形式语言及其句法分析

〔美〕A. V. 阿霍　J. D. 厄尔曼 著

科学出版社

形式语言及其句法分析

〔美〕A. V. 阿霍　J. D. 厄尔曼 著

石青云 译

科 学 出 版 社

1987

内 容 简 介

本书系统介绍形式语言及其句法分析方法. 全书共分七章，其中包括数学预备知识、编译导引、语言理论基础、翻译理论、一般句法分析方法、单路无回溯剖析和回溯量有限制的剖析算法. 书中所述内容是计算机编译理论的重要组成部分，也是句法模式识别的理论基础.

本书论述严谨而富于启发性，各章节末配有由易到难的大量习题，便于读者学习和研讨.

本书可供从事计算机理论、软件工程、模式识别、人工智能等方面工作的科研人员参考，也可作为高等院校有关专业的教材.

Alfred V. Aho Jeffrey D. Ullman
THE THEORY OF PARSING, TRANSLATION, AND COMPILING
VOLUME 1: PARSING
Prentice-Hall, Inc., 1972

形式语言及其句法分析

〔美〕A. V. 阿霍 J. D. 厄尔曼 著
石青云 译
责任编辑 鞠丽娜

科学出版社出版
北京朝阳门内大街 137 号
中国科学院印刷厂印刷
新华书店北京发行所发行 各地新华书店经售

*

1987 年 9 月第 一 版 开本：850×1168 1/32
1987 年 9 月第一次印刷 印张：18 1/2
印数：0001—9,750 字数：484,000

统一书号：15031·840
本社书号：4951·15—8

定价：5.50 元

译 者 的 话

形式语言理论自二十世纪五十年代问世以来，已逐渐成为计算机科学的一个重要领域。其原因不仅在于它提供了计算机程序设计语言的数学模型及编译程序设计的基本理论和方法，而且在于它奠定了模式识别的语言结构法的基础。

这本书的原书名是 A. V. 阿霍和 J. D. 厄尔曼合著的《剖析、翻译和编译理论(卷 1：剖析)》。为使本书书名更能体现该书内容，翻译时，我们将该书书名改为《形式语言及其句法分析》。书中除了讲述形式语言与自动机的关系、语言的闭包性质和可判定性问题等理论性较强的内容以外，一个突出的特点是系统介绍了各种句法分析(剖析)算法，并分析了它们的计算复杂度。这一点对于模式识别尤为重要。当一类模式所对应的形式语言由产生它的文法来描述时，相应的句法分析算法就可以作为该类模式的识别程序。为方便读者系统了解这方面的知识，我们选择了这本书。如果从事计算机软件工作的科研人员和高校师生希望了解作者对编译理论的进一步阐述，请参阅 A. V. 阿霍和 J. D. 厄尔曼合著的《剖析、翻译和编译理论(卷 2：编译)》。

本书结构严谨完整，概念清晰，富于启发性，特别是着重于交待处理各种论题的基本思想。书中选用了很多例子来帮助读者理解基本概念和理论推导，在各章节末尾配有由易到难的大量习题，并给出了进一步研究的问题和有关参考文献的注释。长期以来美国一些大学一直把它作为教科书使用。

本书的译稿曾请中国科学院自动化所戴汝为同志进行了校对，在此一并致谢。

由于译者水平所限，加之时间仓促，书中难免有错误或不妥之

处，敬请读者批评指正。

译　者

于1985年8月

前　言

这本书是为高年级大学生或研究生的编译理论课程写的，可以作为一学期或两学期的教材．书中着重对编译这个实际课题进行理论分析，这样做的出发点有三个方面：

(1) 在一个象计算机科学这样迅速发展的领域，正确有效的教学法是在课程中突出基本思想，而不是强调细节．我们希望在本书中介绍的算法和概念，对下一代计算机和程序设计语言还适用，至少是其中的一部分对编译程序设计以外的领域也能用．

(2) 编译程序的设计已经进展到能把一个编译程序的许多部分分割开来，分别进行设计优化的阶段．这样，为从事优化的人员提供适当的数学工具就很重要．

(3) 一些最有用和效率最高的编译算法，如 LR(k) 剖析法，需要很多的数学基础才能完全理解．因此，对于编译程序的设计者来说，一个基本要求就是要有较好的理论基础．

在不省略与编译有关的较难定理的同时，我们尽量使这本书通俗易懂．为此，列举了大量的例子．每个例子都采用较简单的文法，而不是实际中遇到的复杂文法．希望即使在纯粹理论推导很难懂的情况下，这些例子也足以解释清楚基本概念．

本书的用法

这本书是在给普林斯顿大学和史蒂文斯理工学院的高年级大学生和研究生授课的讲义基础上写成的．以这本书为教材，既讲授过一学期的课，也讲授过两学期的课．作为一学期的课程，不包括本书的第 0，第二，第八章，其余各章的内容大部分要详细讲述，但必须先修包括有限自动机和上下文无关语言的课程．

作为连续两学期的课程，第 I 卷的大部分内容都在第一学期讲授，第 II 卷中除第八章以外，大部分内容在第二学期讲授．两

学期的课程比一学期的课程更注重证明和证明的技巧。

显然，书中有些章节是更为重要的，我们想就第Ⅰ卷中各部分内容的重要程度作一简单评论。一般来说，将大部分证明省略掉大致是可取的。显然把所有主要结果的证明都包括到书中来对于深刻理解论题无疑是必要的，但我们估计编译方面的许多课程很难达到这样的深度，而且只要肤浅地了解一些证明，对于编译也能得到一定程度的理解。

第0章(数学基础)和第一章（编译概貌）几乎都是基本的材料，也许1.3节是例外，这一节包括了句法分析在编译以外的其它应用。

我们相信在第二章(语言理论)中介绍的每个概念和定理都可以在其余九章的一些地方找到用处。但有些材料在编译课中可以省略，如2.2.1节中关于正规表达式方程的较难部分。其次是关于右线性文法的某些材料，尽管有办法得到它们与有限自动机的等价性，但也可以省略掉。此外还可以省略2.4.5节中达到格雷巴赫范式的Rosenkrantz方法。

第三章(翻译)的概念对于本书的其余部分是非常基本的，但3.2.3节关于句法制导翻译的层次这部分内容比较难，可以省略。

我们认为4.1节关于句法分析的回溯方法与4.2节的列表方法相比，是较为次要的。

第五章（单路剖析）的大部分是基本的。我们建议LL文法(5.1节)、LR文法(5.2节)、优先文法(5.3.2节和5.3.4节)和算子优先文法(5.4.3节)应优先考虑。如有必要，其它几节可以省略。

第六章(回溯算法)与第五章的大部分内容相比或与4.2节相比，是较为次要的。如果要给予选择，我们宁愿要6.1节而不要6.2节。

本书的组织

整个句法分析、翻译和编译的理论以两卷的形式出现，即句法分析(第0至第六章)和编译(第七至第十一章)。第二卷的主题是

句法分析程序的优化、确定性句法分析理论、翻译、簿记和代码优化。这两卷形成一个整体，页数是连贯编号的，在第II卷中有两卷的文献目录和索引。

习题和文献注释在每一节（编号为 $i.j$）的末尾。除了公开问题和待研究的问题以外，我们用*号表示习题的难度。单*号的题目要求深入思考来求解，双*号的题目需要更加深入的思考。

以这本书为教材的课程，最好配备一个程序设计实验室，在那里设计和实现几个编译程序的部件。本书一些章节的末尾附有程序设计练习题，可以作为这种程序设计实验室的实验项目。

致谢

在准备这本书的过程中，许多人曾仔细阅读手稿的各个部分并给予很大帮助。我们特别感谢 David Benson, John Bruno, Stephen Chen, Matthew Geller, James Gimpel, Michael Harrison, Ned Horvath, Jean Ichbiah, Brian Kernighan, Douglas McIlroy, Robert Martin, Robert Morris, Howard Siegel, Leah Siegel, Harold Stone 和 Thomas Szymanski, 以及审稿人 Thomas Cheatham, Michael Fischer 和 William McKeeman. 我们也从用过有关讲义的许多学生那里得到重要的意见，如他们中间的 Alan Demers, Nahed El Djabri, Matthew Hecht, Peter Henderson, Peter Maika, Thomas Peterson, Ravi Sethi, Kenneth Sills 和 Steven Squires.

我们还要感谢 Hannah Kresse 和 Dorothy Luciani 对手稿的出色打字，以及在手稿准备过程中贝尔电话实验室所提供的支持帮助。Dennis Ritchie 和 Kenneth Thompson 设计的 PDP-11 计算机的操作系统 UNIX 的运用，加速了手稿中一些部分的准备。

A. V. 阿霍

J. D. 厄尔曼

目　录

第〇章　数学预备知识

为了使讲解清楚和确切，我们需要精确的且严格定义的语言。这一章就来描述在讨论句法分析、翻译和书中其它论题时要用到的语言，包括集合论初步、图论和逻辑的一些基本概念。已经具有这方面基础知识的读者，可以容易地掠过这一章，而把它当作有关定义和记号的参考资料来对待。

0.1　集合论的一些概念

这一节将扼要地复习集合论中一些最基本的概念：关系、函数、次序和普通集合运算。

0.1.1　集合

以下假设有一些当成原子的确定对象。术语"原子"是一个初级概念，换句话说，我们不再给原子下定义。而且，把什么东西叫做原子，依赖于所论及的范围。在许多情况下，把整数或一个字母表中的字母当成原子是合适的。

我们也给"成员资格"一个抽象的记号。若 a 是 A 的成员(元)，就写成 $a\in A$。反之，则写成 $a\notin A$。还假设一个原子没有元，就是说，若 a 是一个原子，则对所论及的范围中一切 x 都有 $x\notin a$。

我们也考虑一些不是原子的基本对象，称之为集合。若 A 是一个集合，则它的成员或元素是满足 $a\in A$ 的那些对象(不一定是原子)。一个集合的每个元或者是原子，或者是其它集合。我们假定，一个集合的每个元，恰好在该集合中出现一次。若 A 的成员数目是有限的，则称 A 为有限集，通常记为 $A=\{a_1, a_2, \cdots, a_n\}$，只要 $a_1, \cdots, a_n$ 都是 A 的元，且当 $i\neq j$ 时，$a_i\neq a_j$。元的次序则是无关

紧要的。例如，我们也可以将上述集合写成 $A=\{a_n, \cdots, a_1\}$。我们约定用符号 ϕ 表示空集，它是一个没有元的集合。要注意，一个原子也没有元，但空集不是原子，原子也不是空集。

表达式 $\#A=n$ 的意思是：集合 A 有 n 个元。

例 0.1

如果把所有非负整数作为原子，那么 $A=\{1, \{2, 3\}, 4\}$ 是一个集合。A 的元是 1，$\{2, 3\}$ 和 4。其中元 $\{2, 3\}$ 也是一个集合，它的元是 2 和 3。但原子 2 和 3 本身不是 A 的元。我们可以等价地写成 $A=\{4, 1, \{3, 2\}\}$。还注意到 $\#A=3$。 □

规定集合的一个有效方式是借助于谓词，即一个语句，其中包含一个或多个未知量，每个未知量可取两值之一：真或假。由一个谓词所规定的集合，是使该谓词为真的那些元素组成的。但我们必须仔细选择规定集合的谓词，不然，所规定的集合就可能不存在。

例 0.2

罗素(Russell) 诤论就可以说明上述现象。令 $P(X)$ 为谓词："X 不是自己的元"，即 $X\notin X$，则可设想规定一个集合 Y，它含有使 $P(X)$为真的所有X，即Y由所有不是自己的元的集合组成。由于最普通的集合似乎都不是自己的元，好象就可以认为集合Y存在。

但是，若Y存在，则应能回答这样的问题："Y 是它自己的元吗?"而这将导致不可能情况。若 $Y\in Y$，则 $P(Y)$ 为假，由集合Y的定义，Y 不是它自己的元。因此，$Y\in Y$ 是不可能的。反之，设 $Y\notin Y$，则由Y的定义，$Y\in Y$。可见，$Y\notin Y$ 蕴含 $Y\in Y$ 且 $Y\in Y$ 蕴含 $Y\notin Y$。而 $Y\in Y$ 与 $Y\notin Y$ 只能有一个为真，两者同时为真是不可能的。因此，只有一条出路，就是承认Y不存在。 □

避免罗素诤论的常规方法是：只用形如"X 在 A中且 $P_1(X)$"的谓词 $P(X)$ 来规定集合。这里A是一已知集合且 P_1 是任一谓词。当对集合A已有了解时，只要写 $P_1(X)$ 来代替 "X 在 A中且 $P_1(X)$"就可以了。

若 $P(X)$是一个谓词，则用 $\{X|P(X)\}$表示使 $P(X)$为真的所有