《基于认知的汉语计算语言学研究》序言

读了袁毓林教授新著的文集《基于认知的汉语计算语言学研究》，使我联想到美国著名人工智能专家 T. Winograd 在 1983 年写的专著《作为认知过程的语言》（Language as a Cognitive Process）。这两本书都试图从认知的角度来研究计算语言学的问题。可惜 Winograd 的专著只写了“句法”（Syntax）部分，没有再继续往下写。几年以前，我在国外曾经遇见 Winograd，问他为什么不继续写“语义学”（Semantics）部分，他回答说，语义学太复杂，不打算继续写下去了，这样，《作为认知过程的语言》这本专著可以说只是写了一半，就半途而废了。从 Winograd 的学识和才气来说，他是完全可以继续写下去的，可是他没有继续写，我感到非常之可惜。毓林的这本文集，着重从认知的角度探讨论元结构和语义标注，基本上都是语义的问题，恰好弥补了 Winograd 专著的不足，令我感到兴奋。

T. Winograd 在他的专著中说，为了从认知的角度来研究语言，应该解决如下两个问题：

第一， 一个人要说话和理解语言，必须具有哪些知识？

第二， 为了在语言交际中使用这些知识，人们的心智（mind）是怎样组织这些知识的？

根据研究计算语言学多年的实践经验，一个人在说话和理解语言时，不仅需要关于语言的知识，而且还需要各种非语言的知识，例如关于外在世界的知识、日常生活中的常识等，这已经是不容争论的问题，事实上，计算语言学研究者也在努力把这些知识形式化，以便计算机处理。但是，要了解人们的心智究竟怎样组织这些知识，却是一个十分困难的问题。认知语言学试图解决这样的问题。

认知语言学是 20 世纪 80 年代才出现的语言学科，如果把 1989 年在德国 Duisburg 召开的国际第一届认知语言学会议作为认知语言学诞生的标志，那么，这门学科至今才有短短 16 年的历史，可以说是非常年轻的学科。其实，在认知语言学产生之前，很早就有人提出了通过语言来揭示人类心智的问题，已经涉及到认知语言学的问题。1933 年，英国数学家 A. M. Turing 就预见到未来的计算机将会对自然语言研究提出新的问题。他在《机器能思维吗》一文中指出：“我们可以期待，总有一天机器会同人在一切的智能领域里竞争起来。但是，以哪一点作为竞争的出发点呢？这是一个很难决定的问题。许多人以为可以把下棋之类的极为抽象的活动作为最好的出发点，不过，我更倾向于支持另一种主张，这种主张认为，最好的出发点是制造出一种具有智能的、可用钱买到的机器，然后，教这种机器理解英语并且说英语。这个过程可以仿效小孩子说话的那种办法来进行。” Turing 提出，检验计算机智能高低的最好办法是让计算机来讲英语和理解英语，他天才地预见到计算机和自然语言将会结下不解之缘。我认为，Turing 这种预见的实质，就是提出了**“语言是认知的窗口”**这个重要命题。这个命题是认知语言学的基础。所以，从认知的角度来研究计算语言学，进行“基于认知的汉语计算语言学研究”，是非常必要的。

毓林在这本文集中，从认知的角度研究了汉语的论元结构和描述框架，并进行了真实文本语义标注的实践，使我们对于汉语的论元结构有了更加深刻的认识。

在20世纪70年代末和80年代初，我在法国格勒诺布尔理科医科大学研制汉-法/英/日/俄/德多语言机器翻译系统FAJRA时，就根据Tesnière的依存语法（grammaire de dépendence），对汉语动词、形容词和部分名词的论元结构进行了初步的探索，当时我提出的论元有：施事、受事、与事、关涉、时刻、时段、时间起点、时间终点、空间点、空间段、空间起点、空间终点、初态、末态、原因、结果、工具、方式、目的、条件、作用、内容、范围、论题、修饰、比较、伴随、判断、陈述、附加等，共30个，其中，施事、受事、与事3个论元是“行

动元”（actants），其他27个论元是“状态元”（circonstants）。我根据机器词典中存储的单词的语法和语义的静态信息以及在句法分析中运算得出的句法功能的动态信息，使用计算机求解了这些论元信息，把汉语自动地翻译成5种外语，顺利了完成了多语言机器翻译实验。可是，我在20多年前对于汉语论元结构的研究，是从依存语法和工程应用的角度出发的，根本没有考虑到这些论元的认知基础。

现在，毓林从认知的角度，根据计算机处理汉语的实际需要，详细地研究了汉语动词论元结构的论元属性、论旨属性、语法特征、语义特征、配位方式，把汉语动词的论元分为施事、感事、致事、主事、受事、与事、结果、对象、系事、工具、材料、方式、场所、源点、终点、范围、命题，共17个。并且使用自立性、使动性、感知性、述谓性、变化性、受动性、渐成性、关涉性、类属性等动态语义特点以及句法特点，来区分这些论元，从而明确地界定了这些论元。毓林的研究，在更深的层次上揭示了汉语论元结构的特性和判断方法，在逻辑上很有魅力，使我们得到一种逻辑上的美感。但是，他提出的这17个论元中，没有表示时间、原因、目的、论题的论元，而这些论元，在真实的文本中是经常出现的；而且毓林提出的命题这个论元，实际上就是句子，显然是不必要的。

也许毓林察觉到了他的这个论元系统的不足，后来他在语料库语义标注的实践中，把他的这17个论元进一步做了扩充。增加了经事、原因、目的、时间、路径、话题、说明等论元，删除了原来的命题论元，共23个，形成了他的“论旨角色标记集”。这个标记集基本上覆盖了我原来的30个论元的标记集，而且更加精炼，每一个论元的区别特征也更加清楚了，我赞同并且非常欣赏毓林的这个标记集。

毓林把他的研究成果应用于新闻语体真实文本的语义标注和信息自动抽取，效果良好，证明了论元结构知识的广泛适用性。他的成功说明了认知语言学对于计算语言学的理论和实践确实是很有吸引力的。计算语言学应该吸取认知语言学的成果，从而促进自身的发展。

认知科学的基础是“物理符号系统假设”。这种假设认为，智能的基础是符号操作，一切认知系统本质上都是符号加工系统，而符号操作就是计算，认知就是计算。

早在80年代初期，著名语言学家J. A. Fodor在《表达》（Representations）一书（MIT Press, 1980）中就说过：“只要我们认为心理过程是计算过程（因此是由表达式定义的形式操作），那么，除了将心智看作别的之外，还自然会把它看作一种计算机。也就是说，我们会认为，假设的计算过程包含哪些符号操作，心智也就进行哪些符号操作。因此，我们可以大致上认为，心理操作跟图灵机的操作十分类似。”Fodor在这里所说的“符号操作”，实际上也就是“规则”，所以，这种说法代表了计算语言学中的基于规则的理性主义观点。这种理性主义的观点，完全被后来兴起的认知语言学继承并进一步发展了。

而在认知语言学产生之前，在计算语言学中的这种基于符号操作规则的理性主义的观点早就受到了学者们的批评。1980年，J. R. Searle在他的论文《心智、大脑和程序》（Minds, Brains and Programmes）（1980，载《行为科学与脑科学》[Behavioral and Brain Sciences], Vol.3）中，提出了所谓“中文屋子”的质疑。他提出，假设有一个懂得英文但是不懂中文的人被关在一个屋子中，在他面前是一组用英文写的指令，说明英文符号和中文符号之间的对应和操作关系的种种规则。这个人要回答用中文书写的几个问题，为此，他首先要根据指令规则来操作问题中出现的中文符号，理解问题的含义，然后再使用指令规则把他的答案用中文一个一个地写出来。这显然是非常困难的而且几乎是不能实现的事情。Searle的批评是非常尖锐的，这样的批评使计算语言学中基于符号操作规则的理性主义的观点受到了普遍的怀疑。

这种理性主义方法的另一个弱点是在实践方面的。计算语言学中的理性主义者往往把自己的目的局限于某个十分狭窄的专业领域之中，他们采用的主流技术是基于规则的句法-语义分析技术，尽管这些应用系统在某些受限的“子语言”中也曾经获得一定程度的成功，但

是，要想进一步扩大这些系统的覆盖面，用它们来处理大规模的真实文本，仍然有很大的困难。因为从自然语言系统所需要装备的语言知识来看，其数量之浩大和颗粒度之精细，都是以往的任何系统所远远不及的。而且，随着系统拥有的知识在数量上和程度上发生的巨大变化，系统在如何获取、表示和管理知识等基本问题上，不得不另辟蹊径。这样，基于统计的经验主义方法就越来越受到计算语言学研究者的欢迎。

毓林的这本文集，尽管其主要内容是讲基于认知的汉语计算语言学研究，但是，他也注意到了计算语言学中基于统计的经验主义方法，他直率地指出了基于统计的语言处理模型的“有用性”和“局限性”，并且认为，“语言信息处理面临的对象既然有如此顽劣的既抗拒规则模型、又抗拒统计模型的属性，那么一种可能的技术途径只能是把规则的方法和统计的方法结合起来”。很多认知语言学家都推崇认知理论而排斥统计方法，而毓林独具慧眼，他重视认知而不排斥统计，主张规则方法和统计方法的结合，这是难能可贵的。

毓林在他的文集中，非常推崇“计算语言学是用计算机和为计算机研究语言的学科”这个关于计算语言学的定义。并且说，这个定义是国际计算语言学界对计算语言学的定义逐步形成的“共识”。这种说法未免有些偏颇。

我认为，科学的定义应该揭示计算语言学这个学科的本质属性，而毓林所推崇的这个定义带有明显的实用色彩，没有反映出计算语言学与计算机科学在理论上的联系，因而也就难以反映这个学科的本质属性。如果一个人在研究语言时，仅仅使用计算机来统计某些语言单位的出现次数，显然还谈不上他是在研究计算语言学，尽管他用计算机研究了语言；同样地，如果一个人仅仅为了在计算机上输入汉字而研究汉字编码，显然也谈不上他是在研究计算语言学，尽管他是在为计算机研究语言。计算语言学是一个独立的学科，它不仅有着严格而系统的理论，而且还有着完善而成熟的方法，计算语言学的这些理论和方法，正如物理学、数学和化学的理论和方法一样，绝不是不学而能的，而是要经过刻苦的学习和反复的实践才能掌握的。如果一个语言学家只是使用计算机来研究语言而不懂计算语言学的基本理论和方法，他只是一个使用计算机的语言学家，还谈不上是一个计算语言学家；如果一个计算机专家为了在计算机上输入汉字来研究汉字编码而不懂得计算语言学的基本理论和方法，他也只是一个为计算机而研究语言的计算机专家，还谈不上是一个计算语言学家。

毓林说他推崇的这个定义已经逐渐成为国际计算语言学界的“共识”，可能与事实不符。我查阅了很多英文文献，并没有发现这个定义，我还查阅了法文、德文、俄文、日文的文献，也没有发现这个定义。可见，这个定义远远还没有成为国际计算语言学的普遍共识。

如果我们把1954年第一次机器翻译实验的成功算做计算语言学的开始，那么，计算语言学这个学科已经有50多年的历史了，在计算语言学创始前后那个充满了理性的年代，计算机科学的先行者Turing和Shannon就非常重视计算机科学的理论和自然语言的联系。Turing提出了著名的Turing实验，认为检验计算机智能高低的最好办法是让计算机来讲英语和理解英语。Shannon在他的《通信的数学理论》（mathematical theory of communication）中，用马尔可夫过程的理论来分析英语，建立了信息论的基础。他们独树一帜的研究都与自然语言有着千丝万缕的联系，他们的远见卓识都为计算语言学播下了科学的种子。50多年来，他们播下的种子早已破土而出，由纤细柔弱的嫩芽长成了枝叶茂密的大树，成为了一门独立的学科。所以，在给计算语言学这个学科下定义时，我们切不可忽视它与计算机科学在理论上的深刻联系，只有这样，才有可能揭示出这个学科的本质属性。

《计算机进展》（Advanced in Computer）是国际计算机科学的权威出版物，这个出版物登载的文章，都是引导计算机科学学术潮流的高质量论文，从中我们可以窥见国际计算机科学的发展方向。

美国计算机科学家 Bill Manaris 在 1999 年出版的《计算机进展》第 47 卷的《从人-机交互的角度看自然语言处理》一文中曾经给“自然语言处理”提出了如下的定义：

“自然语言处理可以定义为研究在人与人交际中以及在人与计算机交际中的语言问题的一门学科。自然语言处理要研制表示语言能力和语言应用的模型，建立计算框架来实现这样的语言模型，提出相应的方法来不断地完善这样的语言模型，根据这样的语言模型设计各种实用系统，并探讨这些实用系统的评测技术。” 这个定义的英文如下：“NLP could be defined as the discipline that studies the linguistic aspects of human-human and human-machine communication, develops models of linguistic competence and performance, employs computational frameworks to implement process incorporating such models, identifies methodologies for iterative refinement of such processes/models, and investigates techniques for evaluating the result systems.”(Bill Manaris: <Natural language processing: A human-computer interaction perspective>, Advances in Computers, Volume 47，1999)

Bill Manaris 关于自然语言处理的这个定义，比较全面地表达了计算机对自然语言的研究和处理的主要内容，说明了自然语言处理不仅要研究表示语言能力（linguistic competence）的模型，而且还要研究表示语言应用（linguistic performance）的模型，涉及到了自然语言处理在理论上的本质问题，因此，这个定义在《计算机进展》上发表以后，逐渐得到国际自然语言处理界的共识。这个定义是针对“自然语言处理”而提出的，而“自然语言处理”与“计算语言学”是如此之接近，在这里，我愿意推荐这个定义给毓林，作为他给计算语言学这个学科下定义的参考。

计算语言学的研究范围涉及到众多的部门，如语音的自动识别与合成、机器翻译、自然语言理解、人机对话、信息检索、文本分类、自动文摘、机器词典、语料加工、算法研究、语言形式模型研究，等等。我们认为，这些部门可以归纳为如下四个大的方向：

- 语言工程方向：把自然语言处理作为面向实践的、工程化的语言软件开发来研究。这一方向的研究一般称为“人类语言技术（Human Language Technique, 简称 HLT）”，或者称为“语言工程”（Language Engineering）。
- 数据处理方向：把自然语言处理作为开发语言研究相关程序以及语言数据处理的学科来研究。这一方向的研究早期的研究有术语数据库的建设、各种机器可读的电子词典的开发，近年来随着大规模语料库的出现，这个方向的研究显得更加重要。
- 人工智能和认知科学方向：把自然语言处理作为在计算机上实现自然语言能力的学科来研究，探索自然语言理解的智能机制和认知机制。这一方向的研究与人工智能以及认知科学关系密切。
- 语言学方向：把自然语言处理作为语言学的分支来研究，它只研究语言及语言处理与计算相关的方面，而不管其在计算机上的具体实现。这个研究方向的最重要的研究领域是语法形式化理论和自然语言处理的数学理论。

我国的计算语言学研究在语言工程方向和数据处理方向已经投入了很多的资金和人力，大多数的计算语言学工作者都在探索这两个方向的问题，硕果累累。但是，对于人工智能和认知科学方向以及语言学方向，投入就比较少，研究的人也不多，显得比较薄弱。毓林的这本文集就是专门探讨这两个方向的各种理论和实践问题的，而且已经取得了煌煌的成绩，令我感到兴奋。我希望有更多的学者能够重视这两个方向的研究，弥补我国计算语言学研究的这些薄弱环节。

计算语言学是语言学、计算机科学和数学的交叉学科。每一个从事计算语言学研究的人，都面临着知识更新的问题。毓林是中文系出身的一个文科学者，为了研究计算语言学，他进行了知识更新的再学习，从他的文集中可以看出，他不但对于计算机科学和数学不是似懂非懂的外行，而且他还熟悉计算语言学本身独有的基本理论和方法，是计算语言学的精研通达的内行，他使用这些理论和方法，把计算机科学和数学的知识与语言研究有机地、巧妙地融为一体。毓林的研究，把文科与理科结合起来，把汉语与外语结合起来，把理论和实践结合

起来，我相信，今后毓林在计算语言学的研究中，一定会做出更加出色的成绩。

冯志伟
于北京后拐棒胡同寓所
2007年11月10日