词汇语义和计算语言学

林杏光 著

语文出版社

题　　词

中文信息处理技术在我国现代化及信息化建设中，越来越起着重要的作用，作为一个高新技术的重点，它已经列入国务院批准的“国家中长期科学技术发展纲要”。十几年来，我国的中文信息处理领域里，在基础理论的研究、新技术的应用、产品的开发以及产业的建立等方面都取得了显著的成绩。

中文信息处理是多学科和跨学科的研究工作，特别需要计算机科学与语言学的密切结合，而且要依靠长期积累的研究成果。过去取得的硕果中凝聚了计算机科学界与语言学界专家们的心血，也证明他们携手合作是大有作为的。

作为语言研究工作者，林杏光同志积极和计算机工作者合作，多年针对中文信息处理的需要进行词汇和语义的研究，取得了许多成果。《词汇语义和计算语言学》这部论著，是他十多年来在现代汉语研究结合中文信息处理研究方面的又一个成果，对促进计算机科学和语言学的进一步沟通和合作很有价值，可喜，可贺，可赞！

陆力为

1998年春节

序

林杏光同志长期从事词汇和语义的研究，最初是结合词典编纂进行的，这个时期的主要著作有先后由商务印书馆出版的《简明汉语义类词典》和《现代汉语实词搭配词典》(主编)以及由中国标准出版社出版的《汉语多用词典》。这几部词典都有特色，可以说都是创新之作，因而受到了学术界的广泛重视和好评，同时引起了从事计算语言学研究的计算机专家的注意。在这以后他走上了一条和计算机专家合作，有意识地根据计算语言学的需要而进行词汇和语义研究的道路，并且取得了一个又一个令人瞩目的新成果，主要有描写格关系的《现代汉语动词大词典》(主编)和基于原则和参数语法的《现代汉语述语动词机器词典》(和清华大学计算机系合作)。

语言科学在我国一直是一门冷门科学。在整个社会重理轻文，重文学轻语言风气的影响下，从事语言教学和研究的人常常感叹“学”不逢时，无人重视，无人支持，难以有所作为。但是冷静下来反躬自问，我们这些长期从事语言教学和研究的人又为社会做了些什么？70年代末，我在北京市语言学会的工作会议上讲过，“要社会为语言学服务，语言学首先要为社会服务”，如果我们为社会服务得好，社会就会重视语言科学，就会支持语言科学

的发展。当今的时代是信息时代，信息处理可以说是热门中的热门，而信息处理实际上就是对语言文字资料的处理，这既离不开计算机科学，也离不开语言科学。现在的情况是语言科学严重滞后，拖了计算机科学的后腿。因此，社会要求语言科学担负起自己应尽的义务，作出自己应有的贡献。在这种情况下，语言学家就不是难以有所作为，而是大有可为，就看你“为”还是“不为”了。

要使自己的研究工作为信息处理服务也不是轻而易举的，特别是对文科出身的语言学家来说，要这样做就有一个知识更新和改变思维方式的问题，有一个再学习和再实践的问题。林杏光同志选择了这样一条道路，经过自己的辛勤劳动，取得了不少成就。现在他把自己在这方面的经验和心得体会，根据多年讲授词汇语义研究课程的内容，写成这部专著，这不仅对有志于从事计算语言学研究的语言专业学生有指导作用，而且对愿意为信息处理贡献自己一份力量的语言学家来说，也有参考价值。这部著作的特色是完全不同于一般的词汇学和语义学专著，重点不在介绍国外哪一派哪一家的理论和体系，而在阐述为信息处理服务的词汇和语义研究需要有什么样的思路，需要具备哪些理论知识，以及本人对词义分类和词语搭配研究的体会与心得，本人对格语法的落实、改进和对现代汉语语义系统的构想。全书不仅介绍了国外有关领域的最新动态，而且还概述了国内有关领域研究的历史和现状，涉及面很广。因为大部分有关的理论和知识都是作者在工作中经过消化了的，所以写得简明扼要，明

白易懂。

为母语教学服务的语言研究也许能做到大概其就可以了，因为学生和读者会自动补充必要的细节，也会自动纠正可能有的失误。但是为计算机处理自然语言服务的语言研究却要求形成明确的规则，不能含糊其词。当然，语言现象是非常复杂的，也不是完全均质的。有些现象出现的条件十分复杂，不过不管怎么复杂，只要有规律可寻，就可以形成明确的规则；有些现象属于例外性质，无法用明确的规则来描写，那就可以穷尽地列举；还有些现象则处于中间状态，那就需要用概率或隶属度的办法来处理。总之，为信息处理服务的语言研究要尽可能做到规则化、系统化、计量化。语言学和计算机科学相结合还有一大好处是研究的成果可以通过计算机自动分析和逆向生成来检验是不是符合语言实际，从而可以避免公说公有理，婆说婆有理，无法判断是非的弊病，大大加强了语言研究的科学性。因此，语言学和计算机处理自然语言的研究相结合是当今语言学发展的大方向，应该大力提倡。

《词汇语义和计算语言学》不仅讲述了和词汇语义研究有关的不少新理论、新方法，而且还讲述了计算语言学最基本的知识以及和计算语言学密切相关的语法领域的一些新理论和新方法，所以想了解和借鉴当代语言学的新理论和新方法的读者可以从中得到不少有用的信息和启迪。

胡明扬

1997年9月

开 场 白

一部著作的正文前面主要用来说明全书的写作目的、写作经过的文章叫序言。在现代汉语词汇中，跟“序言”的意思相同或相近的词语有22个（难免有遗漏）：序言（叙言）、序文（叙文）、序（叙）、代序、自序（自叙）、引言、引论、导语、导言、导论、绪论、弁言（弁，古指男人戴的帽子）、题词、引子、缘起、前言、出版说明、小引、楔子（近代小说加在正文前面的片段）、入话（指径入门的话）、开篇、开场白等；跟“序言”的意思相反或相对的词语有6个（难免有遗漏）：跋、跋文、跋语、后记、书后、附言。序跋和题跋，意思相近。通常说，书前为序，书后为跋。跋，从足，本义是足后的意思，所以它作为一种文体时指写在书后的东西。我国南方有一种曲艺，叫弹词，它是演唱故事的。在演唱故事之前往往先弹唱一段唱词，这段唱词叫开篇。有的戏剧或某些文艺演出开场时有引入本题的道白，这道白叫开场白。现在，开篇、开场白也作为一篇讲话或一部论著指径入门的向导。如下是本书的开场白。

全书分为三部分。第一部分是第一章绪论，讲述研究任务、研究方法、研究态度。第二部分是第二章至第九章，讲述词汇和计算词汇学。第三部分是第十章至第十三章，讲述语义和信息处理用的语义研究。客观事物是有内在逻辑联系的，反映客观事物的语言作品也必然有内在的逻辑联系。因此，读书，在理解内容时要注意分析其内在的逻辑联系。本书内容的逻辑联系，总的是从虚讲到实，即从研究任务、研究方法、研究态度，讲到所研究的具体问题，这是第一条逻辑思路。对所研究的具体问题，是从整体讲到局

部，即从词汇讲到词汇语义，这是第二条逻辑思路。对词汇语义的研究，是从小讲到大，即从语义构成讲到语义聚合，再讲到语义组合，这是第三条逻辑思路。总之，从虚到实，从整体到局部，从小到大，这就是本书安排内容的三条逻辑思路。

书中主要讲述作者自己的研究实践和体会，以词汇、语义和计算语言学为研究重点。本书值得语言研究工作者、计算语言学研究工作者、语文教学和对外汉语教学工作者作为参考，也符合广大青少年学习语文知识和语言信息处理知识的需要。

10多年来，我结合多种类型辞书（义类词典、搭配词典、动词格关系词典、述语动词机器词典）的编纂，结合国内外重大科研课题（亚洲五国合作项目《多国语言机器翻译系统》、中新合作项目《华文报语言文字应用研究》、国家自然科学基金项目《现代汉语述语动词机器词典的研究和建立》、国家“九五”重点攻关项目《中文信息处理技术及产品开发》、国家“九五”社科基金重大课题《信息处理用现代汉语词汇研究》、国家“九七三”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发掘》）的研究，潜心钻研词汇、语义和计算语言学，并将钻研所得用于研究生教学。本书是根据我多年在中国人民大学语言文字研究所给研究生讲授词汇、语义和计算语言学研究课程的内容写成的。书名是我国著名语言学家胡明扬教授提选的。胡明扬先生不但提选书名，而且还撰写序言。在序言中，不但中肯地分析了我的书，热情地给了我许多鼓励，而且还给信息时代的中国语言学指明了方向。谨此对胡明扬先生致以衷心的谢意。

自然语言处理作为人工智能的一个分支，已有40年的发展历程，形成了计算语言学这一跨接语言、信息、认知科学和计算机技术的边缘学科。计算语言学的研究方向是面对整个自然语言处理，而汉语计算语言学注重的目标是中文信息处理。中文信息处理包括汉字信息处理和汉语信息处理。1974年周恩来总理亲自

批准“七四八”工程，标志着中文信息处理技术首开先河。经过20多年的努力，字处理已经陆续取得了相对的突破。目前，词句篇章处理阶段的攻关已经开始，已由黄曾阳先生创立的HNC(概念层次网络)理论赢得了语句理解的突破。在这一阶段的大规模真实文本自动处理中，不但需要计算机的硬件、软件技术，而且需要语言学的研究成果；不但需要懂得语言学的计算机专家，而且需要懂得计算机的语言学专家。计算机要智能化，语言研究要现代化，语言学和计算机科学的结合是历史发展的必然趋势。全国人大常委会副委员长、著名语言学家许嘉璐教授指出：“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。”顺应这一历史发展潮流，10多年来，我在现代汉语研究和中文信息处理的结合研究方面做了一些工作。《词汇语义和计算语言学》反映了我这10多年来所走过的科研历程。表达我热诚希望语言科学和计算机科学进一步沟通和合作的心情，正是我写作本书的一个目的。中国中文信息学会理事长、中国工程院资深院士陈力为教授为本书题词；全国计算语言学专委会首届专委主任鲁川教授给本书的写作提了许多宝贵的意见。谨此对陈力为先生和鲁川先生致以衷心的谢意。

我的研究生苗传江同志为本书的出版做了许多工作。责任编辑程荣先生，不但在出版方面，而且在修改书稿的内容方面，都做了大量精细的工作。我谨在此一并表示衷心的谢意。

林杏光

1998年春节于中国人民大学对外语言文化学院

目　录

第一章

绪　论

第一节　研究任务

一、什么叫人、机两用？

20世纪50年代由六大新技术群(材料、能源、信息、生物、海洋、空间)引起的新产业革命，已经成为全球性的空前巨大的生产力革命。在六大新技术群中，生物革命和信息革命是核心。前者带来农业革命，是人类社会经济可持续发展的根本保证；后者带来信息社会，是社会形态的变革。在信息时代，语言研究应该注意人、机两用。这就是信息时代的语言研究的任务，当然也就是词汇、语义和计算语言学研究的任务。

什么叫人、机两用呢？所谓人、机两用，就是在选择研究课题时，既要考虑人际交流的需要，也要考虑人机对话的需要。

二、为什么要注意人、机两用？

1996年4月15日，世界著名物理学家李政道接受北京电视台的记者采访，他用市场需要、鱼、水三者来比喻应用研究和基础研

究各自的重要性。他认为市场需要鱼，但想要有鱼就得有水。鱼如同应用研究，水就是基础研究。只顾鱼，而无水，则亦无鱼。这就是说，不能只搞应用研究而不进行基础研究，当然只进行基础研究而不搞应用研究也不行。在市场经济社会，不管是应用研究还是基础研究，都务必和市场的需要紧密联系起来。进行语言研究当然也必须考虑市场的需要。

今天，在我们国家，语言研究成果的应用市场主要还是在人际交流方面，即语文教学和对外汉语教学方面，而后者又是重点。在世界上，说汉语的人数约占世界总人口的五分之一。第28届联合国大会于1973年12月18日的全体会议一致通过，把中文作为联合国大会安全理事会的六种工作语文之一（其他五种工作语文是英语、法语、俄语、西班牙语和阿拉伯语）。随着我国改革开放的事业不断发展，对外经贸文化交流日益频繁，学习汉语的人数越来越多。一个学习汉语的热潮，正在全球范围内升温。有些国际有识之士预测：科技和经济的发展中心21世纪将转移到亚太地区，而要和亚洲交往必须学习汉语，有些国家已把学习汉语作为迎接21世纪挑战的一项战略措施。为对外汉语教学研究汉语和中国文化已成为语言学发展新潮流中的奔腾巨浪！我们的语言研究成果应该适应对外汉语教学这一广阔市场的需要。

能使语言研究的成果适应对外汉语教学的需要，就是解决了市场需要的一大半问题，应该感到满意，但不能就此而满足，因为市场的需要还有一个人机交流的方面。这一方面和教学的需要相比，就当前看，还不是大量的，普遍的，但从发展的前景看，它充满着盎然的生机，前途无量。因此，凡是有眼光的语言研究工作者，应该高瞻远瞩地眺望到它的灿烂远景，努力使自己的语言研究去适应人机交流这一市场的需要。注意人机交流的需要是时代赋予语言研究工作者的新的历史使命。

在计算机的研究课题中有许多语言问题需要语言研究工作者

去和计算机研究工作者共同努力，求得解决。在这里谈点 MMT（《多国语言机器翻译系统》）的汉语生成情况。MMT 这个研究项目是中国、日本、马来西亚、印度尼西亚、泰国等五国共同研制的国际合作科研项目，其技术选择采用中间语言转换的方式，即：分析，从表到里地落实到中间语言；生成，由中间语言从里到表地引导到目标语言。汉语生成是 MMT 的一个子课题，它的接口是中间语言表现形式。整个汉语生成过程采用规则控制。通过一系列的生成手段生成出表达原文意思的汉语。有一次进行 263 个句子的汉语生成实验，其结果是，正确的 67 句，如："我有一本书""老师看书""他不是老师""他肯定知道""老师明天肯定去大阪""你明天去京都？""明天去京都的是你""他扮演重要角色""鸟飞去了南方""通常用 0 和 1 表示这 2 的知识""图一指出磁带和磁带机""人们能够制作任何东西""日本中国已缔结协约了""中央处理器是由控制器和译码器构成的""所有国家政府正在忧虑于环境污染问题""构成休息的重要成分是睡眠和松弛"。有错误但还能看懂意思的 111 句，如："数学方面擅长他"（"他擅长数学"或"他在数学方面擅长"）；"我们在估计它中"（应删"中"）；"由第一章和第二章和第三章和第四章形成本文"（一、二个"和"应改为顿号）；"定义人有各种各样"（"人的定义有各种各样"）；"昨天已经学了书兴趣了吗？"（"昨天已经学了有兴趣的书了吗？"）；"爱恋他家时，他哭泣"（"他想家时哭泣"）；"孩子在看夕阳落中"（"孩子们在看日落"）；"我们在开发翻译的计算机中"（"我们在开发能翻译的计算机"）；"他在念中书是女性运动书"（"他在看中文的妇女运动的书"）；"你的问题是在我回答方面深奥的"（"你的问题我回答是困难的"）；"我们得用电车去"（"我们得乘电车去"）；"如果你帮助我，我帮助你"（第二个"我"之后应加"就"）；"他昨天在沿海已经游泳了"（"他昨天已在海滨游泳了"）；"他在站中，他是腰酸腿疼的"（"他站得腰酸腿疼"）；"他是在语言学擅长查日语和查英语的人"（"他是个在语言

学擅长日语和英语的人");"我直到12点在实验室有"("我直到12点还在实验室")。错到看不懂意思的85句,如,"日本经济各国家是发展亲密是合作的""必须把从天然有的能源的人类依靠向能源""可以想我们的国家深度技术的学术领域的正式无效向相当周围的座次""使去技术是改革在新在生成中的这个芽长大和发展"。

以上的统计数字表明,在MMT的汉语生成中存在的语言问题是很多的。产生这些语言问题的原因是多方面的,或是分析错误,或是中间语言错误,或是生成规则错误,或是词典错误,等等。所有这些错误都跟语言研究有关。语言研究工作者有责任和计算机工作者密切合作,去解决这些语言问题。因此,从计算机研究课题中所存在的大量语言问题来看,我们今天的语言研究工作也应当充分注意人机交际的需要问题。

综上所述,当前的汉语研究要特别重视两个"热点":一是计算机的应用,一是对外汉语教学的应用。忽略了这两个"热点"就是舍弃了汉语研究的重点,就是丧失了汉语研究的大好机遇。

有人说,近半个世纪以来,促进我国语言科学发展的动力有三个:对外汉语教学,科学技术的发展,50年代的全国性方言普查。这一说法启迪我们,从语言科学的自身发展来看,我们今天的语言研究也要充分注意对外汉语教学和科学技术的需要问题,即人、机两用的问题。

三、如何为计算机研究语言?

首先要使自己具有"愤悱"的强烈意识。孔子曰:"不愤不启,不悱不发。举一隅不以三隅反,则不复也。"(《论语·述而篇第七》)孔子的意思是说,教导学生,不到他想求明白而不得的时候,不去开导他;不到他想说出来却说不出来的时候,不去启发他;教导他

东方，他却不能由此推知西、南、北三方，便不再教他了。有一位高科技专家据此概括出一种教学方法，叫“愤启悱发”。我从这个“愤启悱发”的教学方法中，领会到一种“愤悱”的学习方法，这就是使自己具有“愤悱”的强烈意识。所谓“愤悱”的强烈意识，就是“想求明白而不得、想说出来却说不出来”。说白了，就是老憋着一股要学习的劲儿，具体到为计算机研究语言来说，就是老憋着一股要为计算机研究语言的劲儿！“愤悱”的强烈意识从何而来？我认为它来自语言研究工作者所肩负的新的历史使命。时代在前进，社会在进步。计算机要智能化，语言研究要现代化。智能化的计算机离不开语言学，现代化的语言研究必然走向和计算机相结合的道路。为计算机研究语言，是信息时代赋予语言研究工作者新的历史使命。当有了“新的历史使命”感的时候，“愤悱”的强烈意识就自然产生了，为计算机研究语言的劲头儿也就有了。

这里有高级阶段和低级阶段之分。高级阶段的标准，就是要使语言研究的成果规则化、形式化、算法化、程序化。只有规则化才能形式化；只有形式化，才能算法化；只有算法化，才能程序化；只有程序化，才能在计算机上实现。按这个高级阶段的要求，语言研究工作者要具有现代语言学、数学和计算机科学的知识。低级阶段的标准，只要求语言研究瞄准计算机的需要，计算机需要什么就研究什么。按低级阶段的要求，就是略懂计算机和不懂数学也可以为计算机研究语言。就我所知，计算机界当前还不是用高级阶段来要求语言研究工作者。只要我们的语言研究成果在方向上符合他们的要求，便于他们进行加工和改造，即便于他们形式化、算法化、自动化，他们就非常欢迎。我就是按低级阶段要求来为计算机进行语言研究的。我将自己定位于低级阶段，目的是便于尽快起步，以免望形式化、算法化、程序化而生畏。但这并不意味着自己永远满足于低级阶段，止步于低级阶段。从低级阶段做起，在做的实际过程中，积极向中文信息界学习，使自己逐步向高级阶段

发展，争取最后达到高级阶段的水平，成为一个真正的计算语言学研究工作者。这就是我的奋斗目标。

我听说词义分类对计算机有用，我就花了好多时间进行词义分类的研究，出版了三部系列性的词义分类词典（《简明汉语义类词典》、《四字语分类写作词典》、《汉语多用词典》）；我听说词语搭配对计算机有用，我就和一些同志合作，花了很多时间进行词语搭配的研究，出版了四部系列性的词语搭配词典（《学生常用词语搭配词典》、《简明汉语搭配词典》、《现代汉语实词搭配词典》、《现代汉语辞海（即词语搭配大典）》）。这两项研究成果对计算机有没有用呢？那就要看怎么要求。如果要求在义类词典中找到一个计算机识别语言所需要的现成的概念分类体系，那是找不到的，但可使计算机工作者减轻一点研究概念分类体系的劳动；如果要求在搭配词典中找到计算机识别语言的形式化搭配规则，那是找不到的，但可使计算机工作者具有研究词语搭配的形式化规则的丰富语料。

美国语言学家菲尔墨（Fillmore）提出格语法，受到人工智能界的赏识。为了适应计算机的需要，我和鲁川组织一批年轻人先后编著和出版了《动词大词典》、《现代汉语动词大词典》，并以此为基础和清华大学计算机系共建《现代汉语述语动词机器词典》。

从义类词典、搭配词典到格关系动词大词典，再到述语动词机器词典，反映了我的语言研究不断向计算机的需要靠拢的发展过程。

选择面向计算机的语言研究课题必须选择“面”。所谓选择“面”，就是选择较大的领域来作研究课题。记得在年轻的时候曾一度为搞科研入不了门而烦恼，恰似在花园门口徘徊，就总是进不到万紫千红的花园里去。有人叫我选个小题目，从点做起，逐步扩大到面。这种办法对我不灵，小点而没有面做基础，深入不下去，无法以小见大，写不出文章来。后来我便从面入手，选一个较大的