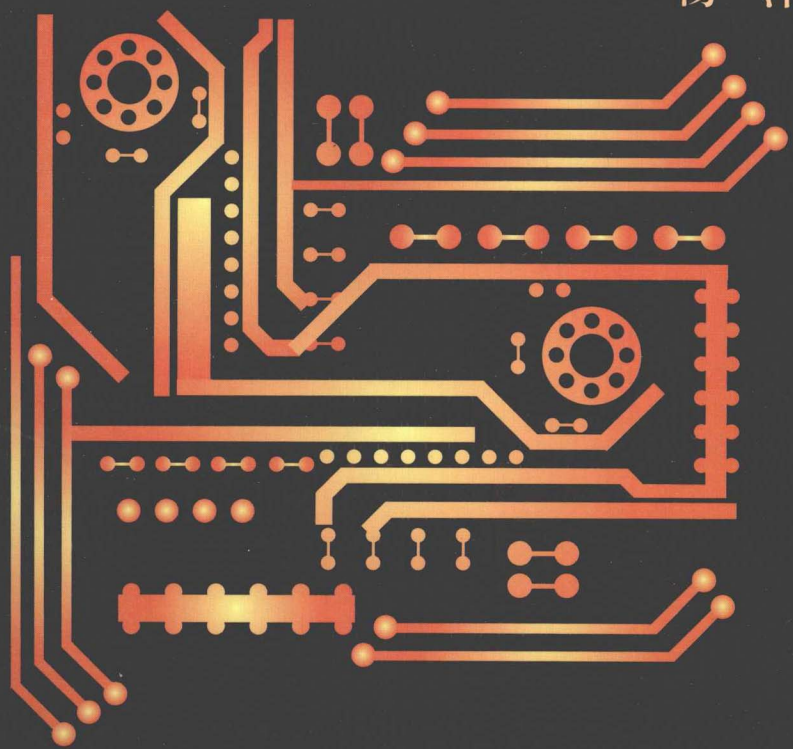
世界著名计算机教材精选

编译器构造

(Java语言版)

Anthony J. Dos Reis 著

杨 萍 等译



COMPILER CONSTRUCTION USING JAVA, JAVACC, AND YACC

清华大学出版社

世界著名计算机教材精选

编译器构造

（Java 语言版）

Anthony J. Dos Reis 著
杨 萍 等译

清华大学出版社
北 京

Anthony J. Dos Reis
Compiler Construction Using Java, JavaCC, and Yacc
EISBN: 978-0-470-94959-7

图书在版编目（CIP）数据

编译器构造：Java 语言版 /（美）赖斯（Reis,A.J.D.）著；杨萍等译. --北京：清华大学出版社，2014
书名原文：Compiler Construction Using Java, JavaCC, and Yacc
世界著名计算机教材精选
ISBN 978-7-302-34055-4

Ⅰ. ①编…　Ⅱ. ①赖…②杨…　Ⅲ. ①编译器－教材　②Java 语言－教材　Ⅳ. ①TP314　②TP312

中国版本图书馆 CIP 数据核字（2013）第 238227 号

责任编辑：龙启铭
封面设计：傅瑞学
责任校对：焦丽丽
责任印制：王静怡

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn, http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 刷 者：北京富博印刷有限公司
装 订 者：北京市密云县京文制本装订厂
经　　销：全国新华书店
开　　本：185mm×260mm　　印　　张：37.5　　字　　数：913 千字
版　　次：2014 年 3 月第 1 版　　印　　次：2014 年 3 月第 1 次印刷
印　　数：1～2000
定　　价：89.00 元

产品编号：041698-01

译 者 序

编译器是计算机系统最核心最基础的支撑软件之一。编译器设计原理与技术相关的知识体系，可以体现从计算机程序设计语言到计算机体系结构相对独立的整机概念，又涉及到形式语言与自动机、数据结构与算法等计算机学科的基础理论，还关系到独特的软件设计方法，不愧为联系计算机科学理论和计算机系统的典范。正如前辈 Alfred V. Aho 和 Jeffrey D.Ullman 在他们的著作中所提到的，在每一个计算机学者的职业生涯中，都会反复用到这些原理和技术。正因为如此，在大多数高等院校的计算机科学与技术专业，编译原理与技术都作为必修的核心专业基础课程之一。

国内外关于编译原理与技术的教材已有不少，也有几部堪称经典之作。然而，如果你是一位想要为本科生开设编译原理与技术相关课程的教师，那么你会发现，选择适合的教材其实是相当不容易的。分析其原因，主要是因为选材非常广泛，不同理念的书籍之间内容差异较大。一些书籍的篇幅过长，它们或者在某些方面过于细节化，或者涉及到许多编译高级话题，有些作者甚至根据个人的研究兴趣选择了某些专题内容。另有一些书籍过于技术化，涉及到的原理部分不成体系，其理念主要是强调如何设计出可用的编译器。还有一些书籍认为基于编译前端的技术已经非常成熟，所以将前端所涉及到的原理和技术完全弱化，主要篇幅是关于编译优化以及后端设计的内容。此外，计算机技术几乎天天都在发展变化，这伴随着编译器技术也在不断的发展变化。那么，本科阶段应该学些什么，应该如何理顺相关理论和技术的脉络，既能使学生轻松掌握相关知识和技能的精髓，又能使教学安排条理有序，并且对学生完成的编译器项目能有一套比较合理的评价体系？这些问题对任何一位相关课程教师来说，都会有不同程度的困惑。

Anthony J. Dos Reis 教授编写的本书能够较好地回答上述问题，为编译器的理论、应用以及编程技术的学习和课程教学设计了一套较为科学的体系，既独特又新颖，或许是符合您开课理念的书籍。该书涵盖了从相关基础到实践技能学习的各个方面，配套有功能强大且灵活的项目评价和辅导材料的软件包、精心设计的项目以及测试用例。在我们看来，这本书很适合作为计算机科学与技术相关专业本科阶段的编译原理与技术课程的教材。该书对于基本原理的讲解易于理解，篇幅也适中，但内容很到位，与书中编译器设计和相关技能的学习紧密关联。本书对于自上而下分析与翻译的原理、方法与技术介绍的十分详细，包含手工构造方法和自动构造方法。本书中使用 JavaCC 自动构造自上而下的分析与翻译程序，对于 JavaCC 的综合性介绍是本书的特色之一，目前其他教科书中还没有相关内容。对于自下而上的分析与翻译，本书的介绍比较简洁，但也涵盖了基本分析原理以及自下而

上分析与翻译程序的自动构造工具 Yacc 等重要内容。本书的内容具有很强的实践性，学生可以渐增地实践不同层次编译器的构造。本书的教学支撑软件包特意提供的关于编译器项目的正确性、运行时间以及代码尺寸等的评价功能，有助于增强实践项目的可操作性。

限于译者水平，译文中难免出现各种错误和疏漏，欢迎广大读者批评指正。

前　言

本书的主要目的是清晰地向读者展示编译器设计和实现的理论，并给出大量的机会使理论付诸实践。理论，固然很重要，然而实践也同样重要。

主要特点

- 提供了若干精心准备的实验项目及其测试用例。这些实验项目能使学生不光懂得理论，而且懂得如何去应用理论。实验项目与课本有机结合，使得教师可以摆脱项目设计的负担。
- 实验项目开始于本书较早章节，学生可以边学理论边去应用。
- 编译器工具（JavaCC、Yacc 和 Lex）为可选主题。
- 整个书是面向 Java 的。实现语言是 Java，主要目标语言类似 Java bytecode，编译器工具生成 Java 代码，以及所用的属性文法在形式上采用了类 Java 的语法。
- 目标语言（一个是面向堆栈的，类似 Java bytecode；另一个是面向寄存器的）非常易学，但功能足以支持现代编译器工程。
- 软件包对于学生和老师来说是梦想成真。它可以自动评估学生的编译器实验项目，包括正确性、运行时间和代码尺寸。它的伟大之处，对于学生来说在于：可以立刻获得关于项目工作的反馈信息；对于教师来说在于：可以容易和准确地评估学生的工作。教师用一条命令就可以生成一份整个班级的报告。软件可运行于三个平台：Microsoft Windows、Linux 和 Macintosh OS X。
- 具体展示了编译技术不仅仅只是用于编译器。在一个实验项目的目标里，将要求学生使用编译技术设计和实现 grep 工具。
- 包含一章解释器相关内容，适用于本书后续章节。
- 包含一章优化的内容，作为入门性课程来说恰到好处。学生不只是简单读懂优化技术——而是去实现各种技术，例如，常量合并，窥孔优化，以及寄存器分配。
- 本书使用类 Java 格式的文法，学生可以容易理解和使用。该格式与 JavaCC 所采用的一致。这样，学生可以快捷、容易地切换到 JavaCC。
- 本书含足够多的理论知识，这使得本书适用于将编译/自动机/形式语言等知识混合起来的课程。本书涵盖了自动机/形式语言课程的多数主题：有穷自动机，栈分析器，正规表达式，正规文法，上下文无关文法，上下文有关文法，非受限文法，Chomsky 层次，以及泵引理。下推自动机，图灵机，可计算性，以及复杂性等在软件包中作为补充讨论。软件包中还包括下推自动机模拟器和图灵机模拟器。
- 覆盖了第一次课或编译方面唯一课程应该涉及的各方面话题。学生不仅要学习理论知识和编译器设计的实践技能，而且还要学习重要的系统概念。

软件包

本教材的软件包具有一些不同寻常的特性。当学生运行编译器所生成的程序时，该软件会产生一个日志文件。日志文件包括时间戳，学生姓名，编译器所生成程序的输出，以及针对编译器所生成程序的评价指标（含正确性、程序大小和执行时间等）。如果输出不正确（表明学生的编译器生成了不正确代码），日志文件用 NOT CORRECT 标记。如果编译后的程序太大，或运行时间太长，日志文件将标记 OVER LIMIT。

日志文件的名字中包括学生的姓名。例如，对于姓 Dos Reis 学生的 S3 项目，日志文件名是 S3.dosreis.log。因为每个日志文件名是唯一的，所以指导老师可以将整个班的所有日志文件存储在同一个目录下。仅一条命令便可以生成全班的报告。

该软件支持两种指令集：栈指令集和寄存器指令集。栈指令集是默认的指令集。要使用寄存器指令集，则要在汇编语言源程序中加入一条指导命令。然后，该软件就会自动更新配置，就变成使用寄存器指令集了。

软件包的三个主要程序是 **a**（汇编程序/链接程序），**e**（运行程序），和 **l**（库 make 程序）。软件包中还包括 **p**（下推自动机模拟程序）和 **t**（图灵机模拟程序）。

本书的软件包从出版商可以得到。编译器工具在网上可以得到。在写书时，JavaCC 的网址是 http://java.net/downloads/javacc，Byacc/j 的网址是 http://byaccj.sourceforge.net/，Jflex 的网址是 http://jflex.de/。

实验项目

本教材给出了不少精心准备的实验项目。源语言根据复杂性的递增分六个层次。学生可以对每个层次编写编译器，翻译到栈指令集。学生也可以对每个层次编写编译器，翻译到寄存器指令集，或者配以某些优化技术。对每个层次，学生可以编写纯解释程序，或是使用中间代码的解释程序。学生使用编译技术可以实现若干不同种类的 grep 程序。对于这些项目，学生可以手工编写代码，或者使用 JavaCC 或 Yacc。有许多章节给出的问题中提供了更多的实验题目。总之，本书包括了足够多的实验项目。

对于每个实验项目，本教材提供了实质性的支持。进一步，许多实验项目是前面实验项目的递增版。这种渐增的方式很好，虽然每个项目都是挑战，但不是那种学生力所不及的挑战。

多数实验项目可在一周时间内完成。这样，学生在一个学期中应该可以做十个或更多的实验项目。

有用的参考资料

对于背景材料，读者可以参考作者的 *An Introduction to Programming Using Java*（Jhones & Bartlett，2010）和 *Assembly Language and Computer Architecture Using C++ and Java*（Course Technology，2004），也推荐 JFLAP（网址 http://www.jflap.org），一种交互式

程序，它可以进行各种类型自动机和文法的实验。

致谢

我要感谢 Robert McNaughton 和 Dean Arden 教授，多年前在 RPI 他们为我传授了形式语言理论的优美；感谢我的学生们使用了该书的初始版本并给予有价值的反馈；感谢 Katherine Guillemette 给予这项工作的支持；感谢我女儿 Laura 对本书的内容和结构的建议，以及我妻子的鼓励。

目　录

第 1 章　字符串、语言和编译器

1.1　概　　述

编译器构建其实是一门工程科学，通过这门科学，人们基本上可以用例行的有章可循的方法来设计和实现快速、可靠和强大的编译器。

学习编译器构建技术的几个理由：

- 编译器构建技术有很广的应用领域。这些技术的使用不局限于编译器。
- 为了有效地编程，需要理解编译过程。
- 语言和语言翻译是最核心的计算，应当熟悉相关理论和实践。
- 不像计算机科学的其他领域，一般情况下不会在工作中有机会学会编译器构建技术。因此，正规学习这些技术是很重要的。

1.2　语言的基本概念

编译器设计理论的学习，从几个重要概念开始。**字母表**（alphabet），是书写语言时所用字符的有限集合。例如，Java 程序设计语言的字母表包括所有可以出现在程序中的字符：大写和小写字母、数字、空白（空格符、制表符、换行符以及回车符），以及所有特殊符号，如=、+、和{。本书的大部分例子都使用很小的字母表，如{b，c}和{b，c，d}。本书避免在字母表中使用字母 a，以免与英语文章中的 a 混淆。

字母表上的字符串（string over a alphabet），是字母表中字符的有限序列。例如，假设字母表是{b，c，d}，那么，

```
cbd
cbcc
c
```

是字母表上字符串的例子。注意，在字母表上的字符串中，字母表上的每个字符可以出现任意多次（包括 0 次），且可以是任意顺序。例如，在字符串 cbcc（一个 3 字符字母表{b，c，d}上的字符串）中，字符 b 出现 1 次，字符 c 出现 3 次，字符 d 没有出现。

字符串的长度（length of a string），是字符串包含字符的个数。我们用一对竖线括起一个串，表示串的长度。例如，|cbcc| 表示串 cbcc 的长度，因此，|cbcc| = 4。

语言（language），是某个字母表上字符串的集合。例如，仅包含三个字符串 cbd、cbcc 和 c 的集合是一个语言。这个集合不是一个十分有趣的语言，但根据定义，它是一个语言。

下面看一下，如何将定义用到真实的语言——程序设计语言 Java——上。考虑一个仅写了一行的 Java 程序：