基于认知的
汉语 HANYU
计算语言学研究

袁毓林 著

北京大学出版社
PEKING UNIVERSITY PRESS

基于认知的
汉语计算语言学研究

袁毓林 著

北京大学出版社
PEKING UNIVERSITY PRESS

图书在版编目(CIP)数据

基于认知的汉语计算语言学研究/袁毓林著. —北京：北京大学出版社，2008.7

ISBN 978-7-301-14052-9

Ⅰ. 基… Ⅱ. 袁… Ⅲ. 汉语-机器翻译-研究 Ⅳ. H085

中国版本图书馆 CIP 数据核字(2008)第 103402 号

书　　名：基于认知的汉语计算语言学研究
著作责任者：袁毓林　著
责 任 编 辑：杜若明
标 准 书 号：ISBN 978-7-301-14052-9/H·2028
出 版 发 行：北京大学出版社
地　　址：北京市海淀区成府路 205 号　100871
网　　址：http://cbs.pku.edu.cn
电 子 信 箱：zpup@pup.pku.edu.cn
电　　话：邮购部 62752015　发行部 62750672　编辑部 62752028
　　出版部 62754962
印　刷　者：北京大学印刷厂
经　销　者：新华书店
　　890 毫米×1240 毫米　A5　15.125 印张　335 千字
　　2008 年 7 月第 1 版　2008 年 7 月第 1 次印刷
定　　价：30.00 元

陆　序

在学术研究领域，袁毓林可以说是一位勤奋的耕耘者。他的论文集《汉语语法研究的认知视野》(商务印书馆)于2004年出版，现在又推出了新的论文集《基于认知的汉语计算语言学研究》。我大略地翻阅了一下全书各篇的内容，论文集的书名"基于认知的汉语计算语言学研究"，点明了该书的基本内容——从认知的视角来研究计算语言学，特别是中文信息处理的问题。正文具体分四部分内容：

第一部分内容，作者取名为"计算理论和语言研究"，包括四篇文章：《计算语言学的理论方法和研究取向》、《基于统计的语言处理模型的有用性和局限性》、《认知科学和汉语计算语言学》和《面向当代科技的语言研究的理论和方法》。计算语言学的研究，大致可以分为两个层面，一个是理论模型的研究，一个是工程研究(或说具体的技术方法研究)。据我所知，袁毓林主要从事理论模型的研究，所以这一部分内容作者主要从宏观的角度介绍说明了计算语言学的理论方法和研究取向；评述了在自然语言处理中已运用过的基于规则和基于统计的两种处理模型，指出处理语言这种复杂的系统"必须走规则和统计相结合的道路"；从认知科学的视角作者把自己认为有价值的并且是可行的计算语言学研究模式介绍给读者，并结合作者自己的研究实践讨论说明了基于认知并面向计算的汉语语法研究的路线；展示了认知语言学和计算语言学相互结合的可能性。这部分内容对有兴趣了解或从事计算语言学研究的人来说，是值得一读的，是很有启迪意义的。

第二部分内容，作者取名为"论元结构和描述框架"，也包括四篇文章：《论元角色的层级关系和语义特征》、《一套汉语动词的论元角色的语法指标》、《汉语谓词的论元结构的描述框架》和《论元结构和句式结构互动的动因、机制和条件——表达精细化对动词配价和句式构造的影响》。袁毓林是我国最早研究配价问题的学者之一，特别是他第一个发表了有关汉语名词配价的研究成果，该成果被广为引用。

以乔姆斯基为代表的生成语法学派所提出的动词论元结构理论与法国依存语法学派特斯尼耶尔提出的动词配价结构理论有相同的一面，当然出发点不同，思考的角度不同，对语言事实解释的广度与深度也不同。十多年来袁毓林一直致力于动词论元结构的研究，在这方面他发表了一系列有分量的文章。我所主持的两个重大科研项目“面向中文信息处理的现代汉语动词论旨结构系统和汉语词语语义分类层级系统研究”（国务院973国家重点基础研究发展规划项目“图像、语音、自然语言理解与知识挖掘”子课题）和“汉语语义知识的形式化模型及语义分类系统研究”（教育部重点研究基地项目），袁毓林都参加了，其中的“汉语动词的题元系统及其语法指标”（包括“题元的层级体系”，“各别题元的定义、示例和句法语义特点”，“不同题元之间的配合关系”，以及“各别题元的语法指标”）就是由袁毓林执笔起草的。因此本书这一部分内容可以说是他对自己在配价问题和动词论元结构研究方面成果的汇集。在这部分内容中，他不仅建立并提出了汉语动词论元角色的层级体系，定义了各个语义角色，并细致描述了各个语义角色在述谓结构中所表现出来的动态性语义特征，同时通过十个各具特色、有代表性的实例（谓词“切、包$_1$、包$_2$、调查、帮忙$_1$、帮忙$_2$、飞$_1$、飞$_2$、吃、专政”）给出了谓词及其论元的句法配置方式，提出了汉语谓词论元结构的描写框架。更值得注意的是，他探讨了谓词论元结构和句式结构（constructions）互动的动因、机制和条件，对汉语谓词所谓“变价”和“论元增容”作了进一步的解释。

第三部分内容，作者取名为“信息抽取和语义标注”，包括五篇文章：《信息抽取的语义知识资源研究》、《用动词的论元结构跟事件模板相匹配——一种由动词驱动的信息抽取方法》、《用逻辑和篇章知识来约束模板匹配——逻辑结构和篇章结构知识在信息抽取中的运用》、《基于论元结构的语义标注的体系和规范》以及《新闻语体真实文本的语义标注的实践》。这部分内容作者主要提出并举例说明了要使计算机有效地自动从真实文本抽取信息，至少要有三种层面的语义知识：话语篇章知识、谓词论元结构知识和句子的逻辑结构知识；为对真实文本进行语义分析和标注，作者细致分析设计了篇章、谓词论元结构、句子逻辑结构这三种层面各自的语义关系，并为这三种层面各自的语义关系设计

并提出了一套可扩充的标记集;作者还以自己设计的这套标记对新闻报道中关于职务调动的真实文本进行了语义关系标注实践。作者标注得相当认真。通过这样的标注实践又有所发现——真实文本中代词或指示词的先行成分(一般称为先行语)常常是隐含的;段落之间的衔接,其形式手段相当缺乏。这就促使大家去进一步思考、探索怎么为计算机自动处理真实文本解决这方面的难题。

第四部分内容,作者取名为"专题研究和个案分析",也包括五篇文章:《容器隐喻和套件隐喻及相关的语法现象——词语同现限制的认知解释和计算分析》、《关于分词规范和规范词表的若干意见》、《中文信息处理中的语言难题问答》、《缓冲式移动通信及其发展方向——一个语言学家的设计思想》和《走向多层面互动的汉语研究》。这部分值得细细阅读的是《容器隐喻和套件隐喻及相关的语法现象——词语同现限制的认知解释和计算分析》和《走向多层面互动的汉语研究》这两篇文章。前一篇文章主要通过对"满"、"全",特别是"满+NP"、"全+NP"在意义、用法上的不平行性的解释,说明语言中的许多现象只有从认知的隐喻的视角来加以解释——用容器隐喻来解释"满"背后的概念结构以及由"满"构成的"满+NP"的使用特点,用套件隐喻来解释"全"背后的概念结构以及由"全"构成的"全+NP"的使用特点,这样才能说得清楚,说得圆满,说得充分,才能有解释力;通过对"满"和"全"又具有一定的可替换性的解释,说明隐喻分析有必要提升到更为抽象的意象图式水平,这样才更有解释力,才能最终解释说明既然"满"、"全"背后的概念结构是属于不同的隐喻范畴,为什么有时又具有可替换性,即才能说明为什么容器隐喻和套件隐喻在语言的实际使用中会出现二者中和化的现象;更积极的意义,还在于正如作者在文章中所指出的,有助于语言的认知解释有可能实现形式化和可计算,从而有可能实现认知和计算的统一("有可能"三个字不是作者说的,是我加的)。后一篇文章是作者为徐杰所编的《词汇语法语音的相互关联——第二届肯特岗国际汉语语言学圆桌会议(2002.11.26—30.)论文集》所写的代前言。文章扼要回顾了20世纪汉语研究的历史,对今后的汉语研究发表了很有见地的看法。作者强调指出,汉语研究必须树立"互动观念",走多层面互动研究之路,而这方面正

是目前汉语学界所缺乏的。文章特别谈到了一段时间来成为人们热门话题的所谓“语法研究三个平面”的问题，作者强调指出，“我们不仅应该分清语法的三个不同的平面，而且应该观察这三个不同的平面之间的互动关系”，并应“引入语言类型学的视野”，“引进语法化这种动态性的概念，来审视语法、语义和语用这三个平面之间的互动关系”，“从而打破共时研究和历时研究之间的藩篱，把语言的共时研究和历时研究沟通起来”，以“推动语言研究走向更为全面、综合和多层面互动的道路”。文章以学界已有的研究成果和作者本人的研究成果具体说明了语法和语音之间、词库结构和句法操作之间的互动关系，以及这种互动所应有的限度。这是很有见地的看法，应引起大家重视。

我虽然只粗粗阅读了一遍，觉得收获良多，推荐大家一读。借此机会我也想发表两点看法，同时也想提出一些意见。

第一点，当今语言研究的走向之一，确实如本书作者所说，要走多层面互动的研究之路。不过这只是“之一”，还应有另一个“之一”，那就是“特征研究”，这也必须重视。从上个世纪七十年代以来，就语言研究说，一个重要的趋向是逐步重视特征的研究和描写。在语言的理论研究和应用研究上都是这样。

先说语言的理论研究，大家知道，在语言研究领域，最早讲特征的是音位学，接着是语义学；语法学里讲语义特征那是七十年代以后的事了。当时把“语义特征”这个概念术语借用到语法学中，为的是做两件事：一件事，用以解释造成同形多义句法格式的原因；另一件事，用以说明在某个句法格式中，为什么同是动词，或同是形容词，或同是名词，而有的能进入，有的不能进入。发展到乔姆斯基的生成语法理论，特征又赋予它新的含义。我们知道，乔姆斯基因为认为结构主义对语言的描写所概括的规则太复杂了，所以他要提出生成语法的观点，以简化语法规则。简约，一直是生成语法学的一个很重要的原则。从 1957 年的由核心句到非核心句的转换，到 1964 年的从深层结构到表层结构的转换，到上个世纪 80 年代初的 GB 理论——只剩下“α 移位”规则，其他都成了原则，再到最简方案及其近几年的论述——众多的原则和移位规则基本都不要了，D-结构，S-结构都没有

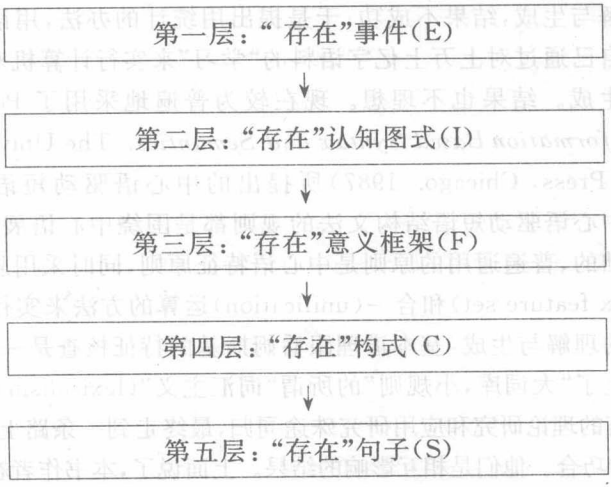
了，似只保留了“原则和参数”理论和“X-bar”结构模式，进一步强调经济原则；而提出了中心词（head）理论和特征核查（feature checking）理论，以及轻动词理论和 VP 空壳理论，注入了新的研究课题——接口（interfaces）的研究。基本的句法运作是从基础部分（即词库）取出带有各种各样的有关语义、句法特征的词项，进行来回合并（Merge），如能通过特征核查（指中心语跟标示语，中心语跟补足语，在特征上吻合），由此生成的词项组合再去跟音韵接口，跟逻辑语义接口，从而最终生成我们所听到或看到的句子。总之，词语的特征的分析和描写放到了非常重要的位置，走上了“大词库，小规则”之路。这里要附带说明的，最近乔姆斯基在 Linguistic Inquiry 杂志 2005 年第 1 期上发表的文章（*Three Factors in Language Design.*）中似乎提到要取消“特征核查”，但他同时在文章中认为，从词库选出词语项，构成词语序列，形成语段后，要通过所谓“探针（probe）”与目标进行相互核查，如果没有发现不可诠释的特征，就转移给语音和语义两个界面接口，由此获得语音和语义相结合的语言形式。这实质上还是需要进行特征核查这一步骤。而所谓“要取消特征核查”，我体会是指在操作手续上要进一步简化。

现在再看自然语言处理与理解这方面的语言应用研究。大家都知道，自然语言处理与理解最早使用规则的方法来实行计算机对句子的理解与生成，结果不成功；于是提出用统计的办法，用语料库让计算机自己通过对上万上亿字语料的“学习”来实行计算机对句子的理解与生成。结果也不理想。现在较为普遍地采用了 Pollard & Sag（*Information Based Syntax and Semantics*. The University of Chicago Press, Chicago. 1987）所提出的中心语驱动短语结构文法——中心语驱动短语结构文法的规则都是围绕中心语展开的，而其最基础的、普遍通用的原则是中心语特征原则，同时采用复杂特征（complex feature set）和合一（unification）运算的方法来实行计算机对句子的理解与生成，基本道理跟乔姆斯基的特征核查是一样的，最终也走上了“大词库，小规则”的所谓“词汇主义”（lexicalism）之路。

语言的理论研究和应用研究殊途同归，最终走到一条路上去，这绝非偶然的巧合。他们是相互影响的结果。上面说了，本书作者在第一部

分内容里，主要从宏观的角度介绍说明了计算语言学的理论方法和研究取向，但作者未注意到“重视特征研究”这一取向，这可能跟作者对于国外近十多年来有关计算语言学方面的文献资料还了解得不全面有关。

第二点，上面说了，作者在第二部分内容里，探讨了谓词论元结构和句式结构(constructions，有人说成“构式”)互动的动因、机制和条件，对汉语谓词所谓“变价”和“论元增容”作了进一步的解释。论述很有新意。但我觉得如果作者能进一步深入思考这样一个问题就好了：“人对客观事物的感知所得最后是怎样用言辞表达出来的?”最近看到王黎在《关于构式和词语的多功能性》(《外国语》2005 年第 4 期)一文中明确提出了这个问题。王黎认为，从人对客观事物的感知所得到最后用言辞表达出来，中间一共可分为五个层面：第一层：是客观世界中所存在的诸多基本的、典型的事件(包括景象等)，诸如“存在事件”、“分配事件”、“事物特征”等(用 E 表示)；第二层：这个事件，如存在事件，被人观察到以后，就相应地在人的认知域里，形成了存在意象(image，用 I 表示)；第三层：这存在意象又激活了人脑里的深层存在意义框架(用 F 表示)；第四层：当这个深层存在意义框架被位于表层的语言表现出来时，就有了存在构式(用 C 表示)；第五层：那存在构式里填上一定的具体的词项，就形成我们在实际语言交际中所听到看到的存在句(用 S 表示)。这五个层面的关系，王黎图示如下：



这当然是一种假设，不能看作是结论，但可引起人们去进一步思考。同时，可以用来更好的解释说明句式的配价问题，也可以对汉语谓词所谓“变价”和“论元增容”作出更好的解释。

袁毓林论文集编就后，要我写序，这已是数个月之前的事了。写序，还是应尽可能做到有的放矢，实事求是。所以我在动笔写之前，一定要先看书稿，了解全书内容。我又比较忙，这样就拖到现在才将序文草就。所言不一定到位，请作者和广大读者批评指正。是为序。

陆俭明

2007-10-08 于北京蓝旗营寓所

冯　序

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 19 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"的这个重要命题。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 refine-

ment 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