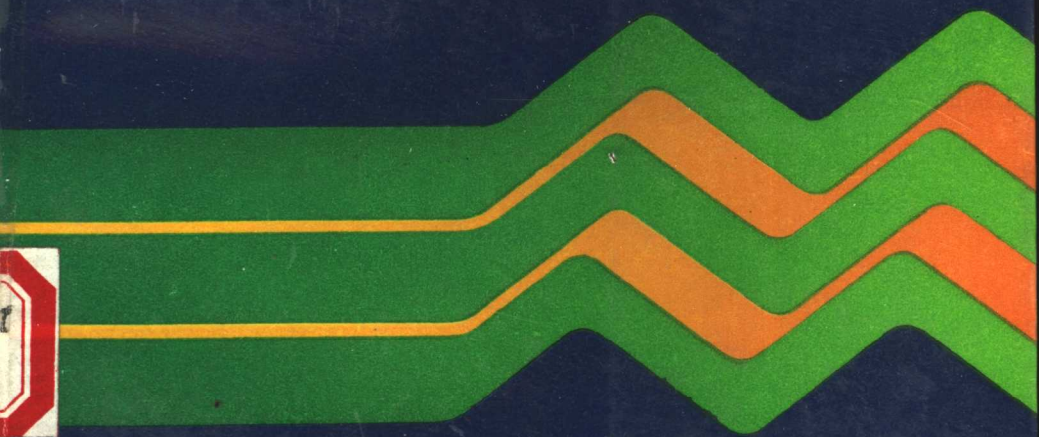
计算语言学导论

陆致极



责任编辑　唐发铙
徐川山
封面设计　王建纲

计算语言学导论
陆致极
上海教育出版社出版发行
（上海永福路123号）
各地新华书店经销　商务印书馆上海印刷厂印刷
开本 850×1156　1/32　印张 16.375　插页 4　字数 356,000
1990年12月第1版　1990年12月第1次印刷
印数 1—1,400 本
ISBN 7-5320-1681-1/G·1636　定价：（软精）6.20 元

读《计算语言学导论》后的随想

——代　序

读书界很需要一部探讨语言科学和信息科学相结合的边缘学科的通俗著作；现在付印的这部《计算语言学导论》在某种程度上(或者说在某一方面，例如在利用计算机对自然语言进行处理方面)满足了这一需要。我认为，这样一部通俗著作应当具备下面的特点：

——它是这门交叉学科的通俗介绍，同时带有适当的学术性或理论性，即它本身不是简单的浅近，解释，而是有一定观点，对所论述的问题有比较深刻的分析；

——它不仅仅作理论上的介绍，而且能提供这种理论在实际上的应用，举出的应用模式带有一定的普遍性和示范性，以便读者可以举一反三，做到理论与实践的结合；

——它的论述应当能够启发读者思考它所提出的论点或实施方案，并且引导读者不迷信书中所揭出的论点，以便除了加深或改变自己原来的认识之外，还能作出自己的独立思考；

——最后的一点也许是最重要的一点：它应当从汉语出发，特别从现代汉语出发，来阐明这个学科的基本理论以及应用，而不是照搬外国书上所列举的外国语言现象作为它所论述的基础。

我读过这部著作得到的印象，是作者力图从以上所表述几点来研讨这门学科的内涵，也许某些论点阐述得不够充分，也许若干论点为已经深造的读者所不能完全同意，但因为作者的论述是认真的，同时也是比较严谨的，所有科学论证都是在取得数据以后提出的，因此我认为这部书在一定程度上满足读书界这方面的需要，是一部值得一读的书。对于文科特别是其中研习语言学的读者，同时对于理工科特别是其中有志于将电子计算机应用到其他学科的读者，都是有益的。

不难看出，作者在这部著作中悄悄地把重点放在上中下三篇里的《中篇：结构篇》，这部分占全书正文的百分之六十（准确地说为287页即58.8%）。开篇前引用了控制论创始人美国学者维纳的一句名言："语言不是生命体所独具的属性，而是生命体和人造机器在一定程度上可以共有的东西。"这句话出自《人有人的用途》第四章《语言的机制和历史》。维纳博士在这里明确提出了语言不仅仅是人与人之间交流信息的手段，而且是"人向机器、机器向人以及机器向机器"交流信息（或者如维纳在他第一部著作《控制论》中所用的术语"通讯"）的手段。这句引言预示着本书作者企图通过这部通俗著作，向读者展示我们这个时代的尖端课题，即人机对话以及人工智能的图景。为此，作者首先从这个出发点介绍了近30年来现代语言学的发展概略，特别是以乔姆斯基为代表的语言学思潮的演变历程——作者对语义理论发展的概括，应当说是简明扼要的，他写道：

"纵观乔姆斯基语义理论的发展历程，不难看到这样的认识发展线索：从初期把语义排斥在语法体系之外，到标准理论时期把语义全盘纳入语法体系，接着长期在句法和语义关系的处理

上下功夫。到70年代后期，在认识到语言意义的复杂性的基础上，始把意义部分地置于语法之外，由此更集中地对语法所能提供的语义信息作深入的研究。”

也许可以认为，语义研究是当代语言学研究过程中的一个薄弱环节；但是如果没有对语义透彻的理解，则对自然语言的理解将不能得到预期的结果。幸而现代语言学的发展及时地、部分地消除了这种缺陷，为自然语言的计算机处理进入一个新领域（语义分析的领域）准备了必要的条件。作者在介绍了当代几种人机对话系统之后，饶有兴味地导入了汉语最简单的人机对话系统雏形(439—440页)——按照这个程序，“对话者”真有一点儿“智能”(“人工智能”)。“当输入一个陈述句时，它在‘初步理解’的基础上，能够‘记忆’句子的基本内容。当向它提问时，它会根据已经‘学习’到的知识，作出判断，然后予以回答。自然，因为输入的程序比较简单，词汇量很小，句子结构也很有限，所以当向机器提问超出了它所能理解的范围时，它只能十分抱歉地说：‘太难，我不懂。’”本书这一部分的描写、分析和程序设计是颇为引人入胜的，它为读者打开了人工智能这个时代话题的前景。

人工智能不是本书的主题，但是本书提供了研究人工智能这个尖端学科的基础知识。人工智能的第一次学术会议是1956年在美国召开的，尽管30多年来这方面的探讨已经有长足的发展，但它至今还没有能够达到理想的高度，还有很多问题需要解决——在这当中，对于语言学研究者来说，则是自然语言内在的歧义性(或多义性)。无论在机器翻译，语言识别，人机对话以及高级人工智能系统中，这个问题是一个很伤脑筋，有时甚至令人沮丧的问题。这一点人工智能的哲学理论家 H. L.德雷福斯在

他的《计算机不能做什么》或《人工智能的极限》一书里慨叹过：在人工智能发展最初阶段，由于这歧义性或多义性，需要作处理（计算机处理）的语言成分“也都含混不清”。

我以为直到现今，自然语言歧义性这个难题，还困扰着研究者。自然本书对此并无深入论述，但本书有关自动翻译、语言识别和人机对话的分析表述中，读者将会感到这个难题。

此外，本书上篇阐述了如何把电子计算机应用于词典学、风格学、方言学、历史音韵学这些方面，有理论分析也有实际的程序设计，应当受到读者的欢迎。至于本书下篇讨论机器在狭义的应用语言学（语言习得学）的作用，篇幅不多，论题没有充分展开。

作者虽是文科出身，但他在自己应用计算机方面的实际经验这个基础上，写出这部学术性的通俗著作，无疑将对文科或理工科读者同样有启发的。

陈原

一九九〇年三月

前　言

计算语言学是使用计算机对自然语言进行处理的一门新兴的边缘性学科。近二十年来，它以自然语言理解为中心课题，已经成了现代科学中最活跃的研究领域之一。

本书是关于计算语言学的一本导论性著作。在本书写作之前，有关计算语言学的发展概况，已经有著作和文章作了介绍（例如冯志伟《数理语言学》、徐志敏《国外人工智能研究概况》等）。因此，本书的任务着重在于探讨如何具体从事这方面的研究，尤其是如何使用计算机对汉语进行处理的问题。

本书分三个部分。上篇计量篇，探讨在词典学、风格学、方言学和历史音韵学等领域内使用计算机进行数量统计和分析的理论和方法。中篇结构篇，集中探讨关于自然语言理解的理论和方法，同时详尽地讨论了用 LISP 和 Prolog 程序语言编写汉语句子分解程序以及人机对话系统的方法和技巧。下篇应用篇，介绍计算机在语言教学中的应用。

本书在写作过程中始终遵循着三个目标：理论性、实践性和探索性。计算语言学是一门正在成长的新学科。近年来发展很快，不断有新的理论和方法在实践中涌现出来。本书虽是导论性著作，但作者在阐述基本知识的同时，力求能触及和反映这一领域的最新理论水准。计算语言学是运用计算机对自然语言进行处理，要从事这方面研究离不开对计算机的了解和使用。由

于作者"来自"中文系，对文科师生初次接触计算机的心情深有体验，因此写作时特别注意实践性。本书介绍了 BASIC、LISP、Prolog 等程序语言的基本特点和功能，详细地讨论了三十余个具体程序的制作问题，而且，其中绝大多数的程序都可以在跟 IBM 微机通用的微型计算机上执行。一个没有计算机方面背景知识的读者，如果在耐心阅读本书和坚持上机实验之后，能独立地开展这方面的研究，那么，作者的目的也就达到了。从这个意义上来说，本书又是一本"实用"计算语言学著作。计算语言学的理论主要是在国外发展起来的。国外的研究自然大多以他们自己的母语为对象(如英语、俄语、日语等)。汉语跟这些语言有较大的差别。直接照搬他们的研究成果，显然无济于事。本书以汉语为主要研究对象，就不能不做探索的工作。本书可以说正是作者近年来在这方面学习和摸索的初步结果的汇集。由于涉及面颇广，而作者学识有限，疏漏和错误之处一定不少，敬祈广大读者和专家不吝赐教。

本书的写作得益于这几年作者在美国伊利诺大学语言学系的学习，特别是下列课程：郑锦全教授的"计算语言学导论"，郑锦全教授和 Jerry Morgan 教授的"计算语言学研究"，Jerry Morgan 教授的"计算机自然语言处理专题"，Erhard W. Hinrichs 助教授的"计算语言学专题"，Robert S. Hart 副教授的"计算机辅助外语教学"，以及 Michael Kenstowicz 教授的"数理语言学导论"。本书中有一些程序直接脱胎于这些课程的示例、作业、考试和学期论文。作者在此谨向这些老师们致以衷心的谢忱。本书的写作也得益于作者近三年来在伊大语言学习实验室承担助研的工作。伊大语言学系及语言学习实验室的一些同学和同事，热情地支持了本书的写作。其中周欣平阅读了本

书中篇部分的初稿，Stephen Helmreich、Atsushi Fukada、Kazumi Hatasa、Dale Gerdemann、Edward Kovach、刘显亲、宋丽梅等为作者提供了资料。此外，在国内的金立鑫同志多次为作者寄来新近出版的有关书籍，正在美国学习的陆丙甫、谢天蔚、吴道平诸友曾来电话关心本书的写作，作者在此一并致以诚挚的谢意。作者还要感谢陈原教授为本书作序。

本书实际写作始于今年五月。在边学习、边工作，有时还需"打工"的情况下，坚持写作，艰巨可想而知。半年来，每当遇到困难或感到十分疲惫的时候，我会想到在大洋彼岸的复旦大学的师友，想到上海现代语言学讨论会的朋友们。他们那种为祖国语言学事业献身的热诚，始终给我以巨大的鼓舞和鞭策。在此，我谨将这本不成熟的作品，献给我在远方的亲爱的老师和朋友们！

作　者

1988年11月24日，感恩节，于美国伊利诺大学，厄巴纳-香槟。

目　录

上篇　计　量　篇

中篇 结 构 篇

下篇　应　用　篇

本书主要计算机程序目录

第一章　导言：计算语言学

第一节　简略的回顾

一颗灿烂的新星已经在现代科学的舞台上升起。在它逼人的光芒中，可以看到将要到来的人类文明又一个新时代的曙光。它就是“计算语言学”(Computational Linguistics)。

计算语言学，顾名思义，就是语言学和计算机相结合的产儿。有着二千多年耕耘历史的语言研究是怎样跟现代科学技术的骄子——电子计算机结下不解之缘的呢？

当第一台电子计算机在 1946 年问世的时候，人们看到，由大量电子管、继电器连结起来的蛛网般的线路，居然能像无数神奇的手同时拨动成千上万个算盘一样，刹那间完成非常复杂的数值运算。惊叹之余，那些早就怀有使语言研究精确化愿望的学者，便想到让这最先进的计算机器来帮助进行语言文字的计量分析工作。比如，统计语言成分(音素、字母、词语)的出现频度；调查作家使用语言的风格特点；在历史比较语言学中，采用数理统计的方法去计算有亲属关系的语言之间的密切程度，等等。这些利用计算机在计量方面的探索和实践，直接孕育了这个戴着“计算”桂冠的边缘性学科。

就在第一台电子计算机诞生不久，英、美两位工程师便提出了用计算机进行语言自动翻译的想法。这两个工程师都是熟谙

破译电报密码的专家。他们想，不是能用类似译码的办法把一种语言翻译成另一种语言吗？所不同者，只是需要编制一部完善的双语言词典，这个想法立即引起了人们极大的兴趣。1952年，在美国麻省理工学院召开了第一次自动翻译会议。1954年初，美国乔治敦大学在国际商用机器公司的协同下，用IBM-701电子计算机进行了世界上第一次自动翻译试验。于是，开始了自动翻译蓬勃发展的十年。

自动翻译的一些初步成就造成了一种假象，一些急于求成的人误以为自动翻译马上要进入商业市场。然而，事情远非那么简单。美国资本家使自动翻译商业化的最初尝试失败后，美国工业界对自动翻译的热情一落千丈。1966年，美国科学院公布了一个题为《语言和机器》的报告（简称ALPAC报告）。这个报告对自动翻译工作采取了全盘否定的态度。自动翻译的上空一时乌云滚滚。不少国家，特别是一直走在自动翻译研究前列的美国，进入了自动翻译的萧条时期。

早期的自动翻译采用的是“词对词”的翻译方式。当时人们认为，好的译文可以通过分别处理词典和重新排列词序两步操作来得到。但这样的处理并不能达到预期的效果，译文质量低劣。因为原语和译语两种语言的差异，不仅只表现在词汇的差异上，还大量地表现在句法结构的不同上。要得到可读的译文，必须重视句法分析。于是，人们从第一代自动翻译系统失败中得到教训：语言理论的研究，特别是语言处理理论的研究，是自动翻译的基础，没有这方面研究的坚实基础，自动翻译就像是沙滩上的楼阁，会瞬息倒坍。

经过自动翻译大起大落的年月，计算语言学成长起来了。它开始从语言的计量方面的研究跨入了语言的结构的研究。