UAlg FCT
UNIVERSIDADE DO ALGARVE
FACULDADE DE CIÊNCIAS E TECNOLOGIA

Compiladores

2020-2021

Projecto 1 – Análise Lexical

Introdução

No 1.º Projeto pretende-se implementar um analisador lexical para a linguagem ALG, cujo compilador será desenvolvido na cadeira de compiladores ao longo do semestre. A linguagem ALG é baseada na linguagem FIR[1].

Usando a ferramenta ANTLR deverão especificar um ficheiro de regras lexicais, com o nome *algLexer.g4* que reconheça os *tokens* da linguagem ALG, de acordo com a especificação que se apresenta de seguida. Cabe a cada grupo decidir que *tokens* serão retornados, e os seus nomes.

Linguagem ALG (especificação do léxico)

A linguagem ALG é uma linguagem imperativa, fracamente tipificada.

Palavras-chave

As palavras indicadas de seguida estão reservadas (palavras-chave), não podendo ser utilizadas como identificadores. Estas palavras têm de ser escritas exatamente como indicado:

- **int bool float string void sizeof null alg true false**
- **while do finally leave restart return if then else write writeln**

Caracteres brancos

São considerados separadores e não representam nenhum elemento lexical:

- mudança de linha (caracter '\n')
- recuo do carreto – *carriage return* - (caracter '\r')
- espaço (caracter ' ')
- tabulação horizontal (caracter '\t').

Comentários

Existem dois tipos de comentários, que também funcionam como elementos separadores:

- explicativos -- começam com ## e acabam no fim da linha; e
- operacionais -- começam com (* e terminam com *), não podendo estar aninhados.

Se as sequências de início fizerem parte de uma cadeia de caracteres, não iniciam um comentário (ver definição das cadeias de caracteres).

Delimitadores e terminadores

Os seguintes elementos lexicais são delimitadores/terminadores:

- , (vírgula)
- ; (ponto e vírgula)

[1] A Linguagem FIR foi uma linguagem usada como projeto da cadeira de Compiladores no Instituto Superior Técnico no ano letivo 2020/2021.

- (e) (delimitadores de expressões).

Identificadores (nomes)

Identificadores correspondem a nomes de variáveis ou nomes de funções. São iniciados por uma letra, seguindo-se 0 (zero) ou mais letras, dígitos ou _ (sublinhado). O comprimento do nome é ilimitado.

Uma letra pode ser uma letra maiúscula ou minúscula do alfabeto latino (a-z e A-Z), mas também qualquer letra maiúscula ou minúscula do Latin-1 Supplement do Standard Unicode. Isto implica que a linguagem ALG suporta nomes de identificadores como por exemplo: pão_com_chouriço.

Literais

São notações para valores constantes de alguns tipos da linguagem (não confundir com constantes, i.e., identificadores que designam elementos cujo valor não pode ser alterado durante a execução do programa).

Inteiros

Um literal inteiro é um número não negativo. Uma constante inteira pode, contudo, ser negativa: números negativos são construídos pela aplicação do operador de negação aritmética unária (-) a um literal positivo.

Literais inteiros decimais são constituídos por sequências de 1 (um) ou mais dígitos de 0 a 9, em que o primeiro dígito não é 0 (zero), exceto no caso do número 0 (zero). Neste caso, é composto apenas pelo dígito 0 (zero).

Reais em vírgula flutuante

Não existem literais negativos (números negativos resultam da aplicação da operação de negação unária).

Um literal real é um literal numérico com parte inteira (ver literal inteiro), podendo ser seguido de uma parte decimal, de uma parte exponencial ou de ambas. A parte inteira é separada da parte decimal (caso a parte decimal exista) por um ponto decimal '.' A parte decimal é constituída por um literal inteiro. No fim do literal real poderá existir uma parte exponencial. Uma parte exponencial é iniciada pelo caracter 'E' ou 'e', podendo ser seguido de um sinal positivo '+' ou negativo '-', e termina com um literal inteiro. Um literal numérico sem . (ponto decimal) nem parte exponencial é considerado apenas um literal do tipo inteiro.

Exemplos: 3.14, 1E3 = 1000 (número inteiro representado em virgula flutuante). 12.34e-24 = 12.34 x 10-24 (notação cientifica).

Cadeias de caracteres

As cadeias de caracteres são delimitadas por plicas (') e podem conter quaisquer caracteres, exceto o caracter UNICODE NULL (\u0000). Nas cadeias, os delimitadores de comentários não têm significado especial.

É possível designar caracteres por sequências especiais (iniciadas por ~), especialmente úteis quando não existe representação gráfica directa. As sequências especiais correspondem aos caracteres LF, CR e HT (\n, \r e \t, respectivamente, em C e ~n, ~r, ~t, respectivamente, em ALG), plica (~'), til (~~).

Operadores e outras sequências relevantes

Os seguintes caracteres e sequências de caracteres serão usados como operadores e/ou parte de expressões da linguagem ALG. A semântica destes operadores e operações será definida mais tarde.

(	)	[	]
+	-	?	%
>	<	>=	<=
==	!=		
~	&&	||	

=			
>>			
@			

Dicas e sugestões

A ferramenta ANTLR permite a especificação de regras gramáticas (o parser) num ficheiro diferente da especificação das regras lexicais. Para isso, o ficheiro algLexer.g4 deverá começar com a instrução:

```
lexer grammar algLexer;
```

No entanto, para conseguirem testar o analisador léxical (lexer) com as ferramentas de depuração do ANTLR, é aconselhado a utilização de um analisador sintático (parser) conjuntamente com o analisador léxical. Para isto poderão criar um ficheiro alg.g4, com uma gramática muito simples, que corresponde a uma expressão seguida de *End Of File*, em que uma expressão pode ser qualquer token da linguagem. Por exemplo:

```
parser grammar alg;

options {tokenVocab=algLexer;}

start : expression EOF;

expression: WHILE …
```

Condições de realização

O projeto deve ser realizado em grupo, de acordo com as inscrições em grupo do laboratório. Projetos iguais, ou muito semelhantes, originarão a reprovação na disciplina. O corpo docente da disciplina será o único juiz do que se considera ou não copiar num projeto.

O 1.º projeto é muito simples, e vale 10% da nota final da componente prática. Esta primeira parte do não será avaliada automaticamente, e, portanto, não será necessária a submissão via *Mooshak*. O ficheiro algLexer.g4 com as regras lexicais deverá ser entregue obrigatoriamente por via eletrónica, através da tutoria, até às **23:59** do dia **26/03/2021**.

Os alunos terão de validar o código juntamente com o docente durante o horário de laboratório correspondente ao turno em que estão inscritos, na semana de 5 a 9 de Abril. Será feita uma breve discussão com cada grupo, e serão feitos pedidos para alterações ligeiras nas regras lexicais. Embora a realização do projeto seja em grupo, a nota do projeto é individual e dependerá da prestação de cada elemento na discussão do projeto.