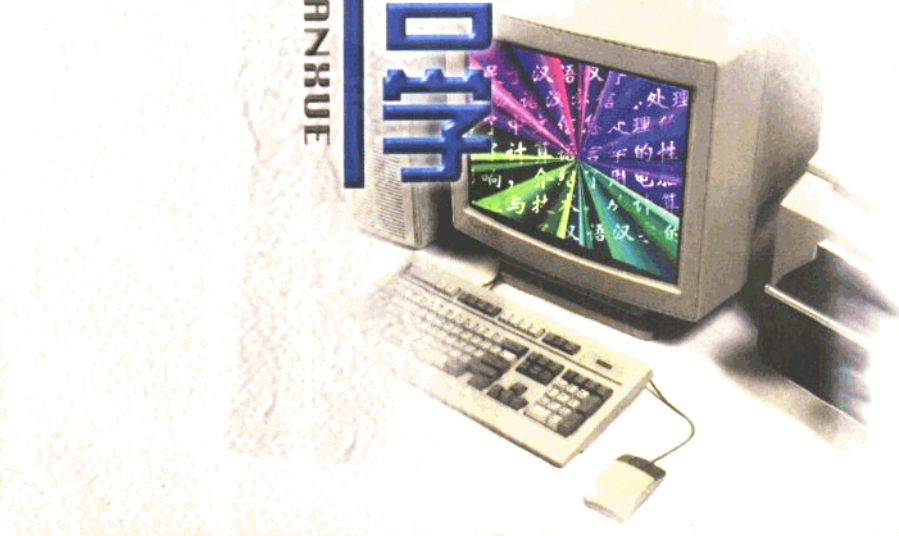
姚亚平 著

江西科学技术出版社

中国计算语言学

ZHONGGUO JISUAN YUYANXUE



序　言

在即将迈入21世纪门槛的时候，我们反观20世纪汉语本身的变化图景和汉语研究的发展脉络，就可发现：无论是汉语结构、汉语生活的历史变化，还是中国语言学研究的学科发展，都受到政治、经济、文化和学术等许多非语言因素的重大影响。这些因素有：社会实践的变化和变迁、中外文化的碰撞和交流、学术传统的沿革和继承，以及在语言运用和语言研究中所出现的科学观念和技术进步，等等。考虑到这一背景，姚亚平同志所著《中国计算语言学》的出版，确实是一件令人高兴的事情。

这部著作描述了语言研究、语言运用和计算机技术结合后所产生的种种现象，介绍了计算语言学的性质、任务、理论、方法与影响。特别要指出的是，它不但探讨了计算机处理一般"语言文字"信息的基本原理，而且探讨了计算机处理"汉语汉字"信息的具体问题与特点，翔实而系统地介绍了我国语言文字信息处理的具体实践和各项成就，勾勒出我国在汉语汉字信息处理方面科技进步、理论探索的轨迹，全面而深入地探讨了中国语言信息处理和汉语研究的有关问题，介绍了汉语汉字信息处理的基本知识、基本原理和基本技能。

这部著作不仅介绍了计算机处理汉语汉字的技术，而且分析了计算机处理汉语汉字的原理，分析了语言文字信息处理技术与实践中的语言学理论，即不但讲了"怎么样"，还讲了"为什么"，从语言学的角度对计算机科学的有关思想与技术作了阐发。此外，此书还论述了语言信息处理给我国语言学研究带来的影响，从一个新的角度对汉语汉字的特点与规律作了新的理解和揭示，对现阶段中国语言学学术观念、学科选题和理论方法作了新的分析与论述。在这两个基础

上，阐述了中国计算语言学的性质、原理、任务、方法及其影响，反映了这一学科领域的最新成果，表现出作者思考的深度和学术的探索。

这部著作在思路和体例上显示了高度的联系性和完整性，它不是孤立地就语言学谈语言学，而是从本世纪新科技革命的宏观背景，介绍了中国计算语言学兴起的社会、文化和学术背景。它从汉语汉字的基本特点以及当代语言交往的社会实践出发，探讨了汉语实践和中国语言学发展的内在规律，既把现代汉语和现代汉字放在计算语言学的视野中进行一番审视，对汉语汉字的性质、特点、结构、功能、属性以及我国的语言研究与文字改革的实践作了一番反思，又按字、词、句等语言单位的大小级别和中文信息处理的历史进程，分别介绍了中国计算语言学的主要理论和技术。见解独到，论述严密，资料翔实，语句通俗。这部著作注重把握语言实践问题与语言研究课题之间的转换关系，揭示出中国计算语言学理论研究和学科发展中历史与逻辑的相互统一。这对语言学的其他研究也是颇有启发意义的。

这部著作通过分析中国计算语言学的研究实例，正确地处理了汉语事实与西方理论的关系。与计算机技术一样，国外的计算语言学的理论和技术比在国内起步早。但是，西方计算机语言信息处理系统是建立在西方语言理论之上的，而西方的语言学理论又是建立在对西方语言自身特点的考察之上的，这种语言理论体系和这种计算机语言信息处理系统在建立之时，并没有考虑我们中国语言和中国文字的事实。这样一来，我们在运用这种理论和这种系统时，就必须注意其适用性。事实上，自《马氏文通》以来，中国语言学发展的历史都证明了我们对国外理论必须有所借鉴，更要有所创造，最重要的是要尊重汉语事实，建立符合汉语实际的中国语言学理论。亚平同志的著作通过对我国语言信息处理实践的考察和对中国计算语言学理论的阐述，也表达了这一基本观点：从汉语事实出发，借鉴古代和国外的一切学术成果，形成自己的理论，进一步指导我们的语言研究和语言运用的实践。

当然，中国计算语言学是一门年轻的学科，许多理论与观点还在

发展变化之中，汉语汉字信息处理的实践还在日新月异地向前推进，亚平同志的这部著作还有待于充实丰富之处，比如在计算语言学的基本理论的探讨上尚有待丰富完善。但从研究的切入点、具体的理论把握以及学科的发展前景来看，本书不失为一部具有开拓意义的学术专著。

亚平同志是在改革开放时期成长起来的年轻学者。他好学深思，刻苦勤奋，多年来勤于钻研，锲而不舍，特别是在担任行政工作后，双肩重负之下，仍执掌教鞭，潜心学术，撰述不倦，硕果累累，这是值得钦佩的。学海无涯，深望亚平同志能继续钻研，以取得更高的成就，为社会主义建设事业作出更大的贡献。

是为序。

胡 裕 树

1996 年 11 月 28 日

目　录

序言 …… 胡裕树
绪　论 …… (1)
第一章　中国计算语言学的兴起 …… (5)
第一节　信息处理的技术革命 …… (5)
一、电脑热：信息处理的科技革命浪潮 …… (5)
二、语言信息处理的科技进步 …… (9)
三、语言信息处理与办公自动化 …… (13)
第二节　信息贮存与复制的技术革命 …… (17)
一、语言文字信息贮存的技术进步 …… (17)
二、语言文字信息印刷的技术进步 …… (20)
三、语言文字信息处理的网络技术 …… (24)
第三节　信息传递的技术革命 …… (25)
一、从电话热到E－mail热：信息传递的科技革命浪潮 … (25)
二、语言信息传递的科技进步 …… (28)
三、语言信息传递技术的发展趋势 …… (31)
第四节　语言信息处理与信息高速公路 …… (34)
一、全球范围的修“路”热 …… (34)
二、信息高速公路的性质与特征 …… (36)
三、信息的处理与传递：信息高速公路的意义与作用 …… (40)
四、我国信息高速公路的铺“路”情况 …… (45)
第二章　中国计算语言学概说 …… (53)
第一节　中国计算语言学的性质和特征 …… (53)
一、利用计算机研究和处理自然语言的新兴学科 …… (53)

二、适应信息化社会需求的应用学科…… (61)
三、计算机技术和语言学相互结合的交叉学科…… (64)
第二节　中国计算语言学的目的和任务 …… (68)
一、汉字汉语状况的调查描写及其特点规律的分析研究
…… (70)
二、汉字汉语特点规律和计算机技术设备匹配关系的建立
…… (71)
三、汉字汉语信息处理与交换的技术标准与规范的建立
…… (76)
四、计算语言学基本理论与学术范式的建立…… (79)
第三节　中国计算语言学的技术与课题 …… (85)
一、中国计算语言学的基础研究…… (85)
二、中国计算语言学的应用研究…… (87)
三、汉字信息处理系统…… (88)
第三章　计算语言学视野下的现代汉字 …… (96)
第一节　电脑与现代汉字 …… (96)
一、计算机的组成与工作原理…… (96)
二、电脑的系统软件和常用命令 …… (100)
三、汉英兼容技术 …… (102)
第二节　现代汉字概说…… (103)
一、汉字的性质与来源 …… (103)
二、现代汉字的字形结构 …… (105)
三、现代汉字的字形功能 …… (112)
第三节　现代汉字的属性…… (116)
一、现代汉字的字量 …… (116)
二、现代汉字的字频 …… (120)
三、现代汉字的字序 …… (121)
四、现代汉字的字形 …… (122)
第四章　汉字的信息处理…… (127)

第一节　信息处理与民族语文支撑能力…… (127)
一、“中文信息处理”与“汉字信息处理” …… (127)
二、汉字信息处理的主要内容 …… (129)
三、汉字信息处理的发展阶段 …… (130)
第二节　汉字编码与汉字输入…… (132)
一、汉字输入的方法类型 …… (132)
二、电脑键盘的匹配与汉字编码 …… (136)
三、五笔字型输入法 …… (143)
第三节　汉字的存贮和汉字的输出…… (154)
一、汉字的存储 …… (153)
二、汉字的输出 …… (156)
第五章　汉语词语的信息处理…… (158)
第一节　语料库:词语信息处理的理论和技术 …… (158)
一、语料库的性质与建设 …… (158)
二、语料库的意义 …… (161)
三、语料库的基本功能 …… (166)
四、语料库的建库原则与开发流程 …… (167)
第二节　汉语词语的词频统计…… (169)
一、词频统计的意义与方法 …… (169)
二、词的分级与词表、词库的建立…… (171)
三、现代汉语的词频统计与汉语言语统计的特殊性 …… (172)
第三节　汉语词语的自动切分…… (174)
一、自动分词的概念与意义 …… (174)
二、汉语自动分词的方法 …… (175)
三、汉语“词”的歧义切分 …… (177)
第四节　汉语词语的自动标注…… (178)
一、汉语“词”的语法自动标注 …… (179)
二、汉语“词”的语义自动标注 …… (180)
三、汉语专有词语与未定义词语的自动切分与标注 …… (184)

第六章　汉语句子和篇章的信息处理…… (189)
第一节　汉语句子的信息处理…… (189)
一、句子的自动切分 …… (189)
二、句法自动标注 …… (190)
三、句法的自动分析 …… (191)
第二节　自然语言的理解…… (191)
一、自然语言理解的原理 …… (191)
二、汉语语句的生成系统 …… (193)
第三节　机器翻译…… (195)
一、我国机器翻译的技术发展 …… (195)
二、机器翻译的原理 …… (200)
三、汉语机器翻译系统的语言理论基础 …… (203)
第七章　中国计算语言学的应用研究…… (207)
第一节　中文信息情报检索…… (207)
一、计算机情报检索的概念 …… (207)
二、计算机情报检索的工作过程 …… (208)
三、计算机情报检索的类型 …… (210)
四、中文"三古"现代化 …… (212)
第二节　计算机辅助教学…… (215)
一、计算机辅助教学的发展与类型 …… (215)
二、计算机辅助教学的优点 …… (218)
三、计算机辅助教学的工作原理与过程 …… (219)
第三节　计算机在语言研究中的运用…… (219)
一、计算机与词典 …… (219)
二、术语数据库 …… (221)
三、方言学研究 …… (224)
第八章　中国计算语言学的学科理论…… (231)
第一节　信息处理的句法理论…… (231)
一、短语结构语法 …… (231)

二、语言串理论 …… (234)
第二节 信息处理的语义理论 …… (235)
一、格语法理论 …… (235)
二、汉语动词的概念分类 …… (237)
三、汉语名词的语义分类 …… (240)
第三节 信息处理的概念理论 …… (244)
一、智能计算机的语言知识的表示 …… (244)
二、语义分类的层级结构 …… (245)
三、复杂特征集的属性描述 …… (245)
四、电子词典 …… (246)
第四节 中国计算语言学的理论影响 …… (250)
一、对语言学基本理论的影响 …… (250)
二、对语言学各分支学科的影响 …… (251)
三、对语言学研究方式的影响 …… (253)
第九章 中国语言学的学科前景 …… (256)
第一节 中国语言学的学科发展结构 …… (256)
一、面向未来在学科的发展结构的位置 …… (256)
二、学科历史总结与现状分析中的发展意识 …… (257)
三、发展的观念是学科发展的思想基础 …… (260)
第二节 中国语言学的学科发展动力 …… (261)
一、问题驱动是中国语言学的发展动力 …… (261)
二、理论与现实的关系是中国语言学学科发展的基本矛盾 …… (264)
三、课题攻关是集结队伍、组织科研的现代方式 …… (267)
第三节 中国语言学的学科发展机制 …… (269)
一、操作:语言技术与语言理论的结合 …… (269)
二、中国语言学的"学""术"之争 …… (272)
三、中国语言学的学科发展目标 …… (275)
第四节 中国语言学的学科发展环境 …… (280)

一、学科问题的综合化与科学研究的综合化 …… (280)
二、学科的自我反省和学科的相互对话 …… (283)
三、世纪之交中国语言学的远大前景 …… (285)
后记…… (289)

绪　论

20世纪人类最伟大的科技进步，恐怕就是计算机所引发的科技革命了。

20世纪汉语应用领域里最重大的科技进步，恐怕也就是运用计算机技术进行中文信息处理了。

在这一宏观背景与现实基础上产生的中国计算语言学正像旭日东升，放射出熠熠夺目的万丈光芒，展示出越来越深刻的理论意义和越来越广阔的应用前景。

随着汉语汉字信息处理技术的相继突破，电脑在我国各个领域的运用迅猛发展，电脑和电脑技术早已不是科研院所和专业人员的独有用品，而正快速地进入各个单位、学校和家庭。广大人民群众了解电脑、掌握电脑的需求不断高涨，在我国形成一浪高过一浪的科技热潮，人们的“写”字和“说”话的手段与方式已经并即将发生重大的变革。

写字说话手段和方式的转换具有革命性的意义。在“字”用笔来写、用打字机来打、用印刷机来印的时代，人们容易产生汉字崇拜、铅字崇拜。在这些时代，文化程度不同的人对文字印刷品的消费水平有很大差异。写字、写书、读书、藏书较多的一小部分人容易产生一种“汉字情结”、“铅字情结”，居高临下地看待芸芸众生。所谓“象牙塔”的墙壁是用汉字与铅字垒筑而成的。语言文字的本来使命是用于人际沟通、文化交流的，但在那个时代，文字由于难“写”、难“打”、难“印”，不能为广大劳动人民所掌握，甚至产生众多的文盲，造成全社会的文化隔膜。以计算机技术为标志的新科技革命推动语言文字的信息处理技术不断向前发展，取得了人类历史上文化传播的空前突

破。计算机科学与文化克服了铅字对读者的选择性。新技术革命最终将引发新文化革命。基于这一点，对语言文字信息处理手段与技术作一番描述和论述，是非常必要的。

本书立足于汉语汉字信息处理的原理，立足于汉语汉字信息处理的成果，详细介绍了中文信息处理的实践与成果，介绍了计算语言学的性质、任务、方法与影响，介绍了用电脑处理语言信息的原理与技术，从计算语言学的角度重新分析了汉语汉字的特点，特别是本书不但探讨了计算机为什么能和怎么样处理"语言"信息，而且探讨了计算机为什么能和怎么样处理"中文"信息，是一部"中国"计算语言学。这样，读者不但比读一般电脑书能更好地对电脑操作有理性把握，而且对汉语汉字的特点也有进一步的认识。

近 10 多年来，我国在汉语汉字信息处理方面不断取得显著成果，并且还有大突破，现在中国计算语言学也有了丰富的实践经验和理论成果，对此作一个全面系统的介绍已具备条件。

本书在勾勒我国在汉语汉字信息处理方面科技进步轨迹的同时，介绍了汉语汉字信息处理的基本知识和技能，帮助人们从一个新的角度了解和运用电脑，认识汉语汉字，推动语言学和计算机学的相互结合。

本书在结构安排上，注重了中文信息处理的内容分割和技术进步的时间统一性，更注意了中国计算语言学的学科结构的完整性与理论结构的逻辑性。

第一、二章主要从人类信息处理和信息传输技术的进步，特别是从电脑热、电话热、办公自动化和信息高速公路在全球迅速兴起的大背景中，介绍了中国计算语言学兴起的必然性，主要介绍中国计算语言学的学科性质、内容与主要研究对象，阐明了中国计算语言学的意义与前景。

这些年来，中文信息处理技术的进步表现在各个方面：在语言单位上，主要表现在字、词两级单位的处理上；在语言应用上，主要表现在情报检索、机器翻译、语言研究、辞书编纂等方面。中国计算语言学

的理论也主要在这些方面得到体现和论述。本书用第三、四章两章来论述中国计算语言学在汉语文字信息处理上的技术与理论。第三章主要把现代汉字放在计算语言学的视野中进行一番审视，对汉字的性质、特点、结构、功能、属性以及我国的文字政策与文字改革实践作了反思，为下几章进一步阐发计算语言学中的汉语信息处理作了准备。

第四、五、六章按字、词、句等语言单位的大小级别和中文信息处理的历程，分别介绍了中国计算语言学的主要理论和技术。第四章主要对汉字信息处理作了简要介绍，阐述汉字编码的原理、汉字存贮和汉字输出的技术，特别是对五笔字型输入法作了全面的介绍和分析，既有一般电脑操作书的实践指导性，又可使读者知其所以然，明白其根据与原理，明白用计算机来处理中文信息的特点、困难与发展历程。第五章主要论述了中国计算语言学在汉语词语的信息处理上的技术与理论，特别是介绍了中国计算语言学在语料库建设上所攻克的难题、所取得的成果，着重论述了语料库语言学的理论观点及其影响。第五章主要对汉语词语的信息处理作了简要介绍。分别介绍了语料库和汉语词频统计、词语的自动切分、词语的自动标注等词语信息处理的理论和技术。第六章初步涉及了汉语句子和篇章的信息处理的有关情况，着重介绍了我国机器翻译的技术发展情况。第七章主要介绍了中国计算语言学在中文信息情报检索、计算机辅助教学以及语言研究等方面的应用研究。第八章讨论了中国计算语言学的学科理论，主要介绍了信息处理的句法理论、语义理论和概念理论等，并概述其理论影响。第九章从中国语言学的学科发展结构、发展动力、发展机制和发展环境等方面论述了中国计算语言学的学科影响，展望了中国语言学的发展未来。

我们的时代正在经历着一场空前的变革。“信息”已经成为新时代的显著标志。时代标志多种多样，例如，用材料标志时代，就有旧石器时代、新石器时代、青铜器时代、铁器时代、高分子时代，等等；用能量标志时代，就有人力时代、畜力时代、蒸汽时代、电力时代、原子能

时代，等等。信息、物质、能量，同为世界三大基本资源；但信息不同于物质，也不同于能量。物质与能量毕竟是有限的，是要被耗尽的，而信息与通信相融合而形成的资源，则具有极大的灵活性，是取之不尽、用之不竭的。现在，信息技术作为时代的一种崭新标志，将即将到来的21世纪的时代特征标示得更加深刻，点燃了新时代黎明的火炬。

美国哈佛大学信息政策研究中心主任安瑟尼·G. 欧廷格(Anthony G. Oettinger)教授有言：

没有物质，就什么东西也不存在；

没有能量，就什么事情也不发生；

没有信息，就什么东西也无意义。

这就是信息的意义。

中国计算语言学就是这样一门揭示汉语汉字信息理论和技术的新兴科学。在迈向21世纪的时候，我们必须面对并解决汉语汉字信息处理这样一个崭新而不容回避的重大课题，为此，让我们掌握计算语言学这样一门学科吧！

第一章　中国计算语言学的兴起

现在，人类社会正在进入信息社会。信息已经和物质、能源一起，成为三大资源之一，影响着人们的生活与社会的发展。就人类活动的本质来说，人们每时每刻都在处理和传递着信息，特别是语言信息。现在，人类处理和传递信息的方方面面，包括信息的获取、记录、整理、复制、交流、传输等等，都在发生前所未有的巨大变革——

第一节　信息处理的技术革命

一、电脑热：信息处理的科技革命浪潮

最近，人们都可以感觉到一种事实，电脑在我国迅速热起来了，而且热浪滚滚，可谓一浪高过一浪。首先是电脑市场“热”。电脑正在以越来越强劲的势头进入成千上万个寻常百姓家。1993 年我国计算机市场销量比上年增长 43%，其中家用电脑连续两年每年增加 6～7 倍；1994 年我国共销售电脑 70 多万台，比 1993 年增长了一半，其中约有 10%为个人购买，1995 年的电脑销售量将突破 100 万台。1995 年电子工业部组织了一次覆盖全国 30 个省、市、自治区的“全国微机市场调查”。在被调查的单位中，微机购买率已达 85.06%，每天使用微机的单位占 79.33%，有 94.96%的单位准备在近 3 年内购置新的微机。

其次是电脑更新“热”。这些年，游戏机、8086、286、386、486、586 一个接一个，让人反应不过来。电脑的品种也花样翻新，层出不穷。人们刚刚将一台台台式电脑高高兴兴地抱回家中，各种笔记本电脑在

市场又悄然红火。1968年，日本东芝公司推出了第一台商品化的笔记本电脑，这种电脑体积小、重量轻、一体化、便携带，为人们提供了一个移动式的“办公室”。1993年中国笔记本电脑的市场开始启动，1994年，随着全球范围笔记本电脑技术在彩色显示屏、电池寿命、扩展性能、可升级性等关键方面取得突破性进展，世界和中国的笔记本电脑市场开始进入稳步发展阶段。有数据表明，全世界笔记本电脑的销售量在逐年增长，1993年是250万台，1994年是300万台，1995年达到500万台，一场由笔记本电脑替代台式电脑的革命正在个人计算机市场蓬勃兴起。

再次，也许是更重要的，电脑观念和计算机教育“热”。这倒不仅仅表现在“电脑”“计算机”已成为当今最普遍、最流行的常用词和时髦词，更表现在人们已普遍认识到计算机的重要性，人人都感觉到要接受计算机教育，要掌握计算机本领。人们广泛认为，信息高速公路的构想之所以会在美国产生，除了先进的技术基础这一原因之外，计算机教育大普及，计算机网络深入日常生活，从而带来了广泛的“网络文化”、“电脑意识”，也是一个重要原因。现在，计算机教育已在全世界迅速、持续、全面地升温。

我国在实现经济信息化的过程中，必须大力发展计算机教育，解决全民普及计算机知识及应用技能的问题，必须尽快提高计算机应用的整体水平，从而使各行业、各层次的人员，不论其年龄、知识背景及专业背景如何，都能掌握和应用计算机，从而解决他们自身专业领域的计算机应用问题，为他们的本职工作或专业服务，使之与国家经济信息化的需要相适应。事实上，我国在大力开展计算机教育、普及计算机知识方面也同样做了大量工作。

计算机教育包括：高校计算机专业培养计算机专门人才，从高校非计算机专业培养计算机应用人才，从在职人员培养计算机应用人才（继续教育），以及从中小学中培养计算机的后备人才。现在，我国各级各类学校都普遍开设了计算机课程，各高校都普遍设立了计算机系或计算机专业。计算机课程已成为学校最受欢迎的课程之一，计

算机系院、计算机专业是每年高考招生录取工作中最热门、最紧俏的系院、专业。这反映出计算机教育大普及的动人情景。

除了学校开设的计算机课程、系科、专业大受学生欢迎之外，面向社会的计算机等级考试也一热再热。普及计算机教育与知识的一个有效途径和重要内容是实施全国范围的计算机等级考试。世界各发达国家为了普及计算机知识，全方位、多层次地培养各行各业计算机应用人员，开展全国范围的定期的计算机各类等级考试。例如美国最权威的教育考试中心 ETS(Educational Testing Service)就面向美国社会推出了"计算机文化考试"、"高级就业计算机科学考试"和"专业领域考试"等三类考试；美国计算机专业人员认证学会 ICCP(Institute for Certification of Computer Professionals)也实施了有关的认证考试；英国计算机学会 BCS(British Computer Society)和 IDPM(Institute of Date Processing Management)分别组织的计算机等级考试，并普及到英联邦及其它国家；日本自 1969 年开始设立"信息处理技术人员考试"，现已成为仅次于日本大学全国统一考试的第二大规模的全国性考试(杨芙清，1994)。

我们国家为了面向 21 世纪，大力发展计算机教育，普及和提高国民的计算机知识和技能的素质，尤其是计算机实际操作的能力，也开设了多种计算机等级考试。1993 年，国家教委考试中心与英国剑桥大学考试委员会在北京签订合作协议，引进 CIT 项目。CIT 是剑桥信息技术的简称，它是由英国剑桥大学考试委员会主办并得到国际认可的信息技术技能培训及资格认证。丁玉章介绍了 CIT 剑桥信息技术培训的三个显著特点：

第一，培训重点明确。其意在提高学员的实际应用能力，整个培训过程贯彻理论知识讲授"少而精，学到手"、实际操作"重能力，作业化"的基本原则，即每一个模块都有明确的技能培训目标，学员要围绕这个技能目标，了解相关的理论知识，完成一项具有实际应用价值的作业。

第二，培训方式新颖。它一改国内传统书面考试的模式，采取指