语料库语义成分标注的若干问题

许小星[1] 亢世勇[1] 孙茂松[2] 刘金凤[1]

（1. 鲁东大学，烟台，264025；2. 清华大学，北京，100084）

摘 要：全面、系统地研究语义角色与句法成分的对应机制，总结语义成分映射为句子成分的规律，是从事自然语言理解研究的学者们追求的主要目标。在大规模标注语料库的基础上进行研究，可以如实反映现代汉语语言现象全貌。在标注过程中遇到许多值得研究的问题，如有些成分不属于某种典型的语义成分，也没有在其他相关论著中出现过，究竟如何为它定性、确定归属是一个难点；“有”字句等特殊句式中主客体之间复杂的语义关系和对同一介词引导的不同的语义成分的判定问题等都是在真实语料标注中遇到的一些难点，如何较好地解决和处理这些难点是今后努力的方向。

关键字：语料库；语义成分

Some Issues in Labeling Corpus Semantic Component

Xu Xiao-xing[1] Kang Shi-yong[1] Sun Mao-song[2] Liu Jin-feng[1]

(1. Ludong University, Yantai, 164025; 2. Tsinghua University, Beijing, 100084)

Abstract: Roundly and systemically researching on the corresponding relation between the semantic component and syntactic component, and summarizing the rule of semantic component mapping into syntactic component constitute the main aims of the scholars who research on the natural language comprehension. On the basis of large scale label corpus, the research faithfully reflects the panorama of the modern Chinese language phenomena. Many problems were encountered during the process of labeling. For example, some components didn't belong to any typical semantic component, and had never been discussed in any other papers. So it was difficult to determine their natures and attributes. Additionally, there were some other difficulties when labeling real corpus, such as the complex semantic relation between subject and object in those special syntaxes (for example, "have"-sentence), the judgment of different semantic components which were booted by the same preposition, and so on. Future efforts are needed to be concentrated on dealing with and solving these problems.

Key words: corpus; semantic component

1. 标注规范概要

语义角色与句法成分的对应问题一直是语法学界关注的一个热点问题。全面、系统地搞清楚现代汉语句子语义成分和句法成分的的对应关系，总结语义成分映射为句子成分的规律是从事自然语言理解研究的学者们追求的主要目标。在大规模标注语料库的基础上进行研究，归纳总结语义成分和句法成分的对应机制，研究语义成分映

基金资助：国家社科规划项目基于标注语料库的现代汉语句子语义结构研究 （05BYY029）

作者简介：许小星（1981-），女，山西朔州人，在读研究生 E-mail.xuxx_2004@163.com

射为句法成分时所受到的各种限制，这样既可以避免过分的泛化，又可以避免丢失语言学家内省工作时不易察觉的事实。

通过参考鲁川、林杏光、陈昌来等几位学者语义成分的分类系统，我们确定了 25 个语义成分（见附录），依据的原则是语义成分的种类应既较为完备，又概括性强，避免繁杂，标注起来易于操作。我们选取的语料来源于人民教育出版社中小学语文课本，目前已经标注完成的语料计 10 余万字。标注的对象是句子中的谓语动块和与其相关的名词性成份之间的句法和语义关系。

标注方法：标记时先用"[]"划出句子的语块，然后在"["后标出该语块的句法成分，再在"]"后标记出该语块的语义成分，谓语语块标记为[P word]V。例：

[S 我]S[D 多[P 想]V[O[D 和小鸟一样]J[P 唱歌]V]啊]！

标注原则：

（1）层次的原则：在句法分析的基础之上进行语义标注。

①当主谓结构、动宾结构充当句子的主语、宾语、补语时，先划出语块并只为其标注句法成分，不标记语义成分。然后在其内部进行二次标注，例：

[S[S 解放军]S[P 进驻山城]V][P 给]V1[J 老百姓]O1+S2[P 吃了]V2[O 一颗定心丸]O2。

②我们对语义成分的标注是建立在句法分析的基础之上的，按照句法上的层次标注语义成分，并不是完全依据词与词之间的语义关系。如"妈妈拿刀切肉"这一句中，"刀"是"拿"的受事，从语义上看也是"切肉"的工具，但只将其标记为"[S 妈妈]S1S2[P 拿刀]V1[P 切肉]V2"，不再将"刀"标记为工具成分。例：

[S 家里的事]D [P[S 你]S[D 多[P 做]V[O 些]]V。

[S 那两个小同志]D [P 连[S 直起腰来的力气]K[D 也[P 没有]V]V 了。

例 1 中不看句法层次的话，"家里的事"应该是"做"的受事，但它处于大主语的位置,不与"做"处于一个句法层次上，就不将其标注为"做"的受事。这是句法层次和词语语义关系之间的矛盾。从语义上来说，例 2"两个小同志"和"力气"之间是领事和客事之间的关系。从句法上看，该句是一个主谓谓语句，"直起腰来的力气也没有了"是整个句子的大谓语，谓语部分是对主语状态的一种陈述，这样"那两个小同志"就不是领事，而应标记为当事。

（2）序列原则：一个句子出现同一层次上的两套或多套主谓结构或谓宾结构时，我们按谓语动词在句中出现的顺序为谓语动词排序，并按此次序为与该谓语动词相关的语义成分排序。如：

[S 老师傅]S1S2S3 [D 还要 [P 试试]V1 [O 他]O1 ，[D 把模型]O2 [D 全部 [P 毁掉]V2 ,[P 让]V3 [J 他]O3+S4 [D 重新 [P 造]V4。

"老师傅"是"试试他""把模型毁掉""让他重新造"三个分句共同的主语，在语义上也是"试试(V1)""毁掉(V2)""让(V3)"三个动作的发出者，为其打上"S1S2S3"的语义标签，与 V1V2V3 相对应,每套结构的语义成分按谓语动词在句中的先后顺序进行标记，"老师傅让他重新造"又是一个兼语句，"他"是"让"的受事，标记为 O3,同时也是"造"的施事，依序排列为 S4。

2. "有"字句的处理

谓语动词"有"连接的主客体的语义关系是比较复杂的。"有"字的词义不同，句子的句法结构不同，"有"字前后项词义的不同，都会影响"有"连接的主客体的语义关系的变化。具体来看：

2.1 N1+有+N2

2.1.1 "有"字表存在的意思。表示某时或某处存在、某人、某物或某种现象。N1 标记为处所或时间，N2 标记为客事。 例：

[S 阶上]P[P 有]V[O 半张被坐皱的报纸]K。

[S 每个时代]H[D 都[P 有]V[O 很多这样有骨气的人]K。

2.1.2 N1 和 N2 之间是领属关系，N2 隶属于 N1，N1 和 N2 之间是整体与部分之间的关系。那么 N1 与 N2 分别标记为领事和分事。例：

[S 它]D1L2[P 小巧玲珑]V1，[S 一双透亮灵活的眼睛下面]P2，[P 有]V2[O 一双又尖又长的嘴]F2。

[春天]D1L2 [P 像]V1 [O 健壮的青年]X1，[P 有]V2 [O 铁一般的{胳膊}@和{腰脚}@]F2。

下两例中，“大师兄”和“一员老将”作为是“班级成员”和“周瑜的部下”的一分子，与“班级成员”和“周瑜手下”构成部分与整体的关系，所以讲“班上/周瑜手下”和“大师兄/一员老将”分别标记为领事和分事。

[S 班上]L[P 有]V [J 一位“能文”的大师兄]F。

[S 周瑜手下]L [P 有]V [O 一员老将]F。

2.1.3 N1 与 N2 之间存在领有关系，既 N1 占有、拥有 N2 或是 N1 具有 N2 所指的某种特性时，N1 和 N2 分别是领事和客事。N1 可以是人也可以是物，N2 可以是实体也可以是抽象体。如：

[S 她]L[P 有]V[O 一个长到二十岁上忽然截瘫了的儿子/n]K。

[S 落日]L[P 有]V[O 落日的妙处]K。

[S 我们中国人]L[D 是 [P 有]V [O 骨气]K 的。

在《动词大词典》中，将“{他}有{大眼睛、情绪}”中的“大眼睛”和“情绪”看作分事；将“{他、兔子}有{词典、窝}”中的“词典”和“窝”看作客事。可以推断该词典在判断“有”的客体是分事还是客事的依据在于客体与主体是不是可相分离的。但他对同是领属动词的“带”的处理与该原则不符，将{经理}+带+{怒气、情绪}的示例中将“怒气、情绪”标为客事。我们对语料标注中发现“有”的名词性宾语的类型有具体名词，也有抽象名词，主客体之间是领属关系还是领有关系界限不明。我们对此的处理趋于简化，当客体与主体之间是整体与部分的关系，将这样的客体定义为分事。像“{}有{印象/经历/念头/办法/理由/见解/坏脾气/信心/缺点/责任/了解/运气/机会…}”这类比较抽象的客体都标记为客事。例：

[S 他]L[P 有]V[O 一段虽是悲痛的却又是丰富的经历/n]K。

2.1.4 “有”字后可以跟非名词性短语已经得到语法界大多数人的肯定，黎锦熙、朱德熙、范晓等都认为形容词、动词充当“有”的宾语，已经具有名词的性质，可以看作“名物化”。N1、N2 之间的关系可以理解为 N2 是 N1 所具有的特点。所以我们同样把 N1 标记为领事，把充当“有”的宾语的形容词或动词也标记为客事，并将其词性标记为 vn 或 an。

[S 机智]L [D 还 [P 有]V [O 什么光荣]K 呢

[D 也许] [S 她在厨房里劳作的情景]L[D 更[P 有]V [O 另外的美]K 吧。

[S 儿子的画]L[P 有]V [O 进步]K。

2.2 N1+有+N2+数量短语

在“N1+有+N2+数量短语”这个句式中的“有”字表示主体达到一定的数量或重量。我们将 N1 标记为当事，将 N2 标记为数量成分。

[S 祖母]D [D 今年]H [D 已 [P 有]V [O 八十五岁]N。

[S 这一次参军[的]h]D [D 就 [P 有]V [O 七个]N。

[S 这个碗]D [P 有]V[O 千斤重]N。

2.3 N1+有+N2+（那么）+形容词

在“N1+有+N2+（那么）+形容词”这样的格式里，大多数语法学家都认为这是比较句。“有”有经过比较得出结论或加以估量的意思，下例中“有六七层楼房那么高”和“有多少米那么高”的表义是一致的，都是指 N1 达到了某个高度、重量或程度。所以我们对其的处理加以简化，“有”后的宾语整个在句法上标记为宾语，语义上标记为数量。[S 十二根大理石的淡青色柱子]D，[P 有]V[O 六七层楼房那么高]N 。

2.4 （N1）+有+N2+V2+（N3）

张豫峰、范晓两位学者认为 N1+有+N2+V2+（N3）格式内部句法成分间的结构关系有两类，一类是述补结构，一类是状中关系。他们认为“有 N2”和“V2（N3）”之间存在一定的语义联系，这就要和连动式相区别。具体来看，“他有个儿子叫小明”中“叫小明”是补充说明前面的“有个儿子”，所以“有个儿子”是“叫小明”的补语；“我有事找他”中“有事”和“找他”之间存在因果关系，“有事”是“找他”的状语。我们不采用这种做法，仍将这两类句式看作兼语式和连动式来处理。在我们看来，标注语义成分之间的关系是建立在句法关系分析的基础之上的。他们所说的第一种情况中，N2 是“有”的宾语又是 V2 的主语，是一个兼语成分，标记为[J。语义上，根据具体的句子作具体的分析。“有个蜘蛛慢慢的爬过来”这一句中，“蜘蛛”是“有”的客事，是“爬过来”的施事。“有个小岛叫光化岛”中的“小岛”是“有”的客事，与“光化岛”构成当事和系事的关系。连动结构中 V1 和 V2 之间的语义关系是多样的，并不因此影响他们句法上的连动关系。但我们标注的是谓语动词和与其相关的名词性成分之间的关系，V1 和 V2

之间的语义关系不在我们考察的范围之内。

[D 忽然[P 有]V1[J 个蜘蛛]K1+S2[D 慢慢地[P 爬过来]V2 。

[S 湖中央]P1[P 有]V1[J 个美丽的小岛/n]K1+D2，[P 叫]V2[O 光化岛]X2。

[S 我们]L1S2[P 有]V1[O 责任]K1[P 解救]V2[O 他们]O2.

3. 语义成分的典型成员和非典型成员

根据原型论的观点，在一个范畴中有原型成员、次原型成员和边缘成员之分。把每一个语义成分看作一个范畴，该语义成分有典型的成员，也有非典型的成员，这些成员，不是我们通常理解上的某种特定语义成分，是边缘的，不典型的。

3.1 处所成分有一些非典型成员，如：

[S 他们的房屋]S，[D 稀稀疏疏的[D 在雨里]P[P 静默着]V。

[S 我]S[D 在好几篇小说中]P[D 都[P 提到过]V[O 一座废弃的古园]O。

[S 她]L[D 自己心里]P[D 也[P 没有]V[O 答案]K。

[S 我]D[D 那时]H[P[S 脾气]D[P 坏到]V[C 极点]P]V。

[S 这同样微妙的神情]D1S2O3[P 好似]V1[O 游丝]K1 一般，[D 飘飘漾漾地[P 合了拢来]V2，[P 绾]V3[C 在一起]P3。

[S 一切真知]D[D 都[D 是[D 从直接经验]P[P 发源]V 的。

我们将处所成分定义为事件发生的场所、境况，动作行为的起点、路径和终点。“在雨里/在空气中/在阳光下”等格式作为动作行为发生的自然环境也是处所的一类。“在小说/诗/作品中写道/提到”中的作品类也可以看作动作行为的环境。“脾气坏到极点”的“极点”，“神情绾在一起”的“一起”，“从直接经验发源”的“发源”可以看作动作行为的起点和终点。

3.2 时间成分的非典型成员有“——下来”格式、“——过后”格式和“有一天”等。

[D 有时[S 一天下来]H，竟[P 有]V[O 十多元收入]K。

“一天下来”是“有收入”这个事件所经历的时间段。

[D 三五排枪过后]H，[S 他们]S[P 投]V[C 出了[O 手榴弹]O。

“——过后”表示前一个事件的结束，“三无排枪过后”是“投手榴弹”这个事件发生的起始时间。

下例也是时间成分的非典型成员。

[D 在我的口哨声中]H，[D 窗外]P[S 小树的叶子]D1D2[P 绿]V1 了，[D 又[P 黄]V2 了。

3.3 方式成分的非典型成员：

林杏光把“当着——的面”作为处所，陈昌来把“当着——的面”作为对象，我们认为将其看作方式更加适合。

在操场上，张剑峰让他们当着全班同学的面把他们丢下的垃圾打扫干净。

这句话中“在操场上”是典型的处所，将“当着全班同学的面”也标记为处所就显得有些牵强了。对象成分是动作行为针对、关涉的对象，与他提出的与事不同之处在于，前者是可有语义成分，后者是必有语义成分。我们的语义成分体系中与事是动作行为针对、替代、交接的对象，包括了陈昌来体系中对象成分的一部分。显然，“当着——的面”不宜再做与事成分。参照下例：

[S 它们]S[D 在私底下]Q[P 嘲笑着]V[O 百合]O。

“在私底下说”和“当着某人的面说”是有共通之处的，它强调的不是动作行为发生的处所和针对的对象，可以理解为一种说话的方式。所以将“当着——的面”标记为方式成分。例：

[S 在中央政治局会议上]E，[S 毛泽东]S[D 当着全体政治局委员的面]Q[D 高声]Q[P 说道]V：

3.4 范围的非典型成员

①“历史上”，“在——阶段中”，“在——过程中”等相关格式

我们把以上格式标记为范围成分。与时间相区别，时间是事件发生、持续、结束的时间，而这些格式表达的是事件所处的背景，可以纳入到范围成分。

[S 历史上]E1[P 没有]V1[J 一个反人民的势力]K1+O2[D 不[D 被人民]S2[P 毁灭]V2 的！/w

[D 原来[S 人]D1D2D3[D 在实践过程中]E，[D 开始[D 只是[P 看到]V1[O 过程中各个事物的现象方面]K1。

②“对/对于——来说”等相关格式

至于[S 暑假]D，[D 对于一个喜欢他的老师的孩子来说]E，[D 又[D 是[D 多么[P 漫长]V！

[D 在一个孩子的眼睛里]E，[S 他的老师]D[D 是[D 多么[P 慈爱]V。

以上我们将这两例都确定为范围成分。先看“对——来说”这个格式，与“在某人看来”同义，强调在特定的视角下，该事件或事态呈现出的特征或状态，并不是事件直接针对的对象。如暑假是否漫长在不同的人来可能会有不同的结论，将“对——来说”这类格式纳入范围应该更合适。“在孩子的眼睛里”同义于“在孩子看来”“对一个孩子来说”，我们也将其纳入范围之列。

类似的格式还有“在——的印象/记忆里”：[D 在我的印象中]E，[S 大写意的国画]D1O2 虽[P 是]V1[O 淡淡几笔]N1，[D 却[D 是[D 极[D 难[P 掌握]V2 的。

4. 对介标的认识

“介词是汉语句子语义成分的重要标志”（鲁川，1987），同一种语义成分可以由不用的介词介引，同一个介词也可以介引不同的语义成分。由同一个介词引导的介词结构在句中做什么语义成分，一方面受介词结构中的宾语的语义特点的影响，一方面受谓语动词的影响，同时要注意介词本身是否有意义上的差别。

4.1 “与”、“和”、“同”：

下例中“商量”属于协同类动词，这类动词要求有施事和共事共同参与完成动作。“同妈妈”是共同参与动作行为的共事成分。

[S 爸爸]S1S2S3[D 从集上]P[P 卖]V1[苇席]O1[回来]V2，[D 同妈妈]Y3[P 商量]V3：

下例是两个比较句，“山西”和“唱歌”作为比较中参照的对象，应该标记为基准。

[S 陕西]D[D 同山西]J，[D 不[D 是[P 差不多]V 吗？

[S 口哨]D[D 与唱歌]J[P 不同]V。

4.2 “向”和“朝”：

“向 / 朝”一般来说是表示动作的方向的介词（《现汉》2006），多数是方向成分的格标。但由于谓语动词的不同，它们介引的语义成分也会不同。

[S 小伙子]S[D 向他的妹妹]A[P 走]V[C 去。

[D 朝里面正在看报的大姑娘]A[P 说]V：

可是 [S 青年人]S，[D 永远 [D 朝着愉快的事情]K [P 想]V。

“走、跑、看”这类动词的方向性较强，“向他妹妹走去”是向着他妹妹所在的位置的那个方向走去，所以“向他的妹妹”标记为方向。“说、问”这类动词并不注重方向性，“朝大姑娘说”实际上是“对大姑娘说”，“向”引入的是说话的对象，应该标记为与事。对于“想”等认知、心理动词来说，“向/朝”介引的则是客事，“朝愉快的事情想”即“想愉快的事情”。

“向”也可以引入处所成分，如下例中“向院子里坐”表达的是“坐在院子里”的意思。

[D 早晨]H1[P 起来]V1，[P 泡]V2[O 一碗浓茶]O2、[D 向院子]P3[D 一[P 坐]V3。

但还有一些句不好判断“向/朝”介引的宾语的语义成分究竟是什么还难以定性，如下例：

[S 她]S[跟着她奶奶]Y[D 一起[P 走]V[O 向新年的幸福中] [C 去。

4.3 “对于/对”：

介词“对于”既是范围成分的格标也是客事成分的格标，我们判断该介词结构是范围成分还是客事成分所采用的方法是这样的，如果该结构可以用“对于——来说”格式替换且意义不变，该成分标记为“范围”，如果不能替换则标记为“与事”。

[D 对于瘫痪病人]E，[S 这]D[D 差不多[P 是]V[O 要命的事]X 。

[S 这]D[D 对于一般见异思迁的人]E，[D 对于一般鄙薄技术工作以为不足道、以为无出路的人]E，[D 也[P 是]V[O 一个极好的教训]X。

[S 他]L[D 对于这项工作的内容和环境]T[P 没有]V[O 规律性的了解]K。

[D 对于他的死]T，[S 我]D[D 是[D 很[P 悲痛]V 的。

4.4 “将”：

“将”作为介词有两个义项，一个是“用、拿”的意思，一个是“把”的意思（《规范词典》）。在例 1 中，“将”取的是第一个意思，引导的是工具成分。在例 2 中取得是第二个意思，“将”引导的是受事成分。

[S 孔乙己]D1S2[P 显出]V1[O 极高兴的样子]K1，[D 将两个指头的长指甲]I2[P 敲着]V2[O 柜台]O2，

[S 我]S[D 将他给我做的紫毛大衣]O[P 铺]V[C 好[O 坐位]O。

4.5 “因为”、“为了/为”：

一般来说，“因为”介引的是原因性成分，“为了/为”介引的是目的性成分。实际中“为了”有时也可以引入原因成分。如：

[S 母亲]D1[D 那时]H1[D 已[D 不[P 年轻]V1，[D 为了我的腿]C，[S 她头上]P2[P 开始]V2[O[P 有了]V[O 白发]K]。

上例中“为了我的腿”是原因还是目的不很明显，参照谓语动词“开始和宾语中的谓语动词“有”来看，正确的理解为因为我的腿使得母亲长出白发，所以将“为了我的腿”标注为原因成分

5. 对其它一些成分的特殊处理

5.1 对“一下”“一些/点/点儿”“些/点儿”的处理：

数量词“一下”用在动词后作虚化补语，表尝试，有试着做或略微作的意思（《规范》），并不是表示动作只做一次的意思，所以我们只将其标记为补语，不进行语义标记。

[S 她]S1S2[D 随便[P 指]V1[C 一下]N1，[D 就[P 喊]V2[O 她的哥哥]O。

“些、点（儿）、一些、一点（儿）、”跟在某些动词和形容词后面，表示稍微的意思，发生的程度上的变化，并不是量的变化，我们不将其标注为数量成分，只做句法标记。

[S 我]S1[P 管保]V1[O[D 比他们]J[S 水式]D[P 好]V]，[D 再[P 深]V2[C 点[P[S 我]S[D 也[D 不[P 怕]V]V2！

5.2 对数量短语语义标注的处理：

当数量短语后可以补出中心语，将数量短语标记为数量，当补不出中心语，我们将之标记为结果。例：

[S 太阳]D1S2[D 也[D 不[P 疲惫]V1，[D 把大树的影子]O2[P [缩小]V[C 成]]V2[C 一团]R2。

[D 不久[D 在她的身子下面]P，[D 就[P 编成了]V[C 一大片]N。

6. 结语

基于语料的语义成分标注是热点也是难点。在对语料的实际标注中，发现语料反映出的语言现象是极其丰富和复杂的。在处理标注中遇到的一些难点时，我们通过翻阅词典、相关语法著作、参考其他语义分类体系，暂拟出一些解决的办法。对于这些不成熟的看法，希望得到同行的批评和指正，使我们的标注规范更加完善，使得基于语料库的句子语义成分研究的最终结论更加真实、更有价值。

参考文献：

[1] 鲁 川. 动词大词典[M]. 北京：中国物资出版社，1994.

[2] 陈昌来. 现代汉语语义平面问题研究[M]. 上海：学林出版社，2001：97~125.

[3] 张豫峰. 表比较的“有”字句[J]. 语文研究，1998，(4)：12～17.

[4] 张豫峰，范晓. “有”字句的后续成分[J]. 语言教学与研究，1996，(4)：22～36.

[5] 鲁 川. 介词是汉语句子语义成分的重要标志[J]. 语言教学与研究，1987，(2)：20～26.