Задание 6

Грамматики

Ключевые слова [1]:*язык, регулярный язык, ДКА, НКА, алгебра регулярных выражений, грамматики, уравнения с регулярными коэффициентами.*

1 Грамматики

Одна из больших проблем науки, которую мы с вами изучаем – определения. Их слишком много и они отличаются друг от друга, хотя в итоге конечно описывают одни и те же классы языков. Я призываю на экзамене пользоваться определениями из книги Серебрякова, хотя при выполнении задания вы можете пользоваться эквивалентными определениями из другой литературы.

Определение 1. Грамматика Γ определяется через

- N – множество нетерминальных символов
- T – множество терминальных символов
- P – множество правил вывода, $P \subseteq (N \cup T)^* \times (N \cup T)^*$.
- S – аксиома, $S \in N$.

Все множества из описания грамматики конечные. При этом, $N \cap T = \varnothing$. Принято обозначение $\Gamma = G(N, T, P, S)$. При описании грамматики приняты следующие соглашения. Нетерминалы обозначают заглавными буквами $A, B, C, \ldots$ терминалы обозначают строчными буквами, смешанные цепочки из $(N \cup T)^*$ обозначают греческими буквами α, β, γ. Слово $w \in T^*$ порождается грамматикой Γ, если существует последовательность правил вывода, начинающаяся с правила вида $S \to \alpha$, в результате применения которых порождается слово w. Под применением правила $\alpha \to \beta$, понимается, что подслово α заменяется на подслово β

[1]минимальный необходимый объем понятий и навыков по этому разделу)

В зависимости от ограничений, налагаемых на правила вывода, получаются разные классы языков. В рамках этого задания нас пока интересует только два последних типа.

- Если на множество правил P не накладывается ограничений, то есть правила имеют вид $\alpha \to \beta$, то грамматика называется грамматикой типа 0 по Хомскому

- Грамматики, в которых правила имеют вид $\alpha A \beta \to \alpha \gamma \beta$, $|\gamma| > 0$ называются грамматиками типа 1 или Контекстно-зависимыми. В качестве исключения грамматике может принадлежать правило $S \to \varepsilon$, но тогда нетерминал S не может встречаться в правых частях.

- Грамматики, в которых правила имеют вид $A \to \alpha$, называются грамматиками типа 2 или Контекстно-Свободными грамматиками.

- Грамматики, в которых правила имеют вид $A \to xB$ или $A \to x$, $x \in T^*$, называются грамматиками типа 3 или праволинейными грамматиками.

В определении КЗ-грамматики существенно, что она является *неукорачивающей*, т.е. правая часть правил всегда длинее левой. Эквивалентное определение из книги Серебрякова гласит, что в КЗ-грамматике все правила, кроме быть может $S \to \varepsilon$, имеют вид $\alpha \to \beta$, $|\alpha| < |\beta|$. Опять-таки, если есть правило $S \to \varepsilon$, то нетерминал S в правых частях правил встречаться не может.

Очень часто грамматиками типа 3 называют грамматики, в которых правила вывода имеют вид $A \to xB$ или $A \to x$, $x \in T$, также допускается правило $S \to \varepsilon$ с всё той же оговоркой, что аксиома не может встречаться в правой части. Такие грамматики называются *праволинейными регулярными* грамматиками.

Упражнение 1. Доказать, что праволинейные грамматики и праволинейные регулярные грамматики эквивалентны, т.е. порождают один и тот же тип языков.

Определение 2. Грамматика типа 3 является *неоднозначной*, если существует более одного вывода хотя бы для одного слово из языка, порождённого грамматикой.

Для КС-грамматик это определение неприемлемо. Вдумчивый читатель может подумать почему, прежде чем переходить к следующему разделу.

Леволинейные грамматики определяются аналогично праволинейным: в них правила имеют вид $A \to Bx$ или $A \to x$.

Приведём пример КС-грамматики. При записи правил используют вспомогательное обозначение $A \to \alpha|\beta$, которое означает, что в грамматике есть два правила: $A \to \alpha$ и $A \to \beta$.

Пример 1. Грамматика G задана правилами:

$$
\begin{aligned}
S &\to aAB \\
A &\to aA|a \\
B &\to bB|b
\end{aligned}
$$

Слово $aabb$ выводится грамматикой. Последовательность применений правил вывода такая:

$$
\begin{aligned}
S &\to aAB \\
&a\underline{A}B \\
A &\to a \\
&aa\underline{B} \\
B &\to bB \\
&aab\underline{B} \\
B &\to b \\
&aabb
\end{aligned}
$$

1.1 Вывод, левый вывод, дерево разбора

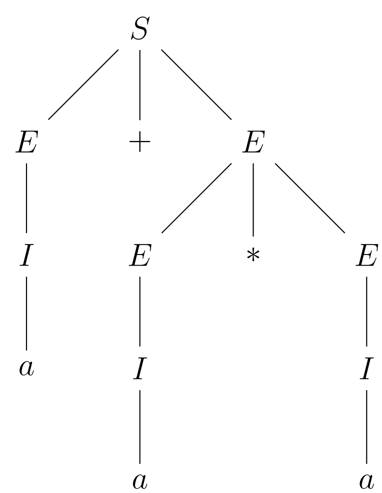
Выводом цепочки α называется такая последовательность применений правил с указанием раскрываемого нетерминала, что применяя правила из неё начиная с аксиомы получается цепочка α. Если цепочка α не содержит нетерминалов, то α принадлежит языку, порождаемому КС-грамматикой. Нам будет удобно пользоваться такими понятиями как левый вывод и правый вывод. *Левым выводом* называют такой вывод, что на каждом его шаге раскрывается самый левый нетерминал в промежуточной цепочке. Вывод в примере 1 является левым. Правый вывод определяется аналогично.

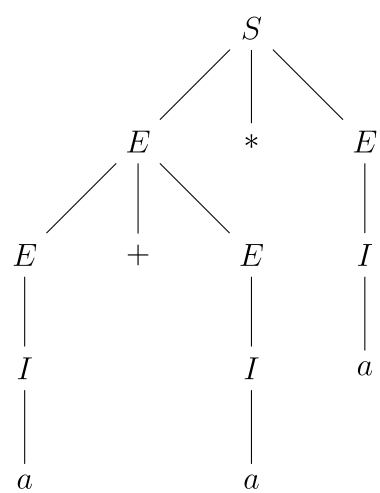
Также мы будем использовать деревья вывода. С формальным определением дерева вывода вы можете познакомиться, например, в книге Хопкрофта, Мотвани и Ульмана, а мы рассмотрим пример деревьев вывода и затем дадим неформальное описание этого понятия.

Пример 2. Грамматика G, $T = \{a, +, *\}$ задана правилами:

$$\begin{aligned} S &\to E \\ E &\to E + E \\ E &\to E * E \\ E &\to I \\ I &\to a \end{aligned}$$

Построим деревья вывода для слова $a + a * a$:





Теперь опишем понятие дерева вывода, которое также называется деревом разбора. Зафиксируем грамматику G. *Деревом вывода* для слова w называется упорядоченное дерево, в корне которого находится аксиома S, каждая вершина помечена нетерминалом, терминалом или пустым словом, если вершина помечена терминалом или ε, то эта вершина является листом, если же вершина помечена нетерминалом A, то для некоторого правила $A \to X_1X_2 \dots X_n \in P$ ($X_i \in N \cup T$) вершины-дети A помечены символами $X_1, X_2 \dots X_n$ слева направо. Листья дерева вывода образуют слово w.

Как мы видим, синтаксически деревья вывода принципиально разные: в первом случае выражение интерпретируется как $a + (a * a)$, а во втором как $(a + a) * a$, что приводит к непредсказуемому результату при выполнении стандартных операций!

Эта проблема приводит нас к новому важному понятию – неоднозначности. Грамматика G является *неоднозначной*, если хотя бы для одного слова существует два различных дерева вывода.

Левый и правый вывод помимо удобства важны тем, что они фактически задают порядок обхода дерева вывода, поэтому каждому левому выводу соответствует ровно одно дерево разбора, а каждому дереву забора соответствует ровно один левый вывод.

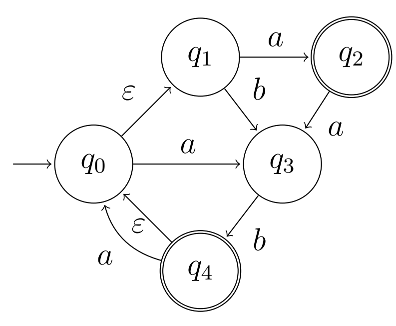
Упражнение 2. Доказать, что грамматика G является однозначной тогда и только тогда, когда каждое слово, порождаемое G имеет ровно один левый(правый) вывод.

2 Задачи

Внимание, все задачи на построение автоматов должны быть снабжены диаграммами!

Задача 1.

Предложите алгоритм построения праволинейной грамматики по автомату $\mathcal{A}$ и докажите его корректность. Постройте по автомату $\mathcal{A}$ регулярную праволинейную грамматику G по предложенному алгоритму. Если вместо построения своего алгоритма, Вы возьмёте его из книжки, укажите это и не списывайте страницами, пожалуйста.



Задача 2. Постройте автомат по грамматике G :

$$S \to abaA \mid abB \mid \varepsilon,\ A \to aB \mid aa,\ B \to bA \mid aS$$

Задача 3. Является ли грамматика G из предыдущей задачи однозначной?

Задача 4. Язык L задан КСГ: $S \to aSa \mid aSb \mid bSa \mid bSb \mid a$.

1. Является ли L регулярным языком?
2. Является ли дополнение L регулярным языком?

Задача 5. Постройте КС-грамматику, порождающую язык

$$L = \{\, a^n b^m \mid n \leqslant m \leqslant 2n \,\}.$$