《中国学术期刊网络出版总库》及CNKI系列数据库入选期刊

二〇一八 第五卷 第一期

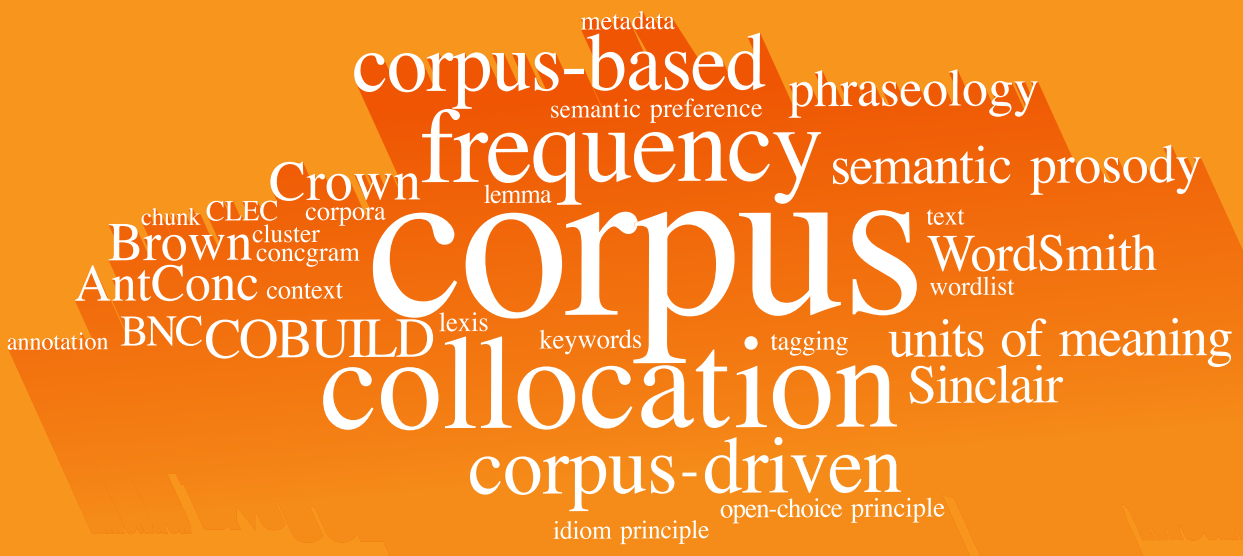
语料库语言学

CORPUS LINGUISTICS

1
Vol. 5 No.1
第5卷 第1期
2018

北京外国语大学中国外语与教育研究中心
中国英汉语比较研究会语料库语言学专业委员会
许家金 主编

语料库语言学



外语教学与研究出版社
FOREIGN LANGUAGE TEACHING AND RESEARCH PRESS

外研社

语料库语言学
（半年刊）

Corpus Linguistics
(Biannual)

主　管：中华人民共和国教育部
主　办：北京外国语大学
承　办：中国外语与教育研究中心
　　中国英汉语比较研究会
　　语料库语言学专业委员会
出　版：外语教学与研究出版社

Administered by the Ministry of Education of China
Directed by Beijing Foreign Studies University
Edited at the National Research Centre for Foreign Language Education and Corpus Linguistics Society of China, China Association for Comparative Studies of English and Chinese
Published by Foreign Language Teaching and Research Press

主　编：许家金
编　校：李晓雨　康　卉

Editor: Xu Jiajin
Proofreaders: Li Xiaoyu and Kang Hui

电　　话：（010）88816828
电子邮箱：bfsucrg@sina.com
投稿网址：http://ylyy.chinajournal.net.cn

本刊地址：北京市西三环北路19号北京外国语大学中国外语与教育研究中心
《语料库语言学》编辑部（100089）

《语料库语言学》

2018年　第5卷　第1期

目　录

学者聚焦

研究论文

书刊评介

CORPUS LINGUISTICS

Volume 5, Number 1, 2018

Table of Contents

基于AMR语料库的汉语谓词语义角色考察*

南京师范大学 宋 丽 闻 媛 葛四嘉 李 斌 周俊生 曲维光

提要：语义角色标注是自然语言处理的基础课题之一，目前自动标注的效果尚未达到实用水平。主要存在两大问题：首先语义角色颗粒度的大小不好确定；其次静态的谓词词典难以覆盖动态的语料标注问题，特别是缺乏对句子的语义角色省略情况的处理机制。因此，本文基于一种新的整句抽象语义表示方法（AMR）来研究谓词的语义角色问题，根据5,000句中文AMR标注语料统计出了谓词词典的动态覆盖情况；并通过与中文命题库（CPB）的对比发现，AMR能更完整地标注谓词的核心语义角色，且在语义关系的设置上做到了颗粒度粗细相融：核心语义关系颗粒度粗，非核心语义关系颗粒度较细，整体表征能力强；此外，允许增补概念的规定解决了语义角色省略的情况。最后得出，AMR作为整句的语义表示方法，在语义角色标注方面具备独特的优势，需要进一步加强AMR语料库的建设，为中文句子语义处理奠定基础。

关键词：抽象语义表示、谓词框架、语义角色、语言知识库

1. 引言

自然语言是人们表达意义的载体，语义的自动分析一直是自然语言处理领域的核心课题。而语义自动分析离不开大规模高质量的语义标注语料，因此，语义表示方法已成为自然语言处理领域的研究热点。在句义表示方面，谓词的事件框架所包含的各种语义关系构成了句子结构的主干，是语言学和自然语言处理关心的重点课题；谓词词典的编写和语料库中谓词语义关系的标注，已经成为相得益彰的研究范式，如FrameNet、PropBank和北大的中文网库。然而，谓词语义关系的抽象程度以及数量和颗粒度问题在学术界依然存在争议。FrameNet（Baker *et al.* 1998）根据语义为每一类动词设置一个框架，使用不同的框架元素标签，抽象程度居中，颗粒度很细，但语义关系数量庞大；北大网库（袁毓林 2007）基于传统的“格”，定义了10种必有论元（主体论元5种，客体论元5种）和11种非

* 本研究得到国家社科基金项目“中文抽象语义库的建构及自动分析研究”（18BYY127）的资助。感谢卜丽君同学在语料标注中的巨大付出。

必有论元（凭借论元5种，环境论元6种），抽象程度低，颗粒度和数量居中；而PropBank（Palmer *et al*. 2005）规定了5种核心语义角色关系（简称“核心语义关系”）和13种非核心语义角色关系（简称“非核心语义关系”），为每一个特定谓词分别建立框架，抽象程度高，但颗粒度很粗，语义关系少。而新近出现的Abstract Meaning Representation（抽象语义表示，简称AMR）句子语义表示方法（Banarescu *et al*. 2013）则采取了分而治之的方法，核心语义关系仍然只有5个，但非核心语义关系多达40个。同时，AMR允许在句中补充出省略的语义角色。所以，对AMR采用的这种新的谓词框架及标注方法的优缺点进行深入的统计分析和理论探讨具有较高的意义。

参考英文AMR的标注体系，Li *et al*.（2016）结合中文的特点，建立了一套适用于中文的AMR标注体系（Chinese AMR，简称CAMR），并已经构建出了中文《小王子》AMR语料库（李斌等 2017a）。CAMR沿用了英文AMR采取的OntoNotes（Yu *et al*. 2008）的标注体系。英文AMR使用PropBank中的谓词义项及核心论元框架，利用核心和非核心语义关系标签来标注概念之间的语义关系。核心语义关系是指谓词与其自身的事件框架中包含的若干语义角色之间的关系，包括5种（arg0-arg4）。非核心语义关系是指核心语义关系之外的语义角色关系，英文AMR规定了40种一般的非核心语义关系，CAMR根据中文标注的需要新增了4种。考虑到与AMR的兼容性，CAMR的非核心语义关系标签仍使用英文单词。CAMR标注过程中采用的谓词语义角色框架词典是从Chinese Proposition Bank（CPB）（Xue & Palmer 2009）标注语料中抽取出来的（Bai & Xue 2016），该词典含有每个谓词在不同义项下的语义角色框架，共收录了24,510个中文谓词（包括动词、形容词等）的26,650个义项的核心语义角色框架。

与传统的语义角色标注体系相比，PropBank的谓词核心框架抽象程度高，能够更好地表示谓词的核心语义角色，但PropBank的谓词框架在语言学界并未得到全面的认可，有学者认为其过于宽泛，不能很好地对语义角色做出分类。此外PropBank设置的非核心语义关系的颗粒度过粗，需要修改和增补。本文基于对CAMR的5,000句标注语料的统计分析，从CAMR使用的谓词框架词典（义项词典）、谓词语义角色标注体系（与CPB标注语料对比）、对汉语中省略核心论元的名词性结构的处理（如“的”字结构）等三方面论述AMR采用的谓词框架的合理性。

2. 语料的基本情况

我们根据Li *et al*.（2016）的中文AMR标注规范，从中文宾州树库（Chinese Penn TreeBank，简称CTB）8.0语料中提取出5,088个中文句子作为语料进行CAMR标注。该语料中的句子来自于微博，内容涉及的领域广泛，话题丰富，且

句子大多比较长，所包含的语义信息丰富。标注前，先人工删减了其中的病句、错句，然后进行了自动分词和人工校对。最终标注完成的语料共包含5,000个中文句子。表1列出了这5,000句语料的基本数据。与中文《小王子》AMR语料相比，这份语料句子更长且结构更为复杂。

表1. 5,000句CAMR标注语料的基本信息

句数	5,000	平均每句字数	34.34
总字数	171,703	平均每句词数	22.46
总词数	112,348	平均每句概念数	18.36
总概念数	91,808	平均每句增加概念数	3.02

3. CAMR使用的谓词框架词典

3.1 现有的3类语义角色标注资源

在学术界，确定语义角色有多种方法，谓词语义关系的数量和颗粒度问题依然存在争议。Xue（2006）指出现有的语义角色标注资源可根据抽象程度的高低进行区分，我们粗略地将其分为以下3类：

（1）抽象程度低，使用通用标签来进行语义成分标注，如agent（施事）、theme（主题）、beneficiary（受益者）。这些标签是从特定谓词或谓词类别中抽取出来的，具有普适性的含义，适用于所有谓词。典型资源有VerbNet（Kipper *et al.* 2000）。北京大学近期也构建了“谓词论旨角色层级分类体系”，划分了8个核心论旨角色（施事、受事、与事等）和18个外围论旨角色（时间、原因、处所等），并据此构建了知识库，这样的语义关系分类颗粒度和数量居中。

（2）抽象程度居中，使用在特定情境中有意义的标签来标注语义角色，标签抽取自特定谓词，且适用于同一类别的相关动词和具有论元结构的名词。典型资源有FrameNet（Baker *et al.* 1998），其根据语义为每一类动词设置一个框架，使用不同的框架元素标签，颗粒度很细，但语义关系数量庞大。

（3）抽象程度高，使用只对特定谓词有意义的标签，如argx（x基于0）来标注语义角色。典型资源有PropBank以及继承其体系而构建出的CPB。PropBank和CPB规定了5种核心语义关系和13种非核心语义关系，为每一个特定谓词分别建立框架，颗粒度很粗，语义关系数量少。

一方面，使用一套抽象程度低且适用于所有谓词的语义标签来构建语义角色

标注资源的方式一直深受语义资源构建者的支持。另一方面，越来越多的学者开始尝试抽象程度高的语义角色标注方式。Bai & Xue（2016）详细介绍了CPB定义语义角色的方法。CPB遵循了PropBank的体系，根据谓词的核心论元（core arguments）和二级论元，或称附加论元（secondary or adjunctive arguments）[1]来定义语义角色，将语义角色区分为核心语义角色和非核心语义角色两种。核心论元有三个重要属性：（1）强制性，若谓词缺少了核心论元，语义就不完整。（2）差异性，不同谓词的核心论元框架各不相同，所以每个谓词的每个义项都有一个自己的核心论元框架。（3）唯一性，多个核心论元不会充当同样的语义角色。而非核心语义角色则是可选的，且不与特定谓词相关。CPB定义了5种核心论元（ARG0-ARG4）和13种附加论元（LOC、TMP等）。

3.2 CAMR使用的谓词框架及词典

CAMR沿用了英文AMR采取的OntoNotes的标注体系，使用CPB中的谓词义项及核心论元框架，利用核心语义关系标签和非核心语义关系标签来标注概念之间的语义关系。核心语义关系是指谓词与其自身的事件框架中包含的若干语义角色之间的关系，包括5种，见表2：

表2. AMR的核心语义关系

核心语义关系	arg0	arg1	arg2	arg3	arg4
中文说明	原型施事	原型受事	间接宾语、工具等	出发点、受益者等	终点

非核心语义关系是指核心语义关系之外的语义角色关系，英文AMR规定了40种一般的非核心语义关系，我们根据中文标注的需要为CAMR新增了4种，共计44种。考虑到与AMR的兼容性，CAMR的非核心语义关系标签仍使用英文单词，见表3：

表3. 中文AMR的非核心语义关系

非核心关系	中文说明	非核心关系	中文说明	非核心关系	中文说明
accompanier	伴随	extent	范围程度	polite	礼貌
*aspect	体	frequency	频率	poss	领属
beneficiary	受益者	instrument	工具	purpose	目的

（待续）

（续表）

非核心关系	中文说明	非核心关系	中文说明	非核心关系	中文说明
cause	起因	li	数字编号	quant	数字
compared-to	参照物	location	处所	range	跨度
consist-of	构成（材料）	manner	方式	source	源
condition	条件	medium	媒介	subevent	子事件
cost	花费	mod	修饰	subset	子集
*cunit	中文特殊量词	mode	语气	superset	父集
degree	程度	name	名称	*tense	时
destination	目的地	ord	序数	time	时间
direction	方向	part-of	部分	topic	话题
domain	陈述	path	路径	unit	度量单位
duration	时长	*perspective	方面	value	值
example	例子	polarity	极性		

注：加*的是中文AMR新增的4种关系[2]，（转引自李斌等（2017b））。

PropBank使用的核心论元结构体系在语言学界一直饱受争议，有学者认为其过于宽泛，不能很好地对语义角色做出分类，也导致AMR采用的谓词框架未能得到学界全面的认可。但我们认为，抽象程度低的谓词框架体系存在两个不可忽视的问题：（1）核心语义角色适用于所有谓词，表示地点、原因、工具等的概念无法进入核心框架，如地点是“遍布”的语义中不可缺少的成分但却不在其核心框架之中。（2）难以标注出语义关系重叠的情况，如“药物缓解了疼痛”中“药物”既可以充当谓词“缓解”的施事，又可以充当它的原因。而抽象程度高的谓词框架体系则可以很好地解决这些问题。

CTB是一个包含分词、词性标记信息，标注了短语结构信息的句法树库。CPB（Xue & Palmer 2009）是一个在CTB标注的句法信息之上补充了谓词语义角色信息的语料库。CAMR采用的谓词语义角色框架词典就是从CPB标注语料中抽取出来的（Bai & Xue 2016），该词典含有每个谓词在不同义项下的核心语义角色框架，共收录了24,510个中文谓词（包括动词、形容词等）共26,650个义项的核心语义角色框架。

下面，我们通过真实数据来探究AMR标注过程中使用的谓词框架及词典对于前途两个问题的处理是否具有明显的优势。

3.3 相关统计分析

3.3.1 表示地点、原因等的概念可充当核心语义角色

首先，抽象程度高的谓词框架体系中每个谓词的每个义项都有一个专属的核心论元框架，如果表示地点、原因等的概念是谓词语义必不可少的成分，则这些概念充当核心语义角色，进入核心论元框架，否则就作为谓词的非核心语义角色。例如表示地点的概念在“遍布-01”的语义中不可缺少，所以处于核心语义角色的地位：

遍布-01
arg0: theme
arg1: **location**

存在类似情况的谓词不在少数。为调查CPB词典中究竟有多少谓词存在这样的情况，我们抽取出了其中对核心语义角色的描述中含有location、cause、condition等关键词的全部义项（关键词根据CAMR的非核心语义关系确定），并人工剔除了无效数据，进行了统计分析。

统计得出，可由 CAMR标注体系规定的非核心语义关系关键词充当核心语义角色的谓词义项共计2,453个，占全部义项的9.20%。其中有147个义项的核心语义角色描述中包含2个以上关键词，占到5.99%，如“引进-01”有5个核心语义角色，其中4个（arg0和arg2-arg4）包含这样的关键词：

引进-01
arg0: agent / **cause**
arg1: entity imported
arg2: **location** arg1 is imported from
arg3: predicate, **purpose**
arg4: **destination**

实际上，由于CPB中对核心语义角色的描述只是用来解释该语义角色与谓词的关系，未必使用统一的单词，如同样表示起源，有时会用source，如“得知-01”；有时又会用origin，如“追溯-04”：

得知-01 arg0: party getting message arg1: message arg2: **source** of message	追溯-04 arg0: agent arg1: **origin** arg2: problem newly found

而在CAMR的标注体系中，表示起源的概念充当非核心语义角色时只用关系source表示。类似的还有location & place、purpose & goal、cause & reason等等。所以我们的程序还未能将表示地点、原因、工具等的概念进入核心框架的情况全部抽取出来。尽管如此，已有近十分之一的义项存在这种现象，可见这种现象是不应忽视的，利用PropBank使用的核心论元结构体系能够更恰当地标注出谓词的语义角色关系。

进一步统计从5,000句标注语料中抽取出的数据可以发现，能够充当核心语义角色，进入核心论元框架的非核心语义关系关键词有accompanier、beneficiary、cause、cost、degree等24种，占全部44种关键词的54.55%。也就是说，有超过一半类别的非核心语义角色是可以充当某些谓词的核心语义角色，进入它们的核心论元框架的。表4列出了出现次数为50以上的非核心语义关系关键词（按出现次数由高到低排列）。

表4. 数量超过50的可充当核心语义角色的非核心语义关系关键词

可充当核心语义角色的非核心关系关键词	cause	location	destination	time	source	name	beneficiary	instrument
数量	1,454	934	140	134	124	80	64	63

注：若谓词的核心语义角色可由不同的非核心语义关系关键词充当，如“泄露-01”的原型施事既可由表示位置的location充当，也可由表示来源的source充当，则分别计入；若谓词有不止一个核心语义角色可由非核心语义关系关键词充当，也分别计入。

从上表可以看出，可充当谓词核心语义角色的非核心语义关系关键词中cause最多，一般充当原型施事，是使得动作产生的原因，说明表示原因的语义角色非常容易进入谓词的核心论元框架。Location次之，且存在大量用place、locative描述的情况未能被抽取出来。表示时间的time也常进入核心论元框架，需要说明的是，核心语义角色的描述中使用的关键词time对应于time和duration两个非核心语义关系。Destination和source表示起止点，一般用于表示时间或地点的起止点，也很可能进入核心论元框架。

3.3.2　一个核心语义角色可表示多种语义关系

其次，对于语义关系重叠的问题，CPB的谓词框架规定了核心论元的最大数量为5，但并未对充当核心论元的概念设限，只要是谓词语义中不可缺少的成分，无论其与谓词之间是何种语义关系，均可充当谓词的核心语义角色，如“药物缓解了疼痛”中“药物”既是“疼痛”的施事，又是它的原因，所以在CPB中“缓解-01”的arg0既可以由表示施事的概念充当，又可以由表示原因的概念充当，也就是说“药物”作为“缓解-01”的arg0，与“缓解”既有施行动作的关系，也有致使关系：

缓解-01
arg0: **cause**, **agent**
arg1: theme

由于CPB词典中对核心语义角色的描述只起到解释其与谓词关系的作用，我们无法确切地统计出有多少谓词框架中存在可以表示多种语义关系的核心语义角色。但仅仅与“缓解-01”一样，arg0与之既有施行动作的关系，又有致使关系的义项就有1,146个，占全部义项总和的4.30%，可见这样的情况较为常见，而CPB词典的谓词语义角色框架可以较好地表示出这样的情况。

4．CAMR遵循的谓词语义角色标注体系

由上节可知，CAMR为表示谓词的语义角色关系，设置了5个核心语义关系标签和44个非核心语义关系标签。并且CAMR遵循AMR的体系，允许重新分析、增补和删减概念，以求更完整地表示句子语义，如“的”字结构“受伤的”省略了谓词“受伤”的施事，则为其补充一个虚节点“person”。我们试图探究这样的谓词语义角色标注体系对于谓词的语义标注是否更加合理有效。

我们获取了规模为26,595句的CPB标注语料库（CoNLL09评测的训练语料），将其与CAMR标注语料进行对比分析。CPB标注语料库是在CTB的句法结构树的基础上，根据PropBank的标签补充标注了语义关系信息的语料库。其核心语义关系标签的设置与CAMR基本相同，唯一的区别在于CAMR中不存在谓词同时与多个概念有相同核心语义关系的情况，若概念为并列结构，则利用and或or来处理，而CPB则利用ARGx和C-ARGx来处理这种情况。但CPB和CAMR在非核心语义关系方面的规定有较大差异，CAMR规定了44种非核心语义关系（见表3），而CPB中仅仅规定了13种（见表5）。

表5. CPB规定的非核心语义关系标签

标签	描述	标签	描述	标签	描述	标签	描述
ADV	adverbial	DIR	direction	FRQ	frequency	PRP	purpose or reason
BNF	beneficiary	DIS	discourse maker	LOC	locative	TMP	temporal
CND	condition	EXT	extent	MNR	manner	TPC	topic
DGR	degree						

4.1 CAMR语料与CPB标注语料在核心论元部分的标注差异

首先统计分析两份语料在核心论元部分的标注差异。根据实际标注中标出的核心论元数与谓词库中谓词的核心论元数的差值，可以将核心论元标注的情况分为全部被标注出来（差值为0）、未全部被标出来（差值小于0）、有人工增补（差值大于0）这3种情况[3]，我们将这3种情况分别称为核心论元标注饱满、核心论元标注不饱满、词典核心论元缺失。

我们分别抽取出了5,000句CAMR标注语料和全部CPB标注语料中的谓词框架信息（由于CAMR语料中未标注到核心语义角色的谓词在抽取时难以与非谓词区分，所以暂时忽略），并进行统计分析。

据统计，CPB标注语料和CAMR 语料中分别出现不同义项的谓词101,326次和19,823次，表6列出了两份语料中实际标注出的核心论元数与CPB词典中谓词的核心论元数的差值不同时谓词（分义项）的数量分布情况。

表6. 两份语料中实际标出的论元数与CPB词典中谓词的论元数的差值情况

语料	差值	-4	-3	-2	-1	0	1	2	累计
CPB标注语料	谓词数	344	1,260	10,060	36,539	**52,735**	**383**	**5**	101,326
	比例	0.34%	1.24%	9.93%	36.06%	**52.04%**	**0.38%**	**0.00%**	100%
CAMR标注语料	谓词数	23	272	1,527	6,862	**11,037**	**99**	**3**	19,823
	比例	0.12%	1.37%	7.70%	34.62%	**55.68%**	**0.50%**	**0.02%**	100%

注：CPB中谓词同时与多个概念有相同核心语义关系的情况以单个计算。

从统计结果可以看出，CAMR中核心论元标注饱满的谓词所占比例较CPB标注语料高出了3.64个百分点，且词典中核心论元缺失的比例也高于CPB语料，而核心论元标注不饱满的比例几乎都比CPB标注语料低，说明CAMR标注体系可以更完整地标注出谓词的核心语义角色框架。究其原因，应该是得益于CAMR允许重新增补概念的规定。由于可以增补概念，CAMR在标注过程中无须局限于原句中的词语，而会尽可能地满足每个谓词的核心论元框架，使得更多谓词的核心论元标注饱满。

从比例上来看，核心论元标注不饱满的谓词还占据较高的比重，我们认为这主要是源于CAMR是句子级别的语义表示方法，所以一些跨句的信息被遗漏。未来我们也试图将CAMR向篇章层级扩展，以更完整地表示出篇章的语义信息。

4.2 CAMR语料与CPB标注语料在非核心语义角色部分的标注差异

CAMR和CPB采用的谓词语义角色框架体系在非核心语义关系方面有明显区别。从标签数量上看，相比于CPB，CAMR使用的非核心语义关系标签数量更丰富，划分更细致，能够较全面地表示出谓词与其语义角色之间的各种语义关系。从标签使用情况上看，我们统计了CPB标注语料和5,000句CAMR语料中谓词的非核心关系标签的使用次数，计算得到两者次数的平均差分别为7,271.53和440.08。

平均差是变量值和平均数的离差绝对值的算术平均数，反映变量之间的差异程度。可见，CPB的13个非核心关系标签使用频率差异很大，设置得过于宽泛，对谓词的语义角色关系区分度不够。根据统计结果可知，标签ADV的使用次数接近其他12个标签使用次数的总和。分析具体语料，发现其中与谓词有语义关系，但语义关系又无法使用其他语义关系标签明确表示的概念，几乎都使用了ADV来标注，如表示否定意义的“不”，表示行为重复的“再”，表示动作次序的“首次”等。此外，表示时间的标签TMP也无法区分时刻、时长、时段等差异较大的概念，说明CPB设置的非核心语义关系标签颗粒度过粗，数量太少，不利于计算机对语义关系的自动分析。当然，如果非核心语义关系标签的数量过于庞杂也会不利于语义分析，还会对标注人员造成过重的负担，例如框架语义学为每一个框架设置专属的语义角色，标签数量繁多，不适于智能化的语义分析。而CAMR设置的44个非核心语义关系标签颗粒度适中，数量合理，具有合适的区分度，所以更适合用于表示谓词的非核心语义关系。

5. CAMR对汉语省略核心论元的名词性结构的处理

AMR允许增补概念，相较于其他句子语义表示方法，如语义依存图等，可

以有效地补充出缺失或省略的信息。汉语中存在“的”字结构、“所”字结构、“所……的”结构等省略谓词核心语义角色的特殊名词性结构，AMR补充概念的这一功能就可以完整地表示出这些结构中谓词的语义信息。比如上节提到的“的”字结构“受伤的”省略了谓词“受伤”的施事（人），我们在标注时会补充一个虚节点“person”来表示被省略的施事，并用核心语义关系的反关系（argx-of，x基于0）标注出这个被省略的成分与谓词的关系“person: arg0-of受伤”。同时，“的”字结构中省略受事的情况也很常见，如“我说的”其实是指“我说的（话）”，于是我们会为其补充虚节点“thing”来表示被省略的受事，并标注出其与谓词的关系“thing: arg1-of说”。此外，“所”字结构中“所”位于谓词前面，与谓词构成名词性结构，如“所说”其实是指“所说（的话）”，省略了谓词“说”的受事，我们为其补充“thing”，标注为“thing: arg1-of说”，但似乎没有省略施事的情况。“所……的”结构是“的”字结构和“所”字结构的结合，如“所共有的”其实是指“共有的（东西）”，省略了谓词“有”的受事，我们为其补充“thing”，标注为“thing: arg1-of有”。

那么，汉语中这些特殊的名词性结构省略的成分究竟充当谓词的什么语义角色呢？“的”字结构是更常省略施事，还是更常省略受事？“所”字结构和“所……的”结构是否存在省略施事的情况？我们试图从真实标注的语料中寻找答案。我们抽取出了语料中与关系argx-of的对应词为“的”“所”和“所……的”且补充了虚节点的所有数据，从中筛选出所有“的”字结构，“所”字结构和“所……的”结构，进行统计分析，来探究进入这3种结构的谓词省略的核心论元的分布情况。

5,000句CAMR标注语料中共出现“的”字结构309次，“所”字结构9次，“所……的”结构7次，虽然数量不高，但它们确实是汉语中重要的语言现象，不可忽略。首先，表7显示了“的”字结构省略的核心论元及对应的关系的数量及分布情况：

表7. “的”字结构省略核心论元的情况统计

省略成分与谓词的关系	省略成分	省略次数	省略次数占比	省略次数在总体中的占比
arg0-of	person	83	51.88%	51.78%
	thing	66	41.25%	
	animal	6	3.75%	
	country	3	1.88%	
	location	1	0.63%	
	organization	1	0.63%	
	累计	160	100%	

（待续）

（续表）

省略成分与谓词的关系	省略成分	省略次数	省略次数占比	省略次数在总体中的占比
arg1-of	thing	126	85.14%	47.90%
	person	17	11.49%	
	animal	2	1.35%	
	and	1	0.68%	
	country	1	0.68%	
	law	1	0.68%	
	累计	148	100%	
arg2-of	thing	1	100%	0.32%
	累计	1	100%	
总计		309		100%

从统计结果可以看出，“的”字结构省略谓词的施事和省略谓词的受事的数量基本持平，省略施事的情况略多于省略受事，只有一例例外，“正如某个D员哥们教育我的一样……”，省略的是核心论元arg2，表示“哥们教育我的（内容）”。统计结果还显示，“的”字结构中绝大部分省略的成分是person或thing，省略的施事大多是指人的概念，指物的次之；省略的受事绝大部分都是指物的概念，但thing的占比很高还有一个可能的原因是AMR规定thing为109个专有名词标签的顶级标签，专名类别不明时则使用顶级标签。

此外，9处“所”字结构和7处“所……的”结构中省略的均是谓词的受事，且均是指物的概念。由于分布情况较为简单，此处就不再列表。虽然由于语料规模较小，数据未必能覆盖全部语言实际，但也能体现出“所”字结构和“所……的”结构省略的成分一般都是谓词的受事。究其原因，这两种结构本身即可表示动作行为涉及的对象或受事，所以可以将受事省略。

通过对真实标注语料中“的”字结构、“所”字结构和“所……的”结构的增补概念情况的统计分析，可以发现，AMR采用的谓词语义角色框架能够较好地补充出汉语特殊结构中的谓词省略的核心论元信息，使得谓词信息更为完整。

6. 结论与未来工作

本文基于5,000句CAMR标注语料，从CAMR使用的谓词框架词典（义项词典）、谓词语义角色标注体系（与CPB标注语料对比）、对汉语中省略核心论元的名词性结构的处理（如“的”字结构、“所”字结构）等三方面分析了AMR采

取的谓词框架的合理性。我们发现：（1）从CPB语料库中抽取出的谓词框架词典抽象程度高，相对于CAMR表示谓词的核心语义角色关系具有明显优势。（2）CAMR在语义角色的设置上做到了颗粒度粗细相融，核心语义关系颗粒度粗，44种非核心语义关系颗粒度较细，对谓词的非核心语义角色关系具有合适的区分度。（3）CAMR允许增补概念的规定有助于更完整地标注核心语义角色，解决了语义角色省略的情况，如“的”字结构、“所”字结构、“所……的”结构。所以，AMR作为整句的语义表示方法，在语义角色标注方面具备独特的优势，需要进一步加强AMR语料库的建设，为中文句子语义处理奠定基础。

当然，要保证标注质量，充分发挥谓词框架的合理性，必须要保证与标注配套的谓词词典的质量。然而我们使用的谓词词典本身还存在不少问题，主要表现在：（1）部分多义词义项不明。有些多义词的不同义项的论元框架描述完全一样，难以区分，影响标注一致性，如“开”的09-14义项的论元框架均为“arg0: agent/cause; arg1: theme”。（2）核心论元不对应。如同样是表示谓词的施事或原因，在“压迫-01”的论元框架中做arg0，但在“取舍-01”的论元框架中做arg1。（3）论元缺失。如“贴补-01”只有一个表示施事的论元，缺少“贴补”的内容和对象。（4）义项缺失。部分谓词义项不全，如“丰富”既可作形容词又可作动词，词典里只收录了做动词的两个义项。词典中存在的这些问题导致标注质量降低，数据可信度降低，且会影响自动分析效果，所以我们计划对词库进行系统性的修改。

此外，我们还意识到名词其实和动词一样有论元结构，谓词的论元框架同样适用于名词，如“信心”一词的词义包含3个不可缺少的成分，分别为有信心的人、有信心的对象和有信心的方面，所以可以为“信心”构建论元框架，包含3个核心论元。关于名词论元框架的问题，现在还处于初探阶段，未来我们将尝试系统地整理名词的论元框架，并在CAMR的标注中体现名词的论元结构，预计将能够更好地表示句子语义。

注释

1. CPB和PropBank中使用的术语存在细微差异，CPB中的core arguments和secondary or adjunctive arguments对应PropBank中的（core）arguments和semantic adjuncts，本文使用的术语“（核心）论元”以PropBank为准。
2. 严格地说，44个非核心关系中cunit，aspect和tense不算语义关系。
3. 谓词库中谓词的核心论元数与实际标注出的核心论元数的差值大于0的情况中也可能包含核心论元有缺漏的情况，但数量很少，几乎可以忽略不计；差值大于0的情况中也可能包含核心论元未被全标出来的情况，同样由于数量很少，暂时忽略。

参考文献

Bai, X. & N. Xue. 2016. Generalizing the semantic roles in the Chinese proposition bank [J]. *Language Resources and Evaluation* 50(3): 643-666.

Baker, C., C. Fillmore & J. Lowe. 1998. The Berkeley FrameNet Project [A]. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics* [C]. 86-90.

Banarescu, L., C. Bonial, S. Cai, M. Georgescu, K. Griffitt, U. Hermjakob, K. Knight, P. Koehn, M. Palmer & N. Schneider. 2013. Abstract meaning representation for sembanking [A]. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse* [C]. Sofia. 178-186.

Kipper, K., H. Dang & M. Palmer. 2000. Class-based construction of a verb lexicon [A]. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence* [C]. Vancouver: 691-696.

Li, B., Y. Wen, W. Qu, L. Bu & N. Xue. 2016. Annotating the *Little Prince* with Chinese AMRs [A]. In *Proceedings of the 10th Linguistic Annotation Workshop* [C]. 7-15.

Palmer, M., D. Gildea & P. Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles [J]. *Computational Linguistics* 31(1): 71-106.

Xue, N. & M. Palmer. 2009. Adding semantic roles to the Chinese Treebank [J]. *Natural Language Engineering* 15(1): 143-172.

Xue, N. 2006. A Chinese semantic lexicon of senses and roles [J]. *Language Resources & Evaluation* 40(3-4): 395-403.

Yu, L., C. Wu & E. Hovy. 2008. OntoNotes: Corpus Cleanup of Mistaken Agreement Using Word Sense Disambiguation [A]. In *Proceedings of the 22nd International Conference on Computational Linguistics: Volume 1* [C]. Stroudsburg: Association for Computational Linguistics. 1057-1064.

李　斌、闻　媛、卜丽君、薛念文，2017a，英汉《小王子》AMR语义图结构的对比分析[J],《中文信息学报》(1)：50-57。

李　斌、闻　媛、宋　丽、卜丽君、曲维光、薛念文，2017b，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》(6)：93-102。

袁毓林，2007，语义角色的精细等级及其在信息处理中的应用 [J],《中文信息学报》(4)：10-20。

通讯地址：210097　江苏省南京市 宁海路122号中大楼南京师范大学文学院（宋丽、闻媛、葛四嘉、李斌）

210023　江苏省南京市 文苑路1号明理楼南京师范大学计算机科学与技术学院（周俊生、曲维光）

语料库语言学

CORPUS LINGUISTICS

要　目

责任编辑：毕　争
责任校对：解碧琰
封面设计：覃一彪　锋尚设计

高等英语教育出版分社宗旨：
推动科研·服务教学·坚持创新
外研社·高等英语教育出版分社
FLTRP Higher English Education Publishing
电话：010-88819595
传真：010-88819400
E-mail：ced@fltrp.com
网址：http://heep.unipus.cn

unipus

heep 微信公众号　iResearch 微信公众号

记载人类文明
沟通世界文化
www.fltrp.com

ISBN 978-7-5213-0470-1

定价：12.00元