

Matheus Floriano Saito Da Silva

Analizador Léxico para a Linguagem C-: Projeto de Implementação utilizando Máquina de Moore

Relatório técnico de atividade prática solicitado pelo professor Rogerio Aparecido Gonçalves na disciplina de Teoria da computação do Bacharelado em Ciência da Computação da Universidade Tecnológica Federal do Paraná.

Universidade Tecnológica Federal do Paraná – UTFPR

Departamento Acadêmico de Computação – DACOM

Bacharelado em Ciência da Computação – BCC

Campo Mourão

Fevereiro / 2025

Resumo

Este relatório apresenta os desenvolvimentos e resultados da atividade prática realizada, cujo objetivo foi projetar e implementar um analisador léxico para a linguagem C- por meio de uma máquina de Moore e linguagem de programação (Python).

Palavras-chave: Moore. Python. analisador léxico. implementar.

Sumário

1	Introdução e objetivos	4
2	Fundamentação	4
3	Materiais	4
4	Projeto do Autômato	4
5	Descrição do funcionamento do Código	5
5.1	Máquina de Moore	5
5.2	Pré-processamento	8
5.3	Processamento	10
6	Testes e Resultados	11
6.1	Correção dos testes	11
6.2	Teste Final e Meu teste	13
7	Conclusão	14
8	Referências	14

1 Introdução e objetivos

Neste relatório será descrito o trabalho 1 da disciplina de Teoria da computação, o qual visa construir um analisador léxico para a linguagem C- criado por meio de uma máquina de Moore e a linguagem Python. Este analisador léxico recebe por entrada um arquivo contendo um código C- e processa este código devolvendo tokens das operações.

2 Fundamentação

Para a realização deste trabalho foi utilizado como base as instruções, o autômato inicial e o código base inicial fornecido pelo professor, além do capítulo 2 do livro do ([LOUDEN, 2004](#)).

3 Materiais

Foi utilizado um Notebook Dell G15 com as especificações:

- Intel Core i5-10500H (@2.50GHz e 12 *thread*)
- 16GB RAM
- SSD NVMe ADATA 512GB
- Pop!_OS 22.04 LTS

E ferramentas :

- JFLAP Versão 7.1
- Python

4 Projeto do Autômato

Foi construído e utilizado como base o seguinte autômato:

5 Descrição do funcionamento do Código

5.1 Máquina de Moore

```
moore = Moore(
states=['q0', 'q1', 'q2', 'q3', 'q4', 'q5', 'q6', 'q7', 'q8', 'q9', 'q10', 'q11', 'q12', 'q13', 'q14',
↪  'q15', 'q16', 'q17', 'q18', 'q19', 'q20', 'q21', 'q22', 'q23',
    'q24', 'q25', 'q26', 'q27', 'q28', 'q29', 'q30', 'q31', 'q32', 'q33', 'q34', 'q35', 'q36',
↪  'q37', 'q38', 'q39', 'q40', 'q41', 'q42', 'q43', 'q44', 'q45', 'q46', 'q47', 'q48',
    'q49', 'q50', 'q51', 'q52', 'q53', 'q54', 'q55', 'q56', 'q57', 'q58', 'q59', 'q60', 'q61',
↪  'q62', 'q63', 'q64', 'q65', 'qID_START', 'qID_CONT', 'qNUM_START',
↪  'qNUM_CONT', 'qCOMS', 'qCOMB', 'qCOME'],

input_alphabet=list(string.ascii_letters) + list(string.digits) + ['+', '!', '-', '*', '/', '<', '>',
↪  '=', '(', ')', '[', ']', '{', '}', ';', ':', '\n', ' '],

output_alphabet=['INT', 'ELSE', 'IF', 'WHILE', 'FLOAT', 'RETURN', 'VOID', 'MINUS', 'PLUS', 'TIMES',
↪  'DIVIDE', 'DIFFERENT', 'LPAREN', 'RPAREN', 'NUMBER', 'ID',
    'LBRACKETS', 'RBRACKETS', 'COMMA', 'LBRACES', 'RBRACES', 'GREATER', 'GREATER_EQUAL',
↪  'LESS', 'LESS_EQUAL', 'EQUALS', 'SEMICOLON', 'ATtribution'],

transitions={
    'q0': {
        'w': 'q2', 'i': 'q1', 'e': 'q18', 'f': 'q23', 'r': 'q29', 'v': 'q36', '-': 'q41', '+': 'q3',
↪  '': '*': 'q42', '/': 'q43',
        '!': 'q52', '(': 'q55', ')': 'q56', '[': 'q57', ']': 'q58', '{': 'q59', '}': 'q60', '<': 'q4',
↪  '>': 'q44', '=': 'q49',
        ';': 'q61', ':': 'q62', '\n': 'q0', ' ': 'q0',

        **{c: 'qID_START' for c in string.ascii_letters if c not in {'w', 'i', 'e', 'f', 'r', 'v'}}},
```

```

    **{c: 'qNUM_START' for c in string.digits}
},

'qID_START': {
    **{c: 'qID_CONT' for c in string.ascii_letters + string.digits}, # Começa um ID
    **{c: 'q0' for c in [' ', '\n', '+', '-', '*', '/', '<', '>', '=', '(', ')', '[', ']', '{',
        ⇨ '}', ';', ','] } # Termina o ID
},

'qID_CONT': {
    **{c: 'qID_CONT' for c in string.ascii_letters + string.digits}, # Continua como ID
    **{c: 'q0' for c in [' ', '\n', '+', '-', '*', '/', '<', '>', '=', '(', ')', '[', ']', '{',
        ⇨ '}', ';', ','] } # Termina o ID
},

'qCOMB': {c: 'qCOMB' for c in string.printable}, # Consome caracteres dentro do comentário
'qCOMB': {'*': 'qCOME', **{c: 'qCOMB' for c in string.printable if c not in ["*"]}}, # Aguarda
⇨ "*"

'qCOME': {'/': 'q0', **{c: 'qCOMB' for c in string.printable if c not in "/"}} # Fecha comentário

'qNUM_START': {c: 'qNUM_CONT' for c in string.digits}, # Começa a ler número
'qNUM_CONT': {c: 'qNUM_CONT' for c in string.digits}, # Continua lendo número
'qNUM_CONT': {c: 'q0' for c in [' ', '\n', '+', '-', '*', '/', '<', '>', '=', '(', ')', '[', ']',
    ⇨ '{', '}', ';', ','] }, # Número termina

'q1': {' ': 'qID_CONT', 'n': 'q10', 'f': 'q13'},
'q2': {' ': 'qID_CONT', 'h': 'q5'},
'q3': {' ': 'q0'},
'q4': {'=': 'q16', ' ': 'q15'},
'q5': {**{c: 'qID_CONT' for c in string.ascii_letters if c != 'i'}, 'i': 'q6'},
'q6': {**{c: 'qID_CONT' for c in string.ascii_letters if c != 'l'}, 'l': 'q7'},
'q7': {**{c: 'qID_CONT' for c in string.ascii_letters if c != 'e'}, 'e': 'q8'},
'q8': {' ': 'q9'},
'q9': {'\n': 'q0'},
'q10': {**{c: 'qID_CONT' for c in string.ascii_letters if c != 't'}, 't': 'q11'},
'q11': {' ': 'q12'},
'q12': {'\n': 'q0'},
'q13': {' ': 'q14'},
'q14': {'\n': 'q0'},
'q15': {'\n': 'q0'},
'q16': {' ': 'q0'},
'q18': {' ': 'qID_CONT', 'l': 'q19'},
'q19': {**{c: 'qID_CONT' for c in string.ascii_letters if c != 's'}, 's': 'q20'},
'q20': {**{c: 'qID_CONT' for c in string.ascii_letters if c != 'e'}, 'e': 'q21'},
'q21': {' ': 'q22'},
'q22': {'\n': 'q0'},
'q23': {' ': 'qID_CONT', 'l': 'q24'},
'q24': {**{c: 'qID_CONT' for c in string.ascii_letters if c != 'o'}, 'o': 'q25'},
'q25': {**{c: 'qID_CONT' for c in string.ascii_letters if c != 'a'}, 'a': 'q26'},
'q26': {**{c: 'qID_CONT' for c in string.ascii_letters if c != 't'}, 't': 'q27'},
'q27': {' ': 'q28'},
'q28': {'\n': 'q0'},
'q29': {' ': 'qID_CONT', 'e': 'q30'},
'q30': {**{c: 'qID_CONT' for c in string.ascii_letters if c != 't'}, 't': 'q31'},
'q31': {**{c: 'qID_CONT' for c in string.ascii_letters if c != 'u'}, 'u': 'q32'},
'q32': {**{c: 'qID_CONT' for c in string.ascii_letters if c != 'r'}, 'r': 'q33'},
'q33': {**{c: 'qID_CONT' for c in string.ascii_letters if c != 'n'}, 'n': 'q34'},
'q34': {' ': 'q35'},
'q35': {'\n': 'q0'},
'q36': {' ': 'qID_CONT', 'o': 'q37'},

```

```

'q37': {**{c: 'qID_CONT' for c in string.ascii_letters if c != 'i'}, 'i': 'q38'},
'q38': {**{c: 'qID_CONT' for c in string.ascii_letters if c != 'd'}, 'd': 'q39'},
'q39': {' ': 'q40'},
'q40': {'\n': 'q0'},
'q41': {' ': 'q0'},
'q42': {' ': 'q0'},
'q43': {'*': 'qCOMS', ' ': 'q0'},
'q44': {'=' : 'q46', ' ' : 'q45'},
'q45': {'\n': 'q0'},
'q46': {' ' : 'q0'},
'q48': {' ': 'q0'},
'q49': {'=' : 'q48', ' ' : 'q50'},
'q50': {'\n': 'q0'},
'q52': {'=' : 'q53'},
'q53': {' ' : 'q0'},
'q55': {' ' : 'q0'},
'q56': {' ' : 'q0'},
'q57': {' ' : 'q0'},
'q58': {' ' : 'q0'},
'q59': {' ' : 'q0'},
'q60': {' ' : 'q0'},
'q61': {' ' : 'q0'},
'q62': {' ' : 'q0'},

},
initial_state='q0',
output_table={
    'q0': '',
    'q1': '',
    'q2': '',
    'q3': 'PLUS',
    'q4': '',
    'q5': '',
    'q6': '',
    'q7': '',
    'q8': 'WHILE',
    'q9': 'WHILE',
    'q10': '',
    'q11': 'INT',
    'q12': 'INT',
    'q13': '',
    'q14': 'IF',
    'q15': 'LESS',
    'q16': 'LESS_EQUAL',
    'q18': '',
    'q19': '',
    'q20': '',
    'q21': 'ELSE',
    'q22': 'ELSE',
    'q23': '',
    'q24': '',
    'q25': '',
    'q26': '',
    'q27': 'FLOAT',
    'q28': 'FLOAT',
    'q29': '',
    'q30': '',
    'q31': '',
    'q32': '',
    'q33': '',

```

```

    'q34': 'RETURN',
    'q35': 'RETURN',
    'q36': '',
    'q37': '',
    'q38': '',
    'q39': 'VOID',
    'q40': 'VOID',
    'q41': 'MINUS',
    'q42': 'TIMES',
    'q43': 'DIVIDE',
    'q44': '',
    'q45': 'GREATER',
    'q46': 'GREATER_EQUAL',
    'q48': 'EQUALS',
    'q49': '',
    'q50': 'ATTRIBUTION',
    'q52': '',
    'q53': 'DIFFERENT',
    'q55': 'LPAREN',
    'q56': 'RPAREN',
    'q57': 'LBRACKETS',
    'q58': 'RBRACKETS',
    'q59': 'LBRACES',
    'q60': 'RBRACES',
    'q61': 'SEMICOLON',
    'q62': 'COMMA',
    'qID_START': 'ID',
    'qID_CONT': 'ID',
    'qNUM_START': 'NUMBER',
    'qNUM_CONT': 'NUMBER',
    'qCOMS': '',
    'qCOMB': '',
    'qCOME': ''
}
)

```

Esta é a representação direta da máquina de Moore do autômato apresentado adaptado ao código por meio da automata-python. O comportamento base é: todos os símbolos lidos retornam a respectiva saída, quando um caractere da tabela ascii não representa a inicial de uma palavra reservada ele é imediatamente considerado ID, quando o caractere faz parte de uma palavra reservada ele só se torna um ID se não completar a palavra reservada, dígitos são reconhecidos e quando detectam um símbolo retornam a q0, comentários são reconhecidos após ter um caractere "*"depois de "/"e param de ser reconhecidos quando é detectado "*"seguido de "/", vale ressaltar também que todas as transições retornam para q0 quando terminam de processar uma entrada.

5.2 Pré-processamento

```

def preprocess_input(input_string):
    formatted_input = ""

```

```
i = 0

while i < len(input_string):
    char = input_string[i]

    # Verifica se é um operador composto
    if i + 1 < len(input_string) and char in ['<', '>', '!', '=']:
        next_char = input_string[i + 1]
        if next_char == '=':
            formatted_input += f"\n{char}{next_char}\n"
            i += 2 # Pula os dois caracteres
            continue

    #verifica se e comentario
    #abertura de comentário
    if i + 1 < len(input_string) and char in ['/']:
        next_char = input_string[i + 1]
        if next_char == '*':
            formatted_input += f"\n{char}{next_char}\n"
            i += 2
            continue

    #fechamento de comentário
    if i + 1 < len(input_string) and char in ['*']:
        next_char = input_string[i + 1]
        if next_char == '/':
            formatted_input += f"\n{char}{next_char}\n"
            i += 2
            continue

    # Se for um delimitador isolado, adiciona espaçamento
    if char in ' (){};,+-*/<>= ![] ':
        formatted_input += f" \n{char} \n"
    elif char == ' ':
        formatted_input += ' ' # Mantém espaços simples
    elif char == '\n':
        formatted_input += '\n' # Mantém novas linhas
    else:
        formatted_input += char # Mantém o caractere

    i += 1 # Avança para o próximo caractere
```

```

formatted_input += '\n'
# print (formatted_input)
return formatted_input.strip()

```

Esta função é responsável por tratar o código diretamente fornecido pelo arquivo, a ideia aqui é simplificar os dados passados para o autômato, neste caso o principal seria passar cada seção do código original em formato de: operação <quebra de linha> como por exemplo um "return(0);" seria transformado em :

```
return ( 0 ) ;
```

Esta função também foi utilizada para tratar um problema com operadores compostos, onde ao separar != por exemplo, teríamos um problema de identificação.

5.3 Processamento

```

def process_input(input_string):
    tokens = []
    current_state = moore.initial_state
    token = ""

    #formatando tokens

    for char in input_string:
        if char in moore.input_alphabet:
            next_state = moore.transitions[current_state].get(char, 'q0')

            if next_state == 'q0': # Finalizou um token
                if moore.output_table.get(current_state):
                    tokens.append(moore.output_table[current_state]) # Acessa
                    ↪ e guarda o output da máquina

                token = "" # Reinicia o token
                current_state = moore.initial_state
            else:
                token += char # Continua construindo o token
                current_state = next_state
        else:
            error_handler.handle_error(f"Unexpected character: {char}")
    return tokens

```

```

# Garante que o último token seja adicionado
if moore.output_table.get(current_state):
    tokens.append(moore.output_table[current_state])
# print(tokens)
return tokens

```

Esta função é responsável por lidar com os tokens sem si, obtendo as saídas da máquina de Moore e as colocando em uma lista de strings onde cada string representa um token.

6 Testes e Resultados

6.1 Correção dos testes

Vale a pena ressaltar que alguns testes estavam errados, e um apenas foi ajustado para funcionar de acordo com a minha implementação como demonstrado a seguir:

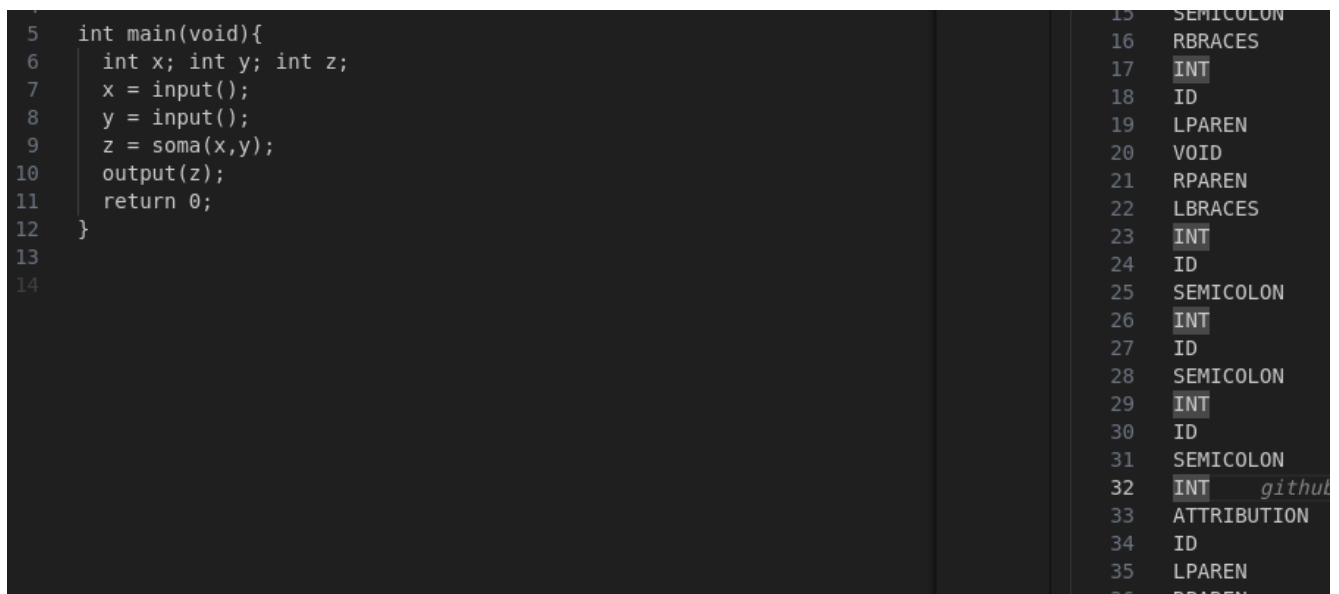


Figura 2 – Teste 004

aqui a linha 32 descreve como INT quando na verdade é um ID.

5	}	51	LPAREN
6		52	RPAREN
7	void main(void){	53	SEMICOLON
8	int x; int y;	54	ID
9	x = input();	55	ATTRIBUTION
10	y = input();	56	ID
11	output(gcd(x,y));	57	LPAREN
12	}	58	RPAREN <i>git</i>
13		59	SEMICOLON
14		60	ID
		61	LPAREN
		62	ID
		63	LPAREN
		64	ID
		65	COMMA
		66	ID
		67	RPAREN
		68	RPAREN
		69	SEMICOLON
		70	RBRACES
		71	

Figura 3 – Teste 005

aqui faltou um RPAREN.

8	void output(int x){	47	SEMICOLON
9		48	IF
10	}	49	LPAREN
11		50	ID
12	int func(int x, int y){	51	DIFFERENT
13	int res;	52	NUMBER
14	res = x + y - 2; <i>github-classroom[bot] [8 weeks ago] • Initial c</i>	53	RPAREN
15	if(res != 0){	54	LBRACES
16	res = -1;	55	ID
17	}	56	ATTRIBUTION
18	return(res);	57	MINUS <i>Not</i>
19	}	58	NUMBER
20		59	SEMICOLON
21	int main(void){	60	RBRACES
22	a = input();	61	RETURN
23	b = input();	62	LPAREN
24		63	ID
25	b[0] = a;	64	RPAREN
26	b[1] = func(a,b);	65	SEMICOLON
27		66	RBRACES
28	output(a);	67	INT
29	return(0);	68	RPAREN

Figura 4 – Teste 006

aqui ocorre uma diferença entre implementações, no meu caso está sendo considerado MINUS, NUMBER, enquanto o teste original considera um número negativo como apenas NUMBER.

2	int b[10];	2	ID
3	int c[3][5];	3	SEMICOLON
4		4	INT
5	b = 10	5	ID
6		6	LBRACKETS
7	int func(int x, int y){	7	NUMBER
8	int res;	8	RBRACKETS
9	res = x + y;	9	SEMICOLON
10	return(res);	10	INT
11	}	11	ID
12		12	LBRACKETS
13	int main(){	13	NUMBER
14	a = input();	14	RBRACKETS
15	b = input();	15	LBRACKETS
16		16	NUMBER
17	b[0] = a;	17	RBRACKETS
18	b[1] = b;	18	SEMICOLON
19	c[0][1] = func(a,b);	19	ID
20		20	ATTRIBUTION
21	output(c[3]);	21	NUMBER
22	return(0);	22	INT
23	}	23	ID

Figura 5 – Teste 007

aqui após o NUMBER tínhamos um SEMICOLON adicional após $b = 10$.

6.2 Teste Final e Meu teste

Após todas as mudanças nos testes como descrito na seção 6.1 o resultado do teste final:

```

===== 11 passed in 0.32s =====
(.venv) saito@pop-os:~/Documents/automaton/anallex-MattSaito$ pytest -v
===== test session starts =====
platform linux -- Python 3.10.12, pytest-8.3.4, pluggy-1.5.0 -- /home/saito/Documents/automaton/.venv/bin/python3
cachedir: .pytest cache
rootdir: /home/saito/Documents/automaton/anallex-MattSaito
configfile: pytest.ini
collected 11 items

anallex_test.py::test_execute[--k] PASSED [ 9%]
anallex_test.py::test_execute[test.c--k] PASSED [ 18%]
anallex_test.py::test_execute[notexists.cm--k] PASSED [ 27%]
anallex_test.py::test_execute[prog-003.cm--k] PASSED [ 36%]
anallex_test.py::test_execute[prog-004.cm--k] PASSED [ 45%]
anallex_test.py::test_execute[prog-005.cm--k] PASSED [ 54%]
anallex_test.py::test_execute[prog-007.cm--k] PASSED [ 63%]
anallex_test.py::test_execute[prog-000.cm--k] PASSED [ 72%]
anallex_test.py::test_execute[prog-001.cm--k] PASSED [ 81%]
anallex_test.py::test_execute[prog-006.cm--k] PASSED [ 90%]
anallex_test.py::test_execute[prog-002.cm--k] PASSED [100%]

===== 11 passed in 0.33s =====

```

Figura 6 – Testes

Também fiz um arquivo de teste que demonstra todas as funções do lexer e a comparação entre código e saída de tokens pode ser vista abaixo:

1	INT	45	LBRACES	88	ID	131	VOID	1	int main()
2	ID	46	RETURN	89	EQUALS	132	ID	2	int x1, x2, x3;
3	LPAREN	47	SEMICOLON	90	ID	133	SEMICOLON	3	int y1, y2;
4	RPAREN	48	RBRACES	91	RPAREN	134	ID	4	void funcaoTeste();
5	LBRACES	49	ELSE	92	LBRACES	135	ATtribution	5	
6	INT	50	LBRACES	93	ID	136	NUMBER	6	x1 = 10 + 20 - 5 * 2 / 1;
7	ID	51	ID	94	ATtribution	137	SEMICOLON	7	y1 = x2;
8	COMMA	52	ATtribution	95	ID	138	RETURN	8	if (x1 < x2) {
9	ID	53	ID	96	RBRACES	139	ID	9	return;
10	COMMA	54	PLUS	97	SEMICOLON	140	ATtribution	10	} else {
11	ID	55	NUMBER	98	IF	141	NUMBER	11	x1 = x1 + 1;
12	SEMICOLON	56	SEMICOLON	99	LPAREN	142	SEMICOLON	12	}
13	INT	57	RBRACES	100	ID	143	Float	13	
14	ID	58	WHILE	101	LESS_EQUAL	144	ID	14	while (x1 != 100) {
15	COMMA	59	LPAREN	102	ID	145	ATtribution	15	x1 = x1 + 1;
16	ID	60	ID	103	RPAREN	146	NUMBER	16	}
17	SEMICOLON	61	DIFFERENT	104	LBRACES	147	SEMICOLON	17	
18	VOID	62	NUMBER	105	ID	148	VOID	18	x2 = 100;
19	ID	63	RPAREN	106	ATtribution	149	ID	19	y2 = 42;
20	LPAREN	64	LBRACES	107	ID	150	ATtribution	20	x1 = x2 > 50;
21	RPAREN	65	ID	108	RBRACES	151	NUMBER	21	if(x3 == x1) {x3 = x1};
22	SEMICOLON	66	ATtribution	109	SEMICOLON	152	SEMICOLON	22	if(x2 <= x3) {x2 = x3};
23	ID	67	ID	110	IF	153	Float	23	if (x1 >= x2) {x1 = x2};
24	ATtribution	68	PLUS	111	LPAREN	154	ID	24	
25	NUMBER	69	NUMBER	112	ID	155	COMMA	25	int ifelse, returnValue, floatNumber, voidPointer;
26	PLUS	70	SEMICOLON	113	GREATER_EQUAL	156	ID	26	ifelse = 42;
27	NUMBER	71	RBRACES	114	ID	157	COMMA	27	returnValue = 99;
28	MINUS	72	ID	115	RPAREN	158	ID	28	floatNumber = 314;
29	NUMBER	73	ATtribution	116	LBRACES	159	COMMA	29	voidPointer = 0;
30	TIMES	74	NUMBER	117	ID	160	ID	30	
31	NUMBER	75	SEMICOLON	118	ATtribution	161	SEMICOLON	31	/* Este e um
32	DIVIDE	76	ID	119	ID	162	RETURN	32	comentario de multiplas linhas */
33	NUMBER	77	ATtribution	120	RBRACES	163	NUMBER	33	Not committed yet
34	SEMICOLON	78	NUMBER	121	SEMICOLON	164	SEMICOLON	34	float a,b,c,e;
35	ID	79	SEMICOLON	122	INT	165	RBRACES	35	
36	ATtribution	80	ID	123	ID	166		36	return 0;
37	ID	81	ATtribution	124	COMMA			37	
38	SEMICOLON	82	ID	125	RETURN			38	
39	IF	83	GREATER	126	ID				
40	LPAREN	84	NUMBER	127	COMMA				
41	ID	85	SEMICOLON	128	Float				
42	LESS	86	IF	129	ID				
43	ID	87	LPAREN	130	COMMA				
44	RPAREN	88	TO	131	VOID				

Figura 7 – MyTest

7 Conclusão

O projeto em si foi realizado com sucesso e os resultados finais foram favoráveis, durante a implementação do projeto foi possível obter mais conhecimento sobre um analisador léxico e sobre como lidar com autômatos em Python.

8 Referências

LOUDEN, K. C. *Compiladores: princípios e práticas*. 2004. [Accessed 27-01-2025]. Citado na página 4.