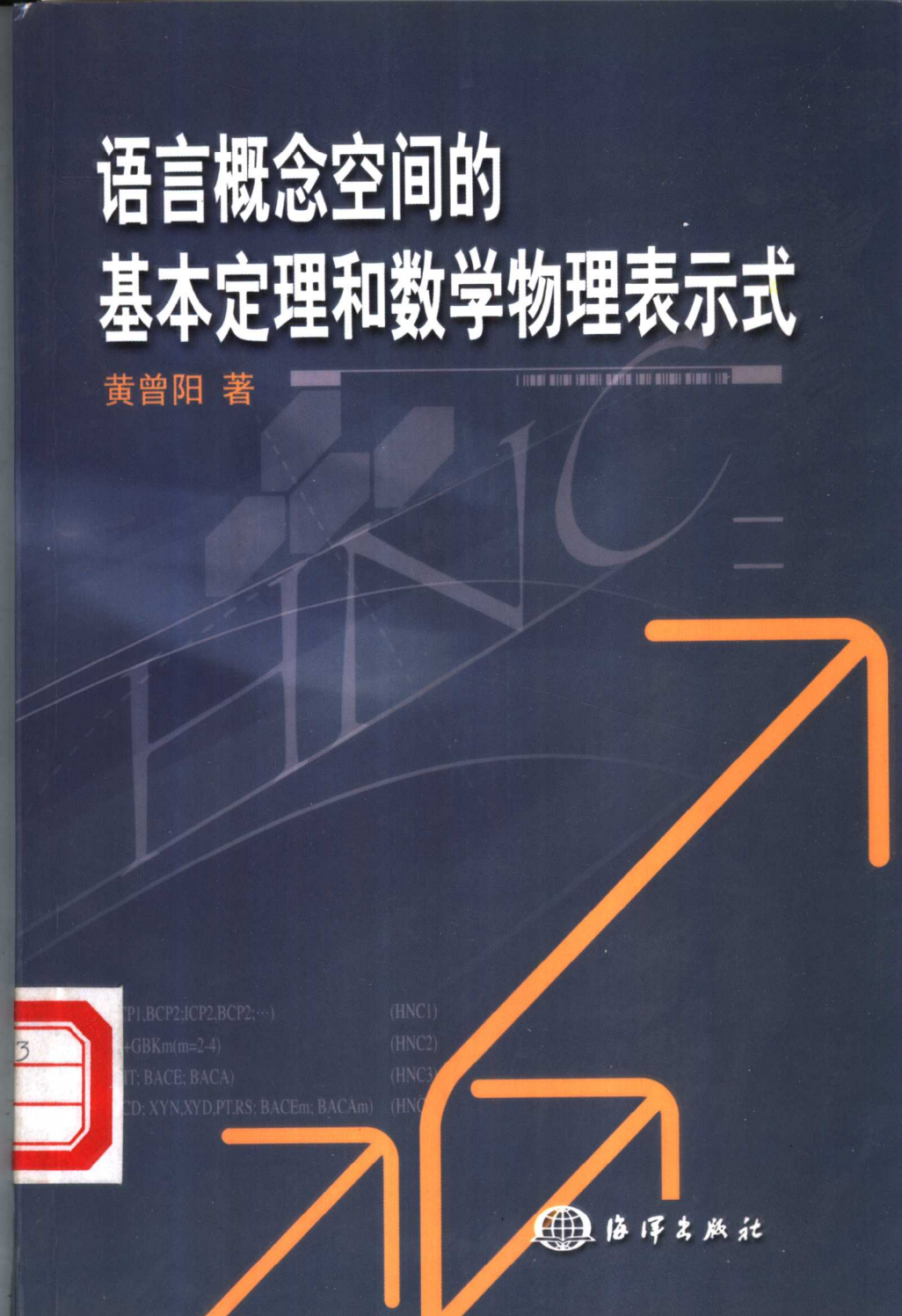
# 语言概念空间的基本定理和数学物理表示式

黄曾阳 著




海洋出版社

# 语言概念空间的
# 基本定理和数学物理表示式

黄曾阳 著

海洋出版社
2004年 北京

**图书在版编目(CIP)数据**

语言概念空间的基本定理和数学物理表示式/黄曾阳著.
北京:海洋出版社,2004.7
ISBN 7-5027-6135-7

Ⅰ.语… Ⅱ.黄… Ⅲ.机器翻译—研究
Ⅳ. H085

中国版本图书馆 CIP 数据核字(2004)第 058139 号

责任编辑:阎 安
责任印制:严国晋

海洋出版社 出版发行
http://www.oceanpress.com.cn
(100081 北京市海淀区大慧寺路 8 号)
北京蓝空印刷厂印刷 各地新华书店经销
2004 年 7 月第 1 版 2004 年 7 月北京第 1 次印刷
开本:850mm×1168mm 1/32 印张:5
字数:120 千字 印数:1~2000 册
定价:16.00 元
**海洋版图书印、装错误可随时退换**

2004 年摄于中国科学院声学研究所德昭楼办公室

# 作 者 简 介

黄曾阳，湖北蕲春人，1935 年出生，1958 年毕业于北京大学物理系。嗣后，一直在中国科学院声学研究所工作，1985 年晋升为研究员。

1988 以前主要从事水声物理、信号处理和声呐系统的研究，先后主持过流体动力噪声、恒虚警处理、光信号傅里叶处理技术、豚类信号分析、鱼雷靶场测试系统、多波束声呐自动判决系统、黄河根石探测系统等项目。1989 年后转向自然语言处理领域，先后在中国科学院的“八五”重大项目、国家计委的“九五”攻关项目、国家“973”项目和“863”项目、中国科学院创新工程中担任自然语言处理的课题负责人。1979 年开始，先后指导过硕士、博士生和博士后共 20 余人。

作者在 40 余年科学探索生涯中的最大贡献是创立了“HNC(概念层次网络)理论”。这一理论的得以创立与作者在幼少年时期接受过系统的国学教育和作者对东西方经典哲学持久不衰的浓厚兴趣密切相关。该理论揭示了人类大脑中交际引擎的运作原理，构造了概念基元、句类、语境单元和语境框架四层级符号体系的数学物理表示式。为交互引擎的研制奠定了坚实的理论基础。

交互引擎是交际引擎的仿制，它将使计算机具有理解自然语言的功能，成为“一目千行，过目不忘”的多语种超级‘才子’(语言超人)。本书描述的交互引擎的问世必将对刚刚诞生的网络世界产生深远的影响。

作者关于 HNC 的第一本专著《HNC(概念层次网络)理论——计算机理解语言研究的新思路》于 1998 年出版以后，引起了学界的特殊关注。2001 年和 2003 年曾先后在武汉和北京召开过两届“HNC 与语言学研讨会”，并出版了会议论文集。HNC 的阶段性技术成果在北京大正语言知识处理研究院(HNC 研究院)正在孵化产品，已取得可喜成绩。

# 编者的话

本书是《HNC(概念层次网络)理论——计算机理解自然语言研究的新思路》(清华大学出版社,1998年出版。以下简称《HNC理论》)的续篇。

HNC理论是一个关于自然语言理解的理论框架,由中国科学院声学研究所黄曾阳研究员创立于1993年。该理论的创立者突破西方语法学和我国传统语法学的束缚,独辟蹊径,坚持通过语言概念空间研究语言现象,揭示了语言概念空间的四层级结构,给出了该空间四层级符号体系的数学物理表示式,并以此为基础形成自然语言理解处理的交互引擎技术。

已故中国工程院院士陈力为先生在《HNC理论》的题词中写到:HNC所走的以语义表达为基础的新路子,对突破汉语理解问题尤其有实际意义。林杏光先生在"编者的话"中这样评价这条新路,"它完全摆脱了我国现有的这套语法学的束缚,而从语言的深层入手,以语义表达为基础,为汉语理解开辟了一条新路"。HNC理论已成为中文信息处理的三大流派之一,这是著名语言学家许嘉璐副委员长在《现状和设想——试论中文信息处理与现代汉语研究》一文中对HNC理论的评价。著名哲学家中国人民大学黄顺基教授在《HNC理论和科学哲学的关系》一文中评价HNC理论"在科学上为认知科学、语言学与人工智能的研究提供了一个全新的理论框架"。中国计算语言学专业委员会第一届主任鲁川先生,在《有关"科学"和"语言"的畅想——浅谈HNC的学科定位》一文中写到,HNC是自然科学跟人文科学交叉的新兴"智慧科学"的一个典型代表。

HNC理论的总体架构也可以划分为语句理论和篇章理论两部

分，每一部分又包含两个理论模式。《HNC 理论》仅涉及语句理论部分，本书则主要涉及篇章理论部分。如果说，1998 年出版的《HNC 理论》是这条新路第一阶段理论探索的总结，那么，本书则是 HNC 第二阶段理论探索基本完成的标志。

篇章理论两模式的建立历经了整整 10 年时间。最初，黄先生认为段落和篇章的联想脉络远比语句层面复杂，“不可能用一组数学物理表示式来表达”，同时还要建立一种计算机记忆和自学习模式。直到 2001 年，萌发了“领域句类”的概念，多年的茫然才出现了转机，语境无限而语境单元有限的概念豁然来临，语境单元和语境框架的数学物理表示式水到渠成，记忆与学习机制自然而然地融进语境单元和语境框架这两个模式的相互转换之中。

这里应该向读者说明的是，《HNC 理论》中阐述的前两个理论模式现在正式命名为概念基元模式和句类模式。概念基元模式的符号体系由高层概念网络和概念延伸结构构成。《HNC 理论》主要给出了高层概念的描述，而本书引进了概念范畴—概念林—概念树—根概念这四个新术语，从而更清楚、更准确地描述了高层概念的有限性和层次性。延伸结构由中层概念基元和底层概念基元灵活地组成，从根概念开始的延伸结构主要描述概念的网络特性。

黄先生在本书中大胆地提出，概念延伸结构采用的“符号再抽象原则”也许就是大脑理解自然语言的核心机制或奥秘。大脑对词语、语句、句群和篇章理解的运作原理（即交际引擎运作原理）就是基于本书给出的四组数学物理表示式。如果计算机配置四层级符号体系，并形成基于四组表示式的三级提升技术，那么，它就可以实现对交际引擎的模拟，具备理解自然语言的功能。

著名的英国物理学家和生物化学家弗朗西斯·克里克教授，曾勇敢地选择了视觉处理作为研究大脑奥秘的突破口，并在 1994 年发表了著名的《惊人的假说》，他认为这一困难的且具有极大魅力的研究课题任重而道远。黄先生选择语言理解作为研究人脑理解机制的突

破口，并创立了HNC理论，这一选择是对网络时代的呼唤更有力的响应。基于该理论的交互引擎有可能诞生于在中华大地，这将是一项对网络时代产生深远影响的重大技术发展。

HNC理论框架完成以后，黄先生应该撰写一部专著，以便系统地、全面地阐述HNC理论，但黄先生正在把主要精力转向三部手册这项巨大的理论工程。于是，我建议出版一本介绍HNC篇章理论的小册子，并得到了黄先生的同意和支持。在HNC理论探索的过程中，黄先生一直笔耕不辍，共写下了约120万字的九个专题序列，1998年出版的《HNC理论》只是其中四个专题的一部分。1999年以后撰写的论文序列有《自然语言理解处理的20项难点及其对策》、语境研究和语料标注序列、HNC机器翻译序列、《概念基元符号体系手册》的论述序列和HNC理论序列。本书选取了HNC理论序列的五篇论文，它们集中反映了近五年来HNC理论研究的最新成果。

由于上述文章的写作时间不同，对同一个概念会出现不同的描述术语，虽然保存了探索过程的原貌，却会给一些读者带来不便。另外，由于编者本人的水平有限，时间比较仓促，还可能出现这样或那样的不当之处，敬请专家和学者多多谅解并不吝指正。

杜燕玲

2004年6月26日

# 序　言

2003年9月 中国科学院声学研究所和北京大正语言知识处理研究院联合举办了第二届HNC与语言学研讨会，我在会上作了《在反思中前进，在碰撞中成长》（简称《反思与碰撞》）的发言，会后将发言稿整理成48 000字符的同名长文。在会议论文集行将出版的时候，论文集主编之一杜燕玲副研究员建议我将该文和我近年应邀所写的另外四篇论文合成一本小册子一同出版。我非常感谢并欣然接受杜主编的这一建议，将这本小册子定名为《语言概念空间的基本定理和数学物理表示式》，并请她承担起已故林杏光教授在《HNC（概念层次网络）理论》（简称《HNC理论》，即本书的前奏）里的编者角色。

在接受建议并给小册子定名的同时 我心里也十分忐忑不安。因为 这一次在一定程度上似乎又在重蹈《HNC理论》出版过程的覆辙。但《HNC理论》有《HNC（概念层次网络）理论导论》的弥补，相信这本小册子将来也会有相应的弥补。因此 我就决心再次献丑了。

所有的自然语言空间（据说 当今世界上还存在6 000种之多）对应着同一语言概念空间，这是HNC理论的第一基本假设。语言空间各有自己的个性，语言研究固然需要面向这些个性，但更需要面向语言空间的共性。更具体地说 对那些寓于个性之中的共性的研究才是语言本体的研究。一个常见的认识误区是：以为语言本体的研究就是思维的研究，这两者是不能画上等号的，语言本体只是思维的一部分，语言是“思维的外壳”这一著名论断实际上应加上“之一”二字，因为 无比丰富的数学符号、图形符号、艺术符号都是思维的外壳。因此 必须承认 语言概念空间只是概念空间的子空间之一，此外 至少还有形象、情感、艺

术、科学等概念空间。建立所有概念空间的统一理论是哲学的任务，这一任务的艰巨性甚至大于物理学的统一场论。认知科学的探索者已经体会到这一探索的艰辛，于是主要采取狐狸式探索，因为他们认为 刺猬式探索在此必然是徒劳无功。但是 如果把研究目标局限于语言概念空间，是否会出现峰回路转的新局面？

这本小册子希望对这个基本问题给出 HNC 的初步答案。

这个初步答案集中体现在《反思与碰撞》一文所阐释的“概念无限而概念基元有限、语句无限而语句概念类型(句类)有限、语境无限而语境单元有限”这一基本论点里，或简称“三限说”。整个 HNC 理论体系的构架是以“三限说”为基础而展开的。概念基元的有限性可称为 HNC 第一定理，句类的有限性可称为 HNC 第二定理，语境单元的有限性可称为 HNC 第三定理。该文并未使用定理一词，而是把它们作为三项需要进行求证的假设来对待。

按成文的时间顺序来说 最早的《HNC 理论与自然语言语句的理解》一文是对第二定理的论述，同时提出了翻译引擎基本原理的研究方向；第二篇《HNC 的发展和展望》是对第一和第二定理的系统论述；第三篇《语义及概念体系在 NLP 中的作用》侧重于第一定理的论述，第四篇《HNC 理论的语言学基础》触及三个定理的全部核心科学问题，最后一篇《反思与碰撞》则侧重第一和第三定理的论述，并对 HNC 理论体系最近五年的发展作了一个全面的概括。因此 编者将此文列为这本小册子的首篇是合理的。

《反思与碰撞》一文给出了语言概念空间的四组表示式——(HNC1)、(HNC2)、(HNC3)和(HNC4)，其中的(HNC1)是概念基元空间之根概念的数学表示式，(HNC2)是句类空间的数学表示式，(HNC3)是语境单元空间的数学表示式，(HNC4)是语境空间的数学表示式。但应该说明：后两个表示式实际上也是相应语言概念空间的物理表示式。至于句类空间的全部物理表示式请参看苗传江博士所著的《HNC(概念层次网络)理论导论》。这些物理表示式还需要扩

展它的符号功能，这项工作的进展将在《HNC探索与实践》网络季刊里即时报道。

以上列四组表示式为依托《反思与碰撞》一文提出了交互引擎的概念，描述了实现交互引擎的三项理论工程、四项技术工程和一项基础工程。交互引擎研究与开发在后工业时代占有特殊地位，它的实现已经不再是一个不切实际的梦想了。这是一个重要信息，也是这本小册子希望向读者传达的基本信息。

最后 作者要向海洋出版社表示最深切的谢意和最钦佩的敬意。谢意而最是因为主事者对我这个已脱离海洋科学战线15年的老朋友依然给予特殊关照与支持，敬意而最是因为主事者同意这本小册子采用作者建议的逗号改革方案。现代汉语逗号的语言功能太多，但众多功能中的小句标记与非小句标记之区分对语句理解处理至为关键，如果逗号仅用于小句标记，而把其他标记功能一律以空格替换，则对于汉语理解处理将产生功德无量的效果。空格是汉字的潜在财富，汉字改革的一些先行者曾经意识到这一点，并建议用于分词标记。这一建议也许由于汉语分词本身就存在众多难题与争议而未获实行，但空格用于逗号改革则简单易行，作者已实际使用了多年，并在HNC内部读物中流行。这一次 海洋出版社接受作者的请求，打破常规，开历史之先河，在正式出版物中第一个试用这一方案，这一举措使作者感受到主事者的战略眼光，无限钦佩之情在多年备受冷落之后不禁油然而生。

黄曾阳

2004年3月1日

# 目　次

# 在反思中前进，在碰撞中成长*

**摘要** 本文是《HNC(概念层次网络)理论》的续篇，概述了HNC理论探索近五年来的基本成果，对HNC的某些传统论述进行了反思与改正，同时给出了HNC未来15年探索的素描。本文正式将HNC的探索目标定位于交互引擎的研发，将语言概念空间描述的原五模式说改为概念基元空间、句类空间、语境单元空间和语境空间的四模式说，给出了这四个模式的数学表示式；将交互引擎的基本模块凝练成句类分析、语境单元萃取和语境生成的三部曲或三个基本环节。为达到交互引擎的彼岸，本文阐释了这一科学征途的三项理论工程、四项技术工程和HNC研发平台的概貌。为说明这一彼岸的理论可行性 本文诠释了HNC的两个基本假设和"概念无限而概念基元有限、语句无限而句类有限、语境无限而语境单元有限"的基本论点。

## 1 引言

HNC理论基本构架的形成已经整整10年了。

HNC这10年的历程是一个幸运的历程。

小托尔斯泰有一部不朽名著叫《苦难的历程》，该名著三部曲的名称是《两姐妹》、《一九一八年》和《阴暗的早晨》。HNC的幸运历程也有一个三部曲，可类比定名为《两仙翁》(按：指语法与逻辑或语言

---

* 本文发表在《第二届HNC与语言学研讨会论文集》，海洋出版社，2004年。

学与哲学)、《一九九七年》和《明媚的晨曦》。

小托尔斯泰完成那"苦难历程"的写作时间是在小说描述事件的20年之后,如果以这个时间为参照 那么 HNC 幸运历程的描述放到2023年以后比较适当,是今天座中年轻才俊进入"知天命"或"从心所欲"的时候,你们到那个时候看着办吧。

但是 我今天还是要说一点,HNC 幸运历程的最大幸运是什么?那就是遇到了众多的伯乐。由于近年我成了一个经常阅读圣经的无神论者 所以 他们成了我心中的科学而非神学的弥赛亚。现在 让我们以热烈的掌声向这些科学弥赛亚表示我们的无限敬意与感谢。有两位科学弥赛亚——陈力为院士和林杏光教授已经离开我们了,让我们以肃穆的沉思表示我们永恒的缅怀。

中国历史最辉煌的年代是唐朝,唐朝最伟大的文豪兼哲人是韩退之(愈)先生,退之先生最富于哲理的名篇之一是《进学解》,《进学解》里最著名的警句之一是:"业精于勤,荒于嬉;行成于思,毁于随。"从某种意义上说 本文是学习《进学解》的一点体会。

我的体会之一是 退之先生说的"思"就是哲学意义的反思,"随"就是现代意义的拒绝创新与改革。"随"的对偶概念是"撞",无"思"无"撞"就不能前进与成长。所以 本文定名为《在反思中前进,在碰撞中成长》。

世俗意义下的反思仅对内而不对外,碰撞仅对外而不对内。本文的反思和碰撞则兼对内外,且都以对内为主。

不过 本文并不直接以反思与碰撞为中心展开论述,而是把两者融合到关于自然语言之电脑理解的以下三个基本问题的阐释里。这三个基本问题是:

(1) 为什么已经具有许多奇妙智能的电脑到今天还不能理解自然语言? 这里的根本障碍到底是什么? 怎样克服这一障碍?

(2) 假定电脑能够理解自然语言,那么 让它成为一个"一目千行,过目不忘"的超级才子(语言超人)存在什么障碍? 怎样克服这一障碍?

(3) 假定电脑能够理解自然语言,那必然有一个从低级到高级的成长过程,那么 这一过程主要阶段或基本步调是什么? 我们怎样推动这些阶段或步调的实现?

围绕着第一个问题的阐释 本文将着重介绍 HNC 理论规划中的三项理论工程,简称三部《手册》——《HNC 概念基元符号体系手册》、《句类知识手册》和《语境单元知识手册》。

围绕着第二个问题的阐释 本文将简要介绍 HNC 技术规划中的四项技术工程——句类分析技术 SCA(Sentences Category Analysis)、语境单元萃取技术 SGUE(Sentences Group Unit Extraction)、语境生成技术 ABS(ABStract)和翻译引擎技术。

围绕着第三个问题的阐释 本文将简略说明 HNC 的一项特定计算工程——计算机自然语言理解度的测试与学习平台,这个平台的首要目标是进行电脑理解度的量化与类化研究,将简称 HNC 研发平台。

如果把这三项阐释所范定的研究目标的实现比做一次重大战役,那么 第一项阐释关系到该战役的战略方针的确定,第二项阐释关系到主战场或主攻方向的选择,第三项阐释关系到关键信息的搜集与分析、先进武器的供应与保证。毋庸论证 这三者对于赢得一场战役的胜利具有决定性意义。

HNC 通过语言概念空间研究语言现象。语言概念空间是存在于人类大脑之中的一个符号体系,这个符号体系既是人类进行语言思维的载体,又是人类进行语言交际的引擎,这个符号体系将简称交际引擎。计算机要获得理解自然语言的能力就必须也拥有一台在功能上类似的引擎,否则计算机就永远不可能成为名副其实的电脑。这个引擎将叫做交互引擎。交互引擎是交际引擎的仿制,仿制就必然存在简化与不足。交际引擎由千亿数量级的神经元构成,交互引擎将由当前电脑的二进制数字符号构成。HNC 的基本定位就是催生交互引擎的问世,这是天方夜谭式的神话么? HNC 不持这种悲观态度。HNC 已经看到 如果能够实施上述三项理论工程、四项技术工

程及其特定计算工程，则完全有可能在15年左右的时间里 使交互引擎在中华大地诞生。

## 2 关于交互引擎的三项理论工程——三部《手册》

本章是本文的重点。将分六节进行论述。2.1节对HNC探索的重要先行者作简要回顾；2.2节介绍HNC理论的基本定位及其两个基本假设；2.3节论述HNC探索预定的三部《手册》，不仅是三项理论探索，而且是三项巨大的理论工程；2.4节是本章的第一重点，介绍第一部手册——《HNC概念基元符号体系手册》的要点，重点阐释了HNC概念基元符号体系设计五项基本原则的延伸原则；2.5节介绍第二部手册——《句类知识手册》的要点，对句类概念的HNC传统论述中的失误作了系统的反思；2.6节是本章的第二重点，提出了交际语境和交互语境的概念，阐释了交互引擎研发的基本思路与框架，对HNC理论体系的整体描述进行了反思，将HNC理论体系的原五模式说改成四层级与三提升说，列举了语境单元萃取和语境生成研究的基本课题，概述了HNC理论体系四组表示式——(HNC1)、(HNC2)、(HNC3)和(HNC4)的物理意义。

### 2.1 对HNC先驱者的简短回顾——交互引擎的背景说明1

已经具有许多奇妙智能的电脑到今天还不能理解自然语言，为什么？这个问题的答案很复杂，但是又似乎很简单，有一个简单答案是这样的：要理解人类语言就必须拥有世界知识，而世界知识浩瀚无垠，需要软件工程师在语言学家的配合下一点一滴地教给电脑，这个工程太浩大了。面对这一浩瀚无垠的困难，人们思考着并实际采取过许多对策，这些对策可以概括成三种基本类型：最大简化策、愚公移山策和最大似然策。

最大简化策的基本思路是：设定一个受限环境，让这个环境的世

界知识十分有限，且语言表达十分规范，此策的代表人物之一是维诺格拉德教授。

愚公移山策的基本思路是：将世界知识划分出一定的类型和层次，用谓词逻辑形式或结构化数据形式写出来交给电脑，前者的代表人物是美国 CYC 计划的主持者 Lenat 教授，后者则代表人物众多。

最大似然策当前最走红，领唱者甚多，自称语料库学派。其核心思想是把言语当做一个随机过程，依托大规模真实语料制造一个计算机的进化环境，让计算机在这个环境里以大体类似于宇宙进化的方式自行进化。

前两种对策基本以所谓理性主义为依托，第三种对策基本以所谓经验主义为依托。

三种对策的基本思路朴素而简明，都具有朴素性的固有力量。

但是 对于“已经具有许多奇妙智能的电脑到今天还不能理解自然语言”这样重大的科学问题 仅仅依靠朴素性的固有力量是不够的，需要进行多侧面多层次的反思。我们至少需要反思以下三个基本问题：第一 为什么开始只有极少世界知识的 1～5 岁正常儿童都是学习自然语言的天才？第二 为什么儿童在完成这一学习的过程中 既不需要学习语法课程，也不需要学习逻辑课程？第三 那高深而且越来越高深的语法学和逻辑学与儿童在潜移默化中获得的基本语法和逻辑知识是否存在某些奇妙的区别？

有两位美国人——乔姆斯基先生和山克先生在 HNC 之前进行过这一反思，还有一位欧洲人维特根斯坦先生在电脑尚未出现之前就进行过这一反思，并为此写了《哲学研究》这一不朽巨著。

乔姆斯基先生对上列三个基本问题的前一半作过刺猬式反思，山克先生则对三者都进行过狐狸式反思。乔先生的答案是：儿童的语言天赋是由于人类大脑里存在一个经过百万年进化而形成的东西，他把这个东西叫做普适语法(UG)。随后他用毕生精力研究这个东西，在 20 多年间写下了参考文献中列出的八部专著[1~8]。山克先