Введение

Языки бывают:

- **Естественными** (“обычные”, “разговорные”).
- **Искусственными.**
- **Формальными.**

Введение

- Любой язык – и естественный и искусственный – обладает набором определенных правил. Они могут быть явно и строго сформулированными (формализованными), а могут допускать различные варианты их использования.
- **Формализованный (формальный)** язык – язык, характеризующийся точными правилами построения выражений и их понимания. Он строится в соответствии с четкими правилами, обеспечивая непротиворечивое, точное и компактное отображение свойств и отношений изучаемой предметной области (моделируемых объектов).

Введение

В математической логике и информатике **формальный язык** — это множество конечных слов (строк, цепочек) над конечным алфавитом. Понятие языка чаще всего используется в теории автоматов и теории алгоритмов. Научная теория, которая имеет дело с этим объектом, называется **теорией формальных языков**.

Введение

Формальный язык может быть определён по-разному, например:

- Простым перечислением слов, входящих в данный язык. Этот способ, в основном, применим для определения конечных языков и языков простой структуры.
- Словами, порождёнными некоторой формальной грамматикой (по иерархии Хомского).
- Словами, порождёнными регулярным выражением.
- Словами, распознаваемыми некоторым конечным автоматом и т.д..

Введение

Формальные языки проще естественных языков. Они создаются одновременно с системой правил построения слов и предложений.

Исключения из правил в формальном языке могут свидетельствовать лишь о противоречивости и некорректности системы грамматических правил.

У формального языка много общего с естественным языком, предложения которого также строятся в соответствии с грамматическими правилами.

Формальные грамматики. Основные понятия

Алфавит — конечное непустое множество символов или букв. Обозначается латинской буквой ***V***. Его элементы называются символами (буквами).

Формальные грамматики. Основные понятия

- **Терминал (терминальный символ)** — объект, непосредственно присутствующий в словах языка, соответствующего грамматике, и имеющий конкретное, неизменяемое значение (обобщение понятия «буквы»).

В формальных языках, используемых на компьютере, в качестве терминалов обычно берут все или часть стандартных символов ASCII — латинские буквы, цифры и специальные символы.

Формальные грамматики. Основные понятия

- **Нетерминал (нетерминальный символ)** — объект, обозначающий какую-либо *сущность* языка (например: формула, арифметическое выражение, команда) и не имеющий конкретного символьного значения.

Формальные грамматики. Основные понятия

Все символы из нетерминального алфавита обозначаются заглавными буквами, а из терминального — строчными буквами.

Формальные грамматики. Основные понятия

Формальная грамматика или просто **грамматика** в теории формальных языков это способ описания формального языка, на основе некоторого конечного алфавита.

Грамматика это объект, состоящий из четвёрки $G=(V_T, V_N, P, S)$, $V_N \cap V_T = \emptyset$ и определяется следующими характеристиками:

V_T - терминальный алфавит, состоящий из терминальных символов.
V_N - нетерминальный или вспомогательный алфавит, состоящий из нетерминальных символов.
P - множество правил или продукций. Любое правило имеет общий вид:

«левая часть» → «правая часть»

где левая часть - непустая последовательность терминалов или нетерминалов, содержащая хотя бы один нетерминал и «правая часть» - любая последовательность терминалов и нетерминалов.
S - стартовый (начальный) символ из набора нетерминалов. Называется аксиомой.

Формальные грамматики. Основные понятия

G=(V_T ,V_N , *P, S*)
V_T={*a* , *b* }, V_N={*S, A, B*}
P={1. S → a
2. S → aA
3. S → bS
4. A → aBb
5. B → b}

Формальные грамматики. Основные понятия

- **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 путем замены некоторой подстроки по одному (любому) из правил.

- **Конечной строкой** является строка, полностью состоящая из терминалов, и следовательно являющаяся **словом** языка.

Формальные грамматики. Основные понятия

G=(V_T ,V_N , *P, S*)
V_T={*a , b* }, V_N={*S, A, B*}
P={1. S → a, 2. S → aA, 3. S → bS, 4. A → aBb 5. B → b}
Вывод данной грамматики:
S ⇒bS ⇒baA ⇒baaBb ⇒baabb

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

Выводам соответствуют так называемые **деревья вывода** - деревья, вершины которых помечены символами алфавита.

…
.

Формальные грамматики. Основные понятия

Слово, не содержащее ни одного символа, называется **пустым словом**. Обозначается при помощи ε (маленького епсилон).

Длина слова ***x***, обозначаемое |***x***|, это число символов в ***x***, причём каждый символ считается столько раз, сколько раз он встречается в данном слове. |ε|=0

Формальные грамматики. Основные понятия

Если ***x*** и ***y*** - слова в алфавите ***V***, то слово ***xy*** - результат приписывания ***y*** в конец ***x***, называется **конкатенацией**. Иногда конкатенацию обозначают как ***x*·*y***.

Если ***V*** — алфавит и ***a*∈*V***, то обозначим ***n*** раз конкатенацию ***a*·*a*·*a*·…·*a*=*a*^{n}**

Формальные грамматики. Основные понятия

Язык в алфавите V - это некоторое множество цепочек в алфавите V.

Пусть дан алфавит V = {a, b}. Вот некоторые языки в алфавите V :

- L_1 = Ø - пустой язык;
- L_2 = {ε} - язык, содержащий только пустую цепочку (заметим, что L_1 и L_2 - различные языки);
- L_3 = { ε, a, b, aa, ab, ba, bb} - язык, содержащий цепочки из a и b, длина которых не превосходит 2;
- L_4 - {aabb, abab, bbaa, baba, abba, baab, aabbbb, aaaabb, aaaabbbb …}-язык, включающий всевозможные цепочки из a и b, содержащие четное число a и четное число b;

Формальные грамматики. Основные понятия

Обозначим:

V^0={ε}

V^+= V \ V^0

V^* = V +V^0

И так, назовём **формальным языком** над алфавитом ***V*** любое подмножество ***L*** ∈V^*.

Называется **формальным языком**, порождаемый грамматикой ***G*** множество **L(G)={x| S⇒ x, x∈V_T^*}**

Формальные грамматики. Основные понятия

Языки могут быть пустыми или нулевыми, конечными, если конечное число слов, и бесконечными.

Две грамматики эквивалентны, если они порождают один и тот же язык.

Классификация Хомского

По иерархии Хомского, грамматики делятся на 4 типа, каждый последующий является более ограниченным подмножеством предыдущего (но и легче поддающимся анализу):

- тип 0: **неограниченные грамматики** - возможны любые правила, любого вида $-\alpha \rightarrow \beta$.
- тип 1: **контекстно-зависимые грамматики** - левая часть правил может содержать один нетерминал, окруженный «контекстом» (последовательности символов, в том же виде присутствующие в правой части); сам нетерминал заменяется непустой последовательностью символов в правой части.
- тип 2: **контекстно-свободные грамматики** - левая часть состоит из одного нетерминала. Правая часть это любая последовательность терминалов и/или нетерминалов.
- тип 3: **регулярные грамматики** - более простые, эквивалентны конечным автоматам. Грамматики и языки, порождаемые ими, называются регулярными или автоматными.
  - **0-й тип ⊃1-й тип ⊃2-й тип ⊃3-й тип**

Применение

- Контекстно-свободные грамматики широко применяются для определения грамматической структуры в грамматическом анализе.
- Регулярные грамматики (в виде регулярных выражений) широко применяются как шаблоны для текстового поиска, разбивки и подстановки, в том числе в лексическом анализе.