Теория формальных языков и грамматик — это обширная область математики, примыкающая к алгебре, математической логике и теории автоматов.

Введем в обиход основные понятия, которые будут использованы в определении формального языка.

***Алфавит*** — конечное непустое множество символов.

Термин символ следует понимать здесь в самом широком смысле. Это может быть буква, цифра или знак препинания. Но символом можно считать и любой другой знак, рассматриваемый как нечто неделимое — служебное слово языка программирования, иероглиф и т. д. Будем обозначать алфавиты буквой Σ (сигма).

Примеры алфавитов:

Σ1 = {0, 1}

Σ2 = {a, b, с}

Σ3 = {A, В, С,..., Z, а, b, с,..., z, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, +, -, \*,..., div, ..., program }

Алфавит — это множество, поэтому при перечислении его элементов использованы фигурные скобки, как это принято в математике. Алфавит Σ1 содержит два символа, алфавит Σ2 — три. Под Σ3 подразумевается алфавит языка Паскаль.

***Цепочка*** над алфавитом Σ — произвольная конечная последовательность символов из Σ.

Примеры цепочек над алфавитом Σ2:

**α** = аbbса

**β** = ab

**γ** = ba

**δ** =с

Цепочки будем обозначать греческими буквами.

***Пустая цепочка*** — цепочка, не содержащая символов {содержащая ноль символов). Обозначается буквой **ε**.

Если **α** и **β** — цепочки, то запись **αβ** означает их конкатенацию (склеивание), то есть **αβ** — это цепочка, образованная приписыванием к цепочке **α** цепочки **β** справа.

Если **α** — цепочка, то **α**n означает цепочку, образованную n-кратным повторением цепочки **α**:

**α**n = **α α** ... **α** **α**.

В частном случае, если а — символ, то аn = а а ... а а.

Будем обозначать **Σ\*** — (бесконечное) множество всех цепочек над алфавитом Σ, включая пустую цепочку; **Σ+** — множество всех цепочек над алфавитом Σ, не включая пустой цепочки. Например, если Σ1 = {0, 1}, то Σ1\* представляет собой множество всех цепочек, которые могут быть составлены из символов 0 и 1. В это множество входят пустая цепочка, все цепочки, состоящие из одного символа, все цепочки, состоящие из двух символов, и т. д.: Σ1\* = { **ε**, 0,1, 00, 01,10,11, 000, 001,...}.

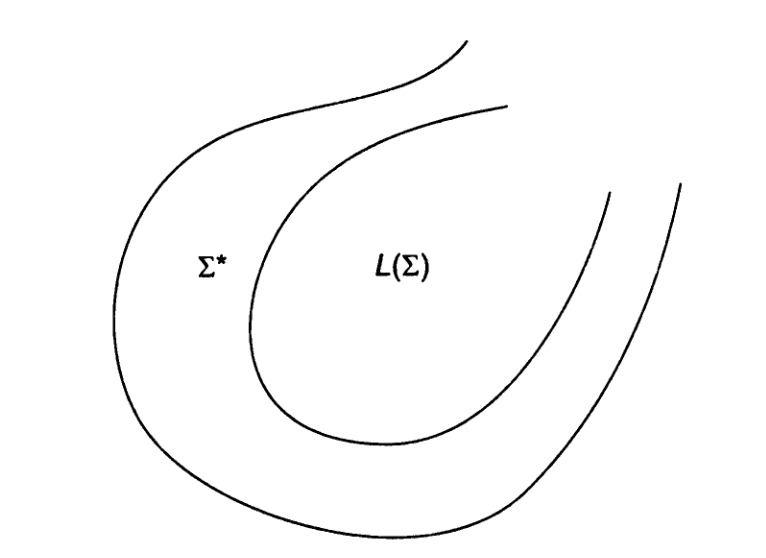
Σ\* = Σ+U { **ε** }, где U — знак операции объединения множеств.

Теперь можно дать и определение формального языка.

***Языком*** над алфавитом Σ называется произвольное множество цепочек, составленных из символов Σ.

Будем обозначать язык над алфавитом (с алфавитом) Σ — L(Σ) или просто L, если алфавит ясен из контекста.

Таким образом, речь идет о том, что язык — это некоторое, тем или иным образом определенное, подмножество множества всех цепочек, которые могут быть построены из символов данного алфавита. L(Σ) ⊆ Σ\*



Язык – подмножество множеств всех цепочек над алфавитом Σ

Принадлежащие языку цепочки называют также *предложениями* языка.

Еще раз отметим, что множество цепочек Σ\* всегда бесконечно, в то время как множество цепочек, образующих язык, может быть и конечным. Практический интерес представляют, конечно, языки, содержащие бесконечное множество цепочек; к числу таких языков относятся и языки программирования.

**Примеры языков**

***Пример 1*.** Определим язык

L1 = {anbn | n ≥ 0}, используя принятую в теории множеств нотацию, как множество всех цепочек, содержащих вначале некоторое количество символов а, а затем такое же количество символов b. Заметим, что L1 включает и пустую цепочку, поскольку n может равняться нулю.

Записанное ранее правило, определяющее язык L1 разделяет все цепочки над алфавитом {а, b}, то есть состоящие из символов а и b, на принадлежащие L1 и не принадлежащие ему.

Примеры цепочек, принадлежащих языку:

**ε** ∈ L1 — пустая цепочка принадлежит L1;

ab ∈ L1 — цепочка из одной буквы а, за которой следует b;

aaabbb ∈ L1.

Цепочки, не принадлежащие языку L1.

aaab ∉ L1 — неодинаковое количество символов а и b

abba ∉ L1 — порядок следования символов не соответствует определению L1

***Пример 2.*** Язык L2 = {anbncn | n ≥ 0} — множество всех цепочек, содержащих вначале некоторое (возможно нулевое) количество символов а, затем такое же количество символов b, затем — столько же символов с.

Например,

aaabbbccc ∈ L2,

в то время как

aaabbccc ∉ L2.

Далеко не всегда удается определить язык, особенно если речь идет о языках, представляющих практический интерес, используя нотацию, примененную при определении L1 и L2. Значительная часть последующего материала будет посвящена рассмотрению порождающих грамматик, позволяющих компактно и однозначно определить обширный класс формальных языков. Пока же дадим словесное описание некоторых представляющих интерес языков в следующих примерах.

***Пример 3.*** Рассмотрим язык правильных скобочных выражений, составленных только из круглых скобок, известный также как язык Дика. Обозначим его L3.

Алфавит языка Дика — это множество из двух символов — открывающей «(» и закрывающей «)» скобок: Σ = { (, ) }. Цепочки, содержащие правильно расставленные скобки, принадлежат языку Дика, все остальные последовательности открывающих и закрывающих круглых скобок — нет.

Например:

(())()() ∈ L3;

()(()))( ∉ L3.

***Пример 4***. Язык L4 — множество всех цепочек, содержащих одинаковое количество символов а и b. Несмотря на простое «устройство», задать язык L4 формулой, подобной формулам для L1, или L2, оказывается затруднительно.

Можно заметить, что рассмотренный ранее язык L1 является подмножеством языка L4: L1 ⊆ L4, поскольку любая цепочка, принадлежащая L1, принадлежит и языку L4. Но не наоборот. Так,

aabb ∈ L1, aabb ∈ L4,

abba ∈ L4, но abba ∉ L1.

***Пример 5.*** В качестве языка L5 рассмотрим множество всех правильных арифметических выражений языка Паскаль, составленных из символов алфавита Σ5 = {а, b, с, +, -, \*, /, (, ) }.

Например,

а\*(b + с) ∈ L5,

но

c++ ∉ L5.