ФОРМАЛЬНЫЙ ЯЗЫК

Формальный язык — **множество** правильно построенных конечных слов (строк, цепочек) над конечным алфавитом. Понятие языка используется в теории автоматов, теории алгоритмов.

Например, если алфавит задан как $\{a,b\}$, а язык L включает в себя все слова над ним, то слово $abba$ принадлежит L. «Пустое» слово или цепочка – это строка нулевой длины. Обозначается как e, ε. Алфавиту пустая цепочка не принадлежит.

Примеры формальных языков

- ✓ Множество однобуквенных цепочек длины $n = \# a^n = |a^n|$, включая пустую цепочку. Обозначается это множество как $\{a^n\}$. Здесь $n \geq 0$ — натуральное число, а «степень» n в выражении a^n означает, что a повторяется n раз подряд. Пример: $a^3 = aaa$.
- ✓ Множество лексически и синтаксически корректных программ языка программирования.

ОПЕРАЦИИ

Некоторые операции могут быть использованы для того, чтобы порождать новые языки из уже имеющихся языков. Предположим, что L_1 и L_2 являются языками, определёнными над некоторым общим алфавитом. Допустимые операции над языками таковы

1. Конкатенация (сцепление) L_1L_2 содержит все слова, удовлетворяющие форме vw, где v — это слово из языка L_1, а w — слово из языка L_2.
2. Объединение $L_1 \cup L_2$ — это новый язык, который содержит все слова, содержащиеся в L_1 или в L_2.
3. Замыкание Клини L_1^* содержит все слова, которые могут быть записаны в форме $w_1w_2...w_n$, где w_i содержится в L_1 и $n \geq 0$. Следует помнить, что это включает и пустое слово ε, так как $n = 0$ допустимо по условию.

Формальная грамматика в теории формальных языков — способ генерации формального языка, то есть выделения некоторого подмножества из множества всех слов некоторого конечного алфавита.

Порождающие грамматики задаются **правилами**, с помощью которых можно построить любое слово языка.

Автоматы-распознаватели позволяют по данному слову определить, входит ли оно в язык или нет.

ТЕРМИНЫ

Терминал (терминальный символ) — литерал, буква алфавита, символ непосредственно присутствующий в словах языка.

Нетерминал (нетерминальный символ) — лингвистическая переменная, не принадлежащая алфавиту.

ПОРОЖДАЮЩИЕ ГРАММАТИКИ

Словами языка, заданного грамматикой, являются все последовательности терминалов, выводимые (порождаемые) из начального нетерминала по правилам вывода.

Чтобы задать грамматику, требуется задать алфавиты терминалов и нетерминалов, набор правил вывода, а также выделить в множестве нетерминалов начальный.

Грамматика определяется следующими характеристиками:

- Σ — набор алфавит терминальных символов.
- N (греческая заглавная буква) — набор алфавит нетерминальных символов.
- P — набор продукций или правил вида: «левая часть» «$\rightarrow$» «правая часть», где: «левая часть» — непустая последовательность терминалов и нетерминалов из множества $(\mathrm{N} \cup \Sigma)^{*} \mathrm{N} (\mathrm{N} \cup \Sigma)^{*}$, содержащая хотя бы один нетерминал, «правая часть» — любая последовательность терминалов и нетерминалов из множества $(\mathrm{N} \cup \Sigma)^{*}$.
- $S \in \mathrm{N}$ — стартовый (или начальный) символ грамматики из набора нетерминалов.

ВЫВОД ЦЕПОЧКИ

Выводом называется последовательность строк, состоящих из терминалов и нетерминалов, где первой идет строка, состоящая из одного стартового нетерминала, а каждая последующая строка получена из предыдущей путём замены некоторой подстроки по одному (любому) из правил. Конечной строкой является строка, полностью состоящая из терминалов, и следовательно являющаяся словом языка.

Существование вывода для некоторого слова является критерием его принадлежности к языку, определяемому данной грамматикой.

ТИПЫ ГРАММАТИК

В иерархии Хомского[1], грамматики делятся на четыре типа, каждый последующий является более ограниченным подмножеством предыдущего (но и легче поддающимся анализу):

- неограниченные грамматики — возможны любые правила
- контекстно-зависимые грамматики — левая часть может содержать один нетерминал, окруженный «контекстом» (последовательности символов, в том же виде присутствующие в правой части); нетерминал левой части правила или продукции заменяется непустой последовательностью символов в правой части. Это «неукорачивающие» грамматики'. Каждое правило или продукции такой грамматики имеет вид $\alpha \to \beta$, где цепочки $\alpha \in (\mathrm{N} \bigcup \Sigma)^* \mathrm{N} (\mathrm{N} \bigcup \Sigma)^*$, $\beta \in (\mathrm{N} \bigcup \Sigma)^+$. Длина правой части правила не меньше длины левой. $|\alpha| \leqslant |\beta|$
- контекстно-свободные грамматики — левая часть состоит из одного нетерминала: $A \to \gamma$, где $\gamma \in (\mathrm{N} \bigcup \Sigma)^*$.
- регулярные грамматики — более простые, эквивалентны конечным автоматам. Каждое правило такой грамматики имеет вид $A \to uBv$, или $A \to u$, то есть в правой части правила может содержаться не более одного вхождения нетерминала.

ПРИМЕНЕНИЕ

Контекстно-свободные грамматики широко применяются для определения *синтаксической конструкций* языка, например, генерации арифметических выражений.

Регулярные грамматики в виде регулярных выражений широко применяются как шаблоны для контекстного поиска, разбивки и подстановки, в том числе в *лексическом анализе*.

Пример — разбор арифметического выражения.

Рассмотрим простой язык, определяющий ограниченное подмножество арифметических формул, состоящих из натуральных чисел, скобок и знаков арифметических действий. В каждом правиле грамматики с левой стороны от стрелки стоит только один нетерминальный символ. Такие грамматики называются контекстно-свободными.

Терминальный алфавит: Σ = {'0','1','2','3','4','5','6','7','8','9','+','-','*','/','(',')'}

Нетерминальный алфавит: $\mathrm{N} =$ { ФОРМУЛА, ЗНАК, ЧИСЛО, ЦИФРА }

Правила:

1. ФОРМУЛА $\to$ ФОРМУЛА ЗНАК ФОРМУЛА (формула есть две формулы, соединенные знаком)

2. ФОРМУЛА $\to$ ЧИСЛО (формула есть число)

[1] Аврам Ноам (Наум) Хомский (часто транскрибируется как Хомски или Чомски, англ. Avram Noam Chomsky; 7 декабря 1928, Филадельфия, штат Пенсильвания, США) - американский лингвист, политический публицист, философ и теоретик.

3. ФОРМУЛА $\rightarrow$ (ФОРМУЛА) (формула есть формула в скобках)

4. ЗНАК $\rightarrow$ + | - | * | / (знак есть плюс или минус, или умножить, или разделить)

5. ЧИСЛО $\rightarrow$ ЦИФРА (число есть цифра)

6. ЧИСЛО $\rightarrow$ ЧИСЛО ЦИФРА (число есть число и цифра)

7. ЦИФРА $\rightarrow$ 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 (цифра есть 0 или 1, или ... 9)

Начальный нетерминал: ФОРМУЛА

ВЫВОД ЦЕПОЧКИ:

Разберем формулу (12+5) с помощью перечисленных правил вывода (номера применяемых правил написаны сверху над стрелкой).

$\text{ФОРМУЛА} \xrightarrow{3} (\text{ФОРМУЛА})$

(ФОРМУЛА) $\xrightarrow{1}$ (ФОРМУЛА ЗНАК ФОРМУЛА)

(ФОРМУЛА ЗНАК ФОРМУЛА) $\xrightarrow{4}$ (ФОРМУЛА + ФОРМУЛА)

(ФОРМУЛА + ФОРМУЛА) $\xrightarrow{2}$ (ФОРМУЛА + ЧИСЛО)

(ФОРМУЛА + ЧИСЛО) $\xrightarrow{5}$ (ФОРМУЛА + ЦИФРА)

(ФОРМУЛА + ЦИФРА) $\xrightarrow{7}$ (ФОРМУЛА + 5)

(ФОРМУЛА + 5) $\xrightarrow{2}$ (ЧИСЛО + 5)

(ЧИСЛО + 5) $\xrightarrow{6}$ (ЧИСЛО ЦИФРА + 5)

(ЧИСЛО ЦИФРА + 5) $\xrightarrow{5}$ (ЦИФРА ЦИФРА + 5)

(ЦИФРА ЦИФРА + 5) $\xrightarrow{7}$ (1 ЦИФРА + 5)

(1 ЦИФРА + 5) $\xrightarrow{7}$ (1 2 + 5)