УДК 519.682.1

**С. Ю. Соловьев**[1]

# ОСОБЕННОСТИ КАНОНИЧЕСКИХ РАЗДЕЛЕННЫХ ГРАММАТИК

Рассматривается класс канонических разделенных грамматик, способных порождать те же языки, что и разделенные грамматики общего вида. Приводятся основные свойства и два характеристических признака канонических грамматик. Предлагается метод унификации нетерминальных символов и доказывается единственность канонического представления разделенной грамматики.

*Ключевые слова*: разделенные грамматики, эквивалентность.

**1. Введение.** В 1974 г. Бруно Курсель [1] доказал существование канонического представления разделенных грамматик в нормальной форме Грейбах. В настоящей работе развивается иной подход к построению канонических грамматик, согласно которому условие нормализации

[1] Факультет ВМК МГУ, проф., д.ф.-м.н., e-mail: soloviev@glossary.ru

заменяется требованием обязательной альтернативности. Каждая каноническая грамматика такого рода содержит наименьшее количество нетерминальных символов по сравнению с прочими эквивалентными ей разделенными грамматиками. В рамках предложенного подхода доказываются два признака и единственность канонического представления; попутно предлагаются новые доказательства некоторых известных [1] утверждений.

**2. Разделенные грамматики.** Строительным материалом любой формальной грамматики являются символы двух непересекающихся алфавитов, первый из которых называется алфавитом терминальных символов (*терминалов*), а второй — алфавитом нетерминальных символов (*нетерминалов*). Из числа нетерминалов выделяется особый *начальный символ*. Везде далее используются обозначения: $\Sigma$ — алфавит терминалов; $N$ — алфавит нетерминалов; $S$ — начальный символ; $D$ — нетерминал из $N$; $N'$ — алфавит $N\setminus\{S\}$; $a$, $b$, $c$ — терминалы из $\Sigma$; $A$, $B$, $C$ — нетерминалы из $N'$; $x$, $y$, $z$ — цепочки терминалов (*предложений*) из $\Sigma^*$; $\alpha$, $\beta$ — цепочки из $(N\cup\Sigma)^*$; $\varphi$, $\psi$ — цепочки из $(N'\cup\Sigma)^+$; $\mu$, $\rho$, $\eta$ — цепочки из $(N'\cup\Sigma)^*$.

*Разделенной грамматикой* [2] называется четверка $G=\langle N,\Sigma,P,S\rangle$, где $P$ — множество *правил вывода* $D\to b\beta$, в которых каждой паре $(D,b)$ отвечает не более одного правила. Правило вывода $D\to b\beta$ называется $D$-*правилом*, а цепочка $b\beta$ — *альтернативой* нетерминала $D$. В примерах принято задавать грамматики правилами вывода, из которых извлекаются остальные члены четверки. В литературе разделенные грамматики также именуются строгими разделенными грамматиками и простыми $LL(1)$-грамматиками. Разделенные грамматики в нормальной форме Грейбах называются $s$-грамматиками [1].

Пусть $G$ — разделенная грамматика и $L$, $L_1$ — некоторые языки, $L, L_1\subseteq\Sigma^*$. Далее используются обозначения: $P(D)$ — множество всех $D$-правил из $P$; $\widehat{P}(N_X)$ — множество всех $D$-правил, где $D\in N_X$ и $N_X\subseteq N$; $R(G,D)$ — множество цепочек-альтернатив $\{\alpha\colon\ D\to\alpha\in P(D)\}$; $x\|L$ — множество цепочек $\{z\colon\ xz\in L\}$; $|\alpha|$ — длина цепочки $\alpha$; $e$ — цепочка нулевой длины; $\times$ — операция произведения языков: $L\times L_1\stackrel{\text{def}}{=}\{xy\colon\ x\in L,\ y\in L_1\}$; $\Rightarrow$ — бинарное отношение непосредственного левого вывода: $\alpha_1\Rightarrow\alpha_2$, если $\alpha_1$ и $\alpha_2$ можно представить в виде $xD\alpha_0$ и $xb\beta\alpha_0$, где $D\to b\beta$ — правило из $P$. В общем случае левый вывод $x_0D_0\beta_0\Rightarrow^+x_n$ имеет вид

$$x_0D_0\beta_0\Rightarrow x_1D_1\beta_1\Rightarrow\ldots\Rightarrow x_{n-1}D_{n-1}\beta_{n-1}\Rightarrow x_n. \tag{1}$$

*Производным языком* для грамматики $G$ и начальной цепочки $\alpha$ будем называть множество предложений $L(G,\alpha)=\{x\colon\ \alpha\Rightarrow^*x\}$. Язык, порождаемый разделенной грамматикой $G$, есть множество $L(G,S)$. В соответствии со сложившейся традицией язык $L(G,S)$ будем называть разделенным языком. Грамматики $G$ и $G_1$ называются *эквивалентными*, если $L(G,S)=L(G_1,S)$.

Примером разделенной грамматики, демонстрирующим многие последующие рассуждения, является грамматика

$$\begin{array}{lllll}
\widehat{g}\colon & S\to & b\,D & \mid c\,A, & L(\widehat{g},S)=\{c,b\,a,b\,b\}\times L(\widehat{g},A),\\
& A\to & a & \mid b\,A \quad \mid c\,C\,B, & L(\widehat{g},A)=\{b,c\,a,c\,b\}^*\times\{a\},\\
& B\to & a & \mid b\,B \quad \mid c\,D, & L(\widehat{g},B)=L(\widehat{g},A),\\
& C\to & a & \mid b, & L(\widehat{g},C)=\{a,b\},\\
& D\to & a\,A & \mid b\,B, & L(\widehat{g},D)=\{a,b\}\times L(\widehat{g},A)=L(\widehat{g},C\,A).
\end{array}$$

Зафиксируем очевидные свойства разделенных грамматик и языков.

1. В разделенной грамматике $G$ все языки $L(G,\varphi)$ являются разделенными.
2. Любой разделенный язык является *префиксным языком*, в котором никакие два предложения $x$ и $y$ не связаны отношением $x=yz$ для некоторого $z\neq e$.
3. Если $L=L_1\times L_2$ и $L_1$ — префиксный язык, то $x\|L=L_2$ для любого $x\in L_1$.
4. В выводе (1) $D_i$-правило, применяемое к $x_iD_i\beta_i$, однозначно определяется терминалом $a$, расположенным в позиции $|x_i|+1$ каждой из цепочек $x_{i+1},\ldots,x_n$.

Лемма 1. *Пусть* $L_1\times\{b\}=L_2\times L_3$, *где* $b$ — *терминал из* $\Sigma$ *и* $L_1,L_2,L_3\subseteq\Sigma^+$. *Пусть, кроме того,* $L_1$ — *префиксный язык, а* $L_3$ — *язык, содержащий более одного предложения. Тогда существует язык* $L_3'\subseteq\Sigma^+$, *такой, что* $L_3=L_3'\times\{b\}$.

Доказательство. Выберем предложение $x \in L_2$. Так как $e \notin L_3$, то все предложения из $L_3$ имеют вид $zb$, и для завершения доказательства достаточно показать, что $b \notin L_3$. Предположим обратное: $b \in L_3$. В этом случае $L_3 \supseteq \{b, yb\}$ для некоторого $y \neq e$; и, следовательно, $L_1 \times \{b\} \supseteq \{xb, xyb\}$ и $L_1 \supseteq \{x, xy\}$, т. е. язык $L_1$ не является префиксным.

Лемма 2. *Пусть в разделенной грамматике $G$ для тройки цепочек $x$, $y$, $\varphi$ выполняются условия $x \neq e$, $y \neq e$, $xy \in L(G, \varphi)$. Тогда существует цепочка $\psi$, такая, что $L(G, \psi) = x \| L(G, \varphi)$.*

Доказательство. Рассмотрим вывод (1) предложения $xy$ ($\varphi \equiv x_0 D_0 \beta_0$ и $xy \equiv x_n$). Обозначим через $k$ наименьший номер $i$, $0 \leqslant i \leqslant n$, для которого цепочку $x_i$ можно представить в виде $xx_i'$. Если $k = n$, то $\psi = y$; в противном случае $\psi = x_k' D_k \beta_k$. Поскольку левый вывод любого предложения $xz$ из $L(G, \varphi)$ (как следует из свойства 4) имеет вид $\varphi \Rightarrow^* x\psi \Rightarrow^* xz$, то $L(G, \psi) = x \| L(G, \varphi)$.

Следствие. Если $L$ — разделенный язык и $xy \in L$, где $x \neq e$, $y \neq e$, то множество $x \| L$ — разделенный язык.

Лемма 3. *Пусть $G$ — разделенная грамматика и $L$ — разделенный язык, в котором предложения $z_1$ и $z_2$ начинаются разными терминалами. Пусть для тройки $\varphi$, $y$ и $L$ имеют место выводы $\varphi \Rightarrow^+ yz$ для всех $z \in L$. Тогда существует цепочка $D\beta$, такая, что $\varphi \Rightarrow^+ yD\beta \Rightarrow^+ yz$ для всех $z \in L$, причем $L \subseteq L(G, D\beta)$.*

Доказательство. Если $y = e$, то утверждение леммы справедливо при $D\beta = \varphi$. Пусть $y \neq e$ и $z$ — некоторая цепочка из $L$. Рассмотрим вывод (1) предложения $yz$ ($\varphi \equiv x_0 D_0 \beta_0$ и $yz \equiv x_n$). В этом выводе на некотором этапе $x_k D_k \beta_k \Rightarrow x_{k+1} D_{k+1} \beta_{k+1}$ выполняются неравенства $|x_k| < |y|$ и $|x_{k+1}| \geqslant |y|$. Как следует из свойства 4, применяемое на этом этапе $D_k$-правило зависит от символов цепочки $y$ и не зависит от $z$. В частности, $x_{k+1} D_{k+1} \beta_{k+1} \Rightarrow^+ yz_1$ и $x_{k+1} D_{k+1} \beta_{k+1} \Rightarrow^+ yz_2$, а поскольку первые символы предложений $z_1$ и $z_2$ различны, то $|x_{k+1}| \leqslant |y|$ и, следовательно, $|x_{k+1}| = |y|$, $y = x_{k+1}$, т. е. утверждение леммы справедливо при $D\beta = D_{k+1}\beta_{k+1}$. Лемма доказана.

Каждую разделенную грамматику эквивалентными преобразованиями можно привести к каноническому виду, позволяющему сравнивать разделенные языки простым сравнением множеств правил вывода. Канонические грамматики должны одновременно удовлетворять шести условиям, которые удобно разбить на две группы.

**3. Квазиканонические разделенные грамматики.** Общие свойства различных подходов к определению канонических грамматик можно формализовать и исследовать в виде отдельного класса грамматик.

*Квазиканонической разделенной грамматикой* (quasicanonical separated grammar; *QCS*-грамматикой) называется разделенная грамматика $G = \langle N, \Sigma, P, S \rangle$, удовлетворяющая следующим условиям.

A1. Основной символ $S$ не встречается в альтернативах правил вывода.

A2. $\forall D \to \psi \in P$ существует вывод вида $S \Rightarrow^* xD\beta \Rightarrow^* x\psi\beta \Rightarrow^* y$.

A3. $\forall A \in N'$ множество $P(A)$ содержит, по крайней мере, два правила.

A4. $\forall A \in N'$ язык $L(G, A)$ нельзя представить как $L_A \times \{a\}$, где $L_A \subseteq \Sigma^+$ и $a \in \Sigma$.

Лемма 4. 1°. *Пусть $G$ — QCS-грамматика, в которой для тройки $\varphi$, $L_0$ и $B$ (где $L_0$ — разделенный язык) выполняется равенство $L(G, \varphi) = L_0 \times L(G, B)$. Тогда $\varphi = \eta A$ для некоторых $\eta$ и $A$.*

2°. *Пусть дополнительно $\varphi = A$. Тогда каждое $A$-правило грамматики $G$ имеет вид $A \to a\rho_a C_a$, причем $L(G, a\rho_a C_a) = L_a \times L(G, B)$, где $L_a = \{x \in L_0 \colon\ x = ax'\}$.*

Доказательство. 1°. Как следует из условий A2 и A3, язык $L(G, B)$ содержит более одного предложения, поэтому $\varphi$ имеет хотя бы один нетерминал и, в частности, либо $\varphi = \varphi' a$, либо $\varphi = \eta A$ для некоторых $\varphi'$, $a$, $\eta$ и $A$. Допустим $\varphi = \varphi' a$. Тогда $L(G, \varphi) = L_0 \times L(G, B)$, что эквивалентно $L(G, \varphi') \times \{a\} = L_0 \times L(G, B)$, и, как следует из леммы 1, $L(G, B) = L_B \times \{a\}$ для некоторого языка $L_B$, а это противоречит условию A4. Таким образом, для цепочки $\varphi$ возможен единственный вариант: $\varphi = \eta A$.

2°. Если $\varphi = A$, то $L(G, A) = L_0 \times L(G, B)$. Выберем произвольное $A$-правило $A \to a\psi_a$, для которого предыдущее равенство принимает вид $L(G, a\psi_a) = L_a \times L(G, B)$. Из первой части доказательства для тройки $a\psi_a$, $L_a$, $B$ следует существование пары $\eta_a$, $C_a$, такой, что $a\psi_a = \eta_a C_a$, а значит $\eta_a$ имеет вид $a\rho_a$, и выбранное $A$-правило есть $A \to a\rho_a C_a$.

Лемма 5. *Пусть $L(G, \varphi) = L(G, \psi)$, где $G$ — QCS-грамматика, а $\varphi$ и $\psi$ — различные цепочки. Тогда $\varphi$ и $\psi$ можно представить как $\rho_\varphi A_\varphi \mu$ и $\rho_\psi A_\psi \mu$, где $A_\varphi$ и $A_\psi$ — различные*

*нетерминалы, связанные одним из трех равенств*: $L(G, A_\varphi) = L(G, A_\psi)$, $L(G, \varphi' A_\varphi) = L(G, A_\psi)$, $L(G, A_\varphi) = L(G, \psi' A_\psi)$ *для некоторых непустых* $\varphi'$ *и* $\psi'$.

Доказательство. Обозначим через $\mu$ наибольший общий постфикс цепочек $\varphi$ и $\psi$. Тогда $\varphi = \rho_\varphi X \mu$, $\psi = \rho_\psi Y \mu$, $X \neq Y$ и $L(G, \rho_\varphi X) = L(G, \rho_\psi Y) = L$. Для символов $X$ и $Y$ возможен один из четырех вариантов:

1) $X \in \Sigma$ и $Y \in \Sigma$, т. е. $L(G, \rho_\varphi a) = L(G, \rho_\psi b)$ для некоторых $a \neq b$;
2) $X \in \Sigma$ и $Y \in N$, т. е. $L(G, \rho_\varphi a) = L(G, \rho_\psi A_\psi)$ для некоторых $a$ и $A_\psi$;
3) $X \in N$ и $Y \in \Sigma$, т. е. $L(G, \rho_\varphi A_\varphi) = L(G, \rho_\psi b)$ для некоторых $A_\varphi$ и $b$;
4) $X \in N$ и $Y \in N$, т. е. $L(G, \rho_\varphi A_\varphi) = L(G, \rho_\psi A_\psi)$ для некоторых $A_\varphi \neq A_\psi$.

Вариант 1 возможен лишь при $L = \emptyset$, что противоречит условию A2. Вариант 2 при $\rho_\psi = e$ противоречит условию A4, а при $\rho_\psi \neq e$ противоречит лемме 4 (часть 1°) для тройки $\rho_\varphi a$, $L(G, \rho_\psi)$ и $A_\psi$. Вариант 3 также противоречит условию A4 и лемме 4 (часть 1°). Рассмотрим единственный возможный вариант 4. При $\rho_\varphi = e$ и/или $\rho_\psi = e$ доказательство леммы очевидно. Пусть $\rho_\varphi \neq e$ и $\rho_\psi \neq e$.

Зафиксируем предложение $x \in L$ и рассмотрим его выводы:

$$\rho_\varphi A_\varphi \Rightarrow^+ y_1 A_\varphi \Rightarrow^+ y_1 z_1 = x, \quad \text{причем} \quad \rho_\varphi \Rightarrow^+ y_1,$$
$$\rho_\psi A_\psi \Rightarrow^+ y_2 A_\psi \Rightarrow^+ y_2 z_2 = x, \quad \text{причем} \quad \rho_\psi \Rightarrow^+ y_2.$$

Как следует из свойства 3 разделенных языков:

$$y_1 \| L = y_1 \| L(G, \rho_\varphi A_\varphi) = L(G, A_\varphi) \quad \text{и} \quad y_2 \| L = y_2 \| L(G, \rho_\psi A_\psi) = L(G, A_\psi).$$

Цепочки $y_1$ и $y_2$ являются префиксами предложения $x$, поэтому между ними возможно одно из трех соотношений: $y_1 = y_2$, $y_1 = y_2 y_0$ или $y_1 y_0 = y_2$, где $y_0 \neq e$.

Если $y_1 = y_2$, то $y_1 \| L = y_2 \| L$, т. е. $L(G, A_\varphi) = L(G, A_\psi)$.

Если $y_1 = y_2 y_0$, то из леммы 2 (для тройки $y_2$, $y_0$, $\rho_\varphi$) вытекает существование цепочки $\varphi'$, такой, что

$$y_2 \| L(G, \rho_\varphi A_\varphi) = \Big( y_2 \| L(G, \rho_\varphi) \Big) \times L(G, A_\varphi) = L(G, \varphi') \times L(G, A_\varphi) = L(G, \varphi' A_\varphi).$$

Из сравнения двух выражений для $y_2 \| L$ следует равенство $L(G, A_\psi) = L(G, \varphi' A_\varphi)$.

Если $y_1 y_0 = y_2$, то (аналогично соотношению $y_1 = y_2 y_0$) $L(G, A_\varphi) = L(G, \psi' A_\psi)$.

**4. Канонические разделенные грамматики.** Лемма 5 позволяет вполне естественно перейти к искомому каноническому представлению разделенных грамматик.

*Канонической разделенной грамматикой* ($CS$-грамматикой) называется $QCS$-грамматика $G$, удовлетворяющая следующим условиям.

A5. $L(G, A) \neq L(G, B)$ для всех $A \neq B$ из $N'$.

A6. $L(G, A) \neq L_0 \times L(G, B)$ для всех $A, B \in N'$ и для всех разделенных языков $L_0$.

Приведенному определению не удовлетворяет $QCS$-грамматика $\widehat{g}$, но удовлетворяет эквивалентная ей грамматика

$$\begin{array}{lllll}
\widehat{\widehat{g}}: & S \to & b\,C\,A & \mid c\,A, & L(\widehat{\widehat{g}}, S) = \{c, b\,a, b\,b\} \times L(\widehat{\widehat{g}}, A), \\
& A \to & a & \mid b\,A \quad \mid c\,C\,A, & L(\widehat{\widehat{g}}, A) = \{b, c\,a, c\,b\}^* \times \{a\}, \\
& C \to & a & \mid b, & L(\widehat{\widehat{g}}, C) = \{a, b\}.
\end{array}$$

Алгоритмическая разрешимость задачи приведения разделенной грамматики к каноническому виду доказана, например, в [3, 4].

Теорема 1. *Если* $L(G, \varphi) = L(G, \psi)$ *для* $CS$*-грамматики* $G$, *то* $\varphi = \psi$.

Доказательство. Допустим в $CS$-грамматике $G$ для некоторых $\varphi \neq \psi$ выполняется равенство $L(G, \varphi) = L(G, \psi)$. Тогда из леммы 5 следует, что в $G$ нарушается либо условие A5, либо условие A6.

Теорема 2. *Если* $G = \langle N, \Sigma, P, S \rangle$ — $CS$*-грамматика и* $A \in N'$, *то* $L(G, A) \neq L_1 \times L_2$, *для любых разделенных языков* $L_1, L_2 \subseteq \Sigma^+$.

Доказательство. Допустим обратное: в некоторой $CS$-грамматике $G$ существует нетерминал $A \in N'$, для которого выполняется равенство $L(G, A) = L_1 \times L_2$. Для языка $L_2$ возможен один из двух вариантов: либо $L_2$ состоит из единственного предложения $xa$, либо $L_2$ содержит не менее двух предложений. Первый случай можно исключить из рассмотрения, так как

$L(G,A) = \{yx\colon\ y \in L_1\} \times \{a\}$, что противоречит условию A4. Во втором случае, не ограничивая общности, можно полагать, что язык $L_2$ содержит предложения $z_1$ и $z_2$, которые начинаются различными символами. Выберем в $L_1$ некоторое предложение $y$. Из леммы 3 (для тройки $A$, $y$, $L_2$) вытекает существование цепочки $B\beta$, где $\beta \in (N' \cup \Sigma)^*$, такой, что $A \Rightarrow^+ yB\beta \Rightarrow^+ yz$ для всех $z \in L_2$, причем $L_2 \subseteq L(G, B\beta)$.

Докажем равенство $L_2 = L(G, B\beta)$. Допустим, что $z \in L(G, B\beta)$. Из $A \Rightarrow^+ yB\beta$ следует утверждение $yz \in L(G,A)$. Из $L(G,A) = L_1 \times L_2$ вытекает, что $yz = y'z'$, где $y' \in L_1$, $z' \in L_2$. Если $y \neq y'$, то $\{y, y'\} \subseteq L_1$, что противоречит свойству 2 префиксных языков. Поэтому $y = y'$ и, следовательно, $z = z'$, $z \in L_2$. Итак, $L(G,A) = L_1 \times L(G, B\beta)$. В зависимости от самого правого символа цепочки $B\beta$ последнее равенство противоречит либо условию A4, либо условию A6.

**5. Первый характеристический признак $CS$-грамматик.** Этот признак позволяет для идентификации канонических разделенных грамматик использовать альтернативную формулировку условия A6, опирающуюся на одно структурное свойство нетерминалов.

Пусть $G = \langle N, \Sigma, P, S\rangle$ — разделенная грамматика и $X$ — фиксированный символ из $N' \cup \Sigma$. Обозначим через $RD(G,X)$ совокупность непустых подмножеств $N_X \subseteq N'$, таких, что $X \notin N_X$, а все правила из $\widehat{P}(N_X)$ имеют вид $A \to \varphi Y$, где $Y = X$ либо $Y \in N_X$. Заметим, что A4 эквивалентно условию "$R(D,a) = \emptyset$ для всех $a \in \Sigma$". Будем рассматривать грамматики, удовлетворяющие условию

A6′. $RD(G,A) = \emptyset$ для всех $A \in N'$.

Т е о р е м а 3. *Если $QCS$-грамматика $G$ удовлетворяет условиям* A5 *и* A6′, *то $G$ является $CS$-грамматикой.*

Д о к а з а т е л ь с т в о. Предположим, что в $QCS$-грамматике $G$, удовлетворяющей условиям теоремы, не выполняется условие A6, т. е. существуют несколько пар $A_i$, $B_i$, для которых $L(G,A_i) = L_{A_iB_i} \times L(G,B_i)$, где $L_{A_iB_i}$ — разделенные языки. Покажем, что такое предположение противоречит условию A6′.

Обозначим через $h(A)$ длину самого короткого предложения языка $L(G,A)$. Из пар $A_i$, $B_i$ выберем пару $A_m$, $B_m$ с наименьшим значением $h(B_i)$. Для этой пары определим непустое множество нетерминалов $N_{B_m}$:

$$N_{B_m} \stackrel{\text{def}}{=} \{A \in N'\colon\ L(G,A) = L_{AB_m} \times L(G,B_m)\},$$

и покажем, что $N_{B_m} \in RD(G,B_m)$.

Для некоторого $A \in N_{B_m}$ выберем произвольное $A$-правило. Как следует из леммы 4 (часть 2) (для тройки $A$, $L_{AB_m}$ и $B_m$), выбранное правило имеет вид $A \to a\rho_a C_a$, причем $L(G, a\rho_a C_a) = L_a \times L(G,B_m)$, где $L_a$ — подмножество предложений из $L_{AB_m}$, начинающихся терминалом $a$. В языке $L_0 = L(G,a\rho_a) \times L(G,C_a) = L_a \times L(G,B_m)$ выберем некоторое предложение $x \in L_0$, которое можно представить как $x = y_1z_1 = y_2z_2$, где $y_1 \in L(G,a\rho_a)$, $y_2 \in L_a$. Из свойства 3 разделенных языков следует, что $y_1 \| L_0 = L(G,C_a)$ и $y_2 \| L_0 = L(G,B_m)$. Цепочки $y_1$ и $y_2$ являются префиксами предложения $x$, поэтому между ними возможно одно из трех соотношений: $y_1 = y_2$, $y_1y_0 = y_2$ или $y_1 = y_2y_0$, где $y_0 \neq e$.

Если $y_1 = y_2$, то $y_1 \| L_0 = y_2 \| L_0$, $L(G,C_a) = L(G,B_m)$ и, как следует из условия A5, $C_a = B_m$.

Если $y_1y_0 = y_2$, то $L(G,C_a) = y_1 \| L_0 = (y_1 \| L_a) \times L(G,B_m) = L'_a \times L(G,B_m)$, где $L'_a$ — некоторый разделенный язык, существующий в силу леммы 2, т. е. $C_a \in N_{B_m}$.

Если $y_1 = y_2y_0$, то $L(G,B_m) = y_2 \| L_0 = L''_a \times L(G,C_a)$, где $L''_a = y_2 \| L(G,a\rho_a)$ — разделенный язык, существующий в силу леммы 2, т. е. $h(B_m) > h(C_a)$, что противоречит выбору $B_m$.

Итак, для $A \in N_{B_m}$ произвольное $A$-правило имеет вид $A \to a\rho_a C_a$, причем $C_a = B_m$ либо $C_a \in N_{B_m}$, что означает $B_m \in RD(G,B_m)$.

**6. Второй характеристический признак $CS$-грамматик.** Этот признак позволяет для идентификации канонических разделенных грамматик использовать альтернативную формулировку условия A5. Следуя [5], введем на множестве нетерминалов отношение структурной эквивалентности.

Пусть $G = \langle N, \Sigma, P, S\rangle$ — грамматика и $E$ — бинарное отношение на некотором подмножестве нетерминалов $N'$. Обозначим символом $\underset{E}{\simeq}$ отношение эквивалентности, полученное из $E$ посредством рефлексивного и транзитивного замыкания отношения $E \cup \{(D,D)\colon\ D \in N'\}$. Доопределим

отношение $\underset{E}{\simeq}$ на множество цепочек следующим образом:

$$\alpha_0 A_1 \alpha_1 \dots A_n \alpha_n \underset{E}{\simeq} \alpha_0 B_1 \alpha_1 \dots B_n \alpha_n \Leftrightarrow \Big(A_i \underset{E}{\simeq} B_i \text{ для } i = 1, \dots, n\Big).$$

Отношение $\underset{E}{\simeq}$ будем называть *отношением структурной эквивалентности* и обозначать его $\underset{E}{\approx}$, если из $A \underset{E}{\simeq} B$ следует, что для любого правила $A \to \varphi$ существует правило $B \to \psi$, для которого $\varphi \underset{E}{\simeq} \psi$. Отношение $\underset{E}{\approx}$ будем называть *тривиальным*, если из $A \underset{E}{\approx} B$ следует $A = B$. Известно, что из $A \underset{E}{\approx} B$ следует $L(G, A) = L(G, B)$. Будем рассматривать грамматики, удовлетворяющие условию

A5′. Все отношения структурной эквивалентности на $N'$ тривиальны.

Т е о р е м а 4. *Если QCS-грамматика $G$ удовлетворяет условиям* A5′ *и* A6, *то $G$ является CS-грамматикой.*

Д о к а з а т е л ь с т в о. Пусть $G$ — $QCS$-грамматика, удовлетворяющая условиям теоремы 4. Обозначим через $E$ множество пар различных нетерминалов $(A, B)$, для которых $L(G, A) = L(G, B)$. Докажем, что $E = \emptyset$, что эквивалентно условию A5.

Предположим, что $E \neq \emptyset$, т. е. $L(G, A) = L(G, B)$ для некоторых $A \neq B$. Поскольку $G$ — $QCS$-грамматика, то любому $A$-правилу $A \to a\eta$ можно взаимно однозначно сопоставить $B$-правило $B \to a\rho$, причем $L(G, a\eta) = L(G, a\rho)$.

Равенство $L(G, a\eta) = L(G, a\rho)$ возможно либо при $a\eta = a\rho$, либо при $a\eta \neq a\rho$. В первом случае соотношение $a\eta \underset{E}{\approx} a\rho$ очевидно. Во втором случае из леммы 5 для цепочек $a\eta$ и $a\rho$ следует, что $a\eta = a\eta_1 A_1 \mu_0$ и $a\rho = a\rho_1 B_1 \mu_0$, причем для пары различных нетерминалов $A_1$ и $B_1$ возможно одно из трех: либо $L(G, A_1) = L(G, B_1)$, либо $L(G, A_1) = L(G, \psi' B_1)$, либо $L(G, \psi' A_1) = L(G, B_1)$. Второе и третье равенства противоречат условию A6, а из первого равенства следует: $(A_1, B_1) \in E$ и $L(G, a\eta_1) = L(G, a\rho_1)$.

Равенство $L(G, a\eta_1) = L(G, a\rho_1)$ рассматривается аналогично, оно возможно либо при $a\eta_1 = a\rho_1$, либо при $a\eta_1 = a\eta_2 A_2 \mu_1$ и $a\rho_1 = a\rho_2 B_2 \mu_1$, где $(A_2, B_2) \in E$ и $L(G, a\eta_2) = L(G, a\rho_2)$.

Генерация пар $a\eta_i$ и $a\rho_i$ не может быть бесконечной, поскольку $|\eta_1|, |\eta_2|, \dots$ есть монотонно убывающая последовательность. На некотором $n$-м этапе будут получены представления цепочек $a\eta$ и $a\rho$ в виде $a\eta_n A_n \mu_{n-1} \dots A_1 \mu_0$ и $a\eta_n B_n \mu_{n-1} \dots B_1 \mu_0$, причем $\{(A_n, B_n), \dots, (A_1, B_1)\} \subseteq E$, т. е. $a\eta \underset{E}{\approx} a\rho$ и $\underset{E}{\approx}$ есть нетривиальное отношение структурной эквивалентности, что противоречит условию A5′.

**7. Единственность $CS$-грамматик.** В каждом классе эквивалентности разделенных грамматик существует ровно одна унифицированная каноническая грамматика. Это утверждение устанавливается в два приема: сначала доказывается, что эквивалентные $CS$-грамматики совпадают с точностью до переименования нетерминалов, а затем излагается метод унификации алфавита нетерминальных символов. Здесь и далее под переименованием понимается применение некоторого взаимно однозначного отображения.

Т е о р е м а 5. *Пусть $G_1 = \langle N_1, \Sigma_1, P_1, S_1\rangle$ и $G_2 = \langle N_2, \Sigma_1, P_2, S_2\rangle$ — эквивалентные CS-грамматики, в которых $S_1 \neq S_2$ и $A \in N_1' \cap N_2' \Leftrightarrow L(G_1, A) = L(G_2, A)$. Тогда $N_1' = N_2'$, $R(G_1, S_1) = R(G_2, S_2)$ и $\widehat{P_1}(N_1') = \widehat{P_2}(N_2')$.*

Д о к а з а т е л ь с т в о. Обозначим: $\Delta_1 = N_1' \setminus N_2'$ и $\Delta_2 = N_2' \setminus N_1'$. Выберем не встречающиеся в $G_1$ и $G_2$ символы $+$, $\div$ и $S$.

*Часть I.* Построим грамматику $G = \langle N, \Sigma, P, S\rangle$ следующим образом:

$$\begin{aligned} N &= \{S\} \cup N_1 \cup N_2 \equiv \{S\} \cup \{S_1, S_2\} \cup N_1' \cup \Delta_2, \qquad \Sigma = \{+, \div\} \cup \Sigma_1, \\ P &= \{S \to + S_1,\ S_1 \to +\} \cup P(S_1) \cup \widehat{P_1}(N_1') \cup \\ &\quad \cup \{S \to \div S_2,\ S_2 \to \div\} \cup P(S_2) \cup \widehat{P_2}(\Delta_2). \end{aligned}$$

По построению для нетерминалов из $N'$ справедливы равенства:

$$\begin{aligned} &L(G, S_1) = L(G_1, S_1) \cup \{+\}, \qquad L(G, A) = L(G_1, A) \text{ для всех } A \in N_1', \\ &L(G, S_2) = L(G_2, S_2) \cup \{\div\}, \qquad L(G, B) = L(G_2, B) \text{ для всех } B \in \Delta_2. \end{aligned} \tag{2}$$

*Часть II.* Покажем, что $G$ — $CS$-грамматика. Условия A1, A2 и A3 вытекают из определения $G$, условие A4 следует из равенств (1). Условие A5 подтверждается следующими выкладками:

$$L(G,S_1)\neq L(G_1,S_2), \qquad L(G,A)\neq L(G,B) \text{ для всех различных } A,B\in N_1'\cup\Delta_2,$$
$$L(G,S_1)\neq L(G,C), \qquad L(G,S_2)\neq L(G,C) \text{ для всех } C\in N_1'\cup\Delta_2.$$

Остановимся на проверке условия A6. Соответствующие неравенства с участием $S_1$ и $S_2$ следуют (для любых $L_0$ и $A\in N_1'\cup\Delta_2$) из уникальности терминалов $\div$ и $+$:

$$L(G,S_1)\neq L_0{\times}L(G,S_2), \qquad L(G,S_2)\neq L_0{\times}L(G,S_1),$$
$$L(G,S_i)\neq L_0{\times}L(G,A_i), \qquad L(G,A_i)\neq L_0{\times}L(G,S_i),\ i{=}1,2.$$

Таким образом, проверка условия A6 сводится к доказательству неравенства $L(G,A)\neq L_0{\times}L(G,B)$ для всех $L_0$ и $A\neq B$ из $N_1\cup\Delta_2$. Предположим обратное: $L(G,A)=L_0\times L(G,B)$ для некоторых $L_0$ и $A\neq B$. Нетерминал $A$ является символом алфавита $N_i$, $i=1$ или $i=2$, поэтому $L(G_i,A)=L_0\times L(G,B)$, что противоречит теореме 2, так как $G_i$ — $CS$-грамматика, а $L_0$ и $L(G,B)$ — разделенные языки (см. свойство 1 разделенных грамматик).

*Часть III.* Покажем справедливость равенства $R(G_1,S_1)=R(G_2,S_2)$. Пусть $a\eta\in R(G_1,S_1)$. Из эквивалентности $G_1$ и $G_2$ следует, что в $R(G_2,S_2)$ существует цепочка $a\rho$, начинающаяся тем же терминалом $a$. По построению в грамматике $G$ имеются правила $S_1\to a\eta$ и $S_2\to a\rho$, причем $L(G,a\eta)=L(G,a\rho)$, и, как следует из теоремы 1, $a\eta=a\rho$.

*Часть IV.* Покажем, что выполняются равенства $N_1'=N_2'$ и $L(G_1,\varphi)=L(G_2,\varphi)$ для любой цепочки $\varphi\in(N_1'\cup\Sigma)^+$. Рассмотрим произвольное предложение $ax$ из $L(G,S_2)\setminus\{\div\}\equiv L(G_2,S_2)$ и его вывод в грамматике $G$: $S_2\Rightarrow_G a\rho\Rightarrow_G^* ax$. Как следует из части III настоящего доказательства, цепочка $\rho$ не содержит символы из $N_2$, а поскольку $P_1\subset P$, то в выводе предложения $ax$ не участвуют нетерминалы $\Delta_2$. Так как грамматика $G$ удовлетворяет условию A2, то $\Delta_2=\emptyset$. Аналогично доказывается $\Delta_1=\emptyset$. Из $N_1'=N_2'$ следует, что $L(G_1,A)=L(G_2,A)$ для всех $A\in N_1'$ и поэтому

$$\begin{aligned}L(G_1,\varphi)=L(G_1,x_0A_1x_1\ldots A_kx_k)&=\{x_0\}{\times}L(G_1,A_1){\times}\cdots{\times}L(G_1,A_k){\times}\{x_k\}\\&=\{x_0\}{\times}L(G_2,A_1){\times}\cdots{\times}L(G_2,A_k){\times}\{x_k\}=L(G_2,\varphi).\end{aligned}$$

*Часть V.* Покажем, что $\widehat{P_1}(N_1')=\widehat{P_2}(N_2')$. Предположим, что в $G_1$ и $G_2$ существуют правила $A\to a\eta$ и $A\to a\rho$, в которых $a\eta\neq a\rho$, но $L(G_1,a\eta)=L(G_2,a\rho)$. Как доказано в части IV $L(G_1,a\eta)=L(G_2,a\eta)$, т. е. в $CS$-грамматике $G_2$ существуют различные цепочки $a\eta$ и $a\rho$, для которых $L(G_2,a\eta)=L(G_2,a\rho)$, что противоречит теореме 1.

Следствие. Пусть $G_1=\langle N_1,\Sigma,P_1,S\rangle$ и $G_2=\langle N_2,\Sigma,P_2,S\rangle$ две эквивалентные $CS$-грамматики, в которых $D\in N_1\cap N_2\Leftrightarrow L(G_1,D)=L(G_2,D)$. Тогда $N_1=N_2$ и $P_1=P_2$.

Пусть $G_{\text{fix}}=\langle N,\Sigma,P,S\rangle$ — разделенная грамматика, удовлетворяющая условию A2, и $\Gamma$ — символ, не встречающийся в $G_{\text{fix}}$. Сопоставим грамматике $G_{\text{fix}}$ эквивалентную ей *унифицированную* грамматику $\overline{G}_{\text{fix}}=\langle\overline{N},\Sigma,\overline{P},\Gamma_\emptyset\rangle$, в которой каждый нетерминал $D$ переименован в $\Gamma_{\overline{D}}$, где $\overline{D}$ — специально построенное множество, исполняющее роль индекса при $\Gamma$.

Для вычисления индексов $\overline{D}$ используется ориентированный граф $\langle N,W\rangle$, помеченные дуги которого порождаются правилами грамматики $G_{\text{fix}}$:

$$(D,B)_{a_m}\in W\Leftrightarrow\Big(D\to a\mu B\eta\in P\quad\&\quad m=|a\mu B|\Big);$$

метка $a_m$ определяется первым символом альтернативы $D$-правила и позицией нетерминала $B$ в этой альтернативе. Переименования нетерминалов в грамматике $G_{\text{fix}}$ порождают графы, изоморфные $\langle N,W\rangle$, причем в этих графах сохраняется разметка ребер. Смежность вершин и разметка ребер в графах, изоморфных $\langle N,W\rangle$, однозначно определяют позиции нетерминалов в правилах вывода, а также задают группы позиций, отвечающих каждому нетерминалу.

В графе $\langle N,W\rangle$ для каждого нетерминала $D$ существует путь от вершины $S$ к вершине $D$. Последовательность меток, расположенных на дугах этого пути, образует $D$-фразу; длина $D$-фразы измеряется количеством дуг. По определению индекс $\overline{D}$ есть совокупность всех $D$-фраз, длина которых не превосходит число $\max\{w(D')\colon\ D'\in N\}+1$, где $w(D')$ — минимальная длина $D'$-фраз. В случае $G_{\text{fix}}=\widehat{\widehat{g}}$ (см. п. 6) унифицированная грамматика $\overline{G}_{\text{fix}}$ получается так:

$N = \{S, A, C\}$; $W = \{(S, C)_{b_2}, (S, A)_{c_2}, (S, A)_{b_3}, (A, A)_{b_2}, (A, C)_{c_2}, (A, A)_{c_3}\}$;
$w(S) = 0$, $w(A) = 1$, $w(C) = 1$; $\max\{0, 1, 1\} + 1 = 2$;
$\overline{S} = \emptyset$, $\overline{A} = \{b_3, c_2, b_3b_2, b_3c_3, c_2b_2, c_2c_3\}$, $\overline{C} = \{b_2, b_3c_2, c_2c_2\}$;
$\overline{G}_{\text{fix}}$: $\Gamma_\emptyset \to b\Gamma_{\overline{C}}\Gamma_{\overline{A}}|c\Gamma_{\overline{A}}$, $\Gamma_{\overline{A}} \to a|b\Gamma_{\overline{A}}|c\Gamma_{\overline{C}}\Gamma_{\overline{B}}$, $\Gamma_{\overline{C}} \to a|b$.

Множество индексов $\overline{N} = \{\overline{D}\colon\ D \in N\}$ позволяет построить следующий граф $\langle \overline{N}, \overline{W}\rangle$, изоморфный $\langle N, W\rangle$:

$$(\overline{D}, \overline{B})_{a_i} \in \overline{W} \Leftrightarrow \exists x\Big(x \in \overline{D} \quad \& \quad xa_i \in \overline{B}\Big).$$

Поэтому CS-грамматики эквивалентны тогда и только тогда, когда они преобразуются в одну и ту же унифицированную грамматику $\overline{G}$.

**8. Заключительные замечания.** В связи с теоремой 1 возникает вопрос о существовании разделенных грамматик, для которых из $L(G, \varphi) \supseteq L(G, \psi)$ следует $\varphi \Rightarrow^* \psi$. Далее, в формулировках теорем 1 и 4 условие A6 может быть ослаблено, а именно, достаточно предположить, что $L_0$ является производным языком заданной $QCS$-грамматики. И наконец, как следует из теорем 3 и 4, наиболее сложные для проверки условия A5 и A6 можно изменять по отдельности, но не одновременно (см. грамматику $\widehat{g}$); вопрос о дополнительном условии, гарантирующем совместно с A5′ и A6′ каноничность $QCS$-грамматик, остается открытым.

## СПИСОК ЛИТЕРАТУРЫ

1. Courcelle B. Une forme canonique pour les grammaires simples deterministes // RAIRO — Theoretical Informatics and Applications. 1974. **8.R1.** P. 19–36.
2. Ахо А., Ульман Дж. Теория синтаксического анализа, перевода и компиляции. Т. 1. М.: Мир, 1978.
3. Серебряков В. А., Соловьев С. Ю. Задача совместимости свойств формальных грамматик // Информационные процессы. 2012. **12**. № 4. С. 408–437.
4. Bastien C., Czyzowicz J., Fraczak W., Rytter W. Prime normal form and equivalence of simple grammars // Implementation and Application of Automata. Lecture Notes in Computer Science. Vol. 3845. Berlin: Springer, 2006. P. 78–89.
5. Paull M. C., Unger S. H. Structural equivalence of context-free grammars // J. Computer and System Sci. 1968. **2**. P. 427–463.

Поступила в редакцию
16.09.15

## SINGULARITIES OF CANONICAL SEPARATED GRAMMARS

**Soloviev S. Y.**

Class canonical separated grammars that generate the same language as the general separated grammar are considered. We present the basic properties and two criterion of canonical grammars. We propose a method of nonterminals unification and prove the uniqueness of the canonical representation separated grammar.

*Keywords*: separated grammars, uniqueness.