基于配价的汉语语义词典*

詹卫东　　北京大学中文系

摘要　本文介绍一个主要是基于配价理论开发的汉语语义词典（Valency_based Chinese Semantic Dictionary，文中简称 VCSD），分为四个方面：第一节简要交代这个研究工作的背景；第二节具体阐述一个汉语语义知识的表达框架；第三节说明这部语义词典的开发过程及目前达到的规模并展示若干样例；最后第四节是评价及对进一步研究工作的设想。

一 背 景

VCSD 是源于开发 TransEasy 汉英机器翻译系统的实际需要而研制的，因此这部语义词典有非常明确的实用目标。TransEasy 系统采用基于规则（rule_based）的路线，语言知识主要由规则库（syntax）加词库（lexicon）两部分组成。最初的词库中对每个汉语词及其对译的英语词都进行了语法信息描述(有关汉语词语语法信息的描述主要来自北京大学计算语言学研究所开发的《现代汉语语法信息词典》)。在这个汉英机器翻译系统开发的过程中，我们感觉到，计算机要分析得到汉语短语、句子的正确的结构，以及在多义词辨义时要能选择准确的译词，仅有语法信息是不够的。很有必要补充更多的能够区分词语间不同用法的信息。这样的信息通常也就是被人们称为“语义”信息的那部分语言知识。相应地，刻画词语这部分信息的知识库也就称为“语义词典”。

事实上，关于语义知识的描述方式（从某种程度上就可以看作是语义词典的组织方式）已经有格语法、配价语法、论旨理论、语义网络、语义特征描写、框架语义学等等不同的策略可供选择。这些方式从理论上和具体操作上讲都有些不同的特点。在我们看来，语义信息跟语法信息类似，也是用来描述一个词跟其他词的组合可能性的（包括能否组合，以及以什么样的关系组合等等）。无论是语法理论也好，语义理论也好，直接目标实际上只有一个，即把任意两个词(以及其他更大或更小的语言成分)之间可能存在的区别描述出来。比如“洗”跟“晾”，这两个词在语法上有共性，都是“动词”，都能作谓语，带宾语，不能受数量词修饰等等（这些实际上统统可以归结为“能否跟某些其他的语言成分组合搭配”）。但这两个词又有区别，比如动词“晾”可以带“阳台上”这样的处所宾语，“洗”则不行。对计算机自然语言处理来讲，“晾”跟“洗”的这些区别需要显性地加以刻画。通常人们就把有关“晾”跟“洗”的上述差别的知识称为语义信息。很显然，语义信息比语法信息的概括度（或抽象度）显得要“低”一些。像已有的格语法、配价语法等等语义描写理论，实际上都是在如何刻画词语之间的这些差别上作文章罢了。各种理论所能期望的最好境界也无非是：以最适合的概括度统摄最琐碎的有可能存在的区别。但现实情况往往不能尽如人意。人们会在“格语法到底应该为一个语言确定几个格”，“配价语法到底应该怎么来算一个动词的价”等等问题上争论不休，难有定论。

由此我们认为，（1）在本体论的层面构建一个人类自然语言的概念体系固然意义重大，但从服务于句法分析这样实际的目的出发，仅在方法论的层面上考虑如何选择有效的语义范畴，来区分那些最需要区分的词语之间的不同搭配特征，似乎更务实；（2）不同语义理论实

* 本文研究工作得到国家 863 项目支持（编号：863-306-03-06-2）

际上应该有共同的追求，即描述任意两个语言成分之间的搭配可能性，譬如动词与名词之间，形容词与名词之间，名词与名词之间，动词与形容词之间，动词与副词之间，……等等。但限于客观条件，我们只能量力而行，集中力量描写某些词语之间的搭配关系。

二 汉语语义知识表示的一个理论框架：广义配价模式

VCSD 的语义知识表达框架基本采用了配价理论的模式，同时极大地简化了一般格语法的语义角色系统。我们把这样的一个语义知识表达框架称为“广义配价模式”（Generalized Valency Mode），包括下面四方面内容。

（一）语义分类体系

要对词语进行语义信息描述，首先需要有一个语义分类体系。语义分类的直接目的跟语法分类一样，实际上也是为了说明一个词的分布（即一个词跟另一些词的搭配可能性），并且可以使得描写词语的语义搭配能够在词对类的概括水平上进行。只不过语义分类可以看作是比抽象的语法分类具体一些罢了。

理想的分类需要满足排他性和完备性，但事实上，目前我们对语义的认识达不到这样的要求，即归入同一个语义类的词，总能找出它们在语义上的差别来，而归入不同语义类的词，也有可能找出它们在语义上的共性来。有鉴于此，我们并不奢望能构造一个四平八稳的语义分类体系，而是力求所建立的语义分类框架线索明晰，这样在针对每个具体的实词操作时能更方便一些。因此，我们强调语义分类跟句法分类保持较高的对应性，同时语义层级的深度和宽度都处在一个比较简单的水平，便于实际的归类操作。比如，跟名、形、动三大语法类相应，我们的语义分类体系最上级也分为“事物”、“性状”、“运动”三个类别。其中“事物”类的语义深度相对深一些（最多也只到 6 层）；“运动”类的语义广度相对宽一些（最底层有 10 个平行的小类，可参见本文附录“汉语实词语义分类体系示意图”。）

（二）动词、形容词、名词的配价数

配价数标示有可能跟一个实词发生依存关系的名词的个数。我们规定配价数的取值在 0 到 3 之间。例见下页表 2“以广义配价模式描述汉语实词语义信息举例”。

（三）实词对其论元成分的语义选择限制

按照格语法的描写框架，一个动词可能搭配的语义角色类型包括“施事”、“受事”、“与事”、“工具”、“处所”、“目的”、“来源”、……等等。这些统统可以称为一个动词的论元（Argument）。我们把论元的概念从动词扩展到形容词和名词，同时把格语法比较复杂的格系统进行简化，归纳成如下的语义关系范畴：

词类 / 词类	名词	形容词	动词	
名词	主体	主体 客体	核心论元	外围论元
			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）

这里的“主体”和“客体”含义相当广泛，比如“主体”包括通常所谓的“施事”（agent）、“当事”（experiencer），还包括了一价名词和有价形容词的配价成分（例见下页表 2）。对大多数动词，我们关注的是“主体”、“客体”、“与事”三个论元的情况，其他论元除非特别必要，一般不作记录（我们区分“处所”与“空间”为不同的论元，详见另文讨论）。

一个实词对其论元成分是有选择限制要求的，这样的信息有助于进行句法语义分析，因此有必要记录在词典中。选择限制可以在语义类的水平上进行，也可以在具体词语的水平上进行，视实际需要和具体词语的用法而定（例见下页表 2）。

值得一提的是，在我们的语义知识表达模式中，论元成分的个数多于配价数。实际上，这正是把格语法和配价理论杂糅到一起的一个结果。格的语义色彩明显，配价则是语法色彩更浓。与其说配价数是个语义范畴，不如说它是个语法范畴更恰当。只不过我们并不关心语义跟语法的差别，甚至走向了它的反面，把二者混到一块儿来加以描写。

（四）动词的配价成分的变化情况

以上有关实词的论元成分的语义选择限制的描述基本可以反映一个实词跟名词搭配的可能的情形，但对动词跟形容词的搭配可能性却没有任何说明，而这样的信息对分析汉语的述补结构又是能够提供帮助的。基于这样的考虑，我们设想通过描述动词的配价成分的变化情况，来间接地描写动词跟形容词的搭配能力。比如动词“走”有一个义项是 1 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”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，这时关注的是“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”的意思。另外，表示“离开”的“走”，其“主体”配价成分也没有“累”的性状变化，但这时“走”可以跟形容词“光”搭配，即有“主体”的数量变化[1]。比如“人都走光了”、“他们中已经走了三个”，等等。

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。下面表 2 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	境况	1	[语义类:人]			
热情	a	品格	2	[语义类:人]	[语义类:人 ｜ 事]		
走 1	v	自移	1	[语义类:人｜动物]		[性状 ｜ 位置]	
走 2	v	自移	1	[语义类:人｜动物]		[数量]	
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	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 [2]）

[1] 所谓主体数量变化，可以理解为作为“类”的主体集合中的元素的个数发生了变化：增加或减少。

[2] 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。

需要说明的是，关于动词的配价成分的变化情况的描述这部分内容目前还只是设想加上一些初步的实验，并没有得到大规模具体的实施。而“广义配价模式”这一框架的前三个部分，我们已经完成了相当规模的实践工作。下面作些具体介绍。

三 基于配价的汉语语义词典的实践

1 实施的方式

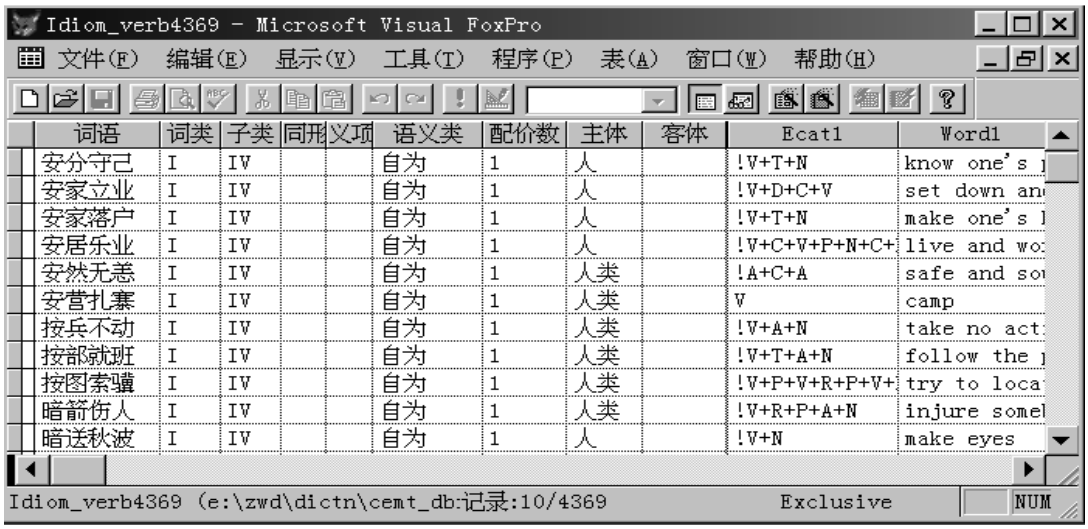
开发这部机器可读（machine-readable）的汉语语义词典，完全是由 TransEasy 汉英机器翻译系统课题组研究人员手工完成的，约 8 个人年的工作量。语义信息的依据则是研究人员个人的语感并辅之以已经出版的传统的面向人的汉语语文词典和百科知识词典。

开发这部语义词典的软件工具则是选用的微软 Visual Foxpro 关系型数据库。

2 目前的规模

目前 VCSD 包括六个数据库：动词、形容词、名词各一个；另外还有三个成语库。目前动词、形容词、名词和成语的记录数分别为 10788、2640、27828、7002，总共 48258 条记录。动词库中 0 价动词 9 个（如“例如”、“变天”等）；1 价动词 4782 个（如“拔河”、“降生”等）；2 价动词 6882 个（如“罢免”、“陷入”等）；3 价动词 115 个（如“拜托”、“赠送”等）。下面给出动词库和成语的部分样例。

3 样例[3]



四 评价及展望

1 对具有一定规模的 VCSD，我们的自我评价是：对提高汉英机器翻译系统的水平确有明显的帮助作用。下面看一个简单的例子。

例句： 安装网络的人正在想主意。

计算机在经过切词和词性标注分析后得到一个结果：

安装/v 网络/n 的/u 人/n 正在/d 想/v 主意/n

如要保证进一步分析的正确性和产生准确反映原文意思的译文，计算机需要对下面这两个问题作出判断：（1）“安装网络的人”是一个 “v n 的 n”歧义结构，即可能按 a. [[v n 的] n]组合，也可能按 b. [v [n 的 n]]组合。在上面这个例句中，正确的分析应该是哪一个？（2）例句中“想”是个多义动词，至少可以有“思考”和“思念”两个意思，在上面例句情形下，应该选择哪一个义项？

词典中对动词“安装”的配价信息描写可以解答问题（1）。因为“安装”的“客体”限制只能是“人为事物”，不能是“人”，因而“安装网络的人”只能按 a 方式组合。词典中对“想”的配价描写则可以帮助判断在上面例句的环境中，“想”应该是“思考”的意思，而不是“想念”的意思，因为后者要求其“客体”配价成分的语义类为“人类”，但这里“客体”是“主意”，语义类属“抽象事物”。经过上述分析，机译系统可以得到如下译文：

Man who installs network system is thinking of idea.

尽管上面这个例子显示了配价信息对正确分析汉语句子和得到比较准确的译文的作用，

[3] 因为我们的词典是面向汉英机器翻译的，所以是双语对译的。Ecat1 表示英语词类 1，Word1 表示译词 1。

但必须认识到，目前的配价词典仍有许多问题解决不了（参见詹卫东[1997]的讨论）。这里我们再在理论层面简要归纳一下我们的看法。事实上，语法（语义）分析要求能描述语言中任意两个成分的搭配关系。宽泛地讲，配价应该是任意两个成分之间的搭配，而不仅仅是局限在动词跟名词之间，尽管人们一开始认识到配价问题是从动名关系开始的。特别是在汉语这样无形态变化的语言中，不同词之间的搭配关系更显得重要，而且描写起来也更困难，比如动词跟动词之间的搭配关系就不好描述（如“游泳治好了他的感冒”，“游泳”、“治好”、“感冒”这些动词之间的搭配关系如何描述？），还有动词跟副词性成分之间的关系也没有纳入描述视野内（如“他认真地看了几遍”可以说，而“他认真地睡觉”就显得语义异常，如何对此进行解释？）。尽管上文提到可以通过描写动词配价成分的变化情况来间接达到分析述补结构复杂的内部关系的目的，但同时我们也注意到，仍有重重困难需要克服，比如“走完了”，就很难判断其中的“走”是表示“离开”还是表示“位移”。凡此种种，都是一种语法理论或语义理论需要去认真面对的问题。配价理论作为一种理论也不例外，对其效用及不足，都应该有清晰的认识，这是将来进一步改进的基础。

2 基于现有的 VCSD，可以进一步展开许多研究工作。比如对语料库进行自动配价标注，可以得到具有深层语义搭配信息的深加工语料库。这样的语料库作为一个语言资源又可以进一步支持多项语言学和计算语言学研究。现在的 VCSD 是基于研究人员语感，纯手工编成的，今后可以在有语义搭配标注的语料库基础上，探索由计算机自动抽取配价信息来编制基于真实语料的汉语配价词典，也可以用同样的方式建立汉英双语配价词典。

致谢：本文研究工作是 TransEasy 汉英机器翻译课题组集体劳动的成果。这个课题由中科院计算所和北大计算语言学研究所共同承担。课题负责人是中科院刘群副研究员和北大计算语言所俞士汶教授。先后参加配价词典开发工作的还有王惠博士、王斌博士、刘颖博士、常宝宝博士、于江生博士、左岩博士等，在此一并表示深深的谢意。

参考文献：

[1] 陈群秀（1996）《信息处理用现代汉语语义分类体系的设计思想》，载罗振声、袁毓林主编《计算机时代的汉语和汉字研究》，清华大学出版社 1996 年版。

[2] 陈小荷（1998）《一个面向工程的语义分类体系》，载《语言文字应用》，1998 年第 2 期；

[3] 刘群等（1997）《一个汉英机器翻译系统的计算模型与语言模型》，载 吴泉源、钱跃良 主编《智能计算机接口与应用进展》，电子工业出版社 1997 年版。（第三届中国计算机智能接口与智能应用学术会议论文集）。

[4] 孙宏林（1994）《信息处理用汉语语义词典的描述方法》，载《现代语言学·第三届全国语言学会议论文集》，语文出版社。

[5] 俞士汶 等（1998）《现代汉语语法信息词典详解》，清华大学出版社 1998 年版。

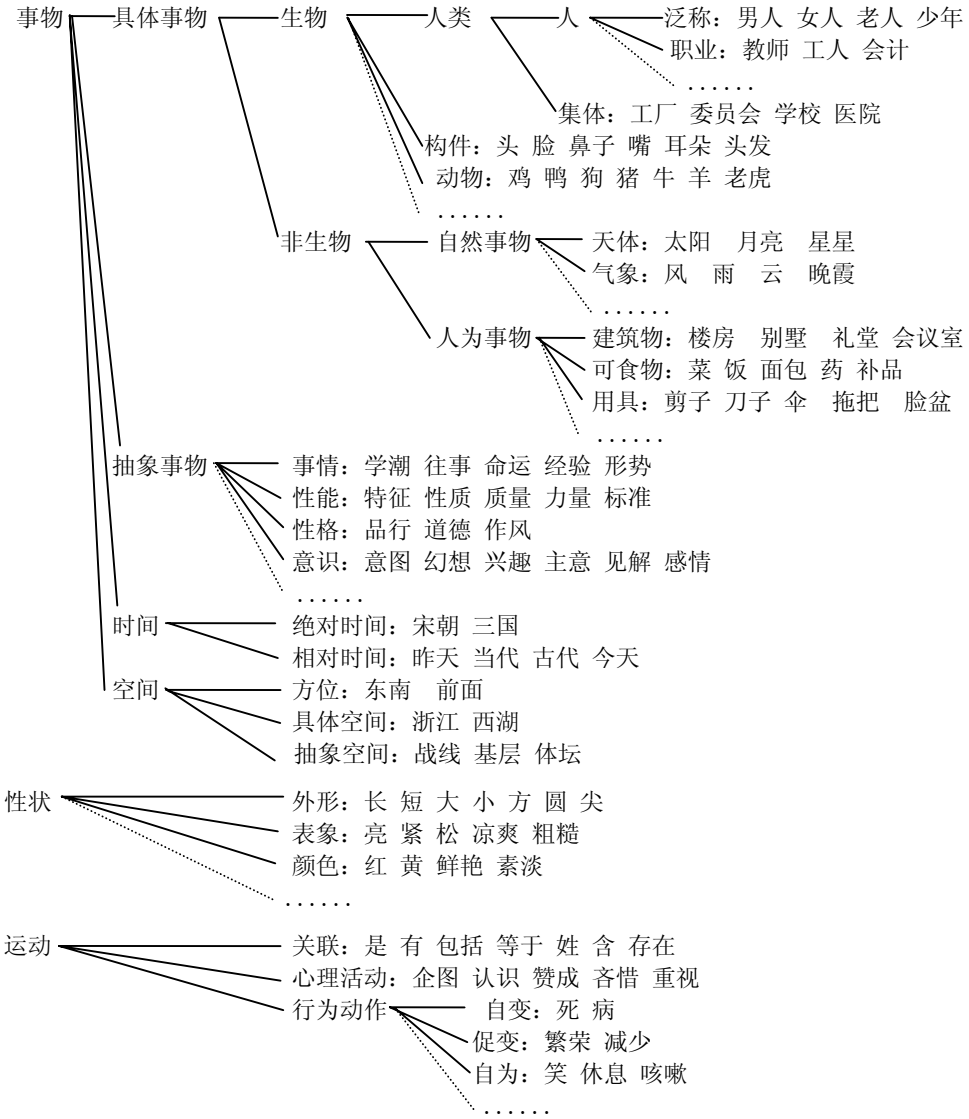
[6] 詹卫东（1997）《词的语义分类在汉英机器翻译中所起的作用以及难以处理的问题》，载陈力为、袁琦 主编《语言工程》，清华大学出版社 1997 年版。（全国第四届计算语言学联合学术会议论文集，JSCL’97）。

[7] 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《中文信息处理应用平台工程》，电子工业出版社 1995 年版。

[8] Fillmore, C. J. 1982. Frame semantics, In Linguistics in the morning calm, The Linguistic Society of Korea ed. Hanshin Publishing Co. Seoul, 111-137.

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

附录： 汉语实词语义分类体系示意图



（“:”后面的是例词；虚线表示大类中还有其他没列出的小类）

Valency-Based Chinese Semantic Dictionary

Weidong Zhan

Dept. of Chinese Language & Literature
Peking University
Beijing , 100871 China

Abstract This paper introduce a Chinese Semantic Dictionary mainly based on Valency theory(VCSD) . The paper consists of four parts: (1) The background of the development of VCSD; (2)A framework of representation of Chinese semantic information; (3) The implementation of VCSD; (4) Evaluation and prospect.