ФГБОУ ВО Московский государственный
университет им. М. В. Ломоносова
Механико-математический факультет

На правах рукописи

Петюшко Александр Александрович

БИГРАММНЫЕ ЯЗЫКИ

Специальность 01.01.09 — дискретная математика
и математическая кибернетика

Диссертация

на соискание учёной степени
кандидата физико-математических наук

Научный руководитель:
доктор физико-математических наук,
профессор Бабин Д. Н.

Москва – 2015

Содержание

Введение

Общая характеристика работы

Актуальность темы

Ещё в начале 20 века выдающимся русским учёным Марковым Андреем Андреевичем (старшим) был создан математический аппарат цепей, впоследствии названных цепями Маркова. Цепи Маркова были опробованы при вычислении переходных вероятностей между соседними буквами (биграммами) в тексте поэмы А. С. Пушкина “Евгений Онегин” [1]. В дальнейшем этот аппарат получил широкое применение для распознавания речи [2] и статистического моделирования естественных языков [3].

Содержательно, биграммный язык — это формальный язык, в котором зафиксированы количества (кратности) биграмм слов языка.

В детерминированном случае для исследования формальных языков биграммы практически не применялись. Во второй половине 70 годов 20 века в результате бурного развития методов генетики для изучения и секвенирования ДНК были опубликованы работы по подсчёту [4] точного числа ДНК-последовательностей, заданных наборами кратностей биграмм и униграмм, а также была получена верхняя асимптотическая оценка числа ДНК-последовательностей [5].

В этой ситуации естественно было бы пойти не от языка к частотам пар букв, а наоборот и изучить формальные языки с фиксированной матрицей частот. Тем самым, получилась возможность увязать свойства языков со свойствами матрицы частот. Ранее, моделированием регулярных языков с фиксированными предельными свойствами частот занимались Д. Н. Бабин и А. Б. Холоденко [6, 7].

Возможны модификации задачи исследований в области формальных

языков, заданных набором кратностей биграмм, такие как изучение спектральных свойств конечных формальных языков, заданных исключительно набором кратностей биграмм (без учёта кратностей униграмм). Или уточнение асимптотики — вместо верхней желательна асимптотически точная оценка по порядку роста.

На данный момент автору неизвестны работы по изучению бесконечных формальных языков, жёстко заданных набором кратностей биграмм. Заметим, что одним из способов задания таких бесконечных языков может служить сохранение частот пар соседних букв. В этом случае становится возможным классифицировать такие формальные языки согласно общепринятой иерархии Н. Хомского формальных языков [8].

Цель работы

- Исследовать на пустоту конечные языки, заданные набором кратностей биграмм.
- Получить аналитическую формулу для мощности конечного биграммнго языка как функцию от матрицы кратностей биграмм.
- Установить точную оценку числа слов в непустом конечном биграммном языке.
- Исследовать бесконечные языки, заданные набором кратностей биграмм, на пустоту, конечность и бесконечность.
- Найти место бесконечных биграммных языков в иерархии Н. Хомского, а также критерии, отделяющие один класс от другого.
- Исследовать биграммные языки не только на прямой, но и на окружности (т.н. биграммные языки “с закольцовыванием”).

- Исследовать взаимосвязь между между m-граммными ($m > 2$) и биграммными языками.

Научная новизна

Полученные в работе результаты являются новыми, получены автором самостоятельно. Среди них:

- Введено понятие как конечных, так и потенциально бесконечных биграммных языков, заданных исключительно матрицей кратностей биграмм.
- Получены условия пустоты, конечности и счётности биграммных языков.
- Приведена точная аналитическая формула для числа слов в конечном биграммном языке, а также точная асимптотическая оценка при длине слова, стремящейся к бесконечности.
- Получены критерии выделения в счётных биграммных языках подклассов из иерархии Н. Хомского: регулярные, контекстно-свободные и контекстно-зависимые языки. Также установлено, что других классов нет.
- Выведена асимптотика числа матриц кратностей биграмм, задающих тот или иной класс формальных языков в иерархии Н. Хомского (как конечных, так и бесконечных).
- Введено понятие биграммных языков с закольцовыванием. Установлена связь между биграммными языками с закольцовыванием и биграммными языками в случае одинакового соответствующего эйлерова графа. Поставлены и решены те же задачи, что и для биграммных языков.

- Предложен метод сведения перечисленных выше задач для языков, заданных кратностями m-грамм при $m > 2$, к соответствующим задачам для биграммных языков.

Основные методы исследования

Основными методами исследования являются: теория автоматов, теория графов, комбинаторика.

Теоретическая и практическая значимость

Работа имеет теоретический характер. Полученные в ней результаты также могут быть использованы в прикладных задачах поиска похожих фрагментов данных в системах хранения в силу хорошей скорости (матрица кратностей биграмм вычисляется за линейное время от длины входного слова) и простоты реализации (для каждого класса из иерархии Н. Хомского существует соответствующий распознаватель).

Апробация результатов

Результаты диссертации докладывались на следующих научно-исследовательских семинарах:

- Общекафедральный семинар “Теория автоматов” под руководством академика, профессора В. Б. Кудрявцева, кафедра Математической теории интеллектуальных систем Механико-математического факультета МГУ им. М. В. Ломоносова (2012–2015 гг, неоднократно)

- Научный семинар “Теория дискретных функций и приложения” под руководством профессора Д. Н. Бабина, Механико-математический факультет МГУ им. М. В. Ломоносова (2009–2015 гг, неоднократно).

Также результаты докладывались на следующих всероссийских и международных конференциях:

- X Международная конференция “Интеллектуальные системы и компьютерные науки”, Москва, Россия, 5–10 декабря 2011.
- Международная научная конференция студентов, аспирантов и молодых учёных “Ломоносов–2012”, Москва, Россия, 9–13 апреля 2012.
- XI Международный семинар “Дискретная математика и ее приложения”, Москва, Россия, 18–23 июня 2012.
- XVII Международная конференция “Проблемы теоретической кибернетики”, Казань, Россия, 16–20 июня 2014.

Структура диссертации

Диссертация состоит из введения, трёх глав, разбитых на параграфы, заключения и списка литературы, содержащего 37 наименований. Общий объем диссертации 121 страница.

Публикации

Результаты автора по теме диссертации опубликованы в 11 печатных работах [27–37], из них 7 [27–33] в научных журналах из перечня, рекомендованного ВАК РФ.

Краткое содержание работы

Во **Введении** описаны структура диссертации и история рассматриваемых в ней вопросов. Обосновываются актуальность темы и научная новизна полученных результатов. Описаны основные результаты диссертации.

В **Главе 1** введены основные понятия, касающиеся биграммных языков. Основным результатом, описанном в **Главе 1**, является классификация счётных биграммных языков согласно иерархии Н. Хомского.

Пусть A, где $|A| = n < \infty$, — конечный алфавит.

Определение 1.1. Биграммой в алфавите A называется двухбуквенное слово $ab \in A^*, a, b \in A$.

Определение 1.2. Обозначим через $\theta_{a_i a_j}(\alpha)$, где $\alpha \in A^*$, отображение $A^* \to \mathbb{N} \cup \{0\}$, сопоставляющее слову α число его подслов $a_i a_j$, т.е. количество различных разложений слова α в виде $\alpha = \alpha' a_i a_j \alpha''$ $(\alpha', \alpha'' \in A^*)$. Само же значение $\theta_{a_i a_j}(\alpha)$ назовём *кратностью биграммы* $a_i a_j$ в слове α.

Таким образом, по каждому слову $\alpha \in A^*$ можно построить квадратную *матрицу кратностей биграмм* $\Theta(\alpha) = (\theta(\alpha))_{i,j=1}^{|A|}$ размера $|A| \times |A|$, в которой на месте (i, j) будет стоять значение $\theta_{a_i a_j}(\alpha)$.

Пример. Пусть $A = \{0, 1\}$, $\alpha = 01011100$.
Тогда матрица биграмм $\Theta(\alpha) = \begin{pmatrix} \theta_{00}(\alpha) & \theta_{01}(\alpha) \\ \theta_{10}(\alpha) & \theta_{11}(\alpha) \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 2 & 2 \end{pmatrix}$.

Пусть Ξ — множество квадратных матриц размера $|A| \times |A|$ с элементами из $\mathbb{N} \cup \{0\}$. Также, здесь и далее через $\Theta(\alpha)$ будем обозначать матрицу биграмм, построенную по конкретному слову α, а через Θ — просто некоторую матрицу из Ξ, при этом будем считать, что на месте (i, j) матрицы Θ стоит значение $\theta_{a_i a_j}$.

Введём понятие простейшего биграммного языка.

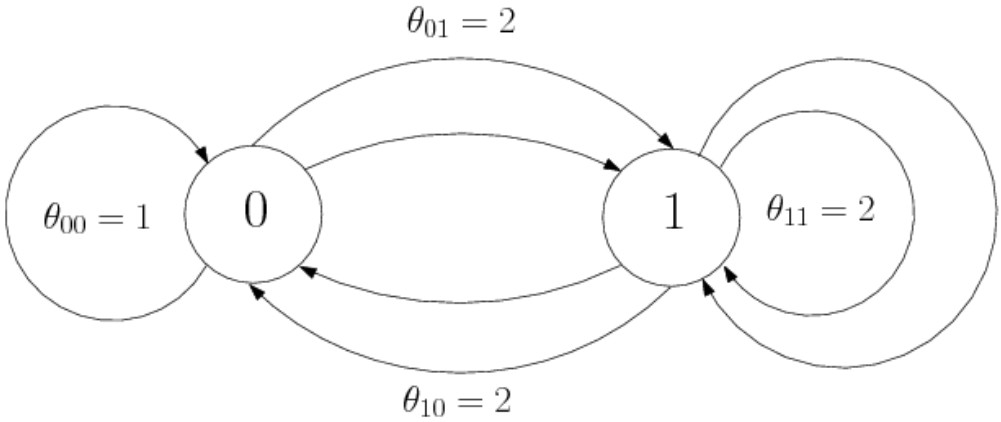
Определение 1.3. Назовём *простейшим биграммным языком* $L(\Theta)$, порождённым матрицей $\Theta \in \Xi$, множество всех слов, имеющих одну и ту же матрицу кратностей биграмм Θ, т.е. $L(\Theta) = \{\beta \in A^* \mid \Theta(\beta) = \Theta\}$.

Лемма 1.1. *Простейший биграммный язык* $L(\Theta)$ *состоит не более чем из конечного числа слов одинаковой длины* $l_\Theta = \sum\limits_{a_i,a_j \in A} \theta_{a_i a_j} + 1$.

Построим по матрице $\Theta(\alpha)$ ориентированный граф $G_{\Theta(\alpha)}$ на плоскости (аналогично строится по произвольной матрице $\Theta \in \Xi$ ориентированный граф G_Θ). Вершины — буквы из алфавита A, ребра соответствуют биграммам с учётом их кратностей.

Пример. $A = \{0, 1\}$, $\alpha = 01011100$.

$$\Theta(\alpha) = \begin{pmatrix} \theta_{00}(\alpha) & \theta_{01}(\alpha) \\ \theta_{10}(\alpha) & \theta_{11}(\alpha) \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 2 & 2 \end{pmatrix}.$$



Напомним важные известные определения [9], которые нам потребуются в дальнейшем.

Определение 1.6, 1.7. *Простым циклом* в ориентированном графе называется цикл, в котором, если два ориентированных ребра имеют одинаковое начало, то они имеют и одинаковый конец (и наоборот). *Элементарным циклом* в ориентированном графе называется простой цикл без повторяющихся рёбер.

Определение 1.10, 1.11. *Полуэйлеров граф* — граф, содержащий эйлеров путь, который не является эйлеровым циклом. *Эйлеров граф* — граф, содержащий эйлеров цикл.

Рассмотрим для начала условие непустоты $L(\Theta)$.

Лемма 1.5. *Для того, чтобы существовало хотя бы одно слово α с данной матрицей кратностей биграмм $\Theta \in \Xi$, достаточно, чтобы построенный по Θ ориентированный граф G_Θ был либо эйлеровым, либо полуэйлеровым.*

Следствие 1.5.1 (Алгоритмическая разрешимость)**.** *Задача определения того, существует ли хотя бы одно слово α с заданной матрицей кратностей биграмм Θ, алгоритмически разрешима.*

Пример. Пусть $A = \{0, 1\}$. Тогда матрица биграмм $\Theta = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ задаёт пустой язык $L(\Theta)$, поскольку в любом его слове есть как буквы "0", так и "1", а при этом переходной биграммы ("01" или "10") — нет.

Рассмотрим язык, в котором отношения $\theta_{ab}(\alpha)/\theta_{cd}(\alpha)$, где $\theta_{cd}(\alpha) > 0$, зависят только от выбора букв $a, b, c, d \in A$, но не зависят от слова α из этого языка.

Определение 1.13. Назовём *биграммным языком*, заданным матрицей кратностей биграмм $\Theta \in \Xi$, язык

$$F_\Theta = \bigcup_{k=1}^{\infty} L(k\Theta),$$

т.е. язык, состоящий из всех таких слов β, что набор кратностей биграмм этих слов $\Theta(\beta)$ кратен набору Θ, а именно, $F_\Theta = \{\beta \in A^* \mid \exists k \in \mathbb{N} : \Theta(\beta) = k\Theta\}$.

Получены условия непустоты, конечности или счетности языка F_Θ в зависимости от типа графа G_Θ:

Теорема 1.8. 1) *Если граф G_Θ — эйлеров, то в биграммном языке F_Θ счётное множество слов;*

2) *Если граф G_Θ — полуэйлеров, то биграммный язык F_Θ совпадает с $L(\Theta)$,*

и в нём конечное ненулевое число слов;

3) *Иначе биграммный язык F_Θ пуст.*

Попробуем теперь выделить классы языков согласно иерархии Н. Хомского [8] среди счётных биграммных языков. Для этого нам потребуется следующее определение:

Определение 1.14. Назовем N ненулевых матриц $\Theta_1, \ldots, \Theta_N$ из Ξ *линейно независимыми*, если не существует ненулевых действительных коэффициентов $c_1, \ldots, c_N \in \mathbb{R}, (c_1, \ldots, c_N) \neq (0, \ldots, 0)$, для которых $\sum_{i=1}^{N} c_i \Theta_i = O$, где O — нулевая матрица из Ξ.

Критерий регулярности выглядит так:

Теорема 1.9. *Пусть матрица биграмм Θ такова, что граф G_Θ является эйлеровым. Тогда:*

1) *Если существует такое разложение Θ в сумму двух ненулевых линейно независимых матриц $\Theta = \Theta_1 + \Theta_2$, что обе матрицы Θ_1 и Θ_2 задают эйлеровы графы G_{Θ_1} и G_{Θ_2}, то язык F_Θ нерегулярен;*

2) *Иначе язык F_Θ регулярен.*

Критерий контекстно-свободности:

Теорема 1.15. *Пусть матрица кратностей биграмм $\Theta \in \Xi$, задающая эйлеров граф, разлагается в сумму не менее двух линейно независимых матриц, также задающих эйлеровы граф. Тогда:*

1) *Если Θ разлагается единственным образом в сумму двух линейно независимых матриц $\Theta = \Theta_1 + \Theta_2$, соответствующих* ***простым*** *эйлеровым циклам, и не разлагается в сумму большего количества линейно независимых матриц, соответствующих эйлеровым циклам, то язык F_Θ — контекстно-свободный;*

2) *Иначе язык F_Θ — не контекстно-свободный.*

Замечание. В п. 1) вышеприведённой Теоремы F_Θ — детерминированный КС-язык (LR).

Выясняется, что все остальные счётные биграммные языки — контекстно-зависимые:

Теорема 1.17. *Бесконечный язык F_Θ, который при этом не является контекстно-свободным — контекстно-зависимый.*

Замечание. В условиях вышеприведённой Теоремы F_Θ — детерминированный КЗ-язык.

Пример. Пусть $A = \{0, 1\}$. Тогда следующие матрицы кратностей биграмм задают регулярные языки F_Θ: $\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$, $\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$ или $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$.

Пример. Пусть $A = \{0, 1\}$. Тогда следующие матрицы кратностей биграмм задают контекстно-свободные языки F_Θ: $\begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$ или $\begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}$.

Пример. Пусть $A = \{0, 1\}$. Тогда следующая матрица кратностей биграмм задаёт контекстно-зависимый язык F_Θ: $\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$.

Глава 2 посвящена мощностным формулам и оценкам для конечных биграммных языков. Также в **Главе 2** рассматривается вопрос о том, какова доля (асимптотически) матриц, задающих тот или иной бигаммный язык из найденной классификации.

Для нахождения мощности языка $L(\Theta)$ нам потребуются следующие определения и лемма.

Определение 2.1. *Матрицей Кирхгофа* $H(\Theta)$ [10], построенной по матрице биграмм $\Theta \in \Xi$, называется квадратная матрица размером $|A| \times |A|$, т.ч. на месте (i, j) стоит элемент $l_{ij} = \begin{cases} -\theta_{a_i a_j}, & i \neq j, \\ \sum_{a_j \neq a_i} \theta_{a_i a_j}, & i = j. \end{cases}$

Замечание. $\det H(\Theta) = 0$.

Лемма 2.2. *Если G_Θ — эйлеров, то все главные миноры $D^{(i,i)}(\Theta)$, полученные вычёркиванием из $H(\Theta)$ i-й строки и i-го столбца, одинаковы (и равны $D(\Theta)$).*

В итоге, мощность N_Θ языка $L(\Theta)$ выражается следующими формулами:

Теорема 2.4. *Пусть задана матрица биграмм Θ, которой соответствует эйлеров или полуэйлеров граф G_Θ, причем для $\forall i \exists j \neq i$, т.ч. $\theta_{a_i a_j} > 0$ или $\theta_{a_j a_i} > 0$. Тогда:*
1) *Если $\exists i'$, т.ч. $\sum_{a_i \in A} \theta_{a_i a_{i'}} > \sum_{a_i \in A} \theta_{a_{i'} a_i}$, то*

$$N_\Theta = \frac{\prod_{a_i \in A} \left(\sum_{a_j \in A} \theta_{a_i a_j} - 1 + \delta_{i'i} \right)!}{\prod_{a_i, a_j \in A} \theta_{a_i a_j}!} D^{(i'i')}(\Theta);$$

где $\delta_{i'i}$ — символ Кронекера, а знак "!" обозначает факториал;
2) *Если $\forall i, j$ $\sum_{a_i \in A} \theta_{a_i a_j} = \sum_{a_i \in A} \theta_{a_j a_i}$, то*

$$N_\Theta = \left(\sum_{a_i, a_j \in A} \theta_{a_i a_j} \right) \frac{\prod_{a_i \in A} \left(\sum_{a_j \in A} \theta_{a_i a_j} - 1 \right)!}{\prod_{a_i, a_j \in A} \theta_{a_i a_j}!} D(\Theta).$$

Несмотря на то, что была получена точная аналитическая формула для мощности $L(\Theta)$, она слишком сложна для практических вычислений. Представляет интерес точная асимптотическая оценка для мощности $L(k\Theta)$ при достаточно большом k.

Определение 2.2. Матрица кратностей биграмм Θ называется *положительной матрицей биграмм*, если $\forall i, j : \theta_{a_i a_j} \in \mathbb{N}$.

Теорема 2.5. *Пусть задана положительная матрица биграмм Θ с эйлеровым графом G_Θ. Тогда при $k \to \infty$ для числа слов β_k, т.ч. $\Theta(\beta_k) = k\Theta$,*

выполняется

$$N_{k\Theta} \sim c_2 * \frac{c_1^k}{k^{n(n-1)/2}},$$

где $c_1 = c_1(\Theta) > 1, c_2 = c_2(\Theta)$, $n = |A|$.

Пример. Пусть $A = \{0, 1\}$. Тогда для положительной матрицы кратностей биграмм $\Theta = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ верна следующая точная асимптотическая оценка: $N_{k\Theta} \sim \frac{1}{\pi}\frac{16^k}{k}$.

Рассмотрим теперь вопрос о том, каких же матриц "больше" (в асимптотическом смысле): задающих пустые, конечные, счётные биграммные языки, а также соотношение между регулярными, контекстно-свободными и контекстно-зависимыми языками.

Пусть Ξ_k — множество матриц размера $|A| \times |A|$, каждый элемент которых $\theta_{ij} \in \mathbb{N} \cup \{0\}, \theta_{ij} \leqslant k, k \in \mathbb{N}$.

$EMPTY(k)$ — количество матриц кратностей биграмм $\Theta \in \Xi_k$, задающих пустые языки F_Θ.

$NONEMPTY(k)$ — количество матриц кратностей биграмм $\Theta \in \Xi_k$, задающих непустые языки F_Θ.

$FINITE(k)$ — количество матриц кратностей биграмм $\Theta \in \Xi_k$, задающих конечные (непустые) языки F_Θ.

$INFINITE(k)$ — количество матриц кратностей биграмм $\Theta \in \Xi_k$, задающих счётные языки F_Θ.

$REG(k)$ — количество матриц кратностей биграмм $\Theta \in \Xi_k$, задающих счётные регулярные языки F_Θ.

$NONREG(k)$ — количество матриц кратностей биграмм $\Theta \in \Xi_k$, задающих счётные нерегулярные языки F_Θ.

$CFL(k)$ — количество матриц кратностей биграмм $\Theta \in \Xi_k$, задающих счётные КС-языки F_Θ, не являющиеся регулярными.

$CSL(k)$ — количество матриц кратностей биграмм $\Theta \in \Xi_k$, задающих счётные КЗ-языки F_Θ, не являющиеся контекстно-свободными.

$ALL(k)$ — общее количество матриц $\Theta \in \Xi_k$.

Замечание. $ALL(k) = (k+1)^{n^2}$.

Замечание. $ALL(k) = EMPTY(k) + NONEMPTY(k)$,
$NONEMPTY(k) = FINITE(k) + INFINITE(k)$,
$INFINITE(k) = REG(k) + NONREG(k)$,
$NONREG(k) = CFL(k) + CSL(k)$.

Тогда взаимные асимптотические соотношения выглядят следующим образом:

Теорема 2.6. 1) *для любого* $k \in \mathbb{N}$ $\frac{1}{n(n-1)} < \frac{INFINITE(k)}{FINITE(k)} < 1$;
2) $\lim_{k\to\infty} \frac{INFINITE(k)}{ALL(k)} = 0$;
3) $\lim_{k\to\infty} \frac{REG(k)}{NONREG(k)} = 0$;
4) $\lim_{k\to\infty} \frac{CFL(k)}{CSL(k)} = 0$.

Следствие 2.6.1. $\lim_{k\to\infty} \frac{NONEMPTY(k)}{ALL(k)} = 0$.

Глава 3 посвящена расширению понятия “биграммный язык”. Сначала достаточно подробно рассматриваются так называемые “биграммные языки с закольцовыванием” (добавляется биграмма, состоящая из последней и первой буквы), а затем показано, как для m-граммных языков свести рассмотренные задачи к таковым для биграммных языков.

Перейдём теперь к рассмотрению языков, которые заданы не на прямой, как выше, а на окружности (будем называть их “с закольцовыванием”). В этом случае для подсчёта биграмм неважно, какая первая буква в слове, а какая — последняя. Определим такие языки.

Определение 3.3. Пусть $\alpha = a_i\alpha', \alpha, \alpha' \in A^*, a_i \in A$. Назовём $\Omega(\alpha)$ *матрицей кратностей биграмм с закольцовыванием* для непустого слова $\alpha \in A^*$ следующую матрицу: $\Omega(\alpha) = \Theta(a_i\alpha' a_i)$.

Таким образом, биграммы подсчитываются не на “линейном” слове, а на слове, начало и конец которого объединены в кольцо. При этом на месте (i, j) матрицы Ω стоит значение $\omega_{a_i a_j}$.

Пример. Пусть $A = \{0, 1\}$, $\alpha = 0101110$.
Тогда матрица биграмм $\Omega_(\alpha) = \begin{pmatrix} \omega_{00}(\alpha) & \omega_{01}(\alpha) \\ \omega_{10}(\alpha) & \omega_{11}(\alpha) \end{pmatrix} = \Theta(01011100) = \begin{pmatrix} 1 & 2 \\ 2 & 2 \end{pmatrix}$.

Определение 3.4. Назовём *простейшим биграммным языком с закольцовыванием* $K(\Omega)$ множество всех слов, имеющих одну и ту же матрицу Ω кратностей биграмм с закольцовыванием, т.е. $K(\Omega) = \{\beta \in A^* \mid \Omega(\beta) = \Omega\}$.

Лемма 3.1. *Простейший биграммный язык с закольцовыванием* $K(\Omega)$ *состоит не более чем из конечного числа слов одинаковой длины* $l_\Omega = \sum\limits_{a_i, a_j \in A} \omega_{a_i a_j}$.

Аналогично Θ построим по матрице $\Omega \in \Xi$ ориентированный граф G_Ω. Условие непустоты $K(\Omega)$ несколько отличается от такового для простейших биграммных языков:

Лемма 3.3. *Для того, чтобы существовало хотя бы одно слово* α *с данной матрицей кратностей биграмм* $\Omega \in \Xi$ *с закольцовыванием, достаточно, чтобы построенный по* Ω *ориентированный граф* G_Ω *был эйлеровым.*

Следствие 3.3.1 (Алгоритмическая разрешимость)**.** *Задача определения по матрице* $\Omega \in \Xi$*, существует ли хотя бы одно слово* α*, имеющее эту матрицу биграмм с закольцовыванием, алгоритмически разрешима.*

Получено важное свойство о связи $K(\Omega)$ с $L(\Theta)$:

Теорема 3.5. *Пусть матрица* $\Omega \in \Xi$ *такова, что соответствующий ориентированный граф* G_Ω *— эйлеров. Тогда существует взаимно-однозначное соответствие между словами языков* $K(\Omega)$ *и* $L(\Theta)$*, где* $\Omega = \Theta$.

Следствие 3.5.1. *Пусть матрица $\Omega \in \Xi$ такова, что соответствующий ориентированный граф G_Ω — эйлеров. Тогда количество слов в языках $K(\Omega)$ и $L(\Theta)$, где $\Omega = \Theta$, одинаково: $|K(\Omega)| = |L(\Theta)|$.*

Таким образом, доказанные выше теоремы о мощности $L(\Theta)$ и его асимптотике без ограничений переносятся и на $K(\Omega)$, где $\Omega = \Theta$.

Аналогично F_Θ, попробуем расширить определение простейших биграммных языков с закольцовыванием для задания счётных языков:

Определение 3.5. Назовём *биграммным языком с закольцовыванием*, заданным матрицей биграмм $\Omega \in \Xi$ с закольцовыванием, язык

$$E_\Omega = \bigcup_{k=1}^{\infty} K(k\Omega),$$

т.е. язык, состоящий из всех таких слов β, что набор кратностей биграмм с закольцовыванием этих слов $\Omega(\beta)$ кратен набору Ω, а именно, $E_\Omega = \{\beta \in A^* \mid \exists k \in \mathbb{N} : \Omega(\beta) = k\Omega\}$.

Получены условия непустоты, конечности или счётности в зависимости от типа графа G_Ω:

Теорема 3.9. 1) *Если ориентированный граф G_Ω — эйлеров, то в биграммном языке E_Θ с закольцовыванием счётное множество слов;*
2) *Если ориентированный граф G_Ω — не эйлеров, то биграммный язык E_Θ с закольцовыванием пуст.*

Замечание. В отличие от биграммных языков F_Θ, биграммный язык с закольцовыванием E_Ω не может быть конечным и непустым одновременно.

Попробуем теперь выделить классы языков согласно иерархии Н. Хомского среди счётных биграммных языков с закольцовыванием, подобно тому как это было сделано для биграммных языков. Выясняется, что, несмотря на

некоторые различия в доказательствах, основные результаты формулируются подобным образом.

Критерий регулярности:

Теорема 3.10. *Пусть матрица биграмм Ω с закольцовыванием такова, что граф G_Ω является эйлеровым. Тогда:*
1) *Если существует такое разложение Ω в сумму двух ненулевых линейно независимых матриц $\Omega = \Omega_1 + \Omega_2$, что обе матрицы Ω_1 и Ω_2 задают эйлеровы графы G_{Ω_1} и G_{Ω_2}, то язык E_Ω нерегулярен;*
2) *Иначе язык E_Ω регулярен.*

Критерий контекстно-свободности:

Теорема 3.11. *Пусть матрица кратностей биграмм $\Omega \in \Xi$ с закольцовыванием, задающая эйлеров граф, разлагается в сумму не менее двух линейно независимых матриц, также задающих эйлеровы граф. Тогда:*
1) *Если Ω разлагается единственным образом в сумму двух линейно независимых матриц $\Omega = \Omega_1 + \Omega_2$, соответствующих **простым** эйлеровым циклам, и не разлагается в сумму большего количества линейно независимых матриц, соответствующих эйлеровым циклам, то язык E_Ω — контекстно-свободный;*
2) *Иначе язык E_Ω — не контекстно-свободный.*

Замечание. В п. 1) вышеприведённой Теоремы E_Ω — детерминированный КС-язык (LR).

Выясняется, что все остальные счётные биграммные языки с закольцовыванием — контекстно-зависимые:

Теорема 3.12. *Бесконечный язык E_Ω, который при этом не является контекстно-свободным — контекстно-зависимый.*

Замечание. В условиях вышеприведённой Теоремы E_Ω — детерминированный КЗ-язык.

Рассмотрим, наконец, языки, заданные не матрицей кратностей биграмм, а набором кратностей m-грамм, где $m > 2$. В общем случае это m-мерная матрица $\overline{\Theta}$ с n^m неотрицательными элементами.

Построим по этому набору граф на плоскости. Для этого воспользуемся конструкциями для т. н. графов де Брёйна [11]. Для этого из любой m-граммы $a_1a_2\dots a_{m-1}a_m$ составим две $(m-1)$-граммы: "левую" $a_1a_2\dots a_{m-1}$ и "правую" $a_2\dots a_{m-1}a_m$. Теперь нанесём на плоскость в качестве вершин ориентированного графа $G_{\overline{\Theta}}$ все получившиеся таким образом $(m-1)$-граммы, а количество ориентированных рёбер между $a_1a_2\dots a_{m-1}$ и $a_2\dots a_{m-1}a_m$ будет равно кратности m-граммы $a_1a_2\dots a_{m-1}a_m$. Заметим, что в случае $m = 2$ вышеописанная процедура приводит к построению ориентированного графа G_Θ, соответствующего матрице биграмм Θ, который был описан в начале данной работы.

Таким образом, мы получаем ориентированный граф $G_{\overline{\Theta}}$, для которого точно также могут ставиться и решаться те же вопросы, что и для биграммных языков, поскольку графовые критерии останутся точно такими же, также как и определение линейной независимости матриц (только теперь матрицы будут m-мерными). В качестве кратностей биграмм в формулах для мощностей нужно подставлять кратность m-грамм. Аналогично, можно рассматривать понятие m-граммного языка с закольцовыванием.

Единственное отличие — теперь вместо n вершин у графа G_Θ, где n — мощность алфавита A, будет n^{m-1} вершин, соответствующих "левым" и "правым" $(m-1)$-граммам.

В **Заключении** представлены основные результаты диссертации.

Благодарности

Автор выражает глубокую благодарность своему научному руководителю — доктору физико-математических наук, профессору Дмитрию Николаевичу Бабину за постановку задачи, постоянное внимание к работе и всестороннюю поддержку, а также заведующему кафедрой академику Валерию Борисовичу Кудрявцеву и всему коллективу кафедры Математической теории интеллектуальных систем за доброжелательную и творческую атмосферу.

1 Биграммные языки

1.1 Начальные определения

Напомним основные понятия, касающиеся формальных языков [12].

Алфавитом называется конечное непустое множество символов. Слово (или строка) над алфавитом A — конечная последовательность символов из A. Пустое слово, то есть слово, не содержащее символов вовсе, будем обозначать через Λ. Обычно символы алфавита обозначаются латинскими буквами a, b, c или цифрами 0, 1, в то время как греческие буквы α, β, γ будут использоваться для обозначения слов.

Конкатенация двух слов — это слово, образованное сочленением этих слов, то есть второе слово приписано сразу за первым без никаких символов между ними (например, пробелов). Например, если задан алфавит $A = \{a, b\}$ и два слова над ним $\alpha = aab, \beta = baba$, то конкатенацией слов α и β будет слово $\gamma = \alpha\beta = aabbaba$.

Обозначим через A^* множество всех конечных слов (включая пустое) над алфавитом A. Тогда для любого слова $\alpha \in A^*$ и пустого Λ будем иметь

$$\Lambda\alpha = \alpha\Lambda = \alpha.$$

Длина слова α, обозначаемая как $|\alpha|$ (или $len(\alpha)$, чтобы не путать с мощностью множества), — это количество букв в этом слове. Очевидно, что $len(\alpha) \geqslant 0$ для любого слова $\alpha \in A^*$.

Пусть k — неотрицательное целое число, а α — некоторое слово над алфавитом A. Тогда α^k — тоже слово над A, которое строится по следующему рекурсивному правилу:

1) $\alpha^0 = \Lambda$,

2) $\alpha^k = \alpha\alpha^{k-1}$ для $k > 0$.

Будем называть α подсловом β, если существуют такие слова $\gamma \in A^*, \delta \in A^*$, что $\beta = \gamma\alpha\delta$.

Формальным языком (или просто языком) L над алфавитом A называется любое множество слов над A, то есть L — это любое подмножество A^*. Пустой язык обозначается через $\varnothing$. Если L — язык, то через $|L|$ будем обозначать мощность множества L.

Конкатенацией L_1L_2 двух языков $L_1, L_2 \subseteq A^*$ называется множество

$$L_1L_2 = \{\alpha_1\alpha_2| \quad \alpha_1 \in L_1, \alpha_2 \in L_2\}.$$

При этом по определению полагается, что для любого языка L над алфавитом A верно

$$L\varnothing = \varnothing L = \varnothing.$$

Подобно тому, как было построено слово α^k для неотрицательного целого числа k, определим k-ую степень языка L в следующей рекурсивной манере:
1) $L^0 = \{\Lambda\}$,
2) $L^k = L^{k-1}L$ для $k > 0$.

Звездой (или замыканием) Клини L^* языка L называется следующее множество:

$$L^* = \bigcup_{i=0}^{\infty} L^i.$$

Таким образом, введенное вначале понятие A^* полностью согласуется с определением замыкания Клини.

Пусть A ($|A| < \infty$) — конечный алфавит.

Определение 1.1. Биграммой в алфавите A называется двухбуквенное слово $ab \in A^*, a, b \in A$ (порядок вхождения букв в биграмму имеет значение, т.е. биграмма ab не равна биграмме ba при $a \neq b$).

Определение 1.2. Обозначим через $\theta_\beta(\alpha)$, где $\beta \in A^*, \alpha \in A^*$, причем β —

непустое слово, отображение $A^* \to \mathbb{N} \cup \{0\}$, сопоставляющее слову α число подслов β в слове α, т.е. количество различных разложений слова α в виде $\alpha = \alpha'\beta\alpha''$ (α' и α'' могут быть пустыми). При длине слова α, меньшей длины слова β, значение $\theta_\beta(\alpha)$ положим равным 0. Само же значение $\theta_\beta(\alpha)$ при данных β и α назовем кратностью β в слове α.

С учетом введенных определений, по каждому слову $\alpha \in A^*$ можно построить квадратную матрицу биграмм $\Theta(\alpha) = (\theta(\alpha))_{i,j=1}^{|A|}$ размера $|A| \times |A|$, в которой на месте (i, j) будет стоять значение $\theta_{a_i a_j}(\alpha)$ (при условии, что все буквы алфавита $A = \{a_1, a_2, ..., a_{|A|}\}$ пронумерованы и нумерация зафиксирована).

Обозначим через Ξ множество квадратных матриц размера $|A| \times |A|$, каждый элемент которых является целым неотрицательным числом. Т.о., $\forall \alpha \in A^*$ имеем $\Theta(\alpha) \in \Xi$. Также, здесь и далее через $\Theta(\alpha)$ будем обозначать матрицу биграмм, построенную по конкретному слову α, а через Θ — просто некоторую матрицу из Ξ, при этом будем считать, что на месте (i, j) матрицы Θ стоит значение $\theta_{a_i a_j}$ (т.е. для произвольной матрицы из Ξ мы опустили зависимость от α как для самой матрицы биграмм, так и для отдельных ее элементов).

Определение 1.3. Назовем простейшим биграммным языком $L(\Theta)$, порожденным матрицей $\Theta \in \Xi$, множество всех слов, имеющих одну и ту же матрицу кратностей биграмм Θ, т.е. $L(\Theta) = \{\beta \in A^* \mid \Theta(\beta) = \Theta\}$.

Пример. Пусть $A = \{0, 1\}$, $\alpha = 01011100$.

Тогда матрица биграмм $\Theta(\alpha) = \begin{pmatrix} \theta_{00}(\alpha) & \theta_{01}(\alpha) \\ \theta_{10}(\alpha) & \theta_{11}(\alpha) \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 2 & 2 \end{pmatrix}$.

1.2 Свойства матрицы биграмм

Лемма 1.1. *Простейший биграммный язык $L(\Theta)$ состоит не более чем из конечного числа слов одинаковой длины*

$$l_\Theta = \sum_{a_i,a_j \in A} \theta_{a_i a_j} + 1.$$

Доказательство. Возьмем произвольное $\alpha \in L(\Theta)$, если таковое имеется (в противном случае язык $L(\Theta)$ пуст и доказывать нечего). Очевидно, что длина слова α будет не меньше 2, иначе в этом слове нельзя было бы выделить ни одной биграммы.

Пусть $\alpha = a_{i_1} a_{i_2} ... a_{i_{l_\Theta}}$, где $a_{i_j} \in A, 1 \leqslant j \leqslant l_\Theta$, а l_Θ ($l_\Theta \geqslant 2$) — это длина слова α. Тогда в слове α можно выделить $(l_\Theta - - - 1)$ биграмму: $a_{i_1} a_{i_2}$, $a_{i_2} a_{i_3}$, ..., $a_{i_{l_\Theta - 1}} a_{i_{l_\Theta}}$. Каждая такая биграмма будет добавлять единицу в соответствующей ячейке матрицы биграмм $\Theta(\alpha) = \Theta$. Значит, сумма всех элементов матрицы Θ есть ни что иное как

$$\sum_{a_i,a_j \in A} \theta_{a_i a_j} = l_\Theta - 1,$$

откуда и следует указанная в условии данной леммы формула.

Т.о., $L(\Theta)$ в случае непустоты состоит из слов одинаковой длины l_Θ, что для конечного алфавита означает конечность языка $L(\Theta)$.

■

Пусть $a \in A$. Обозначим через $n_a^{in}(\alpha), \alpha \in A^*$ количество всех двухбуквенных подслов в α, заканчивающихся на a, т.е. $n_a^{in}(\alpha) = \sum\limits_{b \in A} \theta_{ba}(\alpha)$, что будет равно сумме значений матрицы $\Theta(\alpha)$ в столбце, соответствующем символу a. Аналогичным образом определим и $n_a^{out}(\alpha), \alpha \in A^*$, как количество двухбуквенных подслов в α, начинающихся на a, т.е. $n_a^{out}(\alpha) = \sum\limits_{b \in A} \theta_{ab}(\alpha)$, что будет равно сумме значений матрицы $\Theta(\alpha)$ в строке, соответствующей символу a.

Пример. Рассмотрим то же слово, что и в предыдущем примере, а именно $A = \{0, 1\}$, $\alpha = 01011100$.

Тогда соответствующие значения будут вычисляться как

$n_1^{in}(\alpha) = \theta_{01}(\alpha) + \theta_{11}(\alpha) = 2 + 2 = 4;$
$n_1^{out}(\alpha) = \theta_{10}(\alpha) + \theta_{11}(\alpha) = 2 + 2 = 4;$
$n_0^{in}(\alpha) = \theta_{00}(\alpha) + \theta_{10}(\alpha) = 1 + 2 = 3;$
$n_0^{out}(\alpha) = \theta_{00}(\alpha) + \theta_{01}(\alpha) = 1 + 2 = 3.$

Рассмотрим некоторые свойства матрицы $\Theta(\alpha)$.

Лемма 1.2 (Условие неразрывности)**.** *Пусть задано слово* $\alpha = a_1\beta a_2$, $a_i \in A, i = 1, 2, \beta \in A^*$ *длины не менее 2 (*$len(\alpha) \geqslant 2$*), где* a_1 *и* a_2 *— соответственно первая и последняя буквы этого слова (слово* β *может быть пустым). Тогда матрица* $\Theta(\alpha)$ *обладает следующим свойством:*

$$\forall b \in A \quad n_b^{out}(\alpha) - n_b^{in}(\alpha) = \delta_{ba_1} - \delta_{ba_2}, \tag{1}$$

где δ_{ij} *— символ Кронекера (*$\delta_{ij} = 1$ *при* $i = j$, $\delta_{ij} = 0$ *при* $i \neq j$*).*

Доказательство. Пусть сначала $a_1 \neq a_2$. Рассмотрим три случая.

1) $b \neq a_1, b \neq a_2$. Очевидно, что при букве b, не совпадающей ни с начальной, ни с концевой буквой слова α, каждое такое вхождение b в α будет давать вклад 1 как в значение $n_b^{out}(\alpha)$, так и в $n_b^{in}(\alpha)$; в итоге получим $n_b^{out}(\alpha) = n_b^{in}(\alpha)$. При этом $\delta_{ba_1} = \delta_{ba_2} = 0$, и утверждение леммы выполнено ($0 = 0$).

2) $b = a_1$. Тогда каждое вхождение буквы b в α, будет давать вклад 1 либо одновременно в $n_b^{out}(\alpha)$ и $n_b^{in}(\alpha)$ (b стоит не на первом месте), либо только в $n_b^{out}(\alpha)$ (в случае нахождения b на первом месте). Значит, $n_b^{out}(\alpha) = n_b^{in}(\alpha) + 1$, при этом $\delta_{ba_1} = 1$, а $\delta_{ba_2} = 0$, и утверждение леммы выполнено ($1 = 1$).

3) $b = a_2$. В данном случае рассуждение аналогично п. 2). Каждое вхождение буквы b в α, будет давать вклад 1 либо одновременно в $n_b^{out}(\alpha)$ и $n_b^{in}(\alpha)$ (b стоит не на последнем месте), либо только в $n_b^{in}(\alpha)$ (в случае нахождения b на последнем месте). Значит, $n_b^{out}(\alpha) = n_b^{in}(\alpha) - 1$, при этом $\delta_{ba_1} = 0$, а $\delta_{ba_2} = 1$, и утверждение леммы выполнено (-1 = -1).

Если же $a_1 = a_2$, то случай 1) рассматривается аналогично, а случаи 2) и 3) объединяются в один, где мы имеем $n_b^{out}(\alpha) = n_b^{in}(\alpha)$, $\delta_{ba_1} = \delta_{ba_2} = 1$, и утверждение леммы выполнено ($0 = 0$).

■

Замечание. Доказательство этой несложной леммы, кроме того, можно найти в [22, 23].

Пример. Продолжим рассматривать все то же слово, а именно $A = \{0, 1\}$, $\alpha = 01011100$. Проверим условие неразрывности. Имеем $a_1 = a_2 = 0$. Тогда
$n_0^{out} - n_0^{in} = 3 - 3 = 0$, $\delta_{0a_1} - \delta_{0a_2} = 1 - 1 = 0$, верно;
$n_1^{out} - n_1^{in} = 4 - 4 = 0$, $\delta_{1a_1} - \delta_{1a_2} = 0 - 0 = 0$, верно.

Замечание. Условие неразрывности (1) является необходимым, но не достаточным условием существования хотя бы одного слова α с заданной матрицей биграмм $\Theta \in \Xi$. Например, если $\theta_{00}(\alpha) = \theta_{11}(\alpha) = 1$, $\theta_{01}(\alpha) = \theta_{10}(\alpha) = 0$ в алфавите $A = \{0, 1\}$, то очевидно, что слова α с данным набором кратностей биграмм не существует — хотя бы потому, что где-то в слове должны рядом стоять буквы 1 и 0, и должна быть ненулевая кратность $\theta_{01}(\alpha)$ или $\theta_{10}(\alpha)$ (а у нас обе кратности нулевые). При этом

$$n_0^{in}(\alpha) = n_0^{out}(\alpha) = n_1^{in}(\alpha) = n_1^{out}(\alpha) = 1,$$

и если бы слово α начиналось и заканчивалось на одну и ту же букву, то условие неразрывности было бы выполнено.

Построим по матрице $\Theta(\alpha)$ (или по произвольной матрице $\Theta \in \Xi$) ориентированный граф $G_{\Theta(\alpha)}$ на плоскости. Вершинами у этого графа будут все буквы из алфавита A, при этом ребра будут соответствовать биграммам с учетом их кратностей, т.е. кратность $\theta_{ab}(\alpha)$ будет порождать $\theta_{ab}(\alpha)$ ориентированных ребер $a \to b$. Аналогично, кратность $\theta_{cc}(\alpha)$ будет порождать $\theta_{cc}(\alpha)$ петель $c \to c$.

Пример. $A = \{0, 1\}$, $\alpha = 01011100$.
$\Theta(\alpha) = \begin{pmatrix} \theta_{00}(\alpha) & \theta_{01}(\alpha) \\ \theta_{10}(\alpha) & \theta_{11}(\alpha) \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 2 & 2 \end{pmatrix}$.
Построим граф $G_{\Theta(\alpha)}$ по $\Theta(\alpha)$ — см. Рис. 1.1.

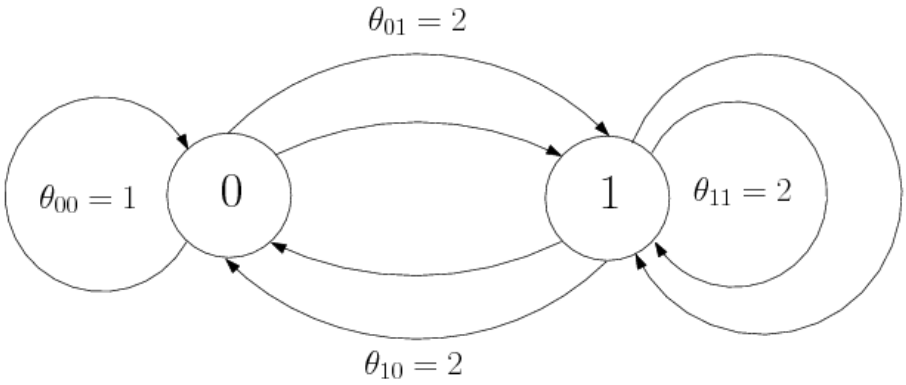


Рис. 1.1: Граф $G_{\Theta(\alpha)}$, построенный по набору $\Theta(\alpha)$

Напомним несколько широко известных понятий, касающихся эйлеровых путей.

Определение 1.4. Путем в ориентированном графе называется такая последовательность ребер этого графа, что конец предыдущего ребра совпадает с началом следующего.

Определение 1.5. Циклом в ориентированном графе называется такой путь, что начало первого ребра в этом пути совпадает с концом последнего.

Определение 1.6. Простым циклом в ориентированном графе называется цикл, в котором, если два ориентированных ребра имеют одинаковое начало, то они имеют и одинаковый конец (и наоборот).

Определение 1.7. Элементарным циклом в ориентированном графе называется простой цикл без повторяющихся ребер.

Определение 1.8. Эйлеровым путем в ориентированном графе называется такой путь, который содержит все ребра этого графа.

Определение 1.9. Эйлеровым циклом в ориентированном графе называется такой цикл, который содержит все ребра этого графа.

Определение 1.10. Полуэйлеров граф — граф, содержащий эйлеров путь, который не является эйлеровым циклом.

Определение 1.11. Эйлеров граф — граф, содержащий эйлеров цикл.

Определение 1.12. Вершина в ориентированном графе называется изолированной, если она не является концом или началом ни для одного ребра этого графа.

Замечание. На самом деле, в каноническом определении полуэйлерова графа не говорится о том, что эйлеров путь не должен являться эйлеровым циклом. Но, следуя такому определению, несложно заметить, что любой эйлеров граф является также и полуэйлеровым, поэтому каждый раз для разграничения данных понятий пришлось бы дополнять фразой „полуэйлеров граф, не являющийся эйлеровым“. Поэтому и было решено для упрощения разграничения данных объектов дать такое определение.

В [9] доказаны следующие важные теоремы, позволяющие достаточно просто проверять ориентированные графы на наличие эйлеровых путей и циклов:

Теорема 1.3. *Ориентированный граф является эйлеровым тогда и только тогда, когда выполняются следующие условия:*
1) *Все вершины, инцидентные ребрам, лежат в одной компоненте связности соответствующего неориентированного графа;*
2) *У всех вершин количество входящих ребер равно количеству исходящих ребер.*

Теорема 1.4. *Ориентированный граф является полуэйлеровым тогда и только тогда, когда выполняются следующие условия:*

1) *Все вершины, инцидентные ребрам, лежат в одной компоненте связности соответствующего неориентированного графа;*
2) *У всех вершин, кроме двух, количество входящих ребер равно количеству исходящих ребер. У оставшихся двух вершин разность количества входящих ребер и количества исходящих ребер равна +1 и -1 соответственно.*

В дальнейшем, там, где будут упоминаться понятия эйлеровых и полуэйлеровых графов, будем иметь в виду, что установить факт, является граф эйлеровым или полуэйлеровым, можно по вышеприведенным двум теоремам.

Замечание. Несложно показать, что условие неразрывности (Лемма 1.2) как раз и является условием 2) в вышеприведенных теоремах.

Лемма 1.5 (Достаточное условие существования)**.** *Для того, чтобы существовало хотя бы одно слово α с данной матрицей кратностей биграмм $\Theta \in \Xi$, достаточно, чтобы построенный по Θ ориентированный граф G_Θ был либо эйлеровым, либо полуэйлеровым.*

Доказательство. По определению, в эйлером и полуэйлеровом графах существует эйлеров путь, т.е. путь, проходящий по всем ребрам орграфа, причем только по одному разу. Пусть такой эйлеров путь задается последовательностью ребер $a_1 \to a_2, a_2 \to a_3, ..., a_{n-1} \to a_n$. Тогда слово $\alpha = a_1 a_2 ... a_{n-1} a_n$ и будет искомым, поскольку в его построении будут участвовать все ребра из G_Θ (а значит, и все ненулевые кратности биграмм из Θ), причем с учетом правильных кратностей.

■

Пример. $A = \{0, 1\}$. Пусть $\Theta = \begin{pmatrix} \theta_{00} & \theta_{01} \\ \theta_{10} & \theta_{11} \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 2 & 2 \end{pmatrix}$.
Построим граф G_Θ по Θ — см. Рис. 1.1. В этом графе можно без труда найти эйлеров путь, например, $1 \to 0 \to 0 \to 1 \to 1 \to 1 \to 0 \to 1$ (мы получили другое слово 10011101, отличное от изначального $\alpha = 01011100$), при этом все

вершины (а именно, “0” и “1”) находятся в одной компоненте связности. Значит, для данного набора Θ есть слово с такой матрицей кратностей биграмм.

Как итог, получаем следующее важное следствие:

Следствие 1.5.1 (Алгоритмическая разрешимость)**.** *Задача определения того, существует ли хотя бы одно слово α с заданной матрицей кратностей биграмм Θ, алгоритмически разрешима.*

Доказательство. Согласно Лемме 1.1 длина l_Θ искомого слова α есть $l_\Theta = 1 + \sum\limits_{a,b \in A} \theta_{ab}$.

Т.о., алгоритм можно предложить следующий. Перебираем все $|A|^{l_\Theta}$ слов длины l_Θ, для каждого такого слова β вычисляем матрицу биграмм $\Theta(\beta)$ и сравниваем с заданной матрицей Θ. Если на каком-то из $|A|^{l_\Theta}$ шагов получили совпадение, значит, искомое слово найдено. Если же за $|A|^{l_\Theta}$ шагов совпадения не было получено, то искомого слова не существует.

Однако даже при небольших значениях мощности алфавита $|A|$ перебор может представлять значительную трудность. Поэтому лучше воспользоваться результатом Леммы 1.5: построить по набору значений Θ ориентированный граф G_Θ и проверить, существует ли в нем эйлеров путь. Очевидно, эта задача алгоритмически разрешима.

■

1.3 Биграммные языки

Более интересен случай, когда мы рассматриваем матрицу биграмм не как абсолютное ограничение, а как задание относительных значений (пропорций) биграмм, то есть случай языка, в котором отношения $\theta_{ab}(\alpha)/\theta_{cd}(\alpha)$, где $\theta_{cd}(\alpha) > 0$, зависят только от выбора букв $a, b, c, d \in A$, но не зависят от слова α из этого языка. Определим такой язык.

Определение 1.13. Назовем биграммным языком, заданным матрицей кратностей биграмм $\Theta \in \Xi$, язык

$$F_\Theta = \bigcup_{k=1}^{\infty} L(k\Theta),$$

т.е. язык, состоящий из всех таких слов β, что набор кратностей биграмм этих слов $\Theta(\beta)$ кратен набору Θ, а именно, $F_\Theta = \{\beta \in A^* \mid \exists k \in \mathbb{N} : \Theta(\beta) = k\Theta\}$, где умножение k на Θ понимается как умножение скаляра на матрицу.

Рассмотрим, какие дополнительные ограничения на матрицу $\Theta(\alpha)$ налагает данное выше определение.

Лемма 1.6 (Условие неразрывности для биграммных языков)**.** *Пусть задано слово $\alpha = a_1\alpha' a_2$, $a_i \in A, i = 1, 2, \alpha' \in A^*$ длины не менее 2 ($len(\alpha) \geqslant 2$, α' может быть пустым). Для того, чтобы в биграммном языке $F_{\Theta(\alpha)}$ существовало хотя бы одно слово $\beta = b_1\beta' b_2$, $b_i \in A, i = 1, 2, \beta' \in A^*$ (β' может быть пустым), т.ч. $\Theta(\beta) = k\Theta(\alpha), k \in \mathbb{N}, k > 1$, необходимо:*

$$\forall b \in A \quad n_b^{out}(\alpha) = n_b^{in}(\alpha), \quad n_b^{out}(\beta) = n_b^{in}(\beta), \tag{2}$$

при этом как в α, так и в β первая буква должна совпадать с последней ($a_1 = a_2$ и $b_1 = b_2$).

Доказательство. Из утверждения Леммы 1.2 получаем, что $\forall b \in A \quad n_b^{out}(\beta) - n_b^{in}(\beta) = \delta_{bb_1} - \delta_{bb_2}$. Если существует такое слово β, т.ч. $\Theta(\beta) = k\Theta(\alpha), k \in \mathbb{N}, k > 1$, то, очевидно, $\forall b \in A \quad n_b^{out}(\beta) = kn_b^{out}(\alpha)$ и $n_b^{in}(\beta) = kn_b^{in}(\alpha)$. Таким образом, $k\left(n_b^{out}(\alpha) - n_b^{in}(\alpha)\right) = \delta_{bb_1} - \delta_{bb_2}$ при $k > 1$. Поскольку $\delta_{bb_1} - \delta_{bb_2} \in \{-1, 0, 1\}$, то последнее равенство выполняется только в случае $n_b^{out}(\alpha) - n_b^{in}(\alpha) = 0$ (и, следовательно, $kn_b^{out}(\alpha) = n_b^{out}(\beta) = n_b^{in}(\beta) = kn_b^{in}(\alpha)$), $\forall b \in A$ и $b_1 = b_2$. Также из условия $0 = n_b^{out}(\alpha) - n_b^{in}(\alpha) = \delta_{ba_1} - \delta_{ba_2}$ $\forall b \in A$ следует, что $a_1 = a_2$.

■

Следствие 1.6.1 (Корректность определения)**.** *Если существует хотя бы одно слово α, соответствующее набору Θ, и при этом выполняется условие неразрывности для биграммных языков (2), то для любого натурального k существует такое слово β_k, что $\Theta(\beta_k) = k\Theta(\alpha) = k\Theta$.*

Доказательство. По Лемме 1.6 слово α представимо в виде $\alpha = a\alpha'a, a \in A, \alpha' \in A^*$ (α' может быть пустым). Тогда искомое слово $\beta_k = a\underbrace{\alpha'a\alpha'a...\alpha'a}_{k-1 \quad \alpha'a}$, где подслово $\alpha'a$ приписано справа от изначального слова α $k-1$ раз. Пусть второй буквой слова α является b (возможно, она одновременно является и последней буквой при пустом α'). Тогда $\Theta(\alpha) = \Theta(ab)+\Theta(\alpha'a)$ („+“ понимается как сложение матриц), $\Theta(\beta_k) = \Theta(ab) + \Theta(\alpha'a) + ... + \Theta(ab) + \Theta(\alpha'a) = k\left(\Theta(ab) + \Theta(\alpha'a)\right) = k\Theta(\alpha) = k\Theta$. ■

Как итог, приведем достаточный признак существования такого β_k, что для любого натурального k выполняется $\Theta(\beta_k) = k\Theta$.

Лемма 1.7 (Достаточное условие существования для биграммных языков)**.** *Для того, чтобы $\forall k \in \mathbb{N}$ существовало слово β_k, т.ч. для заданного набора кратностей биграмм Θ выполнялось $\Theta(\beta_k) = k\Theta$, достаточно, чтобы построенный по Θ ориентированный граф G_Θ являлся эйлеровым графом.*

Доказательство. Возьмем в качестве слова α, имеющего матрицу биграмм Θ, некоторое слово β_1 из условия данной леммы.

Поскольку, согласно Лемме 1.6, первая и последняя буквы α должны совпадать, и при обходе нам нужно пройти через каждое ребро ровно один раз, то для существования α с заданным набором биграмм Θ достаточно, чтобы в графе G_Θ существовал эйлеров цикл (ср. с Леммой 1.5). А поскольку ориентированный граф G_Θ эйлеров, это гарантирует наличие в нем эйлерова

цикла. При этом, согласно Следствию 1.6.1, для получения β_k при $k \in \mathbb{N}$ достаточно пройти по этому циклу k раз.

■

Теорема 1.8 (О числе слов в биграммных языках)**.** *Пусть задан набор биграмм $\Theta \in \Xi$. Тогда:*

1) *Если ориентированный граф G_Θ является эйлеровым, то в биграммном языке F_Θ счетное множество слов;*

2) *Если ориентированный граф G_Θ является полуэйлеровым, то биграммный язык F_Θ совпадает с $L(\Theta)$ и в нем конечное ненулевое число слов, имеющих одинаковую длину;*

3) *Если ориентированный граф G_Θ не является ни эйлеровым, ни полуэйлеровым, то в биграммном языке F_Θ нет ни одного слова.*

Доказательство. 1) Воспользуемся Леммой 1.7. Для каждого $k \in \mathbb{N}$ будет существовать хотя бы одно слово β_k, т.ч. $\Theta(\beta_k) = k\Theta$, и, следовательно, лежащее в F_Θ. Т.к. для каждого k будет не более чем конечное количество таких β_k, а объединение счетного числа конечных множеств счетно, имеем первое утверждение теоремы.

2) По Лемме 1.5 будет существовать хотя бы одно слово α с набором кратностей $\Theta(\alpha) = \Theta$ (т.е. язык $L(\Theta)$ непуст). При этом, если существует более одного такого слова, то все они имеют одинаковую длину $len(\alpha) = l_\Theta$ согласно Лемме 1.1 (и таких слов будет конечное число). С другой стороны, по Лемме 1.6 для существования β_k $\forall k \in \mathbb{N}, k > 1$, т.ч. $\Theta(\beta_k) = k\Theta$, необходимо, чтобы $\forall b \in A \quad n_b^{out}(\alpha) = n_b^{in}(\alpha)$, что очевидным образом влечет за собой равенство у всех вершин орграфа G_Θ количества входящих и исходящих ребер, а это противоречит условию теоремы. Значит, включение $L(\Theta) \subseteq F_\Theta$ в случае полуэйлерова графа G_Θ влечет за собой равенство непустых языков $L(\Theta) = F_\Theta$; при этом в биграммном языке F_Θ конечное ненулевое число слов, имеющих длину l_Θ.

3) В этом случае по Лемме 1.5 нет ни одного слова α с матрицей кратностей биграмм Θ. Согласно Теоремам 1.3 и 1.4 граф G_Θ либо не связен, либо в нем есть вершина с разным числом входящих и исходящих ребер. Тогда для любого целого $k > 1$ граф $G_{k\Theta}$ либо не связен, либо имеет вершину, у которой разность чисел входящих и исходящих ребер по модуля не меньше $k > 1$, и по Лемме 1.5 не существует ни одного слова β_k с матрицей кратностей биграмм $k\Theta$.

Значит, в данном случае язык F_Θ пуст. ■

Пример. $A = \{0, 1\}$. Пусть $\Theta = \begin{pmatrix} \theta_{00} & \theta_{01} \\ \theta_{10} & \theta_{11} \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 2 & 2 \end{pmatrix}$.
Построим граф G_Θ по Θ — см. Рис. 1.1. В этом графе в вершину 0 входит 3 ребра, исходит тоже 3, в вершину 1 входит 4 ребра и исходит тоже 4. Все вершины лежат в одной компоненте связности. Получается, что граф G_Θ эйлеров, т.е. в Теореме 1.8 выполняется условие 1), и, соответственно, в биграммном языке F_Θ счетное множество слов.

Следствие 1.8.1 (Алгоритмическая разрешимость для биграммных языков)**.** *Задача определения того, пуст, конечен или счетен биграммный язык F_Θ с заданной матрицей кратностей биграмм $\Theta \in \Xi$, алгоритмически разрешима.*

Доказательство. В данном случае вариант доказательства (предложенный в Следствии 1.5.1), основанный на переборе, не пройдет, поскольку здесь нужно перебирать бесконечное множество слов разной длины. Поэтому обратимся ко второму варианту доказательства, основанному на рассмотрении эйлеровых путей в графе.

Для этого достаточно построить граф G_Θ и выяснить, существует ли в нем эйлеров цикл (тогда язык счетен), а в отсутствие такового эйлеров путь, не являющийся циклом (тогда язык конечен). В остальных случаях язык F_Θ пуст.

■

1.4 Регулярные биграммные языки

В данном разделе рассмотрим вопрос о регулярности бесконечных языков F_Θ с заданным набором Θ. Рассмотрим сначала общий случай с произвольным числом букв в алфавите A.

Определение 1.14. Назовем N ненулевых матриц $\Theta_1, \ldots, \Theta_N$ из Ξ линейно независимыми, если не существует ненулевых действительных коэффициентов $c_1, \ldots, c_N \in \mathbb{R}, (c_1, \ldots, c_N) \neq (0, \ldots, 0)$, для которых $\sum_{i=1}^{N} c_i \Theta_i = O$, где O — нулевая матрица из Ξ.

Определение 1.15. Назовем приведенной матрицей, соответствующей матрице биграмм Θ в алфавите $A, |A| < \infty$, матрицу $\widehat{\Theta} = \Theta/\text{НОД}(\theta_{ab} \mid \theta_{ab} > 0, a, b \in A)$.

Теорема 1.9. *Пусть $A, |A| < \infty$ — некоторый конечный алфавит. Далее, пусть задана матрица биграмм Θ такая, что соответствующий ей ориентированный граф G_Θ является эйлеровым. Тогда:*
1) *Если существует такое разложение Θ в сумму двух ненулевых линейно независимых матриц $\Theta = \Theta_1 + \Theta_2$, что обе матрицы Θ_1 и Θ_2 задают ориентированные эйлеровы графы G_{Θ_1} и G_{Θ_2}, то язык F_Θ нерегулярен;*
2) *В противном случае язык F_Θ регулярен. При этом для $\forall k \in \mathbb{N}$ существуют ровно l слов $\beta_{k,i}, i = 1..l$, т.ч. $\Theta(\beta_{k,i}) = k\Theta$, а l — число ненулевых элементов в матрице Θ.*

Доказательство. 1) Пусть существует разложение Θ в сумму двух ненулевых линейно независимых матриц $\Theta = \Theta_1 + \Theta_2$, т.ч. обе матрицы Θ_1 и Θ_2 задают ориентированные эйлеровы графы G_{Θ_1} и G_{Θ_2}. На языке графов это значит, что изначальный эйлеров цикл графа G_Θ распадается в сумму двух различных циклов, соответствующих графам G_{Θ_1} и G_{Θ_2}. При этом, поскольку изначально граф G_Θ был связным (с точностью до изолированных вершин), то графы

G_{Θ_1} и G_{Θ_2} имеют хотя бы одну общую вершину. Пусть этой вершиной будет $a \in A$.

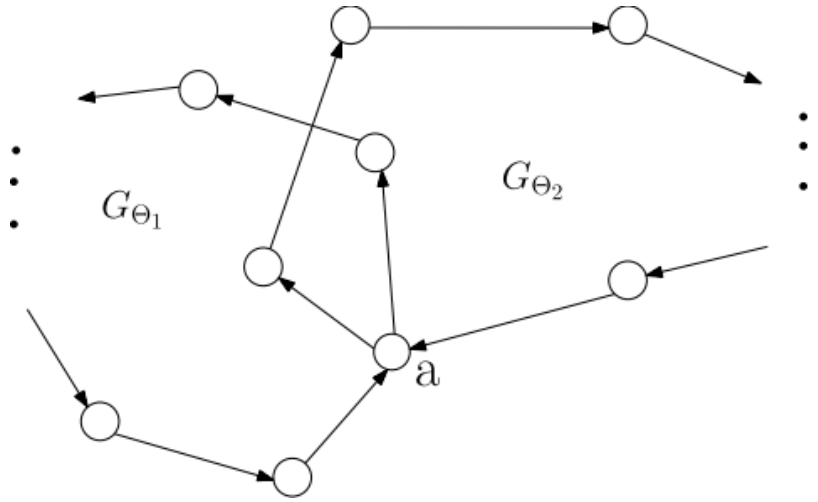


Рис. 1.2: Выделение из графа G_Θ двух линейно независимых эйлеровых циклов для случая 1) теоремы

Пусть в графе G_{Θ_1} эйлеров цикл с началом (и соответственно концом) в общей точке a задается словом $\alpha_1 = a\alpha_1'$, $\alpha_1' \in A^*$, при этом, очевидно, $\Theta(\alpha_1) = \Theta_1$. Аналогично, пусть в графе G_{Θ_2} эйлеров цикл с началом (и соответственно концом) в общей точке a задается словом $\alpha_2 = a\alpha_2'$, $\alpha_2' \in A^*$, при этом $\Theta(\alpha_2) = \Theta_2$. Тогда слово $\alpha' = a\alpha_1'\alpha_2'$ будет задавать изначальный эйлеров цикл в графе G_Θ, т.е. $\Theta(\alpha') = \Theta$. Отметим, что, поскольку матрицы Θ_1 и Θ_2 ненулевые, то и α_1', и α_2' — непустые.

Предположим, что язык F_Θ регулярен, тогда по теореме Клини [13] он представим в некотором конечно-детерминированном инициальном автомате $V_{q_0} = (A, Q, B, \varphi, \psi, q_0)$, где A — входной алфавит, Q — алфавит состояний, B — выходной алфавит (будем считать, что $B = \{0, 1\}$), $\varphi : Q \times A^* \to Q$ — функция переходов, $\psi : Q \times A^* \to B$ — функция выходов, $q_0 \in Q$ — начальное состояние [14]. При этом $\beta \in F_\Theta \Leftrightarrow \psi(q_0, \beta) = 1$. Пусть мощность состояний $|Q| = p$.

Т.к. по Лемме 1.7 для любого $k \in \mathbb{N}$ найдется слово β_k, т.ч. $\Theta(\beta_k) = k\Theta$, то зафиксируем некоторое $s > p$ и возьмем такое слово β, что $\Theta(\beta) = s\Theta$,

при этом слово β представлено в виде $\beta = a\underbrace{\alpha'_1...\alpha'_1}_{s}\underbrace{\alpha'_2...\alpha'_2}_{s}$. Это значит, что нужно сначала пройти s раз по эйлерову циклу графа G_{Θ_1} с началом в общей вершине a, после чего s раз по эйлерову циклу графа G_{Θ_2} с началом все в той же общей вершине a.

Обозначим через $q = \varphi(q_0, a)$ состояние, в которое мы попадем при подаче на вход инициального автомата V_{q_0} первой буквы a слова β. Запишем в ряд состояния, в которые мы будем переходить при последовательной подаче слова α'_1 (как подслова слова β): $q_1 = \varphi(q, \alpha'_1), q_2 = \varphi(q, \alpha'_1\alpha'_1) = \varphi(q_1, \alpha'_1), ..., q_s = \varphi(q, \underbrace{\alpha'_1...\alpha'_1}_{s}) = \varphi(q_{s-1}, \alpha'_1) = \varphi(q_0, a\underbrace{\alpha'_1...\alpha'_1}_{s})$. Поскольку $s > p$, то в этом ряду длины s будут по меньшей мере два повторяющихся состояния, т.е. $\exists i, j \in \mathbb{N}, 1 \leqslant i < j \leqslant s$, т.ч. $q_i = q_j$. Значит, для $\forall m \in \mathbb{N}$ верно тождество $q_j = \varphi(q, \underbrace{\alpha'_1..\alpha'_1}_{i}\underbrace{\alpha'_1..\alpha'_1}_{m(j-i)})$, т.к. мы будем ходить по циклу, подавая одно и то же слово α'_1 по одним и тем же состояниям $q_i, q_{i+1}, ..., q_j = q_i$.

Обозначим через $\beta'_m, m \in \mathbb{N}$ слово $\beta'_m = a\underbrace{\alpha'_1..\alpha'_1}_{s+(m-1)*(j-i)}\underbrace{\alpha'_2..\alpha'_2}_{s}$. Тогда

$$\varphi(q, \underbrace{\alpha'_1..\alpha'_1}_{j}\underbrace{\alpha'_1..\alpha'_1}_{s-j}) = \varphi(q, \underbrace{\alpha'_1..\alpha'_1}_{i}\underbrace{\alpha'_1..\alpha'_1}_{m(j-i)}\underbrace{\alpha'_1..\alpha'_1}_{s-j}) \quad \forall m \in \mathbb{N}.$$

Значит,

$$\varphi(q_0, \beta) = \varphi(\varphi(q_0, a\underbrace{\alpha'_1..\alpha'_1}_{s}), \underbrace{\alpha'_2..\alpha'_2}_{s}) = \varphi(\varphi(q_0, a\underbrace{\alpha'_1..\alpha'_1}_{s+(m-1)(j-i)}), \underbrace{\alpha'_2..\alpha'_2}_{s}) = \varphi(q_0, \beta'_m).$$

Следовательно, и $\varphi(q_0, \widetilde{\beta}) = \varphi(q_0, \widetilde{\beta'_m})$, где под $\widetilde{\delta}$ понимается слово δ без последней буквы, т.к. начиная с первого вхождения непустого подслова α'_2 в слова β и β'_m, мы будем двигаться по автомату V_{q_0} по одинаковым буквам, начиная с одинакового состояния q_s.

Т.о., $\psi(q_0, \beta'_m) = \psi(\varphi(q_0, \widetilde{\beta'_m}), a) = \psi(\varphi(q_0, \widetilde{\beta}), a) = \psi(q_0, \beta) = 1$ и, значит,

$\beta'_m \in F_{\Theta(\alpha)}$ (здесь мы использовали то, что последней буквой слов $\beta'_m, \beta, \alpha'_1$ и α'_2 является a, т.к. в эйлеровом цикле последняя вершина совпадает с начальной, которая в нашем случае является общей вершиной графов G_{Θ_1} и G_{Θ_2}).

При ненулевых и линейно независимых матрицах $\Theta(\alpha_1)$ и $\Theta(\alpha_2)$, т.ч. $\Theta(\alpha_1) + \Theta(\alpha_2) = \Theta_1 + \Theta_2 = \Theta$, имеем: $\exists a_1, a_2, a_3, a_4 \in A, (a_1, a_2) \neq (a_3, a_4)$, т.ч. $\theta_{a_1a_2}(\alpha_1) > 0$ и $\theta_{a_3a_4}(\alpha_2) > 0$ (что, в свою очередь, означает $\theta_{a_1a_2}(\alpha) = \theta_{a_1a_2}(\alpha_1) + \theta_{a_1a_2}(\alpha_2) > 0$ и $\theta_{a_3a_4}(\alpha) = \theta_{a_3a_4}(\alpha_1) + \theta_{a_3a_4}(\alpha_2) > 0$), а также не существует двух коэффициентов $c_1, c_2 \in \mathbb{R}, (c_1, c_2) \neq (0, 0)$, т.ч.

$c_1\theta_{a_1a_2}(\alpha_1) + c_2\theta_{a_1a_2}(\alpha_2) = 0,$

$c_1\theta_{a_3a_4}(\alpha_1) + c_2\theta_{a_3a_4}(\alpha_2) = 0,$

т.е. определитель

$$\begin{vmatrix} \theta_{a_1a_2}(\alpha_1) & \theta_{a_1a_2}(\alpha_2) \\ \theta_{a_3a_4}(\alpha_1) & \theta_{a_3a_4}(\alpha_2) \end{vmatrix} \neq 0 \tag{3}$$

Т.к. для $\forall \gamma \in F_\Theta \; \exists k \in \mathbb{N}$, т.ч. $\Theta(\gamma) = k\Theta$, то для всех $\gamma \in F_\Theta$ выполняется равенство $\frac{\theta_{a_1a_2}(\gamma)}{\theta_{a_3a_4}(\gamma)} = \frac{k\theta_{a_1a_2}}{k\theta_{a_3a_4}} = \frac{\theta_{a_1a_2}}{\theta_{a_3a_4}} = c = const > 0$. Рассчитаем отношение для $\beta'_m \in F_{\Theta(\alpha)}$:

$$\frac{\theta_{a_1a_2}(\beta'_m)}{\theta_{a_3a_4}(\beta'_m)} = \frac{(s + (m-1)(j-i))\theta_{a_1a_2}(\alpha_1) + s\theta_{a_1a_2}(\alpha_2)}{(s + (m-1)(j-i))\theta_{a_3a_4}(\alpha_1) + s\theta_{a_3a_4}(\alpha_2)}.$$

Т.о., это отношение имеет вид отношения двух линейных функций $\frac{ux+v}{zx+t}$ от переменной $x = s + (m-1)(j-i)$. Очевидно, что это отношение будет константным, если постоянные коэффициенты в числителе u, v будут прямо пропорциональны коэффициентам z, t в знаменателе. Значит, $\exists d \in \mathbb{R}, d > 0$, т.ч.

$\theta_{a_1a_2}(\alpha_1) = d\theta_{a_3a_4}(\alpha_1),$

$\theta_{a_1a_2}(\alpha_2) = d\theta_{a_3a_4}(\alpha_2).$

Подставляя эти выражения для расчета определителя (3), получим противоречие (две пропорциональные строки в детерминанте, значит, он нулевой).

Значит, $\beta'_m \notin F_\Theta$, противоречие с предположением о существовании автомата, представляющего множество F_Θ и т.о. регулярности F_Θ.

2) Пусть набор Θ таков, что не существует такого разложения Θ в сумму двух ненулевых линейно независимых матриц $\Theta = \Theta_1 + \Theta_2$, т.ч. обе матрицы Θ_1 и Θ_2 задают ориентированные графы G_{Θ_1} и G_{Θ_2}, которые являются эйлеровыми. Докажем, что в этом случае граф G_Θ будет представлять собой либо множественную петлю (см. Рис. 1.3 a)), либо набор „параллельных“ элементарных циклов (см. на Рис. 1.3 b)) — т. е., будет являться простым циклом.

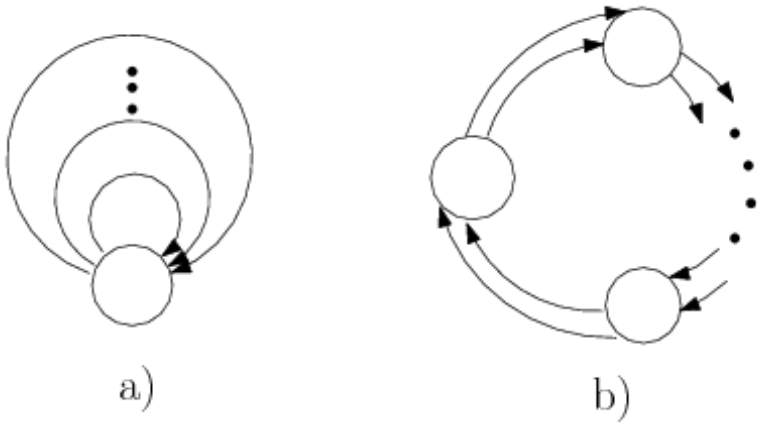


Рис. 1.3: Вид графа G_Θ для случая 2) теоремы: a) Множественная петля b) Набор параллельных элементарных циклов

Предположим, что граф G_Θ имеет отличный от изображенных на Рис. 1.3 a) и Рис. 1.3 b) вид. Значит, это либо цикл без самопересечений, но с петлями (см. Рис. 1.4 a)), либо самопересекающийся цикл (см. на Рис. 1.4 b)), либо комбинация этих двух случаев — самопересекающийся цикл с петлями.

В первом случае он разлагается на сумму цикла без самопересечений и петли (и поэтому противоречие с условием 2) теоремы), во втором — на сумму двух циклов с общей вершиной в точке пересечения (и опять противоречие о невозможности разложения в два различных цикла). Значит, предположение неверно. Более того, по условию неразрывности для биграммных языков (2), количество входящих в любую вершину ребер равно количеству исходящих

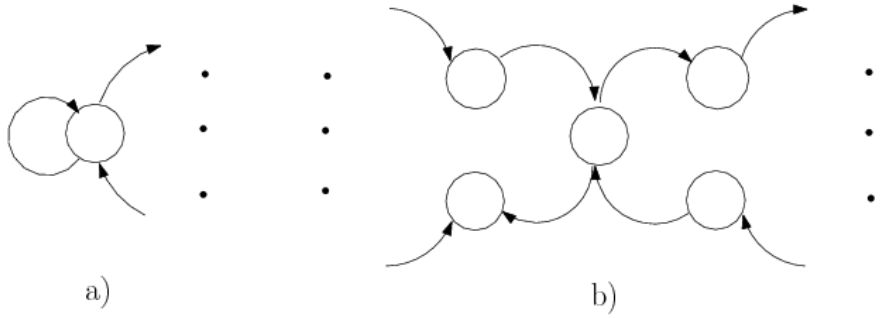


Рис. 1.4: Вид графа G_Θ в случае несовпадении эйлерова цикла с изображенными на Рис. 1.3

ребер, что дает для случая цикла без самопересечений на Рис. 1.3 b) одинаковое количество ребер между любыми двумя соединенными ребрами вершинами. Т.о., граф G_Θ будет представлять собой один из двух видов, представленных на Рис. 1.3.

Заметим, что при умножении матрицы Θ на $\forall k \in \mathbb{N}$, вид графа $G_{k\Theta}$ останется тем же, что и был для G_Θ, поскольку ребер, которые соединяют ранее не связанные вершины, при такой операции не появится, и при этом количество входящих в любую вершину ребер останется равным количеству исходящих ребер, поскольку при этом только увеличиваются в k раз кратности всех ребер.

Т.о., в матрице Θ все ненулевые элементы равны между собой, а любой эйлеров цикл для $G_{k\Theta}, \forall k \in \mathbb{N}$ будет представлять собой m раз повторенный эйлеров цикл для приведенной матрицы $\widehat{\Theta}$ (т.е. в данном случае матрицы, в которой на всех ненулевых местах единицы), где число повторений $m = k * \text{НОД}(\theta_{ab}, a, b \in A)$ (см. пример на Рис. 1.5).

При этом длина цикла для приведенной матрицы $\widehat{\Theta}$ (под длиной цикла будем понимать количество ребер в нем) будет равна в точности числу ненулевых элементов в данной матрице (и, следовательно, в матрице Θ), поскольку разные элементы матрицы соответствуют разным ребрам цикла, и наоборот.

Очевидно, что для $\forall k, k \in \mathbb{N}$ количество различных слов $\beta_k, \Theta(\beta_k) = k\Theta$

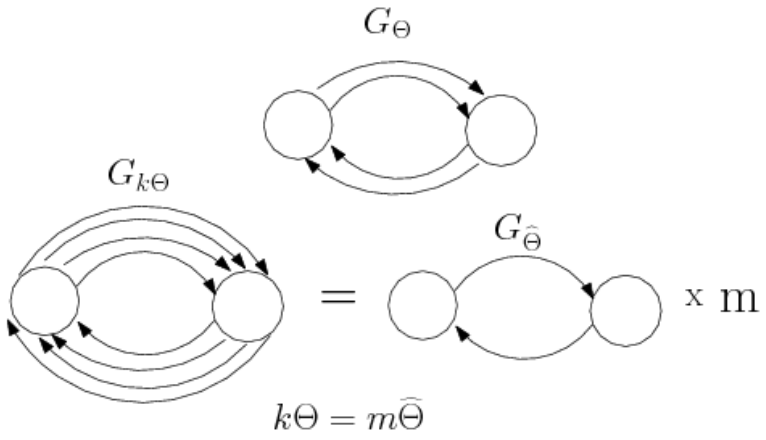


Рис. 1.5: Элементарный эйлеров цикл для k = 2, m = 4, НОД$(\theta_{ab}, a, b \in A) = 2$

будет совпадать с количеством способов выбрать первую букву (и, следовательно, начальное ориентированное ребро) в соответствующем приведенной матрице $\widehat{\Theta}$ цикле, т.к. остальные буквы уже будут однозначно определены самим эйлеровым циклом (при этом ни k, ни НОД$(\theta_{ab}, a, b \in A)$ в этом выборе не играют никакой роли). Т.о., для $\forall k \in \mathbb{N}$ существуют ровно l слов $\beta_{k,i}, i = 1..l$, т.ч. $\Theta(\beta_{k,i}) = k\Theta$, а l — число ненулевых элементов в наборе Θ.

■

Однако данная теорема даёт слишком общие условия на матрицу биграмм. Рассмотрим частный, но часто используемый на практике случай двухбуквенного алфавита.

Следствие 1.9.1. *Пусть $A = \{0, 1\}$. Далее, пусть задан такой набор Θ, что соответствующий ориентированный граф G_Θ является эйлеровым. Тогда:*

1) *Если матрица биграмм Θ имеет вид, не совпадающий ни с одним из перечисленных $M_1^{reg} = \begin{pmatrix} c_1 & 0 \\ 0 & 0 \end{pmatrix}, M_2^{reg} = \begin{pmatrix} 0 & 0 \\ 0 & c_2 \end{pmatrix}, M_3^{reg} = \begin{pmatrix} 0 & c_3 \\ c_3 & 0 \end{pmatrix}$, где $c_i \in \mathbb{N}, 1 \leqslant i \leqslant 3$, то язык F_Θ нерегулярен;*

2) *Если матрица биграмм Θ имеет вид, совпадающий с M_1^{reg} или M_2^{reg}, то язык F_Θ регулярен. При этом для $\forall k \in \mathbb{N}$ существует единственное β_k, т.ч.*

$\Theta(\beta_k) = k\Theta$;

3) *Если матрица биграмм* Θ *имеет вид, совпадающий с* M_3^{reg}, *то язык* F_Θ *регулярен. При этом для* $\forall k \in \mathbb{N}$ *существуют ровно два слова* β_k *и* γ_k, $\beta_k \neq \gamma_k$, *т.ч.* $\Theta(\beta_k) = \Theta(\gamma_k) = k\Theta$.

Пример. $A = \{0, 1\}$. Пусть $\Theta = \begin{pmatrix} \theta_{00} & \theta_{01} \\ \theta_{10} & \theta_{11} \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 2 & 2 \end{pmatrix}$.
Эта матрица задает эйлеров граф G_Θ, как было установлено в одном из предыдущих примеров. При этом вид матрицы биграмм Θ не совпадает ни с одним из M_1^{reg}, M_2^{reg} и M_3^{reg}. Значит, по Следствию 1.9.1 выполняется условие 1), и, соответственно, биграммный язык F_Θ в данном случае нерегулярен.

1.5 Контекстно-свободные биграммные языки

Для начала докажем несложное утверждение о выделении из цикла элементарного (под)цикла.

Лемма 1.10. *В любом цикле на ориентированном графе всегда содержится хотя бы один элементарный цикл.*

Доказательство. Рассмотрим новый ориентированный граф Г, ребрами которого являются ребра заданного в условии данной Леммы цикла, а вершины оставим прежними. Тогда заданный в условии Леммы цикл для нового графа будет эйлеровым, да и сам новый граф также будет эйлеровым.

Выберем любое ориентированное ребро $a_{i_1} \to a_{i_2}$. Если это петля, то доказательство закончено.

Иначе, если выбранное ребро $a_{i_1} \to a_{i_2}$ не является петлей, то $i_1 \neq i_2$. Тогда выберем любое ребро, исходящее из вершины a_{i_2}, это всегда можно сделать, поскольку иначе получается, что в вершину a_{i_2} входит как минимум одно ориентированное ребро, а не выходит ни одного, что противоречит условию на существование эйлерова цикла — см. Теорему 1.3. Пусть этим ребром будет

$a_{i_2} \to a_{i_3}$. Если $i_3 = i_2$ или $i_3 = i_1$, то в качестве искомого элементарного цикла возьмем соответственно $a_{i_2} \to a_{i_2}$ или последовательность ребер $a_{i_1} \to a_{i_2}, a_{i_2} \to a_{i_1}$, которые будут элементарными циклами по определению.

Если же $i_3 \notin \{i_1, i_2\}$, то получим путь без самопересечений $a_{i_1} \to a_{i_2}, a_{i_2} \to a_{i_3}$. Похожим образом и дальше будем выбирать следующее ребро. Пусть, например, мы уже построили путь из последовательности ребер $a_{i_1} \to a_{i_2}, a_{i_2} \to a_{i_3}, \ldots, a_{i_{j-1}} \to a_{i_j}$ такой, что все числа из набора $i_1, i_2, \ldots, i_j, j \geqslant 3$, попарно различны. Выберем следующее ребро $a_{i_j} \to a_{i_{j+1}}$, которое всегда будет существовать, иначе получим противоречие с условием Теоремы 1.3. Тогда если существует такое $1 \leqslant k \leqslant j$, что $i_{j+1} = i_k$, то по определению элементарного цикла таковым будет цикл, заданный последовательностью ребер $a_{i_k} \to a_{i_{k+1}}, a_{i_{k+1}} \to a_{i_{k+2}}, \ldots, a_{i_{j-1}} \to a_{i_j}, a_{i_j} \to a_{i_{j+1}}$.

В противном случае присоединим ребро $a_{i_j} \to a_{i_{j+1}}$ к нашему пути: $a_{i_1} \to a_{i_2}, a_{i_2} \to a_{i_3}, \ldots, a_{i_{j-1}} \to a_{i_j}, a_{i_j} \to a_{i_{j+1}}$. После этого продолжим вышеописанную процедуру нахождения следующего ребра.

Напоследок заметим, что, поскольку граф Γ конечен, равно как и количество в нем ребер, то рано или поздно мы или исчерпаем запас „неприсоединенных“ ребер и доберемся до последнего ребра $a_{i_j} \to a_{i_1}$ (соответственно, исходный цикл в условии данной Леммы и будет элементарным циклом), либо на каком-то шаге $j+1$ получим $i_{j+1} \in \{i_1, i_2, \ldots, i_j\}$ и тогда мы также получаем элементарный цикл, строго вложенный в исходный (см. рассуждение выше). При этом случая, когда последним ребром для „присоединения“ будет некое ребро $a_{i_j} \to a_{i_{j+1}}$, где $i_{j+1} \neq i_1$, быть не может, так как это противоречит определению эйлерова графа (из исходного эйлерова графа мы получили полуэйлеров).

■

Определим операцию вычитания на графах, соответствующих матрицам кратностей биграмм.

Определение 1.16. Разностью двух ориентированных графов G_{Θ_1} и G_{Θ_2}, которым взаимно-однозначно соответствуют такие матрицы кратностей биграмм Θ_1 и Θ_2, что для любых $1 \leqslant i, j \leqslant n$ справедливо $\theta_{1a_ia_j} \geqslant \theta_{2a_ia_j}$, называется такой граф G, которому соответствует матрица кратностей биграмм $\Theta = \Theta_1 - \Theta_2$. Обозначим такую операцию через $G = G_{\Theta_1} \setminus G_{\Theta_2}$.

Замечание. Несложно показать, что результат вычитания G даже для матриц кратностей биграмм Θ_1 и Θ_2, задающих непустые языки $L(\Theta_1)$ и $L(\Theta_2)$ соответственно, может соответствовать как пустому, так и непустому языку $L(\Theta)$ в зависимости от вида матриц Θ_1 и Θ_2.

Замечание. Также легко заметить, что разность двух эйлеровых графов будет одним из перечисленных вариантов: а) эйлеровым графом; б) графом, в котором более одной компоненты связности ребер, каждая из которых представляет собой ориентированный цикл; в) графом без ребер.

Определение 1.17. Расширенной разностью двух ориентированных графов G_{Θ_1} и G_{Θ_2}, которым взаимно-однозначно соответствуют такие матрицы кратностей биграмм Θ_1 и Θ_2, что для любых $1 \leqslant i, j \leqslant n$ справедливо $\theta_{1a_ia_j} \geqslant \theta_{2a_ia_j}$, называется такой граф G, которому соответствует матрица кратностей биграмм $\Theta = \Theta_1 - k\Theta_2$, $k \in \mathbb{N}$, при этом операция разности для графов G_{Θ_1} и $G_{k\Theta_2}$ еще определена, а для G_{Θ_1} и $G_{(k+1)\Theta_2}$ – уже нет. Обозначим такую операцию через $G = G_{\Theta_1} \div G_{\Theta_2}$.

Далее нам потребуется следующее важное утверждение относительно разбиения эйлерова графа на эйлеровы же подграфы.

Введем подобно Определению 1.14 линейно независимых матриц кратностей биграмм понятие независимости эйлеровых циклов.

Определение 1.18. Назовем N эйлеровых графов, заданных ненулевыми матрицами кратностей биграмм $\Theta_1, \ldots, \Theta_N$ из Ξ, независимыми, если матрицы кратностей биграмм $\Theta_1, \ldots, \Theta_N$ линейно независимы.

Теорема 1.11. *Пусть матрица кратностей биграмм $\Theta \in \Xi$, задающая эйлеров граф, разлагается в сумму трех линейно независимых матриц $\Theta = \Theta_1+\Theta_2+\Theta_3$, причем каждая из матриц $\Theta_1, \Theta_2, \Theta_3 \in \Xi$ также задает эйлеров граф. Также известно, что существует такой элементарный эйлеров цикл G_{Θ_0} с матрицей кратностей биграмм Θ_0, что расширенная разность $G_\Theta \div G_{\Theta_0}$ соответствует матрице кратностей биграмм, которая разлагается в сумму 2 линейно независимых матриц кратностей биграмм, задающих эйлеровы циклы. Тогда существуют такие ненулевые матрицы кратностей биграмм $\Theta'_1, \Theta'_2, \Theta'_3 \in \Xi$, задающие эйлеровы графы, что $\Theta = \Theta'_1 + \Theta'_2 + \Theta'_3$, и при этом существуют такие индексы $1 \leqslant i, j, k, l \leqslant n$, что $\theta'_{1a_ia_j} > 0, \theta'_{2a_ia_j} = \theta'_{3a_ia_j} = 0$ и $\theta'_{2a_ka_l} > 0, \theta'_{1a_ka_l} = \theta'_{3a_ka_l} = 0$.*

Доказательство. Для начала перефразируем условие данной теоремы в терминах эйлеровых графов, которое, собственно, и будем в дальнейшем доказывать. Необходимо доказать, что если эйлеров граф разбивается в сумму трех независимых эйлеровых графов, то существует такое разбиение в сумму трех эйлеровых графов, что два из них будут иметь уникальные (т.е. такие, каких нет у двух других) ориентированные ребра.

Возьмем в графе G_Θ такой элементарный цикл G_{Θ_0} с матрицей кратностей биграмм Θ_0, что расширенная разность $G_\Theta \div G_{\Theta_0}$ соответствует матрице кратностей биграмм, которая разлагается в сумму 2 линейно независимых эйлеровых циклов. Это можно сделать по условию теоремы.

Обозначим через $G^1 = G_\Theta \div G_{\Theta_0}$, при этом будем считать, что для некоторого $k \in \mathbb{N}$ операция разности для графов G_Θ и $G_{k\Theta_0}$ еще определена, а для G_Θ и $G_{(k+1)\Theta_0}$ – уже нет.

По построению, так как не можем больше вычитать G_{Θ_0} из G_Θ, то в G_{Θ_0} есть уникальное ребро, которого нет в G^1. Также очевидно, что в G^1 есть уникальное ребро, которого нет в G_{Θ_0}, поскольку в противном случае G^1 был бы эйлеровым циклом в графе G_{Θ_0}, который сам является элементарным, то

есть $G^1 = G_{l\Theta_0}$ для некоторого натурального l, и, следовательно, результат $G_\Theta \div G_{\Theta_0}$ был бы графом без ребер, что противоречит рассматриваемому случаю.

Если G^1 состоит из нескольких (не меньше двух) компонент связности ребер, каждая из которых представляет собой ориентированный цикл, то можем удовлетворить утверждению теоремы следующим образом. Пусть $G^1_1, \ldots, G^1_m$, $m \geqslant 2$ – компоненты связности в G^1, а их матрицы биграмм соответственно равны $\Theta^1_1, \ldots, \Theta^1_m$. По построению, в G_{Θ_0} есть уникальное ребро, которого нет в G^1 (и, следовательно, в любом из $G^1_1, \ldots, G^1_m$). Также в G^1 есть уникальное ребро, которого нет в G_{Θ_0}. Пусть это уникальное ребро из G^1 содержится в некоторой компоненте $G^1_{i_1}$ для некоторого $1 \leqslant i_1 \leqslant m$ (и при этом очевидно не содержится в остальных компонентах). Выделим также некоторую компоненту с графом $G^1_{i_2}$, $1 \leqslant i_2 \leqslant m, i_2 \neq i_1$. Тогда полагая $\Theta'_1 = k\Theta_0 + \sum_{j=1, j\notin\{i_1,i_2\}}^{m} \Theta^1_j$, $\Theta'_2 = \Theta^1_{i_1}$, $\Theta'_3 = \Theta^1_{i_2}$, и учитывая, что в $G_{\Theta'_1}$ и в $G_{\Theta'_2}$ будут уникальные ребра, при этом все три графа $G_{\Theta'_1}$, $G_{\Theta'_2}$ и $G_{\Theta'_3}$ состоят из одной компоненты связности ($G_{\Theta'_2}$ и $G_{\Theta'_3}$ – по построению, $G_{\Theta'_1}$ – так как компоненты связности $G^1_1, \ldots, G^1_m$ возникли из-за того, что мы удалили уникальные ребра $G_{k\Theta_0}$ из изначального односвязного G_Θ, и при присоединении этих ребер обратно к любому подмножеству полученных обособленных компонент связности мы всегда получим одну компоненту связности) и являются эйлеровыми (см. замечание к определению разности для эйлеровых графов), получим утверждение теоремы.

Таким образом, осталось рассмотреть случай односвязного непустого графа G^1. В G^1 есть как минимум одно уникальное ребро, которого нет в G_{Θ_0}. Обозначим все эти уникальные ребра из G^1 как $e_1, \ldots, e_r$ (параллельные ребра будем считать за одно), где $r \geqslant 1$ – их количество. Рассмотрим теперь два случая.

1) Если существует элементарный цикл в G^1 – обозначим соответствующий

этому циклу граф через G^2 с матрицей биграмм Θ^2 – который не содержит хотя бы одно ребро из множества $\{e_1, \ldots, e_r\}$ (пусть этим ребром будет $e_i, 1 \leqslant i \leqslant r$), то в случае односвязного графа $G^3 = G^1 \setminus G^2$ с матрицей биграмм Θ^3 мы получим утверждение теоремы, если в качестве новых трех матриц биграмм возьмем: $\Theta'_1 = k\Theta_0$ (в $G_{\Theta'_1}$ есть уникальные ребра по построению), $\Theta'_2 = \Theta^3$ ($G_{\Theta'_2}$ односвязен и содержит уникальное ребро e_i), $\Theta'_3 = \Theta^2$ ($G_{\Theta'_3}$ – элементарный цикл). В случае же многосвязного графа G^3 пусть компонента связности, содержащая e_i, задается графом G^3_0 с матрицей биграмм Θ^3_0. Тогда в качестве новых трех матриц биграмм возьмем: $\Theta'_1 = k\Theta_0$ (в $G_{\Theta'_1}$ есть уникальные ребра по построению), $\Theta'_2 = \Theta^3_0$ ($G_{\Theta'_2}$ содержит уникальное ребро e_i), $\Theta'_3 = \Theta - k\Theta_0 - \Theta^3_0$ (поскольку граф $G_{\Theta'_3} = G^1 \setminus G^3_0$ является односвязным и эйлеровым по тем же причинам, что и $G_{\Theta'_1}$ в случае с многосвязным G^1).

2) Любой элементарный цикл в G^1 содержит в обязательном порядке все уникальные ребра $e_1, \ldots, e_r$.

Пусть существуют 2 элементарных цикла C_1, C_2 в G^1, заданные последовательностью ребер:

$$C_1 = e_{i_1}\delta_1 e_{i_2}\delta_2 \ldots e_{i_r}\delta_r,$$
$$C_2 = e_{i_1}\sigma_1 e_{g(i_2)}\sigma_2 \ldots e_{g(i_r)}\sigma_r,$$

где буквами $\delta_1, \ldots, \delta_r, \sigma_1, \ldots, \sigma_r$ обозначаются последовательности ребер (возможно, пустые), которые одновременно принадлежат и G^1 и G_{Θ_0}, множества $\{2, \ldots, r\}$ и $\{i_2, \ldots, i_r\}$ совпадают, $g : \{2, \ldots, r\} \mapsto \{2, \ldots, r\}$ – биекция множества $\{2, \ldots, r\}$ на себя.

Докажем, что биекция g является тождественным оператором (то есть в циклах C_1 и C_2 уникальные ребра $e_1, \ldots, e_r$ следуют в одном и том же порядке). В противном случае существует натуральные $p, q \in [2, r]$ такие, что $p < q$, и $i_p = g(i_q)$. Тогда в G^1 содержится цикл

$$C_3 = e_{i_1}\delta_1 \ldots e_{i_p}\sigma_q e_{g(i_{q+1})}\sigma_{q+1} \ldots e_{g(i_r)}\sigma_r.$$

В этом цикле всего различных уникальных ребер из начального множества $\{e_1, \ldots, e_r\}$ меньше чем $p + (r - q) < r$. Значит, в цикле C_3 нет хотя бы одного ребра из множества $\{e_1, \ldots, e_r\}$, а в любом элементарном, содержащемся в нем, не будет также этого ребра. Значит, получаем противоречие с условием данного пункта 2).

Таким образом, g – тождественный оператор и $C_2 = e_{i_1}\sigma_1 e_{i_2}\sigma_2 \ldots e_{i_r}\sigma_r$.

Рассмотрим, например, элементарный цикл C_1. В нем между ребрами из множества $\{e_1, \ldots, e_r\}$ лежат последовательности ребер (возможно, пустые), которые принадлежат элементарному циклу G_{Θ_0}. Рассмотрим, например, последовательность δ_1 между e_{i_1} и e_{i_2}. Если вершина, соответствующая концу ребра e_{i_1}, не принадлежит циклу G_{Θ_0}, то тогда единственная возможность – вершина, соответствующая началу ребра e_{i_2}, совпадает с вершиной, соответствующей концу ребра e_{i_1}, а также последовательность δ_1 — пустая. Также δ_1 будет пустой, если вершина, соответствующая началу ребра e_{i_2}, совпадает с вершиной, соответствующей концу ребра e_{i_1}, хотя и принадлежащей циклу G_{Θ_0} (поскольку иной возможный случай объединить e_{i_1} и e_{i_2} – это весь полный элементарный цикл G_{Θ_0}, но в таком случае получим цикл с самопересечениями, в то время как C_1 — элементарный). Если же вершины, соответствующие началу ребра e_{i_2} и концу ребра e_{i_1}, различны и обе принадлежат циклу G_{Θ_0}, то для получения элементарного (и соответственно без самопересечений) цикла C_1 мы имеем единственную возможность выбрать последовательность ребер δ_1 из G_{Θ_0}.

Таким образом, для любого $1 \leqslant i \leqslant r$ последовательность ребер δ_i определяется единственным образом. Значит, поскольку порядок уникальных ребер из множества $\{e_1, \ldots, e_r\}$ в C_1 и C_2 – одинаков, одинаковы и последовательности соединяющих их ребер: $\delta_1 = \sigma_1, \ldots, \delta_r = \sigma_r$. Значит, и элементарные циклы C_1 и C_2 полностью совпадают.

Получаем, что для данного пункта 2) G^1 – это простой (повторенный не

менее одного раза элементарный) цикл. Это противоречит тому, что матрица $\Theta - k\Theta_0$ (соответствующая графу G^1) разлагается в сумму 2 линейно независимых эйлеровых циклов.

В итоге, мы всегда можем разбить исходный граф на три эйлеровых графа с указанными в условии теоремы свойствами на соответствующие матрицы кратностей биграмм, откуда и следует утверждение теоремы.

■

Лемма 1.12. *Пусть матрица кратностей биграмм $\Theta \in \Xi$, задающая эйлеров граф, разлагается в сумму трех матриц $\Theta = \Theta_1 + \Theta_2 + \Theta_3$, причем каждая из матриц $\Theta_1, \Theta_2, \Theta_3 \in \Xi$ также задает эйлеров граф. При этом существуют такие индексы $1 \leqslant i, j, k, l \leqslant n$, что $\theta_{1a_ia_j} > 0, \theta_{2a_ia_j} = \theta_{3a_ia_j} = 0$ и $\theta_{2a_ka_l} > 0, \theta_{1a_ka_l} = \theta_{3a_ka_l} = 0$. Тогда существуют такие ненулевые матрицы кратностей биграмм $\Theta'_1, \Theta'_2, \Theta'_3 \in \Xi$, задающие эйлеровы графы, что $\Theta = \Theta'_1 + \Theta'_2 + \Theta'_3$, и при этом существуют такие индексы $1 \leqslant i, j, k, l \leqslant n$, что $\theta'_{1a_ia_j} > 0, \theta'_{2a_ia_j} = \theta'_{3a_ia_j} = 0$ и $\theta'_{2a_ka_l} > 0, \theta'_{1a_ka_l} = \theta'_{3a_ka_l} = 0$, а также граф $G_{\Theta'_3}$ имеет общие неизолированные вершины с графами $G_{\Theta'_1}$ и $G_{\Theta'_2}$.*

Доказательство. Как и в доказательстве Теоремы 1.11, будем оперировать не с матрицами кратностей биграмм, а с эйлеровыми графами. В этом случае необходимо доказать, что если эйлеров граф разлагается в сумму трех эйлеровых графов, два из которых обладают уникальными ребрами, то существует такое разбиение исходного графа на три эйлеровых, два из которых имеют уникальные ребра, при котором каждый из графов с уникальными ребрами имеет хотя бы одну общую неизолированную вершину с третьим эйлеровым графом.

Если исходные графы G_{Θ_1}, G_{Θ_2} и G_{Θ_3} обладают этим свойством, то берем в качестве искомых эйлеровых графов исходные, и все доказано.

В противном случае граф G_{Θ_3} имеет общую неизолированную вершину ровно с одним из графов G_{Θ_1} и G_{Θ_2} (не иметь вообще общих неизолированных

вершин граф G_{Θ_3} не может, так как по условию все три графа при суммировании их матриц биграмм дают односвязный граф) , пусть для определенности это будет G_{Θ_1}. Очевидно, что в этом случае все ребра G_{Θ_3} уникальны для графа G_{Θ_2}, и наоборот, поскольку они находятся в разных компонентах связности.

Пусть $\Theta_{13} = \Theta_1 + \Theta_3$, а эйлеров граф, соответствующий этой матрице биграмм – $G_{\Theta_{13}}$ (то есть $G_{\Theta_{13}} = G_\Theta \setminus G_{\Theta_2}$). Выделим в графе G_{Θ_3} какой-нибудь элементарный цикл, которому соответствует граф G_{Θ_0} с матрицей биграмм Θ_0. Согласно Лемме 1.10, это всегда можно сделать.

Обозначим через $G^1 = G_{\Theta_{13}} \div G_{\Theta_0}$, при этом будем считать, что для некоторого $k \in \mathbb{N}$ операция разности для графов $G_{\Theta_{13}}$ и $G_{k\Theta_0}$ еще определена, а для $G_{\Theta_{13}}$ и $G_{(k+1)\Theta_0}$ – уже нет. Граф G^1 всегда будет непустым, так как в графе G_{Θ_1} (и, соответственно, в $G_{\Theta_{13}}$) есть уникальное ребро, которого нет в G_{Θ_3} (и, соответственно, в G_{Θ_0}).

По построению, так как не можем больше вычитать G_{Θ_0} из $G_{\Theta_{13}}$, то в G_{Θ_0} есть уникальное ребро, которого нет в G^1, а также в G_{Θ_2}, поскольку они находятся в разных компонентах связности. Рассмотрим теперь два случая.

1) Граф G^1 состоит из одной компоненты связности ребер, значит, он эйлеров. В этом случае, чтобы удовлетворить утверждению теоремы, необходимо взять в качестве новых матриц биграмм: $\Theta'_3 = \Theta - k\Theta_0 - \Theta_2$ (матрица биграмм графа G^1), $\Theta'_1 = k\Theta_0$ (соответствующий граф имеет общую неизолированную вершину с G^1, а также обладает уникальными ребрами по сравнению с G^1 и G_{Θ_2}), $\Theta'_2 = \Theta_2$ (имеет общую неизолированную вершину с G^1 по предположению, а также обладает уникальными ребрами по сравнению с $G_{\Theta_{13}}$ по условию теоремы, а, значит, и с графами $G_{\Theta'_3}$ и $G_{\Theta'_1}$).

2) Пусть граф G^1 имеет более одной компоненты связности ребер. В таком случае хотя бы одна его компонента связности имеет общую неизолированную вершину с графом G_{Θ_2}, пусть она имеет матрицу кратности биграмм Θ^1, а

соответствующий граф – G_{Θ^1}.

Граф $G^2 = G_{\Theta_{13}} \setminus G_{\Theta^1}$ эйлеров и односвязен, поскольку больше одной компоненты связности G^1 возникли из-за того, что мы удалили уникальные ребра $G_{k\Theta_0}$ из изначального односвязного $G_{\Theta_{13}}$, и при присоединении этих ребер обратно к любому подмножеству полученных обособленных компонент связности мы всегда получим одну компоненту связности. Тогда, чтобы удовлетворить утверждению теоремы, необходимо взять в качестве новых матриц биграмм: $\Theta'_3 = \Theta^1$ (соответствует компоненте связности G_{Θ^1}, которая имеет общую неизолированную вершину с графом G_{Θ_2}), $\Theta'_1 = \Theta_1 + \Theta_3 - \Theta^1$ (соответствует графу G^2, который имеет общую неизолированную вершину с G_{Θ^1} по построению, при этом содержит все ребра из G_{Θ_0}, которые будут уникальными по сравнению с G_{Θ_2} по предположению (не имеют общих неизолированных вершин), и одновременно с этим которые также по построению будет уникальным по сравнению с G_{Θ^1} как отдельной компоненте связности), $\Theta'_2 = \Theta_2$ (имеет общую неизолированную вершину с G_{Θ^1} по построению, а также обладает уникальными ребрами по сравнению с $G_{\Theta_{13}}$ по условию теоремы, а, значит, и с графами $G_{\Theta'_3}$ и $G_{\Theta'_1}$)

■

Для дальнейшего изучения биграммных языков нам потребуются некоторые определения и факты, касающиеся грамматик и контекстно-свободных языков.

Определение 1.19. Грамматикой называется четвёрка $GR = (\Sigma, \Gamma, P, S)$, где

Σ – основной (или терминальный) алфавит, элементы которого называют терминалами (или терминальными символами);

Γ – вспомогательный (или нетерминальный) алфавит, элементы которого называют нетерминалами (нетерминальными или вспомогательными символами); предполагается, что $\Sigma \cap \Gamma = \varnothing$;

$S \in \Gamma$ – выделенный нетерминал, называемый аксиомой (или начальным нетерминалом);
P – конечное множество слов вида $\alpha \to \beta$ (где $\alpha \in (\Sigma \cup \Gamma)^*\Gamma(\Sigma \cup \Gamma)^*, \beta \in (\Sigma \cup \Gamma)^*$), каждое такое слово называют правилом вывода (просто правилом или продукцией); при этом $\to \notin (\Sigma \cup \Gamma)$.

Определение 1.20. Грамматика называется контекстно-свободной (КС-грамматикой), если каждое ее правило имеет вид $R \to \alpha$, где $R \in \Gamma, \alpha \in (\Sigma \cup \Gamma)^*$. Язык называется контекстно-свободным (КС-языком), если некоторая КС-грамматика его порождает.

Замечание. Следует отметить, что началом изучения различных типов грамматик (и соответствующих им языков) послужила революционная работа Н. Хомского [8], описывающая иерархию формальных языков.

Определение 1.21. Автоматом с магазинной памятью (МПА) называется семерка $M = (Q, \Sigma, \Gamma, \delta, q_0, F, \gamma_0)$, где Q — непустое конечное множество состояний, Σ — входной алфавит (с дополнительным символом конца слова $\dashv$), Γ — стековый (или магазинный) алфавит (с дополнительным символом конца стека $\triangledown$), δ — конечное множество команд вида $(q, a, B) \to (q', \gamma, \hookrightarrow)$ или $(q, a, B) \to (q', \gamma, \_)$, (где $q \in Q, a \in \Sigma, B \in \Gamma, q' \in Q, \gamma \in \Gamma^*$), иначе говоря, δ — конечное подмножество множества $(Q \times \Sigma \times \Gamma) \times (Q \times \Gamma^* \times \{\hookrightarrow, \_\})$, $q_0 \in Q$ — начальное состояние, $F \subseteq Q$ — множество заключительных состояний, $\gamma_0 \in \Gamma^*$ — начальное содержимое стека (магазина).

На каждом шаге МПА решает, следует ли сдвинуть указатель на следующую справа ячейку входной ленты ($\hookrightarrow$) или оставить его на месте ($\_$).

Определение 1.22. Автомат допускает слово, записанную на входной ленте, если он дошел до ее конца и оказался при этом в одном из заключительных состояний. Множество всех слов, допускаемых МПА M, называется языком, распознаваемым M, и обозначается $L(M)$.

Для доказательства того, что язык язвляется КС-языком, будем пользоваться следующей теоремой (см. [24], [25] и [26]):

Теорема 1.13. *Классы КС-языков и языков, распознаваемых МПА, в точности совпадают.*

Также нам понадобится лемма Бара-Хиллела, также известная как лемма о накачке для контекстно-свободных языков (см. [18] и [19]). Напомним, что через $|\alpha|$ мы обозначаем длину слова $\alpha \in A^*$.

Лемма 1.14. *Для любого КС-языка L над алфавитом A существуют натуральные числа n_1, n_2 такие, что любое слово $\omega \in L$ при $|\omega| > n_1$ представимо в виде $\omega = \delta_1\mu\delta_2\nu\delta_3$, где $\delta_1, \delta_2, \delta_3 \in A^*$, $|\mu\nu| > 0$, $|\mu\delta_2\nu| \leqslant n_2$ и слово $\delta_1\mu^k\delta_2\nu^k\delta_3 \in L$ для любого $k \geqslant 0$.*

Теорема 1.15. *Рассмотрим матрицу кратностей биграмм $\Theta \in \Xi$, задающую эйлеров граф. При этом пусть эта матрица разлагается в сумму не менее двух линейно независимых матриц, таких, что каждая из матриц разложения задает эйлеров граф. Тогда:*
1) *Если матрица кратностей биграмм Θ разлагается единственным образом в сумму двух линейно независимых матриц $\Theta = \Theta_1 + \Theta_2$, соответствующих простым эйлеровым циклам, и не разлагается в сумму большего количества линейно независимых матриц, соответствующих эйлеровым циклам, то тогда язык F_Θ — контекстно-свободный;*
2) *В противном случае язык F_Θ не является контекстно-свободным.*

Доказательство. 1) Для начала заметим, что в случае единственного разложения G_Θ в сумму двух простых эйлеровых циклов G_{Θ_1} и G_{Θ_2} возможны два случая: либо G_{Θ_1} и G_{Θ_2} имеют единственную общую неизолированную вершину, либо $k > 1$ общих неизолированных вершин $a_{i_1}, \ldots, a_{i_k}$, причем путь $a_{i_1} \to a_{i_2} \to \cdots \to a_{i_{k-1}} \to a_{i_k}$ содержится в обоих этих эйлеровых циклах.

В противном случае, если есть как минимум две общие неизолированные вершины (например, a и b), между которыми у двух простых эйлеровых циклов лежат не параллельные ребра (или пути из параллельных ребер), то существует другое разложение исходной матрицы Θ в сумму двух или более линейно независимых простых эйлеровых циклов.

Пусть для определенности натуральные m_1 и m_2 таковы, что $m_1 \leqslant m_2$, и при этом $\Theta_1 = m_1\Theta_1^0$, $\Theta_2 = m_2\Theta_2^0$, где графы $G_{\Theta_1^0}$ и $G_{\Theta_2^0}$ — элементарные циклы. Тогда для доказательства от противного (см. предыдущий абзац) достаточно предъявить другое разложение G_Θ в два ($m_1 = m_2$) и более ($m_1 < m_2$) простых линейно независимых эйлеровых циклов — см. Рис. 1.6 и Рис. 1.7.

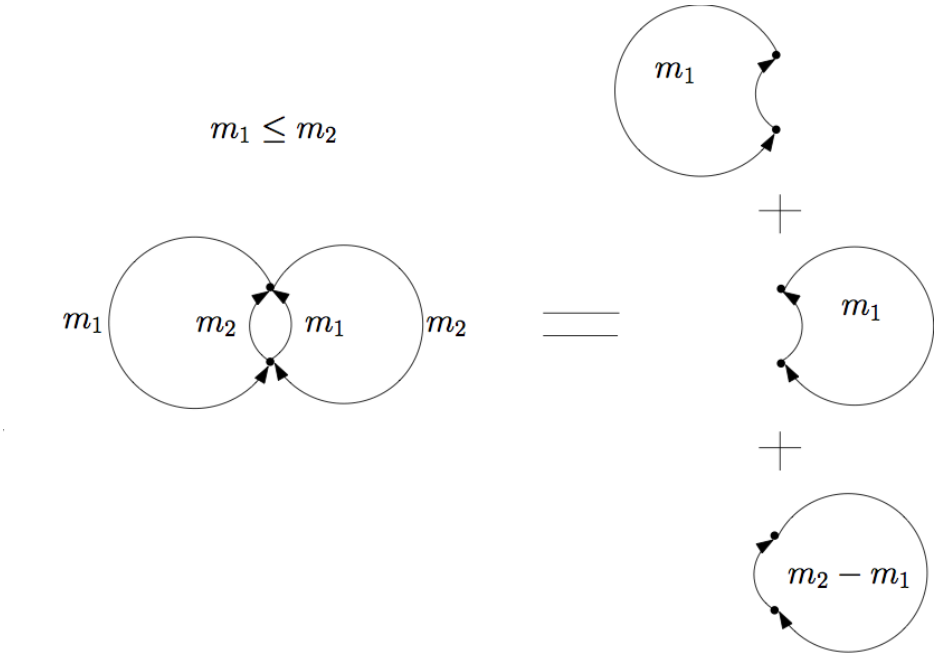


Рис. 1.6: Разложение графа G_Θ в случае “сонаправленных” циклов

Таким образом, случай единственного разложения в сумму двух простых эйлеровых циклов можно представить в общем виде так, как показано на Рис. 1.8.

Согласно Теореме 1.13, достаточно построить МПА, который распознает язык F_Θ.

В качестве входного алфавита для МПА положим $\Sigma = A \cup \{\dashv\}$, и пусть на

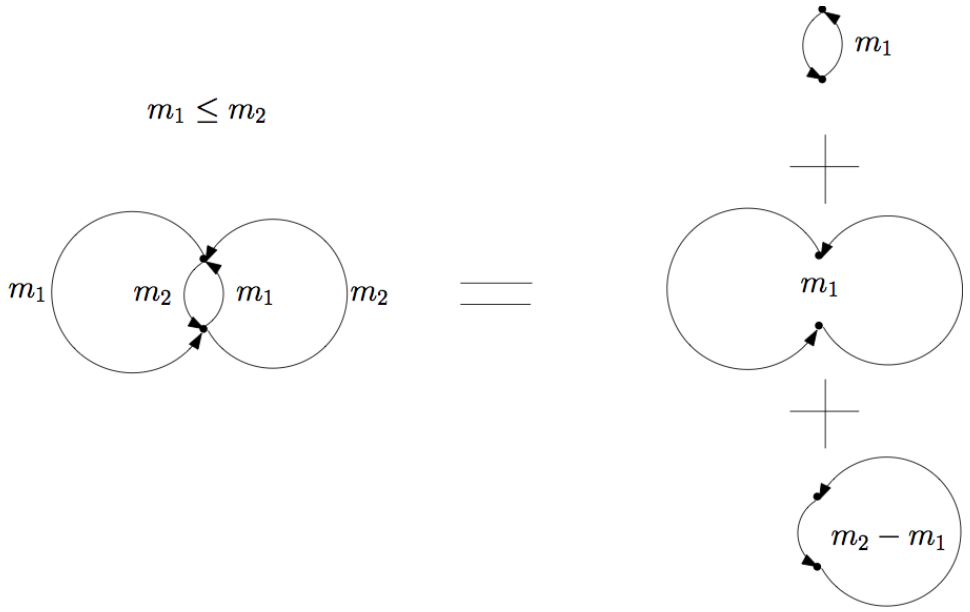


Рис. 1.7: Разложение графа G_Θ в случае “противоположно направленных” циклов

стеке в начальный момент времени находится символ $\gamma_0 = Z \in \Gamma$. Также будем считать, что алфавит стека состоит всего из трех символов: $\Gamma = \{Z, X, \triangledown\}$. Пусть $F = \{f\}$.

Построим правила перехода для распознавания языка F_Θ.

Состояния нашего МПА будем обозначать в виде $q^I(a_{prev}, C, a^1, n_1, a^2, n_2, \delta)$, где a_{prev} — предыдущая буква слова при движении слева направо, $C \in \{1, 2\}$ — номер текущего цикла, $a^i \in A$ — первая буква цикла $i \in \{1, 2\}$, $0 \leqslant n_i \leqslant m_i - 1$ — какой “виток” цикла i в данный момент совершается (начиная с нуля по модулю m_i), $\delta \in \{1, 2\}$ — какой цикл добавляет символ X в магазин (то есть, если $\delta = i$, то при прохождении цикла G_{Θ_i} символ X добавляется в магазин, а при прохождении цикла $G_{\Theta_{3-i}}$ символ X, наоборот, стирается). Также символом Y будем обозначать любой из символов X, Z (для простоты, чтобы не дублировать правила для X и Z).

Для начала рассмотрим ситуацию, когда первая буква a_i $(1 \leqslant i \leqslant k_1)$ входного слова принадлежит только первому циклу. Тогда правила будут следующими:

$(q_0, a_i, Z) \to (q^I(a_i, 1, a_i, 0, c_{k_3}, 0, 1), Z, \hookrightarrow)$,

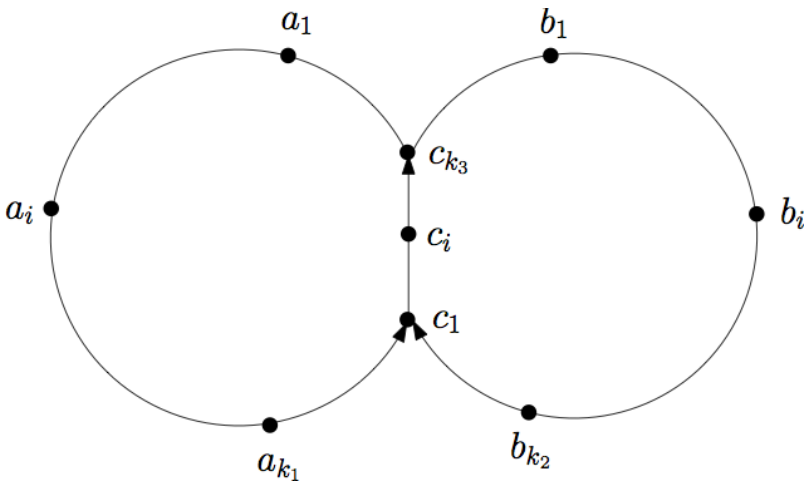


Рис. 1.8: Случай единственного разложения в сумму двух простых эйлеровых циклов

$(q^I(a_i, 1, a_i, n_1, c_{k_3}, n_2, \delta), a_{i+1}, Y) \to (q^I(a_{i+1}, 1, a_i, n_1, c_{k_3}, n_2, \delta), Y, \hookrightarrow)$,

...

$(q^I(a_{k_1-1}, 1, a_i, n_1, c_{k_3}, n_2, \delta), a_{k_1}, Y) \to (q^I(a_{k_1}, 1, a_i, n_1, c_{k_3}, n_2, \delta), Y, \hookrightarrow)$.

Таким образом, мы дошли до общей части циклов. Для нее правила будут такие:

$(q^I(a_{k_1}, C, a_i, n_1, c_{k_3}, n_2, \delta), c_1, Y) \to (q^I(c_1, C, a_i, n_1, c_{k_3}, n_2, \delta), Y, \hookrightarrow)$,

$(q^I(c_1, C, a_i, n_1, c_{k_3}, n_2, \delta), c_2, Y) \to (q^I(c_2, C, a_i, n_1, c_{k_3}, n_2, \delta), Y, \hookrightarrow)$,

...

$(q^I(c_{k_3-1}, 1, a_i, n_1, c_{k_3}, n_2, \delta), c_{k_3}, Y) \to (q^I(c_{k_3}, 1, a_i, n_1, c_{k_3}, n_2, \delta), Y, \hookrightarrow)$ (если мы двигались по первому циклу, то не меняем счетчик n_2),

$(q^I(c_{k_3-1}, 2, a_i, n_1, c_{k_3}, n_2, \delta), c_{k_3}, Y) \to (q^I(c_{k_3}, 2, a_i, n_1, c_{k_3}, n_2 + 1, \delta), Y, \hookrightarrow)$ при $n_2 < m_2 - 1$ (если мы двигались по второму циклу, то увеличиваем счетчик n_2 на единицу),

$(q^I(c_{k_3-1}, 2, a_i, n_1, c_{k_3}, m_2 - 1, 2), c_{k_3}, Y) \to (q^I(c_{k_3}, 2, a_i, n_1, c_{k_3}, 0, 2), XY, \hookrightarrow)$ (дописываем X в магазин и обнуляем счетчик n_2, так как $\delta = 2$),

$(q^I(c_{k_3-1}, 2, a_i, n_1, c_{k_3}, m_2 - 1, 1), c_{k_3}, X) \to (q^I(c_{k_3}, 2, a_i, n_1, c_{k_3}, 0, 1), \Lambda, \hookrightarrow)$ (стираем имеющийся в магазине X и обнуляем счетчик n_2, так как $\delta = 1$),

$(q^I(c_{k_3-1}, 2, a_i, n_1, c_{k_3}, m_2 - 1, 1), c_{k_3}, Z) \to (q^I(c_{k_3}, 2, a_i, n_1, c_{k_3}, 0, 2), XZ, \hookrightarrow)$

(поскольку стереть нечего, то изменяем δ с 1 на 2 и теперь уже при проходе по второму циклу мы дописываем символ X, а не стираем; также обнуляем счетчик n_2).

Теперь у нас есть выбор — либо мы продолжим путь по первому циклу, либо по второму. Рассмотрим отдельно оба варианта.

$(q^I(c_{k_3}, C, a_i, n_1, c_{k_3}, n_2, \delta), b_1, Y) \to (q^I(b_1, 2, a_i, n_1, c_{k_3}, n_2, \delta), Y, \hookrightarrow)$,

$(q^I(b_1, 2, a_i, n_1, c_{k_3}, n_2, \delta), b_2, Y) \to (q^I(b_2, 2, a_i, n_1, c_{k_3}, n_2, \delta), Y, \hookrightarrow)$,

$\ldots$

$(q^I(b_{k_2-1}, 2, a_i, n_1, c_{k_3}, n_2, \delta), b_{k_2}, Y) \to (q^I(b_{k_2}, 2, a_i, n_1, c_{k_3}, n_2, \delta), Y, \hookrightarrow)$,

$(q^I(b_{k_2}, 2, a_i, n_1, c_{k_3}, n_2, \delta), c_1, Y) \to (q^I(c_1, 2, a_i, n_1, c_{k_3}, n_2, \delta), Y, \hookrightarrow)$.

Приведенные выше правила — для второго цикла. Теперь приведем оставшийся набор правил для первого цикла.

$(q^I(c_{k_3}, C, a_i, n_1, c_{k_3}, n_2, \delta), a_1, Y) \to (q^I(a_1, 1, a_i, n_1, c_{k_3}, n_2, \delta), Y, \hookrightarrow)$,

$(q^I(a_1, 1, a_i, n_1, c_{k_3}, n_2, \delta), a_2, Y) \to (q^I(a_2, 1, a_i, n_1, c_{k_3}, n_2, \delta), Y, \hookrightarrow)$,

$\ldots$

$(q^I(a_{i-1}, 1, a_i, n_1, c_{k_3}, n_2, \delta), a_i, Y) \to (q^I(a_i, 1, a_i, n_1 + 1, c_{k_3}, n_2, \delta), Y, \hookrightarrow)$ при $n_1 < m_1 - 1$ (увеличиваем счетчик n_1 на единицу),

$(q^I(a_{i-1}, 1, a_i, m_1 - 1, c_{k_3}, n_2, 1), a_i, Y) \to (q^I(a_i, 1, a_i, 0, c_{k_3}, n_2, 1), XY, \hookrightarrow)$ (обнуляем счетчик n_1 и дописываем в магазин символ X, поскольку $\delta = 1$),

$(q^I(a_{i-1}, 1, a_i, m_1 - 1, c_{k_3}, n_2, 2), a_i, X) \to (q^I(a_i, 1, a_i, 0, c_{k_3}, n_2, 2), \Lambda, \hookrightarrow)$ (стираем имеющийся в магазине X и обнуляем счетчик n_1, так как $\delta = 2$),

$(q^I(a_{i-1}, 1, a_i, m_1 - 1, c_{k_3}, n_2, 2), a_i, Z) \to (q^I(a_i, 1, a_i, 0, c_{k_3}, n_2, 1), XZ, \hookrightarrow)$ (поскольку стереть нечего, то изменяем δ с 2 на 1 и теперь уже при проходе по первому циклу мы дописываем символ X, а не стираем; также обнуляем счетчик n_1).

Осталось дополнить правилом окончания распознавания:

$(q^I(a_i, 1, a_i, 0, c_{k_3}, 0, \delta), \dashv, Z) \to (f, Z, \_)$.

Аналогично строятся правила, если первая буква входного слова принад-

лежит только второму циклу (например, $b_i, 1 \leqslant i \leqslant k_2$). Только в этом случае все состояния будем помечать как $q^{II}(a_{prev}, C, a^1, n_1, a^2, n_2, \delta)$, чтобы правила для этих случаев не пересекались.

Таким образом, осталось рассмотреть случай, когда первая буква входного слова является общей для обоих циклов. Пусть этой буквой будет $c_i, 1 \leqslant i \leqslant k_3$. Заметим, что данный случай отличается от вышеописанных тем, что мы в самом начале не знаем, по какому циклу идем — поэтому можем проставить в качестве значения C любое число (например, 1), все равно мы его затем заменим на верное. Также, началом и концом обоих циклов будет первая буква — c_i.

$(q_0, c_i, Z) \to (q^{III}(c_i, 1, c_i, 0, c_i, 0, 1), Z, \hookrightarrow),$

$(q^{III}(c_i, C, c_i, n_1, c_i, n_2, \delta), c_{i+1}, Y) \to (q^{III}(c_{i+1}, C, c_i, n_1, c_i, n_2, \delta), Y, \hookrightarrow),$

$\ldots$

$(q^{III}(c_{k_3-1}, C, c_i, n_1, c_i, n_2, \delta), c_{k_3}, Y) \to (q^{III}(c_{k_3}, C, c_i, n_1, c_i, n_2, \delta), Y, \hookrightarrow).$

Теперь у нас есть выбор — либо мы продолжим путь по первому циклу, либо по второму. Рассмотрим отдельно оба варианта.

$(q^{III}(c_{k_3}, C, c_i, n_1, c_i, n_2, \delta), a_1, Y) \to (q^{III}(a_1, 1, c_i, n_1, c_i, n_2, \delta), Y, \hookrightarrow),$

$(q^{III}(a_1, 1, c_i, n_1, c_i, n_2, \delta), a_2, Y) \to (q^{III}(a_2, 1, c_i, n_1, c_i, n_2, \delta), Y, \hookrightarrow),$

$\ldots$

$(q^{III}(a_{k_1-1}, 1, c_i, n_1, c_i, n_2, \delta), a_{k_1}, Y) \to (q^{III}(a_{k_1}, 1, c_i, n_1, c_i, n_2, \delta), Y, \hookrightarrow),$

$(q^{III}(a_{k_1}, 1, c_i, n_1, c_i, n_2, \delta), c_1, Y) \to (q^{III}(c_1, 1, c_i, n_1, c_i, n_2, \delta), Y, \hookrightarrow).$

Аналогичные правила выписываются и для движения по второму циклу:

$(q^{III}(c_{k_3}, C, c_i, n_1, c_i, n_2, \delta), b_1, Y) \to (q^{III}(b_1, 2, c_i, n_1, c_i, n_2, \delta), Y, \hookrightarrow),$

$(q^{III}(b_1, 2, c_i, n_1, c_i, n_2, \delta), b_2, Y) \to (q^{III}(b_2, 2, c_i, n_1, c_i, n_2, \delta), Y, \hookrightarrow),$

$\ldots$

$(q^{III}(b_{k_2-1}, 2, c_i, n_1, c_i, n_2, \delta), b_{k_2}, Y) \to (q^{III}(b_{k_2}, 2, c_i, n_1, c_i, n_2, \delta), Y, \hookrightarrow),$

$(q^{III}(b_{k_2}, 2, c_i, n_1, c_i, n_2, \delta), c_1, Y) \to (q^{III}(c_1, 2, c_i, n_1, c_i, n_2, \delta), Y, \hookrightarrow).$

Теперь дополним правилами при движении по общей части циклов от c_1

до c_i:

$(q^{III}(c_1, C, c_i, n_1, c_i, n_2, \delta), c_2, Y) \to (q^{III}(c_2, C, c_i, n_1, c_i, n_2, \delta), Y, \hookrightarrow)$,

$\ldots$

$(q^{III}(c_{i-2}, C, c_i, n_1, c_i, n_2, \delta), c_{i-1}, Y) \to (q^{III}(c_{i-1}, C, c_i, n_1, c_i, n_2, \delta), Y, \hookrightarrow)$.

И завершим важными правилами, в которых обновляются счетчики (сначала для первого цикла, потом для второго):

$(q^{III}(c_{i-1}, 1, c_i, n_1, c_i, n_2, \delta), c_i, Y) \to (q^{III}(c_i, 1, c_i, n_1 + 1, c_i, n_2, \delta), Y, \hookrightarrow)$ при $n_1 < m_1 - 1$ (увеличиваем счетчик n_1 на единицу),

$(q^{III}(c_{i-1}, 1, c_i, m_1 - 1, c_i, n_2, 1), c_i, Y) \to (q^{III}(c_i, 1, c_i, 0, c_i, n_2, 1), XY, \hookrightarrow)$ (обнуляем счетчик n_1 и дописываем в магазин символ X, поскольку $\delta = 1$),

$(q^{III}(c_{i-1}, 1, c_i, m_1 - 1, c_i, n_2, 2), c_i, X) \to (q^{III}(c_i, 1, c_i, 0, c_i, n_2, 2), \Lambda, \hookrightarrow)$ (стираем имеющийся в магазине X и обнуляем счетчик n_1, так как $\delta = 2$),

$(q^{III}(c_{i-1}, 1, c_i, m_1 - 1, c_i, n_2, 2), c_i, Z) \to (q^{III}(c_i, 1, c_i, 0, c_i, n_2, 1), XZ, \hookrightarrow)$ (поскольку стереть нечего, то изменяем δ с 2 на 1 и теперь уже при проходе по первому циклу мы дописываем символ X, а не стираем; также обнуляем счетчик n_1);

$(q^{III}(c_{i-1}, 2, c_i, n_1, c_i, n_2, \delta), c_i, Y) \to (q^{III}(c_i, 2, c_i, n_1, c_i, n_2 + 1, \delta), Y, \hookrightarrow)$ при $n_2 < m_2 - 1$,

$(q^{III}(c_{i-1}, 2, c_i, n_1, c_i, m_2 - 1, 2), c_i, Y) \to (q^{III}(c_i, 2, c_i, 0, c_i, n_2, 2), XY, \hookrightarrow)$,

$(q^{III}(c_{i-1}, 2, c_i, n_1, c_i, m_2 - 1, 1), c_i, X) \to (q^{III}(c_i, 2, c_i, 0, c_i, n_2, 1), \Lambda, \hookrightarrow)$,

$(q^{III}(c_{i-1}, 2, c_i, n_1, c_i, m_2 - 1, 1), c_i, Z) \to (q^{III}(c_i, 2, c_i, 0, c_i, n_2, 2), XZ, \hookrightarrow)$.

Наконец, дополним правилом окончания распознавания:

$(q^{III}(c_i, C, c_i, 0, c_i, 0, \delta), \dashv, Z) \to (f, Z, \_)$.

Напоследок отметим две вещи. Во-первых, если в общей части циклов только одна вершина ($k_3 = 1$), либо в одном из циклов нет уникальных точек ($k_1 = 0$ или $k_2 = 0$, одновременно они не могут быть равны нулю, так как тогда получим два параллельных ребра $c_{k_3} \to c_1$, и наши циклы совпадут), то несложно заметить, что рассуждения выше не перестают быть верными, а

только немного упрощаются правила (мы же рассмотрели наиболее общий случай). Во-вторых, мощность множества состояний МПА для распознавания, который был использован в этом доказательстве, равна $12m_1m_2|A|^3$.

2) Данный случай можно разбить на три подпункта.

2.1) Пусть матрица кратностей биграмм Θ, разлагается в сумму трех линейно независимых матриц $\Theta = \Theta_1 + \Theta_2 + \Theta_3$, причем каждая из матриц $\Theta_1, \Theta_2, \Theta_3 \in \Xi$ также задает эйлеров граф. Также известно, что существует такой элементарный эйлеров цикл G_{Θ_0} с матрицей кратностей биграмм Θ_0, что расширенная разность $G_\Theta \div G_{\Theta_0}$ соответствует матрице кратностей биграмм, которая разлагается в сумму 2 линейно независимых эйлеровых циклов.

В этом случае применим последовательно сначала Теорему 1.11, а затем Лемму 1.12. Таким образом, будем иметь, что существуют такие ненулевые матрицы кратностей биграмм $\Theta'_1, \Theta'_2, \Theta'_3 \in \Xi$, задающие эйлеровы графы, что $\Theta = \Theta'_1 + \Theta'_2 + \Theta'_3$, и при этом существуют такие индексы $1 \leqslant i, j, k, l \leqslant n$, что $\theta'_{1a_ia_j} > 0, \theta'_{2a_ia_j} = \theta'_{3a_ia_j} = 0$ и $\theta'_{2a_ka_l} > 0, \theta'_{1a_ka_l} = \theta'_{3a_ka_l} = 0$, а также граф $G_{\Theta'_3}$ имеет общие неизолированные вершины с графами $G_{\Theta'_1}$ и $G_{\Theta'_2}$.

Пусть $a \in A$ – общая неизолированная вершина у графа $G_{\Theta'_3}$ с графом $G_{\Theta'_1}$, а $b \in A$ – общая неизолированная вершина у графа $G_{\Theta'_3}$ с графом $G_{\Theta'_2}$.

Возьмем по одному слову специального вида из биграммных языков, заданных каждой из матриц $\Theta'_1, \Theta'_2, \Theta'_3$. Пусть $a\alpha a \in L(\Theta'_1)$, $b\beta b \in L(\Theta'_2)$ и $a\gamma a = a\gamma_1 b\gamma_2 a \in L(\Theta'_3)$, где $\alpha, \beta, \gamma, \gamma_1, \gamma_2 \in A^*$, причем $\alpha, \beta, \gamma_1, \gamma_2$ могут быть пустыми, а γ – только в случае $a = b$. Иллюстрация данного случая показана на Рис. 1.9.

Пусть слово $\omega = a\alpha a\gamma_1 b\beta b\gamma_2 a \in L(\Theta)$. Тогда для любого натурального m слово $\omega_m = (a\alpha)^m(a\gamma_1 b\gamma_2)^{m-1}a\gamma_1(b\beta)^m b\gamma_2 a \in L(m\Theta)$ и, следовательно, $\omega_m \in F_\Theta$.

Очевидно, что существует такое натуральное число $m > 0$, что длина средней части $(a\gamma_1 b\gamma_2)^{m-1}a\gamma_1$ слова ω_m будет больше числа n_2 из Леммы 1.14,

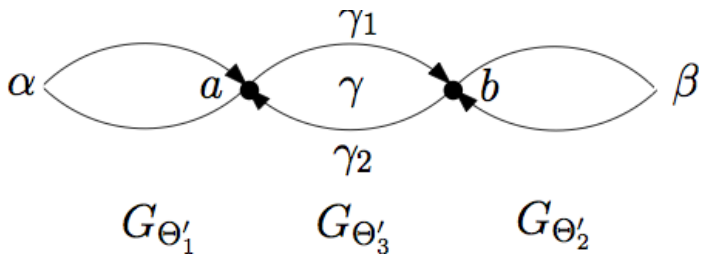


Рис. 1.9: Случай трех эйлеровых циклов, два из которых имеют уникальные ребра

при этом $|\omega_m| > n_1$. Пусть слово $\omega_m = \delta_1\mu\delta_2\nu\delta_3$, где $|\mu\nu| > 0$, $|\mu\delta_2\nu| \leqslant n_2$.

Тогда уникальные биграммы a_ia_j и a_ka_l будут в слове ω_m отстоять друг от друга не менее чем на длину средней части $(a\gamma_1 b\gamma_2)^{m-1}a\gamma_1$, то есть не менее чем на $n_2 + 1$. Значит, в разбиении $\omega_m = \delta_1\mu\delta_2\nu\delta_3$ в подслове $\mu\delta_2\nu$ не могут одновременно находится обе эти уникальные биграммы.

Согласно Лемме 1.14, если язык F_Θ – контекстно-свободный, то в нем содержится и слово $\omega_0 = \delta_1\delta_2\delta_3$ (подставляем значение $k = 0$). Из условия $|\mu\nu| > 0$ следует, что при вычеркивании подслов μ и ν из слова ω_m количество биграмм в нем, с одной стороны, уменьшится как минимум на 2 (если вычеркнута одна буква, не конечная и не начальная), а с другой стороны, увеличится на 1 (если только одно из μ и ν – непустое) или на 2 (если оба μ и ν – непустые). Если же вычеркивается конечная или начальная буква, то тогда количество биграмм уменьшается на 1, и не прибавляется вовсе.

Если ни в μ, ни в ν не входила ни одна из уникальных биграмм (имеется в виду, что уникальная биграмма была ни подсловом в μ или в ν, ни связывала эти подслова с остальной частью слова ω_m), то получим, что уникальных биграмм в ω_0 как минимум не уменьшилось, а других — как минимум уменьшилось на одну (так как суммарно биграмм стало как минимум меньше на одну). В этом случае $\omega_0 \notin F_\Theta$ и противоречие с утверждением Леммы 1.14.

Теперь расмотрим случай, когда в одном из μ и ν есть уникальная биграмма (пусть для определенности a_ia_j в μ). После вычеркивания μ и ν мы можем не добавить биграмм (если оба μ и ν — концевые), добавить одну (когда ν

пустое, либо одно из μ и ν — концевое) или две (когда ν непустое и оба из μ и ν не лежат на концах слова) биграммы.

Если не добавляем биграмм, то очевидно, что число биграмм $a_i a_j$ уменьшится на 1, а число биграмм $a_k a_l$ не изменится. Значит, $\omega_0 \notin F_\Theta$.

Если добавляем только одну биграмму, то число биграмм $a_i a_j$ как максимум останется прежним, биграмм $a_k a_l$ — как минимум не уменьшится, а всего биграмм станет меньше. Значит, $\omega_0 \notin F_\Theta$.

Если же добавляем две биграммы, то возможны два случая: а) число биграмм $a_i a_j$ как максимум останется прежним, биграмм $a_k a_l$ — как минимум не уменьшится, а всего биграмм станет меньше; б) число биграмм $a_i a_j$ увеличится на 1, а число биграмм $a_k a_l$ — останется таким же. В любом из этих случаев $\omega_0 \notin F_\Theta$ (так как кратность всех биграмм должна уменьшатся или увеличиваться в одно и то же число раз, а здесь получаем, что кратность одних увеличивается, а других — уменьшается или остается таким же). Значит, $\omega_0 \notin F_\Theta$.

Таким образом, утверждение Леммы 1.14 никогда не выполняется для $\omega_m \in F_\Theta$, и, следовательно, F_Θ – не контекстно-свободный язык.

2.2) Существуют два разных разложения матрицы кратностей биграмм Θ в сумму двух линейно независимых матриц $\Theta = \Theta_1 + \Theta_2 = \Theta_3 + \Theta_4$, где различные матрицы $\Theta_1, \Theta_2, \Theta_3, \Theta_4$ соответствующих простым эйлеровым циклам, и при этом Θ не разлагается в сумму большего количества линейно независимых матриц, соответствующих эйлеровым циклам.

Для начала установим пару важных следствий из разложения $\Theta = \Theta_1 + \Theta_2 = \Theta_3 + \Theta_4$ в пары линейно независимых простых эйлеровых циклов.

Во-первых, если для одного разложения (допустим, это разложение $\Theta = \Theta_1 + \Theta_2$) оба простых цикла содержат ребро e, то это же ребро будут содержать оба простых цикла для другого разложения (в данном случае это $\Theta = \Theta_3 + \Theta_4$). Это так, поскольку, двигаясь по направлению ребра e, мы рано или поздно

придем в вершину, из которой исходит два непараллельных ребра e_1 и e_2, $e_1 \neq e_2$. Для любого разложения Θ в любом простом цикле и из любой вершины исходят только одно ребро (возможно, с соответствующей кратностью). Значит, эти два непараллельных ребра принадлежат разным простым циклам, и, следовательно, путь до них, начиная с ориентированного ребра e, содержал ребра, принадлежащие обоим циклам (в данном случае соответствующих матрицам Θ_3 и Θ_4).

Во-вторых, если a — общая неизолированная вершина для циклов Θ_1 и Θ_2, то она же будет общей неизолированной и для циклов Θ_3 и Θ_4. Если a принадлежит общему ребру e для Θ_1 и Θ_2, то по доказанному выше a принадлежит тому же общему ребру e для Θ_3 и Θ_4. Если же a не принадлежит никакому общему ребру для Θ_1 и Θ_2, то, очевидно, что в вершину a входят два непараллельных ребра, и выходят два непараллельных ребра. Очевидно, что эти четыре ребра не могут входить в один простой цикл, а должны принадлежать двум различным простым эйлеровым циклам. Значит, вершина a будет общей и для Θ_3 и Θ_4.

Рассмотрим произвольное разложение, пусть это будет $\Theta = \Theta_1 + \Theta_2$. Уже неоднократно показывалось, что как в G_{Θ_1}, так и в G_{Θ_2} должны быть так называемые уникальные ребра, которых нет в цикле-дополнении до G_Θ. Рассмотрим произвольный простой цикл из другого разложения, пусть это будет G_{Θ_3}. С одной стороны, он должен содержать уникальные ребра как из G_{Θ_1}, так и из G_{Θ_2}, ведь в противном случае он содержал бы ребра только из одного простого цикла (например, из G_{Θ_1}) и таким образом совпадал бы с ним. С другой стороны, G_{Θ_3} не может содержать все уникальные ребра ни G_{Θ_1}, ни G_{Θ_2}, так как в противном случае по доказанному он также содержит общие ребра для G_{Θ_1} и G_{Θ_2} и, следовательно, совпадает с одним из простых циклов из первого разложения.

Значит, оба простых цикла из одного разложения содержат уникальные

ребра для любого из простых циклов второго разложения (и наоборот). Теперь можно свести этот случай к п. 2.1).

Пусть $a \in A$ – общая неизолированная вершина у простого цикла G_{Θ_1} с простыми циклами G_{Θ_2}, G_{Θ_3} и G_{Θ_4} (то, что такая вершина существует, было показано выше).

Возьмем по одному слову специального вида из биграммных языков, заданных каждой из матриц $\Theta_1, \Theta_2, \Theta_3, \Theta_4$. Пусть $a\alpha a \in L(\Theta_1)$, $a\beta a \in L(\Theta_2)$, $a\gamma a \in L(\Theta_3)$ и $a\delta a \in L(\Theta_4)$, где $\alpha, \beta, \gamma, \delta \in A^*$. Данный случай проиллюстрирован на Рис. 1.10.

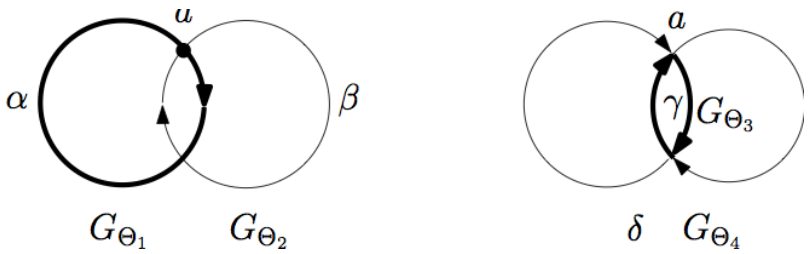


Рис. 1.10: Случай неединственного разложения в сумму двух простых эйлеровых циклов

Тогда слово $\omega = a\alpha a\delta a\beta a\gamma a \in L(2\Theta)$. Тогда для любого натурального m слово $\omega_m = (a\alpha)^m(a\delta)^m(a\beta)^m(a\gamma)^m a \in L(2m\Theta)$ и, следовательно, $\omega_m \in F_\Theta$.

Очевидно, что существует такое натуральное число $m > 0$, что длина каждой из частей слова ω_m, соответствующей любому простому циклу из разложения Θ, будет больше числа n_2 из Леммы 1.14, то есть

$$\min(m|a\alpha|, m|a\beta|, m|a\gamma|, m|a\delta|) > n_2,$$

и при этом $|\omega_m| > n_1$. Пусть слово $\omega_m = \delta_1\mu\delta_2\nu\delta_3$, где $|\mu\nu| > 0$, $|\mu\delta_2\nu| \leqslant n_2$.

Тогда подслово $\mu\delta_2\nu$ слова ω_m полностью содержится либо в части $\varepsilon_1 = (a\alpha)^m(a\delta)^m(a\beta)^m a$, либо в части $\varepsilon_2 = (a\delta)^m(a\beta)^m(a\gamma)^m a$, причем длины средних частей ε_1 и ε_2 больше, чем длина $\mu\delta_2\nu$. Пусть для определенности подслово $\mu\delta_2\nu$ содержится в ε_1.

Далее, по доказанному выше, крайние части $(a\alpha)^m a$ и $(a\beta)^m a$ слова ε_1 содержат уникальные биграммы (по отношению друг к другу и по отношению к средней части $(a\delta)^m a$). Теперь, повторяя рассуждения из п. 2.1) об отсутствии взаимных корреляций для кратностей уникальных и неуникальных биграмм для слова ε_1 при стирании частей μ и ν (в то время как в остальной части слова ω_m, а именно $(a\gamma)^m a$, все осталось неизменным), приходим к выводу, что слово $\omega_0 = \delta_1\delta_2\delta_3 \notin F_\Theta$.

Получаем противоречие с Леммой 1.14, и, значит, язык F_Θ — не контекстно-свободный.

2.3) Пусть матрица кратностей биграмм Θ, разлагается в сумму трех линейно независимых матриц $\Theta = \Theta_1 + \Theta_2 + \Theta_3$, причем каждая из матриц $\Theta_1, \Theta_2, \Theta_3 \in \Xi$ также задает эйлеров граф. При этом не существует такого элементарного эйлерова цикла G_{Θ_0} с матрицей кратностей биграмм Θ_0, что расширенная разность $G_\Theta \div G_{\Theta_0}$ соответствует матрице кратностей биграмм, которая разлагается в сумму 2 линейно независимых эйлеровых циклов.

В данном случае для любого элементарного цикла G_{Θ_0} расширенная разность $G_\Theta \div G_{\Theta_0}$ всегда будет представлять некий простой эйлеров цикл G_1.

Сведем этот случай к предыдущему в п. 2.2). Для этого заметим, что существует как минимум два разных разложения Θ в сумму двух матриц, соответствующих простым эйлеровым циклам. Одно такое разложение следует из условия: пусть $\Theta_1 = k\Theta_0$ (полагаем, что для некоторого $k \in \mathbb{N}$ операция разности для графов G_Θ и $G_{k\Theta_0}$ еще определена, а для G_Θ и $G_{(k+1)\Theta_0}$ – уже нет), $\Theta_2 = \Theta - k\Theta_0$ (соответствует простому циклу G_1).

Пусть $\Theta_2 = m\Theta_2', m \in \mathbb{N}$, где граф $G_{\Theta_2'}$ — элементарный. Из двух матриц Θ_0 и Θ_2' невозможно с помощью линейных комбинаций составить 3 линейно независимых матрицы (вспомним разложение $\Theta = \Theta_1 + \Theta_2 + \Theta_3$). Значит, существует по крайней мере еще один элементарный цикл с матрицей Θ_3', не совпадающий с первыми двумя ($\Theta_3' \neq \Theta_0, \Theta_3' \neq \Theta_2'$). Очевидно, что в этом

случае граф $G_{\Theta} \div G_{\Theta'_3}$ будет простым эйлеровым. Полагаем, что для некоторого $l \in \mathbb{N}$ операция разности для графов G_{Θ} и $G_{l\Theta'_3}$ еще определена, а для G_{Θ} и $G_{(l+1)\Theta'_3}$ – уже нет. Тогда в качестве нового разбиения исходной матрицы возьмем $\Theta = \Theta_3 + \Theta_4$, где $\Theta_3 = l\Theta'_3, \Theta_4 = \Theta - l\Theta'_3$. По построению новые матрицы Θ_3 и Θ_4, соответствующие простым эйлеровым циклам, отличны от матриц Θ_1 и Θ_2.

Значит, имеем два различных разбиения исходной матрицы кратностей биграмм в сумму двух простых циклов: $\Theta = \Theta_1 + \Theta_2 = \Theta_3 + \Theta_4$, к которым применим рассуждения из п. 2.2) (заметим, что там мы нигде не использовали тот факт, что Θ не разлагается в сумму $k \geqslant 3$ линейно независимых матриц, соответствующих эйлеровым циклам). Получаем, что язык F_{Θ} — не контекстно-свободный.

■

Замечание. В п. 1) Теоремы 1.15 F_{Θ} — детерминированный КС-язык (LR).

Следствие для двухбуквенного алфавита:

Следствие 1.15.1. *Пусть $A = \{0, 1\}$. Далее, пусть задан такой набор Θ, что соответствующий ориентированный граф G_{Θ} является эйлеровым. Тогда:*

1) *Если матрица биграмм Θ имеет вид, не совпадающий ни с одним из перечисленных $M_1^{reg}, M_2^{reg}, M_3^{reg}$ (см. Следствие 1.9.1), $M_1^{cfl} = \begin{pmatrix} c_1 & c_2 \\ c_2 & 0 \end{pmatrix}, M_2^{cfl} = \begin{pmatrix} 0 & c_3 \\ c_3 & c_4 \end{pmatrix}$, где $c_i \in \mathbb{N}, 1 \leqslant i \leqslant 4$, то язык F_{Θ} — не контекстно-свободный;*

2) *Иначе язык F_{Θ} — контекстно-свободный.*

1.6 Контекстно-зависимые биграммные языки

В данном разделе нам понадобятся следующие определения и теорема.

Определение 1.23. Грамматика называется контекстно-зависимой (КЗ-грамматикой), если каждое ее правило имеет вид $\alpha R\beta \to \alpha\gamma\beta$, где $R \in \Gamma$, а $\alpha, \beta \in (\Sigma \cup \Gamma)^*, \gamma \in (\Sigma \cup \Gamma)^+$, или $S \to \Lambda$, но тогда S не входит в правую часть никакого правила. Язык называется контекстно-зависимым (КЗ-языком), если некоторая КЗ-грамматика его порождает.

Как и в случае с КС-языками, истоки данного понятия восходят к работе Н. Хомского по описанию иерархии формальных языков [8].

Определение 1.24 (Сильное определение)**.** Линейно-ограниченный автомат — это такая недетерминированная машина Тьюринга, что:
1) входной алфавит содержит два специальных символа (левый и правый), которые обозначают конец и начало входного слова;
2) машина не может писать поверх этих специальных символов;
3) считывающая головка не может сместиться ни влево от левого специального символа, ни вправо от правого специального символа.

Другими словами, вместо потенциально бесконечной входной ленты для вычислений, в данном случае мы ограничены размерами входного слова (плюс две ячейки под левый и правый специальный символы).

Определение 1.25 (Слабое определение)**.** Линейно-ограниченный автомат — это такая недетерминированная машина Тьюринга, что:
1) входной алфавит содержит два специальных символа (левый и правый), которые обозначают конец и начало входного слова;
2) машина не может писать поверх этих специальных символов;
3) размер непрерывной рабочей области работы считывающей головки — не более чем некоторая линейная функция от длины входного слова.

Используя расширение входного алфавита, можно показать [20], что сильное и слабое определения линейно-ограниченного автомата эквивалентны в

том смысле, что они приводят к одинаковым вычислительным способностям соответствующих классов автоматов.

Следующая теорема [21] связывает КЗ-языки и линейно-ограниченные автоматы.

Теорема 1.16. *Классы КЗ-языков и языков, распознаваемых линейно-ограниченным автоматом, в точности совпадают.*

Теорема 1.17. *Бесконечный язык F_Θ, который при этом не является контекстно-свободным — контекстно-зависимый.*

Доказательство. Покажем, что для любой матрицы $\Theta \in \Xi$, соответствующей эйлерову графу G_Θ, существует некий линейно-ограниченный автомат, распознающий в точности F_Θ. Тогда по Теореме 1.16 F_Θ — КЗ-язык.

Не выписывая подробно уравнения для машины Тьюринга, покажем, как построить искомый линейно-ограниченный автомат из небольшого набора достаточно простых линейно-ограниченных автоматов.

Далее будем считать, что входное слово $\alpha, |\alpha| = l > 0$, записано на ленте слева направо, и добавлять символы мы будем справа от входного слова.

1) Предположим, что среди n^2 биграмм матрицы Θ (полагаем, как и ранее, что мощность алфавита для входного слова есть $|A| = n$) есть $m < n^2$ нулевых элементов. Тогда первый автомат будет проверять, что кратность этих биграмм в слове α также равна нулю.

Это можно несложно сделать, двигая головку слева направо, и проверяя, что во входном слове нет биграммы, которой также не было в исходной матрице биграмм Θ. Если это не так, то переходим в некоторое фиктивное состояние, не принадлежащее множеству конечных, и там и остаемся.

Заметим, что для поиска биграмм (для проверки на отсутствие или подсчета, как в следующем пункте) нам необходимо каким-то образом запоминать только что просмотренную букву, чтобы при чтении следующей мы знали,

какую именно биграмму мы только что просмотрели. Это можно сделать, "вкладывая" в состояние линейно-ограниченного автомата при просмотре входного слова информацию о только что просмотренной букве (например, с помощью специальной индексации подмножества состояний, отвечающих за перемещение по входному слову).

2) В противном случае, убедившись в отсутствии биграмм, которых не было в исходной матрице биграмм Θ, будем иметь дело только с оставшимися $n^2 - m$ биграммами, которые должны быть ненулевыми.

Теперь для каждой оставшейся биграммы посчитаем ее кратность. Кратность будем записывать справа от входного слова, для простоты в 1-ичной системе (с помощью специального символа $|_i, i = 1 \dots n^2 - m$: сколько таких символов записано, такова и кратность соответствующей биграммы). Подсчет кратностей биграмм аналогичен описанному в 1): проходим по входному слову слева направо, как только находим соответствующую биграмму, идем направо до упора и записываем $|_i$, после чего возвращаемся на прежнее место (для простоты поиска можно заменить букву входного слова a_j на, например, b_j, а после нахождения этого места совершить обратную замену).

Таким образом, для подсчета кратностей биграмм нам понадобится область размером в $l - 1$, также для простоты можно добавлять специальные символы (пробелы) между участками, обозначающими кратности разных биграмм — их будет не больше $2(n^2 - m) \leqslant 2n^2$. Значит, размер непрерывной рабочей области с учетом входного слова и двух ограничивающих его символов будет не более $(l + 2) + (l - 1) + 2n^2 = 2l + 2n^2 + 1$. Член $2n^2 + 1$ — не зависит от длины слова α. В итоге размер рабочей области — не более чем линейная функция от длины входного слова.

3) Пусть символ $|_i$ соответствует биграмме $a_{i_1}a_{i_2}$, а ее кратность в исходной матрице биграмм — $\theta_{i_1 i_2} > 0$. Теперь для каждого $i = 1 \dots n^2 - m$ будем делить число, соответствующее числу подряд идущих $|_i$, на $\theta_{i_1 i_2} > 0$. Для

этого каждый раз будем стирать по $\theta_{i_1 i_2}$ символов $|_i$ и на их месте писать один специальный символ $!_i$. Если мы можем стереть некоторое число символов $|_i$, но при этом их меньше $\theta_{i_1 i_2}$, то переходим в некоторое фиктивное состояние, не принадлежащее множеству конечных, и там и остаемся, поскольку в этом случае кратность биграммы в слове $a_{i_1} a_{i_2}$ не кратно изначальному $\theta_{i_1 i_2}$, и, следовательно, слово $\alpha \notin F_\Theta$.

В противном случае, если мы стерли все символы $|_i$ группами по $\theta_{i_1 i_2}$ символов (и оставили соответственно k_i символов $!_i$), то деление нацело завершено, и переходим к следующей биграмме.

4) Наконец, проверим, что все $k_i, i = 1 \dots n^2 - m$, равны между собой. Это можно сделать следующим образом: берем первый символ $!_1$ и стираем по одному символу $!_i$ для всех $i, i = 1 \dots n^2 - m$. Если для какого-то $j = 2 \dots n^2 - m$ такого символа не нашлось, то $k_1 \neq k_j$, то переходим в некоторое фиктивное состояние, не принадлежащее множеству конечных, и там и остаемся, так как все кратность всех биграмм должна увеличиться в одно и то же число раз. После того, как удалили все $!_1$, проверяем, что на ленте справа больше вообще нет символов $!_i$ для всех $i, i = 2 \dots n^2 - m$. Если есть, то опять переходим в некоторое фиктивное состояние, не принадлежащее множеству конечных, и там и остаемся, поскольку слово α не принадлежит языку F_Θ по той же причине.

Если же мы такой операцией удалили абсолютно все символы $!_i$ для всех $i, i = 1 \dots n^2 - m$, то переходим в конечное состояние, поскольку в этом случае $k_1 = \dots = k_{n^2 - m} = k$, и слово α принадлежит языку $L(k\Theta)$, и, следовательно, языку F_Θ.

■

Замечание. В условиях Теоремы 1.17 F_Θ — детерминированный КЗ-язык.

Следствие для двухбуквенного алфавита:

Следствие 1.17.1. *Пусть* $A = \{0, 1\}$. *Далее, пусть задан такой набор* Θ, *что соответствующий ориентированный граф* G_Θ *является эйлеровым. Тогда матрица биграмм* Θ *обязательно имеет вид, совпадающий с одним из перечисленных:* $M_1^{reg}, M_2^{reg}, M_3^{reg}$ *(см. Следствие 1.9.1),* M_1^{cfl}, M_2^{cfl} *(см. Следствие 1.15.1),* $M_1^{csl} = \begin{pmatrix} c_1 & c_2 \\ c_2 & c_3 \end{pmatrix}$, *где* $c_i \in \mathbb{N}, 1 \leqslant i \leqslant 3$. *При этом если матрица биграмм* Θ *имеет вид* M_1^{csl}, *то язык* F_Θ *— контекстно-зависимый, и не является ни регулярным, ни контекстно-свободным.*

Напоследок приведем несколько простых примеров разных классов бесконечных биграммных языков.

Пример. Пусть $A = \{0, 1\}$. Тогда следующие матрицы кратностей биграмм задают регулярные языки F_Θ: $\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$, $\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$ или $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$.

Пример. Пусть $A = \{0, 1\}$. Тогда следующие матрицы кратностей биграмм задают контекстно-свободные языки F_Θ: $\begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$ или $\begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}$.

Пример. Пусть $A = \{0, 1\}$. Тогда следующая матрица кратностей биграмм задаёт контекстно-зависимый язык F_Θ: $\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$.

2 Мощность и асимптотические оценки

2.1 Мощность конечного биграммного языка

Рассмотрим вопрос о том, сколько существует слов с данным набором Θ. Для начала рассмотрим случай двухбуквенного алфавита.

Теорема 2.1. *Для алфавита $A = \{0, 1\}$ и матрицы биграмм $\Theta \in \Xi$, задающей эйлеров или полуэйлеров граф G_Θ, число слов N_Θ с заданной матрицей биграмм Θ:*

1) *При $\theta_{01} > \theta_{10}$ количество $N_\Theta = C^{\theta_{11}}_{\theta_{11}+\theta_{10}} C^{\theta_{00}}_{\theta_{00}+\theta_{10}}$;*

2) *При $\theta_{01} < \theta_{10}$ количество $N_\Theta = C^{\theta_{11}}_{\theta_{11}+\theta_{01}} C^{\theta_{00}}_{\theta_{00}+\theta_{01}}$;*

3) *При $\theta_{01} = \theta_{10}$ количество $N_\Theta = C^{\theta_{00}}_{\theta_{00}+\theta_{01}} C^{\theta_{11}}_{\theta_{11}+\theta_{01}} \left(\frac{\theta_{01}}{\theta_{00}+\theta_{01}} + \frac{\theta_{01}}{\theta_{11}+\theta_{01}} \right)$;*

(здесь под C^k_n понимается число сочетаний из n по k, т.е. $C^k_n = \frac{n!}{k!(n-k)!}$).

Доказательство. Прежде всего отметим, что согласно Лемме 1.5 гарантируется хотя бы одно слово α с матрицей кратностей биграмм Θ.

Возьмем любое слово α, удовлетворяющее набору биграмм Θ, т.е. $\Theta(\alpha) = \Theta$. Из условия неразрывности (1) следует, что, поскольку величины $\theta_{00}(\alpha)$ и $\theta_{11}(\alpha)$ дают одинаковый вклад в подсчет как $n^{in}_a(\alpha)$, так и $n^{out}_a(\alpha)$, $a \in \{0, 1\}$, то величины $\theta_{01}(\alpha)$ и $\theta_{10}(\alpha)$ отличаются не более чем на 1.

1) Проделаем следующую операцию: заменим в слове α все участки вида $\underbrace{1..1}_{k}$ и $\underbrace{0..0}_{l}$, где $k, l \in \mathbb{N}$, $k, l \geqslant 2$, на одиночные буквы 1 и 0 соответственно. При этом получим слово α', в котором будет $\theta_{00}(\alpha') = \theta_{11}(\alpha') = 0$, т.е. буквы 0 и 1 будут перемежаться друг за другом. Кратности биграмм „01“ и „10“ не изменятся, т.е. $\theta_{01}(\alpha') = \theta_{01}(\alpha)$, $\theta_{10}(\alpha') = \theta_{10}(\alpha)$. С учетом замечания в начале доказательства можно сделать вывод, что $\theta_{01}(\alpha') = \theta_{10}(\alpha') + 1$. Т.о., получим, что слово α' имеет единственное возможное представление: $\alpha' = 0101...01$, где число нулей равно числу единиц и равно $\theta_{01}(\alpha)$.

Теперь проделаем обратную операцию: посчитаем количество разных слов,

которые можно построить по единственному α', чтобы добиться искомых кратностей $\Theta(\alpha)$. Для этого необходимо приписать около каждой буквы 0 слова α' не более $\theta_{00}(\alpha)$ букв 0 (легко заметить, что общее количество приписываемых букв равно в точности $\theta_{00}(\alpha)$) так, чтобы суммарно приписать ровно $\theta_{00}(\alpha)$ букв 0; и независимо от этого проделать аналогичную процедуру с буквами 1: приписать около каждой буквы 1 слова α' не более $\theta_{11}(\alpha)$ букв 1 так, чтобы суммарно приписать ровно $\theta_{11}(\alpha)$ букв 1. Пусть при этом мы построим слово α''. При этом мы добьемся того, что $\theta_{01}(\alpha'') = \theta_{01}(\alpha')$ и $\theta_{10}(\alpha'') = \theta_{10}(\alpha')$ (поскольку не меняем исходный порядок чередования 0 и 1 в α'), $\theta_{00}(\alpha'') = \theta_{00}(\alpha)$ и $\theta_{11}(\alpha'') = \theta_{10}(\alpha)$ (по построению α''). Т.о., количество различных способов построить слово α'' и будет искомым числом N_Θ.

Поскольку добавление букв 0 и букв 1 производится независимо друг от друга, то для подсчета N_Θ достаточно будет сделать следующее: посчитать, сколькими способами можно разместить на $\theta_{01}(\alpha)$ местах $\theta_{00}(\alpha)$ нулей, и перемножить с числом способов размещения на $\theta_{01}(\alpha)$ местах $\theta_{11}(\alpha)$ единиц (напомним, что число нулей и число единиц в α', а значит и количества „мест“ под размещение, одинаково и равно $\theta_{01}(\alpha)$).

Заметим, что число способов разместить k объектов на n местах можно посчитать следующим способом. Представим, что между местами для размещения расположены т.н. „перегородки“, т.о. перегородок всего $n-1$ штук. Если каким-то образом „перемешать“ перегородки с самими объектами, выстроенными в один ряд, получим какое-то размещение k объектов по n местам. Т.о., любая перестановка из k объектов и $n-1$ перегородки будет давать какое-то размещение (возможно. эти размещения будут повторяться). Чтобы посчитать число уникальных размещений k объектов по n местам, необходимо посчитать количество способов выбрать из $k+n-1$ объекто-перегородок $n-1$ перегородку, а это известное комбинаторное число — число сочетаний из $k+n-1$ по $n-1$, т.е. C_{k+n-1}^{n-1}.

Значит, число способов размещения на $\theta_{01}(\alpha)$ местах $\theta_{11}(\alpha)$ единиц равно $C^{\theta_{01}-1}_{\theta_{11}+\theta_{01}-1}$, а число способов размещения на $\theta_{01}(\alpha)$ местах $\theta_{00}(\alpha)$ нулей равно $C^{\theta_{01}-1}_{\theta_{00}+\theta_{01}-1}$. В итоге, учитывая $\theta_{01}(\alpha) = \theta_{10}(\alpha) + 1$, искомое число N_Θ есть произведение $N_\Theta = C^{\theta_{10}}_{\theta_{11}+\theta_{10}} C^{\theta_{10}}_{\theta_{00}+\theta_{10}}$, а с учетом тождества для биноминальных коэффициентов $C^k_n = C^{n-k}_n$ имеем $N_\Theta = C^{\theta_{11}}_{\theta_{11}+\theta_{10}} C^{\theta_{00}}_{\theta_{00}+\theta_{10}}$.

2) Рассмотрение данного случая полностью аналогично случаю 1). Имеем $\theta_{01}(\alpha') = \theta_{10}(\alpha') - 1$, и слово α' имеет единственное возможное представление: $\alpha' = 101...010$, где число нулей равно числу единиц и равно $\theta_{10}(\alpha)$. Т.о., в полученной на предыдущем шаге формуле нужно заменить θ_{10} на θ_{01}: $N_\Theta = C^{\theta_{11}}_{\theta_{11}+\theta_{01}} C^{\theta_{00}}_{\theta_{00}+\theta_{01}}$.

3) В данном случае после проведения процедуры замены $\underbrace{1..1}_{k}$ и $\underbrace{0..0}_{l}$, где $k, l \in \mathbb{N}$, $k, l \geqslant 2$, на одиночные буквы 1 и 0 соответственно получим $\theta_{01}(\alpha') = \theta_{10}(\alpha')$ и два возможных представления слова α': а) $\alpha' = 101...101$ (количество 1 равно $\theta_{01}(\alpha) + 1$, количество 0 равно $\theta_{01}(\alpha)$); б) $\alpha' = 010...010$ (количество 1 равно $\theta_{01}(\alpha)$, количество 0 равно $\theta_{01}(\alpha) + 1$).

В случае а) число способов разместить на $\theta_{01}(\alpha) + 1$ местах $\theta_{11}(\alpha)$ единиц равно $C^{\theta_{01}}_{\theta_{11}+\theta_{01}}$, а число способов размещения на $\theta_{01}(\alpha)$ местах $\theta_{00}(\alpha)$ нулей равно $C^{\theta_{01}-1}_{\theta_{00}+\theta_{01}-1}$. В случае б) число способов разместить на $\theta_{01}(\alpha)$ местах $\theta_{11}(\alpha)$ единиц равно $C^{\theta_{01}-1}_{\theta_{11}+\theta_{01}-1}$, а число способов размещения на $\theta_{01}(\alpha) + 1$ местах $\theta_{00}(\alpha)$ нулей равно $C^{\theta_{01}}_{\theta_{00}+\theta_{01}}$. Объединяя эти случаи и упрощая (учитывая тождества $C^k_n = C^{n-k}_n$ и $C^k_{n-1} = \frac{n-k}{n} C^k_n$), получим:

$$
\begin{aligned}
N_\Theta = C^{\theta_{01}}_{\theta_{11}+\theta_{01}} C^{\theta_{01}-1}_{\theta_{00}+\theta_{01}-1} + C^{\theta_{01}-1}_{\theta_{11}+\theta_{01}-1} C^{\theta_{01}}_{\theta_{00}+\theta_{01}} = \\
= C^{\theta_{11}}_{\theta_{11}+\theta_{01}} C^{\theta_{00}}_{\theta_{00}+\theta_{01}-1} + C^{\theta_{11}}_{\theta_{11}+\theta_{01}-1} C^{\theta_{00}}_{\theta_{00}+\theta_{01}} = \\
= \frac{\theta_{01}}{\theta_{00}+\theta_{01}} C^{\theta_{11}}_{\theta_{11}+\theta_{01}} C^{\theta_{00}}_{\theta_{00}+\theta_{01}} + \frac{\theta_{01}}{\theta_{11}+\theta_{01}} C^{\theta_{11}}_{\theta_{11}+\theta_{01}} C^{\theta_{00}}_{\theta_{00}+\theta_{01}} = \\
= C^{\theta_{11}}_{\theta_{11}+\theta_{01}} C^{\theta_{00}}_{\theta_{00}+\theta_{01}} \left(\frac{\theta_{01}}{\theta_{00}+\theta_{01}} + \frac{\theta_{01}}{\theta_{11}+\theta_{01}} \right).
\end{aligned}
$$

■

Пример. $A = \{0, 1\}$. Пусть $\Theta = \begin{pmatrix} \theta_{00} & \theta_{01} \\ \theta_{10} & \theta_{11} \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 2 & 2 \end{pmatrix}$.
Т.к. $\theta_{01} = \theta_{10}$, то искомое число слов с данным набором по доказанной выше теореме равно $N_\Theta = C^{\theta_{00}}_{\theta_{00}+\theta_{01}} C^{\theta_{11}}_{\theta_{11}+\theta_{01}} \left(\frac{\theta_{01}}{\theta_{00}+\theta_{01}} + \frac{\theta_{01}}{\theta_{11}+\theta_{01}} \right) = C^1_3 C^2_4 \left(\frac{2}{3} + \frac{1}{2} \right) = 12 + 9 = 21$. Это действительно так, поскольку с данным набором Θ существует ровно 21 слово: 11100101, 11001101, 11001011, 10011101, 10011011, 10010111, 11101001, 11011001, 11010011, 10111001, 10110011, 10100111, 00111010, 00110110, 00101110, 01110010, 01100110, 01001110, 01110100, 01101100, 01011100.

Однако в случае произвольного алфавита A нам понадобятся следующие определение и сопутствующая лемма.

Определение 2.1. Матрицей Кирхгофа $H(\Theta)$ [10], построенной по матрице биграмм $\Theta \in \Xi$, называется квадратная матрица размером $|A| \times |A|$, т.ч. на месте (i, j) стоит элемент

$$l_{ij} = \begin{cases} -\theta_{a_i a_j}, & i \neq j, \\ \sum_{a_j \neq a_i} \theta_{a_i a_j}, & i = j. \end{cases}$$

Замечание. Матрица Кирхгофа для любой матрицы Θ имеет нулевой определитель ($\det H(\Theta) = 0$), поскольку, очевидно, сумма всех столбцов $H(\Theta)$ есть нулевой столбец.

Лемма 2.2. *Если матрица биграмм $\Theta \in \Xi$ такова, что соответствующий ориентированный граф G_Θ является эйлеровым, то все главные миноры $D^{(i,i)}(\Theta)$, полученные вычеркиванием из $H(\Theta)$ i-й строки и i-го столбца, одинаковы.*

Доказательство. Пусть $|A| = n$. Без ограничения общности докажем, что $D^{(1,1)}(\Theta) = D^{(2,2)}(\Theta)$. Запишем

$$D^{(1,1)}(\Theta) = \begin{vmatrix} \sum_{a_j \neq a_2} \theta_{a_2 a_j} & -\theta_{a_2 a_3} & \cdots & -\theta_{a_2 a_n} \\ -\theta_{a_3 a_2} & \sum_{a_j \neq a_3} \theta_{a_3 a_j} & \cdots & -\theta_{a_3 a_n} \\ \vdots & \vdots & \ddots & \vdots \\ -\theta_{a_n a_2} & -\theta_{a_n a_3} & \cdots & \sum_{a_j \neq a_n} \theta_{a_n a_j} \end{vmatrix}.$$

Прибавим к первой строке все остальные строки (определитель при этом не изменится). Тогда на позиции $s-1$ ($s>2$) в первой строке будет стоять $\sum_{a_j \neq a_s} \theta_{a_s a_j} - (\sum_{a_t \neq a_s} \theta_{a_t a_s} - \theta_{a_1 a_s})$. По Лемме 1.2 с учетом того, в эйлеровом цикле первая буква совпадает с последней, получаем $\forall s \sum_{a_j \neq a_s} \theta_{a_s a_j} = \sum_{a_t \neq a_s} \theta_{a_t a_s}$. Значит, минор теперь будет таким:

$$D^{(1,1)}(\Theta) = \begin{vmatrix} \theta_{a_1 a_2} & \theta_{a_1 a_3} & \cdots & \theta_{a_1 a_n} \\ -\theta_{a_3 a_2} & \sum_{a_j \neq a_3} \theta_{a_3 a_j} & \cdots & -\theta_{a_3 a_n} \\ \vdots & \vdots & \ddots & \vdots \\ -\theta_{a_n a_2} & -\theta_{a_n a_3} & \cdots & \sum_{a_j \neq a_n} \theta_{a_n a_j} \end{vmatrix}.$$

Теперь прибавим к первому столбцу все остальные столбцы (определитель при этом не изменится). Тогда на позиции $s-1$ ($s>3$) в первом столбце будет стоять $\sum_{a_j \neq a_s} \theta_{a_s a_j} - (\sum_{a_t \neq a_s} \theta_{a_s a_t} - \theta_{a_s a_1})$. Получим

$$D^{(1,1)}(\Theta) = \begin{vmatrix} \sum_{a_j \neq a_1} \theta_{a_1 a_j} & \theta_{a_1 a_3} & \cdots & \theta_{a_1 a_n} \\ \theta_{a_3 a_1} & \sum_{a_j \neq a_3} \theta_{a_3 a_j} & \cdots & -\theta_{a_3 a_n} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{a_n a_1} & -\theta_{a_n a_3} & \cdots & \sum_{a_j \neq a_n} \theta_{a_n a_j} \end{vmatrix}.$$

Умножим первую строку и первый столбец на (-1); определитель опять не изменится. Получим

$$D^{(1,1)}(\Theta) = \begin{vmatrix} \sum_{a_j \neq a_1} \theta_{a_1 a_j} & -\theta_{a_1 a_3} & \cdots & -\theta_{a_1 a_n} \\ -\theta_{a_3 a_1} & \sum_{a_j \neq a_3} \theta_{a_3 a_j} & \cdots & -\theta_{a_3 a_n} \\ \vdots & \vdots & \ddots & \vdots \\ -\theta_{a_n a_1} & -\theta_{a_n a_3} & \cdots & \sum_{a_j \neq a_n} \theta_{a_n a_j} \end{vmatrix} = D^{(2,2)}(\Theta).$$

■

Т.о., в случае эйлерова графа G_Θ можно рассматривать величину $D(\Theta)$, которая равна любому из миноров $D^{(i,i)}(\Theta)$ для $\forall i \in \mathbb{N}, i = 1..|A|$.

Пусть $a \in A$, $\theta_a(\alpha)$ — количество букв a (униграмм) в слове α, а $\Delta = (\theta_{a_1}, \theta_{a_2}, ..., \theta_{a_{|A|}})$ — вектор кратностей униграмм.

В работе [4] доказана следующая

Теорема 2.3. *Число слов в алфавите A с заданными вектором униграмм Δ ($\theta_{a_i} > 0$ $\forall i = 1..|A|$) и матрицей биграмм Θ есть в точности число*

$$N_{\Delta,\Theta} = \frac{\prod_{a_i \in A}(\theta_{a_i} - 1)!}{\prod_{a_i, a_j \in A} \theta_{a_i a_j}!} \det(\theta_{a_i} * \delta_{ij} - \theta_{a_i a_j})_{i,j=1}^{|A|},$$

где δ_{ij} — символ Кронекера.

Замечание. Для доказательства данной теоремы в работе [4] используются: теорема Кирхгофа (также известная как матричная теорема об остовных деревьях) из [15] и так называемая теорема "B.E.S.T.", которая устанавливает число эйлеровых циклов в ориентированном графе (см. [16, 17]).

В нашем случае нет данных о количестве униграмм в дополнение к количеству биграмм Θ. Поэтому предыдущую теорему необходимо немного переработать, а именно, верна следующая

Теорема 2.4. *Пусть задана матрица биграмм Θ, которой соответствует эйлеров или полуэйлеров граф G_Θ, причем для $\forall i \exists j \neq i$, т.ч. $\theta_{a_i a_j} > 0$ или*

$\theta_{a_j a_i} > 0$. *Тогда:*

1) *Если* $\exists i'$, *т.ч.* $\sum_{a_i \in A} \theta_{a_i a_{i'}} > \sum_{a_i \in A} \theta_{a_{i'} a_i}$, *то*

$$N_\Theta = \frac{\prod_{a_i \in A} (\sum_{a_j \in A} \theta_{a_i a_j} - 1 + \delta_{i'i})!}{\prod_{a_i, a_j \in A} \theta_{a_i a_j}!} D^{(i'i')}(\Theta);$$

где $\delta_{i'i}$ — *символ Кронекера;*

2) *Если* $\forall i, j$ $\sum_{a_i \in A} \theta_{a_i a_j} = \sum_{a_i \in A} \theta_{a_j a_i}$, *то*

$$N_\Theta = \left(\sum_{a_i, a_j \in A} \theta_{a_i a_j} \right) \frac{\prod_{a_i \in A} (\sum_{a_j \in A} \theta_{a_i a_j} - 1)!}{\prod_{a_i, a_j \in A} \theta_{a_i a_j}!} D(\Theta).$$

Доказательство. Прежде всего отметим, что согласно Лемме 1.5 существует хотя бы одно слово α с матрицей кратностей биграмм Θ.

Далее заметим, что условие $\forall i \exists j \neq i$, т.ч. $\theta_{a_i a_j} > 0$ или $\theta_{a_j a_i} > 0$, означает, что каждая буква из алфавита A встретится хоть раз в написании слова, соответствующего матрице биграмм Θ, т.е. $\theta_{a_i} > 0$ $\forall i = 1..|A|$, и, значит, выполняется условие Теоремы 2.3.

1) Пусть $\exists i'$, т.ч. $\sum_{a_i \in A} \theta_{a_i a_{i'}} > \sum_{a_i \in A} \theta_{a_{i'} a_i}$. Это означает, что граф G_Θ является полуэйлеровым и для любого слова $\alpha \in A^*$, т.ч. $\Theta(\alpha) = \Theta$, буква $a_{i'}$ будет являться последней буквой слова α (согласно Лемме 1.2).

Более того, теперь можно вычислить однозначно любую униграмму θ_{a_i}. Если $a_i \neq a_{i'}$, то $\theta_{a_i} = \sum_{a_j \in A} \theta_{a_i a_j}$. При этом $\theta_{a_{i'}} = \sum_{a_j \in A} \theta_{a_{i'} a_j} + 1$, поскольку нужно учесть, что из последней буквы не исходит соответствующее ребро, а посчитать в кратности униграммы эту последнюю букву мы должны.

Теперь преобразуем детерминант из формулировки Теоремы 2.3. Подставляя выведенные выше выражения для униграмм, получим, что

$$\det(\theta_{a_i} * \delta_{ij} - \theta_{a_i a_j}) = \begin{vmatrix} \ddots & \vdots & \vdots & \vdots & \iddots \\ \cdots & \sum\limits_{a_j \neq a_{i'-1}} \theta_{a_{i'-1} a_j} & -\theta_{a_{i'-1} a_{i'}} & -\theta_{a_{i'-1} a_{i'+1}} & \cdots \\ \cdots & -\theta_{a_{i'} a_{i'-1}} & \sum\limits_{a_j \neq a_{i'}} \theta_{a_{i'} a_j} + 1 & -\theta_{a_{i'} a_{i'+1}} & \cdots \\ \cdots & -\theta_{a_{i'+1} a_{i'-1}} & -\theta_{a_{i'+1} a_{i'}} & \sum\limits_{a_j \neq a_{i'+1}} \theta_{a_{i'+1} a_j} & \cdots \\ \iddots & \vdots & \vdots & \vdots & \ddots \end{vmatrix}.$$

Прибавим к i'-му столбцу сумму остальных столбцов (при этом определитель не изменится). Получим:

$$\det(\theta_{a_i} * \delta_{ij} - \theta_{a_i a_j}) = \begin{vmatrix} \ddots & \vdots & 0 & \vdots & \iddots \\ \cdots & \sum_{a_j \neq a_{i'-1}} \theta_{a_{i'-1} a_j} & 0 & -\theta_{a_{i'-1} a_{i'+1}} & \cdots \\ \cdots & -\theta_{a_{i'} a_{i'-1}} & 1 & -\theta_{a_{i'} a_{i'+1}} & ... \\ \cdots & -\theta_{a_{i'+1} a_{i'-1}} & 0 & \sum_{a_j \neq a_{i'+1}} \theta_{a_{i'+1} a_j} & \cdots \\ \iddots & \vdots & 0 & \vdots & \ddots \end{vmatrix}.$$

Разложим определитель по элементам i'-го столбца. Поскольку определитель матрицы равен сумме произведений элементов столбца на их алгебраические дополнения, а единственный ненулевой элемент в столбце находится в строке i' и равен 1, то получим

$$\det(\theta_{a_i} * \delta_{ij} - \theta_{a_i a_j}) = 1 * (-1)^{i'+i'} \begin{vmatrix} \ddots & \vdots & \vdots & \iddots \\ \cdots & \sum\limits_{a_j \neq a_{i'-1}} \theta_{a_{i'-1} a_j} & -\theta_{a_{i'-1} a_{i'+1}} & \cdots \\ \cdots & -\theta_{a_{i'+1} a_{i'-1}} & \sum\limits_{a_j \neq a_{i'+1}} \theta_{a_{i'+1} a_j} & \cdots \\ \iddots & \vdots & \vdots & \ddots \end{vmatrix} = \\ = D^{(i',i')}(\Theta).$$

Объединяя полученные выражения для униграмм и для определителя, подставим их в формулу Теоремы 2.3 и получим:

$$N_\Theta = \frac{\prod_{a_i \in A} (\sum_{a_j \in A} \theta_{a_i a_j} - 1 + \delta_{i'i})!}{\prod_{a_i, a_j \in A} \theta_{a_i a_j}!} D^{(i'i')}(\Theta).$$

2) Пусть $\forall i, j \sum_{a_i \in A} \theta_{a_i a_j} = \sum_{a_i \in A} \theta_{a_j a_i}$. Это означает, что граф G_Θ является эйлеровым и первая буква совпадает с последней. Поскольку в эйлеровом цикле любая буква может быть первой (и соответственно последней), то нужно рассмотреть все варианты для каждой из букв a_i быть на последнем месте.

Пусть буква $a_{i'}$ — последняя (и соответственно первая) при прохождении эйлерова цикла. Тогда можно вычислить однозначно кратность любой униграммы θ_{a_i}. Если $a_i \neq a_{i'}$, то $\theta_{a_i} = \sum_{a_j \in A} \theta_{a_i a_j}$. При этом $\theta_{a_{i'}} = \sum_{a_j \in A} \theta_{a_{i'} a_j} + 1$. Таким образом, пользуясь результатами предыдущего пункта, получим число слов с данной матрицей биграмм Θ и последней буквой $a_{i'}$

$$N_{\Theta, a_{i'}} = \left(\sum_{a_j \in A} \theta_{a_{i'} a_j} \right) \frac{\prod_{a_i \in A} (\sum_{a_j \in A} \theta_{a_i a_j} - 1)!}{\prod_{a_i, a_j \in A} \theta_{a_i a_j}!} D^{(i'i')}(\Theta).$$

Обратим внимание на то, что в данном случае можно вынести множитель $\sum_{a_j \in A} \theta_{a_{i'} a_j}$ из факториала, поскольку значение $\sum_{a_j \in A} \theta_{a_i a_j} - 1$, согласно условию теоремы, всегда неотрицательное и факториал определен.

Полное же число слов с данной матрицей биграмм Θ — это сумма по всем возможным последним буквам величин $N_{\Theta,a_{i'}}$: $N_\Theta = \sum_{a_{i'}\in A} N_{\Theta,a_{i'}}$.

Поскольку соответствующий граф эйлеров, то согласно Лемме 2.2 $\forall i,j = 1..|A|$ $D^{(i,i)}(\Theta) = D^{(j,j)}(\Theta) = D(\Theta)$.

Подставляя, получим утверждение теоремы

$$N_\Theta = \sum_{a_{i'}\in A} N_{\Theta,a_{i'}} = \sum_{a_{i'}\in A} \left(\sum_{a_j\in A} \theta_{a_{i'}a_j} \right) \frac{\prod_{a_i\in A}(\sum_{a_j\in A}\theta_{a_ia_j}-1)!}{\prod_{a_i,a_j\in A}\theta_{a_ia_j}!} D^{(i'i')}(\Theta) =$$

$$= \left(\sum_{a_i,a_j\in A} \theta_{a_ia_j} \right) \frac{\prod_{a_i\in A}(\sum_{a_j\in A}\theta_{a_ia_j}-1)!}{\prod_{a_i,a_j\in A}\theta_{a_ia_j}!} D(\Theta).$$

■

2.2 Асиптотика мощности $L(k\Theta)$

Рассмотрим вопрос об асимптотическом росте числа слова с матрицей биграмм $k\Theta$ при $k \to \infty$.

Определение 2.2. Матрица биграмм Θ называется положительной матрицей биграмм, если все элементы этой матрицы — положительные числа (т.е. нет нулей).

Теорема 2.5. *Пусть задана положительная матрица биграмм Θ, т.ч. соответствующий ориентированный граф G_Θ — эйлеров. Тогда при $k \to \infty$ для числа слов β_k, т.ч. $\Theta(\beta_k) = k\Theta$, выполняется*

$$N_{k\Theta} \sim c_2 * \frac{c_1^k}{k^{n(n-1)/2}},$$

где $c_1 = c_1(\Theta) > 1, c_2 = c_2(\Theta)$ — некоторые константы, зависящие только от изначальной матрицы биграмм Θ, а $n = |A|$ — мощность алфавита.

Доказательство. Поскольку ориентированный граф G_Θ — эйлеров, то можно воспользоваться п. 2 из формулировки Теоремы 2.4:

$$N_{k\Theta} = \left(\sum_{a_i,a_j \in A} k\theta_{a_i a_j} \right) \frac{\prod_{a_i \in A} (\sum_{a_j \in A} k\theta_{a_i a_j} - 1)!}{\prod_{a_i,a_j \in A} (k\theta_{a_i a_j})!} D(k\Theta) =$$

$$= \left(\frac{\sum_{a_i,a_j \in A} k\theta_{a_i a_j}}{\prod_{a_i \in A} (\sum_{a_j \in A} k\theta_{a_i a_j})} \right) \left(\frac{\prod_{a_i \in A} (\sum_{a_j \in A} k\theta_{a_i a_j})!}{\prod_{a_i,a_j \in A} (k\theta_{a_i a_j})!} \right) (D_k),$$

где D_k — главный минор матрицы Кирхгофа, построенной по матрице биграмм $k\Theta$.

Оценим по отдельности каждый из трех сомножителей в вышеприведенной формуле. Первый сомножитель имеет вид

$$\frac{\sum_{a_i,a_j \in A} k\theta_{a_i a_j}}{\prod_{a_i \in A} (\sum_{a_j \in A} k\theta_{a_i a_j})} = \frac{c'_1 k}{c'_2 k^n} = c'_3 k^{1-n},$$

где c'_1, c'_2, c'_3 — некоторые константы, зависящие только от матрицы биграмм Θ.

Т. к. D_k — это определитель матрицы размером $(n-1) \times (n-1)$, при этом, поскольку при переходе от Θ к $k\Theta$ все кратности биграмм просто умножились на k, то при переходе от $D(\Theta)$ к D_k каждая из $n-1$ строк просто умножилась на k, что по свойству определителя означает $D_k = k^{n-1} D(\Theta)$. С другой стороны, определитель $D(\Theta)$ зависит только от матрицы биграмм Θ. Значит,

$$D_k = c'_4 k^{n-1},$$

где c'_4 — некоторая константа, зависящая только от матрицы биграмм Θ.

Для оценки последнего сомножителя воспользуемся формулой Стирлинга

для факториала: $n! \sim \sqrt{2\pi n}(\frac{n}{e})^n$:

$$\frac{\prod_{a_i \in A}(\sum_{a_j \in A} k\theta_{a_i a_j})!}{\prod_{a_i,a_j \in A}(k\theta_{a_i a_j})!} \sim$$

$$\sim \frac{\prod_{a_i \in A}\sqrt{2\pi \sum_{a_j \in A} k\theta_{a_i a_j}}(\sum_{a_j \in A} k\theta_{a_i a_j})^{\sum_{a_j \in A} k\theta_{a_i a_j}}/e^{\sum_{a_j \in A} k\theta_{a_i a_j}}}{\prod_{a_i,a_j \in A}\sqrt{2\pi k\theta_{a_i a_j}}(k\theta_{a_i a_j})^{k\theta_{a_i a_j}}/e^{k\theta_{a_i a_j}}} =$$

$$= c_5' \frac{\prod_{a_i,a_j \in A} e^{k\theta_{a_i a_j}}}{\prod_{a_i,a_j \in A} e^{k\theta_{a_i a_j}}} \frac{k^{n/2}}{k^{n^2/2}} \frac{\prod_{a_i \in A}(\sum_{a_j \in A} k\theta_{a_i a_j})^{\sum_{a_j \in A} k\theta_{a_i a_j}}}{\prod_{a_i,a_j \in A}(k\theta_{a_i a_j})^{k\theta_{a_i a_j}}} =$$

$$= c_5' \frac{1}{k^{(n^2-n)/2}} \frac{\prod_{a_i,a_j \in A}(\sum_{a_l \in A} k\theta_{a_i a_l})^{k\theta_{a_i a_j}}}{\prod_{a_i,a_j \in A}(k\theta_{a_i a_j})^{k\theta_{a_i a_j}}},$$

где c_5' — некоторая константа, зависящая только от матрицы биграмм Θ.

Рассмотрим частное $\frac{\sum_{a_l \in A} k\theta_{a_i a_l}}{k\theta_{a_i a_j}}$. Т.к. мы имеем положительную матрицу биграмм, то все биграммы в любой ее строке отличны от нуля, поэтому, очевидно, имеет место $\frac{\sum_{a_l \in A} k\theta_{a_i a_l}}{k\theta_{a_i a_j}} = 1 + \sigma_{ij}$, где σ_{ij} — некоторая положительная константа. Значит,

$$\frac{\prod_{a_i \in A}(\sum_{a_j \in A} k\theta_{a_i a_j})!}{\prod_{a_i,a_j \in A}(k\theta_{a_i a_j})!} \sim c_5' \frac{1}{k^{(n^2-n)/2}} \prod_{a_i,a_j \in A} (1+\sigma_{ij})^{k\theta_{a_i a_j}} =$$

$$= c_5' \frac{1}{k^{(n^2-n)/2}} \left(\prod_{a_i,a_j \in A} (1+\sigma_{ij})^{\theta_{a_i a_j}} \right)^k = c_5' \frac{1}{k^{(n^2-n)/2}} c_1^k,$$

где c_1 — некоторая константа, зависящая только от матрицы биграмм Θ, при этом $c_1 > 1$ (как произведение положительных степеней чисел, больших единицы).

Соберем воедино:

$$N_{k\Theta} \sim c_3' k^{1-n} c_4' k^{n-1} c_5' \frac{1}{k^{(n^2-n)/2}} c_1^k = c_2 * \frac{c_1^k}{k^{n(n-1)/2}},$$

где c_2 — некоторая константа, зависящая только от матрицы биграмм

Θ. ■

Замечание. Похожую оценку можно найти в работе [5], однако там авторы ограничились только верхней асимптотической оценкой.

Пример. Пусть $A = \{0, 1\}$. Тогда для положительной матрицы кратностей биграмм $\Theta = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ верна следующая точная асимптотическая оценка: $N_{k\Theta} \sim \frac{1}{\pi}\frac{16^k}{k}$.

2.3 Асимптотика количества матриц, задающих определенный класс биграммных языков

Рассмотрим вопрос о том, каких матриц "больше" (в асимптотическом смысле) — тех, которые задают пустые, конечные или счётные биграммные языки; также интересен этот вопрос в случае бесконечных биграммных языков: каких "больше" — задающих регулярные, контекстно-свободные или контекстно-зависимые биграммные языки.

Замечание. Если две разные матрицы кратностей биграмм Θ_1 и Θ_2, $\Theta_1 \neq \Theta_2$, задают непустые языки $L(\Theta_1)$ и $L(\Theta_2)$, то очевидно, что $L(\Theta_1) \cap L(\Theta_2) = \varnothing$.

Обозначим через Ξ_k множество матриц размера $|A| \times |A|$, каждый элемент которых представляет собой неотрицательное целое число, не превосходящее натуральное $k > 0$. Все соотношения теперь будем рассматривать с учетом того, что все матрицы кратностей биграмм Θ принадлежат Ξ_k. Также будем считать, что исходный алфавит A зафиксирован и $|A| = n > 1$.

Через $EMPTY(k)$ обозначим количество матриц кратностей биграмм $\Theta \in \Xi_k$, задающих пустые языки F_Θ.

Через $NONEMPTY(k)$ обозначим количество матриц кратностей биграмм $\Theta \in \Xi_k$, задающих непустые языки F_Θ.

Через $FINITE(k)$ обозначим количество матриц кратностей биграмм $\Theta \in \Xi_k$, задающих конечные (непустые) языки F_Θ.

Через $INFINITE(k)$ обозначим количество матриц кратностей биграмм $\Theta \in \Xi_k$, задающих счетные языки F_Θ.

Через $REG(k)$ обозначим количество матриц кратностей биграмм $\Theta \in \Xi_k$, задающих счетные регулярные языки F_Θ.

Через $NONREG(k)$ обозначим количество матриц кратностей биграмм $\Theta \in \Xi_k$, задающих счетные нерегулярные языки F_Θ.

Через $CFL(k)$ обозначим количество матриц кратностей биграмм $\Theta \in \Xi_k$, задающих счетные КС-языки F_Θ, не являющиеся регулярными.

Через $CSL(k)$ обозначим количество матриц кратностей биграмм $\Theta \in \Xi_k$, задающих счетные КЗ-языки F_Θ, не являющиеся контекстно-свободными.

Через $ALL(k)$ обозначим общее количество матриц $\Theta \in \Xi_k$.

Замечание. Очевидно, что $ALL(k) = (k+1)^{n^2}$.

Замечание. Справедливы следующие тождества:
$ALL(k) = EMPTY(k) + NONEMPTY(k)$,
$NONEMPTY(k) = FINITE(k) + INFINITE(k)$,
$INFINITE(k) = REG(k) + NONREG(k)$,
$NONREG(k) = CFL(k) + CSL(k)$.

Теорема 2.6. *С учетом введенных выше обозначений верны следующие соотношения:*
1) *для любого* $k \in \mathbb{N}$ $\frac{1}{n(n-1)} < \frac{INFINITE(k)}{FINITE(k)} < 1$;
2) $\lim_{k\to\infty} \frac{INFINITE(k)}{ALL(k)} = 0$;
3) $\lim_{k\to\infty} \frac{REG(k)}{NONREG(k)} = 0$;
4) $\lim_{k\to\infty} \frac{CFL(k)}{CSL(k)} = 0$.

Доказательство. 1) Для начала рассмотрим эйлеровы графы, отличные от одиночных (возможно, кратных) петель.

Заметим, что любой эйлеров граф превращается в полуэйлеров путем изъятия одного ребра, соединяющего различные вершины, из эйлерова графа. Это будет так по определению и свойству эйлеровых графов, так как при удалении одного ребра из эйлерова графа новых компонент связности не может возникнуть (в любом эйлеровом графе для создания двух и более компонент связности нужно удалить как минимум два ребра). Также, во всех вершинах, кроме тех двух, которые были началом и концом удаленного ребра, сумма входящих и сумма исходящих ребер не поменялась. При этом, если была удалено ребро $a_i \to a_j$, то в получившемся полуэйлеровом графе началом эйлерова пути будет a_j, а концом — a_i.

Необходимо заметить, что для двух разных эйлеровых графов полуэйлеровы графы, полученные такой операцией „удаления“ одного ребра, также будут разными. Также будут разными полуэйлеровы графы, полученные из одного эйлерова графа, если удалять ребра, соединяющие разные пары вершин. При этом каждому полуэйлерову графу будет соответствовать некоторый единственный эйлеров граф, получающийся добавлением одного ребра.

Обозначим через n_i^k число эйлеровых графов, построенных по матрицам из Ξ_k, в которых ровно i разных пар вершин соединены ребрами (пары упорядочены, т.е. (a_i, a_j) и (a_j, a_i) считаются разными парами). Тогда минимальное количество пар вершин для эйлерова графа равно 2 (минимальный цикл на двух вершинах), а максимальное — $n(n-1)$ (каждая пара вершин между собой соединена в двух противоположных направлениях). Каждому эйлерову графу с i парами соединенных вершин соответствует ровно i различных полуэйлеровых графов, которые получаются операцией „удаления“ одного ребра, соединяющего различные вершины эйлерова графа.

Значит, число эйлеровых графов без одиночных петель равно $\sum_{i=2}^{n(n-1)} n_i^k$, а общее число полуэйлеровых — $\sum_{i=2}^{n(n-1)} i n_i^k$.

Теперь рассмотрим эйлеровы графы, являющиеся одиночными петлями.

Очевидно, из них операцией „удаления“ любого ребра (в данном случае — петли) мы не получим полуэйлерова графа. Всего различных одиночных (возможно, кратных) петель, построенных по матрицам из Ξ_k, будет равно количеству букв n в алфавите A, помноженному на максимальное количество k кратных петель в одной вершине, т.е. kn.

Таким образом, легко получить оценку снизу:

$$\frac{INFINITE(k)}{FINITE(k)} = \frac{\sum_{i=2}^{n(n-1)} n_i^k + nk}{\sum_{i=2}^{n(n-1)} in_i^k} > \frac{\sum_{i=2}^{n(n-1)} n_i^k}{\sum_{i=2}^{n(n-1)} in_i^k} \geqslant \frac{1}{n(n-1)}.$$

Для оценки сверху расчитаем величину n_2^k. Всего вариантов выбрать неупорядоченную пару различных вершин $\frac{n(n-1)}{2}$ (эйлеров цикл на двух вершинах содержит как некоторое ребро $a_i \to a_j$, так и обязательно $a_j \to a_i$ при $i \neq j$). В каждой вершине независимо может быть от 0 до k петель, при этом эйлеров цикл на двух вершинах также может повторяться от 1 до k раз. Таким образом, $n_2^k = \frac{n(n-1)}{2} k(k+1)^2$.

Легко проверить, что для любых натуральных k будет верно соотношение $n_2^k + nk < 2n_2^k$. Поэтому верхняя оценка выводится следующим образом:

$$\frac{INFINITE(k)}{FINITE(k)} = \frac{\sum_{i=3}^{n(n-1)} n_i^k + n_2^k + nk}{\sum_{i=3}^{n(n-1)} in_i^k + 2n_2^k} < \frac{\sum_{i=3}^{n(n-1)} n_i^k + 2n_2^k}{\sum_{i=3}^{n(n-1)} in_i^k + 2n_2^k} \leqslant$$
$$\leqslant \frac{\sum_{i=3}^{n(n-1)} n_i^k + 2n_2^k}{3\sum_{i=3}^{n(n-1)} n_i^k + 2n_2^k} \leqslant 1.$$

В заключение заметим, что количество эйлеровых графов в точности соответствует количеству матриц кратностей биграмм, задающих счетные языки F_Θ, а количество полуэйлеровых графов — количеству матриц кратностей биграмм, задающих конечные (непустые) языки F_Θ.

2) Оценим сверху количество $INFINITE(k)$.

В эйлеровом графе количество входящих ребер всегда равно количеству выходящих для любой из n вершин — значит, можно записать систему линейных однородных уравнений $i = 1, \ldots, n$:

$$\sum_{j=1}^{n} \theta_{a_i a_j} = \sum_{j=1}^{n} \theta_{a_j a_i}.$$

В этой системе n уравнений (по одному для каждой вершины), при этом n^2 неизвестных $\theta_{a_i a_j}$, $i, j = 1, \ldots, n$. Значит, размерность пространства решений данной системы линейных однородных уравнений равна разности размерности всего пространства переменных (n^2) и ранга матрицы, соответствующей этой системе (n).

Поскольку $\forall i, j = 1, \ldots, n$ верно $0 \leqslant \theta_{a_i a_j} \leqslant k$, то число решений есть не более чем произведение количества возможностей проварьировать каждую переменную (k+1 возможность) для каждой из базисных переменных (которых по доказанному выше $n^2 - n$).

Количество решений системы не меньше количества эйлеровых графов (поскольку вышеприведенная система не учитывает связность графа), что, в свою очередь, равно $INFINITE(k)$. Также учтем, что $ALL(k) = (k+1)^{n^2}$. В итоге получим

$$\lim_{k\to\infty} \frac{INFINITE(k)}{ALL(k)} \leqslant \lim_{k\to\infty} \frac{(k+1)^{n^2-n}}{(k+1)^{n^2}} = \lim_{k\to\infty} \frac{1}{(k+1)^n} = 0.$$

3) Согласно доказательству п. 2) Теоремы 1.9, множество матриц кратностей биграмм для регулярных языков F_Θ задает либо кратные одиночные петли, либо повторяющиеся циклы вида $a_{i_1} \to a_{i_2} \to \ldots \to a_{i_s} \to a_{i_1}$, где $i_{j_1} \neq i_{j_2}$ для $j_1 \neq j_2$, $s \leqslant n$.

Всего одиночных некратных петель равно числу букв алфавита A, т.е. n штук. Число различных элементарных циклов длины $s \geqslant 2$ есть также функция от n и не зависит от k. Т.о., суммарное количество одиночных

некратных петель и элементарных циклов любой длины от 2 до n есть некая функция $f(n)$, не зависящая от k.

Каждую из таких одиночных некратных петель или элементарных циклов можно изменять следующим образом, оставаясь в рамках Ξ_k: умножать количество петель (или количество ребер для любых двух соединенных ребром вершин) на одно и то же число k_1, где $1 \leqslant k_1 \leqslant k$. В итоге общее число матриц из Ξ_k, задающих регулярные языки F_Θ, есть $REG(k) = kf(n)$.

Для оценки числа матриц из Ξ_k, задающих нерегулярные языки F_Θ, возьмем любой элементарный цикл, содержащий все n вершин, например, $a_1 \to a_2 \to \ldots \to a_n \to a_1$. Для того, чтобы перевести данный эйлеров граф в разряд графов, задающих нерегулярные языки, достаточно добавить в какую-либо из вершин петлю, возможно, кратную. Всего возможностей разместить петли в вершинах такого графа будет $k^{n+1} - 1 > k^n$ для $k > 1$.

Всего же матриц из Ξ_k, задающих нерегулярные языки F_Θ, будет очевидно не меньше, чем таких циклов с петлями. Значит, $NONREG(k) > k^n$, и

$$\lim_{k\to\infty} \frac{REG(k)}{NONREG(k)} \leqslant \lim_{k\to\infty} \frac{kf(n)}{k^n} = \lim_{k\to\infty} \frac{f(n)}{k^{n-1}} = 0.$$

4) Для начала докажем следующий факт: если F_Θ — КС-язык, не являющийся регулярным при $|A| = n > 1$, то в соответствующем графе G_Θ не более одной петли (возможно, кратной).

Действительно, пусть это не так, т. е. в графе G_Θ как минимум две петли на разных вершинах. Т. к. эти петли должны быть соединены (и притом циклом, иначе граф G_Θ — не эйлеров), то получим, что в граф G_Θ распадается в сумму трех независимых эйлеровых циклов (две петли и цикл, их соединяющий). Тогда согласно Теореме 1.15, язык F_Θ — не КС-язык, и получаем противоречие.

Далее, т. к. F_Θ — КС-язык, не являющийся регулярным при $|A| = n > 1$, то в графе G_Θ как минимум две вершины являются неизолированными, т. к. в противном случае граф G_Θ был бы петлей (возможно, кратной) и,

соответственно, F_Θ был бы регулярным.

Значит, для любого КС-языка F_Θ, не являющийся регулярным, мы можем выбрать в соответствующем графе G_Θ неизолированную вершину, в которой нет петли. Пусть этой вершиной будет a_i. Тогда добавив в эту вершину петлю $a_i \to a_i$, мы получим три независимых цикла (два уже были, т. к. F_Θ — КС-язык, не являющийся регулярным, и один мы добавили — $a_i \to a_i$). Значит, новый граф будет соответствовать уже не КС, а КЗ-языку (см. Теоремы 1.15 и 1.17).

Таким образом, мы для любого КС-языка F_Θ, не являющегося регулярным, путем "наращивания" петли $a_i \to a_i$ (сначала один раз, потом еще и т. д. вплоть до k кратных петель $a_i \to a_i$) в соответствующем графе G_Θ будем получать граф, соответствующий КЗ-языку, который не является контекстно-свободным. Значит, для любого k каждой матрице кратностей биграмм из Ξ_k, соответствующей КС-языку, не являющемуся регулярным, можно поставить в соответствие как минимум k матриц кратностей биграмм, соответствующих КЗ-языкам, которые не являются контекстно-свободными.

В итоге получим, что

$$\lim_{k\to\infty} \frac{CFL(k)}{CSL(k)} \leqslant \lim_{k\to\infty} \frac{CFL(k)}{kCFL(k)} = \lim_{k\to\infty} \frac{1}{k} = 0.$$

■

Следствие 2.6.1.

$$\lim_{k\to\infty} \frac{NONEMPTY(k)}{ALL(k)} = 0.$$

Доказательство. Согласно п. 1) Теоремы 2.6, $INFINITE(k)$ и $FINITE(k)$ имеют одинаковые порядки роста по k. Так как согласно п. 2) $\lim_{k\to\infty} \frac{INFINITE(k)}{ALL(k)} = 0$, то и $\lim_{k\to\infty} \frac{FINITE(k)}{ALL(k)} = 0$. С учетом того, что $NONEMPTY(k) = INFINITE(k) + FINITE(k)$, получим

$$\lim_{k\to\infty} \frac{NONEMPTY(k)}{ALL(k)} = \lim_{k\to\infty} \frac{INFINITE(k)}{ALL(k)} + \lim_{k\to\infty} \frac{FINITE(k)}{ALL(k)} = 0.$$

■

3 Расширение понятия биграммных языков

3.1 Свойства матрицы биграмм с закольцовыванием

Пусть задан конечный алфавит $A = \{a_1, a_2, ..., a_n\}$, а его мощность $|A| = n$.

Определение 3.1. Элементарной матрицей кратностей биграмм $\Theta_{ij} \in \Xi$ будем называть матрицу из пространства матриц биграмм Ξ, имеющую единственный ненулевой элемент на месте (i, j): $\theta_{a_i a_j} = 1, \theta_{a_k a_l} = 0, (k, l) \neq (i, j), 1 \leqslant i, j, k, l \leqslant n$.

Определение 3.2. Назовем $\Omega(\alpha)$ матрицей кратностей биграмм с закольцовыванием для непустого слова $\alpha \in A^*$ следующую матрицу:
1) при однобуквенном слове $\alpha = a_t, 1 \leqslant t \leqslant n$, $\Omega(\alpha) = \Theta_{tt}$,
2) при длине слова α не меньше 2, то есть $\alpha = a_i \alpha_1 a_j, 1 \leqslant i, j \leqslant n$, где $\alpha_1 \in A^*$ (в том числе α_1 может быть пустым), $\Omega(\alpha) = \Theta(\alpha) + \Theta_{ji}$.

Следующее определение эквивалентно приведённому выше:

Определение 3.3. Пусть $\alpha = a_i \alpha', \alpha, \alpha' \in A^*, a_i \in A$. Назовём $\Omega(\alpha)$ матрицей кратностей биграмм с закольцовыванием для непустого слова $\alpha \in A^*$ следующую матрицу: $\Omega(\alpha) = \Theta(a_i \alpha' a_i)$.

В дальнейшем будем пользоваться тем определением, которое удобнее.

Очевидно, что матрица $\Omega(\alpha) = (\omega_{ij}(\alpha))_{i,j=1}^n$ размера $n \times n$ лежит в том же пространстве матриц Ξ, что и матрица биграмм $\Theta(\alpha)$. Пусть на месте (i, j) матрицы $\Omega(\alpha)$ будет стоять значение $\omega_{a_i a_j}(\alpha)$. Также, здесь и далее через $\Omega(\alpha)$ будем обозначать матрицу биграмм с закольцовыванием, построенную по конкретному слову α, а через Ω — просто некоторую матрицу из Ξ, при этом будем считать, что на месте (i, j) матрицы Ω будет стоять значение $\omega_{a_i a_j}$ (т.е. для произвольной матрицы из Ξ мы опустили зависимость от α как

для самой матрицы биграмм с закольцовыванием, так и для отдельных ее элементов).

Таким образом, матрица биграмм с закольцовыванием — это матрица биграмм за исключением единичной добавки в одной ячейке, которая отвечает за биграмму, связывающую последнюю букву слова с первой (отсюда и название — „с закольцовыванием“, поскольку мы как бы считаем биграммы не просто на слове, а на слове, начало и конец которого объединены в кольцо).

Определение 3.4. Назовем простейшим биграммным языком с закольцовыванием $K(\Omega)$ множество всех слов, имеющих одну и ту же матрицу Ω кратностей биграмм с закольцовыванием, т.е. $K(\Omega) = \{\beta \in A^* \mid \Omega(\beta) = \Omega\}$.

Рассмотрим основные свойства матрицы биграмм с закольцовыванием, которые во многом повторяют аналогичные утверждения для матрицы биграмм.

Лемма 3.1. *Простейший биграммный язык с закольцовыванием $K(\Omega)$ состоит не более чем из конечного числа слов одинаковой длины*

$$l_\Omega = \sum_{a_i, a_j \in A} \omega_{a_i a_j}.$$

Доказательство. Возьмем произвольное $\alpha \in K(\Omega)$, если таковое имеется (в противном случае язык $K(\Omega)$ пуст и доказывать нечего).

Если $\exists t, 1 \leqslant t \leqslant n$, т.ч. $\Omega = \Theta_{tt}$, то α — однобуквенное слово (иначе по определению в матрице было бы либо больше ненулевых элементов, либо $\omega_{a_t a_t} > 1$), причем $\alpha = a_t$. Т.о., $K(\Omega) = \{a_t\}$ и $l_\Omega = 1$.

В противном случае α — не менее чем двухбуквенное слово, а именно пусть $\alpha = a_s \alpha_1 a_t, 1 \leqslant s, t \leqslant n$, где $\alpha_1 \in A^*$. Тогда согласно Лемме 1.1 длину α можно вычислить как сумму кратностей всех его биграмм ($\sum_{a_i, a_j \in A} \theta_{a_i a_j}$)

плюс 1, что совпадает с условием данной леммы, т.к.

$$len(\alpha) = \sum_{a_i,a_j \in A} \theta_{a_i a_j}(\alpha) + 1 = \sum_{a_i,a_j \in A,(i,j)\neq(t,s)} \theta_{a_i a_j}(\alpha) + (\theta_{a_t a_s}(\alpha) + 1) =$$

$$= \sum_{a_i,a_j \in A,(i,j)\neq(t,s)} \omega_{a_i a_j}(\alpha) + \omega_{a_t a_s}(\alpha) = \sum_{a_i,a_j \in A} \omega_{a_i a_j} = l_\Omega.$$

Т.о., $K(\Omega)$ в случае непустоты состоит из слов одинаковой длины, что для конечного алфавита означает конечность языка $K(\Omega)$.

■

Пусть $a \in A$. Обозначим за $m_a^{in}(\alpha), \alpha \in A^*$ следующую величину: $m_a^{in}(\alpha) = \sum\limits_{b \in A} \omega_{ba}(\alpha)$, что будет равно сумме значений матрицы $\Omega(\alpha)$ в столбце, соответствующем букве a. Аналогичным образом определим и $m_a^{out}(\alpha), \alpha \in A^*$ как $m_a^{out}(\alpha) = \sum\limits_{b \in A} \omega_{ab}(\alpha)$, что будет равно сумме значений матрицы $\Omega(\alpha)$ в строке, соответствующей букве a.

Лемма 3.2 (Условие неразрывности)**.** *Пусть задано непустое слово $\alpha \in A^*$. Тогда матрица $\Omega(\alpha)$ обладает следующим свойством:*

$$\forall i, 1 \leqslant i \leqslant n, m_{a_i}^{in}(\alpha) = m_{a_i}^{out}(\alpha).$$

Доказательство. Если слово α однобуквенное, то имеется только единственное ненулевое значение в матрице $\Omega(\alpha)$, стоящее на диагонали, и утверждение данной леммы очевидно.

Если же $\alpha = a_i \alpha_1 a_j, 1 \leqslant i, j \leqslant n$, где $\alpha_1 \in A^*$, то согласно Лемме 1.2

$$\forall b \in A \quad n_b^{out}(\alpha) - n_b^{in}(\alpha) = \delta_{ba_i} - \delta_{ba_j},$$

где δ_{ij} — символ Кронекера.

Рассмотрим четыре случая.

1) $b \neq a_i, b \neq a_j$. В этом случае $n_b^{out}(\alpha) = n_b^{in}(\alpha)$. При этом очевидно, что $\omega_{ba}(\alpha) = \theta_{ba}(\alpha)$ и $\omega_{ab}(\alpha) = \theta_{ab}(\alpha)$ для любой буквы a из A. Тогда получим, что $m_b^{in}(\alpha) = \sum\limits_{a \in A} \omega_{ab}(\alpha) = \sum\limits_{a \in A} \theta_{ab}(\alpha) = n_b^{in}(\alpha)$ и $m_b^{out}(\alpha) = \sum\limits_{a \in A} \omega_{ba}(\alpha) = \sum\limits_{a \in A} \theta_{ba}(\alpha) = n_b^{out}(\alpha)$, и, значит, $m_b^{out}(\alpha) = m_b^{in}(\alpha)$.

2) $b = a_i, b \neq a_j$. В этом случае $n_b^{out}(\alpha) = n_b^{in}(\alpha) + 1$. При этом $\omega_{ba}(\alpha) = \theta_{ba}(\alpha)$ для любой буквы a из A, однако $\omega_{ab}(\alpha) = \theta_{ab}(\alpha)$ для $a \neq a_j$ и $\omega_{a_j b}(\alpha) = \theta_{a_j b}(\alpha) + 1$.

В итоге получим, что $m_b^{in}(\alpha) = \sum\limits_{a \in A, a \neq a_j} \omega_{ab}(\alpha) + \omega_{a_j b}(\alpha) = \sum\limits_{a \in A, a \neq a_j} \theta_{ab}(\alpha) + \theta_{a_j b}(\alpha) + 1 = n_b^{in}(\alpha) + 1$ и $m_b^{out}(\alpha) = \sum\limits_{a \in A} \omega_{ba}(\alpha) = \sum\limits_{a \in A} \theta_{ba}(\alpha) = n_b^{out}(\alpha)$, и, значит, $m_b^{out}(\alpha) = m_b^{in}(\alpha)$.

3) $b = a_j, b \neq a_i$. В этом случае $n_b^{in}(\alpha) = n_b^{out}(\alpha) + 1$. При этом $\omega_{ab}(\alpha) = \theta_{ab}(\alpha)$ для любой буквы a из A, однако $\omega_{ba}(\alpha) = \theta_{ba}(\alpha)$ для $a \neq a_i$ и $\omega_{ba_i}(\alpha) = \theta_{ba_i}(\alpha) + 1$.

В итоге получим, что $m_b^{out}(\alpha) = \sum\limits_{a \in A, a \neq a_i} \omega_{ba}(\alpha) + \omega_{ba_i}(\alpha) = \sum\limits_{a \in A, a \neq a_i} \theta_{ba}(\alpha) + \theta_{ba_i}(\alpha) + 1 = n_b^{out}(\alpha) + 1$ и $m_b^{in}(\alpha) = \sum\limits_{a \in A} \omega_{ab}(\alpha) = \sum\limits_{a \in A} \theta_{ab}(\alpha) = n_b^{in}(\alpha)$, и, значит, $m_b^{out}(\alpha) = m_b^{in}(\alpha)$.

4) $b = a_i = a_j$. В этом случае $n_b^{out}(\alpha) = n_b^{in}(\alpha)$. При этом $\omega_{ab}(\alpha) = \theta_{ab}(\alpha)$ и $\omega_{ba}(\alpha) = \theta_{ba}(\alpha)$ для $a \neq b$, однако $\omega_{bb}(\alpha) = \theta_{bb}(\alpha) + 1$. В итоге получим, что $m_b^{out}(\alpha) = \sum\limits_{a \in A, a \neq b} \omega_{ba}(\alpha) + \omega_{bb}(\alpha) = \sum\limits_{a \in A, a \neq b} \theta_{ba}(\alpha) + \theta_{bb}(\alpha) + 1 = n_b^{out}(\alpha) + 1$ и $m_b^{in}(\alpha) = \sum\limits_{a \in A, a \neq b} \omega_{ab}(\alpha) + \omega_{bb}(\alpha) = \sum\limits_{a \in A, a \neq b} \theta_{ab}(\alpha) + \theta_{bb}(\alpha) + 1 = n_b^{in}(\alpha) + 1$, и, значит, $m_b^{out}(\alpha) = m_b^{in}(\alpha)$.

■

Следует отметить, что, так же как и в случае с обычными биграммными языками, условие неразрывности является необходимым, но не достаточным условием непустоты языка $K(\Omega)$, заданного матрицей биграмм с закольцовыванием Ω. Несложно привести пример, когда условие неразрывности выполняется для некоторой матрицы биграмм с закольцовыванием Ω, но при этом в

языке $K(\Omega)$ нет ни одного слова. Например, если $\Omega = \begin{pmatrix} \omega_{00} & \omega_{01} \\ \omega_{10} & \omega_{11} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, то очевидно, что условие неразрывности выполняется: $\omega_0^{in} = \omega_0^{out} = 1$ и $\omega_1^{in} = \omega_1^{out} = 1$, однако очевидно, что в языке $K(\Omega)$ нет ни одного слова, поскольку раз уж в слове есть буквы „1“ и „0“, то должна быть и ненулевая кратность биграммы, связывающей их („10“ или „01“, возможно, как связь последней буквы слова с первой буквой).

Аналогичным образом, как и в случае с обычными биграммными языками, построим по матрице $\Omega(\alpha)$ (или по произвольной матрице $\Omega \in \Xi$) ориентированный граф $G_{\Omega(\alpha)}$ (соответственно, G_Ω) на плоскости. Вершинами у этого графа будут все буквы из алфавита A, при этом ребра будут соответствовать биграммам с учетом их кратностей, т.е. кратность $\omega_{ab}(\alpha)$ будет порождать $\omega_{ab}(\alpha)$ ориентированных ребер $a \to b$. Аналогично, кратность $\omega_{cc}(\alpha)$ будет порождать $\omega_{cc}(\alpha)$ петель $c \to c$.

Пример. $A = \{0, 1\}$, $\beta = 0101110$.
$\Omega(\beta) = \begin{pmatrix} \omega_{00}(\beta) & \omega_{01}(\beta) \\ \omega_{10}(\beta) & \omega_{11}(\beta) \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 2 & 2 \end{pmatrix}$.
Соответствующий ориентированный граф $G_{\Omega(\beta)}$ изображен на Рис. 3.1. Несложно заметить, что граф $G_{\Omega(\beta)}$ будет совпадать с графом $G_{\Theta(\alpha)}$ на Рис. 1.1.

Лемма 3.3 (Достаточное условие существования)**.** *Для того, чтобы существовало хотя бы одно слово α с данной матрицей кратностей биграмм $\Omega \in \Xi$ с закольцовыванием, достаточно, чтобы построенный по Ω ориентированный граф G_Ω был эйлеровым.*

Доказательство. По определению, в эйлером графе существует эйлеров цикл, т.е. путь, проходящий по всем ребрам орграфа, причем только один раз, и имеющий в качестве финальной вершины вершину, совпадающую с исходной.

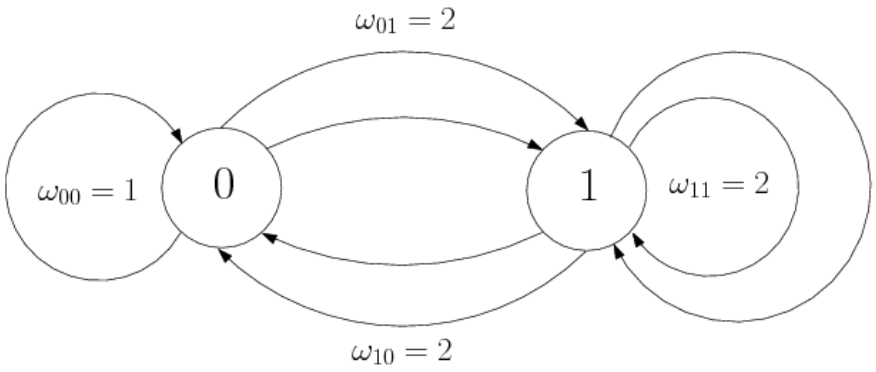


Рис. 3.1: Граф $G_{\Omega(\beta)}$, построенный по матрице биграмм с закольцовыванием $\Omega(\beta)$

Если этот цикл состоит из одного ребра, то он обязан быть некоторой петлей $a_i \to a_i$. В этом случае в языке $K(\Omega)$ содержится единственное слово $\alpha = a_i$.

В противном случае пусть такой эйлеров цикл длины $k > 1$ задается последовательностью ребер $a_{i_1} \to a_{i_2}, a_{i_2} \to a_{i_3}, ..., a_{i_{k-1}} \to a_{i_k}, a_{i_k} \to a_{i_1}$. Тогда слово $\alpha = a_{i_1}a_{i_2}a_{i_3}...a_{i_{k-1}}a_{i_k}$ и будет искомым, поскольку при его построении будут участвовать все ребра из G_Ω (а значит, и все ненулевые кратности биграмм из Ω), причем с учетом правильных кратностей. Заметим, что, хотя для последнего ребра в цикле $a_{i_k} \to a_{i_1}$ нет соответствующей пары букв в слове α, стоящих рядом, это компенсируется тем, что последняя и первая буквы, записанные рядом, как раз и дают это ребро: $a_{i_k}a_{i_1}$.

■

Как итог, получаем следующее важное следствие:

Следствие 3.3.1 (Алгоритмическая разрешимость)**.** *Задача определения по матрице $\Omega \in \Xi$, существует ли хотя бы одно слово α, имеющее эту матрицу биграмм с закольцовыванием, алгоритмически разрешима.*

Доказательство. Согласно Лемме 3.1 длина l_Ω искомого слова α есть $l_\Omega = \sum\limits_{a_i, a_j \in A} \omega_{a_i a_j}$.

Значит, алгоритм можно предложить следующий. Перебираем все $|A|^{l_\Omega}$ слов длины l_Ω, для каждого такого слова β вычисляем матрицу биграмм закольцовыванием $\Omega(\beta)$ и сравниваем с заданной матрицей Ω. Если на каком-то из $|A|^{l_\Omega}$ шагов получили совпадение, значит, искомое слово найдено. Если же за $|A|^{l_\Omega}$ шагов совпадения не было получено, то искомого слова не существует.

Однако, даже при небольших значениях мощности алфавита $|A|$ перебор может представлять значительную трудность. Поэтому лучше воспользоваться результатом Леммы 3.3: построить по набору значений Ω ориентированный граф G_Ω и проверить, существует ли в нем эйлеров цикл. Очевидно, эта задача алгоритмически разрешима.

■

Отметим еще один интересный момент, касающийся матрицы биграмм с закольцовыванием. В случае с обычной матрицей биграмм не всегда по Θ можно было определить точное количество каждой буквы из алфавита A в слове из языке $L(\Theta)$.

Напомним определение вектора униграмм. Пусть $a \in A$, $\theta_a(\alpha)$ — количество букв a (униграмм) в слове α, а $\Delta(\alpha) = (\theta_{a_1}(\alpha), \theta_{a_2}(\alpha), \ldots, \theta_{a_n}(\alpha))$ — вектор кратностей униграмм.

Лемма 3.4. *Пусть матрица $\Omega \in \Xi$ такова, что соответствующий ориентированный граф G_Ω — эйлеров. Тогда для любого слова α из языка $K(\Omega)$ можно определить вектор униграмм $\Delta(\alpha) = (\theta_{a_1}(\alpha), \theta_{a_2}(\alpha), \ldots, \theta_{a_n}(\alpha))$ однозначно, причем*

$$\theta_{a_1}(\alpha) = \sum_{b \in A} \omega_{ba_1} = \sum_{b \in A} \omega_{a_1 b}, \ldots, \theta_{a_n}(\alpha) = \sum_{b \in A} \omega_{ba_n} = \sum_{b \in A} \omega_{a_n b}.$$

Доказательство. Если матрица $\Omega \in \Xi$ такова, что соответствующий ориентированный граф G_Ω — эйлеров, то согласно Лемме 3.3 язык $K(\Omega)$ непуст.

Пусть произвольное слово $\alpha \in K(\Omega)$, $\Omega(\alpha) = \Omega$. Тогда согласно Лемме 3.2 имеем

$$\forall i, 1 \leqslant i \leqslant n, \sum_{b \in A} \omega_{ba_i} = m_{a_i}^{in}(\alpha) = m_{a_i}^{out}(\alpha) = \sum_{b \in A} \omega_{a_i b}.$$

Несложно заметить, что для любой буквы a, которая не является первой в слове α, ее количество в слове α можно посчитать как $\theta_a(\alpha) = \sum\limits_{b \in A} \theta_{ba}(\alpha) = \sum\limits_{b \in A} \omega_{ba}$.

Если же буква a является первой в слове α, то ее количество в слове α можно посчитать как $\theta_a(\alpha) = \sum\limits_{b \in A} \theta_{ba}(\alpha) + 1 = \sum\limits_{b \in A} \omega_{ba}$.

Учитывая все вышеприведенное, получаем утверждение данной леммы.

■

Теорема 3.5. *Пусть матрица $\Omega \in \Xi$ такова, что соответствующий ориентированный граф G_Ω — эйлеров. Тогда существует взаимно-однозначное соответствие между словами языков $K(\Omega)$ и $L(\Theta)$, где $\Omega = \Theta$.*

Доказательство. Если матрица $\Omega \in \Xi$ такова, что соответствующий ориентированный граф G_Ω — эйлеров, то согласно Лемме 3.3 язык $K(\Omega)$ непуст.

Возьмем любое слово $\alpha \in K(\Omega)$, $\Omega(\alpha) = \Omega$. Пусть $\alpha = a\alpha_1$, где $\alpha_1 \in A^*$ (заметим, что α_1 может быть пустым). Тогда слово $\beta = \alpha a$ будет иметь матрицу биграмм $\Theta(\beta)$, такую что $\Theta(\beta) = \Omega(\alpha) = \Omega$. Таким образом, сопоставим любому слову $\alpha \in K(\Omega)$ некоторое слово β из $L(\Theta)$, где $\Omega = \Theta$.

Если матрица $\Theta \in \Xi$ такова, что соответствующий ориентированный граф G_Θ — эйлеров, то согласно Лемме 1.5 язык $L(\Theta)$ непуст.

Возьмем любое слово $\beta \in L(\Theta)$, $\Theta(\beta) = \Theta$. Пусть $\beta = a_i\alpha_1 a_j$, где $a_i, a_j \in A, \alpha_1 \in A^*$ (заметим, что α_1 может быть пустым). Согласно Лемме 3.2 и тому, что $\forall a, b \in A\ \theta_{ab} = \omega_{ab}$ имеем $n_{a_i}^{in}(\beta) = m_{a_i}^{in}(\beta) = m_{a_i}^{out}(\beta) = n_{a_i}^{out}(\beta)$. С другой стороны, согласно Лемме 1.2 $n_{a_i}^{out}(\beta) - n_{a_i}^{in}(\beta) = \delta_{a_i a_i} - \delta_{a_i a_j}$. Получаем, что $0 = 1 - \delta_{a_i a_j}$, т.е. $a_i = a_j$ и первая буква β совпадает с последней.

Пусть $a_i = a_j = a$ и, таким образом, $\beta = a\alpha_1 a$. Тогда слово $\alpha = a\alpha_1$ будет иметь матрицу биграмм $\Omega(\alpha)$ с закольцовыванием, такую что $\Omega(\alpha) = \Theta(\beta) = \Theta$. Таким образом, сопоставим любому слову $\beta \in L(\Theta)$ некоторое слово α из $K(\Omega)$, где $\Omega = \Theta$.

Как итог, получаем взаимно-однозначное соответствие между словами языков $K(\Omega)$ и $L(\Theta)$, где $\Omega = \Theta$.

■

В качестве следствия из этой теоремы получим важные утверждения о мощности языка $K(\Omega)$.

Следствие 3.5.1. *Пусть матрица $\Omega \in \Xi$ такова, что соответствующий ориентированный граф G_Ω — эйлеров. Тогда количество слов в языках $K(\Omega)$ и $L(\Theta)$, где $\Omega = \Theta$, одинаково: $|K(\Omega)| = |L(\Theta)|$.*

Теорема 3.6. *Для алфавита $A = \{0, 1\}$ и матрицы биграмм $\Omega \in \Xi$ с закольцовыванием, задающей эйлеров ориентированный граф G_Ω, мощность языка $K(\Omega)$ есть:*

$$|K(\Omega)| = C^{\omega_{00}}_{\omega_{00}+\omega_{01}} C^{\omega_{11}}_{\omega_{11}+\omega_{01}} \left(\frac{\omega_{01}}{\omega_{00}+\omega_{01}} + \frac{\omega_{01}}{\omega_{11}+\omega_{01}} \right);$$

(здесь под C^k_n понимается число сочетаний из n по k, т.е. $C^k_n = \frac{n!}{k!(n-k)!}$).

Доказательство. Согласно Следствию 3.5.1 количество слов в языках $K(\Omega)$ и $L(\Theta)$, где $\Omega = \Theta$, одинаково. Значит, найдя количество слов в языке $L(\Theta)$, где $\Omega = \Theta$, мы найдем количество слов языке $K(\Omega)$.

Согласно Лемме 3.2 и тому, что $\forall a, b \in A\ \theta_{ab} = \omega_{ab}$ имеем $\sum\limits_{b \in A} \theta_{ba} = \sum\limits_{b \in A} \omega_{ba} = \sum\limits_{b \in A} \omega_{ab} = \sum\limits_{b \in A} \theta_{ab}$. Значит, $\theta_{01} + \theta_{11} = \theta_{10} + \theta_{11}$, т.е. $\theta_{01} = \theta_{10}$.

Таким образом, согласно Теореме 2.1 (условие 3) этой теоремы) число слов в языке $L(\Theta)$ есть

$$N_\Theta = C^{\theta_{00}}_{\theta_{00}+\theta_{01}} C^{\theta_{11}}_{\theta_{11}+\theta_{01}} \left(\frac{\theta_{01}}{\theta_{00}+\theta_{01}} + \frac{\theta_{01}}{\theta_{11}+\theta_{01}} \right).$$

Меняя $\forall a, b \in A\ \theta_{ab}$ на ω_{ab} в этой формуле, получим утверждение данной теоремы.

■

Пример. $A = \{0, 1\}$. Пусть $\Omega = \begin{pmatrix} \omega_{00} & \omega_{01} \\ \omega_{10} & \omega_{11} \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 2 & 2 \end{pmatrix}$.
Искомое число слов с данной матрицей биграмм с закольцовыванием по доказанной выше теореме равно $|K(\Omega)| = C^{\omega_{00}}_{\omega_{00}+\omega_{01}} C^{\omega_{11}}_{\omega_{11}+\omega_{01}} \left(\frac{\omega_{01}}{\omega_{00}+\omega_{01}} + \frac{\omega_{01}}{\omega_{11}+\omega_{01}} \right) = C^1_3 C^2_4 \left(\frac{2}{3} + \frac{1}{2} \right) = 12 + 9 = 21$. Это действительно так, поскольку с данной матрицей биграмм Ω с закольцовыванием существует ровно 21 слово: 1110010, 1100110, 1100101, 1001110, 1001101, 1001011, 1110100, 1101100, 1101001, 1011100, 1011001, 1010011, 0011101, 0011011, 0010111, 0111001, 0110011, 0100111, 0111010, 0110110, 0101110.

Теорема 3.7. *Пусть задана матрица биграмм Ω, которой соответствует эйлеров ориентированный граф G_Ω, причем для $\forall i \exists j \neq i$, т.ч. $\omega_{a_i a_j} > 0$. Тогда:*

$$|K(\Omega)| = \left(\sum_{a_i, a_j \in A} \omega_{a_i a_j} \right) \frac{\prod_{a_i \in A} (\sum_{a_j \in A} \omega_{a_i a_j} - 1)!}{\prod_{a_i, a_j \in A} \omega_{a_i a_j}!} D(\Omega);$$

где $D(\Omega)$ — любой из главных миноров матрицы Кирхгофа $H(\Omega)$, полученных вычеркиванием из $H(\Omega)$ i-ой строки и i-го столбца, $1 \leqslant i \leqslant n$.

Доказательство. Согласно Следствию 3.5.1 количество слов в языках $K(\Omega)$ и $L(\Theta)$, где $\Omega = \Theta$, одинаково. Значит, найдя количество слов в языке $L(\Theta)$, где $\Omega = \Theta$, мы найдем количество слов языке $K(\Omega)$.

Согласно Лемме 3.2 и тому, что $\forall a, b \in A\ \theta_{ab} = \omega_{ab}$ имеем $\sum\limits_{b \in A} \theta_{ba} = \sum\limits_{b \in A} \omega_{ba} = \sum\limits_{b \in A} \omega_{ab} = \sum\limits_{b \in A} \theta_{ab}$. Значит, мы можем воспользоваться утверждением 2) Теоремы 2.4, в которой мощность языка $L(\Theta)$ определяется как

$$N_\Theta = \left(\sum_{a_i, a_j \in A} \theta_{a_i a_j} \right) \frac{\prod_{a_i \in A} (\sum_{a_j \in A} \theta_{a_i a_j} - 1)!}{\prod_{a_i, a_j \in A} \theta_{a_i a_j}!} D(\Theta).$$

Меняя $\forall a, b \in A\ \theta_{ab}$ на ω_{ab} в этой формуле, а также Θ на Ω, получим утверждение данной теоремы.

■

3.2 Биграммные языки с закольцовыванием

Аналогично изучению биграммных языков, рассмотрим случай, когда мы рассматриваем матрицу биграмм с закольцовыванием не как абсолютное ограничение, а как задание относительных значений (пропорций) ω_{ab}, то есть случай языка, в котором отношения $\omega_{ab}(\alpha)/\omega_{cd}(\alpha)$, где $\omega_{cd}(\alpha) > 0$, зависят только от выбора букв $a, b, c, d \in A$, но не зависят от слова α из этого языка. Определим такой язык.

Определение 3.5. Назовем биграммным языком с закольцовыванием, заданным матрицей биграмм $\Omega \in \Xi$ с закольцовыванием, следующий язык при $k \in \mathbb{N}$:

$$E_\Omega = \bigcup_{k=1}^{\infty} K(k\Omega),$$

т.е. язык, состоящий из всех таких слов β, т.ч. матрица биграмм $\Omega(\beta)$ с закольцовыванием этих слов кратна набору Ω, а именно $E_\Omega = \{\beta \in A^* \mid \exists k \in \mathbb{N} : \Omega(\beta) = k\Omega\}$, где умножение k на Ω понимается как умножение скаляра на матрицу.

Лемма 3.8 (Корректность определения)**.** *Если существует хотя бы одно непустое слово α, имеющее матрицу биграмм с закольцовыванием Ω, то существует слово β_k для любого натурального k, т.ч. $\Omega(\beta_k) = k\Omega(\alpha) = k\Omega$.*

Доказательство. Пусть $\alpha \in K(\Omega)$, $\Omega(\alpha) = \Omega$.

Если $\alpha = a_t, 1 \leqslant t \leqslant n$, то в качестве слова β_k возьмем $\beta_k = (a_t)^k = \underbrace{a_t...a_t}_{k \quad a_t}$.
В этом случае $\Omega(\beta_k) = \Theta(\beta_k) + \Theta_{tt} = (k-1)\Theta_{tt} + \Theta_{tt} = k\Theta_{tt} = k\Omega(\alpha) = k\Omega$.

В противном случае пусть $\alpha = a_i\alpha_1 a_j, 1 \leqslant i, j \leqslant n, \alpha_1 \in A^*$ (α_1 может быть пустым). В качестве слова β_k возьмем $\beta_k = (\alpha)^k = \underbrace{a_i\alpha_1 a_j a_i\alpha_1 a_j ... a_i\alpha_1 a_j}_{k \quad a_i\alpha_1 a_j}$.
Тогда будем иметь $\Omega(\beta_k) = \Theta(\beta_k) + \Theta_{ji} = k\Theta(\alpha) + (k-1)\Theta(a_j a_i) + \Theta_{ji} = k\Theta(\alpha) + (k-1)\Theta_{ji} + \Theta_{ji} = k(\Theta(\alpha) + \Theta_{ji}) = k\Omega(\alpha) = k\Omega$.

■

Следствие 3.8.1 (Достаточное условие)**.** *Для того, чтобы $\forall k \in \mathbb{N}$ существовало слово β_k, т.ч. для заданной матрицы биграмм Ω с закольцовыванием выполнялось условие $\Omega(\beta_k) = k\Omega$, достаточно, чтобы построенный по Ω ориентированный граф G_Ω являлся эйлеровым графом.*

Доказательство. Очевидно, что в случае $k = 1$ будем иметь $\beta_1 \in K(\Omega)$. Согласно Лемме 3.3, для непустоты языка $K(\Omega)$ достаточно, чтобы ориентированный граф G_Ω был эйлеровым.

С другой стороны, согласно Лемме 3.8, для получения остальных слов $\beta_k, k > 1$, достаточно иметь хотя бы одно непустое слово $\alpha \in A^*$, т.ч. $\Omega(\alpha) = \Omega$. Таким образом, беря в качестве α слово β_1, получим утверждение данного следствия.

■

Теорема 3.9 (О мощности биграммных языков с закольцовыванием)**.** *Пусть задана матрица биграмм $\Omega \in \Xi$ с закольцовыванием. Тогда*

1) *Если ориентированный граф G_Ω является эйлеровым, то в биграммном языке E_Θ с закольцовыванием счетное множество слов;*

2) *Если ориентированный граф G_Ω не является эйлеровым, то в биграммном языке E_Θ с закольцовыванием нет ни одного слова.*

Доказательство. 1) Воспользуемся Следствием 3.8.1. Для каждого $k \in \mathbb{N}$

будет существовать хотя бы одно слово β_k, т.ч. $\Omega(\beta_k) = k\Omega$, и, следовательно, лежащее в E_Ω. Т.к. для каждого натурального k будет не более чем конечное количество таких β_k, а объединение счетного числа конечных множеств счетно, имеем первое утверждение теоремы.

2) В этом случае по Лемме 3.3 нет ни одного слова α, удовлетворяющего набору Ω. Согласно Теоремам 1.3 и 1.4 граф G_Ω либо не связен, либо в нем есть вершина с разным числом входящих и исходящих ребер. Тогда для любого целого $k > 1$ граф $G_{k\Omega}$ либо не связен, либо имеет вершину, у которой разность чисел входящих и исходящих ребер по модуля не меньше $k > 1$, и по Лемме 3.3 не существует ни одного слова β_k с матрицей кратностей биграмм $k\Omega$ с закольцовыванием.

Значит, в данном случае язык E_Ω пуст. ■

Пример. $A = \{0, 1\}$. Пусть $\Omega = \begin{pmatrix} \omega_{00} & \omega_{01} \\ \omega_{10} & \omega_{11} \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 2 & 2 \end{pmatrix}$.
Построим граф G_Ω по Ω — см. Рис. 3.1. В этом графе в вершину 0 входит 3 ребра, исходит тоже 3, в вершину 1 входит 4 ребра и исходит тоже 4. Все вершины лежат в одной компоненте связности. Получается, что граф G_Ω эйлеров, т.е. в Теореме 3.9 выполняется условие 1), и, соответственно, в частотном языке E_Ω счетное число слов.

Следствие 3.9.1 (Алгоритмическая разрешимость)**.** *Задача определения по матрице биграмм $\Theta \in \Xi$ с закольцовыванием, пуст или счетен биграммный язык с закольцовыванием E_Ω, алгоритмически разрешима.*

Доказательство. В данном случае вариант доказательства (предложенный в Следствии 3.3.1), основанный на переборе, не пройдет, поскольку здесь нужно перебирать бесконечное множество слов разной длины. Поэтому обратимся ко второму варианту доказательства, основанному на рассмотрении эйлеровых циклов в графе.

Для этого достаточно построить граф G_Ω и выяснить, существует ли в нем эйлеров цикл. Если существует — значит, язык счетен. В противном случае биграммный язык с закольцовыванием E_Ω пуст.

■

3.3 Регулярные, контекстно-свободные и контекстно-зависимые биграммные языки с закольцовыванием

В начале данного раздела для примера подробно докажем теорему о регулярности биграммных языков с закольцовыванием, аналогичную Теореме 1.9. После чего приведем формулировки теорем, аналогичных Теоремам 1.15 и 1.17, и очень кратко укажем на различия в доказательствах, поскольку основная часть доказательства останется той же, что и в случае с биграммными языками.

Определение 3.6. Назовем приведенной матрицей, соответствующей матрице биграмм с закольцовыванием Ω в алфавите $A, |A| < \infty$, матрицу $\widehat{\Omega} = \Omega/\text{НОД}(\omega_{ab} \mid \omega_{ab} > 0, a, b \in A)$.

Теорема 3.10. *Пусть $A, |A| < \infty$ — некоторый конечный алфавит. Далее, пусть задана матрица биграмм Ω с закольцовыванием такая, что соответствующий ей ориентированный граф G_Ω является эйлеровым. Тогда:*
1) *Если существует такое разложение Ω в сумму двух ненулевых линейно независимых матриц $\Omega = \Omega_1 + \Omega_2$ такое, что обе матрицы Ω_1 и Ω_2 задают ориентированные эйлеровы графы G_{Ω_1} и G_{Ω_2}, то биграммный язык с закольцовыванием E_Ω нерегулярен;*
2) *В противном случае язык E_Ω регулярен. При этом для $\forall k \in \mathbb{N}$ существуют ровно l слов $\beta_{k,i}, i = 1..l$, т.ч. $\Omega(\beta_{k,i}) = k\Omega$, а l — число ненулевых элементов в матрице Ω.*

Доказательство. 1) Пусть существует разложение Ω в сумму двух ненулевых линейно независимых матриц $\Omega = \Omega_1 + \Omega_2$, т.ч. обе матрицы Ω_1 и Ω_2 задают ориентированные графы G_{Ω_1} и G_{Ω_2}, которые являются эйлеровыми. На языке графов это значит, что эйлеров цикл графа G_Ω распадается в сумму двух различных циклов, соответствующих графам G_{Ω_1} и G_{Ω_2}. При этом, поскольку изначально граф был связным (с точностью до изолированных вершин), то графы G_{Ω_1} и G_{Ω_2} имеют хотя бы одну общую вершину. Пусть этой вершиной будет $a, a \in A$.

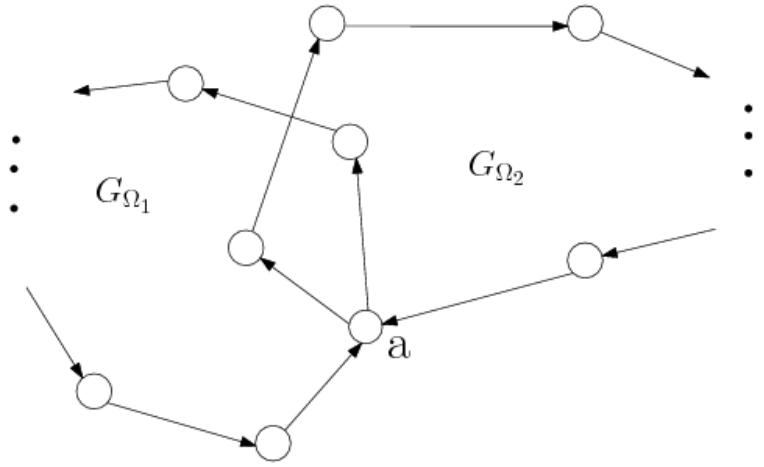


Рис. 3.2: Выделение из графа G_Ω двух линейно независимых эйлеровых циклов для случая 1) теоремы

Пусть в графе G_{Ω_1} эйлеров цикл с началом (и соответственно концом) в общей точке a задается словом $\alpha_1 = a\alpha_1'$, $\alpha_1' \in A^*$, при этом, очевидно, $\Omega(\alpha_1) = \Omega_1$. Аналогично, пусть в графе G_{Ω_2} эйлеров цикл с началом (и соответственно концом) в общей точке a задается словом $\alpha_2 = a\alpha_2'$, $\alpha_2' \in A^*$, при этом $\Omega(\alpha_2) = \Omega_2$. Тогда слово $\alpha' = \alpha_1\alpha_2 = a\alpha_1' a\alpha_2'$ будет задавать изначальный эйлеров цикл в графе G_Ω, т.е. $\Omega(\alpha') = \Omega$. Отметим, что, поскольку матрицы Ω_1 и Ω_2 ненулевые, то и α_1, и α_2 — непустые.

Предположим, что язык E_Ω регулярен, тогда по теореме Клини он представим в некотором конечно-детерминированном инициальном автомате $V_{q_0} = (A, Q, B, \varphi, \psi, q_0)$, где A — входной алфавит, Q — алфавит состояний,

B — выходной алфавит (будем считать, что $B = \{0, 1\}$), $\varphi : Q \times A^* \to Q$ — функция переходов, $\psi : Q \times A^* \to B$ — функция выходов, $q_0 \in Q$ — начальное состояние. При этом $\beta \in E_\Omega \Leftrightarrow \psi(q_0, \beta) = 1$. Пусть мощность состояний $|Q| = p$.

Т.к. по Следствию 3.8.1 для любого $k \in \mathbb{N}$ найдется слово β_k, т.ч. $\Omega(\beta_k) = k\Omega$, то зафиксируем некоторое $s > p$ и возьмем такое слово β, что $\Omega(\beta) = s\Omega$, при этом слово β представлено в виде $\beta = \underbrace{\alpha_1...\alpha_1}_{s}\underbrace{\alpha_2...\alpha_2}_{s}$. Это значит, что нужно сначала пройти s раз по эйлерову циклу графа G_{Ω_1} с началом в общей вершине a, после чего s раз по эйлерову циклу графа G_{Ω_2} с началом все в той же общей вершине a.

Обозначим $q = q_0$. Запишем в ряд состояния, в которые мы будем переходить при последовательной подаче слова α_1 (как подслова слова β): $q_1 = \varphi(q, \alpha_1), q_2 = \varphi(q, \alpha_1\alpha_1) = \varphi(q_1, \alpha_1), ..., q_s = \varphi(q, \underbrace{\alpha_1...\alpha_1}_{s}) = \varphi(q_{s-1}, \alpha_1)$. Поскольку $s > p$, то в этом ряду длины s будут по меньшей мере два повторяющихся состояния, т.е. $\exists i, j \in \mathbb{N}, 1 \leqslant i < j \leqslant s$, т.ч. $q_i = q_j$. Значит, для $\forall m \in \mathbb{N}$ верно тождество $q_j = \varphi(q, \underbrace{\alpha_1..\alpha_1}_{i}\underbrace{\alpha_1..\alpha_1}_{m(j-i)})$, т.к. мы будем ходить по циклу, подавая одно и то же слово α_1 по одним и тем же состояниям $q_i, q_{i+1}, ..., q_j = q_i$.

Обозначим через $\beta'_m, m \in \mathbb{N}$ слово $\beta'_m = \underbrace{\alpha_1..\alpha_1}_{s+(m-1)*(j-i)}\underbrace{\alpha_2..\alpha_2}_{s}$. Тогда

$$\varphi(q, \underbrace{\alpha_1..\alpha_1}_{j}\underbrace{\alpha_1..\alpha_1}_{s-j}) = \varphi(q, \underbrace{\alpha_1..\alpha_1}_{i}\underbrace{\alpha_1..\alpha_1}_{m(j-i)}\underbrace{\alpha_1..\alpha_1}_{s-j}) \quad \forall m \in \mathbb{N}.$$

Значит,

$$\varphi(q_0, \beta) = \varphi(\varphi(q_0, \underbrace{\alpha_1..\alpha_1}_{s}), \underbrace{\alpha_2..\alpha_2}_{s}) = \varphi(\varphi(q_0, \underbrace{\alpha_1..\alpha_1}_{s+(m-1)(j-i)}), \underbrace{\alpha_2..\alpha_2}_{s}) = \varphi(q_0, \beta'_m).$$

Обозначим через $\widetilde{\delta}$, где $\delta \in A^*$ — непустое слово, слово δ без последней

буквы. Следовательно, и $\varphi(q_0, \widetilde{\beta}) = \varphi(q_0, \widetilde{\beta'_m})$, т.к. начиная с первого вхождения непустого подслова α_2 в слова β и β'_m, мы будем двигаться по автомату V_{q_0} по одинаковым буквам, начиная с одинакового состояния q_s.

Пускай последней буквой непустого слова α_2 является буква $b \in A$. Тогда имеем, что $\psi(q_0, \beta'_m) = \psi(\varphi(q_0, \widetilde{\beta'_m}), b) = \psi(\varphi(q_0, \widetilde{\beta}), b) = \psi(q_0, \beta) = 1$ и, значит, $\beta'_m \in E_{\Omega(\alpha)}$ (здесь мы использовали то, что последней буквой слов β'_m, β и α_2 является b).

При ненулевых и линейно независимых матрицах $\Omega(\alpha_1)$ и $\Omega(\alpha_2)$, $\Omega(\alpha_1) + \Omega(\alpha_2) = \Omega_1 + \Omega_2 = \Omega$, имеем: $\exists a_1, a_2, a_3, a_4 \in A, (a_1, a_2) \neq (a_3, a_4)$, т.ч. $\omega_{a_1a_2}(\alpha_1) > 0$ и $\omega_{a_3a_4}(\alpha_2) > 0$ (что, в свою очередь, означает $\omega_{a_1a_2}(\alpha) = \omega_{a_1a_2}(\alpha_1) + \omega_{a_1a_2}(\alpha_2) > 0$ и $\omega_{a_3a_4}(\alpha) = \omega_{a_3a_4}(\alpha_1) + \omega_{a_3a_4}(\alpha_2) > 0$), а также не существует двух коэффициентов $c_1, c_2 \in \mathbb{R}, (c_1, c_2) \neq (0, 0)$, т.ч.

$c_1\omega_{a_1a_2}(\alpha_1) + c_2\omega_{a_1a_2}(\alpha_2) = 0,$

$c_1\omega_{a_3a_4}(\alpha_1) + c_2\omega_{a_3a_4}(\alpha_2) = 0.$

т.е. определитель

$$\begin{vmatrix} \omega_{a_1a_2}(\alpha_1) & \omega_{a_1a_2}(\alpha_2) \\ \omega_{a_3a_4}(\alpha_1) & \omega_{a_3a_4}(\alpha_2) \end{vmatrix} \neq 0 \tag{4}$$

Т.к. для $\forall \gamma \in E_\Omega \ \exists k \in \mathbb{N}$, т.ч. $\Omega(\gamma) = k\Omega$, то выполняется равенство $\frac{\omega_{a_1a_2}(\gamma)}{\omega_{a_3a_4}(\gamma)} = \frac{k\omega_{a_1a_2}}{k\omega_{a_3a_4}} = \frac{\omega_{a_1a_2}}{\omega_{a_3a_4}} = c = const > 0$. Рассчитаем отношение для $\beta'_m \in E_{\Omega(\alpha)}$:

$$\frac{\omega_{a_1a_2}(\beta'_m)}{\omega_{a_3a_4}(\beta'_m)} = \frac{(s + (m-1)(j-i))\omega_{a_1a_2}(\alpha_1) + s\omega_{a_1a_2}(\alpha_2)}{(s + (m-1)(j-i))\omega_{a_3a_4}(\alpha_1) + s\omega_{a_3a_4}(\alpha_2)}.$$

Т.о., это отношение имеет вид отношения двух линейных функций $\frac{ux+v}{zx+t}$ от переменной $x = s + (m-1)(j-i)$. Очевидно, что это отношение будет константным, если постоянные коэффициенты в числителе u, v будут прямо пропорциональны коэффициентам z, t в знаменателе. Значит, $\exists d \in \mathbb{R}, d > 0$, т.ч.

$\omega_{a_1a_2}(\alpha_1) = d\omega_{a_3a_4}(\alpha_1)$,

$\omega_{a_1a_2}(\alpha_2) = d\omega_{a_3a_4}(\alpha_2)$.

Подставляя эти выражения для расчета определителя (4), получим противоречие (две пропорциональные строки в детерминанте, значит, он нулевой). Значит, $\beta'_m \notin E_\Omega$, противоречие с предположением о существовании автомата, представляющего множество E_Ω и т.о. регулярности E_Ω.

2) Пусть набор Ω таков, что не существует такого разложения Ω в сумму двух ненулевых линейно независимых матриц $\Omega = \Omega_1 + \Omega_2$, т.ч. обе матрицы Ω_1 и Ω_2 задают ориентированные графы G_{Ω_1} и G_{Ω_2}, которые являются эйлеровыми. Докажем, что в этом случае граф G_Ω будет из себя представлять либо множественную петлю (см. Рис. 1.3 a)), либо набор „параллельных“ элементарных циклов (см. на Рис. 1.3 b)) — т. е. будет являться простым циклом.

Предположим, что граф G_Ω будет иметь отличный от изображенных на Рис. 1.3 a) и Рис. 1.3 b) вид. Значит, это будет либо цикл без самопересечений, но с петлями (см. Рис. 1.4 a)), либо самопересекающийся цикл (см. на Рис. 1.4 b)), либо комбинация этих двух случаев — самопересекающийся цикл с петлями.

В первом случае он разлагается на сумму цикла без самопересечений и петли (и поэтому противоречие с условием 2) теоремы), во втором — на сумму двух циклов с общей вершиной в точке пересечения (и опять противоречие о невозможности разложения в два различных цикла). Значит, предположение неверно. Более того, по условию неразрывности для биграммных языков с закольцовыванием (Лемма 3.2), количество входящих в любую вершину ребер равно количеству исходящих ребер, что дает для случая цикла без самопересечений на Рис. 1.3 b) одинаковое количество ребер между любыми двумя соединенными ребрами вершинами. Т.о., граф G_Ω будет из себя представлять один из двух видов, представленных на Рис. 1.3.

Заметим, что при умножении матрицы Ω на $\forall k \in \mathbb{N}$, вид графа $G_{k\Omega}$ останется тем же, что и был для G_Ω, поскольку ребер, которые соединяют ранее не связанные вершины, при такой операции не появится, и при этом количество входящих в любую вершину ребер останется равным количеству исходящих ребер, поскольку при этом только увеличиваются в k раз кратности всех ребер графа G_Ω.

Т.о., в матрице Ω все ненулевые элементы равны между собой, а любой эйлеров цикл для $G_{k\Omega}, \forall k \in \mathbb{N}$ будет из себя представлять m раз повторенный эйлеров цикл для приведенной матрицы $\widehat{\Omega}$ (т.е. в данном случае матрицы, в которой на всех ненулевых местах единицы), где число повторений $m = k * \text{НОД}(\omega_{ab}, a, b \in A)$ (см. пример на Рис. 3.3).

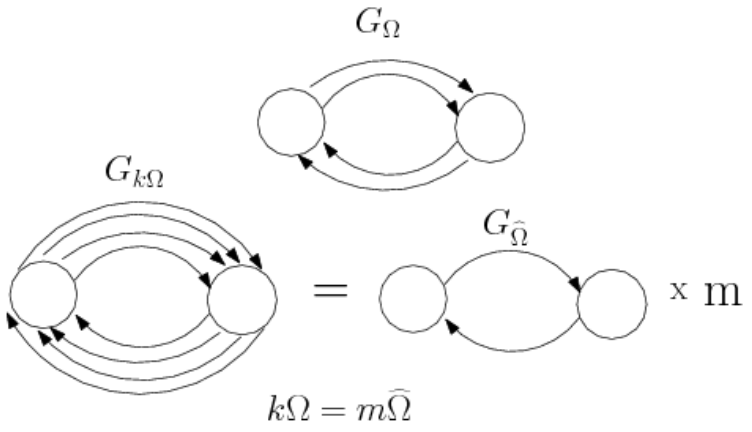


Рис. 3.3: Элементарный эйлеров цикл для k = 2, m = 4, НОД$(\omega_{ab}, a, b \in A) = 2$

При этом длина цикла для приведенной матрицы $\widehat{\Omega}$ (под длиной цикла будем понимать количество ребер в нем) будет равна в точности числу ненулевых элементов в данной матрице (и, следовательно, в матрице Ω), поскольку разные элементы матрицы соответствуют разным ребрам цикла, и наоборот.

Очевидно, что для $\forall k, k \in \mathbb{N}$ количество различных слов $\beta_k, \Omega(\beta_k) = k\Omega$ будет совпадать с количеством способов выбрать первую букву (и, следовательно, начальное ориентированное ребро) в соответствующем приведенной матрице $\widehat{\Omega}$ цикле, т.к. остальные буквы уже будут однозначно определены

самим эйлеровым циклом (при этом ни k, ни НОД$(\omega_{ab}, a, b \in A)$ в этом выборе не играют никакой роли). Т.о., для $\forall k \in \mathbb{N}$ существуют ровно l слов $\beta_{k,i}, i = 1..l$, т.ч. $\Omega(\beta_{k,i}) = k\Omega$, а l — число ненулевых элементов в наборе Ω.

■

Теорема 3.11. *Рассмотрим матрицу кратностей биграмм с закольцовыванием $\Omega \in \Xi$, задающую эйлеров граф. При этом пусть эта матрица разлагается в сумму как минимум двух линейно независимых матриц, таких, что каждая из матриц разложения задает эйлеров граф. Тогда:*
1) *Если матрица кратностей биграмм с закольцовыванием Ω разлагается единственным образом в сумму двух линейно независимых матриц $\Omega = \Omega_1 + \Omega_2$, соответствующих простым эйлеровым циклам, и не разлагается в сумму большего количества линейно независимых матриц, соответствующих эйлеровым циклам, то тогда язык E_Ω — контекстно-свободный;*
2) *В противном случае язык E_Ω не является контекстно-свободным.*

Доказательство. 1) В данном случае немного изменяется процедура построения МПА, распознающего слово из языка E_Ω.

В качестве состояния обозначим $q^I(a_{prev}, C, a^1, n_1, a^2, n_2, \delta)$, где $a_{prev} \in A$ — предыдущая буква слова при движении слева направо, $C \in \{1, 2\}$ — номер текущего цикла, $a^i \in A$ — последняя (а не первая, как в доказательстве Теоремы 1.15) буква цикла $i \in \{1, 2\}$, $0 \leqslant n_i \leqslant m_i - 1$ — какой "виток" цикла i в данный момент совершается (начиная с нуля по модулю m_i), $\delta \in \{1, 2\}$ — какой цикл добавляет символ X в магазин (то есть, если $\delta = i$, то при прохождении цикла G_{Ω_i} символ X добавляется в магазин, а при прохождении цикла $G_{\Omega_{3-i}}$ символ X, наоборот, стирается).

Тогда в случае с первой буквой входного слова, принадлежащей исключительно первому либо, наоборот, исключительно второму циклу, достаточно исправить начальные правила:

$(q_0, a_i, Z) \to (q^I(a_i, 1, a_{i-1}, 0, c_{k_3}, 0, 1), Z, \hookrightarrow), i > 1,$

$(q_0, a_1, Z) \to (q^I(a_i, 1, c_{k_3}, 0, c_{k_3}, 0, 1), Z, \hookrightarrow)$,

$(q_0, b_i, Z) \to (q^{II}(b_i, 2, c_{k_3}, 0, b_{i-1}, 0, 2), Z, \hookrightarrow), i > 1$,

$(q_0, b_1, Z) \to (q^{II}(b_i, 2, c_{k_3}, 0, c_{k_3}, 0, 2), Z, \hookrightarrow)$.

Подобным же образом изменятся и начальные правила при первой букве входного слова, общей для первого и второго циклов:

$(q_0, c_i, Z) \to (q^{III}(c_i, 1, c_{i-1}, 0, c_{i-1}, 0, 1), Z, \hookrightarrow), i > 1$,

$(q_0, c_1, Z) \to (q^{III}(c_i, 1, a_{k_1}, 0, b_{k_2}, 0, 1), Z, \hookrightarrow)$.

Также во всех правилах, где обновляются счетчики циклов $n_i, i \in \{1, 2\}$, необходимо теперь учитывать, что само обновление нужно будет производить не при достижении первой буквы соответствующего цикла, как мы делали раньше, а при достижении последней буквы (которую мы в данном случае "запоминаем" в индексе состояния).

И, наконец, обновим правила окончания распознавания:

$(q^I(a_{i-1}, 1, a_{i-1}, 0, c_{k_3}, 0, \delta), \dashv, Z) \to (f, Z, \_), 1 < i \leqslant k_1$,

$(q^I(c_{k_3}, C, c_{k_3}, 0, c_{k_3}, 0, \delta), \dashv, Z) \to (f, Z, \_)$,

$(q^{II}(b_{i-1}, 2, c_{k_3}, 0, b_{i-1}, 0, \delta), \dashv, Z) \to (f, Z, \_), 1 < i \leqslant k_2$,

$(q^{II}(c_{k_3}, C, c_{k_3}, 0, c_{k_3}, 0, \delta), \dashv, Z) \to (f, Z, \_)$,

$(q^{III}(c_{i-1}, C, c_{i-1}, 0, c_{i-1}, 0, \delta), \dashv, Z) \to (f, Z, \_), 1 < i \leqslant k_3$,

$(q^{III}(a_{k_1}, 1, a_{k_1}, 0, b_{k_2}, 0, \delta), \dashv, Z) \to (f, Z, \_)$,

$(q^{III}(b_{k_2}, 2, a_{k_1}, 0, b_{k_2}, 0, \delta), \dashv, Z) \to (f, Z, \_)$.

2) В данном случае следует обратить внимание только на пп. 2.1)–2.2) Теоремы 1.15, где мы при вычеркивании μ и ν считали корреляцию для количества биграмм (уникальных и неуникальных). В случае биграммных языков с закольцовыванием попросту не нужно рассматривать отдельно случай концевых μ и ν (раньше при их зачеркивании мы только уменьшали количество биграмм, не увеличивая), поскольку даже при вычеркивании последней или первой буквы в ω_m мы уменьшаем на 2 количество биграмм, при этом увеличивая на 1, что было подробно разобрано.

П. 2.3) Теоремы 1.15 остается полностью без изменений. ■

Замечание. В п. 1) Теоремы 3.11 E_Ω — детерминированный КС-язык (LR).

Теорема 3.12. *Бесконечный язык E_Ω, который при этом не является контекстно-свободным — контекстно-зависимый.*

Доказательство. В данном случае по сравнению с доказательством Теоремы 1.17 при подаче слова $\alpha = a_i\alpha_1 a_j, \alpha, \alpha_1 \in A^*$, необходимо также принимать во внимание и биграмму $a_j a_i$ при подсчете кратности $\omega_{a_j a_i}$.

1) При проверке $m < n^2$ биграмм на их отсутствие во входном слове $\alpha = a_i\alpha_1 a_j$ нужно также проверить, что биграмма $a_j a_i$ не входит в список отсутствующих. Для этого сначала проверяем все остальные биграммы в слове α, двигаясь по нему слева направо, а при достижении правого края слова "запоминаем" (с помощью специальной индексации состояний) крайнюю правую букву (a_j) и двигаемся теперь уже налево вплоть до левого края слова α, считываем первую букву (a_i) и теперь уже проверяем, есть ли биграмма $a_j a_i$ в списке отсутствующих для языка E_Ω.

2) При подсчете любой кратности $\omega_{a_{i_1} a_{i_2}}$, необходимо после подсчета при прохождении входного слова $\alpha = a_i\alpha_1 a_j$ слева направо, "запомнить" крайнюю правую букву слова α, после чего возвратиться налево к началу слова α и сравнить биграмму $a_j a_i$ с текущей подсчитываемой биграммой $a_{i_1} a_{i_2}$. Если биграммы совпадают, то дописываем символ $|_s$, соответствующий биграмме $a_j a_i$.

Все остальные моменты доказательства, в том числе пп. 3)–4), остаются без изменений по сравнению с Теоремой 1.17 . ■

Замечание. В условиях Теоремы 3.12 E_Ω — детерминированный КЗ-язык.

3.4 m-граммный язык

Рассмотрим, наконец, языки, заданные не матрицей кратностей биграмм, а набором кратностей m-грамм, где $m > 2$. В общем случае это m-мерная матрица $\overline{\Theta}$ с n^m неотрицательными элементами.

Построим по этому набору граф на плоскости. Для этого воспользуемся конструкциями для т. н. графов де Брёйна [11]. Для этого из любой m-граммы $a_1a_2\ldots a_{m-1}a_m$ составим две $(m-1)$-граммы: "левую" $a_1a_2\ldots a_{m-1}$ и "правую" $a_2\ldots a_{m-1}a_m$. Теперь нанесём на плоскость в качестве вершин ориентированного графа $G_{\overline{\Theta}}$ все получившиеся таким образом $(m-1)$-граммы, а количество ориентированных рёбер между $a_1a_2\ldots a_{m-1}$ и $a_2\ldots a_{m-1}a_m$ будет равно кратности m-граммы $a_1a_2\ldots a_{m-1}a_m$. Заметим, что в случае $m = 2$ вышеописанная процедура приводит к построению ориентированного графа G_Θ, соответствующего матрице биграмм Θ, который был описан в начале данной работы.

Таким образом, мы получаем ориентированный граф $G_{\overline{\Theta}}$, для которого точно также могут ставиться и решаться те же вопросы, что и для биграммных языков, поскольку графовые критерии останутся точно такими же, также как и определение линейной независимости матриц (только теперь матрицы будут m-мерными). В качестве кратностей биграмм в формулах для мощностей нужно подставлять кратность m-грамм. Аналогично, можно рассматривать понятие m-граммного языка с закольцовыванием.

Единственное отличие — теперь вместо n вершин у графа G_Θ, где n — мощность алфавита A, будет n^{m-1} вершин, соответствующих "левым" и "правым" $(m-1)$-граммам.

3.5 Возможные области применения

Полученные в работе результаты могут быть использованы в прикладных задачах поиска похожих фрагментов данных в системах хранения в силу хорошей скорости (матрица кратностей биграмм вычисляется за линейное время от длины входного слова) и простоты реализации (для каждого класса из иерархии Н. Хомского существует эффективный соответствующий распознаватель). Также представляет практическую ценность возможность подсчитывать мощность языка, заданного матрицей кратностей биграмм, как по точной аналитической формуле, так и использовать точную асимптотическую оценку для числа слов этого языка при росте их длины.

Заключение

Основными результатами работы являются:

1. Получение аналитических формул, а также точных асимптотических оценок мощности для простейших биграммных языков.

2. Получение простых графовых критериев для выделения подклассов бесконечных биграммных языков согласно иерархии Н. Хомского: регулярные, контекстно-свободные и контекстно-зависимые. Доказательство, что других подклассов нет.

3. Введение понятия биграммных языков с закольцовыванием и установление взаимно-однозначного соответствия с биграммными языками при одинаковом эйлеровом графе.

4. Получение простых графовых критериев для выделения подклассов бесконечных биграммных языков с закольцовыванием согласно иерархии Н. Хомского: регулярные, контекстно-свободные и контекстно-зависимые. Доказательство, что других подклассов нет.

5. Сведение как задач о мощности, так и задач о выделении среди бесконечных языков подклассов согласно иерархии Н. Хомского для языков, заданных m-граммами (при $m > 2$), к решённым задачам в случае биграммных языков.

Полученные автором результаты являются связующим звеном между классической математикой и программированием. В дальнейшем планируется продолжить работу в данной области.

Список литературы

[1] А. А. Марков. Пример статистического исследования над текстом “Евгения Онегина”, иллюстрирующий связь испытаний в цепь. // Известия Императорской Академии наук, серия VI. – 1913. – Т. 10. – № 3. – С. 153–162.

[2] E. P. Giachin. Phrase bigrams for continuous speech recognition. // Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on. – IEEE, 1995. – Vol. 1. – P. 225–228.

[3] U. Essen and V. Steinbiss. Cooccurrence smoothing for stochastic language modeling. // Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 International Conference on. – IEEE, 1992. – Vol. 1. – P. 161–164.

[4] J. P. Hutchinson and H. S. Wilf. On Eulerian Circuits and Words with Prescribed Adjacency Patterns. // Journal of Combinatorial Theory, Series A. – 1975. – Vol. 18. – № 1. – P. 80–87.

[5] K. H. Kim and F. Roush. Words with prescribed adjacencies. // Journal of Combinatorial Theory, Series B. – 1979. – Vol. 26. – № 1. – P. 85–97.

[6] Д. Н. Бабин. Частотные регулярные языки. // Интеллектуальные системы. – 2014. – Т. 18. – № 1. – С. 205–211.

[7] Д. Н. Бабин и А. Б. Холоденко. Об автоматной аппроксимации естественных языков. // Интеллектуальные системы. – 2008. – Т. 12. – № 1–4. – С. 125–136.

[8] N. Chomsky. Three models for the description of language. // Information Theory, IRE Transactions on. – 1956. – Vol. 2. – № 3. – P. 113–124.

[9] О. Оре. Теория графов. – М.: Наука, 1980.

[10] F. R. K. Chung. Spectral Graph Theory. – American Mathematical Soc., 1997.

[11] N. G. de Bruijn. A Combinatorial Problem. // Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen. Series A. – 1946. – Vol. 49. – № 7. – P. 758–764.

[12] G. Rozenberg and A. Salomaa. Handbook of formal languages, vol. 1: word, language, grammar. – Springer-Verlag, 1997

[13] S. C. Kleene. Representation of events in nerve nets and finite automata. // In Shannon, Claude E.; McCarthy, John. Automata Studies, Princeton University Press. – 1956. – P. 3–41.

[14] В. Б. Кудрявцев, С. В. Алешин и А. С. Подколзин. Введение в теорию автоматов. – М.: Наука, 1985.

[15] G. Kirchhoff. Über die Auflösung der Gleichungen, auf welche man bei der Untersuchung der linearen Verteilung galvanischer Ströme geführt wird. // Annalen der Physik. – 1847. – Vol. 72. – P. 497–508.

[16] T. van Aardenne-Ehrenfest and N. G. de Bruijn. Circuits and trees in oriented linear graphs. // Simon Stevin: Wis-en Natuurkundig Tijdschrift. – 1951. – Vol. 28. – P. 203–217.

[17] C. A. B. Smith and W. T. Tutte. On unicursal paths in a network of degree 4. // American Mathematical Monthly. – 1941. – Vol. 48. – P. 233–237.

[18] Y. Bar-Hillel, M. Micha Perles and Eli Shamir. On formal properties of simple phrase-structure grammars. // Zeitschrift fur Phonetik, Sprachwissenschaft, und Kommunikationsforschung. – 1961. – Vol. 14. – № 2. – P. 143–172.

[19] Y. Bar-Hillel. Language and Information: Selected Essays on their Theory and Application. – Addison-Wesley, 1964.

[20] John E. Hopcroft and Jeffrey D. Ullman. Introduction to Automata Theory, Languages, and Computation. – Addison-Wesley, 1979.

[21] Sige-Yuki Kuroda. Classes of languages and linear-bounded automata. // Information and Control. – 1964. – Vol. 7. – № 2. – P. 207–223.

[22] Н. В. Смирнов, О. В. Сарманов и В. К. Захаров. Локальная предельная теорема для чисел переходов в цепи Маркова и ее применения. // Докл. АН СССР. – 1966. – Т. 167. – № 6. – С. 1238–1241.

[23] Н. В. Смирнов. Теория вероятностей и математическая статистика. Избранные труды. – М.: Наука, 1970.

[24] N. Chomsky. Context-free Grammars and Pushdown Storage. // MIT Res. Lab. Electron. Quart. Prog. Report. – 1962. – Vol. 65. – P. 187–194.

[25] R. J. Evey. Application of Pushdown-Store Machines. // Proceedings of the November 12–14, 1963, Fall Joint Computer Conference. – ACM, 1963. – P. 215–227.

[26] M. P. Schützenberger. On Context-Free Languages and Push-Down Automata. // Information and Control. – 1963. – Vol. 6. – № 3. – P. 246–264.

Публикации автора по теме диссертации

[27] А. А. Петюшко. Частотные языки. // Интеллектуальные системы в производстве. – 2012. – Т. 19. – № 1. – С. 192–201.

[28] А. А. Петюшко. О частотных языках на биграммах. // Интеллектуальные системы. – 2013. – Т. 17. – № 1–4. – С. 363–365.

[29] А. А. Петюшко. О биграммных языках. // Дискретная математика. – 2013. – Т. 25. – № 3. – С. 64–77.

[30] А. А. Петюшко. О мощности биграммных языков. // Дискретная математика. – 2014. – Т. 26. – № 2. – С. 71–82.

[31] А. А. Петюшко. О контекстно-свободных биграммных языках. // Интеллектуальные системы. Теория и приложения. – 2015. – Т. 19. – № 2. – С. 187–208.

[32] A. A. Petushko. On bigram languages. // Discrete Mathematics and Applications. – 2014. – Vol. 23. – № 5-6. – P. 463–477.

[33] A. A. Petushko. On cardinality of bigram languages. // Discrete Mathematics and Applications. – 2014. – Vol. 24. – № 3. – P. 153–162.

[34] А. А. Петюшко. О частотных языках на биграммах. // Материалы X Международной конференции “Интеллектуальные системы и компьютерные науки”. – М.: Изд-во мех.-мат. фак-та МГУ, 2011. – С. 287–289.

[35] А. А. Петюшко. О мощности языка, заданного матрицей кратностей биграмм. // Материалы Международной конференции студентов, аспирантов и молодых учёных “Ломоносов-2012”. – URL: `http://lomonosov-msu.ru/archive/Lomonosov_2012/1793/32063_f0f1.pdf`.

[36] А. А. Петюшко. Об асимптотических оценках для биграммных языков. // Материалы XI Международного семинара, посвящённого 80-летию со дня рождения академика О.Б.Лупанова “Дискретная математика и ее приложения”. – М.: Изд-во мех.-мат. фак-та МГУ, 2012. – С. 363–365.

[37] А. А. Петюшко. О биграммных языках с закольцовыванием. // Материалы XVII Международной конференции “Проблемы теоретической кибернетики”. – Казань: Отечество, 2014. – С. 226–229.