Теорія побудови трансляторів

Лекція 3.

ФОРМАЛЬНІ МОВИ І ГРАМАТИКИ

Визначення формальних мов і граматик

Математичні моделі, що використовують подання текстів у вигляді послідовності символів, називають **формальними мовами і граматиками**.

Кінцева множина символів, неподільних у даному розгляді, називається *словником* чи *алфавітом*, а символи, що входять у множину, *–буквами алфавіту*.

Наприклад, алфавіт *A = {2, b, c, +, !}* містить 5 букв, а алфавіт *B* = {00, 01, 10, 11} містить 4 букви, кожна з яких складається з двох символів.

Послідовність букв алфавіту називається *словом* чи *ланцюжком* у цьому алфавіті. Число букв, що входять у слово, називається його *довжиною*.

Наприклад, слово в алфавіті *A a= 2bc* має довжину l(*a*) = 3, а слово в алфавіті *B b* = 0000110010 має довжину l *(b)* = 5.

Визначення формальних мов і граматик

Формальною граматикою Г, що породжує множину символів, називається наступна сукупність чотирьох об'єктів: Г = { $V_т$, V_A, I ∈ V_A, R }, де V_T – термінальний алфавіт (словник); букви цього алфавіту називаються термінальними символами; з них будуються ланцюжки породжувані граматикою; V_A – нетермінальний, допоміжний алфавіт (словник); букви цього алфавіту використовуються при побудові ланцюжків; вони можуть входити в проміжні ланцюжки, але не повинні входити в результат породження; I - початковий символ граматики $I \in V_a$; R - множина правил чи виводу правил вигляду, що $\alpha \to \beta$, де α і β - ланцюжки, побудовані з букв алфавіту $V_т \cup V_A$, що називають повним алфавітом (словником) граматики Г.

До множини правил граматики можуть також входити правила з порожньою правою частиною вигляду $E \to \$$.

Визначення формальних мов і граматик

Нехай $\tau \rightarrow \gamma$ – правило граматики Г і $\alpha \rightarrow \chi'\tau\,\chi''$ – ланцюжок символів, причому $\chi', \chi'' \in (V_T \cup V_A)^*$. Тоді ланцюжок $\beta \rightarrow \chi'\gamma\,\chi''$ може бути отриманий з ланцюжка α шляхом застосування правила $\tau \rightarrow \gamma$. У цьому випадку говорять, що ланцюжок β безпосередньо виведений з ланцюжка α і позначають $\alpha \Rightarrow \beta$.

Множина кінцевих ланцюжків термінального алфавіту V_T граматики Г, виведених з початкового символу I, називається *мовою, породжуваною граматикою* Г, і позначається $L(\Gamma)$:

$$L(\Gamma) = \{\varpi \in V_T^* \mid <I> \Rightarrow^* \varpi\}.$$

Приклад 1

Задана граматика $\Gamma_{.1}$. Потрібно визначити мову, породжувану цією граматикою:

$$\Gamma_1: \quad V_T = \{a, b, c, d\},\ V_A = \{I, B, C\}$$
$$R = \{\ I \rightarrow aB$$
$$B \rightarrow Cd$$
$$B \rightarrow dc$$
$$C \rightarrow \$\}.$$

Побудуємо всі виводи в цій граматиці:

$I \Rightarrow aB \Rightarrow aCd \Rightarrow ad,$

$I \Rightarrow aB \Rightarrow adc.$

Отже мова $L(\Gamma_{.1}) = \{adc, ad\}$.

ПОРОЖНЯ МОВА

Задано граматику Γ_2. Потрібно визначити мову, породжувану цією граматикою .

$$\Gamma_2 : V_T = \{a, b\},\ V_A = \{I, A\}, R = \{\ I \rightarrow aA,\ A \rightarrow bA\}.$$

Спроба побудови виведення в цій граматиці призводить нас до ланцюжка:

$$I \Rightarrow aA \Rightarrow abA \Rightarrow abbA \Rightarrow \ \ldots\ ,$$

який виявляється нескінченним. Іншими словами, Γ_2 породжує порожню мову. Якщо мова, породжувана граматикою Г, не містить жодного кінцевого ланцюжка (кінцевого слова), то вона називається *порожньою*.

Для того, щоб мова $L(\Gamma)$ не була порожньою, у множині R повинне бути хоча б одне правило вигляду $r = \chi \rightarrow \psi$, *де* $\psi \in V_m^*$ і повинен існувати вивід $I \Rightarrow^* \chi$.

Типи формальних мов і граматик. Граматики типу 0.

Граматики типу 0, які називаються граматиками загального вигляду, не мають ніяких обмежень на правила породження. Будь-яке правило:

$$r = \eta \rightarrow \psi$$

може бути побудоване з використанням довільних ланцюжків $\eta, \psi \in (V_T \cup V_A)^*$. Наприклад, $CCLW \rightarrow WTxSAb \rightarrow xrtHD$.

Типи формальних мов і граматик. Граматики типу 1.

Граматики типу 1, що називаються контекстно-залежними граматиками, не допускають використання будь-яких правил. Правила виводу в таких граматиках повинні мати вигляд:

$$\chi_1 A \chi_2 \rightarrow \chi_1 \omega \chi_2,$$

де χ_1, χ_2 – ланцюжки, можливо порожні, з множини $(V_т \cup V_А)^*$, символ $A \in V_А$ і ланцюжок $w \in (V_т \cup V_А)^*$. Ланцюжки χ_1 і χ_2 залишаються незмінними при застосуванні правила, тому їх називають контекстом (відповідно лівим і правим), а граматику – контекстно-залежною.

Типи формальних мов і граматик. Граматики типу 2.

Граматики типу 2 називають контектно-вільними (КВ) граматиками, або бесконтекстними граматиками.

Правила виводу таких граматик мають вигляд:

$A \rightarrow \alpha,$

де $A \in V_{\text{А}}$ і $\alpha \in (V_{\text{т}} \cup V_{\text{А}})^*$.

Очевидно, що ці правила виходять із правил граматики типу 1 за умови $\chi_1 = \chi_2 = \$$. Оскільки контекстні умови відсутні, то правила КВ-граматик виходять простіші, ніж правила граматик типу 1. Саме такі граматики використовують для опису мов програмування.

Типи формальних мов і граматик. Граматики типу 3.

Граматики типу 3 називають автоматними граматиками (А - граматиками). Правила виводу в таких граматиках мають вигляд:

$A \rightarrow a$, або $A \rightarrow aB$, або $A \rightarrow B\,a$,

де а $\in V_т$, $A, B \in V_А$, причому граматика може мати тільки правила вигляду $A \rightarrow aB$ – правосторонні правила, або тільки вигляду $A \rightarrow Ba$ – лівосторонні правила.

Виведення у КВ- граматиках і правила побудови дерева виведення

Правила побудови дерева виведення можна сформулювати так:

1) Як початок чи вершину кореня дерева візьмемо вершину, яку позначимо початковим символом граматики I; ця вершина утворить нульовий ярус дерева;

2) Якщо при виведенні ланцюжка на черговому кроці використовується правило граматики $A \rightarrow \alpha$ і вершина, позначена нетерміналом A, розташована на ярусі з номером *k-1*, то до побудованого дерева потрібно додати стільки вершин, скільки міститься символів у ланцюжку α, розташувати ці вершини на ярусі k , позначити їх символами ланцюжка α і з'єднати ці вершини дугами з вершиною A. Результатом виведення є множина кінцевих вузлів – листів, що виписуються при обході дерева ліворуч – униз – праворуч – нагору.

Приклад 2

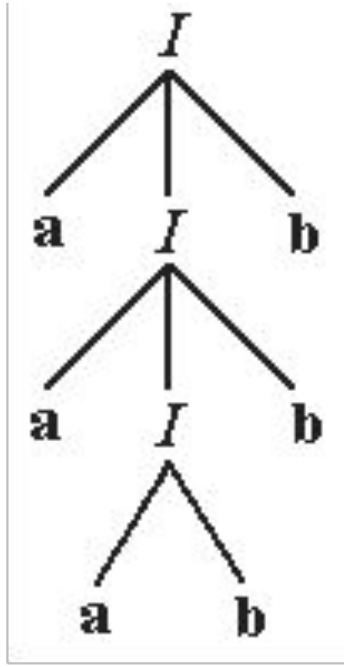
Задано граматику Γ_3:

$$V_m = \{a, b\},\ V_a = \{I\},$$

$$R = \{I \rightarrow aIb,$$

$$I \rightarrow ab\ \},$$

яка породжує мову $L(\Gamma_3) = \{aa...abb...b\}$, де a і b повторюються по n разів ($n=1,2,...$).



Виведення ланцюжка *aaabbb*

Синтаксичний розбір

Послідовність номерів правил граматики Г, застосування яких дозволяє побудувати вивід розглянутого ланцюжка σ з початкового символу граматики, називається синтаксичним розбором σ.

Наприклад, у граматиці Γ_4:

$$V_m = \{ i, +, *, (,) \},\ V_a = \{I, T, P\}$$

$$R = \{ (1)\ I \rightarrow I + T$$
$$(2)\ I \rightarrow T$$
$$(3)\ T \rightarrow T * P$$
$$(4)\ T \rightarrow P$$
$$(5)\ P \rightarrow (I)$$
$$(6)\ P \rightarrow i \},$$

правила якої пронумеровані, вивід

$$I \Rightarrow I + T \Rightarrow T + T \Rightarrow T * P + T \Rightarrow P * P + T \Rightarrow i * P + T \Rightarrow i * i + T \Rightarrow$$
$$i * i + P \Rightarrow i * i + i$$

має синтаксичний розбір [1, 2, 3, 4, 6, 6, 4, 6].

Ліве і праве виведення

Якщо при побудові виведення ланцюжка α при кожнім застосуванні правила заміняється самий лівий нетермінальний символ, то такий виведення називається лівим, або лівостороннім виведенням α. Якщо при побудові виведення α, завжди заміняється самий правий нетермінальний символ проміжного ланцюжка, то виведення називається правим, або правостороннім виведенням α.

Неоднозначні й еквівалентні граматики

Існують граматики, в яких один и той самий ланцюжок може бути отриманий за допомогою різного виведення.

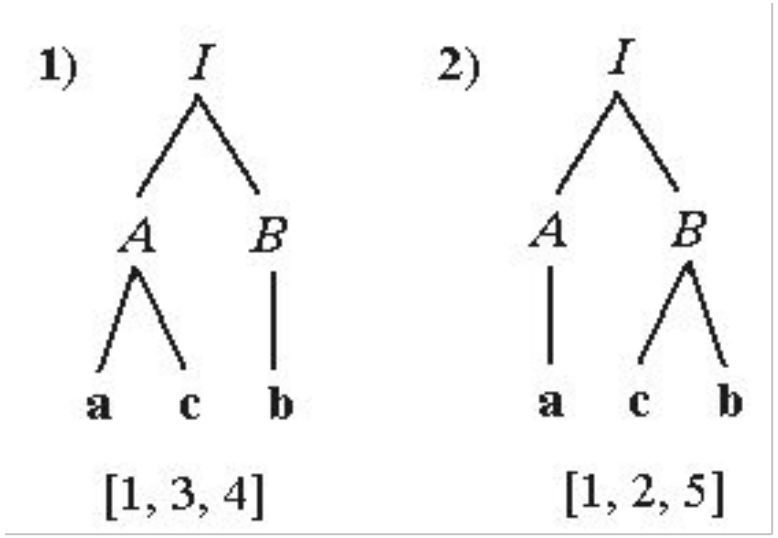
Наприклад, у граматиці Γ_5: $V_T = \{a, b, c, d\}$, $V_A = \{I, A, B\}$, $R = \{1.\ I \to AB,$
$2.\ A \to a,$
$3.\ A \to ac,$
$4.\ B \to b,$
$5.\ B \to cb\}.$

Перше виведення цього ланцюжка має вигляд:

$I \Rightarrow AB \Rightarrow Ab \Rightarrow acb,$

а друге можна отримати так: $I \Rightarrow AB \Rightarrow Acb \Rightarrow acb.$

Цим виведенням відповідають різні синтаксичні дерева і розбори.



Синтаксичні дерева і розбори виразу *acb*

Побудова граматик і граматики, що описують основні конструкції мов програмування

Основою створення правил граматики є спосіб виділення структури заданої множини ланцюжків. Цей спосіб передбачає розчленування ланцюжків, що входять у задану множину, на їх частини таким чином, щоб виявити частини ланцюжків, які повторюються і частини, що входять у всі ланцюжки в незмінному вигляді. Таке розчленування на частини являє собою виявлення структури ланцюжків заданої множини.

Для кожного виявленого елемента структури вводиться позначення. Множина таких позначень складає основу словника нетермінальних символів деякої граматики. Наступним кроком побудови є виявлення послідовностей, у яких елементи структури можуть входити в задані ланцюжки. Такі послідовності є основою для побудови правил граматики.

Рекомендації з побудови граматик

1. Ланцюжку, що складається з заданих символів *abc*, відповідає правило:

$$I \rightarrow abc.$$

2. Ланцюжку, що починається з заданого символу *a*, відповідає правило:

$$I \rightarrow aA.$$

3. Ланцюжку, що закінчується заданим символом *a*, відповідає правило:

$$I \rightarrow Aa.$$

4. Ланцюжку, що починається і закінчується заданими символами *a*, *b*, відповідає правило:

$$I \rightarrow aAb.$$

5. Ланцюжку, що містить у середині символ *a*, відповідає правило:

$$I \rightarrow AaB.$$

6. Ланцюжку заданої довжини l =2 відповідають правила:

$$A \rightarrow aB \text{ і } B \rightarrow a.$$

7. Ланцюжку, що складається з повторюваних символів *a*, відповідають правила:

$$A \rightarrow aA \; i \; A \rightarrow a.$$

8. Ланцюжку, що складається із символів, що чергуються, *a* і *b*, відповідають правила:

$$A \rightarrow aB \; i \; B \rightarrow bA.$$

Опис списків

Послідовності символів і послідовності символів з роздільниками часто називають списками.

1. Позначимо елемент послідовності a. Найпростіша послідовність може складатися з одного елемента a. Всі інші послідовності можуть бути отримані шляхом приписування до вже побудованої послідовності ще одного елемента. Якщо позначити побудовану частину послідовності нетермінальним символом R, а послідовність символом L, то одержимо правила граматики у вигляді:

$$\Gamma_6: \quad L \rightarrow aR,$$

$$R \rightarrow aR,$$

$$R \rightarrow \$,$$

Опис списків

2. У попередній задачі передбачалося, що список L повинен містити хоча б один елемент. Якщо ж допустити, що множина ланцюжків, породжених правилами граматики, може включати порожній символ, то до побудованих правил потрібно додати ще одне правило $L \rightarrow \$$. У цьому випадку набір правил має вигляд :

$$\Gamma_7: \quad L \rightarrow aR,$$
$$R \rightarrow aR,$$
$$R \rightarrow \$,$$
$$L \rightarrow \$.$$

Опис списків

3. Розглянемо побудову списку, між елементами якого повинні стояти роздільники. Виберемо як роздільник кому. Найпростіший список, як і в попередньому випадку, складається з одного елемента, а побудова списку з декількох елементів може бути виконана приписуванням до вже побудованої частини списку роздільника з елементом списку. Правила, що відповідають цій побудові, мають вигляд:

$$\Gamma_8: \quad L \rightarrow aR,$$

$$R \rightarrow ,aR,$$

$$R \rightarrow \$.$$

Опис списків

4. Якщо список з роздільниками може бути порожнім, то наведений вище набір правил потрібно доповнити ще одним правилом з порожньою правою частиною. У результаті одержимо:

Γ_9: $L \rightarrow aR,$

$R \rightarrow ,aR,$

$R \rightarrow \$,$

$L \rightarrow \$.$

Порядок побудови правил граматики

1) виписати кілька прикладів із заданої множини ланцюжків;

2) проаналізувати структуру ланцюжків, виділяючи початок, кінець, що повторюються, або символи з групи символів;

3) ввести позначення для складних структур, що складаються з груп символів; такі позначення є нетермінальними символами шуканої граматики;

4) побудувати правила для кожної з виділених структур, використовуючи для завдання повторюваних структур рекурсивні правила;

5) об’єднати всі правила;

6) перевірити за допомогою виведення можливість одержання ланцюжків з різною структурою.

Приклад 3

Побудувати граматику для мови L ,термінальний словник якого $V_m = \{*, |\}$, а ланцюжки, що утворюють мову, мають наступну структуру:

а) кожен ланцюжок починається буквою * і закінчується двома буквами **.

б) між початком і кінцем ланцюжків можуть бути або непорожня послідовність паличок, або кілька таких послідовностей, розділених символами *.

1. Спочатку побудуємо кілька ланцюжків заданої мови, що можуть бути подані в наступному вигляді:

* |||**,

* |*|*|**,

* ||*||||*|||||** ,

* |||*|*||*||||||** .

Приклад 3

2. Розглядаючи наведені ланцюжки, можна виділити наступні їх структурні компоненти:

- початок ланцюжка (символ *);
- кінець ланцюжка (символи **);
- непорожня група паличок;
- послідовність груп паличок, розділених зірочками.

3. Позначимо групу паличок символом *A*, а послідовність груп паличок символом *B*.

Приклад 3

4. Виділені структури можна розглядати як списки. Так послідовність паличок являє собою список без роздільників, елементом якого є паличка. Правила граматики, що задає такий список, мають такий вигляд:

$$1.A \rightarrow |R,$$

$$2.R \rightarrow |R,$$

$$3.R \rightarrow \$.$$

Приклад 3

5. Послідовність груп паличок, розділених зірочкою, являє собою список з роздільником. Елементом такого списку є група паличок *A*, а роздільником – зірочка. Правила граматики, що задає такий список, можна записати так:

$$B \rightarrow AR_1,$$

$$R_1 \rightarrow *AR_1,$$

$$R_1 \rightarrow \$.$$

З огляду на те, що кожен ланцюжок мови повинен мати початок і кінець, і , вибираючи як початковий символ граматики *I*, одержуємо правило, що визначає загальний вигляд ланцюжка:

$$I \rightarrow *B**.$$

Приклад 3

6. Поєднуючи побудовані правила, остаточно одержимо схему шуканої граматики у вигляді:

$$\Gamma_{10}: R = \{ I \rightarrow *B**,$$

$$B \rightarrow AR1,$$

$$R1 \rightarrow *AR1,$$

$$R1 \rightarrow \$,$$

$$A \rightarrow | R,$$

$$R \rightarrow | R,$$

$$R \rightarrow \$ \}$$

Приклад 3

7. За допомогою правил побудованої граматики:

Γ_{10}: $R = \{ 1.\ I \rightarrow *B**,\quad 2.\ B \rightarrow AR1,\quad 3.R1 \rightarrow *AR1,\quad 4.\ R1 \rightarrow \$,\ 5.A \rightarrow |\,R,$

$6.R \rightarrow |\,R,\ 7.\ R \rightarrow \$ \}$

можна отримати, наприклад, ланцюжок: * | * | * | **.

Номера правил				
2	5	7	3	5
$I \Rightarrow *B**$	$\Rightarrow *AR1**$	$\Rightarrow *\|\ R\ R1**$	$\Rightarrow *\|\ R_1**$	$\Rightarrow *\|\ *AR_1**$
Номера правил				
7	2	5	4	
$\Rightarrow *\|\ *\|\ R\ R_1**$	$\Rightarrow *\|\ *\|\ R_1**$	$\Rightarrow *\|\ *\|\ AR_1**$	$\Rightarrow *\|\ *\|\ *\|R_1**$	$\Rightarrow *\|\ *\|\ *\|**$

Приклад 4

Побудувати правила граматики для опису цілих чисел.

Γ_{11}: *1.* $N \rightarrow DR$,

2. $R \rightarrow DR \mid 0R$

3. $R \rightarrow \$$,

4. $D \rightarrow 1 \mid \ldots \mid 9$.

Виведення для числа 127.

$N \rightarrow DR \rightarrow 1R \rightarrow 1DR \rightarrow 12R \rightarrow 12DR \rightarrow 127R \rightarrow 127$

Синтаксичний розбір: [1, 4.2, 2, 4.3, 3, 4.8, 3].

Приклад 5

Побудувати правила граматики для опису ідентифікатора.

Γ_{12}: $R = \{$ *1.* $I \rightarrow CA$,

2. $A \rightarrow CA|DA$,

3. $A \rightarrow C|D$,

4. $A \rightarrow \$$,

5. $D \rightarrow 0 \mid 1 \mid \ldots \mid 9$,

6. $C \rightarrow a \mid b \mid c \mid \ldots \mid z \mid \_ \}$.

Виведення для ідентифікатора ***a0_d:***

I → CA → aA → aDA → a0A → → a0CA → a0_A → a0_CA → a0_dA → a0A

Синтаксичний розбір: [6.1, 2.2, 5.1, 2.2,6.28, 2.2, 6.4, 4].

Дякую за увагу