现代汉语语料的句子级语义标注*

苗传江　刘智颖
北京语言大学语言信息处理研究所，北京 100083
miaochj@blcu.edu.cn　liuzhiying@blcu.edu.cn

摘要：本文讨论了一种标注现代汉语语料的方案，它有两个特点：一是采取自上而下的标注方式，即先标注大的语言单位，再标注小的语言单位；二是对句子进行语义标注，标注了句子及句内子句的语义类型和它们的下一级语义构成成分。按此方案建立的语料库是现代汉语句子语义研究和处理的重要资源。

关键词：句子语义　语料库　句类　语义块　HNC 理论　现代汉语

A Corpora of Modern Chinese Tagged Semantically on Sentence Level

MIAO Chuanjiang　LIU Zhiying
Language Information Processing Center, Beijing Language & Culture University, Beijing 100083
miaochjiang@blcu.edu.cn　liuzhiying@blcu.edu.cn

ABSTRACT: A scheme of building modern Chinese corpora is discussed in this paper, which has two characteristic features. One is that the sentences are tagged in a top-down way, which means that the larger units are tagged first, and then the smaller ones. The other is that the sentences are tagged semantically, which means that the semantic sentence category and the semantic chunks of the sentences and clauses are analyzed and tagged. A corpora constructed in this way is a very useful resource for studying and processing modern Chinese sentence meaning.

Keywords: sentence meaning, corpora, semantic sentence category, semantic chunk, HNC Theory, modern Chinese

语料库是语言本体研究和语言信息处理的重要资源，在规模足够大的前提下，语料库的价值取决于它的加工状况，也就是对其中的语料做了怎样的标注。本文讨论一种对现代汉语语料进行标注的方案，它有两个特点：一是采取自上而下的标注方式；二是在句子级进行语义标注。

1　自上而下的语料标注方式

从已有的语料库加工工作来看，对现代汉语语料库进行标注的一般步骤是：先分词，然后标注词性、词义，再做短语捆绑等。这一步骤可以称为自下而上的语料标注方式，所谓自下而上就是先标注小的语言单位，再标注大的语言单位。分词、词性和词义标注是词一级的标注，词是句子中最小的能够独立运用的语言单位。短语捆绑是短语级的标注，短语是词的上一级语言单位。再往上就应该依次是句子、句群、段落和篇章级的标注了。目

* 本文得到国家重点基础研究发展规划（973）项目 G1998030506 和北京语言大学青年研究项目的资助。

前，经过加工的较大规模的现代汉语语料库，主要是做了分词和词性标注，词义标注和短语捆绑尚处于初始阶段。我们认为，语料标注还应该采取自上而下的方式，也就是先标注大的语言单位，再标注小的语言单位。

为什么需要对语料进行自上而下的标注呢？可从两方面来看。一方面，从语言自身的机制和人脑的语言认知模式来看，语言不是语音或文字符号的线性序列，而是有层次结构的，语言的分析理解和生成表达的过程绝不只是简单的组词成句，而是遵循着自上而下的宏观结构的指导。另一方面，从计算机对自然语言的处理来看，只靠词汇层面的微观知识是不够的，还必须有句子及其以上层面的宏观框架知识。人脑和电脑对语言的理解处理都需要自下而上和自上而下两种方式，因此，语言研究也就有自下而上和自上而下两个方面。例如，从字词出发，研究它们组合成短语、句子的规律，这属于自下而上的语言研究；从句子的框架结构出发，研究它对句子内短语和字词的管辖约束作用，或者研究话语结构对句子的影响，这些属于自上而下的语言研究。自下而上和自上而下的语言研究，需要以这两种不同方式来对语料库进行加工。当然，这两种语料标注方式并不是对立的，而是互补的，它们就像从上下两端出发的相向运动，最后应该相遇和衔接起来。但目前尚缺少采取自上而下方式进行标注的语料库，应该填补这一空白。

那么，要对现代汉语语料进行自上而下的标注，它的“上”应该是什么呢？也就是说，应该从哪一级语言单位的标注做起呢？自然应该从句子一级做起，其依据有两个。第一，句子是语言实现其功用的基本单位，是语言理解和表达的基本单位，也是计算机处理自然语言的基本单位。第二，中文信息处理正在从字处理阶段转向句处理阶段（许嘉璐 2000），实现对句子的理解处理是当前阶段的中心目标。所以，从句子一级对现代汉语语料进行自上而下的标注，符合语言本体研究和语言信息处理的迫切需要。

2 句子级语义标注的基本内容

对自然语言的理解处理最终要靠语义，所以句子级的语料标注应以语义为主导。HNC理论（黄曾阳 1998）建立了自然语言句子语义的表述模式，我们以该理论模式为指导来对句子的语义进行自上而下的标注，第一阶段的标注内容主要有三项：（1）句类，即句子的语义类型；（2）语义块，即句子的下一级语义构成成分；（3）句蜕，即包含在语义块内的句子。

2.1 句类标注

HNC 定义的句类是指句子的语义类型。HNC 根据作用效应链的六个环节加上判断，划分出 7 大句类：作用句、过程句、转移句、效应句、关系句、状态句和判断句，又划分了它们的子类，共有 57 种，称为基本句类。这 57 种基本句类是句子语义的基元类型，可以用它们来描述任何句子的语义类型。自然语言中一个句子的语义类型，可能是某一种基本句类，也可能是某两种或多种基本句类的组合。例如：

（1） 美英联军已经占领了巴格达。（基本作用句 XJ）

（2） 国际田联有条件同意贾亚辛格参加亚运会。（反应句 X20J）

（3） 边防部队送给养路工人一车蔬菜。（一般转移句 T0J）

（4） 容祖儿彩排时收到歌迷血书。（一般接收句 T1J）

（5） 墨西哥总统福克斯与他的女发言人喜结连理。（双向关系句 R1J）

（6） 中小型商店难以对抗大型连锁超市。（单向关系句 R310J）

（7） 他们对青岛啤酒的质量给予了很高的评价。（一般判断句 D01J）

（8） 台舆论谴责岛内外台独势力同流合污。（反应+信息转移句 X21T3*22J）

（9） 俄罗斯反对美英对伊拉克动武。（单向关系+主动反应句 X21R311*21J）

后面的括号里给出的就是各例句所属的句类，其中的符号是句类的编码。前 7 个句子的句类都是 57 种基本句类中的某一种，后 2 个句子的句类则是两种基本句类的组合。基本作用句和反应句是作用句的子类，一般转移句和一般接收句是转移句的子类，双向关系句和单向关系句是关系句的子类，一般判断句是判断句的子类。

我们就是根据 HNC 确立的 57 种基本句类及其组合，来对句子的语义类型进行标注的。

2.2 语义块标注

HNC 定义的语义块是指句子的下一级语义构成成分。不同的句类需要配置不同的语义块，例如，反应句需要配置三个语义块，分别是反应、反应者、反应引发者及其表现；而一般转移句需要配置四个语义块，分别是转移、转移发出者、转移接收者和转移内容，如下面的例句所示（“‖”是语义块之间的分隔符，下同）：

国际田联‖有条件同意‖贾亚辛格参加亚运会。（反应句 X20J）

反应者‖ 反应 ‖ 反应引发者及其表现

边防部队 ‖ 送给 ‖ 养路工人 ‖ 一车蔬菜。（一般转移句 T0J）

转移发出者‖ 转移 ‖ 接收者 ‖ 转移内容

HNC 构造了语义块的表示式，语义块的表示式加在一起，就构成句类的表示式。57 种基本句类都有明确的表示式，上面两种基本句类的表示式如下：

X20J=X2B+X20+XBC（反应者+反应+反应引发者及其表现）

T0J=TA+T0+TB+TC（转移发出者+转移+接收者+转移内容）

表示式中的语义块是由句类决定的，确定了句类，也就确定了句子中应该有几个什么样的语义块，这类语义块称为主语义块。主语义块可以分为两类，一类是包含述语动词的，它决定句类，称为特征语义块，如上例中的反应 X20 和转移 T0，其余的是另一类，称为广义对象语义块。除了主语义块以外，还有一类语义块，它们不是由句类决定的，称为辅语义块，如“国际田联上周三有条件同意贾亚辛格参加亚运会”里的“上周三”就是个辅语义块。HNC 把辅语义块归纳为 7 种：方式、工具、途径、比照、条件、因、果。

语义块标注就是标出句子中的各个主辅语义块。标注语义块是分析句子的语义构成成分，描述句子的语义结构，它不同于语法上的句子成分分析。下面对此做几点说明。

第一，语义块是句类的函数。句子的主语义块的数量和含义是由句类决定的，这是语义块是句类函数的基本内涵，已如上述。另外，语义块的主辅之分，也要视句类而定，这也体现了语义块与句类之间的函数关系。例如

（1）美军将从海上进攻伊拉克。 （2）红军从井冈山出发。

这两个句子中都有表示空间的语义成分，但第一个句子的“从海上”是辅语义块，而第二个句子的“从井冈山”则是主语义块，这是因为：第一个句子是基本作用句，对这种句类来说，表示空间的成分不是必需的，应该是辅语义块；而第二个句子是自身转移句，对这种句类来说，表示转移起点、终点或途经的空间成分就是必需的，应该是主语义块。

第二，句法位置不影响语义块。语义块是句子的语义构成单位，它不会因句法位置的不同而不同。例如，

（1）他‖向领导‖报告了‖这里的情况。　　（2）他‖把这里的情况‖报告了‖领导。

（3）他‖把这里的情况‖向领导‖报告了。　　（4）这里的情况‖由他‖向领导‖报告。

这是转移句的子类信息转移句，其中“报告”是表示信息转移的特征语义块，“他”是转移发出者，“领导”是转移接收者，“这里的情况”是转移内容，不管它们的句法位置在不同的句子里怎样变化，其语义块角色都保持不变。再如，

（1）台上坐着主席团。　　（2）主席团坐在台上。　　（3）主席团台上坐。

这是状态句的一种子类，这种句类有三个主语义块，一是表示状态的特征语义块，二是状态体现者，三是状态存在的空间。在这三个句子中，“坐”都是状态，“主席团”都是状态体现者，“台上”都是状态空间，而与它们的句法位置无关。

第三，语义块可以分离。在句子中，同一个语义块的两部分可能分离到两个句法位置，这时候它们仍然是一个语义块，而不是变成两个。例如，

李四被张三打断了腿。

这是个基本作用句，该句类有三个主语义块：基本作用（打断）、作用者（张三）和被作用者（李四的腿），“李四”和“腿”在句子中分置两处，但仍然属于同一个语义块。在语料标注中，要把发生了分离的语义块明确标示出来。

第四，特征语义块的复合构成。特征语义块包含述语动词，但特征语义块不同于述语，其表现之一就是特征语义块具有复合构成形式，例如，

中国政府‖对国有企业的经营机制‖进行了深刻的改造。（基本作用句 XJ）

李鹏总理‖对法国的支持‖表示衷心的感谢。（反应句 X20J）

中国与莫斯科‖向美国‖施加了一定的压力。（基本作用句 XJ）

白人极右势力‖对新南非‖持仇视态度。（反应句 X20J）

这四个句子的特征语义块都在最后，它们都采用了复合构成的形式。第一个句子中，特征语义块的核心不是“进行”，而是“改造”，它决定句子的句类是基本作用句，“中国政府”和“国有企业的经营机制”分别是作用者和被作用者。对后面三个句子也要按同样的方式进行语义分析，它们的特征语义块的核心分别是“感谢、压力、仇视”。在语法学上，一般只把“进行、表示、施加、持”看作述语，把它们后面的成分看作宾语，由此可以看出特征语义块和述语的不同之处。

第五，句类转换中的语义块。先看下面的例句，

李小姐的办事能力‖得到‖张先生的赏识。

这是个承受句（作用句的子类之一），可以把它看作是由反应句“张先生赏识李小姐的办事能力”转换来的，这种情况就称为句类转换。在语义分析和理解过程中，必须揭示出转换句类的两个语义块之间隐含的语义关系，拿这个例子来说，就是要揭示出“得到”的前后两个语义块之间存在着“张先生‖赏识‖李小姐的办事能力”这样的语义关系。在语料标注

中，也要对此做出明确的标示。

由以上几点可以说明，语义块是句子的语义构成成分，而不是句法成分，标注语义块就是对句子的语义构成进行分析。

2.3 句蜕标注

句蜕的意思是指句子蜕化为语义块或语义块的一部分，也就是语义块中包含的句子。句蜕的基本形式有两种，如下面的例句所示：

（1）俄罗斯‖反对‖{美国|攻打|伊拉克}。

（2）<生产|信息技术产品|的工厂>‖都转移到了‖国外。

（3）<经济危机|造成|的后遗症>‖也减轻了。

（4）这些话‖似乎表示了‖<他|对奴隶的生活境况|的同情>。

例句（1）的第三个语义块“美国攻打伊拉克”是个句蜕，蜕化前后句子的基本形式没有变化，这种句蜕称为原型句蜕。例句（2）（3）（4）中，尖括号内的语义块是句蜕，它们分别是由下面的句子蜕化来的，蜕化的方式是把句子的某一个语义块作为中心语，其他的语义块作为修饰语，这种句蜕称为要素句蜕。

工厂生产信息技术产品。

经济危机造成后遗症。

他同情奴隶的生活境况。

在语言理解中，不论是原型句蜕，还是要素句蜕，都应该作为句子来处理，要确定它的句类和各个语义块。因此，在我们的语料中，对句蜕都要作为句子来分析，要标明其句类和语义块，如例句（4）中的句蜕是个反应句，其语义块分别是：表示反应的特征语义块“同情”、反应者“他”和反应引发者及其表现“奴隶的生活境况”。（在上面的例句中，花括号内是原型句蜕，尖括号内是要素句蜕，单竖线是句蜕内语义块的分隔符，下同。）

上述三项标注内容都是语义层面的，句类是对句子总体语义的描述，语义块是对句子语义构成成分的描述，句蜕是对语义块内包含的子句的描述，对这三项内容的依次标注，也体现了自上而下的语料标注方式。

3 已进行的工作及其总结

按照上述思路，我们已经制定了具体而详细的语料标注规范，并且已经标注了 20 万字的语料。在标注这些语料的过程中，我们及时讨论和解决遇到的问题，并据此对标注规范进行必要的修改或补充，使之不断完善。现在，我们的标注规范已经确定下来了，相信不会再有大的改动。我们已进行的工作可以证明以下两点：

第一，在 HNC 理论的指导下，采取自上而下的方式，对现代汉语语料进行句子级的语义标注，是可行的。对句子进行语义标注的基础是掌握 HNC 的句类和语义块思想，并学会和熟悉用它们来对句子进行语义分析。我们曾先后对十几个人进行过培训，结果是理想的：有一定语言学基础的人（可以以语言类专业的本科毕业生为参照），经过两个星期的讲习之后，一般都能掌握这种句子语义分析方法，胜任这种语料标注工作。

第二，这种句子级语义标注的结果可以获得比较好的一致性。对在标注语料的过程中遇到的问题，我们经过讨论后，绝大部分都能取得一致的意见。对不同人标注的语料进行比较，结果表明一致性是令人满意的。

在我们已标注的 20 万字语料中，有 10 万字是真实的连续文本语料，另外 10 万字是由单独的句子组成的语料，共有 5500 多个句子，其中 80%以上的句子来自真实语料。下面选录一段连续文本的标注语料，以直观地展现这种语料标注的结果。

!0T21R411*22J+Cn ~这天早上，‖小学生们‖都带着‖自己的暑假作业，

!31T2bJ+Cn ~{!31T2bY9*11J+Re ~按地区|集合}后，‖整队走向‖学校。

!2P01J+Cn ~8:30‖举行‖开学式，

!0X20J 校长‖希望‖{SP10*21J+Ms 他们|~以新的精神状态|开始|新的学习生活}。

!0T2bJ 接着，他们‖回到‖各自的教室，

!0D01J 班主任‖不仅要确认‖{T2bS*11J 全班同学|是否到齐}，

!31T31J 还要询问‖{PS041*21J 他们的暑假生活|过得|是否充实愉快}，

!31T19J 观察和检查‖{X21J 他们|是否做好了|新学期的学习[准备|]}，

!31113T31Y30*21J 然后向他们‖提出‖新的要求。

（选自《光明日报》2000 年 09 月 04 日《各国开学第一天》）

上面每一行是一个句子，行首标出了该句的句类代码和辅语义块，根据句类代码就知道句类的表示式，也就确定了这个句子有几个什么样的主语义块，再根据句类代码前的代码“!mn”确定各主语义块在句中出现的次序，句子中的辅语义块则以“~”标出，这样就可以确定句子中各个语义块的角色了。例如第一个句子，T21R411*22J 是句类代码，它对应的表示式是 TA+T21R411+T2C，表明有三个主语义块，前面的!0 表明这三个语义块在句子中出现的次序跟表示式中相同，句类代码后的+Cn 表示句子中有一个条件辅块，就是以“~”标出的“这天早上”。对句蜕也以同样的方式标在句蜕部分的前面。

我们的目标是建立一个包含 10 万个句子（约 200 万字）的语料库，要实现这个目标，需要做好以下四方面的工作：

第一，提高标注的效率。语义标注的难度较大，工作量也很大，应该尽量开发辅助工具，减少人工劳动量。HNC 句类分析系统（晋耀红 1998）的结果之一就是确定句类和语义块，我们计划利用该系统来辅助标注。

第二，保证标注结果的一致性。确保一致性的关键举措有二：一是对参加标注者进行有效的培训，保证他们具有较高的标注水平；二是制定有效的检查方法，并开发相应的工具来帮助发现问题。在第一个方面，我们已有丰富的经验；在第二个方面，我们将利用 HNC 词语知识库（苗传江 2001a）的内容来对标注语料进行检查和校对。

第三，选择有代表性的语料，建成平衡性良好的语料库。

第四，开发语料管理工具，实现对标注内容的有效检索。

4 对语料进行句子级语义标注的意义

标注语料、建立加工语料库的意义可以概括为两个方面：一是为语言研究搭建平台、

开辟园地，因为在建设语料库的过程中要对大量的真实的语言材料进行系统的分析和标注，必然要研究解决很多的理论和工程问题，从而让语言研究的理论和方法得到全面的检验；二是为语言本体研究和应用研究积累丰富而宝贵的素材和资源。我们对现代汉语语料进行句子级的语义标注，这项工作也具有上述两方面的意义，标注语料的过程就是对分析句子语义的理论和方法进行研究的过程，标注后的语料库又形成语言研究和中文信息处理的宝贵资源，利用这一资源至少可以开展以下研究：

（1） 对各种句子语义类型（句类）特点的深入研究，以便针对不同的句类制定不同的理解处理策略。

（2） 句蜕研究，句蜕是造成句子复杂的主要原因之一，对句蜕分析能力的提高是语句理解处理的关键环节之一。

（3） 句群和段落的语义结构模式研究，语料库中对句子语义的分析和描述为这一研究建立了必要的基础。

（4） 句子语义的认知研究，我们标注了句子的语义类型和句子语义的构成成分，它们组成句子的语义结构，我们相信这一结构与人脑的语句理解模式有直接关联。

希望本文讨论的语料标注方案得到有关专家和部门的支持，早日建成进行了句子级语义标注的现代汉语语料库，为现代汉语句子的语义研究和理解处理做出贡献。

参 考 文 献

[1] 黄曾阳. 1997. HNC 理论概要. 中文信息学报, 4

[2] 黄曾阳. 1998. HNC(概念层次网络)理论. 清华大学出版社

[3] 黄曾阳. 1999. HNC 理论与自然语言语句的理解. 中国基础科学, 2-4

[4] 晋耀红. 1998. 基于 HNC 理论的句类分析系统的设计与实现. 中国科学院声学研究所硕士学位论文

[5] 林杏光. 2002. 为 NLP 创立模式，用 HNC 研究汉语. 汉语学习, 3

[6] 刘志文, 庄咏璆, 郝惠宁, 萧友芙. 1998. 自然语言语句的 HNC 表示. 语言文字应用, 2

[7] 刘智颖. 2001. 论广义对象语义块的分离. 见: [15]

[8] 苗传江. 1998. HNC 理论的句类. 见: 1998 中文信息处理国际会议论文集. 清华大学出版社

[9] 苗传江. 2001a. HNC 自然语言表述模式与知识库建设. 见: [15]

[10] 苗传江. 2001b. HNC 句类知识研究. 中国科学院声学研究所博士学位论文

[11] 苗传江. 2002. HNC 语料标注规范. 内部资料

[12] 唐兴全. 2002. 现代汉语复杂句蜕块研究. 北京语言大学硕士学位论文

[13] 许嘉璐. 2000. 现状和设想——试论中文信息处理与现代汉语研究. 中国语文, 6

[14] 张普. 1991. 信息处理用现代汉语语义分析的理论和方法. 中文信息学报, 3

[15] 张全, 萧国政(主编). 2001. HNC 与语言学研究. 武汉理工大学出版