所区别；（3）词与词的组合通常不是

孤立的，而是以类与类的范畴组合模式，呈现出系统规律性，不同的短语因其组成成分的范畴（类属）不同而有所区别；（4）词与词组合成短语后，短语之间又可以继续组合，表现为层级组合的特点，复杂短语因结构层次差异而有所区别。上述这四个要素，可以说是组织一个语言的短语结构文法体系的基础，作为句法描写的核心信息，可以在一定程度上，实现“用有限手段刻画无限的语言实例”的句法学目标。

不过，上述理论图式，是针对普通的规则性的短语结构设计的。如果从上述四个要素角度审视构式，就不难发现，典型的构式在这些要素上均与普通短语结构形成对立，即“去关系、去中心、去范畴、去层次”。即使没有达到完全去除这四个特征，也至少是弱化的形式，可以用“弱关系、弱中心、弱范畴、弱层次”来说明构式相对于短语的不同特点。下面图 1 是普通短语结构的内部层次构造模式（1—a）和构式的词语序列组合构造模式（1—b）的对比示意图。



(a)　短语结构示例



(b)　构式示例

图 1　普通短语结构与构式的结构模式对比示意图

图 1—a 的短语结构是语言单位的常规组合模式，体现了语法成分组合的层次性和递归性（recursiveness）。树结构的叶子节点（图中编号为 1，2，3，4）中，最左端和最右端的两个节点是 np（名词性短语）范畴，且 np 是可以扩展（自嵌套）的，如：阿 Q 生命中的三个瞬间→阿 Q 短暂生命中的三个惊心动魄的瞬间→阿 Q 令人惋惜的短暂生命中的三个不得不提的惊心动魄的瞬间……。图 1—b 是典型的构式例子，也由 4 个节点（词语）组合而成，但跟短语结构相比，这 4 个节点（尤其是位居两端的 1、4 两个节点）均无扩展性，因而也不再需要用树结构来表达层次性，其结构模式可以简单地看作是四个词语组成的线性序列。内部成分的结构关系，谁是中心成分，成分是什么语法范畴，同时都弱化了，在语法形式和语法意义的配对关系中，整个组合的“极端范例义、高程度义”得到凸显。这个语义，虽然可以溯源于“n+中+的+n”这一词类范畴组合的内部结构关系并由定中结构关系推导出来[②]，但在该构式实例的使用中，这个理论上假设的可能存在的推导过程已不再重要，如图 1—b 所示的树结构可以隐去，只留下表层的线性序列结构，而该结构模式又可以对应到像“公子的公子、教授的教授”[③]“一切的一切”“万一的万一”“但是的但是”等形式，这些用例相比于“n+中+的+n”格式，“中”已经脱落，且同形成分从名词 n 已经扩展到了数量词“一切”、副词“万一”、连词“但是”等，可见范畴约束已经弱化（或者说是泛化）了。不难体会到，这类表达形式的整个结构有一个整体的语义，独立于词存在。整个结构没有突出的中心成分，结构的形式特征（同形成分复现）和结构整体的语义高度固化（entrenchment），从形式到意义，不需再经过像普通短语结构那样的组合推导过程。

从“形式——意义”配对的角度来说，传统的词这一级语法单位也具有“构式性”。不过，作为语法单位的“构式”，跟“词”

还是有区别的，而且从语言工程角度讲，也有区分的必要：传统的语言单位分为性质不同的两类，一类是词这样的基础的“形——义”配对单位，无组合性和递归性；一类是短语这样的单位，由词组合得到的单位，有组合性，也有递归性。对语言单位做这样的划分，从语言工程上讲，是使得语言资源建设的效率比较高，易于维护和管理。其核心是：词的实例有限，可以逐项罗列；短语的实例无限，需要基于“组合性原则”，以有限范畴的组合规则模式，以简驭繁地描述无限实例。构式既不同于词，也不同于短语：构式有组合性，但无递归性。从实例层面来看，构式往往有固定的线性模式（不强调内部层次），有有限能产性，有一定的变形形式（参见第 3 节），其语义由内部成分词义和整体的构式语义共同构成。从语言工程角度来说，构式需要像词一样，逐条罗列，描述其有限能产性和变形形式，以及形式和意义之间的捆绑方式。

2.2 从树库标注实践中看构式与短语的对立

从树库（Treebank）语料中可以抽取组合规则（詹卫东 2013；Zhan 2016），如上节图 1—a 例句对应的规则就有：np→np !np；np →!sp u<的>；sp→!np f<中>等[④]。从规模约 130 万字的北大中文树库语料中抽取到这样的规则 1930 条，按频次从高到低排序后，前 446 条规则（占规则数量的 23.1%）覆盖了树库中 99%的语料。排在后面频次为 1 的规则共 441 条。除去其中的复句（fj）、整句（zj）规则 149 条，我们逐一检查了剩下的 292 条短语型规则，其中包含三种情况：有 100 条（占 34.25%）属于标注人员的误标[⑤]，有 63 条（占 21.58%）是常规短语组合[⑥]，剩下的 129 条（占 44.17%）则是非常规的组合模式，例如：

表 1　树库低频规则（非常规组合模式）示例

规则	频次	实例
dj→qp wco !np	1	dj(qp(两本)wco(，)!np(一块钱))
dj→vp !np	1	dj(vp(算一卦)!np(一块钱))

vp→sp !v	1	vp(sp(整体上)!v(看))
vp→!vp c tp	1	vp(!vp(今天开始)c(还是)tp(明天))
vp→!v np vp u	1	vp(!v(想)np(家)vp(想)u(的))
ap→!a u v u	1	ap(!a(多)u(了)v(去)u(了))

在树库加工过程中标注得到的如表 1 所示的组合模式，实际上就是语言中的所谓边缘（peripheral）现象。这些组合跟常规短语结构不同，往往是低频的实例，一般结构整体的长度比较有限，其组成成分可以有限替换，但缺乏像常规短语结构那样的递归扩展能力，结构体中的成分有一些明显的形式特征，如同形重复（“想家想的”“多了去了”），同时结构整体有明显的独立于词语的结构意义，或者说结构整体的意义解读容易让人产生结构中有成分省略的感觉，因为现有的词语的词义不足以解释整个结构所表达的意义。如表 1 中“算一卦一块钱”这个例子中的“算一卦”跟“一块钱”之间就隐含了“花费、消耗”的关系语义；“想家想的”这个例子整体上有“解释原因”的语义⑦，等等。

从 2.1 节提到的句法结构四要素的角度看，表 1 中这些例子都跟普通常规短语不同，具有“弱关系、弱中心、弱范畴、弱层次”的特点，是短语中的特殊的一类组合体，也即我们认为的构式语法单位。从树库语料标注的工程实践出发，也可以体会到，语言系统中的组合单位，可以划分出两类，一类是突出组合关系、层次性、强范畴性、强中心的递归性短语；另一类就是弱范畴、弱中心化、弱内部关系的线性化的构式。这也正印证了 Ronald Langacker 的说法“Language is a mixture of regularity and idiosyncrasy.”（语言是共性规则与个性习语的混合体，Langacker 1987：411）。共性规则（即短语结构）适合以短语功能范畴（如 np、vp 等）的树结构层级组织方式来描述，而个性习语（即构式）则更适合以常项加变项的线性序列形式来描述。

3　构式的形式表征：线性组合模型

绝大多数构式都可以常项加变项的线性序列方式来表示，其中常项是固定的词语，变项一般可以用词类范畴（如 n，v 等）或短语范畴（如 np、vp 等）来表示。上文提到构式成分有去范畴化或范畴泛化的特点，指的是有些构式的变项成分可以由不同短语范畴的语法单位来充当。下面表 2 的例子分别展示了变项由词、短语范畴和跨范畴成分表示的情况。

表 2　构式由常项和变项成分组合表示示例

构式形式	示例	常项成分	变项成分
n1+一+把+n2+一+把[⑧]	**鼻涕** 一 把 **泪** 一 把	一，把	n1，n2
一+百+个+vp	一百个 **不同意**	一，百，个	vp
X+就+X+吧	**等等** 就 **等等** 吧 **晚点儿** 就 **晚点儿** 吧 **差生** 就 **差生** 吧	就，吧	X（vp，ap，n…）

构式的实例在使用中也不是都严格遵循构式形式的要求，存在不同的形变，因此构式的形式表征也要考虑如何应对构式的变体形式（variation）。在构式知识库构建和构式语料标注实践中，我们发现主要有三种变异情况，分别采用了三种处理方式来应对：

（一）对于构式的常项成分有变异，且常项成分与变项成分连续出现，没有破坏构式整体的连续性的情况，在构式知识库中设置“构式变体”字段，穷尽性地列举一个构式的所有变体形式。

（二）对于构式整体出现对举或复用情况的，在构式知识库中设置“是否可扩展”字段，描述一个构式是限于单用，还是可以并列扩展使用。

（三）对于构式的变项成分有可能跨句法结构层级使用的情况，对这类构式的描述，更适合在短语结构规则中设置“构式激活成分”（Construction-Evoking Element，CEE，参见 Fillmore et al. 2012），在短语结构分析过程，如果碰到短语结构体内部包含 CEE，并且短

语结构间关系满足特定条件时，就激活该结构的构式解读。

下面通过对一些实例的分析来展开说明上述三种情况。首先看常项成分有变体的例子。

（1）a 该发生的，您拦也拦不住。

b 该发生的，您再拦也拦不住。

c 该发生的，您怎么拦也拦不住。

d 该发生的，您再怎么拦也拦不住。

例 1a 的构式形式为“v+也+v+不+X”，例 1b—d 均是这一构式的变体用例。在“构式变体”字段，可以填入“再+v+也+v+不+X”|“怎么+v+也+v+不+X”|“再+怎么+v+也+v+不+X”，穷尽性地列举全部变体形式[⑨]。

在目前构式库收录的 1066 个构式条目中，有 250 条填写了构式变体（占 23.45%）。不同构式的变体数量不一，变体间差异程度也不太一样。下面例 2 的构式基本形式为“a+就+a+在+X”，其变体形式跟例 1 相比，情况要更复杂一些。

（2）a 学习文言文难就难在一些实词和虚词上。

b 学习文言文难主要就难在一些实词和虚词上。

c 很多学生觉得文言文难，就是难在一些实词和虚词上。

d 文言文难，很多学生觉得就是难在一些实词和虚词上。

上面例 1 中的构式变体形式都还是一个完整的语法单位。例 2b 中的情况也是如此，在常项“就”之前多了一个词“主要”。这个成分也可以作为一个常项成分加入到“a+就+a+在+X”构式的变体形式中，不过，相对于例 1 来看，像例 2b 中的“主要”这样可添加的成分似乎较多，如“大概、可能、也许、肯定、很大程度上……”等等。不难体会到，这个构成的变项 a 和其后的“就+a+在+X”之间，结合相对比较松散。例 2c 中，第一个变项 a 后面有逗号，就足以说明这一点。例 2d 则除了逗号，还多出一个片段“很多学生觉得”把“a”和“就+a+在+X”分到了两个小句中。

可见，例 2 中“a+就+a+在+X”这个构式还没有凝固到铁板一块的程度，其中主要表达解释性状原因语义的部分是“就”后面的“a+在+X”，这个部分是相对独立的一个语块，可以跟前面的变项“a”分离使用。针对例 2 这样的构式变体形式，一方面是跟例 1 一样，尽量在“构式变体”字段列举出可能的用例模式，另一方面，构式库中要同时收录“a+在+X”作为一个独立的构式条目，并描述其跟“a+就+a+在+X”构式之间有同义关系。在计算机分析实际语料的句子时，按照最长匹配原则，“a+就+a+在+X”构式形式（及其变体形式）具有优先匹配权，只有在无法完全匹配时（如例 2c、2d），才退而求其次，选择跟“a+在+X”构式形式进行匹配。此外，也可以采取上面提到的第（三）种处理方式，通过构式激活成分设置，在常规短语结构分析过程中，识别“a……就+a+在+X”这样的远距离组合（参见下文例 4 的讨论）。

下面再看构式有扩展用法的例子。

（3）有人会问，这不是《教育法》和《科技进步法》早已规定的吗？有什么新意呢？它新就新在财政部门认真对待全国人大、政协“两会”代表、委员的意见上，新在他们转变作风、行动迅速上。

例 3 中构式的基本形式是“a+就+a+在+X”（同例 2），但其中“a+在+X”这一部分扩展了，在例句中出现了两次。在构式知识库中设置“组块扩展”特征来描述。缺省情况下，构式组块扩展特征值为“是”，即均可并列扩展。但也有一些构式没有扩展用法。例如：“**一个不留神**，摔了个大跟头。”其中“一个不留神”的构式形式为“一+个+vp”，对应“一个不留神、一个没站稳、一个手软……”这类表示出乎意料，突然发生（且造成不如意结果）的状况。这个构式没有扩展用例。实际语料中出现的能匹配“一+个+vp”的复现用例，都不是这个构式的用例，如“一个愿打，一个愿挨。”“一个使劲骂一个偷东西的孩子，还有一个在边上帮腔。”虽然例句中也有“一+个+vp”模式的重复出现，但或者是并列结构用法，或者

是在不同结构层次上的语法成分，都不是像“一个不留神”那样的构式用法。在构式知识库中，“一+个+vp”构式的“组块扩展”特征值取值为“否”，这样可以帮助计算机避免把这些例句误判为“一+个+vp”构式的实例。

最后看一组变项成分跨句法结构层次的构式例子。

（4）a　再多打几份工也要供孩子上学

b　你奉献得再多，那些人也觉得不够。

例 4a 是 CCL–CxnBank 中的复句型构式“再+vp1+也+vp2”的实例。例 4b 虽然从语义上讲也跟例 4a 一样，表达了反转关系（adversative），但例 4b 从形式上很难描写为线性序列的组合模式，其中的常项成分“……再……也……”，更像是插入到两个小句中来起到关联作用。例 4b 可以分解为命题 1：你奉献得多；命题 2：那些人觉得（你奉献得）不够。“再……也……”的作用是给两个命题赋值为转折关系（相当于“虽然……但是……”）。

考察更多的用例会发现，用“再+vp1+也+vp2”这个线性序列模式来描述这个反转关系构式过于简化了。实际用例中“……再……也……”关联的复句型构式可以用图 2 来示意：

图 2　“……再……也……”构式内部成分示意图

该构式中常项“再”比较确定，“也”所在的位置上可以出现一批副词成分如“都、总、还……”等等，而且还可能连用（如“仍然也……”）。另外图 2 中的变项成分 X、vp1、Y 也可能不是一个完整的连续的句法结构体，因此用线性组合形式来表示这个构式并不适用。针对例 4b 这种情况，可以按照短语结构规则分析方法对句子结构进行层次分析，同时识别其中的“再、也”等 CEE 成分，在句子组成成分满足特定条件的情况下，将“反转关系”这一构式语

义赋值给句子中相应的部分。换言之，如果句子被识别为构式实例，就不再是单独处理“再”和“也”的词义与句子整体语义的关系，而是把“再……也……”关联起来，将其激活的构式义与句子中其他部分的语义进行整合（关于具体分析方法，将另文讨论）。

4　构式的语义表征：释义模板与释义框架相结合

原则上，作为跟词、短语各有一定相似性的语法单位，构式的语义既需要向词的释义那样，逐条来定义“形式——意义”之间的“固定”联系，也需要像短语的语义表征那样，从内部成分的语义组合性角度，得到构式整体的语义，并将构式语义跟它所在的句子的其他部分的语义整合为完整的句义。

一般句子的语义分析遵循组合性原则（Principle of Compositionality），主要包括词汇的语义与句子结构语义的组合（Jurafsky & Martin, 2000: chapter 15.1, 15.2）。从计算的角度看构式语义的分析，也同样要遵循这一原则。构式语义的表征方式，一种思路是将构式近似地转写为常规短语组合，即采用线性的释义模板方式，给出一个构式对应的同义短语结构形式，然后将这个释义模板交给语义分析器，对其语义做组合性分析。这个思路相当于尽量借用传统的短语结构分析方式来处理构式的语义。另一种思路是针对难以用线性的释义模板来表示构式语义的情况，可以用复杂特征结构（feature structure），即释义框架来对构式语义进行描述，这种方式可以更细致更全面地描写构式的意义，同时也可以跟整句的语义分析融为一体。下面还是通过两个实例来说明这两种语义表征方式。

（5）贝多芬十一岁时，就已经显露了他的音乐天才，被认为是莫扎特第二。

例 5 中“莫扎特第二”是“n+第二”构式的实例。在 CCL-CxnBank

构式库中，该构式的释义模板描述为“像+n+一样”或“很+像+n”（两个模板任选）。这意味着例 5 可以被替换（改写）为“贝多芬十一岁时，就已经显露了他的音乐天才，被认为是很像莫扎特”，其中“很像莫扎特”是常规短语结构，就可以调用已有的针对常规短语的语义分析设计的模块来处理了。用释义模板来表征构式语义的做法，适用于像“n+第二”这样的语义相对简单的构式。下面的例子包含的语义比较复杂，需要由多个命题组合表达。

（6）有一种胜利叫撤退

例 6 是“有+一+种+X+叫+Y”构式的实例。这个构式带有鲜明的修辞色彩，来自介绍新概念新信息的格式，如果作为常规短语组合，其释义模板为：Y+是+一+种+X。比如“有一种中药材叫麻黄草”，意思就是“麻黄草是一种中药材”。Y 和 X 是种属关系，即从内涵来讲，Y 是 X 的下位概念，从外延来讲，Y 是 X 的子集。但作为构式的“有+一+种+X+叫+Y”，显然不再适合用释义模板来表示其语义，例 6 中的“撤退”无法理解为跟“胜利”有种属关系。“撤退”并非“胜利”的下位概念。例 6 的实际语义是：（1）“撤退”是表面现象；（2）“胜利”才是实质；（3）不要看表面，要看实质。这可以视为作为构式的“有+一+种+X+叫+Y”所具有的实际意义。

“有+一+种+X+叫+Y”构式脱胎于表达客观知识的短语组合格式，从表达功能角度说，可以称为“发布新知”构式，不过，跟原有的客观知识不同，作为构式的用法，总是发布“主观的的新知”，而且越新奇越好，在书面本文中，这个构式的实例往往是用于标题，可以很好地起到用奇谈怪论吸引眼球的效果。要完整地描述这些表达功能，仅用线性的释义模板就不够了。可以采用释义框架来表征。如下面图 3 所示（P 代表命题，@Y 表示引用变项 Y，@X 表示引用变项 X）：

$$\begin{bmatrix} P1: @Y\text{是表象} \\ P2: @X\text{是实质} \\ P3: \text{不要看表面要看实质} \end{bmatrix}$$

图 3　“有+一+种+X+叫+Y”构式的释义框架

值得注意的是，像例 6 这样的表义功能复杂的构式，其具体的语义解释往往要在上下文语境中才能最终确定下来。比如例 6 的语义虽然可以概括为图 3 所示的三个命题义的组合，但是说话人这样说的时候，到底是在肯定“撤退”行动，还是在否定（讽刺）“撤退”行动呢？就不清楚了。只能在具体的语用环境中，根据交际参与者的立场，以及所描述的“撤退”事件的具体前因后果来解读了。假如撤退的主体是以撤退为手段，诱敌深入，待敌人进入埋伏圈后一举歼灭敌人，在这种情景下说出例 6，那么就是在积极评价“撤退”这种计谋，最终带来“胜利”的成果。假如撤退的主体是在无力对抗的情况下出于报名而选择撤退，同时还嘴硬，为了面子，把撤退行动说成是胜利转移，那么在这种情况下，例 6 就是在讽刺撤退的主体，对其做否定的评价。目前，要对构式实例所在的语境进行形式化的表征还很难做到。而像例 6 这样的构式，其真实表达意图的识解又需要从语境特征中获取线索，为了更好地研究构式语义跟语境之间的互动关系，除了在构式知识库中采用框架释义方式表征其基本语义外，就还需要面向构式的真实用例进行标注，包括标注构式相关的评价、立场、情感等主观态度语义（参见下文第 5 节）。

5　构式资源构建：知识库与语料标注互动

CCL-CxnBank 构式知识库的描述信息分为五大部分：（1）构式基本信息；（2）构式内部成分信息描述；（3）构式整体句法功能描述；（4）构式整体语义功能描述；（5）构式所在语境特征描述。另外设置一个“参考文献”表，收集每个具体构式的研究文献。在上述五大部分中，目前填写信息比较全面的是“构式基本信息”数据表，包括构式的形式表征、释义模板、构式用例，构式形式特征和语义特征、构式类型等。CCL-CxnBank 现收录构式超过 1000 条。

这 1000 余条构式的基本信息表已经填写完成，其余部分的信息填写工作仍在进行当中。下面图 4 概括展示了构式库的五部分信息内容框架。

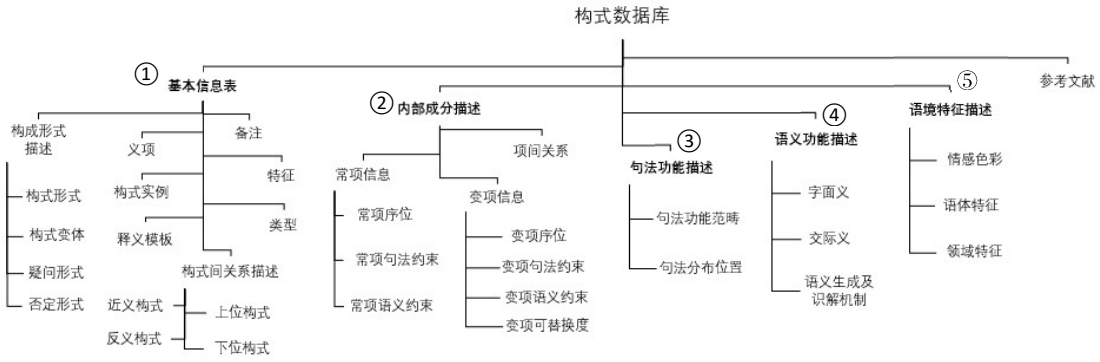


图 4　北京大学现代汉语构式数据库描述框架示意图

国际上已有一些语言的构式库在建设当中，但规模都不大，根据 Lyngfelt 等（2018）的介绍以及在一些构式库网站的资料显示，像英语 FrameNet 构式库中描写了 73 个构式的信息，标注了 1481 句（平均每条构式 20 多个例句）；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，等等。这些构式库大都参考了 FrameNet 构式库的设计框架。关于构式库中的信息内容，Fillmore 等（2012）以英语 rate（比率）构式为例做了说明，包括 7 部分：（1）构式名称（有助于称说和记忆）；（2）构式形式表征（母节点符号+儿子节点符号）；（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）。

CCL–CxnBank 对汉语构式的形式特征和语义特征做了比较详细的描写，比如“n+中+的+n”构式，其特征就标记为“复现、主观大量”，前者表示构式形式中有同形成分重复出现，后者表示构式语义中有主观大量的特征。对这个构式而言，主观大量又可进一步具体化为指 n 的典型性更为极端。限于篇幅，这里不展开讨论构式

库特征描写的细节内容。感兴趣者可以访问构式库统计信息页面查询具体的特征分布情况（http://ccl.pku.edu.cn/ccgd/stat/）。

下面仅举两例说明 CCL-CxnBank 在描述构式句法功能分布情况的做法。

（7）a 作者在这里打的是**一场又一场苦而又苦**的攻坚战。

b 平行或几乎平行长江修起的铁路**一条又一条**。

例 7 中包含了两个构式用例：“一 q 又 一 q”和“a 而 又 a”。前者的整体功能范畴是 qp（数量短语），可以用作定语修饰 np（如 7a），也可以作谓语，有一定的述谓性（如 7b）。后者的整体功能是 ap，但这个 ap 已经包含程度义，不能再受程度性状语的修饰。在构式库中，除了通过构式整体功能类标为 qp、ap 反映一个构式的主要分布特征外，还要以“作定语”“受程度状语修饰”等句法特征来详细描写一个构式在不同句法位置上的分布能力。图 4 中“句法功能描述”节点下面包含的“句法功能范畴”和“句法分布位置”这两部分，就是对构式的句法分布特征的细节描述。

从语言本体研究的角度考虑，CCL-CxnBank 构式库还设计了一些字段，考察一个构式对应的不同句法形式。比如一个构式是否有对应的疑问形式、否定形式用法等等。这些信息在图 4“基本信息表”中“构成形式描述”节点下。下面例 9 中的“连”字构式，就缺少对应的特指疑问句形式。

（8）a 张三也买了这本书

b 谁也买了这本书？

c 张三也买了哪本书？

（9）a 连张三也买了这本书

b *连谁也买了这本书？

c *连张三也买了哪本书？

例 8a 是普通的主谓句，可以有对应的特指疑问句形式 8b，8c。例 9a 是“连”字构式，没有对应的特指疑问句形式。例 9a 不能像例 8a 那样，从肯定句变换为对应的特指疑问句形式，例 9b、9c 均

为不合法的句子[10]。类似地，像“n+那+个+a+啊”（兴致那个高啊）、“n+倒+不+是+n”（坏人倒不是坏人）、“np+有+我+呢”（剩下的事有我呢）等构式也都没有疑问形式。这些跟构式用法有关的形式信息，CCL-CxnBank 中均做了描述。

下面再概要介绍 CCL-CxnBank 构式语料标注的设计方案，通过基于网页的标注程序，在网络环境下，由标注人员通过浏览器对构式例句进行标注，标注结果以 XML 文件形式存放在服务器端，标注后的语料可提供检索服务。现阶段已实现的标注界面功能包括两部分：（1）对构式实例的边界及内部成分的标注（如图 5 所示）；（2）对构式在语境中使用时的主观态度语义进行特征标注（如图 6 所示）。

图 5　构式实例常项、变项成分标注

图 6　构式实例主观态度义特征标注

图 5 中的例句为“一个人要是没有奋斗目标，别说干事业，连吃饭走道都打不起精神”。构式条目为“别+说+X+，+连+Y+都+Z”，其中 X、Y、Z 是变项成分，“别说、连、都”是常项成分。句子中“一个人要是没有奋斗目标”是构式外成分，即形成构式所在的语境。

我们从构式库中选取了 50 个构式，以短语型构式为主，每个构式在北大 CCL 语料库[11]中收集 100 个左右的例句，经过一些文字校对和整理，得到 5202 句作为标注对象。因为例句主要是以简单的字符串模式匹配方式，通过构式形式与语料中的例句进行匹配获取的，其中有少量的例句并不是构式用例，例如“有什么好 vp 的”构式，可以匹配得到例句“有什么好玩的地方”，但该句里的“有什么好玩的”是普通短语结构，它作“地方”的定语，并非构式用例。经统计，在标注句子中，真实构式用例为 4777 句，占比 91.83%。目前经过一轮主观态度义信息标注，结果如表 3 所示。

表 3　构式标注语料库主观态度义特征标注结果统计

总句数	评价		立场		情感	强度		
5202	正面	负面	接受	拒绝		极	很	不很
构式句	1141 23.89%	2223 46.53%	1204 25.20%	2111 44.19%	1238 25.92%	785 16.43%	1731 36.24%	1022 21.39%
4777 91.83%	3364 70.42%		3315 69.40%			3538 74.06%		

构式用例的标注是语言资源构建中一项难度较高的任务。无论是形式层面的构式成分标注，还是语义层面的构式主观态度义标注，都还存在不少问题。针对前者，我们已经探索了一些自动标注方法（黄海斌等 2020），随着标注语料的增多，未来有望通过机器学习训练更好的标注器；针对后者，后续的工作思路是将构式信息的标注和整句的句法结构标注以及句中谓词的论元角色标注等结合起来，形成整句丰富完整的句法和语义标注信息，并在此基础上进一步分

析构式内部成分与其所处语境成分间的互动关系。沿着这个思路，可以将汉语构式的资源建设跟已有的树库资源和语义角色标注资源融为一体。

6 结语

通过 CCL-CxnBank 构式语言资源建设工作，我们认识到，构式是语言系统中对常规短语结构的必要补充。相对短语的系统性来说，目前对构式的认识还并不系统，尤其是从语言工程角度来讲，相比已有的树库和语义角色标注语料库，构式资源库的标注规范，还有待更深入的研究。已有的一些构式资源构建工作（Lyngfelt *et al.* 2018）也能说明这一点。构式作为一个整体，更像是词汇系统，其内部组织方式的系统性还难以达到像短语结构那样的程度。

虽然构式主义的语法观比较强调整体性而相对忽视内部成分的组合语义分析，但是构式的整体“形——义”配对观念并不应该成为舍弃还原主义分析方法的理由。语义的组合原则仍然适用于构式的分析，在“形式——意义”之间建立联系，仍然需要考虑把形式进行成分分解，然后再做意义的建构。构式用例的内部成分标识和构式整体语义信息标注的设计，应该充分考虑这一基本原则，同时还应该结合语义的另一种分析路径，即情境性分析原则（Principle of contextuality），将构式成分语义的分析跟它所在语境的特征标注结合起来。

构式知识库和构式标注语料库，应该跟已有的语言资源兼容，跟包括树库（Treebank）、命题语义角色库（Propbank）和框架语义库（FrameNet）等形式的语言资源融为一体。充分利用已有词库和短语结构语法体系下的语料标注成果，以这种方式融合形成的新的语言数据资源，将会更有价值。

致谢：感谢北京大学中文系研究生黄海斌、唐乾桐、陈龙、

王佳骏在编写 CCL-CxnBank 网页程序、整理构式语料、统计标注数据方面做的工作。北京大学中文系现代汉语专业多位研究生先后参与了 CCL-CxnBank 构式库的填写、校对工作。在此一并致谢。

附　注

① CCL-CxnBank 构式库的网址：http://ccl.pku.edu.cn/ccgd。

② 由定中结构（语法）关系对应的“整体——部分”语义关系可以进一步细化为：“在一个具有 n 特征的集合中凸显 n 特征程度最高的成员”。

③ 学术界对陈寅恪先生的美誉。

④ 规则的表示形式为上下文无关文法（context free grammar）。箭头（→）左边为根节点，右边为子节点。箭头的含义为“由……组成”，即根节点代表的语言单位由子节点代表的语言单位组成。sp 代表处所词性短语，u 代表助词（如“的”），f 代表方位词（如“中”）。叹号!标记其后成分为中心语。

⑤ 如“从 2004 年开始”，其中“从 2004 年”是介词性短语（pp），“开始”是动词（v），整个结构标注为时间词性短语（tp），于是有组合规则：tp→pp !vp。按照标注规范，这个短语组合规则应该是 vp→pp !vp，即构成动词性短语（语义上表时间），而不是 tp。

⑥ 如“我们对这个问题比对那个问题更加重视。”其中“比对那个问题”是“介词（比）+介词性短语（对那个问题）”的组合模式，是介宾结构中的低频组合规则：pp→!p pp。

⑦ 例如在对话中，“甲：你眼睛怎么肿了？乙：熬夜熬的。”乙说的这句是“v+n+v+的”构式的实例，有“解释原因”的表达功能。

⑧ 这个构式有一个变体形式“一+把+n1+一+把+n2”，如“鼻涕一把泪一把”也可以说“一把鼻涕一把泪”。

⑨ 这个构式形式中的变项成分 v 和 X 的具体条件，还要在构式知识库的变项句法信息字段进一步详细描写。

⑩ 要对例 9a 中的宾语“这本书”发问，只能把它前移到句首，移出“连”字结构的管辖范围，形成“哪本书连张三也买了？”这样的特指疑问句。

11 http://ccl.pku.edu.cn:8080/ccl_corpus

参考文献

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》2020 年第 9 期，Vol.34,No.9，1—8 页。
詹卫东 （2013） 基于大规模中文树库的汉语句法知识获取研究，郑秋豫主编《语言资讯与语言类型》（第四届国际汉学会议论文集），台湾中研院，2013 年 11 月出版，239—267 页。
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》2017 年第 1 期，230—238 页。
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》，商务印书馆，北京。
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》2018 年第 4 期（总第 192 期） 2—11 页。
张 娟 （2013） 国内汉语构式语法研究十年，《汉语学习》2013 年第 2 期，65—77 页。
朱德熙 （1982） 《语法讲义》，商务印书馆，北京。
朱德熙 （1985） 《语法答问》，商务印书馆，北京。
Boas, H.C., Sag, I.A. (eds), 2012, Sign–Based Construction Grammar. Stanford, CA: CSLI Publications.
Bonial, Claire, Bianca Badarau, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Tim O’Gorman, Martha Palmer, Nathan Schneider, 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)
Chomsky, Noam, 1956, Three models for the description of language. *Transactions on Information Theory*, Vol. IT–2 No. 3 (1956), pp. 113 – 124.
Croft, William, 2001, Radical Construction Grammar : Syntactic Theory in Typological Perspective, Oxford University Press.
Croft, W. & Cruse, D. A. 2004, *Cognitive linguistics*. Cambridge: Cambridge University Press.
Fillmore, Charles J., P. Kay, M. O’Connor, 1988, Regularity and idiomaticity in grammatical constructions: the case of let alone. *Language*, Vol.64, No.3, pp.501 – 538.
Fillmore, Charles J., Russell R. Lee–Goldman, and Russell Rhodes, 2012, The FrameNet Constructicon, Boas, H.C. and Sag, I.A. (eds.) *Sign–based Construction Grammar, CLSI Publications*, Stanford University. 2012. pp.309–372.
Goldberg, A. E. 1995, *A Construction Grammar Approach to Argument Structure*, [M], The University of Chicago Press. 1995.

Goldberg, A. E. 2013, Constructionist approaches, Hoffmann, T. & Trousdale G., eds., *The Oxford Handbook of Construction Grammar*, Oxford University Press. 2013. chapter 2.

Hoffmann, Thomas, G. Trousdale, eds. 2013, *The Oxford Handbook of Construction Grammar*, Oxford University Press, 2013

Hoffmann, Thomas, 2017, The Renaissance of Constructions: from constructions to construction grammars, Barbara Dancygier, ed. *The Cambridge Handbook of Cognitive Linguistics*. Cambridge: Cambridge University Press, 2017.

Jurafsky, Daniel & James H. Martin, 2000, *Speech and Language Processing: An Introduction to Nautural Lanugage Processing, Computational Linguistics, and Speech Recognition*, ISBN:0130950696, Pearson Education, Inc.., Chapter 15.1, 15.2.

Kay, P., Fillmore, C.J., 1999, Construction Grammar and Linguistic Generalizations: The What's X Doing Y? Construction. *Language*, Vol.75, No.1, pp.1–33.

Langacker, Ronald, 1987, Foundations of Cognitive Grammar, Volume I. Theoretical Prerequisites. Stanford: Stanford University Press.

Li, C. N., Thompson, S. A., 1981, Mandarin Chinese: a Functional Reference Grammar. California: University of California Press.

Lyngfelt, Benjamin, Lars Borin, Kyoko Ohara, Tiago Timponi Torrent, eds., 2018, *Constructicography: Constructicon development across languages*, John Benjamins Publishing Company, Constructional Approaches to Language, ISSN: 1573–594X, 2018.

Zhan,Weidong, 2016, Peking University Treebank, in Rint Sybesma eds., *Encyclopedia of Chinese Language and Linguistics*, Volume 3, pp.332–336. Brill Publishing House, The Netherlands.

Form and Meaning Representation In Chinese Constructicography

ZHAN Weidong

Abstract: The linguistic knowledge of constructions needs to be transformed into computer readable data resources to support natural language processing applications. This paper introduces a Chinese

constructicon (CCL–CxnBank) and a corpus annotation platform for the description of actual usages of constructions in contexts. CCL–CxnBank is an online repository that contains more than 1,000 constructions, as well as the linguistic descriptions of their various features and the relations between constructions, such as synonymy, antonymy and hyponymy etc. Based on our practice of constructicography, we hold that constructions differ from phrases in that they are not recursive. We propose that the formal representation of a given construction should be linear, while its meaning should be represented through paraphrase templates and semantic frames. In the future, the knowledge resources of constructions, words, and phrases should be linked together to form a comprehensive linguisitc database for Chinese information processing.

Keywords: Chinese constructicon, Constructicography, Construction grammar, Form and meaning representation, Language engineering.

（100871　北京，北京大学中国语言文学系　zwd@pku.edu.cn）