## Tradutores - Analisador Semântico

Pedro Lucas Pinto Andrade - 160038316

Ciência da Computação, Universidade de Brasília, Brasília DF, Brasil pedrolpandrade@gmail.com

## 1 Motivação

O funcionamento de um tradutor é dividido em diversas etapas, sendo uma delas a etapa de análise. A análise também é dividida em quatro tipos principais: análise léxica, análise sintática, análise semântica e geração de código intermediário, cada uma possuindo um papel específico na execução do tradutor [ALSU07]. As fases têm como objetivo: ler um arquivo de entrada e separar o código em tokens seguindo as regras especificadas pelas expressões regulares (léxico); analisar se os tokens estão organizados conforme o conjunto de regras da gramática da linguagem (sintático); verificar se a semântica do código está condizente com o definido para a linguagem (semântico); e gerar o código intermediário que é utilizado para, posteriormente, gerar o código de máquina (geração de código intermediário).

Nesse trabalho, a linguagem escolhida é uma versão simplificada de C chamada C-IPL [Nal21], com a adição de um tipo novo de dados para listas (*list*), além de primitivas para manipular os dados nessas listas.

Com essas operações para manipulação dos elementos, uma lista se torna uma estrutura de dados versátil, que permite implementar diversas outras estruturas de dados por meio de abstrações, como pilhas ou filas. A adição da lista ao subconjunto da linguagem que foi selecionado para o trabalho permite armazenar e manipular dados de formas variadas.

## 2 Descrição

Para realizar a análise léxica, foram criadas definições (Tabela 1) com o uso de expressões regulares para separar o código de entrada em tokens, com o auxílio do gerador Flex. As expressões regulares determinam padrões para: identificadores; tipos de dados; operadores, como aritméticos, binários, lógicos e os operadores utilizados para manipular as listas; delimitadores, como parênteses, chaves, aspas e delimitadores de comentários; e marcadores utilizados para a formatação, como espaços e quebras de linha. Sempre que um lexema é lido e reconhecido por uma das expressões regulares, uma função correspondente dessa expressão é executada, informando qual o tipo do token e utilizando seu tamanho para calcular a sua posição no código de entrada.

Para esse posicionamento em relação à linha e coluna, são utilizadas variáveis globais para armazenar em qual posição está o *token* que foi lido por último,

calculadas com o auxílio da variável *yylen*, incluída pelo Flex, que guarda o tamanho do lexema lido. O posicionamento também é utilizado para a impressão de erros léxicos quando um lexema não segue as regras do conjunto de definições regulares.

Após ler e reconhecer cada lexema, o analisador léxico monta e envia os tokens para o analisador sintático, que cuida da próxima etapa da tradução.

O analisador sintático utiliza os tokens recebidos do léxico para verificar se o código está seguindo a sintaxe definida pelo conjunto de regras da gramática da linguagem C-IPL. Para isso, é construído um analisador, com o auxílio do Bison [DS21], que utiliza tabelas LR(1) canônicas para interpretar o código conforme as regras da gramática.

O analisador sintático é responsável por armazenar todas as unidades léxicas relevantes na tabela de símbolos, como declarações de variáveis e de funções, assim como sua posição no código e seu escopo. O analisador léxico também pode ser responsável por armazenar alguns dos *tokens* na tabela, mas, para simplificar o funcionamento do tradutor nesse trabalho, todas as inserções na tabela foram mantidas na fase sintática.

Os registros na tabela de símbolos serão úteis durante a fase semântica da análise, onde o programa poderá utilizar as informações armazenadas para determinar se as operações do código de entrada estão seguindo as regras semânticas definidas para a linguagem C-IPL, como verificar se uma variável está sendo utilizada dentro de seu escopo, se uma função está recebendo todos os parâmetros necessários, entre outras.

A tabela de símbolos foi implementada utilizando uma lista de símbolos [Aab04]. Um símbolo é representado por uma estrutura que contém os campos relevantes para a tabela (nome, posicionamento em linha e coluna, escopo, escopo pai e se é uma variável ou função) e campos relevantes para o funcionamento da lista (um ponteiro para o próximo elemento). Dois ponteiros globais são utilizados para auxiliar na implementação da lista, um para o começo e outro para o último elemento inserido.

Uma pilha é utilizada para determinar o escopo das declarações que são armazenadas na tabela de símbolos. Ao encontrar uma abertura de escopo, um novo valor é empilhado, para que todas as declarações dentro desse bloco de código recebam esse valor na tabela. Ao encontrar o fechamento do escopo, o valor é desempilhado. Um contador global é utilizado para garantir que o valor de cada escopo é único. O valor de escopo pai é uma forma de determinar qual símbolo, se existir, foi responsável por abrir um escopo onde uma determinada variável foi declarada. Isso auxilia na impressão da tabela de símbolos, deixando mais claro onde cada declaração ocorreu.

A árvore sintática é construída pelo analisador sintático para exibir quais foram as derivações da gramática realizadas para obter o código C-IPL de entrada, mostrando assim como o código está dentro da sintaxe definida para a linguagem. Entretanto, a árvore construída não é a de derivação completa, mas sim a abstrata. Portanto, algumas regras são omitidas de forma a facilitar a leitura da sintaxe.

Cada nó da árvore é uma estrutura (Lista 1.1) com seu nome e ponteiros para seus nós filhos, para que seja possível encadear e exibir todos em formato de árvore ao final da execução. Além disso, cada um também possui uma flag para determinar se o mesmo é um nó vazio. Os nós vazios são criados nas derivações vazias e não são necessários para a impressão, mas sim para o funcionamento correto da árvore. Algumas regras da gramática possuem o intuito de gerar uma lista de outras regras com um número indeterminado de elementos, como a regra de declarações locais, por exemplo. Com a montagem da árvore, isso acaba gerando uma sequência de nós repetidos que dificulta a leitura. Para resolver isso, os nós repetidos são tratados e todos os filhos são impressos como filhos de um único nó [Zan11].

Os nós também possuem outras informações úteis para a análise semântica, como a sua sigla (que é utilizado para identificar cada nó), o seu tipo, a que função está atrelado (no caso de uma chamada de função), sua posição no código em linha e coluna e um identificador (nos casos onde um identificador é utilizado explicitamente na regra do nó).

```
typedef struct Node{
    char* name;
    char* sigla;
    int empty;
    struct Node** child;
    char* type;
    char* functionName;
    int line;
    int column;
    char* id;
} t node;
```

Lista 1.1. Estrutura do nó

O analisador semântico utiliza a árvore sintática e a tabela de símbolos para verificar se a semântica do código C-IPL está correta. Para isso, algumas regras semânticas são verificadas:

- Existência de uma definição correta da main;
- Parâmetros corretos (número e tipo) em chamadas de funções;
- Existência de um identificador utilizado em um escopo válido;
- Tipos de operandos válidos para operadores;

Para verificar as regras semânticas uma estratégia de 2 passagens foi adotada, pois algumas regras são verificadas mais facilmente durante a análise sintática, enquanto outras são verificadas mais facilmente utilizando a árvore sintática e a tabela de símbolos completas. Durante a análise sintática e a análise semântica os tipos dos nós da árvore são preenchidos para que seja possível realizar a verificação da semântica da linguagem. Alguns tipos são preenchidos diretamente, como o de constantes, enquanto outros são preenchidos com base nos nós filhos, como o de expressões.

Durante a análise sintática, são verificados os números e os tipos de parâmetros em uma chamada de função, se os identificadores utilizados foram declarados em um escopo válido e se os tipos dos operandos de um operador seguem os definidos pela semântica.

Para verificar os parâmetros, ao encontrar uma declaração de função a análise sintática registra na tabela de símbolos quais são os parâmetros necessários para chamar a função. Ao encontrar uma chamada de função, a análise semântica armazena os tipos dos parâmetros passados, busca na tabela de símbolos onde está a declaração da função e compara os tipos dos parâmetros passados com os parâmetros armazenados. Caso o número seja discrepante, um erro é exibido. Caso um tipo seja discrepante, o analisador verifica se é possível converter o argumento passado para o tipo esperado. Caso não seja possível converter entre os tipos, um erro é exibido.

Em C-IPL um identificador só pode ser utilizado após sua declaração. Ao encontrar um identificador sendo utilizado na análise sintática, a análise semântica busca na tabela de símbolos se existem declarações desse identificador e, caso sejam encontrados, percorre a pilha procurando quais escopos ainda estão válidos. Um escopo é válido se o mesmo ainda não foi fechado, o que implica que o ponto atual do código está dentro desse escopo. Caso a declaração encontrada faça parte de um dos escopos válidos, o identificador pode ser usado. Caso nenhuma declaração satisfaça isso, um erro é exibido.

A verificação dos tipos dos operandos de um operador é a etapa mais complexa da análise semântica. Ao encontrar uma das regras que precisam dessa verificação, a análise sintática passa o nó dessa regra para a função que realiza essa etapa. Essa função verifica qual é a classe do nó por meio de sua sigla para determinar quais são as regras semânticas que devem ser observadas em cada caso. A verificação então utiliza o tipo dos nós filhos e outras informações armazenadas para verificar se o nó em questão está seguindo a semântica determinada pela linguagem. Alguns casos são mais simples, como em expressões de soma ou multiplicação, onde é preciso verificar se os elementos são do tipo simples, enquanto outros são mais complexos, como em operações de map e filter, onde é preciso verificar se um dos parâmetros é uma função unária com os tipos de parâmetro e de retorno corretos.

Durante o preenchimento dos tipos de cada nó e durante essa verificação dos tipos dos operandos podem ocorrer casos onde é necessário converter o tipo de um nó. Na linguagem só é possível converter entre os tipos inteiro e ponto flutuante. Os tipos de lista não podem ser convertidos. São vários casos onde a conversão é realizada, como: em operações aritméticas caso hajam operandos de tipos diferentes, onde inteiros são convertidos para ponto flutuantes; em uma construção de lista, onde os elementos são convertidos para o tipo básico da lista; em uma chamada de função, onde os argumentos são convertidos para os tipos definidos na declaração da função; no map e filter, onde os elementos da lista que serão passados para a função unária são convertidos para o tipo do parâmetro da função; em operações relacionais, onde os inteiros são convertidos para ponto flutuante caso haja um ponto flutuante na expressão; em uma atribuição, onde os

elementos são convertidos para o tipo da variável; em operações com a constante *nil*, onde ela é convertida para um tipo de lista; e outras. A constante *nil* é o único caso em C-IPL onde ocorre uma unificação de tipo. Isso é anotado na árvore como uma conversão, apesar de não ser possível converter para um tipo de lista, apenas para armazenar essa informação. A etapa de análise semântica não realiza a conversão, mas apenas anota na árvore onde ela deve ser realizada utilizando um nó simbólico.

Após o término da análise sintática e da construção completa da árvore sintática e da tabela de símbolos, é analisado se existe uma definição correta da *main*. Para isso, a tabela de símbolos é percorrida em busca de declarações de uma função com o nome *main*. Caso não exista ou exista mais de uma, um erro é exibido.

### 3 Arquivos de teste

Estão disponíveis quatro arquivos de teste do analisador dentro da pasta /tests/. Os arquivos são divididos em arquivos sem erros:

```
- test_correct1.c;
- test_correct2.c;
```

E arquivos com erros semânticos:

- test\_wrong1.c Erros sintáticos: for inesperado (linha 12 coluna 8); read inesperado (linha 75 coluna 9); Erros semânticos: tipo de operando inválido (linha 36 coluna 15); função não declarada (linha 39 coluna 27); função não declarada (linha 40 coluna 29); tipo de operando inválido (linha 40 coluna 34); tipo de operando inválido (linha 43 coluna 15); sem definição da função main (linha -1 coluna -1);
- test\_wrong2.c Erros sintáticos: id inesperado (linha 16 coluna 20); ponto e vírgula inesperado (linha 35 coluna 40); Erros semânticos: tipo de operando inválido (linha 8 coluna 50); tipo de operando inválido (linha 30 coluna 15); tipo de operando inválido (linha 31 coluna 16); número de parâmetros incorreto (linha 33 coluna 9); redefinição da função main (linha 40 coluna 9);

# 4 Instruções para compilação e execução

Um arquivo Make foi incluído para facilitar a compilação do tradutor. Basta utilizar o comando

\$ make semantico

na pasta principal.

Caso ocorra algum problema com o arquivo Make, é possível compilar executando os seguintes comandos em uma janela do terminal:

Após compilar, basta executar os seguintes comandos para rodar os arquivos de teste:

```
$ make run_correct1
$ make run_correct2
$ make run_wrong1
$ make run wrong2
```

Para executar o analisador em outros arquivos, basta utilizar o seguinte comando:

```
$ ./tradutor teste.c
```

trocando a palavra teste pelo nome de um arquivo na linguagem C-IPL.

As versões dos sistemas utilizados para compilação foram: Flex (flex 2.6.4), Bison (GNU Bison 3.7.5 compilado do código fonte), Make (GNU Make 4.2.1), GCC (11.2.0 compilado do código fonte), OS (Linux 5.8.0-63-generic 20.04.1-Ubuntu).

### Referências

- [Aab04] A. A. Aaby. Compiler construction using flex and bison. https://www.admb-project.org/tools/flex/compiler.pdf. Acessado em 14 Set 2021, Fevereiro 2004.
- [ALSU07] A. Aho, M. Lam, R. Sethi, and J. Ullman. Compilers: Principles, Techniques, and Tools. Pearson Addison Wesley, 2nd edition, 2007.
- [DS21] C. Donnelly and R. Stallman. Bison The Yacc-compatible Parser Generator. https://www.gnu.org/software/bison/manual/bison.pdf. Acessado em 14 Set 2021, Setembro 2021.
- [Hec21] R. Heckendorn. A grammar for the C- programming language (version s21). http://marvin.cs.uidaho.edu/Teaching/CS445/c-Grammar.pdf. Acessado em 14 Set 2021, Março 2021.
- [Nal21] C. Nalon. Trabalho prático descrição da linguagem. https://aprender3.unb.br/mod/page/view.php?id=464034. Acessado em 14 Set 2021, Agosto 2021.
- $[Zan11] \quad B. \quad V. \quad Zanden. \quad Building \quad Abstract \quad Syntax \quad Trees. \\ \quad http://web.eecs.utk.edu/~bvanderz/cs461/notes/parse\_tree/. \quad Acessado \\ \quad em \quad 15 \ Set \ 2021, \ Fevereiro \ 2011.$

#### A Tabelas léxicas

A Tabela 1 mostra os tokens criados para o analisador léxico, além das expressões regulares utilizadas para definir cada um. A Tabela 2 mostra os lexemas de cada token e seu valor de atributo.

## B Gramática da linguagem

A gramática foi criada com base na gramática para a linguagem C- [Hec21] e nas regras padrões de C das operações que fazem parte da linguagem reduzida definida para o trabalho.

- 1.  $program \rightarrow declarationList$
- 2.  $delarationList \rightarrow declarationList \ declaration \ | \ declaration \ | \ declarationList \ statement \ | \ statement$
- 3.  $declaration \rightarrow varDeclaration \mid funcDeclaration$
- 4.  $varDeclaration \rightarrow \mathbf{TYPE} \; \mathbf{ID} \; \mathbf{DELIM\_SEMICOLLON} \; | \; \mathbf{TYPE} \; \mathbf{LIST} \; \; \mathbf{TYPE} \; \mathbf{ID} \; \mathbf{DELIM} \; \; \mathbf{SEMICOLLON} | \;$
- 5. funcDeclaration → TYPE ID DELIM\_PARENT\_L parameters DE-LIM\_PARENT\_R bodyStatement | TYPE LIST\_TYPE ID DELIM\_PARENT\_L parameters DELIM\_PARENT\_R bodyStatement
- 6.  $parameters \rightarrow parameterList \mid \epsilon$
- 7. parameterList → parameterList **DELIM\_COMMA** parameterSimple | parameterSimple
- 8.  $parameterSimple \rightarrow \mathbf{TYPE} \; \mathbf{ID} \mid \mathbf{TYPE} \; \mathbf{LIST} \; \; \mathbf{TYPE} \; \mathbf{ID}$
- 9. statement → bodyStatement | ifStatement | loopStatement | returnStatement | listStatement DELIM\_SEMICOLLON | writeOp DELIM\_SEMICOLLON | readOp DELIM\_SEMICOLLON | expressionStatement
- 10.  $bodyStatement \rightarrow \mathbf{DELIM\_CUR\_BRACKET\_L}$  statementList  $\mathbf{DE-LIM\_CUR\_BRACKET\_R}$
- 11.  $localDeclaration \rightarrow localDeclaration \ varDeclaration \ | \epsilon$
- 12.  $statementList \rightarrow statementList\ localDeclaration\ statement\ |\ \epsilon$
- 13. ifStatement → IF\_KEY DELIM\_PARENT\_L
  simpleExpression DELIM\_PARENT\_R statement |
  IF\_KEY DELIM\_PARENT\_L simpleExpression
  DELIM\_PARENT\_R statement ELSE\_KEY statement
- 14. loopStatement → FOR\_KEY DELIM\_PARENT\_L expression DE-LIM\_SEMICOLLON simpleExpression DELIM\_SEMICOLLON expression DELIM PARENT R statement
- 15. returnStatement → RETURN\_KEY expression DELIM\_SEMICOLLON
- 16.  $expression \rightarrow ID ASSIGN OP expression | simple Expression$
- 17.  $simpleExpression \rightarrow logicBinExpression$

Tabela 1. Definições dos tokens no código Flex.

Token	Expressão regular	Exemplo de lexema
DIGIT	[0-9]	0, 6
INT	{DIGIT}+	1955
FLOAT	{DIGIT}+[.]{DIGIT}+	11.05
ID	[a-zA-Z ][a-z0-9A-Z]*	main
TYPE	int float	int
LIST TYPE	list	list
STRING	\"(\\. [^"\\])*\"	"texto"
NULL CONST	nil	nil
PLUS OP	+	+
MINUS OP	_	_
DIV_OP	/	/
MUL OP	*	*
LOGIC OP	&&	, &&
BINARY OP		==,>
ASSIGN OP		
EXCLA OP	[i]	!
IF KEY	if	if
ELSE KEY	else	else
FOR KEY	for	for
RETURN KEY	return	return
INPUT KEY	read	read
OUTPUT KEY	write	write
OUTPUTLN KEY	writeln	writeln
ASSIGN LISTOP	:	:
HEADER LISTOP	?	?
TAILDES LISTOP	%	%
MAP LISTOP	»	»
FILTER LISTOP	«	<b>«</b>
DELIM PARENT L	(	(
DELIM_PARENT_R	)	)
DELIM BRACKET L	[	
DELIM_BRACKET_R		
DELIM_CUR_BRACKET_L	{	[{
DELIM CUR BRACKET R	}	}
DELIM_COMMA	[,]	,
DELIM SEMICOLLON	[;]	;
DELIM_SQUOTE	[']	,
DELIM_DQUOTE	["]	"
SINGLE_COMMENT	//[^\n]*	// texto
MULTI_COMMENT	\/\*[^(\*\/)]*\*\/	/* texto */
FORMAT_BLANKSPACE		espaço em branco
FORMAT_NEWLINE	\\n	quebra de linha
FORMAT_TAB	$\setminus t$	tab

 ${\bf Tabela~2.~Tokens~e~seus~lexemas~e~valores~de~atributo~correspondentes}$ 

Lexemas	Token	Valor de atributo
Inteiro	INT	Ponteiro para a tabela de símbolos
Decimal	FLOAT	Ponteiro para a tabela de símbolos
Id	ID	Ponteiro para a tabela de símbolos
int	TYPE	INT
float	TYPE	FLO
list	LIST TYPE	LIST
Cadeia de caracteres	_	Ponteiro para a tabela de símbolos
	NULL CONST	Fonteiro para a tabela de simbolos
nil	PLUS OP	DI LIC
+		PLUS
-	MINUS OP	MINUS
/  *	DIV_OP	DIV
	MUL_OP	MUL
&&	LOGIC_OP	AND
	LOGIC_OP	OR
<	BINARY_OP	LT
< <= > >= ==	BINARY_OP	LE
>	BINARY_OP	GT
>=	BINARY_OP	GE
	BINARY_OP	EQ
!=	BINARY OP	DIF
	ASSIGN OP	_
!	EXCLA OP	_
if	$\overline{\text{IF KEY}}$	IF
else	ELSE KEY	ELS
for	FOR KEY	FOR
return	RETURN KEY	RET
read	INPUT KEY	-
write	OUTPUT KEY	BASE
writeln	OUTPUTEN KEY	LN
«	LIST OP	FILT
»	LIST OP	MAP
?	LIST OP	HEAD
· %	LIST OP	TAIL
	LIST OP	CONS
	<del>-</del>	
	DELIM_PARENT_L DELIM_PARENT_R	LEFT
)	DELIM_PARENT_R DELIM_BRACKET_L	RIGHT
		LEFT
	DELIM_BRACKET_R	RIGHT
{	DELIM_CUR_BRACKET_L	LEFT
}	DELIM_CUR_BRACKET_R	RIGHT
,	DELIM_COMMA	-
;	DELIM_SEMICOLLON	-
, , , , , , , , , , , , , , , , , , ,	DELIM_SQUOTE	-
"	DELIM_DQUOTE	-
//texto	SINGLE_COMMENT	-
/*texto*/	DOUBLE_COMMENT	-
Espaço em branco	FORMAT_BLANKSPACE	-
$\setminus n$	FORMAT NEWLINE	-
\t	FORMAT TAB	-

- 18. logicBinExpression → logicBinExpression LOGIC\_OP logicUnExpression | logicUnExpression
- 19.  $logicUnExpression \rightarrow \mathbf{EXCLA}$  **OP**  $logicUnExpression \mid binExpression$
- 20.  $binExpression \rightarrow binExpression$  BINARY\_OP  $listExpression \mid listExpression$
- 21.  $sumExpression \rightarrow sumExpression \ sumOP \ mulExpression \ | \ mulExpression$
- 22.  $mulExpression \rightarrow mulExpression mulOP factor | factor$
- 23.  $sumOP \rightarrow PLUS_OP \mid MINUS_OP$
- 24.  $mulOP \rightarrow \mathbf{MUL} \quad \mathbf{OP} \mid \mathbf{DIV} \quad \mathbf{OP}$
- 25. factor → ID | functionCall | DELIM\_PARENT\_L simpleExpression DELIM PARENT R | listExpression | constant
- 26.  $constant \rightarrow INT \mid MINUS\_OP INT \mid FLOAT \mid MINUS\_OP FLOAT \mid NULL CONST$
- 27. functionCall → ID DELIM\_PARENT\_L
  parametersPass DELIM\_PARENT\_R
- 28. parametersPass → parametersPass **DELIM\_COMMA** simpleExpression | simpleExpression | ε
- 29.  $writeOp \rightarrow write \mid writeln$
- 30. write → OUTPUT\_KEY DELIM\_PARENT\_L STRING
  DELIM\_PARENT\_R | OUTPUT\_KEY DELIM\_PARENT\_L
  simpleExpression DELIM\_PARENT\_R
- 31. writeln → OUTPUTLN\_KEY DELĪM\_PARENT\_L STRING
  DELĪM\_PARENT\_R | OUTPUTLN\_KEY
  DELĪM\_PARENT\_L simpleExpression DELĪM\_PARENT\_R
- 32.  $readOp \rightarrow INPUT\_KEY DELIM\_PARENT\_L ID$ DELIM PARENT R
- 33.  $expressionStatement \rightarrow expression DELIM SEMICOLLON$
- 34.  $listExpression \rightarrow listAssign$
- 35.  $listAssign \rightarrow listAssign$  **ASSIGN LISTOP**  $listHeader \mid listHeader$
- 36.  $listHeader \rightarrow \textbf{HEADER}$  LISTOP  $listHeader \mid listTailDestructor$
- 37.  $listTailDestructor \rightarrow TAI\overline{LDES}$  LISTOP  $listTailDestructor \mid listMap$
- 38.  $listMap \rightarrow listMap \text{ MAP\_LISTOP } listFilter \mid listFilter$
- 39.  $listFilter \rightarrow listFilter$  **FILTER LISTOP**  $sumExpression \mid sumExpression$