1 Регулярные языки и ДКА

Опр. Пусть Σ – произвольное конечное непустое множество, которое будем называть *алфавитом*, а его элементы — *символами* этого алфавита. Множеством *слов* алфавита Σ назовем множество

$$\Sigma^* = \bigcup_{n \in \mathbb{N}} \Sigma^n$$

Для $n \geq 0$ слово $w = (x_1, \ldots, x_n) \in \Sigma^n$ будем обозначать $x_1 x_2 \ldots x_n$, а число $|w| = n$ будем называть *длиной* слова w. Будем считать, что $\Sigma^0 = \{e\}$, где e — обозначение для пустого слова, т.е. слова, в котором нет символов (в частности, $|e| = 0$).

Опр. Для слов $u = x_1 x_2 \ldots x_n \in \Sigma^n$ и $v = y_1 y_2 \ldots y_m \in \Sigma^m$, *конкатенацией* слов u и v называется слово $w = uv = x_1 x_2 \ldots x_n y_1 y_2 \ldots y_m$ ($|w| = n + m$).

Аналогично определяется конкатенация произвольного конечного числа слов $u_1, u_2, \ldots, u_k$ из Σ^*. При $k = 0$ считаем $u_1 u_2 \ldots u_k = e$.

Пример. На словах могут быть определены другие содержательные синтаксические операции. В качестве примера рассмотрим операцию *обращения*: для слова $w \in \Sigma^*$, слово $w^R \in \Sigma^*$ определяется индукцией по длине: $e^R = e$, $(ua)^R = au^R$, где $u \in \Sigma^*$, $a \in \Sigma$.

Опр. *Языком* алфавита Σ называется произвольное множество слов $L \subseteq \Sigma^*$

Опр. Пусть $L_1, L_2, L \subseteq \Sigma^*$ – некоторые языки. Определим над ними следующие операции:

1. $L_1 \cup L_2 = \{w \in \Sigma^* \mid w \in L_1 \text{ или } w \in L_2\}$
2. $L_1 \cap L_2 = \{w \in \Sigma^* \mid w \in L_1 \text{ и } w \in L_2\}$
3. $L_1 \setminus L_2 = \{w \in \Sigma^* \mid w \in L_1 \text{ и } w \notin L_2\}$
4. $\overline{L} = \Sigma^* \setminus L$
5. $L_1 \circ L_2 = L_1 L_2 = \{w \in \Sigma^* \mid w = uv,\ u \in L_1, v \in L_2\}$
6. $L^* = \{w \in \Sigma^* \mid w = u_1 u_2 \ldots u_n,\ n \geqslant 0,\ u_1, u_2, \ldots, u_n \in L\}$

Полезно также определить язык $L^+ = L \circ L^*$, состоящий из всевозможных "нетривиальных" конкатенаций слов из L. В случае, когда $e \notin L$, имеет место соотношение $L^+ = L^* \setminus \{e\}$. Например, для $L = \Sigma$ язык Σ^+ состоит из всех слов алфавита Σ, имеющих ненулевую длину (то есть отличных от e).

Опр. Пусть Σ – конечный алфавит, и пусть символы $\varnothing$, $\cup$, $\circ$, *, (,) не лежат в алфавите Σ. Определим по индукции множество *регулярных выражений* алфавита Σ:

1. $\varnothing$ – регулярное выражение алфавита Σ;
2. a – регулярное выражение алфавита Σ для любого $a \in \Sigma$;
3. если α и β – регулярные выражения алфавита Σ, то $(\alpha \cup \beta)$, $(\alpha \circ \beta)$, α^* также являются регулярными выражениями алфавита Σ;
4. других регулярных выражений алфавита Σ нет.

Опр. Пусть α – регулярное выражение алфавита Σ. Язык $L(\alpha) \subseteq \Sigma^*$ определяется по регулярному выражению α индукцией по сложности α:

1. $L(\varnothing) = \emptyset$;
2. $L(a) = \{a\},\ a \in \Sigma$;
3. $L((\alpha \cup \beta)) = L(\alpha) \cup L(\beta)$;
4. $L((\alpha \circ \beta)) = L(\alpha) \circ L(\beta)$;
5. $L(\alpha^*) = L(\alpha)^*$.

Опр. Язык $L \subseteq \Sigma^*$ называется *регулярным*, если существует регулярное выражение α алфавита Σ такое, что

$$L = L(\alpha).$$

Опр. *Детерминированным конечным автоматом (ДКА)* над алфавитом Σ называется упорядоченная пятерка

$$M = (Q, \Sigma, s, F, \delta), \quad \text{где}$$

Q – конечное множество *состояний*, $s \in Q$ – *начальное состояние*, $F \subseteq Q$ – множество *заключительных состояний*, а $\delta : Q \times \Sigma \to Q$ – *функция перехода*.

Опр. *Конфигурацией* ДКА $M = (Q, \Sigma, s, F, \delta)$ называется произвольная пара (q, w), где $q \in Q$, $w \in \Sigma^*$.

Опр. Отношение $\vdash_M$ перехода (за один шаг) определим на множестве конфигураций ДКА M следующим образом: $(q_1, u) \vdash_M (q_2, v)$, если $u = av$ для некоторого $a \in \Sigma$, $\delta(q_1, a) = q_2$

Опр. Отношение $\vdash_M^*$ перехода за несколько (включая ноль) шагов определяется как рефлексивное и транзитивное замыкание $\vdash_M$, т.е. $(q, u) \vdash_M^* (r, v)$, если существует $n \in \mathbb{N}$ и слова $u_1, u_2, \ldots, u_n \in \Sigma^*$, т.ч.

$$(q, u) = (q_1, u_1) \vdash_M (q_2, u_2) \vdash_M \ldots \vdash_M (q_n, u_n) = (r, v)$$

для некоторых $q_1, q_2, \ldots, q_n \in Q$. При $n = 0$ $(q, u) = (r, v)$.

Опр. Слово $w \in \Sigma^*$ *распознается* ДКА M, если $(s, w) \vdash_M^* (q, e)$ для некоторого заключительного состояния $q \in F$.

Для ДКА M определим язык $L(M) = \{w \in \Sigma^* \mid w \text{ распознается ДКА } M\}$.

Опр. Язык $L \subseteq \Sigma^*$ называется *автоматным*, если найдется ДКА M, т.ч.

$$L = L(M).$$

Опр. *Недетерминированным конечным автоматом* (НКА) над алфавитом Σ называется упорядоченная пятерка

$$M = (Q, \Sigma, s, F, \Delta), \quad \text{где}$$

Q – конечное множество *состояний*, $s \in Q$ – *начальное состояние*, $F \subseteq Q$ – множество *заключительных состояний*, а $\Delta \subseteq Q \times (\Sigma \cup \{e\}) \times Q$ – отношение перехода.

Множеством *конфигураций* НКА M называется множество $Q \times \Sigma^*$.

Опр. Отношение $\vdash_M$ перехода на множестве конфигураций НКА M определяется так: $(q_1, u) \vdash_M (q_2, v)$, если $u = av$ для некоторого $a \in \Sigma \cup \{e\}$ и $(q_1, a, q_2) \in \Delta$.

Рефлексивное и транзитивное замыкание $\vdash_M$ обозначается $\vdash_M^*$ и определяется для НКА так же, как и в случае ДКА. Точно так же, слово $w \in \Sigma^*$ *распознается* НКА M, если $(s, w) \vdash_M^* (q, e)$ для некоторого заключительного состояния $q \in F$.

Одним из основных результатов теории формальных языков является теорема о равенстве классов регулярных и автоматных языков. Перед доказательством теоремы рассмотрим некоторые вспомогательные утверждения.

Предложение 1. *Для языка $L \subseteq \Sigma^*$ следующие условия эквивалентны:*

1. *$L = L(M)$ для некоторого ДКА M;*
2. *$L = L(M)$ для некоторого НКА M.*

Доказательство. $(1 \Rightarrow 2)$

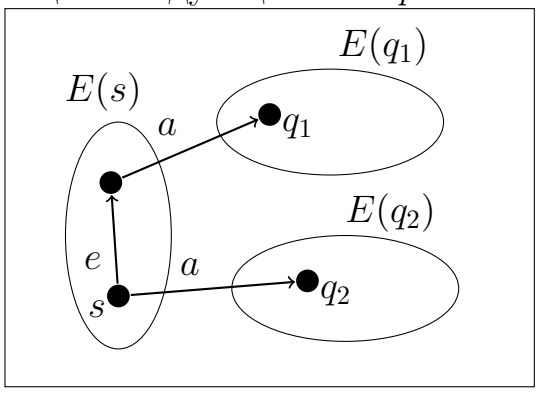
ДКА – частный случай НКА. Пусть $M = (Q, \Sigma, s, F, \delta)$ – ДКА, определим НКА $M' = (Q', \Sigma, s', F', \Delta)$, полагая $Q' = Q$, $s' = s$, $F' = F$. Отношение перехода зададим как график функции перехода

$$\Delta = \big\{(q, a, \delta(q, a)) \mid q \in Q\big\}$$

По построению, $L(M') = L(M)$. Действительно, легко доказать индукцией по длине слова w, что $w \in L(M')$ тогда и только тогда, когда $w \in L(M)$.

$(2 \Rightarrow 1)$

Пусть $M = (Q, \Sigma, s, F, \Delta)$ – НКА, определим ДКА $M' = (Q', \Sigma, s', F', \delta)$ с помощью следующего *алгоритма детерминизации.*



1. для каждого $q \in Q$ определим *e-орбиту* $E(q) \leftrightharpoons \left\{r \in Q \mid (q, e) \vdash_M^* (r, e)\right\}$ – множество состояний, в которые можно попасть из q с помощью скачков (e-переходов);
2. $Q' \leftrightharpoons P(Q) = \{X \mid X \subseteq Q\}$, $s' \leftrightharpoons E(s)$, $F' \leftrightharpoons \{X \subseteq Q \mid X \cap F \neq \varnothing\}$;
3. для всякого $X \subseteq Q$ и всякого $a \in \Sigma$, определим

$$\delta(X, a) \leftrightharpoons \bigcup \{E(r) \mid (q, a, r) \in \Delta \text{ для некоторого } q \in X\}.$$

По построению, $L(M') = L(M)$. Действительно, легко доказать индукцией по длине слова w, что $w \in L(M')$ тогда и только тогда, когда $w \in L(M)$. $\square$

Замечание. Не все состояния Q' являются достижимыми в ДКА M'. Такие состояния не влияют на язык, распознаваемый автоматом M'.

Теорема 1 (о совпадении классов регулярных и автоматных языков)**.** *Для языка $L \subseteq \Sigma^*$ следующие условия эквивалентны:*

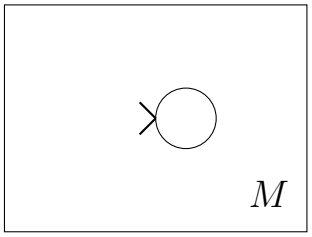
1. *$L = L(\alpha)$ для некоторого регулярного выражения α алфавита Σ;*
2. *$L = L(M)$ для некоторого ДКА M.*

Т.о., классы регулярных и автоматных языков совпадают.

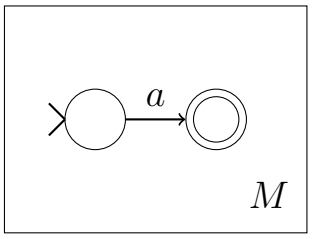
Доказательство. $(1 \Rightarrow 2)$

Пусть $L = L(\alpha)$ для некоторого регулярного выражения α. Учитывая предложение 1, достаточно показать, что существует НКА M, т.ч. $L(\alpha) = L(M)$. Используем индукцию по сложности выражения α.

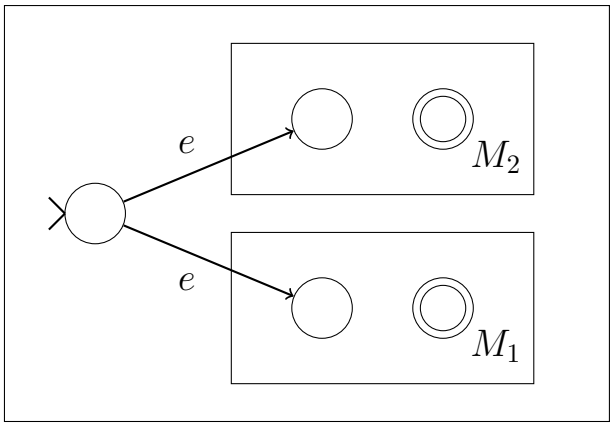
1. $\alpha = \varnothing$



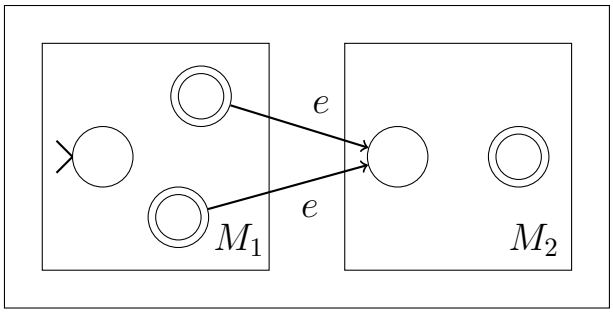
2. $\alpha = a$, $a \in \Sigma$



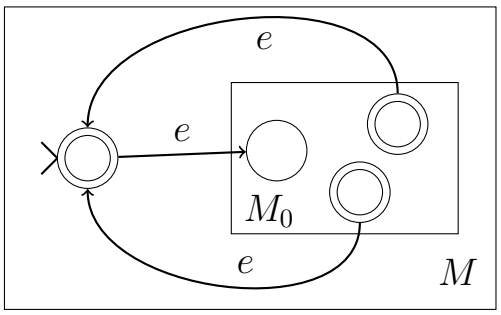
3. $\alpha = (\alpha_1 \cup \alpha_2)$, где $L(\alpha_1) = L(M_1)$, $L(\alpha_2) = L(M_2)$ (не нарушая общности, можно считать, что множества состояний автоматов M_1 и M_2 не пересекаются). Тогда $L(\alpha) = L(M)$ для автомата M, соединяющего автоматы M_1 и M_2 *параллельно*:



4. $\alpha = (\alpha_1 \circ \alpha_2)$, где $L(\alpha_1) = L(M_1)$, $L(\alpha_2) = L(M_2)$ (не нарушая общности, можно считать, что множества состояний автоматов M_1 и M_2 не пересекаются). Тогда $L(\alpha) = L(M)$ для автомата M, соединяющего автоматы M_1 и M_2 *последовательно*:



5. $\alpha = \alpha_0^*$, где $L(\alpha_0) = L(M_0)$. Тогда $L(\alpha) = L(M)$, где автомат M строится по автомату M_0 следующими образом:



$(2 \Rightarrow 1)$

Пусть M – КА (не обязательно ДКА) и $Q = \{q_1, q_2, \ldots, q_n\}$ $(n \geq 1)$, $s = q_1$, $F = \{q_{i_1}, \ldots, q_{i_k}\} \subseteq Q$. Для каждой пары $i, j \in \{1, \ldots, n\}$ и $m \leq n$ определим регулярное выражение $R(i, j, m)$ индукцией по m. Регулярные выражения $R(i, j, m)$ соответствуют маршрутам из q_i в q_j с использованием состояний $q_1, \ldots, q_m$ в качестве промежуточных.

1. $R(i, j, 0)$ определяется по отношению перехода Δ автомата M. Если из q_i в q_j нет переходов, то $R(i, j, 0) \leftrightharpoons \varnothing$. Если существуют переходы (q_i, a_1, q_j), $\ldots$, $(q_i, a_s, q_j) \in \Delta$, $a_1, \ldots, a_s \in \Sigma \cup \{e\}$, то $R(i, j, 0) \leftrightharpoons (a'_1 \cup \ldots \cup a'_s)$, где $a' \leftrightharpoons a$ для всех $a \in \Sigma$ и $e' \leftrightharpoons \varnothing^*$.

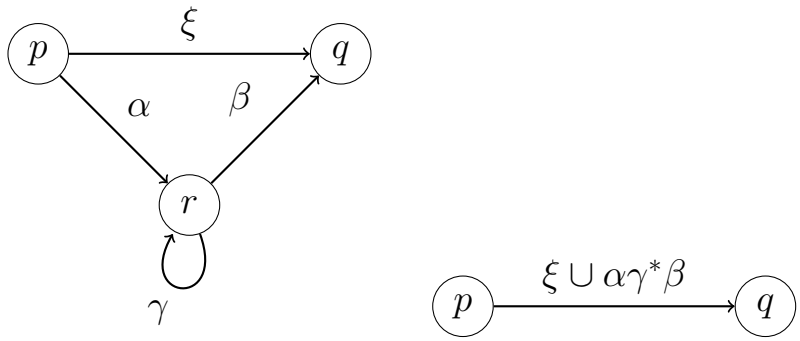
2. Для $m > 0$ имеем
$$R(i, j, m) \leftrightharpoons R(i, j, m-1) \cup \\ \cup R(i, m, m-1) R(m, m, m-1)^* R(m, j, m-1).$$

Т.о., для регулярного выражения $\alpha_M = \cup\{R(1, i, n) \mid q_i \in F\}$ получаем $L(\alpha_M) = L(M)$. □

Практическая реализация алгоритма построения по автомату соответствующего ему регулярного выражения использует идею этого доказательства. Пусть $L = L(M)$ для некоторого КА (согласно предложению 1 достаточно рассмотреть случай, когда M – НКА). Не нарушая общности, считаем, что в M только одно заключительное состояние, причём в начальное состояние нет входящих дуг, а из конечного состояния нет исходящих дуг. Алгоритм построения по КА M регулярного выражения α, т.ч. $L(\alpha) = L(M)$ заключается в сборке "метаавтомата" переходы в котором маркируются регулярными выражениями для каждой пары p и q оставшихся состояний.

Последовательно выполняя процедуру удаления для всех состояний, кроме начального и конечного, получим "метаавтомат" с единственной дугой содержащей регулярное выражение α, для которого $L(M) = L(\alpha)$.

Пример. Рассмотрим схему работы алгоритма при удалении одного состояния:



Теорема 2. *Класс регулярных языков замкнут относительно объединения, пересечения, разности, дополнения, конкатенации и итерации ("звездочки" Клини).*

Доказательство. Замкнутость относительно объединения, конкатенации и "звездочки" Клини очевидно следует из определения регулярности. Рассмотрим остальные операции.

- Пересечение: $L_1 \cap L_2 = \overline{\overline{L}_1 \cup \overline{L}_2}$
- Разность: $L_1 \backslash L_2 = L_1 \cap \overline{L}_2$
- Дополнение: используем автоматность языка L. Пусть $L = L(M)$ для ДКА $M = (Q, \Sigma, s, F, \delta)$, тогда $\overline{L} = L(\overline{M})$, где $\overline{M} = (Q, \Sigma, s, Q \backslash F, \delta)$.

□

Пусть Σ – некоторый конечный непустой алфавит и пусть $L \subseteq \Sigma^*$ – произвольный язык алфавита Σ.

Опр. Бинарное отношение $\approx_L$ задаётся на множестве Σ^* следующим образом:

$$\forall x, y \in \Sigma^* \quad x \approx_L y \Leftrightarrow \forall z \in \Sigma^* \quad xz \in L \Leftrightarrow yz \in L$$

Заметим, что $\approx_L$ является отношением эквивалентности на Σ^*. Т.е. можно определить систему классов эквивалентности (фактор-множество) $\Sigma^*/_{\approx_L}$ следующим образом

$$\forall x \in \Sigma^* \quad [x]_{\approx_L} = \{y \in \Sigma^* \mid x \approx_L y\}$$

Далее покажем, что фактор-множество конечно т. и т. т., когда L – регулярный язык.

Пусть $L \subseteq \Sigma^*$ – некоторый регулярный язык и пусть $M = (Q, \Sigma, s, F, \delta)$ – ДКА, такой что $L = L(M)$. На множестве слов определим бинарное отношение $\sim_M$ следующим образом:

$$\forall x, y \in \Sigma^* \quad x \sim_M y \Leftrightarrow (s, x) \vdash_M^* (q, e) \wedge (s, y) \vdash_M^* (q, e)$$

Т.к. $\sim_M$ является отношением эквивалентности на Σ^*, то можно говорить о классах эквивалентности $[x]_{\sim_M}$ и фактор-множестве $\Sigma^*/_{\sim_M}$.

Предложение 1. *Если $L \subseteq \Sigma^*$ – регулярный язык и M – ДКА, такой что $L = L(M)$, то*

$$\forall x, y \in \Sigma^* \quad x \sim_M y \Rightarrow x \approx_L y$$

Доказательство. Следует из определения этих эквивалентностей. □

Таким образом, отношение $\sim_M$ является утончением отношения $\approx_L$ в случае, когда $L = L(M)$, т.е.

$$\forall x \in \Sigma^* \quad [x]_{\sim_M} \subseteq [x]_{\approx_L}$$

Так как количество элементов в фактор-множестве $\Sigma^*/_{\sim_M}$ не превосходит количества элементов в множестве Q (конечного по определению ДКА) и так как $|\Sigma^*/_{\approx_L}| \leq |\Sigma^*/_{\sim_M}|$, то $\Sigma^*/_{\approx_L}$ конечно.

Теорема 3. *Пусть $L \subseteq \Sigma^*$ – язык, для которого $\Sigma^*/_{\approx_L}$ – конечное множество. Существует такой ДКА $M = (Q, \Sigma, s, F, \delta)$, что $L = L(M)$ и $|Q| = |\Sigma^*/_{\approx_L}|$.*

Доказательство. Построим для L *канонический* ДКА $M_L = (Q_L, \Sigma, s_L, F_L, \delta_L)$. Полагаем

$$Q_L \leftrightharpoons \Sigma^*/_{\approx_L}, \quad s_L \leftrightharpoons [e]_{\approx_L}, \quad F_L \leftrightharpoons \{[x]_{\approx_L} \mid x \in L\},$$

а для произвольных $x \in \Sigma^*$ и $a \in \Sigma$ определим функцию перехода следующим образом:

$$\delta_L([x]_{\approx_L}, a) \leftrightharpoons [xa]_{\approx_L}$$

Необходимо убедиться в том, что

1. такое определение корректно (т.е. не зависит от выбора представителя в классе эквивалентности)
2. $L(M_L) = L$.

Доказательство пункта 1) следует из определения. Корректность функции перехода

$$\forall x, y \in \Sigma^* \quad \forall a \in \Sigma \quad x \approx_L y \Rightarrow xa \approx_L ya,$$

следует из определения отношения $\approx_L$ при $z = az'$ и произвольном z'.

Чтобы проверить 2), покажем, что

$$\forall x, y \in \Sigma^* \quad ([x]_{\approx_L}, y) \vdash^*_{M_L} ([xy]_{\approx_L}, e)$$

индукцией по длине y. В случае $y = e$ утверждение очевидно. Шаг индукции $y = ay'$, где $a \in \Sigma$, следует из отределения функции перехода

$$([x]_{\approx_L}, ay') \vdash_{M_L} ([xa]_{\approx_L}, y') \vdash^*_{M_L} ([xay']_{\approx_L}, e)$$

Таким образом, для любого слова $x \in \Sigma^*$, x распознаётся ДКА M_L т. и т. т., когда $([e]_{\approx_L}, x) \vdash^*_{M_L} ([x]_{\approx_L}, e)$ и $[x]_{\approx_L} \in F_L$, т.е. $x \in L$. □

Теорема 4 (Майхилл-Нероуд). *Язык $L \subseteq \Sigma^*$ является регулярным т. и т. т., когда $\Sigma^*/_{\approx_L}$ – конечное множество.*

Доказательство. Если $\Sigma^*/_{\approx_L}$ конечно, то существует канонический ДКА M_L, построенный в доказательстве предыдущей теоремы и распознающий язык L, т.е. язык $L \subseteq \Sigma^*$ регулярен.

Если L – регулярный язык, то он распознаётся некоторым ДКА M, а значит $|\Sigma^*/_{\approx_L}| \leq |\Sigma^*/_{\sim_M}| \leq |Q_M|$, откуда $\Sigma^*/_{\approx_L}$ конечно.

□

Непосредственно из доказательства теоремы Майхилла-Нероуда следует, что канонический ДКА M_L для регулярного языка L имеет наименьшее возможное число состояний среди ДКА, распознающих язык L. Однако построение канонического ДКА для L нельзя назвать конструктивным ввиду сложности определения отношения $\approx_L$. Опишем алгоритм, который по произвольному ДКА M находит ДКА M', распознающий тот же самый язык и имеющий наименьшее возможное число состояний (*алгоритм минимизации*).

Пусть $L \subseteq \Sigma^*$ – регулярный язык и $M = (Q, \Sigma, s, F, \delta)$ – ДКА, для которого $L = L(M)$.

Опр. Конфигурация $(q, w) \in Q \times \Sigma^*$ ДКА M называется "хорошей" если $(q, w) \vdash_M^* (q', e)$ для некоторого $q' \in F$.

Опр. Для состояний $p, q \in Q$

$p \equiv q \overset{def}{\Longleftrightarrow} \forall z \in \Sigma^*$ (p, z) – "хорошая" конфигурация $\Leftrightarrow$

(q, z) – "хорошая" конфигурация

Нетрудно убедиться, что отношение $\equiv$ является отношением эквивалентности на Q.

Так как для ДКА M отношение $\sim_M$ является утончением $\approx_L$ на Σ^* и количество классов в $\Sigma^*/_{\sim_M}$ равно числу достижимых состояний в M, то отношение $\equiv$ на множестве $Q_{\text{дост.}} \subseteq Q$ всех достижимых состояний из Q позволяет определить ДКА $M' = (Q', \Sigma, s', F', \delta')$ следующим образом: полагаем $Q' \rightleftharpoons Q_{\text{дост.}}/_{\equiv}$, $s' \rightleftharpoons [s]_{\equiv}$, $F' \rightleftharpoons \{[q]_{\equiv} \mid q \in F \cap Q_{\text{дост}}\}$, $\delta'([q]_{\equiv}, a) \rightleftharpoons [\delta(q, a)]_{\equiv}$ для $q \in Q_{\text{дост.}}$, $a \in \Sigma$. По построению, $L(M') = L(M)$ и M' имеет наименьшее возможное число состояний.

Упражнение. Доказать корректность.

Разберем практическую реализацию описанного алгоритма. Для произвольного $n \geq 0$, рассмотрим отношение $\equiv_n$ на множестве $Q_{\text{дост.}}$ достижимых состояний M, определенное следующим образом:

$p \equiv_n q \overset{def}{\Longleftrightarrow} \forall z \in \Sigma^*,\ |z| \leq n$

(p, z) – "хорошая" конфигурация $\Leftrightarrow$ (q, z) – "хорошая" конфигурация

Можно заметить, что $\forall m, n \in \mathbb{N}$, т.ч. $m \geqslant n$, $p \equiv_m q \Rightarrow p \equiv_n q$, т.е. $\equiv_m$ – утончение $\equiv_n$. Т.о., $p \equiv_0 q \Leftrightarrow (p, q \in F \vee p, q \in Q \setminus F)$ — базисный шаг индуктивного определения отношения $\equiv_n$, а индукционный переход определяется следующим образом:

$$\forall p, q \in Q, \forall n \geq 1 \quad p \equiv_n q \Leftrightarrow \begin{cases} p \equiv_{n-1} q \\ \forall a \in \Sigma \ \delta(p, a) \equiv_{n-1} \delta(p, a) \end{cases}$$

Упражнение. Проверить эквивалентность.

Т.о., можно построить последовательность отношений эквивалентности $\equiv_0$, . . . , $\equiv_{n-1}, \equiv_n$, где каждое следующее является утончением предыдущего. Вследствие конечности Q найдется $n \in \mathbb{N}$, т.ч. $\equiv_n = \equiv_{n+1} = \ldots$ (наступает стабилизация), а значит $\equiv = \equiv_n$.

Следующее утверждение дает одно полезное необходимое условие регулярности языка.

Теорема 5 (о накачке)**.** *Пусть $L \subseteq \Sigma^*$ – регулярный язык. Тогда существует $n_0 \geq 1$, т.ч. $\forall w \in L$ $\big(|w| \geq n_0 \Rightarrow \exists x, y, z \in \Sigma^*$, т.ч. $w = xyz$, $|xy| \leq n_0$, $y \neq e$ и $xy^k z \in L \ \forall k \geq 0\big)$.*

Доказательство. Следует из автоматности языка L. Пусть $M = (Q, \Sigma, s, F, \delta)$ – ДКА, т.ч. $L = L(M)$ и пусть $n_0 = |Q|$. Рассмотрим произвольное слово $w \in L$, $|w| \geq n_0$, т.е. $w = a_1 a_2 \ldots a_{n_0} a_{n_0+1} \ldots a_m$ $m \geq n_0$. Для $q_1 = s$ имеем

$$(q_1, a_1 a_2 \ldots a_m) \vdash_M (q_2, a_2 \ldots a_m) \vdash_M \ldots$$
$$\vdash_M (q_{n_0}, a_{n_0} \ldots a_m) \vdash_M (q_{n_0+1}, a_{n_0+1} \ldots a_m) \vdash_M (q_{m+1}, e), \quad q_{m+1} \in F$$

Т.к. $|Q| = n_0$, то $\exists i, j \in 1, \ldots, n_0 + 1$, т.ч. $i < j$ и $q_i = q_j$. Полагаем $x \leftrightharpoons a_1 \ldots a_{i-1}$, $y \leftrightharpoons a_i \ldots a_{j-1}$, $z \leftrightharpoons a_j \ldots a_m$. Легко убедиться, что $xy^k z \in L \ \forall k > 0$. □

Следствие. *Язык $L = \{a^n b^n \mid n \geq 0\}$ не является автоматным.*

Упражнение. Доказать следствие.

2 Контекстно-свободные языки

Опр. *Контекстно-свободной грамматикой* (КС-грамматикой) называется упорядоченная четверка $G = (V, \Sigma, S, R)$, где V – конечное множество символов, $\Sigma \subset V$ – множество *терминальных символов* (символы из $V \backslash \Sigma$ называются *нетерминальными*), $S \in V \backslash \Sigma$ – *начальный* нетерминальный символ, $R \subseteq (V \backslash \Sigma) \times V^*$ – конечное множество *правил*, называемых также *продукциями*.

Замечание. Для $(A, v) \in R$ используем обозначение $A \to_G v$, т.е. нетерминал A в грамматике G порождает слово v.

Опр. Для КС-грамматики $G = (V, \Sigma, S, R)$ и слов $u, v \in V^*$, отношение $u \Rightarrow_G v$ означает, что $\exists x, y, z \in V^*$ и $A \in V \backslash \Sigma$, т.ч. $u = xAy$, $v = xzy$ и $(A \to_G z) \in R$.

Опр. Отношение $\Rightarrow_G^*$ есть рефлексивное и транзитивное замыкание отношения $\Rightarrow_G$, т.е. $\forall u, v \in V^*$ $u \Rightarrow_G^* v \overset{def}{\Longleftrightarrow}$ либо $u = v$, либо $\exists n \geq 0$ $w_1, \ldots, w_n \in V^*$, т.ч. $u \Rightarrow_G w_1 \Rightarrow_G \ldots \Rightarrow_G w_n \Rightarrow v$

Опр. Для КС-грамматики $G = (V, \Sigma, S, R)$, язык $L(G) = \{w \in \Sigma^* \mid S \Rightarrow_G^* w\}$ называется языком, порожденным грамматикой G.

Опр. Язык $L \subset \Sigma^*$ называется *контекстно-свободным*, если $L = L(G)$ для некоторой КС-грамматики G.

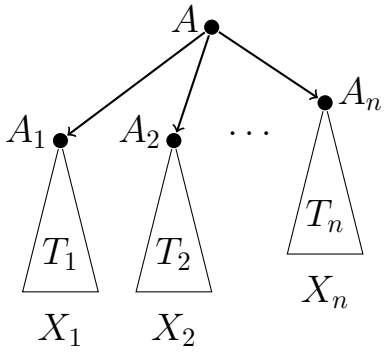
Пример. Рассмотрим КС-грамматику $G = (V, \Sigma, S, R)$, где $V = \{S, a, b\}$, $\Sigma = \{a, b\}$, $R = \{S \to aSb, S \to e\}$. Тогда $L(G) = \{a^n b^n \mid n \geq 0\}$.

Теорема 6. *Всякий регулярный язык является контекстно-свободным.*

Доказательство. Так как из регулярности следует автоматность, соответствующая КС-грамматика строится по детерминированному конечному автомату $M = (Q, \Sigma, s, F, \delta)$, распознающему данный язык: в качестве множества нетерминальных символов возьмем Q, в качестве стартового нетерминала s, а множество правил состоит из продукций вида $q \to ap$, где $\delta(q, a) = p$, а также продукций вида $q \to e$, где $q \in F$. □

Пусть $G = (V, \Sigma, S, R)$ – КС-грамматика. Множество *деревьев разбора* для G, а также понятия корня, листьев, результата и высоты дерева определяются индуктивно (корень и листья понимаются как вершины дерева, а результат как слово из Σ^*):

1. $\bullet\ a$, $a \in \Sigma$ (корень, лист и результат такого дерева равен a, высота равна 0);
2. $A\bullet \to \bullet e$, где $(A \to e) \in R$ ($A \in V$ – корень, e – лист и результат, высота такого рана 1);
3. пусть $T_1, \ldots, T_n$ ($n \geqslant 1$) – деревья разбора для G с корнями $A_1, \ldots, A_n$, множествами листьев $Y_1, \ldots, Y_n$, результатами $X_1, \ldots, X_n \in \Sigma^*$ и пусть $A \to A_1 \ldots A_n$ – продукция из R, тогда



– дерево разбора для G, в котором A – корень, $Y_1 \cup Y_2 \cup \ldots \cup Y_n$ – листья, $X_1 \cdot X_2 \cdot \ldots \cdot X_n$ – результат, а высота равна максимуму из высот деревьев $T_1, \ldots, T_n$, увеличенному на 1;

4. других деревьев разбора для G нет.

Таким образом, всякое дерево разбора есть ориентированный упорядоченный размеченный граф, являющийся деревом.

Теорема 7. Пусть $G = (V, \Sigma, R, S)$ – КС-грамматика, тогда для любого нетерминала $A \in V \backslash \Sigma$ и любого слова $w \in \Sigma^*$ следующие условия эквивалентны

1. $A \Rightarrow_G^* w$ (слово w порождается в грамматике G из нетерминала A);

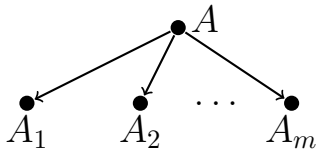
2. существует дерево разбора для G с корнем A и результатом w.

Доказательство. $(1 \Leftarrow 2)$

Пусть существует дерево разбора. Проведем рассуждение индукцией по его высоте h. Для $h = 1$ возможны следующие случаи:

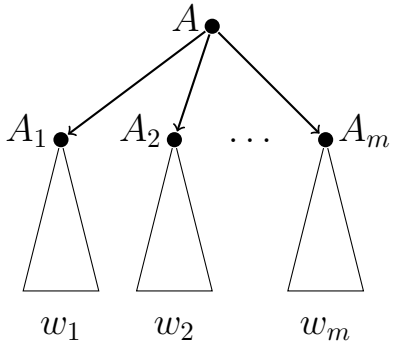
а) $A\bullet \to \bullet e \Rightarrow (A \to e) \in R \Rightarrow A \Rightarrow_G^* e$;

б)



т.е. $A_1, A_2, \ldots, A_m \in \Sigma$, $(A \to A_1 A_2 \ldots A_m) \in R \Rightarrow A \Rightarrow_G^* A_1 A_2 \ldots A_m$

Индукционный переход от случая $h \leq n$ к $h = n + 1$:



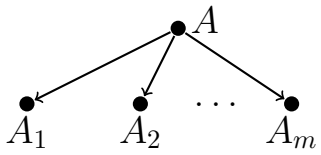
среди $A_1, A_2, \ldots, A_m$ есть нетерминалы $A_{i_1}, \ldots, A_{i_k}$ с соответствующими порождениями $A_{i_1} \Rightarrow_G^* w_{i_1}$, …, $A_{i_k} \Rightarrow_G^* w_{i_k}$. Для слова $w = w_1 w_2 \ldots w_n$ построим искомое порождение $A \Rightarrow_G^* w$:

$$A \Rightarrow_G A_1 \ldots A_n \Rightarrow_G A_1 \ldots A_{i_1 - 1} w_{i_1} A_{i_1 + 1} \ldots A_n \Rightarrow_G \ldots \Rightarrow_G w$$

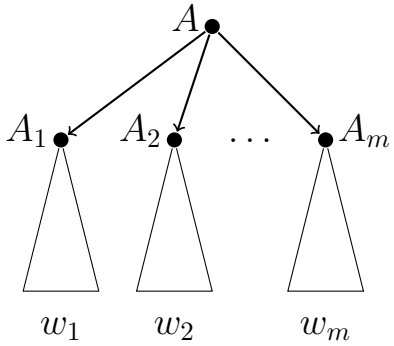
$(1 \Rightarrow 2)$

Пусть $A \Rightarrow_G^* w$, т.е. $A \Rightarrow_G u_1 \Rightarrow_G u_2 \Rightarrow \ldots \Rightarrow_G u_n \Rightarrow_G w$ для некоторого $n \geq 0$ и $u_1, \ldots, u_n \in V^*$. Построим дерево разбора индукцией по n.

Для $n = 0$ и $A \Rightarrow_G w$ либо $w = e$, тогда $A\bullet \longrightarrow \bullet e$ – дерево, либо $w = A_1 \ldots A_m \in \Sigma^*$ $m \geq 1$ и $(A \to A_1 \ldots A_m) \in R$ и дерево разбора имеет вид



Индукционный переход $n \to n+1$. Пусть $A \Rightarrow_G u_1 \Rightarrow_G u_2 \Rightarrow_G \ldots \Rightarrow_G u_n \Rightarrow_G u_{n+1} \Rightarrow_G w$. Рассмотрим слово $u_1 \in V^*$: $u_1 = A_1 \ldots A_m$, где $A_{i_1}, \ldots, A_{i_k}$ – нетерминалы, участвующие в порождении w. Причем для каждого нетерминала A_{i_j} длина вывода $A_{i_j} \Rightarrow_G \ldots \Rightarrow_G w_{i_j}$ не превосходит n, т.е. имеется соответствующее дерево разбора. Объединив их, получим полное дерево разбора



□

Опр. Пусть $G = (V, \Sigma, S, R)$ – КС-грамматика. Для $u, v \in V^*$, $u \Rightarrow_{L,G} v$ означает, что $u = xAy$, $v = xzy$ для некоторых $x \in \Sigma^*$, $y, z \in V^*$ и $A \in V \setminus \Sigma$, т.ч. $(A \to z) \in R$. Аналогично, $u \Rightarrow_{R,G} v$ означает, что $u = xAy$, $v = xzy$ для некоторых $x, z \in V^*$, $y \in \Sigma^*$ и $A \in V \setminus \Sigma$, т.ч. $(A \to z) \in R$.

Как обычно, отношения $\Rightarrow^*_{L,G}$ и $\Rightarrow^*_{R,G}$ определяются как рефлексивные и транзитивные замыкания отношений $\Rightarrow_{L,G}$ и $\Rightarrow_{R,G}$, соответственно. Непосредственно из определений и свойств КС-грамматик вытекает

Предложение 2. Пусть $G = (V, \Sigma, S, R)$ – КС-грамматика. Для любых $u \in V^*$ и $v \in \Sigma^*$, следующие условия эквивалентны:

1) $u \Rightarrow^*_G v$;

2) $u \Rightarrow^*_{L,G} v$;

3) $u \Rightarrow^*_{R,G} v$.

Из этого предложения и установленной ранее теоремы получаем

Следствие. Для КС-грамматики $G = (V, \Sigma, S, R)$ и $w \in \Sigma^*$, следующие условия эквивалентны:

1. $w \in L(G)$;

2. $S \Rightarrow^*_G w$;

3. $S \Rightarrow^*_{L,G} w$;

4. $S \Rightarrow^*_{R,G} w$;

5. существует дерево разбора для G с корнем S и результатом w.

С помощью деревьев разбора устанавливается теорема о накачке для контекстно-свободных языков, дающая полезное необходимое условие принадлежности этому классу и позволяющая доказывать, что конкретный язык не является контекстно-свободным.

Для контекстно-свободной грамматики $G = (V, \Sigma, S, R)$, назовем ее *шириной* натуральное число

$$\varphi(G) = \max\{n \mid (A \to A_1 \dots A_n) \in R\}.$$

Теорема 8 (о накачке)**.** Пусть $L \subseteq \Sigma^*$ – язык, порождаемый контекстно-свободной грамматикой $G = (V, \Sigma, S, R)$. Тогда для любого $w \in L$ такого, что $|w| > \varphi(G)^{|V \setminus \Sigma|}$, существуют $u, v, x, y, z \in \Sigma^*$ такие, что $w = uvxyz$, $vy \neq e$ и, для любого $k \geqslant 0$, $uv^k xy^k z \in L$.

Доказательство. В качестве вспомогательного утверждения потребуется

Лемма. Если T – дерево разбора для КС-грамматики G, имеющее высоту h и результат w, то $|w| \leqslant \varphi(G)^h$.

Доказательство. Индукция по h. □

Следствие. Если $w \in L(G)$ и $|w| > \varphi(G)^h$, то в любом дереве разбора для G с результатом w, существует путь из корня в один из листьев, длина которого $> h$.

Итак, пусть слово $w \in L(G)$ таково, что $|w| > \varphi(G)^{|V \setminus \Sigma|}$. Выберем дерево разбора T для G, имеющее результат w и наименьшее число листьев среди всех таких деревьев (считаем также, что в G нет правил вида $A \to A$). По следствию из леммы, в дереве T существует путь (из корня в один из листьев) длины большей, чем $|V \setminus \Sigma|$, а значит, в этом пути как минимум $|V \setminus \Sigma| + 2$ вершин (узлов). По определению деревьев разбора, все эти вершины, кроме последней, являются нетерминалами. Таким образом, данный путь содержит как минимум $|V \setminus \Sigma| + 1$ нетерминальных символов. Следовательно, существует символ $A \in V \setminus \Sigma$, встречающийся на этом пути (как минимум) дважды. Пусть поддерево дерева T, порождаемое нижним вхождением A, имеет результат x, а поддерево дерева T, порождаемое верхним вхождением A, имеет результат vxy. Слова $u, z \in \Sigma^*$ определяются из условия $w = uvxyz$. По выбору дерева T, $vy \neq e$.

Из всего этого следует, что $uv^k xy^k z \in L(G)$ для всех $k \geqslant 0$.

□

Пользуясь установленной выше теоремой о накачке, легко получаем

Пример. Язык $\{a^n b^n c^n \mid n \geqslant 0\}$ не является контекстно-свободным.

Следующая теорема содержит положительные результаты о свойствах замкнутости класса контекстно-свободных языков.

Теорема 9. Класс контекстно-свободных языков замкнут относительно операций объединения, конкатенации и итерации ("звездочки" Клини).

Доказательство. Пусть $L_1 \subseteq \Sigma_1^*$ и $L_2 \subseteq \Sigma_2^*$ – произвольные КС-языки, то есть $L_1 = L(G_1)$ и $L_2 = L(G_2)$ для некоторых КС-грамматик $G_1 = (V_1, \Sigma_1, S_1, R_1)$ и $G_2 = (V_2, \Sigma_2, S_2, R_2)$. Можно считать, что $(V_1 \setminus \Sigma_1) \cap (V_2 \setminus \Sigma_2) = \varnothing$. Выберем новый символ $S \notin V_1 \cup V2$. По грамматикам G_1 и G_2 непосредственно определяются грамматики, порождающие, соответственно, языки $L_1 \cup L_2$, $L_1 \circ L_2$ и $(L_1)^*$:

а) $L_1 \cup L_2 = L(G_\cup)$ для КС-грамматики $G_\cup = (V_1 \cup V_2 \cup \{S\}, \Sigma_1 \cup \Sigma_2, S, R_\cup)$, где $R_\cup = R_1 \cup R_2 \cup \{S \to S_1, S \to S_2\}$;

б) $L_1 \circ L_2 = L(G_\circ)$ для КС-грамматики $G_\circ = (V_1 \cup V_2 \cup \{S\}, \Sigma_1 \cup \Sigma_2, S, R_\circ)$, где $R_\circ = R_1 \cup R_2 \cup \{S \to S_1 S_2\}$;

в) $(L_1)^* = L(G_*)$ для КС-грамматики $G_* = (V_1 \cup \{S\}, \Sigma_1, S, R_*)$, где $R_* = R_1 \cup \{S \to e,\ S \to S S_1\}$.

□

В отличие от класса регулярных языков, класс контекстно-свободных языков не замкнут относительно операций пересечения и дополнения. А именно, легко убедиться, что языки $\{a^n b^n c^m \,|\, m, n \geqslant 0\}$ и $\{a^m b^n c^n \,|\, m, n \geqslant 0\}$ являются контекстно-свободными. Пересечением этих двух языков является язык $\{a^n b^n c^n \,|\, n \geqslant 0\}$, который не является контекстно-свободным. Как следствие, поскольку операция пересечения выражается через операции объединения и дополнения по формуле $L_1 \cap L_2 = \overline{\overline{L_1} \cup \overline{L_2}}$, класс контекстно-свободных языков не замкнут относительно операции дополнения.

Опр. *Автоматом с магазинной памятью* (pushdown automaton) называется упорядоченная шестерка

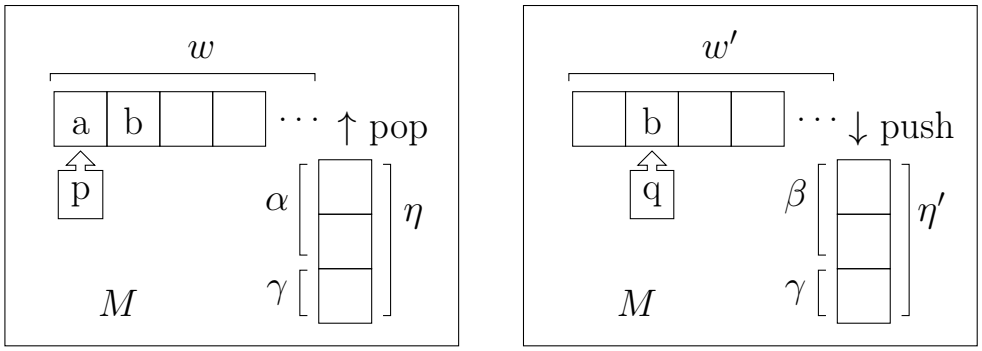
$$M = (Q, \Sigma, \Gamma, s, F, \Delta), \quad \text{где}$$

Q – конечное множество *состояний*, Σ – *внешний алфавит*, Γ – *внутренний алфавит*, $s \in Q$ – *начальное состояние*, $F \subseteq Q$ – множество *заключительных состояний*, а $\Delta \subseteq \big(Q \times (\Sigma \cup \{e\}) \times \Gamma^*\big) \times (Q \times \Gamma^*)$ – конечное *отношение перехода*.

Опр. *Конфигурацией* автомата с магазинной памятью M называется упорядоченная тройка $C = (q, w, \eta)$, где $q \in Q$ – текущее состояние, $w \in \Sigma^*$ – текущее входное слово, $\eta \in \Gamma^*$ – текущее состояние магазина (стека).

Опр. Для конфигураций C и C' автомата M будем говорить, что конфигурация C' получена из C за один шаг работы автомата M (обозн. $C \vdash_M C'$), если $C = (p, w, \eta)$, $C' = (q, w', \eta')$, и справедливо следующее: $w = aw'$, $\eta = \alpha\gamma$, $\eta' = \beta\gamma$ для некоторой пятерки $\big((p, a, \alpha), (q, \beta)\big) \in \Delta$ (здесь $\gamma \in \Gamma^*$ – неиспользуемое содержимое стека).

Как обычно, $\vdash_M^*$ есть рефлексивное и транзитивное замыкание отношения $\vdash_M$.



Опр. Слово $w \in \Sigma^*$ *распознается* автоматом с магазинной памятью M, если $(s, w, e) \vdash_M^* (q, e, e)$ для некоторого заключительного состояния $q \in F$.

Для автомата M, как обычно, определим язык

$$L(M) = \{w \in \Sigma^* \mid w \text{ распознаётся автоматом } M\}.$$

Будем говорить, что язык $L \subseteq \Sigma^*$ *распознается автоматом с магазинной памятью*, если $L = L(M)$ для некоторого автомата с магазинной памятью M.

Справедлива следующая

Теорема 10. Пусть $L \subseteq \Sigma^*$ – произвольный язык. Следующие условия эквивалентны:

1) L – контекстно-свободный язык;

2) L распознается некоторым автоматом с магазинной памятью.

Перед тем, как приступить к (довольно длинному) доказательству этой важной теоремы, установим одно из ее следствий.

Предложение 3. Пересечение контекстно-свободного языка с регулярным языком является контекстно-свободным языком.

Доказательство. Пусть для $L_1, L_2 \subseteq \Sigma^*$ справедливы равенства $L_1 = L(M_1)$, $L_2 = L(M_2)$, где $M_1 = \langle Q_1, \Sigma, \Gamma_1, s_1, F_1, \Delta\rangle$ и $M_2 = \langle Q_2, \Sigma, s_2, F_2, \delta\rangle$ – автомат с магазинной памятью и детерминированный конечный автомат, соответственно. Определим автомат с магазинной памятью $M = (Q, \Sigma, \Gamma, s, F, \Delta)$, для которого $L(M) = L_1 \cap L_2$. Полагаем $Q = Q_1 \times Q_2$, $\Gamma = \Gamma_1$, $s = (s_1, s_2)$, $F = F_1 \times F_2$. Наконец, определим отношение перехода Δ как объединение множеств

$$\{(((q_1, q_2), a, \alpha), ((q_1', \delta(q_2, a)), \beta)) \,|\, q_i, q_i' \in Q_i, a \in \Sigma, ((q_1, a, \alpha), (q_1', \beta)) \in \Delta_1\}$$

и

$$\{(((q_1, q_2), e, \alpha), ((q_1', q_2), \beta)) \mid q_1, q_1' \in Q_1, q_2 \in Q_2, ((q_1, e, \alpha), (q_1', \beta)) \in \Delta_1\}.$$

□

Итак, приступим к доказательству теоремы. Установим сначала импликацию 1) ⇒ 2). Пусть $L = L(G)$ для некоторой контекстно-свободной грамматики $G = (V, \Sigma, S, R)$, где $\Sigma = \{a_1, \ldots, a_m\}$, $R = \{A_1 \to v_1, \ldots, A_n \to v_n\}$ (здесь $A_i \in V \setminus \Sigma$, $v_i \in V^*$). Определим автомат с магазинной памятью

$$M_G = (\{s, q\}, \Sigma, V, s, \{q\}, \Delta),$$

в котором отношение перехода Δ состоит из команд

$((s, e, e), (q, S))$,
$((q, e, A_1), (q, v_1))$,
...
$((q, e, A_n), (q, v_n))$,
$((q, a_1, a_1), (q, e))$,
...
$((q, a_m, a_m), (q, e))$.

Покажем, что $L(M_G) = L(G)$. Справедлива следующая

Лемма. Для любого $w \in \Sigma^*$ и любого $\alpha \in (V \setminus \Sigma)V^* \cup \{e\}$, $S \Rightarrow_{L,G}^* w\alpha$ тогда и только тогда, когда $(q, w, S) \vdash_{M_G}^* (q, e, \alpha)$.

Из этой леммы, полагая $\alpha = e$, получаем требуемое равенство.

Доказательство. Пусть $S \Rightarrow_{L,G}^* w\alpha$, то есть $S \Rightarrow_{L,G} u_1 \Rightarrow_{L,G} \ldots \Rightarrow_{L,G} u_n = w\alpha$ для некоторого $n \geqslant 0$, $u_1, \ldots, u_n \in V^*$. Индукция по n: если $n = 0$, то $w = e$, $\alpha = S$, то есть $(q, e, S) \vdash_{M_G}^* (q, e, s)$. Пусть теперь $S \Rightarrow_{L,G} u_1 \Rightarrow_{L,G} \ldots \Rightarrow_{L,G} \Rightarrow_{L,G} u_n \Rightarrow_{L,G} u_{n+1} = w\alpha$ для некоторого $n \geqslant 0$ и $u_1, \ldots, u_{n+1} \in V^*$. Рассмотрим последний переход в этом порождении: пусть $u_n = xA\beta$, $u_{n+1} = x\gamma\beta$, где $x \in \Sigma^*$, $\beta \in V^*$ и $(A \to \gamma) \in R$. По индукционному предположению, $S \Rightarrow_{L,G}^* xA\beta$ влечет $(q, x, S) \vdash_{M_G}^* (q, e, A\beta)$. Так как $u_{n+1} = x\gamma\beta = w\alpha$, то $w = xy$, $\gamma\beta = y\alpha$. Отсюда $(q, xy, S) \vdash_{M_G}^* (q, y, A\beta) \vdash_{M_G} (q, y, \gamma\beta) = (q, y, y\alpha) \vdash_{M_G}^* (q, e, \alpha)$.

Наоборот, пусть $(q, w, S) \vdash_{M_G}^* (q, e, \alpha)$. Индукция по числу n переходов, обрабатывающих нетерминалы из стека. Если $n = 0$, то $w = e, \alpha = S$, то есть $S \Rightarrow_{L,G}^* S$. Пусть теперь $(q, w, S) \vdash_{M_G}^* (q, e, \alpha)$ с использованием $n + 1$ перехода, обрабатывающего нетерминалы из стека. Рассмотрим последний такой переход: пусть $(q, w, S) \vdash_{M_G}^* (q, y, A\beta) \vdash_{M_G} (q, y, \gamma\beta) \vdash_{M_G}^* (q, e, \alpha)$, где $w = xy$ для некоторых $x, y \in \Sigma^*$, и $(A \to \gamma) \in R$. По индукционному предположению, $S \Rightarrow_{L,G}^* xA\beta$, а значит и $S \Rightarrow_{L,G}^* x\gamma\beta$. Но, так как в последних

переходах $(q, w, S) \vdash^*_{M_G} (q, y, \gamma\beta) \vdash^*_{M_G} (q, e, \alpha)$ использовались только переходы, обрабатывающие терминалы из стека, получаем, что $y\alpha = \gamma\beta$, то есть $S \Rightarrow^*_{L,G} x\gamma\beta = xy\alpha = w\alpha$. □

Установим теперь импликацию 2) ⇒ 1). Пусть $L = L(M)$ для некоторого автомата с магазинной памятью $M = (Q, \Sigma, \Gamma, s, F, \Delta)$.

Опр. Автомат с магазинной памятью $M = (Q, \Sigma, \Gamma, s, F, \Delta)$ называется *простым*, если для любого перехода $((q, a, \beta), (p, \gamma)) \in \Delta$, т.ч. $q \neq s$, выполняются условия $\beta \in \Gamma$ и $|\gamma| \leqslant 2$.

Лемма. Для любого автомата с магазинной памятью M существует простой автомат с магазинной памятью M', для которого $L(M) = L(M')$.

Доказательство. Построим по автомату $M = (Q, \Sigma, \Gamma, s, F, \Delta)$ эквивалентный ему простой автомат $M = (Q', \Sigma, \Gamma', s', F', \Delta')$ следующим образом. Пусть s', f' и Z – новые символы. Полагаем $F' \leftrightharpoons \{f'\}$, $\Gamma' \leftrightharpoons \Gamma \cup \{Z\}$ и $Q' \leftrightharpoons Q \cup \{s', f'\}$ (в дальнейшем Q' может быть дополнено новыми состояниями). Помещаем в Δ', помимо всех элементов Δ, переходы $((s', e, e), (s, Z))$ и $((f, e, Z), (f', e))$ для всех $f \in F$. Далее, заменяем в Δ' все переходы, нарушающие условия простоты, по следующей схеме:

а) все переходы вида $((q, a, \beta), (p, \gamma))$, т.ч. $\beta = B_1 B_2 \dots B_n$, $n \geqslant 2$, заменяем на последовательности переходов

$((q, e, B_1), (q_{B_1}, e))$,
$((q_{B_1}, e, B_2), (q_{B_1 B_2}, e))$,
$\dots$
$((q_{B_1 B_2 \dots B_{n-1}}, a, B_n), (p, \gamma))$

(здесь q_{B_1}, …, $q_{B_1 B_2 \dots B_{n-1}}$ – новые состояния, которые добавляются в Q');

б) далее, все переходы вида $((q, a, \beta), (p, \gamma))$, т.ч. $\gamma = C_1 C_2 \dots C_m$, $m \geqslant 2$, заменяем на последовательности переходов

$((q, a, \beta), (r_1, C_m))$,
$((r_1, e, e), (r_2, C_{m-1}))$,
$\dots$
$((r_{m-1}, e, e), (p, C_1))$

(здесь r_1, …, r_{m-1} – новые состояния, которые добавляются в Q');

в) наконец, все переходы вида $((q,a,e),(p,\gamma))$, т.ч. $q \neq s'$, заменяем на множества переходов вида $((q,a,A),(p,\gamma A))$ для всех $A \in \Gamma'$.

Множество правил R контекстно-свободной грамматики, порождающей язык $L(M)$, определяется по соответствующему M простому автомату M' следующим образом. Помещаем в R правило $S \to \langle s, Z, f' \rangle$, а также правила вида $\langle q, e, q \rangle \to e$ для всех $q \in Q'$. Далее, сопоставляем каждому переходу типа $((q,a,B),(p,C))) \in \Delta'$ множество правил вида $\langle q, B, r \rangle \to a\langle p, C, r \rangle$, $r \in Q'$, а каждому переходу типа $((q,a,B),(p,C_1C_2))) \in \Delta'$ – множество правил вида $\langle q, B, r \rangle \to a\langle p, C_1, r' \rangle \langle r', C_2, r \rangle$, $r, r' \in Q'$.

Используя индуктивные рассуждения, несложно убедиться в том, что справедлива следующая

Лемма. Для любых $w \in \Sigma^*, p, q \in Q', A \in \Gamma'$,
$\langle q, A, p \rangle \Rightarrow^*_{G_M} w$ тогда и только тогда, когда $(q, w, A) \vdash^*_{M'} (p, e, e,)$.

Как следствие, $S \Rightarrow^*_{G_M} w$ т. и т.т., когда $S \to \langle s, Z, f' \rangle \Rightarrow^*_{G_M} w$ т. и т.т., когда (вследствие леммы) $(s, w, Z) \vdash^*_{M'} (f', e, e)$, то есть $w \in L(M') = L(M)$.

□

3 Нормальная форма Хомского

Опр. Контекстно-свободная грамматика $G = (V, \Sigma, S, R)$ находится в *нормальной форме Хомского*, если все правила в R имеют вид $A \to BC$ для некоторых $A, B, C \in V$.

Теорема 11. Для любой контекстно-свободной грамматики $G = (V, \Sigma, S, R)$ существует контекстно-свободная грамматика $G' = (V', \Sigma, S, R')$, находящаяся в нормальной форме Хомского, для которой $L(G') = L(G) \setminus (\Sigma \cup \{e\})$.

Доказательство. Помещаем в V' все символы из V, а в R' – все правила из R. Затем проводим следующие преобразования:

а) удаляем из R' *"длинные" правила* вида $A \to B_1 B_2 \dots B_n$, $n > 2$, и добавляем вместо каждого такого правила множество правил $A \to B_1 A_1$, $A_1 \to B_2 A_2$, …, $A_{n-2} \to B_{n-1} B_n$ (здесь $A_1, \dots, A_{n-2}$ – новые нетерминальные символы, которые добавляются в V');

б) для удаления из R' *"e-правил"* вида $A \to e$ поступаем следующим образом. Определим множество $\mathcal{E} = \{B \in V \mid B \Rightarrow^*_G e\}$ (оно определяется конструктивно как наименьшая неподвижная точка последовательности $\mathcal{E}_0 = \varnothing$, $\mathcal{E}_{n+1} = \mathcal{E}_n \cup \{B \in V \mid (B \to \beta) \in R$ для некоторого $\beta \in (\mathcal{E}_n)^*\}$ для $n \geqslant 0$). Теперь удаляем из R' все правила вида $A \to e$ и для каждого правила из R' вида $A \to BC$ или $A \to CB$, где $B \in \mathcal{E}$, добавляем правило $A \to C$;

в) для удаления из R' "1-*правил*" вида $A \to B$ поступаем следующим образом. Для каждого $A \in V$ (в том числе и для $A \in \Sigma$) определим множество $\mathcal{D}(A) = \{B \in V \mid A \Rightarrow_G^* B\}$ (оно определяется конструктивно как наименьшая неподвижная точка последовательности $\mathcal{D}_0(A) = \{A\}$, $\mathcal{D}_{n+1}(A) = \mathcal{D}_n(A) \cup \{C \in V \mid (B \to C) \in R$ для некоторого $B \in \mathcal{D}_n(A)\}$ для $n \geqslant 0$). Теперь удаляем из R' все правила вида $A \to B$ и для каждого правила из R' вида $A \to BC$, добавляем (конечное) множество правил вида $A \to B'C'$, где $B' \in \mathcal{D}(B) \setminus \{B\}$, $C' \in \mathcal{D}(C) \setminus \{C\}$. Корме того, добавляем в R' (конечное) множество правил вида $S \to BC$ для всех правил $A \to BC$ из R, таких, что $A \in \mathcal{D}(S) \setminus \{S\}$.

Индукцией по длине слова $w \in \Sigma^*$, $|w| \geqslant 2$, легко убедиться, что $w \in L(G')$ тогда и только тогда, когда $w \in L(G)$.

□

4 Грамматики, машины Тьюринга и вычислимо перечислимые языки

Опр. *Грамматикой* (системой преобразований, rewriting system) называется упорядоченная четверка $G = (V, \Sigma, S, R)$, где V – конечное множество символов, $\Sigma \subset V$ – множество *терминальных символов* (символы из $V \backslash \Sigma$ называются *нетерминальными*), $S \in V \backslash \Sigma$ – *начальный* нетерминальный символ, $R \subseteq (V^* \times (V \backslash \Sigma) \times V^*) \times V^*$ – конечное множество *правил*, называемых также *продукциями*.

Замечание. Для правила $((u, A, v), \alpha) \in R$ используем обозначение $uAv \to_G \alpha$.

Опр. Для грамматики $G = (V, \Sigma, S, R)$ и слов $\alpha, \beta \in V^*$, отношение $\alpha \Rightarrow_G \beta$ означает, что, существуют $\exists x, y, u, v, \gamma \in V^*$ и $A \in V \backslash \Sigma$, т.ч. $\alpha = xuAvy$, $\beta = x\gamma y$ и $(uAv \to_G \gamma) \in R$.

Опр. Отношение $\Rightarrow_G^*$ есть рефлексивное и транзитивное замыкание отношения $\Rightarrow_G$, т.е. $\forall u, v \in V^*$ $u \Rightarrow_G^* v \overset{def}{\Longleftrightarrow}$ либо $u = v$, либо $\exists n \geq 0$ $w_1, \ldots, w_n \in V^*$, т.ч. $u \Rightarrow_G w_1 \Rightarrow_G \ldots \Rightarrow_G w_n \Rightarrow v$

Опр. Для грамматики $G = (V, \Sigma, S, R)$, язык $L(G) = \{w \in \Sigma^* \mid S \Rightarrow_G^* w\}$ называется языком, порожденным грамматикой G.

Теорема 12. Для любой грамматики $G = (V, \Sigma, S, R)$ существует грамматика $G' = (V', \Sigma, S', R')$ такая, что $L(G) = L(G')$ и все правила в R' имеют вид

$$uAv \to u\gamma v,$$

где $u, v, \gamma \in (V')^*, A \in V' \setminus \Sigma$.

Доказательство. Помещаем в V' все элементы из V, в R' – все правила из R, и проводим следующие преобразования. Не нарушая общности, можно считать, что все правила в R' имеют вид $A_1A_2\dots A_m \to B_1B_2\dots B_n$, где все $A_1,\dots,A_m$ являются нетерминалами из $V'\setminus\Sigma$. Действительно, достаточно добавить в V' новые нетерминальные символы N_a для всех $a\in\Sigma$, заменить в R' все правила вида $u\to v$ на правила вида $N(u)\to N(v)$, где $N(C_1\dots C_k)=N(C_1)\dots N(C_k)$, причем $N(C)=C$, если $C\in V\setminus\Sigma$, и $N(C)=N_C$, если $C\in\Sigma$. Кроме того, нужно добавить в R' правила вида $N_a\to a$ для всех $a\in\Sigma$.

Рассмотрим, например, случай правила $A_1A_2\dots A_m\to B_1B_2\dots B_n$, в котором $m\leqslant n$. Заменяем его на множество правил

$A_1A_2\dots A_m\to A'_1A_2\dots A_m$,
$A'_1A_2\dots A_m\to A'_1A'_2A_3\dots A_m$,
$\dots$
$A'_1A'_2\dots A'_{m-1}A_m\to A'_1A'_2\dots A'_{m-1}B_m\dots B_n$,
$A'_1A'_2\dots A'_{m-1}B_m\dots B_n\to B_1A'_2\dots A'_{m-1}B_m\dots B_n$,
$\dots$
$B_1B_2\dots A'_{m-1}B_m\dots B_n\to B_1B_2\dots B_{m-1}B_m\dots B_n$.

(здесь $A'_1,\dots,A'_{m-1}$ – новые нетерминальные символы, которые добавляются в множество V').

Случай $m>n$ рассматривается аналогично. □

Опр. Грамматика $G=(V,\Sigma,S,R)$ называется *контекстно-зависимой* (или *неукорачивающейся*), если все правила в R (кроме, возможно, правила $S\to e$) имеют вид

$$uAv\to u\gamma v,$$

где $u,v\in V^*, A\in V\setminus\Sigma$, и $\gamma\in V^+$. В случае, когда в R входит правило $S\to e$, нетерминал S не должен содержаться в правых частях правил из R.

Язык называется *контекстно-зависимым*, если он порождается некоторой контекстно-зависимой грамматикой.

Предложение 4. Всякий контекстно-зависимый язык является вычислимым.

Доказательство. Пусть $G=(V,\Sigma,S,R)$ — контекстно-зависимая грамматика. Укажем алгоритм, который по любому слову $w\in\Sigma^*$ определяет, верно ли, что $w\in L(G)$. Имеется конечное множество упорядоченных наборов $(u_0,u_1,\dots,u_n)$ попарно различных слов $u_0,u_1,\dots,u_n\in V^*$, $n\geqslant 0$, таких, что $|u_0|\leqslant|u_1|\leqslant\dots\leqslant|u_n|\leqslant|w|$. Таким образом, достаточно проверить, существует ли в этом (вычислимом) множестве такой набор, для которого $u_0=S$, $u_n=w$, при этом u_{i+1} получается из u_i получается по одному из правил из R для всех $i<n$. □

Приведем пример вычислимого языка, не являющегося контекстно-зависимым. Пусть $\{\,w_n\,|\,n\in\omega\,\}$ — вычислимая нумерация всех слов алфавита $\{a,b\}$, и пусть

$\{G_n \,|\, n \in \omega\}$ — вычислимая нумерация всех контекстно-зависимых грамматик с алфавитом терминальных символов $\{a, b\}$. Определим язык $L_c \subseteq \{a, b\}^*$ следующим образом: для всякого $n \in \omega$, $w_n \in L_c$ тогда и только тогда, когда $w_n \notin L(G_n)$. Из построения непосредственно следует, что L_c — вычислимый язык, при этом $L_c \neq L(G_n)$ для всех $n \in \omega$.