【美】西摩·金诗伯格著

陈力行　译

上下文无关语言的数学理论

山东大学出版社

上下文无关语言的数学理论

[美] 西摩·.金斯伯格 著

陈 力 行 译

山东大学出版社

上下文无关语言的数学理论

[美] 西摩·.金斯伯格 著

陈 力 行 译

*

山东大学出版社出版
山东省新华书店发行
山东大学印刷厂印刷

*

开本850×1168毫米 1/32 10印张 字数：268千字
1986年12月第一版 1986年12月第一次印刷
印数：1000
统一书号：13338·14 定价：2.00元

内 容 简 介

本书着重阐述了上下文无关文法，上下文无关语言，正规语言，有界语言，派生树，歧义性，有穷状态接受器，下推接受器，上下文无关语言的运算，可解性和不可解性等基本概念和理论，并包含有较好的习题和思考题，内容丰富，推理严谨，是一本好教材，适合于计算机科学，语言学，数学，逻辑学，甚至理论生物学大学生、研究生和有关工作者参阅。

序　言

上下文无关语言的概念首先是由*chomsky*于1959年在[*Ch*3]+中提出的。他企图寻求诸如英语、法语等自然语言的合理的数学模型。在1959—1960年间一些研究工作者写出了许多发展这一理论的文章〔*Ch*3、4；*CM*；*BGS*；*BPS*〕。在六十年代的后期，发现“类*ALGOL*”语言，也就是说，用*Backus*范式（用来描述广泛使用的程序设计语言*ALGOL* 60的元语言）定义的语言和上下文无关语言是等同的。从那时以后，上下文无关语言的理论研究发展十分迅速。自然语言和计算机的关系或程序设计语言的研究都在这方面作了不少的工作。剩下的一些工作是那些对固有的问题、方法和结果感兴趣的数学工作者或逻辑学工作者们完成的。这个发展产生了计算机科学，特别是程序设计技术的一些理论性结果。例如，我们用下推接受器描述上下文无关语言。（下推接受器是在编译的分析方面使用的装置）。我们有这样的事实：一个广义时序机（原始转换装置）把上下文无关语言映射到上下文无关语言。并且我们得到确定任意的上下文无关语言是否岐义的递归不可解性。

上下文无关语言和自动机的结合，其中的每一个作为另一个的问题的工具和源泉，比过去更令人感兴趣。实际上，如上面所

+参看本书末尾的“参考文献和有关文章”中所列出的 N. Chomsky 的第三篇文章。在本书正文中使用同样的简写形式，例如，[Ch3]，就是“文献”中相应的条目前面的记号。

举的前两例，常常难以确切的说它是自动机理论的课题还是语言理论的课题。就这一点而论，我把上下文无关语言的主要内容看作是自动机理论的逻辑延续。特别，我在加利福尼亚大学洛杉矶分校开设了一个两学期的课程，课程名称叫自动机理论和上下文无关语言，第一学期使用一个较早的课本〔*Gi*1〕，第二学期使用的是本书的前身。上下文无关部分是为受过数学训练，具有大学毕业水平或者对数学，特别是逻辑学；或者对计算机科学，特别是程序设计；或者对计算机工程，特别是自动机有兴趣的学生安排的。不需多说，这些学生们的评论对消除错误和使证明严格化是很有帮助的。

这本教材的目的是提供上下文无关语言的重要的特征，为的是把读者送到这个领域的前沿。如果一个项目（1）在另外的理论的发展中被反复的使用，或者是（2）看来似乎与自然语言或是程序设计语言的某些方面有关，则是最为重要的。因此像时序语言和亚线性语言的论题，虽然有它们自身的兴趣，但是几乎没有讨论到。

主要内容被划分为六章（另外还有一章复习必要的数学预备知识）。第一章介绍诸如上下文无关文法，上下文无关语言，派生树，和岐义性等基本概念，并且提出一些关于这些问题的基本事实。第二章讨论正规集合，有穷状态接受器，和下推接受器。给出了正规集合用右线性（或左线性）文法的描述，和上下文无关语言用下推接受器的描述。第三章检查保上下文无关语言的运算。本章还有上下文无关语言和正规语言的交是正规语言的证明，和广义时序机把一个上下文无关语言映射到上下文无关语言。第四章则涉及到可解性和不可解性。主要结果是下述各条的递归不可解性：（1）两个上下文无关文法是否生成同一个上下文无关语言，（2）是否存在一个广义时序机把任意给定的上下文无关语言非平凡地映射到另一个任意给定的上下文无关语言，和（3）任意上下文无关语言是否是岐义的。第五章讨论有界语

言和半线性集合。特别，证明了关于把上下文无关语言映射到半线性集合上的 *Parikh* 的结果，描述了有界上下文无关语言，并导出了证明某个集合为非语言的方法。第六章研究先天岐义性的课题。给出了一个有界上下文无关语言是先天岐义的代数的充分必要条件，也给出了确定任意上下文无关语言是否先天岐义的递归不可解性。

关于课文内容的历史参考资料列在各章的末尾。给出了一部分收编了各章没有引用的文章的综合书目。此外，还包含有相当数量的习题。它们不仅能使对课文有更好的理解，并且包括了这个理论的一些次要论题。(我们要预先告诉读者不少的习题或者是较复杂的，或者是较难的。)许多习题是从公开出版的文献中或为众所周知但未公开出版的文献中取来的。关于这些结果（或题目）的参考文献往往是尽可能的随即给出。另外一些习题是在和 *Sheila A. Greibach*, *Thomas N. Hibbard*, *Gene F. Rose*, *Edwin H. Spanier*, 和 *Joseph S. Ullian* 的交谈过程中提出的。

一些未解决的，也就是说，尚待解决的问题以及思考题分布在全书各处，以指示这一学科另外一些发展方向。如果这些问题和课题中的某些被读者们所研究，这将是对我的劳动的广泛报赏。

感谢 *Gene F. Rose* 和 *Michael Harrsion* 对原稿的细心的阅读，我也很感激我的同事 *Thomas N. Hibbard* 关于第二章的一些有益的谈话。我感谢 *W.E.Singletary* 和 *H·P·Edmundson* 对关于正文进一步解释的建议。最后，我愿意向系统发展公司表示更深切的谢意，他们提供了所需的复制和打印经费以及曾经并将继续使我能够坚持这一领域的工作的机会和气氛。

Seymour Ginsburg

预 备 知 识

在这里，把本书中要用到的集合论的或代数的某些概念和性质给以简要的回顾。我建议把这部分内容推迟到需要的时候再读。

具有性质 P 的那些元素的集合记为 $\{x \mid x$ 具有性质 $P\}$。

其中不含有元素的集合叫做**空集**，用 ϕ 表示。

对于集合 A 和 B，如果 A 的各个元素都是 B 的元素，就称 A 是 B 的一个**子集**，记为 $A \subseteq B$ 或者 $B \supseteq A$。

如果 $A \subseteq B$ 并且 $B \subseteq A$，就称 A 和 B **相等**，记为 $A = B$。

集合族 $\{A_\alpha \mid \alpha\}$ 的**集合并**或**并**是至少属于 A_α 之一的元素的集合，记为 $\bigcup_\alpha A_\alpha$. 没有集合的并被定义为空集。

集合族 $\{A_\alpha \mid \alpha\}$ 的**集合交**或**交**是属于所有 A_α 的元素的集合，记为 $\bigcap_\alpha A_\alpha$。

如果涉及到有穷(无穷)多个集合 $A_1, \cdots, A_n (A_1, \cdots, A_n, \cdots)$，则集合并记为 $A_1 \cup A_2 \cup \cdots \cup A_n$ 或者 $\bigcup_{1 \le i \le n} A_i$ 或者 $\bigcup_{i=1}^{n} A_i$ $(\bigcup_{i \ge 1} A_i$ 或者 $\bigcup_{i=1}^{\infty} A_i)$，而集合交记为 $A_1 \cap \cdots \cap A_n$ 或者 $\bigcap_{1 \le i \le n} A_i$ 或者 $\bigcap_{i=1}^{n} A_i$ $(\bigcap_{i \ge 1} A_i$ 或者 $\bigcap_{i=1}^{\infty} A_i)$。

如果 $A \cap B = \phi$，集合 A 和 B 是**无交的**。

对于集合 A 和 B，$A - B$ 是那些属于 A 而不属于 B 的元素的

集合。

如果$A \subseteq B$，则A对于B的**补**是集合$B - A$。通常B是可理解的。取补的动作即称为取补。

有穷集合A的元素的个数用$\#(A)$表示。

如果A是一个集合，则A的所有子集合的族用2^A表示，即$2^A = \{B \mid B \subseteq A\}$。

一个由单个元素a构成的集合有时就记为a，而不写成$\{a\}$。

一个**字母表**是一个有穷非空集合。除非作相反声明，Σ总是用来作为一个**字母表**。①

一个字母表Σ上的一个**字**（其长度$k > 0$）是Σ中元素的有穷序列$x_1, \cdots, x_k$。一个字$x_1, \cdots, x_k$通常的写法是不加逗号。即写为$x_1 \cdots x_k$。

长度为零的字叫做**空字**，用ε表示。

一个非空字是一个长度$k \geq 1$的字。

字w的长度用$|w|$表示。

如果$k \geq 1$且各个x_i在Σ中，则$x_1 \cdots x_k$是一个Σ**字**。

如果$r = s$并且对每一个i**都有**$x_i = y_i$，**就说**Σ字$x_1 \cdots x_r$和$y_1 \cdots y_s$是相等的，记为$x_1 \cdots x_r = y_1 \cdots y_s$。

如果$x = x_1 \cdots x_r$和$y = y_1 \cdots y_s$都是字，则$x \cdot y$是字$x_1 \cdots x_r y_1 \cdots y_s$，叫做$x$和$y$的**连结**或（复合）乘积。我们通常用$xy$代替$x \cdot y$。

字母表Σ上所有字，包括ε的集合用Σ^*表示。二元组$(\Sigma^*, \cdot)$叫做（由Σ产生的）自由半群，其中，是乘积运算，我们通常用Σ^*代替$(\Sigma^*, \cdot)$。

如果X和Y是Σ^*的字集，则X和Y的（复合）乘积是字集$\{xy \mid x$在X内，y在Y内$\}$，记作$X \cdot Y$或XY。

---

① 这个记号不要和$\sum_{i=m}^{n} x_i$混淆，通常这是数$x_m, \cdots, x_n$的和，即$x_m + \cdots + x_n$。

如果 $A \subseteq \Sigma^*$，则 $A^0 = \{\varepsilon\}$，并且对每一个 $n \geq 1$，$A^n = A^{n-1}A$。A **星号**是集合 $\bigcup_{n=0}^{\infty} A^n$，记为 A^*。

因而，对一个字 w，$w^0 = \varepsilon$，对每一个 $n \geq 1$ $w^n = w^{n-1}w$，且 $w^* = \{w^n \mid n \geq 0\}$。

设 x 和 y 是 Σ^* 中的字。如果对 Σ^* 中某 u 及 v 有 $y = uxv$。则 x 是 y 的**子字**。如果对 Σ^* 中某个 v 有 $y = xv$，则 x 是 y 的**初始**子字。如果对 Σ^* 中某个 u 有 $y = ux$，则 x 是 y 的**结尾子字**。

集合的有穷序列 $A_1, \cdots, A_n$ 的**笛卡儿乘积**是所有 n 元组 $(a_1, \cdots, a_n)$ 的集合，用 $A_1 \times A_2 \times \cdots \times A_n$ 表示，其中各个 a_i 分别在 A_i 中。

集合 A 上的一个**关系** R 是 $A \times A$ 的一个子集 R。我们写成 xRy 或者 "xRy 成立" 来指明 (x,y) 在关系 R 中。我们把 xRy 不成立记为 $x\not Ry$。

非空集合 A 上的关系 R 是**等价关系**，如果（1）对 A 中所有元素 x 有 xRx，（2）只要有 xRy 就有 yRx，和（3）只要 xRy 和 yRz 就有 xRz。

如果 R 是 A 上的等价关系，则各个集 $[x] = \{y \mid xRy\}$ 集合叫做由 R 生成的**等价类**，这里的 x 在 A 中。

集合 A 上的一个**偏序**是 A 上的一个关系 $\leqslant$，使得（1）对 A 中所有 x 有 $x \leqslant x$，（2）如果 $x \leqslant y$ 且 $y \leqslant z$，则 $x \leqslant z$，和（3）如果 $x \leqslant y$ 且 $y \leqslant x$，则 $x = y$，即 x 是 y。

二元组（A，$\leqslant$）称之为**偏序集合**。

设（A，$\leqslant$）是一个偏序集合。x_0 是 A 中**极小元**，如果 $x \leqslant x_0$ 隐含 $x_0 = x$。x_0 是 A 中**极大元**，如果 $x_0 \leqslant x$ 隐含 $x_0 = x$。A 中元素 x 和 y 是不可比较的，如果 $x \leqslant y$ 和 $y \leqslant x$ 都不成立。

集合 A 上的一个**全序**是 A 上的一个偏序 $\leqslant$，使得对 A 中所有的 x 和 y，或是 $x \leqslant y$，或是 $y \leqslant x$。

一个**函数**（或**映射**，**变换**，或**运算**）是一个三元组（A，B，

f），其中A和B是非空集合，并且对A中每一个元素x，在B中有一个元素与之对应，这个元素用$f(x)$表示。（A，B，f）也叫做**A到B内的一个函数f**。

函数（A，B，f）通常用符号f表示，并且称为**定义**在A上。

设f是A到B内的函数。如果$E\subseteq A$，则$f(E)=\{f(x)\mid x$在E中$\}$。如果$E\subseteq B$，则$f^{-1}(E)=\{x\mid f(x)$在E中$\}$。如果$f(A)=B$，则f称作**A到B上的函数**。

函数f是一对一的，如果只要$x_1\neq x_2$就有$f(x_1)\neq f(x_2)$。

设f是A到B内的函数，g是C到D内的函数，$B\subseteq C$。则gf是对于A中各个x，由$gf(x)=g[f(x)]$定义的A到D内的函数。

只要写出gf以及函数（A，B，f）和（C，D，g），即假定有$B\subseteq C$。

一个集合是**可数的**，如果存在一个集合A到正整数集合的一对一函数f。A是**可数无穷的**，如果它是可数的且为无穷的。

如果一个集合不是可数的，就称为**不可数的**。

如果从一个不可数的集合去掉可数个元素，则得到的集合仍然是不可数的。

如果存在整数n_0和n_1，使得对于每个整数$n\geq n_0$，有$x_{n+n_1}=x_n$，则称元素的无穷序列$\{x_n\}$是终极周期的。非负整数集合B称为**终极周期的**，如果B或是有穷，或是$y_1,y_2-y_1,\cdots,y_{n+1}-y_n,\cdots$是终极周期序列，其中$y_1,y_2,\cdots$是$B$的元素，按其大小依增大顺序排列。

一个**有穷有向图**是一个二元组(V,D)，其中

1　V是（结点的）有穷非空集合。

2　D是$V\times V$的元素的有穷序列$(p_1,q_1),\cdots,(p_r,q_r)$[各个$(p_i,q_i)$都是**从$p_i$到$q_i$的有向线段**]。

有穷有向图内的一条**路径**π是D中的有向线段

$$(p_{i_1}, q_{i_1}), \cdots, (p_{i_k}, q_{i_k})$$

的非空有穷序列，具有性质：对于$1 \leqslant j \leqslant k-1$，有$q_{i_j} = p_{i_{j+1}}$。$\pi$也叫做**从$p_{i_1}$到$q_{i_k}$（长度为k）的路径**，各个$p_{i_j}$和$q_{i_j}$都叫做**路径中的结点**。等价地，也可以说对于各个i，路径包含结点p_i和q_i。

如果（V，D）是一个有穷有向图，其中结点p和q的各对至多有一个i，使得$(p, q) = (p_i, q_i)$，则各条路径(p_{i_1}, q_{i_1})，$\cdots$，(p_{i_k}, q_{i_k})，通常写成p_{i_1}，q_{i_1}，q_{i_2}，$\cdots$，q_{i_k}。

如果π_1是路径(p_{i_1}, q_{i_1})，$\cdots$，(p_{i_k}, q_{i_k})，π_2是路径$(p_{i_{k+1}}, q_{i_{k+1}})$，$\cdots$，$(p_{i_t}, q_{i_t})$，并且$q_{i_k} = p_{i_{k+1}}$，则$\pi_1\pi_2$是路径$(p_{i_1}, q_{i_1})$，$\cdots$，$(p_{i_t}, q_{i_t})$。

已知路径π_1和π_2。π_1是π_2的**结尾子路径**，如果$\pi_1 = \pi_2$，或者对于某条路径π_3，有$\pi_2 = \pi_3\pi_1$。

一个**有穷有向附标图**是一个四元组（V, D, E, f），其中

1　（V, D）是一个有穷有向图，且D是序列(p_1, q_1)，$\cdots$，(p_r, q_r)。

2　E是（**标记的**）有穷非空集。

3　f是从$\{1, \cdots, r\}$到E内的函数。[$f(i)$是(p_i, q_i)的标记。]

设（V, D, E, f）是一个有穷有向附标图，且$E \subseteq \Sigma^*$。路径(p_{i_1}, q_{i_1})，$\cdots$，(p_{i_k}, q_{i_k})的标记是字$f(i_1)\cdots f(i_k)$。

一棵**具有附标结点的有穷有根有向树**是一个五元组（V, D, H, f, S_0），其中

1　（V, D）是一个有穷有向图，且D是序列(p_1, q_1)，$\cdots$，(p_r, q_r)。

2　H是（**结点名称**的）有穷非空集。

3　f 是从 V 到 H 内的函数。[$f(v)$ 是 v 的**结点名称**。]

4　S_0 在 V 中（根）。

5　对于 $i \neq j$，有 $q_i \neq q_j$。

6　对于 V 内各个 $q \neq S_0$，有一条 S_0 到 q 的路径。

7　在 V 中不存在 q 有 q 到 q 的路径。

设 $\mathscr{R}^n = \{(x_1, \cdots x_n) \mid$ 各 x_i 是一个有理数$\}$。$\mathscr{R}^n$ 的每一个元素是一个**向量**。对各个有理数 c 和 $\mathscr{R}^n$ 中各个 $x = (x_1, \cdots, x_n)$ 和 $y = (y_1, \cdots, y_n)$，令 $x+y = (x_1 + y_1, \cdots, x_n + y_n)$ 和 $cx = (cx_1, \cdots, cx_n)$。$(\mathscr{R}^n, +)$ 是**有理数上所有 n 元有理数组的向量空间。**

对于 $\mathscr{R}^n$ 中 $(x_1, \cdots, x_n)$ 和 $(y_1, \cdots, y_n)$，如果对于各个 i，都有 $x_i = y_i$，则记为 $(x_1, \cdots, x_n) = (y_1, \cdots, y_n)$。

$\mathscr{R}^n$ 中元素的集合 $\{z_1, \cdots, z_k\}$ 是**线性无关的**，如果只要 $c_1 z_1 + \cdots + c_k z_k = (0, \cdots, 0)$ 则必有各个 $c_i = 0$。否则 $\{z_1, \cdots, z_k\}$ 是**线性相关的。**

$\mathscr{R}^n$ 中每一个 $n+1$ 个向量的集合都是线性相关的。

设 $\{z_1, \cdots, z_n\}$ 是 $\mathscr{R}^n$ 中线性无关向量的集合。则 $\mathscr{R}^n$ 中各个 x 都是 z_i 的唯一线性组合，即存在唯一的一组实数 $c_1, \cdots, c_n$ 使得 $x = c_1 z_1 + \cdots + c_n z_n$。

如果 $\{z_1, \cdots, z_k \mid k < n\}$ 是 $\mathscr{R}^n$ 的线性无关向量的集合，则在 $\mathscr{R}^n$ 中存在向量 $z_{k+1}, \cdots, z_n$，使得 $\{z_1, \cdots, z_n\}$ 是线性无关的。再加强些，如果 $\{z_1, \cdots, z_k \mid k < n\}$ 是 $\mathscr{R}^n$ 的线性无关向量的集合，且 $\{z'_1, \cdots, z'_n\}$ 是 $\mathscr{R}^n$ 的 n 个线性无关的向量，则存在 $n-k$ 个向量 z_i'，譬如 $z'_{i_1}, \cdots, z'_{i_{n-k}}$，使得 $\{z_1, \cdots, z_k, z'_{i_1}, \cdots, z'_{i_{n-k}}\}$ 是线性无关的。

如果 $\{z_i = (z_{i1}, \cdots, z_{in}) \mid 1 \leqslant i \leqslant n\}$ 是 $\mathscr{R}^n$ 的线性无关向量的集合，则①

---

①我们假定读者已熟悉许多大学代数教本，例如［BDW］中所讲到的初等行列式理论.

$$\begin{vmatrix} z_{11} \cdots\cdots z_{1n} \\ \cdots\cdots\cdots\cdots \\ z_{n1} \cdots\cdots z_{nn} \end{vmatrix} \neq 0$$

如果 $t_1, \cdots, t_r$ 是非负整数，则 $\max\{t_i \mid 1 \leqslant i \leqslant r\}$ 是 t_i 中最大的数，而 $\min\{t_i \mid 1 \leqslant i \leqslant r\}$ 是最小的数。

对于整数 m 和 n，如果存在一个整数 k 使得 $n = km$，就说 m **能除尽** n。

对于整数 $t_1, \cdots, t_r$，t_i 的 **最大公约数** 是能除尽各个 t_i 的最大整数 m。

整数 $t_1, \cdots, t_r$ **是互素的**，如果它们的最大公约数是 1。

目　　录