汉语工程词论

卞成林 著

han yu
gong cheng
ci lun

山东大学出版社

汉语工程词论

卞成林　著

山东大学出版社

图书在版编目（CIP）数据

汉语工程词论/卞成林著 .—济南：山东大学出版社，1999.12
ISBN 7-5607-1981-3

Ⅰ.汉…
Ⅱ.卞…
Ⅲ.汉语-词汇学-研究
Ⅳ.H13

中国版本图书馆 CIP 数据核字（2000）第 13598 号

山东大学出版社出版发行
（山东省济南市山大南路 27 号　邮政编码：250100）
山东省新华书店经销
济南市市中印刷五厂印刷
850×1168 毫米　1/32　9 印张　234 千字
2000 年 2 月第 1 版　2000 年 2 月第 1 次印刷
印数：1－1200 册
定价：20.00 元

序

随着社会的发展和科学技术的进步，在语言学研究中，计算语言学的建立和发展早已被人们公认为是现代科学发展的必然结果。目前计算机在社会各行各业中的普遍应用，更为计算语言学的研究不断地提出了许多崭新的课题。

事实早已说明，计算机的应用绝对离不开自然语言的参与，一切信息的计算机处理过程都与自然语言有着密不可分的关系。因此，对汉语信息处理来说，面临的将是要逐步解决字的处理、词的处理、语句的处理和篇章的处理等问题，而在这些问题当中，很明显，在解决了字处理的基础上，词的处理就成了大家关注的核心和关键。只有解决了信息处理中的词的问题，才能更好地进一步向语句处理等问题上迈进。

面对社会的需求，汉语词汇研究者决不能袖手旁观，卞成林老师对汉语工程词的研究就是在这种形势下开始进行的。卞成林老师不仅在汉语词汇研究方面具有相当深的学术功底，而且对计算机及其应用方面也掌握了相当全面的学术理论和实践经验，因此他在工程词的研究上，自身就具备了相当坚实的基础和条件。现在他的研究成果《汉语工程词论》一书就要出版了，该书的特点，不仅内容丰富，而且著述非常扎实、严谨，全书以理论和实际紧密结合为原则，对每一个问题的阐述，每一个理论观点的论证，都能做到有根有据，这就为该书理论的建立和具体应用打下了极为有利的基础。此外，还有以下三个方面也应在此提及。

第一，该书第一次明确地界定了“工程词”的概念，指出工程词是在词汇词、语法词的基础上出现的信息工程中的词，它既包括了词汇词和语法词，也包括着一些非词汇词和语法词的成分，该书从理论上阐明了确定为工程词的原则和标准，以及工程词所包括的内容和范围。更可贵的是，该书还通过大量的资料数据，提出了形成汉语工程词语的生成规律，探讨和基本解决了机器理解自然语言的未登录词，以及机器词典收词范围和语义网络等问题。这一系列的理论观点新颖而且独到，在学术研究和实际应用方面都具有重要的意义。第二，该书的写作采用了借助于计算机的软硬件技术的方法，在大量占有语料和穷尽式分析具体词汇现象的基础上，深入探讨了工程词的词素构词能力，词语结构方式的分布，汉语通用词的词长及其分布等等，从而总结出了汉语工程词的特点。这样得出的结论，依据扎实可靠，极具说服力。第三，该书在阐述自己理论观点的同时，也为汉语信息处理中词处理问题的研究提供了大量翔实可靠的数据资料，书中大量的图表就足以说明了这一点。这是具有非常实际应用价值的事情。

综上所述可以看出，目前在我国信息处理领域的有关词处理的研究中，如此有理论、有实际、全面系统进行分析研究的专著，《汉语工程词论》还当属第一本。因此我相信，该书的问世一定会为词处理问题，以至计算语言学的发展等起到积极的推动作用。

当然，对任何研究来说，都会是只有开始而没有终止的，卞成林对汉语工程词的研究也是如此。希望卞成林老师能够在此基础上继续奋力前进，继续将该课题和与该课题有关的一切问题深入细微地探讨下去。

葛本仪写于山东大学

1999.12.18

目　录

第一章　基于信息工程的工程词研究

本论题所涉及的工程词是指基于信息工程的词，它以汉语信息处理的词汇问题为研究对象，并对汉语信息处理中的语言问题提出词汇学对策。

第一节　信息工程中的语言问题

一、术语的由来

“工程词”这一术语有三个由来：

1．吕叔湘先生关于词汇的词和语法的词的区分

吕先生在讨论词和短语的区分问题时，指出：“一般人心目中的词是不太长不太复杂的语音语义单位，大致跟词典里的词目差不多，这可以叫做‘词汇的词’”；而有些组合单位可以连成很长的一串，像“袖珍英汉词典/大型彩色纪录片/同步稳相回旋加速器/多弹头分导重入大气层运载工具”等，这一类单位称作“语法的词”。因为“词这个东西，不光是语法单位，也是词汇单位。二者有时候一致，有时候不一致，因为所用的标准不同”[①]。

这里，吕先生根据语法分析的需要，用不同的标准，对汉语

的词汇单位作了新的区分，为我们在词汇单位中另立一新的分析单位打开了思路。

2.《信息处理用现代汉语分词规范》提出的“分词单位”

80年代末开始，在众多研究者的理论和实践的基础上，产生了《信息处理用现代汉语分词规范（中华人民共和国国家标准GB13715)》（以下称作“规范”)。“规范”明确提出信息处理不能以词典中所收的词为主要依据，应视具体用途的不同，会有所变化。“规范”用了一个新的术语——“分词单位”，来取代词汇学中的“词”（即“词典词”)，“分词单位”指“汉语信息处理使用的、具有确定的语义或语法功能的基本单位。它包括本规则限定的词和词组”[②]。

这里，“规范”已开始在传统的词汇单位之外加进了信息处理中的一些需要特殊对待的语言单位。

3.词的三个划分

语言学界在多年关于词的讨论中，已明确了“词”在不同的用途中应有不同的“定义”。即“语法词”、“词汇词”和“连写词”的划分。语法词是从语法角度确定的词，以不可扩展性为其主要特征。词汇词是从词汇角度确定的基本单位，以“专指义”或“专门意义”为其主要特征。不管语法上是词还是短语，只要其实际意义不能简单地从字面或者说从构成词的词素义获得，就属于词汇或语汇的范围。词典收词也往往包括这样的单位——成语、惯用语、熟语等。连写词则是从拼音文字角度确定的基本书写单位。凡是连写在一起的就是一个连写词。连写词的主要特点是语音上结合较紧密，经常能形成一个节拍群，《汉语拼音正词法基本规则》就是对连写词的一些具体说明。

汉语信息处理中，从使用的角度，将汉语的词语分成“语法词”、“逻辑词”和“工程词”三大类型。“语法词”就是“完全符合汉语语法规则的严格定义的词”；“逻辑词”是“以概念意义

为单位的词语”；“工程词”专指“在信息处理工程领域由于特殊的需要而规定的一种处理单位”[③]。

1996年4月在烟台召开的第二届全国现代汉语词汇学术研讨会上，山东大学葛本仪先生更进一步提出，应该将现代汉语的词分成三个方面，即词汇词、语法词和工程词。词汇词是指语言中静态的、规范的词。语法词则包括：a. 词汇词；b. 在词汇词基础上，附加某些语法成分构成的语言单位，如：“朋友们”、“坐着”、“说了”、“来过”、“屋里”等等。工程词包括词汇词和语法词，同时考虑信息处理过程中的语音、语义、语法、语用特征，将某些非词汇词、非语法词也当作词汇单位来处理。

二、信息工程与信息工程中的语言问题

20世纪是信息的社会，人们对世界把握的质量是以对信息量的掌握为标准的，信息，信息处理，语言信息处理，这是当前高技术领域的研究热点。语言信息的处理量和处理水平已经成为一个国家和民族能否步入信息社会的入场券，是现代文明的重要标志。美国的尼葛洛庞蒂在他的《数字化生存》（[美] 尼葛洛庞蒂著《数字化生存》，海南出版社1996年版）一书中更进一步认为，比特（信息的计量单位）作为信息的DNA，正迅速取代原子（atom）而成为人类社会的基本要素。

人类信息的传播是以语言和文字作为载体的，信息处理首先是语言文字的处理，然后才谈得上语言文字所负载的知识信息的处理。计算机对语言文字的处理，在学术界称为“语言信息处理”，即“用计算机对自然语言（书面和口语两种形式）的音、形、义等信息进行处理。即对字、词、句、篇章的输入、输出、识别、分析、理解、生成等的操作与加工”，“汉语信息处理”则是“用计算机对汉语的音、形、义等信息进行处理”[④]，是语言信息处理的组成部分，有时又称“中文信息处理”。

以计算机作为工具的语言信息处理，涉及到数学、计算机科学、语言学等多种知识，是一个横跨理科、工科、文科等学科的系统工程。因此，“信息处理”经常被代之以“信息工程”这一术语。

计算机处理信息的过程中，自然语言是不可或缺的媒介，人与机器的对话，依然是基于自然语言的对话，这一对话过程可以表述为：

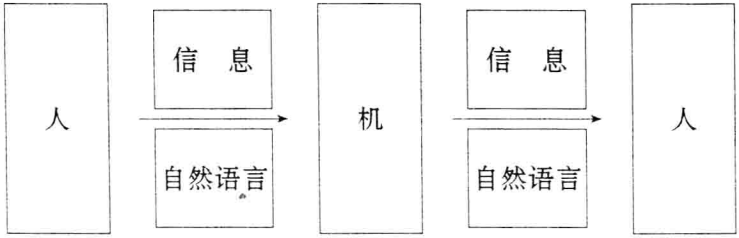

图 1.1 人与机器对话过程图

其中，信息要借助语言输入机器，机器的处理结果又要借助语言输出。虽然电子计算机软件中，早已设计了诸如 BASIC、PASCAL、COBOL、C 等程序设计语言，但这些程序设计语言在两个方面注定了与自然语言的密不可分：第一，任何一个程序设计语言都是在自然语言的基础上由人工设计完成的，因而也被称之为“人工语言”；第二，这些人工语言与自然语言一样，都遵循着形式语言的规律和法则。美国语言学家乔姆斯基的形式语言理论，既适用于自然语言，也适用于程序设计语言，从理论上说明了自然语言和人工语言的密切关系。因此，信息的计算机处理过程与自然语言有着密不可分的关系，自然语言的研究在当代就必须与信息工程结合起来了。

然而，自然语言与信息工程的结合一开始就是非常困难的。困难的原因在于，人是具有高级智能的动物。人的智能体现在人

们能够根据某种特定的环境产生某种行动，或者根据某一特定的前提产生某种结论。人的语言生活属于高级智能的一个重要方面，作为交际工具的人际自然语言，结构复杂多样，语义表达千变万化，而这正是人际语言最“自然”的一面，它所对应的是生动活泼、丰富多彩的人际生活。计算机对人类智能的模拟被称作人工智能，人工智能中，自然语言的信息处理是重要的方面。但是人类的智能活动是基于知识、经验的认知活动；机器的智能处理则是基于分析的模式匹配。自然语言进入计算机，要求结构简单、意义单一、结构与语义整齐对应，显然，这与自然语言本身的特点是矛盾的。

为解决这一矛盾，人们首先想到的是仿照自然语言来设计人工语言，以消除自然语言中随处可见的歧义，使结构与意义一一对应。但是，就计算机的发展趋势来说，智能化的计算机应该会说话、会学习、会思维，信息处理的高级阶段是自然语言的自动处理，特别是大规模真实文本的自动处理，而不是重新设计若干套专供计算机使用的人工语言。为达到这一目标，一方面需要计算机硬件和软件的研究和开发，另一方面，需要语言学理论对计算机所需要的语言知识进行深入的研究。由于自然语言是一个充分复杂的信息系统，而计算机只能处理形式化的知识。让计算机理解自然语言，完全用自然语言自由地进行人机对话在现阶段还是非常困难的事情。较为实际可行的办法是对自然语言加以一定的限制，然后把受限的语言规则教给计算机。智能化的程度越高，受限的语言规则会越复杂，受限语言也越逼近自然语言。

计算机处理自然语言，其过程是：首先，将自然语言形式化，表述为若干套类似数学表达式的规则系统[5]；其次，建立语言单位与规则系统之间的运算关系，使得规则系统是可以推导和运算的；第三，根据自然语言形式化算法，设计计算机程序，通过计算机程序使计算机模拟出自然语言，以实现机器自动翻译、

人机对话、大规模真实文本的自动检索等等。可见，要想让计算机处理自然语言，首要的任务是把语言学知识形式化，从而为语言信息处理提供一套可以运算、推演的自然语言规则系统和一部规范化、标准化的词典。

可以说，语言信息处理技术每前进一步，都需要语言文字学理论的支撑，处处要求语言学提供规律、理论和假设，而且，语言信息处理越是向高层次发展，就越需要语言文字学研究的深入。面对人机和人际两个系统的自然语言研究，对所研究的对象需要重新审视，重新定位。例如，语音识别、语音合成技术，要求语音学研究除了更加细致地描写自然语言的音素、音位，还要同时着力于“言语声”可变性的研究，如各语音单元在连续语流中所发生的音间过渡、协同发音、语音简缩、口语风格等等[⑥]。机器翻译、人机对话工程则要求语言学研究不仅提供一部词典，而且还要解决自然语言的歧义问题、上下文环境的识别与确定问题、代词所指问题、句子中相关词的照应问题、省略成分的复原问题、连词的作用范围问题，等等。

综上所述，信息处理中的诸多语言问题可以归结为一点，即自然语言的模式化。纷繁复杂的自然语言只有模式化了才能进行形式化表示，然后才能通过计算机程序模拟出来。自然语言的模式化包括结构模式化和语义模式化。结构模式指句法结构模式、词语搭配模式、组合限制模式等等；语义模式则指词语意义的特征集（义素）、语义的逻辑演算规则等等。自然语言模式化的目标，要求语言学在传统的研究方法之外，寻求用工程的方法研究自然语言的语音、词汇、语法、语义和语用等问题。本文所关注的则是自然语言特别是汉语信息处理中的词汇问题。

三、汉语信息处理中的词汇问题

自然语言信息处理的内容包括计算机软硬件的开发和探讨自

然语言的受限规则两大范畴。就汉语而言，是开发计算机的汉语言支撑能力和研究受限汉语的规则。

由于计算机的汉语言支撑能力的开发与受限汉语规则的研究都不是一件容易的事，不能一蹴而就，并且两者的研究又互相促进也互相制约，因此，汉语信息处理的研究与发展可能呈现出四个阶段：字处理阶段、词处理阶段、语句处理阶段、篇章处理阶段。这四个阶段主要是就汉语信息处理的发展水平而言的，而不是单纯时间上的划分。虽然各个阶段研究内容的侧重点不同，但在技术和成果方面，一些阶段之间是互相渗透的。比如：词处理阶段为语句处理阶段打下了基础，语句处理阶段还将发展词处理阶段的成果。总的来看，汉语信息处理的四个发展阶段中，词处理是基础，语句处理是中心。

信息处理中以词处理为基础，语句处理为中心，是与汉语认知心理相关的，心理学家用下面的四个例子说明了中文认知中的心理过程[⑦]：

A. 中文的言忍矢口分木斤

B. 研的中阅究文读

C. 的阅读研究中文

D. 通知中文的阅读

从A到D，人们对字符序列的理解越来越轻松，其认知过程可以这样来表述：人首先要对视觉空间的信息进行知觉分析，A例改变了平常用来帮助人们进行知觉分析以辨认个别单字的资料，因而需要较长的时间才能看懂；知觉分析的单位不是孤立的汉字，而是词汇单位，词汇分析是汉语认知过程中最基本的操作，B例很难分析成有意义的词汇单位，所以无法进行进一步的分析理解工作；分析句子中各个词之间的关系，并从中推论出句子的含意也是汉语认知过程中的一个重要步骤，C例中，知觉和词汇信息分析的问题都不存在，但是，这些“的”、“阅读”、“研究”、

“中文”几个词之间的排列不符合中文句子的组织结构规则，所以造成了理解上的困难。造成D例理解上的困难的原因是字符序列中语义关系的障碍，“通知”与“中文阅读”之间不能构成语义上的联系，换成像“研究”、“了解”等词语，这一句的意义才能被接受，可见，“语义分析”是汉语认知过程中非常关键的一步。

据此，我们把阅读认知的心理过程分成四个步骤：字符知觉→词汇分析→关系分析→语义分析，这四个步骤与汉语信息处理的四个发展阶段有着某种程度的对应关系：

字符知觉→词汇分析→关系分析→语义分析

字处理→词处理→语句处理→篇章处理

词汇分析是认知过程的基础，惟其如此，在现阶段的汉语信息处理应用领域（汉字识别，汉语语音识别及合成，全文信息检索及文本自动分类，文本自动校对，音转字等等）中，汉语的词汇平面是主要支撑平台，几乎没有什么应用技术可以游离于这个平面之外而存在。摆在我们面前的，不仅是单纯的学术问题，而是一项颇具规模的语言工程。

与印欧语言相比，汉语书面语的特点是：第一，不实行分词连写；第二，形态变化不丰富。基于印欧语言的信息处理可以直接在“句法－语义”层面上进行，词汇分析主要是词义的分析；而汉语的词汇分析则分成两个步骤：词形辨认（分词）、词义分析，词义分析必须建立在词形的正确辨认上。例如：“波兰工人也罢工了”一句，计算机的切分可以是：

a. 波/兰/工/人/也/罢/工/了

b. 波/兰/工人/也/罢/工/了

c. 波兰/工人/也罢/工了

d. 波兰/工人/也/罢工/了

……

借助词库，计算机一般不会有 a 和 b 两种辨认结果，但即使提供一部相当完备的词库，计算机也很难分辨出 c 和 d 哪一个才是正确的。在信息处理界，汉语的自动分词问题已经成为阻碍汉语信息处理发展的“瓶颈”问题了。

如上所述，汉语自动分词的前提是需要一个基于现代汉语真实文本的词库，与一般的辞书不同，这个词库至少要满足如下条件：

第一，标准化。即符合现代汉语规范化的标准，符合汉语实际的词汇单位。

第二，大容量。受印刷、使用等多方面的限制，一般辞书的收词数量都是很有限的，计算机为建设大容量词库提供了现代化的科技手段。

第三，生成性。词汇系统是随着社会发展而不断产生新词的开放系统，任何一个词库都不可能是穷尽的，因而每一个具体的词库还要提供汉语新词产生的规则系统。

第四，可控性。词汇系统的变动不居注定了词的数量是一个天文数字，而且这个数字每日每时都在改变着，因此，无限膨胀的词库还必需受到控制，否则不仅人脑无法接受，以高速度、大容量见长的现代电脑也是无法承受的。

总之，词库建设的关键是解决现代汉语的“词”的定义问题，有了关于词的明确、统一和科学的标准，我们才能名正言顺地对语言单位实施取舍，实现可控。但是，长期以来，“词是什么”、“什么是词”，一直是语言学家们议而不决或议而难决的东西。随着信息处理技术的发展，我们必须从工程的角度找到一个贯穿始终的、“一揽子”的解决方案，哪怕一下子还远非尽善尽美。

《信息处理用现代汉语分词规范（中华人民共和国国家标准 GB13715)》是汉语词库建设的国家标准。“规范”明确提出信息

处理不能以词典中所收的词为主要依据[8]，因为：

①词典词中只有词，较少通用的语；

②由于历史的原因，词典词偏重人文科学，而信息处理应社会科学和自然科学并重；

③词典词中的词条是历代积累而成的，有相当一部分已经不用了。相反，信息处理应有大量的新词新语；

④词典词中没有表示语法意义的重叠性的词，而信息处理必须收录这些词；

⑤词典词中没有通用的人名、地名，而信息处理必须收录相当的数量；

⑥词典词没有频度，无法按计算机内存的需要依频度分段使用，而信息处理中的词应能做到这一点。

当然，“规范”所规范的不是“词”，仅是供信息处理用的“分词单位”，它在明确了“词典词”和“分词单位”的区别后，并没有提出词的辨认上的最高指导原则。许多问题的处理难免会有前后矛盾或主观限定，具体的使用者在实际操作时不得不寻求其他的理论指导。北京大学计算语言所新近完成的《现代汉语语法信息库》所收词条就是以朱德熙《语法讲义》关于词的定义为基础，并“从计算机处理实际文本的需要出发，从提高计算机处理效率的角度考虑，词典中确实包含了以下 7 类语言成分”：

前接成分（h）：阿，老，超

后接成分（k）：儿，子，性，员，器

语素字　（g）：柿，衣，失，遥，郝

非语素字（x）：鹌，枇，蚣

成语　　（i）：胸有成竹，八拜之交

习用语　（l）：总而言之，由此可见

简称略语（j）：三好，全总

因此，在其报告中“将电子词典中登录的各种语言成分笼统地叫