Разработка программ с помощью Objective Caml
Developing Applications with Objective Caml

Emmanuel Chailloux
Pascal Manoury
Bruno Pagano

1 сентября 2012 г.

Оглавление

Глава 1

Как заполучить Objective CAML

Часть I

Основы языка

Глава 2

Функциональное программирование

Глава 3

Императивное программирование

Глава 4

Функциональный и императивный стиль

Глава 5

Графический интерфейс

Глава 6

Приложения

Часть II

Средства разработки

Глава 7

Компиляция и переносимость

Глава 8

Библиотеки

Глава 9

Автоматический сборщик мусора

Глава 10

Средства анализа программ

Глава 11

Средства лексического и синтаксического анализа

11.1 Введение

Определение и реализация средств лексического и синтаксического анализа являлись важным доменом исследования в информатике. Эта работа привела к созданию генераторов лексического и синтаксического анализа `lex` и `yacc`. Команды `camllex camlyacc`, которые мы представим в этой главе, являются их достойными наследниками. Два указанных инструмента стали de–facto стандартными, однако существуют другие средства, как например потоки или регулярные выражения из библиотеки `Str`, которые могут быть достаточны для простых случаев, там где не нужен мощный анализ.

Необходимость подобных инструментов особенно чувствовалась в таких доменах как компиляция языков программирования. Однако и другие программы могут с успехом использовать данные средства: базы данных, позволяющие определять запросы или электронная таблица, где содержимое ячейки можно определить как результат какой–нибудь формулы. Проще говоря, любая программа, в которой взаимодействие с пользователем осуществляется при помощи языка, использует лексический и синтаксический анализ.

Возьмём простой случай. Текстовый формат часто используется для хранения данных, будь то конфигурационный системный файл или данные табличного файла. Здесь, для использования данных, необходим лексический и синтаксический анализ.

Обобщая, скажем что лексический и синтаксический анализ преобразует линейный поток символов в данные с более богатой структурой:

последовательность слов, структура записи, абстрактное синтаксическое дерево программы и т.д.

У каждого языка есть словарный состав (лексика) и грамматика, которая описывает каким образом эти составные объединяются (синтаксис). Для того, чтобы машина или программа могли корректно обрабатывать язык, этот язык должен иметь точные лексические и синтаксические правила. У машины нет «тонкого чувства» для того чтобы правильно оценить двусмысленность натуральных языков. По этой причине к машине необходимо обращаться в соответствии с чёткими правилами, в которых нет исключений. В соответствии с этим, понятия лексики и семантики получили формальные определения, которые будут кратко представлены в данной главе.

11.2 План главы

Данная глава знакомит нас со средствами лексического и синтаксического анализа, которые входят в дистрибутив Objective CAML. Обычно синтаксический анализ следует за лексическим. В первой части мы узнаем о простом инструменте лексического анализа из модуля `Genlex`. После этого ознакомимся с формализмом рациональных выражений и тем самым рассмотрим более детально определение множества лексических единиц. А так же проиллюстрируем их реализацию в модуле `Str` и инструменте `ocamllex`. Во второй части мы определим грамматику и рассмотрим правила создания фраз языка. После этого рассмотрим два анализа фраз: восходящий и нисходящий. Они будут проиллюстрированы использованием `Stream` и `ocamlyacc`. В приведённых примерах используется контекстно–независимая грамматика. Здесь мы узнаем как реализовать контекстный анализ при помощи Stream. В третьей части мы вернёмся к интерпретатору BASIC (см. стр **??**) и при помощи `ocamllex` и `ocamlyacc` добавим лексические и синтаксические функции анализа языка.

11.3 Лексика

Синтаксический анализ это предварительный и необходимый этап обработки последовательностей символов: он разделяет этот поток в последовательность слов, так называемых лексические единицы или лексемы.

11.3.1 Модуль `Genlex`

В данном модуле имеется элементарное средство для анализа символьного потока. Для этого используются несколько категорий предопределённых лексических единиц. Эти категории различаются по типу:

```
# type token = Kwd of string
    | Ident of string
    | Int of int
    | Float of float
    | String of string
    | Char of char ;;
```

Таким образом мы можем распознать в потоке символов целое число (конструктор `Int`) и получить его значение (аргумент конструктора `int`). Распознаваемые символы и строки подчиняются следующим общепринятым соглашением: строка окружена символами ("), а символ окружён (’). Десятичное число представлено либо используя запись с точкой (например 0.01), либо с мантиссой и экспонентой (на пример 1E-2). Кроме этого остались конструкторы `Kwd` и `Ident`.

Конструктор `Ident` предназначен для определения идентификаторов. Идентификатором может быть имя переменной или функции языка программирования. Они состоят из любой последовательности букв и цифр, могут включать символ подчёркивания (_) или апостроф (’). Данная последовательность не должна начинаться с цифры. Любая последовательность следующих операндов тоже будет считаться идентификатором: +,*,> или -. И наконец, конструктор `Kwd` определяет категорию специальных идентификаторов или символов.

Категория ключевых слова — единственная из этого множества, которую можно сконфигурировать. Для того, чтобы создать лексический анализатор, воспользуемся следующей конструкцией, которой необходимо передать список ключевых слов на место первого аргумента.

```
# Genlex.make_lexer ;;
- : string list -> char Stream.t -> Genlex.token Stream.t = <fun>
```

Тем самым получаем функцию, которая принимает на вход поток символов и возвращает поток лексических единиц (с типом `token`).

Таким образом, мы без труда реализуем лексический анализатор для интерпретатора BASIC. Объявим множество ключевых слов:

```
# let keywords =
    [ "REM"; "GOTO"; "LET"; "PRINT"; "INPUT"; "IF"; "THEN"; "-";
    "!"; "+"; "-"; "*"; "/"; "%";
```

```
    "="; "<"; ">"; "<="; ">="; "<>";
    "&"; "|" ] ;;
```

При помощи данного множества, определим функцию лексического анализа:

```
# let line_lexer l = Genlex.make_lexer keywords (Stream.of_string l) ;;
val line_lexer : string -> Genlex.token Stream.t = <fun>
# line_lexer "LET x = x + y * 3" ;;
- : Genlex.token Stream.t = <abstr>
```

Приведённая функция `line_lexer`, из входящего потока символов создаёт поток соответствующих лексем.

11.3.2 Использование потоков

Мы также можем реализовать лексический анализ «в ручную» используя потоки.

В следующем примере определён лексический анализатор арифметических выражений. Функции lexer передаётся поток символов из которого она создаёт поток лексических единиц с типом `lexeme Stream.t` [1]. Символы пробел, табуляция и переход на новую строку удаляются. Для упрощения, мы не будем обрабатывать переменные и отрицательны целые числа.

```
# let rec spaces s =
    match s with parser
      [<'' ' ; rest >] -> spaces rest
      | [<''\t' ; rest >] -> spaces rest
      | [<''\n' ; rest >] -> spaces rest
      | [<>] -> ();;
val spaces : char Stream.t -> unit = <fun>
# let rec lexer s =
   spaces s;
   match s with parser
     [< ''(' >] -> [< 'Lsymbol "("; lexer s >]
     | [< '')' >] -> [< 'Lsymbol ")"; lexer s >]
     | [< ''+' >] -> [< 'Lsymbol "+" ; lexer s >]
     | [< ''-' >] -> [< 'Lsymbol "-" ; lexer s >]
     | [< ''*' >] -> [< 'Lsymbol "*" ; lexer s >]
     | [< ''/' >] -> [< 'Lsymbol "/" ; lexer s >]
```

[1] тип `lexeme` определён на стр. **??**

```
    | [< ''0'..'9' as c;
      i,v = lexint (Char.code c - Char.code('0')) >]
      ->[<'Lint i ; lexer v>]
  and lexint r s =
    match s with parser
      [< ''0'..'9' as c >]
      -> let u = (Char.code c) - (Char.code '0') in lexint (10*r + u) s
      | [<>] -> r,s ;;
val lexer : char Stream.t -> lexeme Stream.t = <fun>
val lexint : int -> char Stream.t -> int * char Stream.t = <fun>
```

Функция `lexint` предназначена для анализа той части потока символов, которая соответствует числовой постоянной. Она вызывается, когда функция `lexer` встречает цифры. В этом случае функция `lexint` поглощает все последовательные цифры и выдаёт соответствующее значение полученного числа.

11.3.3 Регулярные выражения

Оставим ненадолго практику и рассмотрим проблему лексических единиц с теоретической точки зрения.

Лексическая единица является словом. Слово образуется при конкатенации элементов алфавита. В нашем случае алфавитом является множество символов ASCII.

Теоретически, слово может вообще не содержать символов (пустое слово [2]) или состоять из одного символа.

Теоретические исследования конкатенации элементов алфавита для образования лексических элементов (лексем) привели к созданию простого формализма, известного как регулярные выражения.

Определение

Регулярные выражения позволяют определить множества слов. Пример такого множества: идентификаторы. Принцип определения основан на некоторых теоретико–множественный операциях. Пусть M и N — два множества слов, тогда мы можем определить:

- объединение M и N, записываемое $M|N$.
- дополнение M, записываемое $\wedge M$: множество всех слов, кроме тех, которые входят в M.

[2] традиционно, такое слово обозначается греческой буквой эпсилон: ε

- конкатенация M и N: множество всех слов созданных конкатенацией слова из M и слова из N. Записывается просто MN.

- мы можем повторить операцию конкатенации слов множества M и тем самым получить множество слов образованных из конечной последовательности слов множеств M. Такое множество записывается $M+$. Он содержит все слова множества M, все слова полученные конкатенацией двух слов множества M, трёх слов, и т.д. Если мы желаем чтобы данное множество содержало пустое слово, необходимо писать $M*$.

- для удобства, существует дополнительная конструкция $M?$, которая включает все слова множества M, а так же пустое слово.

Один единственный символ ассоциируется с одно элементным множеством. В таком случае выражение *a| b| c* описывает множество состоящее из трёх слов *a*, *b* и *c*. Существует более компактная запись: *[abc]*. Так как наш алфавит является упорядоченным (по порядку кодов ASCII), можно определить интервал. Например, множество цифр запишется как $[0-9]$. Для группировки выражений можно использовать скобки.

Для того, чтобы использовать в записи сами символы–операторы, как обычные символы, необходимо ставить перед ними обратную косую черту: \. Например множество $(\backslash *)*$ обозначает множество последовательностей звёздочек.

Пример

Пусть существует множество из цифр (0,1,2,3,4,5,6,7,8,9), символы плюс $(+)$ и минус $(-)$, точки (.) и буквы E. Теперь мы можем определить множество чисел *num*. Назовем *integers* множество определённое выражением $[0-9]+$. Множество неотрицательных чисел *unum* определяется так:

$$integers?(.integers)?(E(\backslash + | -)?integers)?$$

Множество отрицательных и положительных чисел записывается:

$$unum| - unum \quad \text{или} \quad -?unum$$

Распознавание

Теперь, после того как множество выражений определено, остаётся проблема распознавания принадлежности строки символов или одной из её

подстрок этому множеству. Для решения данной задачи необходимо реализовать программу обработки выражений, которая соответствует формальным определениям множества. Для регулярных выражений такая обработка может быть автоматизирована. Подобная автоматизация реализована в модуле `Genlex` из библиотеки `Str` и инструментом `ocamllex`, которые будут представлены в следующих двух параграфах.

Библиотека `Str`

В данном модуле имеется абстрактный тип `regexp` и функция `regexp`. Указанный тип представляет регулярные выражения, а функция трансформирует регулярное выражение, представленное в виде строки символов, в абстрактное представление.

Модуль `Str` содержит несколько функций, которые используют регулярные выражения и манипулируют символьными строками. Синтаксис регулярных выражений библиотеки `Str` приведён в таблице 11.1.

Таблица 11.1: Регулярные выражения

.	любой символ, кроме \
$*$	ноль или несколько экземпляров предыдущего выражения
$+$	хотя бы один экземпляр предыдущего выражения
?	ноль или один экземпляр предыдущего выражения
[..]	множество символов
.	любой символ, кроме \
	интервал записывается при помощи - (пример $[0-9]$)
	дополнение записывается при помощи $\wedge$ (пример $[\wedge A-Z]$)
$\wedge$	начало строки (не путать с дополнением $\wedge$)
$	конец строки
|	вариант
(..)	группировка в одно выражение (можно ссылаться на это выражение)
i	числовая константа i ссылается на i–ый элемент группированного выражения
\	забой, используется для сопоставления зарезервированных символов в регулярных выражениях

Пример

В следующем примере напишем функцию, которая переводит дату из английского формата во французский. Предполагается, что входной файл

состоит из строк, разбитых на поля данных и элементы даты разделяются точкой. Определим функцию, которая из полученной строки (строка файла), выделяет дату, разбивает её на части, переводит во французских формат и тем самым заменяет старую дату на новую.

```
# let french_date_of d =
  match d with
   [mm; dd; yy] -> dd^"/"^mm^"/"^yy
   | _ -> failwith "Bad date format" ;;
val french_date_of : string list -> string = <fun>

# let english_date_format = Str.regexp "[0-9]+\.[0-9]+\.[0-9]+" ;;
val english_date_format : Str.regexp = <abstr>

# let trans_date l =
  try
  let i=Str.search_forward english_date_format l 0 in
  let d1 = Str.matched_string l in
  let d2 = french_date_of (Str.split (Str.regexp "\.") d1) in
    Str.global_replace english_date_format d2 l
  with Not_found -> l ;;
  val trans_date : string -> string = <fun>

# trans_date
"..............06.13.99............" ;;
- : string = "..............13/06/99............"
```

11.3.4 Инструмент `Ocamllex`

`ocamllex` — это лексический генератор созданный для Objective CAML по модели `lex`, написанном на языке C. При помощи файла, описывающего элементы лексики в виде множества регулярных выражений, которые необходимо распознать, он создаёт файл–исходник на Objective CAML. К описанию каждого лексического элемента можно привязать какое–нибудь действие, называемое семантическое действие. В полученном коде используется абстрактный тип `lexbuf` из модуля `Lexing`. В данном модуле также имеется несколько функций управления лексическими буферами, которые могут быть использованы программистом для того чтобы определить необходимые действия.

Обычно, файлы, описывающие лексику, имеют расширение `.mll`. Для того, чтобы из файла `lex_file.mll` получить файл на Objective CAML,

необходимо выполнить следующую команду:

```
ocamllex lex_file.mll
```

После этого, мы получим файл `lex_file.ml`, содержащий код лексического анализатора. Теперь, данный файл можно использовать в программе на Objective CAML. Каждому множеству правил анализа соответствует функция, которая принимает лексический буфер (типа `Lexing.lexbuf`) и затем возвращает значение, определённое семантическим действием. Значит, все действия для определённого правила должны создавать значение одного и того же типа.

Формат у файла для `ocamllex` следующий:

```
{
  header
}
let ident = regexp
          ...
rule ruleset1 = parse
                  regexp { action }
                | ...
                | regexp { action }
and ruleset2 = parse
          ...
and ...
{
  trailer-and-end
}
```

Части «заголовок» и «продолжение-и-конец» не являются обязательными. Здесь вставляется код Objective CAML, определяющий типы данных, функции и т.д. необходимые для обработки данных. В последней части используются функции, которые используют правила анализа множества лексических данных из средней части. Серия объявлений, которая предшествует определению правил, позволяет дать имя некоторым регулярным выражениям. Эти имена будут использоваться в определении правил.

Пример

Вернёмся к нашему интерпретатору BASIC и усовершенствуем тип возвращаемых лексических единиц. Таким образом, мы можем воспользо-

ваться функцией `lexer`, (см. стр. ??) которая возвращает такой же тип результата (`lexeme`), но на входе она получает буфер типа `Lexing.lexbuf`.

```
  {
    let string_chars s =
    String.sub s 1 ((String.length s)-2) ;;
  }

let op_ar = ['-' '+' '*' '\%' '/']
let op_bool = ['!' '\&' '|']
let rel = ['=' '<' '>']

rule lexer = parse
    [' '] { lexer lexbuf }
  | op_ar { Lsymbol (Lexing.lexeme lexbuf) }
  | op_bool { Lsymbol (Lexing.lexeme lexbuf) }
  | "<=" { Lsymbol (Lexing.lexeme lexbuf) }
  | ">=" { Lsymbol (Lexing.lexeme lexbuf) }
  | "<>" { Lsymbol (Lexing.lexeme lexbuf) }
  | rel { Lsymbol (Lexing.lexeme lexbuf) }
  | "REM" { Lsymbol (Lexing.lexeme lexbuf) }
  | "LET" { Lsymbol (Lexing.lexeme lexbuf) }
  | "PRINT" { Lsymbol (Lexing.lexeme lexbuf) }
  | "INPUT" { Lsymbol (Lexing.lexeme lexbuf) }
  | "IF" { Lsymbol (Lexing.lexeme lexbuf) }
  | "THEN" { Lsymbol (Lexing.lexeme lexbuf) }
  | '-'? ['0'-'9']+ { Lint (int_of_string (Lexing.lexeme lexbuf)) }
  | ['A'-'z']+ { Lident (Lexing.lexeme lexbuf) }
  | '"' [^ '"']* '"' { Lstring (string_chars (Lexing.lexeme lexbuf
    )) }
```

После обработки данного файла командой `ocamllex` получим функцию `lexer` типа `Lexing.lexbuf -> lexeme`. Далее, мы рассмотрим каким образом подобные функции используются в синтаксическом анализе (см. стр. ??).

11.4 Синтаксис

Благодаря лексическому анализу, мы в состоянии разбить поток символов на более структурированные элементы: лексические элементы. Теперь необходимо знать как правильно объединять эти элементы в син-

таксически корректные фразы какого-нибудь языка. Правила синтаксической группировки определены посредством грамматических правил. Формализм, произошедший из лингвистики, был с успехом перенят математиками, занимающимися теориями языков, и специалистами по информатике. На странице **??** мы уже видели пример грамматики для языка BASIC. Здесь мы снова вернёмся к этому примеру, чтобы более углублённо ознакомится с базовыми концепциями грамматики.

11.4.1 Грамматика

Говоря формальным языком, грамматика основывается на четырёх элементах:

1. Множество символов, называемых терминалами. Эти символы являются лексическими элементами языка. В Бэйсике к ним относятся символы операторов, арифметических отношений и логические ($+$, &, $<$, $\leq$, $\cdots$), ключевые слова языка **(GOTO, PRINT, IF, THEN,** $\cdots$**)**, целые числа (элемент `integer`) и переменные (элемент `variable`).

2. Множество нетерминальных символов, которые представляют синтаксические компоненты языка. Например, программа на языке Бэйсик состоит из строк. Таким образом мы имеем компоненту Line, которая в свою очередь состоит из выражений (Expression), и т.д.

3. Множество так называемых порождающих правил. Они описывают каким образом комбинируются терминальные и нетерминальные символы, чтобы создать синтаксическую компоненту. Строка в Бэйсике начинается с номера, за которой следует инструкция. Смысл правила следующий:

   $$\text{Line} ::= \textit{integer}\ \text{Instruction}$$

   Одна и та же компонента может быть порождена несколькими способами. В этом случае варианты разделяются символом | как в:

   $$\begin{array}{rcl} \text{Instruction} & ::= & \text{LET variable} = \text{Instruction} \\ & | & \text{GOTO integer} \\ & | & \text{PRINT Expression} \\ \text{etc} & & \end{array}$$

Среди всех нетерминальных символов различают один, называемый начальным. Правило, которое порождает эту аксиому является порождающим правилом всего языка.

11.4.2 Порождение и распознавание

С помощью порождающих правил можно определить принадлежит ли последовательность лексем языку.

Рассмотрим простой язык, описывающий арифметические выражения:

Exp	::=	*integer*	(R1)
	|	Exp + Exp	(R2)
	|	Exp * Exp	(R3)
	|	(Exp)	(R4)

здесь *(R1) (R2) (R3) (R4)* — имена правил. По окончании лексического анализа выражение $1 * (2 + 3)$ становится последовательностью следующих лексем:

*integer * (integer + integer)*

Для того, чтобы проанализировать эту фразу и убедиться в том что она принадлежит языку арифметических выражений, воспользуемся правилами справа налево: если часть выражения соответствует правому члену какого-нибудь правила, мы заменяем это выражение соответствующим левым членом. Этот процесс продолжается до тех пор, пока выражение не будет редуцировано до аксиомы. Ниже представлен результат такого анализа [3]:

integer * (integer + integer)	(R1)	Exp * (integer + integer)
	(R1)	Exp * (Exp + integer)
	(R1)	Exp * (Exp + Exp)
	(R2)	Exp * (Exp)
	(R4)	Exp * Exp
	(R3)	Exp

Начиная с последней линии, которая содержит лишь Exp и следуя стрелкам, можно определить каким образом было полученное выражение исходя из аксиомы Exp. Соответственно данная фраза является правильно сформированной фразой языка арифметических выражений, определённого грамматикой.

[3] анализируемая часть выражения подчёркнута, а так же указано используемое правило

Преобразование грамматики в программу, способную распознать принадлежность последовательности лексем языку, который определён грамматикой, является более сложной проблемой, чем проблема использования регулярных выражений. Действительно, в соответствии с математическим результатом любое множество (слов), определённое формализмом регулярных выражений, может быть определено другим формализмом: детерминированные конечные автоматы. Такие автоматы легко реализуются программами, принимающими поток символов. Подобный результат для грамматик в общем не существует. Однако, имеются менее строгие (?) (weaker) результаты устанавливающие эквивалентность между определёнными классами грамматик и более богатыми автоматами: автомат со стеком. Здесь мы не станем вдаваться ни в детали этих результатов, ни в точное определение таких автоматов. Однако, мы можем определить какие классы грамматики могут использоваться в средствах генерации синтаксических анализаторов или для реализации напрямую анализатора.

11.4.3 Нисходящий анализ

Разбор выражения $1 * (2 + 3)$ в предыдущем параграфе не является единственным: мы с таким же успехом могли бы начать редуцирование *integer*, то есть воспользоваться правилом *(R2)* редуцирования $2 + 3$. Эти два способа распознавания являются двумя типами анализа: восходящий анализ (справа налево) и нисходящий слева направо. Последний анализ легко реализуется при помощи потоков лексем модуля `Stream`. Средство `ocamlyacc` использует восходящий анализ, при котором применяется стек, как это уже было проиллюстрировано синтаксическим анализатором программ на Бэйсике. Выбор анализа не просто дело вкуса, в зависимости от используемой для спецификации языка формы грамматики, можно или нет применять нисходящий анализ.

Простой случай

Каноническим примером нисходящего анализа является префиксная запись арифметических выражений, определяемая как:

$$\begin{array}{rcl} \text{Exp} & ::= & \text{integer} \\ & | & + \text{ Exp Exp} \\ & | & * \text{ Exp Exp} \end{array}$$

В данном случае достаточно знать первую лексему, для того чтобы определить какое правило может быть использовано. При помощи по-

добной предсказуемости нет необходимости явно управлять стеком, достаточно положиться на рекурсивный вызов анализатора. И тогда при помощи `Genlex` и `Stream` очень просто написать программу реализующую нисходящий анализ. Функция `infix_of` из полученного префиксного выражения возвращает его инфиксный эквивалент:

```
# let lexer s =
    let ll = Genlex.make_lexer ["+";"*"]
    in ll (Stream.of_string s);;
val lexer : string -> Genlex.token Stream.t = <fun>

# let rec stream_parse s =
    match s with parser
      [<'Genlex.Ident x>] -> x
    | [<'Genlex.Int n>] -> string_of_int n
    | [<'Genlex.Kwd "+"; e1=stream_parse; e2=stream_parse>] -> "("^
        e1^"+"^e2^")"
    | [<'Genlex.Kwd "*"; e1=stream_parse; e2=stream_parse>] -> "("^
        e1^"*"^e2^")"
    | [<>] -> failwith "Parse error" ;;
val stream_parse : Genlex.token Stream.t -> string = <fun>

# let infix_of s = stream_parse (lexer s) ;;
val infix_of : string -> string = <fun>

# infix_of "* +3 11 22";;
- : string = "((3+11)*22)"
```

Однако не стоит забывать о некоторой примитивности лексического анализа. Советуем периодически добавлять пробелы между различными лексическими элементами.

```
# infix_of "*+3 11 22";;
- : string = "*+"
```

Случай посложней

Синтаксический анализ при помощи потоков предсказуем, он облагает грамматику двумя условиями:

- В правилах грамматики не должно быть левой рекурсии. Правило называется рекурсивным слева, если его правый член начинается с

нетерминального символа, который является левой частью правила. Например: `Exp ::= Exp + Exp`

- Не должно существовать правил начинающихся одним и тем же выражением.

Грамматика арифметических выражений, приведённая на стр. 40, не подходит для нисходящего анализа: они не удовлетворяют ни одному из условий. Для того, чтобы применить нисходящий анализ необходимо переформулировать грамматику таким образом, чтобы удалить левую рекурсию и неопределённость правил. Вот полученный результат:

$$
\begin{array}{rcl}
\text{Exp} & ::= & \text{Atom NextExpr} \\
\text{NextExpr} & ::= & + \text{ Atom} \\
 & | & - \text{ Atom} \\
 & | & * \text{ Atom} \\
 & | & / \text{ Atom} \\
 & | & \varepsilon \\
\text{Atom} & ::= & \text{integer} \\
 & | & (\text{ Expr })
\end{array}
$$

Заметьте использование пустого слова ε в определении `NextExpr`. Оно необходимо, если мы хотим чтобы просто целое число являлось выражением.

Следующий анализатор есть просто перевод вышеуказанной грамматики в код. Он реализует абстрактное синтаксическое дерево арифметических выражений.

```
# let rec rest = parser
    [< 'Lsymbol "+"; e2 = atom >] -> Some (PLUS,e2)
  | [< 'Lsymbol "-"; e2 = atom >] -> Some (MINUS,e2)
  | [< 'Lsymbol "*"; e2 = atom >] -> Some (MULT,e2)
  | [< 'Lsymbol "/"; e2 = atom >] -> Some (DIV,e2)
  | [< >] -> None
 and atom = parser
     [< 'Lint i >] -> ExpInt i
   | [< 'Lsymbol "("; e = expr ; 'Lsymbol ")" >] -> e
 and expr s =
   match s with parser
     [< e1 = atom >] ->
       match rest s with
           None -> e1
```

```
        | Some (op,e2) -> ExpBin(e1,op,e2) ;;
val rest : lexeme Stream.t -> (bin_op * expression) option = <fun>
val atom : lexeme Stream.t -> expression = <fun>
val expr : lexeme Stream.t -> expression = <fun>
```

Сложность использования нисходящего анализа заключается в том, что грамматика должна быть очень ограниченной формы. Если язык выражен естественно с использованием левой рекурсии (как в инфиксных выражениях), то не всегда легко определить эквивалентную грамматику, то есть определяющую такой же язык, которая бы удовлетворяла требованиям нисходящего анализа. По этой причине, средства `yacc` и `ocamlyacc` реализуют восходящий анализ, который разрешает определение более естественных грамматик. Однако, мы увидим, что даже в этом случае существуют ограничения.

11.4.4 Восходящий анализ

Мы уже вкратце представили на странице **??** принципы восходящего анализа: сдвиг и вывод [4]. После каждого подобного действия, состояние стека изменяется. Из этой последовательности можно вывести правила грамматики, в случае если грамматика это позволяет, как в примере с нисходящим анализом. Опять же, сложности возникают из-за неопределённости правил, когда невозможно выбрать между продвинутся или сократить. Проиллюстрируем действие восходящего анализа и его недостатки на все тех же арифметических выражениях в постфиксном и инфиксном написании.

Положительная сторона

Упрощённая постфиксная грамматика арифметических выражений выглядит так:

$$\begin{array}{lcll} \text{Exp} & ::= & \textit{integer} & \text{(R1)} \\ & | & \text{Exp Exp +} & \text{(R2)} \\ & | & \text{Exp Exp -} & \text{(R3)} \end{array}$$

Данная грамматика является двойственной по отношению к префиксной: для того чтобы точно знать какое правило следует применить, необходимо дождаться окончания анализа. В действительности анализ по-

[4] Обычно на русский язык термины **shift** и **reduce** переводят как сдвиг и вывод, но в данной книге для термина **reduce** будет также использоваться перевод «сокращение» — прим. пер.

добных выражений схож с вычислением при помощи стека. Только вместо проталкивания результата вычисления, проталкиваются грамматические символы. Если в начале стек пустой, то после того как ввод закончен, необходимо получить стек содержащий лишь нетерминальную аксиому. Приведём изменение стека: если мы продвигаемся, то проталкивается текущий нетерминальный символ; если сокращаем, то первые символы стека соответствуют правому члену (в обратном порядке) правила и тогда мы заменяем эти элементы соответствующими нетерминальными элементами.

В таблице 11.2 приведён восходящий анализ выражения 1 2 + 3 * 4 +. Считываемая лексическая единица подчёркивается для более удобного чтения. Конец потока помечается символом $.

Таблица 11.2: Восходящий анализ

Действие	Вход	Стек
	1 2 + 3 * 4 + $	[]
Сдвиг		
	2 + 3 * 4 + $	[1]
Сократить *(R1)*		
	2 + 3 * 4 + $	[Exp]
Сдвиг		
	+ 3 * 4 + $	[2 Exp]
Сдвиг, Сократить *(R1)*		
	+ 3 * 4 + $	[Exp Exp]
Сдвиг, Сократить *(R2)*		
	3 * 4 + $	[Exp]
Сдвиг, Сократить *(R1)*		
	* 4 + $	[Exp Exp]
Сдвиг, Сократить *(R3)*		
	4 + $	[Exp]
Сдвиг, Сократить *(R1)*		
	+ $	[Exp Exp]
Сдвиг, Сократить *(R2)*		
	$	[Exp]

Отрицательная сторона

Вся трудность перехода от грамматики к программе распознающей язык заключается в определении действия, которое необходимо применить. Проиллюстрируем эту проблему на трёх примерах, приводящих к трём неопределённостям.

Первый пример есть грамматика выражений использующих операцию сложения:

$$\begin{array}{lcll} \text{E0} & ::= & \textit{integer} & \text{(R1)} \\ & | & \text{E0} + \text{E0} & \text{(R2)} \end{array}$$

Неопределённость данной грамматики проявляется при использовании правила *R2*. Предположим следующую ситуацию:

Действие	Вход	Стек
:		
	$\underline{+}$	E0 + E0 ...
:		

В подобном случае невозможно определить необходимо сдвинуть и протолкнуть в стек + или сократить в соответствии с правилом *(R2)* оба E0 и присутствующий в стеке +. Подобная ситуация называется конфликтом сдвиг-вывод (shift/reduce). Она является следствием того, что выражение *integer* + *integer* + *integer* может быть выведено справа двумя способами.

Первый вариант:

$$\begin{array}{ll} \text{E0} & (\textit{R2})\ \text{E0} + \underline{\text{E0}} \\ & (\textit{R1})\ \underline{\text{E0}} + \textit{integer} \\ & (\textit{R2})\ \text{E0} + \underline{\text{E0}} + \textit{integer} \end{array}$$

Второй вариант:

$$\begin{array}{ll} \text{E0} & (\textit{R2})\ \text{E0} + \underline{\text{E0}} \\ & (\textit{R1})\ \text{E0} + \text{E0} + \underline{\text{E0}} + \textit{integer} \\ & (\textit{R2})\ \text{E0} + \underline{\text{E0}} + \textit{integer} \end{array}$$

Выражения, полученные двумя выводами, могут показаться одинаковыми с точки зрения вычисления выражения,

(*integer* + *integer*) + *integer* и *integer* + (*integer* + *integer*)

но разными для конструкции синтаксического дерева (см. рис. **??** на стр. **??**)

Второй пример грамматики, порождающей конфликт между сдвиг-вывод, содержит такую же неопределённость: явное заключение в скобки. Но в отличии от предыдущего случая, выбор сдвиг-вывод изменяет смысл выражения. Пусть есть грамматика:

$$\begin{array}{lcl} \text{E1} & ::= & \textit{integer}\ (\textit{R1}) \\ & | & \text{E1} + \text{E1}\ (\textit{R2}) \\ & | & \text{E1} * \text{E1}\ (\textit{R3}) \end{array}$$

Здесь мы снова получаем предыдущий конфликт как в случае с + так и для *, но к этому добавляется другой, между + и *. Опять же, одно и то же выражение может быть получено двумя способами, так как у него существует два вывода справа:

integer + *integer* * *integer*

Первый вариант:

$$\begin{array}{ll} \text{E1} & (\textit{R3})\ \text{E0} * \underline{\text{E1}} \\ & (\textit{R1})\ \underline{\text{E1}} * \textit{integer} \\ & (\textit{R2})\ \text{E1} + \underline{\text{E1}} * \textit{integer} \end{array}$$

Второй вариант:

$$\begin{array}{ll} \text{E1} & (\textit{R2})\ \text{E1} + \underline{\text{E1}} \\ & (\textit{R2})\ \text{E1} + \text{E1} * \underline{\text{E1}} \\ & (\textit{R1})\ \text{E1} + \underline{\text{E1}} * \textit{integer} \end{array}$$

В данном выражении обе пары скобок имеют разный смысл:

(*integer* + *integer*) * *integer* $\neq$ *integer* + (*integer* * *integer*)

Подобную проблему, мы уже встречали в выражениях Basic (см. стр. **??**). Она была разрешена при помощи приоритетов, которые присваиваются операторам: сначала редуцируется правило (R3), затем (R2), что соответствует заключению в скобки произведения.

Данную проблему выбора между + и * можно решить изменив грамматику. Для того, введём два новых терминальных символа: член T (*term*) и множитель F (*factor*). Отсюда получаем:

$$\begin{array}{lcll} \text{E} & ::= & \text{E} + \text{T} & (\textit{R1}) \\ & | & \text{T} & (\textit{R2}) \\ \text{T} & ::= & \text{T} + \text{F} & (\textit{R3}) \\ & | & \text{F} & (\textit{R4}) \\ \text{F} & ::= & \text{T} + \textit{integer} & (\textit{R5}) \end{array}$$

После этого, единственный способ получить *integer* + *integer* * *integer*: посредством правила (*R1*).

Третий и последний случай касается условных конструкций языка программирования. На пример в Pascal существует две конструкции: `if .. then` и `if .. then .. else`. Пусть существует следующая грамматика:

$$\begin{array}{lcll} \text{Instr} & ::= & \textit{if}\ \text{EXP}\ \textit{then}\ \text{Instr} & (R1) \\ & - & \textit{if}\ \text{EXP}\ \textit{then}\ \text{Instr}\ \textit{else}\ \text{Instr} & (R2) \\ & - & \text{etc}\ \ldots & \end{array}$$

И в следующей ситуации:

Действие	Вход	Стек
:		
	else	[Instr *then* Epx *if* ...]

Невозможно определить соответствуют ли элементы стека правила (*R1*) и в этом случае необходимо сократить или соответствуют первому `Instr` правила (*R2*) и тогда необходимо сдвинуть.

Кроме конфликтов сдвиг-вывод, восходящий анализ вызывает конфликт вывод-вывод.

Мы представим здесь инструмент `ocamlyacc`, который использует подобную технику может встретить указанные конфликты.

11.5 Пересмотренный Basic

Теперь, используя совместно `ocamllex` и `ocamlyacc`, заменим функцию `parse` для Бэйсика, приведённую на странице **??**, на функции полученные при помощи файлов спецификации лексики и синтаксиса языка.

Для этого, мы не сможем воспользоваться типами лексических единиц, в таком виде как они были определены. Необходимо определить более точные типы, чтобы различать операторы, команды и ключевые слова.

Так же, нам понадобится изолировать в отдельном файле `basic_types.mli` декларации типов, относящиеся к абстрактному синтаксису. В нем будут содержатся декларации типа `sentences`, а так же других типы необходимые этому.

11.5.1 Файл `basic_parser.mly`

Заголовок

Данный файл содержит вызовы деклараций типов абстрактного синтаксиса и две функции перевода строк символов в их эквивалент абстрактного синтаксиса.

```
%{
open Basic_types ;;

let phrase_of_cmd c =
 match c with
   "RUN" -> Run
 | "LIST" -> List
 | "END" -> End
 | _ -> failwith "line : unexpected command"
;;

let bin_op_of_rel r =
 match r with
   "=" -> EQUAL
 | "<" -> INF
 | "<=" -> INFEQ
 | ">" -> SUP
 | ">=" -> SUPEQ
 | "<>" -> DIFF
 | _ -> failwith "line : unexpected relation symbol"
;;

%}
```

Декларации

Здесь содержится три части: декларация лексем, декларация правил ассоциативности и приоритетов, декларация стартового символа `line`, которая соответствует анализу линии программы или команды.

Ниже представлены лексические единицы:

```
%token <int> Lint
%token <string> Lident
%token <string> Lstring
```

```
%token <string> Lcmd
%token Lplus Lminus Lmult Ldiv Lmod
%token <string> Lrel
%token Land Lor Lneg
%token Lpar Rpar
%token <string> Lrem
%token Lrem Llet Lprint Linput Lif Lthen Lgoto
%token Lequal
%token Leol
```

Имена деклараций говорят сами за себя и они описаны в файле `basic_lexer.mll` (см. стр. **??**).

Правила приоритета операторов схожи со значениями, которые определяются функциями `priority_uop` и `priority_binop`, которые были определены грамматикой Бейсика (см. стр. **??**).

```
%right Lneg
%left Land Lor
%left Lequal Lrel
%left Lmod
%left Lplus Lminus
%left Lmult Ldiv
%nonassoc Lop
```

Символ `Lop` необходим для обработки унарных минусов. Он не является терминальным символом, а «псевдо–терминальным». Благодаря этому, получаем перегрузку операторов, когда в двух случаях использования одного и того же оператора, приоритет меняется в зависимости от контекста. Мы вернёмся к этому случаю, когда будем рассматривать правила грамматики.

Здесь нетерминалом является `line`. Полученная функция возвращает дерево абстрактного синтаксиса, которое соответствует проанализированной линии.

```
%start line
%type <Basic_types.phrase> line
```

Правила грамматики

Грамматика делится на 3 нетерминальных элемента: `line` для линии, `inst` для инструкции и `exp` для выражений. Действия, которые привязаны к каждому правилу лишь создают соответствующую часть абстрактного синтаксиса.

```
 %%
line :
    Lint inst Leol { Line {num=$1; inst=$2} }
  | Lcmd Leol { phrase_of_cmd $1 }
  ;

inst :
    Lrem { Rem $1 }
  | Lgoto Lint { Goto $2 }
  | Lprint exp { Print $2 }
  | Linput Lident { Input $2 }
  | Lif exp Lthen Lint { If ($2, $4) }
  | Llet Lident Lequal exp { Let ($2, $4) }
  ;

exp :
    Lint { ExpInt $1 }
  | Lident { ExpVar $1 }
  | Lstring { ExpStr $1 }
  | Lneg exp { ExpUnr (NOT, $2) }
  | exp Lplus exp { ExpBin ($1, PLUS, $3) }
  | exp Lminus exp { ExpBin ($1, MINUS, $3) }
  | exp Lmult exp { ExpBin ($1, MULT, $3) }
  | exp Ldiv exp { ExpBin ($1, DIV, $3) }
  | exp Lmod exp { ExpBin ($1, MOD, $3) }
  | exp Lequal exp { ExpBin ($1, EQUAL, $3) }
  | exp Lrel exp { ExpBin ($1, (bin_op_of_rel $2), $3) }
  | exp Land exp { ExpBin ($1, AND, $3) }
  | exp Lor exp { ExpBin ($1, OR, $3) }
  | Lminus exp %prec Lop { ExpUnr(OPPOSITE, $2) }
  | Lpar exp Rpar { $2 }
  ;
  %%
```

Данные правила не нуждаются в особых комментариях, кроме следующего:

```
exp :
  ...
  | Lminus exp %prec Lop { ExpUnr(OPPOSITE, $2) }
```

Это правило касается использования унарного минуса -. Ключевое слово `%prec` означает, что указанная конструкция получает приоритет от `Lop` (в данном случае наивысший).

11.5.2 Файл `basic_lexer.mll`

Лексический анализ содержит лишь одно множество: `lexer`, которое точно соответствует старой функции `lexer` (см. стр. **??**).

Семантическое действие, которое связано с распознаванием лексических единиц, возвращает результат соответствующего конструктора. Необходимо загрузить файл синтаксических правил, так как в нем декларируется тип лексических единиц. Добавим так же функцию, которая удаляет кавычки вокруг строк.

```
{
 open Basic_parser ;;

 let string_chars s = String.sub s 1 ((String.length s)−2) ;;
}

rule lexer = parse
    [' ' '\t'] { lexer lexbuf }

 | '\n' { Leol }

 | '!' { Lneg }
 | '&' { Land }
 | '|' { Lor }
 | '=' { Lequal }
 | '%' { Lmod }
 | '+' { Lplus }
 | '−' { Lminus }
 | '*' { Lmult }
 | '/' { Ldiv }

 | ['<' '>'] { Lrel (Lexing.lexeme lexbuf) }
 | "<=" { Lrel (Lexing.lexeme lexbuf) }
 | ">=" { Lrel (Lexing.lexeme lexbuf) }

 | "REM" [^ '\n']* { Lrem (Lexing.lexeme lexbuf) }
 | "LET" { Llet }
```

```
| "PRINT" { Lprint }
| "INPUT" { Linput }
| "IF" { Lif }
| "THEN" { Lthen }
| "GOTO" { Lgoto }

| "RUN" { Lcmd (Lexing.lexeme lexbuf) }
| "LIST" { Lcmd (Lexing.lexeme lexbuf) }
| "END" { Lcmd (Lexing.lexeme lexbuf) }

| ['0'-'9']+ { Lint (int_of_string (Lexing.lexeme lexbuf)) }
| ['A'-'z']+ { Lident (Lexing.lexeme lexbuf) }
| '"' [^ '"']* '"' { Lstring (string_chars (Lexing.lexeme lexbuf)
    ) }
```

Заметьте, что нам пришлось изолировать символ =, который используется одновременно в выражениях и приравниваниях.

Для двух рациональных выражений необходимо привести определённые объяснения. Линия комментариев соответствует выражению `("REM"[^'\n']*)`, где за ключевым словом `REM` следует какое угодно количество символов и затем перевод строки. Правило, которое соответствует символьным строкам, `('\ [^ '\]* '\)`, подразумевает последовательность символов, отличных от "и заключённых в кавычки ".

11.5.3 Компиляция, компоновка

Компиляция должна быть реализована в определённом порядке. Это связано с взаимозависимостью деклараций лексем. Поэтому в нашем случае, необходимо выполнить команды в следующем порядке:

```
ocamlc -c basic_types.mli
ocamlyacc basic_parser.mly
ocamllex basic_lexer.mll
ocamlc -c basic_parser.mli
ocamlc -c basic_lexer.ml
ocamlc -c basic_parser.ml
```

После чего получим файлы `basic_lexer.cmo` и `basic_parser.cmo`, которые можно использовать в нашей программе.

Теперь, у нас есть весь необходимый арсенал, для того чтобы переделать программу.

Удалим все типы и функции параграфов «лексический анализ» (стр. ??) и «синтаксический анализ» (стр. ??) для программы Бэйсик. Также в функции `one_command` (стр. ??) заменим выражение:

```
match parse (input_line stdin) with
```

на

```
match line lexer (Lexing.from_string ((input_line stdin)^"\n")) with
```

Заметьте, что необходимо поместить в конце линии символ конца '\n', который был удалён функцией `input_line`. Это необходимо, потому что данный символ используется для указания конца командной линии (`Leol`).

11.6 Exercises

11.7 Резюме

В данной главе были описаны различные средства лексического и синтаксического анализа Objective CAML. По порядку описания:

- модуль `Str` для фильтрования рациональных выражений
- модуль `Genlex` для быстрого создания простых лексических анализаторов
- `ocamllex` представитель семейства `lex`
- `ocamlyacc` представитель семейства `yacc`
- потоки, для построения нисходящих анализаторов, в том числе и контекстных

При помощи инструментов `ocamllex` и `ocamlyacc` мы переделали синтаксический анализ Бэйсика, который проще поддерживать, чем анализатор представленный на стр. ??.

11.8 To Learn More

Глава 12

Взаимодействие с языком C

Глава 13

Приложения

Часть III

Устройство программы

Глава 14

Модульное программирование

Глава 15

Объектно-ориентированное программирование

Глава 16

Сравнение моделей устройств программ

Глава 17

Приложения

Часть IV

Параллелизм и распределение

Глава 18

Процессы и связь между процессами

Глава 19

Параллельное программирование

Глава 20

Распределённое программирование

Глава 21

Приложения

Часть V

Разработка программ с помощью Objective CAML

Часть VI

Приложения