语义角色句法实现的词汇语义制约信息库的建设及其应用*

周明海，亢世勇
鲁东大学 中文信息处理研究所，山东 烟台 264025
E-mail: freer516@163.com; kangsy64@163.com

摘 要：词汇语义制约了语义角色的句法实现。我们以语义角色为纲、以句中动词为中心，抽取了标注信息比较成熟的《中小学语文课本标注语料库》中必有论元块的核心词，在标注义类、句法语义格式等信息的基础上建立了《语义角色句法实现的词汇语义制约信息库》。目前该库共有施事、受事、当事、共事、客事、系事、领事、与事、结果、致事、分事等11类49727条信息。文章最后探讨了该库在本体及应用方面特别是语义角色自动标注方面的作用。

关键词：语义角色；句法成分；词汇语义；信息库；应用

Construction of Information Database for Lexical Semantics Constraining Syntactic Realization of Semantic Roles and Its Application

Zhou Minghai, Kang Shiyong
Institute of Chinese Information Processing, Ludong University, Yantai 264025
E-mail: freer516@163.com; kangsy64@163.com

Abstracts: Lexical semantics constrain the mapping from semantic roles to syntactic elements. Taking semantic roles as an outline and centering on verbs of sentences, we extracted core words of indispensable argument chunks from the *Tagged Corpus of Chinese Textbooks for Primary and Middle Schools* which is well processed. Based on tagging sense categories, semantic syntactic structures and other information, *Construction of Information Database for Lexical Semantics Constraining Syntactic Realization of Semantic Roles and its Application* was built. At present, the database has 11 categories such as agent, patient, theme, partner, objective, relative, possessor, dative, result, effect, ofpart etc., including 49727 pieces of information. Finally, functions of the database for the theory study and its application, especially tagging semantic roles automatically, are discussed.

Keywords: semantic role; syntactic element; lexical semantics; information database; application

1 前言

语义角色的句法实现就是要研究句法与语义接口问题或者说深层的语义格如何映射为表层的句法成分的问题，也称联接理论和映射理论。该理论旨在研究句法和语义的对应规律，以及题元角色同句法论元联接的制约条件，并对它们做出合适的形式化描写。目前中文信息处理领域语义角色自动标注特征研究大多注重了形式方面，意义很少有涉及，从邵艳秋、穗志方、吴云芳（2009）实验的结果来看，语义特征对语义角色的标注有着一定的作用，是句法语义接口研究的一个方向。

为了进一步研究词汇语义在语义角色标注中的作用，我们建立了《语义角色句法实现的词汇语义制约信息库》（下文简称《信息库》），该库旨在通过对真实文本句子中的核心词汇语义信息的准确标注，在词汇语义层面上建立起句法关系与谓词-论元结构之间的内在联系，为进行大规模真实文本句子的语义角色自动标注提供训练和测试语料库。

* 本文承国家社科规划项目“基于大规模标注语料库的现代汉语句子语义结构系统研究（05BYY029）”和863项目“智能感知与先进计算技术专题”（项目编号：2007AA01Z173）子课题“构建汉语句法语义标注库”的资助。

2 语义角色句法实现中的词汇语义研究

现代意义上的句法语义关系的探讨应在结构主义的鼎盛时期，乔姆斯基在《句法结构》提到"语言的形式和语义之间，不可否认地存在着某些对应关系，虽然这些对应关系不是十分贴切的……"出于建立语法理论的特殊需要，当时他和结构主义学者一样也不去考虑意义，但句法语义关系的探讨正孕育其中。

词汇语义在语义角色句法实现中的作用从 Gruber、Fillmore 到 Chomsky 都有所涉及，但不及下面学者的观点明确。

俄罗斯学者阿普列祥（1974）在界定"语义配价"时指出，语义配价是从词汇意义中直接引出的，它不同于语法意义，语法意义是许多词共有的一类语法范畴的体现。

1989 年 Bresnan 和 Kenerva 提出了词汇功能语法，其中词汇映射理论集中探讨了词汇语义如何实现为句法的问题，他们认为论元结构是连接词汇语义和句法结构的纽带，词汇的语义特征只有赋予到论旨角色上才能映射为句法成分，即论旨角色是词汇语义和句法成分之间的桥梁。

Levin 和 Hovav（1996）在《The Handbook of Contemporary Semantic Theory》提出了谓词的词汇语义表达与其论元的句法表达形式之间的映射（mapping）关系是完全可以预测的假设，他们认为既然一个论元的语义角色是由选择它的谓词的意义决定的，那么谓词的意义在句子的句法结构上就成为决定因素。

莫斯科语义学派则认为动词所要求的句法成分深层次上来源于动词词汇语义，动词语义预示了其句法搭配特点，动词表层句法特征可由其深层语义推导，在特定的意义上说明了题元的语义实质，即"语义决定论"。

在国内句法和语义关系的研究虽然较早，但将词汇语义与句法语义联系起来大约是 20 世纪 90 年代的事了。

龚群虎（1996）指出把句法—语义信息放入带主目的词中，不但可以解释句子的意思，而且可能实现句法和语义的接口。

陆俭明（2006）在《句法语义接口问题》中进一步明确地指出了句法语义接口研究的词汇语义方向，他说："句法语义的接口问题可以有不同的研究、探索的思路，但都不可忽视词语特征的研究。词语携带了丰富的句法语义信息，在很大程度上决定了它所在的句子的句法语义结构。反过来，句子之所以表现出不同于其他句子的句法语义结构，也正是因为其中所包含的某些关键词语不同。因此，重视词语的句法、语义的特征的研究与描写，将是解决好句法语义接口问题的重要一步"。

王葆华（2006）认为动词为一个句子提供了复杂的句法和语义信息，决定了一个句子可能的句法结构或句法框架，也决定了那些与之共现的名词性成分的语义选择限制，动词的词汇语义在一定程度上决定了论元的句法表达。

孙道功、李葆嘉（2009）依据词汇语义(词汇义征，义场) —句法范畴义征—句法语义(句法范畴—角色)的思路，基于词汇义征和范畴义征的分析，初步揭示了动核结构中"词汇语义—句法语义"，即词汇单位如何凭借范畴义征转化为句法范畴的衔接机制。

在汲取国内外研究成果基础上，结合我们这几年句法语义标注的经验，我们认为词汇语义决定了语义角色，语义角色可以映射为句法成分，词汇语义一定程度上制约、决定了句法成分。由于句子中动词词义和名词词义互相缠绕、互相依赖，两个词汇语义单位组合的前提是有共同的义素，所以我们的词汇语义包括动词义和名词词义。下面图 1 是词汇语义和句法语义关系图。鉴于目前词汇语义特征提取的困难，我们选用《同义词词林》（下文简称《词林》）的义类体系作为词汇语义标注的依据。

图 1 中句法配位指语义角色可以映射为句法成分，语义角色和句法成分对应关系的抽象表达序列就是精简的句模，在《信息库》中我们用“句法语义格式”来称谓。句法填位指词汇语义可

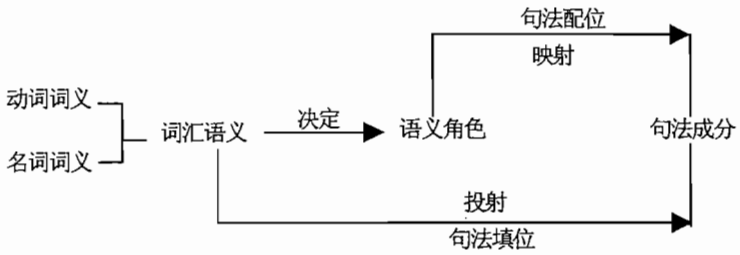

图 1 词汇语义与句法语义关系图

以投射为句法成分。决定语义角色和句法成分的词汇语义包括动词的词义和名词的词义。

3 《语义角色句法实现的词汇语义制约信息库》的建设

3.1 基础语料

我们选用鲁东大学国家社科规划项目形成的《中小学语文课本标注语料库》中比较成熟的人教版初中、高中课本语料作为构建《信息库》的基础语料，约 70 万字。该语料库标注了词性、语块、中心词、句法成分、语义角色等信息，为核心词汇的提取提供了方便。

3.2 《信息库》建设的具体步骤

（1）以“。”、“？”、“！”、“：”和“；”为界符，将近 70 万字的语料分句，生成《语义角色句法实现的词汇语义制约例句库》，共有 15835 个句子。

（2）以语义角色为纲、以动词为中心，提取句子中含该语义角色短语块的中心词（《信息库》中称为名核）、谓语短语块的中心词（《信息库》中称为动核）、中心词词性和句法语义格式等信息，语义角色在在介词短中还需要提取格标。一个语义角色建立一张表。

（3）对提取的中心词进行，《词林》义类自动标注。

（4）人工选取、修正具体语境下词的义类信息，对于《词林》没有的词采取最大相似度算法进行计算、标注。

（5）形成《语义角色句法实现的词汇语义制约信息库》（如表 1）。

表 1《语义角色句法实现的词汇语义制约信息库》样例

ID	句子号	名核	名词	名核词林	名性	动核	动词	动核词林	动性	格标	句法语义格式
1	(*4*)	她/r	她	Aa04	r	老/a	老	Eb36	a		SDPV
2	(*4*)	身体/n	身体	Dd17	n	好/a	好	Ib10	a		SDPV
3	(*9*)	老人/n	老人	Ab02	n	挺/v	挺	Ib03	v		JDPV

表 1 中“句子号”指我们分句后的句子序号，方便我们回到语料，在语境中进行判断。“名核”指带有语义角色标记的名词短语块的中心词，“动核”指句子动词短语块的中心词。在编程提取中心词时我们遵循朱德熙先生关于向心结构、离心结构的理论，对于“的”字结构做主宾语我们也提取出来，以备后续研究。“名核词林”、“动核词林”指名核和动核在《词林》的义类标记，“名词”和“动词”指不带有词性的主要短语中心词，“名性”、“动性”指名核的词性和动核的词性，由于《词林》中词性和义类有着一个大致的对应关系，对词义消歧起着很大作用，这里我

们也将其包含在库中。句法语义格式是指名核和动核的句法语义框架，即名词短语块和动词短语块“[ ]”内外的标记，如 SDPV 中 S 是主语的意思，D 指 S 的语义角色为当事，P 是动词，V 是动词 P 的语义角色，在句子中的格式是[S]D[P]V，PV 是固定的。

3.3 《信息库》现在的规模

目前已经处理了施事、受事、当事、共事、客事、系事、领事、与事、结果、致事、分事共 11 类 49727 条信息，具体情况如表 2。

表 2 《语义角色句法实现的词汇语义制约信息库》规模

语义角色	句法成分（格式）	个数	语义角色	句法成分（格式）	个数
施事	施事主语（SSPV）	16350	系事	系事宾语（PVOX）	3752
	施事宾语（PVOS）	195	领事	领事主语（SLPV、PVSL）	1059
	施事状语（DSPV）	186		领事兼语（JLPV、PVJL）	27
受事	受事宾语（PVOO）	9598		领事宾语（OLPV、PVOL）	21
	受事主语（SOPV）	737	与事	与事状语（DTPV）	830
	受事状语（DOPV）	252		与事宾语（PVOT）	388
当事	当事主语（SDPV、SDCV、PVSD）	10228		与事补语（PVCT）	74
	当事兼语（JDPV、JDCV）	544		与事兼语（PVJT）	27
	当事宾语（CVOD、ODPV、PVOD）	37	结果	结果宾语（PVOR）	303
	当事状语（DDPV）	22		结果主语（SRPV）	6
共事	共事状语（DYPV）	295	致事	致事宾语（PVOZ）	59
客事	客事宾语（PVOK、OKPV）	4068	分事	分事宾语（PVOF）	36
	客事兼语（PVJK、JKPV）	391		分事状语（DFPV）	2
	客事主语（SKPV）	217		分事主语（SFPV）	2
	客事状语（DKPV）	16			
	客事补语（PVCK）	5			

4 《信息库》的作用

《信息库》的作用体现在本体和应用两个方面，下面分别介绍。

4.1 本体方面

《信息库》以语义角色为纲标注了词汇语义、词性、句法格式等信息，利用这些信息可以研究句法、语义、词汇两两之间的关系以及三者的对应关系，深化句法语义接口研究。对每一类语义角色进行细致深入地研究可以发现新问题，比如共事义类的填位顺序如下：

A 人>D 抽象事物>B 物>E 特征>H 活动>I 现象与状态>C 时间与空间>G 心理活动

按常理来看，共事应只有“A 人”来充当，但 A 类却只占 53.90%， D、B、E、H、I、C、G 类还有相当一部分，这是为什么？是不是我们的标注有问题？先看下例：

[S 我/r 的/u 心/n]S [D 跟/p 别人/r 的/u 心/n]Y，[D 都/d[D 是/v[D 紧紧地/z[P 连接/v]V[C 在/p 一起/s]P 的/y]U。

上例中“跟别人的心”显然不是针对的对象，不能标成与事，标成其他类也不合适，那标成共事能否说得通呢？句中的两个“心”都参与到动词主载的事件中了，不违背“共事是事件中共

同参与动作行为的角色”的定义，只是存在典型与否的问题，这样，我们将由 A 人类充任的共事称为典型共事，由其他类充当的称为非典型共事。需要指出的是我们的《中小学语文课本标注语料库》是多位本科生、研究生历时两三年标注成的，在这种情况下还有如此大的一致性，这说明这样标注一定程度上符合人们的认知。

4.2 应用方面

该《信息库》可以用于语义角色自动标注、词义消歧、文本蕴含自动识别和扩展等方面，下面分别介绍：

4.2.1 语义角色自动标注

目前语义角色自动标注大多采用形式特征的方法，利用语义信息的标注还处于起步阶段，我们建设《信息库》进行标注是在这个方面的一个尝试。下面我们在中心词、句法成分、词义标注都确定的情况下，利用《信息库》进行了语义角色自动标注实验：

（1）预备资源。预备资源包括两部分，一部分是语义角色备选库，一部分是语义角色句法实现词汇语义制约规则二维表。

语义角色备选库以句法成分为纲，把句法语义格式首字母相同的放在同一个库中，以初步确定标注时首先进入哪一部分，减少回溯率

第一类为 S1 类：SSPV、SOPV、SDPV、SDCV、SKPV、SLPV、SRPV、SFPV；第二类为 D 类：DSPV、DOPV、DDPV、DKPV、DTPV、DFPV；第三类为 J 类：JDPV、JDCV、JKPV、JLPV；第四类为 O1 类：ODPV、OKPV、OLPV；第五类为 P 类：PVOS、PVOO、PVSD、PVOD、PVOK、PVJK、PVCK、PVOX、PVSL、PVJL、PVOL、PVOT、PVCT、PVJT、PVOR、PVOZ、PVOF；第六类为 C 类：CVOD；

根据谓语后句法成分的不同，我们对第五类再细分：①PVO 类：PVOS、PVOO、PVOD、PVOK PVOX、PVOL、PVOT、PVOR、PVOZ、PVOF；②PVS 类 PVSD、PVSL；③PVJ 类：PVJK、PVJL、PVJT；④PVC 类 PVCK、PVCT。

由于汉语中经常出现“雪下了”和“下雪了”这样主宾互换的句子，针对这种情况，我们增加了两个语义角色备选子库，S2 类：SSPV、SOPV、SDPV、SKPV、SLPV、SRPV、SFPV；O2 类：PVOS、PVOO、PVOD、PVOK、PVOL、PVOR、PVOF，这两个子库是对应的，即同一语义角色可以出现主宾两种位置时这两个子库才会发挥作用，目前这两个库只用在实例搭配阶段。

《信息库》建成后我们将动核义类和名核义类搭配情况生成规则二维表，包括中类和大类两种。每个句法语义格式建两张搭配规则表，共 30 类，60 张，表 3 是系事（PVOX）的中类二维规则表。

（2）具体流程

1）根据句法成分进入语义角色备选库；

2）谓核、名核如果跟库中已有词完全相同的，则标注成相同的语义角色，对于 S 类、O 类找不到的，反向调用语义角色备选子库。

3）动核和名核在语义角色备选库中如果有一个是一样的（这里指不在同一行中，采取动核优先），则查找库中有没有和另一个词相近的词（即查找《词林》义类，先小类、再中类、最后为大类），如果有则标注相同的语义角色，如果出现了多个候选角色，则标注成频率高的。

4）如果动核、名核都是库中的未登陆词则调用二维规则表，先中类后大类，标注成频率高的。

根据上述算法我们对普通话测试 50 篇中的一篇进行了测试，该篇共有 62 个需要标注的核心语义角色，我们标注正确 55 个，正确率为 88.71%。

表3 系事二维规则表

名核义类 / 动核义类	A	B	C	D	E	F	G	H	I	J	K	L	Total
A	0	0	0	0	0	0	0	0	0	0	0	0	0
B	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	0	0	0	0	0	0	0	0	0	0	0	0
D	1	0	2	0	1	0	0	0	0	0	0	0	4
E	0	0	1	1	0	0	0	0	0	0	0	0	2
F	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	1	0	0	0	0	0	0	0	0	0	0	1
H	70	8	1	10	5	0	0	2	1	0	0	0	97
I	55	102	12	94	31	0	7	6	4	0	0	0	311
J	631	821	217	1158	216	7	44	115	53	23	32	0	3317
K	4	3	2	10	0	0	0	0	0	0	1	0	20
L	0	0	0	0	0	0	0	0	0	0	0	0	0
Total	761	935	235	1273	253	7	51	123	58	23	33	0	3752

4.2.2 词义消歧资源库

多义词的词义消歧就是利用词义消歧技术来解决如何在给定上下文语境内外中确定多义词义项的问题。《信息库》标注了词汇对的义类信息，其可以作为一个词义消歧实例库，同时也可以作为词义消歧训练、测试库。

4.2.3 文本蕴含识别、扩展库

中文信息处理领域的文本蕴含是一个比较宽泛的概念，如果一个连贯的文本 T 可以推出假设 H，我们将这种关系称为文本蕴含，目前文本蕴含的识别主要集中在词汇和语法转换层面，《信息库》在句法语义相同的情况下为文本蕴含提供了更有效的词汇信息，便于识别和扩展。

参 考 文 献

[1] Апресян, Ю.Д. Лексическая семантика-синонимические средства языка, Наука, Москва, 1974.

[2] Shalon Lappin, ed. The Handbook of Contemporary Semantic Theory[M]. Oxford: Blackwell, 1996.

[3] 龚群虎. 论句法语义一体化分析中词义的地位[J]. 语文研究, 1996, (4).

[4] 霍花. 俄语题元问题刍议[J]. 中国俄语教学, 2008, (3).

[5] 亢世勇, 许小星, 马永腾. 施事、受事句法实现的义类制约[A]. 词汇语义学的新进展[C]. 新加坡: 新加坡东方语言信息处理学会出版, 2010.

[6] 陆俭明. 句法语义接口问题[J]. 外国语, 2006, (3).

[7] 诺姆. 乔姆斯基著, 邢公畹等译. 句法结构[M]. 北京: 中国社会科学出版社, 1979.

[8] 潘海华. 词汇映射理论在汉语句法研究中的应用[J]. 现代外语, 1997, (4).

[9] 邵艳秋, 穗志方, 吴云芳. 基于词汇语义特征的中文语义角色标注研究[J]. 中文信息学报, 2009, (6).

[10] 孙道功, 李葆嘉. 动核结构的“词汇语义—句法语义”衔接研究[J]. 语言文字应用, 2009, (1).

[11] 王葆华. 动词的词汇语义与论元表达之关系—兼谈动词意义的原型效应和家族相似性[J]. 汉语学报, 2006, (1).

[12] 张家骅. 莫斯科语义学派的配价观[J]. 外语学刊, 2003, (4).