Глава 4

КОНТЕКСТНО-СВОБОДНЫЕ ГРАММАТИКИ

§ 4.1. Упрощение контекстно-свободных грамматик

В этой главе мы опишем некоторые основные упрощения КС-грамматик и докажем несколько важных теорем о нормальных формах Хомского и Грейбах. Мы также покажем, что существуют алгоритмы для определения, является ли язык, порождаемый КС-грамматикой, пустым, конечным или бесконечным.

Будет определено так называемое свойство самовложенности КС-грамматик и показано, что КС-язык нерегулярен тогда и только тогда, когда каждая КС-грамматика, порождающая его, обладает этим свойством.

Наконец, мы рассмотрим специальные типы КС-грамматик, такие, как последовательные и линейные грамматики.

Формальное определение КС-грамматики допускает структуры, которые в некотором смысле являются “расточительными”. Например, словарь может включать нетерминалы, которые не могут использоваться в выводе хоть какой-нибудь терминальной цепочки; или в множестве правил не запрещено иметь такое правило, как $A \to A$. Мы докажем несколько теорем, которые показывают, что каждый КС-язык может порождаться КС-грамматикой специального вида. Более того, будут даны алгоритмы, которые для любой КС-грамматики находят эквивалентную КС-грамматику в одной из заданных форм.

Прежде всего мы докажем результат, который важен сам по себе. Будем предполагать, что КС-грамматики, рассматриваемые в этой главе, не содержат ε-правил.

Теорема 4.1. *Существует алгоритм для определения, является ли язык, порождаемый данной КС-грамматикой, пустым.*

Доказательство. Пусть $G = (V_N, V_T, P, S)$ — контекстно-свободная грамматика. Предположим, что $S \overset{*}{\Rightarrow} x$ для некоторой терминальной цепочки x. Рассмотрим дерево вывода, представляющее этот вывод. Предположим, что в этом дереве есть путь с узлами n_1 и n_2, имеющими одну и ту же метку A. Пусть узел n_1 расположен ближе к корню S, чем узел n_2 (рис. 4.1, *а*).

Поддерево с корнем n_1 представляет вывод $A \overset{*}{\Rightarrow} x_1$ цепочки x_1. Аналогично поддерево с корнем n_2 представляет вывод $A \overset{*}{\Rightarrow} x_2$ цепочки x_2. Заметим, что x_2 является подцепочкой цепочки x_1, которая, впрочем, может совпадать с x_1. Кроме того, цепочка $x = x_3x_1x_4$, где $x_3, x_4 \in \Sigma^*$, причем одна из них или обе могут быть пустыми цепочками. Если в дереве с корнем S мы заменим поддерево с корнем n_1 поддеревом с корнем n_2, то получим дерево (см. рис. 4.1, *б*), представляющее вывод $S \overset{*}{\Rightarrow} x_3x_2x_4$. Так мы исключили, по крайней мере, один узел (n_1) из исходного дерева вывода.

Если в полученном дереве имеется путь с двумя узлами, помеченными одним и тем же нетерминалом, процесс может быть повторен с деревом вывода $S \overset{*}{\Rightarrow} x_3x_2x_4$. Фактически процесс может повторяться до тех пор, пока в очередном дереве имеется путь, в котором находятся два узла, помеченных одинаково.

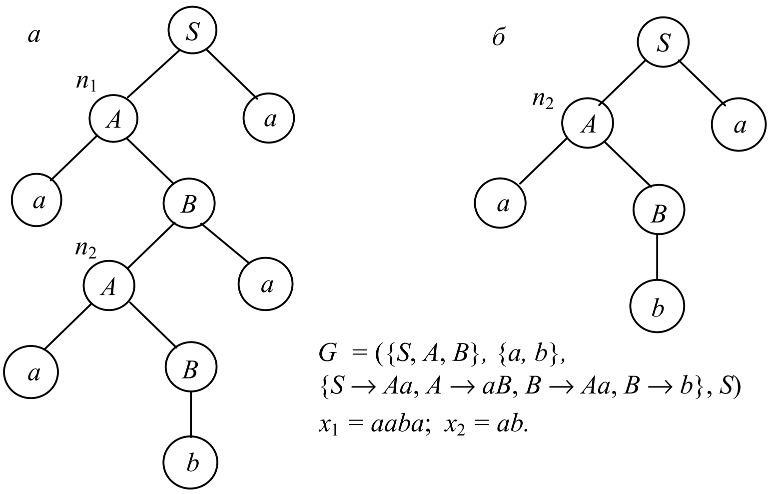

Рис. 4.1.

Поскольку каждая итерация исключает один узел или более, а дерево конечно, то процесс в конце концов закончится. Если в грамматике G имеется m нетерминалов, то в полученном дереве все ветви будут иметь длину (подразумевается, что длина ветви измеряется числом дуг, ее составляющих) не больше m, ибо в противном случае на длинной ветви неминуемо встретились бы два узла с идентичными метками.

Итак, если грамматика G порождает какую-нибудь цепочку вообще, то существует вывод (другой) цепочки, дерево которого не содержит ни одной ветви, длиннее m.

Алгоритм, определяющий, является ли язык $L(G)$ пустым, можно организовать следующим образом. Сначала надо построить коллекцию деревьев, представляющих выводы в грамматике G:

Шаг 1. Начать коллекцию с единственного дерева, представленного только корнем — узлом с меткой S.

Шаг 2. Добавить к коллекции любое дерево, которое может быть получено из дерева, уже имеющегося в коллекции, посредством применения единственного правила, если образующееся дерево не имеет ни одной ветви, длиннее m, и если такого еще нет в коллекции. Поскольку число таких деревьев конечно, то процесс в конце концов закончится.

Ш а г 3. Теперь язык $L(G)$ непуст, если в построенной коллекции есть хотя бы одно дерево, представляющее вывод терминальной цепочки. Иначе язык $L(G)$ пуст.

Требуемый алгоритм построен.

Существование алгоритма для определения, порождает ли данная КС-грамматика пустой язык, является важным фактом. Мы будем использовать его при упрощении КС-грамматик.

Как увидим в дальнейшем, никакого такого алгоритма для более сложных грамматик, например для контекстно-зависимых, не существует.

Теорема 4.2. *Для любой контекстно-свободной грамматики* $G = (V_N, V_T, P, S)$, *порождающей непустой язык, можно найти эквивалентную контекстно-свободную грамматику* G_1, *в которой для любого нетерминала* A *существует терминальная цепочка* x, *такая, что* $A \overset{*}{\underset{G_1}{\Rightarrow}} x$.

Д о к а з а т е л ь с т в о. Для каждого нетерминала $A \in V_N$ рассмотрим грамматику $G_A = (V_N, V_T, P, A)$. Если язык $L(G_A)$ пуст, то мы удалим A из алфавита V_N, а также все правила, использующие A в правой или левой части правила. После удаления из G всех таких нетерминалов мы получим новую грамматику: $G_1 = (V_N^1, V_T, P_1, S)$, где V_N^1 и P_1 — оставшиеся нетерминалы и правила. Ясно, что $L(G_1) \subseteq L(G)$, поскольку вывод в G_1 есть также вывод в G.

Предположим, что существует терминальная цепочка $x \in L(G)$, но $x \notin L(G_1)$. Тогда вывод $S \overset{*}{\underset{G}{\Rightarrow}} x$ должен включать сентенциальную форму вида $\alpha_1 A \alpha_2$, где $A \in V_N \setminus V_N^1$, т.е. $S \overset{*}{\underset{G}{\Rightarrow}} \alpha_1 A \alpha_2 \overset{*}{\underset{G}{\Rightarrow}} x$. Однако тогда должна существовать некоторая терминальная цепочка x_1, такая, что $A \overset{*}{\underset{G}{\Rightarrow}} x_1$, — факт, противоречащий предположению о том, что $A \in V_N \setminus V_N^1$. Что и требовалось доказать.

***Определение* 4.1.** Нетерминалы из V_N^1 принято называть *продуктивными.*

В дополнение к исключению нетерминалов, из которых невозможно вывести ни одной терминальной цепочки, мы можем также исключать нетерминалы, которые не участвуют ни в каком выводе.

Теорема 4.3. *Для любой данной контекстно-свободной грамматики, порождающей непустой язык L, можно найти контекстно-свободную грамматику, порождающую язык L, такую, что для каждого ее нетерминала A существует вывод вида* $S \overset{*}{\Rightarrow} x_1 A x_3 \overset{*}{\Rightarrow} x_1 x_2 x_3$, *где* $x_1, x_2, x_3 \in V_T^*$.

Д о к а з а т е л ь с т в о. Пусть $G_1 = (V_N, V_T, P, S)$ — произвольная cfg, удовлетворяющая условиям теоремы 4.2. Если $S \overset{*}{\underset{G_1}{\Rightarrow}} \alpha_1 A \alpha_2$, где $\alpha_1, \alpha_2 \in V^*$, то существует вывод $S \overset{*}{\underset{G_1}{\Rightarrow}} \alpha_1 A \alpha_2 \overset{*}{\underset{G_1}{\Rightarrow}} x_1 A x_3 \overset{*}{\underset{G_1}{\Rightarrow}} x_1 x_2 x_3$, поскольку терминальные цепочки могут быть выведены из A и из всех нетерминалов, появляющихся в α_1 и α_2. Мы можем эффективно построить множество V_N' всех нетерминалов A, таких, что будет существовать вывод $S \overset{*}{\underset{G_1}{\Rightarrow}} \alpha_1 A \alpha_2$, следующим образом.

Для начала поместим S в искомое множество. Затем последовательно будем добавлять к этому множеству любой нетерминал, который появляется в правой части любого правила из P, определяющего нетерминал, уже имеющийся в этом множестве. Процесс завершается, когда никакие новые элементы не могут быть добавлены к упомянутому множеству.

Положим $G_2 = (V_N', V_T, P', S)$, где P' — множество правил, оставшихся после исключения всех правил из P, которые используют символы из $V_N \setminus V_N'$ слева или справа. G_2 — требуемая грамматика.

Покажем, что $L(G_1) = L(G_2)$ и G_2 удовлетворяет условию теоремы.

I. $L(G_1) \subseteq L(G_2)$. Пусть $x \in L(G_1)$, т.е. $S \underset{G_1}{\overset{*}{\Rightarrow}} x$. Очевидно, что все нетерминалы, встречающиеся в сентенциальных формах этого вывода достижимы, т.е. принадлежат алфавиту V_N', и соответственно в нем участвуют только правила из P'. Следовательно, $S \underset{G_2}{\overset{*}{\Rightarrow}} x$ и $x \in L(G_2)$.

II. $L(G_2) \subseteq L(G_1)$. Это очевидно, так как $P' \subseteq P$.

Из I и II следует, что $L(G_1) = L(G_2)$.

Если $A \in V_N'$, то существует вывод вида $S \underset{G_2}{\overset{*}{\Rightarrow}} \alpha_1 A \alpha_2$, и поскольку все нетерминалы продуктивны, то $S \underset{G_2}{\overset{*}{\Rightarrow}} \alpha_1 A \alpha_2 \underset{G_2}{\overset{*}{\Rightarrow}} x_1 A x_3 \underset{G_2}{\overset{*}{\Rightarrow}} x_1 x_2 x_3$, где $x_1, x_2, x_3 \in V_T^*$.

Что и требовалось доказать.

***Определение* 4.2.** Контекстно-свободные грамматики, удовлетворяющие условию теоремы 4.3, принято называть *приведенными*.

***Определение* 4.3.** Вывод в контекстно-свободной грамматике назовем *левосторонним*, если на каждом его шаге производится замена крайнего левого вхождения нетерминала. Более формально: пусть $G = (V_N, V_T, P, S)$ — контекстно-свободная грамматика. Вывод в грамматике G вида $S \Rightarrow \alpha_1 \Rightarrow \alpha_2 \Rightarrow \ldots \Rightarrow \alpha_n$ — левосторонний, если для $i = 1, 2, \ldots, n-1$ имеет место $\alpha_i = x_i A_i \beta_i$, $x_i \in V_T^*$, $A_i \in V_N$, $\beta_i \in V^*$, а $A_i \to \gamma_i \in P$. Наконец, $\alpha_{i+1} = x_i \gamma_i \beta_i$, т.е. α_{i+1} выведено из α_i заменой A_i на γ_i.

Для обозначения одного шага или нескольких шагов левостороннего вывода будем использовать значок $\underset{\mathrm{lm}}{\Rightarrow}$ или $\underset{\mathrm{lm}}{\overset{*}{\Rightarrow}}$ соответственно.

Лемма 4.1. *Пусть* $G = (V_N, V_T, P, S)$ — *контекстно-свободная грамматика. Если* $S \underset{G}{\overset{*}{\Rightarrow}} x$, *где* $x \in V_T$, *то существует и левосторонний вывод* $S \underset{\mathrm{lm}}{\overset{*}{\Rightarrow}} x$.

Доказательство. Индукцией по длине вывода l докажем более общее утверждение: если для любого нетерминала $A \in V_N$ существует вывод $A \underset{G}{\overset{*}{\Rightarrow}} x$, то существует и левосторонний вывод $A \underset{\mathrm{lm}}{\overset{*}{\Rightarrow}} x$. Утверждение леммы будет следовать как частный случай при $S = A$.

База. Пусть $l = 1$. Для одношагового вывода утверждение выполняется тривиальным образом.

Индукционная гипотеза. Предположим, что утверждение справедливо для любых выводов длиной $l \le n$ ($n \ge 1$).

Индукционный переход. Докажем, что оно справедливо и для $l = n + 1$. Пусть $A \Rightarrow \alpha \overset{*}{\Rightarrow} x$ — вывод длиной $n + 1$ и пусть $\alpha = B_1 B_2 \ldots B_m$, где $B_i \in V^*$, $1 \le i \le m$.

Очевидно, что вывод имеет вид $A \Rightarrow B_1 B_2 \ldots B_m \overset{*}{\Rightarrow} x_1 x_2 \ldots x_m$, причем $B_i \overset{l_i}{\Rightarrow} x_i$, $l_i \le n$, $1 \le i \le m$. Заметим, что некоторые B_i могут быть терминалами, и в этом случае $B_i = x_i$ и вывод не занимает никаких шагов. Если же $B_i \in V_{\mathrm{N}}$, то согласно индукционному предположению $B_i \underset{\mathrm{lm}}{\overset{*}{\Rightarrow}} x_i$. Таким образом, мы можем выстроить левосторонний вывод $A \Rightarrow B_1 B_2 \ldots B_m \underset{\mathrm{lm}}{\overset{*}{\Rightarrow}} x_1 B_2 \ldots B_m \underset{\mathrm{lm}}{\overset{*}{\Rightarrow}} x_1 x_2 \ldots \underset{\mathrm{lm}}{\overset{*}{\Rightarrow}} x_1 x_2 \ldots x_m = x$, воспользовавшись частичными левосторонними выводами для тех B_i, которые являются нетерминалами, применяя их в последовательности слева направо. Что и требовалось доказать.

Теорема 4.4. *Любой контекстно-свободный язык может быть порожден контекстно-свободной грамматикой, не содержащей цепных правил, т.е. правил вида $A \to B$, где A и B — нетерминалы.*

Доказательство. Пусть $G = (V_{\mathrm{N}}, V_{\mathrm{T}}, P, S)$ — cfg и $L = L(G)$. Мы построим новое множество правил P_1, прежде всего включив в него все нецепные правила из P. Затем мы добавим в P_1 правила вида $A \to \alpha$ при условии, что существует вывод вида $A \underset{G}{\overset{*}{\Rightarrow}} B$, где A и B — нетерминалы, а $B \to \alpha$ — нецепное правило из P.

Заметим, что мы легко можем проверить, существует ли вывод $A \underset{G}{\overset{*}{\Rightarrow}} B$, поскольку, если $A \underset{G}{\Rightarrow} B_1 \underset{G}{\Rightarrow} B_2 \underset{G}{\Rightarrow} \ldots \underset{G}{\Rightarrow} B_m \underset{G}{\Rightarrow} B$ и некоторый нетерминал появляется дважды в этом выводе, то мы можем найти более короткую последовательность цепных правил, которая дает результат $A \underset{G}{\overset{*}{\Rightarrow}} B$. Таким образом, достаточно рассматривать только те цепные выводы, длина которых меньше, чем число нетерминалов в V_{N}.

Мы теперь имеем модифицированную грамматику $G_1 = (V_{\mathrm{N}}, V_{\mathrm{T}}, P_1, S)$.

I. Покажем, что $L(G_1) \subseteq L(G)$. Действительно, если $A \to \alpha \in P_1$, то $A \underset{G}{\overset{*}{\Rightarrow}} \alpha$. Следовательно, если терминальная цепочка x выводится в G_1, то она выводима и в G.

II. Покажем теперь, что $L(G) \subseteq L(G_1)$.

Пусть $x \in L(G)$. Рассмотрим левосторонний вывод $S = \alpha_0 \underset{G}{\Rightarrow} \alpha_1 \underset{G}{\Rightarrow} \ldots \underset{G}{\Rightarrow} \alpha_n = x$. Если $\alpha_i \underset{G}{\Rightarrow} \alpha_{i+1}$ для $0 \le i < n$ посредством нецепного правила, то $\alpha_i \underset{G_1}{\Rightarrow} \alpha_{i+1}$. Предположим, что $\alpha_i \underset{G}{\Rightarrow} \alpha_{i+1}$ посредством цепного правила, но что $\alpha_{i-1} \underset{G}{\Rightarrow} \alpha_i$ с помощью нецепного правила при условии, конечно, что $i \ne 0$.

Предположим также, что $\alpha_{i+1} \underset{G}{\Rightarrow} \alpha_{i+2} \underset{G}{\Rightarrow} \ldots \underset{G}{\Rightarrow} \alpha_j$ все посредством цепных правил, а $\alpha_j \underset{G}{\Rightarrow} \alpha_{j+1}$ при помощи нецепного правила. Тогда все α_{i+1}, α_{i+2}, ..., α_j

одинаковой длины, и поскольку вывод — левосторонний, то нетерминал, заменяемый в каждой из них, должен быть в одной и той же позиции. Но тогда $\alpha_i \underset{G_1}{\Rightarrow} \alpha_{j+1}$ посредством одного из правил из $P_1 \setminus P$. Следовательно, $x \in L(G_1)$.

Из утверждений I и II следует $L(G_1) = L(G)$. Что и требовалось доказать.

§ 4.2. Нормальная форма Хомского

Докажем первую из двух теорем о нормальных формах КС-грамматик. Каждая из них утверждает, что все КС-грамматики эквивалентны грамматикам с ограничениями на вид правил.

Теорема 4.5 — нормальная форма Хомского. *Любой КС-язык может быть порожден грамматикой, в которой все правила имеют вид $A \to BC$ или $A \to a$ (A, B, C — нетерминалы, a — терминал).*

Доказательство. Пусть G — КС-грамматика и $L = L(G)$. В соответствии с теоремой 4.4 мы можем найти эквивалентную cfg $G_1 = (V_N, V_T, P, S)$, такую, что множество ее правил P не содержит ни одного цепного правила. Таким образом, если правая часть правила состоит из одного символа, то этот символ — терминал, и это правило уже находится в приемлемой форме.

Теперь рассмотрим правило в P вида $A \to B_1B_2 \ldots B_m$, где $B_i \in V$, $i = 1, 2, \ldots, m$, $m \geq 2$. Если $B_i \in V_T$, заменим его на новый нетерминал C_i, $C_i \notin V_N$, и создадим новое правило для него вида $C_i \to B_i$, которое имеет допустимую форму, поскольку B_i — терминал. Правило $A \to B_1B_2 \ldots B_m$ заменяется правилом $A \to C_1C_2 \ldots C_m$, где $C_i = B_i$, если $B_i \in V_N$.

Пусть пополненное множество нетерминалов — V_N^2, а пополненное множество правил — P_2.

Рассмотрим грамматику $G_2 = (V_N^2, V_T, P_2, S)$. Пока не все ее правила удовлетворяют нормальной форме Хомского (НФХ). Покажем, что она эквивалентна грамматике G_1.

I. Докажем, что $L(G_1) \subseteq L(G_2)$. Пусть $S \overset{*}{\underset{G_1}{\Rightarrow}} x$. Один шаг этого вывода в грамматике G_1, на котором используется правило $A \to B_1B_2 \ldots B_m \in P$, равносилен в грамматике G_2 применению нового правила: $A \to C_1C_2 \ldots C_m \in P_2$ и нескольких правил вида $C_i \to B_i \in P_2$, о которых шла речь. Поэтому имеем $S \overset{*}{\underset{G_2}{\Rightarrow}} x$.

II. Докажем, что $L(G_2) \subseteq L(G_1)$. Индукцией по длине вывода l покажем, что если для любого $A \in V_N$ существует вывод $A \overset{l}{\underset{G_2}{\Rightarrow}} x$, где $x \in V_T^*$, то $A \overset{*}{\underset{G_1}{\Rightarrow}} x$.

База. Пусть $l = 1$. Если $A \underset{G_2}{\Rightarrow} x$, $A \in V_N$, $x \in V_T^*$, то согласно построению грамматики G_2 использованное правило $A \to x \in P_2$ имеется также и во множестве правил P. Действительно, $|x|$ не может быть больше единицы, так как такое пра-

правило не могло бы быть в множестве правил P_2. Следовательно, x — просто терминал, и $A \to x \in P$, а тогда $A \underset{G_1}{\Rightarrow} x$.

Индукционная гипотеза. Предположим, что утверждение выполняется для всех $1 \le l \le n$ $(n \ge 1)$.

Индукционный переход. Пусть $A \overset{l}{\underset{G_2}{\Rightarrow}} x$, где $l = n + 1$. Этот вывод имеет вид: $A \underset{G_2}{\Rightarrow} C_1C_2 \ldots C_m \overset{n}{\underset{G_2}{\Rightarrow}} x$, где $m \ge 2$.

Очевидно, что $x = x_1x_2 \ldots x_m$, причем $C_i \overset{l_i}{\underset{G_2}{\Rightarrow}} x_i$, $l_i \le n$, $i = 1, 2, \ldots, m$. Если $C_i \in V_N^2 \setminus V_N$, то существует только одно правило из множества правил P_2, которое определяет этот нетерминал: $C_i \to a_i$ для некоторого $a_i \in V_T$. В этом случае $a_i = x_i$. По построению правило $A \to C_1C_2 \ldots C_m \in P_2$, используемое на первом шаге вывода, обязано своим происхождением правилу $A \to B_1B_2 \ldots B_m \in P$, где $B_i = C_i$, если $C_i \in V_N$, и $B_i = a_i$, если $C_i \in V_N^2 \setminus V_N$. Для $C_i \in V_N$ мы имеем выводы $C_i \overset{l_i}{\underset{G_2}{\Rightarrow}} x_i$, $l_i \le n$, и по индукционному предположению существуют выводы $B_i \overset{*}{\underset{G_1}{\Rightarrow}} x_i$. Следовательно, $A \overset{*}{\underset{G_1}{\Rightarrow}} x$. При $A = S$ имеем как частный случай $x \in L(G_1)$.

Итак, мы доказали промежуточный результат: любой контекстно-свободный язык может быть порожден контекстно-свободной грамматикой, каждое правило которой имеет форму $A \to a$ или $A \to B_1B_2 \ldots B_m$, где $m \ge 2$; $A, B_1, B_2, \ldots, B_m$ — нетерминалы; a — терминал.

Очевидно, что все правила при $m \le 2$ имеют такой вид, какого требует нормальная форма Хомского. Остается преобразовать правила для $m \ge 3$ к надлежащему виду. Если $G_2 = (V_N^2, V_T, P_2, S)$ — такая cfg, модифицируем ее, добавляя некоторые дополнительные нетерминалы и заменяя некоторые ее правила. Именно: для каждого правила вида $A \to B_1B_2 \ldots B_m \in P_2$, где $m \ge 3$, мы создаем новые нетерминалы $D_1, D_2, \ldots, D_{m-2} \notin V_N^2$ и заменяем правило $A \to B_1B_2 \ldots B_m \in P_2$ множеством правил $\{A \to B_1D_1, D_1 \to B_2D_2, \ldots, D_{m-3} \to B_{m-2}D_{m-2}, D_{m-2} \to B_{m-1}B_m\}$. Пусть V_N^3 — новый нетерминальный словарь, а P_3 — новое множество правил.

Рассмотрим контекстно-свободную грамматику $G_3 = (V_N^3, V_T, P_3, S)$. Докажем, что она эквивалентна грамматике G_2.

III. Докажем, что $L(G_2) \subseteq L(G_3)$. Пусть $S \overset{*}{\underset{G_1}{\Rightarrow}} x$. Один шаг этого вывода в грамматике G_2, на котором используются правила вида $A \to a$ или $A \to B_1B_2$, является и шагом вывода в грамматике G_3, так как по построению эти правила также входят в грамматику G_3.

Шаг вывода в грамматике G_2, на котором используется правило $A \to B_1B_2 \ldots B_m \in P_2$, $m \ge 3$, равносилен последовательному применению правил $A \to B_1D_1$, $D_1 \to B_2D_2, \ldots, D_{m-3} \to B_{m-2}D_{m-2}$, $D_{m-2} \to B_{m-1}B_m \in P_3$. Поэтому имеем $S \overset{*}{\underset{G_3}{\Rightarrow}} x$.

IV. Докажем, что $L(G_3) \subseteq L(G_2)$. Индукцией по длине вывода l покажем, что если для любого $A \in V_N$ существует вывод $A \overset{l}{\underset{G_3}{\Rightarrow}} x$, $x \in V_T^*$, то $A \overset{*}{\underset{G_2}{\Rightarrow}} x$.

База. Пусть $l = 1$. Если $A \underset{G_3}{\Rightarrow} x$, $A \in V_N$, $x \in V_T^*$, то согласно построению G_3 использованное правило $A \to x \in P_3$ содержится также и во множестве правил P_2, так как в этом случае $x \in V_T$, а тогда $A \overset{*}{\underset{G_2}{\Rightarrow}} x$.

Индукционная гипотеза. Предположим, что утверждение выполняется для всех $1 \le l \le n$ $(n \ge 1)$.

Индукционный переход. Пусть $A \overset{l}{\underset{G_3}{\Rightarrow}} x$, где $l = n + 1$. Этот вывод имеет следующий вид: $A \underset{G_3}{\Rightarrow} B_1D_1 \underset{G_3}{\Rightarrow} B_1B_2D_2 \underset{G_3}{\Rightarrow} \dots \underset{G_3}{\Rightarrow} B_1B_2 \dots B_{m-2}D_{m-2} \underset{G_3}{\Rightarrow} B_1B_2 \dots B_m \overset{*}{\underset{G_3}{\Rightarrow}} x$. Очевидно, что $x = x_1x_2 \dots x_m$, где $B_i \overset{l_i}{\underset{G_3}{\Rightarrow}} x_i$, $l_i \le n$, $i = 1, 2, \dots, m$. По индукционной гипотезе $B_i \overset{*}{\underset{G_2}{\Rightarrow}} x_i$. Следовательно, $A \overset{*}{\underset{G_2}{\Rightarrow}} x$. В частности, при $A = S$ получаем $S \overset{*}{\underset{G_2}{\Rightarrow}} x$. Утверждение IV доказано, а вместе с ним доказано равенство $L(G_2) = L(G_3)$, и сама теорема.

Пример 4.1. Рассмотрим грамматику $G = (\{S, A, B\}, \{a, b\}, P, S), \{a, b\}, S)$, в которой $P = \{S \to bA,\ S \to aB,\ A \to a,\ A \to aS,\ A \to bAA,\ B \to b,\ B \to bS,\ B \to aBB\}$.

Построим эквивалентную грамматику в нормальной форме Хомского. Во-первых, два правила, а именно: $A \to a$ и $B \to b$, уже имеют требуемый вид. Нет никаких цепных правил, так что мы можем начать с замены терминалов в правых частях остальных правил на новые нетерминалы и построения правил для них. Правило $S \to bA$ заменяется двумя правилами $S \to C_1A$, $C_1 \to b$. Аналогично правило $S \to aB$ заменяется правилами $S \to C_2B$, $C_2 \to a$. Вместо $A \to aS$ вводятся правила $A \to C_3S$, $C_3 \to a$. Правило $A \to bAA$ заменяется тремя новыми $A \to C_4D_1$, $C_4 \to b$, $D_1 \to AA$. Правило $B \to bS$ заменяется правилами $B \to C_5S$, $C_5 \to b$. Правило $B \to aBB$ заменяется правилами $C_6 \to a$, $B \to C_6D_2$, $D_2 \to BB$.

Итак, мы получили эквивалентную грамматику в НФХ:

$G_1 = (\{S, A, B, C_1, C_2, C_3, C_4, C_5, C_6, D_1, D_2\}, \{a, b\}, P_1, S), \{a, b\}, S)$, где

$$P_1 = \{S \to C_1A, S \to C_2B, A \to C_3S\ A \to C_4D_1, A \to a, B \to C_5S, B \to C_6D_2, B \to b,$$
$$C_1 \to b, C_2 \to a, C_3 \to a, C_4 \to b, C_5 \to b, C_6 \to a, D_1 \to AA, D_2 \to BB\}.$$

§ 4.3. Нормальная форма Грейбах

***Определение* 4.4.** Говорят, что контекстно-свободная грамматика $G = (V_N, V_T, P, S)$ представлена в *нормальной форме Грейбах*, если каждое ее правило имеет вид $A \to a\alpha$, где $a \in V_T$, $a \in V_N^*$.

Для доказательства того, что всякая контекстно-свободная грамматика может быть приведена к нормальной форме Грейбах, нам потребуется обосновать эквивалентность используемых при этом преобразований.

Лемма 4.2. *Пусть $G = (V_N, V_T, P, S)$ — контекстно-свободная грамматика, $A \to \alpha_1 B \alpha_2 \in P$ и $\{B \to \beta_i \mid B \in V_N, \beta_i \in V^+, i = 1, 2, \dots, m\}$ — множество всех B-порождений, т.е. правил с нетерминалом B в левой части. Пусть грамматика $G_1 = (V_N, V_T, P_1, S)$ получается из грамматики G отбрасыванием правила $A \to \alpha_1 B \alpha_2$ и добавлением правил вида $A \to \alpha_1 \beta_i \alpha_2$, $i = 1, 2, \dots, m$. Тогда $L(G) = L(G_1)$.*

Доказательство.

I. Очевидно, что $L(G_1) \subseteq L(G)$. Пусть $S \overset{*}{\underset{G_1}{\Rightarrow}} x$, $x \in V_T^*$. Использование в этом выводе правила $A \to \alpha_1 \beta_i \alpha_2 \in P_1 \setminus P$ равносильно двум шагам вывода в грамматике G: $A \underset{G}{\Rightarrow} \alpha_1 B \alpha_2 \underset{G}{\Rightarrow} \alpha_1 \beta_i \alpha_2$. Шаги вывода в грамматике G_1, на которых используются другие правила из множества правил P, являются шагами вывода в грамматике G. Поэтому $S \overset{*}{\underset{G}{\Rightarrow}} x$.

II. Очевидно, что $L(G) \subseteq L(G_1)$. Пусть $S \overset{*}{\underset{G}{\Rightarrow}} x$. Если в этом выводе используется правило $A \to \alpha_1 B \alpha_2 \in P \setminus P_1$, то рано или поздно для замены B будет использовано правило вида $B \to \beta_i \in P$. Эти два шага вывода в грамматике G равносильны одному шагу вывода в грамматике G_1: $A \underset{G_1}{\Rightarrow} \alpha_1 \beta_i \alpha_2$. Шаги вывода в грамматике G, на которых используются другие правила из множества P, являются шагами вывода в грамматике G_1. Поэтому $S \overset{*}{\underset{G_1}{\Rightarrow}} x$. Что и требовалось доказать.

Лемма 4.3 — об устранении левой рекурсии. *Пусть $G = (V_N, V_T, P, S)$ — контекстно-свободная грамматика, $\{A \to A\alpha_i \mid A \in V_N, \alpha_i \in V^*, i = 1,2,\dots, m\}$ — множество всех леворекурсивных A-порождений, $\{A \to \beta_j \mid j = 1,2,\dots, n\}$ — множество всех прочих A-порождений.*

Пусть $G_1 = (V_N \cup \{Z\}, V_T, P_1, S)$ — контекстно-свободная грамматика, образованная добавлением нетерминала Z к V_N и заменой всех A-порождений правилами:

1) $A \to \beta_j$, $A \to \beta_j Z$, $j = 1, 2, \dots, n$;

2) $Z \to \alpha_i$, $Z \to \alpha_i Z$, $i = 1, 2, \dots, m$.

Тогда $L(G_1) = L(G)$.

Доказательство. Прежде всего заметим, что посредством левосторонних выводов при использовании одних лишь A-порождений порождаются регулярные множества вида $\{\beta_1, \beta_2, \dots, \beta_n\}\{\alpha_1, \alpha_2, \dots, \alpha_m\}^*$, и это является в точности множеством, порождаемым правилами грамматики G_1, имеющими A или Z в левых частях.

I. Докажем, что $L(G) \subseteq L(G_1)$. Пусть $x \in L(G)$. Левосторонний вывод $S \overset{*}{\underset{G}{\Rightarrow}} x$ мы можем перестроить в вывод $S \overset{*}{\underset{G_1}{\Rightarrow}} x$ следующим образом: каждый раз, когда в левостороннем выводе встречается последовательность шагов:

$$tA\gamma \underset{G}{\Rightarrow} tA\alpha_{i_1}\gamma \underset{G}{\Rightarrow} tA\alpha_{i_2}\alpha_{i_1}\gamma \underset{G}{\Rightarrow} \dots \underset{G}{\Rightarrow} tA\alpha_{i_p}\dots\alpha_{i_2}\alpha_{i_1}\gamma \underset{G}{\Rightarrow} t\beta_j\alpha_{i_p}\dots\alpha_{i_2}\alpha_{i_1}\gamma$$

($t \in V_T^*$, $\gamma \in V^*$), заменим их последовательностью

$$tA\gamma \underset{G_1}{\Rightarrow} t\beta_j Z\gamma \underset{G_1}{\Rightarrow} t\beta_j\alpha_{i_p}Z\gamma \underset{G_1}{\Rightarrow} \dots \underset{G_1}{\Rightarrow} t\beta_j\alpha_{i_p}\dots\alpha_{i_2}Z\gamma \underset{G_1}{\Rightarrow} t\beta_j\alpha_{i_p}\dots\alpha_{i_2}\alpha_{i_1}\gamma.$$

Полученный таким образом вывод является выводом цепочки x в грамматике G_1, хотя и не левосторонним. Следовательно, $x \in L(G_1)$.

II. Докажем, что $L(G_1) \subseteq L(G)$. Пусть $x \in L(G_1)$. Рассмотрим левосторонний вывод $S \overset{*}{\underset{G_1}{\Rightarrow}} x$, и перестроим его в вывод в грамматике G следующим образом. Всякий раз, как Z появляется в сентенциальной форме, мы приостанавливаем левосторонний порядок вывода и вместо того, чтобы производить замены в цепочке β, предшествующей Z, займемся замещением Z с помощью правил вида $Z \to \alpha Z$. Далее, вместо того, чтобы производить подстановки в цепочке α, продолжим использовать соответствующие правила для Z, пока, наконец, Z не будет замещено цепочкой, его не содержащей. После этого можно было бы заняться выводами терминальных цепочек из β и α. Результат этого, уже не левостороннего, вывода будет тем же самым, что и при исходном левостороннем выводе в грамматике G_1.

В общем случае вся последовательность шагов этого перестроенного участка вывода, в которых участвует Z, имеет вид

$$tA\gamma \underset{G_1}{\Rightarrow} t\beta_j Z\gamma \underset{G_1}{\Rightarrow} t\beta_j\alpha_{i_p}Z\gamma \underset{G_1}{\Rightarrow} \ldots \underset{G_1}{\Rightarrow} t\beta_j\alpha_{i_p} \ldots \alpha_{i_2}Z\gamma \underset{G_1}{\Rightarrow} t\beta_j\alpha_{i_p} \ldots \alpha_{i_2}\alpha_{i_1}\gamma.$$

Очевидно, что такой же результат может быть получен в грамматике G:

$$tA\gamma \underset{G}{\Rightarrow} tA\alpha_{i_1}\gamma \underset{G}{\Rightarrow} tA\alpha_{i_2}\alpha_{i_1}\gamma \underset{G}{\Rightarrow} \ldots \underset{G}{\Rightarrow} tA\alpha_{i_p} \ldots \alpha_{i_2}\alpha_{i_1}\gamma \underset{G}{\Rightarrow} t\beta_j\alpha_{i_p} \ldots \alpha_{i_2}\alpha_{i_1}\gamma.$$

Таким образом, $L(G_1) = L(G)$. Что и требовалось доказать.

Теорема 4.6 — нормальная форма Грейбах. *Каждый контекстно-свободный язык может быть порожден контекстно-свободной грамматикой в нормальной форме Грейбах.*

Доказательство. Пусть $G = (V_N, V_T, P, S)$ — контекстно-свободная грамматика в нормальной форме Хомского, порождающая контекстно-свободный язык L. Пусть $V_N = \{A_1, A_2, \ldots, A_m\}$.

Первый шаг построения состоит в том, чтобы в правилах вида $A_i \to A_j\gamma$, где γ — цепочка нетерминалов новой грамматики, всегда было $j > i$. Этот шаг выполняется последовательно для $i = 1, 2, \ldots, m$ следующим образом.

При $i = 1$ правило для A_1 может иметь вид $A_1 \to a$, $a \in V_T$, и тогда оно не нуждается в преобразованиях, либо оно имеет вид $A_1 \to A_jA_k$, A_j, $A_k \in V_N$. Если $j > 1$, то правило уже имеет требуемый вид. В противном случае оно леворекурсивно, и в соответствии с леммой 4.3 может быть заменено правилами вида $A_1 \to \beta$, $A_1 \to \beta Z_1$, $Z_1 \to A_k$, $Z_1 \to A_kZ_1$, $\beta = a$, $a \in V_T$, или $\beta = BC$, причем $B \neq A_1$.

Предположим, что для $i = 1, 2, \ldots, k$ правила вида $A_i \to A_j\gamma$ были преобразованы так, что $j > i$.

Покажем, как добиться выполнения этого условия для A_{k+1}-порождений. Если $A_{k+1} \to A_j\gamma$ есть правило, в котором $j < k + 1$, то мы образуем новые правила, подставляя вместо A_j правую часть каждого A_j-порождения согласно лемме 4.2. В результате, если в позиции A_j окажется нетерминал, то его номер будет больше j. Повторив этот процесс самое большее $k-1$ раз, получим порождения вида $A_{k+1} \to A_p\gamma$, $p \geq k + 1$. Порождения с $p = k + 1$ затем преобразуются согласно лемме 4.3 введением новой переменной Z_{k+1}.

Повторив описанный процесс для каждого нетерминала исходной грамматики, мы получим правила только одного из трех следующих видов:

$A_k \to A_p\gamma$, где $p > k$

$A_k \to a\gamma$, где $a \in V_T$

$Z_k \to X\gamma$, где $X \in V_T \cup V_N$, $\gamma \in (V_N \cup \{ Z_1, Z_2, \dots, Z_m\})^*$.

Отметим, что крайний левый символ правой части правила для A_m по необходимости является терминалом, так как нетерминала с большим номером не существует. Крайний левый символ в правой части правила для A_{m-1} может быть терминалом либо нетерминалом A_m. В последнем случае мы можем построить новые правила, заменяя A_m правыми частями A_m-порождений согласно лемме 4.2. Эти новые правила будут иметь правые части, начинающиеся с терминального символа.

Подобным же образом преобразуем правила для $A_{m-2}, A_{m-3}, \dots, A_1$ до тех пор, пока правые части каждого из этих правил не будут начинаться с терминала.

Остается преобразовать правила для новых переменных $Z_1, Z_2, \dots, Z_m$. Правые части этих правил начинаются с терминального символа либо с нетерминала исходной грамматики. Пусть имеется правило вида $Z_i \to A_k\gamma$. Достаточно еще раз применить к нему преобразования, описанные в лемме 4.2, заменив A_k правыми частями A_k-порождений, чтобы получить требуемую форму правил, поскольку правые части правил для A_k уже начинаются с терминалов. На этом построение грамматики в нормальной форме Грейбах, эквивалентной исходной грамматике G, завершается. Что и требовалось доказать.

Пример 4.2. Преобразуем грамматику $G = (\{A_1, A_2, A_3\}, \{a, b\}, P, A_1)$, где $P = \{A_1 \to A_2A_3,\ A_2 \to A_3A_1,\ A_2 \to b,\ A_3 \to A_1A_2,\ A_3 \to a\}$, к нормальной форме Грейбах.

Шаг 1. Поскольку правые части правил для A_1 и A_2 начинаются с нетерминалов с большими номерами и с терминала, то мы начинаем с правила $A_3 \to A_1A_2$ и подставляем цепочку A_2A_3 вместо A_1. Заметим, что $A_1 \to A_2A_3$ является единственным правилом с A_1 в левой части. В результате получаем следующее множество правил

$$\{A_1 \to A_2A_3,\ A_2 \to A_3A_1,\ A_2 \to b,\ A_3 \to A_2A_3A_2,\ A_3 \to a\}.$$

Поскольку правая часть правила $A_3 \to A_2A_3A_2$ начинается с нетерминала с меньшим номером, мы подставляем вместо первого вхождения A_2 либо A_3A_1, либо b. Таким образом, правило $A_3 \to A_2A_3A_2$ заменяется на $A_3 \to A_3A_1A_3A_2$ и $A_3 \to bA_3A_2$. Новое множество есть

$$\{A_1 \to A_2A_3,\ A_2 \to A_3A_1,\ A_2 \to b,\ A_3 \to A_3A_1A_3A_2,\ A_3 \to bA_3A_2,\ A_3 \to a\}.$$

Теперь применим лемму 4.3 к правилам $A_3 \to A_3A_1A_3A_2$, $A_3 \to bA_3A_2$ и $A_3 \to a$. Введем символ Z_3 и заменим правило $A_3 \to A_3A_1A_3A_2$ правилами $A_3 \to bA_3A_2Z_3$, $A_3 \to aZ_3$, $Z_3 \to A_1A_3A_2$ и $Z_3 \to A_1A_3A_2Z_3$. Теперь мы имеем множество:

$$\{A_1 \to A_2A_3,\ A_2 \to A_3A_1,\ A_2 \to b,\ A_3 \to bA_3A_2,\ A_3 \to a,$$
$$A_3 \to bA_3A_2Z_3,\ A_3 \to aZ_3,\ Z_3 \to A_1A_3A_2Z_3,\ Z_3 \to A_1A_3A_2\}.$$

Шаг 2. Все правила с A_3 слева начинаются с терминалов. Они используются для замены A_3 в правиле $A_2 \to A_3A_1$, а затем правила для A_2 используются для того, чтобы заменить A_2 в правиле $A_1 \to A_2A_3$. Получаем:

$$\begin{array}{llll}
A_3 \to bA_3A_2, & A_2 \to bA_3A_2A_1, & A_1 \to bA_3A_2A_1A_3, & Z_3 \to A_1A_3A_2Z_3, \\
A_3 \to a, & A_2 \to bA_3A_2Z_3A_1, & A_1 \to bA_3A_2Z_3A_1A_3, & Z_3 \to A_1A_3A_2. \\
A_3 \to bA_3A_2Z_3, & A_2 \to aA_1, & A_1 \to aA_1A_3, & \\
A_3 \to aZ_3; & A_2 \to aZ_3A_1, & A_1 \to aZ_3A_1A_3, & \\
 & A_2 \to b; & A_1 \to bA_3; &
\end{array}$$

Шаг 3. Два правила для Z_3 заменяются на десять новых в результате подстановки в них вместо A_1 правых частей правил для A_1:

$$\begin{array}{ll}
Z_3 \to bA_3A_2A_1A_3A_3A_2Z_3, & Z_3 \to bA_3A_2A_1A_3A_3A_2, \\
Z_3 \to bA_3A_2Z_3A_1A_3A_3A_2Z_3, & Z_3 \to bA_3A_2Z_3A_1A_3A_3A_2, \\
Z_3 \to aA_1A_3A_3A_2Z_3, & Z_3 \to aA_1A_3A_3A_2, \\
Z_3 \to aZ_3A_1A_3A_3A_2Z_3, & Z_3 \to aZ_3A_1A_3A_3A_2, \\
Z_3 \to bA_3A_3A_2Z_3, & Z_3 \to bA_3A_3A_2.
\end{array}$$

Окончательно, получаем следующее множество правил эквивалентной грамматики в нормальной форме Грейбах:

$$\begin{array}{llll}
A_3 \to bA_3A_2, & A_2 \to bA_3A_2A_1, & A_1 \to bA_3A_2A_1A_3, & Z_3 \to bA_3A_2A_1A_3A_3A_2Z_3, \\
A_3 \to a, & A_2 \to bA_3A_2Z_3A_1, & A_1 \to bA_3A_2Z_3A_1A_3, & Z_3 \to bA_3A_2Z_3A_1A_3A_3A_2Z_3, \\
A_3 \to bA_3A_2Z_3, & A_2 \to aA_1, & A_1 \to aA_1A_3, & Z_3 \to aA_1A_3A_3A_2Z_3, \\
A_3 \to aZ_3; & A_2 \to aZ_3A_1, & A_1 \to aZ_3A_1A_3, & Z_3 \to aZ_3A_1A_3A_3A_2Z_3, \\
 & A_2 \to b; & A_1 \to bA_3; & Z_3 \to bA_3A_3A_2Z_3, \\
 & & & Z_3 \to bA_3A_2A_1A_3A_3A_2, \\
 & & & Z_3 \to bA_3A_2Z_3A_1A_3A_3A_2, \\
 & & & Z_3 \to aA_1A_3A_3A_2, \\
 & & & Z_3 \to aZ_3A_1A_3A_3A_2, \\
 & & & Z_3 \to bA_3A_3A_2.
\end{array}$$

§ 4.4. Разрешимость конечности контекстно-свободных языков

В теореме 4.2 было показано, что из контекстно-свободной грамматики можно исключить те нетерминалы, которые не порождают терминальных цепочек. Фактически можно добиться большего. Мы можем протестировать, является ли язык, порождаемый из данного нетерминала, конечным или бесконечным, и исключить те нетерминалы, не являющиеся начальным нетерминалом грамматики, из которых можно породить только конечное число терминальных цепочек. При доказательстве этого утверждения, мы покажем два результата (теоремы 4.7 и 4.8), очень интересные и сами по себе.

Теорема 4.7 — “*uvwxy*”. *Пусть L — контекстно-свободный язык. Существуют постоянные p и q, зависящие только от языка L, такие, что если существует $z \in L$ при $|z| > p$, то цепочка z представима в виде $z = uvwxy$, где $|vwx| \le q$, причем v, x одновременно не пусты, так что для любого целого $i \ge 0$ цепочка $uv^iwx^iy \in L$.*

Доказательство. Пусть $G = (V_N,\ V_T,\ P,\ S)$ — какая-нибудь контекстно-свободная грамматика в нормальной форме Хомского для языка L. Если $\#V_N = k$, положим $p = 2^{k-1}$ и $q = 2^k$. Докажем теорему для этих значений p и q.

Заметим, что дерево вывода любой терминальной цепочки в грамматике G является бинарным. Поэтому, если в нем нет пути, длиннее j, то выводимая терминальная цепочка не длиннее 2^{j-1}.

Пусть существует $z \in L$, причем $|z| > p = 2^{k-1}$. Тогда самый длинный путь в дереве вывода цепочки z длиннее k, ибо в противном случае $|z| \le 2^{k-1}$, и это противоречило бы предположению, что $|z| > p$.

Рассмотрим самый длинный путь (R) в дереве вывода z (от корня до листа). В нем должны быть два узла: n_1 и n_2, удовлетворяющие следующим условиям:

1) они имеют одинаковые метки, скажем, $A \in V_N$;

2) узел n_1 ближе к корню, чем узел n_2;

3) часть пути R от узла n_1 до листа[5] имеет длину, равную самое большее $k + 1$.

Чтобы убедиться в том, что такие узлы всегда можно найти, пройдем R от листа в сторону корня. Из первых $k + 2$ пройденных узлов только один имеет терминальную метку. Остальные $k + 1$ узлов не могут быть помечены разными нетерминалами.

Рассмотрим поддерево T_1 с корнем n_1. Его результат, являющийся подсловом слова z, обозначим через z_1. В поддереве T_1 не может быть пути, длиннее $k + 1$, так как R является самым длинным путем во всем дереве T. Действительно, пусть $R = Sn_1 + n_1a$. Если допустить, что в T_1 существует другой, более длинный, путь, скажем n_1b, то путь $R' = Sn_1 + n_1b$ окажется длиннее R, так как n_1b длиннее n_1a. Однако это противоречило бы первоначальному предположению, что R является самым длинным путем во всем дереве T. Потому $|z_1| \le 2^k = q$. Эти рассуждения иллюстрирует рис. 4.2.

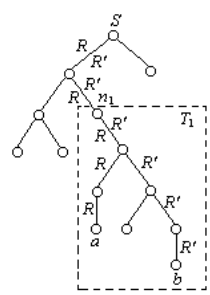

Рис. 4.2.

[5] Очевидно, что самый длинный путь в дереве вывода всегда содержит лист.

Обозначим через T_2 поддерево с корнем n_2, а его результат — через z_2. Ясно, что цепочка z_1 представима в форме $z_1 = z_3z_2z_4$, где z_3 и z_4 одновременно не пусты. Действительно, если первое правило, используемое в выводе z_1, имеет вид $A \to BC$, то поддерево T_2 должно быть полностью в пределах либо поддерева с корнем B, либо поддерева с корнем C.

Рис. 4.3 иллюстрирует три случая: (*а*) когда B есть корень поддерева T_2 ($z_3 = \varepsilon$), (*б*) C есть корень поддерева T_2 ($z_4 = \varepsilon$), (*в*) корень поддерева T_2 расположен внутри поддерева B ($z_3 \neq \varepsilon$, $z_4 \neq \varepsilon$).

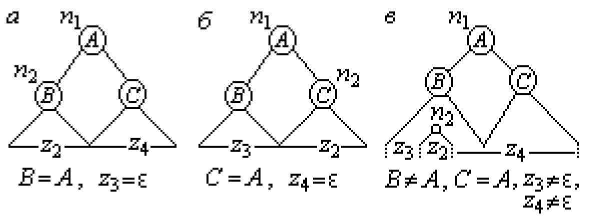

Рис. 4.3.

Теперь мы знаем, что $A \overset{*}{\underset{G}{\Rightarrow}} z_3Az_4$ и, само собой разумеется, что $A \overset{*}{\underset{G}{\Rightarrow}} z_2$. Поэтому $A \overset{*}{\underset{G}{\Rightarrow}} z_3{}^iz_2z_4{}^i$ для любого $i \geq 0$ и цепочка z представима в виде $z = uz_3z_2z_4y$ для некоторых $u, y \in V_{\text{Т}}^*$.

Чтобы закончить доказательство, положим $v = z_3$, $w = z_2$ и $x = z_4$.

Теорема 4.8. *Существует алгоритм для определения, порождает ли данная контекстно-свободная грамматика G конечный или бесконечный язык.*

Доказательство. Пусть p и q — константы, определяемые теоремой 4.7. Если $z \in L(G)$ и $|z| > p$, то $z = uvwxy$ при некоторых $u, v, w, x, y \in V_{\text{Т}}^*$, $|v| + |x| > 0$, и для любого $i \geq 0$ цепочка $uv^iwx^iy \in L(G)$. Следовательно, если в языке $L(G)$ существует цепочка длиной больше p, то язык $L(G)$ бесконечен.

Пусть язык $L = L(G)$ бесконечен. Тогда в нем имеются сколь угодно длинные цепочки и, в частности, цепочка длиной больше $p + q$. Эта цепочка может быть представлена как $uvwxy$, где $|vwx| \leq q$, $|v| + |x| > 0$, и цепочка $uv^iwx^iy \in L$ для любого $i \geq 0$. В частности, при $i = 0$ цепочка $uwy \in L$ и $|uwy| < |uvwxy|$.

Убедимся в том, что $|uwy| > p$. Так как $p + q < |uvwxy|$ и $q \geq |vwx|$, то $p < |uy| \leq |uwy|$. Если $|uwy| > p + q$, то эту процедуру можно повторять снова до тех пор, пока мы не найдем цепочку в языке L, длина которой (l) не будет удовлетворять неравенству $p < l \leq p + q$.

Таким образом, язык L бесконечен тогда и только тогда, когда он содержит цепочку длиной l, $p < l \leq p + q$. Поскольку мы можем проверить, имеется ли данная цепочка в данном контекстно-свободном языке L (см. теорему 2.2 о рекурсивности контекстно-зависимых грамматик), то мы просто должны проверять все цепочки в интервале длин между p и $p + q$ на принадлежность языку $L(G)$. Если такая цепочка имеется, то ясно, что язык L бесконечен; если в языке L нет цепочек длиной больше p, то язык L конечен. Что и требовалось доказать.

В теореме 4.2 доказывалось, что из контекстно-свободной грамматики можно исключить все нетерминалы, из которых не выводится ни одной терминальной цепочки. Теперь мы докажем возможность исключения нетерминалов, из которых выводится только конечное число терминальных цепочек.

Теорема 4.9. *Для всякой контекстно-свободной грамматики* G_1 *можно найти эквивалентную ей контекстно-свободную грамматику* G_2*, такую, что если* A *— нетерминал грамматики* G_2*, не являющийся начальным нетерминалом, то из* A *выводимо бесконечно много терминальных цепочек.*

Доказательство. Если язык $L(G_1) = \{x_1, x_2, ..., x_n\}$ конечен, то утверждение очевидно. Действительно, положим $G_2 = (\{S\}, V_T, P_2, S)$, где $P_2 = \{S \to x_i \mid i = 1, 2, ..., n\}$. В этой грамматике совсем нет нетерминалов, отличных от S.

Пусть теперь грамматика $G_1 = (V_N, V_T, P_1, S)$ и язык $L(G_1)$ бесконечен. Рассмотрим грамматики $G_A = (V_N, V_T, P_1, A)$ для всех $A \in V_N$. Так как существует алгоритм, позволяющий узнать, бесконечен ли порождаемый грамматикой G_A язык, то весь словарь V_N мы можем разбить на две части: $V_N = \{A_1, A_2, ..., A_k\} \cup \{B_1, B_2, ..., B_m\}$, где A_i $(i = 1, 2, ..., k)$ — нетерминалы, порождающие бесконечно много терминальных цепочек, причем начальный нетерминал S среди них, поскольку язык L бесконечен; B_j $(j = 1, 2, ..., m)$ — нетерминалы, порождающие конечные множества терминальных цепочек.

Построим грамматику $G_2 = (\{A_1, A_2, ..., A_k\}, V_T, P_2, S)$, где

$$P_2 = \{A_i \to u_1u_2 ... u_r \mid \exists\, A_i \to C_1C_2 ... C_r \in P_1,$$
$$(1)\ u_i = C_i, \text{ если } C_i \in V_T \cup \{A_1, A_2, ..., A_k\},$$
$$(2)\ C_i \overset{*}{\underset{G_1}{\Rightarrow}} u_i,\ u_i \in V_T^*, \text{ если } C_i \in \{B_1, B_2, ..., B_m\}\}.$$

Короче говоря, правила P_2 получаются из правил P_1 посредством отбрасывания всех правил с нетерминалами B_j в левых частях, а в оставшихся правилах для нетерминалов A_i все вхождения нетерминалов B_j в правых частях надо заменить какими-нибудь их терминальными порождениями. Поскольку число таких терминальных порождений конечно, то и число получающихся правил в P_2 тоже конечно.

Покажем теперь, что $L(G_1) = L(G_2)$.

I. $L(G_1) \subseteq L(G_2)$. Индукцией по длине вывода l докажем, что если $A_i \overset{l}{\underset{G_1}{\Rightarrow}} w$, то $A_i \overset{*}{\underset{G_2}{\Rightarrow}} w$, $w \in V_T^*$ $(i = 1, 2, ..., k)$.

База. Пусть $l = 1$ и пусть $A_i \underset{G_1}{\Rightarrow} w$. При этом применялось правило $A_i \to w \in P_1$, где $w \in V_T^*$. Но это же правило есть в P_2 по построению. Поэтому $A_i \underset{G_2}{\Rightarrow} w$.

Индукционная гипотеза. Предположим, что утверждение выполняется для всех выводов длиной $l \le n$ $(n \ge 1)$.

Индукционный переход. Рассмотрим вывод длиной $l = n + 1$, причем $A_i \underset{G_1}{\Rightarrow} C_1C_2 ... C_r \overset{n}{\underset{G_1}{\Rightarrow}} w_1w_2 ... w_r$, где $C_p \overset{l_p}{\underset{G_1}{\Rightarrow}} w_p$, $w_p \in V_T^*$, $l_p \le n$. На первом шаге применяется правило $A_i \to C_1C_2 \dots C_r \in P_1$. Возьмем во множестве правил P_2 соответствующее правило, которое получается из данного заменой в нем всех нетерминалов типа B на соответствующие w_p, т.е. правило $A_i \to u_1u_2 \dots u_r \in P_2$, в котором $u_p = w_p$, если $C_p \in V_T \cup \{B_1, B_2, \dots, B_k\}$, $u_p = C_p$, если $C_p \in \{A_1, A_2, \dots, A_k\}$, $p = 1, 2, \dots, r$.

Таким образом, имеем $A_i \underset{G_2}{\Rightarrow} u_1u_2 \dots u_r$, причем здесь все $u_p = w_p$, кроме тех u_p, которые равны $C_p \in \{A_1, A_2, \dots, A_k\}$. Но для них $u_p = C_p \overset{l_p}{\underset{G_1}{\Rightarrow}} w_p$, $l_p \le n$, и по индукционному предположению $C_p \overset{*}{\underset{G_2}{\Rightarrow}} w_p$. Поэтому $A_i \underset{G_2}{\Rightarrow} u_1u_2 \dots u_r \overset{*}{\underset{G_2}{\Rightarrow}} w_1w_2 \dots w_r$.

Итак, из $A_i \overset{*}{\underset{G_1}{\Rightarrow}} w$ следует вывод $A_i \overset{*}{\underset{G_2}{\Rightarrow}} w$. Поскольку $S \in \{A_1, A_2, \dots, A_k\}$, то $L(G_1) \subseteq L(G_2)$.

II. $L(G_2) \subseteq L(G_1)$. Пусть $\alpha \underset{G_2}{\Rightarrow} \beta$. Покажем, что $\alpha \overset{*}{\underset{G_1}{\Rightarrow}} \beta$. Шаг вывода $\alpha \underset{G_2}{\Rightarrow} \beta$ предполагает применение правила вида $A_i \to u_1u_2 \dots u_r \in P_2$. Его существование обусловлено существованием правила $A_i \to C_1C_2 \dots C_r \in P_1$, такого что либо $C_i = u_i$, если $u_i \in V_T$, либо C_i — нетерминал типа B и $C_i \overset{*}{\underset{G_1}{\Rightarrow}} u_i$ $(i = 1, 2, \dots, r)$. Следовательно, $A_i \underset{G_1}{\Rightarrow} C_1C_2 \dots C_r \overset{*}{\underset{G_1}{\Rightarrow}} u_1u_2 \dots u_r$.

Таким образом, применение одного правила $A_i \to u_1u_2 \dots u_r \in P_2$ в выводе $\alpha \underset{G_2}{\Rightarrow} \beta$ равносильно применению нескольких правил из множества P_1, позволяющих в цепочке α заменить A_i на $u_1u_2 \dots u_r$, что дает β. Итак, каждый шаг вывода терминальной цепочки в грамматике G_2 может быть заменен несколькими шагами вывода в грамматике G_1, т.е. $L(G_2) \subseteq L(G_1)$.

Из рассуждений I и II следует, что $L(G_1) = L(G_2)$.

Пример 4.3. Рассмотрим грамматику $G_1 = (\{S, A, B\}, \{a, b, c, d\}, P_1, S)$, где $P_1 = \{S \to ASB, S \to AB, A \to a, A \to b, B \to c, B \to d\}$.

Легко видеть, что A порождает только цепочки a и b, а B порождает только цепочки c и d. Однако, S порождает бесконечно много цепочек.

Правило $S \to ASB$ заменяется правилами $S \to aSc$, $S \to aSd$, $S \to bSc$, $S \to bSd$. Аналогично, правило $S \to AB$ заменяется правилами $S \to ac$, $S \to ad$, $S \to bc$, $S \to bd$.

Новая грамматика есть $G_2 = (\{S\}, \{a, b, c, d\}, P_2, S)$, где

$$P_2 = \{S \to aSc, S \to aSd, S \to bSc, S \to bSd, S \to ac, S \to ad, S \to bc, S \to bd\}.$$

§ 4.5. Свойство самовставленности

***Определение* 4.5.** Говорят, что контекстно-свободная грамматика G является *самовставленной*, если существует нетерминал A, такой, что $A \overset{*}{\underset{G}{\Rightarrow}} \alpha_1 A \alpha_2$, где $\alpha_1, \alpha_2 \in V^+$. Говорят также, что нетерминал A является *самовставленным*.

Заметим, что именно свойство самовставленности является причиной появления цепочек вида uv^iwx^iy. Возможно, некоторые понимают, что это свойство самовставленности отличает строго контекстно-свободные языки от регулярных множеств. Но отметим и то, что просто из-за свойства самовставленности грамматики порождаемый ею язык не обязан быть регулярным.

Например, грамматика $G = (\{S\}, \{a, b\}, P, S)$, где $P = \{S \to aSa, S \to aS, S \to bS, S \to a, S \to b\}$, порождает регулярное множество. Действительно, $L(G) = \{a, b\}^+$.

В этом параграфе будет показано, что контекстно-свободная грамматика, которая не является самовставленной, порождает регулярное множество. Следовательно, контекстно-свободный язык не регулярен тогда и только тогда, когда *все* его грамматики — самовставленные.

Теорема 4.10. *Пусть G — несамовставленная контекстно-свободная грамматика. Тогда $L(G)$ — регулярное множество.*

Доказательство. Нетрудно убедиться в том, что если исходная грамматика не является самовставленной, то и эквивалентная ей грамматика в нормальной форме — тоже несамовставленная. В частности, это так для нормальной формы Грейбах. Поэтому, если G — несамовставленная грамматика, то мы можем найти грамматику $G_1 = (V_N^1, V_T, P_1, S_1)$ в нормальной форме Грейбах, эквивалентную грамматике G, которая тоже будет несамовставленной. Хотя это утверждение не очевидно, его легко доказать. Ясно, что применение подстановок, описанных в лемме 4.2 не вводит самовставленности. Что касается преобразований по исключению левой рекурсии, описанных в лемме 4.3, то следует доказать, что нетерминал Z самовставлен, только если нетерминал A — самовставлен. Кроме того, по теореме 4.2 терминальная цепочка может быть выведена из каждого нетерминала.

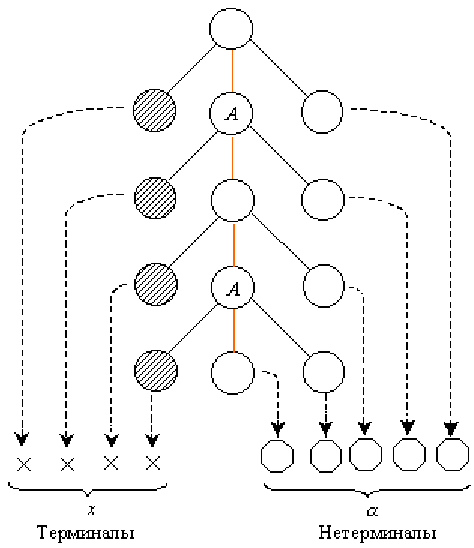

Рис. 4.4.

Рассмотрим левосторонний вывод в грамматике G_1 некоторой сентенциальной формы $x\alpha$. Если G_1 имеет m нетерминалов и l — длина самой длинной правой части правил, то никакая сентенциальная форма не может иметь больше, чем ml нетерминалов. Чтобы убедиться в этом, предположим, что в некоторой сентенциальной форме α левостороннего вывода появляется больше, чем ml нетерминалов. В дереве вывода α рассмотрим путь от корня к крайнему левому листу, помеченному нетерминалом (рис. 4.4). Узлы одного уровня представляют правую часть одного правила грамматики, породившего эти узлы. Все узлы, расположенные справа от этого пути, также как и упомянутый лист, еще не раскрывались с помощью правил. Именно они и образуют цепочку α, состоящую из нетерминалов. На каждом уровне таких узлов не больше $l-2$, кроме самого нижнего. На нижнем же уровне их не больше $l-1$.

Предположим, что в нашем дереве вывода k уровней. Тогда на всех уровнях узлов, порождающих α, не больше, чем $(l-2)(k-1)+l-1$. Всего таких узлов на всех k уровнях по предположению больше ml. Следовательно,
$(l-2)(k-1)+l-1 \geq ml+1$, $k \geq (ml-l+2)/(l-2)+1 = ml/(l-2) > ml/l = m$,
это, естественно, предполагает, что $l > 2$.

Итак, уровней в дереве вывода α (длина пути, о котором шла речь) больше m, т.е. по крайней мере их $m+1$. Следовательно, на этом пути найдутся, по крайней мере, два узла, помеченных одним и тем же нетерминалом A. В этом случае существует левосторонний вывод вида $A \underset{G_1}{\overset{*}{\Rightarrow}} zA\beta$, где $z \in V_T^+$, $\beta \in V_N^{1+}$, т.е. $z \neq \varepsilon$, $\beta \neq \varepsilon$ $(l > 2)$. А это значит, что A — самовставленный нетерминал, что противоречит условию теоремы.

Теперь, если в любой сентенциальной форме самое большее ml нетерминалов, мы можем построить грамматику типа 3: $G_2 = (V_N^2, V_T, P_2, S_2)$, порождающую язык $L(G)$ следующим образом. Нетерминалы грамматики G_2 соответствуют цепочкам нетерминалов грамматики G_1, длина которых не больше ml, т.е. $V_N^2 = \{[\alpha] \mid \alpha \in V_N^{1+}, |\alpha| \leq ml\}$. При этом $S_2 = [S]$. Если $A \rightarrow b\alpha \in P_1$, то для всех нетерминалов из словаря V_N^2, соответствующих строкам, начинающимся на A, в множество правил P_2 мы включаем правила вида $[A\beta] \rightarrow b[\alpha\beta]$ при условии, что $|\alpha\beta| \leq ml$.

Из построения должно быть очевидно, что грамматика G_2 моделирует все левосторонние выводы в грамматике G_1, так что $L(G_2) = L(G_1)$. Действительно, индукцией по длине вывода легко показать, что $S \underset{G_1}{\overset{*}{\Rightarrow}} x\alpha$ посредством левостороннего вывода тогда и только тогда, когда $[S] \underset{G_2}{\overset{*}{\Rightarrow}} x[\alpha]$. Здесь $x \in V_T^+$ — закрытая, а $\alpha \in V_N^{1*}$ — открытая часть данной сентенциальной формы.

I. Докажем сначала, что если $S \underset{G_1}{\overset{l}{\Rightarrow}} x\alpha$, то $[S] \underset{G_2}{\overset{*}{\Rightarrow}} x[\alpha]$.

База. Пусть $l = 1$. Имеем $S \underset{G_1}{\Rightarrow} x\alpha$, $S \rightarrow x\alpha \in P_1$, $x \in V_T$, $\alpha \in V_N^{1*}$. Следовательно, существует правило $[S] \rightarrow x[\alpha] \in P_2$ и потому $[S] \underset{G_2}{\Rightarrow} x[\alpha]$.

Индукционная гипотеза. Предположим, что аналогичное утверждение имеет место при всех $l \le n$ $(n \ge 1)$.

Индукционный переход. Докажем утверждение при $l \le n + 1$. Пусть $S \underset{G_1}{\overset{n}{\Rightarrow}} x'A\beta \underset{G_1}{\Rightarrow} x'b\alpha'\beta = x\alpha$, т.е. $x = x'b$, $\alpha = x'\beta$. По индукционной гипотезе из существования вывода $S \underset{G_1}{\overset{n}{\Rightarrow}} x'A\beta$ следует, что $[S] \underset{G_2}{\overset{*}{\Rightarrow}} x'[A\beta]$, а поскольку на последнем шаге вывода использовано правило $A \to b\alpha' \in P_1$, то существует правило $[A\beta] \to b[\alpha'\beta] \in P_2$, с помощью которого можно завершить имеющийся вывод $[S] \underset{G_2}{\overset{*}{\Rightarrow}} x'[A\beta] \underset{G_2}{\Rightarrow} x'b[\alpha'\beta] = x[\alpha]$.

II. Докажем теперь, что если $[S] \underset{G_2}{\overset{l}{\Rightarrow}} x[\alpha]$, то $S \underset{G_1}{\overset{*}{\Rightarrow}} x\alpha$.

База. Пусть $l = 1$. Имеем $[S] \underset{G_2}{\Rightarrow} x[\alpha]$. Существует $[S] \to x[\alpha] \in P_2$, $x \in V_T$, $\alpha \in V_N^{1^*}$, которое обусловлено существованием правила $S \to x\alpha \in P_1$, и потому $S \underset{G_1}{\Rightarrow} x\alpha$.

Индукционная гипотеза. Предположим, что аналогичное утверждение имеет место при всех $l \le n$ $(n \ge 1)$.

Индукционный переход. Докажем утверждение при $l \le n + 1$.

Пусть $[S] \underset{G_2}{\overset{n}{\Rightarrow}} x'[A\beta] \underset{G_2}{\Rightarrow} x'b[\alpha'\beta] = x[\alpha]$. По индукционной гипотезе из существования вывода $[S] \underset{G_2}{\overset{n}{\Rightarrow}} x'[A\beta]$ следует, что $S \underset{G_1}{\overset{*}{\Rightarrow}} x'A\beta$. На последнем шаге вывода использовано правило $[A\beta] \to b[\alpha'\beta] \in P_2$, существование которого обусловлено существованием правила $A \to b\alpha' \in P_1$, которое можно использовать для завершения имеющегося вывода $S \underset{G_1}{\overset{*}{\Rightarrow}} x'A\beta \underset{G_1}{\Rightarrow} x'b\alpha'\beta = x\alpha$.

Из рассуждений I и II при $\alpha = \varepsilon$ получаем $L(G_1) = L(G_2)$. Таким образом, язык $L(G)$ — регулярен. Что и требовалось доказать.

§ 4.6. ε-Правила в контекстно-свободных грамматиках

Ранее мы показали, что на правила КС-грамматик можно накладывать некоторые *ограничения*, не сужая класс языков, которые могут порождаться. Теперь мы рассмотрим *расширения* КС-грамматик, которые разрешают использовать правила вида $A \to \varepsilon$ для *любого* нетерминала. Такое правило называется ε-*правилом* или ε-*порождением*. Многие описания синтаксиса языков программирования допускают такие порождения. Мы покажем, что язык, порождаемый КС-грамматикой с ε-правилами, — всегда КС-язык.

Понятия, касающиеся деревьев вывода для КС-грамматик, непосредственно переносятся на эти расширенные грамматики. Просто разрешается использовать обозначение ε в качестве метки узла. Ясно, что такой узел должен быть листом.

Теорема 4.11. *Если L — язык, порождаемый грамматикой $G = (V_N, V_T, P, S)$, и каждое правило в P имеет вид $A \to \alpha$, где A — нетерминал, а $\alpha \in V^*$ ($\alpha = \varepsilon$ допустимо), то L может быть порожден грамматикой, в которой каждое правило имеет вид $A \to \alpha$, где A — нетерминал, а $\alpha \in V^+$, либо $S \to \varepsilon$ и, кроме того, начальный нетерминал грамматики S не появляется в правой части никакого правила.*

Доказательство. При помощи тривиального расширения леммы 2.1 мы можем предположить, что S не появляется справа ни в одном правиле в P. Для любого нетерминала $A \in V_N$ мы можем решить, существует ли вывод $A \underset{G}{\overset{*}{\Rightarrow}} \varepsilon$, поскольку если такой вывод существует, то существует и дерево вывода, ветви которого не длиннее, чем число нетерминалов грамматики G (этот аргумент использовался в теореме 4.1).

Пусть $A_1, A_2, ..., A_k$ — те нетерминалы из словаря V_N, из которых цепочка ε может быть выведена, а $B_1, B_2, ..., B_m$ — те нетерминалы, из которых цепочка ε не выводима. Мы построим новое множество правил P_1 следующим образом.

Если $S \underset{G}{\overset{*}{\Rightarrow}} \varepsilon$, то в P_1 включим правило $S \to \varepsilon$. Никакие правила вида $A \to \varepsilon$ в P_1 не включаются.

Если $A \to C_1C_2 \dots C_r \in P$, $r \geq 1$, то в P_1 включаются правила вида $A \to \alpha_1\alpha_2 \dots \alpha_r$, где $\alpha_i = C_i$, если $C_i \in V_T \cup \{B_1, B_2, ..., B_m\}$, либо $\alpha_i = C_i$ или $\alpha_i = \varepsilon$, если $C_i \in \{A_1, A_2, ..., A_k\}$, однако *не все* $\alpha_i = \varepsilon$. Другими словами, преобразования на шаге 3 состоят в том, что в правой части A-правила каждое вхождение A альтернативно либо подменяется на ε, либо остается, как есть. Вхождения других символов не затрагиваются. При этом не допускается, чтобы правая часть обратилась в ε.

Ясно, что новая грамматика $G_1 = (V_N, V_T, P_1, S)$ отличается от грамматики G только набором правил, причем все они имеют требуемый вид.

I. Докажем, что $L(G_1) \subseteq L(G)$. Пусть $\alpha \underset{G_1}{\Rightarrow} \beta$ и при этом применяется правило $A \to \alpha_1\alpha_2 \dots \alpha_r \in P_1$. Его применение эквивалентно применению правила $A \to C_1C_2 \dots C_r \in P$, из которого оно было получено, и нескольких правил из множества правил P для выводов $C_i \underset{G}{\overset{*}{\Rightarrow}} \varepsilon$, если $\alpha_i = \varepsilon$.

II. Докажем теперь, что $L(G) \subseteq L(G_1)$. Индукцией по числу шагов l в выводе докажем, что если $A \underset{G}{\overset{l}{\Rightarrow}} w$ и $w \neq \varepsilon$, то $A \underset{G_1}{\overset{*}{\Rightarrow}} w$ для $A \in V_N$.

База. Пусть $l = 1$. Очевидно, что вывод $A \underset{G}{\Rightarrow} w$ есть также вывод $A \underset{G_1}{\Rightarrow} w$.

Индукционная гипотеза. Предположим, что утверждение выполняется для всех выводов длиной $l \leq n$ $(n \geq 1)$.

Индукционный переход. Пусть $A \underset{G}{\overset{n+1}{\Rightarrow}} w$. Более детально этот вывод имеет следующий вид: $A \underset{G}{\Rightarrow} C_1C_2 \dots C_r \underset{G}{\overset{n}{\Rightarrow}} w_1w_2 \dots w_r$, причем $C_i \underset{G}{\overset{l_i}{\Rightarrow}} w_i$, $l_i \leq n$. Если $w_i \neq \varepsilon$, то по индукционному предположению $C_i \underset{G_1}{\overset{*}{\Rightarrow}} w_i$. Кроме того, по построению из правила $A \to C_1C_2 \ldots C_r \in P$ получается правило $A \to \alpha_1\alpha_2 \ldots \alpha_r \in P_1$, где $\alpha_i = C_i$, если $w_i \neq \varepsilon$, или $\alpha_i = \varepsilon$, если $w_i = \varepsilon$. Следовательно, $A \overset{*}{\underset{G_1}{\Rightarrow}} w$.

Из рассуждений I и II следует, что $L(G_1) = L(G)$. Что и требовалось доказать.

Из теоремы 4.11 непосредственно следует тот факт, что единственная разница между контекстно-свободной грамматикой с правилами вида $A \to \varepsilon$ и грамматиками без ε-правил состоит в том, что первая может порождать пустое предложение. Далее мы будем называть cfg с ε-правилами просто cfg, зная, что эквивалентная грамматика без ε-правил (за исключением быть может $S \to \varepsilon$) может быть найдена.

§ 4.7. Специальные типы контекстно-свободных языков и грамматик

Здесь мы рассмотрим несколько ограниченных классов КС-языков.

***Определение* 4.6.** Говорят, что контекстно-свободная грамматика $G = (V_N, V_T, P, S)$ — *линейна*, если каждое ее правило имеет вид $A \to uBv$ или $A \to u$, где $A, B \in V_N$, $u, v \in V_T^*$. Если $v = \varepsilon$, то грамматика называется *праволинейной*, если $u = \varepsilon$, то она *леволинейна*.

Язык, который может порождаться линейной грамматикой, называется *линейным* языком.

Не все контекстно-свободные языки являются линейными языками. Заметим, что ни одна цепочка, выводимая в линейной грамматике, не имеет более одного нетерминала.

Пример 4.4. Грамматика $G = (\{S\}, \{0, 1\}, P, S)$, где $P = \{S \to 0S1, S \to \varepsilon\}$, является линейной грамматикой, которая порождает язык $L = \{0^n1^n \mid n \geq 0\}$.

***Определение* 4.7.** Говорят, что грамматика $G = (V_N, V_T, P, S)$ — *последовательная*, если нетерминалы $A_1, A_2, \ldots, A_k$ из словаря V_N можно упорядочить так, что если $A_i \to \alpha \in P$, то α не содержит ни одного нетерминала A_j с индексом $j < i$.

Язык, порождаемый последовательной грамматикой, называется *последовательным языком*.

Пример 4.5. Грамматика $G = (\{A_1, A_2\}, \{0, 1\}, P, A_1)$, где $P = \{A_1 \to A_2A_1, A_1 \to A_2, A_2 \to 0A_21, A_2 \to \varepsilon\}$, является последовательной грамматикой, которая порождает язык $L = \{0^n1^n \mid n \geq 0\}^*$.

***Определение* 4.8.** Если контекстно-свободный язык L над алфавитом V_T есть подмножество языка $w_1^* w_2^* \ldots w_k^*$[6] для некоторого k, где $w_i \in V_T^*$, $i = 1, 2, \ldots, k$, то говорят, что L — *ограниченный язык*.

[6] Строго говоря, $w_1^*w_2^* \ldots w_k^*$ следовало бы записывать в виде $\{w_1\}^*\{w_2\}^* \ldots \{w_k\}^*$. Но и без скобок не возникает никаких недоразумений.

Пример 4.6. Язык $\{(ab)^n c^n (dd)^* \mid n \geq 1\}$ является ограниченным языком. Здесь $k = 3$, а $w_1 = ab$, $w_2 = c$, $w_3 = d$.

***Определение* 4.9.** Говорят, что контекстно-свободная грамматика $G = (V_N, V_T, P, S)$ — *неоднозначна*, если в языке $L(G)$ существует цепочка с двумя или более различными левосторонними выводами.

Если все грамматики, порождающие некоторый контекстно-свободный язык, неоднозначны, то говорят, что этот язык существенно неоднозначен.

Существенно неоднозначные контекстно-свободные языки существуют. Классическим примером такого языка является язык $L = \{a^i b^j c^k \mid i = j$ или $j = k\}$. Основная причина, по которой язык L существенно неоднозначен, состоит в том, что любая cfg, порождающая язык L, должна порождать те цепочки, для которых $i = j$, при помощи процесса, который отличается от процесса порождения тех цепочек, для которых $j = k$. Невозможно не порождать некоторые из тех цепочек, для которых $i = j = k$, посредством обоих процессов. Строгое доказательство этого факта весьма сложно (см., например, [1]).

Известно, что проблема распознавания существенной неоднозначности КС-языков алгоритмически неразрешима.

Пример 4.7. Рассмотрим грамматику G из примера 4.1, которая имеет следующие правила:

$$\begin{aligned} P = \{&S \to bA, && S \to aB, && \\ &A \to a, && A \to aS, && A \to bAA, \\ &B \to b, && B \to bS, && B \to aBB\}. \end{aligned}$$

Цепочка *aabbab* имеет следующие два левосторонних вывода:

$S \Rightarrow aB \Rightarrow aaBB \Rightarrow aabB \Rightarrow aabbS \Rightarrow aabbaB \Rightarrow aabbab$,

$S \Rightarrow aB \Rightarrow aaBB \Rightarrow aabSB \Rightarrow aabbAB \Rightarrow aabbaB \Rightarrow aabbab$.

Следовательно, грамматика G — неоднозначная. Однако язык состоит из цепочек, содержащих равное число букв a и b, и может быть порожден однозначной грамматикой $G_1 = (\{S, A, B\}, \{a, b\}, P, S)$, где P состоит из правил

$$S \to aBS,\ S \to aB,\ S \to bAS,\ S \to bA,\ A \to bAA,\ A \to a,\ B \to aBB,\ B \to b.$$