Доклады Академии наук СССР
1974. Том 215, № 2

УДК 51:801 *МАТЕМАТИКА*

А. Л. СЕМЕНОВ

РЕГУЛЯРНОСТЬ ЯЗЫКОВ, k-ЛИНЕЙНЫХ ПРИ РАЗЛИЧНЫХ k

(Представлено академиком П. С. Новиковым 6 VII 1973)

Амаром и Пуцолу в работе (1) была поставлена следующая проблема: будет ли язык, k-линейный при всяком k, регулярным. В настоящей заметке доказывается, что язык, k-линейный для двух различных k, будет регулярным. Определения и свойства КС-грамматик и конечных автоматов считаются известными (2).

Введем некоторые обозначения. Фиксируем алфавит $\Sigma=\{a_1,\ldots,a_n\}$. Длину цепочки u будем обозначать $|u|$, множество всех цепочек длины p — Σ^p, пустую цепочку — λ, $\Sigma^*=\{\lambda\}\cup\Sigma\cup\Sigma^2\cup\ldots$, k — всегда рациональное число, равное p/q, k_i для $i=1,2$ — рациональное число, равное p_i/q_i, $p, q>0$.

Следующие определения имеются в работе (1).

Определение 1. КС-грамматика $\langle\Sigma, V, I, P\rangle$ называется k-линейной, если ее правила имеют вид:

$$A\to uBv,$$
$$A\to w,$$

где $q|u|=p|v|$; $A, B\in V$; $u, v, w\in\Sigma^*$.

Определение 2. Язык, порождаемый k-линейной грамматикой, называется k-линейным языком, или просто k-языком.

Определение 3. k-автоматом называется упорядоченная пятерка $\mathfrak{A}=\langle\Sigma, S, F, f, \delta\rangle$, где Σ — алфавит, $S=\{s_1,\ldots,s_m\}$ — множество состояний, F — множество заключительных состояний, f: $(\Sigma\cup\{\lambda\})^{p+q-1}\to S$ — функция выбора начального состояния, δ: $\Sigma^p\times S\times\Sigma^q\to S$ — функция перехода. Функция δ распространяется на $\bigcup_{l=0}^{\infty}\Sigma^{pl}\times S\times\Sigma^{ql}$ следующим образом:

$$\delta(\lambda, s, \lambda)=s,$$
$$\delta(u'u, s, vv')=\delta(u', \delta(u, s, v), v').$$

Всякое натуральное число d можно единственным образом представить в виде

$$d=np+\theta+nq, \tag{1}$$

где n, θ натуральные, $\theta<p+q$. Следовательно, всякую цепочку x в алфавите Σ можно единственным образом представить в виде

$$x=uwv,\ \text{где}\ q|u|=p|v|,\quad |w|<p+q. \tag{1$'$}$$

Определение 4. Цепочка x, представленная в виде (1′), допускается автоматом $\mathfrak{A}=\langle\Sigma, S, F, f, \delta\rangle$, если $\delta(u, f(w), v)\in F$.

Определение 5. Язык L представим в автомате $\mathfrak{A}$, если он состоит из всех цепочек, допускаемых автоматом $\mathfrak{A}$.

Таким образом, поведение автомата $\mathfrak{A}$ может быть неформально описано следующим образом: автомат сначала читает цепочку w и переходит в некоторое начальное состояние, затем он читает цепочки u и v кусками длины p и q соответственно с помощью двух головок, расходящихся от w в разные стороны. Цепочка допускается, если после ее прочтения автомат оказывается в заключительном состоянии.

Из результатов работы ([1]) непосредственно следует, что язык k-линеен тогда и только тогда, когда он представим в k-автомате. В той же работе доказано, что класс k-языков при фиксированном k замкнут относительно объединения, пересечения и дополнения и что регулярные языки являются k-линейными для любого k.

Лемма 1. *Пусть L — k-язык, x, y — фиксированные цепочки, $\tilde{L}$ — множество цепочек t, для которых $xty \in L$. Тогда $\tilde{L}$ — k-язык.*

Доказательство. Докажем лемму сначала для случая $x=\lambda$; $y=a\in\Sigma$. Представим язык L в виде конечного объединения языков L_w, отвечающих различным w, встречающимся в разложении (1′) цепочек из L, $L=\bigcup\limits_w L_w$. В силу замкнутости класса k-языков относительно объединения, лемму достаточно доказать для L_w. При фиксированном w автомат начинает работу над цепочками u и v, находясь всегда в одном и том же состоянии s_0. Язык L_w представим в k-автомате $\mathfrak{A}=\langle\Sigma, S, F, f, \delta\rangle$, среди состояний которого имеется «начальное» состояние s_0 и «тупиковое» состояние $s_\varnothing$ такое, что для любых u, v $\delta(u, s_\varnothing, v)=s_\varnothing$, $s_\varnothing \notin F$ и функция выбора начального состояния определена так:

$$f(w)=s_0, \quad f(z)=s_\varnothing \text{ для } z\in(\Sigma\cup\{\lambda\})^{p+q-1}-\{w\}.$$

Дальнейшее доказательство зависит от того, равно ли w пустой цепочке.

Первый случай: $w=\lambda$. Определим автомат $\mathfrak{B}=\langle\Sigma, S', F', f', \delta'\rangle$ следующим образом:

$S'=\Sigma^p\times S\times\Sigma^{q-1}\cup\{s_\varnothing\}$;

$f'(z)=(u_0, s_0, v_0')$, если $z=u_0v_0'$, $|u_0|=p$, $|v_0'|=q-1$;

$f'(z)=s_\varnothing$, если $|z|\neq p+q-1$;

$\delta'(u_2, (u_1, s, v_1'), a_iv_2')=(u_2, \delta(u_1, s, v_1'a_i), v_2')$, $|u_2|=p$, $|v_2'|=q-1$;

$(u, s, v')\in F'$, если $\delta(u, s, v'a)\in F$.

Второй случай: $w=\overline{w}a_{i_0}$. Определим $\mathfrak{B}=\langle\Sigma, S', F', f', \delta'\rangle$ следующим образом:

$$S'=\Sigma^p\times S\times\Sigma^{q-1}\cup\{s_0'\}\cup\{s_\varnothing\};$$

$$f'(\overline{w})=s_0'; \quad f'(z)=s_\varnothing, \quad \text{если } z\neq\overline{w};$$

$$\delta'(u, s_0', a_iv')=s_\varnothing, \quad \text{если } a_i\neq a_{i_0};$$

$$\delta'(u, s_0', a_{i_0}v')=(u, s_0, v');$$

$$\delta'(u_1, (u, s, v'), a_iv_1')=(u_1, \delta(u, s, v'a_i), v_1');$$

$$(u, s, v')\in F', \text{ если } \delta(u, s, v'a)\in F; \; s_0'\in F, \text{ если } a_{i_0}=a.$$

Таким образом, автомат $\mathfrak{B}$ с помощью внутренних состояний запоминает $q-1$ последнюю букву цепочки v, которая читается автоматом $\mathfrak{A}$ на некотором шаге, и цепочку u, которая читается автоматом $\mathfrak{A}$ на том же шаге, для того чтобы при появлении первой буквы следующей цепочки v смоделировать переход автомата $\mathfrak{A}$. Из определения автомата $\mathfrak{B}$ непосредственно вытекает представимость в нем $\tilde{L}$.

Доказательство для случая $x=a\in\Sigma$, $y=\lambda$ аналогично. Утверждение леммы легко получить теперь, используя индукцию по длине x и y.

Определение 6. k-язык L называется k^λ-языком, если длины всех его цепочек кратны $p+q$.

Следствие. *Пусть L — k-язык. Тогда $L=\bigcup\limits_{i=1}^{N} L_ie_i$, где L_i — k^λ-язык, $e_i\in\Sigma^*$.*

Автомат $\mathfrak{A}=\langle\Sigma, S, F, f, \delta\rangle$, в котором представим k^λ-язык, будем записывать в виде $\langle\Sigma, S, F, s_0, \delta\rangle$, где $s_0=f(\lambda)$. Этому автомату можно поставить в соответствие обычный автомат $\overline{\mathfrak{A}}=\langle\Sigma^p\times\Sigma^q, S, F, s_0, \bar{\delta}\rangle$, где $\bar{\delta}(s, (u, v))=\delta(u, s, v)$. Для автомата $\overline{\mathfrak{A}}$ существует приведенный, т. е. автомат, допускающий тот же язык, и такой, что любое состояние этого автомата достижимо из начального и любые два состояния неэквивалентны. Приведенному автомату соответствует k-автомат, в котором представим язык L. Последний автомат будем также называть приведенным.

Определение 7. Язык называется k_1, k_2-линейным языком, или k_1, k_2-языком, если он k_1-линеен и k_2-линеен; k_1^λ, k_2-язык — это k_1^λ-язык, являющийся k_2-языком.

Лемма 2. *Предположим, что L является k_1^λ, k_2-языком. Пусть L представим в некотором приведенном k_1-автомате $\mathfrak{A}=\langle\Sigma, S, F, s_0, \delta\rangle$ и $s\in S$.*

Тогда язык $\tilde{L}$, представимый в k_1-автомате $\langle\Sigma, S, \{s\}, s_0, \delta\rangle$, k_1^λ, k_2-линеен.

Доказательство. Если у автомата $\mathfrak{A}$ одно состояние, то лемма очевидна. Пусть число состояний равно $m>1$. Так как автомат $\mathfrak{A}$ приведенный, то для всякого состояния s', отличного от s, существует пара цепочек u и v, для которых $|u|=lp$, $|v|=lq$, $l<m$, и одно из состояний $\delta(u, s, v)$ и $\delta(u, s', v)$ принадлежит, а другое не принадлежит F. Для найденных u и v пусть

$$L(s')=\{t\in\Sigma^*\,|\,utv\in L\Leftrightarrow\delta(u, s, v)\in F\}.$$

Язык $L(s')$ содержит все цепочки, приводящие из s_0 в s, не содержит ни одной цепочки, приводящей из s_0 в s', и k_2-линеен в силу леммы 1 и замкнутости k-линейных языков относительно дополнения. Очевидно, что $\tilde{L}=\bigcap\limits_{s'\neq s}\tilde{L}(s')$. Следовательно, $\tilde{L}-k_1$, k_2-линейный язык.

В дальнейшем предполагается, что $k_1<k_2$ и l — натуральное число, делящееся на $M=(p_1q_2-p_2q_1)p_2$. Определим числа

$$n=\frac{lq_1(p_2+q_2)}{p_1q_2-p_2q_1},\quad l'=\frac{l(p_1+q_1)}{p_2},\quad n'=\frac{n(p_1+q_1)}{p_2+q_2}.$$

Легко проверить, что $n>l$, $n'>l'$ и выполняются равенства $(n-l)p_1=(n'-l')p_2$, $l(p_1+q_1)=l'p_2$. Отсюда число $d=n(p_1+q_1)$ можно представить в виде

$$d=(n-l)p_1+lp_1+lq_1+(n-l)q_1,\tag{2}$$

$$d=(n'-l')p_2+l'p_2+l'q_2+(n'-l')q_2.\tag{3}$$

Следовательно, имеют место следующие представления слова x длины d:

$$x=\tilde{u}uv\tilde{v},\tag{2'}$$

$$x=\tilde{u}yzt,\tag{3'}$$

где сомножители имеют длины, равные слагаемым в представлениях (2) и (3), при этом $y=uv$, $\tilde{v}=zt$. Таким образом, в начале чтения слова x k_1-автомат находится между u и v, а k_2-автомат — между v и $\tilde{v}$.

Определим эквивалентность, которую будем называть отдаленной эквивалентностью, на состояниях приведенного k-автомата $\mathfrak{A}=\langle\Sigma, S, F, f, \delta\rangle$ с числом состояний m

$$s_1\sim s_2 \overset{\mathrm{def}}{=} \forall u, v\quad q|u|=p|v|\geqslant pqm(m-1)\Rightarrow\delta(u, s_1, v)=\delta(u, s_2, v).$$

Классы отдаленной эквивалентности обозначим S_i.

Лемма 3. *Пусть s_1 и s_2 не отдаленно эквивалентны. Тогда существуют такие множества $A=(t_2)^*t_1$, $B=z_1(z_2)^*$, где $t_i\in(\Sigma^p)^*$, $z_i\in(\Sigma^q)^*$, что для всех $u\in A$, $v\in B$, если $q|u|=p|v|$, то $\delta(u, s_1, v)\neq\delta(u, s_2, v)$.*

Доказательство. Из определения отдаленной эквивалентности вытекает существование u и v вида $u=u_nu_{n-1}\ldots u_1$, $v=v_1v_2\ldots v_n$, где $|u_i|=p$, $|v_i|=q$, $n\geqslant m(m-1)$, для которых $\delta(u, s_1, v)\neq\delta(u, s_2, v)$. Рассмотрим последовательность пар состояний $(s_1^{(0)}, s_2^{(0)})=(s_1, s_2)$; $(s_1^{(1)}, s_2^{(1)})=(\delta(u_1, s_1^{(0)}, v_1),\ \delta(u_1, s_2^{(0)}, v_2))$; ...; $(s_1^{(n)}, s_2^{(n)})=(\delta(u_n, s_1^{(n-1)}, v_n), \delta(u_n, s_2^{(n-1)}$

$v_n)$). $s_1^{(i)} \neq s_2^{(i)}$ для всякого i. Так как $n \geqslant m(m-1)$, то в последовательности встретятся две одинаковые пары. Пусть, например, $s_1^{(i)} = s_1^{(j)}$, $s_2^{(i)} = s_2^{(j)}$, $j > i$. Тогда

$$A = (u_j \ldots u_{i+1})^* u_i \ldots u_1, \quad B = v_1 \ldots v_i (v_{i+1} \ldots v_j)^*,$$

очевидно, удовлетворяет условию.

Следствие. *Пусть r — фиксированное натуральное число и для всех i и j из $u_{ij} \in U_{ij}$, $v_{ij} \in V_{ij}$, $|u_{ij}| = rp$, $|v_{ij}| = rq$ следует $\delta(u_{ij}, s_1, v_{ij}) = \delta(u_{ij}, s_2, v_{ij})$. Тогда $s_1 \sim s_2$.*

Пусть L — $k_1{}^\lambda$, k_2-язык, представимый в k_1-автомате $\mathfrak{A} = \langle \Sigma, S, F, s_0, \delta \rangle$. По лемме 2, для всякого состояния s_h автомата $\mathfrak{A}$ существует k_2-автомат $\mathfrak{B}_h = \langle \Sigma, T_h, F_h, f_h, \delta_h \rangle$, в котором представим тот же язык, что и в k_1-автомате $\langle \Sigma, S, s_h, s_0, \delta \rangle$.

Лемма 4. *Пусть u, v, u', v' — цепочки в алфавите Σ такие, что $|u| = |u'| = lp$; $|v| = |v'| = lq$; $l = Mr$; $M = (p_1 q_2 - p_2 q_1) p_2$; r натуральное. Пусть далее для всех i, j, h из равенств*

$$\tilde{u} u v \tilde{v} = xyzt, \quad \tilde{u} u' v' \tilde{v} = xy' zt,$$

левые части которых есть разложения (2′), а правые части есть разложения (3′) и $\tilde{u} \in U_{ij}$, $\tilde{v} \in V_{ij}$, следует, что $\delta_h(y, f_h(\lambda), z) = \delta_h(y', f_h(\lambda), z)$.

Тогда $\delta(u, s_0, v) \sim \delta(u', s_0, v')$.

Доказательство. Так как $\delta_h(y, f_h(\lambda), z) = \delta_h(y', f_h(\lambda), z)$, то $\delta_h(xy, f_h(\lambda), zt) = \delta_h(xy', f_h(\lambda), zt)$ для всех h. Следовательно, $\delta(\tilde{u} u, s_0, v \tilde{v}) = \delta(\tilde{u} u', s_0, v' \tilde{v})$. Отсюда и из следствия леммы 3 сразу вытекает требуемое утверждение.

Лемма 5. *Для всякого S_i множество R_i цепочек uv таких, что $|u| = lp_1$, $|v| = lq_1$, $\delta(u, s_0, v) \in S_i$, регулярно.*

Доказательство. Тройке i, j, h поставим в соответствие разбиение $(\Sigma^{M(p_1+q_1)})^*$ на конечное число классов C_α. Цепочка x принадлежит классу C_α, если $\delta_h(x, f_h(\lambda), v_{ij}) = t_\alpha \in T_h$. Докажем, что C_α регулярно. Множество P всех цепочек uv таких, что $|u| = lp_2$, $|v| = lq_2$, $u \in \Sigma^*$, $v \in V_{ij}$, k_2-линейно. Множество Q_α цепочек uv таких, что $\delta_h(u, f_h(\lambda), v) = t_\alpha$, k_2-линейно. Следовательно, $L_\alpha = P \cap Q_\alpha$ k_2-линейно. Тогда C_α — это множество всех u в разложении (1′) цепочек из L_α. Нетрудно построить конечный автомат, допускающий это множество.

Для каждой тройки i, j, h выберем какой-нибудь один класс и рассмотрим пересечение таких классов во всех i, j, h. Таких пересечений конечное число, все они регулярны, их объединение есть $(\Sigma^{M(p_1+q_1)})^*$ и каждое из них целиком лежит в некотором R_i. Регулярность R_i доказана.

Теорема. *Пусть L — k_1, k_2-язык. Тогда L регулярен.*

Доказательство. В силу следствия из леммы 1, теорему достаточно доказать для $k_1{}^\lambda$, k_2-языка. Для всякого множества R_i цепочек длины, кратной $M(p_1+q_1)$, переводящих $\mathfrak{A}$ в одно из состояний S_i, определим множество D_i пар цепочек (u, v) таких, что $m(m-1)p_1 < |u| \leqslant (M + m(m-1))p_1$, $m(m-1)q_1 < |v| \leqslant (M + m(m-1))q_1$ и $\delta(u, s, v) \in F$ для некоторого состояния s из S_i, а значит, и для всех состояний из S_i. Множество E_i цепочек вида uxv, где $x \in R_i$, $(u, v) \in D_i$ регулярно. Множество G всех цепочек y таких, что $y \in L$, $|y| \leqslant m(m-1)(p_1+q_1)$, конечно. Очевидно, $L = (\cup E_i) \cup G$ регулярен. Теорема доказана.

Автор благодарит А. А. Мучника и А. Маслова за помощь в работе.

Московский государственный университет
им. М. В. Ломоносова

Поступило
5 VII 1973

ЦИТИРОВАННАЯ ЛИТЕРАТУРА

[1] *V. Amar, G. Putzolu*, Information and Control, v. 8, № 1 (1965). [2] *А. В. Гладкий*, Формальные грамматики и языки, М., 1973.