Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное образовательное
учреждение высшего образования «Национальный исследовательский
университет ИТМО»

Факультет инфокоммуникационных технологий

Математическая лингвистика

Практическая работа №3

Выполнили:

студент группы К34422

Малаев Степан Геннадьевич

Проверил:

доцент практики, КТН

Болгова Екатерина Владимировна

Санкт-Петербург

2025

Ход работы.

1. Задана грамматика G с терминальным алфавитом T = {a, b}, нетерминальным алфавитом N = {S, A, C}, аксиомой S, множеством правил P = {S → ABC, A → aAa, B → bBb, C → cCc, A → a, B → a}. Существует ли в этой грамматике последовательность правил, позволяющая вывести цепочку a3b2 из цепочки AbB?

Аналитически, невозможно вывести строку a3b2 из AbB исходя из особенностей грамматики.

Правила для A и для B устроены рекурсивно так, что при каждом применении правила вида X → xXx добавляются по одному терминальному символу с каждой стороны. Это означает, что базовое правило например, A → a, порождает строку длины 1, а каждое последующее применение рекурсии увеличивает длину на 2. Таким образом, любая строка, выведенная из A, имеет нечётную длину.

Целевая цепочка a^3b^2 содержит чётное число букв b. Чтобы получить ровно 2 b, необходимо иметь правило, которое добавляет один b или завершает вывод B с чётным количеством b. Но рекурсивное правило B → bBb добавляет b парами, всегда сохраняя нечётность, а базовое правило B → a порождает символ, отличный от требуемого b.

Таким образом, нет такой последовательности правил.

Практическое решение с использованием BFS также не дало результат.

2. Задана грамматика G с терминальным алфавитом T = {a, b} и нетерминальным алфавитом N = {S, A, F}; P = {S → aAa, S → bAb, A → aAa, A → bAb, S → aa, S → bb, A → aa, A → bb}. Выберите правила этой грамматики, которые нужно использовать при выводе цепочки a4b4a4 в грамматике G (постройте полный вывод, заканчивающийся цепочкой a4b4a4).

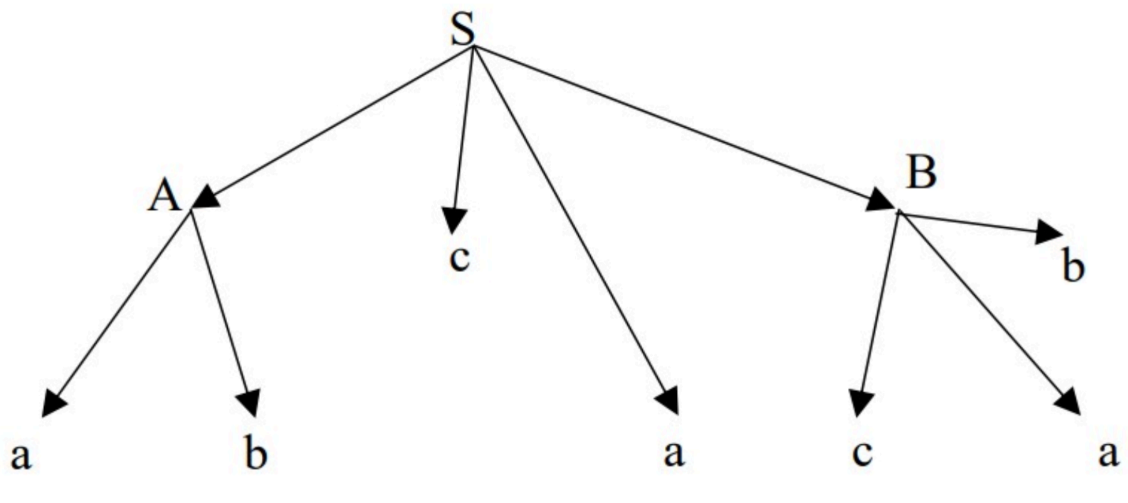
Применяя следующую последовательность правил:

1. $S \to aAa$
2. $A \to aAa$
3. $A \to aAa$
4. $A \to aAa$
5. $A \to bAb$
6. $A \to bb$

Можно прийти к следующему выводу:

1. A_5 = “bb”
2. A_4 = “b” + “bb” + “b” = “bbbb”
3. A_3 = “a” + “bbbb” + “a” = “abbbba”
4. A_2 = “a” + “abbbba” + “a” = “aabbbbaa”
5. A_1 = “a” + “aabbbbaa” + “a” = “aaabbbbaaa”
6. S = “a” + “aaabbbbaaa” + “a” = “aaaabbbbaaaa”

3. Задана грамматика G с аксиомой S, терминальным алфавитом T = {a, b, c} и нетерминальным алфавитом N = {S, D, A}; P = {$S \to AcaD$, $A \to DAa$, $A \to ab$, $D \to cab$}. Задано дерево вывода.



Среди приведённых утверждений выберите истинное:

a) Это полное дерево вывода цепочки abcacab в грамматике G

b) Это полное дерево вывода цепочки abaca в грамматике G

c) Это полное дерево вывода цепочки AcBb в грамматике G

d) Это полное дерево вывода цепочки SBcab в грамматике G

e) Это полное дерево вывода цепочки abAS в грамматике G

f) Это дерево не является деревом полного вывода в грамматике G

В рассматриваемой грамматике G с нетерминалами {S, D, A} и терминалами {a, b, c} символа B нет ни в множестве терминалов, ни в множестве нетерминалов. Следовательно, дерево, в котором фигурирует B, не может соответствовать ни одному правилу данной грамматики. Поэтому такое дерево не является деревом полного вывода для грамматики G.

Утверждение f является верным.

4. Рассмотрим высказывание Рано встаёт охрана. Смысл не изменится, если наречие рано переместить в последнюю позицию: Встаёт охрана рано. В тех же двух позициях можно вместо рано употребить другие наречия: весело, бодро, лениво, медленно, быстро. Рассмотрим множество предложений, образующихся путём добавления к словосочетанию встаёт охрана в первой или последней позиции одного (и только одного) из названных шести наречий. Постройте КС-грамматику, которая порождала бы в точности это множество предложений.

Построим контекстно-свободную грамматику:

T = {рано, весело, бодро, лениво, медленно, быстро, встаёт, охрана}

N = {S, ADV, PH}, где:

- S — аксиома, описывает всё предложение.
- ADV — нетерминал для одного из шести наречий.
- PH — нетерминал для словосочетания “встаёт охрана”.

S → ADV PH | PH ADV

ADV → рано | весело | бодро | лениво | медленно | быстро

PH → встаёт охрана

Таким образом, каждое из шести наречий порождается ровно один раз либо слева, либо справа.

5. Покажите, что грамматика F, имеющая продукции вида S → bA | aB, A → a | aS | bAA, B → b | bS | aBB порождает язык L {a, b}* составленный из строк, содержащих равное число ⊂ символов a и символов b. (Воспользуйтесь методом индукции и докажите, что для любой сентенциальной формы общее число элементов a и A равно общему числу элементов b и B.)

Мы введём функцию f для символов следующим образом:

- f(a) = 1
- f(b) = -1

Приписываем также значения для нетерминалов, чтобы инвариант сохранялся:

- f(S) = 0
- f(A) = 1
- f(B) = -1

Для произвольной сентенциальной формы w определим сумму

$$F(w) = \sum_{x \in w} f(x).$$

Если w состоит только из терминальных символов, то

$$F(w) = \#(a) - \#(b).$$

Нужно доказать, что для любой сентенциальной формы, полученной из S, справедливо F(w) = 0, что эквивалентно условию #(a) = #(b).

База индукции:

Начнём с аксиомы: w = S.

Так как f(S)=0, то F(S)=0.

Индуктивный шаг:

Докажем, что замена нетерминала согласно продукции сохраняет значение F.

1. Для S:

$$S \to bA, F(bA) = f(b) + f(A) = (-1) + 1 = 0$$

$$S \to aB, F(aB) = f(a) + f(B) = 1 + (-1) = 0$$

2. Для A:

$$A \to a, F(a) = 1, f(A) = 1$$

$$A \to aS, F(aS) = f(a) + f(S) = 1 + 0 = 1$$

$$A \to bAA, F(bAA) = f(b) + f(A) + f(A) = -1 + 1 + 1 = 1$$

3. Для B:

$$B \to b, F(b) = -1, f(B) = -1.$$

$$B \to bS, F(bS) = -1 + 0 = -1$$

$$B \to aBB, F(aBB) = 1 + (-1) + (-1) = -1$$

Таким образом, каждая замена нетерминала сохраняет значение F. По индукции, для любой сентенциальной формы w, полученной из S, выполнено F(w) = 0. А поскольку w состоит только из терминальных символов a и b, это означает, что

$$\#(a) - \#(b) = 0 \quad \Longrightarrow \quad \#(a) = \#(b).$$

То есть грамматика F порождает только такие строки, в которых число символов a равно числу символов b.

Для дополнительной уверенности была проведена практическая проверка с перебором выводов грамматики F с ограничением на 20 шагов.
Результатом вышло 250952 терминальных цепочек и все из них сбалансированы.

6. Покажите, что грамматика F, определенная в задаче 5, неоднозначна.

Грамматика порождает язык строк с равным числом символов a и b. При этом конструкции A → aS и B → bS позволяют встраивать поддеревья, а правила A → bAA и B → aBB допускают разбиение вывода на два независимых вывода. Таким образом, для некоторых строк существует возможность группировать подвыражения по-разному.

Доказательство:
Найдём строку

$$w \in \{a, b\}^*$$

Такую, что существует две различные последовательности применений правил, приводящие к w.

Например, пусть w — некоторая сбалансированная строка, w = abaabb.
Тогда один вариант вывода может начинаться с S → aB с последующим использованием продукции B → bS и затем развёрткой поддерева через S → bA и A → aS или A → a. Или же сначала использовать S → bA и затем применять правило A → aS с последующей заменой S → aB и выбором другого способа завершения.
В результате структура дерева оказывается различной, хотя итоговая терминальная цепочка w совпадает.

Проверим с практической стороны, найдем все терминальные цепочки и сохраним их информацию путей, если для какой-либо терминальной строки обнаруживается более 1 пути вывода, тогда грамматика неоднозначна.

Результатом вышло 5 строк на ограничение в максимум 7 шагов, все строки имеют 2 различных вывода:

- bbaaba
- bbabaa
- aabbab
- aababb

7. Грамматика H, имеющая продукции вида:

- S → aB | bAS | bA
- A → bAA | a
- B → aBB | b

a) Является ли данная грамматика однозначной?

b) Являются ли грамматики F и H эквивалентными?

a) Можно заметить альтернативные продукции для стартового символа S в грамматике H. S → aB и S → bA, аналогичные продукции, как в грамматике F. При этом добавлено правило S → bAS. Оно позволяет интерполировать дополнительное появление S внутри вывода, что может привести к тому, что одна и та же строка будет получаться либо напрямую, либо с дополнительной вложенной структурой через S → bAS, затем вывод для S завершается так же, как и в первом варианте.

Например попробуем вывести некоторую сбалансированную строку w. Тогда можно рассмотреть два варианта:

1. W получается напрямую посредством применения продукции S → bA или S → aB.
2. В процессе вывода используется правило S → bAS, а потом поддерево, отвечающее за последний S, выводится тем же способом, что и в первом варианте.

Поскольку выбор между использованием правила S → bA и S → bAS никак не влияет на конечное равновесие символов a и b, можно построить два различающихся по структуре дерева вывода для одной и той же терминальной строки. Таким образом, грамматика H неоднозначна.

b) Чтобы грамматики были эквивалентны, они должны порождать один и тот же язык. В обоих случаях по предыдущим доказательствам для F, язык, порождаемый грамматикой является множеством строк над {a, b}, в которых число символов a равно числу символов b.

В грамматике F правило S → bA | aB обеспечивает начальное парное распределение символов, а рекурсивные правила в A и B, включающие альтернативы с S, гарантируют, что при каждом шаге разность между количеством a и b сохраняется.

Грамматика H имеет аналогичные продукции. Она содержит и те же базовые альтернативы S → aB и S → bA как в F, а дополнительное правило S → bAS позволяет вводить дополнительную рекурсию, но при этом все продукции для A и B, то есть A → bAA | a и B → aBB | b, устроены так, что вклад символов остаётся таким же, как и в грамматике F.

Таким образом, обе грамматики порождают язык

$$L = \{w \in \{a, b\}^* \mid \#(a) = \#(b)\}$$

Множество строк, в которых число символов a равно числу символов b. Следовательно, несмотря на то что грамматика H неоднозначна, она порождает тот же язык, что и грамматика F.

8. Определите контекстно свободные грамматики, которые порождали бы следующие языки:

a) Все строки - элементы множества {0,1}*, такие, что в каждой из них непосредственно справа от каждого символа 0 стоит символ 1.
b) Все строки - элементы множества {0, 1}*, такие, что результаты чтения этих строк символов слева направо и справа налево совпадают.
c) Все строки - элементы множества {0,1}*, которые содержат символов 0 вдвое больше, чем символов 1.

a) Эквивалентно, что строка никогда не содержит подстроку 00. Лучший способ описать это считать 0 1 единым блоком, а 1 может встречаться и сама по себе.

S → ε | 1S | 01S, где:

- S → ε дает пустую строку ε.
- S → 1S дает одиночный символ 1.
- S → 01S дает блок 01.

b) Стандартная грамматика для палиндромов

$S \to \varepsilon \mid 0 \mid 1 \mid 0S0 \mid 1S1$, где:

- $S \to \varepsilon$ дает пустую строку ε.
- $S \to 0 \mid 1$ дает либо один символ 0 или 1.
- $S \to 0S0 \mid 1S1$ обрамляет текущий палиндром парой одинаковых символов слева и справа.

c) Лучшим решением будет рассмотреть тройки длины 3, в каждой из которых ровно две 0 и одна 1. Тогда все строки языка — это произвольная конкатенация таких троек в любом порядке, причем допускается и пустая строка.

$S \to \varepsilon \mid XS$, где

- $S \to \varepsilon$ дает пустую строку ε.
- $S \to XS$ дает сколько угодно тройных блоков X подряд.

$X \to 001 \mid 010 \mid 100$

Конкатенация этих блоки разными способами дает все возможные строки, в которых общее число 0 ровно в 2 раза превосходит общее число 1.

Вывод.

В ходе выполнения данной практической работы изучалась методика анализа контекстно-свободных грамматик, включая способы вывода строк, построение дерева вывода и проверку корректности произведенных преобразований.

Кроме того, рассматривались доказательства правильности и неоднозначности грамматик, применяя метод индукции для сохранения инварианта, а также проведения практических экспериментов перебора возможных цепочек.

В итоге установлены основные подходы к построению грамматик под заданные языковые ограничения и показано, как рассуждения о структурах правил позволяют подтвердить свойства порождаемых языков.

Ссылки.

Репозиторий с исходным кодом, использованный в данной лабораторной работе: [ссылка]