2 АВТОМАТЫ С МАГАЗИННОЙ ПАМЯТЬЮ И КОНТЕКСТНО-СВОБОДНЫЕ ЯЗЫКИ

2.1 Контекстно-свободные грамматики и языки

В этой главе мы рассмотрим еще один способ задания языков — задание при помощи порождающих грамматик.

Начнем с определения контекстно-свободных грамматик.

Определение 2.1.1 (КС-грамматики). *Контекстно-свободная грамматика (сокращенно КС-грамматика) G — это четверка $\langle A, Q, S, P\rangle$, где:*

1. *A — (конечный) алфавит. A называют еще алфавитом терминалов, а его элементы — терминалами.*
2. *Q — (конечный) алфавит. Q называют еще алфавитом нетерминалов, а его элементы — нетерминалами.*
3. *$S \in Q$ — начальный нетерминал, иногда называемый аксиомой.*
4. *P — конечное множество правил. Его элементы — правила.*

Правила имеют вид:

$R \to u$, где $R \in Q$, $u \in (A \bigcup Q)^*$, то есть u — слово в алфавите терминалов и нетерминалов. R называется при этом левой частью правила, u — правой частью правила.

КС-грамматика G определяет отношение выводимости на множестве слов в алфавите $(A \bigcup Q)$.

Определение 2.1.2 (вывода в КС-грамматике). *Пусть $v, w \in (A \bigcup Q)^*$ и $G = \langle A, Q, S, P\rangle$ — КС-грамматика. Будем говорить, что слово v выводимо из слова w за один шаг в грамматике, и писать*

$$w \Rightarrow_G v \quad \text{или} \quad w \Rightarrow^1_G v,$$

тогда и только тогда, когда можно представить w в виде xRy, а v в виде xuy, где $x, y, u \in (A \bigcup Q)^*$, $R \in Q$, и правило $R \to u$ принадлежит G. Если при этом $x \in A^*$, то будем говорить, что v левовыводимо из w за один шаг в грамматике G и писать

$$w \mid\!\Rightarrow_G v \quad \text{или} \quad w \mid\!\Rightarrow^1_G v.$$

Будем писать

$$v \Rightarrow_G^0 v \quad \text{и} \quad v \mid\Rightarrow_G^0 v.$$

Если $n > 0$, *то будем говорить, что слово* v *выводимо из слова* w *за* $n+1$ *шаг в грамматике* G, *и писать*

$$w \Rightarrow_G^{n+1} v,$$

тогда и только тогда, когда существует такое слово $x \in (A \bigcup Q)^*$, *что одновременно*

$$w \Rightarrow_G^n x \quad \text{и} \quad x \Rightarrow_G v.$$

Если при этом

$$w \mid\Rightarrow_G^n x \quad \text{и} \quad x \mid\Rightarrow_G v,$$

то будем говорить, что слово v *левовыводимо из слова* w *за* $n+1$ *шаг в грамматике, и писать*

$$w \mid\Rightarrow_G^{n+1} v.$$

Будем говорить, что слово v *выводимо из слова* w *в грамматике* G, *и писать*

$$w \Rightarrow_G^* v,$$

если либо $w = v$, *либо существует* $k > 0$, *что* $w \Rightarrow_G^k v$. *Если* $k > 0$ *и* $w \Rightarrow_G^k v$, *то будем писать*

$$w \Rightarrow_G^+ v.$$

Будем говорить, что слово v *левовыводимо из слова* w *в грамматике* G, *и писать*

$$w \mid\Rightarrow_G^* v,$$

если либо $w = v$, *либо существует* $k > 0$, *что* $w \mid\Rightarrow_G^k v$. *Если* $k > 0$ *и* $w \mid\Rightarrow_G^k v$, *то будем писать*

$$w \mid\Rightarrow_G^+ v.$$

Букву G *под стрелкой будем иногда опускать, если ясно, о какой грамматике идет речь.*

Выводом слова $w \in A^*$ *в* G *будем называть конечную последовательность слов* $w_0, w_1, w_2, \ldots, w_n \in (A \bigcup Q)^*$ *такую, что:*

1. $w_0 = S$,
2. $w_n = w$,
3. $w_i \Rightarrow_G^1 w_{i+1}$ *для* $i = 0, \ldots, n-1$.

Ясно, что выводимое слово может иметь в G несколько, или даже бесконечно много выводов. В частности, вывод называется левым, если для всякого $i = 0, \ldots, n-1$

$$w_i \mid\Rightarrow^1_G w_{i+1},$$

то есть w_{i+1} левовыводимо из w_i за один шаг.

Иногда по аналогии с левыми рассматривают правые выводы, в которых применение правила всегда производится к самому правому нетерминалу.

Определение 2.1.3 (псевдоязыка и языка грамматики). *Множество всех слов $v \in (A \bigcup Q)^*$, выводимых из S в G, будем называть псевдоязыком КС-грамматики $G = \langle A, Q, S, P\rangle$ и обозначать через $PL(G)$.*

Языком $L(G)$ КС-грамматики $G = \langle A, Q, S, P\rangle$ будем называть множество всех слов $w \in A^*$, выводимых из S в G.

Если $G = \langle A, Q, S, P\rangle$ — КС-грамматика, то язык $L(G)$ называется КС-языком. Две КС-грамматики называются эквивалентными, если их языки совпадают.

Приведем пример.

Пример 2.1.1. *Пусть $G = \langle \{0, 1, a\}, \{S\}, S, P\rangle$ — КС-грамматика, где P — следующее множество правил:*

$$S \to 0SS1, S \to a$$

Слова $0aa1$ и $0a0aa11$ принадлежат $L(G)$. Действительно:

$$S \Rightarrow_G 0SS1 \Rightarrow_G 0aS1 \Rightarrow_G 0aa1,$$

$$S \Rightarrow_G 0SS1 \Rightarrow_G 0aS1 \Rightarrow_G 0a0SS11 \Rightarrow_G 0a0Sa11 \Rightarrow_G 0a0aa11.$$

Первый вывод является левым, а второй не является ни правым, ни левым.

Лемма 2.1. *Пусть $G = \langle A, Q, S, P\rangle$ — КС-грамматика,*

$$u, v, u', v' \in (A \bigcup Q)^*,$$

$$u \Rightarrow^n_G u',$$

$$v \Rightarrow^m_G v'.$$

Тогда $uv \Rightarrow^{n+m}_G u'v'$.

Доказательство. Существуют выводы:

$$u \Rightarrow_G u_1 \Rightarrow_G \dots \Rightarrow_G u_n = u', \quad v \Rightarrow_G v_1 \Rightarrow_G \dots \Rightarrow_G v_m = v'.$$

Пусть $n = 0$, то есть $u = u'$. Индукцией по m покажем, что тогда $uv \Rightarrow_G^m u'v'$. Действительно, если $m = 0$, то $v = v'$, то есть $uv = u'v'$, значит, $uv \Rightarrow_G^0 u'v'$. Если же $m > 0$, то, по определению, $v_0 = v$ представляется в таком виде xyz, где $x, z \in (A \bigcup Q)^*$ и $y \in Q$, что $v_1 = xwz$ для некоторого такого $w \in (A \bigcup Q)^*$, что $y \to w \in P$. Тогда очевидно, что uv представляется в виде $uxyz$ и что $uv = uxyz \Rightarrow_G uxwz = uv_1$. По индукции, так как $v_1 \Rightarrow_G^{m-1} v'$, имеем $uv_1 \Rightarrow_G^{m-1} uv' = u'v'$. Сопоставляя эти два вывода, получим, что $uv \Rightarrow_G^m u'v'$.

Пусть теперь $n > 0$. По определению, $u_0 = u$ представляется в таком виде xyz, где $x, z \in (A \bigcup Q)^*$ и $y \in Q$, что $u_1 = xwz$ для некоторого такого $w \in (A \bigcup Q)^*$, что $y \to w \in P$. Ясно, что тогда $u_0v = xyzv \Rightarrow_G xwzv = u_1v$. Но $u_1 \Rightarrow_G^{n-1} u_n = u'$, значит, по индукции, $u_1v \Rightarrow_G^{n+m-1} u'v'$. Сопоставляя, получим $uv = u_0v \Rightarrow_G^{n+m} u'v'$. ∎

Лемма 2.2. *Пусть $uv \Rightarrow_G^k w$ в КС-грамматике $G = \langle A, Q, S, P \rangle$, где $u, v, w \in (A \bigcup Q)^*$. Тогда существуют $u', v' \in (A \bigcup Q)^*$ и такие натуральные числа m и n, что $0 \le m, n \le k$; $m + n = k$; $w = u'v'$ и $u \Rightarrow_G^n u'$, $v \Rightarrow_G^m v'$.*

Доказательство. Если $k = 0$, то $w = uv$, и лемма доказана, так как $v \Rightarrow_G^0 v$, $u \Rightarrow_G^0 u$.

Пусть $k > 0$. Тогда существует $w' \in (A \bigcup Q)^*$ такое, что $uv \Rightarrow_G w'$, $w' \Rightarrow_G^{k-1} w$. Тогда слово uv представляется в виде xyz, где $x, z \in (A \bigcup Q)^*$ и $y \in Q$, что $w' = xsz$ для некоторого такого $s \in (A \bigcup Q)^*$, что $y \to s \in P$.

Пусть $|x| \ge |u|$. Тогда $x = ux'$ для подходящего x'. Тогда $v = x'yz \Rightarrow_G x'sz$, по определению. С другой стороны, $ux'sz = w' \Rightarrow_G^{k-1} w$, и по индукции, существуют такие $u', v' \in (A \bigcup Q)^*$ и натуральные числа m, n, что $0 \le m \le k-1$, $0 \le n \le k-1$, $m + n = k - 1$, $w = u'v'$, $u \Rightarrow_G^n u'$, $x'sz \Rightarrow_G^m v'$. Так как $v = x'yz \Rightarrow_G x'sz$, то $v \Rightarrow_G^{m+1} v'$ и лемма доказана, так как $0 < m + 1 \le k$.

Аналогично рассматривается случай, когда $|x| < |u|$, то есть $|z| \ge |v|$. ∎

Следствие 2.3. *Пусть $G = \langle A, Q, S, P \rangle$ — КС-грамматика, $R \in Q$, $R \to \epsilon_1\epsilon_2 \dots \epsilon_n \in P$, $\epsilon_1, \epsilon_2, \dots, \epsilon_n \in (A \bigcup Q)^*$ и для некоторого $w \in (A \bigcup Q)^*$, $R \Rightarrow_G \epsilon_1\epsilon_2 \dots \epsilon_n \Rightarrow_G^* w$. Тогда для некоторых $w_1, w_2, \dots, w_n \in (A \bigcup Q)^*$, слово w представляется в виде $w = w_1w_2 \dots w_n$ и $\epsilon_i \Rightarrow_G^* w_i$.*

Доказательство. Очевидное следствие леммы. ∎

Лемма 2.4. *Пусть* $G = \langle A, Q, S, P\rangle$ — *КС-грамматика;* $R, X \in Q$; $R \to \alpha X\beta \in P$; $\alpha, \beta \in (A \bigcup Q)^*$, *и пусть* $X \to \theta_1$, $X \to \theta_2, \ldots, X \to \theta_n$ — *все правила из* P *с левой частью* X, *где* $\theta_1, \theta_2, \ldots, \theta_n \in (A \bigcup Q)^*$. *Если* $u \in (A \bigcup Q)^*$, $v \in A^*$, $u \Rightarrow^*_G v$, *то* $u \Rightarrow^*_{G'} v$, *где* $G' = \langle A, Q, S, P'\rangle$,

$$P' = (P \setminus \{R \to \alpha X\beta\}) \bigcup \{R \to \alpha\theta_1\beta,\ R \to \alpha\theta_2\beta, \ldots, R \to \alpha\theta_n\beta\}.$$

Доказательство. Пусть, в условиях леммы, $G'' = \langle A, Q, S, P''\rangle$, где

$$P'' = P \bigcup \{R \to \alpha\theta_1\beta,\ R \to \alpha\theta_2\beta, \ldots, R \to \alpha\theta_n\beta\}.$$

Пусть $u \in (A \bigcup Q)^*$, $v \in A^*$, $u \Rightarrow^*_G v$. Тогда $u \Rightarrow^*_{G''} v$, то есть, для некоторого m, $u \Rightarrow^m_{G''} v$. Значит, существуют $u_0, u_1, \ldots, u_m \in (A \bigcup Q)^*$ такие, что

$$u = u_0 \Rightarrow_{G''} u_1 \Rightarrow_{G''} u_2 \Rightarrow_{G''} \cdots \Rightarrow_{G''} u_m = v.$$

По определению, для всякого $j = 0, 1, \ldots, m-1$, u_j представляется в виде xYz, где $x, z \in (A \bigcup Q)^*$, а $Y \in Q$, так, что для некоторого правила $Y \to y \in P''$, $y \in (A \bigcup Q)^*, u_{j+1} = xyz$. Это правило $Y \to y \in P''$ обозначим p_{i_j}.

Пусть среди $p_{i_0}, p_{i_1}, \ldots, p_{i_{m-1}}$ нет правила $R \to \alpha X\beta$ из условия леммы. Тогда, по определению,

$$u = u_0 \Rightarrow_{G'} u_1 \Rightarrow_{G'} u_2 \Rightarrow_{G'} \cdots \Rightarrow_{G'} u_m = v$$

и $u \Rightarrow^*_{G'} u_m = v$, что доказывает лемму.

Иначе пусть k, $0 \le k < m$, — максимальное такое, что p_{i_k} есть $R \to \alpha X\beta$. Тогда по построению u_k представляется в виде xRz, а u_{k+1} — в виде $x\alpha X\beta z$ и $u_{k+1} \Rightarrow^*_{G'} v$ (в силу максимальности). Применяя лемму 2.2 четыре раза, получим, что v представляется в виде $v_1v_2v_3v_4v_5$ так, что:

$$x \Rightarrow^*_{G'} v_1;\quad \alpha \Rightarrow^*_{G'} v_2;\quad X \Rightarrow^*_{G'} v_3;\quad \beta \Rightarrow^*_{G'} v_4;\quad z \Rightarrow^*_{G'} v_5.$$

Так как $v \in A^*$, то и $v_3 \in A^*$, значит, $v_3 \ne X$. Тогда для некоторого l, $1 \le l \le n$,

$$X \Rightarrow_{G'} \theta_l \Rightarrow^*_{G'} v_3.$$

По лемме 2.1,

$$\alpha X\beta \Rightarrow^*_{G'} \alpha\theta_l\beta \Rightarrow^*_{G'} v_2v_3v_4.$$

Но $R \Rightarrow_{G'} \alpha\theta_l\beta$, значит, $R \Rightarrow^*_{G'} v_2v_3v_4$. Тогда, по лемме 2.1,

$$u_k = xRz \Rightarrow^*_{G'} v_1v_2v_3v_4v_5 = v.$$

Значит,

$$u = u_0 \Rightarrow^*_{G''} u_k \Rightarrow^*_{G'} v$$

с меньшим на 1 числом применений правила $R \to \alpha X\beta$ и, по индукции, $u = u_0 \Rightarrow^*_{G'} v$, что и требовалось доказать. ∎

По определению, всякий левый вывод является выводом. Оказывается, что можно всегда ограничиться рассмотрением только левых выводов.

Лемма 2.5. *Пусть* $G = \langle A, Q, S, P\rangle$ *— КС-грамматика,* $u \in (A \bigcup Q)^*$, $v \in A^*$, $u \Rightarrow^*_G v$. *Тогда* v *левовыводимо из* u, *то есть* $u \mid\Rightarrow^*_G v$.

Доказательство. Пусть $u \Rightarrow^k_G v$. Если $k = 0$, то, по определению, $u \mid\Rightarrow^0_G v$, и лемма доказана.

Пусть $k > 0$. Тогда, по определению, для всякого $i = 0, 1, \ldots, k-1$ существуют u_i, x_i, y_i, z_i в $(A \bigcup Q)^*$ и $X_{j_i} \in Q$ такие, что:

1. $u = u_0$;
2. $u_{i-1} \Rightarrow_G u_i$ для $i = 1, 2, \ldots, k-1$;
3. $u_{k-1} \Rightarrow_G v$;
4. $u_i = x_iX_{j_i}z_i$ для $i = 0, 1, \ldots, k-1$;
5. $u_i = x_{i-1}y_{i-1}z_{i-1}$ для $i = 1, 2, \ldots, k-1$;
6. $X_{j_i} \to y_i \in P$ для $i = 0, 1, \ldots, k-1$.

Рассмотрим минимальное n, $0 \le n \le k-1$, такое, что $x_n \notin A^*$. Если такого n не существует, то, по определению,

$$u = u_0 \mid\Rightarrow_G u_1 \mid\Rightarrow_G \cdots \mid\Rightarrow_G u_{k-1} \mid\Rightarrow_G v,$$

то есть $u \mid\Rightarrow^*_G v$, и лемма доказана.

В противном случае ясно, что

$$u \mid\Rightarrow^*_G u_n \ \text{ и } \ u_n \Rightarrow^{k-n}_G v.$$

Тогда по лемме 2.2, существуют такие $u', v' \in A^*$ и натуральные числа l, l', что $0 \le l \le k-n$, $0 \le l' \le k-n$, $l+l' = k-n$, $x_n \Rightarrow^l_G u'$, $X_{j_n}z_n \Rightarrow^{l'}_G v'$, и $v = u'v'$. Так как ни x_n, ни $X_{j_n}z_n$ не являются словами в алфавите A, а $v = u'v' \in A^*$, то $l > 0$ и $l' > 0$. Значит, $l < k-n$ и $l' < k-n \le k$. Тогда, по индукции, $x_n \mid\Rightarrow^*_G u'$, $X_{j_n}z_n \mid\Rightarrow^*_G v'$.

Видно, что лемма 2.1 справедлива и для левых выводов (доказательство проходит буквально). Значит, из предыдущего следует, что $u_n = x_n X_{j_n} z_n \mid\Rightarrow_G^* u'v' = v$. Но, так как $u \mid\Rightarrow_G^* u_n$, то $u \mid\Rightarrow_G^* v$, что и требовалось доказать. ∎

Оказывается, что все регулярные языки являются КС-языками.

Теорема 2.6. *Если R — регулярный язык, то R — КС-язык.*

Доказательство. Пусть $J = \langle A, Q, q_0, F, P\rangle$ — задающий R ДКА. Построим по нему КС-грамматику $G = \langle A, Q', S, P'\rangle$ такую, что

$$R = L(J) = L(G).$$

Положим $Q' = Q$, $S = q_0$.

Без ограничения общности будем считать, что для всякого $q \in Q$ и всякого $a \in A$ существует $q' \in Q$, что $\langle q, a, q'\rangle \in P$.

Для каждого правила $\langle q, a, q'\rangle \in P$ включим в P' правило $q \to aq'$. Для всякого $q \in F$ включим в P' правило $q \to e$, где e — пустое слово.

Других правил в P' нет.

Покажем, что $L(G) = L(J) = R$, чем будет доказана теорема.

В параграфе 1.4 замечено, что $w \in L(J)$ тогда и только тогда, когда существует $q'' \in F$ такое, что $(q_0, w) \mid\Rightarrow_J^+ (q'', e)$. С другой стороны, по определению, $w \in L(G)$ тогда и только тогда, когда $q_0 \Rightarrow_G^+ w$. Это эквивалентно тому, что существует $q'' \in F$ такое, что $q_0 \Rightarrow_G^+ wq''$, поскольку для всякого $q'' \in F$ и только для них $q'' \Rightarrow_G e$.

Покажем, что для произвольных $q, q' \in Q$, $w \in A$,

$$(q, w) \mid\Rightarrow_J^+ (q', e)$$

тогда и только тогда, когда

$$q \Rightarrow_G^+ wq'.$$

Из этого утверждения будет следовать $L(G) = L(J)$.

Итак, пусть $(q, w) \mid\Rightarrow_J (q', e)$. По определению это значит, что $w \in A$ и $\langle q, w, q'\rangle \in P$. Но тогда $q \to wq' \in P'$, значит, $q \Rightarrow_G wq'$. Аналогично показывается, что из $q \Rightarrow_G wq'$ следует $(q, w) \mid\Rightarrow_J (q', e)$.

Пусть теперь $(q, w) \mid\Rightarrow_J^+ (q', e)$.

Покажем, что $q \Rightarrow_G^+ wq'$. Из того, что $(q, w) \mid\Rightarrow_J^+ (q', e)$, следует по определению, что существуют $a_0, a_1, \ldots, a_n \in A$ и $q_{i_0}, q_{i_1}, \ldots, q_{i_{n+1}} \in Q$ такие, что

1. $w = a_0 a_1 \dots a_n$;

2. $q_{i_0} = q, \ \ q_{i_{n+1}} = q'$;

3. $(q_{i_k}, a_k a_{k+1} \dots a_n) \models_J (q_{i_{k+1}}, a_{k+1} a_{k+2} \dots a_n)$ для всех таких k, что $0 \le k \le n$.

Из (3) следует, что $(q_{i_k}, a_k) \models_J (q_{i_{k+1}}, e)$.

Значит, по доказанному выше, при введенных обозначениях, для всех k, $0 \le k \le n$,

$$q_{i_k} \Rightarrow_G a_k q_{i_{k+1}}.$$

По лемме 2.1, тогда $q \Rightarrow_G^+ wq'$.

Аналогично показывается, что из

$$q \Rightarrow_G^+ wq'$$

следует

$$(q, w) \models_J^+ (q', e).$$

Теорема доказана. ∎

Однако не всякий КС-язык регулярен. Чтобы продемонстрировать это, рассмотрим язык из упражнения 1.3.2, нерегулярность которого доказана в следствии 1.10.

Итак, язык из упражнения 1.3.2 нерегулярен. Однако он является КС-языком, ибо задается следующей КС-грамматикой:

$$G = \langle \{0, 1\}, \{S\}, S, \{S \to 0S1, S \to e\} \rangle.$$

Упражнения

Упражнение 2.1.1. *Являются ли следующие языки контекстно-свободными:*

1) язык в алфавите $\{0, 1\}$*, составленный из всех слов с равным числом* 0 *и* 1;

2) язык, составленный из всевозможных последовательностей правильных расстановок скобок.

Упражнение 2.1.2. *Предложите способ определения принадлежности данного слова языку, определенному данной КС-грамматикой.*

2.2 Нормальная форма Хомского

В этом параграфе мы покажем,что правила КС-грамматик могут быть всегда представлены в упрощенном виде.

Определение 2.2.1 (нормальной формы Хомского). *Будем говорить, что КС-грамматика*

$$G = \langle A, Q, S, P \rangle$$

находится в нормальной форме Хомского, если каждое ее правило имеет один из следующих видов:

1. $S \to e$, где e — пустое слово.
2. $R \to a$, где $R \in Q$, $a \in A$.
3. $R \to R_1R_2$, где $R \in Q$ и $R_1, R_2 \in Q \setminus \{S\}$.

Теорема 2.7. *Для всякой КС-грамматики $G = \langle A, Q, S, P \rangle$ существует эквивалентная ей КС-грамматика $G' = \langle A, Q', S', P' \rangle$ в нормальной форме Хомского.*

Для доказательства теоремы нам потребуются промежуточные леммы и определения.

Определение 2.2.2 (грамматики без пустых правил). *Грамматика $G = \langle A, Q, S, P \rangle$ называется грамматикой без e-правил, если она либо вообще не имеет правил вида $R \to e$ ни для какого $R \in Q$, либо имеет ровно одно такое правило $S \to e$ для начального символа $S \in Q$, причем S не встречается в правых частях правил из P.*

Лемма 2.8. *По всякой КС-грамматике $G = \langle A, Q, S, P \rangle$ может быть построена эквивалентная грамматика $G' = \langle A, Q', S', P' \rangle$ без e-правил.*

Доказательство. Рассмотрим множество $E \subseteq Q$:

$$E = \{R \in Q \mid \ R \Rightarrow^*_G e\}.$$

Это множество E может быть построено не более чем за $|Q|$ шагов так.

Положим

$$E_0 = \{R \in Q \mid \ R \Rightarrow_G e\}.$$

Если E_i уже построено, определим

$$E_{i+1} = \{R \in Q \mid \ R \to w \in P, \ w \in E_i^*\} \bigcup E_i.$$

Ясно, что $E_{|Q|} = E_{|Q|+l}$ для всякого $l \geq 0$. Положим $E = E_{|Q|}$.

Построим теперь P'.

Всякое правило из P может быть, для некоторых $m \geq 0$, $R \in Q$, представлено в виде:

$$R \to w_0 R_{i_1} w_1 R_{i_2} \dots w_{m-1} R_{i_m} w_m,$$

где

$$R_{i_1}, R_{i_2}, \dots, R_{i_m} \in E,$$

$$w_0, w_1, \dots, w_m \in ((Q \setminus E) \bigcup A)^*.$$

Включим в P' для каждого такого правила из P всевозможные правила вида

$$R \to w_0 v_1 w_1 v_2 w_2 \dots w_{m-1} v_m w_m,$$

где v_j либо равно R_{i_j}, либо равно e.

Затем включим в Q' все нетерминалы из Q и новый нетерминал S'.

Теперь исключим из P' все правила с пустой правой частью, и включим правило $S' \to S$ в P.

Наконец, если $S \in E$, включим в P' правило $S' \to e$.

Формальное доказательство леммы следует проводить подобно доказательству леммы 2.4 индукцией с использованием лемм 2.1 и 2.2. То есть сначала следует рассмотреть грамматику с множеством правил $P \bigcup P'$, а затем постепенно элиминировать все правила из P. Аналогичный метод применен и для доказательства леммы 2.9 ниже. ∎

Правила вида $R \to R'$, где R, R' — нетерминалы, мы будем называть цепными правилами.

Лемма 2.9. *По всякой КС-грамматике $G = \langle A, Q, S, P \rangle$ может быть построена эквивалентная грамматика $G' = \langle A, Q', S', P' \rangle$ без цепных правил.*

Доказательство. В силу леммы 2.8 можно считать, что G является грамматикой без e-правил.

Для каждого $R \in Q$ построим множество

$$H(R) = \{R' \in Q \mid \; R' \Rightarrow_G^* R\}$$

следующим образом. Пусть $H_0(R) = \{R\}$. Если $H_i(R)$ уже построено, то положим

$$H_{i+1}(R) = \{R'' \in Q \mid \; R'' \to R' \in P, \; R' \in H_i(R)\} \bigcup H_i(R).$$

Ясно, что

$$H_{|Q|}(R) = H_{|Q|+l}(R)$$

для всяких R и $l \geq 0$. Значит, мы вправе положить $H(R) = H_{|Q|}(R)$.

Пусть теперь $P'' \subseteq P$ и P'' содержит все цепные правила, и только их. Пусть $P' = P \setminus P''$, то есть P' — все нецепные правила из P.

Для каждого правила $R \to w$ из P' и для каждого $R' \in H(R)$ включим в P' правило $R' \to w$.

Положим $S' = S$, $Q' = Q$.

Покажем, что построенная грамматика эквивалентна исходной. Для этого рассмотрим сначала грамматику

$$G'' = \langle A, Q, S, P \bigcup P' \rangle.$$

Пусть $u \in (A \bigcup Q)^*$, $v \in A^*$ и $u \Rightarrow_G^* v$. Тогда $u \Rightarrow_{G''}^* v$, то есть, для некоторого m, $u \Rightarrow_{G''}^m v$. Значит, существуют

$$u_0, u_1, \ldots, u_m \in (A \bigcup Q)^*$$

такие, что

$$u = u_0 \Rightarrow_{G''} u_1 \Rightarrow_{G''} u_2 \Rightarrow_{G''} \cdots \Rightarrow_{G''} u_m = v.$$

По определению, для всякого $j = 0, 1, \ldots, m-1$, u_j представляется в виде xYz, где $x, y \in (A \bigcup Q)^*$, а $Y \in Q$ так, что для некоторого правила $Y \to y \in P''$, где $y \in (A \bigcup Q)^*$, $u_{j+1} = xyz$. Это правило $Y \to y$ из P'' обозначим p_{i_j}.

Пусть среди

$$p_{i_0},\ p_{i_1}, \ldots, p_{i_{m-1}}$$

нет цепных правил. Заметим, что всякое нецепное правило из P содержится и в P', поскольку $R \in H(R)$ для всякого $R \in Q$. Получаем, что

$$u = u_0 \Rightarrow_{G'} u_1 \Rightarrow_{G'} u_2 \Rightarrow_{G'} \cdots \Rightarrow_{G'} u_m = v.$$

То есть $u \Rightarrow_{G'}^* v$ и лемма доказана ($u = S$ — частный случай).

Иначе пусть k, $0 \leq k < m$, — максимальное число такое, что p_{i_k} — цепное правило. Пусть p_{i_k} есть $R \to R'$ для некоторых $R, R' \in Q$. Заметим, что при этом $R \in H(R')$. По определению, тогда u_k представляется в виде xRy для $x, y \in (Q \bigcup A)^*$, $u_{k+1} = xR'y$, и $u_{k+2} = xwy$ для некоторого $w \in (A \bigcup Q)^*$ и нецепного (в силу максимальности) $R' \to w \in P$. Так как $R \in H(R')$, то, по построению, $R \to w \in P'$, то есть $R \Rightarrow_{G'} w$.

В силу максимальности, по ранее доказанному,

$$u_{k+2} = xwy \Rightarrow^*_{G'} v.$$

По лемме 2.2, тогда слово v представляется в виде $v = v_1v_2v_3$ так, что

$$x \Rightarrow^*_{G'} v_1, \quad w \Rightarrow^*_{G'} v_2, \quad y \Rightarrow^*_{G'} v_3.$$

Сопоставляя, имеем $R \Rightarrow^*_{G'} v_2$ и, по лемме 2.1,

$$u_k = xRy \Rightarrow^*_{G'} v_1v_2v_3 = v.$$

То есть

$$u \Rightarrow^*_{G''} u_k \Rightarrow^*_{G'} v$$

с меньшим числом применений цепных правил, значит, по индукции, $u \Rightarrow^*_{G'} v$ и лемма доказана ($u = S$ — частный случай). ∎

Заметим, что новые e-правила при устранении цепных правил не возникают.

Теперь мы готовы доказать теорему о нормальной форме Хомского.

Доказательство теоремы. Мы вправе предположить,что грамматика $G = \langle A, Q, S, P\rangle$ не содержит цепных правил и является грамматикой без e-правил.

Рассмотрим множество нетерминальных символов

$$Q_A = \{R_a \mid \ a \in A\}.$$

Рассмотрим грамматику $G'' = \langle A, Q'', S, P''\rangle$, где

$$Q'' = (Q \bigcup Q_A), \quad P'' = \{R_a \to a \mid \ a \in A\} \bigcup P''',$$

а P''' получено из P заменой в неодноэлементных правых частях правил из P терминалов $a \in A$ на соответствующие им нетерминалы $R_a \in Q_A$.

Таким образом, в G'' всякое правило из P'' имеет один из видов:

1. $S \to e$,
2. $R \to a$, где $R \in Q''$ и $a \in A$,
3. $R \to R_{i_1} \dots R_{i_m}$, где $R \in Q''$, $R_{i_1}, \dots, R_{i_m} \in (Q'' \setminus \{S\})$, а $m \geq 2$.

Чтобы преобразовать G'' в грамматику в нормальной форме Хомского, осталось лишь преобразовать правила со слишком длинной нетерминальной правой частью.

Итак, включим в Q' все нетерминалы из Q'', а в P' все правила из P'', кроме правил вида:

$$R \to R_{i_1} R_{i_1} \dots R_{i_m},$$

где $R, R_{i_1}, \dots R_{i_m} \in Q''$, и $m > 2$.

Для каждого же правила указанного вида из P'', включим в Q' новые нетерминалы

$$N_{i_2},\ N_{i_3}, \dots, N_{i_{m-1}},$$

а в P' новые правила:

$$R \to R_{i_1} N_{i_2},$$

$$N_{i_j} \to R_{i_j} N_{i_{j+1}}$$

для $2 \leq j \leq m-2$ и правило

$$N_{i_{m-1}} \to R_{i_{m-1}} R_{i_m}.$$

Положив $S' = S$, легко видеть, используя лемму 2.4, что грамматика $G' = \langle A, Q', S', P' \rangle$ обладает требуемыми в условиях теоремы свойствами. ∎

Пусть $G = \langle A, Q, S, P \rangle$ — КС-грамматика. Символ $R \in Q$ называется недостижимым, если для всех $\alpha, \beta \in (Q \bigcup A)^*$ неверно, что $S \Rightarrow_G^* \alpha R \beta$.

Символ $R \in Q$ называется непродуктивным, если ни для какого $\alpha \in A^*$ неверно, что $R \Rightarrow_G^* \alpha$.

Символ называется бесполезным, если он либо непродуктивен, либо недостижим.

Грамматика G называется *приведенной*, если Q не содержит бесполезных символов, кроме, быть может, S.

Теорема 2.10. *Пусть $G = \langle A, Q, S, P \rangle$ — КС-грамматика. Можно преобразовать ее к эквивалентной приведенной КС-грамматике $G' = \langle A, Q', S, P' \rangle$.*

Доказательство. Построение эквивалентной G приведенной грамматики будет производиться в два этапа. Сначала будет построено множество J непродуктивных символов грамматики. Ясно, что все выводы всех слов языка $L(G)$ получаются без участия непродуктивных символов. Таким образом, мы можем не меняя языка грамматики выбросить все правила, содержащие непродуктивные символы, из P, получив множество правил P'', и соответственно выбросить J из Q, получив множество нетерминалов Q''. Если при этом Q'' стало пустым, добавим к нему символ S (таким образом получив грамматику вообще без правил, что соответствует случаю, когда все символы непродуктивны, и язык грамматики пуст). Итак, грамматика $G'' = \langle A, Q'', S, P'' \rangle$ эквивалентна G и все ее символы продуктивны.

Теперь построим множество I недостижимых символов в грамматике G''. Так как в выводах в грамматике G'' участвуют лишь достижимые символы, то можно без ущерба выбросить все правила, содержащие недостижимые символы, из P'', получив множество правил P', и соответственно выбросить I из Q'', получив множество нетерминалов Q'. Отметим, что в данном случае аксиома S не будем выброшена, ибо она по определению всегда достижима. Итак, грамматика $G' = \langle A, Q', S, P' \rangle$ эквивалентна грамматике $G'' = \langle A, Q'', S, P'' \rangle$, а значит, и грамматике $G = \langle A, Q, S, P \rangle$, и не содержит недостижимых символов. Заметим, что она не содержит и непродуктивных символов. Действительно, все символы из Q'' в грамматике G'' были продуктивными. Допустим, что некоторый символ $R \in Q' \subseteq Q''$ непродуктивен в грамматике G'. Так как R был продуктивен в G'', то существуют $R' \in (Q'' \setminus Q')$ и $x, y \in (Q'' \bigcup A)^*$ такие, что $R \Rightarrow_{G''} xR'y$. Но из последнего следует, что R' достижим, так как $R \in Q'$. Получили противоречие с тем, что $R' \notin Q'$.

Итак, все символы грамматики $G' = \langle A, Q', S, P' \rangle$ достижимы и продуктивны.

Для завершения доказательства осталось построить множества J и I.

Построение множества $J \subseteq Q$ непродуктивных символов грамматики.

Построим множество J' продуктивных символов и положим $J = Q \setminus J$. Чтобы построить J', построим последовательность

$$J_0,\ J_1, \ldots, J_{|Q|}$$

подмножеств Q так:

$$J_0 = \emptyset, \quad J_{i+1} = J_i \bigcup \{R \in Q \mid \ (R \to \alpha) \in P, \quad \alpha \in (A \bigcup J_i)^*\}.$$

Другими словами, на первом шаге мы получим множество "непосредственно продуктивных символов"J_1, то есть множество таких нетерминалов, из которых терминальная цепочка выводима за один шаг применением правила из P.

На каждом следующем шаге мы добавляем множество тех нетерминалов, из которых за один же шаг применением правила из P выводима цепочка, составленная лишь из терминалов, и таких нетерминалов, продуктивность которых уже ранее установлена.

Ясно, что для всякого $i = 1, 2, \ldots, |Q|$, J_i содержит только продуктивные символы. С другой стороны, если на некотором шаге не добавилось ни одного продуктивного символа, то ни на каком последующем также ни одного символа не добавится. Таким образом, данный процесс построения имеет смысл делать не более $|Q|$ раз. И наконец, всякий продуктивный нетерминал на некотором шаге i попадет во множество J_i, значит, всякий продуктивный нетерминал содержится во множестве $J_{|Q|}$.

Итак, можно положить $J' = J_{|Q|}$.

Построение множества $I \subseteq Q''$ недостижимых символов грамматики G''.

Построим множество I' достижимых символов грамматики G'' и положим $I = Q'' \setminus I'$.

Чтобы построить I', построим последовательно подмножества

$$I_0, I_1, I_2, \ldots, I_{|Q''|}$$

множества Q'' так:

$$I_0 = \{S\},$$

$$I_{i+1} = I_i \bigcup \{R \in Q \mid \ (R' \to \alpha R \beta) \in P'', \ \ R' \in I_i, \ \ \alpha, \beta \in (A \bigcup Q'')^*\}.$$

Здесь процесс прямо обратный процессу построения продуктивных символов. На первом шаге мы получим множество "непосредственно выводимых символов"I_1, то есть множество таких нетерминалов, которые достижимы (в некотором контексте) из аксиомы S за один шаг применением правила из P''.

На каждом следующем шаге мы добавляем множество тех нетерминалов, которые за один же шаг применением правила из P'' достижимы из нетерминалов, достижимость которых уже ранее установлена.

Соображения, аналогичные рассмотренным в предыдущем случае, показывают, что можно взять $I' = I_{|Q''|}$. ∎

Теорема 2.11. *Всякая КС-грамматика G может быть преобразована к эквивалентной приведенной КС-грамматике G' в нормальной форме Хомского.*

Доказательство. Построим по G эквивалентную грамматику G'' в нормальной форме Хомского (теорема 2.7). Далее используем конструкцию теоремы 2.10 и строим по G'' эквивалентную приведенную грамматику G'. Очевидно, что G' — приведенная грамматика в нормальной форме Хомского. ∎

Упражнения

Упражнение 2.2.1. *Докажите теорему 2.11 формально.*

Упражнение 2.2.2. *Приведите к нормальной форме Хомского грамматики* $G = \langle\{a, 0, 1\}, \{R, R', S, S'\}, S, P\rangle$ *со следующими наборами правил* P*:*

1.

$$S \to RR,$$
$$R \to e,$$
$$R \to a,$$
$$R \to R',$$
$$R' \to S,$$
$$R' \to SaR.$$

2.

$$S \to SS0S,$$
$$S \to 01S1S,$$
$$S \to 1,$$
$$S \to e.$$

3.

$$S \to SS'SS',$$
$$S \to S'1S,$$
$$S \to 0,$$
$$S' \to SS'S,$$
$$S' \to S'SS',$$
$$S' \to S'1SS0S',$$
$$S' \to S1,$$
$$S' \to e.$$

2.3 Нормальная форма Грейбах

Хотя теорема о нормальной форме Хомского позволяет представлять правила грамматики в упрощенном виде, она дает нам мало информации о выводах в грамматике. Для многих методов работы с грамматиками оказываются важными различные свойства выводов. Например, если в грамматике $G = \langle A, Q, S, P\rangle$ для некоторых $R \in Q$ и $w \in (A \bigcup Q)^*$ верно $R \Rightarrow_G^* Rw$, то R называется *леворекурсивным* символом, а G — *леворекурсивной* грамматикой. Многие алгоритмы работы с грамматиками могут работать лишь с нелеворекурсивными грамматиками.

Теорема 2.12. *По всякой КС-грамматике $G = \langle A, Q, S, P\rangle$ может быть построена эквивалентная ей КС-грамматика $G' = \langle A, Q', S, P'\rangle$, не являющаяся леворекурсивной.*

Доказательство. Без ограничения общности можно считать, что G находится в нормальной форме Хомского (в силу теоремы 2.7).

Рассмотрим некоторое упорядочение R_1, R_2, $R_3, \ldots$, R_n элементов Q. Рассмотрим также новые нетерминалы $R'_1, R'_2, \ldots R'_n \notin Q$. Возьмем в качестве Q' множество $Q \bigcup \{R'_1, R'_2, \ldots R'_n\}$.

Будем строить теперь множество правил P'. Для этого построим последовательность множеств правил $P_0, P_1, P_2, \ldots P_n$, где $P_0 = P$, и положим $P' = P_n$. Мы стремимся удовлетворить следующие условия:

1. Грамматики $\langle A, Q', S, P_i\rangle$ эквивалентны для всех $0 \le i \le n$.

2. В грамматиках $\langle A, Q', S, P_i\rangle$ для $0 \le i \le n$ символы R'_1, $R'_2, \ldots$, R'_n не являются леворекурсивными.

3. Во множестве P_i для произвольного такого j, что $1 \le j \le n$, и для всякого правила $R_j \to \alpha$, либо α начинается с терминала из A, либо с такого нетерминала R_k, что $k > min(j, i)$. В частности, нетерминалы $R_1, R_2, \ldots, R_i$ в грамматике $\langle A, Q', S, P_i\rangle$ не леворекурсивны.

4. S не встречается в правых частях правил из P_i.

5. Грамматика $\langle A, Q', S, P_i\rangle$ является грамматикой без e-правил и без цепных правил.

Из (2) и (3) очевидно следует, что $G' = \langle A, Q', S, P'\rangle$ не левокурсивна, ибо $P' = P_n$. Очевидно, что условия (1) — (5) выполнены для P_0, ибо G находится в нормальной форме Хомского.

Пусть теперь P_{i-1} построено для $1 \le i \le n$. Построим P_i.

По предположению индукции в P_{i-1} все правила $R_j \to \alpha$ для $1 \le j \le n$ либо начинаются с терминала, либо с такого R_k, что $k > min(i-1, j)$.

Рассмотрим сначала в качестве P_i множество P_{i-1}. Ясно, что пункты (1),(2),(4),(5) гипотезы для P_i выполнены, а пункт (3) может нарушаться двояко: либо правилами вида $R_i \to R_i\alpha$, либо правилами вида $R_j \to R_i\alpha$ для $j > i$.

Если $R_i = S$, то, в силу пункта (4), обе возможности невозможны, и построение P_i завершено.

Иначе (т.е. если R_i отлично от S) рассмотрим все правила с левой частью R_i. Пусть это правила:

$$R_i \to R_i\alpha_1, R_i \to R_i\alpha_2, \ldots, R_i \to R_i\alpha_m,$$

$$R_i \to \beta_1, R_i \to \beta_2, \ldots, R_i \to \beta_l,$$

где

$$\alpha_1, \ldots, \alpha_m, \beta_1, \ldots, \beta_l \in (A \bigcup Q)^*,$$

$$\beta_1, \ldots, \beta_l \text{ не начинаются с } R_i.$$

Следовательно, $\beta_1, \ldots, \beta_l$ могут начинаться либо с терминала, либо с некоторого такого R_j, что $j > i$ (по индукции).

Заменим в P_i все правила с левой частью R_i на правила:

$$R_i \to \beta_1, R_i \to \beta_2, \ldots, R_i \to \beta_l,$$

$$R_i \to \beta_1 R_i', R_i \to \beta_2 R_i', \ldots, R_i \to \beta_l R_i',$$

$$R_i' \to \alpha_1, R_i' \to \alpha_2, \ldots, R_i' \to \alpha_m,$$

$$R_i' \to \alpha_1 R_i', R_i' \to \alpha_2 R_i', \ldots, R_i' \to \alpha_m R_i'.$$

Эта замена сохранила эквивалентность грамматик.

Действительно, рассмотрим $u \in (A \bigcup Q)^*$ и $w \in A^*$ такие, что $u \Rightarrow^+ w$ в нашей грамматике до последней замены. Покажем, что $u \Rightarrow^+ w$ и в грамматике после последней замены. Рассмотрим левый вывод w из u в старой грамматике, то есть такие $n > 1$ и $x_1, x_2, \ldots, x_n \in (A \bigcup Q)^*$, что $x_1 = u$, $x_n = w$ и $x_j \Rightarrow x_{j+1}$ для $j = 1, 2, \ldots, n-1$ (такой вывод существует в силу леммы 2.5).

Утверждение будем доказывать индукцией по n.

Если $n = 1$, то на этом единственном шаге в этом выводе ни одно из правил вида $R_i \to R_i\alpha_j$ не использовалось, и тот же самый вывод останется корректным и в новой грамматике.

Пусть $n > 1$. Если в выводе ни одно из правил вида

$$R_i \to R_i\alpha_j \tag{2}$$

не использовалось, то тот же самый вывод останется корректным и в новой грамматике.

Если это неверно, то пусть k — наименьшее такое, что в рассматриваемом выводе при переходе от x_k к x_{k+1} использовалось правило (2) при некотором j. Пусть для некоторых v и v' выполняется равенство $x_k = vR_iv'$. Существуют такие слова w', w'', w''' и такие натуральные числа $n', n''' < n$ и $n'' \leq n$, что $w = w'w''w'''$, $v \Rightarrow^{n'} w'$, $v' \Rightarrow^{n'''} w'''$, $R_i \Rightarrow^{n''} w''$. Очевидно, что $u \Rightarrow^* vR_iv'$ в новой грамматике. Покажем только, что $R_i \Rightarrow^* w''$ и в грамматике после замены (остальные два вывода преобразуются по индукции).

Итак, рассмотрим левый вывод длины $n'' \leq n$ слова w'' из R_i в старой грамматике (так как исходный вывод был левым, то очевидно, что существует левый вывод w'' из R_i указанной длины). Очевидно, что для некоторого $t \geq 0$ и для некоторых $1 \leq j_1, j_2, \ldots, j_t \leq m$; $1 \leq s \leq l$, этот вывод имеет вид

$$R_i \Rightarrow R_i\alpha_{j_1} \Rightarrow R_i\alpha_{j_2}\alpha_{j_1} \Rightarrow \cdots \Rightarrow R_i\alpha_{j_t}\alpha_{j_{t-1}} \ldots \alpha_{j_1} \Rightarrow$$

$$\beta_s\alpha_{j_t}\alpha_{j_{t-1}} \ldots \alpha_{j_1} \Rightarrow^r w'',$$

где $r = n'' - t - 1 < n$. Значит, по индукции, в новой грамматике

$$\beta_s\alpha_{j_t}\alpha_{j_{t-1}} \ldots \alpha_{j_1} \Rightarrow^* w''.$$

Осталось показать, что

$$R_i \Rightarrow^* \beta_s\alpha_{j_t}\alpha_{j_{t-1}} \ldots \alpha_{j_1}$$

в новой грамматике. Но действительно, в новой грамматике имеет место вывод:

$$R_i \Rightarrow \beta_sR'_i \Rightarrow \beta_s\alpha_{j_t}R'_i \Rightarrow \beta_s\alpha_{j_t}\alpha_{j_{t-1}}R'_i \Rightarrow \cdots \Rightarrow \beta_s\alpha_{j_t}\alpha_{j_{t-1}} \ldots \alpha_{j_1}.$$

Итак, мы доказали, что наша замена не уменьшает язык грамматики. Аналогично можно показать, что эта замена и не увеличивает языка грамматики. Здесь надо рассматривать правые выводы в новой грамматике, преобразуя их в выводы в старой грамматике.

Ясно также, что условия (1), (2), (4) гипотезы индукции теперь выполнены, ибо R'_i ранее в правилах не встречался. Гипотеза (5) может нарушаться, если среди $\alpha_1, \alpha_2, \ldots, \alpha_m$ есть хотя бы один нетерминал (появятся цепные правила). Отметим, что по (5) индукционного предположения все $\alpha_1, \alpha_2, \ldots, \alpha_m$ непусты. Применим стандартную процедуру избавления от цепных правил. Отметим, что при этом условия (1), (2), (4) не нарушатся. Итак, теперь условия (1), (2), (4), (5) гипотезы выполнены для P_i.

Заметим, что $\alpha_1, \ldots, \alpha_m$, $\beta_1, \ldots, \beta_l$ все непусты, по (5) индукционного предположения для P_{i-1}. Значит условие (3) может теперь нарушаться единственным способом: правилами вида $R_j \to R_i\alpha$ для $j > i$. Каждое такое правило в P_i заменим на правила:

$$R_j \to \beta_1\alpha, \quad R_j \to \beta_2\alpha, \ldots, R_j \to \beta_l\alpha,$$

$$R_j \to \beta_1 R_i'\alpha, \quad R_j \to \beta_2 R_i'\alpha, \ldots, R_j \to \beta_l R_i'\alpha.$$

Ясно, что это преобразование сохранило эквивалентность грамматик. Ясно также, что теперь все пункты (1) — (5) выполнены для P_i. Напомним, что, во-первых, все $\beta_1, \ldots, \beta_l$ непусты, а, во-вторых, они начинаются либо с терминалов, либо с таких R_k, что $k > i$. ∎

Определение 2.3.1 (нормальной формы Грейбах). *КС-грамматика $G = \langle A, Q, S, P\rangle$ называется грамматикой в нормальной форме Грейбах, если G есть грамматика без e-правил, и каждое правило из P имеет один из следующих видов:*

1. $S \to e$,

2. $R \to x\alpha$, где $x \in A$, $R \in Q$ и $\alpha \in Q^*$.

Теорема 2.13. *Всякая КС-грамматика $G = \langle A, Q, S, P\rangle$ может быть приведена к эквивалентной ей грамматике $G'' = \langle A, Q'', S'', P''\rangle$ в нормальной форме Грейбах.*

Доказательство. Воспользуемся построениями теоремы 2.12 и преобразуем грамматику $G = \langle A, Q, S, P\rangle$ к грамматике $G' = \langle A, Q', S, P'\rangle$ со свойствами (1) — (5) как в доказательстве теоремы 2.12 ($P' = P_n$). В частности, согласно (3) из доказательства теоремы 2.12 грамматика $G' = \langle A, Q', S, P'\rangle$ отличается тем свойством, что для произвольного нетерминала R_j, где $1 \leq j \leq n$, и для произвольного правила $R_j \to \alpha$, либо α начинается с терминала, либо с такого R_k, что $k > j$, либо же это правило имеет вид $S \to e$. (Напомним, что множество Q' состоит из нетерминалов $R_1, \ldots, R_n, R_1', \ldots, R_n'$).

В частности, правые части всех правил вида $R_n \to \alpha$ начинаются с терминала.

Построим последовательность множеств $P_1', P_2', \ldots, P_n'$ так:

$P_1' = P'$. Если P_i' уже построено, то найдем максимальное j, $1 \leq j \leq n$, такое, что в P_i' есть хотя бы одно правило вида $R_j \to \alpha$, в котором α начинается с нетерминала.

Если такого j не существует, положим $P^* = P_i'$, и прекратим построение множеств P_k'.

Иначе рассмотрим, для этого, все правила с левой частью R_j и правой частью, начинающейся с нетерминала

$$R_j \to R_{k_1}\alpha_1, R_j \to R_{k_2}\alpha_2, \ldots, R_j \to R_{k_r}\alpha_r. \tag{3}$$

Для каждого такого правила

$$R_j \to R_{k_v}\alpha_v,$$

где $1 \leq v \leq r$, включим в P'_{i+1} правила

$$R_j \to \beta_{1,k_v}\alpha_v, \ldots, R_j \to \beta_{l,k_v}\alpha_v,$$

где

$$R_{k_v} \to \beta_{1,k_v}, \ldots, R_{k_v} \to \beta_{l,k_v}$$

есть все правила с левой частью R_{k_v} (по индукции, все

$$\beta_{1,k_v}, \ldots, \beta_{l,k_v}$$

уже начинаются с терминалов).

Теперь включим в P'_{i+1} все правила из P'_i, кроме правил (3).

Пусть $P^* = P'_n$. Множество P^* будет, очевидно, построено не более чем за n шагов. Ясно, что

1) грамматики $\langle A, Q', S, P'\rangle$ и $\langle A, Q', S, P^*\rangle$ эквивалентны;

2) в P^* правая часть произвольного правила вида $R_j \to \alpha$ (где $j = 1, 2, \ldots, n$) либо начинается с терминала, или само это правило имеет вид $S \to e$.

Рассмотрим теперь в P^* правила с левой частью R'_j для $j = 1, 2, \ldots, n$. Так как мы их не меняли в последнем построении, то они те же, что и в P'. Тогда, по построению P', для всякого R'_j и всякого правила $R'_j \to \alpha$, это α либо начинается с терминала, либо с одного из R_k. Заменим в P^* каждое правило $R'_j \to R_k\alpha$ на правила

$$R'_j \to \beta_1\alpha, \ldots, R' \to \beta_r\alpha,$$

где

$$R_k \to \beta_1, \ldots, R_k \to \beta_r$$

есть все правила с левой частью R_k в P^*.

В полученном таким образом множестве P^{**} правая часть любого правила либо начинается с терминала, либо это есть правило $S \to e$. Ясно также, что грамматики $\langle A, Q', S, P'\rangle$ и $\langle A, Q', S, P^{**}\rangle$ эквивалентны.

Рассмотрим в качестве Q^* такое множество $\{T_a \mid a \in A\}$, что

$$Q' \bigcap \{T_a \mid a \in A\} = \emptyset.$$

Пусть

$$Q'' = Q' \bigcup Q^*.$$

Заменим в каждом правиле $R \to x\alpha \in P^{**}$, где $x \in A$, каждый терминал a в α на T_a. Получим множество правил P^{***}. В качестве P'' рассмотрим множество правил

$$P^{***} \bigcup \{T_a \to a \mid a \in A\}.$$

Ясно, что:

1) грамматики $\langle A, Q, S, P\rangle$, и $\langle A, Q'', S, P''\rangle$ эквивалентны;

2) грамматика $\langle A, Q'', S, P''\rangle$ находится в нормальной форме Грейбах. ∎

Упражнения

Упражнение 2.3.1 (*). *Будем говорить, что КС-грамматика $G = \langle A, Q, S, P\rangle$ находится в сильной нормальной форме Грейбах, если*

1) G находится в нормальной форме Грейбах,

2) длина правой части любого правила не превосходит 3.

Покажите, что произвольная КС-грамматика может быть приведена к эквивалентной КС-грамматике в сильной нормальной форме Грейбах.

Упражнение 2.3.2. *Приведите следующую грамматику в нормальную форму Грейбах: $G = \langle\{a, b\}, \{S, S'\}, S, P\rangle$, где P — следующее множество правил:*

$$S \to Sa,\ S \to b,\ S \to SS',\ S \to S'S,\ S \to SS'abS,\ S' \to S.$$

Упражнение 2.3.3. *Приведите к нормальной форме Грейбах грамматики из упражнения 2.2.2.*

2.4 Индекс грамматики и языка

В этом параграфе мы более подробно остановимся на специальных свойствах КС-грамматик, нужных для технических целей.

Рассмотрим некоторую КС-грамматику $G = \langle A, Q, S, P\rangle$. Символ $R \in Q$ называется *самовставимым* (в G), если для некоторых $x, y \in A^*$, где $|xy| \neq 0$, имеет место $R \Rightarrow_G^* xRy$. Если в Q есть хотя бы один самовставимый символ, мы будем называть G *грамматикой с самовставлением.*

Если добавить в определение самовставимого символа требование, что одновременно и x, и y непусты, то есть что $x, y \in A^+$, получим определение символа, самовставимого по Хомскому (или *сильно самовставимого*). Соответственно, грамматику G будем называть *грамматикой с самовставлением по Хомскому* (или с *сильным самовставлением*), если в Q есть хотя бы один самовставимый по Хомскому символ.

Два нетерминала $R, R' \in Q$ будем называть *сводимыми*, если существуют $x, y, x', y' \in A^*$, что $R \Rightarrow_G^* xR'y$ и $R' \Rightarrow_G^* x'Ry'$.

Ясно, что если $R, R', R'' \in Q$, а R и R' сводимы и R и R'' сводимы, то R' и R'' сводимы. Таким образом, отношение сводимости является отношением эквивалентности на Q. Класс эквивалентности, содержащий R, будем обозначать через $D(R)$. Нетерминалы R и R' будем называть *вставимыми*, если они сводимы и самовставимы.

Определение 2.4.1 (высота символа). *Пусть $R \in Q$.*

1). Высота R есть 1 (в записи $h(R) = 1$) тогда и только тогда, когда для произвольного $R' \in Q$, из того, что существуют такие $x, y \in (A \bigcup Q)^*$, что $R \to xR'y \in P$, следует, что R и R' сводимы.

2). Пусть условие 1) настоящего определения не выполнено для R. Рассмотрим множество

$$U = \{R' \mid R' \in (Q \backslash D(R));\ R'' \in D(R);\ x, y \in (A \bigcup Q)^*;\ R'' \to xR'y \in P\}.$$

Положим

$$h(R) = \max\{h(R') + 1 \mid R' \in U\}.$$

Лемма 2.14. *Определение 2.4.1 корректно определяет высоту всякого $R \in Q$. Высота произвольного символа $R \in Q$ не превосходит $|Q|$.*

Доказательство. Фактически в определении прямо описан способ вычисления высоты символа. При этом сначала высота 1 будет приписана всем символам, из которых за один шаг достижимы лишь сводимые с ними символы. Затем на каждом шаге будет приписываться высота символам, из которых за один шаг достижимы лишь сводимые с ними символы и символы, которым уже приписана высота на предыдущих шагах. ∎

Высотой рассматриваемой грамматики G, в записи $h(G)$, будем называть $\max\{h(R) \mid R \in Q\}$. *Высотой* КС-языка L будем называть минимальную высоту КС-грамматики, задающей этот язык L.

Заметим, что два нетерминала одинаковой высоты не обязаны быть сводимыми.

Теорема 2.15. *Пусть $G = \langle A, Q, S, P\rangle$ — КС-грамматика без самовставления. Тогда $L(G)$ конечен.*

Доказательство. Если $L(G) = \emptyset$, то теорема доказана.

Без ограничения общности можно считать, что G — приведенная. Докажем индукцией по высоте G, что $L(G)$ конечен.

Базис. Пусть $h(G) = 1$. Если $L(G) = \{e\}$, где e — пустое слово, то теорема доказана. Так как G — приведенная грамматика без самовставления, то ясно, что для произвольного $R \in Q$ и произвольного правила $R \to \alpha \in P$ либо $\alpha \in Q$, либо $\alpha \in A^*$. Заметим, что так как $h(G) = 1$, то если для $R, R' \in Q$ правило $R \to R' \in P$, то R и R' сводимы, то есть $R' \in D(R)$. Ясно, что из приведенности G следует, что $S \in D(R)$.

Рассмотрим в качестве P^* множество правил, в которых всякое вхождение всякого нетерминала $R \in Q$ заменено на S, после чего правило $S \to S$ выброшено, если оно было. Ясно, что в грамматике $G^* = \langle A, \{S\}, S, P^*\rangle$, эквивалентной G, правая часть всякого правила пуста или состоит из терминалов. Следовательно $L(G) = L(G^*)$ — конечный язык.

Шаг индукции. Пусть теорема доказана для меньших высот, и $h(G) = i + 1$. Пусть $R_1, \dots, R_k \in Q$ — все нетерминалы высоты 1. Рассмотрим грамматики

$$G_1 = \langle A, Q, R_1, P_1\rangle, \dots, G_k = \langle A, Q, R_k, P_k\rangle,$$

где P_i для $i = 1, \dots, k$ — множество правил из P с левыми частями $R' \in D(R_i)$. Ясно, что для произвольного $x \in (A \bigcup Q)^*$ и для произвольного $i = 1, \dots, k$ из $R_i \Rightarrow^*_G x$ следует $R_i \Rightarrow^*_{G_i} x$, и наоборот. Так как грамматики $G_1, \dots, G_k$ имеют высоту 1 то, по индукции, $L(G_1), \dots, L(G_k)$ являются конечными языками.

Рассмотрим грамматику

$$G' = \langle A, (Q \setminus \{R_1, \dots, R_k\}), S, P'\rangle,$$

где P' получено выбрасыванием из P всех правил, содержащих или R_1, или R_2, или $\dots$, или R_k, и добавлением, для каждого правила

$$R \to \alpha_1 R_{i_1} \alpha_2 \dots \alpha_m R_{i_m} \alpha_{m+1} \in P,$$

где $R \in (Q \setminus \{R_1, \dots, R_k\})$ и

$$\alpha_1, \alpha_2, \dots, \alpha_m, \alpha_{m+1} \in ((Q \setminus \{R_1, \dots, R_k\}) \bigcup A)^*,$$

всевозможных правил вида

$$R \to \alpha_1 x_1 \alpha_2 \dots \alpha_m x_m \alpha_{m+1},$$

где

$$x_1 \in L(G_{i_1}), \dots, x_m \in L(G_{i_m}).$$

Имеют место следующие утверждения.

1. G и G' эквивалентны.
2. $h(G') = h(G) - 1$.
3. G' — грамматика без самовставления.

По индукции, язык $L(G')$ конечен. Значит, конечен и $L(G)$. ■

КС-грамматика $G = \langle A, Q, S, P \rangle$ называется сжатой, если все символы $R \in (Q \setminus \{S\})$ самовставимы.

Теорема 2.16. *Всякая такая КС-грамматика $G = \langle A, Q, S, P \rangle$, что $L(G) \neq \emptyset$, может быть преобразована в эквивалентную сжатую КС-грамматику $G' = \langle A, Q', S, P' \rangle$.*

Доказательство. Так как $L(G) \neq \emptyset$, то можно без ограничения общности считать, что G — приведенная грамматика без e-правил.

Пусть $(Q \setminus \{S\})$ содержит несамовставимые символы. Пусть тогда $R \in (Q \setminus \{S\})$ — несамовставимый символ минимальной высоты, то есть для всякого $R' \in Q$ такого, что $R \to xR'y \in P$ для некоторых $x, y \in (A \bigcup Q)^*$, либо $h(R) = h(R')$, либо R' — самовставимый символ.

Ясно, что из приведенности G и несамовставимости R следует, что для всех $x, y \in (A \bigcup Q)^*$ и для всех $R', R'' \in D(R)$ из того, что $R' \to xR''y \in P$ следует, что $x = y = e$. Заменим во всех правилах из P все вхождения всех нетерминалов из $D(R)$ на R, и, выбросив правило $R \to R$, получим новое множество правил P_1.

Рассмотрим множество $Q_1 = ((Q \setminus D(R)) \bigcup \{R\})$. Ясно, что:

1. Грамматика $G_1 = \langle A, Q_1, S, P_1 \rangle$ приведена.
2. G_1 эквивалентна $G = \langle A, Q, S, P \rangle$.
3. $R \in Q_1$ — несамовставимый символ.
4. Если для некоторых $R' \in Q_1$ и $x, y \in (A \bigcup Q_1)^*$ правило $R \to xR'y \in P_1$, то $h(R') < h(R)$.
5. Если $R' \in Q_1$ несамовставим в G_1, то R' несамовставим в G.

Пусть $P_2 = P_1 \setminus P(R)$, где

$$P(R) = \{R \to \alpha_1, R \to \alpha_2, \dots, R \to \alpha_m\}$$

есть все правила с левой частью R из P_1, а

$$\alpha_1, \ldots, \alpha_m \in (A \bigcup Q_1)^*.$$

Для всякого правила

$$R' \to \beta_1 R \beta_2 R \ldots \beta_l R \beta_{l+1} \in P_2,$$

где $l > 0$, а

$$\beta_1, \ldots, \beta_{l+1} \in ((Q_1 \setminus \{R\}) \bigcup A)^*,$$

выбросим его из P_2 и включим в P_2 всевозможные правила вида

$$R' \to \beta_1 x_1 \beta_2 x_2 \ldots \beta_l x_l \beta_{l+1},$$

где

$$x_1, \ldots, x_l \in \{\alpha_1, \ldots, \alpha_m\}.$$

Рассмотрим грамматику

$$G_2 = \langle A, (Q_1 \setminus \{R\}), S, P_2 \rangle.$$

Ясно, что G приведена и эквивалентна G, но число несамовставимых нетерминалов у нее меньше, чем у G. Применив этот процесс не более $|Q|$ раз, получим искомую сжатую грамматику. ∎

В 2.1 было дано определение вывода в грамматике.

Определение 2.4.2 (индекса грамматики). *1. Индексом вывода*

$$s = \langle w_0, w_1, \ldots, w_n = w \rangle$$

слова $w \in L(G)$ в КС-грамматике $G = \langle A, Q, S, P \rangle$ будем называть максимум по $i = 0, \ldots, n-1$ числа вхождений нетерминальных символов в w_i.

2. Индексом слова $w \in L(G)$ в G будем называть минимальный индекс вывода w в G.

3. Индексом КС-грамматики $G = \langle A, Q, S, P \rangle$ будем называть максимальный индекс в G слова w среди $w \in L(G)$. Если индекс слов $w \in L(G)$ неограничен, то говорим, что индекс G бесконечен.

4. Индексом КС-языка L будем называть минимальный индекс задающей его КС-грамматики. Итак, индекс КС-языка вправе быть бесконечным.

Чтобы показать, что действительно существуют КС-грамматики и КС-языки бесконечного индекса, введем в рассмотрение так называемые камешковые игры на графах.

Определение 2.4.3 (графа). *Ориентированным графом будем называть произвольное множество γ вместе с определенным на γ двухместным предикатом $C \subseteq \gamma^2$ (то есть вместе с некоторым множеством упорядоченных пар элементов G). Элементы множества γ называются вершинами графа, а элементы множества C — ребрами или дугами.*

Вершина $a \in \gamma$ называется предшественником $b \in \gamma$, если дуга $\langle a, b\rangle$ принадлежит C.

Опишем сначала игру в n камешков на графе $\langle \gamma, C\rangle$ неформально. Игрок имеет n камешков, которые он может класть на вершины графа, все предшественники которых уже накрыты камешками. Игрок также может снять камешек с любой вершины. Всякое такое действие называется ходом. Все вершины, которые игроку удается накрыть камешками, называются выигранными в n-игре, или n-вершинами.

Приведем формальное определение.

Определение 2.4.4 (позиции в игре). *Пусть дан граф $\langle \gamma, C\rangle$ и натуральное число n. Позицией в n-игре или n-позицией называется произвольное такое конечное множество $p \subseteq \gamma$, что $|p| \le n$. Определим отношение достижимости на n-позициях. Пусть p и q — n-позиции. Будем говорить, что позиция q следует из позиции p в n-игре, и писать $p \Rightarrow_n q$, если существует $a \in (p \bigcup q)$, что имеет место одно из двух:*

1. $p = (q \bigcup \{a\})$;

2. $q = (p \bigcup \{a\})$, *причем все предшественники a содержатся в p.*

Будем говорить, что позиция q достижима из позиции p в n-игре, и писать $p \Rightarrow_n^* q$, если существует такая конечная последовательность n-позиций

$$q_0, q_1, \ldots, q_m,$$

что $q_0 = p$, $q_m = q$ и $q_i \Rightarrow_n q_{i+1}$ для всякого $i = 0, 1, \ldots m-1$. Произвольная такая конечная последовательность n-позиций $q_1, \ldots, q_m$, что $q_i \Rightarrow_n q_{i+1}$ для всякого $i = 1, \ldots m-1$, называется n-трассой.

$a \in \gamma$ называется n-вершиной, если $\emptyset \Rightarrow_n^* \{a\}$.

Рассмотрим теперь множество всех натуральных чисел N и такую функцию $f : N^2 \to N$, что

$$f(m,n) = \begin{cases} n+1, & \text{если } [n/2] = m \\ 0, & \text{в остальных случаях.} \end{cases}$$

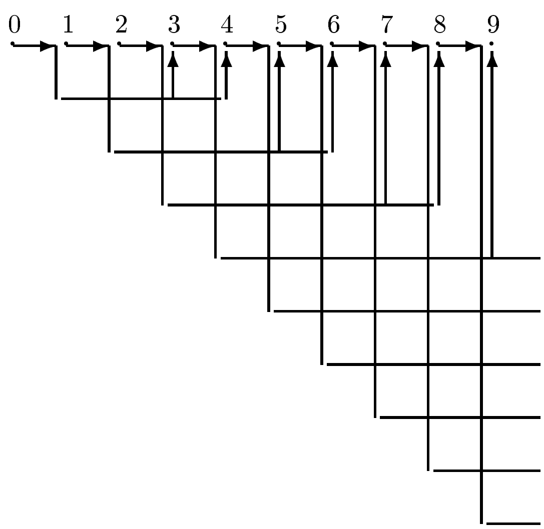
Напомним, что $[n/2]$ — деление n нацело на 2.

Рассмотрим граф

$$\Delta = \langle N, (\{\langle i, i+1\rangle \mid \ i \in N\} \bigcup \{\langle i, j+1\rangle \mid \ i \in N,\ j \in N,\ f(i,j) \neq 0\})\rangle.$$

Δ называется графом Кфури.

Приведем схематическое изображение начального отрезка этого графа:



Лемма 2.17. *Для произвольного натурального n множество n-вершин графа Кфури Δ конечно.*

Доказательство. Доказательство леммы 2.17 непосредственно следует из леммы 2.18: любая n-точка в Δ содержится среди $0, 1, \ldots, \rho(n)$, где ρ — функция из леммы 2.18. ∎

Пусть p — произвольная n-позиция графа Кфури Δ для некоторого n, а $m \in N$. Множество $E(p,m)$ всех таких натуральных чисел i, что либо $i \leq m$, либо существует $j \in p$, что $0 \leq i - j \leq m$, называется m-окрестностью p.

Лемма 2.18. *Существует функция $\rho : N \to N$, что если p,q — n-позиции в графе Кфури Δ и $p \Rightarrow_n^* q$, то $q \subseteq E(p, \rho(n))$.*

Доказательство. Пусть $\rho : N \to N$ определена индуктивно следующим образом:

$\rho(1) = 1$;

$\rho(k+1) = 4(k+1)(\rho(k)+1)+1$.

Утверждается, что эта функция удовлетворяет условиям леммы 2.18.

Действительно, рассмотрим произвольную n-трассу, начинающуюся с p и заканчивающуюся на q.

Докажем утверждение леммы индукцией по n.

При $n = 1$ утверждение очевидно, ибо существует в точности два корректных хода: удалить единственный камешек, или же положить его на 0.

Допустим, что лемма доказана для меньших чисел. Предположим, что лемма неверна для n. Значит, существуют такие позиция p, содержащая не более n натуральных чисел, и n-трасса t, начинающаяся с p, что в t в некоторой позиции (без потери общности можно считать, что в последней) камешек кладется на некоторое натуральное число $k \notin E(p, \rho(n))$. Зафиксируем это k. Итак, для всякого $j \in p$, либо $j > k$, либо же $k - j > \rho(n)$. В частности, существует $j \in (p \bigcup \{0\})$ такое, что $j < k, k-j > \rho(n)$, и множество $\{j+1, j+2, \ldots, k\}$ не содержит элементов из p. Зафиксируем это j.

По определению игры в камешки и по определению Δ, чтобы достичь k, следует последовательно пройти в позициях все $\{j+1, j+2, \ldots, k\}$.

Пусть $p_0 = p$, а p_l — позиция, стоящая на l-ом элементе трассы t (считая с 0). Пусть $t_0 = t$, а t_l — это t без первых l элементов. Таким образом, t_l есть n-трасса, начинающаяся с p_l, для $l < len(t)$ (где $len(t)$ обозначает длину t). Пусть $m =](k-j-1)/2[+j+1$, и $F = \{m, m+1, \ldots, k\}$, где, для рационального числа a, $]a[$ обозначает наименьшее целое, большее или равное a. Другими словами, F содержит вторую половину последовательности $\{j+1, j+2, \ldots, k\}$. Из определений следует, что F обладает следующими свойствами:

- никакой элемент p не принадлежит F;
- $[k/2] < m$.

Легко видеть, что существует натуральное число $l < len(t)$ со следующими свойствами:

- для всякого такого i, что $l \leq i < len(t)$, p_i содержит элемент из F;
- p_{l-1} не содержит никакого элемента F;
- для всякого $j \in F$ существует i, что $l \leq i < len(t)$, и p_i содержит j.

Зафиксируем это l. В соответствии с определениями, чтобы достичь

$$m, m+1, \ldots, k,$$

следует иметь камешки на

$$[(m-1)/2], [m/2], \ldots, [(k-1]/2]$$

соответственно. Пусть

$$F' = \{[(m-1)/2], [m/2], \ldots, [(k-1)/2]\}.$$

Значит, для всякого $j \in F'$ существует такое i, что $l \leq i < len(t)$ и p_i содержит j.

Сформулируем очевидное замечание.

Лемма 2.19. *Пусть p' — некоторая позиция в n-игре на Δ, $k \in p'$ — натуральное число, $m < k$ и некоторая n-трасса t', начинающаяся с p', обладает следующими свойствами:*

- *в любой позиции q трассы t' существует натуральное число, большее или равное k;*
- *некоторая позиция q трассы t' содержит m.*

Тогда существует такая $(n-1)$-трасса t'', начинающаяся с позиции $p' \setminus \{j \mid \ j \geq k\}$, что m содержится в некоторой позиции q' трассы t''.

Доказательство. Надо просто взять n-трассу t' и преобразовывать ее в $(n-1)$-трассу t'' так. В каждом элементе трассы выкинуть все камешки, лежащие на числах, больших или равных k. Если очередной элемент трассы t' получен путем снятия некоторого камешка, и этот камешек лежит и в соответствующей позиции трассы t'', то также снимем его. Если очередной элемент трассы t' получен путем помещения некоторого камешка на вершину $l < k$, то также положим в трассе t'' камешек на эту вершину. Так как все предшественники этой вершины в n-позиции были накрыты камешками, то и в $(n-1)$-позиции они накрыты камешками, ибо на числах меньших k трассы по построению совпадают.

Наконец, если очередной элемент трассы t' получен путем помещения некоторого камешка на вершину $l \geq k$ или снятия его с такой вершины, то в трассе t'' этот ход пропустим. ∎

Так как всякое $j \in F'$ меньше, чем m, мы в состоянии применить замечание, сформулированное выше. Итак, существует $(n-1)$-трасса t', начинающаяся с позиции $p' = p_l \setminus \{j \mid j \geq k\}$, такая что, в противоречие с индукционным предположением, некоторая позиция q' этой трассы такова, что q' не содержится в $E(p', \rho(n-1))$. Последнее вытекает из того, что множество F' по мощности больше, чем $E(p', \rho(n-1))$. Значит, F' содержит элемент не из $E(p', \rho(n-1))$. И действительно,

$$|E(p', \rho(n-1))| \leq n(\rho(n-1)+1),$$

тогда как

$$|F'| = [|F|/2] \geq (k-j)/4 > \rho(n)/4 > n(\rho(n-1)+1).$$

Этим противоречием доказана лемма. ∎

Определение 2.4.5 (гомоморфизма графов). *Пусть* $\langle \gamma_1, C_1 \rangle$ *и* $\langle \gamma_2, C_2 \rangle$ *— два графа,* $h : \gamma_1 \to \gamma_2$ *— произвольное отображение.* h *называется гомоморфизмом, если для произвольных* $a, b \in \gamma_1$*, если* $\langle a, b \rangle \in C_1$*, то* $\langle h(a), h(b) \rangle \in C_2$*. Гомоморфизм называется разнозначным, если для произвольных* $a, b \in \gamma_1$*, если* $a \neq b$*, то* $h(a) \neq h(b)$*.*

Рассмотрим теперь контекстно-свободную грамматику

$$G = \langle \{(,), a\}, \{S\}, S, P \rangle,$$

где P — следующее множество правил:

$$S \to (SS), S \to a.$$

Рассмотрим граф $\Theta = \langle L(G), \theta \rangle$, где θ состоит из таких пар слов $\langle u, v \rangle$, что для некоторого w либо v есть (uw), либо v есть (wu).

Лемма 2.20. *Существует разнозначный гомоморфизм графов* Δ *и* Θ*, то есть разнозначный гомоморфизм* $h : N \to L(G)$*.*

Доказательство. Определим отображение h так:
$h(0) = a$;
$h(k+1) = (h([k/2])h(k))$.

В частности,

$$h(1) = (aa), h(2) = (a(aa)), h(3) = ((aa)(a(aa))), \dots$$

Прямо по определению устанавливается, что h — разнозначный гомоморфизм графов Δ и Θ. ∎

Лемма 2.21. *Пусть $\langle\gamma_1, C_1\rangle$ и $\langle\gamma_2, C_2\rangle$ — графы, $h : \gamma_2 \to \gamma_1$ — разнозначный гомоморфизм, n — натуральное число. Если $a \in \gamma_2$ и $h(a)$ — n-вершина в графе $\langle\gamma_1, C_1\rangle$, то a — n-вершина в графе $\langle\gamma_2, C_2\rangle$.*

Доказательство. Пусть $p \subseteq \gamma_1$. Через $h^{-1}(p)$ обозначим следующее подмножество множества γ_2:

$$h^{-1}(p) = \{\epsilon \in \gamma_2 \mid \ h(\epsilon) \in p\}.$$

Заметим, что если $p \subseteq \gamma_1$ — n-позиция в графе $\langle\gamma_1, C_1\rangle$, то $h^{-1}(p)$ — n-позиция в графе $\langle\gamma_2, C_2\rangle$, то есть $|h^{-1}(p)| \le n$. Действительно, пусть $|h^{-1}(p)| > n$. Тогда, в силу разнозначности,

$$|\{h(\epsilon) \mid \epsilon \in h^{-1}(p)\}| = |h^{-1}(p)| > n.$$

Но, по построению,

$$\{h(\epsilon) \mid \epsilon \in h^{-1}(p)\} \subseteq p,$$

то есть $|p| > n$, что противоречит тому, что p — n-позиция.

Далее, пусть p, q — n-позиции в графе $\langle\gamma_1, C_1\rangle$, $p \Rightarrow_n q$. Покажем, что тогда либо $h^{-1}(p) \Rightarrow_n h^{-1}(q)$, либо $h^{-1}(p) = h^{-1}(q)$. Действительно, пусть $h^{-1}(p) \neq h^{-1}(q)$. Доказательство распадается на два случая. Первый — когда $p \subseteq q$, второй — когда $q \subseteq p$. По определению, один из этих двух случаев имеет место.

Первый случай. Итак, $p \subseteq q$. Тогда, по определению, существует $\epsilon \in \gamma_1$ такой, что $q = (p \bigcup \{\epsilon\})$, то есть q получена из p помещением камешка на элемент ϵ. Тогда, по определению,

$$r = \{\theta \in \gamma_1 \mid \ \langle\theta, \epsilon\rangle \in C_1\} \subseteq p.$$

Поскольку $h^{-1}(p) \neq h^{-1}(q)$, то существует (в силу разнозначности h — единственный) $\eta \in \gamma_2$ такой, что $h(\eta) = \epsilon$. Итак, $h^{-1}(q) = (h^{-1}(p) \bigcup \{\eta\})$. По замеченному ранее, $h^{-1}(q)$ и $h^{-1}(p)$ — n-позиции. Чтобы доказать, что $h^{-1}(p) \Rightarrow_n h^{-1}(q)$, осталось заметить, что

$$r' = \{\tau \in \gamma_2 \mid \langle\tau, \eta\rangle \in C_2\} \subseteq h^{-1}(p).$$

Но это следует из того, что h — гомоморфизм. Действительно,

$$\langle h(\tau), h(\eta)\rangle \in C_1$$

для всякого $\tau \in r'$, значит, $h(\tau) \in r$, так как $h(\eta) = \epsilon$, но тогда $\tau \in h^{-1}(p)$.

Второй случай. Итак, $q \subseteq p$. Тогда, по определению, существует $\epsilon \in \gamma_1$ такой, что $p = q \bigcup \{\epsilon\}$, то есть q получена из p снятием камешка с элемента ϵ. . Поскольку $h^{-1}(p) \neq h^{-1}(q)$, то существует (в силу разнозначности h — единственный) $\eta \in \gamma_2$ такой, что $h(\eta) = \epsilon$. Итак, $h^{-1}(p) = (h^{-1}(q) \bigcup \{\eta\})$. Так как ранее замечено, что $h^{-1}(q)$ и $h^{-1}(p)$ — n-позиции, то, по определению, $h^{-1}(p) \Rightarrow_n h^{-1}(q)$.

Теперь применением индукции легко показать, что если p, q — n-позиции в графе $\langle\gamma_1, C_1\rangle$ и $p \Rightarrow_n^* q$, то либо $h^{-1}(p) \Rightarrow_n^* h^{-1}(q)$, либо $h^{-1}(p) = h^{-1}(q)$.

Допустим теперь, что $a \in \gamma_2$, $h(a)$ — n-вершина в графе $\langle\gamma_1, C_1\rangle$. Тогда $\emptyset \Rightarrow_n^* \{h(a)\}$. Так как h — гомоморфизм, то $h^{-1}(\emptyset) = \emptyset$. С другой стороны, $h^{-1}(\{h(a)\}) = \{a\}$. По доказанному выше, тогда $\emptyset \Rightarrow_n^* \{a\}$. Значит, a — n-вершина в графе $\langle\gamma_1, C_1\rangle$. ∎

Лемма 2.22. *Для произвольного $n \in N$ существует $u \in L(G)$, не являющееся n-вершиной в графе Θ.*

Доказательство. Следствие лемм 2.17, 2.21, 2.20. ∎

Теорема 2.23. *Приведенная выше КС-грамматика G имеет бесконечный индекс.*

Доказательство. Следствие лемм 2.21 и 2.24. ∎

Определение 2.4.6 (трассы). *Пусть $G = \langle A, Q, S, P\rangle$ — КС-грамматика. Трассой в G будем называть конечную последовательность троек*

$$\langle w_1, i_1, p_{j_1}\rangle, \ldots, \langle w_m, i_m, p_{j_m}\rangle,$$

где $w_k \in (A \bigcup Q)^*$, i_k — натуральное число, не превосходящее числа вхождений нетерминалов в w_k, и $p_{j_k} \in P$, причем выполнены условия:

1. $w_1 = S$.

2. Пусть $k < m$ и

$$w_k = \alpha_1 R_{l_1} \alpha_2 \ldots \alpha_s R_{l_s} \alpha_{s+1},$$

где

$$\alpha_1, \ldots, \alpha_{s+1} \in A^*,$$

$$R_{l_1}, \dots R_{l_s} \in Q.$$

Тогда p_{j_k} *имеет вид* $R_{l_{i_k}} \to \beta$, *где* $\beta \in (A \bigcup Q)^*$ *и*

$$w_{k+1} = \alpha_1 R_{l_1} \alpha_2 \dots \alpha_{i_k} \beta \alpha_{i_k+1} R_{l_{i_k+1}} \dots R_{l_s} \alpha_{s+1}.$$

Другими словами, трасса состоит из троек, первый элемент которых — слово, полученное к этому шагу вывода, второй — порядковый номер слева нетерминала в этом слове, который будет заменяться на этом шаге вывода, и третий — правило, которое будет использовано на этом шаге.

Ясно, что всякому выводу

$$S = w_0, w_1, w_2, \dots, w_n \in A^*$$

в грамматике G можно, вообще говоря, неоднозначно, сопоставить трассу

$$\langle w_0, 1, p_{j_0} \rangle, \langle w_1, i_1, p_{j_1} \rangle, \dots, \langle w_n, 0, p_{j_n} \rangle$$

в этой грамматике.

Приведем пример.

Рассмотрим КС-грамматику

$$G = \langle \{0\}, \{S\}, S, \{S \to 0, S \to SS, S \to e\} \rangle.$$

В ней выводимо слово 00. Рассмотрим следующий вывод слова 00 в G:

$$S, SS, SSS, SS, 0S, 00.$$

Приведем примеры двух трасс в G, соответствующих данному выводу слова 00 в G.

Первая трасса.

$$\langle S, 1, S \to SS \rangle, \langle SS, 1, S \to SS \rangle, \langle SSS, 1, S \to e \rangle,$$

$$\langle SS, 1, S \to 0 \rangle, \langle 0S, 1, S \to 0 \rangle.$$

Вторая трасса.

$$\langle S, 1, S \to SS \rangle, \langle SS, 2, S \to SS \rangle, \langle SSS, 3, S \to e \rangle,$$

$$\langle SS, 1, S \to 0 \rangle, \langle 0S, 1, S \to 0 \rangle.$$

Лемма 2.24. *Рассмотрим грамматику*

$$G = \langle \{(,),a\}, \{S\}, S, P\rangle,$$

где P — следующее множество правил:

$$S \to (SS), \quad S \to a.$$

Рассмотрим граф $\Theta = \langle L(G), \theta\rangle$, где θ состоит из таких пар слов $\langle u, v\rangle$, что для некоторого w либо v есть (uw), либо v есть (wu).

Если w — слово индекса n в G, то в графе Θ w является $(n+1)$-вершиной.

Доказательство. Рассмотрим произвольный вывод индекса n слова w и его произвольную трассу

$$\langle S, 1, p_{j_0}\rangle, \langle w_1, i_1, p_{j_1}\rangle, \ldots, \langle w_{m-1}, i_{m-1}, p_{j_{m-1}}\rangle, \langle w, 0, 1\rangle.$$

Построим по этой трассе последовательность конечных множеств

$$F_0, F_1, F_2, \ldots, F_m,$$

элементами которых являются вершины графа Θ. При этом окажется, что построенные множества являются $(n+1)$-позициями и предыдущее множество получается из последующего при помощи нескольких (не более трех) ходов. Так как начальное множество будет содержать только слово w, то это будет доказывать, что w является $(n+1)$-вершиной.

Очевидным образом по трассе для каждого вхождения нетерминала S в слово вывода определяется терминальное слово, порождаемое этим вхождением в рассматриваемом выводе. Именно, если это вхождение не заменяется в этой трассе, то оно порождает то же слово, что и соответствующее вхождение в слово из следующего элемента трассы. Если оно заменяется на a, то оно порождает a. Если же оно заменяется на (SS), а полученные два новых вхождения нетерминала в слово из следующего элемента трассы порождают терминальные слова u и v соответственно, то рассматриваемое вхождение порождает слово (uv).

Для построения этих конечных множеств будем обрабатывать данную трассу поэлементно в обратном порядке. Положим $F_0 = \emptyset$ и будем считать, что элемент $\langle w, 0, 1\rangle$ трассы мы обработали. Пусть мы обработали элементы

$$\langle w_{s+1}, i_{s+1}, p_{j_{s+1}}\rangle, \ldots, \langle w, 0, 1\rangle$$

трассы и построили F_{s+1}. При этом мы предполагаем, что для каждого вхождения нетерминала в w_{s+1} множество F_{s+1} содержит терминальное слово, порождаемое этим вхождением.

Если при переходе от w_s к w_{s+1} использовалось правило $S \to a$, то для получения F_s из F_{s+1} кладем камушек на a. Если же использовалось правило $S \to (SS)$, а заменяемое вхождение нетерминала порождает терминальное слово u, то кладем камушек на u. После этого снимаем камушки с терминальных слов, порождаемых вхождениями нетерминала, полученными после использования этого правила, если эти терминальные слова не порождаются никакими вхождениями нетерминала в слово w_s.

Итак, $\emptyset \Rightarrow^*_{n+1} \{w\}$, то есть w — $(n+1)$-вершина графа Θ. ∎

Доказательство того, что язык $L(G)$ также имеет бесконечный индекс, предоставляется читателю в качестве упражнения (см. упражнение 2.4.7). Идейно доказательство следует доказательству леммы 2.24, однако здесь имеются некоторые технические сложности. Наводящие соображения приведены в упражнениях 2.4.8— 2.4.10.

Упражнения

Упражнение 2.4.1. *Покажите, что равновысокие нетерминалы не обязаны быть сводимыми.*

Упражнение 2.4.2 (**). *Докажите, что если G — КС-грамматика, то либо $L(G)$ регулярен, либо G — грамматика с сильным самовставлением.*

Упражнение 2.4.3. *Пусть $G = \langle A, Q, S, P\rangle$ — приведенная КС-грамматика. Покажите, что $h(S) = h(G)$.*

Упражнение 2.4.4 (*). *Приведите пример КС-языка высоты 3. Докажите это.*

Упражнение 2.4.5 (**). *Приведите пример КС-языка L такого, что если G — КС-грамматика в нормальной форме Грейбах и $L = L(G)$, то G не является сжатой.*

Упражнение 2.4.6 (**). *КС-грамматику $G = \langle A, Q, S, P\rangle$ будем называть древесной, если $D(R) = \{R\}$ для всякого $R \in Q$. Всякий ли КС-язык задается древесной грамматикой?*

Упражнение 2.4.7 (**). *Покажите, что если G — грамматика из леммы 2.24, то $L(G)$ имеет бесконечный индекс (см. упражнения 2.4.8– 2.4.10).*

Упражнение 2.4.8. *Пусть G — произвольная КС-грамматика конечного индекса. Постройте по ней эквивалентную сжатую КС-грамматику конечного индекса.*

Упражнение 2.4.9 (**). *Пусть $G' = \langle A, Q', S', P' \rangle$ — сжатая приведенная КС-грамматика, эквивалентная грамматике G из леммы 2.24. Покажите, что для произвольного нетерминала $R \in Q'$ существуют $x, y \in A^*$ такие, что если $z \in A^*$ и $R \Rightarrow^*_{G'} z$, то $xzy \in L(G')$.*

Упражнение 2.4.10 (**). *Пусть G' как в упражнении 2.4.9. Покажите, что для произвольного нетерминала $R \in Q'$ существует натуральное число n и $\alpha_1, \ldots, \alpha_{n+1} \in \{), (\}$, что если $x \in A^* = \{(,), a\}^*$ и $R \Rightarrow^*_{G'} x$, то существуют*

$$\beta_1, \beta_2, \ldots, \beta_n \in L(G'),$$

что

$$x = \alpha_1 \beta_1 \alpha_2 \ldots \alpha_n \beta_n \alpha_{n+1}.$$

2.5 Автоматные грамматики

Этот параграф посвящен соотношению контекстно-свободных и регулярных языков.

Согласно теореме 1.3, всякий регулярный язык распознается некоторым недетерминированным конечным автоматом.

Следующая теорема уже фактически подробно доказана нами (см. теорему 2.6) в чуть другой формулировке. Тем не менее мы напомним конструкцию.

Теорема 2.25. *Пусть $I = \langle A, Q, q_0, F, P \rangle$ — недетерминированный конечный автомат. Можно построить такую КС-грамматику*

$$G = \langle A, Q, q_0, P' \rangle,$$

что $L(I) = L(G)$.

Доказательство. Построим P' явно по P. Для каждого правила $\langle q, a, q' \rangle$ НКА I, где $q, q' \in Q, a \in A$, включим в P' правило $q \to aq'$. Если при этом $q' \in F$, то включим в P' также правило $q \to a$. Если $q_0 \in F$, включим в P' также правило $q_0 \to e$, где e-пустое слово.

Легко видеть, что требуемая КС-грамматика построена. ∎

В действительности мы доказали, что всякий регулярный язык задается некоторой грамматикой специального вида, так называемой автоматной грамматикой.

Определение 2.5.1 (праволинейности и автоматности). *КС-грамматика $G = \langle A, Q, S, P \rangle$ называется автоматной, если всякое правило $p \in P$ имеет один из видов:*

1. $R \to aR'$, где $R, R' \in Q$, $a \in A$.

2. $R \to a$, где $R \in Q$, $a \in A$.

3. $S \to e$, где e-пустое слово.

Грамматика G называется праволинейной, если всякое правило $p \in P$ имеет один из трех видов:

1. $R \to \alpha R'$, где $R, R' \in Q$, $\alpha \in A^*$, $\alpha \neq e$.

2. $R \to \alpha$, где $R \in Q$, $\alpha \in A^*$, $\alpha \neq e$.

3. $S \to e$, где e-пустое слово.

Соответственно, КС-язык называется праволинейным (автоматным), если он задается праволинейной (автоматной) грамматикой.

Основной теоремой данного параграфа является:

Теорема 2.26. *Пусть $G = \langle A, Q, S, P \rangle$ — праволинейная грамматика. Можно построить регулярное выражение r в алфавите A такое, что $L(r) = L(G)$.*

Прежде, чем перейти к доказательству, отметим некоторые очевидные следствия.

Следствие 2.27. *Пусть I — недетерминированный конечный автомат. Можно построить регулярное выражение r в том же алфавите такое, что $L(I) = L(r)$.*

Следствие 2.28. *Пусть G — праволинейная грамматика. Можно построить такую автоматную грамматику G' с тем же алфавитом терминалов, что $L(G) = L(G')$.*

Итак, все три способа задания языков: регулярными выражениями, конечными автоматами и праволинейными грамматиками, эквивалентны.

Перейдем теперь к доказательству теоремы 2.26.

Доказательство теоремы 2.26. Наша цель — сопоставить всякому слову $\alpha \in (A \bigcup Q)^*$ такое регулярное выражение $r(\alpha)$, что для любого $\beta \in A^*$, $\alpha \Rightarrow_G^* \beta$ эквивалентно тому, что $\beta \in L(r(\alpha))$.

Тогда в качестве искомого r можно взять регулярное выражение $r(S)$.

Если $\alpha \in A^*$, то построение $r(\alpha)$ очевидно.

Если для некоторых

$$\alpha_1, \alpha_2, \ldots, \alpha_m \in (A \bigcup Q)^*$$

выражения

$$r(\alpha_1), r(\alpha_2), \ldots, r(\alpha_m)$$

уже построены, то в качестве $r(\alpha_1 \alpha_2 \ldots \alpha_m)$ можно взять

$$r(\alpha_1) r(\alpha_2) \ldots r(\alpha_m).$$

Построим для $R \in Q$ выражения $r(R)$ индукцией по высоте R.

Пусть для всех символов $R' \in Q$ меньшей высоты, чем высота R, выражения $r(R')$ уже построены. Рассмотрим символ R.

Если R не является самовставимым, то все правила с левой частью R имеют вид $R \to \alpha$, где α составлено из терминалов и таких нетерминалов, высота которых меньше высоты R. Пусть

$$R \to \alpha_1, \ldots, R \to \alpha_n$$

будут все правила с левой частью R. По индукции, $r(\alpha_1), \ldots, r(\alpha_n)$ можно считать построенными. Положим

$$r(R) = r(\alpha_1) + r(\alpha_2) + \cdots + r(\alpha_n).$$

Итак, $r(R)$ построено.

Предположим теперь, что R — самовставимый нетерминал.

Напомним, что для всякого R' из $D(R)$ существуют такие непустые слова $x, y \in A^+$, что $R \Rightarrow_G^* xR'$, $R' \Rightarrow_G^* yR$ (непустота x, y следует из праволинейности).

Пусть $R' \in D(R)$ и

$$R' \to \alpha_1, R' \to \alpha_2, \ldots, R' \to \alpha_n$$

будут все такие правила с левой частью R', что

$$\alpha_1, \ldots, \alpha_n \in (A \bigcup (Q \setminus D(R)))^*,$$

то есть $\alpha_1, \ldots, \alpha_n$ не содержат нетерминалов из $D(R)$. Положим $r_0(R') = r(\alpha_1) + \cdots + r(\alpha_n)$.

Определим теперь для произвольных $R', R'' \in D(R)$ и произвольного $D \subseteq D(R)$, такие регулярные выражения $\rho(R', R'', D)$, что язык $L(\rho(R', R'', D))$, задаваемый $\rho(R', R'', D)$, состоит из всех таких слов x в алфавите A, для которых в грамматике G существует вывод xR'' из R', не содержащий нетерминалов из D, кроме входящих в начальное или заключительное слово этого вывода. В частности, если R' и R'' совпадают, то язык $L(\rho(R', R'', D))$ содержит пустое слово.

Другими словами, мы стремимся построить регулярные выражения $\rho(R', R'', D)$, задающие язык, "соединяющий"в грамматике G нетерминалы R' и R'', но без промежуточного участия нетерминалов из D.

Построим регулярное выражение $\rho(R', R'', D)$ обратной индукцией по мощности множества D.

Базис. Пусть $D = D(R)$. Пусть

$$R' \to \beta_1 R'', R' \to \beta_2 R'', \ldots R' \to \beta_m R''$$

будут все правила из P с левой частью R' и правой частью, содержащей R''. Тогда можно положить

$$\rho(R', R'', D(R)) = r(\beta_1) + \cdots + r(\beta_m),$$

если $R' \neq R''$, и

$$\rho(R', R'', D(R)) = r(\beta_1) + \cdots + r(\beta_m) + e,$$

если $R' = R''$. Отметим, что

$$\beta_1, \ldots, \beta_m \in A^*.$$

Индукционный шаг. Пусть $F = D(R) \setminus D$.

Пусть $R' \in F$. Определим сперва $\rho(R', R', D)$. Ясно, что

$$\rho(R', R', D) = (\rho(R', R', D \bigcup \{R'\}))^*$$

и по индукции определено. Теперь

$$\rho(R', R'', D) = \rho(R', R', D)\rho(R', R'', D \bigcup \{R'\})$$

и также определено по индукции.

Пусть теперь $R' \notin F$, то есть $R' \in D$. Этот случай сводится к предыдущему, ибо ясно, что

$$\rho(R', R'', D) = \rho(R', R'', D(R)) + \sum_{X \in F} (\rho(R', X, D(R))\rho(X, R'', D)).$$

$\rho(R', R'', D(R))$ и $\rho(R', X, D(R))$ определены по индукции, а $\rho(X, R'', D)$ — предыдущий случай, ибо $X \in F$.

Итак, мы построили регулярные выражения $\rho(R', R'', D)$ для произвольных $D \subseteq D(R)$ и $R', R'' \in D(R)$.

Теперь

$$r(R) = \sum_{Y \in D(R)} (\rho(R, Y, \emptyset) r_0(Y)).$$

∎

Рассмотрим пример праволинейной грамматики

$$G = \langle \{0, 1, 2\}, \{S, T\}, S, P \rangle$$

со следующим множеством правил P:

$$S \to 0S, S \to 1T, S \to 0,$$

$$T \to 2T, T \to 1S, T \to 2.$$

Преобразуем ее к эквивалентному регулярному выражению $r(S)$. Ясно, что все нетерминалы грамматики самовставимы, и что $D(S) = \{S, T\}$. Поэтому $r_0(S) = 0$; и $r_0(T) = 2$. Далее

$$\rho(S, S, \{S, T\}) = r(0) = (0 + e);$$
$$\rho(T, T, \{S, T\}) = r(2) = (2 + e);$$
$$\rho(S, T, \{S, T\}) = r(1) = 1;$$
$$\rho(T, S, \{S, T\}) = r(1) = 1;$$

$$\rho(S, S, \{T\}) = (0 + e + e)^* = (0 + e)^*;$$
$$\rho(S, T, \{T\}) = (0 + e)^* 1;$$
$$\rho(T, T, \{S\}) = (2 + e + e)^* = (2 + e)^*;$$
$$\rho(T, S, \{S\}) = (2 + e)^* 1;$$

$$\rho(S, S, \{S\}) = (0 + e) + 1(2 + e)^* 1;$$
$$\rho(S, T, \{S\}) = 1 + 1(2 + e)^*;$$
$$\rho(T, T, \{T\}) = (2 + e) + 1(0 + e)^* 1;$$

$$\rho(T, S, \{T\}) = 1 + 1(0 + e)^*;$$

$$\rho(S, S, \emptyset) = (\rho(S, S, \{S\}) + e)^* = ((0 + e) + 1(2 + e)^*1)^*;$$
$$\rho(S, T, \emptyset) = \rho(S, S, \emptyset)\rho(S, T, \{S\}) = ((0 + e) + 1(2 + e)^*1)^*(1 + 1(2 + e)^*);$$
$$\rho(T, T, \emptyset) = (\rho(T, T, \{T\}) + e)^* = ((2 + e) + 1(0 + e)^*1)^*;$$
$$\rho(T, S, \emptyset) = \rho(T, T, \emptyset)\rho(T, S, \{T\}) = ((2 + e) + 1(0 + e)^*1)^*(1 + 1(0 + e)^*).$$

Наконец,

$$r(S) = \rho(S, S, \emptyset)r_0(S) + \rho(S, T, \emptyset)r_0(T) =$$
$$((0 + e) + 1(2 + e)^*1)^*0 + ((0 + e) + 1(2 + e)^*1)^*(1 + 1(2 + e)^*)2.$$

Конец примера.

Итак, мы показали, что всякий праволинейный язык регулярен (и наоборот).

Очевидно, что индекс всякого регулярного языка равен 1. Однако, существуют и нерегулярные языки индекса 1. Для доказательства этого рассмотрим язык $L = \{0^i1^i \mid i \in N\}$ из следствия 1.10. В названном следствии леммы о разрастании для регулярных языков показано, что L нерегулярен. Однако L порождается грамматикой

$$G = \langle\{0, 1\}, \{S\}, S, P\rangle$$

со следующим множеством правил

$$P = \{S \to 0S1, \quad S \to e\}.$$

Очевидно, что индекс L равен 1.

Упражнения

Упражнение 2.5.1. *Рассмотрим праволинейную грамматику*

$$G = \langle\{a, b, c\}, \{S, T, T'\}, S, P\rangle$$

со следующим множеством правил P :

$$S \to aS, \quad S \to bT, \quad S \to abS, \quad S \to baT,$$
$$T \to bT', \quad T \to cS, \quad T \to aS,$$
$$T' \to aT, \quad T' \to e, \quad T' \to a.$$

Постройте регулярные выражения r *и детерминированный конечный автомат* J *такие, что* $L(G) = L(J) = L(r)$.

Упражнение 2.5.2. *Всякий ли регулярный язык задается такой праволинейной грамматикой G, в которой нет двух различных сводимых нетерминалов?*

Упражнение 2.5.3. *Грамматику*

$$G = \langle A, Q, S, P \rangle$$

будем называть леволинейной, если всякое правило $p \in P$ имеет один из следующих видов:

$R \to R'\alpha$, *где* $R, R' \in Q$, $\alpha \in A^*$, $\alpha \neq e$;

$R \to \alpha$, *где* $R \in Q$, $\alpha \in A^*$, $\alpha \neq e$;

$S \to e$, *где* e *— пустое слово.*

Докажите, что по всякой леволинейной грамматике можно построить эквивалентную ей праволинейную.

2.6 Автоматы с магазинной памятью

Подобно тому, как регулярные языки распознаются конечными автоматами, КС-языки распознаются автоматами с магазинной памятью.

Содержательно, автомат с магазинной памятью или МП-автомат — это конечный автомат, дополнительно оборудованный магазином (стеком). Стек (магазин) — это дополнительная память автомата специального вида. Существует специальный алфавит магазинных символов автомата. В каждый момент времени стек содержит некоторое слово в этом алфавите. В начальный момент времени стек содержит пустое слово. На каждом такте работы МП-автомат обозревает некоторый префикс слова из стека и выбирает следующий шаг в зависимости не только от обозреваемого символа на ленте, но и от этого префикса (ограниченной длины). Сам такт, кроме сдвига головки вправо на ленте и перехода в новое состояние, состоит еще в замене обозреваемого префикса слова, записанного в стеке, на новое слово (новый префикс).

Дадим формальное определение.

Определение 2.6.1 (МП-автомата). *МП-автомат — это шестерка*

$$J = \langle A, Q, \Gamma, q_0, F, P \rangle,$$

где:

- A *— конечный входной алфавит;*
- Q *— конечный алфавит состояний;*

- Γ — *конечный магазинный алфавит;*
- $q_0 \in Q$ — *начальное состояние;*
- $F \subseteq Q$ — *множество заключительных состояний;*
- $P : (A^* \times Q \times \Gamma^*) \to FS(Q \times \Gamma^*)$ — *частичная функция переходов с конечной областью определения, где* $FS(Q \times \Gamma^*)$ — *множество всех конечных подмножеств* $(Q \times \Gamma^*)$.

Определение 2.6.2 (работы МП-автомата). *Конфигурацией МП-автомата*

$$J = \langle A, Q, \Gamma, q_0, F, P\rangle$$

называется тройка $\langle w, q, s\rangle$, *где* $w \in A^*$, $q \in Q$, $s \in \Gamma^*$.

Пусть $J = \langle A, Q, \Gamma, q_0, F, P\rangle$ — *МП-автомат. Рассмотрим его конфигурации*

$$\langle w, q, s\rangle, \quad \langle w', q', s'\rangle.$$

Будем говорить, что конфигурация $\langle w', q', s'\rangle$ *следует за конфигурацией* $\langle w, q, s\rangle$ *и писать* $\langle w, q, s\rangle \Rightarrow_J \langle w', q', s'\rangle$, *если существуют такие слова* $a \in A^*$ *и* $, x, y, z \in \Gamma^*$, *что выполнены условия:*

1. $w = aw'$;

2. $s = xz$, $s' = yz$;

3. $\langle q', y\rangle \in P(a, q, x)$.

Будем писать

$$\langle w, q, s\rangle \Rightarrow_J^n \langle w', q', s'\rangle,$$

если существует такая конечная последовательность конфигураций

$$\langle w_0, q_0, s_0\rangle, \ldots, \langle w_n, q_n, s_n\rangle,$$

где $n \geq 0$, *что*

$$\langle w_i, q_i, s_i\rangle \Rightarrow_J \langle w_{i+1}, q_{i+1}, s_{i+1}\rangle$$

для $i = 0, \ldots, n-1$, *и*

$$\langle w_0, q_0, s_0\rangle = \langle w, q, s\rangle, \quad \langle w_n, q_n, s_n\rangle = \langle w', q', s'\rangle.$$

Если для некоторого n

$$\langle w, q, s\rangle \Rightarrow_J^n \langle w', q', s'\rangle,$$

то пишем

$$\langle w, q, s\rangle \Rightarrow_J^* \langle w', q', s'\rangle.$$

Будем говорить, что слово $w \in A^*$ *распознается МП-автоматом* J*, если существует такое* $q \in F$*, что*

$$\langle w, q_0, e\rangle \Rightarrow_J^* \langle e, q, e\rangle,$$

где e *— пустое слово.*

Множество всех слов $w \in A^*$*, которые распознаются МП-автоматом* J*, будем называть языком, распознаваемым автоматом* J *и обозначать* $L(J)$.

Теорема 2.29. *Пусть* $G = \langle A, Q, S, P\rangle$ *— КС-грамматика. Можно построить такой МП-автомат* $J = \langle A, Q', \Gamma, q_0, F, P'\rangle$*, что* $L(G)=L(J)$*, при этом правила автомата* J *имеют очень специальный вид, а именно:*

$$P' : ((A \bigcup \{e\}) \times Q' \times (\Gamma \bigcup \{e\})) \to FS(Q' \times \Gamma^*).$$

Доказательство. Положим $Q' = \{q, q_0\}$, $\Gamma = A \bigcup Q$, $F = \{q\}$, а P' определим следующим образом:

$$P'(e, q_0, e) = \{\langle q, S\rangle\}, \quad P'(a, q, a) = \{\langle q, e\rangle\}$$

для всех $a \in A$. Кроме того для всякого правила $R \to \alpha \in P$, где $R \in Q$ и $\alpha \in (A \bigcup Q)^*$, включим $\langle q, \alpha\rangle$ в $P'(e, q, R)$.

Разобьем доказательство теоремы на ряд лемм.

Лемма 2.30. *Пусть* $J = \langle A, Q', \Gamma, q_0, F, P\rangle$ *— произвольный МП-автомат. Пусть, для некоторых* $\alpha, \beta \in \Gamma^*$*,* $u, v \in A^*$ *и* $q, q', q'' \in Q'$

$$\langle u, q, \alpha\rangle \Rightarrow_J^* \langle e, q', e\rangle, \quad \langle v, q', \beta\rangle \Rightarrow_J^* \langle e, q'', e\rangle.$$

Тогда

$$\langle uv, q, \alpha\beta\rangle \Rightarrow_J^* \langle e, q'', e\rangle.$$

Доказательство. Очевидно. ∎

Лемма 2.31. *Пусть* G *и* J *— КС-грамматика и МП-автомат, рассматриваемые в теореме 2.29. Пусть для некоторых* $\beta \in \Gamma^*$ *и* $w \in A^*$

$$\beta \Rightarrow_G^* w.$$

Тогда

$$\langle w, q, \beta\rangle \Rightarrow_J^* \langle e, q, e\rangle.$$

Доказательство. Пусть $\beta \Rightarrow_G^m w$. Докажем лемму индукцией по этому m.

Базис. Если $m = 0$, то $\beta = w \in A^*$ и утверждение очевидно, ибо

$$\langle w, q, w\rangle \Rightarrow_J^{|w|} \langle e, q, e\rangle.$$

Пусть $m = 1$. Тогда существуют такие $w', w'', w''' \in A^*$ и $R \in Q$, что $\beta = w'Rw'''$, $w = w'w''w'''$ и $R \to w'' \in P$. Ясно, что

$$\langle w', q, w'\rangle \Rightarrow_J^{|w'|} \langle e, q, e\rangle,$$

$$\langle w'', q, w''\rangle \Rightarrow_J^{|w''|} \langle e, q, e\rangle,$$

$$\langle w''', q, w'''\rangle \Rightarrow_J^{|w'''|} \langle e, q, e\rangle.$$

Но $\langle q, w''\rangle \in P'(e, q, R)$, следовательно,

$$\langle w'', q, R\rangle \Rightarrow_J \langle w'', q, w''\rangle.$$

Значит,

$$\langle w'', q, R\rangle \Rightarrow_J^{|w''|+1} \langle e, q, e\rangle.$$

Сопоставляя эти утверждения, по лемме 2.30 получим, что

$$\langle w, q, \beta\rangle = \langle w'w''w''', q, w'Rw'''\rangle \Rightarrow_J^* \langle e, q, e\rangle.$$

Индукционный шаг. Пусть утверждение доказано для выводов, длина которых меньше m.

Так как $m > 1$, существует такое $\beta' \in (A \bigcup Q)^*$, что

$$\beta \Rightarrow_G \beta' \Rightarrow_G^{m-1} w.$$

Тогда существуют такие $\alpha_1, \alpha_2, \delta \in (A \bigcup Q)^*$ и $R \in Q$, что

1. $\beta = \alpha_1 R \alpha_2$,
2. $\beta' = \alpha_1 \delta \alpha_2$,
3. $R \to \delta \in P$.

Значит, существуют такие $w', w'', w''' \in A^*$ и $n', n'', n''' \leq m - 1$, что $w = w'w''w'''$ и

$$\alpha_1 \Rightarrow_G^{n'} w',$$

$$\delta \Rightarrow_G^{n''} w'',$$

$$\alpha_2 \Rightarrow_G^{n'''} w'''.$$

Так как $\langle q, \delta \rangle \in P'(e, q, R)$, то

$$\langle w'', q, R \rangle \Rightarrow_J \langle w'', q, \delta \rangle.$$

По индукции:

$$\langle w', q, \alpha_1 \rangle \Rightarrow_J^* \langle e, q, e \rangle,$$

$$\langle w'', q, \delta \rangle \Rightarrow_J^* \langle e, q, e \rangle,$$

$$\langle w''', q, \alpha_2 \rangle \Rightarrow_J^* \langle e, q, e \rangle.$$

Поэтому

$$\langle w'', q, R \rangle \Rightarrow_J^* \langle e, q, e \rangle.$$

Сопоставляя эти утверждения, по лемме 2.30 получим, что

$$\langle w, q, \beta \rangle = \langle w'w''w''', q, \alpha_1 R \alpha_2 \rangle \Rightarrow_J^* \langle e, q, e \rangle.$$

∎

Пусть теперь $w \in L(G)$. Тогда $S \Rightarrow_G^* w$. Тогда, по лемме 2.31,

$$\langle w, q, S \rangle \Rightarrow_J^* \langle e, q, e \rangle.$$

По определению,

$$\langle e, q_0, e \rangle \Rightarrow_J \langle e, q, S \rangle.$$

По лемме 2.30,

$$\langle w, q_0, e \rangle \Rightarrow_J^* \langle e, q, e \rangle,$$

значит, $w \in L(J)$. Итак, мы доказали, что $L(G) \subseteq L(J)$.

Докажем теперь, что $L(J) \subseteq L(G)$.

Лемма 2.32. *Пусть G и J — КС-грамматика и МП-автомат, рассматриваемые в теореме 2.29. Пусть для некоторых $\beta \in \Gamma^*$ и $w \in A^*$*

$$\langle w, q, \beta \rangle \Rightarrow_J^* \langle e, q, e \rangle.$$

Тогда $\beta \Rightarrow_G^* w$.

Доказательство. Пусть

$$\langle w, q, \beta \rangle \Rightarrow_J^m \langle e, q, e \rangle.$$

Если $m = 0$, то $\beta = w = e$ и $e \Rightarrow_G^0 e$.

Базис. Пусть $m = 1$. Тогда, по определению P', $|\beta| = 1$.

Случай 1. $\beta = a \in A$. Тогда $w = a$ и $a \Rightarrow_G^0 a$.

Случай 2. $\beta = R \in Q$. Из определения P', ясно, что $w = e$ и $R \to e \in P$. Значит, $R \Rightarrow_G e$ и $\beta \Rightarrow_G^* e$.

Индукционный шаг. Пусть утверждение доказано для меньших m. Из

$$\langle w, q, \beta \rangle \Rightarrow_J^m \langle e, q, e \rangle$$

следует существование таких $w' \in A^*$ и $\beta' \in \Gamma^*$, что

$$\langle w, q, \beta \rangle \Rightarrow_J \langle w', q, \beta' \rangle \Rightarrow_J^{m-1} \langle e, q, e \rangle.$$

По индукции, $\beta' \Rightarrow_G^* w'$. Ясно, что $|w'| \leq |w|$.

Рассмотрим отдельно два случая.

Случай 1. $|w'| < |w|$. Тогда, по определению P', существуют такие $a \in A$ и $\beta' \in \Gamma^*$, что $w = aw'$ и $\beta = a\beta'$. Следовательно, $\beta = a\beta' \Rightarrow_G^* aw' = w$.

Случай 2. $|w'| = |w|$. Тогда $w = w'$. Из определения P' следует, что существуют такие $R \in Q$ и $\alpha_1, \alpha_2 \in \Gamma^*$, что:

$\beta = R\alpha_1$,

$\beta' = \alpha_2\alpha_1$,

$R \to \alpha_2 \in P$.

Но тогда $R \Rightarrow_G \alpha_2$ и, значит, $\beta = R\alpha_1 \Rightarrow_G \alpha_2\alpha_1 = \beta'$. Но из $\beta \Rightarrow_G \beta'$ и $\beta' \Rightarrow_G^* w' = w$ следует, что $\beta \Rightarrow_G^* w$. ∎

Если теперь $w \in L(J)$, то

$$\langle w, q_0, e \rangle \Rightarrow_J^* \langle e, q, e \rangle.$$

Но, если

$$\langle w, q_0, e \rangle \Rightarrow_J \langle w', q', \alpha' \rangle$$

для некоторых $w' \in A$, $\alpha' \in \Gamma^*$ и $q' \in Q'$, то из определения P' следует, что $q' = q$, $w' = w$, $\alpha' = S$, то есть имеем

$$\langle w, q_0, e \rangle \Rightarrow_J \langle w, q, S \rangle \Rightarrow_J^* \langle e, q, e \rangle.$$

По лемме 2.32 тогда $S \Rightarrow_G^* w$, то есть $w \in L(G)$. Итак, $L(J) \subseteq L(G)$. ∎

Определение 2.6.3 (полуканонического МП-автомата). *Пусть $J = \langle A, Q, \Gamma, q_0, F, P \rangle$ — МП-автомат. J будем называть полуканоническим МП-автоматом, если P есть частичное отображение $A^* \times Q \times (\Gamma \bigcup \{e\})$ в $FS(Q \times \Gamma^*)$.*

Заметим, что МП-автомат, строящийся в теореме 2.29, полуканонический.

Лемма 2.33. *Пусть $J = \langle A, Q, \Gamma, q_0, F, P\rangle$ — МП-автомат. Можно построить такой полуканонический МП-автомат*

$$J' = \langle A, Q', \Gamma', q_0', F', P'\rangle,$$

что $L(J) = L(J')$.

Доказательство. Итак, P есть частичное отображение $(A^* \times Q \times \Gamma^*) \to FS(Q \times \Gamma^*)$. Рассмотрим минимальное из таких натуральных чисел n, что если для некоторых $w \in A^*$, $q \in Q$ и $\alpha \in \Gamma^*$, $P(w, q, \alpha)$ определено, то $|\alpha| \le n$, то есть n — это максимальная глубина, на которую J просматривает стек за один шаг. Положим

$$Q' = \{\langle\alpha, q\rangle|\ q \in Q,\ \ \alpha \in (\Gamma \bigcup \{e\})^n\}.$$

Таким образом, $Q' \subseteq ((\Gamma \bigcup \{e\})^n \times Q)$. Положим $\Gamma' = \Gamma$. Определим частичную функцию P' так:

Для каждых $q \in Q$, $\epsilon \in \Gamma$, $\alpha \in (\Gamma \bigcup \{e\})^{n-1}$ включим пару $\langle\langle\epsilon\alpha, q\rangle, e\rangle$ в $P'(e, \langle\alpha, q\rangle, \epsilon)$.

Для каждых $q \in Q$, $\alpha \in (\Gamma \bigcup \{e\})^n$ и $w \in A^*$ таких, что $P(w, q, \alpha)$ определена, и для каждой пары $\langle q', \alpha'\rangle \in P(w, q, \alpha)$, включим пару $\langle\langle e, q'\rangle, \alpha'\rangle$ в $P'(w, \langle\alpha, q\rangle, e)$.

Пусть q_0' есть $\langle e, q_0\rangle$ и F' есть $\{\langle e, q\rangle \mid\ q \in F\}$. $L(J') = L(J)$. Действительно, автомат J' отличается от J лишь тем, что прежде чем применить переход, "перекачивает"нужную часть стека в состояние. ∎

Теорема 2.34. *Пусть $J = \langle A, Q, \Gamma, q_0, F, P\rangle$ — МП-автомат. Можно построить такую КС-грамматику $G = \langle A, Q', S, P'\rangle$, что $L(J) = L(G)$.*

Доказательство. В силу леммы 2.33, можно считать, что автомат J — полуканонический, значит, P есть частичное отображение

$$(A^* \times Q \times (\Gamma \bigcup \{e\})) \to FS(Q \times \Gamma^*)$$

Положим

$$Q' = \{\langle q, \alpha, q'\rangle \mid\ q, q' \in Q;\ \alpha \in (\Gamma \bigcup \{e\})\}.$$

Множество правил P' определим так:

1. Включим в P' правила $S \to \langle q_0, e, q\rangle$ для всех $q \in F$.
2. Для произвольных $q, q', q'' \in Q$ и $\epsilon \in \Gamma$ включим в P' правило:

$$\langle q, \epsilon, q''\rangle \to \langle q, e, q'\rangle\langle q', \epsilon, q''\rangle.$$

3. Для произвольных таких $w \in A^*$, $q \in Q$ и $\epsilon \in (\Gamma \bigcup \{e\})$, что $P(w, q, \epsilon)$ определено, и для всякой пары $\langle q', \alpha\rangle \in P(w, q, \epsilon)$, где $q' \in Q$,, $\alpha \in \Gamma^*$, $|\alpha| = n > 0$ и $\alpha = \epsilon_1\epsilon_2 \dots \epsilon_n$, включим в P' всевозможные правила вида:

$$\langle q, \epsilon, q_{i_n}\rangle \to w\langle q', \epsilon_1, q_{i_1}\rangle\langle q_{i_1}, \epsilon_2, q_{i_2}\rangle \dots \langle q_{i_{n-1}}, \epsilon_n, q_{i_n}\rangle$$

для всевозможных $q_{i_1}, q_{i_2}, \dots, q_{i_n} \in Q$.

Если же $\langle q', e\rangle \in P(w, q, \epsilon)$, то включим в P' правило $\langle q, \epsilon, q'\rangle \to w$.

На этом построении грамматики $G = \langle A, Q', S, P'\rangle$ завершено.

Читателю предлагается индукцией по n доказать два утверждения для произвольных $q, q' \in Q$, $\epsilon \in (\Gamma \bigcup \{e\})$ и $w \in A^*$:

1. Если $\langle q, \epsilon, q'\rangle \Rightarrow_G^n w$, то

$$\langle w, q, \epsilon\rangle \Rightarrow_J^* \langle e, q', e\rangle;$$

2. Если $\langle w, q, \epsilon\rangle \Rightarrow_J^n \langle e, q', e\rangle$, то

$$\langle q, \epsilon, q'\rangle \Rightarrow_G^* w.$$

Из этих двух утверждений следует, что для произвольного $w \in A^*$ $S \Rightarrow_G^* w$ тогда и только тогда, когда для некоторого $q \in F$

$$\langle w, q_0, e\rangle \Rightarrow_J^* \langle e, q, e\rangle.$$

Значит, $L(J) = L(G)$, это и требовалось доказать. ∎

В литературе рассматривают также частный случай МП-автоматов, так называемые детерминированные МП-автоматы или ДМП-автоматы. Их отличие от МП-автоматов вообще состоит в том, что если МП-автомат

$$J = \langle A, Q, \Gamma, q_0, F, P\rangle$$

есть ДМП-автомат, то для произвольных $w, w', w'' \in A^*$; $q, q', q'' \in Q$; $s, s', s'' \in \Gamma^*$, если $\langle w, q, s\rangle \Rightarrow_J \langle w', q', s'\rangle$ и $\langle w, q, s\rangle \Rightarrow_J \langle w'', q'', s''\rangle$, то $w' = w''$, $q' = q''$ и $s' = s''$. Этого результата можно достичь простыми ограничениями на функцию P. В отличие от конечных автоматов, детерминированный вариант которых позволяет моделировать недетерминизм, МП-автоматы не сводятся к ДМП. Существует МП-автомат J такой, что, для произвольного ДМП-автомата J', $L(J) \neq L(J')$. В силу теоремы 2.34, класс ДМП-автоматов не позволяет распознавать все КС-языки. Рассмотрение ДМП-автоматов выходит за рамки этой книги.

Упражнения

Упражнение 2.6.1. *Постройте МП-автомат J такой, что $L(J) = L(G)$ для КС-грамматики $G = \langle \{0,1\}, \{S\}, S, P\rangle$ со следующим множеством правил P :*

$$\{S \to e, S \to 0S1, S \to SS\}.$$

Язык $L(G)$ называется двупорожденным языком Дика.

Упражнение 2.6.2. *Пусть J — МП-автомат. Пусть M — это множество слов, которые могут появиться в магазине, при условии, что J начинает работу, когда в магазине записано пустое слово. Покажите, что язык M регулярен.*

Упражнение 2.6.3. *Пусть J и M как в предыдущем упражнении. Пусть M конечен. Покажите, что $L(J)$ регулярен.*

Упражнение 2.6.4. *Пусть $J = \langle A, Q, \Gamma, q_0, F, P\rangle$ — МП-автомат. Его называют каноническим, если P есть отображение $(A \times Q \times \Gamma) \to FS(Q \times \Gamma^*)$. Покажите, что по всякому МП-автомату I можно построить такой канонический МП-автомат J, что $L(I) = L(J)$.*

2.7 Алгоритмические проблемы для КС-языков

Начнем с проблемы принадлежности слова КС-языку.

Теорема 2.35. *Существует алгоритм, который по данным КС-грамматике G и терминальному слову w определяет, принадлежит ли w к $L(G)$.*

Доказательство. По теореме 2.13 можно эффективно преобразовать G в КС-грамматику $G' = \langle A, Q', S', P'\rangle$ в нормальной форме Грейбах так, что $L(G) = L(G')$.

По теореме 2.29, можно построить по G' полуканонический МП-автомат $J = \langle A, \{q_0, q\}, \Gamma, q_0, F, P'\rangle$, что $L(G') = L(J) = L(G)$.

Из доказательства теоремы 2.29 легко видеть, что

$$\langle w, q_0, e\rangle \Rightarrow_J \langle w, q, S'\rangle$$

и что, так как G' находится в нормальной форме Грейбах, из двух подряд идущих тактов работы автомата хотя бы один сдвигает головку на входной ленте (кроме, быть может, первых двух тактов). Иными словами, из

$$\langle w, q, \alpha\rangle \Rightarrow_J \langle w', q, \beta\rangle \Rightarrow_J \langle w'', q, \gamma\rangle$$

следует, что $|w''| < |w|$. Ясно тогда, что из $w \in L(J)$, то есть из

$$\langle w, q_0, e\rangle \Rightarrow_J^* \langle e, q, e\rangle$$

следует, что

$$\langle w, q_0, e\rangle \Rightarrow_J^n \langle e, q, e\rangle$$

для некоторого $n \leq 2|w| + 1$. Зафиксируем это n.

Рассмотрим всевозможные последовательности вида:

$$\langle w, q_0, e\rangle, \langle v_1, q, \alpha_1\rangle, \langle v_2, q, \alpha_2\rangle, \ldots, \langle v_n, q, \alpha_n\rangle,$$

где $v_i \in (A \bigcup \{e\})^*$, $\alpha_i \in (\Gamma \bigcup \{e\})^n$ для $i = 1, 2, \ldots, n$, $v_n = \alpha_n = e$, $n \leq 2|w| + 1$.

Ясно, что если $w \in L(G)$, то среди этого конечного числа последовательностей есть такая, что выполнены условия:

$$\langle w, q_0, e\rangle \Rightarrow_J \langle v_1, q, \alpha_1\rangle;$$

$$\langle v_i, q, \alpha_i\rangle \Rightarrow_J \langle v_{i+1}, q, \alpha_{i+1}\rangle$$

для $i = 1, 2, \ldots, n - 1$.

Верно и обратное. Таким образом, проверка $w \in L(J) = L(G)$ сводится к перебору конечного числа последовательностей, описанных выше, и проверке условий для каждой из этих последовательностей. ∎

Теорема 2.36. *Существует алгоритм, который по данной КС-грамматике $G = \langle A, Q, S, P\rangle$ определяет, пуст ли язык $L(G)$.*

Доказательство. По теореме 2.10 можно эффективно преобразовать G к приведенной КС-грамматике $G' = \langle A, Q', S, P'\rangle$, что $L(G') = L(G)$.

Очевидно, что пустота $L(G) = L(G')$ равносильна пустоте P'. ∎

Рассмотрим теперь проблему пустоты пересечения двух КС-языков. Выясняется, что она неразрешима. Точнее, если G и G' — две КС-грамматики, то проблема определения, пусто ли $L(G) \bigcap L(G')$ неразрешима.

Не обладая точным определением алгоритма, мы не можем доказать неразрешимость, то есть несуществование алгоритма решения задачи. Точное определение алгоритма дано в главе 4. В главе 5 доказана неразрешимость некоторых проблем. В частности, в главе 5 доказывается, что проблема соответствий Поста (ПСП) неразрешима. В пределах данной главы неразрешимость всех проблем будет доказываться сведением проблемы соответствий Поста к данным проблемам. Таким образом, если данная проблема оказалась бы разрешимой, то таковой была бы и ПСП, а это не так.

Сформулируем проблему соответствий Поста.

ПСП. *Дан алфавит A и конечное множество пар слов*

$$U = \{(a_1, b_1), \dots (a_n, b_n)\}$$

в этом алфавите. Пару слов $(a, b) \in (A^*)^2$ будем называть U-парой, если либо $(a, b) \in U$, либо существуют такие пары $(a', b') \in U$ и $(a'', b'') \in (A^*)^2$, что $a = a'a''$, $b = b'b''$, и (a'', b'') является U-парой.

Для заданных конечного множества U пар слов в алфавите A определить, существует ли такое $a \in A^*$, что пара (a, a) является U-парой.

Ясно, что по каждой паре слов $(a, b) \in (A^*)^2$ можно определить, является ли (a, b) U-парой для данного U. Однако, как будет показано в главе 5, не существует алгоритма, который по произвольному конечному множеству пар U определяет, разрешима ли ПСП для данного U.

Теорема 2.37. *Неразрешима проблема определения по данным двум КС-грамматикам $G = \langle A, Q, S, P\rangle$ и $G' = \langle A', Q', S', P'\rangle$, пусто ли пересечение $L(G) \bigcap L(G')$.*

Доказательство. Рассмотрим произвольный алфавит A_1 и произвольное конечное множество пар слов U в алфавите A_1. Пусть

$$U = \{(a_1, b_1), \dots, (a_n, b_n)\}. \tag{4}$$

Рассмотрим такие новые попарно различные буквы $x_1, x_2, \dots, x_n$, что

$$(\{x_1, \dots, x_n\} \bigcap A_1) = \emptyset.$$

Пусть

$$A = A' = (A_1 \bigcup \{x_1, \dots, x_n\}), \quad Q = \{S\}, \quad Q' = \{S'\}.$$

Рассмотрим два множества правил P и P':

$$P = \{S \to a_i S x_i,\ S \to a_i x_i \mid\ i = 1, \dots, n\};$$

$$P' = \{S' \to b_i S' x_i,\ S' \to b_i x_i \mid\ i = 1, \dots, n\}.$$

Очевидно, что если

$$a_{i_1} a_{i_2} \dots a_{i_m} =\ b_{i_1} b_{i_2} \dots b_{i_m}, \tag{5}$$

то
$$v = a_{i_1} a_{i_2} \dots a_{i_m} x_{i_m} x_{i_{m-1}} \dots x_{i_2} x_{i_1} \in (L(G) \bigcap L(G')).$$
Наоборот, если $(L(G) \bigcap L(G')) \neq \emptyset$ и $v \in (L(G) \bigcap L(G'))$, то существуют $m,\ i_1, i_2, \dots, i_m$ такие, что (5) выполняется.

Итак, если бы проблема
$$(L(G) \bigcap L(G')) = \emptyset?$$
была бы разрешима, то была бы разрешима и ПСП, что неверно. ∎

Оказывается, что для КС-грамматик неразрешима и проблемы эквивалентности и тотальности.

Теорема 2.38. *Не существует алгоритма, который, по данной КС-грамматике $G = \langle A, Q, S, P \rangle$ определял бы, выполняется ли равенство $L(G) = A^*$.*

Доказательство. Рассмотрим вновь произвольный алфавит A_1 и произвольное конечное множество пар слов U в этом алфавите. Пусть выполняется (4). Рассмотрим вновь КС-грамматики G и G' и языки $L(G)$ и $L(G')$. Наша цель — показать, что их дополнения тоже являются КС-языками. Ясно, что объединение этих дополнений совпадает с A^* тогда и только тогда, когда пересечение $L(G) \bigcap L(G')$ пусто.

Заметим, что КС-грамматика $\langle A, Q_3, S_3, P_3 \rangle$, где для $S_3 \notin (Q_1 \bigcup Q_2)$
$$Q_3 = \{S_3\} \bigcup Q_1 \bigcup Q_2,\ P_3 = P_1 \bigcup P_2 \bigcup \{S_3 \to S_1,\ S_3 \to S_2\},$$
задает $L(G_1) \bigcup L(G_2)$ для $G_1 = \langle A, Q_1, S_1, P_1 \rangle$ и $G_2 = \langle A, Q_2, S_2, P_2 \rangle$.

Итак, для завершения доказательства теоремы осталось построить КС-грамматики, задающие $A^* \setminus L(G)$ и $A^* \setminus L(G')$. В силу произвольности U, достаточно рассмотреть только первый язык. Достаточно построить МП-автомат, задающий $A^* \setminus L(G)$. Мы предлагаем это сделать читателю в качестве упражнения. Прокомментируем только идею его работы. Автомат J стремится доказать, что слово w на входе не имеет вида
$$a_{i_1} a_{i_2} \dots a_{i_m} x_{i_m} x_{i_{m-1}} \dots x_{i_2} x_{i_1}.$$
Фактически он устроен так, что пытается показать, что входное слово имеет описанный вид, и если это так, то завершает работу в состоянии, которое не является заключительным. В случае же любой неудачи, автомат освобождает стек, дочитывает до конца входную ленту, и заканчивает работу в заключительном состоянии.

Автомат J имеет четыре состояния. В начальном состоянии он кладет некоторый новый символ, называемый дном стека, в стек и переходит в следующее состояние, в котором переписывает побуквенно входное слово в свой стек пока не встретит одну из букв x_i. После этого автомат J переходит в третье состояние, в котором он, видя входную букву x_i, если верхушка стека есть a_i, съедает эту букву и эту верхушку стека и остается в этом третьем состоянии. Если в третьем состоянии обозревается входная буква x_i и при этом некоторая верхушка стека не является суффиксом a_i, хотя и имеет длину, не превосходящую длины этого слова a_i, либо длина стека (до его дна) меньше длины слова a_i, то автомат переходит в четвертое состояние, которое является единственным заключительным состоянием этого автомата и в котором автомат побуквенно съедает как входное слово, так и магазинное слово. Если же в третьем состоянии автомат J видит в качестве входной буквы одну из букв алфавита A_1, он тоже переходит в четвертое состояние. ∎

Следствие 2.39. *(1) Не существует алгоритма, который по произвольной КС-грамматике G и регулярному выражению r определяет, совпадает ли $L(G)$ с $L(r)$.*

(2) Проблема эквивалентности КС-грамматик неразрешима.

Доказательство. (1) следует из теоремы 2.38, ибо A^* — регулярное множество.

(2) следует из (1) и теоремы 2.25 о том, что по всякому регулярному выражению r можно (причем, конечно, эффективно) построить КС-грамматику G такую, что $L(r) = L(G)$. ∎

Некоторые другие разрешимые и неразрешимые проблемы приведены в упражнениях.

Упражнения

Упражнение 2.7.1. *Докажите теорему 2.38. В частности, формально задайте автомат J и докажите корректность его работы.*

Упражнение 2.7.2. *Покажите, что для КС-грамматики неразрешима проблема регулярности задаваемого ею языка.*

Упражнение 2.7.3. *Покажите, что для КС-грамматики неразрешима проблема, является ли дополнение языка данной грамматики КС-языком.*

Упражнение 2.7.4. *Докажите разрешимость проблемы пустоты пересечения двух языков, один из которых задан КС-грамматикой, второй — регулярным выражением.*

2.8 Свойства замкнутости для КС-языков

Как обычно, для доказательства того, что некоторая операция над языками не выводит за класс КС-языков, мы будем указывать явный способ построения по КС-грамматикам, задающим аргументы операции, КС-грамматики, задающей результат.

Теорема 2.40. *Класс КС-языков замкнут относительно операций объединения, конкатенации и итерации языков.*

Доказательство. Все три утверждения очевидны. Действительно, пусть $G_1 = \langle A, Q_1, S_1, P_1 \rangle$ и $G_2 = \langle A, Q_2, S_2, P_2 \rangle$ — две КС-грамматики. Без ограничения общности можно считать, что $Q_1 \bigcap Q_2 = \emptyset$. Тогда язык $L(G) = (L(G_1) \bigcup L(G_2))$ задается грамматикой:

$$\langle A, Q_1 \bigcup Q_2 \bigcup \{S\}, S, P_1 \bigcup P_2 \bigcup \{S \to S_1, S \to S_2\} \rangle.$$

Язык $L(G_1)L(G_2)$ задается грамматикой:

$$\langle A, Q_1 \bigcup Q_2 \bigcup \{S\}, S, P_1 \bigcup P_2 \bigcup \{S \to S_1 S_2\} \rangle.$$

Наконец, язык $L(G_1)^*$ задается грамматикой:

$$\langle A, Q_1 \bigcup \{S\}, S, P_1 \bigcup \{S \to SS_1, S \to e\} \rangle.$$

Везде выше $S \notin Q_1 \bigcup Q_2$. ∎

Теорема 2.41. *Класс КС-языков замкнут относительно операции пересечения с регулярными множествами.*

Доказательство. Пусть $G = \langle A, Q, S, P \rangle$ — произвольная КС-грамматика, $I = \langle A, Q'', q'', F'', R'' \rangle$ — произвольный конечный автомат (напомним, что согласно теореме 1.3 произвольное регулярное множество может быть задано некоторым конечным автоматом).

Наша цель — построить недетерминированный МП-автомат, распознающий язык $L(G) \bigcap L(I)$.

Согласно теореме 2.29, по G можно построить МП-автомат

$$J = \langle A, Q', \Gamma, q', F', P' \rangle,$$

что $L(G) = L(J)$. При этом

$$P' : ((A \bigcup \{e\}) \times Q' \times (\Gamma \bigcup \{e\})) \to (Q' \times \Gamma^*).$$

Будем строить МП-автомат

$$J' = \langle A, Q''', \Gamma''', q''', F''', P''' \rangle,$$

который будет моделировать "совместную"работу автоматов I и J. Идея состоит в том, чтобы, всякий раз, когда в автомате J с входной ленты считывается буква, осуществлялся также переход в соответствующее состояние в автомате I.

Формально, положим:

$$Q''' = (Q' \times Q''),\ \Gamma''' = \Gamma,\ q''' = \langle q', q'' \rangle,\ F''' = F' \times F'',$$

а P''' определим так. Для произвольных $q' \in Q'$, $q'' \in Q''$, $\epsilon \in \Gamma \bigcup \{e\}$, $\alpha \in \Gamma^*$, $a \in A$:

если $\langle q', \alpha \rangle \in P'(e, q', \epsilon)$, то включим пару $\langle \langle q', q'' \rangle, \alpha \rangle$ в $P'''(e, \langle q', q'' \rangle, \epsilon)$;

если $\langle q', \alpha \rangle \in P'(a, q', \epsilon)$ и $\langle q'', a, q'' \rangle \in R''$, то включим пару $\langle \langle q', q'' \rangle, \alpha \rangle$ в $P'''(a, \langle q', q'' \rangle, \epsilon)$.

Легко видеть, что $L(J') = L(I) \bigcap L(J)$, чем доказана теорема. ∎

К сожалению, относительно других естественных операций класс КС-языков оказывается незамкнутым. Основные из этих операций — пересечение и разность.

В действительности уже из материала параграфа 2.7 можно заключить, что эти операции плохи для КС-языков.

Действительно, рассмотрим алфавит A. Для произвольного языка L в алфавите A можно рассмотреть операцию дополнения

$$\overline{L} = (A^* \setminus L).$$

Пусть G — КС-грамматика, r — регулярное выражение. Мы знаем, что $\overline{L(r)}$ является регулярным языком. Если бы мы могли эффективно строить $\overline{L(G)}$, то можно было бы свести проблему эквивалентности G и r к проблеме пустоты $\overline{L(G)} \bigcap L(r)$ и $\overline{L(r)} \bigcap L(G)$ в силу теоремы 2.41 настоящего параграфа. Однако исходная проблема совпадения $L(G)$ и $L(r)$ неразрешима согласно следствию 2.39.

Таким образом, класс КС-языков не замкнут относительно операции разности с регулярными языками по меньшей мере в эффективном смысле.

В действительности мы увидим, что дополнение КС-языка не обязано быть КС-языком, то есть класс КС-языков не замкнут относительно операции разности с регулярными языками и в обычном смысле.

Но для этого нам необходим инструмент доказательства того, что некоторый язык не является КС-языком. Этим инструментом является теорема о разрастании для КС-языков, сформулированная и доказанная ниже как теорема 2.43.

2.9 Теорема о разрастании для КС-языков

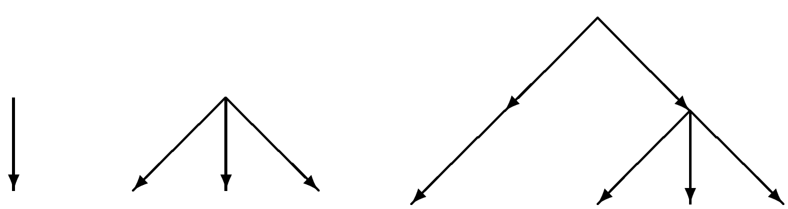
Для доказательства теоремы нам понадобится понятие дерева вывода слова в грамматике.

Пусть $G = \langle A, Q, S, P \rangle$ — КС-грамматика, $w \in L(G)$.

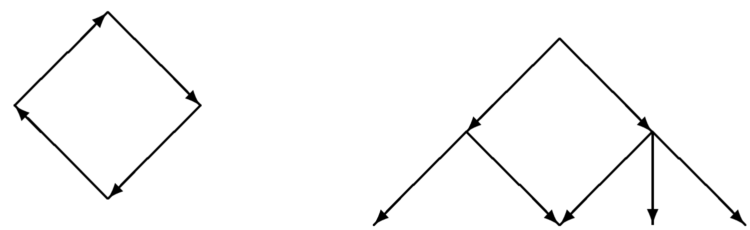
Определение 2.9.1 (дерева). *Деревом называется ориентированный граф, все вершины которого достижимы, причем единственным образом, из одной фиксированной вершины, называемой корнем дерева. Вершины, из которых не исходит ни одной дуги, называются висячими вершинами или листьями. Если дуга исходит из некоторой вершины a в некоторую вершину b, то b называется сыном, а a — отцом. a называется предком, а b — потомком, если существует натуральное n и последовательность $c_1, c_2, \ldots, c_n$ вершин дерева такая, что выполнены условия:*

1. $a = c_1$;
2. $b = c_n$;
3. c_i *является отцом* c_{i+1} *для* $i = 1, 2, \ldots, n-1$.

Следующие графы являются деревьями.



Следующие графы не являются деревьями.



Примем соглашение, что деревья изображаются на плоскости сверху вниз (то есть корень находится сверху), как в приведенных примерах. При этом соглашении на вершинах дерева устанавливается порядок, определенный так:

Определение 2.9.2 (порядка вершин дерева). *Вершина a дерева меньше вершины b, если имеет место одно из трех:*

1. a является предком b;

2. Существует вершина c, что a и b являются сыновьями c и a расположен левее b;

3. Существуют вершины c, d, e, что a и b являются потомками c и d соответственно, c, d являются сыновьями e и c расположен левее d.

Определение 2.9.3 (дерева вывода). *Деревом псевдовывода в КС-грамматике $G = \langle A, Q, S, P\rangle$ называется дерево, все невисячие вершины которого помечены нетерминалами из Q, а все висячие вершины (листья) — терминалами из A, пустым словом или нетерминалами из Q, причем корень дерева помечен аксиомой S. При этом если*

$$\epsilon_1, \epsilon_2, \ldots, \epsilon_r \in (A \bigcup Q \bigcup \{e\})$$

являются метками всех сыновей некоторой вершины с меткой $\theta \in Q$, взятые в описанном выше порядке, то

$$\theta \to \epsilon_1\epsilon_2\ldots\epsilon_r \in P.$$

Метки всех листьев псевдодерева вывода, считанные с дерева в указанном выше порядке и сконкатенированные в слово из $(A \bigcup Q)^*$, называются кроной дерева псевдовывода.

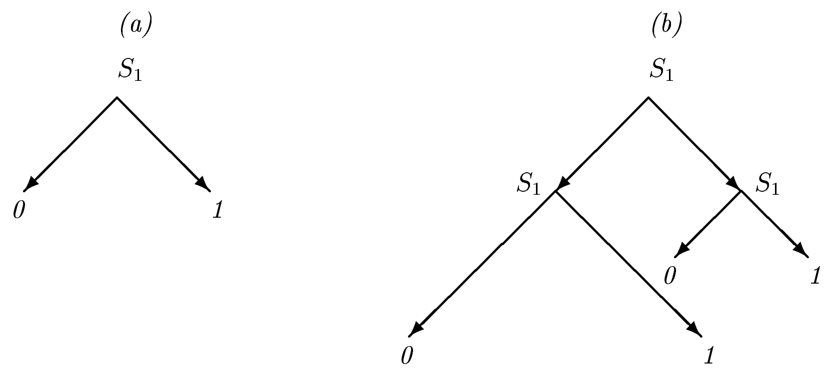
Дерево псевдовывода в КС-грамматике $G = \langle A, Q, S, P\rangle$ называется деревом вывода, если все его висячие вершины (листья) помечены терминалами из A или пустым словом (но не нетерминалами).

Приведем пример.

Пример 2.9.1. *Пусть* $G_1 = \langle \{0,1\}, \{S_1\}, S_1, P_1 \rangle$, *где* P_1 *— следующее множество правил:*

$$S_1 \to 01, \quad S_1 \to S_1 S_1.$$

Тогда, например, следующие объекты являются деревьями выводов в грамматике G_1 *:*



Говорят, что дерево вывода в грамматике G является *деревом вывода слова* $w \in A^*$, если w является кроной этого дерева вывода. Так, дерево (a) в примере 2.9.1 является деревом вывода слова 01 в грамматике G_1, а дерево (b) — деревом вывода слова 0101 в G_1.

Теорема 2.42. $w \in L(G)$ *тогда и только тогда, когда существует дерево вывода* w *в* G.

Доказательство. Пусть $G = \langle A, Q, S, P \rangle$ — КС-грамматика и пусть t — дерево вывода слова $w \in A^*$. Будем строить последовательность псевдодеревьев вывода $t_0, t_1, t_2, \ldots$ и последовательность слов $w_0, w_1, \ldots$ из $(A \bigcup Q)^*$ так. Обозначим t через t_0 и w через w_0. Если дерево псевдовывода t_i и слово w_i уже построены и t_i содержит более одной вершины, то рассмотрим произвольную вершину t_i, все потомки которой являются листьями. Отбросив все потомки этой вершины, получим дерево псевдовывода t_{i+1}. Его крону обозначим w_{i+1}. Из определений видно, что $w_{i+1} \Rightarrow_G^1 w_i$. Отметим, что на этом шаге число вершин в псевдодереве уменьшилось. Если на некотором шаге построенное t_{i+1} состоит только из корня, то ясно, что $w_{i+1} = S$. По построению, w_{i+1}, $w_i, \ldots$, w_0 — вывод слова w в грамматике G.

Наоборот, пусть $S = w_0$, $w_1, \ldots$, $w_n = w$ — вывод слова $w \in A^*$ в грамматике G. Построим последовательность t_0, $t_1, \ldots$, t_n псевдодеревьев вывода в грамматике G так, что t_n — дерево вывода и что w_i есть крона t_i для всех $i = 0, 1, \ldots, n$. Через t_0 обозначим дерево, состоящее из одного только корня с меткой S. Пусть t_i уже построено, и крона t_i есть w_i. По определению, w_{i+1} получается из w_i путем замены некоторого вхождения некоторого нетерминала R в слово w_i на такое слово $v \in (A \bigcup Q)^*$, что $R \to v \in P$. Пусть

$$v = \epsilon_1 \epsilon_2 \ldots \epsilon_k,$$

где

$$\epsilon_1, \epsilon_2, \ldots, \epsilon_k \in (A \bigcup Q).$$

Найдем в псевдодереве t_i лист, соответствующий тому самому вхождению, которое заменялось (такой лист есть, ибо w_i — крона t_i). Построим дерево t_{i+1} из дерева t_i, приписав к найденному листу сыновей с метками

$$\epsilon_1, \epsilon_2, \ldots, \epsilon_k$$

слева направо. Если же $k = 0$, то есть $v = e$, то припишем к этому листу единственного сына с меткой e. Получим по определению дерево псевдовывода с кроной w_{i+1}. Очевидно, что крона t_n есть $w_n = w$, то есть t_n есть дерево вывода слова w в грамматике G. ∎

Определение 2.9.4 (пути и высоты дерева). *Путем в дереве называется произвольная последовательность его вершин $\epsilon_1, \epsilon_2, \ldots, \epsilon_m$ такая, что ϵ_i является отцом ϵ_{i+1} для $i = 1, \ldots, m-1$. Высотой дерева называется максимальная длина пути в дереве.*

В примере 2.9.1 высоты деревьев (a) и (b) равны 1 и 2.

Теперь мы готовы доказать главный результат параграфа:

Теорема 2.43 (о разрастании для КС-языков). *Пусть L — КС-язык. Существует такое натуральное число n, что если $w \in L$ и $|w| > n$, то w можно представить в виде $xyzuv$ так, что $x, y, z, u, v \in A^*$, $|yu| > 0$, $|yzu| \leq n$ и $xy^m zu^m v \in L$ для всех натуральных $m = 0, 1, 2, 3, \ldots$*

Доказательство. Пусть $G = \langle A, Q, S, P \rangle$ — такая КС-грамматика, что $L(G) = L$. Без ограничения общности можно считать, что G находится в нормальной форме Грейбах. Пусть k — максимальная длина правой части правил из G.

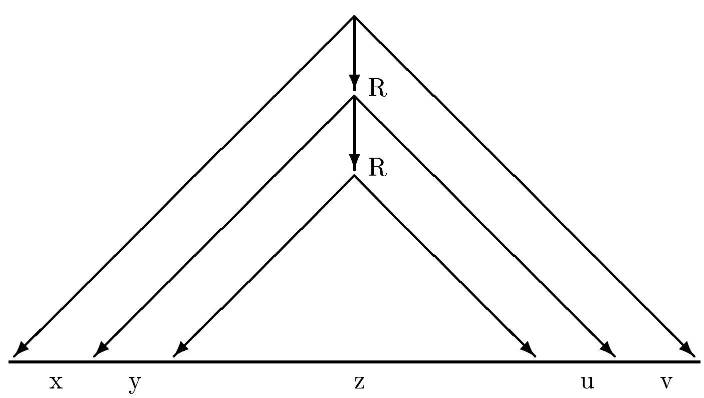
Пусть $w \in L$ и имеется дерево вывода w в G высоты, большей $|Q|+1$. По определению это значит, что существует путь в этом дереве вывода, длина которого превосходит $|Q|+1$. Рассмотрим произвольный конкретный путь длины $|Q|+2$, который завершается листом (очевидно, такой есть).

Первые $|Q|+1$ вершины на этом пути не висячие, следовательно, они помечены нетерминалами из Q. Ясно, что некоторый нетерминал $R \in Q$ встречается в этом пути дважды. Это изображено на приводимом ниже рисунке.

Обозначим, как указано на рисунке, через z слово, считанное с листьев поддерева, порожденного вторым вхождением нетерминала, через yzu — слово, считанное с листьев поддерева, порожденного первым вхождением нетерминала, а через x и v — соответственно начало и конец слова, то есть то, что считано со всего дерева вывода w в G, соответственно, левее и правее поддерева, порожденного первым вхождением нетерминала R.

Покажем, что выбранные x, y, z, u, v удовлетворяют условиям теоремы, если положить $n = k^{|Q|+1} + 1$.

Действительно, то, что $|yzu| \leq n = k^{|Q|+1} + 1$, следует из мощностных соображений (на каждом шаге можно ветвиться не более, чем на k ветвей, а шагов не более $|Q| + 1$).



То, что $|yu| > 0$, следует из того, что G находится в нормальной форме Грейбах (то есть правая часть всякого правила начинается с терминала, значит, $y \neq e$).

С другой стороны, можно укоротить дерево, склеив вершины, соответствующие первому и второму вхождениям R в наш путь. Тогда мы получим дерево вывода слова xzv в G.

Затем можно вырезать поддерево, порожденное вторым вхождением R и на его место вклеить поддерево, порожденное первым вхождением R. Получим дерево

Очевидно, что это дерево является деревом вывода в G слова

$$xy^2zu^2v.$$

Таким образом, $xy^2zu^2v \in L(G) = L$. Аналогично показывается, что $xy^mzu^mv \in L(G) = L$ для любого натурального m.

Итак, утверждение теоремы доказано для $n = k^{|Q|+1} + 1$ в предположении, что для w есть дерево вывода высоты, большей $|Q| + 1$. Пусть w — произвольное слово длины $|w| > n = k^{|Q|+1} + 1$. Очевидно, что не существует дерева вывода w в G высоты, меньшей или равной $|Q| + 1$ (из комбинаторных соображений). ∎

Следствие 2.44. *Язык*

$$\{a^ib^ic^i \mid \quad i = 1, 2, \dots\}$$

не является контекстно-свободным.

Доказательство. Разбором случаев легко показать, что никакое слово вида $a^mb^mc^m$ с положительным m нельзя представить в виде $xyzuv$ так, чтобы $|yu| > 0$ и $xy^2zu^2v \in L$, то есть представлялось в виде $a^{m'}b^{m'}c^{m'}$. ∎

Следствие 2.45. *Класс КС-языков не замкнут относительно операции пересечения.*

Доказательство. Легко показать, что языки

$$\{a^ib^ic^j \mid \quad i, j = 1, 2, \dots\}, \quad \{a^jb^ic^i \mid \quad i, j = 1, 2, \dots\}$$

контекстно-свободны.

Однако язык

$$\{a^ib^ic^i \mid \quad i = 1, 2, \dots\}$$

не является КС-языком, как показано в следствии 2.44. ∎

Следствие 2.46. *Класс КС-языков не замкнут относительно левой разности с регулярными множествами и дополнения.*

Доказательство. Операция дополнения выражается через операцию левой разности языка всевозможных слов в алфавите A и данного языка:

$$\overline{L} = (A^* \setminus L).$$

A^* — регулярный язык. Значит, если класс КС-языков не замкнут относительно операции дополнения, то он не замкнут и относительно левой разности с регулярными множествами.

С другой стороны, очевидно тождество:

$$(L_1 \bigcap L_2) = \overline{\overline{L_1} \bigcup \overline{L_2}}.$$

Так как класс КС-языков замкнут относительно операции объединения, то из его замкнутости относительно дополнения и из предыдущего тождества следовала бы его замкнутость относительно пересечения, а это противоречило бы следствию 2.45. ∎

Упражнения

Упражнение 2.9.1. *Покажите, что язык*

$$\{0^i 1^i 0^i \mid \ i = 1, 2, \ldots\}$$

не является КС-языком.

Упражнение 2.9.2. *Покажите, что всякий КС-язык в однобуквенном алфавите регулярен.*

Упражнение 2.9.3 (*). *Приведите пример языка, не являющегося контекстно-свободным, но удовлетворяющего теореме о разрастании для КС-языков.*

Предметный указатель