7. Контекстно-свободные грамматики и языки

Разделы:

- Контекстно-свободные грамматики
- Деревья разбора
- Неоднозначность в грамматикаих и языках
- Эквивалентность МПА и КСГ
- ДМПА и КСГ

Неформальное описание КСГ

- КСГ представляет собой формальную запись рекурсивных определений языков
- КСГ, как и любая грамматика, состоит из переменных (нетерминальных символов), которые представляют собой наборы строк, или языки
- Только один нетерминальный символ для представления класса строк языка L
- Имеются правила построения строк каждого класса
- При построении используются символы алфавита и уже построенные строки из разных классов

Неформальное описание КСГ

Table Text Size

LHS		RHS
P	→	ε
P	→	0
P	→	1
P	→	0P0
P	→	1P1

Формальное определение КСГ

- КСГ $G = (V_T, V_N, P, S)$ с известными нам обозначениями
- Специфичным для КСГ является вид правил вывода (или **продукций**)
- Каждая продукция в КСГ имеет вид $\alpha \to \beta$, где $|\alpha|=1$
- Иначе говоря, в LHS любой продукции всегда ровно один нетерминал

Формальное определение КСГ

Table Text Size

LHS		RHS
E	→	I
E	→	E+E
E	→	E*E
E	→	{E}
I	→	a
I	→	b
I	→	Ia
I	→	Ib
I	→	I0
I	→	I1

Порождения

- Для того чтобы убедиться, что данные строки принадлежат языку некоторого нетерминала, мы применяем продукции КСГ
- Есть два способа действий
  1. Применяют продукции от RHS к LHS - **обратное порождение**, или **рекурсивный вывод**
  2. Продукции применяются от LHS к RHS
     - Аксиома разворачивается с использованием одной из его продукций
     - Это называется **порождением**, или **выводом**

Порождения

- Строка $a*(a+b00)$ принадлежит языку нетерминала E
- Доказательство можно отразить с помощью порождения, начиная от этого символа

$E \Rightarrow E * E \Rightarrow I * E \Rightarrow a * E \Rightarrow$
$a * (E) \Rightarrow a * (E + E) \Rightarrow$
$a * (I + E) \Rightarrow a * (a + E) \Rightarrow$
$a * (a + I) \Rightarrow a * (a + I0) \Rightarrow$
$a * (a + I00) \Rightarrow a * (a + b00)$

Порождения

LHS		RHS
E	⟶	I
E	⟶	E+E
E	⟶	E*E
E	⟶	{E}
I	⟶	a
I	⟶	b
I	⟶	Ia
I	⟶	Ib
I	⟶	I0
I	⟶	I1

Production	Derivation
	E
E→E*E	E*E
E→{E}	E*{E}
E→E+E	E*{E+E}
E→I	E*{E+I}
I→I0	E*{E+I0}
E→I	I*{E+I0}
E→I	I*{I+I0}
I→I0	I*{I+I00}
I→a	a*{I+I00}
I→a	a*{a+I00}
I→b	a*{a+b00}

Порождения

- Обратное порождение для первого :
  - $a * (a + b00) <= a * (a + I00) <=$
    $a * (a + I0) <= a * (a + I) <=$
    $a * (a + E) <= a * (I + E) <=$
    $a * (E + E) <= a * (E) <= a * E <=$
    $I * E <= E * E <= E$
- Иногда используется еще один способ записи обратного порождения:
  - $E =>_R E * E =>_R I * E =>_R a * E =>_R a * (E)$
    $=>_R a * (E + E) =>_R a * (I + E)$
    $=>_R a * (a + E) =>_R a * (a + I)$
    $=>_R a * (a + I0) =>_R a * (a + I00)$
    $=>_R a * (a + b00)$
- Мы можем использовать сокращенную запись с использованием известного оператора:
  - $E =>^* a * (a + b00)$
  - $E =>_R^* a * (a + b00)$

Порождения

- Если на каждом шаге порождения мы заменяем крайний левый нетерминал, такое порождение - **левое**
- Если на каждом шаге порождения мы заменяем крайний правый нетерминал, то такое порождение - **правое**
- Следующее порождение – правое для той же строки, что и в предыдущем примере
- $E \Rightarrow E * E \Rightarrow E * (E) \Rightarrow E * (E + E) \Rightarrow$
  $E * (E + I) \Rightarrow E * (E + I0) \Rightarrow E * (E + I00) \Rightarrow$
  $E * (E + b00) \Rightarrow E * (I + b00) \Rightarrow$
  $E * (a + b00) \Rightarrow I * (a + b00) \Rightarrow a * (a + b00)$

Языки КСГ

$$L(G) = \left(w \in V_T^* \middle| S \Rightarrow_G^* w \right)$$

- Если язык задается некоторой КСГ, то он называется **контекстно-свободным языком**
- Любую строку, которая участвует на каждом шаге порождения, мы называем **сентенциальной формой** (СФ)
- КСЯ представляет собой множеств СФ, состоящих исключительно из терминальных символов

Деревья разбора

- Деревья разбора для КСГ G – это деревья со следующими свойствами:
  - Корень помечается аксиомой грамматики
  - Каждый внутренний узел помечен нетерминалом из V_N
  - Каждый лист отмечен либо нетерминалом, либо терминалом, либо ε. В последнем случае, лист должен быть единственным потомком родительского узла
  - Если внутренний узел помечен A и его прямые потомки отмечены слева направо $X_1, \ldots, X_k$, то $A \to X_1, \ldots, X_k$ является продукцией из P

Деревья разбора

Table Text Size

LHS		RHS
E	→	I
E	→	E+E
E	→	E*E
E	→	{E}
I	→	a
I	→	b
I	→	Ia
I	→	Ib
I	→	I0
I	→	I1

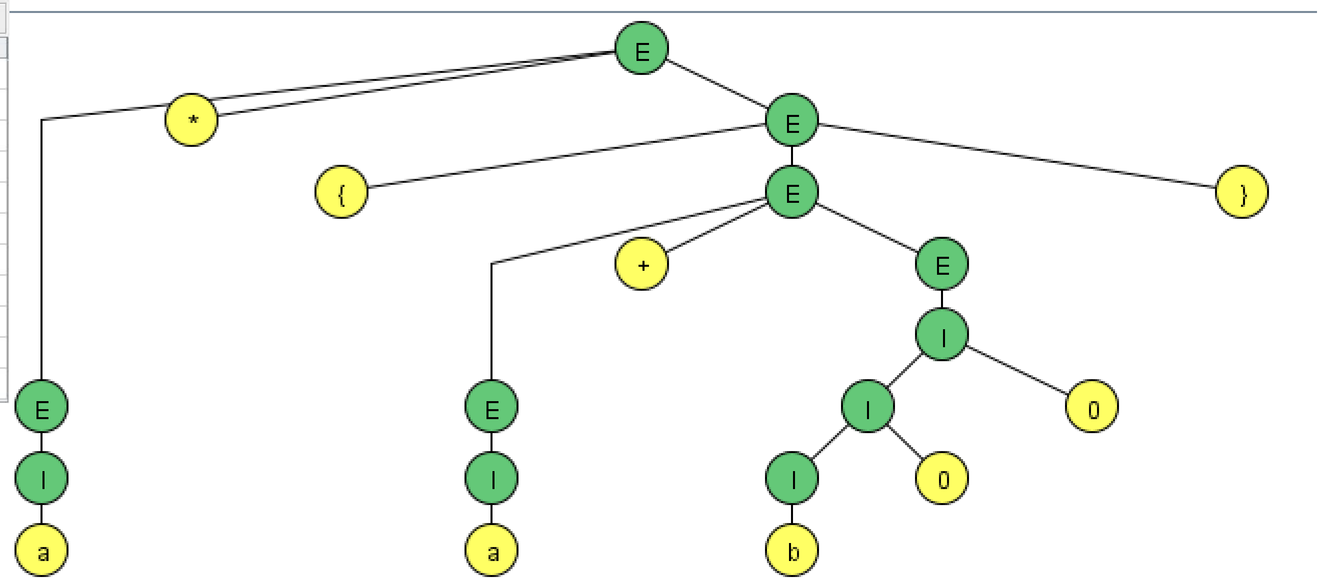
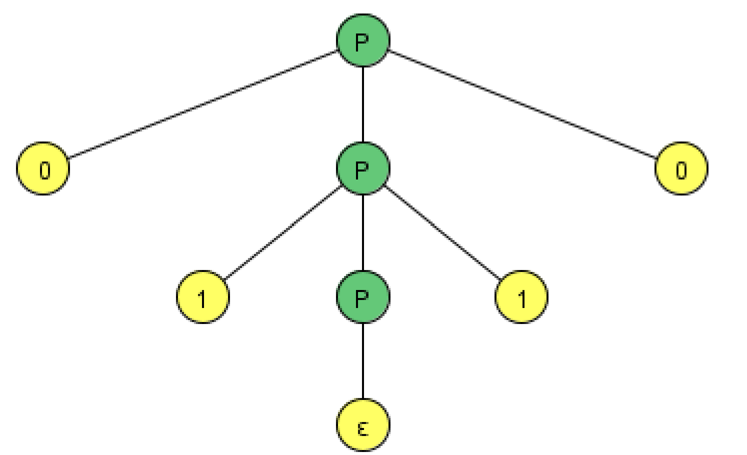
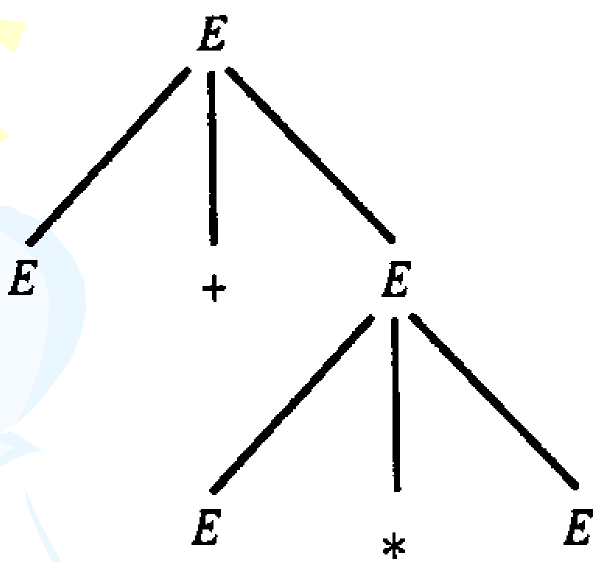


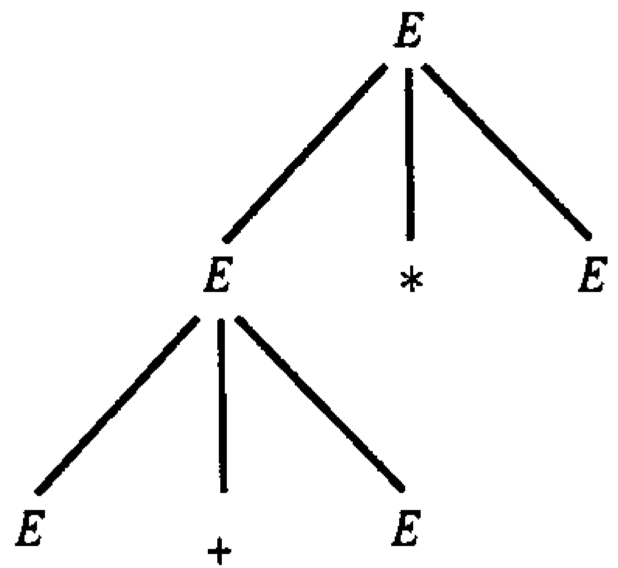
Table Text Size

LHS		RHS
P	→	ε
P	→	0
P	→	1
P	→	0P0
P	→	1P1



Неоднозначности



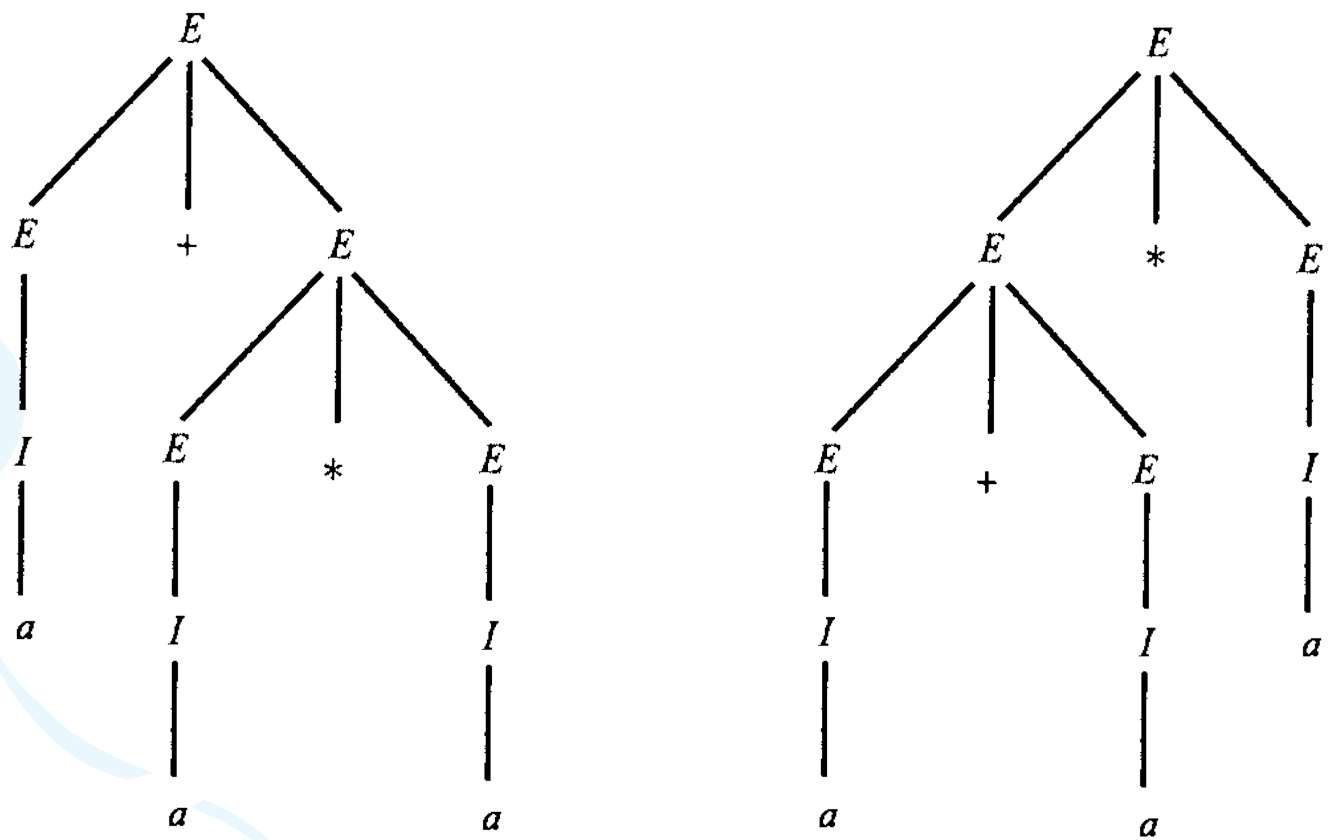


- **СТУДЕНТАМ:** Покажите, что составляет кроны деревьев разбора

Неоднозначности

- КСГ является **неоднозначной**, если найдется хотя бы одна строка *w* из V_T^*, для которой существуют два дерева разбора с корнем *S* и кроной *w*
- Если каждая строка имеет не более одного дерева разбора в грамматике, то она **однозначна**
- Иными словами, неоднозначность происходит не от количества порождений, а от существования двух или более деревьев разбора

Неоднозначности



•**СТУДЕНТАМ:** Покажите, что составляет кроны деревьев разбора

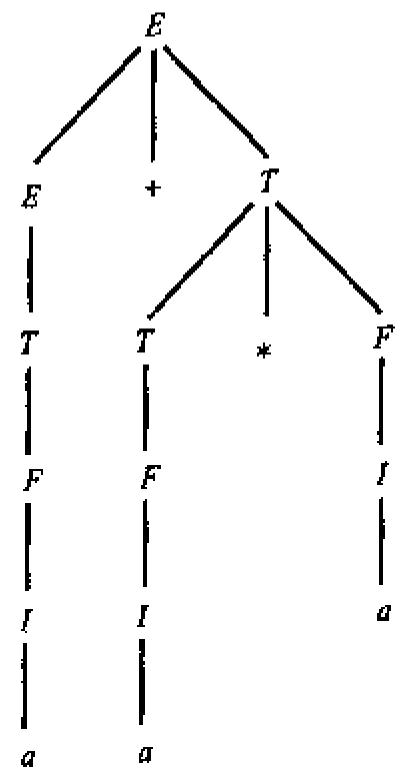
Неоднозначности

- **Сомножитель**, или **фактор** (*factor*) – это выражение, которое не может быть разделено ни одной из операций
  - Обозначим сомножитель символом *F*
- **Слагаемое**, или **терм** (*term*) – это выражение, которое не может быть разделено операцией сложения
  - Обозначим терм символом *T*
- **Выражение** (*expression*) – это любые выражения, которые могут быть разделены знаками операций
  - Оставим для выражения обозначение символом *E*

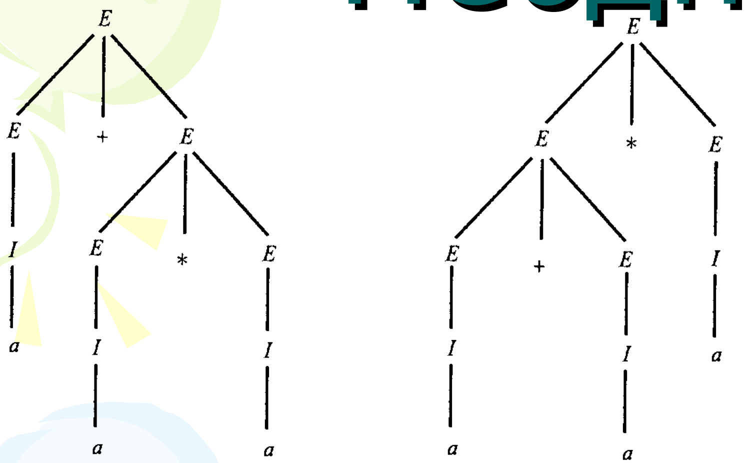
Неоднозначности

Table Text Size

LHS		RHS
E	→	T
E	→	E+T
T	→	F
T	→	T*F
F	→	I
F	→	{E}
I	→	a
I	→	b
I	→	Ia
I	→	Ib
I	→	I0
I	→	I1



Неоднозначности



- По кронам деревьев разбора мы можем получить два левых порождения
  - $E \Rightarrow E+E \Rightarrow I+E \Rightarrow a+E \Rightarrow a+E*E \Rightarrow a+I*E \Rightarrow a+a*E \Rightarrow a+a*I \Rightarrow a+a*a$
  - $E \Rightarrow E*E \Rightarrow E+E*E \Rightarrow I+E*E \Rightarrow a+E*E \Rightarrow a+I*E \Rightarrow a+a*E \Rightarrow a+a*I \Rightarrow a+a*a$
- Эти левые порождения **различаются**
- **Теорема 7.1.** Для каждой КСГ G и w из V_T^* строка w имеет два дерева разбора тогда и только тогда, когда w имеет два разных левых порождения

Неоднозначности

- КСЯ *L* называется **существенно неоднозначным**, если все его грамматики неоднозначны
- Если хотя бы одна грамматика однозначна, то и КСЯ тоже однозначен
- Рассмотрим язык, неоднозначность которого можно доказать, хотя мы это сделаем неформальным способом

$$L = \{a^n b^n c^m d^m \mid n \geq 1, m \geq 1\} \cup \{a^n b^m c^m d^n \| n \geq 1, m \geq 1\}$$

•**СТУДЕНТАМ:** сформулируйте словесное описание языка *L*

Неоднозначности

$$L = \left\{a^n b^n c^m d^m \middle| n \geq 1, m \geq 1\right\} \cup \left\{a^n b^m c^m d^n \middle\| n \geq 1, m \geq 1\right\}$$

Table Text Size

LHS		RHS
S	→	AB
S	→	C
A	→	aAb
A	→	ab
B	→	cBd
B	→	cd
C	→	aCd
C	→	aDd
D	→	bDc
D	→	bc

Неоднозначности

$$L = \{a^n b^n c^m d^m \mid n \geq 1, m \geq 1\} \cup \{a^n b^m c^m d^n \parallel n \geq 1, m \geq 1\}$$

- Левое порождение #1 (*using* JFLAP)

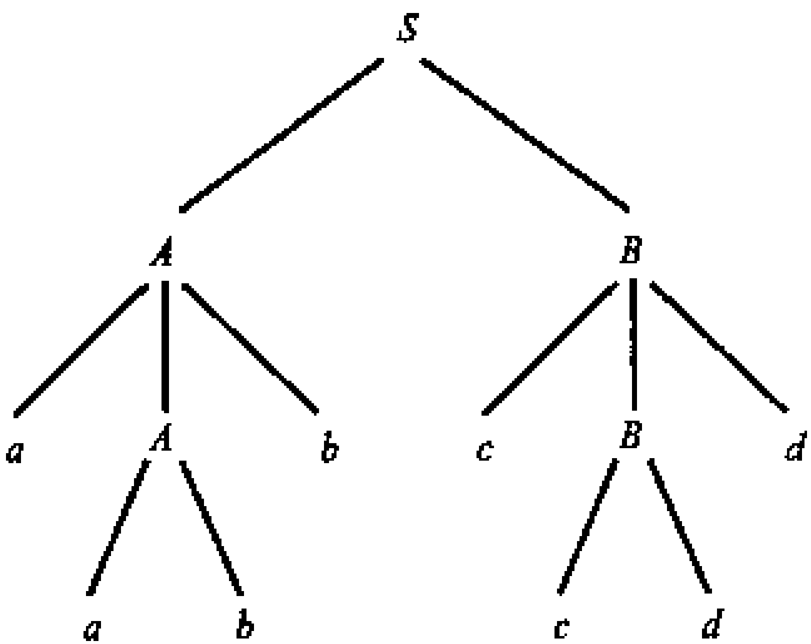
Production	Derivation
	S
S→AB	AB
A→aAb	aAbB
A→ab	aabbB
B→cBd	aabbcBd
B→cd	aabbccdd

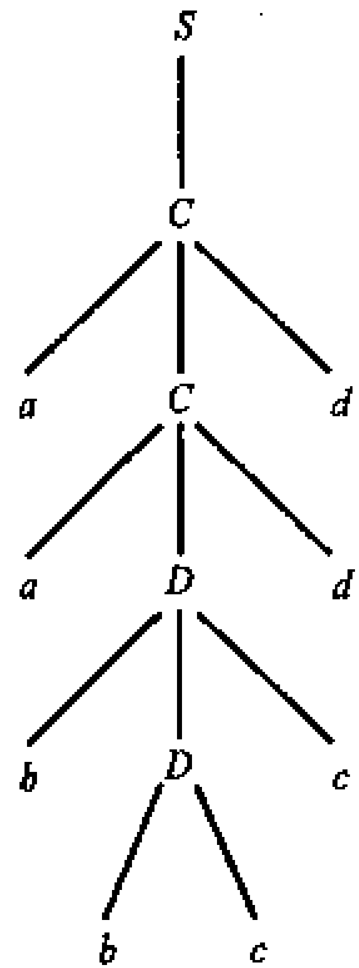
- Левое порождение #2 (*using derivation symbol*)

S=>C=>aCd=>aaDdd=>aabDcdd=>aabbccdd

Неоднозначности

$$L = \{a^n b^n c^m d^m | n \geq 1, m \geq 1\} \cup \{a^n b^m c^m d^n \| n \geq 1, m \geq 1\}$$





Эквивалентность МПА и КСГ

Пусть по грамматике G строится МПА, имитирующий ее левые порождения. Любую СФ, порождаемую таким образом, которая не является терминальной, можно записать в виде $xA\alpha$, где A – крайний левый нетерминал, x – строка терминалов слева от A, α – строка терминалов и нетерминалов справа. $A\alpha$ называется **остатком**, или **хвостом** (*tail*) этой СФ. У терминальной СФ остатком, очевидно, является ε.

МПА имитирует последовательность «левопорождаемых» СФ, используемых в грамматике для порождения терминальной СФ w. Хвост $A\alpha$ появляется в магазине с нетерминалом A на вершине. При этом x представляется прочитанными на входе символами, а суффикс СФ w после x еще не прочитан.

Пусть МПА находится в конфигурации $(q, y, A\alpha)$, представляющей нашу строку $xA\alpha$. Он «угадывает» продукцию, используемую для расширения A. Допустим, это продукция $A \to \beta$. Переход МПА состоит в том, что A на вершине магазина заменяется на β, тем самым достигается МО $(q, y, \beta\alpha)$. У этого МПА всего одно состояние q.

Теперь $(q, y, \beta\alpha)$ может не оказаться представленной следующей левопорождаемой СФ, поскольку у β может быть терминальный префикс. Более того, β может вообще не иметь нетерминалов, а у α может оказаться терминальный префикс. Все терминалы требуется удалить из начала СФ $\beta\alpha$ до появления следующего нетерминала в магазине. Эти терминалы сравниваются со следующими входными символами для того, чтобы убедиться в правильности наших предположений о левом порождении входной строки w. Если предположение неправильно, соответствующая ветвь вычислений МПА отбрасывается.

Если таким образом нам удается угадать левое порождение w, то рано или поздно мы дойдем до левопорождаемой строки w. В этот момент все нетерминалы в магазине расширены, и все терминалы совпали с входными символами. Магазин опустошен, и МПА допускает по пустому магазину.

Эквивалентность МПА и КСГ

- Пусть $G = (V_T, V_N, Q, S)$ – КСГ
- Построим МПА $P = (\{q\}, V_T, V_T \cup V_N, \delta, q, Z_0, S)$, который допускает $L(G)$ по пустому магазину
- Функция переходов определена так:
  1. $\delta(q, \varepsilon, A) = \{(q, \beta) \mid A \text{ -> } \beta$ – продукция $G\}$ для всех нетерминалов A
  2. $\delta(q, a, a) = \{(q, \varepsilon)\}$ для всех терминалов a

Эквивалентность МПА и КСГ

- Преобразуем КСГ арифметических выражений в МПА
- Множеством входных символов для МПА является $\{a, b, 0, 1, (,), +, *\}$
- Объединив эти символы с нетерминалами E и I, мы получим магазинный алфавит
- Функция переходов:
  - $\delta(q, \varepsilon, I) = \{(q, a), (q, b), (q, Ia), (q, Ib), (q, I0), (q, I1)\}$ – по правилу 1
  - $\delta(q, \varepsilon, E) = \{(q, I), (q, E+E), (q, E*E), (q, (E))\}$ – по правилу 1
  - $\delta(q, a, a) = \delta(q, b, b) = \delta(q, 0, 0) = \delta(q, 1, 1) = \delta(q, (, () = \delta(q,),)) = \delta(q, +, +) = \delta(q, *, *) = \{(q, \varepsilon)\}$ – по правилу 2

Эквивалентность МПА и КСГ

- **Теорема 7.2.** Если МПА P построен по КСГ G, то $N(P) = L(G)$
- Конструкция эквивалентной КСГ использует переменные, которые могут обозначаться составным символом
- Каждая из которых представляет событие, состоящее в следующем:
  - Окончательное удаление элемента X из магазина
  - Изменение состояния, скажем, с p на q, когда X заменяется на ε в магазине
    - Символ $[pXq]$
- **Теорема 7.3.** Пусть P – МПА. Тогда существует КСГ G, для которой $L(G) = N(P)$

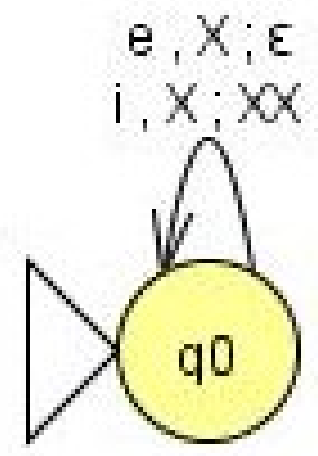
Эквивалентность МПА и КСГ

- Преобразуем в грамматику МПА

$$P_N = \left(\{q_0\}, \{i, e\}, \{X\}, \delta_N, q_0, Z_0, \varnothing\right)$$

$$\delta_N\left(q_0, i, X\right) = \{(q_0, XX)\}$$

$$\delta_N\left(q_0, e, X\right) = \{(q_0, \varepsilon)\}$$



Эквивалентность МПА и КСГ

- В КСГ только два нетерминала
  - S – аксиома
  - $[qXq]$ – единственная тройка, которую можно собрать из состояний и магазинных символов нашего МПА
- Продукции
  1. Единственная продукция для аксиомы $S \to [qXq]$
  2. Вследствие того, что $\delta_N(q, i, X)$ содержит (q, XX), получаем продукцию $[qXq] \to i[qXq][qXq]$
  3. Из того, что $\delta_N(q, e, X)$ содержит (q, ε), получаем продукцию $[qXq] \to e$
- Результат: $S \to A$, $A \to iAA$, $A \to e$
- Т.к. S порождает то же, что и A, то результат можно упростить: $S \to iSS$, $A \to e$

ДМПА и КСГ

Существуют КСЯ вроде L_{wwr}, которые не могут допускаться по заключительному состоянию никаким ДМПА. Интуитивное доказательство сравнительно просто. Если P – ДМПА, допускающий язык L_{wwr}, то при чтении последовательности символов 0 он должен вталкивать их в магазин или нечто другое, что может подсчитать их количество. Можно записывать один Y для каждых 00 на входе и использовать состояние для запоминания четности/нечетности количества символов 0.

Пусть P прочитал n символов 0, а затем увидел на входе 110^n. Он должен проверить, что после пары символов 11 находятся n символов 0. Для этого он должен опустошить свой магазин. Теперь он прочитал строку 0^n110^n. Если далее он видит такую же строку, то он должен ее принимать, т.к. весь вход имеет вид ww^R, где $w = 0^n110^n$. Однако если P видит вход 0^m110^m, где $m \neq n$, он должен ее **отвергнуть**. Т.к. его магазин пуст, он не помнит, каким было значение n, и не способен допустить L_{wwr}.

Резюме: языки, допускаемые ДМПА по заключительному состоянию, включают РЯ как собственное подмножество, но сами образуют собственное подмножество КСЯ.

Мощность ДМПА можно уточнить, заметив, что все языки, допускаемые ими, имеют однозначные грамматики. Однако класс языков ДМПА не совпадает с существенно неоднозначными КСЯ. В частности, L_{wwr} имеет однозначную КСГ (S->0S0, S->1S1, S->ε), не являясь при этом языком ДМПА.

ДМПА и КСГ

- **Теорема 7.4.** Если $L = N(P)$ для ДМПА P, то L имеет однозначную КСГ

- **Теорема 7.5.** Если $L = L(P)$ для ДМПА P, то L имеет однозначную КСГ

Дополнительные источники

- Автомат с магазинной памятью - http://ru.wikipedia.org/wiki/Автомат_с_магазинной_памятью
- Введение в схемы, автоматы и алгоритмы - http://www.intuit.ru/studies/courses/1030/205/info
- Гилл, А. Введение в теорию конечных автоматов / А. Гилл. – М.: Наука, 1966. – 272 с.
- Контекстно-свободная грамматика - http://ru.wikipedia.org/wiki/Контекстно-свободная_грамматика
- Короткова, М.А. Математическая теория автоматов. Учебное пособие / М.А. Короткова. – М.: МИФИ, 2008. – 116 с.

Дополнительные источники

- Кузнецов, А.С. Теория вычислительных процессов [Текст] : учеб. пособие / А. С. Кузнецов, М. А. Русаков, Р. Ю. Царев ; Сиб. федерал. ун-т. - Красноярск: ИПК СФУ, 2008. – 184 с.
- Молчанов, А. Ю. Системное программное обеспечение. 3-е изд. / А.Ю. Молчанов. – СПб.: Питер, 2010. – 400 с.
- Серебряков В. А., Галочкин М. П., Гончар Д. Р., Фуругян М. Г. Теория и реализация языков программирования — М.: МЗ-Пресс, 2006 г., 2-е изд. - http://trpl7.ru/t-books/TRYAP_BOOK_Details.htm
- Теория автоматов / Э. А. Якубайтис, В. О. Васюкевич, А. Ю. Гобземис, Н. Е. Зазнова, А. А. Курмит, А. А. Лоренц, А. Ф. Петренко, В. П. Чапенко // Теория вероятностей. Математическая статистика. Теоретическая кибернетика. — М.: ВИНИТИ, 1976. — Т. 13. — С. 109–188. — URL http://www.mathnet.ru/php /getFT.phtml?jrnid=intv&paperid=28&what=fullt&option_lang=