Modern Linguistics 现代语言学, 2013, 1, 138-143
http://dx.doi.org/10.12677/ml.2013.13026 Published Online November 2013 (http://www.hanspub.org/journal/ml.html)

Building a Token Corpus of Canonical Semantic Sentence Types for Modern Chinese Based on HNC Theory*

Yan Jiang[1], Chuanjiang Miao[1], Xiaodie Liu[2]

[1]Department of Chinese and Bilingual Studies, Hong Kong Polytechnic University, Hong Kong
[2]Institute of Chinese Information Processing, Beijing Normal University, Beijing
Email: ctyjiang@polyu.edu.hk, ctcjmiao@polyu.edu.hk, lxdtg1@yahoo.com.cn

Received: Jul. 2nd, 2013; revised: Aug. 22nd, 2013; accepted: Sep. 6th, 2013

Abstract: A token corpus of semantic sentence types provides elementary resource for the study of sentences from semantic perspectives. We have built such a corpus using the theory of HNC (Hierarchical Network of Concepts). As HNC contains a complete system of sentence semantic types, it provides a good theoretical framework for meaning-based sentence analysis. The corpus we built contains token sentences taken from real-life whose semantic structures are labeled using XML techniques (Extensible Markup Language). The search function for token sentence is also provided. We will use this token corpus as a basis for further developments so as to accumulate resources for meaning-based sentence analysis.

Keywords: Token Corpus; Semantic Sentence Type; HNC (Hierarchical Network of Concepts) Theory; Modern Chinese

基于 HNC 的现代汉语句子基本语义类型例句库建设*

蒋　严[1]，苗传江[1]，刘小蝶[2]

[1]香港理工大学中文及双语学系，香港
[2]北京师范大学中文信息处理研究所，北京
Email: ctyjiang@polyu.edu.hk, ctcjmiao@polyu.edu.hk, lxdtg1@yahoo.com.cn

收稿日期：2013 年 7 月 2 日；修回日期：2013 年 8 月 22 日；录用日期：2013 年 9 月 6 日

摘　要：句子语义类型例句库是开展基于语义的句子研究所需要的基础资源。我们以 HNC (Hierarchical Network of Concepts，概念层次网络)理论为指导建立句子语义类型例句库，该理论建立了完整的句子语义类型体系，为基于语义的句子研究提供了良好的理论框架。我们已经建立了一个现代汉语句子基本语义类型的例句库，为每个类型配备了典型而真实的例句，并且采用 XML (Extensible Markup Language，可扩展标记语言)技术标注了每个例句的语义结构，还提供了例句查询功能。我们将以这个例句库为基础，逐步扩展，为基于语义的句子研究不断积累资源。

关键词：语料库；句子语义类型；HNC(概念层次网络)理论；现代汉语

1. 建立句子语义类型例句库的意义

简而言之，建立句子语义类型例句库的基本意义是，为基于语义的句子研究积累基础资源。

对句子的研究，有句法、语义、语用三个维度，语义维的研究目前还比较薄弱。句子语义的涵义是什么？对此有很多不同的认识[1]，但不管如何界定，句子语义都是语言理解和生成的基础，对人脑的语言运用是如此，对电脑的自然语言处理也是如此。因此，加强语义维的句子研究，是非常必要的。

*本研究得到香港理工大学中文及双语学系编号 4-ZZ7S 的研究项目资助，项目全名为“现代汉语句子基本语义类型例句库的构建”。

[1]可参阅徐烈炯的专著《语义学》[12]。

基于语义的句子研究，就是把语义维作为句子研究的出发点和着眼点，或者说，是用语义维来统摄句子研究。这就需要从语义维形成句子研究的纲目，需要从语义维划分句子的类型，形成句子语义类型系统。有了这个系统，就可以全面深入地开展语义维的句子研究[2]。

假如已经有了适当的句子语义类型系统，要想基于这个系统很好地开展句子研究，就需要建立一个例句库，为这个系统中的各种句子语义类型提供丰富的例句，以便通过语言实例对各种类型加以描述和说明，帮助研究者具体地理解和把握这个系统。这样的例句库是基于一个句子语义类型系统开展句子研究的基础资源。

如果要从真实文本中获得句子语义类型的例句，可以先逐一标注文本中每个句子的语义类型，然后从中抽取各种类型的例句。对文本做这种标注，实际上是在建立句子语义标注的语料库，这是基于语义的句子研究所迫切需要的基础资源。对文本做句法标注的语料库建设，以树库为代表，已有丰硕成果。自从宾州树库(http://www.cis.upenn.edu/~treebank/)建立以来，国内外有很多机构都建立了树库，为句法研究和基于句法的自然语言处理积累了丰富而宝贵的资源。而做句子语义标注的语料库就还很少见[3]，基于句子语义的语言研究和自然语言处理所需的资源还十分稀缺，亟待加强建设。建立句子语义类型例句库将会为改善这种状况做出贡献。

2. HNC 理论与句子语义类型研究

句子语义类型研究是个相当艰难和复杂的课题，尽管如此，也已有不少学者作了很多探索，取得了很多成果，其中比较有代表性的应该说是基于格语法的句模研究及其后继发展[4]。本文不对句子语义类型研究的各类成果进行具体的比较分析，只是对我们基于 HNC 理论所从事的相关研究做简要的介绍和展示。

HNC(Hierarchical Network of Concepts，概念层次网络)理论[5]基于概念联想脉络建立了自然语言的表述和处理模式，其表述模式有词汇、语句和句群三个层面。HNC 理论在语句层面建立的表述模式的中心内容是：一方面，构建了句子语义的框架表示式；另一方面，发现了句子语义的 57 种基本类型，称为基本句类。这 57 种基本句类是句子语义的基元类型，每种基本句类都有确定的表示式。基本句类及其表示式可用于描述各种句子的语义类型及其框架语义结构。例如，下面是两种基本句类，一是一般反应句，二是信息转移句。

一般反应句

表示式：X2B + X2 + XBC (反应者 + 反应 + 反应引发者及其表现)

例句：

张先生喜欢李小姐温柔体贴。

张先生对李小姐的这种做法很反感。

他们为儿子的出色表现感到骄傲。

信息转移句

表示式：TA + T3 + TB + T3C (转移发出者 + 信息转移 + 转移接收者 + 转移内容)

例句：

张三告诉李四王五结婚了。

张三把王五要结婚的消息告诉了李四。

张三向李四透露了王五要结婚的消息。

一般反应句是表达心理反应的句子，它有三个语义构成成分：一是反应者(X2B)，如例句中的“张先生”和“他们”；二是反应行为(X2)，如例句中的“喜欢”、“很反感”和“感到骄傲”；三是反应引发者及其表现，也就是让反应者做出反应的原因，如例句中的“李小姐温柔体贴”、“李小姐的这种做法”和“儿子的出色表现”。

信息转移句，顾名思义，就是对信息进行转移的句子，信息的转移包括输入、输出和传递等。这种句类有四个语义构成成分：一是转移发出者(TA)，即对信息进行转移的行为发出者，如例句中的“张三”；二是信息转移行为，如例句中的“告诉”和“透露”；三是信息转移的接收者，如例句中的“李四”；四是转移内容，也就是被转移的信息，如例句中的“王五结婚了”和“王五要结婚的消息”。

以上两种句类是句子语义的两种基元类型，自然语言中的有些句子是这两种句类的组合，换言之，这些句子的语义类型和语义结构就可以用一般反应句

[2] 关于汉语句子语义类型的较早探讨，可参阅蒋严和潘海华的论文《汉语语句的类型表达》[5]。

[3] 跟句子语义标注有关的研发工作，有代表性的当属 FrameNet (https://framenet.icsi.berkeley.edu/fndrupal/)。

[4] 可参阅鲁川的专著《汉语语法的意合网络》[6]和朱晓亚的专著《现代汉语句模研究》[13]。

[5] 关于 HNC 理论的系统而详细的内容，请参阅该理论创立者黄曾阳先生的著作[1-4]。

和信息转移句的组合来描述，例如像下面这样的句子：

张三祝贺李四乔迁新居。

老师批评小明经常迟到。

这两个句子的语义类型就是一般反应句和信息转移句的组合，也就是表达由心理反应引发信息转移的句子，其语义结构也就由来自这两种句类的三个语义成分构成：一是反应者兼转移发出者，如例句中的“张三”和“老师”；二是心理反应和信息转移行为，如例句中的“祝贺”和“批评”；三是反应引发者及其表现，如例句中的“李四乔迁新居”和“小明经常迟到”，其中的反应引发者同时也是信息转移的接收者，如例句中的“李四”和“小明”。

由上可见，HNC 理论的句类就是句子的语义类型，句类表示式就是句子语义框架的构成模式。句类表示式的构成成分称为语义块，语义块是句类的函数，也就是说，语义块的涵义依句类的不同而不同。

运用 HNC 理论的语句表述模式和句类体系，可以对句子的语义做出适当的分析和描述。例如，下面的句子都是主谓宾结构，但它们的句类是不同的，如句后的括号内所示。句类不同，其语义块的涵义就不同，这些句子的主宾语的语义角色也就各不相同。比如，“张先生”和“李小姐”在句(1)中分别是作用的发出者和承受者，在句(2)中则分别是反应者和反应引发者，在句(6)中则是关系的双方。

(1) 张先生踩到了李小姐。(基本作用句)

(2) 张先生爱上了李小姐。(一般反应句)

(3) 李小姐收到一封情书。(接收句)

(4) 张先生换了工作单位。(效应句)

(5) 张先生买了花园洋房。(交换句)

(6) 李小姐嫁给了张先生。(关系句)

(7) 李小姐有了两个孩子。(状态句)

(8) 张先生开始了新生活。(过程句)

可见，根据句类及其语义块的涵义，可以对句中动词所关涉的体词成分的语义角色做出适当的描述和解释。例如，在“他吃食堂”中，“食堂”的语义角色是什么呢？这个句子的句类是物转移句，“吃”是把食物转移到体内，在这类转移句中，常常要说明食物的制作者或提供者，“食堂”就是这个角色，而“吃”的东西不言自明，常常就省略掉了。“吃食堂”是说吃食堂的饭菜，而不是说吃的处所，因为吃食堂可以不在食堂里吃。

根据句类，也可以描述和解释句子的句法特征。例如，什么样的句子可以用“把”字句和“被”字句呢？只有作用句、转移句和思维判断句才可以，它们是 57 种基本句类中的三个大类。

HNC 理论的句类体系已有系统而清晰的阐释，它的 57 种基本句类的含义都已有明确的描述[7]。HNC 句类体系的描写能力已得到充分的检验，其重要体现之一是 HNC 汉语词语知识库的建设[10]。该知识库内有 80,000 多个词语，其中 25,000 多个是动词，每个动词所能形成的句子的语义类型和语义结构都用 HNC 的句类体系做了比较充分的描述，这是对该体系描写能力的很好的说明。

上面对 HNC 理论做了一点管窥式的介绍，目的是想说明，从可行性和可操作性的角度看，要开展基于语义的汉语句子研究，以 HNC 理论的句类体系为指导，是个较为理想的选择。

3. 建立现代汉语句子基本语义类型例句库的目标

我们基于 HNC 理论的基本句类建立现代汉语句子基本语义类型的例句库，要实现以下四个目标：

1) 为每个基本句类表示式配备丰富的例句。

HNC 理论的基本句类及其表示式就是现代汉语句子的基本语义类型。HNC 理论的基本句类有 57 种，每种基本句类至少有一个表示式，所有基本句类的表示式一共有 136 个。要为每个句类表示式都配备例句，例句要足够丰富，能够用来充分地说明该表示式的基本涵义和特征。

2) 每个句类表示式的例句要符合规范性、真实性、典型性和全面性的要求。

所谓规范性，是指例句要符合现代汉语的基本规范，不能有不规范的字词用法或语法表达等方面的问题。

所谓真实性，是指例句应来源于真实语料，而不能是生造的，至少不能让人觉得是生造的。上文所举“张先生爱上了李小姐”这样的例句，显然就不符合真实性的要求。例句应来源于真实语料，这并不是说只能原封不动地采用真实语料中的句子，而是可以对原始的句子进行必要的编辑和修改，以使之更符合规范性和典型性的要求。

所谓典型性，是指例句应该能够简明地体现出它所描述的句类表示式在某个方面的涵义或特征。来源于真实语料的原始句子常常不够简明，需要把妨碍典型性的次要成分精简掉。当然，精简的结果要符合规范性和真实性的要求。

所谓全面性，是指每个基本句类各方面的基本特征都应该有例句来体现。句类的特征称为句类知识，它有多方面的丰富内容[6]，这里不能详细阐释，只能做点举例性的说明。例如语句格式知识，是指语义块在句子中出现的顺序。句类表示式描述了一个句类应该有几个什么样的语义块，但没有限定这些语义块在自然语言语句中出现的顺序。句子中语义块出现的顺序不同，也就是语句格式的不同。例如物转移句有四个语义块，分别是转移发出者、转移行为、转移接收者和被转移物，它们在具体的句子中可以有多种出现顺序，如“张三送给李四一本书”、“张三送了一本书给李四”、“张三把那本书送给了李四”、“那本书由张三送给了李四”等。不同的句类所能采用的语句格式是不同的，应该为一种句类所能采用的各种语句格式都配备例句，这样才符合全面性的要求。

3) 制定规范，标注每个例句的语义结构。

根据 HNC 的句类表示式，句子的框架语义结构是由语义块构成的，标注语义结构也就是标明句子中的各个语义块。除了句类表示式中定义的语义块以为，句子中还会出现其他的语义块，HNC 理论称之为辅语义块，有 7 种类型[7]。对例句中出现的辅语义块，也要标明其类型。

4) 开发查询工具，方便找出所需的例句。

句子基本语义类型例句库的使用者很可能并不熟悉 HNC 理论，甚至完全不懂句类和语义块等基本概念，因此，例句库的查询工具必须能以简明直观的方式向用户提供适当的说明和帮助，引导用户找到所需的例句。

4. 句子语义类型例句的收集和标注方法

前文说过，要从真实文本中获得句子语义类型的例句，可以先标注文本中各个句子的语义类型，形成标注语料库，然后从该语料库中抽取所需的例句。但这样做的成本是很高的，因为句子的各种语义类型在真实文本中的分布是不均匀的，要想各种类型的句子都有足够多的数量，需要标注一个大规模的平衡语料库才行。因此，我们暂时没有采用这种方式。

我们的做法是，利用 HNC 词语知识库中动词的句类知识到语料库中查找例句。HNC 词语知识库[10]中有 25,000 多个动词，每个动词都描述了它所能形成的句子所属的句类，也就是它能形成哪种语义类型的句子，并根据句类知识描述了这些句子的基本特点。要查找某个基本句类的例句时，先到这个库中查找能形成这种基本句类的句子的动词，从中挑选出一些常用动词，然后到语料库中查找包含这些动词的句子，从中挑选出符合需要的句子作为该基本句类的例句。要对从语料库中查找到的例句进行筛选的原因是：一方面，一个动词形成的句子所属的句类可能不止一种；另一方面，作为例句的句子要符合上文所说典型性和规范性的要求。从语料库中找到满意的句子并不容易，常常需要对原始的句子进行必要的修改，才能得到一个适当的例句。

我们使用的语料库含有 22.2 万篇文本，共计 2.15 亿字，其中的文本大部分是书面语，在文体和题材等方面都有比较好的平衡性。

对例句的标注，我们采用了 XML(Extensible Mark-up Language，可扩展标记语言)技术。XML 是万维网联盟(World Wide Web Consortium, W3C)研发的一系列标准技术(http://www.w3.org/standards/xml/)，已得到广泛的支持和应用，采用该技术有两大方面的好处：一是保证语料标注结果的数据文档有良好的通用性；二是语料的标注、使用和维护都可以得到诸多便利。例如，可以用已有的功能强大的 XML 工具[8]标注语料，而不必自己开发标注工具；可以用 XML 的 Schema 技术(http://www.w3.org/standards/xml/schema)定义语料标注文档的数据结构，以保证其有效性和规范性；可以用 XQuery 技术(http://www.w3.org/stan dards/xml/query)对标注语料进行查询；可以用 XML 的 Trans-formation 技术(http://www.w3.org/standards/xml/trans-formation)把语料转换为 HTML 等文档格式。

利用 XML 技术对例句进行标注，其基础是要制定一个 Schema 文档，用来描述文档的数据结构，作为标注的规范。我们已经制定了这样的规范，根据该

[6] 可参阅苗传江的专著《HNC(概念层次网络)理论导论》[7]中的有关内容以及苗传江的其他相关文章[8-11]。

[7] 见黄曾阳的专著《HNC(概念层次网络)理论——计算机理解语言研究的新思路》[2]第 55 页。

[8] 例如 XMLSpy (http://www.altova.com/xmlspy.html)。

规范对例句进行标注，下面是标注结果的一个简单示例。

<sentence code = "X20">
<fk type = "Cn">在现代都市社会里</fk>，
<jk type = "1">人们</jk>
<ek>常常忽略</ek>
<jk type = "2">眼前的幸福和美丽</jk>。
</sentence>

一个<sentence>元素就是一个例句，它的属性 code 表示例句所属句类的代码，根据这个代码就可以确定该句类的表示式。<sentence>元素的子元素<ek>、<jk>和<fk>就是语义块，即句子的语义构成成分。包含在句类表示式中的语义块分为两类，内含述语动词的标为<ek>，其他的标为<jk>。句类表示式之外的语义块，即辅语义块，标为<fk>。<jk>和<fk>的属性 type 用于标明其具体类型。根据这些子元素的名称及其 type 属性的值，就可以确定句子中各个语义块的角色和涵义，从而描述清楚句子的语义结构。

5. 现代汉语句子基本语义类型例句库的发展

目前，现代汉语句子基本语义类型例句库的建设已取得初步成果，体现在以下三个方面：

1) 已为 57 种基本句类的每个表示式都配备了典型的例句。57 种基本句类的表示式总共有 136 个，其例句总共有 1286 个句子。例如，下面是为一般反应句(该句类只有一个表示式)所配备的例句：

- 人们常常忽略近在咫尺的幸福和美。
- 政治家们对一般公众的切身利益很少顾及。
- 王凤梅一直在为儿子的婚事发愁。
- 每一个炎黄子孙都为有光荣历史的祖国而感到骄傲。
- 雷诺夫人为自己奋斗了 9 年的夙愿变成现实而深感欣慰。
- 江泽民对厄立特里亚政府坚持“一个中国”的政策，不与台湾发生官方关系的原则立场表示赞赏。
- 面对政治腐败、民生凋敝的现实，他对“如花的祖国”感到失望和悲痛。
- 艾黎仍对儿时打雪仗的情景记忆犹新。
- 报纸所报道的新闻要真正为人民群众所喜闻乐见。
- 我们的工作始终不为领导所重视。
- 领导干部要这么做才能为群众所尊重。
- 有些父母常常为拒绝了孩子而感到愧疚。
- 虹鳟鱼对水温非常挑剔。

2) 每个例句都以上文所述的 XML 方式标注了其语义结构。

3) 已经开发了一个简单的例句查询工具。利用这个查询工具，可以方便地找出各个句类表示式的例句。在显示查询到的例句时，以不同的背景颜色直观地标示出语义块的不同类型。查询界面上有必要的提示和说明，帮助使用者理解句类和语义块等的基本知识。查询工具是用微软的 ASP.NET 技术和 Visual C# 语言开发的，是一个基于网络的应用程序。

以这个例句库为基础，我们将从以下四个方面进一步发展，努力为基于语义的句子研究积累基础资源。

一是向混合句类扩展。混合句类是两种或多种基本句类的组合，如前文所举一般反应句和信息转移句的混合，我们将为汉语中常见的混合句类配备例句。

二是向概念体系扩展。句子的基本语义类型对应着基本的概念范畴，比如反应句就对应着“反应”这一概念范畴，每个概念范畴在概念体系中都包涵若干子概念范畴，比如“反应”就包含“喜爱与反感、信任与怀疑、同意与反对”等子概念，每个子概念在一种语言中可能有多个词语。针对每种句类所涵盖的概念范畴，我们将为每类子概念所对应的常用词语都配备例句。

三是向真实的连续文本扩展。也就是分析和标注真实文本中每个句子的语义类型和语义结构，这样标注而成语料库，可称之为“语义树库”。

四是向英汉对照扩展。句子的语义类型是与语种无关的，但语句的具体表达方式则有很多方面的语种差异，有了基于句子语义类型的英汉对照语料库，就可以开展基于句子语义的英汉对比和翻译等方面的研究。

参考文献 (References)

[1] 黄曾阳 (1997) HNC 理论概要. *中文信息学报*, **4**, 11-20.
[2] 黄曾阳 (1998) HNC(概念层次网络)理论——计算机理解语言研究的新思路. 清华大学出版社, 北京.

[3] 黄曾阳 (2004) 语言概念空间的基本定理和数学物理表示式. 海洋出版社, 北京.
[4] 黄曾阳 (2010) HNC 理论全书. 中国科学院声学研究所内部资料.
[5] 蒋严, 潘海华 (1998) 汉语语句的类型表达. In: 黄昌宁, Ed., 1998 *中文信息处理国际会议论文集*, 清华大学出版社, 北京, 323-329.
[6] 鲁川 (2001) 汉语语法的意合网络. 商务印书馆, 北京.
[7] 苗传江 (2005) HNC(概念层次网络)理论导论. 清华大学出版社, 北京.
[8] 苗传江 (2006) 基于 HNC 句类体系的句子语义研究. *语言文字应用*, **1**, 126-133.
[9] 苗传江 (2007) 现代汉语句子的语义类型. *语文建设*, **11**, 56-58.
[10] 苗传江, 刘智颖 (2010) 基于 HNC 的现代汉语词语知识库建设. *云南师范大学学报*(*哲学社会科学版*), **4**, 15-18.
[11] 苗传江, 刘智颖 (2003) 现代汉语语料的句子级语义标注. In: 孙茂松, 陈群秀, Eds., *语言计算与基于内容的文本处理*, 清华大学出版社, 北京, 325-331.
[12] 徐烈炯 (1995) 语义学(修订本). 语文出版社, 北京.
[13] 朱晓亚 (2001) 现代汉语句模研究. 北京大学出版社, 北京.