

# 《语言学中的数学方法》<sup>1</sup> 导读

冯志伟

## 一、 在语言学中使用数学方法的学术背景

法国数学家 J. Hadamard (阿达玛) 曾经说过: “语言学是数学和人文科学之间的桥梁”。Hadamard 不愧是一位具有远独特创见的学者, 他用自己的慧眼, 早就清楚地看出语言学在人文科学中是最容易与数学建立联系的学科。

然而, 在人类的科学发展历史上, 学者们是经过相当漫长的过程, 才逐渐认识到语言学 and 数学之间的这种亲密的关系的。

传统语言学的目的在于规定正确的读和写的种种规则, 这样的语言学有点像法律。历史语言学用谱系树的方法来表示不同语言之间的亲属关系, 这样的语言学一如生物学。结构语言学着力于研究语言的结构, 力图找出语言中各个要素之间的结构关系, 这样的语言学则酷似化学。那么, 语言学和数学究竟有什么关系呢?

语言学和数学都是有相当长历史的古老学科。语言学历来被看做典型的人文科学, 数学则被许多人看成是最重要的自然科学。在学校的教育中, 语文和数学被认为是两门最基础的学科, 成为了任何一个受教育者的必修课。它们似乎成了学校教育的两个极点: 一个极点是作为文科代表者的语文, 一个极点是作为理科代表者的数学, 在一般人看来, 语文和数学似乎是两个风马牛不相及的学科, 很少有人想到, 这两门表面上如此不同的学科之间竟然还存在着深刻的内在联系。

可是, 一些有远见卓识的学者却慧眼独具, 敏锐地看出了语言和数学之间的联系。

早在 19 世纪中叶, 就有人提出过用数学来研究语言现象的想法。例如, 1847 年, 俄国数学家 В. Я. Буляковский (Buljakovski, 布里亚柯夫斯基) 认为可以用概率论来进行语法、词源及语言历史比较的研究。1894 年, 瑞士语言学家 De Saussure (索绪尔) 指出, “在基本性质方面, 语言中的量和量之间的关系可以用数学公式有规律地表达出来”, 后来, 他在其名著《普通语言学教程》(1916 年) 中又指出, 语言学好比一个几何系统, “它可以归结为一些待证的定理”。1904 年, 波兰语言学家 Baudouin de Courtenay (博杜恩·德·古尔特内) 认为, 语言学家不仅应该掌握初等数学, 而且还要掌握高等数学。他表示坚信, 语言学将日益接近精密科学, 语言学将根据数学的模式, 一方面 “更多地扩展量的概念”, 一方面 “将发展新的演绎思想的方法”。1933 年, 美国语言学家 L. Bloomfield (布龙菲尔德) 提出了一个著名的论点: “数学不过是语言所能达到的最高境界”。

当时, 学者们不仅提出了这些想法, 还有人用数学方法对语言进行了实际的研究。1851 年, 英国数学家 A. De Morgan (德摩根) 曾把词长作为文章风格的一个特征进行过统计研究。1867 年, 苏格兰学者 L. Campbell (坎贝尔) 用统计方法来确定 Plato (柏拉图) 著作的执笔时期。1881 年, 德国学者 Dittinberger (迪丁贝尔格) 进一步用统计方法把 Plato 著作的执笔时期分为前期、中期和后期三个阶段。1887 年, 美国学者 C. Mendenhall (门登霍尔) 对不同时期的英国文学作品进行过统计分析, 特别是研究了 Shakespeare (莎士比亚) 的作品。1898 年, 德国学者 F. W. Kaeding (凯定) 编制了世界上第一部频度词典《德语频度词典》,

---

<sup>1</sup> Mathematical Methods in Linguistics, 《语言学中的数学方法》, 世界图书出版公司, 2009 年 3 月出版, 书号: ISBN 978-7-5602-9287-0/H•1068

用以改进速记的方法。

1913年，俄国数学家А. А. Марков (Markov, 马尔可夫) 采用概率论方法研究过《欧根·奥涅金》中的俄语元音和辅音字母序列的生成问题，提出可马尔可夫随机过程论，后来成了数学一个独立的分支，对现代数学产生了深远的影响。语言结构中所蕴藏的数学规律，成了Марков创造性思想的源泉。1935年，美国语文学家G. F. Zipf (齐夫) 发表了Zipf定律 (Zipf's law)，用数学方法描述频度词典中单词的序号分布规律。同年，加拿大学者E. Vardar Bake (贝克) 提出了词的分布率的概念，认为词典在选词时，应当以分布率为主要标准，频度为辅助标准。1941年，英国数学家G. U. Yule (尤勒) 发表了《文学词语的统计分布》一书，大规模地使用概率和统计方法来研究语言。

然而，不论是Вуляковскый, Saussure, Baudouin 和 Bloomfield 的想法和信念也好，还是Марков等学者的实际研究也好，都没有对语言学本身发生显著的影响。这是由当时的社会实践的要求所决定的，因为当时的语言学，主要是为语言教学、文献翻译、文学创作和社会历史研究服务的，在这样的社会实践要求下，语言学还没有很大的必要与数学建立直接的联系。就是像Марков从语言符号序列的观察和分析中发现随机过程的卓越成就，在语言学界也鲜为人知。语言学仍然沿着自己传统的道路，孤立于数学之外，迟缓地发展着。

与此同时，有一些杰出的学者开始从计算机和通讯的角度来关注语言问题，取得了突破性的成就。

在计算机出现以前，英国数学家A. M. Turing (图灵, 1912-1954) 就预见到未来的计算机将会对自然语言研究提出新的问题。

他在1950年发表的《机器能思维吗》一文中指出：“我们可以期待，总有一天机器会同人在一切的智能领域里竞争起来。但是，以哪一点作为竞争的出发点呢？这是一个很难决定的问题。许多人以为可以把下棋之类的极为抽象的活动作为最好的出发点，不过，我更倾向于支持另一种主张，这种主张认为，最好的出发点是制造出一种具有智能的、可用钱买到的机器，然后，教这种机器理解英语并且说英语。这个过程可以仿效小孩子说话的那种办法来进行。”<sup>2</sup>A. M. Turing 提出，检验计算机智能高低的最好办法是让计算机来讲英语和理解英语，他天才地预见到计算机和自然语言将会结下不解之缘，显示他不同凡响的洞察力。

在第二次世界大战刚结束时那个充满了理智的时代，计算机的研制正在紧锣密鼓地进行着，这时，有三项基础性的研究特别值得注意：

- 一项是A. M. Turing (图灵) 算法计算模型的研究，
- 第二项是N. Chomsky 形式语言理论的研究，
- 第三项是C. E. Shannon (香农) 概率和信息论模型的研究。

20世纪50年代提出的自动机理论来源于A. M. Turing 在1936年提出的算法计算模型，这种模型被认为是现代计算机科学的基础。

Turing 的工作首先导致了McCulloch-Pitts (麦克洛克-皮特) 的神经元 (neuron) 理论。一个简单的神经元模型就是一个计算的单元，它可以用命题逻辑来描述。接着，Turing 的工作导致了Kleene (克林) 关于有限自动机和正则表达式的研究，这些研究都与语言的形式化描述有密切关系。Turing 是一个数学家，他的算法计算模型是针对形式语言的，但与数学有着密切的关系。

1948年，C. E. Shannon 把离散马尔可夫过程的概率模型应用于描述语言的自动机。1956年，N. Chomsky (乔姆斯基) 从Shannon 的工作中吸取了有限状态马尔可夫过程的思想，

---

<sup>2</sup> A. M. Turing, Can A Machine Think?, Mind 50, 1950, 亦见 The World of Mathematics (edited by J. K. Newman, pp.2099)

首先把有限状态自动机作为一种工具来刻画语言的语法,并且把有限状态语言定义为由有限状态语法生成的语言。这些早期的研究工作产生了“形式语言理论”(formal language theory)这样的研究领域,采用代数和集合论把形式语言定义为符号的序列。Chomsky 在研究自然语言的时候首先提出了上下文无关语法(Context-Free Grammar),但是,Backus(巴库斯)和 Naur(瑙尔)等在描述 ALGOL 程序语言的工作中,分别于 1959 年和 1960 年也独立地发现了这种他们提出的巴库斯-瑙尔范式(Backus-Naur normal form)与 Chomsky 的上下文无关语法是等价的。这些研究把数学、计算机科学与语言学巧妙地结合起来,大大地促进了学者们采用数学方法来研究语言的数学面貌。

N. Chomsky 在他的研究中,把计算机程序设计语言与自然语言置于相同的平面上,用统一的观点进行研究和界说。

Chomsky 在《自然语言形式分析导论》一文中,从数学的角度给语言提出了新的定义,指出:“这个定义既适用于自然语言,又适用于逻辑和计算机程序设计理论中的人造语言”<sup>3</sup>。在《语法的形式特性》一文中,他专门用了一节的篇幅来论述程序设计语言,讨论了有关程序设计语言的编译程序问题,这些问题,是作为“组成成分结构的语法的形式研究”<sup>4</sup>,从数学的角度提出来,并从计算机科学理论的角度来探讨的。他在《上下文无关语言的代数理论》一文中提出:“我们这里要考虑的是各种生成句子的装置,它们又以各种各样的方式,同自然语言的语法和各种人造语言的语法二者都有着密切的联系。我们将把语言直接地看成在符号的某一有限集合  $V$  中的符号串的集合,而  $V$  就叫做该语言的词汇……,我们把语法看成是对程序设计语言的详细说明,而把符号串看成是程序。”<sup>5</sup>在这里乔姆斯基把自然语言和程序设计语言放在同一平面上,从数学和计算机科学的角度,用统一的观点来加以考察,对“语言”、“词汇”等语言学中的基本概念,获得了高度抽象化的认识。

这个时期的另外一项基础研究工作是用于语音和语言处理的概率算法的研制,这是 C. E. Shannon 的另一个贡献。Shannon 把通过诸如通信信道或声学语音这样的媒介传输语言的行为比喻为噪声信道(noisy channel)或者解码(decoding)。Shannon 还借用热力学的术语“熵”(entropy)作为测量信道的信息能力或者语言的信息量的一种方法,并且他采用手工方法来统计英语字母的概率,然后使用概率技术首次测定了英语字母的熵为 4.03 比特。

这些研究与数学和统计学有着密切的关系,属于信息论(information theory)的基础性研究。

Turing, Chomsky 和 Shannon 这三位学者的研究,为在语言学中采用数学方法提出了明确的思路,为语言学和数学的结合奠定了坚实的理论基础。

另一方面,社会实践的迫切要求进一步推动了语言学和数学的结合。

20 世纪以来,由于科学技术突飞猛进的发展,科技文献的数量与日俱增,世界各国每天出版的科技文献以数十万计,科技文献的这种增长情况被形容为“信息爆炸”(information explosion)。面对浩如烟海的科技文献,科技工作者为了了解外国的研究成果,取得科技情报,不得不花费大量的人力、物力来做难以数计的翻译工作,大大地影响了科研工作的效率。

1946 年,世界上第一台电子计算机研制成功,紧接着,在 20 世纪 50 年代初期,人们

---

<sup>3</sup> N. Chomsky and G. A. Miller, Introduction to the formal analysis of natural languages, Wiley, New York, 1963.

<sup>4</sup> N. Chomsky, Formal properties of grammar, Wiley, New York, 1963.

<sup>5</sup> N. Chomsky and M. P. Schützenberger, The algebraic theory of context-free grammars, in *Computer Programming and Formal Systems*, North-Holland Publishing Company, Amsterdam, 1963.

就开始考虑把这些工作交给电子计算机去做,利用电子计算机把一种形式的信息转换成另一种形式的信息,也就是将原始信息转换成结果信息,这就提出了机器翻译、机器自动作文摘以及机器自动检索科技文献等信息加工问题。

在用计算机将一种语言 A 翻译为另一种语言 B 时,除了确定语言 A 中的每一个词在语言 B 中相应的等价物之外,还必须分析语言 A 的句子结构和语义结构,并把翻译出来的词作某种变化,按照语言 B 的结构把它们配置起来,最后生成语言 B。这样,人们就得“教会”计算机自动地分析和生成句子。但是我们知道,任何一个问题要用计算机自动地来解决,首先就要使该问题所涉及的现象能够用数学语言来描述,也就是要把所考虑的问题“数学化”。所以,为了进行机器翻译,首先就要采用数学语言来描写语言现象,对传统语言学中的各种概念用数学的方法进行严格的分析,建立语言的数学模型 (mathematical model)。

用计算机自动做文摘和检索时,要求把科技文献的信息储存在计算机中,计算机按照人们的要求,在其所储存的信息的范围内,对人们提出的问题自动地进行文摘和检索。在计算机中用以储存信息的语言,在内容上应该是严格的、精确的,在形式上应该适于计算机储存形式的要求,这当然也要用精密的数学方法来加以描述。

由于自动化技术和计算技术的发展,人们正在迅速解决生产过程自动化问题,争取在不远的将来,用自然语言来进行“人机对话”(man-machine dialogue),让计算机能理解自然语言,这就要求将自然语言代码化,变为计算机所能理解的形式,自动地从自然语言的外部形态中,抽出它所表示的语义内容,并将计算机所理解到的语义内容,根据“人机对话”的要求,由计算机组织成相应的语句,回答人所提出的问题。

另外,由于通讯技术的发展,要求对负载信息的语言寻找最佳编码方法,要求提高信道的传输能力,以便在保持意义不变的前提下,最大限度地压缩所传输的文句,在单位时间内传输最多的信息,这就需要对语言的统计特性进行精密的研究。

语音识别和语音合成的研究也需要使用信号处理的技术,而信号处理技术离不开数学。

随着信息技术的进步和网络的发展,因特网(Internet)逐渐变成一个多语言的网络世界。目前,在因特网上除了使用英语之外,越来越多地使用汉语、西班牙语、德语、法语、日语、韩国语等英语之外的语言。从 2000 年到 2005 年,因特网上使用英语的人数仅仅增加了 126.9%,而在此期间,因特网上使用俄语的人数增加了 664.5%,使用葡萄牙语的人数增加了 327.3%,使用中文的人数增加了 309.6%,使用法语的人数增加了 235.9%。因特网上使用英语之外的其他语言的人数增加得越来越多,英语在因特网上独霸天下的局面已经打破,因特网确实已经变成了多语言的网络世界,因此,网络上的不同语言之间的自动翻译自然也就越来越迫切了。

在当今的信息时代,科学技术的发展日新月异,新的信息、新的知识如雨后春笋地不断增加,信息爆炸的情况更加严重。现在,世界上出版的科技刊物达 165000 种,平均每天有大约 2 万篇科技论文发表。专家估计,我们目前每天在因特网上传输的数据量之大,已经超过了整个 19 世纪的全部数据的总和;我们在新的 21 世纪所要处理的知识总量将要大大地超过我们在过去 2500 年历史长河中所积累起来的全部知识总量。随着知识突飞猛进的增长,翻译市场供不应求的局面也就越来越严重了。根据国际权威机构对于世界翻译市场的调查显示,全世界翻译市场的规模在 1999 年只是 104 亿美元,在 2003 年为 172 亿美元,而在 2005 年则达到了 227 亿美元。随着因特网应用范围的扩大和国际电子商务市场的日渐成熟,到 2007 年,只是网页的翻译业务将达到 17 亿美元的规模。目前,我国翻译能力严重不足,我国翻译市场的规模尽管已经超过了 100 亿人民币,但是现有的国内翻译公司只能消化 10% 左右,由于无法消化大量从国际上传来的信息流,我们的信息不灵,就有可能使我们在国际

竞争中失去大量的机会。在这样的情况下，机器翻译、信息检索、信息挖掘、自动文摘等自然语言处理（Natural Language Processing，简称 NLP）的研究显得更加迫切，自然语言处理成为了当代语言学中最引人注意的一个新兴学科。

美国计算机科学家 Bill Manaris（玛纳利斯）在 1999 年出版的《计算机进展》（Advanced in Computers）第 47 卷的《从人-机交互的角度看自然语言处理》一文中曾经给自然语言处理提出了如下的定义：

“自然语言处理可以定义为研究在人与人交际中以及在人与计算机交际中的语言问题的一门学科。自然语言处理要研制表示语言能力（linguistic competence）和语言应用（linguistic performance）的模型，建立计算框架来实现这样的语言模型，提出相应的方法来不断地完善这样的语言模型，根据这样的语言模型设计各种实用系统，并探讨这些实用系统的评测技术。”

这个定义比较全面，我们认同这样的定义。我们认为，计算机对自然语言的研究和处理，一般应经过如下四个方面的过程：

第一，把需要研究的问题在语言学上加以形式化，建立语言的形式化模型，使之能以一定的数学形式，严密而规整地表示出来；这个过程可以叫做“形式化”。

第二，把这种严密而规整的数学形式表示为算法，这个过程可以叫做“算法化”；

第三，根据算法编写计算机程序，使之在计算机上加以实现，建立各种实用的自然语言处理系统；这个过程可以叫做“程序化”。

第四，对于所建立的自然语言处理系统进行评测，使之不断地改进质量和性能，以满足用户的要求；这个过程可以叫做“实用化”。

显而易见，要为自然语言处理建立形式化、算法化、程序化、实用化的语言模型，都离不开数学，都必须使用数学方法来分析和描述语言，语言学与数学的结合已经到了迫在眉睫的地步了。

上面我们只是分析了语言学与数学结合的必要性，那么，语言学与数学的结合是否有可能呢？我们认为，不论从语言本身的性质来看，还是从当前科学技术发展的水平来看，都是有可能的。

语言本身的性质来看，正如 De Saussure（索绪尔）指出的，语言是一个符号系统，它可以同交通信号灯这样的符号系统相类比，只不过比交通信号灯复杂得多。每一种语言都是“能指”（即符号的物质表达）与“所指”（即概念或对象）的统一体，它为不同平面上的一定的结构规律制约着。音位学支配着语音的结合，形态学支配着构词和变词，句法学支配着词的组合。因此，我们在研究语言时，可以只管它的结构，至于这种语言是口说的或是手写的，还是用莫尔斯电码编了码的，对于研究者来说都是无关紧要的。这正如在下棋时，棋局的结构是重要的，而用木头的棋子或是用象牙的棋子则是无关紧要的一样。这样，我们就可以把语言看成是一个抽象的符号系统，这种抽象的符号系统，当然可以用数学来加以研究。

从科学技术当前的发展水平来看，也为用数学来研究语言提供了理论和方法。现代数学日新月异地发展，20 世纪以来迅速发展着的概率论、数理统计、信息论、集合论、数理逻辑、图论、格论和抽象代数等数学部门，为用数学思想和方法研究语言提供了有力的武器。

现代语言学也逐渐向精密化方向发展，在传统语言学内，出现了 O. Jespersen（叶斯泊森）的“分析句法”，在结构语言学内，L. Bloomfield（布龙菲尔德），Z. Harris（哈里斯）和 C. Hockett（霍凯特）等人提出了以替换和分布为手段，以辨别语素、分析层次为目标的一套严格的语言研究法。这些语言学派，在其语言观方面难免有片面之处，就是其具体方法本身，也有许多故弄玄虚、徒滋纷扰的地方。但是，由于采用了比过去的语言学更加严格的精密方法，在某些方面，对于用数学思想和数学方法来研究语言也有一定的启示作用。

另外，本世纪以来，控制论（cybernetics）逐渐成熟起来。控制论是研究机器与机器之

问、人与人之间、人与机器之间的信息的传输、接收、储存、加工和利用的一门综合学科，而语言是人类最重要的交际工具，是信息的最主要的负荷者，对语言进行精密的研究，有助于控制论的发展，而控制论采用的一些方法，特别是模拟方法，也可以作为建立语言模型的借鉴。

近年来，计算机科学发展迅速，语言学与计算机科学日益接近并互相渗透。计算机科学中使用的高级程序语言要尽量与人们的自然语言相接近，而其高级的程度，恰恰就是依这种程序语言与自然语言相接近的程度而定的，越接近自然语言就越高级，也就越便于人们掌握和使用。这样，计算机科学中对程序语言结构和编译技术的研究，就可以作为用数学方法研究自然语言的参考。

目前，人工智能已经成为国内外科技界十分关注的一个领域。自然语言是人类最重要的一种智能，人工智能所探讨的有关人类智能活动的一般规律，对使用数学方法来研究语言，有着一般性的指导作用。

因此，我们可以说，使用数学方法来研究语言，不但是必要的，而且也是可能的。生活在信息时代的语言学家，应当面对信息时代的需要，努力进行知识更新的再学习，改进自己的知识结构。本书恰好可以满足这样的要求。

我国学者早在 50 年代就关注到语言学中数学问题的研究。最早关注这个问题并且认识到其深远意义的人是 50 年代的大学生冯志伟。他于 1957 年在北京大学求学期间，就敏锐地注意到语言的数学描述问题，1959 年他毅然从理科转到中文系语言专业，试图用公理化的方法来研究语言的结构，他的研究生毕业论文就是“数学方法在语言学中的应用”。可惜他的研究很少人能够理解，一些人甚至认为他是离经叛道的古怪学生，他成为了一个孤掌难鸣、应者寥寥的孤独者，随着 1966 年的“文革”浩劫，他的研究生毕业论文被中途腰斩了，冯志伟也被迫改行，于 1968 年离开了北京大学。离开北京大学之后，由于资料缺乏，信息闭塞，他的研究条件变得异常地艰苦，但冯志伟仍然一如既往地坚持着他的探索，1982 年秋天，冯志伟应母校北京大学的邀请，在北京大学中文系汉语专业开设了“语言学中的数学问题”的选修课。这是国内首次在高等学校全面地、系统地讲述语言学中的数学问题的课程，受到学生们的欢迎。北京大学前任校长、著名数学家丁石孙教授在他的专著《数学与教育》一书中，对冯志伟的这门课程作了如下的评价：“1982 年，北京大学中文系开设了《语言学中的数学问题》，这是给汉语专业学生开的选修课程，许多同学对这门学科产生了很大的兴趣，经过一个学期的学习，同学们初步认识了现代数学的发展给语言学注入了生机，觉得获益匪浅，对语言学这门古老的学科分支的发展充满了信心，而且这一举动冲击了相当多的人的旧概念，使闭塞的中国学术界认识到，即使在人文科学教育中，数学也在逐渐起作用。”<sup>6</sup>在北京大学讲稿的基础之上，冯志伟写出了我国第一部系统地用数学方法研究语言的专著，书名叫做《数理语言学》，于 1985 年 8 月由上海知识出版社出版，全书分代数语言学、统计语言学和应用数理语言学三部分，简明扼要地论述数理语言学的基本原理和方法。他在 1991 年又出版了《数学与语言》，分七章论述了数学与语言的关系：语言符号的随机性与统计学，语言符号的离散性与集合论，语言符号的递归性与公理化方法，语言符号的层次性与图论，语言符号的非单元性与复杂特征的运算，语言符号的模糊性与模糊数学。著名数学家陈省身在扉页上为该书题词：“我们赞赏数学，我们需要数学”<sup>7</sup>。冯志伟的艰苦探索终于得到了我国学术界的认可。

现在我们进入了信息网络时代，冯志伟数十年前的对于语言学中数学方法的探索已经成为陈年旧事，依稀地存留在一些人的记忆中。随着我国自然语言处理研究的进一步发展，越

---

<sup>6</sup> 丁石孙、张祖贵，数学与教育，湖南教育出版社，1989 年，第 88 页。

<sup>7</sup> 冯志伟，数学与语言，湖南教育出版社，1991 年。

来越多的学者开始关注语言学中数学方法的研究，数学方法在语言研究中的应用越来越广泛，就是在传统的语言学研究中，也开始采用数学的方法，不再认为使用数学方法来研究语言是一种离经叛道的古怪行为。我们认为，在语言研究中采用数学方法，现在已经得到了我国语言学界的普遍认同。随着自然语言处理研究的发展，数学已经成为语言学研究的最重要的一种工具。今天，现代语言学的研究，特别是面向计算机的语言学研究，离开了数学将寸步难行。正是由于这种在观念上的巨大变化，今天我们才有可能理直气壮向读者们推荐这本《语言学中的数学方法》，并且建议在语言研究中使用这样的方法。本书是专门为语言学工作者写的，讲数学问题时都紧紧扣住语言，深入浅出，实例丰富，作者还精心设计了大量的练习，书末附有练习答案选，正好满足了语言学工作者更新知识的迫切需要，是一本不可多得优秀读物。

## 二、 本书主要内容

本书的作者有三位，他们都是语言学家。本书第一作者 Barbara H. Partee（帕蒂）是美国马萨诸萨大学（University of Massachusetts）著名的语言学和哲学教授，国际上资深的蒙塔鸠语法（Montague grammar）研究专家和形式语义学（formal semantics）奠基人之一，蒙塔鸠（Montague）英年早逝，Barbara H. Partee 首先提出蒙塔鸠语法这个术语，使得蒙塔鸠的卓越成果得以传于后世。她于 1965 年师从当代著名语言学家 N. Chomsky 获得博士学位，1986 年担任美国语言学会主席，1984 年和 1989 年先后当选为美国文理科学院和国家科学院院士，她也是荷兰皇家文理科学院外籍院士。多年来她一直担任国际形式语义学刊物的编委。



Barbara H. Partee

本书第二作者 Alice ter Meulen（默梭），出生于荷兰的阿姆斯特丹（Amsterdam），获美国斯坦福大学（Stanford University）博士学位，1985-86 年在荷兰格隆尼根大学（University of Groningen）讲学，现在美国印地安纳大学（Indiana University）哲学和语言学系担任教授。



Alice ter Meulen

本书第三作者 Robert E. Wall（华尔）在美国德克萨斯大学（University of Texas）语言学系任教。

由于本书是这三位语言学家针对语言学的需要写的数学方法方面的著作，完全从语言学的角度来讲述数学，因此特别适合于那些想学习语言学中的数学方法的语言学专业的教师和

学生阅读。

在20世纪80年代中期, Barbara H. Partee 曾出版过《语言学家的数学基础》(Fundamentals of Mathematics for Linguists), Robert E. Wall 曾出版过《数理语言学导论》(Introduction to mathematical Linguistics)。由于学科的迅速发展, 他们深切地感到这两部著作的内容已经比较陈旧, 打算重新编写一本新的著作来论述语言学中的数学问题, 正在这个时候, Alice ter Meulen 从荷兰格隆尼根大学回到美国, 加入了他们的编写工作, 最后, 他们三人同心协力, 写成了这本《语言学中的数学方法》。目前, 本书是这个领域中最有影响、最受读者欢迎的英文著作。

这本书主要是为学习语言学的学生写的, 具有高中数学知识的读者都不难理解本书的内容。本书大多数章节的后面都有练习, 书末附有练习的答案。作者的这些独具匠心的安排, 不仅便于读者深入理解本书的内容, 并可帮助读者把数学的概念应用到语言学的研究中去。

本书的读者群是: 语言学专业的本科高年级大学生和研究生, 对于计算语言学、逻辑程序设计和人工智能有兴趣的计算机科学专业的大学生和研究生, 对于语言学和形式语义学有兴趣的数学研究者和逻辑学研究者, 对于语言学中的数学方法有兴趣的其他读者。

本书包括 A, B, C, D, E 五篇。A 篇讲述集合论, B 篇讲述逻辑和形式系统, C 篇讲述抽象代数, D 篇讲述作为形式语言的英语, E 篇讲述形式语言、形式语法和自动机。建议读者从 A 篇开始, 一篇一篇地仔细阅读, 反复推敲, 认真做练习, 逐步深入下去, 就可以升堂入室, 了解到语言学中使用的主要的数学方法。

当然, 由于篇幅的限制, 本书不可能把目前在语言学中所使用的全部数学方法都包揽无遗。例如, 在统计自然语言处理中使用的概率和统计方法, 在声学语音学中使用的信号处理和声波理论的数学方法, 在机器翻译中使用的自动剖析 (parsing) 方法等, 本书都没有涉及。关于这些方面的内容, 读者可以阅读另外的一些著作。

本书共二十二章, 各章的篇名如下:

#### A 篇 集合论

第一章 集合论的基本概念

第二章 关系和函数

第三章 关系的性质

第四章 无限性

#### B 篇 逻辑和形式系统

第五章 逻辑和形式系统的基本概念

第六章 命题逻辑

第七章 谓词逻辑

第八章 形式系统, 公理化与模型理论

#### C 篇 代数

第九章 代数的基本概念

第十章 运算结构

第十一章 格

第十二章 布尔代数与赫廷 (Heyting) 代数

#### D 篇 作为形式语言的英语

第十三章 基本概念

第十四章 广义量词

第十五章 内涵性

#### E 篇 形式语言、形式语法与自动机

第十六章 基本概念

- 第十七章 有限自动机，正则语言与 3 型语法
- 第十八章 下推自动机，上下文无关语法与语言
- 第十九章 机器，递归可枚举语言与 0 型语法
- 第二十章 线性有界自动机，上下文有关语言与 1 型语法
- 第二十一章 介于上下文无关与上下文有关之间的语言
- 第二十二章 转换语法

下面我们分别介绍各篇、各章的内容。

A 篇《集合论》分为四章，分别讲述集合论的基本概念，关系和函数，关系的性质，无限性。

第一章《集合论的基本概念》首先解释了什么叫做集合，集合的规范说明方法，讨论了罗素悖论 (Russell's Paradox) 与递归规则，接着介绍集合论中的等同与基数，子集合，幂集，集合的并、交、差、补等运算，最后介绍了集合论中的几个重要的定律：幂等律，交换律，分布律，等同律，求补律，德摩根律与一致性原则。本章采用文氏图 (Venn diagram) 来形象地表示集合的运算原理和定律，直观性强，简明易懂。

第二章《关系和函数》首先解释有序对与卡氏积的概念，接着说明  $A$ ， $B$  上的关系  $R$  是卡氏积  $A \times B$  上的一个子集合，这个子集合要满足  $R \subseteq A \times B$  这样的性质，从集合论的角度定义了关系，然后又从关系的定义引入了函数的概念。最后，举例说明了函数的组合。本章对于关系和函数的定义都是从集合论的角度来进行的，向读者具体地说明了如何应用集合论的方法来观察和研究数学问题。

第三章《关系的性质》首先讨论关系的自反性、对称性、传递性和连通性，并且采用图 (diagram) 来表示这些性质，接着讨论关系  $R$  的逆  $R^{-1}$  和关系的补  $R'$ ，分析在  $R^{-1}$  和  $R'$  中，关系  $R$  的自反性、对称性、传递性和连通性的变化情况。在这些讨论的基础上，引入等价关系、等价类与划分的概念。最后，把序 (order) 看成是一种二元关系，并讨论了偏序 (Partial orders) 和偏序集 (partially ordered sets)。

第四章《无限性》深入地讨论无限性问题。首先引入等价集合与基数的概念，讨论集合可枚举性 (也就是可数性) 的概念，并分析不可枚举集合，最后讨论“无限” (infinite) 与“无界” (unbounded) 的区别。

A 篇后面的附录 A 介绍如何用集合论来对于数的系统进行再构建。首先用集合论来描述自然数，接着把这样的描述扩充到所有的整数，再扩充到所有的实数，并进一步扩充到所有的有理数。通过这样的扩充，我们可以体会到集合论在表示数学系统的结构方面确实具有强大的威力，当然也就自然地联想到如何用集合论这个强有力的数学工具来描述语言的结构。

B 篇《逻辑和形式系统》也分为四章，分别讲述逻辑和形式系统的基本概念，命题逻辑，谓词逻辑，形式系统，公理化与模型理论。

第五章《逻辑和形式系统的基本概念》首先讨论形式系统与模型，接着讨论自然语言与形式语言的异同，并说明它们的句法与语义，最后简要地介绍命题逻辑和谓词逻辑。

第六章《命题逻辑》专门讨论命题逻辑问题。命题逻辑的句法比较简单，本章对命题逻辑的句法做了简短的介绍之后，着重讨论命题逻辑的语义，介绍了真值与真值表，讨论了否定、合取、析取、条件、双条件的语义；接着讨论重言命题、矛盾命题与相依命题，介绍逻辑等价、逻辑结论与逻辑定律，并讨论了命题逻辑中的自然演绎 (包括条件证明与间接证明)，最后介绍荷兰逻辑学家 Evert Beth (贝思) 提出的描述语义表的贝思方法，叫做贝思表 (Beth tableaux)。

第七章《谓词逻辑》专门讨论谓词逻辑问题。首先介绍谓词逻辑的句法和谓词逻辑的语义，引入了逻辑量词的概念，讨论量词定律与前束范式 (prenex normal form)，分析谓词逻

辑中的自然演绎,介绍谓词逻辑中的贝思表,最后讨论形式证明与非形式证明,并举例说明数学证明中的非形式风格,指出这种非形式风格的证明并不是草率的证明,而是精练的证明。

第八章《形式系统,公理化与模型理论》内容丰富。首先介绍形式系统的句法方面的问题,引入递归的概念,然后讨论公理系统与推导,引入扩展公理系统的概念,在此基础上介绍挪威数学家 Axil Thue 提出的半图厄 (semi-Thue) 系统<sup>8</sup>,并介绍了作为归纳证明基础的皮亚诺 (Peano) 公理,讨论了数学归纳法。本章还讨论了形式系统的语义方面的问题,介绍了模型理论,分析了理论与模型,讨论了一致性、完备性、独立性和同构的概念,并以美国华裔数理逻辑学家 Wang Hao (王浩) 在《数理逻辑》一书中的 L 系统为实例,进一步说明形式系统的基本概念。接着讨论了关于顺序关系的公理、关于符号串毗连的公理以及集合论的公理化等问题。本章最后介绍公理化逻辑,从讨论命题逻辑的公理化问题开始,介绍一致性的证明与独立性的证明;接着讨论谓词逻辑的公理化问题,介绍关于完备性的证明、可判定性,并介绍了奥地利数学家 Kurt Gödel (哥德尔) 在 1931 年提出了著名的哥德尔 (Gödel) 不完备性定理,说明任何形式系统都不能完备地刻画数学理论,总有某些问题从形式系统的公理出发是不能解答的,因此,可证明的命题固然是真的,但是,真的命题却不一定都是可证明的;对于形式系统来说,真的命题除了可证明之外,还需要进一步的思想能动性以及超穷工具,从而把人们对数学真理的认识推向了新的层次。本章最后讨论了高阶逻辑的问题。

B 篇后面的附录 B.I 列出了各种不同的逻辑符号与联结词,附录 B.II 简单地介绍了克林 (Kleene) 的三值逻辑。

C 篇《代数》也分为四章,介绍抽象代数的基本理论,分别讨论了抽象代数的基本概念,运算结构,格,布尔代数与赫廷 (Heyting) 代数。

第九章《代数的基本概念》介绍代数的定义,运算的性质,并介绍了几个特殊的元(如幺元、逆元、零元、极大元、极小元等),最后讨论映射 (maps) 与同型 (morphisms)。

第十章《运算的结构》讨论了几个代数结构,包括群 (groups), 子群 (subgroups), 半群 (semigroups), 幺半群 (monoids), 整域 (integral domains), 最后讨论不同的代数结构之间的同型问题,例如,群与幺半群之间的同态 (homomorphism) 问题,两个整域之间的同构 (isomorphism) 问题。

第十一章《格》专门讨论“格论”(lattice theory)。首先介绍部分有序集 (posets)、对偶性与图的概念,接着引入格 (lattices)、半格 (semilattices) 与子格 (sublattices) 的概念,讨论了格论中的同型问题,最后讨论筛选格 (filters)、理想格 (ideals)、有补格 (complemented lattices)、分配格 (distributive lattices) 与模块格 (modular lattices)。

第十二章介绍布尔代数与赫廷 (Heyting) 代数。首先从格论的角度引入了两种特殊的代数,一种是布尔代数 (Boolean algebra), 一种是赫廷代数 (Heyting algebra), 讨论了布尔代数的性质,介绍了布尔代数的模式以及布尔代数的集合表示方法,接着介绍赫廷代数,最后介绍 Saul Kripke (克里普克) 在 1965 年提出的克里普克语义学 (Kripke semantics)。

D 篇《作为形式语言的英语》以英语为实例讨论形式语言的特性。本篇只有三章,分别讨论形式语言的一些基本概念,广义量词和内涵性。

第十三章《基本概念》讨论组成性原则 (compositionality) 与 Lambda ( $\lambda$ ) 抽象。在组成

---

<sup>8</sup> 早在 1984 年,冯志伟在《生成语法的公理化方法》(载《生成语法讨论会文集》,1984 年,哈尔滨) 的论文中,把乔姆斯基的形式语法同数学中的半图厄 (semi-Thue) 系统相比较,指出了乔姆斯基的形式语法在实质上就是一个半图厄 (semi-Thue) 系统。

性原则方面，讨论了命题逻辑的组成性说明，谓词逻辑的组成性说明，自然语言与组成性原则。在 Lambda ( $\lambda$ ) 抽象方面，讨论了类型理论， $\lambda$  抽象的句法与语义，然后以形式语言 TL 模式 (schema for Typed Language, 简称“TL 模式”) 为样本，说明了 TL 的句法和语义。最后介绍 Lambda ( $\lambda$ ) 演算及其在语言学上的应用。

第十四章《广义量词》讨论广义逻辑量词问题。首先介绍逻辑学中的限定词和量词的概念，说明了量词的条件，分析了限定词与量词的性质，并且把限定词解释为集合 A 与集合 B 之间关系，最后讨论上下文与量词化问题。

第十五章《内涵性》与本书的其他章很不相同。这一章既不介绍语言分析的各种数学工具，也不介绍如何在语言研究中应用这些数学工具，而是讨论语言哲学中的一些困难的问题。首先讨论了奥地利数理逻辑学家 Frege (弗雷格) 提出的两个问题，从而说明了自然语言中存在着“不透明性”(opacity)，并且强调地说明，不透明性这样的现象是自然语言本身固有的，它会影响语言交际的效率，接着分析了不透明性的形式，讨论了索引 (indices) 与可达性 (accessibility) 关系，并且用索引的方法来解释英语动词的时态和时间副词的语义。最后讨论索引性 (indexicality) 问题。

E 篇《形式语言、形式语法与自动机》是本书最重要的部分，篇幅最长，包括七章。本篇系统地、全面地讨论了与计算机自然语言处理关系最为密切的形式语言、形式语法和自动机等问题。这一篇也是我们的这个导读论述的重点，因此，我们的介绍也稍微详细一些。

第十六章《基本概念》主要讨论形式语言、形式语法和自动机的基本概念。形式语法用于生成语言，自动机用于识别语言。因此，语法和自动机是刻画语言的有效手段。语言中句子的结构可以使用树形图 (tree) 来形式化地表示。如下图所示：

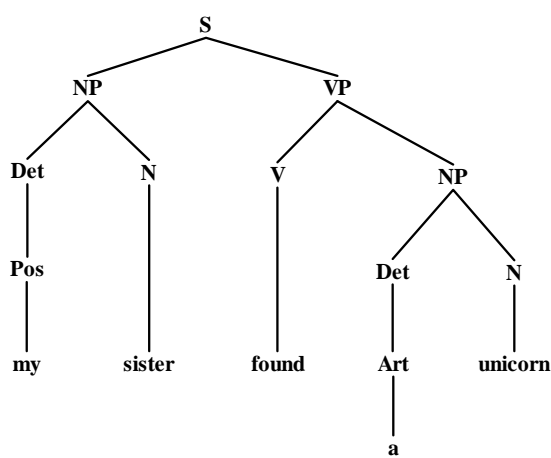


图 1 树形图

这个树形图表示了关于句子 My sister found a unicorn (我妹妹发现了一个独角兽) 的句法结构的三种信息：

- (i) 关于句子成分的层级分组的信息；
- (ii) 关于各个成分的语法类型的信息；
- (iii) 关于句子中的各个成分从左到右的顺序的信息。

从图 1 可以看出，标有 S (Sentence) 的最大成分由 NP (Noun Phrase) 和 VP (Verb Phrase) 两个成分组成，而 NP 又由 Det (Determiner) 和 N (Noun) 组成，……等等。S 表示句子，NP 名词词组，VP 动词词组，Det 表示限定词，N 表示名词，V 表示动词，Art 表示冠词，

Pos 表示物主代词。从图 1 还可以看出, 在句子这个成分中, 名词词组 NP 前于动词词组 VP, 在名词词组 NP 中, 限定词 Det 前于名词 N, ……等等。

树形图本身由结点 (node) 和连接结点的枝 (branch) 组成。每一个结点有一个标记 (label), 这个标记是从语法范畴 (如 S, NP, VP, ……等) 和符号串元素 (如 my, sister, ……等) 的有限集合中选出的。习惯上把树形图看成是在书页上竖直正立的, 标有 S 的结点在顶上, 标有符号串元素的结点在底处。由于枝总是从较高的结点连接到较低的结点, 因此, 它是有固定的连接方向的, 这个方向在一般情况下不必用箭头标出, 只需要规定在树形图的竖直方向上, 枝总是从较高的结点向较低的结点延伸就行了。如果枝用箭头而不是用线段画出, 那么, 结点与结点之间的相对的竖直位置就成了树形图的无关特征。

树形图中有两种关系最为重要: 一种是支配关系 (dominance), 一种是前于关系 (precedence)。本章用数学方法描述了这两种关系。

我们说结  $x$  支配结  $y$ , 如果在树中, 有从  $x$  延伸到  $y$  的一系列的枝把它们连接起来, 这时, 系列中所有的枝从  $x$  到  $y$  要有同一个方向。例如图 1 中, 标有 VP 的结支配着标有 Art 的结, 因为连接它们的一系列的枝都一律地从较高的结点 VP 降到较低的结点 Art。但是, 标有 VP 的结点不支配标有 Pos 的结点, 因为连接这两个结点的路径首先要从 VP 升到 S, 然后再从 S 通过 NP 和 Det 降到 Pos。

如果  $x$  与  $y$  是相异的,  $x$  支配  $y$ , 而且,  $x$  与  $y$  之间没有另一个相异的结点, 那么,  $x$  直接支配  $y$ 。在上面的树形图中, 标有 VP 的结点直接支配标有 V 的结点, 但不直接支配标有 found 的结点。一个结点叫做直接支配它的结点的子结点或直接后裔, 而被同一个结点直接支配的相异的结点叫做姊妹结点。在上面的树形图中, 标有 VP 的结点有两个子结点, 即标有 V 结点和最右边的标有 NP 的结点, V 和 NP 这两个结点是姊妹结点。在支配关系中是极小元的结点, 即不被任何其它的结支配的结点, 叫做根 (root)。在上面的树形图中, 标有 S 的结点就是根。在支配关系是极大元的结点, 也就是被其它的结点支配而它本身不支配任何其它相异的结点的那些结点, 叫做叶 (leaf)。在上面的树形图中, 标有符号串元素 my, sister, ……等的结点都是叶。由于树形图通常是从上到下画出的, 所以, 根总是在顶部, 叶总是在底部。

对于每一个合格的树, 应满足单根条件: 在每一个成分结构树中, 恰好只有一个结点是支配每一个结点的, 这个结点就是根。

树中的两个结点, 只有当它们之间没有支配关系时, 才能够在从左到右的方向上排序。上面的树形图中, 标有 V 的结点前于标有 NP 的结点以及所有被 NP 支配的结点, 但它不能前于或后于标有 S, VP, V 及 found 等的结点, 即那些支配 V 以及被 V 支配的结点。可见, 在树形图中, 从上到下的支配关系同从左到右的前于关系是相互排斥的。

另外, 在树形图中还应该排除某些病态的结构:

树形图不同于偏序集的一个重要特征是不存在一个以上的枝进入的结点, 也就是说, 每一个结点最多只有一个直接支配它的结点。图 2(a)中, 结点 d 有两个直接支配它的结点 b 和 c, 因此, 这个图不是树形图。

树形图的另一个特性是它的枝不容许交叉, 图 2(b)中, 枝是交叉的, 因此, 这个图也不是树形图。

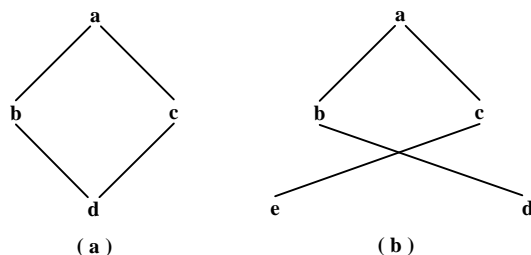


图2 病态的结构

为了排除这两种病态结构，提出了如下的非交条件 (nontangling condition)：

在任何的成分结构树中，对于任何的结点  $x$  与  $y$ ，如果  $x$  前于  $y$ ，则由  $x$  支配的所有的结点前于由  $y$  支配的所有的结点。

图 2(a)的图不满足这个条件，因为  $b$  前于  $c$ ， $b$  支配  $d$ ， $c$  支配  $d$ ，所以， $d$  应该前于  $d$ ，但这是不可能的，因为前于关系具有“反自反”(irreflexivity)的特性。图 2(b)中， $b$  前于  $c$ ， $b$  支配  $d$ ， $c$  支配  $e$ ，根据非交条件， $d$  应该前于  $e$ ，但实际上与此相反。

这一章还描述了树形图中的标记 (labeling)。

在图 1 中，每一个结都有一个相应的标记，我们用标记函数  $L$  来表示结与标记之间的配对情况。这个标记函数的定义域是树中结的集合，而其值域是有限的语法范畴和符号串元素的集合。

语法范畴  $S$ ， $NP$ ， $VP$  等，适于描写各种自然语言，它们对于各种自然语言是通用的。而符号串元素则因语言而异，数量又很多，因此，我们把标记函数  $L$  分为  $L_1$  和  $L_2$  两部分： $L_1$  把叶的结点映入符号串元素  $F$ ， $L_2$  把非叶的结点映入语法范畴  $G$ ，并且， $G$  与  $F$  是不相交的。

本章还介绍了乔姆斯基层级 (Chomsky hierarchy)。根据语法中重写规则的限制，Chomsky 把形式语法分为四种：0 型语法，1 型语法，2 型语法，3 型语法。0 型语法又叫做递归可枚举语法 (recursively enumerable grammar)，1 型语法又叫做上下文有关语法 (context sensitive grammar)，2 型语法又叫做上下文无关语法 (context free grammar)，3 型语法又叫做有限状态语法 (finite state grammar)。每一个有限状态语法都是上下文无关的，每一个上下文无关语法都是上下文有关的，每一个上下文有关语法都是递归可枚举的。这样，我们可把由 0 型语法生成的语言叫 0 型语言 (type 0 language)，把由上下文有关语法、上下文无关语法、有限状态语法生成的语言分别叫做上下文有关语言 (context-sensitive language)、上下文无关语言 (context-free language)、有限状态语言 (finite state language)，也可以分别叫做 1 型语言 (type 1 language)、2 型语言 (type 2 language)、3 型语言 (type 3 language)。由于从限制 1 到限制 3 的限制条件是逐渐增加的，因此，不论对于语法或对于语言来说，都有

$$0 \text{ 型} \supseteq 1 \text{ 型} \supseteq 2 \text{ 型} \supseteq 3 \text{ 型},$$

这就是关于形式语言和形式语法的乔姆斯基层级。

本章最后还讨论了自动机 (automata)。

从自然语言处理的角度来说，我们可以设计一种装置来检验输入符号串，从而识别该符号串是不是语言中的成立句子，如果是成立的句子，这个装置就接收它，如果是不成立的句子，这个装置就不接收它。这种装置就是自动机，它的目的在于识别语言，从而判定某个输入符号串是不是语言中成立的句子，这样，我们可以把自动机看成是语言的识别机。从形式语言理论的角度抽象地说，假如我们把某一个字母表  $A$  上的符号串输入自动机，自动机可以判定这个符号串是不是属于这个字母表  $A$  上的语言，如果是就接收它，如果不是就拒绝

它。因此，自动机是一种识别语言的抽象机器。

第十七章《有限自动机，正则语言与 3 型语法》讨论用于识别正则语言的有限自动机，正则语言以及用于生成正则语言的 3 型语法。

一个有限自动机由控制器、读数头及输入带子所组成。带子两端是无限长的，它被分成一些方格，输入符号串从左到右记录在带子上，每个符号占一个方格。不包含输入符号的方格，用一个特殊的空白符号“#”表示。在任何一个给定的时刻，有限自动机可处于有限数目的状态之中的一个状态上，在这些状态中，如果自动机总是从某个状态开始处理输入符号串，那么，这个状态就叫做初始状态。在有限自动机开始工作时，读数头处于输入带子的最左的非空白符号上。如图 3 所示：

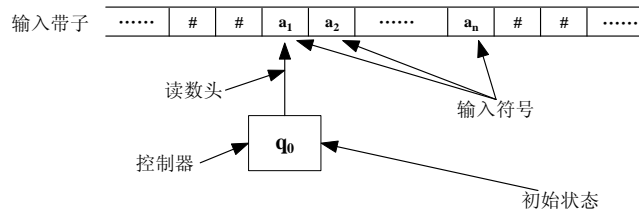


图 3 有限自动机示意图

有限自动机的工作由程序来控制，程序由有限数目的指令组成，每一条指令是形式为  $(q_i, x, q_j)$  的三元组，其中， $q_i$  和  $q_j$  是自动机的状态， $x$  是输入符号。这条指令的含义是：如果自动机处于状态  $q_i$ ，读数头在输入带子上读到符号  $x$  后，向右移动一个方格，并进入状态  $q_j$ 。如果在这个新的状态  $q_j$  又存在另一条指令，则继续执行这条指令，仿此一直进行到自动机不再有可执行的指令为止，这时，自动机也就停止了。另外，当自动机读完输入符号串的最右符号并进入空白时，它也必须停止。空白#不能作为输入字母表中的符号，因此，没有形式为  $(q_i, \#, q_j)$  的指令。如果自动机还没有读完输入带子的全部符号就在中途停止了，那么，我们就说，自动机锁住了，这说明，自动机不能接收输入带子上的符号串。如果自动机读完了整个的输入符号串，并在右边的第一个空白符号#上停住了，那么，输入符号串接收或拒绝依赖于自动机所停止的状态，如果这个状态是最后状态，则输入符号串被接收，否则就被拒绝。

有限自动机可以分为确定性有限自动机与非确定性有限自动机，本章讨论了确定性有限自动机与非确定性有限自动机的等价问题。

接着，本章讨论正则语言，说明了正则语言与有限自动机的关系，指出正则语言也就是有限自动机语言 (finite automata language, 简称 fal)，并讨论了有限自动机语言 (fal) 的抽吸定理 (Pumping Theorem)。最后，讨论正则语言的性质，指出自然语言右线性语法的不足之处，说明英语不是有限自动机语言，因此，3 型语法和有限自动机都不能很好地描述英语。

第十八章《下推自动机，上下文无关语法与上下文无关语言》讨论用于识别上下文无关语言的下推自动机，上下文无关语言以及用于生成上下文无关语言的上下文无关语法。

下推自动机 (pushdown automata) 有一个特殊的后进先出栈，这个栈实际上是一个符号暂存器，在栈中只有在栈顶才能把符号加进或抹去，当把一个符号加到栈顶时，原先处于栈最顶端的符号变成了从栈顶算起的第二个符号，而原先从栈顶算起的第二个符号变成了第三个符号，……等等；当一个符号从栈顶抹去时，原先从栈顶算起的第二个符号变成了最顶端的符号，而原先从栈顶算起的第三个符号变成了第二个符号，……等等，总之，最后加入栈顶的符号，也就是要最先抹去的符号，符号的加进或抹去，遵守着“后进先出” (last in, first

out) 的原则。

如果下推自动机处于如下的格局，输入符号串就被接收：

- i. 输入符号串中的所有符号都已经读完；
- ii. 下推自动机处于最后状态；
- iii. 在后进先出栈变空。

本章接着讨论上下文无关语法与上下文无关语言的关系，说明上下文无关语言可以用上下文无关语法来生成，并讨论了上下文无关语言的抽吸定理、上下文无关语言的闭包特性以及上下文无关语言的可判定性问题。

一般认为，自然语言基本上是一种上下文无关语言，因而可以使用上下文无关语法来生成，也可以使用下推自动机来识别。但是，作者最后提出这样的问题：“自然语言是上下文无关的吗？”

计算语言学家 Shieber 于 1983 年发现，在瑞士德语中存在着词序的交叉对应现象，也就是存在着如图 4 所示的如下的符号串：

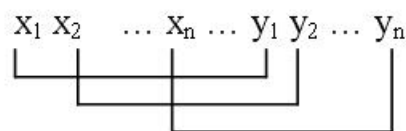


图 4 词序的交叉对应

在图 4 的符号串中， $x_1$  与  $y_1$  对应， $x_2$  与  $y_2$  对应，...， $x_n$  与  $y_n$  对应，上下文无关语法描述不了这样的语言现象。

例如，在瑞士德语中有这样的句子：

Jan säid das mer d'chind em Hans es huus lönd hälfed aastriiche  
约翰 说 我们 小孩 汉斯 房屋 让 帮助 粉刷  
(约翰说，我们让小孩帮助汉斯粉刷房屋)

其中，d'chind (小孩) 与动词 lönd (让) 相对应，Hans (汉斯) 与动词 hälfed (帮助) 相对应，huus (房屋) 与动词 aastriiche (粉刷) 相对应。瑞士德语中这样的语言现象是不能用上下文无关语法来描述的。这样看来，自然语言并不是上下文无关的，它的性质似乎介于上下文无关与上下文有关之间。

第十九章《图灵机，递归可枚举语言与 0 型语法》讨论识别递归可枚举语言的图灵机，递归可枚举语言以及生成递归可枚举语言的 0 型语法。

图灵 (Turing) 机是英国数学家 A. M. Turing (图灵) 最早提出的，所以，就用这位数学家的名字来命名它。

图灵机由控制器、分成方格的输入带子以及每次可扫描输入带子上一个方格的读写头所组成，这个读写头除了能在带子上读，还可以在带子上写，另外，这个读写头还能向左或向右运动。

如有限自动机一样，图灵机的计算从初始状态  $q_0$  开始，这时，读写头在最左的非空白符号上。输入带子的左边和右边也是无限地长的，除输入符号串之外，其它所有的方格都记以空白符号 #。

图灵机的指令是形式为  $(q_i, a_j, q_k, X)$  的四元组，其中， $q_i$  与  $q_k$  是状态 (也可能相同)， $a_j$  是字母表中的符号， $X$  或者是字母表中的符号，或者是特殊符号 L (表示向左运动) 或 R (表示向右运动)。这个指令可解释如下：如果图灵机在状态  $q_i$  扫描符号  $a_j$ ，然后进入状态  $q_k$  (可能与状态  $q_i$  相同)，如果  $X$  是字母表中的一个符号，则用这个符号来替换  $a_j$ 。如果  $X$

是 L 或 R, 那么, 符号  $a_j$  不改变, 而读写头向左或向右移动一个方格。

当 X 的值等于 R 时, 读写头向右移动一个方格, 当 X 的值等于 L 时, 读写头向左移动一个方格, 被读的符号有时也可以是空白, 这样, 控制器就可以超出输入符号串的范围而到达带子的任何一个部分。

当图灵机到达的指令中不能再继续有四元组时, 图灵机就停止。如果图灵机在最后状态停止, 它就接收了输入符号串, 如果图灵机停止在非最后状态, 则输入符号串被拒绝。

由于图灵机不必像有限自动机那样一定要在空白符号停止, 这样, 它就可能会把某一个计算没完没了地进行下去而停不下来。如果对于某一给定的输入符号串, 图灵机的计算永不停止, 那么, 我们也认为这个输入符号串被拒绝。

本章给出了图灵机的形式定义, 对于图灵机的等价性进行了阐释, 讨论了非限定语法与图灵机的关系, 介绍了逻辑学家 Alonzo Church(丘奇)提出的丘奇假设(Church's hypothesis)。最后讨论递归集合与递归可枚举集合, 给出了通用图灵机的概念, 并讨论图灵机的停机问题。第二十章《线性有界自动机, 上下文有关语言与 1 型语法》讨论识别上下文有关语言的线性有界自动机, 上下文有关语言以及生成上下文有关语言的 1 型语法。

识别能力介于后进先出自动机与图灵机之间的自动机是线性有界自动机 (linear bounded automata 简称 lba), 它能识别上下文有关语言。

线性有界自动机是加了如下限制的图灵机, 这种限制是: 带子上的输入符号串是左右都有界的, 读写头在其向左或向右的运动中, 任何时候都绝对不能超出这个界限。

本章讨论了线性有界自动机与上下文有关语言的关系, 上下文有关语言与递归集合的关系。最后讨论上下文有关语言的闭包特性与判定特性。

综上所述, 我们认识到, 刻画语言的有效手段是语法和自动机。语法的基本作用是通过有限数目的重写规则来生成语言中无限数目的句子, 而自动机的基本作用则是识别输入的语言句子是不是成立。

乔姆斯基和 S. Y. Kuroda(库洛达)等人把四种类型的语法分别与图灵机、线性有界自动机、下推自动机及有限自动机等四种类型的自动机相联系, 并且证明了关于语言语法的生成能力和语言自动机的识别能力的等价性的四个重要结果, 即:

- 1) 若一语言 L 能用图灵机来识别, 则它就能用 0 型语法生成, 反之亦然;
- 2) 若一语言 L 能用线性有界自动机识别, 则它就能用 1 型(上下文有关)语法生成, 反之亦然;
- 3) 若一语言 L 能用下推自动机识别, 则它就能用 2 型(上下文无关)语法生成, 反之亦然;
- 4) 若一语言 L 能用有限自动机识别, 则它就能用 3 型(有限状态)语法生成, 反之亦然。

上述理论提出了关于语言的生成过程和语言的识别过程的极为重要的见解, 对于编译技术很有用处, 在计算机科学界产生了很大的影响, 对于自然语言处理具有重要的理论意义和实用价值。

第二十一章《介于上下文无关与上下文有关之间的语言》介绍几种描述自然语言的语法, 它们的生成能力是介于上下文无关语法与上下文有关语法之间的。目前大多数的研究人员都认为, 在关于形式语言和形式语法的乔姆斯基层级中, 自然语言处于上下文无关语言与上下文有关语言之间, 而且它更接近于上下文无关语言。在自然语言句法分析方面, 大量的工作都可以使用上下文无关语法来处理, 而且已经获得了显著的成果, 而在另一方面, 又有一些事实足以说明(例如, 瑞士德语中词序的交叉对应现象), 自然语言似乎是上下文无关语法不能描述的。鉴于这种情况, 学者们提出了一些介于上下文无关与上下文有关之间的语法来描

述自然语言。本章介绍了 Hopcroft (霍普特罗夫特) 与 Ullman (乌尔曼) 1979 年提出的索引语法 (Indexed grammars), Joshi (尤喜) 等在 1975 年提出的树邻接语法 (Tree adjoining grammars), Pollard 在 1984 年提出的中心词语法 (head grammars)。最后本章还讨论了 Bar-Hillel (巴希勒) 早在 1953 年就提出的范畴语法 (category grammars), 说明范畴语法与上下文无关语法在生成能力方面是弱等价的。

第二十二章《转换语法》简要地介绍了 Chomsky 在 1965 年提出的转换语法 (transformational grammars) 的标准理论 (Standard theory), 并介绍了在转换语法的标准理论的基础之上, Peters 和 Richie 的工作。他们在 1973 年用数学方法证明, 使用基于上下文有关的转换语法生成的语言正好是递归可枚举语言, 因此, 由上下文有关语法改进了的转换语法在生成能力方面是与 0 型语法弱等价的。这些使用数学方法研究自然语言问题的重要成果, 不但在自然语言处理中具有指导意义, 而且对于理论语言学的研究, 也是很有启发的。

E 篇后面的附录 E-I-1 给出了乔姆斯基 (Chomsky) 层级的一个详细的图示, 附录 E-I-1 给出了各类形式语言的闭包特性总表和可判定性的总表, 附录 E-II 介绍语义自动机。

书末附有全书的大部分练习答案, 供读者参考, 最后是各篇的参考文献和全书索引。

### 三、 推荐阅读资料

1. Gross, M. and Lentin, A. Introduction to Formal Grammars, Springer Verlag, 1970.
2. D. Jurafsky, J. Martin, Speech and Language Processing, Prentice Hall, 2000, 中译本: 自然语言处理综论 (冯志伟, 孙乐 译), 电子工业出版社, 2005 年。
3. 冯志伟, 数理语言学, 上海知识出版社 1985 年版。
4. 冯志伟, 数学与语言, 湖南教育出版社, 1991 年版。