

Контекстно-свободные грамматики и магазинные автоматы

Контекстно-свободной грамматикой (сокращённо КСГ) называется четвёрка $G = \langle V, \Sigma, R, S \rangle$, где

- V — внешний алфавит;
- $\Sigma \subset V$ — внутренний алфавит;
- $R \subseteq (V \setminus \Sigma) \times V^*$ — множество правил;
- $S \in V \setminus \Sigma$ — стартовый символ.

Символы, принадлежащие Σ , называем *терминальными* символами, а символы из $V \setminus \Sigma$ — *нетерминальными*. Стартовый символ по определению всегда является нетерминальным. Множество правил R представляет из себя множество пар вида $\langle A, \alpha \rangle$, где A — нетерминальный символ, а α — произвольное слово алфавита V . Будем использовать следующую форму записи: если пара $\langle A, \alpha \rangle$ принадлежит множеству правил R , то мы говорим, что правило $A \rightarrow \alpha$ входит в набор правил грамматики G . Таким образом, мы воспринимаем R как конечный список записей вида "терминальный символ \rightarrow слово".

Для $w, v \in V^*$ пишем $w \Rightarrow_G v$, если $w = u_1 A u_2$, $v = u_1 x u_2$ и в R присутствует правило $A \rightarrow x$. Другими словами, v получается из w заменой вхождения символа A на слово x .

Пишем $w \Rightarrow_G^* v$, если существует конечная последовательность

$$w = u_0 \Rightarrow_G u_1 \Rightarrow_G \dots \Rightarrow_G u_n = v$$

Эту последовательность называем *выводом*. В случае существования такого вывода мы говорим, что слово v *выводится* в грамматике G из слова w . Говорим, что слово $v \in \Sigma^*$ *порождается* грамматикой G , если оно выводится из слова S (состоящего из единственного стартового символа). Множество всех слов из Σ^* называется *языком, порождаемым грамматикой G* и обозначается $L(G)$. Всякий язык, порождаемый некоторой контекстно-свободной грамматикой, называется контекстно-свободным языком (сокращённо КСЯ).

Как будет доказано ниже, все регулярные языки являются контекстно-свободными. Однако существуют контекстно-свободные языки, не являющиеся регулярными: например, язык $\{a^n b^n : n \in \mathbb{N}\}$, порождаемый грамматикой с правилами $S \rightarrow aSb, S \rightarrow e$.

Рассмотрим вывод

$$w = u_0 \Rightarrow_G u_1 \Rightarrow_G \dots \Rightarrow_G u_n = v$$

и пусть $w = w_0 w_1 \dots w_k$. Тогда для каждого $i \leq n$ верно равенство $u_i = u_i^0 u_i^1 \dots u_i^k$ для некоторых слов u_i^0, \dots, u_i^k , причём при $i < j$ и $l \in [0, k]$ $u_i^l \Rightarrow_G^* u_j^l$. Действительно, при $i = 0$ это верно. Пусть это верно для некоторого $i < n$. Слово u_{i+1} получается из u_i заменой вхождения некоторого терминального символа A на слово $x \in V^*$ по правилу грамматики $A \rightarrow x$. Этот символ входит в одно из слов u_i^1, \dots, u_i^k : пусть он входит в слово u_i^l . Тогда $u_i^l = yAz$ для некоторых $y, z \in V^*$ и мы можем положить $u_{i+1}^l = yxz$, $u_{i+1}^m = u_i^m$ для $m \neq l$. Заметим также, что описанная конструкция даёт следующий факт: для $0 \leq i < j \leq n$ и $l \in [0, k]$ существует вывод u_j^l из u_i^l длины, не превосходящей $j - i$.

Положив $i = 0$ и $j = n$, получаем такое следствие: если $w_0 \dots w_k \Rightarrow v$, то существует разложение $v = v_0 \dots v_k$, такое что $w_l \Rightarrow_G^* v_l$ для любого l от 0 до k . Обратное тоже верно. Пусть $w = w_0 \dots w_k$, $v = v_0 \dots v_k$ и $w_i \Rightarrow_G^* v_i$ для каждого $i \in [0, k]$. Пусть для $i \in [0, k]$

$$w_i = u_0^i \Rightarrow_G u_1^i \Rightarrow_G \dots \Rightarrow_G u_{n_i}^i = v_i$$

есть вывод v_i из w_i . Тогда

$$\begin{aligned} w_0 \dots w_k &= u_0^0 u_0^1 \dots u_0^k \Rightarrow_G^* u_1^0 u_1^1 \dots u_1^k \Rightarrow_G \dots \Rightarrow_G u_{n_0}^0 u_0^1 \dots u_0^k = \\ &= u_{n_0}^0 u_0^1 \dots u_0^k \Rightarrow_G u_{n_0}^0 u_1^1 u_0^2 \dots u_0^k \Rightarrow_G \dots \Rightarrow_G u_{n_0}^0 u_{n_1}^1 u_0^2 \dots u_0^k = \\ &\quad \dots \quad \dots \quad \dots \\ &= u_{n_0}^0 \dots u_{n_{k-1}}^{k-1} u_0^k \Rightarrow_G u_{n_0}^0 \dots u_{n_{k-1}}^{k-1} u_1^k \Rightarrow_G \dots \Rightarrow_G u_{n_0}^0 \dots u_{n_{k-1}}^{k-1} u_{n_k}^k = v_0 \dots v_k \end{aligned}$$

есть вывод v из w .

Вывод

$$w = u_0 \Rightarrow_G u_1 \Rightarrow_G \dots \Rightarrow_G u_n = v$$

называется *левым* (или *левосторонним*), если для любого $i < n$ слово u_{i+1} получается из слова u_i заменой самого левого вхождения нетерминального символа согласно одному из правил грамматики.

Лемма 1 Пусть $G = \langle V, \Sigma, R, S \rangle$, $w \in V^*$ и $v \in \Sigma^*$. Если v выводима из w в G , то существует левый вывод v из w .

Доказательство. Докажем лемму индукцией по длине вывода v из w . Если есть вывод длины 1, то он, очевидно, левый. Пусть теперь

$$w = u_0 \Rightarrow_G u_1 \Rightarrow_G \dots \Rightarrow_G u_n = v$$

вывод длины $n > 1$ и для всех случаев, когда имеется вывод длины, меньшей n , существует левый вывод. Тогда $w = yAz$ и $u_1 = yxz$ для некоторых $A \in V \setminus \Sigma$, $x, y, z \in V^*$ и правила $A \rightarrow x$ из R . Из рассмотрений перед леммой следует, что существуют $v_1, v_2 \in \Sigma^*$, такие что v_1 выводится из y , v_2 выводится из xz и в обоих случаях выводы имеют длину, меньшую n . По предположению индукции существуют левые выводы

$$\begin{aligned} y = s_0 \Rightarrow_G s_1 \Rightarrow_G \dots \Rightarrow_G s_k = v_1 \\ xz = t_0 \Rightarrow_G t_1 \Rightarrow_G \dots \Rightarrow_G t_m = v_2 \end{aligned}$$

Но тогда

$$\begin{aligned} w = xAz = s_0Az \Rightarrow_G s_1Az \Rightarrow_G \dots \Rightarrow_G s_kAz \Rightarrow_G s_kxz = \\ = s_k t_0 \Rightarrow_G s_k t_1 \Rightarrow_G \dots \Rightarrow_G s_k t_m = v_1 v_2 = v \end{aligned}$$

есть левый вывод v из w . \square

Левые выводы нам понадобятся при рассмотрении магазинных автоматов. А пока перейдём к дальнейшему изучению контекстно-свободных языков.

Теорема 1 Множество контекстно-свободных языков замкнуто относительно операций объединения, конкатенации и звёздочки Клини.

Доказательство. Пусть $G_1 = \langle V_1, \Sigma, R_1, S_1 \rangle$ и $G_2 = \langle V_2, \Sigma, R_2, S_2 \rangle$ — контекстно-свободные грамматики, такие что $V_1 \cap V_2 = \Sigma$ и $L_1 = L(G_1)$, $L_2 = L(G_2)$. Пусть $S \notin V_1 \cup V_2$.

1. Рассмотрим грамматику $G = \langle V, \Sigma, R, S \rangle$, где $V = V_1 \cup V_2 \cup \{S\}$, а R состоит из элементов $R_1 \cup R_2$ и ещё двух правил: $S \rightarrow S_1$, $S \rightarrow S_2$. Тогда $L(G) = L_1 \cup L_2$.

2. Рассмотрим грамматику $G = \langle V, \Sigma, R, S \rangle$, где $V = V_1 \cup V_2 \cup \{S\}$, а R состоит из элементов $R_1 \cup R_2$ и правила $S \rightarrow S_1 S_2$. Тогда $L(G) = L_1 L_2$.
3. Рассмотрим грамматику $G = \langle V, \Sigma, R, S \rangle$, где $V = V_1 \cup \{S\}$, а R состоит из элементов R_1 и двух дополнительных правил: $S \rightarrow S_1 S$, $S \rightarrow e$. Тогда $L(G) = L_1^*$.

Все три утверждения достаточно очевидны. Строгое доказательство основывается на рассмотрении выводов и выкладках, подобных выкладкам перед леммой 1. Оно оставляется желающим в качестве упражнения. \square

Следствие 1 *Все регулярные языки являются контекстно-свободными.*

Доказательство. Достаточно заметить, что все базисные языки контекстно-свободны и применить теорему 1. \square

Для грамматики G через $\varphi(G)$ обозначим максимум длины слов, находящихся в правых частях правил этой грамматики.

Теорема 2 (о накачке для КСЯ) *Пусть $G = \langle V, \Sigma, R, S \rangle$ — контекстно-свободная грамматика. Тогда для любого $w \in L(G)$, такого что $|w| > \varphi(G)^{|V|}$, существуют $u, v, x, y, z \in \Sigma^*$, такие что $w = uvxyz$, $vu \neq e$ и для любого $i \in \mathbb{N}$ $uv^i x u^i z \in L(G)$.*

Доказательство. Пусть $w \in L(G)$, $|w| > \varphi(G)^{|V|}$ и

$$S = u_0 \Rightarrow_G u_1 \Rightarrow_G \dots \Rightarrow_G u_n = w$$

вывод слова w минимально возможной длины.

Под *цепочкой* будем понимать произвольную конечную последовательность элементов множества V (то есть, по сути, слово алфавита V). Для каждого i от 0 до n сопоставим каждой букве слова u_i некоторую цепочку.

Слово u_0 состоит из одной единственной буквы S , которой мы сопоставляем цепочку длины 1, состоящую из того же самого символа S . Далее действуем по индукции. Пусть $i < n$ и всем буквам слова u_i цепочки уже сопоставлены. Имеем $u_i = yAz$ и $u_{i+1} = yxz$, где $x, y, z \in V^*$, $A \in V \setminus \Sigma$ и правило $A \rightarrow x$ присутствует в грамматике. Буквам слова

u_{i+1} , лежащим в частях y и z разложения yxz , сопоставляем те же самые цепочки, что и аналогичным буквам, лежащим в тех же самых частях слова u_i . Если же какая-то буква $a \in V$ лежит в части x , то сопоставляем ей цепочку, состоящую из цепочки, сопоставленной символу A слова u_i , к которой в конце дописан символ a .

Проделав описанный выше индуктивный переход $n - 1$ раз, получаем правило, сопоставляющее каждой букве слова $w = u_n$ некоторую цепочку. Все символы этой цепочки, кроме последнего, являются нетерминальными, а последний — терминальным и совпадающим с той самой буквой, которой она сопоставлена.

У нас имеется ровно $|w|$ цепочек¹, сопоставленных буквам слова w . Поскольку каждая из них стала продолжением цепочки длины 1 на первом шаге построения мультимножества цепочек, то все они являются продолжениями каких-то цепочек длины 2, а общее количество этих цепочек длины 2 не превосходит $\varphi(G)$ (с учётом одинаковых). Далее, некоторые из получающихся по ходу построения цепочек длины 2 в ходе дальнейших построений превращаются в цепочки длины 3, причём каждая цепочка длины 2 служит началом не более чем для $\varphi(G)$ цепочек длины 3, возникающих по ходу построения. Следовательно, всего цепочек длины 3, возникающих по ходу построения, вместе с цепочками длины 2, не получающимися продолжения в процессе построения цепочек, не более чем $\varphi(G)^2$. Рассуждая аналогично, получаем, что возникающих по ходу построения цепочек длины 4 вместе с не получающимися продолжения цепочками длины ≤ 3 не более чем $\varphi(G)^3$ и т. д.

Таким образом, если все цепочки имеют длину $\leq |V| + 1$, то их общее количество не превосходит $\varphi(G)^{|V|}$. Однако всего цепочек столько же, сколько букв в слове w , то есть больше чем $\varphi(G)^{|V|}$. Значит, одна из цепочек $A_0A_1 \dots A_ka$, сопоставленных буквам слова w имеет длину $k+1 > |V|+1$. Последний символ в этой цепочке — терминальный, осталь-

¹Отметим, что совокупность рассматриваемых на каждом шаге цепочек является не множеством, а мультимножеством, то есть множеством, в котором могут встречаться повторяющиеся элементы. К примеру, пусть u_1 получается из $u_0 = S$ согласно правилу $S \rightarrow AA$. Тогда как первой, так и второй буквам слова u_1 сопоставляется одна и та же цепочка SA . Мы считаем их разными цепочками, поскольку они сопоставлены разным буквам, несмотря на то, что длины обоих цепочек и последовательности букв в них совпадают. Говоря другими словами, совокупность цепочек, сопоставленных буквам слова u_1 , в данном случае является двухэлементным мультимножеством $\{SA, SA\}$ а не одноэлементным множеством $\{SA\}$.

ные — нетерминальные и последовательность нетерминальных символов $S = A_1, \dots, A_k$ содержит $k > |V|$ членов, являющихся элементами множества V . Следовательно, $A_i = A_j = A$ для некоторых $A \in V \setminus \Sigma$ и $1 \leq i < j \leq k$.

Из описания процесса построения цепочек заключаем, что существуют $0 \leq s < t \leq n$, такие что $u_s = u''Az''$, $u_t = u'v'Ay'z'$ для некоторых $u', v', y', z', u'', z'' \in V^*$, причём $A \Rightarrow_G^* v'Ay'$ и существуют $u, v, x, y, z \in \Sigma^*$, такие что $u'' \Rightarrow_G^* u' \Rightarrow_G^* u$, $v' \Rightarrow_G^* v$, $A \Rightarrow_G^* x$, $y' \Rightarrow_G^* y$, $z'' \Rightarrow_G^* z' \Rightarrow_G^* z$ и $w = uvxyz$. Если $vy = e$, то исходный вывод слова w длины n можно сократить, убрав из него шаги, соответствующие выводам $A \Rightarrow_G^* v'Ay'$ и $v' \Rightarrow_G^* v$, $y' \Rightarrow_G^* y$; однако n — минимально возможная длина вывода слова w . Следовательно, $vy \neq e$.

Осталось показать, что $uv^i xy^i z \in L(G)$ для любого $i \in \mathbb{N}$. Из сказанного выше имеем $S \Rightarrow_G^* uAz$ и $A \Rightarrow_G^* vAy$. Таким образом, для любого i получаем

$$S \Rightarrow_G^* uAy \underbrace{\Rightarrow_G^* uvAy \Rightarrow_G^* uv^2Ay^2z \Rightarrow_G^* \dots \Rightarrow_G^* uv^i Ay^i z}_{i \text{ раз}} \Rightarrow_G^* uv^i xy^i z,$$

что, собственно, и требуется. \square .

Следствие 2 Язык $\{a^p : p \text{ — простое число}\}$ не является контекстно-свободным.

Доказательство. Аргумент, аналогичный аргументу, использованному при доказательстве нерегулярности этого языка через теорему о накачке для регулярных языков. \square

Следствие 3 Язык $\{a^n b^n c^n : n \in \mathbb{N}\}$ не является контекстно-свободным.

Доказательство. Пусть этот язык совпадает с $L(G)$ для некоторой контекстно-свободной грамматики $G = \langle V, \{a, b, c\}, R, S \rangle$. Положим $n = \varphi(G)^{|V|}$ и рассмотрим слово $a^n b^n c^n$, которое, очевидно, принадлежит нашему языку. По теореме о накачке для КСЯ существует представление $a^n b^n c^n = uvxyz$ этого слова, такое что $vy \neq e$ и $uxz, uv^2xy^2z \in L(G)$. Однако если слово vy содержит не более двух различных букв алфавита $\{a, b, c\}$, то количества букв a, b, c в слове uxz не равны, а если vy содержит все три буквы, то для одного из слов v, y первая и последняя

буквы не совпадают, из-за чего в слове uv^2xy^2z либо буква a встречается после буквы b , либо буква b после буквы c . В обоих случаях имеем противоречие. \square

Следствие 4 *Класс контекстно-свободных языков не замкнут относительно пересечений.*

Доказательство. Языки $L_1 = \{a^n b^n c^k : n, k \in \mathbb{N}\}$ и $L_2 = \{a^k b^n c^n : k, n \in \mathbb{N}\}$ контекстно-свободны. Действительно, первый является конкатенацией контекстно-свободного языка $\{a^n b^n : n \in \mathbb{N}\}$ и регулярного языка, задаваемого регулярным выражением c^* ; второй — конкатенацией регулярного языка, задаваемого регулярным выражением a^* , и контекстно-свободного языка $\{b^n c^n : n \in \mathbb{N}\}$. Однако $L_1 \cap L_2 = \{a^n b^n c^n : n \in \mathbb{N}\}$. \square

Следствие 5 *Класс контекстно-свободных языков не замкнут относительно взятия дополнений.*

Доказательство. Следует из отсутствия замкнутости относительно пересечений, ибо $L_1 \cap L_2 = \Sigma^* \setminus ((\Sigma^* \setminus L_1) \cup (\Sigma^* \setminus L_2))$ для любых языков $L_1, L_2 \in \Sigma^*$. \square

Магазинным автоматом (сокращённо МА) называется шестёрка $M = \langle Q, \Sigma, \Gamma, \Delta, s, F \rangle$, где

- Q — множество состояний автомата;
- Σ — алфавит;
- Γ — алфавит стека (мы считаем, что в Γ присутствует специальный символ t , ”маркирующий” начало (дно) стека);
- $\Delta \subseteq (Q \times (\Sigma \cup \{e\}) \times \Gamma) \times (Q \times \Gamma^*)$ — множество команд автомата. Если $\langle \langle q, a, A \rangle, \langle p, \alpha \rangle \rangle \in \Delta$, то мы говорим, что команда $qaA \rightarrow p\alpha$ присутствует в списке команд автомата. На множество команд наложено ограничение: если в списке команд присутствует команда $qaA \rightarrow p\alpha$, то при $A \neq t$ должно выполняться $\alpha \in (\Gamma \setminus \{t\})^*$, а если $A = t$, то $\alpha = \beta t$ для некоторого $\beta \in (\Gamma \setminus \{t\})^*$;
- $s \in Q$ — стартовое состояние;

- $F \subseteq Q$ — множество *конечных состояний*.

Существует естественное наглядное представление магазинных автоматов в виде устройств, состоящих из ленты, стековой памяти и управляющего устройства с двумя указателями. Мы полагаем, что читателям оно хорошо известно² и в дальнейшем будем неоднократно иметь его в виду (в частности, при изложении доказательств).

Конфигурация автомата M — тройка $\langle w, q, \alpha \rangle$, где $w \in \Sigma^*$, $q \in Q$ и $\alpha = \beta t$ для $\beta \in (\Gamma \setminus \{t\})^*$.

Для конфигураций пишем $\langle w, q, \alpha \rangle \vdash_M \langle v, p, \beta \rangle$, если $w = xv$ (где $x \in \Sigma \cup \{e\}$), $\alpha = A\gamma$, $\beta = \delta\gamma$ и в Δ присутствует команда $qx A \rightarrow p\delta$. Введённое обозначение формализует понятие ”шаг работы автомата”. Пишем $\langle w, q, \alpha \rangle \vdash_M^* \langle v, p, \beta \rangle$, если существует последовательность конфигураций

$$\langle w, q, \alpha \rangle = C_0 \vdash_M C_1 \vdash_M \dots \vdash_M C_k = \langle v, p, \beta \rangle$$

Говорим, что слово $w \in \Sigma^*$ *распознаётся* автоматом M , если выполнено $\langle w, s, t \rangle \vdash_M^* \langle e, q, t \rangle$ для одного из конечных состояний $q \in F$. Множество всех слов, распознаваемых автоматом M , называется *языком*, *распознаваемым автоматом M* , и обозначается $L(M)$.

Теорема 3 *Язык является контекстно-свободным тогда и только тогда, когда он распознаётся магазинным автоматом.*

Доказательство. Необходимость. Пусть $L = L(G)$ для $G = \langle V, \Sigma, R, S \rangle$. Определим автомат $M = \langle Q, \Sigma, \Gamma, \Delta, s, F \rangle$. Алфавит Σ уже задан. Полагаем $Q = \{s, q\}$, $F = \{q\}$, $\Gamma = V \cup \{t\}$. В Δ поместим следующие команды:

1. $set \rightarrow q(Sm)$;
2. $qeA \rightarrow qx$ для каждого правила $A \rightarrow x$ из R ;
3. $qaa \rightarrow qe$ для каждого символа $a \in \Sigma$.

Покажем, что для любых $w \in \Sigma^*$ и $\alpha \in (V \setminus \Sigma)V^* \cup \{e\}$ справедлива эквивалентность

$$S \Rightarrow_G^* w\alpha \iff \langle w, s, t \rangle \vdash_M^* \langle e, q, \alpha t \rangle$$

²По крайней мере, оно хорошо известно тем, кто посещал лекции по ТА. Ну а тем, кто не посещал... никто не обещал, что им будет легко.

Импликацию слева направо докажем индукцией по длине левого вывода $w\alpha$ из S . Для вывода длины 0 она очевидна из первого пункта описания Δ . Пусть

$$S = u_0 \Rightarrow_G u_1 \Rightarrow_G \dots \Rightarrow_G u_n \Rightarrow_G u_{n+1} = w\alpha$$

Тогда $u_n = yAz$ и $u_{n+1} = yxz$ для $y \in \Sigma^*$, $A \in V \setminus \Sigma$ и $z \in V^*$, причём в R присутствует правило $A \rightarrow x$. По индукционному предположению имеем $\langle y, s, m \rangle \vdash_M^* \langle e, q, Azm \rangle$. Из пункта 2 описания Δ справедливо $\langle e, q, Azm \rangle \vdash_M \langle e, q, xzm \rangle$, откуда $\langle y, s, m \rangle \vdash_M^* \langle e, q, xzm \rangle$. Далее, $w = yv$ для некоторого $v \in \Sigma^*$, такого что $xz = v\alpha$; следовательно, $\langle w, s, m \rangle \vdash_M^* \langle v, q, v\alpha m \rangle$. Остаётся заметить, что $\langle v, q, v\alpha m \rangle \vdash_M^* \langle e, q, \alpha m \rangle$ из третьего пункта описания Δ и мы имеем всё, что требуется.

Для доказательства импликации справа налево воспользуемся индукцией по числу шагов работы автомата. В силу $s \neq q$ оно заведомо больше нуля. Если оно равно 1, то $w = e$, $\alpha = S$ и утверждение очевидно. Пусть теперь

$$\langle w, s, m \rangle = C_0 \vdash_M C_1 \vdash_M \dots \vdash_M C_n \vdash_M C_{n+1} = \langle e, q, \alpha m \rangle$$

Пусть k — максимальное число от 1 до n , такое что на шаге $C_k \vdash_M C_{k+1}$ автомат выполняет команду $qeA \rightarrow qx$ для $A \in V \setminus \Sigma$ и $x \in V^*$ (на последующих шагах он выполняет команды вида $qaa \rightarrow qe$). Тогда $C_k = \langle v, q, A\beta m \rangle$, $C_{k+1} = \langle v, q, x\beta m \rangle$, $w = uv$ и $x\beta = v\alpha$ для некоторых $u, v \in \Sigma^*$ и $\beta, x \in V^*$. Имеем $\langle u, s, m \rangle \vdash_M^* \langle e, q, A\beta m \rangle$ за k шагов работы автомата и, по индукционному предположению, $S \Rightarrow_G^* uA\beta$. Но в грамматике присутствует правило $A \rightarrow x$ и, значит, $S \Rightarrow_G^* uA\beta \Rightarrow_G ux\beta = w\alpha$.

Таким образом, эквивалентность доказана. Положив теперь в ней $\alpha = e$, получаем $L(M) = L(G)$.

Достаточность. Пусть теперь задан автомат $M = \langle Q, \Sigma, \Gamma, \Delta, s, F \rangle$. Определим грамматику $G = \langle V, \Sigma, R, S \rangle$. Алфавит Σ уже задан. Положим $V = \Sigma \cup \{S\} \cup (Q \times \Gamma \times Q)$. В R поместим следующие правила:

1. $S \rightarrow \langle s, m, q \rangle$ для всех $q \in Q$;
2. $\langle q, A, p \rangle \rightarrow a \langle r, B_1, p_1 \rangle \langle p_1, B_2, p_2 \rangle \dots \langle p_{k-1}, B_k, p_k \rangle$ для всех $p_1, \dots, p_k \in Q$ и $B_1, \dots, B_k \in \Gamma$, таких что $p_k = p$ при $k > 0$ и в Δ присутствует команда $qaA \rightarrow r(B_1 B_2 \dots B_k)$;
3. $\langle q, m, q \rangle \rightarrow e$ для всех $q \in F$.

Отметим следующие свойства введённой грамматики.

1. Если $S \Rightarrow_G^* x$ для $x \in V^*$ и x получено в результате левого вывода, то $x = yz$, где $y \in \Sigma^*$, а $z \in (V \setminus \Sigma)^*$, причём t фигурирует в середине последнего символа из z (если, конечно, $z \neq e$) и не фигурирует в середине никакого другого символа из z . Свойство очевидно, поскольку в каждом из правил грамматики G правая часть принадлежит $\Sigma^*(V \setminus \Sigma)^*$ и при левом выводе каждый переход заключается в замене самого левого нетерминального символа.
2. Если для $w \in \Sigma^*$ и $A_1, \dots, A_k \in \Gamma \setminus \{t\}$ имеет место $\langle w, s, t \rangle \vdash_M^* \langle e, q, A_1 A_2 \dots A_k t \rangle$, то $S \Rightarrow_G^* w \langle q, A_1, p_1 \rangle \langle p_1, A_2, p_2 \rangle \dots \langle p_{k-1}, A_k, p_k \rangle \langle p_k, t, r \rangle$ для произвольных $p_1, \dots, p_k, r \in Q$. Это свойство легко доказать индукцией по числу шагов работы автомата, производящего преобразование конфигураций $\langle w, s, t \rangle \vdash_M^* \langle e, q, A_1 A_2 \dots A_k t \rangle$.
3. Если в грамматике G из S левым выводом выводится слово $w \langle q, A_1, p_1 \rangle \langle p_1, A_2, p_2 \rangle \dots \langle p_{k-1}, A_k, p_k \rangle \langle p_k, t, r \rangle$, где все символы и состояния такие же, как в предыдущем пункте, то $\langle w, s, t \rangle \vdash_M^* \langle e, q, A_1 \dots A_k t \rangle$. Свойство доказывается индукцией по длине левого вывода, при доказательстве учитывается свойство 1.

Для завершения доказательства остаётся применить свойства 2 и 3 к случаю $k = 0$, приняв во внимания третий пункт описания правил грамматики и тот факт, что каждое слово алфавита Σ допускает левый вывод. \square

Следствие 6 *Пересечение контекстно-свободного языка с регулярным языком есть контекстно-свободный язык.*

Доказательство. Пусть $M = \langle Q, \Sigma, \Gamma, \Delta, s, F \rangle$ — магазинный автомат и $M_1 = \langle Q_1, \Sigma, \Delta_1, s_1, F_1 \rangle$ — недетерминированный конечный автомат. Без ограничения общности можно считать, что $\langle q, e, q \rangle \in \Delta_1$ для каждого $q \in Q_1$.

Построим новый магазинный автомат $M' = \langle Q', \Sigma, \Gamma', \Delta', s', F' \rangle$. Положим в нём $Q' = Q \times Q_1$, $\Gamma' = \Gamma$, $s' = \langle s, s_1 \rangle$ и $F' = F \times F_1$. В Δ' поместим команды $\langle q, q_1 \rangle x A \rightarrow \langle p, p_1 \rangle \alpha$ для всех $q, p \in Q$, $q_1, p_1 \in Q_1$, $x \in \Sigma \cup \{e\}$, $A \in \Gamma$ и $\alpha \in \Gamma^*$, таких что $\langle q_1, x, p_1 \rangle \in \Delta_1$ и в Δ присутствует команда $qx A \rightarrow p\alpha$.

Нетрудно показать, что $L(M') = L(M) \cap L(M_1)$. Доказательство этого факта оставляется читателю в качестве упражнения. \square

Магазинный автомат M называется *детерминированным*, если для каждой конфигурации $\langle w, q, \alpha \rangle$ существует не более одной конфигурации $\langle v, p, \beta \rangle$, такой что $\langle w, q, \alpha \rangle \vdash_M \langle v, p, \beta \rangle$. Другими словами, магазинный автомат детерминирован, если в списке его команд не встречается команд с одинаковыми левыми частями и не встречается ситуация, когда присутствуют команды $qaA \rightarrow p\alpha$, $qeA \rightarrow r\beta$ для $a \in \Sigma$.

Контекстно-свободный язык L называется *детерминировано контекстно-свободным*, если язык $L\$$ распознаётся детерминированным магазинным автоматом (здесь $\$$ — символ, не входящий в алфавит Σ). Причина приписывания символа $\$$ заключается в том, чтобы разрешить детерминированным магазинным автоматам "распознавать" слова алфавита Σ , останавливаясь после их прочтения в одном из конечных состояний, но, возможно, не с пустым стеклом. Это сделано для того, чтобы расширить класс рассматриваемых языков, включив в него языки типа $a^* \cup \{a^n cb^n : n \in \mathbb{N}\}$.

Теорема 4 *Класс детерминированных контекстно-свободных языков замкнут относительно взятия дополнений.*

Доказательство. Пусть $M = \langle Q, \Sigma, \Gamma, \Delta, s, F \rangle$ — детерминированный магазинный автомат, распознающий язык $L\$$. Дополним (в случае необходимости) множества команд и состояний этого автомата таким образом, чтобы любое слово языка $\Sigma^*\$$ могло быть прочитано этим автоматом из любого состояния с любым стеклом до конца (то есть чтобы для любых $w \in \Sigma^*\$, \alpha \in (\Gamma \setminus \{t\})^*$ и $q \in Q$ существовало $p \in Q$, такое что $\langle w, q, \alpha t \rangle \vdash_M^* \langle e, p, t \rangle$).

Для этого сначала добавим в Q новые состояния q_0, q_1, q_2 , ранее там не лежавшие. Объявим состояние q_2 конечным, а состояния q_0 и q_1 не конечными. Затем добавим в Δ команды, позволяющие автомату, находясь в состоянии q_1 , читать буквы алфавита Σ , не производя изменений в стеке, а также команду, переводящую автомат из состояния q_1 в состояние q_0 при прочтении символа $\$$. Добавим также команды, позволяющие автомату, находясь в состоянии q_0 , освобождать стек (если стек не пуст), не меняя состояния.

Затем для каждого случая, попадающего в одну из описанных ниже двух категорий, добавим в Δ команды, гарантирующие корректную работу автомата во всех ситуациях.

1. Состояние неопределённости. Допустим, что для некоторых символов $a \in \Sigma \cup \{\$\}$ и $A \in \Gamma$, состояния $q \in Q$ и слова α , равного Am при $A \neq m$ и m при $A = m$, не существует конфигурации C , такой что $\langle a, q, \alpha \rangle \vdash_M C$ (то есть, читая на ленте символ a , видя на верхушке стека A и находясь в состоянии q , автомат не может сделать ни одного шага работы). Добавим в Δ команду $qaA \rightarrow q_1A$ при $a \neq \$$ и команду $qaA \rightarrow q_0A$ при $a = \$$ для устранения этой ситуации.
2. Тупиковые состояния. Допустим, что для некоторых символов $a \in \Sigma \cup \{\$\}$ и $A \in \Gamma$, состояния $q \in Q$ и слова α , равного Am при $A \neq m$ и m при $A = m$, если $\langle a, q, \alpha \rangle \vdash_M \langle v, p, \beta \rangle$, то $v = a$ и $\beta = \gamma\alpha$. Другими словами, автомат, видя на ленте символ a , на верхушке стека символ A и находясь в состоянии q , перестаёт читать буквы слова и начинает бесконечную работу со стеком, не опускаясь в нём ниже уже достигнутого уровня. Дополним Δ командой, уводящей автомат из этого тупикового состояния, сделав то же самое, что и в предыдущем случае.

Наконец, если в исходном (до внесения каких-либо изменений) автомате присутствовали команды $q\$A \rightarrow p\alpha$, то заменим каждую из них на команду $q\$A \rightarrow q_2\alpha$ при $p \in F$ и на $q\$A \rightarrow q_0\alpha$ при $p \notin F$. Также добавим команды, переводящие (не читая буквы слова) автомат из состояния q_2 в состояние q_0 в случае, если стек не пуст.

Ясно, что после внесённых изменений автомат останется детерминированным и будет читать до конца любое слово языка $\Sigma^*\$$ (останавливаясь в его конце с пустым стеком). Действительно, детерминированность очевидна по построению. Кроме того, при чтении слова невозможны состояния неопределённости и тупиковые состояния. А если автомат не может дочитать до конца какое-то слово, то он либо попадает в состояние неопределённости, либо, остановив чтение слова и начав работать со стеком, через некоторое время достигает минимального уровня стека и попадает в тупиковое состояние. Отметим ещё, что после внесения изменений автомат будет распознавать в точности те же самые слова языка $\Sigma^*\$$, что и исходный автомат.

В заключение доказательства остаётся лишь поменять местами множества конечных и неконечных состояний (то есть объявить все неконечные состояния конечными и наоборот). С учётом того, что после всех замен автомат дочитывает до конца любое слово языка $\Sigma^*\$$, останавли-

ваясь с пустым стеком, распознаваемость этого слова равносильна конечности состояния, в котором автомат находится после прочтения слова³. Это наблюдение завершает доказательство. \square

Следствие 7 *Класс детерминированных контекстно-свободных языков является собственным подклассом класса контекстно-свободных языков.*

Можно также привести пример КСЯ, не являющегося детерминированным КСЯ, в явном виде. Таковым, например, будет язык $L = \{a^n b^m c^k : m \neq n \vee m \neq k\}$. То, что L есть КСЯ, очевидно: для него легко выписать грамматику в явном виде. Однако $\Sigma^* \setminus L$ не является КСЯ, как было доказано ранее.

Замечание 1 Можно показать, что класс детерминированных контекстно-свободных языков не замкнут относительно пересечений, объединений, конкатенаций и навешивания звёздочки.

Замечание 2 Если не ограничиваться магазинной памятью, а брать устройства более общего типа (например, машину Тьюринга), то все контекстно-свободные языки вычислимы (распознаваемы на этом общем устройстве). Машина Тьюринга распознаёт слова КСЯ за время $O(n^3)$, где n — длина слова. Для детерминированных КСЯ оценку можно улучшить.

³На самом деле это не совсем верно, то есть вывод, по большому счёту, ошибочен. Впрочем, возникающая трудность легко устранима. Желающие могут самостоятельно найти ошибку и пути её исправления (заработав тем самым лишний балл на экзамене).