
MP 8 – A Parser for PicoML

CS 421 – Fall 2009

Revision 1.1

Assigned Tuesday November 3, 2009

Due Tuesday November 17, 2009 23:59 PM

Extension 48 hours (20% penalty)

1 Change Log

- 1.1 Fixed an example parse given in the text (“Similarly, [2;3] parses to ...”) Added explanatory text regarding `_` patterns to the `try_with` problem Corrected name of minor included file to sync up with latest version of grader

1.0 Initial Release.

2 Overview

In this MP, we will deal with the process of converting PicoML code into an abstract syntax tree using a parser. We will use the *occamlyacc* tool to generate our parser from a description of the grammar. This parser will be combined with the lexer and type inferencer from previous MPs to make an interactive PicoML interpreter (well, it does not interpret yet), where you can type in PicoML expressions and see a proof tree of the expression’s type:

```
Welcome to the Student parser
```

```
> let x = 5;;
val x : int

final environment:

{x=>int}

proof:
  {} |= 5 : int

>
```

To complete this MP, you will need to be familiar with describing languages with BNF grammars, adding attributes to return computations resulting from the parse, and expressing these attribute grammars in a form acceptable as input to *occamlyacc*.

3 Given Files

mp8-skeleton.mly: You should copy the file **mp8-skeleton.mly** to **mp8.mly**. The skeleton contains some pieces of code that we have started for you, with triple dots indicating places where you should add code.

picomlIntPar.ml: This file contains the main body of the PicoML executable. It essentially connects your lexer, parser, and type inference code and provides a friendly prompt to enter PicoML expressions.

picomllex.ml: This file contains the ocamllex specification for the lexer. It is a modest expansion to the lexer you wrote for MP7.

mp8common.ml: This file includes the types of expressions and declarations. It also contains the type inferencing code. Appropriate code from this file will automatically be called by the interactive loop defined in **picomlInt-Par.ml**.

4 Overview of `ocamlyacc`

Take a look at the given `mp8-skeleton.mly` file. The grammar specification has a similar layout to the lexer specification of MP7. It begins with a header section (where you can add raw OCaml code), then has a section for directives (these start with a `%` character), then has a section that describes the grammar (this is the part after `%%`). You will only need to add to the last section.

4.1 Example

The following is the `exp` example from class (lecture 19 LR Parsing, slides 52 – 59):

```
%token <string> Id_token
%token Left_parenthesis Right_parenthesis
%token Times_token Divide_token
%token Plus_token Minus_token
%token EOL
%start main
%type <expr> main
%%
exp:
  term                    { Term_as_Expr $1 }
  | term Plus_token exp   { Plus_Expr ($1, $3) }
  | term Minus_token exp  { Minus_Expr ($1, $3) }
term:
  factor                  { Factor_as_Term $1 }
  | factor Times_token term { Mult_Term ($1, $3) }
  | factor Divide_token term { Div_Term ($1, $3) }
factor:
  Id_token                { Id_as_Factor $1 }
  | Left_parenthesis exp Right_parenthesis { Parenthesized_Expr_as_Factor $2 }
main:
  exp EOL                  { $1 }
```

Recall from lecture that the process of transforming program code (i.e., as ASCII text) into an *abstract syntax tree* (AST) has two parts. First, the *lexical analyzer* (lexer) scans over the text of the program and converts the text into a sequence of *tokens*. The type of tokens in general may be a preexisting OCaml type, or a user-defined type created for the purpose. In the case where `ocamlyacc` is used, the type should be named `token` and the datatype `token` is created implicitly by the `%token` directives. These tokens are then fed into the *parser* created by entry points in your input, which builds the actual AST.

The first five lines in the example above define the sorts of tokens of the language. These directives are converted by `ocamlyacc` into an OCaml disjoint type declaration defining the type `token`. Notice that the `Id.token` token has data associated with it (this corresponds to writing `type token = ... | Id.token of string` in OCaml). The sixth line says that the start symbol for the grammar is the nonterminal called `main`. After the `%%` directive comes the important part: the productions. The format of the productions is fairly self-explanatory. The above specification

describes the following extended BNF grammar:

$$\begin{aligned} S &::= E \text{ eol} \\ E &::= T \quad | T + E \quad | T - E \\ T &::= F \quad | F * T \quad | F / T \\ F &::= id \quad | (E) \end{aligned}$$

An important fact about *ocamlyacc* is that **each production returns a value** that is to be put on the stack. We call this the *semantic value* of the production. It is described in curly braces by the *semantic action*. The semantic action is actual OCaml code that will be evaluated when this parsing algorithm reduces by this production. The result of this code is the semantic value, and it is placed on the stack to represent the nonterminal.

What do \$1, \$2, etc., mean? These refer to the positional values on the stack, and are replaced in the OCaml code by the semantic values of the subexpressions on the right-hand side of the production. Thus, the symbol \$1 refers to the semantic value of the first subexpression on the right-hand side, and so on. As an example, consider the following production:

```
exp:
  ...
  | term Plus_token exp                { Plus_Expr ($1, $3) }
```

When the parser reduces by this rule, \$1 holds the semantic value of the `term` subexpression, and \$3 holds the value of the `exp` subexpression. The semantic rule generates the AST representing the addition of the two, and the result becomes the semantic value for this production and is put on the stack to replace the top three items.

Also note that when tokens have associated data (like `Id_token`, which has a string), that associated data is treated as the semantic value of the token:

```
factor:
  Id_token                { Id_as_Factor $1 }
```

Thus, the above \$1 corresponds to the string component of the token, and not the token itself.

4.2 More Information

Here is a website you should check out if you would like more information or an alternate explanation of *ocamlyacc* usage:

- <http://caml.inria.fr/pub/docs/manual-ocaml/manual026.html>

5 Compiling

A `Makefile` is provided for this MP. After you make changes to `mp8.mly`, all you have to do is type `gmake` (or possibly `make` if you are using a non-linux machine) and the two needed executables will be rebuilt.

5.1 Running PicoML

The given `Makefile` builds executables called `picomlIntPar` and `picomlIntParSol`. The first is an executable for an interactive loop for the parser built from your solution to the assignment, and the second is one built from the standard solution. If you run `./picomlIntPar` or `./picomlIntParSol`, you will get an interactive screen, much like the OCaml interactive screen. You can type in PicoML expressions followed by a double semicolon, and they will be parsed and their types inferred and displayed:

Welcome to the Solution parser

```
> 3;;
```

```
val _ : int
```

```
final environment:
```

```
{}
```

```
proof:
```

```
{ } |= 3 : int
```

```
> let x = 3 + 4;;
```

```
val x : int
```

```
final environment:
```

```
{x=>int}
```

```
proof:
```

```
{ } |= (3+4) : int
```

```
|--{} |= ((+))(3) : int -> int
```

```
| |--{} |= (+) : int -> int -> int
```

```
| |--{} |= 3 : int
```

```
|--{} |= 4 : int
```

```
> let f = fun y -> y * x;;
```

```
val f : int -> int
```

```
final environment:
```

```
{f=>int -> int,x=>int}
```

```
proof:
```

```
{x=>int} |= (fun y -> (y*x)) : int -> int
```

```
|--{y=>int,x=>int} |= (y*x) : int
```

```
|--{y=>int,x=>int} |= ((*)(y) : int -> int
```

```
| |--{y=>int,x=>int} |= (*) : int -> int -> int
```

```
| |--{y=>int,x=>int} |= y : int
```

```
|--{y=>int,x=>int} |= x : int
```

```
> f 5;;
```

```
val _ : int
```

```
final environment:
```

```
{f=>int -> int,x=>int}
```

```
proof:
```

```
{f=>int -> int,x=>int} |= (f)(5) : int
```

```
|--{f=>int -> int,x=>int} |= f : int -> int
```

```
|--{f=>int -> int,x=>int} |= 5 : int
```

>

Notice the accumulation of values in the (type) environment as expressions are entered. To reset the environment, you must quit the program (with CTRL+C) and start again.

6 Important Notes

- The BNFs below for PicoML's grammar are ambiguous, and it is just a description of the *concrete* syntax of PicoML. You are also provided with a table listing the associativity/precedence attributes of the various language constructs. You are supposed to use the information given in this table in order to create a grammar that generates the same language as the given one, but that is unambiguous and enforces the constructs to be specified as in the table. Your actual *ocaml yacc* specification will consist of the latter grammar.
- **The BNFs do not show the stratification needed to eliminate ambiguity. That is your job!** This will likely involve reorganizing things.
- For this MP, you must implement precedence using stratification. *ocaml yacc* has some shortcut directives (`%left`, `%right`) for defining operator precedence, but these are off-limits for this MP!
Do not use them!
- Even though the work in this MP is split into several problems, you should really have the overall view on how the disambiguated grammar will look like, because precedence makes the choices for the productions corresponding to the language constructs conceptually interdependent. You might want to read through all the expression types first and try to organize your stratification layers before starting. 90 percent of your intellectual effort in this MP will consist of disambiguating the grammar properly.

Stratification means breaking something up into layers. In the example 4.1, we could have expressed the grammar more succinctly by

$$\begin{aligned} S &::= E \text{ eol} \\ E &::= id \mid E + E \mid E - E \mid E * E \mid E / E \mid (E) \end{aligned}$$

This grammar, while compact, and comprehensible to humans, is highly ambiguous for the purposes of parsing. To render it unambiguous, we introduced intermediate non-terminals (layers, or strata) to express associativity and precedence of operators. You will need to perform similar transformations on the description given here to remove ambiguities and avoid shift-reduce or reduce-reduce conflicts.

7 Problem Setting

The concrete syntax of PicoML that you will need to parse is the following:

```
<main> ::= <exp> ;;
        | let IDENT = <exp> ;;
        | let rec IDENT = <exp> ;;

<exp> ::= IDENT
        | BOOL | INT | FLOAT | STRING | UNIT
        | ( <exp> )
        | ( <exp> , <exp> )
        | let IDENT = <exp> in <exp>
```

```

| let rec IDENT = <exp> in <exp>
| <exp> <infix> <exp>
| <exp> && <exp>
| <exp> || <exp>
| [ ]
| [ <list_contents> ] /* sugar for non-empty lists, extra credit */
| if <exp> then <exp> else <exp>
| <exp> <exp>
| fun IDENT -> <exp>
| raise <exp>
| try <exp> with n1 -> e1 | ... /* extra credit */

```

<list_contents> ::= <nonempty sequence of expressions separated by double colons>

IDENT refers to an identifier token (only one token, takes a string as argument). <infix> refers to some infix identifier token (one for each infix operator). (<infix> is not an explicit syntactic category that you must parse. It is only used here to express <exp>.)

The nonterminals in this grammar are `main`, `exp`, and `list_contents`, with `main` being the start symbol.

The rest of the symbols are terminals, and their representations in OCaml are elements of the type `token`, defined at the beginning of the file `mp8.mly`. Our OCaml representation of terminals is not always graphically identical to the one shown in the above grammar; we have used concrete syntax in place of tokens for the terminals. For example, `::` is represented by `DCOLON` and `+` by `PLUS`. Our OCaml representation of the identifier tokens (IDENT) is achieved by the constructor `IDENT` that takes a string and yields a token, as constructed by the lexer from MP7.

Recall that identifying the tokens of the language is the job of `lexer`, and the parser (that you have to write in this MP) takes as input a *sequence of tokens*, such as `(INT 3) PLUS (INT 5)` and tries to make sense out of it by transforming it into an abstract syntax tree, in this case `AppExp (AppExp (BinOpExp (PlusOp) , ConstExp (IntConst 3)) , ConstExp (IntConst 5))`. The abstract syntax trees into which you have to parse your sequences of tokens are given by the following OCaml types (metatypes, to avoid confusion with PicoML types), present in the file `mp8common.ml`:

```

(* declarations *)
type declaration =
  Let of string option * exp
  | LetRec of string * exp

(* expressions for PicoML *)
type const =
  BoolConst of bool
  | IntConst of int
  | FloatConst of float
  | StringConst of string
  | NilConst
  | UnitConst

type exp =
  VarExp of string
  | ConstExp of const
  | BinOpExp of string
  | IfExp of exp * exp * exp
  | AppExp of exp * exp
  | FunExp of string * exp
  | LetInExp of string * exp * exp

```

```

| LetRecInExp of string * exp * exp
| RaiseExp of exp
| TryWithExp of exp * (int option * exp) * ((int option * exp) list)

```

Thus each sequence of tokens should either be interpreted as an element of metatype `exp` or declaration, or should yield a parse error. Note that the metatypes `exp` and `declaration` contain abstract, and not concrete syntax. Recall from MP5 that our abstract syntax encodes any operator of PicoML (except for `&&` and `||`, which are syntactic sugar for corresponding if/then/else expressions) using application. This is why `3 + 4` parses to

```
AppExp (AppExp (BinOpExp "+", ConstExp (IntConst 3)), ConstExp (IntConst 4)).
```

Similarly, `[2;3]` parses to

```
AppExp (AppExp (BinOpExp "::", ConstExp (IntConst 2)), AppExp (AppExp (BinOpExp "::",
ConstExp (IntConst 3)), ConstExp (NilConst)))
```

If we do not specify the precedence and associativity of our language constructs and operators, the parsing function is not well-defined. For instance, how should `if true then 3 else 2 + 4` be parsed? Depending on how we “read” the above sequence of tokens, we get different results:

- If we read it as the sum of a conditional and a number, we get the same thing as if it were: `(if true then 3 else 2) + 4`
- If we read it as a conditional having a sum in its false branch, we get `if true then 3 else (2 + 4)`

The question is really which of the sum and the conditional binds its arguments tighter, that is, which one has a higher precedence (or which one has precedence over the other). In the first case, the conditional construct has a higher precedence; in the second, the sum operator has a higher precedence.

Another source of ambiguity arises from associativity of operators: how should `true && true && false` be parsed?

- If we read it as the conjunction between true and a conjunction, we get `true && (true && false)`
- If we read it as a conjunctions between a conjunction and false, we get `(true && true) && false`

In the first case, `&&` is right-associative; in the second, it is left-associative.

The desired precedence and associativity of the language constructs and operators (which impose a unique parsing function) are given below, where a left-associative operator is preceded by “left”, a right-associative operator by “right”, and precedence decreases downwards on the lines (thus two items listed on the same line have the same precedence).

```

left _ _      (application is left associative, and binds tighter than anything else)
raise_
right **      (** is right associative, binds tighter than anything but app. or raise)
left * left *. left / left /.
left + left +. left - left -. left ^
right ::
left = left < left > left <= left >=
left _&&_
left _||_
if_then_else_
fun_->_
let_=in_
let_rec=_in_
try_with_->_|_|_ ..., where | is right associative

```

Above, the underscores are just a graphical indication of the places where the various syntactic constructs expect their “arguments”. For example, the conditional has three underscores, the first for the condition, the second for the then branch, and the third for the else branch.

8 Problems

At this point, your assignment for this MP should already be fairly clear. The following problems just break your assignment into pieces and are meant to guide you towards the solution. A word of warning is however in order here: The problem of writing a parser is *not* a modular one, because the parsing of each language construct depends on all the other constructs. Adding a new syntactic category may well force you to go back and rewrite all the categories already present. Therefore you should approach the set of problems as a whole, and always keep in mind the precedences and associativities given for the PicoML constructs. You are allowed, and even encouraged, to add to your grammar new nonterminals (together with new productions) in addition to the one that we require (`main`). In addition, you may find it desirable to rewrite or reorganize the various productions we have given you. The productions given are intended only to be enough to allow you to start testing your additions. Also, it is allowed that you define the constructs in an order that is different from the one we have given here. For instance, we have gathered the requirements according to the intended semantics of the constructs (e.g., grouping arithmetic operators together and list operators together); you may rather want to group the constructs according to their precedence; you are absolutely free to do that. However, we require that the non-terminal `main` that we introduced in the problem statement be present in your grammar and that it produces exactly the same set of strings as described by the grammar in Section 7, obeying the precedences and associativities also described in that section.

1. (5 pts) In the file `mp8.mly` add the integer, unit, boolean, float, and string constants.

```
> let x = "hi";;
val x : string

final environment:

{x=>string}

proof:
  {} |= "hi" : string

> let x = true;;
val x : bool

final environment:

{x=>bool,x=>string}

proof:
  {x=>string} |= true : bool

>
```

2. (5 pts) Add parentheses. You won;t be able to rone the second example until you add infix binary operators.

```
> (3);;
val _ : int

final environment:

{}
```

```

proof:
  {} |= 3 : int

> (3 + (4 * 6));;
val _ : int

final environment:

{}

proof:
  {} |= (3+(4*6)) : int
  ...
>

```

3. (5 pts) Add pairs. Note that unlike OCaml, PicoML requires opening and closing parentheses around pairs.

```

> (3, 9);;
val _ : int * int

final environment:

{}

proof:
  {} |= (3,9) : int * int
  ...

```

4. (12 pts) Add infix operators. You will need to heed the precedence and associativity rules given in the table above. Start by adding arithmetical and string operators.

```

> 3 + 4 * 2;;
val _ : int

final environment:

{}

proof:
  {} |= (3+(4*2)) : int
  ...

> 2.0 *. 3.1 *. 2.5;;
val _ : float

final environment:

{}

```

```
proof:
  {} |= ((2.*.3.1)*.2.5) : float
  ...
```

5. (12 pts) Add comparison operators.

```
> "a" < "b";;
val _ : bool
```

```
final environment:
```

```
{}
```

```
proof:
  {} |= ("a"<"b") : bool
  ...
```

```
> 3 = 5;;
val _ : bool
```

```
final environment:
```

```
{}
```

```
proof:
  {} |= (3=5) : bool
  ...
```

```
>
```

6. (10 pts) Add :: (list consing).

```
> 3 :: 2 :: 1 :: [];;
val _ : int list
```

```
final environment:
```

```
{}
```

```
proof:
  {} |= (3::(2::(1::[]))) : int list
  ...
```

7. (8 pts) Add let_in_ and let_rec_in.

```
> let x = 3 in (x + x);;
val _ : int
```

```
final environment:
```

```

{}

proof:
  {} |= let x = 3 in (x+x) : int
  ...

> let rec x = 3::x in 5;;
val _ : int

final environment:

{}

proof:
  {} |= let rec x = (3::x) in 5 : int
  ...

```

8. (20 pts) Add `fun->` and `if.then.else.`

```

> fun x -> if x then 3 else 4;;
val _ : bool -> int

final environment:

{}

proof:
  {} |= (fun x -> if x then 3 else 4) : bool -> int
  ...

```

9. (10 pts)

Add application.

```

> (fun x -> x + x +3) 4;;
val _ : int

final environment:

{}

proof:
  {} |= ((fun x -> ((x+x)+3)))(4) : int
  ...

```

10. (11 pts) Add `&&` and `||`. Note that $e_1 \&\& e_2$ should parse the same as `if e_1 then e_2 else false`, and $e_1 || e_2$ should parse the same as `if e_1 then true else e_2` . In each case, the appropriate OCaml constructor to use is `IfExp`.

```

> true || false && true;;
val _ : bool

final environment:

{}

proof:
  {} |= if true then true else if false then true else false : bool
  ...

```

11. (5 pts) Add `raise`. Notice from one of the examples below that `raise` binds more tightly than `+` but less tightly than application.

```

> let n = raise 3;;
val n : 'a

final environment:

{n=>'a}

proof:
  {} |= raise 3 : 'a
  |--{} |= 3 : int

> raise n + 3;;
val _ : int

final environment:

{n=>int}

proof:
  {n=>int} |= (raise n+3) : int
  ...

> raise (fun x -> x) n;;
val _ : 'a

final environment:

{n=>int}

proof:
  {n=>int} |= raise ((fun x -> x)) (n) : 'a
  |--{n=>int} |= ((fun x -> x)) (n) : int
  ...

```

9 Extra Credit

12. (5 pts) Add syntactic sugar for lists to your expressions. More precisely, add the following expressions to the grammar:

- $\text{exp} \rightarrow [\text{list_contents}]$

where *list_contents* is a non-empty sequence of expressions separated by semicolons. It has to be the case that semicolon binds less tightly than any other language construct or operator.

```
> let x = [1; 2; 3];;
val x : int list
```

```
final environment:
```

```
{x=>int list}
```

```
proof:
```

```
{} |= (1::(2::(3::[]))) : int list
...
```

13. (10 pts) Add `try_with_`. Be sure to notice how the expression is parsed in the second example: pipes are associated with the right-most preceding try-with (the ambiguity this fixes is analogous to the dangling-else problem.)

Valid patterns have the form $n \rightarrow e$, where n is to be represented by `Some` wrapped around an integer, or $_ \rightarrow e$, where $_$ is represented is to be represented by `None`.

```
> try "hi" with 1 -> "one" | 2 -> "two";;
val _ : string
```

```
final environment:
```

```
{}
```

```
proof:
```

```
{} |= try "hi" with (1 -> "one" | 2 -> "two") : string
...
```

```
> try "hi" with 1 -> "one" | 2 -> try "two" with 3 -> "three" | 4 -> "four";;
val _ : string
```

```
final environment:
```

```
{}
```

```
proof:
```

```
{} |= try "hi" with (1 -> "one" | 2 -> try "two" with (3 -> "three" | 4 -> "four")) : st
...
```

10 Additional tests

1. Can you pass this test? Make sure your parser parses the expression as in the example.

```
> 3 - 4 - 2 * 9 > 10 && true;;  
val _ : bool
```

```
final environment:
```

```
{}
```

```
proof:
```

```
{ } |= if ((3-4)-(2*9))>10 then true else false : bool  
...
```

2. This one?

```
> if true then 1 else 0 + 2;;  
val _ : int
```

```
final environment:
```

```
{}
```

```
proof:
```

```
{ } |= if true then 1 else (0+2) : int  
...
```

3. This one?

```
> (fun x -> ()) 3;;  
val _ : unit
```

```
final environment:
```

```
{}
```

```
proof:
```

```
{ } |= ((fun x -> ())) (3) : unit  
...
```