6 КОНТЕКСТНО СВОБОДНЫЕ ГРАММАТИКИ

Начнем с определения контекстно-свободных грамматик.

Определение 6.0.1 (КС-грамматики). *Контекстно-свободная грамматика (сокращенно КС-грамматика) G — это четверка $\langle \Sigma, Q, S, P \rangle$, где:*

1. *Σ — (конечный) алфавит. Σ 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 $R \in Q$, $u \in (\Sigma \cup Q)^*$, 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 $(\Sigma \cup Q)$.

Определение 6.0.2 (вывода в КС-грамматике). *Пусть $v, w \in (\Sigma \cup Q)^*$ и $G = \langle \Sigma, Q, S, P \rangle$ — КС-грамматика. Будем говорить, что слово v выводимо из слова w за один шаг в грамматике, и писать*

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

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

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

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 $x \in (\Sigma \cup Q)^*$, *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 $w \in \Sigma^*$ в G будем называть конечную последовательность слов $w_0, w_1, w_2, \ldots, w_n \in (\Sigma \cup Q)^*$ 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 $i = 0, \ldots, n-1$

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

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

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

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

Приведем пример.

Пример 6.0.1. 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 6.1. *Пусть* $G = \langle \Sigma, Q, S, P\rangle$ *— КС-грамматика,*

$$u, v, u', v' \in (\Sigma \cup Q)^*,$$

$$u \Rightarrow_G^n u',$$

$$v \Rightarrow_G^m v'.$$

Тогда $uv \Rightarrow_G^{n+m} u'v'$.

Доказательство. Существуют выводы:

$$u \Rightarrow_G u_1 \Rightarrow_G \cdots \Rightarrow_G u_n = u', \;\; v \Rightarrow_G v_1 \Rightarrow_G \cdots \Rightarrow_G v_m = v'.$$

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 $x, z \in (\Sigma \cup Q)^*$ и $y \in Q$, что $v_1 = xwz$ для некоторого такого $w \in (\Sigma \cup Q)^*$, 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 $x, z \in (\Sigma \cup Q)^*$ и $y \in Q$, что $u_1 = xwz$ для некоторого такого $w \in (\Sigma \cup Q)^*$, что $y \to w \in P$. Ясно, что тогда $u_0 v = xyzv \Rightarrow_G xwzv = u_1 v$. Но $u_1 \Rightarrow_G^{n-1} u_n = u'$, значит, по индукции, $u_1 v \Rightarrow_G^{n+m-1} u'v'$. Сопоставляя, получим $uv = u_0 v \Rightarrow_G^{n+m} u'v'$. ∎

Приведём ещё один пример.

Пример 6.0.2. Пусть $G = \langle \{0, 1, \}, \{S\}, S, P\rangle$ — КС-грамматика, где P — следующее множество правил:

$$S \to 0S1, \; S \to 01, \; S \to SS.$$

В дальнейшем мы будем писать короче:

$$S \to 0S1 \mid 01 \mid SS.$$

Определение 6.0.4. *Рассмотрим непустое слово* $w = a_1a_2\dots a_m$, *составленное из нулей и единиц. Через* $o_i(w)$ *обозначим число нулей в слове* $a_1a_2\dots a_i$, *а через* $z_i(w)$ *— число единиц в слове* $a_1a_2\dots a_i$. *В непустом слове* w, *составленном из нулей и единиц, скобки расставлены правильно, если*

$o_{|w|}(w) = z_{|w|}(w)$

и для каждого $i = 1, 2, \dots, |w|$ *выполняется следующее условие:*

$o_i(w) \geqslant z_i(w)$.

Здесь, как обычно, $|w|$ *обозначает длину (число букв) слова* w. *Если* $w = a_1a_2\dots a_m$, *то* $|w| = m$.

Теорема 6.2. $L(G)$ *состоит из тех и только тех слов, в которых скобки расставлены правильно.*

Доказательство. Докажем сначала, что любое непустое слово w, в котором скобки расставлены правильно, выводимо в грамматике G. Это будем доказывать индукцией по длине слова w. Если слово имеет длину 2, оно есть 01 и его вывод есть S, 01.

Каждое непустое слово, в котором скобки расставлены правильно, начинается на 0 и оканчивается на 1.

Если $o_i(w) > z_i(w)$ для каждого $i = 1, 2, \dots, |w| - 1$, то либо $w = 01$, либо $w = 0a1$, а в непустом слове a скобки расставлены правильно. В самом деле,

$$o_i(a) = o_{i+1}(w) - 1 \geqslant z_{i+1}(w) = z_i(a)$$

для $i = 1, \dots, |a| - 1$ и

$$o_{|a|}(a) = o_{|a|+1}(w) - 1 = o_{|w|}(w) - 1 = z_{|w|}(w) - 1 = z_{|a|}(a).$$

По индукционному предположению, слово a имеет вывод

$$S,\ d_1, \dots, d_s = a.$$

Поэтому

$$S,\ 0S1,\ 0d_11, \dots, 0d_s1 = 0a1 = w$$

есть вывод w.

Если же $o_i(w) = z_i(w)$ для некоторого $i \in \{1, 2, \dots, |w| - 2\}$, то $w = w_1w_2$, $|w_1| = i$, а в непустых словах w_1 и w_2 скобки расставлены правильно. Действительно,

$$o_i(w_2) = o_{i+|w_1|}(w) - o_{|w_1|}(w_1) \geqslant z_{i+|w_1|}(w) - z_{|w_1|}(w_1) = z_i(w_2).$$

При этом

$$o_{|w_2|}(w_2) = o_{|w_2|+|w_1|}(w) - o_{|w_1|}(w_1) = z_{|w_2|+|w_1|}(w) - z_{|w_1|}(w_1) = z_{|w_2|}(w_2).$$

По индукционному предположению, имеются выводы

$$S,\ d_1, \dots, d_s = w_1, \quad S,\ b_1, \dots, b_r = w_2.$$

По лемме 6.1, имеется вывод $SS, \dots, w$ слова w из SS. Но тогда $S,\ SS, \dots, w$ является выводом w в грамматике G .

Покажем теперь, что в каждом выводимом слове скобки расставлены правильно. Для этого индукцией по длине вывода из S докажем, что после вычеркивания всех букв S каждое слово вывода оказывается либо пустым, либо словом, в котором скобки расставлены правильно.

Следующее слово вывода после такого вычеркивания либо совпадает с предыдущим, либо получается из предыдущего вставкой 01 вначале, в конец слова или между двумя какими-то буквами. Понятно, что такая вставка не нарушает правильную расстановку скобок. ∎

Приведём ещё один пример.

Пример 6.0.3. Рассмотрим язык

$$L = \{0^i1^i \mid i = 1, 2, \dots\}$$

в алфавите $\{0, 1\}$. Этот язык не является автоматным. Для доказательства нам понадобится одно свойство автоматных языков, называемое *теоремой о разрастании.*

Как выясняется, класс регулярных множеств довольно беден. Однако внимание к этому классу объясняется, в частности, тем, что регулярные языки можно очень быстро и просто распознавать при помощи конечных автоматов.

Когда мы имеем дело с каким-то языком (заданным некоторым способом), нам важно знать, регулярен ли он. Для доказательства регулярности достаточно записать задающее язык регулярное выражение. Однако у нас пока нет способа доказывать нерегулярность языка.

Следующая лемма дает нам мощный инструмент для доказательства нерегулярности языков. Фактически эта лемма описывает строение автоматного языка.

Лемма 6.3 (о разрастании). *Пусть L — автоматный язык в алфавите Σ. Существует такое натуральное число $k > 0$, что всякое такое слово $w \in L$, что $|w| > k$, можно представить в виде $w = xyz$, где x, y и z — слова в алфавите Σ, так что выполняются следующие условия:*

1. $0 \leqslant |x| < k$;
2. $0 < |y| < k$;
3. $xy^i z \in L(r)$ *для всякого* $i = 0, 1, 2, \ldots$.

Доказательство. Пусть $M = \langle \Sigma, Q, q_0, Q_{\text{закл}}, P \rangle$ — задающий L детерминированный конечный автомат.

Положим $k = |Q| + 1$, где $|Q|$ обозначает число элементов Q,

$$Q = \{q_0, q_1, \ldots, q_{|Q|-1}\}$$

.

Пусть $w \in L = L(M)$, $|w| > k$. Рассмотрим работу автомата M на слове w. Пусть при этой работе автомат проходит состояния

$$q_0 = q_{i_1}, q_{i_2}, \ldots, q_{i_k}, q_{i_{k+1}}, \ldots, q_{i_t},$$

где $t = |w| + 1$, $q_{i_t} \in Q_{\text{закл}}$. Значит, среди $i_1, i_2, \ldots, i_k$ найдутся два одинаковых индекса, поскольку $0 \leqslant i_j \leqslant |Q| - 1$, а $k = |Q| + 1$.

Пусть, например, $i_j = i_m$, $1 \leqslant j < m \leqslant k$. Представим слово w в виде xyz, где x имеет длину $j-1$, y — длину $m-j$, z — длину $t-m$.

Покажем, что такое представление удовлетворяет условиям леммы. Действительно,

$$0 < |y| < m \leqslant k = |Q| + 1,$$

т.е. условие (2) леммы выполнено. Затем,

$$0 \leqslant |x| < j < m \leqslant k = |Q| + 1,$$

т.е. условие (1) леммы также выполнено. Далее, видно, что автомат M распознает и слово xz, и все слова вида xy^iz. Для слова xz автомат M в силу детерминизма пройдет последовательность состояний

$$q_0 = q_{i_1}, \dots, q_{i_j}, q_{i_{m+1}}, \dots, q_{i_t}.$$

Для слова xy^iz автомат M в силу детерминизма пройдет

$$q_0 = q_{i_1}, \dots, q_{i_j},$$

затем i раз последовательность состояний

$$q_{i_{j+1}}, q_{i_{j+2}}, \dots, q_{i_m} = q_{i_j},$$

и затем последовательность

$$q_{i_{m+1}}, \dots, q_{i_t}.$$

∎

Применим лемму для доказательства неавтоматности языков.

Следствие 6.4. *Язык*

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

не является автоматным.

Доказательство. Пусть, напротив, язык L автоматен. По лемме, существует тогда константа $k > 0$, что всякое такое слово $w \in L$, что $|w| > k$, можно представить в виде $w = xyz$, где $0 < |y| < k$, что, в частности, слово $xyyz$ тоже принадлежит L.

Рассмотрим отдельно три случая.

1. y составлено только из нулей. Тогда слово $xyyz$ содержит больше нулей, чем единиц, и не принадлежит L по определению.

2. y составлено только из единиц. Аналогично предыдущему.

3. y составлено и из нулей, и из единиц. Тогда, так как xyz содержит в начале только нули, а потом только единицы, то y начинается с 0 и кончается на 1. Значит, слово $xyyz$ содержит подслово 10. Но тогда слово $xyyz$ не принадлежит L, ибо никакое слово с подсловом 10 не принадлежит L.

Во всех трех случаях получили противоречие.

Другое доказательство. Возьмём слово $0^{2k}1^{2k}$. Это слово лежит в L и длина его больше k. По теореме о разрастании, это слово можно представить в виде xyz так, что $0 < |y| < k$, $|x| < k$ и слово $xyyz$ тоже принадлежит L. Так как $|xy| < 2k$, то y состоит только из нулей. Поэтому в слове $xyyz$ нулей больше, чем единиц, и $xyyz$ не принадлежит L по определению. ∎

С другой стороны, язык

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

задаётся грамматикой

$$S \to 01 \mid 0S1.$$

В самом деле, если слово выводимо в этой грамматике, оно имеет вид S, 0^iS1^i или 0^i1^i. Это легко доказывается индукцией по длине вывода. Наоборот, любое слово вида 0^i1^i выводимо в этой грамматике. Это легко доказывается индукцией по i.

Таким образом, класс кс-языков содержит неавтоматный язык.

Покажем, что каждый автоматный язык является кс-языком.

Теорема 6.5. *Если R — автоматный язык, то R — КС-язык.*

Доказательство. Пусть $M = \langle \Sigma, Q, q_0, Q_{\text{закл}}, P\rangle$ — задающий R детерминированный конечный автомат. Построим по нему такую КС-грамматику $G = \langle \Sigma, Q', S, P'\rangle$, что

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

Положим $Q' = Q$, $S = q_0$.

Без ограничения общности будем считать, что для всякого $q \in Q$ и всякого $a \in \Sigma$ 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 $q \in Q_{\text{закл}}$ включим в P' правило $q \to \lambda$, где λ — пустое слово.

Других правил в P' нет.

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

По определению, $w \in L(M)$ тогда и только тогда, когда существует такое $q'' \in Q_{\text{закл}}$, что $(q_0, w) \mid\Rightarrow^+_M (q'', e)$. С другой стороны, по определению, $w \in L(G)$ тогда и только тогда, когда $q_0 \Rightarrow^+_G w$. Это эквивалентно тому, что существует $q'' \in Q_{\text{закл}}$ такое, что $q_0 \Rightarrow^+_G wq''$, поскольку для всякого $q'' \in Q_{\text{закл}}$ и только для них $q'' \Rightarrow_G e$.

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

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

тогда и только тогда, когда

$$q \Rightarrow^+_G wq'.$$

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

Итак, пусть $(q, w) \mid\Rightarrow_M (q', e)$. По определению это значит, что $w \in \Sigma$ 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 $(q, w) \mid\Rightarrow_M (q', e)$.

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

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

1. $w = a_0 a_1 \ldots a_n$;
2. $q_{i_0} = q, \quad q_{i_{n+1}} = q'$;
3. $(q_{i_k}, a_k a_{k+1} \ldots a_n) \mid\Rightarrow_M (q_{i_{k+1}}, a_{k+1} a_{k+2} \ldots a_n)$ для всех таких k, что $0 \leqslant k \leqslant n$.

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

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

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

По лемме 6.1, тогда $q \Rightarrow_G^+ wq'$.

Аналогично показывается, что из

$$q \Rightarrow_G^+ wq'$$

следует

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

Теорема доказана. ∎

Приведём пример.

Пример 6.0.4. Рассмотрим язык L в алфавите $\{0,1\}$, содержащий те и только те слова, в которых чётное число единиц и ровно два нуля. Этот язык задаётся следующим конечным детерминированным автоматом M:

	0	1
начальное $q00$	$q10$	$q01$
заключительное $q20$	q	$q21$
q	q	q
$q10$	$q20$	$q11$
$q01$	$q11$	$q00$
$q21$	q	$q20$
$q11$	$q21$	$q10$

Кс-грамматика, задающая тот же язык, содержит правила:

$$q_{00} \to 0q_{10} \mid 1q_{01},$$
$$q_{20} \to 1q_{21} \mid \lambda,$$
$$q_{10} \to 0q_{20} \mid 1q_{11},$$
$$q_{01} \to 0q_{11} \mid 1q_{00},$$
$$q_{21} \to 1q_{20},$$
$$q_{11} \to 0q_{21} \mid 1q_{10},$$