Теория автоматов и формальных языков

Введение

Лектор: Екатерина Вербицкая

Санкт-Петербургский государственный электротехнический университет «ЛЭТИ»

8 сентября 2016г.

О чем этот курс?

Теория автоматов и формальных языков изучает:

- Математические модели для описания языков
- Абстрактные машины для работы с языками

Также рассматриваются:

- Подходы к описанию синтаксиса языков
- Подходы к описанию “смысла” программ и предложений
- Принципиальные ограничения механизмов для работы с языками

Какие бывают языки?

Какие бывают языки?

- Естественные
  - Русский, английский...

Какие бывают языки?

- Естественные
  - Русский, английский...
- Искусственные
  - Эсперанто, ложбан...
  - Клингонский, эльфийский...

Какие бывают языки?

- Естественные
  - Русский, английский...
- Искусственные
  - Эсперанто, ложбан...
  - Клингонский, эльфийский...
  - C++, Python, Java, C#, Haskell, OCaml, Perl, Coq, Agda...

Где можно встретить языки?

В повседневной жизни:

- при разговоре, в переписке
- на заборах, на стенах гробниц
- в собственной голове при формулировке мыслей...

Где можно встретить языки?

В повседневной жизни:

- при разговоре, в переписке
- на заборах, на стенах гробниц
- в собственной голове при формулировке мыслей...

При работе с различными **языковыми процессорами**:

- текстовыми редакторами
- компиляторами, интерпретаторами, трансляторами
- средами разработки...

Где можно встретить языки?

В повседневной жизни:

- при разговоре, в переписке
- на заборах, на стенах гробниц
- в собственной голове при формулировке мыслей...

При работе с различными **языковыми процессорами**:

- текстовыми редакторами
- компиляторами, интерпретаторами, трансляторами
- средами разработки...

Все нуждаются в **формализованном представлении** языка

Два аспекта спецификации языка программирования

- Синтаксис — правила построения программ из символов
  - Форма
- Семантика — правила истолкования программ
  - Смысл

Пример: английский язык

You know nothing, Jon Snow

- Синтаксис
  - ...
  - Порядок слов в предложении: подлежащее, сказуемое, все остальное
  - Обращение обособляется запятыми
  - ...
- Семантика
  - Говорящий обращается к Джону Сноу, утверждая, что Джон ничего не знает.

Пример: язык арифметических выражений

$1 * (2 + 3)/4 - 5$

- Синтаксис
  - **Терм**: последовательность цифр или любое **выражение** в скобках
  - **Слагаемое**: последовательность **термов**, соединненых знаками умножения и деления
  - **Выражение**: последовательность **слагаемых**, соединенных знаками сложения и вычитания (перед первым **слагаемым** может стоять минус)
- Семантика
  - Значение арифметического выражения

Пример: язык арифметических выражений

$1 * (2 + 3)/4 - 5$

- Синтаксис
  - **Терм**: последовательность цифр или любое **выражение** в скобках
  - **Слагаемое**: последовательность **термов**, соединненых знаками умножения и деления
  - **Выражение**: последовательность **слагаемых**, соединенных знаками сложения и вычитания (перед первым **слагаемым** может стоять минус)
- Семантика
  - Значение арифметического выражения
    - -3.75
    - -4

Что такое язык?

Что такое язык?

Язык — множество строк

Что такое множество?

Что такое множество?

Множество — набор уникальных элементов

Что такое множество?

Множество — набор уникальных элементов

- $x \in X$: x — элемент множества X (x принадлежит X)
- $x \notin X$: x не является элементом множества X (x не принадлежит X)
- Уникальность, неупорядоченность: $\{13, 42\} = \{42, 13\} = \{42, 13, 42\}$
- Универсальное множество (универсум $\mathcal{U}$): множество всех мыслимых объектов
  - $\mathbb{N} = \{1, 2, 3, \dots\}$
  - $\mathbb{Z} = \{\dots, -2, -1, 0, 1, 2, \dots\}$
  - $\mathbb{Q} = \{m/n \mid m, n \in \mathbb{Z}; n \neq 0\}$

Подмножества

A является **подмножеством** B тогда и только тогда, когда все элементы A являются элементами B

$$A \subseteq B \iff \forall x : x \in A \Rightarrow x \in B$$

Подмножества

A является **подмножеством** B тогда и только тогда, когда все элементы A являются элементами B

$$A \subseteq B \iff \forall x : x \in A \Rightarrow x \in B$$

- $\{13, 42\} \subseteq \{7, 13, 37, 42, 99\}$
- $\{1, 3, 5, \ldots\} \subseteq \mathbb{N}$
- $\mathbb{N} \subseteq \mathbb{Z}$
- $\forall A : A \subseteq A$

- Пустое множество ($\varnothing$): множество без элементов
  - $\forall x : x \notin \varnothing$
  - $\forall A : \varnothing \subseteq A$

Подмножества

Множества A и B **равны** тогда и только тогда, когда A является подмножеством B и B является подмножеством A

$$A = B \iff A \subseteq B \text{ и } B \subseteq A$$

Подмножества

Множества A и B **равны** тогда и только тогда, когда A является подмножеством B и B является подмножеством A

$$A = B \iff A \subseteq B \text{ и } B \subseteq A$$

A является **строгим подмножеством** B тогда и только тогда, когда A является подмножеством B, но они не равны друг другу

$$A \subset B \iff A \subseteq B \text{ и } A \neq B$$

Подмножества

Множества A и B **равны** тогда и только тогда, когда A является подмножеством B и B является подмножеством A

$$A = B \iff A \subseteq B \text{ и } B \subseteq A$$

A является **строгим подмножеством** B тогда и только тогда, когда A является подмножеством B, но они не равны друг другу

$$A \subset B \iff A \subseteq B \text{ и } A \neq B$$

- $\forall x : \varnothing \subset \{x\}$
- $\mathbb{N} \subset \mathbb{Z}$
- $\mathbb{Z} \not\subset \mathbb{N}$
- $\forall A : A = A$ и $A \not\subset A$

Множество всех подмножеств (powerset)

Множество всех подмножеств множества A состоит из всех подмножеств A

$$2^A = \{B \mid B \subseteq A\}$$

- $\forall A : \varnothing \in 2^A$
- $\forall A : A \in 2^A$
- $A = \{0, 1\} \Rightarrow \{\varnothing, \{0\}, \{1\}, \{0, 1\}\}$

Множество всех подмножеств (powerset)

Множество всех подмножеств множества A состоит из всех подмножеств A

$$2^A = \{B \mid B \subseteq A\}$$

- $\forall A : \varnothing \in 2^A$
- $\forall A : A \in 2^A$
- $A = \{0, 1\} \Rightarrow \{\varnothing, \{0\}, \{1\}, \{0, 1\}\}$

Сколько элементов может быть в множестве всех подмножеств?

Операции над множествами

- Объединение: $A \cup B = \{x \mid x \in A \text{ или } x \in B\}$
- Пересечение: $A \cap B = \{x \mid x \in A \text{ и } x \in B\}$
- Разность: $A \setminus B = \{x \mid x \in A \text{ и } x \notin B\}$
- Дополнение: $\overline{A} = \{x \mid x \in \mathcal{U} \text{ и } x \notin A\} = \mathcal{U} \setminus A$

Что такое строки?

Что такое строки?

Строка — последовательность символов

Что такое строки?

Строка — последовательность символов

Не очень формально

Алфавит

- **Алфавит** (Σ) — конечное множество (атомарных, неделимых) *символов*
  - $\{a, b, c, \ldots, z\}$
  - $\{\alpha, \beta, \gamma, \ldots, \omega\}$
  - $\{0, 1\}$
  - $\{$ <u>let</u>, <u>in</u>, <u>where</u>, $\ldots\}$

Цепочка

- **Цепочка (предложение, слово, *строка*)** — любая конечная последовательность символов алфавита
  - cat
  - $\kappa\alpha\tau$
  - 011000110110000101110100
  - `main = putStrLn . show . inc 2 where inc = \x -> x + 1`
- **Пустая цепочка** ε — цепочка, не содержащая ни одного символа
  - ε не является символом алфавита

Конкатенация строк

- **Конкатенация строк** α **и** β $(\alpha \cdot \beta = \alpha\beta)$ — результат приписывания строки β в конец строки α
  - $\forall \alpha\beta\gamma : (\alpha \cdot \beta) \cdot \gamma = \alpha \cdot (\beta \cdot \gamma)$
  - $\forall \alpha : \alpha \cdot \varepsilon = \varepsilon \cdot \alpha = \alpha$

Пример: арифметические выражения

- Алфавит $\Sigma = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, +, -, *, /, (,)\}$
- $1 * (2 + 3)/4 - 5 =$
  $1 \cdot *(2 + 3)/4 - 5 =$
  $1 * (2 + 3) \cdot /4 - 5 =$
  $1 \cdot * \cdot (\cdot 2 \cdot + \cdot 3 \cdot) \cdot / \cdot 4 \cdot - \cdot 5 =$
  $1 * (2 + 3)/4 - 5 \cdot \varepsilon$
- Является ли ε арифметическим выражением?

Операции над строками

- **Обращение (реверс) цепочки** a^R — цепочка, символы которой записаны в обратном порядке
  - Если $x = abc$, то $x^R = cba$
  - $\varepsilon^R = \varepsilon$
- $n-$**я степень цепочки** a^n — конкатенация n повторений цепочки
  - $a^0 = \varepsilon$
  - $a^n = a \cdot a^{n-1} = a^{n-1} \cdot a$
- **Длина цепочки** $|a|$ — количество составляющих ее символов
  - $|babb| = 4$
  - $|babb|_a = 1, |babb|_b = 3, |babb|_c = 0$
  - $|\varepsilon| = 0$

Формальный язык

- Σ — алфавит
  - $\Sigma = \{0, 1\}$
- Σ^* — множество, содержащее все цепочки в алфавите Σ, включая пустую цепочку
  - $\Sigma^* = \{\varepsilon, 0, 1, 00, 11, 01, 10, 000, 001, 011, ...\}$
  - Сколько может быть элементов в Σ^*?

Формальный язык

- Σ — алфавит
  - $\Sigma = \{0, 1\}$
- Σ^* — множество, содержащее все цепочки в алфавите Σ, включая пустую цепочку
  - $\Sigma^* = \{\varepsilon, 0, 1, 00, 11, 01, 10, 000, 001, 011, ...\}$
  - Сколько может быть элементов в Σ^*?
- $\Sigma^+ = \Sigma^* \setminus \{\varepsilon\}$
  - $\Sigma^+ = \{0, 1, 00, 11, 01, 10, 000, 001, 011, \dots\}$
  - Сколько может быть элементов в Σ^+?

Формальный язык

- Σ — алфавит
  - $\Sigma = \{0, 1\}$
- Σ^* — множество, содержащее все цепочки в алфавите Σ, включая пустую цепочку
  - $\Sigma^* = \{\varepsilon, 0, 1, 00, 11, 01, 10, 000, 001, 011, ...\}$
  - Сколько может быть элементов в Σ^*?
- $\Sigma^+ = \Sigma^* \setminus \{\varepsilon\}$
  - $\Sigma^+ = \{0, 1, 00, 11, 01, 10, 000, 001, 011, \dots\}$
  - Сколько может быть элементов в Σ^+?
- Σ — подмножество множества всех цепочек в этом алфавите.
  - Для любого языка L (в алфавите Σ) справедливо $L \subseteq \Sigma^*$
  - $L = \{0, 00, 000, \dots\} \subset \{0, 1\}^*$
  - $L = \{0, 0101, 011011011, \dots\} \subset \{0, 1\}^*$

Метаязык

- Язык, на котором дано описание языка
  - Естественный язык

Метаязык

- Язык, на котором дано описание языка
  - Естественный язык
  - Язык металингвистичесих формул Бэкуса (БНФ)

Метаязык

- Язык, на котором дано описание языка
  - Естественный язык
  - Язык металингвистичесих формул Бэкуса (БНФ)
  - Синтаксические диаграммы

Метаязык

- Язык, на котором дано описание языка
  - Естественный язык
  - Язык металингвистичесих формул Бэкуса (БНФ)
  - Синтаксические диаграммы
  - Грамматики...

БНФ — Бэкуса-Наура форма

- **Символ** — элементарное понятие языка
  - $+$ означает сложение в языке арифметических выражений
- **Метапеременная** — сложное понятие языка
  - Переменной $<$выражение$>$ можно обозначить выражение
- **Формула**
  - $<$определяемый символ$>$::=$<$посл.1 $>$ | … | $<$посл.n $>$
  - В правой части формулы — альтернатива конкатенаций строк, составленных из символов и метапеременных
- Пример: число
  - $<$число$>$::=$<$цифра$>$ | $<$цифра$><$число$>$
  - $<$цифра$>$::= 0 | 1 | … | 9

Расширенная форма Бэкуса Наура (EBNF)

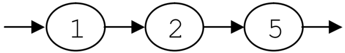
- Более емкие операции
- **Итерация**
  - $<x> ::= \{<y>\}$ эквивалентно: $<x> ::= \varepsilon \mid <y><x>$
- **Условное вхождение**
  - $<x> ::= [<y>]$ эквивалентно: $<x> ::= \varepsilon \mid <y>$
- Скобки для группировки
  - $(<x> \mid <y>)\ <z>$ эквивалентно: $<x><z> \mid <y><z>$

Пример: арифметические выражения

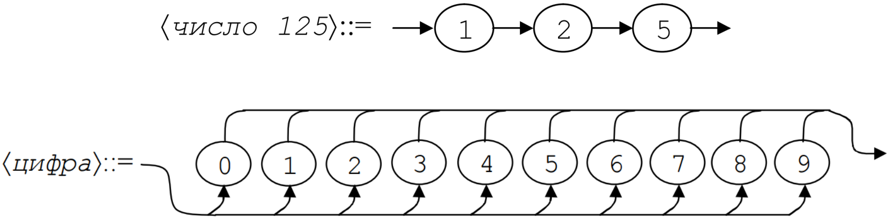
- **Терм**: последовательность цифр или любое **выражение** в скобках
- **Слагаемое**: последовательность **термов**, соединненых знаками умножения и деления
- **Выражение**: последовательность **слагаемых**, соединенных знаками сложения и вычитания (перед первым **слагаемым** может стоять минус)

$$
\begin{aligned}
< \mathit{expr} > \quad &::= \quad [-] < \mathit{factor} > \{\, (+ \,|\, -) \; < \mathit{factor} > \} \\
< \mathit{factor} > \quad &::= \quad < \mathit{term} > \{\, (* \,|\, /) \; < \mathit{term} > \} \\
< \mathit{term} > \quad &::= \quad < \mathit{number} > \; |'(' < \mathit{expr} > ')'
\end{aligned}
$$

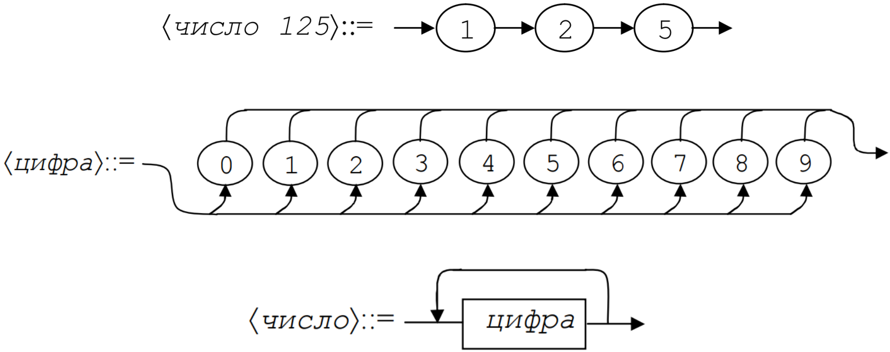
Синтаксические диаграммы Вирта

⟨*число 125*⟩::= 

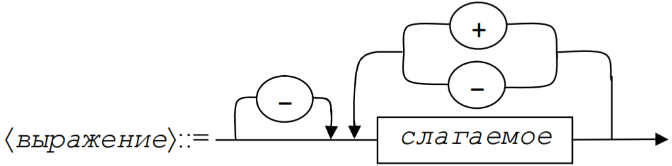
Синтаксические диаграммы Вирта



Синтаксические диаграммы Вирта

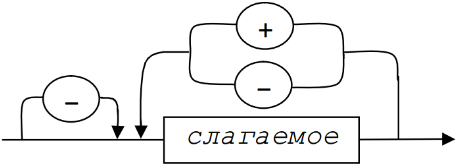


Синтаксические диаграммы Вирта

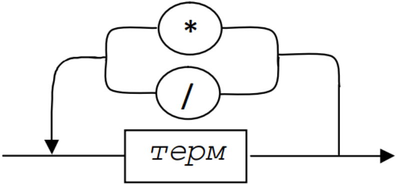


Синтаксические диаграммы Вирта

⟨выражение⟩::=

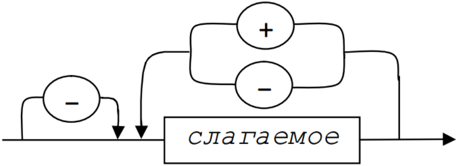


⟨слагаемое⟩::=

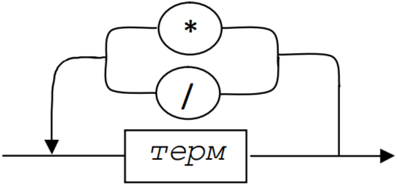


Синтаксические диаграммы Вирта

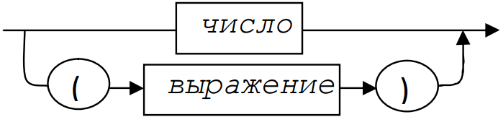
⟨выражение⟩::=



⟨слагаемое⟩::=



⟨терм⟩::=



Описание языка: формальная грамматика

- **Порождающая грамматика** G — это четверка $\langle V_T, V_N, P, S \rangle$
  - V_T — алфавит терминальных символов (терминалов)
  - V_N — алфавит нетерминальных символов (нетерминалов)
    - $V_T \cap V_N = \emptyset$
    - $V ::= V_T \cup V_N$
  - P — конечное множество правил вида $\alpha \rightarrow \beta$
    - $\alpha \in V^* V_N V^*$
    - $\beta \in V^*$
  - S — начальный нетерминал грамматики, $S \in N$

Пример: язык чисел в двоичной системе счисления

$V_T = \{0, 1\}; V_N = \{S, N, A\}$

$$\begin{array}{rcl}
S & \to & 0 \\
S & \to & N \\
S & \to & -N \\
N & \to & 1A \\
A & \to & 0A \\
A & \to & 1A \\
A & \to & \varepsilon
\end{array}$$

Пример: язык чисел в двоичной системе счисления

$V_T = \{0, 1\}; V_N = \{S, N, A\}$

$$\begin{array}{lcl} S & \rightarrow & 0 \\ S & \rightarrow & N \\ S & \rightarrow & -N \\ N & \rightarrow & 1A \\ A & \rightarrow & 0A \\ A & \rightarrow & 1A \\ A & \rightarrow & \varepsilon \end{array} \qquad \begin{array}{lcl} S & \rightarrow & 0 \,|\, N \,|\, -N \\ N & \rightarrow & 1A \\ A & \rightarrow & 0A \,|\, 1A \,|\, \varepsilon \end{array}$$

Пример: язык чисел в двоичной системе счисления

$V_T = \{0, 1\}; V_N = \{S, N, A\}$

$$\begin{array}{rcl} S & \to & 0 \\ S & \to & N \\ S & \to & -N \\ N & \to & 1A \\ A & \to & 0A \\ A & \to & 1A \\ A & \to & \varepsilon \end{array} \qquad \begin{array}{rcl} S & \to & 0 \,|\, N \,|\, -N \\ N & \to & 1A \\ A & \to & 0A \,|\, 1A \,|\, \varepsilon \end{array} \qquad \begin{array}{rcl} S & \to & 0 \,|\, [-]N \\ N & \to & 1A \\ A & \to & (0 \,|\, 1)A \,|\, \varepsilon \end{array}$$

Отношение непосредственной выводимости

- $\alpha \to \beta \in P$
- $\gamma, \delta \in V^*$
- $\gamma\alpha\delta \Rightarrow \gamma\beta\delta$: $\gamma\beta\delta$ **непосредственно выводится** из $\gamma\alpha\delta$ при помощи правила $\alpha \to \beta$

Отношение выводимости

- $a_0, a_1, a_2, \ldots, a_n \in V^*$
- $a_0 \Rightarrow a_1 \Rightarrow a_2 \Rightarrow \cdots \Rightarrow a_n$
- $a_0 \overset{*}{\Rightarrow} a_n$: a_n **выводится** из a_0
- $S \Rightarrow -N \Rightarrow -1A \Rightarrow -11A \overset{*}{\Rightarrow} -1101A \Rightarrow -1101$

Отношение выводимости

- $a_0, a_1, a_2, \ldots, a_n \in V^*$
- $a_0 \Rightarrow a_1 \Rightarrow a_2 \Rightarrow \cdots \Rightarrow a_n$
- $a_0 \overset{*}{\Rightarrow} a_n$: a_n **выводится** из a_0
- $S \Rightarrow -N \Rightarrow -1A \Rightarrow -11A \overset{*}{\Rightarrow} -1101A \Rightarrow -1101$
- $\forall a \in V^*. a \overset{*}{\Rightarrow} a$
- $a_0 \overset{+}{\Rightarrow} a_n$: вывод использует хотя бы одно правило грамматики
- $a_0 \overset{k}{\Rightarrow} a_n$: вывод происходит за k шагов

Язык, порождаемый грамматикой $G = \langle V_T, V_N, P, S \rangle$

- $L(G) = \{\omega \in V_T^* \mid S \overset{*}{\Rightarrow} \omega\}$

Эквивалентность грамматик

- Грамматики G_1 и G_2 **эквивалентны**, если $L(G_1) = L(G_2)$

Эквивалентность грамматик

- Грамматики G_1 и G_2 **эквивалентны**, если $L(G_1) = L(G_2)$

$$
\begin{aligned}
V_T &= \{0, 1\} \\
V_N &= \{S, N, A\}
\end{aligned}
$$

$$
\begin{aligned}
S &\to 0 \,|\, N \,|\, -N \\
N &\to 1A \\
A &\to 0A \,|\, 1A \,|\, \varepsilon
\end{aligned}
$$

Эквивалентность грамматик

- Грамматики G_1 и G_2 **эквивалентны**, если $L(G_1) = L(G_2)$

$$\begin{aligned} V_T &= \{0, 1\} \\ V_N &= \{S, N, A\} \\ \\ S &\to 0 \,|\, N \,|\, -N \\ N &\to 1A \\ A &\to 0A \,|\, 1A \,|\, \varepsilon \end{aligned}$$

$$\begin{aligned} V_T &= \{0, 1\} \\ V_N &= \{S, A\} \\ \\ S &\to 0 \,|\, 1A \,|\, -1A \\ A &\to 0A \,|\, 1A \,|\, \varepsilon \end{aligned}$$

- **Контекстно-свободная грамматика** — грамматика, все правила которой имеют вид $A \to \alpha, A \in V_N, \alpha \in V^*$

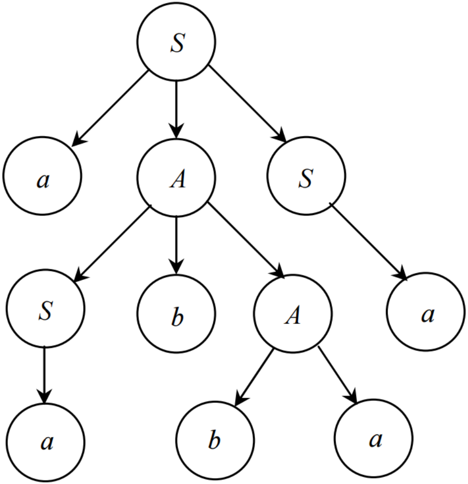
Дерево вывода

Дерево является **деревом вывода** для $G = \langle V_N, V_T, P, S \rangle$, если:

- Каждый узел помечен символом из алфавита V
- Метка корня — S
- Листья помечены терминалами, остальные узлы — нетерминалами
- Если узлы $n_0, \ldots, n_k$ — прямые потомки узла n, перечисленные слева направо, с метками $A_0, \ldots, A_k$; метка n — A, то $A \to A_0 \ldots A_k \in P$

Пример дерева вывода

$$G = \langle \{S, A\}, \{a, b\}, \{S \to aAS | a, A \to SbA | ba | SS\}, S \rangle$$
$$S \Rightarrow aAS \Rightarrow aSbAS \Rightarrow aabAS \Rightarrow aabbaS \Rightarrow aabbaa$$



Вывод и дерево вывода

Теорема

Пусть $G = \langle V_N, V_T, P, S \rangle$ *— КС-грамматика*
Вывод $S \overset{*}{\Rightarrow} \alpha$, *где* $\alpha \in V^*, \alpha \neq \varepsilon$ *существует* $\Leftrightarrow$ *существует дерево вывода в грамматике* G *с результатом* α