Глава 1. Элементы теории формальных языков и грамматик

1.7. Канонические формы КС-грамматик

Разнообразие форматов продукций КС-грамматик усложняет их теоретические исследования и доказательства ряда важных теорем теории формальных языков и грамматик. Поэтому естественно стремление исследователей ограничить форматы продукций без изменения порождаемых языков и ухудшения выразительных свойств грамматик.

Рассмотрим две канонические формы КС-грамматик: нормальная форма Хомского и нормальная форма Грейбах.

1.7.1. Нормальная форма Хомского

КС-грамматика $G = (V_T, V_N, P, S)$ называется грамматикой в *нормальной форме Хомского*, если множество продукций содержит продукции только следующего вида:

$A \to BC$, где $A, B, C \in V_N$;

$A \to a$, где $A \in V_N$, $a \in V_T$;

$S \to \varepsilon$, если $\varepsilon \in L(G)$, причем S не должен встречаться в правых частях других продукций.

Для ε-свободных КС-грамматик существует альтернативное определение грамматики в нормальной форме Хомского, которое отличается тем, что в нем отсутствует ε-продукция $S \to \varepsilon$ и в продукции вида $A \to BC$ нетерминалы B и C могут быть начальными символами грамматики.

Любая КС-грамматика может быть преобразована в эквивалентную КС-грамматику в нормальной форме Хомского. Преобразование представлено алгоритмом 1.7.

Алгоритм 1.7. Преобразование КС-грамматики к нормальной форме Хомского
Вход: КС-грамматика $G = (V_T, V_N, P, S)$
Выход: КС-грамматика $G' = (V_T, V'_N, P', S')$ в нормальной форме Хомского

Шаг 1. Удалить ε-продукции (см. раздел 1.6.6). Если в грамматике нет ε-продукций, а начальный символ S входит в правые части продукций, то пополнить грамматику новым начальным символом S' и включить в нее продукцию $S' \to S$.

Шаг 2. Удалить цепные продукции (см. раздел 1.6.7).

Шаг 3. Удалить бесполезные символы (см. раздел 1.6.1), поскольку они могут появиться при выполнении шагов 1 и 2.

Шаг 4. Для продукций вида $A \to \alpha$, $|\alpha| > 1$, правые части которых включают в себя подстроки терминалов, каждому терминалу a из правой части поставить в соответствие новый нетерминал A_a и новую продукцию $A_a \to a$. В результате правая часть такой продукции будет состоять только из нетерминалов.

Шаг 5. Для продукций вида $A \to B_1B_2\ldots B_m$, $m > 2$, поставить в соответствие совокупность продукций вида

$A \to B_1B'_1$, $B'_1 \to B_2B'_2$, …, $B'_{m-1} \to B_{m-1}B_m$,

где B'_1, B'_2, …, B'_{m-1} – новые нетерминалы, не содержащиеся более ни в одной продукции. Заметим, что $B_1B_2\ldots B_m$ после выполнения шага 4 могут быть только нетерминалами.

Рассмотрим в качестве примера грамматику со следующими продукциями:

$S \to ASB \mid \varepsilon$

$A \to aA \mid \varepsilon$

$B \to bB \mid b$

Шаг 1. Удаление ε-продукций

$S' \to \varepsilon \mid S$

$S \to ASB \mid SB \mid AB \mid B$

$A \to aA \mid a$

$B \to bB \mid b$

Шаг 2. Удаление цепных продукций

$S' \to \varepsilon \mid ASB \mid SB \mid AB \mid bB \mid b$

$S \to ASB \mid SB \mid AB \mid bB \mid b$

$A \to aA \mid a$

$B \to bB \mid b$

Шаг 3. Бесполезных символов нет.

Шаг 4. Замена терминалов на нетерминалы в продукциях вида $A \to \alpha,\ |\alpha| > 1$

$S' \to \varepsilon \mid ASB \mid SB \mid AB \mid A_bB \mid b$

$S \to ASB \mid SB \mid AB \mid A_bB \mid b$

$A \to A_aA \mid a$

$B \to A_bB \mid b$

$A_a \to a$

$A_b \to b$

Шаг 5. Замена продукций вида $A \to B_1B_2\ldots B_m$, $m > 2$ на совокупность продукций. Имеем две такие продукции: $S' \to ASB$ и $S \to ASB$. Добавим новый нетерминал D для обозначения подстроки SB и продукцию $D \to SB$. В результате получим

$S' \to \varepsilon \mid AD \mid SB \mid AB \mid A_bB \mid b$

$S \to AD \mid SB \mid AB \mid A_bB \mid b$

$A \to A_aA \mid a$

$B \to A_bB \mid b$

$A_a \to a$

$A_b \to b$

$D \to SB$

1.7.2. Нормальная форма Грейбах

КС-грамматика $G = (V_T, V_N, P, S)$ называется грамматикой в *нормальной форме Грейбах* (Sheila Greibach), если множество продукций содержит продукции только следующего вида:

$A \to a\beta$, где $A \in V_N$, $a \in V_T$, $\beta \in V_N^*$;

$S \to \varepsilon$, если $\varepsilon \in L(G)$, причем S не должен встречаться в правых частях других продукций.

В ряде источников иногда на строку β накладывают ограничение $|\beta| \le 2$, т. е. β представляет собой строку из не более, чем двух нетерминалов. Это продукции вида $A \to a$, $A \to aB$, $A \to aBC$, где $A, B, C \in V_N$, $a \in V_T$. Данный вариант мы рассматривать не будем из-за достаточно сложного формального алгоритма преобразования к такой форме.

Существует альтернативное определение нормальной формы Грейбах, в котором $\beta \in (V_T \cup V_N)^*$, т. е. β – произвольная строка терминалов и нетерминалов (включая и строку длины 0). Данное определение ослабляет ограничение на строку β, поэтому ряд авторов называют его как *ослабленная нормальная форма Грейбах*.

Следует обратить внимание на то, что КС-грамматика в нормальной форме Грейбах не имеет левой рекурсии.

Любая КС-грамматика может быть преобразована в эквивалентную КС-грамматику в нормальной форме Грейбах. Преобразование представлено алгоритмом 1.8.

Алгоритм 1.8. Преобразование КС-грамматики к нормальной форме Грейбах
Вход: КС-грамматика $G = (V_T, V_N, P, S)$
Выход: КС-грамматика $G' = (V_T, V'_N, P', S')$ в нормальной форме Грейбах

Шаг 1. Удалить ε-продукции (см. раздел 1.6.6).

Шаг 2. Устранить левую рекурсию (см. разделы 1.6.3 и 1.6.4). В результате все продукции (кроме $S \to \varepsilon$, если она есть) будут иметь вид $A_i \to a\beta$ или $A_i \to A_j\beta$, $i < j$, где $A_i, A_j \in V_N$, $a \in V_T$, $\beta \in (V_T \cup V_N)^*$.

Шаг 3. Определить частичный линейный порядок на множестве V_N в соответствии со следующим правилом: если существует продукция вида $A \to B\alpha$, где $A, B \in V_N$, $\alpha \in (V_T \cup V_N)^*$, то $A < B$. Пусть $V_N = \{A_1, A_2, \dots, A_m\}$. Тогда все A_m-продукции имеют вид $A_m \to \alpha_1, \alpha_2, \dots, \alpha_k$, причем каждая строка $\alpha_1, \alpha_2, \dots, \alpha_k$ начинается с терминала.

Шаг 4. $i := m - 1$.

Шаг 5. Если $i > 0$, для продукций вида $A_i \to A_j\beta$, $j > i$ выполнить замену вхождений нетерминала A_j в соответствии с продукциями $A_j \to \alpha_1, \alpha_2, \dots, \alpha_k$ (каждая строка $\alpha_1, \alpha_2, \dots, \alpha_k$ начинается с терминала), получив продукции $A_i \to \alpha_1\beta \mid \alpha_2\beta \mid \dots \mid \alpha_k\beta$. В противном случае (если $i = 0$) перейти к шагу 7.

Шаг 6. $i := i - 1$ и вернуться к шагу 5.

Шаг 7. Построена эквивалентная КС-грамматика в ослабленной нормальной форме Грейбах. Осталось выполнить требование $\beta \in V_N^*$. В каждой продукции вида $A \to a\beta$, в которой строка β включает в себя подстроки терминалов, каждому терминалу b из строки β поставить в соответствие новый нетерминал A_b и новую продукцию $A_b \to b$.

Рассмотрим данное преобразование на примере. Воспользуемся грамматикой, не содержащей левую рекурсию (шаги 1 и 2 алгоритма выполнены):

$E \to T \mid TE'$

$E' \to +T \mid +TE'$

$T \to F \mid FT'$

$T' \to \times F \mid \times FT'$

$F \to (E) \mid i$

Шаг 3. Упорядочим нетерминалы. В соответствии с продукциями $E \to T \mid TE'$, имеем $E < T$. Из продукций $T \to F \mid FT'$ следует, что $T < F$. Больше продукций вида $A \to B\alpha$ нет. Для нетерминалов E' и T' порядок не определен, поэтому считаем, что они находятся в отношении < с остальными нетерминалами. В результате получаем порядок $E < T < F < E' < T'$. Заметим, что правая часть каждой продукции с нетерминалами T', E' и F в левой части начинается с терминала.

Шаги 4, 5, 6. Для продукций $T \to F \mid FT'$ выполним замену вхождений нетерминала F:

$T \to (E) \mid i \mid (E)T' \mid iT'$

Для продукций $E \to T \mid TE'$ выполним замену вхождений нетерминала T:

$E \to (E) \mid i \mid (E)T' \mid iT' \mid (E)E' \mid iE' \mid (E)T'E' \mid iT'E'$

В результате получим эквивалентную грамматику в ослабленной нормальной форме Грейбах:

$E \to (E) \mid i \mid (E)T' \mid iT' \mid (E)E' \mid iE' \mid (E)T'E' \mid iT'E'$

$E' \to +T \mid +TE'$

$T \to (E) \mid i \mid (E)T' \mid iT'$

$T' \to \times F \mid \times FT'$

$F \to (E) \mid i$

Шаг 7. В продукциях

$$E \to (E) \mid (E)T' \mid (E)E' \mid (E)T'E'$$

$$T \to (E) \mid (E)T'$$

$$F \to (E)$$

терминалу) сопоставим нетерминал B и добавим соответствующую новую продукцию:

$$E \to (EB \mid i \mid (EBT' \mid iT' \mid (EBE' \mid iE' \mid (EBT'E' \mid iT'E'$$

$$E' \to +T \mid +TE'$$

$$T \to (EB \mid i \mid (EBT' \mid iT'$$

$$T' \to \times F \mid \times FT'$$

$$F \to (EB \mid i$$

$$B \to)$$