Введение в синтаксический анализ

Слова, языки и грамматики

Автор: Григорьев Семён

Санкт-Петербургский государственный университет
Математико-Механический факультет
Кафедра системного программирования

26 октября 2011г.

Слова, языки и грамматики

А. Е. Пентус, М. Р. Пентус "Теория формальных языков"

Формальные языки

Definition

Алфавитом называется конечное непустое множество. Его элементы называются ***символами*** (***буквами***).

Формальные языки

Definition

Алфавитом называется конечное непустое множество. Его элементы называются ***символами*** (***буквами***).

Definition

Словом (***цепочкой, строкой***) в алфавите Σ называется конечная последовательность элементов Σ .

Формальные языки

Пример

- Алфавит: $\Sigma = \{a, b, c\}$
- a – слово в алфавте Σ
- $abbc$ – слово в алфавте Σ

Формальные языки

Definition

Слово, не содержащее ни одного символа(то есть последовательность длины 0), называется ***пустым словом*** и обозначается ε.

Формальные языки

Definition

Слово, не содержащее ни одного символа(то есть последовательность длины 0), называется ***пустым словом*** и обозначается ε.

Definition

Длина слова ω, обозначаемая $|\omega|$, есть число символов в ω, причём каждый символ считается столько раз, сколько раз он встречается в ω.

Формальные языки

Пример

- $|abbc| = 4$
- $|\varepsilon| = 0$

Формальные языки

Definition

Если x и y — слова в алфавите Σ, то слово xy (результат приписывания слова y в конец слова x) называется ***конкатенацией*** слов x и y. Иногда конкатенацию слов x и y обозначают $x \cdot y$.

Definition

Если x и y — слова в алфавите Σ, то слово xy (результат приписывания слова y в конец слова x) называется ***конкатенацией*** слов x и y. Иногда конкатенацию слов x и y обозначают $x \cdot y$.

Definition

Если x — слово и $n \in \mathbb{N}$, то через x^n обозначается слово $\underbrace{x \cdot x \cdot \ldots \cdot x}_{n \text{ раз}}$.

По определению $x^0 \rightleftharpoons \varepsilon$ (знак $\rightleftharpoons$ читается "равно по определению"). Всюду далее показатели над словами и символами, как правило, являются натуральными числами.

Пример

- По принятым соглашениям:
  - $ba^3 = baaa$
  - $(ba)^3 = bababa$

Definition

Говорят, что слово x – ***префикс*** слова y (обозначение $x \sqsubset y$), если $y = xu$ для некоторого слова u.

Пример: $\varepsilon \sqsubset baa$, $b \sqsubset baa$, $ba \sqsubset baa$, $baa \sqsubset baa$

Definition

Говорят, что слово x – ***префикс*** слова y (обозначение $x \sqsubset y$), если $y = xu$ для некоторого слова u.

Пример: $\varepsilon \sqsubset baa, b \sqsubset baa, ba \sqsubset baa, baa \sqsubset baa$

Definition

Говорят, что слово x – ***суффикс*** слова y (обозначение $x \sqsupset y$), если $y = ux$ для некоторого слова u.

Definition

Говорят, что слово x – ***подслово*** (substring) слова y, если $y = uxv$ для некоторых слов u и v.

Definition

Через $|w|_a$ обозначается количество вхождений символа a в слово w.

Формальные языки

Definition

Если $L \subseteq \Sigma^*$, то L называется ***языком*** (или ***формальным языком***) над алфавитом Σ.

Порождающие грамматики

Порождающие грамматики

Definition

Порождающей грамматикой (***грамматикой типа 0***)называется четвёрка $G \rightleftharpoons \{N, \Sigma, P, S\}$, где N и Σ – конечные алфавиты, $N \cap \Sigma = \varnothing$, $P \subset (N \cup \Sigma)^+ \times (N \cup \Sigma)^*$, P конечно и $S \in N$.

Здесь:

- Σ – ***основной алфавит*** (***терминальный алфавит***), его элементы называются ***терминальными символами*** или ***терминалами***
- N — ***вспомогательный алфавит*** (***нетерминальный алфавит***), его элементы называются ***нетерминальными символами***, ***нетерминалами***
- S — ***начальный символ***
- Пары $(\alpha, \beta) \in P$ называются ***правилами подстановки***, просто ***правилами*** или ***продукциями*** и записываются в виде $\alpha \to \beta$

Порождающие грамматики. Пример

Пусть даны множества:

- $N = \{S\}$
- $\Sigma = \{a, b, c\}$
- $P = \{S \to acSbcS, cS \to \varepsilon\}$

Тогда (N, Σ, P, S) является порождающей грамматикой.

Порождающие грамматики

Будем обозначать:

- элементы множества Σ – строчным и буквам и из начала латинского алфавита
- элементы множества N – заглавными латинскими буквами

Обычно грамматику задают в виде списка правил, подразумевая, что алфавит N составляют все заглавные буквы, встречающиеся в правилах, а алфавит Σ – все строчные буквы, встречающиеся в правилах. При этом правила порождающей грамматики записывают в таком порядке, что левая часть первого правила есть начальный символ S.

Порождающие грамматики

Для обозначения n правил с одинаковыми левыми частями $\alpha \to \beta_1, ..., \alpha \to \beta_n$ часто используют сокращённую запись $\alpha \to \beta_1 | ... | \beta_n$.

Порождающие грамматики

Definition

Пусть дана грамматика G. Пишем $\varphi \underset{G}{\Rightarrow} \psi$, если $\varphi = \eta\alpha\theta$, $\psi = \eta\beta\theta$ и $(\alpha \to \beta) \in P$ для некоторых слов α, β, η, θ в алфавите $N \cup \Sigma$.

Remark

Когда из контекста ясно, о какой грамматике идёт речь, вместо $\underset{G}{\Rightarrow}$ можно писать просто $\Rightarrow$.

Порождающие грамматики. Пример

Пусть $G = \langle S, a, b, c, S \to acSbcS, cS \to \varepsilon, S \rangle$. Тогда $cSacS \underset{G}{\Rightarrow} cSa$.

Порождающие грамматики

Definition

Если $\omega_0 \underset{G}{\Rightarrow} \omega_1 \underset{G}{\Rightarrow} \ldots \underset{G}{\Rightarrow} \omega_n$, где $n \geq 0$, то пишем $\omega_0 \overset{*}{\underset{G}{\Rightarrow}} \omega_n$ (другими словами, бинарное отношение $\overset{*}{\underset{G}{\Rightarrow}}$ является рефлексивным, транзитивным замыканием бинарного отношения $\underset{G}{\Rightarrow}$, определённого на множестве $(N \cup \Sigma)^*$). При этом последовательность слов $\omega_0, \omega_1, \ldots, \omega_n$ называется ***выводом*** (derivation) слова ω_n из слова ω_0 в грамматике G. Число n называется ***длиной*** (***количеством шагов***) этого вывода.

Remark

В частности, для всякого слова $\omega \in (N \cup \Sigma)^*$ имеет место $\omega \overset{*}{\underset{G}{\Rightarrow}} \omega$ (так как возможен вывод длины 0)

Порождающие грамматики. Пример

Пусть $G = \langle \{S\}, \{a, b\}, \{S \to aSa, S \to b\}, S \rangle$. Тогда $aSa \overset{*}{\underset{G}{\Rightarrow}} aaaaSaaaa$.

Длина этого вывода – 3.

Порождающие грамматики

Definition

Язык, ***порождаемый грамматикой*** G, – это множество $L(G) \rightleftharpoons \{\omega \in \Sigma^* | S \overset{*}{\underset{G}{\Rightarrow}} \omega\}$. Будем также говорить, что грамматика G ***порождаёт*** язык $L(G)$.

Remark

Существенно, что в определение порождающей грамматики включены два алфавита – Σ и N. Это позволяет нам "отсеять"часть слов, получаемых из начального символа. А именно, отбрасывается каждое слово, содержащее хотя бы один символ, не принадлежащий алфавиту Σ.

Порождающие грамматики. Пример

Если $G = \langle \{S\}, \{a, b\}, \{S \to aSa, S \to bb\}, S \rangle$, то $L(G) = \{a^n bba^n | n \geq 0\}$.

Порождающие грамматики

Definition

Две грамматики эквивалентны, если они порождают один и тот же язык.

Пример
Грамматика $S \to abS, S \to a$ и грамматика $T \to aU, U \to baU, U \to \varepsilon$ эквивалентны.

Классы грамматик

Классы грамматик

Definition

Контекстной грамматикой *(контекстно-зависимой грамматикой, грамматикой непосредственно составляющих, НС-грамматикой, грамматикой типа 1) (context-sensitive grammar, phrase-structure grammar)* называется порождающая грамматика, каждое правило которой имеет вид $\eta A\theta \to \eta\alpha\theta$, где $A \in N$, $\eta \in (N \cup \Sigma)^*$, $\theta \in (N \cup \Sigma)^*$, $\alpha \in (N \cup \Sigma)^+$.

Definition

Контекстно-свободной грамматикой *(КС-грамматикой, бесконтекстной грамматикой, грамматикой типа 2) (context-free grammar)* называется порождающая грамматика, каждое правило которой имеет вид $A \to \alpha$, где $A \in N$, $\alpha \in (N \cup \Sigma)^*$.