为 NLP 创立模式，用 HNC 研究汉语

林 杏 光

中国人民大学对外语言文化学院

北京 100872

提要：本文介绍了 HNC 创立 NLP 模式的几点思路，论述了为什么要用 HNC 研究汉语和如何用 HNC 研究汉语的问题。结论：用 8 个网络研究词语，用 57 个基本句类研究语句，这是用 HNC 研究汉语的一种方法。

关键词：NLP HNC 模式 网络 句类 词语 语句

Application of HNC Theory to Chinese Study---A Paradigm of NLP

Xingguang Lin

The School of Foreign Language and Culture,

Renmin University of China,

Beijing, 100872, China

ABSTRACT: In this paper we study the HNC theory as a paradigm of NLP, We answer the two questions, namely, why should we use the HNC theory to Chinese study, and how do we employ HNC theory to study Chinese. Our main conclusion is to study the word by using the 8 concept networks and study the sentence by using the 57 basic categories of sentence, which is our method of study Chinese by HNC theory.

Keywords: NLP, HNC, Paradigm, Network, Category of sentence, Word, Sentence

0 解题

NLP 是“Natural Language Processing ”的速写，是“自然语言处理”的英文对译。HNC 是“Hierarchical Network of Concepts”的速写，是“概念层次网络”的英文对译。本文标题前半截的意思是，要让计算机处理自然语言，就要为计算机创立理解语言的模式；本文标题后半段的含义是，HNC 是 NLP 的一种模式，它既可以用来处理语言，也可以用来研究语言，汉语研究工作者应积极开展基于 HNC 的汉语研究。

1 为 NLP 创立模式

中文信息处理，目前正处在词语和语句处理的阶段。许嘉璐先生的《现状和设想》将中文信息处理的研究现状分为三个流派：第一个流派是以传统的计算语言学为基本理论的流派，第二个流派是 HNC 理论，第三个流派是基于内涵模型论的语义分析。我希望三个流派互相学习，取长补短，团结奋进，为中国能成为未来中文信息处理技术发展的中坚作出各自的贡献。这里试对第二个流派为 NLP 创立模式的思路谈几点认识。

1.1 模式的定位问题

我很重视定位问题，因为不管是研究什么，定位都非常重要。“理解”，这个概念很难定义，不同的学科对“理解”有自己的特殊认识。就同在 NLP 这一科学领域，对“理解”也有不同的定位。

1．据说有位女大学生在一位男大学生的书里夹了一个纸条：“花开堪折直须折，莫待无花空折枝。”男大学生好不容易才“理解”到女大学生夹纸条的意思是暗示他及时求爱，机不可失。

2．传说有对恋人在女方家里相会，分手时天下起雨来。女方说：“下雨啦！”女方的意思是别走了，住在这里吧！男方却对女方的话“理解”成“下雨怎么办呢？”于是回答说：“我有雨伞。”说罢走了，使女方很生气。

3．有一位外国朋友问一位外出回来的中国同志：“您到哪里去了？”中国同志回答说：“我打酱油去了。”外国朋友很惊异地问：“您为什么打酱油，它有什么罪，您为什么打它？”中国同志解释说：“‘打’是买的意思，打酱油就是买酱油。”

4．美国会反对这个提议。这个“会”可理解为上连“国会”，也可理解为下连“会反对”。但在具体的上下文中只能有一种理解。

5．前面来了一个人，（这个人）山东人长相，（这个人）秃脑袋，（脑袋）剃得挺光亮。括号内是代词所代或省略了的内容。

上述 5 个例子，可分为两类：1 和 2 为一类，其余为另一类。HNC 不要求计算机具有理解第一类问题的能力，而是将书面语言的“理解”定位于第二类。第二类问题包括多义词义项的选择、词或短语的切分、代词所代或省略内容的确定。HNC 将这些问题概括为三重模糊：词的多义模糊、语义块构成的分合模糊、指代冗缺模糊。HNC 将消解三重模糊作为书面语言处理的初步标准，我认为这个定位是恰当的。因为目前计算机的硬件和软件水平有可能实现这个定位；另方面这个定位一旦实现，在语言信息处理史上是一个很大的进步，它将促使机器翻译、智能检索、信息过滤、搜索引擎等语言信息处理应用领域产生实质性的进展。

1.2 在什么空间创立模式

多年来，我和中文信息界的一些朋友们总是在语言空间上打转转儿。我们这么想，人之

所以能理解语言，是因为人拥有语言知识和生活常识。要让计算机理解语言首先就要让计算机拥有语言知识。什么是语言知识呢？我们想无非是语法吧，语义吧，再加上语用吧。这种复制性的惯常思维总是设计不出便于计算机理解语言的好模式，原因是语言是一个无限的不确定集，计算机把握不住。HNC 以扩散性的求异思维，创新地跳到概念空间去创立语言理解模式。世界万物所占据的空间是物质空间，和物质空间相对应的是概念空间。语言空间是整个物质空间的一个子物质空间，跟语言空间相对应的一个小概念空间是整个概念空间的一个子概念空间。HNC 正是在这个子概念空间上创立理解模式。这个子概念空间不是一般所说的逻辑结构，而是反映语言内容的概念空间。反映语言内容的概念空间，其最基本的基元是有限的、确定的，那就是 HNC 所发现的作用效应链的作用、效应、过程、转移、关系、状态等 6 个环节加上判断。在这样的概念空间上创立的理解模式是有限的、确定的、封闭性的，它和众多的语言空间存在多种相互映射的形式，可以将无限的、不确定的语言描述到概念空间所创立的理解模式上去，计算机就有可能通过理解模式去把握无限的、不确定的语言内容。

1.3 创立什么样的模式

黄曾阳先生在《HNC 理论与自然语言语句的理解》这篇论文中指出：“自然语言理解的本质是概念联想脉络激活、扩展、浓缩、转换与存储的全过程运作。激活运作的要点是语句的理解；扩展与浓缩运作的要点是段落与篇章的理解；转换与存储的运作要点是记忆与学习。语句的理解显然是自然语言理解的基础，但这不等于说，任何一种形式的语句理解处理算法都可以成为自然语言理解的基础。要取得这一资格，就必须把语句理解定位于概念联想脉络运作全过程的激活。”黄先生的这段论述表明，HNC 创立的理解模式是以概念联想脉络为纲的激活、扩展、浓缩、转换、存储的语言感知过程的理解模式。

HNC 预计创立 5 个理解模式：词汇层面的理解模式、语句层面的理解模式、句群和篇章层面的理解模式、短期记忆和长期记忆的形成及相互转换的模式、自学习模式。目前，已经创立了两个理解模式，即概念表述模式和句类表述模式。这两个模式的目标是使计算机理解句子的类别（语义结构）和句子中各个词语的含义。

自然语言理解的“理解”就是“认知”。要解决“认知”问题，不可能脱离语言的“理解”。要理解语言就要通过词语和语句的网络形式。HNC 不但认识到了这一点，而且还提出了一整套使词语和语句网络化的策略，而这些策略正是研究汉语的新思路和新方法。

2 用 HNC 研究汉语

日本在第五代计算机（又称智能计算机）计划中没有突破自然语言理解这一科学难题之后，紧接着又由日本通产省下属的ＣＩＣＣ（国际信息合作中心）发起组织亚洲五国（日本、中国、泰国、印度尼西亚、马来西亚）专家研制《多国语言机器翻译系统》（简称ＭＭＴ）。作为项目的主持国，日本投资６０个亿，整整搞了 8 年，最终没有取得应有的效果。但这 8 年的合作对我们还是有帮助的，主要是开阔了我们的学术视野，引发了我们深沉的思考。我认识到，如果自然语言理解继续沿着国内外普遍采用的语法分析、语法语义分析、语料统计

等技术路线搞下去，计算机将很难理解人类的语言。正当我寻求新的技术路线的时候，１９９５年在国家科委召开的第４２次香山科学创新会上见到了黄曾阳先生，他使我高兴地看到了 HNC。从１９９５年开始，我一直在学习、考察、观看 HNC。我在 HNC 身上看到了什么呢？一句话，我看到了自然语言理解的新思路。这新思路既可以用来处理语言，也可以用来研究语言。这里谈谈用 HNC 研究汉语的问题。

2.1 为什么要用 HNC 研究汉语？

近些年来，愈来愈多研究语言的学者感到，纯语言学的单一语言研究前景不妙。我认为这种学术心态应该把它放到语言科学历史发展的潮流中去思考。中国五部著名的古典小说之一《三国演义》一开头就说："当今天下，合久必分，分久必合。"这说的是整个人类社会发展的规律。何勇、王海龙两位先生在《文化人类学与当代语言学研究论纲》一文中认为语言科学的发展也有这个趋势。远古时期，语言学是人类文化的结晶。在人类学术史上，几乎是每一位深刻的思想家、学问家，都对人类语言这一深刻文化命题进行过思考，做过深刻的研究。反过来，有见地的语言学家，他们的目光也不限于语言学本身，在他们的语言研究中涉及很多科学文化问题，并对这些问题进行科学的解释，提出了崭新的理论。这是语言学和人类文化综合的现象。随着科学的发展，学科的分工日益细化，于是把一些综合的学问分割开来，成为许多小的独立的学科。语言学从广大的文化背景下脱离出来，成为狭隘的语言学，限于书面语、口语、书写符号的探讨。这是语言学和人类文化分的现象。进入当代，语言学和人类文化分久又合，现在又强调站在宏观和阔大的角度去探讨语言现象。以上就是语言科学发展过程中的合－分－合的现象，后一个"合"不是前一个"合"的简单的循环，而是在更新更高境界中的新的综合。当前出现语言学和社会科学、自然科学结合的势头，正是语言学在新的历史条件下的更高境界的综合。

顺应语言学科发展的新形势，已经有不少语言学者认识到，信息时代的汉语研究主旋律应面向中文信息处理。如同建造房子必须按照建筑蓝图来施工一样，为中文信息处理服务的汉语研究理应按照汉语理解的模式来进行。HNC 是面向整个自然语言理解（包括汉语理解）的一种理论模式，针对这样的模式研究出来的汉语研究成果势必能更有效地为中文信息处理服务，所以我主张用 HNC 来研究汉语。

2.2 如何用 HNC 研究汉语？

HNC 的词语和语句层面的理解模式，已实现了从理论思路向技术转换的基本过程，并通过了专家鉴定。为此，本文试从词语和语句的角度来探讨如何用 HNC 研究汉语的问题。

2.2.1 词语的研究

HNC 将词语跟网络相联系，跟句类挂钩，用符号体系映射到和语言空间相对应的子概念空间的联想脉络上去，为计算机理解词语的意义提供可能的条件。为此，HNC 设计了 8 个网络：＜１＞基本物具体概念语义网络。＜２＞基本概念语义网络。＜３＞主体基元概念语义网络。＜４＞扩展基元概念语义网络。＜５＞语言逻辑概念语义网络。＜６＞基本逻辑概念

网络。＜7＞综合类语义网络。＜8＞“语法”网络。试举例说明 HNC 的 8 个网络对相应概念的描述：

基本物具体概念语义网络。“土地”ＪＷ５３８８（ＪＷ代表基本物具体概念语义网络，“５”代表这一网络第一个层次的 5 号节点“宏观基本物”，“３”代表这一网络第二个层次的 3 号节点“固态物”，“８”代表这一网络第三个层次的 8 号节点“土”，“８”代表这一网络第四个层次的 8 号节点“土地”)。将“土地”的符号变成文字说明即是，土地＝基本物具体概念语义网络ＪＷ：宏观基本物＋固态物＋土＋土地。

基本概念语义网络。“次序”Ｊｇ００（Ｊ代表基本概念语义网络，g 代表五元组的静态，“０”代表这一网络第一个层次的一号节点“序及广义空间”，“０”代表这一网络第二个层次的一号节点“序”)。将“次序”的符号变成文字说明即是，次序＝基本概念语义网络Ｊ：静态＋序及广义空间＋序。

主体基元概念语义网络。“好处”ｇ３２１（内定省略了主体基元概念语义网络符号(，g 代表五元组的静态，“３”代表这一网络第一个层次的三号节点“效应”，“２”代表这一网络第二个层次的二号节点“利害”，“１”代表这一网络第三个层次的一号节点“利”)。将“好处”的符号变成文字说明即是，好处＝主体基元概念语义网络℃：静态＋效应＋利害＋利。

扩展基元概念语义网络。“策略”ｒ８３０（内定省略了扩展基元概念语义网络符号(，r 代表五元组的效应，“８”代表这一网络第一个层次的八号节点“思维活动”，“３”代表这一网络第二个层次的三号节点“策划和设计”，“０”代表这一网络第三个层次的一号节点“策划与决策”)。将“策略”的符号变成文字说明即是，策略＝扩展基元概念语义网络(：效应＋思维活动＋策划和设计＋策划与决策。

语言逻辑概念语义网络。“被”ｌ００（ｌ代表语言逻辑概念语义网络，“０”代表这一网络第一个层次的一号节点“主语义块标志符”，“０”代表这一网络第二个层次的二号节点“特征语义块标志符”)。将“被”的符号变成文字说明即是，被＝语言逻辑概念语义网络ｌ：主语义块标志符＋特征语义块标志符。

基本逻辑概念网络。“势必”Ｊｌｕｖ１２ｃ３３（ｊｌ代表基本逻辑概念网络，ｕＶ代表五元组属性的动态，“１”代表“基本判断”，“２”代表“判断的客观势态”，“ｃ”表示对比性概念，“３”表示分三级，“３”表示第三级)。将“势必”的符号变成文字说明即是，势必＝基本逻辑概念网络ｊｌ：属性的动态＋基本判断＋判断的客观势态＋对比性概念＋分三级＋第三级。

综合概念语义网络。“方法”ｓｇ２２(方法＝综合概念语义网络ｓ:静态＋手段＋方法)。

“语法”概念网络。“到底”ｆ４２０ｂ（到底＝“语法”概念网络ｆ：语句格式＋疑问＋一般疑问＋问内容)。

汉语研究工作者如能用 HNC 的网络来描述汉语的词语，并推出工程性的研究成果，一方面可加速 HNC 工程化的进程，另一方面也可为汉语的词语研究开拓一条新路子。

2.2.2 语句的研究

HNC 从哲学的高度思考世界万物及其运动发展演变的规则，思考如何使语言的表述具有完备性，创造了作用效应链。这作用效应链的“作用－效应－过程－转移－关系－状态”等 6 个环节加上“判断”就是自然语言所表述的完备内容。HNC 以作用效应链的六个环节和判

断作为句子语义分类的标准，划分出7个句类基元，即：作用句（例：有关部门撤销了责任人的职务）；效应句（例：欧洲爆发了革命）；过程句（例：演出开始了）；转移句（例：这批货物将运往天津港）；关系句（例：少数大国主宰着世界的命运）；状态句（例：晚会并不精彩）；判断句（例：中国认为人权是具体的，也是历史的）。

基本概念的演绎系统需具备三个条件：相容性（无矛盾）、完备性（概括无余）、独立性（彼此有联系但不可取代）。7个句类基元是自然语言所表述的完备内容，HNC以7个句类基元为基本概念进行演绎，演绎出57个基本句类。

HNC的57个基本句类是：＜1＞基本作用句；＜2＞作用效应句；＜3＞约束句；＜4＞一般承受句；＜5＞主动承受句；＜6＞被动承受句；＜7＞特殊承受句；＜8＞一般反应句；＜9＞主动反应句；＜10＞被动反应句；＜11＞作用反应句；＜12＞一般免除句；＜13＞主动免除句；＜14＞被动免除句；＜15＞一般转移句；＜16＞物转移句；＜17＞信息转移句；＜18＞自身转移句；＜19＞针对性接收句；＜20＞对等交换句；＜21＞控制交换句；＜22＞先入交换句；＜23＞先出交换句；＜24＞基本替代句；＜25＞扩展替代句；＜26＞扩展双向替代句；＜27＞基本主从关系句；＜28＞扩展主从关系句；＜29＞扩展双向关系句；＜30＞一般判断句；＜31＞块扩判断句；＜32＞双对象效应句；＜33＞传输句；＜34＞接收句；＜35＞交换句；＜36＞变换句；＜37＞双向关系句；＜38＞一般效应句；＜39＞基本效应句；＜40＞一般过程句；＜41＞基本过程句；＜42＞素描句；＜43＞因果果因句；＜44＞双向替代句；＜45＞一般状态句；＜46＞基本状态句；＜47＞两换位状态句；＜48＞三换位状态句；＜49＞相互比较判断句；＜50＞标准比较判断句；＜51＞是否判断句；＜52＞存在判断句；＜53＞简明判断句；＜54＞参照比较判断句；＜55＞集内比较判断句；＜56＞势态判断句；＜57＞简明状态句。我们虽未能用数学的方法证明57个基本句类的“完备”性，但由于它是7个句类基元演绎出来的结果，同时迄今为止我们还不能在主要的几种自然语言中找到这57个基本句类覆盖不了的语句，因此，我们称这57个基本句类具有穷尽语句的“完备”性。

HNC的每一个句类都有自己的表示式，如“基本作用句”的表示式是：XJ＝A＋X＋B；B＝XB＋YB＋YC。“作用效应句”的表示式是：XYJ＝A＋XY＋B＋YC；YC＝（E）＋EC。“结束句”的表示式是：X4J＝A＋X4＋X4B，X4B＝XB＋YB，＝X4BB＋X4BC；X401J＝X4＋X4B，X4B＝X4BB＋X4BC%，X402J＝X4B＋X4。其余各句类表示式从略。

HNC每一句类表示式的符号都代表一定的内容，而且每一个句类都包含句类的知识。如“基本作用句”：XJ代表作用句。J是ju的声母，表示句子。A是作用者语义块，X是作用特征语义块，B是作用对象语义块。XJ（基本作用句）＝A（作用者语义块）＋X（作用特征语义块）＋B（作用对象语义块），例：张三打了李四。B（作用对象语义块）有要素的复合构成，包括三个部分：XB（作用对象）、YB（效应对象）、YC（效应内容）。B语义块复合构成的三个部分有7种组合格式：＜1＞只有作用对象。例：张三打了（李四）。＜2＞只有效应对象。例：他们打扫（教室）。＜3＞只有效应内容。例：他打扫（卫生）。＜4＞作用对象＋效应对象。例：上级撤消了（他的＋职务）。＜5＞作用对象＋效应内容。例：他正打扫（房间的＋卫生）。＜6＞效应对象＋效应内容。例：剧团大力推动（改革的＋进程）。

＜7＞作用对象＋效应对象＋效应内容。例：他打扫（我们＋房间的＋卫生）。在汉语中B语义块复合构成的各个部分的排列顺序是：作用对象－效应对象－效应内容。作用对象必是具体的，效应内容必是抽象的，效应对象有两可的模糊性。同一词语在不同的情况下有可能是不同的构成成分，如“垃圾”，我们打扫（垃圾），是作用对象；我们打扫广场的（垃圾），是效应对象。从作用对象语义块B的复合构成可以看到，传统语法学所说的定语，在 HNC 句类语义块系统中兵分两路：一部分充当语义块要素的复合构成；一部分作为语义块要素的修饰成分构成语义块的复合构成。如“张三打断李四受伤的那条腿。”其中的“李四”作为语义块要素和“腿”构成复合语义块，“受伤的”“那条”作为修饰成分和“腿”构成复合语义块。另外，复合构成的语义块可以产生分离现象。如“张三打断李四的腿”，“李四的腿”是复合构成的语义块。“李四被张三打断了腿”，“李四”和“腿”还是一个复合构成的语义块，但已分离开了。

汉语研究工作者如能用 HNC 的句类语义块系统来研究汉语的句子，并推出工程性的研究成果，一方面可加速 HNC 产业化的进程，另方面也可为汉语的语句研究开辟一条新的途径。

3 结语

２００１年的“两会”期间，朱镕基总理所作的《关于国家第十个五年计划社会发展和国家经济发展纲要的报告》提出一个要求：“促进自然科学和社会科学的交叉融合。”我写作本文的目的就是为了促进计算机科学和语言学的进一步交叉融合。许嘉璐先生多次指出：“中文信息处理需要‘两栖’学者，这就要求计算机工作者和语言学工作者很好地结合起来。”许先生还进一步指出：“语言研究和计算机技术一结合，所带来的不仅是中文信息处理事业的顺利发展，而且有可能引发语言研究的一场革命。”本文提出用 HNC 研究汉语的目的正是想达到许嘉璐先生所说的效果。本文的结论是：用 8 个网络研究词语，用 5 7 个基本句类研究语句，这是用 HNC 研究汉语的一种方法。汉语研究工作者应积极开展基于 HNC 的汉语研究。如何用 8 个网络研究词语？如何用 5 7 个基本句类研究语句？这两个问题尚待进一步具体论述。

本文所谈到的对 HNC 的有关看法是我个人的理解。如有理解不当之处，请批评指正。

参考文献

[1] 许嘉璐：《现状和设想（试论中文信息处理与现代汉语研究）》（《中国语文》2000 年 6 期）
[2] 黄曾阳：《HNC（概念层次网络）理论》（清华大学出版社 1998 年 11 月出版
[3] 黄曾阳：《HNC 理论与自然语言语句的理解》（《中国基础科学》1999 年 2--4 期）