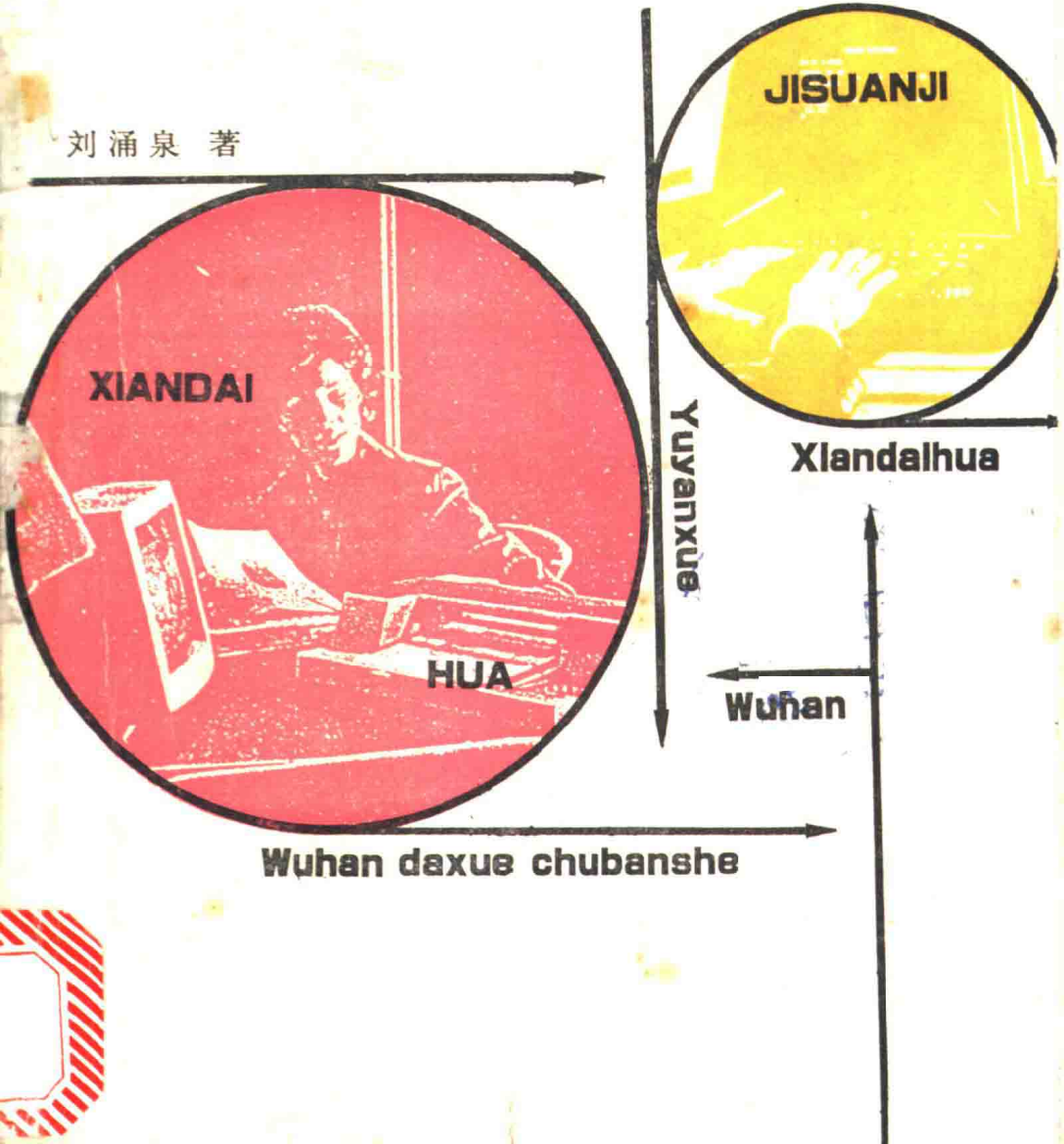
语言学现代化和计算机

刘涌泉 著



语言学现代化和计算机
刘涌泉　著
武汉大学出版社出版
（武昌珞珈山）
新华书店湖北发行所发行　武汉大学印刷厂印刷
*
850×1168 毫米　1/32　7.875印张　193千字
1986年8月第一版　1986年8月第一次印刷
印数：1—5,000
统一书号：9279·10　定价：精2.75元
平1.85元

目　录

前　言

这里汇集的论文绝大多数曾在国内外不同刊物或会议录上发表过。其中《俄汉机器翻译中的词序问题及其解决办法》(1959)是机器翻译界第一篇较为深入地研究词序问题的文章，为不同类型语言之间的翻译确定了有效的解决途径。《机器翻译和文字改革》(1963)从语言信息处理角度阐述了文字改革的必要性，并且比较早地探讨了汉字编码、词汇切分等问题。

整个论文集可以分作下面这样几个部分：(1)总体性文章——论述了语言学的发展阶段、应用语言学的重大意义、计算语言学的兴起、语言学必须现代化、语言工程的开发等问题。(2)机器翻译——介绍了机器翻译发展概况和前景，并着重阐述了我国机器翻译研究的方法和成就。(3)中文信息处理——论述了中文信息处理的定义、范围和原理，并且主要讨论了汉字信息处理系统各方面的问题。同时还以相当的篇幅阐述了中文信息处理和信息化社会的关系。另外，还提出了建立最佳的现代化语言信息系统的设想(这些设想，实际上也就是第五代或新一代计算机所要解决的课题)。最后，郑重指出：汉字进入计算机不是完事大吉；必须重视“词”在中文信息系统中的作用，并解决好词的切分问题，否则，词的问题将成为中文信息处理前进道路上的拦路虎。(4)文字改革——这是我国语言规划(Language Planning)中的重要课题。这里阐述了汉字改革和科技革命的关系，并且着重讨论了文字改革对中文信息处理各方面的重要意义。(5)术语工作——语言规划中的另一项重要课题。术语是科技语言的基本要

素，是科技交流中最基本的载信者。术语工作搞得不好，势必影响科技革命。为此，提出了术语工作现代化的问题以及解决这一问题的方法和措施。

读者不难看出，有好几篇文章原来是国际会议上的报告。为了向不同会议的代表介绍我国情况，因而其中某些内容有重复。个别重复较多的地方，收入本集时作了删节。

最后，还想着重指出，语言学的现代化问题已迫在眉睫，而且解决这一问题的条件已经成熟（微处理机的迅猛发展大大加快了语言学现代化的进程）。我们已经落后了一大截，如不尽快赶上，什么情报工作自动化、印刷排版现代化、办公室自动化……都将落为空谈，因为后面这几个“化”都同语言学现代化有直接关系。因此说，语言学的现代化不是一件小事，而是关系我国四个现代化进展快慢的大问题，关系到信息化社会在我国何时成为现实的大问题。作者的知识和能力有限，但是愿为这一美好的事业贡献一点力量。不足之处，请读者批评指正。

刘涌泉

1985年10月于北京

I

语言学必须现代化*

——电子计算机和语言学

准备谈三个问题：一、语言学为什么要现代化？计算机向语言学提出了哪些要求？二、语言学能否现代化？计算机给语言学提供了哪些条件？三、语言学如何现代化？同计算机结合，要做哪些准备？

一、语言学为什么要现代化？

1. 语言在发展，语言的应用在扩大，语言学也必须相应发展。

语言是人类最重要的交际工具。人们在运用语言的过程中不断改善和扩大它的交际职能。这样也就促成了语言的变化和发展。

语言的发展又总是与社会的发展密切相关的。最初的语言，只有口语形式，后来出现了文字，语言的应用扩大了，从时间上说，它能流传到后代，从空间上说，它能流传到远方。后来又出现了印刷术，语言的应用得到进一步扩大，不仅能流传到后代和远方，而且能比较迅速地大量复制，广为传播。电话和录音机的

---

* 原载《中国语文》1978年4期

出现，是交际手段的又一次飞跃，从此语言不但可以保留其书面形式，而且又可以保留其语音形式，迅速传递到远方。电视、传真复印、模式识别、各种语音分析合成仪器、电子计算机等先进技术和设备的出现，更把语言的应用提高到一个新水平。

语言应用范围每扩大一步，都会出现一些新问题。为了解决这些新问题，又必须时常采用一些新的方法和新的工具，而新方法和新工具的应用也会促进语言学的发展。比如，有了电话以后，语言学中就出现了清晰度问题的研究和言语压缩问题的研究，前者的目的在于保证通讯质量，后者则是为了提高通讯线路的经济性。解决这样的问题，语音学先前的那种描写（例如，发a时口张开，舌头放平，气流通过口腔不受任何阻碍……）已经不够用了。为了提供语音的各种物理参数，以满足工程技术方面的需要，开展语音实验研究的任务便提上了日程。又如，为了解决科技情报工作现代化问题，语言学中出现了机器翻译这样一个新课题，而为了解决这个新课题，必须对语言进行精密描写，要求语言学开辟一个新的战场，即与机器打交道的战场。

语言学的发展是时代的要求。我们现今处在一个科学技术大发展的时代，如今已不再是利用马和鸽子等动物来传递消息的时代，也不再是仅用笔和墨水书写的时代。如今通过卫星，世界上任何角落都能快速传递信息（很多国家之间已建立情报检索网，几分钟就可到别国查完一个课题）；如今语言要命令机器进行工作（掌握一百多个语词的机器人已有不少，现正研究能理解言语的机器人）；如今深入海底、高入太空都要求进行快速而有效的通信；如今语言逐步深入人机对话的交际过程，成了人机对话的基础。……时代不同了，语言学显然不能停顿在熟知的研究课题上，也不能局限在原有的研究范围内。语言学必须随着语言应用范围的扩大而相应发展，因而必须现代化。

2. 电子计算机的出现加速了语言学现代化的进程。

电子计算机的出现，引起了科学技术的巨大变化，同时也给语言学展示了新的发展前景。计算机这种神奇的机器不仅能以惊人的速度完成复杂的数值运算，也能高速度地完成复杂的逻辑运算。由于计算机具有逻辑运算功能，人们便越来越多地用它来代替人的脑力劳动，进行语言的各种自动加工。到目前为止，人们给计算机提出的语言自动加工的任务计有：(1)编辑各类索引，(2)编词表（包括逆序词表），(3)统计语音、词或词组的频率，(4)检索文献，(5)查找数据（目前国际上已建立各种数据库一百余个），(6)进行程序教学，(7)根据一定规则进行语言分析，(8)绘制方言地图，(9)自动翻译，(10)识别语音，(11)识别文字，(12)进行言语合成，(13)自动标目（文献），(14)自动作文摘，(15)自动判断和证明。①

为了完成这些任务，计算机向语言学提出了两个要求。第一，要求语言学家武装它的“头脑”以发展它的智力。例如，人们给计算机提出做翻译的任务，计算机便要求语言学家赋予它翻译外语的能力。这时，语言学家就要编出适合它使用的词典和语法。这些词典和语法一旦存入它的“头脑”（存储器）之中，它就变成一个翻译员了。它的翻译本领大小决定于给它的词典和语法的好坏。第二，要求给它增添“翅膀”以赋予它听觉（识别口语）和更强的视觉（识别文字），赋予它说话能力（合成言语）和听写能力（语音打字）。这样，语言学家又需要提供各种物理参数和语言概率性等方面的数据，以便同有关的专家、工程师一道共同解决语音、文字的输入输出问题。

计算机的这两个要求都是围绕着发展人工智能这个总任务而提出的。正如人的智力总是借助语言来表达一样，计算机的人工智能也都是以语言为基础的。离开了语言或它的变体，人工智能也就无从谈起。研究人工智能，用计算机的术语来说，就是给计算机提供各种软件（程序系统）。

总之，语言学的现代化，无论从哪一个角度来说，都是势在必行。不现代化，就要影响其他方面的发展，就要影响科学技术规划中一系列重要课题（情报工作自动化、通讯线路最佳化、语音控制自动化、汉字信息处理自动化等）的顺利完成。这样，也势必要影响到四个现代化的进程。

二、语言学能否现代化

答案是肯定的。首先，从语言谈起。语言不是杂乱无章的，也不是不可捉摸的，语言是一个具有形、音、义的符号系统。这个系统中的成分（词）以及这些成分的组合（句）都有其物质基础——语音以及反映在书面上的书写符号。由于有物质作基础，并有一定规律可循，因而就完全有可能设计一套程序让机器进行各种加工。换句话说，语言的符号性、规律性和可译性都是语言学能够现代化的前提条件。

另外，语言学本身的发展也提供了这种可能性。语言学经过传统语言学、历史语言学、描写语言学、结构语言学、形式语言学几个阶段，现在又进入计算语言学阶段。近二三十年，语言学的发展上有一个明显的特点，这就是日益向精密科学靠拢。语言学过去只是同文学、人类学、历史学、考古学、哲学和文化史等有联系，而今天除了这些之外，它同数学、生理学、物理学、电子学、信息论、控制论、符号学、计算技术、通讯技术、自动化技术等学科建立了越来越密切的关系。客观现实要求语言学同这些学科共同解决一些边缘性课题。在这种联系和合作中，语言学从其他学科引进了不少新的概念、方法和工具。而这些新概念、新方法、新工具的引进，大大有利于语言学的现代化，使语言科学中出现了许多新名目：数理语言学、统计语言学、代数语言学、算法语言学、计算语言学、工程语言学、机器语言学，等等。这些新名目反映了现代化进程中的繁荣景象，同时也标志着语言学

更加成熟、更加深化、更加现代化。

我们再从技术上来看一看语言学能否现代化。语言的自动加工，不同于数值运算的地方主要有以下两点：

1. 一般来说，同数值运算相比，语言自动加工的程序复杂，数据也多。这样，便要求存储器（特别是内存）的容量大。

2. 语言自动加工时输入输出的数据多，因此，输入输出装置要更为先进，才能满足语言加工的需要。

现代的计算机能不能满足语言自动加工的需要？它能给语言学的现代化提供哪些条件呢？

1. 速度：每秒能运算几百万次，几千万次，甚至一亿五千万次。

2. （内）存储量：2048兆字节。正在研制的激光存储器具有更大的发展前途。

3. 输入设备：一种字体的英文光学自动阅读器国外已研制成功，每秒能识别2000—3000字符；多种字体的目前各国正在大力研制。机械式的汉字输入装置已有数种，光电式的尚处于研制阶段[2]。

4. 输出设备：外文的和汉字的高速输出问题已得到解决，如有的汉字信息处理机每秒已可印出1200汉字。

如果拿这些条件同1959年我们进行俄汉机器翻译试验时的条件比较一下，就会看出，目前的条件已经好得多了：那时比较大、比较快的104机每秒运算速度仅一万次，比现在最快的计算机慢一万多倍，内存仅有2048字，比现在的大存储器小二十万倍。适合文字加工的输入输出设备那时根本没有。

上面主要谈的是硬件的情况，至于软件方面这些年来也有很大进展，程序设计语言不断增多和改进，越来越有利于语言的自动加工。由此可见，语言学现代化所需要的技术条件也已基本具备，并且日益完善。

三、语言学如何现代化

语言学同计算机结合，首要的一条就是要求语言形式化，因为只有形式化，才能算法化、自动化。我们所说的形式，既包括语言单位本身的形式，又包括它的位置和结构关系。所谓形式化，就是一切从形式出发，而不是从意义出发。要知道，机器只认识形式，不懂意义。当然，这绝对不是说要完全抛弃意义。人在分析语言时总是要参考意义的，因为在许多场合完全靠形式行不通。但是，可以供机器依据的绝不是纯粹的意义，而是形式化了的意义。例如，某些词之间具有语义搭配关系，这时我们只要给它们分好类，在分辨它们时只要依靠它们的类号（语义特征组别）就可以了。③

语言的形式化是分层次进行的。语法的形式化相对来说比较简单，人们已做了一定工作，而语义的形式化则是一个复杂的问题，人们进行的工作还不多。语义形式化问题解决得好坏，将大大影响语言自动加工的成效。④

由此可见，单就语言形式化这一点来考虑，语言学现代化实现起来就有不少困难。下面结合着这个问题简单地谈一谈为了使用计算机完成语言学现代化，究竟该做哪些准备工作。

1. 思想上的准备：对语言和语言学研究方法要有一个正确认识。这方面的问题很多，这里只谈两点：比如“语言是人类最重要的交际工具”，这的确是最概括的最正确的定义。但是不能因此而排斥或批判那些从某种角度给语言所下的定义，例如把语言看作是一个符号系统（结构学派和符号学的定义）、一个集合（数理语言学）、一个代码（信息论）等等。其实这些定义不但同那个最概括的定义丝毫没有矛盾，相反倒是对它的一些具体补充和说明。又比如我们现在强调同计算机结合，就把形式化的分析方法说成是唯一的万能的方法，从而排斥其他一切方法，这就

不是正确态度。反之，如果给形式化的分析方法扣上形式主义帽子而加以否定，也不是正确态度。

这类问题需要得到正确解决，否则语言学现代化的实现就要遇到障碍。

2. 语言研究本身的准备：语言研究工作是一切语言自动加工的基础，没有一个坚实的基础，各项任务就要落空。语言是一个相当复杂的系统，要对它进行科学的、精密的、形式化的描写，的确不是一件轻而易举的事。面向人的语言学已进行了长时期的研究，至今尚存在不少问题。面向机器的语言学近二三十年才发展起来，现在水平还相当低，这是不难理解的。现在有很多课题（例如机器翻译，言语分析和合成还）未得到解决，究其原因，主要不是由于技术条件不够，而是由于语言研究不够。说得明确一些，就是语言学未搞上去而拖了后腿。

目前机器翻译研究已处于实际应用的前夕，世界上已存在十来个初步应用的机器翻译系统，据学者们估计，到1980—1990年，机器翻译的产品将流通于世。在朝这个目标迈进的过程中，一个国家语言研究的基础如何，目前投入的力量如何，同到达这个目标的时间有很大关系。为此，我们有必要大大加强机器语言学的研究。

3. 技术方面的准备：上面我们谈到，技术条件已基本具备并日益完善，这是指整个世界而言的。具体到我们国家，还有一些特殊情况。比如说，汉字信息处理问题还是一个亟待解决的问题。各种编码系统有待统一和完善，各种机械式输入装置尚须改进，光电式输入装置正在研制。这个问题不解决，有关汉字自动加工的问题，诸如自动排版、编索引、编词表、搞统计、查文献以及其他语言研究自动化课题，都会遇到很大障碍。拿词汇统计来说，英语词汇统计调查比汉语词汇统计要省事得多，输入输出可以利用一般的计算机外围设备。⑤而汉语词汇统计，如利用一

般的计算机外围设备输入输出，则首先需要把汉字转换成电报码或拼音字母，同时还要解决词的界限问题以及同音问题（如转换成拼音的话），加工完了输出之后还要再转换成汉字。

为了创造更有利的技术条件，语言学家应该更多地关心和参与汉字信息处理等工作。

4．干部队伍的准备：从上面的叙述中，不难看出，建立机器语言学不是一个简单的任务。它要求有一支队伍，而且是一支既懂语言又掌握数学、物理学等自然科学知识的队伍。为此，有必要在有条件的高等学校设立数理语言学或应用语言学专业，一方面学习语言学方面的有关课程，另一方面学习数学、物理学等方面的有关课程。争取早日培养出大量全面发展的机器语言学人材，已经是刻不容缓的任务了。

最后，有必要再强调一下：在科学技术日益成为巨大生产力的年代，作为人机对话基础的语言必将在不断扩大的人机对话领域内发挥越来越大的作用；同样，与现代科学技术关系极为密切的机器语言学（如果不把它算作自然科学一支的话），也必将随着现代科学技术的发展而在国防建设和整个国民经济建设中占有越来越显著的地位。这就是语言学现代化的意义所在，也是语言学现代化和四个现代化的关系所在。

附　注

① 1—8已实际应用或基本上能实际应用，9—12仍处于试验研究阶段，13—15可以说尚处于探索阶段。

② 光电式输入装置目前还没有真正过关，因此输入问题在语言自动加工的整个过程中还是一个薄弱环节。但是，对于英汉机器翻译这样的课题来说，由于可以利用近年来国外发行的磁带形式的书本作为直接输入资料，因而这个问题也算部分解决了。

③ 例如，处于主语结构中的 all, every 这类英语词，翻译成汉语

时谓语前应加“都”字，如 All reactionaries are paper tigers（一切反动派都是纸老虎）。机器只须依靠 all 的语义特征，便可在谓语前自动加上“都”字。

④ 这个问题在机器翻译研究史上反映得非常清楚：人们通常把词对词的翻译系统称为第一代机器翻译，把具有语法（尤其是句法）分析能力的称作第二代机器翻译，把具有语义分析能力的系统称为第三代机器翻译。

⑤ 例如，我们利用中国科学院计算研究所同志编的程序在通用计算机上对英语冶金文献的词汇进行过统计，除穿孔打字输入占用时间较多以外，二十多分钟便可打印出几大本数据：1，按字母顺序编排的词表，2，按频率高低编排的词表，3，按字母个数编排的词表，4，带有每个词的出处的词表，5，字母顺序逆排的词表。每个词表同时能告诉我们词的总数。

Ⅱ

语言学的新发展*

我想谈三个问题：1)从历史上看语言学的发展，2)电子计算机的出现对语言学发展的影响，3)语言学的新发展是时代的要求。主要谈第二个问题。

一、从历史上看语言学的发展

人们常常把语言的研究分为两个时期：一是科学以前的时期，从古代起到十八世纪，二是科学时期，从十九世纪起到现在。仔细考察一下十九世纪到现在这一段历史，不难看出，有三方面的因素对语言学的发展有重要作用。

1. 科学上的重要发现、发明，以及各种学说和社会思潮对语言学的影响：

（1）历史比较语言学的产生，跟整个社会科学从十八世纪抽象的唯理主义之转向历史主义有联系。历史比较语言学的产生和发展使语言的研究转入了历史的科学的轨道。

（2）历史比较语言学史中起承前启后作用的自然主义学派，更为突出地说明当时自然科学（达尔文学说）对语言学的影响。他们把语言看作有机体，并把植物中的谱系原理应用到语言

* 本文原是中国语言学会成立大会上的发言。后来收入《把我国语言科学推向前进》，湖北人民出版社，1981。

学中，建立了所谓语言“系谱树”。

（3） 马尔学派是本世纪二十年代末在苏联语言学界出现的一个以庸俗社会学观点解释语言现象的学派。拿马克思主义关于社会关于革命的学说硬往语言上套，提出语言是上层建筑，有阶级性等谬论，致使苏联语言学一度步入歧途。

（4） 我国左倾路线的影响，尤其是“四人帮”时期的种种谬论和破坏行为给我国语言学的发展带来了严重危害。总起来说，一句话，不少作品的内容空的多，实的少。

在这里还可顺便提一下：语言学不受重视的问题长期得不到解决，也是同社会思潮有关的。但是，语言学是基础科学之一，这一点不仅对人来说如此，就是对机器的人工智能来说，也是如此。不认识这一点，迟早要吃大亏的。

2. 社会需要是语言学发展的巨大推动力：

（1） 美国描写语言学主要是在调查无文字语言的基础上形成的。据说美洲印第安语年年总有一种或几种语言消亡，为了抢时间收集这些语言的资料，美国人类学家和语言学家发起了语言调查。他们在调查研究美洲印第安人的语言时，发现印欧语传统的研究方法完全不适用，因而发展了以语言外部形式特征为重点的结构主义描写原则。

（2） 通讯科学的发展引起了实验语音学的巨大发展。比如，有了电话以后，语言学中就出现了清晰度问题的研究和言语压缩问题的研究，前者的目的在于保证通讯质量，后者则是为了提高通讯线路的经济性。解决这样的问题，语言学先前的那种描写（例如，发a时口张开，舌头放平，气流通过口腔不受任何阻碍……）已经不够用了。为了提供语言的各种物理参数，以满足工程技术方面的需要，语音实验研究得到了迅速发展。

（3） 机器翻译的产生和复兴更有力地证明社会需要是科学发展的巨大推动力。克服语言障碍，是一个老问题。人们曾想通