

NPRG045 Project Specification: OPythn

Marcel Goh

March 24, 2019

1 Introduction

The OPythn project aims to implement a working subset of the Python programming language by means of a bytecode compiler and interpreter, written in OCaml. Users will be able to interact with OPythn via a top-level read-eval-print loop or compile OPythn source code to bytecode. OPythn will run on Unix operating systems, including macOS and Ubuntu.

2 Lexical Conventions

OPythn reads source code as ASCII characters, which are fed into a lexer. This section describes how the tokenises source code input.

2.1 Line and block structure

The end of a line is represented by the `NEWLINE` token. In general, simple statements cannot be split over multiple lines. However, an expression in parentheses, square brackets, or curly braces can be written over multiple lines without ending the statement. The `#` character can be used to indicate comments in source code. Any letters from `#` to the end of the line will be ignored by the lexer. A line that contains only whitespaces or comments is ignored except during interactive evaluation, when a blank line is used to indicate the end of a multi-line statement.

In OPythn, the amount of leading whitespace at the beginning of a line indicates the indentation level of the line. A tab character counts for exactly four spaces during this computation. Indentation levels of consecutive lines are used to generate `INDENT` and `DEDENT` tokens. This is done by means of a stack using the following algorithm, as described by the standard Python reference [1].

Algorithm 1 Insertion of `INDENT` and `DEDENT` tokens (as part of general lexing procedure)

```

1: procedure LEX(OPythn source code)
2:   token_list  $\leftarrow$  create a new empty list
3:   stack  $\leftarrow$  create a new empty stack
4:   stack.push(0)
5:   while there are lines to be read do
6:     line  $\leftarrow$  the current line
7:     num_sp  $\leftarrow$  the number of leading spaces in line
8:     if curr > stack.peek() then stack.push(curr) token_list.add(INDENT)
9:     else if curr < stack.peek() then
10:      count  $\leftarrow$  0
11:      while curr < stack.peek() do
12:        stack.pop()
13:        count  $\leftarrow$  count + 1
14:      end while
15:      for i  $\leftarrow$  0, count do
16:        token_list.add(DEDENT)
17:      end for
18:    end if
19:    tokenise the rest of the line and add tokens to token_list
```

```

20:   end while
21:   return token_list
22: end procedure

```

2.2 Keywords and identifiers

The following are OPythn keywords and cannot be used in ordinary identifiers:

int	float	str	bool	def	return
True	False	None	and	or	not
if	elif	else	for	in	while
break	continue	class	is	del	import
in	from	as	global	nonlocal	lambda

Any string that is not a keyword and contains uppercase and lowercase letters, underscores, and (except for the first character) numbers is a valid identifier. Case is significant. Some names are reserved for use by the standard library, which is described in Section 3.

2.3 Literals

Strings are enclosed in matching single quotes or double quotes. Triple-quoted strings are not supported, nor are formatted string literals. The backslash character \ is used to escape characters that normally have a special meaning.

```

string: """ stringitem* """ | ''' stringitem '''
stringitem: stringchar | stringescapeseq
stringchar: <source character besides '\\', '\n', '"', '>
stringescapeseq: '\\> <source character>

```

There are two types of numeric literals in OPythn: integers and floating-point numbers. Leading zeroes in a non-zero decimal are not allowed. Underscores are also not allowed in numeric literals, unlike in Python 3.

```

integer: decinteger | bininteger | octinteger | hexinteger
decinteger: nonzerodigit digit* | '0'+
bininteger: '0' ('b'|'B') bindigit+
octinteger: '0' ('o'|'O') octdigit+
hexinteger: '0' ('x'|'X') hexdigit+
nonzerodigit: '1'...'9'
digit: '0'...'9'
bindigit: '0' | '1'
octdigit: '0'...'7'
hexdigit: digit | 'a'...'f' | 'A'...'F'

floatnumber: pointfloat | exponentfloat
pointfloat: [digit+] fraction | digit+ '.'
exponentfloat: (digit+ | pointfloat) exponent
fraction: '.' digit+
exponent: ('e'|'E') ['+' | '-'] digit+

```

2.4 Operators and delimiters

All symbolic operators and delimiters are inherited from standard Python except the `@` and `@=` symbols. (They are used for matrix multiplication and decorators, both of which are not features of OPythn.)

3 Language

The core of OPythn is designed to be lightweight and minimal. Basic types and operations, control structures, and elementary data structures are included, while more complex Python constructions, such as list comprehensions, generators, and coroutines are omitted.

3.1 Simple and compound types

Every value in OPythn is of a certain type. There are four primitive types (`int`, `float`, `str`, and `bool`) and all of them are immutable. Integers are represented with 63 bits and can range between -4,611,686,018,427,387,904 and 4,611,686,018,427,387,903 inclusive. Strings contain 8-bit characters.

OPythn has three compound types: `list`, `dict`, and `tuple`. Of these, only tuples are immutable. Lists are zero-indexed and accessing an element via its index takes constant time. Dictionaries support values of different types, with the restriction that keys must be of an primitive type or `NoneType`. For example, `d = {2: 'bonjour', False: 3.3}` is a valid OPythn dictionary.

3.2 Numeric and boolean operations

Ordinary arithmetic operators for addition, multiplication, subtraction, division, modulus, and exponentiation are supported between types `int` and `float`. When performing integer division, OPythn takes the floor of the quotient, so `-13 // 5` gives `-3`, and the modulus operator always gives a positive result, so `-13 % 5` returns `2`.

OPythn supports the bitwise operations `|`, `^`, `&`, `<<`, `>>`, and `~` on integers. Numeric values can be compared using the operators `<`, `<=`, `>`, and `>=` (note that chained comparisons are not allowed), and any values can be tested for equality using `==`, `!=`, `is`, and `is not`. Any object can be tested for truth value. The objects `None`, `False`, `0`, `0.0`, `''`, `[]`, and `{}` are considered false; any other object is considered true.

3.3 String operations

The following useful methods are defined on strings, along with all of the methods on sequence types described in the following subsection.

`find()` `isalpha()` `isnumeric()` `lower()` `split()` `strip()` `upper()`

3.4 Operations on sequence types

Strings, lists, and tuples are considered sequence types in OPythn and support a number of useful operations.

- They can be indexed into using the postfix `[]` operator. The expression inside must be an integer. Negative indices are allowed. For example, if `x` is the list `[1,2,3,4]`, then `x[-2]` will return `3`. Indexing into a list is the normal way to mutate a value. For example, running the expression `x[1] = 'grape'` changes `x` into the list `[1,'grape',3,4]`. This cannot be done with tuples because they are immutable.

- The number of elements in the sequence (characters, in the case of strings), can be determined using the built-in function `len()`.
- To determine if a sequence `l` contains the element `e`, one may use the expression `e in l` which returns the corresponding boolean. When used with a `for` loop, this syntax allows the user to iterate through all the elements in a list. To see if an element does not belong to a list, one may use `not in`, but of course this cannot be used to create a `for` loop. The `enumerate` object does not exist in OPython.
- Sequence objects can be sliced using the operator `:`. For example, if `a = [1, True, 3, 'hi', 5]`, then `a[1:3]` would give the list `[True, 3]`. This can be used with the `del` keyword to delete a group of adjacent elements in a sequence. Slicing in step increments is not supported and a `slice` object cannot be explicitly created.
- Sequences of the same type can be concatenated using the `+` operator.
- A sequence can be explicitly converted into a list by calling `list()` on it. (Calling `list()` on a list returns the same list.)

These functions and methods are defined on lists specifically. The special method `range(i, j)` is also included, which creates a list of integers from `i`, inclusive, to `j`, exclusive.

```
append()  count()  extend()  index()
insert()  pop()    remove()  reverse()
```

3.5 Operations on dictionaries

The following operations are defined on dictionaries.

- Accessing an element can be done using the syntax `dict[key]`. New entries can be added by running `dict[new_key] = new_value`, but duplicate keys are not allowed, so if an entry already exists with the same key, it will be replaced with the new key-value pair.
- An entry can be deleted using the syntax `del dict[key]` and the entire dictionary can be deleted by calling `del` on the whole dictionary.
- The user can determine if a key is present in a dictionary using `in` or `not in`.

Note that two dictionaries are equal if they contain the same keys and values. The following methods are defined on dictionary objects.

```
clear()  items()  keys()  values()
```

3.6 Control structures

OPython inherits the control structures `if`, `for`, and `while` from Python. Like in Python, consecutive lines at the same indentation level belong to the same block. Conditional expressions of the form `<expr> if <condition> else <expr>` are also valid.

3.7 Functions

OPython supports higher-order functions and nested functions. In OPython, functions cannot modify variables outside their scope, unless one of the keywords `global` or `nonlocal` is used. OPython supports anonymous functions using the keyword `lambda`. OPython functions take positional arguments only, and specifying default values is not allowed.

The following functions comprise OPython's standard library and are included in a normal installation. Some of these functions are implemented in OPython itself while others are implemented in OCaml.

abs()	bin()	bool()	chr()	divmod()	filter()
float()	hash()	hex()	input()	int()	isinstance()
issubclass()	iter()	len()	map()	max()	min()
next()	oct()	open()	ord()	pow()	print()
range()	repr()	reversed()	round()	str()	sorted()
sum()	super()	type()			

3.8 Classes and objects

OPython is object-oriented and allows the user to define classes. However, nested classes and multiple inheritance are not supported. When creating a class, the following special methods can be added to indicate how the class should behave: `__init__`, `__eq__`, `__ne__`, `__str__`.

OPython does not allow private attributes and methods. In general, an object's attributes and methods can be accessed and mutated by functions outside the class' definition. All Python values are objects, but the user is not allowed to subclass built-in types.

3.9 Error handling

OPython does not include support for exceptions. When an error occurs, the program terminates/returns to the REPL and a stack trace is printed.

3.10 Grammar

This is the complete OPython grammar specification:

```
# Start symbols
single_input: NEWLINE | simple_stmt | compound_stmt NEWLINE
file_input: (NEWLINE | stmt)* ENDMARKER

funcdef: 'def' NAME parameters ':' suite
classdef: 'class' NAME ['(' [arglist] ')'] ':' suite
parameters: '(' [arglist] ')'
vararglist: arg (',' arg)*
arg: NAME

stmt: simple_stmt | compound_stmt
simple_stmt: small_stmt (';' small_stmt)* [';'] NEWLINE
small_stmt: (expr_stmt | del_stmt | flow_stmt | import_stmt
            global_stmt | nonlocal_stmt)
expr_stmt:
compound_stmt: if_stmt | while_stmt | for_stmt | funcdef |
              classdef
global_stmt: 'global' NAME (',' NAME)*
nonlocal_stmt: 'nonlocal' NAME (',' NAME)*

# Control structures
```

```

if_stmt: 'if' test ':' suite ('elif' test ':' suite)*
        ['else' ':' suite]
while_stmt: 'while' test ':' suite ['else' ':' suite]
for_stmt: 'for' exprlist 'in' testlist ':' suite
        ['else' ':' suite]

# Tests and expressions
test: or_test ['if' or_test 'else' test]
or_test: and_test ('or' and_test)*
and_test: not_test ('and' not_test)*
not_test: 'not' not_test | comparison
comparison: expr (comp_op expr)*
comp_op: '<' | '>' | '=' | '>=' | '<=' | '!=' | 'in' | 'not in' |
        'is' | 'is not'
expr: xor_expr ('|' xor_expr)*
xor_expr: and_expr ('^' and_expr)*
and_expr: shift_expr ('&' shift_expr)*
shift_expr: arith_expr (('<<' | '>>') arith_expr)*
arith_expr: term (('+' | '-') term)*
term: factor (('*' | '/' | '%' | '//') factor)*
factor: ('+' | '-' | '~') factor | power
power: atom_expr ['**' factor]
atom_expr: atom trailer*
atom: (NAME | NUMBER | STRING+ | '...' | 'None' |
        'True' | 'False')
trailer: '(' [arglist] ')' | '[' subscript ']' | . NAME
subscriptlist: subscript (',' subscript)* [',']
subscript: test | [test] ':' [test] [sliceop]
sliceop: ':' [test]
exprlist: expr (',' expr)* [',']
testlist: test (',' test)* [',']
dictorsetmaker: ( ((test ':' test | '**' expr)
                   ((',' (test ':' test)* [','])) |
                   (test (',' test)* [','])) )
arglist: argument (',' argument)* [',']

argument (test | test '=' test)

```

4 Implementation

Both the OPython bytecode compiler and interpreter will be implemented in OCaml. This section gives a rough overview of the process by which OPython source code is handled and executed.

4.1 Input/Output

The OPython program can be run directly in a Unix terminal via the command `opythn`. When run without arguments, this command launches a read-eval-print loop into which the user can enter

commands. In interactive mode, the interpreter evaluates a valid command as soon as it is received:

```
OPytn (Interactive), version 0.0.0
:? for help, :q to quit
]=> print('Hello, world!')
Hello, world!
```

Alternatively, OPytn source code can be defined in a separate `.opy` file and provided to the interpreter as a program argument. In this, mode, OPytn immediately compiles and interprets the source code directly.

4.2 Lexical Analysis

Upon reading source code, the OPytn front-end passes the data as a string to a series of functions that lex the code into tokens. The lexer will be created with the help of `ocamllex`, a program that generates a deterministic finite automaton in OCaml. This resulting lexer matches regular expressions in the string to convert chunks of characters into the correct tokens.

The lexer will produce tokens of the following OCaml datatype:

```
type token =
  NEWLINE | INDENT | DEDENT
| NAME of string (* identifier *)
| KWORD of string
| NUMBER of string
| STRING of string
| OP of string
| DELIM of string
```

4.3 Parsing

The tokens produced by the lexer is then fed into the parser, which will be implemented according to the grammar rules outlined in Section 2.3 and using `menhir`, an OCaml parser generator. The result will be an abstract syntax tree that represents the structure and semantics of the OPytn program.

Typed values are represented as follows:

```
type py_prim =
  INT of int
| FLOAT of float
| STR of string
| BOOL of bool
| NONE
```

There will be a separate type for OPytn compound types and then a general `py_val` type to capture both of these possibilities. To represent an OPytn list, we will need to create a resizing array type in OCaml. To represent a dictionary, we use an OCaml hashtable.

```
type py_comp =
  LIST of py_val Resizable.t (* dynamic array type *)
| DICT of (py_immut, py_val) Hashtbl.t
| TUPLE of py_val array
```



```

| TYPE of ...
| FUN of ...
| OBJ of ...
and type py_val =
  IMMUT of py_prim
| MUT of py_comp

```

4.4 Bytecode

OPython evaluates an abstract syntax tree by generating intermediary bytecode, which is evaluated by a virtual stack machine, as in many conventional Python implementations. As an example of what this looks like, consider this simple algorithm for integer multiplication in terms of addition:

```

x = int(input('Enter the multiplicand: '))
y = int(input('Enter the multiplier: '))

acc = 0
while y > 0:
    if y % 2 == 0:
        x *= 2
        y //= 2
    else:
        acc += x
        y -= 1

print('The product is: ', acc)

```

The disassembled output produced by the CPython interpreter looks like this:

0	LOAD_NAME	0 (int)
2	LOAD_NAME	1 (input)
4	LOAD_CONST	0 ('Enter the multiplicand: ')
6	CALL_FUNCTION	1
8	CALL_FUNCTION	1
10	STORE_NAME	2 (x)
12	LOAD_NAME	0 (int)
14	LOAD_NAME	1 (input)
16	LOAD_CONST	1 ('Enter the multiplier: ')
18	CALL_FUNCTION	1
20	CALL_FUNCTION	1
22	STORE_NAME	3 (y)
24	LOAD_CONST	2 (0)
26	STORE_NAME	4 (acc)
28	SETUP_LOOP	58 (to 88)
30	LOAD_NAME	3 (y)
32	LOAD_CONST	2 (0)

```

34 COMPARE_OP          4 (>)
36 POP_JUMP_IF_FALSE  86

38 LOAD_NAME           3 (y)
40 LOAD_CONST          3 (2)
42 BINARY_MODULO
44 LOAD_CONST          2 (0)
46 COMPARE_OP          2 (==)
48 POP_JUMP_IF_FALSE  68

50 LOAD_NAME           2 (x)
52 LOAD_CONST          3 (2)
54 INPLACE_MULTIPLY
56 STORE_NAME          2 (x)

58 LOAD_NAME           3 (y)
60 LOAD_CONST          3 (2)
62 INPLACE_FLOOR_DIVIDE
64 STORE_NAME          3 (y)
66 JUMP_ABSOLUTE       30

68 LOAD_NAME           4 (acc)
70 LOAD_NAME           2 (x)
72 INPLACE_ADD
74 STORE_NAME          4 (acc)

76 LOAD_NAME           3 (y)
78 LOAD_CONST          4 (1)
80 INPLACE_SUBTRACT
82 STORE_NAME          3 (y)
84 JUMP_ABSOLUTE       30
86 POP_BLOCK

88 LOAD_NAME           5 (print)
90 LOAD_CONST          5 ('The product is: ')
92 LOAD_NAME           4 (acc)
94 CALL_FUNCTION        2
96 POP_TOP
98 LOAD_CONST          6 (None)
100 RETURN_VALUE

```

OPython compiled source will look largely the same, but will not be converted into bitstrings. Instead, an `instr` datatype will be used to capture the same information as CPython's two-byte instructions. A different type tag will be used for each bytecode instruction.

```

type instr =
    LOAD_NAME of int
  | LOAD_CONST of int
  (* ... *)

```

```
| RETURN_VALUE
```

The advantage of this representation is that the interpreter can easily pattern-match on instruction tags. When expressed in this way, the above bytecode would be represented as the following `instr` array.

```
[| LOAD_NAME 0;
   LOAD_NAME 1;
   LOAD_CONST 0;
   CALL_FUNCTION 1;
   CALL_FUNCTION 1;
   STORE_NAME 2;
   (* ... *)
   RETURN_VALUE |]
```

4.5 Virtual Machine

Finally, the bytecode is interpreted by a stack-based virtual machine that produces the desired output. We can define a datatype specifically for stack items, so that any sort of data can be at the top of the stack. Then the stack can be represented as an `item list`, because we will only ever have to access the first few elements in the (linked) list. Many bytecode instructions access and manipulate the stack directly. For example, the instruction `POP_TOP` removes the top element of the stack.

Additionally, the interpreter has access to an array of names and an array of constants, both of which were created upon code generation. For example, the instruction `LOAD_NAME 1` accesses the name array at index 1 and pushes the associated object onto the stack; in our case above, this was the function `input`. Likewise, the instruction `LOAD_CONST 5` accesses the constant array at the index 5 and pushes its value onto the stack.

Control-flow logic is implemented using the stack using certain instructions that tell the interpreter to jump to other parts of the bytecode. For example, at line 46 of the example bytecode, the top two elements on the stack are the values `0` and `y % 2`. The `COMPARE_OP 2` instruction checks if the top two elements of the stack are equal, and then pushes a boolean value onto the stack. The next instruction, `POP_JUMP_IF_FALSE 68`, pops this boolean off the stack and jumps to line 68 if the boolean was false. Loops are implemented in a similar way. The instruction `SETUP_LOOP n` designates the next `n` instructions as a block, and a test has to be run every loop to determine if the block should be exited via a `JUMP` instruction.

The evaluation procedure is performed by a recursive function `eval : int -> item list -> int`, in a context where the instruction, name, and constant arrays are defined. The `eval` function has as its inputs a line number and the current stack, and each time it is called, it performs the appropriate stack manipulations before calling itself again with the next line to evaluate. When `eval` runs out of instructions to evaluate (i.e. reaches the last instruction in the array with no `JUMP` command), the program halts and an exit code is returned.

5 Timeline

The OPython project aims to be complete by the end of June 2019.

5.1 April 2019

The month of April will be dedicated to implementing a vertical slice of the entire OPythn project. To accomplish this, most of the lexer and parser will have to be written, but during this step, OPythn will not support functions and objects yet. Basic types, variable assignment, and control structures will be supported. Ideally, by the end of April, the intermediary compiler should be able to produce bytecode for the small snippet of code in Section 4.4. If possible, during this step we will also be able to implement part of the virtual machine as well, to make it easier to test the features of the language that will be added later on.

References

- [1] Guido van Rossum. *The Python Language Reference*. Python Software Foundation, 2019.