ФГБОУ ВО Московский государственный
университет им. М. В. Ломоносова

На правах рукописи

Петюшко Александр Александрович

БИГРАММНЫЕ ЯЗЫКИ

Специальность 01.01.09 — дискретная математика
и математическая кибернетика

Автореферат

диссертации на соискание учёной степени
кандидата физико-математических наук

Москва — 2015

Работа выполнена на кафедре Математической теории интеллектуальных систем Механико-математического факультета ФГБОУ ВО «Московский государственный университет имени М. В. Ломоносова».

Научный руководитель:	**Бабин Дмитрий Николаевич** доктор физико-математических наук, профессор
Официальные оппоненты:	**Чечкин Александр Витальевич** доктор физико-математических наук, профессор кафедра “Математика-1”, ФГБОУ ВО “Финансовый университет при Правительстве Российской Федерации”
	Холоденко Александр Борисович кандидат физико-математических наук, ООО “Центр прикладной соционики”
Ведущая организация:	**ФГБОУ ВО “Московский государственный университет информационных технологий, радиотехники и электроники”**

Защита диссертации состоится 25 марта 2016 г. в 16 ч. 45 м. на заседании диссертационного совета Д.501.001.84, на базе ФГБОУ ВО МГУ имени М.В. Ломоносова, по адресу: Российская Федерация, 119991, Москва, ГСП-1, Ленинские горы, д.1, ФГБОУ ВО МГУ имени М.В. Ломоносова, Механико-математический факультет, аудитория 14-08.

С диссертацией можно ознакомиться в Фундаментальной библиотеке ФГБОУ ВО МГУ имени М.В. Ломоносова, по адресу: Москва, Ломоносовский проспект, д. 27, сектор А, `http://mech.math.msu.su/~snark/index.cgi`, `http://istina.msu.ru/dissertations/11791569/`.

Автореферат разослан 25 февраля 2016 г.

Учёный секретарь
диссертационного совета Д.501.001.84
на базе ФГБОУ ВО МГУ,
д.ф.-м.н., профессор **Шафаревич Андрей Игоревич**

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].

Возможны модификации задачи исследований в области формальных

[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, А. Б. Холоденко. Об автоматной аппроксимации естественных языков. // Интеллектуальные системы. – 2008. – Т. 12. – № 1–4. – С. 125–136.

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 мощности языка.

Новизна работы состоит в изучении бесконечных биграммных языков, заданных кратностями биграмм. Одним из способов задания таких бесконечных языков может служить выполнение свойства сохранения относительных частот пар соседних букв для всех слов из языка. В этом случае становится возможным классифицировать такие формальные языки согласно общепринятой иерархии Н. Хомского формальных языков[7].

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.

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

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 этого числа.
- 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:

- Научный 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 [1 – 11], из них 7 [1 – 7] в научных журналах из перечня, рекомендованного ВАК РФ. Список публикаций автора приводится в конце автореферата.

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 $\theta_{a_ia_j}$.

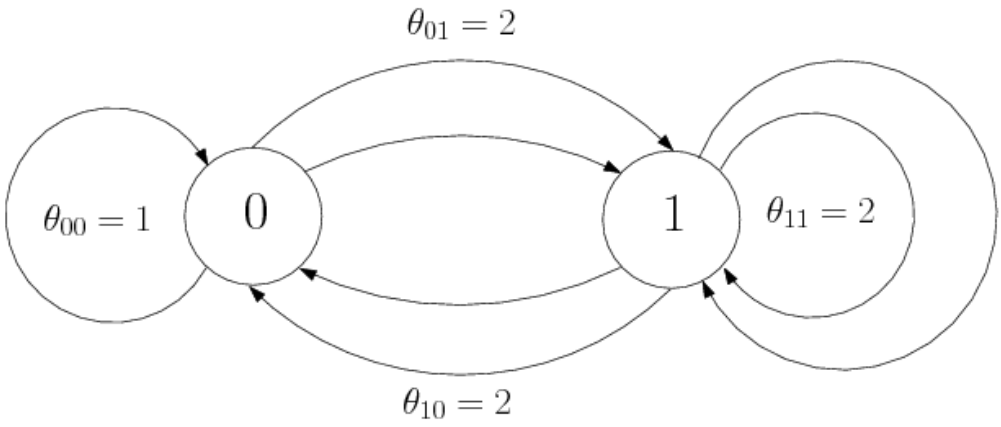
Введём понятие простейшего биграммного языка.

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* $l_\Theta=\sum\limits_{a_i,a_j\in A}\theta_{a_ia_j}+1$.

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[8], которые нам потребуются в дальнейшем.

[8] О. Оре. Теория графов. – М.: Наука, 1980.

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.*

Выделим классы языков согласно иерархии Н. Хомского[9] среди счётных биграммных языков. Для этого нам потребуется следующее определение:

Определение 1.14. Назовём 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.*

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

Замечание. В п. 1) 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 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 $D(\Theta)$).*

[10] F. R. K. Chung. Spectral Graph Theory. – American Mathematical Soc., 1997.

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

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.

Выделим 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) 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 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 $a_1a_2\ldots a_{m-1}a_m$ 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.

В **Заключении** представлены основные результаты диссертации.

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.

[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.

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.

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.

Список работ автора по теме диссертации

Из списка ВАК:

1. А. А. Петюшко. Частотные языки. // Интеллектуальные системы в производстве. – 2012. – Т. 19. – № 1. – С. 192–201.

2. А. А. Петюшко. О частотных языках на биграммах. // Интеллектуальные системы. – 2013. – Т. 17. – № 1–4. – С. 363–365.

3. А. А. Петюшко. О биграммных языках. // Дискретная математика. – 2013. – Т. 25. – № 3. – С. 64–77.

4. А. А. Петюшко. О мощности биграммных языков. // Дискретная математика. – 2014. – Т. 26. – № 2. – С. 71–82.

5. А. А. Петюшко. О контекстно-свободных биграммных языках. // Интеллектуальные системы. Теория и приложения. – 2015. – Т. 19. – № 2. – С. 187–208.

6. A. A. Petushko. On bigram languages. // Discrete Mathematics and Applications. – 2014. – Vol. 23. – № 5-6. – P. 463–477.

7. A. A. Petushko. On cardinality of bigram languages. // Discrete Mathematics and Applications. – 2014. – Vol. 24. – № 3. – P. 153–162.

Не из списка ВАК:

8. 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.

9. 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`.

10. 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.

11. А. А. Петюшко. О биграммных языках с закольцовыванием. // Материалы XVII Международной конференции “Проблемы теоретической кибернетики”. – Казань: Отечество, 2014. – С. 226–229.