计算机辅助外语教学

黄人杰 编著

上海交通大学出版社

计 算 机 辅 助 外 语 教 学

黄人杰　编著

上海交通大学出版社

沪新登字 205 号

内 容 简 介

计算机技术进入外语教学和研究，促使外语教学和研究发生日新月异的巨大变化。本书介绍计算机技术在外语教学与研究中的主要应用与成果，内容包括：计算机语料库；计算机辅助语言学习；计算机辅助词典编纂；计算机与语言统计；计算机与语言测试；计算机试题库；机器翻译；文书处理、教科书编写、语音处理、计算机控制语言实验室以及教学管理等。本书可作大专院校外语专业的教材，也可供外语教学与研究工作者阅读参考。

计 算 机 辅 助 外 语 教 学

出版：上海交通大学出版社
（淮海中路 1984 弄 19 号）
发行：新华书店上海发行所
印刷：上海交通大学印刷厂
开本：787×1092（毫米） 1/32
印张：5.125
字数：111000
版次：1992 年 12 月 第 1 版
印次：1992 年 12 月 第 1 次
印数：1—1850
科目：276—321
ISBN7—313—01065—6/TP・39
定价：1.50 元

目　　录

第一章　引论

1.1　计算机的发展和应用

1946 年世界上第一台电子计算机出现在美国的宾夕法尼亚(Pennsylvania)。这台叫做 ENIAC 的机器是一个庞然大物，占地 120 平方米，重达 30 吨，用了 19000 个电子管，价格昂贵。它的计算速度每秒仅 5000 次，却已使当时人们惊叹不已。与今天的水平来比，这是一个非常原始的装置，但正是它，开创了计算机时代。计算机问世被誉为人类文明发展史上的一个里程碑，是继蒸汽机、电、原子能等之后又一个重大的技术突破。

从那时起，40 多年来，电子计算机技术突飞猛进，发展之快令人眼花缭乱。当年轰动一时的 ENIAC 早已成为过眼云烟。今天的计算机，无论运算、处理速度，存储能力，还是应用范围都远远胜过当年。据统计，每隔 5～8 年，计算机的运算速度可提高 10 倍，而与此同时，体积和价格却大幅度下降。

这一令人惊讶的发展是由于不断采用新技术、新设计、新工艺和新材料的结果。从第一代电子管计算机(1946～1958)，经过第二代晶体管计算机(1958～1964)，第三代集成电路计算机(1964～1971)，到今天的第四代大规模集成电路计算机(1971 以后)。同时，计算机的软件技术也有相应的飞速发展。现在，美、日等国都在投下大量资金，竞相研制第五代超大规模集成电路计算机。这将是一种能够理解自然语言、识别文字图像，有学习、推理和判断能力的新型的智能计

算机。

微型计算机的出现和迅猛发展为普及计算机的应用提供了坚实的物质技术条件，由于微机价格低廉，已经涌进公司、商店、教室、银行、工厂，甚至大量进入千家万户。计算机再也不是普通人无法接触的神秘装置，而开始应用在日常生活的各个领域。

计算机的应用，已从初期局限于数值运算，扩大到自动控制、信息处理、辅助教学等方面。今天的计算机功能不再局限于"计算"，即数值运算，而更多地用于非数值运算。所以有些科学家在最近召开的国际学术会议上提出，计算机应该改名为信息处理机(Information processor)，这不是没有理由的。

信息处理也叫数据处理，就是由计算机对数据化了的信息进行加工、分析、整理。它与科学运算不同，面对的往往是众多的数据和大量的逻辑判断，并不进行复杂的数值运算，例如图书检索、工资管理等。计算机在外语教学与外语研究中的应用也属于这一范畴。字母、单词、句子、文章等都可以看作为数据，计算机的任务就是根据语言学家和语言教师的需要对文字数据作不同的加工处理。

是哪些特点使得计算机在人类活动的各个领域，包括语言研究和语言教学中，得到广泛应用呢？

首先是计算机的高速运算能力和数据处理能力。现代大型计算机的运算速度已经可以达到每秒几十亿次。这对那些计算量非常大而时间又很紧迫的工作尤为重要，例如导弹轨道的计算与调整。语言学家依靠计算机的高速处理能力，可从上百万以至上千万词的语料库中迅速检索到所需的语言材料。用人工这是无法做到的。

其次是计算机的巨大存储能力。计算机可以存储各种原

始数据、中间数据及经过处理后的数据，也可以存储程序。由于存储介质的改进，计算机的存储量几乎不受限制，这个特点对于语言研究很重要。语言研究所面对的数据量常常是很大的，因而大存储量就成了前提条件。

第三是运算准确可靠。现代计算机技术的发展已可以使计算机有非常高的准确度与可靠度。

1.2 计算机硬件

计算机系统由硬件及软件两大部分组成。计算机的电子、机械方面的物质设备称之为硬件，通常指下列四部分：

(1)输入设备是用户把需要处理的数据及如何处理这些数据的指令(程序)告诉计算机的设备。最常用的输入设备是计算机输入键盘(keyboard)。键盘上的英文字母及数字、标点的排列方式及顺序和普通的西文打字机完全一样。每个已经学会使用打字机的人，只要在手法的轻重上稍作改进，就可以得心应手地使用计算机输入键盘。但它的键数较多，主要是增加了一些专用功能键。

键盘一般和显示器结合使用。人们用键盘打字时，显示器屏幕上就会显示相应的内容。如果发现打字错误或需要对文本进行编辑时，可利用键盘上的功能键及时修改。必要时修改可以反复进行，直到屏幕上显示出用户感到满意的文本为止。但是利用键盘输入文字数据时，即使是一个熟练的打字员，其输入速度与计算机中央处理器的运算速度相比，也不过像破牛车的速度和宇宙火箭速度相比，输入成了计算机系统的“瓶颈”。当需要输入大量数据时，这一矛盾更为突出。

光学字符阅读器(OCR，Optical Character Reader)的出现大大改善了输入技术。OCR 可以将印刷文本或打印文本

自动直接读入计算机，速度很快，每小时可输入几十页。这不仅成倍加快输入速度，而且其误识率一般亦较低，但是 OCR 价格昂贵，还不能普遍采用。一种比较简单的光学符号阅读(OMR，Optical Mark Reader)，由于价格比较低廉，已在自动阅卷、人口普查、企业管理等方面广泛应用。OMR 只能识别在规定位置是否有规定符号(例如用铅笔划的一横或涂的一圈)，但这对需要大量数据输入、而数据又可以用定位作记号方法解决的场合，是非常合适的。例如人口普查时，只要在调查卡上表示不同性别、年龄、职业等的位置上打上记号，OMR 就可根据在规定位置有否记号，而将人口普查数据自动输入计算机。

已经有人在研制语声识别装置 (Speech Recognition Unit)，希望计算机能直接接受人的有声语言。语声识别技术尚待发展完善，目前还没有成熟的产品供人使用，只有不多的试验设备，效果尚不太理想。

(2)存储器是计算机中存放程序和数据的装置。它可以根据用户的要求，随时写入程序和数据，并保存在里边，一旦用户需要，又可以立即读出。它的作用相当于人脑的记忆功能。存储可分为内存(主存)和外存(辅存)两种。内存的存取方便、速度快，但存储量有一定限制。内存存储量的大小是计算机性能的一个重要指标。外存的存储量随存储介质及存储方式而异。目前常用的是软盘、硬盘、磁带。一张 5 英寸软盘可以储存 256KB 到 1.2MB，如果用来存储英文语料，放得下四万到二十万个英文词。硬盘的存储量就更大了，微机上的温盘(Winchester)的容量从 10MB 到 300MB。用户可根据工作性质选配使用。一些面向学生的计算机辅助语言学习软件用软盘就足够了，但如果想用计算机进行大规模的语言统

计，那么最好配备有硬盘或磁带机。

(3)中央处理器是计算机的心脏，它负责执行程序，处理数据。采用大规模集成电路后，中央处理器往往只是一块集成电路芯片。

(4)输出设备和输入设备一样，是人机交际的通道。它将数据及程序以用户需要的形式输出，供给用户使用。输出可分为两种形式：硬拷贝及软拷贝，为此分别要用打印机及显示器。

打印机的作用有点像平常见到的打字机，当然它结构要复杂得多，打印速度很快，快的一秒钟可以打印一行(64～132个字符)。而且打印质量好，清晰美观，可以与印刷出来的文字媲美。使用硬拷贝输出主要是为了可以较长时间保留资料。

显示器很像一台电视机，实际上一些比较简单的家用计算机，为降低价格，就把家用电视机连接在计算机上，当作显示器使用。当然专用显示器的分辨率要高得多，这样就可在屏幕上产生高清晰度的字符与图像。显示器输出的内容虽然不能长期保留，但使用方便、效率高，是用得最频繁的输出设备。

1.3 计算机软件

计算机软件的作用是有效地管理和使用计算机硬件，从而充分发挥其功能。软件分为系统软件和应用软件两大类。系统软件包括计算机语言、编译程序、操作系统等。通常随计算机硬件一起由厂商提供。应用软件是用户为完成某一项特定的运算工作或数据处理而专门编写的程序。

计算机语言是人机交际的工具。人通过计算机语言把自

己的意图告诉计算机。计算机在懂得人的意图后进行工作，并把结果用计算机语言告诉人。在处理过程中，人机还可以通过计算机语言进行会话。最早出现的计算机语言称为机器语言(Machine language)。用这种语言时，人需要把自己的指令或数据编成由一系列 0 与 1 组成的二进制代码，因为计算机只能识别 0 与 1 两种状态。用这种面向机器的语言来设计程序或输入数据是一件十分繁难的工作，容易出错，校对也异常艰难。同时不同型号的计算机由于内部设计不同，机器语言也互不通用。

后来出现了汇编语言(Assembly language) 使人们摆脱用 0 与 1 来编写指令，而使用一些比较容易记忆的数学符号如+、−等。随着计算机技术的发展，出现了高级程序设计语言 (High-level programming language)。这是一种比较接近数学公式和自然语言的人工语言。这种人工语言的词汇是有限的、语法规则是严格的。这样，程序的编写和数据的输入比较直观，不容易出错，提高了工作效率。高级程序设计语言的出现使得一般科技人员、管理人员只要经过短期培训就可应用计算机了。但这种面向用户的语言，计算机是无法直接"懂得"的。为此，还需要有一个编译程序或解释程序把高级语言翻译成机器语言以便计算机执行，而这种工作是自动完成的，不需要用户的介入。高级语言还有一个优点，它在各种不同型号的计算机上有一定通用性，也就是说，用高级语言编写的程序只需稍作改动，就可以在别的计算机上运行，这有点像自然语言中的方言。

目前在各个领域中使用的计算机高级语言不下几十种。比较有名的有 FORTRAN，ALGOL，COBOL，LISP，SNOBOL，PL/1，PASCAL，C，BASIC 等，它们各具特

色，各有自己的适用范围。例如 FORTRAN，ALGOL 常用于科学和工程计算，COBOL 是一种商用数据处理语言。目前在语言研究和语言教学中使用得较多的是 BASIC，SNOBOL，PROLOG。BASIC 是会话式高级语言，简单易学，为一般人，尤其为初学者所乐于采用。SNOBOL 有很强的字符串处理能力，是为文字数据处理而设计的，对语言研究十分有效。

随着计算机应用范围的扩大和深入，还出现了一些为在某一专门领域使用而设计的专用语言，如用于程序教学的 PILOT 语言，用于机器翻译的 COMIT 语言。这些语言在特定范围使用十分简易方便，当然它的通用性也就差些了。

用户可以针对自己的需要，利用上述高级程序设计语言编写各种应用程序，例如项目分析程序，词汇检索程序等。用于同一目的的一组相关程序放在一起，成为解决某一类问题的程序组合，称为软件包 (Package)。比较有名的用于文字语言处理的软件包有 COCOA，OXEYE。前者可以用于词语索引、词汇统计等，后者的功能包括自动句法分析等。上海交通大学科技外语系编制、上海交通大学出版社出版的 CEST 软件包是一套计算机辅助英语教学程序。一些应用程序及应用软件包已经商品化，用户可在市场上买到。这样，用户可根据自己的工作需要挑选合适的计算机软件。这些软件的使用当然比自己动手用高级语言编制专门的应用程序要省时省力得多了。

1.4 计算机与语言研究

随着计算机的语言文字处理能力的扩大及提高，计算机越来越受到语言工作者的青睐。计算机进入了语言研究的各

个领域，计算机不再是语言研究室和语言学家书房的稀客了。计算机科学与语言学结合形成了一门新学科——计算语言学(Computational linguistics)。这是一门以计算机作为工具来研究和处理语言文字的学科。

描写语言学认为：语言规律或语言现象是从客观存在的语言素材中进行定性定量分析后总结出来的。语言规律来自对客观语言材料的如实描写。对通常的语言研究过程进行观察后可以发现，它大致有三个环节：(1)语料收集，从浩如烟海的语言材料中收集语言学家感兴趣的有研究价值的原始语料。(2)语料整理，对收集到的语料按一定的方法整理分类。研究人员在这两个环节上所付出的时间、物力、精力往往要占80～90%。(3)在分析语料基础上发现规律，提出自己的观点。

当然，这三个环节不能截然分开、相反，是相互交叉、相互渗透并相互促进，出现在整个研究过程中，直至研究工作结束。此外，不同的研究课题，这三个环节的比重也是不同的。有些研究工作，例如对某一语法规律的探索，尤其是对一些罕见语法现象的探索，收集原始语料的工作将比较繁重，有时甚至要延续达几年。因为只有这样，才能积累足够数量语料，从而得出有说服力的结论。还有些研究工作，例如语言现象的频率统计，其语料整理的工作量十分大。在这两个研究环节上语言学家可以求助于计算机。语料库的发展及语言研究软件包的出现使语料的收集及整理工作变得方便了。这样，语言学家可以把自己的精力集中在研究过程中的创造性工作上，从而加速研究进程，提高研究质量。

那么，计算机可以在哪些类型的语言研究工作中发挥作用呢？回答这个问题前，让我们先看一看计算机是怎样处理

语言文字的。实际上，计算机都有一个自己的规定的字符集，这个字符集的元素包括数字(0,1,2,……9)、拉丁字母(a，b，.....y，z)、标点(• ，?,！ 等)、空格以及一些常用的符号(&，+,−,%等)。凡字符集里的单个元素或由这些元素组成的字符串，计算机都可直接进行识别、比较、排序等处理。由不同字母组成的单词，如 if 和 of,of 和 off,或虽由相同字母组成但排列次序不同的词，如 form 和 from,计算机可以很容易地识别，更长的字符串，如词组、短语、句子甚至文章，只要是由上述字符集元素所组成的，计算机当然也一样能处理。但是，这一切处理都是根据字符串组成的异同而进行的，也就是说，凡是最终能在字符串的组成或排列上有明显区别，即有明显的词形特征的，处理起来就比较方便，否则比较困难，甚至无法处理。例如要计算机从文章中寻找出 study 一词是方便的，但要计算机指出这个 study 在句中是动词还是名词，就比较困难了。这时我们就要借助其他手段或规则，如前面是 the，my 等，则是名词，如是 can，will 则是动词。这实际上仍是词形上的识别，即借助邻近词的词形来识别。

象形文字如汉字的计算机处理将会遇到困难，它不能直接进入计算机，因为汉字不像英文是由上述字符集元素直接组成的。这时就需要对汉字进行编码，也就是说，把汉字用一定的规则编成由上述字符集元素组成的码，然后输入计算机处理。汉字编码方案多达几百种，大致可分为三类：音码、形码和音形结合码。无论哪一种码。都是用 26 个字母及 10 个数字的不同组合来代表不同的汉字。目前还没有一种方案已得到全社会的一致认可。

现在我们可以知道了，对于那些最终可以归结到词形异同上的语言现象的处理，计算机是驾轻就熟的，这也是计算机

在语言研究上最能发挥作用并取得极大成功的领域，例如词汇统计、自动检索、词典及常用词表的编制等。某些语言，例如俄语、德语，语法形态变化丰富，则计算机自动句法分析就较容易，相反，某些语法形态变化贫乏的语言用计算机进行自动句法分析就有一定局限性，例如英语自动句法分析的正确率就难以达到 100%。至于更深层次上的语言研究，例如语义层次上的研究在目前的条件下使用计算机尚有一定困难，但随着人们对语言本质的认识逐步深入及计算机技术的进展，计算机在语言研究中的作用会进一步深化与扩大。

鉴于计算机在语言材料的收集、存储、检索、分类等方面的作用，它已成了语言学家的强有力的工具，并将在语言研究中发挥日益重大的作用。

1.5 计算机与语言教学

当代科学技术发展日新月异，推动着社会生产力的发展，同时也深刻影响社会生活的各个方面。由于社会进步和技术发展，知识更新的速度和社会信息量都在加速增长。有人测算，进入 80 年代后，人类知识量每三年就会翻一翻。理工科大学毕业生要适应社会需要，就要不断及时更新自己的知识。在这种咄咄逼人的形势下，教育事业面临重大挑战。旧的教育体制已不能完全满足社会需要。

从个人角度看，现代化生产要求生产者接受终身教育，受教育不再限于青少年时期。因为不这样，就无法适应由于科学技术发展而对劳动者所提出的新要求。从社会角度看，现代生产要求扩大教育面。随着生产发展，单纯体力型的工作比重越来越小，各种职业岗位都要求越来越高的智力和越来越丰富的知识。

社会进步对教育事业的挑战要求教育发生三个战略转变:(1)由"封闭型"学校向"开放型"学校转变。受教育人数的增加和受教育年限的延长使传统的"封闭型"学校无论在设备或师资上均无法满足社会需要。电视、无线电、录音机、教育卫星等现代教育技术使这一转变得以实现。现在一节课听课学生已可以不再受教室大小的限制,通过电波、磁带、录像带等使成千上万个学生都能听课,接受教育。(2)从课堂讲授向个别讲授转变。个别讲授将有可能真正做到因材施教,极大提高教学效率。(3)从学校教学向个人自学转变。学生的学习可不受时间、地点、师资等因素的制约,学生的学习积极性得到充分的发挥。

向"开放型"学校的过渡将使受教育人数大幅度增加,在量上满足教育发展的需要;而向个别讲授及个人自学的过渡将会导至教育上质的飞跃,使教学效率和培养速度显著提高。后两个转变的技术基础就是现代电子计算机。计算机的高速运算能力和巨大存储能力,使教师的部分工作由计算机执行成为可能。一个学生面对一台微机或一台终端,使用计算机辅助教学软件,就好像面对一位"私人教师"学习。这种机助教学可以在学生自己选定的时间、地点,在计算机课件的指导下学习自己选定的科目。当然,后两个转变是一个长期的过程,比向"开放式"学校过渡要复杂得多,艰巨得多。这不仅因为有些技术问题尚未解决,也因为有些机助教学规律有待探索。可喜的是近年来计算机辅助教学,包括计算机辅助语言学习已有重大进展。

那么计算机可以进入语言教学的哪些环节呢?我们观察一个称职的语言教师的全部教学活动,可以发现他们是:(1)教学计划的制订者:根据教学大纲和教材编写教案,计划好各