

Классификация грамматик и языков по Хомскому

грамматики классифицируются по виду их правил вывода

Четыре типа грамматик:

тип 0, тип 1, тип 2, тип 3

Язык, порождаемый грамматикой типа k ($k=0,1,2,3$), является языком *типа k* .

$$G = \langle T, N, P, S \rangle$$

Тип 0

Любая порождающая грамматика является грамматикой *типа 0*.

На вид правил грамматик этого типа не накладывается никаких дополнительных ограничений.

Класс языков типа 0 совпадает с классом рекурсивно перечислимых языков (распознаваемых МТ).

Грамматика с ограничениями на вид правил вывода

Тип 1

Грамматика $G = \langle T, N, P, S \rangle$ называется *неукорачивающей*, если левая часть каждого правила из P не короче правой части (т. е. для любого правила $\alpha \rightarrow \beta \in P$ выполняется неравенство $|\alpha| \leq |\beta|$).

В виде исключения в неукорачивающей грамматике допускается наличие правила $S \rightarrow \varepsilon$, при условии, что S (начальный символ) не встречается в правых частях правил.

Грамматикой *типа 1* будем называть неукорачивающую грамматику.

Тип 1 в некоторых источниках определяют с помощью так называемых контекстно-зависимых грамматик.

Грамматика $G = \langle T, N, P, S \rangle$ называется *контекстно-зависимой* (КЗ), если каждое правило из P имеет вид $\alpha \rightarrow \beta$, где $\alpha = \xi_1 A \xi_2$, $\beta = \xi_1 \gamma \xi_2$, $A \in N$, $\gamma \in (T \cup N)^+$, $\xi_1, \xi_2 \in (T \cup N)^*$.

В виде исключения в КЗ-грамматике допускается наличие правила $S \rightarrow \varepsilon$, при условии, что S (начальный символ) не встречается в правых частях правил.

КЗ-грамматика удовлетворяет определению неукорачивающей.

Неукорачивающие и КЗ-грамматики определяют один и тот же класс языков.

Тип 2

Грамматика $G = \langle T, N, P, S \rangle$ называется *контекстно-свободной (КС)*, если каждое правило из P имеет вид $A \rightarrow \beta$, где $A \in N, \beta \in (T \cup N)^*$.

В КС-грамматиках допускаются правила с пустыми правыми частями.

Язык, порождаемый контекстно-свободной грамматикой, называется *контекстно-свободным языком*.

Грамматикой *типа 2* будем называть контекстно-свободную грамматику.

Любую КС-грамматику можно преобразовать в эквивалентную неукорачивающую КС-грамматику. (т.е. КС, удовлетворяющую также и определению неукорачивающей)

Тип 3

Грамматика $G = \langle T, N, P, S \rangle$ называется *праволинейной*, если каждое правило из P имеет вид $A \rightarrow wB$ либо $A \rightarrow w$, где $A \in N, B \in N, w \in T^*$.

Грамматика $G = \langle T, N, P, S \rangle$ называется *леволинейной*, если каждое правило из P имеет вид $A \rightarrow Bw$ либо $A \rightarrow w$, где $A \in N, B \in N, w \in T^*$.

Праволинейные и левوليнейные грамматики определяют один и тот же класс языков. Такие языки называются *регулярными*. Право- и левوليнейные грамматики тоже называют регулярными.

Регулярная грамматика является грамматикой *типа 3*.

Автоматной грамматикой называется праволинейная (левوليнейная) грамматика, такая, что каждое правило с непустой правой частью имеет вид: $A \rightarrow a$ либо $A \rightarrow aB$ (для левوليнейной, соответственно, $A \rightarrow a$ либо $A \rightarrow Ba$), где $A \in N$, $B \in N$, $a \in T$.

Для любой регулярной (автоматной) грамматики G существует неукорачивающая регулярная (автоматная) грамматика G' , такая что $L(G) = L(G')$.

Праволинейные и левوليнейные грамматики определяют один и тот же класс языков. Такие языки называются *регулярными*. Право- и левوليнейные грамматики тоже называют регулярными.

Регулярная грамматика является грамматикой *типа 3*.

Автоматной грамматикой называется праволинейная (левوليнейная) грамматика, такая, что каждое правило с непустой правой частью имеет вид: $A \rightarrow a$ либо $A \rightarrow aB$ (для левوليнейной, соответственно, $A \rightarrow a$ либо $A \rightarrow Ba$), где $A \in N$, $B \in N$, $a \in T$.

Для любой регулярной (автоматной) грамматики G существует неукорачивающая регулярная (автоматная) грамматика G' , такая что $L(G) = L(G')$.

Иерархия Хомского

Справедливы следующие соотношения:

- 1) любая регулярная грамматика является КС-грамматикой;
- 2) любая неукорачивающая КС-грамматика является КЗ-грамматикой;
- 3) любая неукорачивающая грамматика является грамматикой типа 0.

Неукорачивающие Регулярные \subset Неукорачивающие КС \subset КЗ \subset Тип 0

(Запись $A \subset B$ означает, что A является собственным подклассом класса B)

Справедливы следующие соотношения для языков:

- каждый регулярный язык является КС-языком, но существуют КС-языки, которые не являются регулярными, например:

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

- каждый КС-язык является КЗ-языком, но существуют КЗ-языки, которые не являются КС-языками, например:

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

- каждый КЗ-язык является языком типа 0 (т. е. рекурсивно перечислимым языком), но существуют языки типа 0, которые не являются КЗ-языками, например: язык, состоящий из записей самоприменимых алгоритмов Маркова в некотором алфавите.

Иерархия классов языков



Тип 3 (Регулярные) \subset Тип 2 (КС) \subset Тип 1 (КЗ) \subset Тип 0

Проблема «Можно ли язык, описанный грамматикой типа k ($k = 0, 1, 2$), описать грамматикой типа $k + 1$?» является алгоритмически неразрешимой.

Язык $L_{a,b} = \{a, b\}$. Какого он типа? Обычно требуется указать максимально возможный тип.

Ответ: типа 3

$S \rightarrow a \mid b$ — грамматика типа 3, порождающая данный язык.

($L_{a,b}$ является также языком типа 2, 1, 0 в силу иерархии Хомского)

(1) **Примеры грамматик и языков**

$$S \rightarrow ABCS \quad | \quad ABc$$

$$BA \rightarrow AB$$

$$CA \rightarrow AC$$

$$CB \rightarrow BC$$

$$Cc \rightarrow cc$$

$$Bc \rightarrow bc$$

$$Bb \rightarrow bb$$

$$Ab \rightarrow ab$$

$$Aa \rightarrow aa$$

Тип 1. Неукорачивающая, но не КЗ

Язык: $\{a^n b^n c^n \mid n > 0\}$

Примеры грамматик и языков

(2)

$$S \rightarrow aSb \mid ab$$

Язык: $\{a^n b^n \mid n > 0\}$

(3)

$$S \rightarrow aS \mid a$$

Язык: $\{a^n \mid n > 0\}$

Иерархия классов Хомского



Задача распознавания

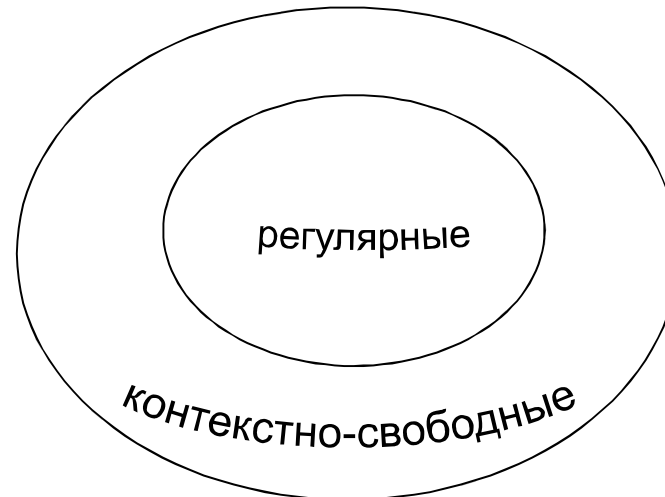
Даны грамматика G и цепочка x

$x \in L(G)$?

Для грамматик типа 1 (а также типов 2 и 3) по классификации Хомского задача распознавания разрешима, т.е. существует общий алгоритм, отвечающий на вопрос: $x \in L(G)$?

Контекстно-свободные грамматики и языки

КС-грамматики позволяют выразить такие свойства языков программирования, как скобочные структуры, последовательность описаний и операторов и др. Но не могут задавать контекстно-зависимые свойства, например, соответствие числа формальных и фактических параметров при вызове функции. Для КС-грамматик существуют эффективные алгоритмы анализа, поэтому они применяются в трансляции, контекстные условия проверяются на этапе семантического анализа



Левый (левосторонний) вывод цепочки $\beta \in (V_T)^*$ из $S \in V_N$ в КС-грамматике $G = (V_T, V_N, P, S)$:

в этом выводе каждая очередная сентенциальная форма получается из предыдущей заменой самого левого нетерминала.

Правый (правосторонний) вывод цепочки $\beta \in (V_T)^*$ из $S \in V_N$ в КС-грамматике $G = (V_T, V_N, P, S)$:

в этом выводе каждая очередная сентенциальная форма получается из предыдущей заменой самого правого нетерминала.

Рассмотрим пример грамматики:

$$G = (\{a,b,+ \}, \{S,T\}, \{S \rightarrow T \mid T+S; T \rightarrow a \mid b\}, S)$$

можно построить выводы для цепочки $a+b+a$:

$$(1) \quad S \rightarrow T+S \rightarrow T+T+S \rightarrow T+T+T \rightarrow a+T+T \rightarrow a+b+T \rightarrow a+b+a$$

$$(2) \quad S \rightarrow T+S \rightarrow a+S \rightarrow a+T+S \rightarrow a+b+S \rightarrow a+b+T \rightarrow a+b+a$$

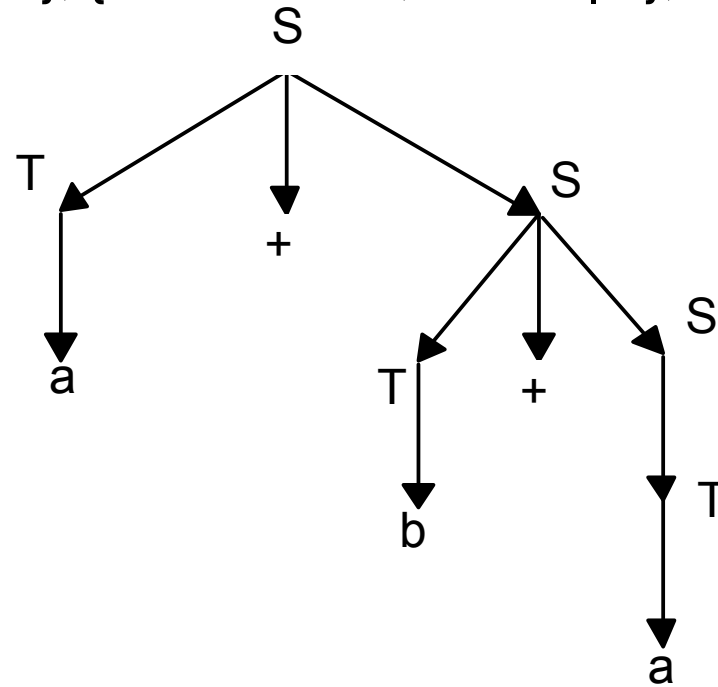
$$(3) \quad S \rightarrow T+S \rightarrow T+T+S \rightarrow T+T+T \rightarrow T+T+a \rightarrow T+b+a \rightarrow a+b+a$$

Здесь (2) - левосторонний вывод, (3) - правосторонний, а (1) не является ни левосторонним, ни правосторонним

Определение: упорядоченное ориентированное дерево называется **деревом вывода** (или **деревом разбора**) в КС-грамматике $G = (T, N, P, S)$, если выполнены следующие условия:

- (1) каждая вершина дерева помечена символом из множества $N \cup T \cup \{\varepsilon\}$, при этом корень дерева помечен символом S ; листья - символами из $T \cup \{\varepsilon\}$;
- (2) если вершина дерева помечена символом A , а ее непосредственные потомки - символами a_1, a_2, \dots, a_n , где каждое $a_i \in T \cup N$, то $A \rightarrow a_1 a_2 \dots a_n$ - правило вывода в этой грамматике;
- (3) если вершина дерева помечена символом A , а ее единственный непосредственный потомок помечен символом ε , то $A \rightarrow \varepsilon$ — правило вывода в этой грамматике.

Пример дерева вывода для цепочки $a+b+a$ в грамматике $G =$
 $(\{a,b,+ \}, \{S,T\}, \{S \rightarrow T \mid T+S; T \rightarrow a \mid b\}, S)$:



- (1) $S \rightarrow T+S \rightarrow T+T+S \rightarrow T+T+T \rightarrow a+T+T \rightarrow a+b+T \rightarrow a+b+a$
- (2) $S \rightarrow T+S \rightarrow a+S \rightarrow a+T+S \rightarrow a+b+S \rightarrow a+b+T \rightarrow a+b+a$
- (3) $S \rightarrow T+S \rightarrow T+T+S \rightarrow T+T+T \rightarrow T+T+a \rightarrow T+b+a \rightarrow a+b+a$

КС-грамматика G называется **неоднозначной**, если существует хотя бы одна цепочка $\alpha \in L(G)$, для которой может быть построено два или более различных деревьев вывода.

Это утверждение эквивалентно тому, что цепочка α имеет два или более разных левосторонних (или правосторонних) выводов.

В противном случае грамматика называется **однозначной**.

Утв. Проблема определения, является ли заданная КС-грамматика однозначной, является **алгоритмически неразрешимой**.

Язык, порождаемый грамматикой, называется **неоднозначным**, если он не может быть порожден никакой однозначной грамматикой.

Утв. Проблема определения, порождает ли данная КС-грамматика однозначный язык (т.е. существует ли эквивалентная ей однозначная грамматика), является **алгоритмически неразрешимой**.

- Пример неоднозначного языка:

$$L = \{a^n b^n c^m \mid n > 0, m > 0\} \cup \{a^n b^m c^m \mid n > 0, m > 0\}$$

Вопросы и задачи

1. Перечислить классы грамматик и классы языков.
2. Каким классам принадлежит данная грамматика? Каким классам принадлежит язык, порождаемый данной грамматикой?

(a) $S \rightarrow AB$	(б) $S \rightarrow AB$	(в) $S \rightarrow aB$
$AB \rightarrow BA \mid bb$	$Ab \rightarrow bb$	$B \rightarrow bB \mid A$
$BA \rightarrow aa \mid ba$	$B \rightarrow b$	$A \rightarrow \varepsilon \mid Ba$

3. Сколько деревьев вывода существует для цепочки ааааа в КС-грамматике

$$S \rightarrow SS \mid a \quad ?$$

4. Построить левый и правый выводы для цепочки ааа для грамматики

$$S \rightarrow SAS \mid \varepsilon$$
$$A \rightarrow aS \mid Sa$$

5. Однозначна ли грамматика из задачи 4? Однозначен ли порождаемый ею язык?

Еще задачи на классификацию можно найти на стр. 94 пособия

<http://cmcmsu.info/download/formal.grammars.and.languages.2009.pdf>

[Волкова И.А., Вылиток А.А., Руденко Т.В. Формальные грамматики и языки. Элементы теории трансляции]