

# Теория автоматов и формальных языков

## Синтаксический анализ

**Автор:** Екатерина Вербицкая

Санкт-Петербургский государственный электротехнический университет «ЛЭТИ»

4 декабря 2020

Задачи синтаксического анализа: по последовательности символов

- Определить, корректна ли она с точки зрения грамматики языка
  - ▶ Если нет, сообщить об ошибке
- Построить для нее некоторое структурное представление
  - ▶ Дерево вывода
  - ▶ Абстрактное синтаксическое дерево
  - ▶ Вычислить семантику (реже)

Терминалами грамматики являются символы строки

$$E \rightarrow E + T \mid T$$

$$T \rightarrow T * F \mid F$$

$$F \rightarrow (E) \mid N \mid Id$$

$$N \rightarrow 0 \mid 1 \mid \dots \mid 9 \mid 0N \mid 1N \mid \dots \mid 9N$$

$$Id \rightarrow a \mid b \mid \dots \mid z \mid aId \mid \dots \mid zId$$

Недостатки безлексерного синтаксического анализа:  
грамматики очень быстро перестают быть читаемыми

$$E \rightarrow Sp\ E\ Sp\ +\ Sp\ T\ Sp\ |\ Sp\ T\ Sp$$

$$T \rightarrow Sp\ T\ Sp\ *\ Sp\ F\ Sp\ |\ Sp\ F\ Sp$$

$$F \rightarrow Sp\ (Sp\ E\ Sp)\ Sp\ |\ Sp\ N\ Sp\ |\ Sp\ Id\ Sp$$

$$N \rightarrow 0\ |\ 1\ |\ \dots\ |\ 9\ |\ 1N\ |\ 2N\ |\ \dots\ |\ 9N$$

$$Id \rightarrow a\ |\ b\ |\ \dots\ |\ z\ |\ a\ Id\ |\ \dots\ |\ z\ Id$$

$$Sp \rightarrow \varepsilon\ |\ \sqcup\ Sp$$

## Недостатки безлексерного синтаксического анализа: работает медленно

Большую часть времени синтаксический анализатор тратит на разбор регулярных фрагментов языка, что чревато накладными расходами

$$\begin{aligned} N &\rightarrow 0 \mid 1 \mid \dots \mid 9 \mid 1N \mid 2N \mid \dots \mid 9N \\ Id &\rightarrow a \mid b \mid \dots \mid z \mid aId \mid \dots \mid zId \end{aligned}$$

Преобразование последовательности символов в  
последовательность *лексем* (или *токенов*)

Лексема — минимальная смысловая единица языка

"(123 + 45) \* 6" →

LBR, NUM 123, PLUS, NUM 45, RBR, MULT, NUM 6

- Ключевое слово
- Число
  - ▶ Целое
  - ▶ Целое в двоичной системе
  - ▶ С плавающей точкой
  - ▶ ...
- Идентификатор
- Оператор

# Лексический анализ и регулярные языки

Часто языки разных типов лексем являются регулярными языками

Лексеры часто реализуются при помощи регулярных выражений

NUM :: 0 | [1-9][0-9]\*

IDENT :: [a-z\_][a-z\_0-9]\*

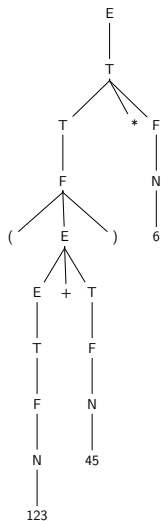
Лексеры также часто пропускают пробельные символы и выделяют комментарии



Терминалы грамматики — лексемы

$$E \rightarrow E \text{ PLUS } T \mid T$$
$$T \rightarrow T \text{ MULT } F \mid F$$
$$F \rightarrow \text{LBR } E \text{ RBR} \mid \text{NUM} \mid \text{IDENT}$$

# Конкретный синтаксис vs абстрактный синтаксис


$$\begin{aligned} E &\rightarrow E \text{ PLUS } T \mid T \\ T &\rightarrow T \text{ MULT } F \mid F \\ F &\rightarrow \text{LBR } E \text{ RBR} \mid N \end{aligned}$$
$$e :: e + e \mid e * e \mid \text{num}$$
