Рыбин Сергей Витальевич

ЯЗЫКИ И ГРАММАТИКИ

1. ВВЕДЕНИЕ

Из школьного опыта хорошо известно, что представляет собой грамматический разбор предложения. При таком разборе определяется, какое слово является подлежащим, какое используется в роли сказуемого, какие слова играют роль определения, дополнения, обстоятельства и т. д. При разборе мы имеем дело с грамматическими категориями: предложение, группа существительного, группа сказуемого, существительное, глагол, наречие и т. д. При этом используются слова, составляющие предложение. В качестве примера рассмотрим грамматический разбор предложения:

«Маленький Саша учится хорошо».

Грамматический разбор подразумевает использование правил некоторой грамматики. Эти правила будем представлять в следующем виде (см. рис. 1), где символ $\rightarrow$ означает «можно заменить на». Разбором нашего предложения будет следующая последовательность шагов (см. рис. 2).

... что представляет собой грамматический разбор предложения...

Впервые понятие грамматики было формализовано лингвистами при изучении естественных языков. Предполагалось, что это может помочь при их автоматической трансляции.

На интуитивном уровне язык можно определить как некоторое множество пред-

$\langle$предложение$\rangle \rightarrow \langle$*группа существительного*$\rangle\langle$*группа сказуемого*$\rangle$
$\langle$*группа существительного*$\rangle \rightarrow \langle$*прилагательное*$\rangle\langle$*существительное*$\rangle$
$\langle$*группа сказуемого*$\rangle \rightarrow \langle$*глагол*$\rangle\langle$*наречие*$\rangle$
$\langle$*прилагательное*$\rangle \rightarrow$ *маленький*
$\langle$*существительное*$\rangle \rightarrow$ *Саша*
$\langle$*глагол*$\rangle \rightarrow$ *учится*
$\langle$*наречие*$\rangle \rightarrow$ *хорошо*

Рис. 1

$\langle$предложение$\rangle$
$\langle$*группа существительного*$\rangle\langle$*группа сказуемого*$\rangle$
$\langle$*прилагательное*$\rangle\langle$*существительное*$\rangle\langle$*группа сказуемого*$\rangle$
Маленький $\langle$*существительное*$\rangle\langle$*группа сказуемого*$\rangle$
Маленький Саша $\langle$*группа сказуемого*$\rangle$
Маленький Саша $\langle$*глагол*$\rangle\langle$*наречие*$\rangle$
Маленький Саша учится $\langle$*наречие*$\rangle$
Маленький Саша учится хорошо

Рис. 2

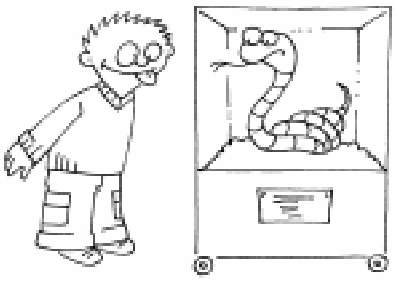
ложений допустимой структуры. Правила, определяющие допустимость той или иной конструкции языка, составляют синтаксис языка. Эти рассуждения можно отнести и к искусственным языкам программирования. Именно применительно к ним были достигнуты наилучшие результаты в этом направлении.

Если бы все языки состояли из небольшого конечного числа предложений, то проблемы синтаксического анализа не существовало. Можно было бы просто перечислить все предложения данного языка. Но так как большинство языков содержит неограниченное (или достаточно большое) число правильно построенных предложений, то возникает проблема разработки формальных средств описания и анализа языка (формальных грамматик).

Для формального описания языка необходимо задать набор символов (алфавит)

Ноам Хомский, американский лингвист и общественный деятель, родился в Филадельфии 7 декабря 1928 г. Его отец эмигрировал из России незадолго до начала Первой мировой войны. С 1945 г. Хомский изучал в Пенсильванском университете лингвистику, математику и философию. Там же, в 1955 г. получил докторскую степень, однако большая часть исследований, положенных в основу его первой монографии, была выполнена в Гарвардском университете в 1951–1955 г.

Ноам Хомский – создатель системы грамматического описания, известной как *генеративная (порождающая) грамматика*; соответствующее направление лингвистики часто называют генеративизмом. Уже несколько десятилетий он является самым знаменитым лингвистом мира; ему и его теории посвящено множество статей и монографий, а развитие лингвистики в последней трети XX в. иногда называют «хомскианской революцией». Работы Хомского оказали влияние не только на лингвистику, но и инициировали развитие новых идей в философии, психологии, музыке. Например, Леонард Бернстайн использовал теорию Хомского для анализа музыкального ряда. Хомский также входит в первую десятку самых цитируемых в научном мире авторов. За резкость и бескомпромиссность суждений его часто называют «Че Гевара в лингвистике».



Правила, определяющие допустимость той или иной конструкции языка...

языка и правила, по которым из символов строятся их последовательности (предложения), принадлежащие данному языку.

2. ФОРМАЛЬНОЕ ОПРЕДЕЛЕНИЕ ГРАММАТИКИ

В рассмотренном выше примере конкретной грамматики имелись:

1) грамматические термины (*⟨группа существительного⟩*, *⟨группа сказуемого⟩* и т. д.), которые в теории формальных грамматик принято называть *нетерминальными символами* или *нетерминалами*;

2) слова, составляющие предложения языка, – они называются *терминальными символами, терминалами* или *примитивами*;

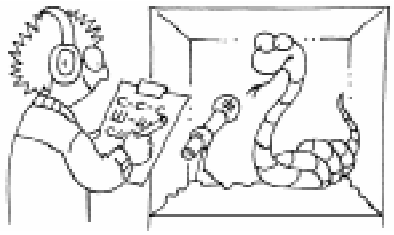
3) правила замены, левые и правые части которых состоят из терминальных и нетерминальных символов;

4) главный грамматический термин, который называют *начальным символом, аксиомой* или *целью* грамматики, с которого начинается разбор (вывод) любого предложения языка (в нашем примере таким символом является термин *⟨предложение⟩*).

Перейдем к формальным определениям. Основные идеи формальных грамматик были заложены в работах Н. Хомского.

Определение

Конечное множество символов $A = \{a_1, a_2, \ldots, a_n\}$ называют *алфавитом (словарем)*, а сами символы a_i – *буквами*. *Цепочкой* в алфавите A называют последовательность символов из этого алфавита. Пустую цепочку принято обозначать символом ε. Множество всех цепочек алфа-



Для формального описания языка необходимо задать набор символов ...

вита A обозначают A^*. Символом $A^+ \subset A^*$ будем обозначать множество, содержащее все непустые цепочки конечной длины в алфавите A.

Длиной цепочки $\alpha \in A^*$ называют число составляющих ее символов и обозначают $|\alpha|$. Например, если $\alpha = abcdabcd$, то $|\alpha| = 8$. Длина пустой цепочки ε по определению равна нулю.

Определение

Грамматикой G называют четверку объектов $\{V_T, V_N, P, S\}$, где

1) V_T – алфавит терминальных символов;

2) V_N – алфавит нетерминальных символов, причем $V_T \cap V_N = \varnothing$;

3) Множество P – конечное множество *правил (продукций)*, каждое из которых имеет вид:

$$\alpha \to \beta, \quad \alpha \in (V_T \cup V_N)^+, \ \beta \in (V_T \cup V_N)^*;$$

4) S – начальный символ грамматики, $S \in V_N$.

Замечание

1. Для того чтобы различать терминальные и нетерминальные символы, принято обозначать терминальные символы строчными, а нетерминальные символы прописными буквами латинского алфавита.

2. Для записи нескольких правил с одинаковыми левыми частями

$$\alpha \to \beta_1, \alpha \to \beta_2, \ldots, \alpha \to \beta_n$$

часто используют сокращенную запись

$$\alpha \to \beta_1 \,|\, \beta_2 \,|\ldots|\, \beta_n.$$

3. Иногда при описании грамматики словари терминальных и нетерминальных символов не указывают. В таком случае обычно предполагается, что грамматика содержит только те терминальные и нетерминальные символы, которые встречаются в правилах вывода.

Мы представили грамматику, но не определили язык, который она порождает. Для этого нам потребуется ввести понятие вывода.

Определение

Цепочка $\beta \in (V_T \cup V_N)^*$ *непосредственно выводима* из цепочки $\alpha \in (V_T \cup V_N)^+$ в грамматике $G = \{V_T, V_N, P, S\}$ (обозначают $\alpha \to \beta$), если

$$\alpha = \zeta_1 \gamma \zeta_2, \ \beta = \zeta_1 \delta \zeta_2, \quad \zeta_1, \zeta_2 \delta \in (V_T \cup V_N)^*,$$
$$\gamma \in (V_T \cup V_N)^+$$

и правило вывода $\gamma \to \delta$ содержится во множестве P.

Пример 1

Рассмотрим грамматику G_1, где множество правил вывода имеет вид:

$$S \to aAb, \ \ aA \to aaAb, \ \ A \to \varepsilon \ .$$

Цепочка $aaAbb$ непосредственно выводима из цепочки aAb применением второго правила.

Определение

Цепочка $\beta \in (V_T \cup V_N)^*$ *выводима* из цепочки $\alpha \in (V_T \cup V_N)^+$ в грамматике $G = \{V_T, V_N, P, S\}$ (обозначают $\alpha \Rightarrow \beta$) если существуют последовательность цепочек $\gamma_0, \gamma_1, \ldots, \gamma_n$, такая, что

$$\alpha = \gamma_0 \to \gamma_1 \to \cdots \to \gamma_{n-1} \to \gamma_n = \beta.$$

Последовательность $\gamma_0, \gamma_1, \ldots, \gamma_n$ называют *выводом* длины n.

Пример 2

В грамматике G_1 из примера 1 $S \Rightarrow aaaAbbb$, так как существует вывод

$$S \to aAb \to aaAbb \to aaaAbbb,$$

при этом длина вывода равна 3.

Определение

Языком, порождаемым грамматикой G, называют множество

$$L(G) = \{\alpha \in V_T \,|\, S \Rightarrow \alpha\},$$

то есть $L(G)$ – это все цепочки в алфавите

V_T, которые выводимы из начального символа S с помощью правил вывода P.

Такие цепочки языка называют *сентенциальными формами*.

Пример 3

Рассмотрим грамматику G_2, где множество правил вывода имеет вид:

$$S \to aSb, \quad S \to ab \ .$$

При применении первого правила всегда остается один нетерминальный символ. После второго правила получаем сентенциальную форму. Таким образом, очевиден единственный порядок применения – несколько раз использовать первое правило, а затем один раз применить второе. Тогда

$$L(G_2) = \left\{ \underbrace{aa\ldots a}_{n} \underbrace{bb\ldots b}_{n} \mid n > 0 \right\} = \left\{ a^n b^n \mid n > 0 \right\}.$$

Задача 1

Покажите, что для грамматики G_1 из примера 1 $L(G_1) = L(G_2)$.

Определение

Грамматики G_1 и G_2, порождающие один и тот же язык ($L(G_1) = L(G_2)$), называют *эквивалентными*.

Приведенные примеры грамматик очень просты: было почти очевидно, какие цепочки в них можно вывести. В общем случае определить, что же порождается грамматикой, бывает очень трудно. Дадим более сложный пример.

Задача 2

Покажите, что грамматика G, определяемая следующими правилами вывода:

$$S \to aSBC,\ S \to aBC,\ CB \to BC,\ aB \to ab,$$
$$bB \to bb,\ bC \to bc,\ cC \to cc,$$

порождает язык

$$L(G) = \left\{ a^n b^n c^n \mid n > 0 \right\}.$$

3. КЛАССИФИКАЦИЯ ГРАММАТИК

Накладывая различные ограничения на правила вывода, можно выделить классы грамматик. Рассмотрим классификацию, предложенную Н. Хомским.

• Тип G_0 (**свободные или неограниченные грамматики**). Грамматику G называют *грамматикой типа 0*, если на ее правила вывода не накладывается никаких ограничений.

• Тип G_1 (**контекстно-зависимые, контекстные или НС-грамматики**). Грамматику G называют *грамматикой типа 1*, если для каждого ее правила

$$\alpha \to \beta, \quad |\beta| \geq |\alpha|.$$

Часто, вместо термина «контекстно-зависимая грамматика», употребляют сокращение csg (context-sensitive grammar).

Пример 4

Очевидно, что грамматики из примеров 2 и 3, а также задачи 2 являются контекстно-зависимыми, поскольку правые части их правил не короче левых частей. Заметим, что для грамматики из примера 1 этому условию не удовлетворяет последнее правило:

$$A \to \varepsilon \ .$$

• Тип G_2 (**контекстно-свободные или КС-грамматики**). Грамматику G называют *грамматикой типа 2*, если каждое ее правило имеет вид:

$$A \to \beta, \quad A \in V_N, \ \beta \in (V_N \cup V_T)^+.$$

Вместо термина «контекстно-свободная грамматика» часто используют аббревиатуру cfg (context-free grammar).

Замечание.

Правило вида $A \to \beta$ позволяет заменить A на β *независимо от контекста*, в котором появляется A.

Пример 5

1. Грамматика G_2 из примера 3 является контекстно-свободной грамматикой.

2. Грамматика G, где множество правил имеет вид

$$S \to aAb \mid accb,\ A \to cSc,$$

является контекстно-свободной грамматикой. Грамматика порождает язык

$$L(G) = \left\{ (ac)^n (cb)^n \mid n > 0 \right\}.$$

• Тип G_3 (**автоматные или регулярные грамматики**). Грамматику G называют *грам-*

матикой типа 3, если каждое ее правило имеет вид:

$$A \to aB \mid a, \quad a \in V_T, A, B \in V_N.$$

Пример 6

Грамматика G, где множество правил имеет вид

$$S \to aA \mid bB, \quad A \to aA \mid a, \quad B \to b,$$

является регулярной грамматикой. Грамматика G порождает язык

$$L(G) = \{a^n b^2 \mid n > 1\}.$$

Классы грамматик типа G_0, G_1, G_2 и G_3 образуют *иерархию Хомского*.

Определение

Язык называют *контекстным* (*контекстно-свободным, регулярным*), если он порождается некоторой контекстной (контекстно-свободной, регулярной) грамматикой. Контекстно-свободные языки называют также *алгебраическими* языками.

Замечание

Двигаясь от грамматики G_0 к грамматике G_3 можно заметить, что в то время как ограничения на правила вывода усиливаются, описательные возможности языков уменьшаются.

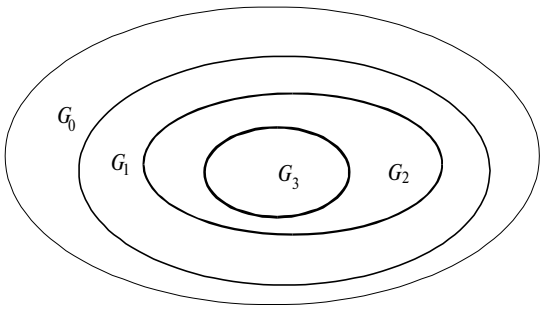


Рис. 1

Для грамматик справедливо следующее очевидное утверждение (см. рис. 1):

1. Любая регулярная грамматика является контекстно-свободной грамматикой.
2. Любая контекстно-свободная грамматика является контекстно-зависимой грамматикой.
3. Любая контекстно-зависимая грамматика является грамматикой типа 0.

Аналогичными свойствами обладают и языки, описываемые этими грамматиками:

1. Каждый регулярный язык является контекстно-свободным языком, но существуют контекстно-свободные языки, которые не являются регулярными (например,

$$L = \{a^n b^n \mid n > 0\}.$$

2. Каждый контекстно-свободный язык является контекстно-зависимым языком, но существуют контекстно-зависимые языки, которые не являются контекстно-свободным языками, например

$$L = \{a^n b^n c^n \mid n > 0\}.$$

3. Каждый контекстно-зависимый язык является языком типа 0.

Заметим, что если язык задан грамматикой типа m, то это не значит, что не существует грамматики типа $m_1 > m$, описывающей тот же язык. Поэтому, когда говорят о языке типа m, обычно имеют в виду максимально возможный номер m.

Рыбин Сергей Витальевич, кандидат физико-математических наук, доцент кафедры ВМ-2 СПбГЭТУ «ЛЭТИ».