PASCALMAISPRESQUE
Introduction to language theory and compiling
Project – Part 1

Gilles GEERAERTS Mathieu SASSOLAS Léonard BRICE

Info-F403 – 2023-2024

Rien n'est si insupportable à l'homme que d'être dans un plein repos, sans passions, sans affaires, sans divertissement, sans application. — Nothing is as unbearable for a man as to be completely at rest, with no passions, no business, no diversion, no work.

Blaise Pascal, *Pensées*

1 Introduction

In this project, you are requested to design and write a compiler for PASCALMAISPRESQUE, a simple imperative language. The grammar of the language is given in Figure 1 (page 2), where reserved keywords have been typeset using `typewriter font`. In addition, [VarName] and [Number] are lexical units, which are defined as follows. A [VarName] identifies a variable, which is a string of digits and letters, starting with a lowercase letter (this is case sensitive). A [Number] represents a numerical constant, and is made up of a string of digits only. The minus sign can be generated using rule [17].

Finally, comments are allowed in PASCALMAISPRESQUE. There are two kinds of comments: Short comments, which are all the symbols appearing after a a double star $\star\star$ up to the end of the line, and long comments, which are all the symbols occurring between two double apostrophes $''$ keywords (this is NOT a quotation mark!). There is no possibility of nested comments. Observe that comments do not occur in the rules of the grammar: they must be ignored by the scanner, and will not be transmitted to the parser.

Figure 2 shows an example of PASCALMAISPRESQUE program.

2 Assignment - Part 1

In this first part of the assignment, you must produce the *lexical analyzer* of your compiler, using the JFlex tool reviewed during the practicals.

Please adhere strictly to the instructions given below, otherwise we might be unable to grade your project, as automatic testing procedures will be applied.

2.1 Environment and provided files

The lexical analyzer will be implemented in JAVA 1.8. It must recognise the different lexical units of the language, and maintain a symbol table. To help you, several JAVA classes are provided on the UV:

- The `LexicalUnit` class contains an enumeration of all the possible lexical units;

[1] <Program> → `begin` <Code> `end`
[2] <Code> → ε
[3] → <InstList>
[4] <InstList> → <Instruction>
[5] → <Instruction> `...` <InstList>
[6] <Instruction> → <Assign>
[7] → <If>
[8] → <While>
[9] → <For>
[10] → <Print>
[11] → <Read>
[12] → `begin` <InstList> `end`
[13] <Assign> → [VarName] `:=` <ExprArith>
[14] <ExprArith> → [VarName]
[15] → [Number]
[16] → `(` <ExprArith> `)`
[17] → `-` <ExprArith>
[18] → <ExprArith> <Op> <ExprArith>
[19] <Op> → `+`
[20] → `-`
[21] → `*`
[22] → `/`
[23] <If> → `if` <Cond> `then` <Instruction>
[24] → `if` <Cond> `then` <Instruction> `else` <Code>
[25] <Cond> → <Cond> `and` <Cond>
[26] → <Cond> `or` <Cond>
[27] → `{` <Cond> `}`
[28] → <SimpleCond>
[29] <SimpleCond> → <ExprArith> <Comp> <ExprArith>
[30] <Comp> → `=`
[31] → `<`
[32] <While> → `while` <Cond> `do` <Instruction>
[33] <Print> → `print(`[VarName]`)`
[34] <Read> → `read(`[VarName]`)`

Figure 1: The PASCALMAISPRESQUE grammar.

```
1  '' Euclid's algorithm ''
2
3  begin
4    read(a)...
5    read(b)...
6    while 0 < b do
7      begin
8        c := b...
9        while b < a+1 do      ** Computation of modulo
10         a := a-b...
11       b := a...
12       a := c
13     end...
14   print(a)
15 end
```

Figure 2: An example PASCALMAISPRESQUE program.

- The `Symbol` class implements the notion of token. Each object of the class can be used to associate a value (a generic Java `Object`) to a `LexicalUnit`, and a line and column number (position in the file). The code should be self-explanatory. If not, do not hesitate to ask questions to the teacher, or to the teaching assistants.

2.2 Expected outcome

You must hand in all files required to compile and evaluate your code, as well as the proper documentation, including a **PDF report**. This must be structured into five folders as follows, and a `Makefile` must be provided.

`src/` Contains all source files required to evaluate and compile your work:

- the JFlex source file `LexicalAnalyzer.flex` for your lexical analyzer;
- the generated file `LexicalAnalyzer.java`;
- provided files `Symbol.java` and `LexicalUnit.java` *without modification*.
- a `Main.java` file calling the lexical analyzer that reads the file given as argument and writes on the standard output stream the sequence of matched lexical units and the content of the symbol table (see below for details on the expected output).

`doc/` Contains the JAVADOC and the PDF report.

Note that the documentation of your code will be taken into account for the grading, especially for Parts 2 and 3, so Part 1 is a good occasion to setup JAVADOC for your project.

The PDF report should contain all regular expressions, and present your work, with all the necessary justifications, choices and hypotheses, as well as descriptions of your example files. Such report will be particularly useful to get you partial credit if your tool has bugs.

Bonus: Everyday programming languages can handle nested comments. Explain what technical difficulties arise if you are to handle those.

`test/` Contains all your example PASCALMAISPRESQUE files. It is necessary to provide relevant example files of your own.

`dist/` Contains an executable JAR called `part1.jar`. The command for running your executable should therefore be: `java -jar part1.jar sourceFile.pmp`

`more/` Contains all other files.

`Makefile` At the root of your project (i.e. not in any of the aforementioned folders). There is an example Makefile on the UV.

You will compress your root folder (in the *zip* format—no *rar* or other format), **which is named according to the following regexp**: `Part1_Surname1(_Surname2)?.zip`, where `Surname1` and, if you are in a group, `Surname2` are the last names of the student(s) (in alphabetical order); you are allowed to work in a group of maximum two students.

The *zip* file shall be submitted on Université Virtuelle before **October 23th, 2023, 23:59**, Brussels Grand-Place time.

2.3 Output of the program

The format of the output of the program must be:

1. First, the sequence of matched lexical units. You must use the *toString()* method of the provided Symbol class to print individual tokens;
2. Then, the word *Variables*, to clearly separate the symbol table from the sequence of tokens;
3. Finally, the content of the symbol table, formatted as the sequence of all recognised variables, in lexicographical (alphabetical) order. There must be one variable per line, together with the number of the line of the input file where this variable has been encountered for the first time (the variable and the line number must be separated by at least one space).

For instance, on the following input:

```
read(b)
```

your executable must produce exactly, using the `toString()` method of the `Symbol` class, the following output for the sequence of tokens (an example for the symbol table is given hereunder):

```
token: read           lexical unit: READ
token: (              lexical unit: LPAREN
token: b              lexical unit: VARNAME
token: )              lexical unit: RPAREN
```

Note that the *token* is the matched input string (for instance `b` for the third token) while the *lexical unit* is the name of the matched element from the `LexicalUnit` enumeration (`VARNAME` for the third token).

Also, for the example in Figure 2, the symbol table must be displayed as:

```
Variables
a 4
b 5
c 8
```

An example input PASCALMAISPRESQUE file with the expected output is available on Université Virtuelle to test your program.

3 Frequently Asked Questions

Here are some questions that we have gotten in the previous years which might help you during your project. If you have further questions please ask in the forum on the UV or via email.

- Q: What should happen if the PASCALMAISPRESQUE file can not be correctly tokenized?

  A: Throw a (useful) error message such as "Unknown symbol detected" or "Long comment not closed" for example. You can assume that you will not encounter a PASCALMAISPRESQUE file that has nested comments.

- Q: What about whitespaces?

  A: Whitespaces are not necessary beyond separating keywords, but they are nice for readability. All extra whitespaces must be ignored. For example:

```
(if abbb then ... x:=3
```

```
( if abbb then...x:=3
```

```
(
    if          abbb

    then
        ...
        x := 3
```

  must all yield the same matching of lexical units, namely:

```
token: (            lexical unit: LPAREN
token: if           lexical unit: IF
token: abbb         lexical unit: VARNAME
token: then         lexical unit: THEN
token: ...          lexical unit: DOTS
token: x            lexical unit: VARNAME
token: :=           lexical unit: ASSIGN
token: 3            lexical unit: NUMBER
```

  However, there are cases where spaces are required to isolate a keyword: `ifabbbthen` should be scanned as one variable name, and therefore, `(ifabbbthen...x:=3` would not produce the same result as above.

- Q: What is the use of a `Makefile`?

  A: You need to provide a `Makefile` because your executable might not run on our machine, because of different JAVA versions (unavoidable because some use Windows, Linux, MacOS). A working `Makefile` allows us to quickly produce an executable from your code. It takes a considerable amount of time if we need to figure out how to turn your code into an executable by hand. In addition, having a Makefile can streamline the testing and debugging of the code, so it is worthwhile to have for your sake as well.

- Q: Can I add features to this language?

  A: Personal initiative is encouraged for this project: do not hesitate to explain why some features would be hard to add at this point of the project, or to add relevant features to PASCALMAISPRESQUE. Ask us before, just to check that the feature in question is both relevant and doable.