一个汉语语义知识表达框架：广义配价模式*

詹卫东
北京大学中文系 100871

摘要 本文在已有的自然语言语义分析理论基础上，结合汉语信息处理的实践，提出了一个增强的汉语语义知识表达框架：广义配价模式，尝试在三个层级上描述汉语实词的语义知识，并把对实词的语义信息描述扩展到对短语的语义搭配信息的描述。

关键词 语义信息 广义配价 中文信息处理

A Framework of Chinese Semantic Representation: Generalized Valence Mode

Zhan Weidong
Dept. of Chinese, Peking University, Beijing China 100871

Abstract In this paper, the author puts forward an augmented framework of Chinese semantic representation, namely Generalized Valence Mode. Based on this framework, not only semantic information about lexical entries can be recorded on three levels, but also semantic collocation information about phrases should be described.

Keywords semantic information, generalized valence mode, Chinese information processing

一

关于自然语言的语义分析或语义知识表达方法，大致已经有语义场理论、义素分析理论、配价理论、格语法、论旨理论、概念依存理论、语义网络、蒙太格语法，等等。就大规模描述自然语言语义知识的工程实践而言，以英语和其他主要欧洲语言为描述对象的研究，影响较大的有 WordNet、MindNet、FrameNet 等等。中文信息处理方面则有“信息处理用现代汉语语义分类体系”、“现代汉语述语动词机器词典”、“董氏语义知识词典”等比较有代表性的研究工作。

本文尝试在这些研究工作提供的广阔背景空间下，宏观地考虑语义知识的效用及性质问题，并结合我们开发汉英机器翻译系统的实际经验，提出了一个汉语语义知识的表达框架：广义配价模式。

以下第二小节阐述我们对语义知识的效用和性质的总体认识；第三小节展开说明广义配价模式的具体内容；最后第四小节是结语。

二

就我们研制汉英机器翻译系统的实践经验来讲，语义知识的作用可以概括为两个主要的方面：（1）帮助句法分析得到正确的结果；（2）为发现句子中各个语言成分所对应的概念之间的意义关系提供支持（这可以类比为回答新闻中的五个 W 和一个 How 的问题）。比如，汉

* 本文的研究工作得到国家 863 课题支持（项目编号：863-306-zd03-03-4）。

语中有这样两个短语，a.“修理自行车的后胎”和 b.“修理自行车的师傅”。这两个短语都对应着线性序列模式 M：“V N 的 N”，这个序列在句法结构上至少有两种组合的可能性，即 A.[V [[N 的] N]] 和 B.[[[V N] 的] N]。要判断一个符合 M 形式的具体短语到底是按 A 切分还是按 B 切分，光靠组成这个短语的词的句法属性信息是不够的，需要有关于“修理”、“自行车”、“后胎”、“师傅”等词语的语义知识，即“后胎”是“修理”的【对象】，同时又是“自行车”的【部分】；“师傅”是“修理”动作的【发出者】，等等。基于这些语义知识，就可以判断出例 a 应该按 A 式切分，例 b 应该按 B 式切分。并且在得到句法分析的正确结果的同时，还能得到 a 和 b 的语义解释，a 中“修理”跟“自行车的后胎”之间是“动作——对象”的关系，“自行车”跟“后胎”之间是“整体——部分”的关系；而 b 中“修理”跟“师傅”之间是“动作——动作发出者”的关系，等等。

再来看语义知识的性质。实际上，语义知识跟句法知识一样，都是记录一个语言成分的搭配能力的。比如“修理”这个语言成分是“动词”（v），这个句法知识意味着，“修理”可以向后跟名词搭配形成述宾结构，可以向前跟“不”搭配形成状中结构，不能跟数量词“一个”搭配，等等；而“修理”是 2 价的，这个语义知识则意味着，“修理”可以跟两个名词性成分发生搭配关系。名词“师傅”的语义类是“人”，这个语义知识意味着“师傅”这个名词可能跟“修理”这个动词搭配并成为“修理”动作的发出者，“自行车”这个名词的语义类是“用具”，这个语义知识意味着“自行车”这个名词能跟“修理”这个动词搭配，但不是“修理”动作的发出者，而是动作的对象。不难看出，尽管语义知识跟句法知识一样都是反映一个语法成分的搭配性质的，但语义知识的抽象度比句法知识低。比如在句法层面，“修理”作为 v 可以跟任意一个 n 搭配构成述宾结构，而在语义层面，“修理”的对象则受到更多的限制，比如“修理”不能跟“气泡”、“感想”、“馄饨”等等搭配构成述宾结构，虽然这些词的词性都是 n。

三

从上面第二小节的扼要分析不难看出，要对自然语言的句子进行正确的分析，离不开关于语言成分的搭配能力的知识。从理论上讲，一个语言知识描述系统，应该把语言中任意两个成分间的搭配可能性纳入描述范围之内。对此，大部分信息实际上是由句法知识来承担的。以往的语义研究，注意力主要是放在了“动——名”语义搭配方面。

本文则在此基础上，把对搭配知识的描写加以扩展，形成了一个覆盖面更大的汉语语义知识表达框架：广义配价模式。这个框架的主要思想包括两个方面。一是除了描写“动——名”语义搭配性质，还有必要描写“名——名”、“形——名”、“动——形”等成分间的语义搭配性质；二是在描写实词的语义搭配信息之外，还有必要描写短语的语义搭配信息，因为在从词到短语的过程中，语言成分的搭配性质会发生显著的变化，记录这些变化对分析是有作用的。上述两个方面也就是在纵横两个方向上对已有的配价理论进行拓展。

在广义配价模式下，汉语“名、动、形”三大类主要实词的语义知识可以在下面这三个层级上展开描述。

（1）词语本身所属的语义类。

在词典中，对每一个具体的实词，需要描述其本身所属的**语义类**属性。比如名词“太阳”语义上属于[天体]类、形容词“红”语义上属于[颜色]类、动词“拿”语义上属于[搬移]类，等等。跟对词进行句法功能分类的目的类似，对词语进行语义分类，实际上也可看作是在刻画一个词的搭配能力。譬如，就名词来说，同属一个语义类的名词一般可以搭配形成无标记联合结构，如“太阳月亮”可以构成联合结构（sun and moon），其中的组成成分“太阳”跟“月亮”属于相同语义类。而不同语义类的名词一般不大能够搭配形成无标记联合结

构，而是倾向于形成偏正结构，如“阳光城市”构成偏正结构（sunny city），其中的组成成分“阳光”跟“城市”属于不同语义类。此外，语义类属性的另一个作用是使得搭配可以在词对类的概括水平上进行描述。比如“支持”这个动词可以搭配许多“个别”的名词，像“支持现任总统、支持张三、支持这个计划”等等。如果在更为概括的语义类层面上加以描述，就不需要一一罗列能跟“支持”搭配的对象，而是通过把这些个别名词概括为语义上属于“人类”、“抽象事物”等的办法来加以描述。我们希望词语语义分类的结果达到这样的功效，即不同语义类别的词语在句法上应该有显著的差异表现，而同属一个语义类的词语在句法上应该有显著的相似表现。不过，跟词类的情形一样，要达到这样的目标光靠分类是很困难的，对此可以在更多的维度上对一个词的语义性质进行区分。于是下面引入第二层级的语义知识。

（2）词语配价成分的个数，角色类型，及其选择限制。

这个层级是具体而直接地来刻画一个实义词跟其他实义词的搭配可能性。从理论上讲，词语的配价成分这个概念对动词、形容词、名词都适用，不过从重要性上讲，动词更为突出，动词的配价成分的情况也最为复杂。限于篇幅，这里我们主要以动词为例来说明如何组织这一层级的语义知识。有关论述对名词和形容词同样适用。

对动词来讲，**配价数**（即配价成分的个数）的取值范围为：0、1、2、3。**角色类型**分成两部分：一部分是核心语义角色，包括主体、客体、与事、工具、处所等；一部分是外围语义角色，包括空间和时间等。**对配价成分的选择限制**则要求指明一个动词对其配价成分有哪些条件限制（主要是从语义方面来进行限制）。

关于上述三个概念，有必要作些说明。

本文把配价数的取值范围定在 0—3 之间，像“例如”这样的动词，不跟体词性成分搭配，配价数为 0，像“送”这样的动词，可以联系三个体词性成分（如“张三送李四一本书”），配价数为 3。其他动词的配价数介于这二者之间，比如“咳嗽”只能跟一个名词发生搭配关系（“张三咳嗽”），是 1 价动词。“看”可以同时跟两个名词发生关系（如“张三看报纸”），是 2 价动词。但这样处理仅仅是一种简化的方式。关于配价成分应该如何记数，牵涉到对汉语动词整个配价系统的认识，同时也涉及到具体实施时的工作量问题。目前我们暂时走保守路线，先简单地把配价数定在 0—3 的范围内，操作上则以句法结构形式为主要依据。随着实践经验的积累，将来根据需要，可以进一步对配价数的取值进行调整。

再看角色类型。所谓“核心”和“外围”，只是表述上的区分，我们并不在术语（或语义范畴）的意义上使用这两个概念（即在规则中并不用到它们）。格语法对不同类型格的区分是所谓的“必选格”和“可选格”。在我们看来，就对汉语句法分析的制约作用而言，区分“必选”跟“可选”的语义角色类型，并没有特别大的意义。对“主体”、“客体”等等语义角色类型概念，本文不作严格定义。“主体”大致包括一般所说的“施事”、“当事”等语义格，“客体”大致包括一般所说的“受事”、“对象”、“目的”、“结果”……等等语义格。此外，跟有些语言学家的看法不同，我们不认为所有动词后表示空间方位的宾语成分都能算作动词的处所格。比如“喜欢南京”中的“南京”不能认为是处所宾语，也不能算作“喜欢”的处所类型的配价成分。理由是这里的“南京”在句法表现上跟“喜欢巴金的小说”里的“巴金的小说”是平行的，都是受事宾语，而不是跟处所宾语的性质接近。还有一点需要说明的是，我们认为应该区分动词对应的动作过程所在的“空间”和动作相关的物体所在的“处所”两个概念。在我们的表述框架中，前者是外围的角色类型“空间”，在句法形式上是以“在+宾语”介词结构做状语来标示的（如“在教室里唱歌”）；后者是核心的角色类型“处所”。在句法形式上除了可能以介词结构做状语来标示外（如“从屋里搬了出来”），还可以处所宾语来标示（如“（字）写在黑板上”、“（书）放桌上”），特别是以宾语来标示的情况，一定是标示“处所”，而非“空间”；一般来说，动作总是在一定的“空间”中发生，因此动词“空

间”属性的缺省值为“+”。实际表达中“空间”跟“处所”也可能重合（即有歧义）。比如“在院子里堆白菜”（比较“白菜堆在院子里”），“堆”的动作过程所在“空间”是“院子里”，所涉及的物体“白菜”所在的“处所”也是“院子里”。此外，一个动词能否带“处所”宾语，是由其词汇内容（lexical content）决定的。像通常说的“吃食堂”，“食堂”是“吃”的处所宾语，只是一种简单的处理方式。跟“睡床上”这种真正的“处所”比起来，“吃食堂”中的“食堂”的处所意味并不那么强就可以明显地暴露出来。与其说“食堂”是一种处所，还不如说它是一种用餐的“方式”更符合实际。因为，大学生几乎都“吃食堂”，但并不见得“饭都在食堂里”。事实上，有不少学生是“吃食堂的”，同时“他（她）们是把饭买回去，在宿舍里吃的”。而“睡床上”，决定了动作“睡”的主体所处的位置一定在“床上”。这种差别还体现在形式上，就是“吃”并不能带像“床上”、“大厅里”等含方位词的处所短语（如不说“吃大厅里”）。进一步深究“吃”和“睡”的词义差别，不难发现，“吃”的词义内容中，它的“主体”或者“客体”所在的空间位置并不是突显的（salient），而“睡”的词义内容中，它的“主体”状态最需要强调，同时主体所在空间位置也是突显的。像“睡”这样的主体位置突显动词还有如“来”、“去”、“回”、“坐”等，此外还有像“搬”、“写”、“晾”等客体位置突显的动词，这些动词都关注“主体”或“客体”的空间位置（即“处所”）。它们是能带真正的处所宾语的那一类动词。而像“吃”这样的动词，本文认为它不带处所宾语。对这种差别，也可以通过词语配价成分的变化来描述。这就是下面“广义配价模式”第三层级的语义知识。

最后，关于配价成分的条件限制，需要说明的是，搭配限制可以在句法层面描述，也可以在语义层面描述；可以是精确限制，也可以是概括限制；可以从正面来要求配价成分应符合某种条件，也可以从反面来要求配价成分应避免哪些情况。表达形式可以根据具体动词的不同情况灵活选择。如果强调精确度，就最好是词对词的搭配。如果要强调概括度，就描写词对类或者类对类的搭配。比如“后胎”是 1 价构件类名词，其“主体”名词限于“自行车”、“三轮车”、“汽车”等等带“车”字的交通工具词语。要描写“后胎”的配价性质，就可以强调精确度，用“汉字”特征来描述，而不用语义类特征来描述（参见下页表 2）。而动词“吃”能搭配的名词很广泛，要描写“吃”的配价性质，就可以概括描写为跟[可食物]类名词搭配。再比如动词“洗”，一方面我们可以从正面描写“洗”的客体配价成分应该是“具体事物”，同时也可以从反面限制“洗”的客体配价成分应该避免像“天体”类这样的“具体事物”（参见下页表 2）。

以上简要勾勒了对动词语义搭配性质进行描写的一般情况。形容词和名词也都可以按类似的方式进行描写。下表给出了目前我们描述三大类实词之间语义关系用到的范畴：

词类 / 词类	名词	形容词	动词	
名词	主体	主体 客体	核心语义角色	外围语义角色
			主体 客体 与事 工具 处所	空间 时间

（表 1：汉语实词语义关系范畴）

（3）词语配价成分的变化情况。

实际上，这里所谓的词语**配价成分的变化情况**，主要是指动词配价成分的变化情况。引入这个层级的语义信息，目的仍然是为了刻画动词的搭配能力。具体而言是描述动词跟补语成分（通常是形容词或动词）的搭配能力。比如“走”有一个义项是一价位移类动词，它可以搭配一个名词性成分，即“主体”配价，并且对“主体”配价的选择要求是“人”或“动物”等可移动的物体。这是上面两个层级描述的语义信息。如果考虑“走”的“主体”配价成分的变化情况，不难发现，经历了“走”这个过程，“主体”可能发生的变化是位置变化

以及性状变化，而这在形式上刚好对应着汉语结构系统中的述补结构。比如“走远了”、“走近了”是表示“主体”的位移变化的；“走累了”、“走跛了”则是表示“主体”的性状变化的，等等。此外，“走”还有一个义项是“离开”，在这个义项下，配价成分的变化情况就跟表示位移的“走”的情况有显著的不同，它的“主体”没有“远”、“近”等位移变化。这并不是说主体在物理意义上没有移动，而是指在说汉语者的心理认知中并不关注位移，而是关注“在场”还是“缺席”，比如这个义项下的“走”可以出现在“他已经走掉了”中，“走”跟补语动词“掉”搭配。这种情况下的“走”就不是“走路位移”的意思。另外，表示“离开”的“走”，其“主体”配价成分也没有“累”、“跛”等性状变化，但这时“走”可以跟形容词“光”搭配，即有“主体”的数量变化。比如“人都走光了”、“他们中已经走了三个”，等等。人之所以能够准确判断出“走”的不同意义，正是依靠了不同的“走”所出现的不同上下文，或者说就是不同的搭配。要让计算机也能对此进行准确分析，就要求事先记录动词“走”跟形容词“远、近、累、光”及动词“跛”等等的搭配关系。对此，我们是统一放在广义配价模式的第三个层级，即配价成分的变化层面来加以记录的。比如对于表示“位移”义的“走”，其配价成分的变化情况为：[主体变化：性状|位移]；而对表示“离开”义的“走”，其配价成分的变化情况则描述为：[主体变化：数量]。跟上面描述词语的配价成分的选择限制一样，描述动词配价成分的变化情况，也可以在不同的概括程度上进行。上面这样记录“走”的配价成分变化是在语义类的层面上进行概括描述。实际上也可以在具体词的层面上精确记录，比如对表“位移”义的“走”，还可以记录它的具体搭配对象“遍”（如“走遍了大江南北”）；对表“离开”义的“走”，也还可以记录它的具体搭配对象“掉”（如“他早就走掉了”）等等。值得指出的是，用这种方式，仍有无法区分的情况，比如“走完了”，“走”跟动词补语“完”搭配，其中的“走”既可能是表“位移”的意思，也可能是表“离开”的意思。对此，在广义配价模式的描述框架内目前还没有办法刻画。

实际上，动词配价成分的变化是普遍的现象。上面我们以 1 价动词为例进行了简单说明，对于 2 价、3 价动词，都可以用同样的方式刻画它们的配价成分的变化情况。这些变化在句法上大多是由汉语的述补结构来表达的。通过描述动词配价成分的变化情况，有可能为分析汉语的述补结构，特别是其中补语成分的语义指向提供一条线索。限于篇幅，我们就不多举例说明了。下表给出了一些例子，可以大致反映在“广义配价模式”的描述框架下记录汉语实词语义信息的一般面貌。

词语	词性	广义配价模式					
		语义类	配价数	论元角色选择限制		配价成分变化	
				主体	客体	主体变化	客体变化
大衣	n	服饰	0				
父亲	n	人	1	[语义类:人]			
后胎	n	构件	1	[汉字:*车]			
高兴	a	性质	1	[语义类:人]			
热情	a	态度	2	[语义类:人]	[语义类:人 | 事]		
洗	v	促变	2	[语义类:人]	[语义类:具体物¯天体…]	[性状]	{[性状],[原形:¯干]}
晾	v	促变|搬移	2	[语义类:人]	[语义类:具体物¯天体…]	[性状]	[性状 | 位置]

（表 2：以广义配价模式描述汉语实词语义信息举例 [1]）

以上是广义配价模式描述汉语实词语义知识的三个层级。除此之外，广义配价模式还特别强调在短语一级描述语言成分的语义信息。理由是在从词向短语的组合过程中，一个成分的语义搭配能力可能发生变化，描述这种变化，有助于计算机得到正确的分析结果。比如，

[1] 表中“*”表示通配符；“-”表示逻辑“非”；“|”表示逻辑“或”。为简化起见，省略了“处所”、“工具”等角色类型。“原形”跟“汉字”不同，是指某个特定的词本身。“汉字”指包含某个字的所有词。

动词“带”带上补语“来”之后组成 vp“带来”，其组合能力就发生了一些变化。“带”一般只能搭配表示具体物体的名词，如“带钢笔”、“带钱”等等。“带来”的搭配范围则扩展了，可以说“带来了一线希望”、“十月革命一声炮响，带来了马克思列宁主义”等等。我们可以把这种变化记录在“带来”的语义属性信息描述中，即“带来”的客体配价成分的选择限制放宽。这是语义搭配能力拓宽的情况。再比如动词“穿”带上介词补语“在”后，整个述补式 vp“穿在”可以搭配的客体成分也发生了变化，不能搭配普通名词成分，只能搭配处所成分。比如既可以说“穿西装”，也可以说“穿身上”，但只能说“穿在身上”，不能说“穿在西装”。这可以看作是语义搭配能力收缩的情况。在广义配价模式下，要显性地对短语“带来”、“穿在”等的这些语义搭配性质的变化进行描述。这有助于正确分析像“穿在身上的衣服”这样的短语。它的结构应该是:[[穿在身上 的] 衣服](译文:the clothes dressed in one's body）不能分析成： [[穿 在] [身上的 衣服]] (错误译文：wearing clothes of one's body)。后一种错误分析正是允许“穿在”跟普通名词搭配造成的。

四

广义配价模式最基本的思想是分层次描写实词的语义性质。这跟以词类信息加属性特征描述来刻画词语的语法性质的思想是一致的。目前我们已经在一个汉英机器翻译系统的语言知识库中手工完成了对 4 万多汉语实词（名、动、形）的语义类和配价信息描述，对动词配价成分变化情况的描写也已经开始了实验性的探索。我们希望在积累了更多的实践经验后，能更进一步完善这一语义描述框架,以及在具体贯彻实施方面探索由机器辅助抽取词语广义配价语义信息的可行性。恳请同行专家对本文的研究工作提出批评意见，帮助我们提高。

参考文献：

[1] Bake, C.F., C.J. Fillmore, and John B. Lowe. (1998) The Berkeley FrameNet Project. In Proceedings of COLING'98, 86-90.

[2] Miller,G., et al. (1990) Introduction to WordNet: an on-line lexical database. In International Joural of Lexicography 3, No.4, 235-244.

[3] Richardson,S.D.,William B.Dolan, and Lucy Vanderwende. (1998) MindNet: acquring and structuring semantic information from text. In Proceedings of COLING'98, 1098-1102.

[4] 陈群秀、张普（1995）《信息处理用现代汉语语义分类体系：属性分类》，载陈力为、袁琦 主编（1995）《中文信息处理应用平台工程》，电子工业出版社。

[5] 董振东（1998）《语义关系的表达和知识系统的建造》，载《语言文字应用》1998 年第 3 期。

[6] 李临定（1990）《现代汉语动词》，中国社会科学出版社 1990 年版。

[7] 林杏光 主编（1994）《现代汉语述语动词机器词典》，北京语言学院出版社 1994 年版。

[8] 鲁川（1995）《现代汉语的语义网络》，载《中文信息处理应用平台工程》，电子工业出版社 1995 年版。

[9] 沈阳、郑定欧（1995）主编《现代汉语配价语法研究》，北京大学出版社 1995 年版。

[10] 王惠、詹卫东、刘群（1998）《<现代汉语语义词典>的概要与设计》，载《1998 中文信息处理国际会议论文集》，清华大学出版社 1998 年版。

[11] 袁毓林（1998）《汉语动词的配价研究》，江西教育出版社 1998 年版。

[12] 詹卫东、常宝宝、俞士汶（1998）《基于词组本位语法的语义模型》，载新加坡《中文与东方语言信息处理学会学报》Vol.8,No.1,1998。

[13] 张普（1995）《信息处理用现代汉语语义分析的理论与方法》，载陈力为、袁琦 主编（1995）《中文信息处理应用平台工程》，电子工业出版社。