Теоретическая информатика, осень 2020 г. Лекция 5. Определения грамматик. Примеры построения грамматик. Действия над языками, выразимые в грамматиках. Ограничения выразительной мощности грамматик: лемма о накачке*

Александр Охотин

1 октября 2020 г.

Содержание

1 Определения грамматик

Определение 1. *(Формальная) грамматика — это четвёрка $G = (\Sigma, N, R, S)$, состоящая из следующих компонентов.*

- *Конечное множество* ***символов*** Σ — алфавит *определяемого языка.*
- *Конечное множество N — это множество определяемых в грамматике свойств строк, которым всякая строка над алфавитом Σ обладает или не обладает. Обозначаются обычно буквами $A, B, C, \ldots$*

  В информатике элементы N традиционно называют «нетерминальными символами» («нетерминалами») — отсюда буква N — поскольку пути в дереве разбора на них не заканчиваются. В лингвистике они называются синтаксическими категориями.
- *Конечное множество R* правил грамматики, *каждое из которых описывает возможную структуру строк со свойством $A \in N$ в виде конкатенации $u_0 B_1 u_1 \ldots B_\ell u_\ell$, где*

*Краткое содержание лекций, прочитанных студентам 2-го курса факультета МКН СПбГУ в осеннем семестре 2020–2021 учебного года. Страница курса: `http://users.math-cs.spbu.ru/~okhotin/teaching/tcs_fl_2020/`.

$B_1, \ldots, B_\ell$ ($\ell \geqslant 0$) — все нетерминальные символы, на которые ссылается правило, а любые символы, написанные между ними, образуют строки $u_0, u_1, \ldots, u_\ell$.

$$A \to u_0 B_1 u_1 \ldots B_\ell u_\ell \qquad (A \in N,\ \ell \geqslant 0,\ B_1, \ldots, B_\ell \in N,\ u_0, u_1, \ldots, u_\ell \in \Sigma^*)$$

- Начальный символ $S \in N$, *от «sentence», обозначает множество всех синтаксически правильных строк, определяемых в грамматике.*

Общий вид правил часто записывается как $A \to X_1 \ldots X_\ell$, где $X_1, \ldots, X_\ell \in \Sigma \cup N$ ($\ell \geqslant 0$) — все символы и нетерминальные символы, конкатенация которых записана в правиле. Дальнейшее обозначение: $A \to \alpha$, где $\alpha = X_1 \ldots X_\ell$ — строка над алфавитом $\Sigma \cup N$.

Несколько правил $A \to \alpha_1, \ldots, A \to \alpha_m$ для одного и того же нетерминального символа $A \in N$ записывают возможную структуру строк с этим свойством. Это такой же оператор выбора, как и в регулярных выражениях, и потому такие правила обычно записываются в одну строчку.

$$A \to \alpha_1 \mid \ldots \mid \alpha_m$$

Пример 1. *Грамматика для языка Дика записывается как $G = (\Sigma, N, R, S)$, где $\Sigma = \{a, b\}$, $N = \{S\}$ и $R = \{S \to aSb,\ S \to SS,\ S \to \varepsilon\}$. Такую грамматику записывают следующим образом.*

$$S \to aSb \mid SS \mid \varepsilon$$

Дерево разбора строки $w = abaabb$ в соответствии с этой грамматикой приведено на рис. **??**.

1.1 Определение через дерево разбора

Дерево разбора — это уже само по себе формальное определение.

Дерево разбора строки $w \in \Sigma^*$. Потомки каждой внутренней вершины упорядочены, откуда вытекает порядок на листьях. Листья слева направо помечены символами строки w. Каждая внутренняя вершина помечена некоторым нетерминальным символом $A \in N$, и если $X_1, \ldots, X_\ell$ — её потомки, то в грамматике должно быть правило $A \to X_1 \ldots X_\ell$. Для правила $A \to \varepsilon$ вершина помечена символом A, не имеет потомков, но всё равно рассматривается в качестве внутренней.

Язык, задаваемый грамматикой, $L(G) \subseteq \Sigma^*$, состоит из всех строк, для которых есть дерево разбора.

1.2 Определение через логический вывод

Логическая система, в которой выводятся все верные высказывания вида «строка w имеет свойство A», обозначаемые через $A(w)$.

Пример 1(D). *В грамматике из примера 1 определён один нетерминальный символ S, задающий все правильно вложенные строки. Поэтому высказывание вида $S(w)$ можно прочитать как «w — правильно вложенная строка».*

Утверждение о правильной вложенности строки $w = abaabb$ выводится так.

$$\frac{}{S(\varepsilon)}(S \to \varepsilon) \qquad \frac{S(\varepsilon)}{S(ab)}(S \to aSb) \qquad \frac{S(ab)}{S(aabb)}(S \to aSb) \qquad \frac{S(ab), S(aabb)}{S(abaabb)}(S \to SS)$$

Этот вывод можно записать в виде дерева доказательства, разделив $S(ab)$ для двух разных подстрок ab.

$$\dfrac{\dfrac{\dfrac{S(\varepsilon)}{S(ab)}}{S(aabb)} \quad \dfrac{S(\varepsilon)}{S(ab)}}{S(abaabb)}$$

Это тот же объект, что и дерево разбора.

Определение 1(D). *Для грамматики $G = (\Sigma, N, R, S)$, высказывания имеют вид «строка w имеет свойство A», где $w \in \Sigma^*$ и $A \in N$, и обозначаются через $A(w)$.*

Пусть $A \to u_0 B_1 u_1 \dots B_\ell u_\ell$ — правило, в котором $B_1, \dots, B_\ell \in N$, где $\ell \geqslant 0$ — это все нетерминальные символы, на которые оно ссылается, а $u_0, u_1, \dots, u_\ell \in \Sigma^*$ — символы между ними. Это правило позволяет сделать следующий логический вывод, для любых строк $v_1, \dots, v_\ell$, где над чертой — посылки, *а под чертой* — следствие.

$$\frac{B_1(v_1), \dots, B_\ell(v_\ell)}{A(u_0 v_1 u_1 \dots v_\ell u_\ell)}(A \to u_0 B_1 u_1 \dots B_\ell u_\ell) \qquad (\textit{for all } v_1, \dots, v_\ell \in \Sigma^*)$$

Вывод высказывания $A(u)$ — это последовательность таких шагов вывода, где в качестве посылок на каждом шаге используются ранее выведенные высказывания: $I_j \subseteq \{\, A_i(u_i) \mid i \in \{1, \dots, j-1\} \,\}$, для всех j.

$$\frac{I_1}{A_1(u_1)}, \quad \frac{I_2}{A_2(u_2)}, \quad \dots \quad \frac{I_{z-1}}{A_{z-1}(u_{z-1})}, \quad \frac{I_z}{A(u)}$$

Если такой вывод существует, это обозначается через $\vdash A(u)$.

Тогда, для всех $A \in N$, определяется $L_G(A) = \{\, w \mid\vdash A(w) \,\}$. Язык, задаваемый грамматикой — $L(G) = L_G(S) = \{\, w \mid\vdash S(w) \,\}$.

1.3 Определение через перезапись строк

Перезапись *сентенциальных форм* — строк над алфавитом $\Sigma \cup N$. Каждая такая строка — это схема предложения, в которой каждое вхождение каждого нетерминального символа $A \in N$ означает *некоторую неуказанную строку со свойством A*.

Перезапись начинается с S, что означает *любое правильное предложение*. На каждом шаге перезаписи некоторый нетерминальный символ A заменяется на правую часть правила для A, так что получается более определённая сентенциальная форма. Перезапись продолжается, пока не останутся только символы из Σ — то есть, предложение языка.

Пример 1(R). *Для грамматики из примера 1, строка abaabb получается перезаписью так.*

$$S \Longrightarrow SS \Longrightarrow aSbS \Longrightarrow abS \Longrightarrow abaSb \Longrightarrow abaaSbb \Longrightarrow abaabb$$

Определение 1(R) (Хомский)**.** *Пусть $G = (\Sigma, N, R, S)$ — грамматика. Если $\eta A \theta$ — строка над совмещённым алфавитом символов и нетерминальных символов $\Sigma \cup N$ (где $A \in N$ и $\eta, \theta \in (\Sigma \cup N)^*$), и если $A \to \alpha$ — некоторое правило для A, тогда строка $\eta A \theta$ может быть* ***перезаписана в*** *$\eta\alpha\theta$* ***за один шаг****, что обозначается следующим образом.*

$$\eta A \theta \Longrightarrow \eta \alpha \theta \qquad (\text{для всех } A \to \alpha \in R \text{ and } \eta, \theta \in (\Sigma \cup N)^*)$$

Последовательность строк $\alpha_0, \alpha_1, \ldots, \alpha_n$ над алфавитом $\Sigma \cup N$, где $n \geqslant 0$, называется цепочкой перезаписи*, если всякая строка α_i может быть перезаписана за один шаг в строку α_{i+1}. Это обозначается так.*

$$\alpha_0 \Longrightarrow \alpha_1 \Longrightarrow \ldots \Longrightarrow \alpha_n$$

Тогда говорится, что α_0 перезаписывается в α_n за n шагов (обозначение: $\alpha_0 \Longrightarrow^n \alpha_n$). Также вводятся обозначения для перезаписи за ноль и более шагов ($\alpha \Longrightarrow^* \beta$), за один и более шагов ($\alpha \Longrightarrow^+ \beta$), и за не более чем n шагов ($\alpha \Longrightarrow^{\leqslant n} \beta$).

Язык, задаваемый грамматикой.

$$L_G(A) = \{\, w \mid w \in \Sigma^*,\ A \Longrightarrow^+ w \,\} \qquad (A \in N)$$
$$L(G) = L_G(S)$$

1.4 Определение через языковые уравнения

Гинзбург и Райс [1962]: грамматика представляется в виде системы уравнений, где языки — неизвестные. Всякий нетерминальный символ в грамматике $G = (\Sigma, N, R, S)$ становится *переменной*, принимающей значение языка над Σ. Значение переменной $A \in N$ в решении системы — это и есть множество строк, имеющих свойство A.

Рис. 1: Сеймур Гинзбург (1928–2004).

Эта *система языковых уравнений* содержит одно уравнение вида $A = \varphi_A$ для каждой переменной A, где φ_A — выражение с аргументами-языками, содержащее в себе все правила для нетерминального символа A. При преобразовании всякий оператор выбора переходит в операцию *объединения множеств*, конкатенация в правилах становится конкатенацией языков в уравнении, и всякий символ $a \in \Sigma$, использованный в правиле, становится одноэлементным языком $\{a\}$.

Пример 1(E). *Грамматика в примере 1 представляется следующим языковым уравнением, где объединение трёх конкатенаций в правой части соответствует трём правилам для S.*

$$S = \underbrace{\underbrace{(\{a\} \cdot S \cdot \{b\})}_{S \to aSb} \cup \underbrace{(S \cdot S)}_{S \to SS} \cup \underbrace{\{\varepsilon\}}_{S \to \varepsilon}}_{\varphi_S}$$

Одно из решений этого уравнения — язык Дика.

Уравнение можно прочитать так: «w — правильно вложенная строка тогда и только тогда, когда она или имеет вид $w = aub$, где u — правильно вложенная, или представляется в виде конкатенации $w = uv$ двух правильно вложенных строк u, v, или это пустая строка $w = \varepsilon$».

Это не единственное решение; есть, например, $S = \Sigma^*$. Язык, задаваемый грамматикой, определяется с помощью *наименьшего решения* относительно частичного порядка поэлементного включения. Для данного уравнения, язык Дика — наименьшее решение.

Теорема 1 (Гинзбург и Райс [1962])**.** *Пусть $X_1, \ldots, X_n$ — переменные, и для каждой переменной X_i, пусть $\varphi_i(X_1, \ldots, X_n)$ — выражение, содержащее переменные, любые постоянные языки, а также операции объединения и конкатенации. Тогда система n уравнений $X_i = \varphi_i(X_1, \ldots, X_n)$, где $i \in \{1, \ldots, n\}$, имеет решения, и среди них есть* ***наименьшее решение*** *$(L_1, \ldots, L_n)$ — то есть такое, что всякое решение $(L'_1, \ldots, L'_n)$ должно удовлетворять $L_i \subseteq L'_i$ для всех i.*

Определение 1(E) (Гинзбург и Райс [1962])**.** *Пусть $G = (\Sigma, N, R, S)$ — грамматика. Соответствующая система языковых уравнений такова:*

$$A = \underbrace{\bigcup_{A \to u_0 B_1 u_1 \ldots B_\ell u_\ell \in R} \{u_0\} \cdot B_1 \cdot \{u_1\} \cdot \ldots \cdot B_\ell \cdot \{u_\ell\}}_{\varphi_A} \qquad (\textit{for all } A \in N)$$

Пусть наименьшее решение имеет вид $A = L_A$, для всех $A \in N$. Тогда $L(G)$ определяется как L_S.

1.5 Равносильность четырёх определений

Они равносильны, т.е., определяют те же значения $L_G(A)$ и $L(G)$.

Теорема 2. *Пусть $G = (\Sigma, N, R, S)$ — грамматика, как в определении 1. Для всякого нетерминального символа $A \in N$ и для всякой строки $w \in \Sigma^*$, следующие утверждения равносильны:*

(T). существует дерево разбора w из A;

(D). высказывание $A(w)$ выводимо ($\vdash A(w)$);

(R). A можно перезаписать в w за один и более шагов ($A \Longrightarrow^{+} w$);

(E). w принадлежит A-компоненту наименьшего решения системы языковых уравнений ($w \in L_A$).

На лекции доказывалась равносильность определений через логический вывод и через перезапись строк: в одну сторону — индукцией по длине вывода, в другую — индукцией по длине последовательности перезаписи.

2 Примеры построения грамматик

2.1 Задание регулярных конструкций

Теорема 3. *Для всякого регулярного языка есть грамматика.*

Доказательство. Пусть $\mathcal{A} = (\Sigma, Q, Q_0, \delta, F)$ — произвольный NFA.

Строится грамматика $G = (\Sigma, N, R, S)$, где $N = \{ A_q \mid q \in Q \}$ и $S = A_{q_0}$. Цель: чтобы всякий нетерминальный символ A_q задавал множество всех строк, принимаемых автоматом из состояния q.

$$L_G(A_q) = \{ w \mid \delta(q, w) \in F \}$$

Для всякого перехода в автомате вводится следующее правило.

$$A_q \to aA_r \qquad (a \in \Sigma,\ r \in \delta(q, a))$$

Для принимающих состояний автомата, соответствующий нетерминальный символ в грамматике задаёт пустую строку.

$$A_q \to \varepsilon \qquad (q \in F)$$

Тогда $L(G) = L(\mathcal{A})$. □

Можно доказать иначе, используя регулярное выражение вместо NFA.

Другое доказательство теоремы 3. Пусть α — произвольное регулярное выражение.

Строится грамматика $G = (\Sigma, N, R, S)$, где $N = \{ A_\varphi \mid \varphi$ — подвыражение $\alpha \}$ и $S = A_\alpha$. Цель: чтобы всякий нетерминальный символ A_φ задавал тот же язык, что и регулярное выражение φ.

$$L_G(A_\varphi) = L(\varphi)$$

Построение — индукцией по структуре выражения. Для одиночного символа $a \in \Sigma$ соответствующий нетерминальный символ его и задаёт.

$$A_a \to a$$

Для нетерминального символа, соответствующего регулярному выражению $\varnothing$, можно вообще не задавать правил; а можно, например, задать такое правило, которое всё равно не определит ни одной строки.

$$A_\varnothing \to A_\varnothing$$

Если регулярное выражение — это конкатенация двух выражений, то правило задаёт конкатенацию.

$$A_{(\varphi\psi)} \to A_\varphi A_\psi$$

Если выражение — это выбор между двумя подвыражениями, то в грамматике два правила.

$$A_{(\varphi \mid \psi)} \to A_\varphi \mid A_\psi$$

Наконец, повторение (звёздочку) можно выразить следующими правилами.

$$A_{\varphi^*} \to A_\varphi A_{\varphi^*} \mid \varepsilon$$

Доказательство правильности — опять же индукцией по структуре выражения. □

2.2 Сравнение числа объектов

Простейшая нетривиальная вещь, которую могут описать грамматики — соотнесение числа подстрок в одной части строки с числом подстрок в другой.

Пример 2. *Язык* $\{ a^n b^n \mid n \geqslant 0 \}$ *задаётся следующей грамматикой.*

$$S \to aSb \mid \varepsilon$$

Дерево разбора устанавливает связь между соответствующими a и b.

Тем же способом можно связать между собой подстроки произвольного вида, описываемого регулярным языком. В следующем примере строка содержит два списка чисел, и грамматика описывает, что число элементов в обоих списках должно быть одинаковым.

Пример 3. *Алфавит* $\Sigma = \{\texttt{0}, \texttt{1}, \texttt{,}, \texttt{;}\}$.

$$S \to A\texttt{,}S\texttt{,}A \mid A\texttt{;}A$$
$$A \to \texttt{0}A \mid \texttt{1}A \mid \texttt{0} \mid \texttt{1}$$

В следующем примере граница между сравниваемыми кусками внешне не видна.

Пример 4. *Язык* $\{ a^m b^{m+n} a^n \mid m, n \geqslant 0 \}$ *описывается следующей грамматикой, в которой строка* $a^m b^{m+n} a^n$ *представлена как конкатенация* $a^m b^m$ *и* $b^n a^n$.

$$S \to AB$$
$$A \to aAb \mid \varepsilon$$
$$B \to bBa \mid \varepsilon$$

Следующий пример — вариант языка Дика, где условие правильной вложенности снимается, достаточно совпадения количества символов a и b, записанных в произвольном порядке.

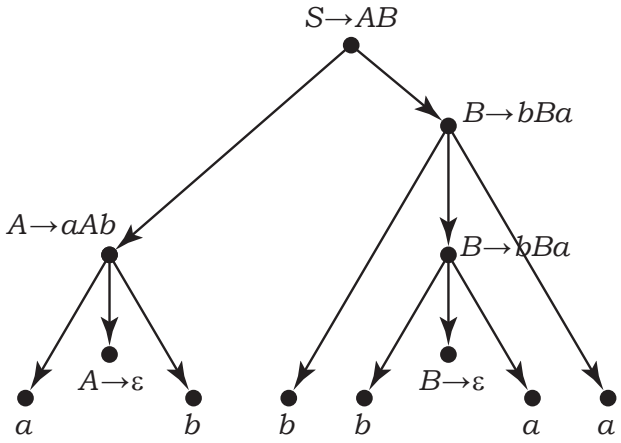

Рис. 2: Дерево разбора строки *abbbaa* по грамматике в примере 4.

Пример 5. *Двухсторонний язык Дика* $L = \{ w \mid w \in \{a,b\}^*,\ |w|_a = |w|_b \}$ *задаётся следующей грамматикой.*

$$S \to SS \mid aSb \mid bSa \mid \varepsilon$$

Доказательство. То, что приведённая грамматика действительно задаёт именно этот язык, надо ещё доказать.

С одной стороны, нужно доказать, что всякая строка, определяемая грамматикой, лежит в языке L. Это без затей доказывается индукцией по высоте дерева разбора. Базовый случай: пустая строка задаётся грамматикой и лежит в L. Переход, правило $S \to SS$: если строка w задаётся грамматикой по этому правилу, то $w = uv$, причём u и v задаются деревьями меньшей высоты; тогда $u, v \in L$ по предположению индукции и, стало быть $|u|_a = |u|_b$ и $|v|_a = |v|_b$, откуда следует $|uv|_a = |uv|_b$. Аналогично для двух других правил.

В другую сторону, пусть $|w|_a = |w|_b$. Надо доказать, что w задаётся грамматикой. Индукция по длине строки. Если $w = \varepsilon$, то задаётся. Если $w = axb$, то по предположению индукции x задаётся, и тогда w задаётся по правилу $S \to aSb$. Случай $w = bxa$ — аналогично, используя правило $S \to bSa$.

Пусть w начинается с a и заканчивается на a. Рассматривается функция $f\colon \{0, 1, \ldots, |w|\} \to \mathbb{Z}$, где $f(i)$ определяется как разность между количеством символов a и b среди первых i символов строки w. Тогда $f(1) = 1$, $f(|w| - 1) = -1$, и значения функции на двух последовательных шагах отличаются на 1. Стало быть, существует i, для которого $f(i) = 0$, что даёт разложение $w = uv$, где $|u| = i$ и $|u|_a = |u|_b$, откуда также следует, что $|v|_a = |v|_b$. По предположению индукции, строки u, v задаются грамматикой, и тогда w получается по правилу $S \to SS$.

Случай w, начинающейся и заканчивающейся на b, рассматривается аналогично. □

2.3 Более сложные конструкции

Конкатенация работает в каком-то смысле как квантор существования, и с её помощью можно описать, что где-то в строке — то есть, при каком-то разложении строки на подстроки — выполняется некое условие.

Пример 6. *Язык* $\overline{\{\, ww \mid w \in \{a,b\}^* \,\}}$ *описывается следующей грамматикой.*

$$
\begin{aligned}
S &\to AB \mid BA \mid O \\
A &\to XAX \mid a \\
B &\to XBX \mid b \\
X &\to a \mid b \\
O &\to XXO \mid X
\end{aligned}
$$

(см. рис. 3)

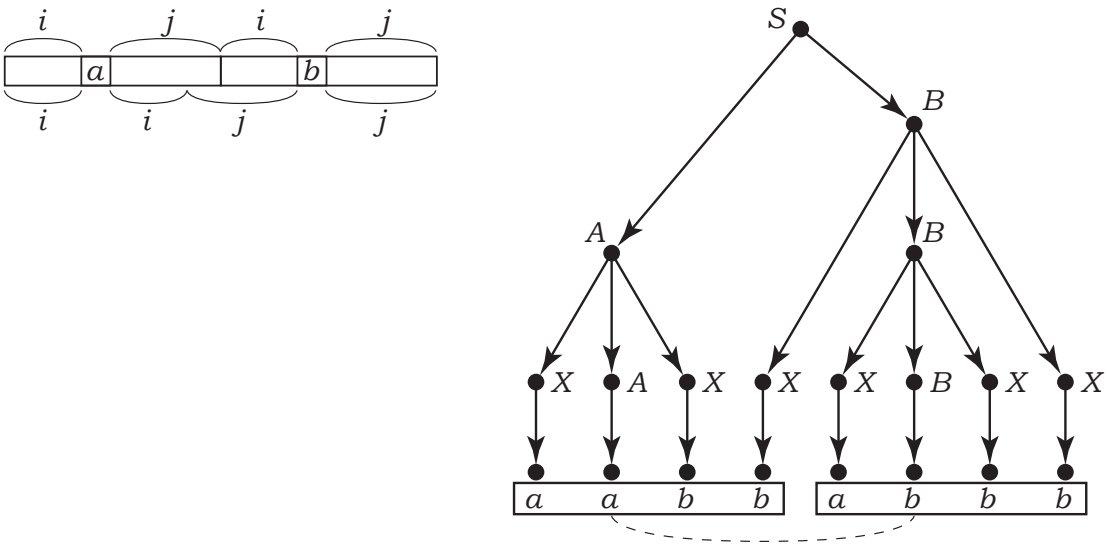

Рис. 3: (слева) Как грамматика в примере 6 определяет строки вида uv, где $|u| = |v|$ и $u \neq v$; (справа) дерево разбора строки *aabbabbb*.

3 Действия над языками, выразимые в грамматиках

Класс языков, задаваемых грамматиками, замкнут относительно объединения, конкатенации, повторения; все три действия прямо выражаются, как во втором доказательстве теоремы 3. Замкнут относительно обращения — в произвольной грамматике достаточно записать в обратном порядке правые части всех правил. Замкнут относительно циклического сдвига (просто, но неочевидно).

Замкнутость относительно пересечения с регулярными языками доказывается так.

Теорема 4 (Бар-Хиллель, Перлес и Шамир [1961])**.** *Для всякой грамматики G и для всякого регулярного языка K, язык $L(G) \cap K$ описывается некоторой грамматикой.*

Это, в частности, означает, что при написании грамматики можно задавать регулярные ограничения на форму подстрок одновременно с определением общей структуры предложения — и конструктивное доказательство служит примером введения таких ограничений в существующую грамматику.

Например, если G описывает грамотные предложения на естественном языке, а K задаёт все последовательности слогов, в которых выдержан стихотворный размер и есть рифма, то пересечение даст грамматику для стихов.

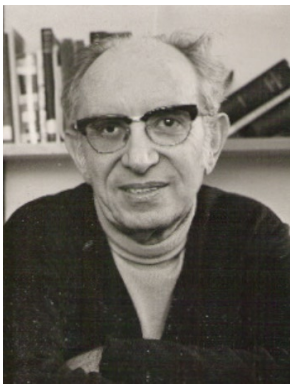
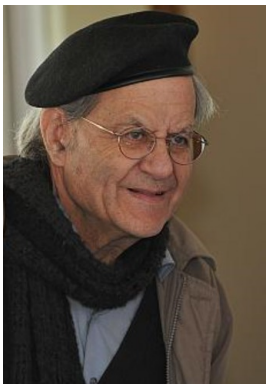


Рис. 4: Иегошуа Бар-Хиллель (1915–1975), Миха Перлес (род. 1936), Элияху Шамир (род. 1934).

Доказательство. Пусть $G = (\Sigma, N, R, S)$ — грамматика, пусть $M = (\Sigma, Q, q_0, \delta, F)$ — DFA. В новой грамматике $G' = (\Sigma, N', R', S')$ нетерминальные символы имеют вид $A_{q,q'}$, где $A \in N$ — одна из нетерминальных символов из G, а q и q' — любые два состояния M. Цель: чтобы $A_{q,q'}$ обозначало множество всех строк со свойством A (определённом в исходной грамматике) на которых DFA, начав вычисление в состоянии q, закончит читать в состоянии q'.

$$L_{G'}(A_{q,q'}) = L_G(A) \cap \{\, w \mid \delta(q, w) = q' \,\}$$

$$N' = \{S'\} \cup \{\, A_{q,q'} \mid A \in N,\ q, q' \in Q \,\}$$

Деревья разбора в новой грамматике имеют такую же структуру, как в исходной, но дополнительно содержат данные, необходимые для моделирования работы DFA.

Пусть $A_{q,q'} \in N'$ и пусть

$$A \to u_0 B^{(1)} u_1 \dots B^{(\ell)} u_\ell \tag{1}$$

— правило для A в исходной грамматике. Правило определяет строки вида $u_0 v_1 u_1 \dots v_\ell u_\ell$, где всякая подстрока v_i имеет свойство $B^{(i)}$. В новой грамматике соответствующее правило должно также описывать вычисление DFA на той же строке $u_0 v_1 u_1 \dots v_\ell u_\ell$, начинающееся в состоянии q и заканчивающееся в состоянии q'. Пусть $p_0, p_1, \dots, p_\ell \in Q$ — последовательность промежуточных состояний, где каждое состояние p_i достигается перед чтением соответствующей подстроки u_i. Тогда $p_0 = q$ и $\delta(p_\ell, u_\ell) = q'$. Для всякой такой строки автомат начинает чтение каждой подстроки v_i в состоянии $\delta(p_{i-1}, u_{i-1})$, и должен закончить её читать в состоянии p_i. В новой грамматике это обеспечивается ссылкой на нетерминальный символ $B^{(i)}_{\delta(p_{i-1},u_{i-1}),p_i}$, помеченный двумя состояниями из предполагаемой последовательности. Тогда новая грамматика содержит следующее правило.

$$A_{q,q'} \to u_0 B^{(1)}_{\delta(p_0,u_0),p_1} u_1 B^{(2)}_{\delta(p_1,u_1),p_2} u_2 \dots B^{(\ell)}_{\delta(p_{\ell-1},u_{\ell-1}),p_\ell} u_\ell \tag{2}$$

Правила для S'.

$$S' \to S_{q_0,q} \qquad\qquad (q \in F)$$

□

Построение напоминает прямое произведение двух DFA. Возможно ли прямое произведение двух произвольных грамматик, задающее пересечение языков? Нет — потому что две грамматики определяют для одной строки деревья разбора различной структуры, и такие два дерева не объединить в одно. Далее будет доказано, что класс языков, задаваемых грамматиками, вообще не замкнут относительно пересечения.

Ещё один пример преобразования грамматики.

Теорема 5. *Пусть $G = (\Sigma, N, R, S)$ — грамматика. Тогда существует грамматика, задающая язык* $\text{prefixes}(L) = \{\, u \mid \exists v : uv \in L(G) \,\}$.

Доказательство. $G' = (\Sigma, N \cup N', R \cup R', S')$, где $N' = \{\, A' \mid A \in N \,\}$. Цель построения: $L_{G'}(A) = L_G(A)$ и $L_{G'}(A') = \text{prefixes}(L_G(A))$. Для каждого A используются все старые правила из R, а для A' определяются следующие новые правила.

Если для A было правило $A \to X_1 \dots X_\ell \in R$, задающее конкатенации $u_1 \dots u_\ell$, то A' должен уметь задавать все префиксы этих конкатенации. Во-первых, будут префиксы вида $u_1 \dots u_k$, в которые первые k подстрок входят целиком, а остальные целиком не входят.

$$A' \to X_1 \dots X_{k-1} X_k \qquad (A \to X_1 \dots X_\ell \in R,\ k \in \{0, \dots, \ell\})$$

Во-вторых, возможны префиксы вида $u_1 \dots u_{k-1} x$, состоящие из первых $k-1$ подстрок целиком, продолженных некоторым префиксом строки u_k. Такие префиксы задаются следующим правилом.

$$A' \to X_1 \dots X_{k-1} X'_k \qquad (A \to X_1 \dots X_\ell \in R,\ k \in \{1, \dots, \ell\})$$

□

Упражнение 1. *Доказать, что для всякой грамматики G существует грамматика G', задающая* циклический сдвиг *языка $L(G)$, то есть, $L(G') = \{\, vu \mid uv \in L(G) \,\}$.*

Упражнение 2. *Доказать, что для всякого регулярного языка L над двусимвольным алфавитом существует грамматика, задающая множество всех перестановок всех строк из L.*

$$\{\, a_{i_1} \dots a_{i_\ell} \mid a_1 \dots a_\ell \in L,\ (i_1, \dots, i_\ell) \text{ — перестановка} \,\}$$

4 Ограничения выразительной мощности грамматик

4.1 Лемма о накачке

Лемма 1 (Лемма о накачке для грамматик: Бар-Хиллель, Перлес и Шамир [1961])**.** *Для всякого языка $L \subseteq \Sigma^*$, задаваемого грамматикой, существует такое число $p \geqslant 1$, что для всякой строки $w \in L$ длины не менее чем p ($|w| \geqslant p$) существует разбиение $w = xuyvz$, где $|uv| > 0$ и $|uyv| \leqslant p$, для которого выполняется $xu^i yv^i z \in L$ для всех $i \geqslant 0$.*

Доказательство. Пусть $G = (\Sigma, N, R, S)$ — грамматика, задающая язык L. Пусть $m = \max_{A \to \alpha \in R} |\alpha|$. Тогда число p определяется как $p = m^{|N|+1}$.

Пусть $w \in L$ — строка длины не менее чем p. Доказательство основано на анализе структуры дерева разбора w. Внутренняя вершина s в дереве называется *точкой ветвления*, если в этом месте листья разделяются не менее чем на две непустых группы; иными словами, дело *не* обстоит так, что у одного потомка s все листья, а у остальных ни одного.

В этом дереве строится путь из корня, на каждом шаге выбирается наибольшее поддерево данной вершины. Разделение гарантировано, коль скоро в текущем поддереве не менее чем m листьев. В каждой точке ветвления число листьев уменьшается не более чем в m раз, и потому, после прохождения ℓ точек ветвления, в текущем поддереве останется не менее $\frac{|w|}{m^\ell}$ листьев. Поскольку в строке не менее чем $m^{|N|+1}$ символов, можно проделать не менее чем $|N| + 1$ нетривиальных разбиений, и потому путь будет содержать не менее чем $|N| + 1$ точек ветвления.

Среди нижних $|N| + 1$ точек ветвления на этом пути где-то повторится дважды некоторая метка $A \in N$. На отрезке между этими двумя экземплярами A какие-то поддеревья ответвляются направо, какие-то — налево. Пусть u — строка листьев в левых поддеревьях, а v — строка листьев в правых. Поскольку на этом пути есть не менее одной точки ветвления (верхнее A — одна из этих точек), хотя бы одна из строк u и v должна быть непустой.

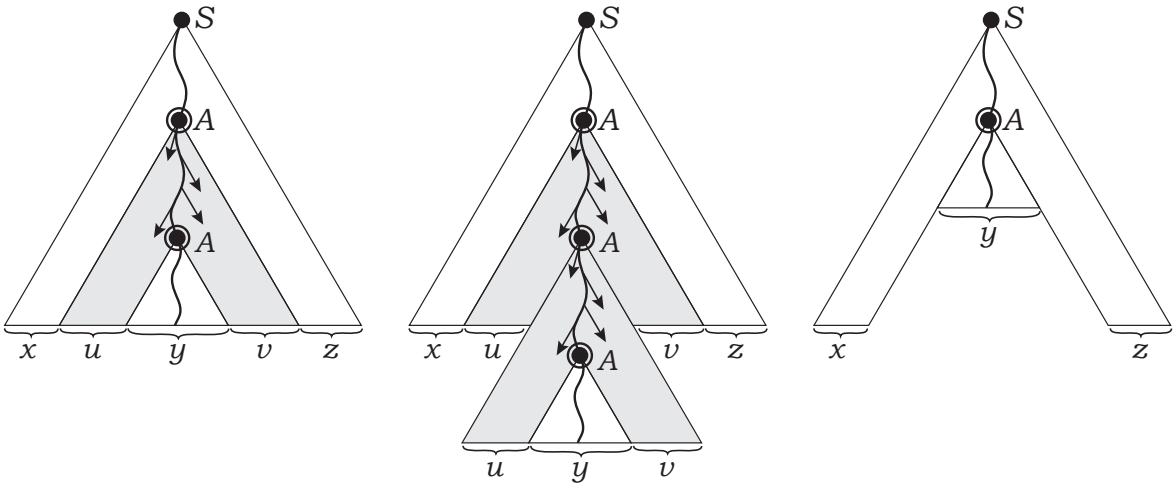

Рис. 5: (слева) Дерево разбора строки $w = xuyvz$ в лемме о накачке; (посередине) Накачанное дерево разбора xu^2yv^2z; (справа) Накачивание 0 раз: дерево разбора xu^0yv^0z.

Пусть $w = xuyvz$ — разбиение всей строки, в котором обозначены эти подстроки u и v. Такое разбиение показано на рис. 5(слева).

Участок дерева между двумя указанными экземплярами A можно повторить 0 и более раз, получая деревья разбора для строки xu^iyv^iz. Повторение 2 раза показано на рис. 5(посередине), а повторение 0 раз — на рис. 5(справа). □

Список литературы

[1961] Y. Bar-Hillel, M. Perles, E. Shamir, "On formal properties of simple phrase-structure grammars", *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung*, 14 (1961), 143–177.

[1962] S. Ginsburg, H. G. Rice, "Two families of languages related to ALGOL", *Journal of the ACM*, 9 (1962), 350–371.