На правах рукописи

БОРИСОВ Александр Евгеньевич

Закономерности в словах стохастических КС-языков с двумя классами нетерминальных символов. Вопросы экономного кодирования

01.01.09 - дискретная математика и математическая кибернетика

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата физико-математических наук

Нижний Новгород - 2006

Работа выполнена в Нижегородском государственном университете им. Н.И. Лобачевского.

Научный руководитель:	доктор физико-математических наук, доцент Л.П. Жильцова
Официальные оппоненты:	доктор физико-математических наук, профессор М.Ю. Мошков кандидат физико-математических наук, доцент В.В. Носков
Ведущая организация:	Московский государственный университет им. М.В. Ломоносова (механико-математический факультет)

Защита состоится 21 сентября 2006 г. в 14:40, на заседании диссертационного совета Д 212.166.06 в Нижегородском государственном университете им. Н. И. Лобачевского по адресу: 603950, г. Нижний Новгород, пр. Гагарина, 23, конференц-зал ННГУ, корп. 2.

С диссертацией можно ознакомиться в фундаментальной библиотеке ННГУ.

Автореферат разослан 25 июля 2006 г.

Ученый секретарь
диссертационного совета,
кандидат физико-математических
наук, доцент В. И. Лукьянов

Общая характеристика работы.

Актуальность темы. Вопросы оптимального кодирования сообщений (слов), порождаемых некоторым вероятностным источником, возникают в задачах, связанных с передачей и хранением информации. Под оптимальным кодированием, как правило, понимают кодирование, которое дает максимальную степень сжатия информации (отношение длины исходного слова к длине его кода). Возможная степень сжатия информации определяется вероятностными и структурными свойствами источника информации. Хорошо известны и широко используются на практике алгоритмы побуквенного кодирования, которые учитывают только частоты букв. Наиболее известными являются алгоритмы Хаффмана, Фано, Шеннона, алгоритм арифметического кодирования, которые применяются при отсутствии априорных знаний о структуре кодируемых слов (сообщений).

Хорошо известны также словарные методы сжатия, основанные на учете часто повторяемых фрагментов в кодируемом тексте. Здесь следует выделить алгоритмы Лемпеля-Зива LZ77, LZ78 и их многочисленные модификации.

Задача кодирования, учитывающего синтаксические свойства сообщений, была впервые рассмотрена К.Шенноном в его работе "Математическая теория связи". Им был рассмотрен вероятностный источник сообщений с конечным числом состояний. Такой источник фактически соответствует неразложимому марковскому процессу с конечным числом состояний.

Вопросы кодирования слов с учетом структурных свойств (синтаксиса) рассматривались в работах А.А.Маркова. Он поставил задачу кодирования не на всем множестве слов алфавита, а на некотором подмножестве слов, описываемом синтаксическими правилами. Для описания синтаксиса рассматривались в основном регулярные источники. Марков показал, что учет синтаксиса регулярных языков позволяет увеличить степень сжатия и уменьшить вычислительную сложность алгоритмов кодирования.

Ближайшим обобщением класса языков, порожденных регулярными источниками, являются языки, порожденные стохастическими контекстно-свободными (КС-) грамматиками. Л.П. Жильцовой были изучены свойства слов языков, порожденных неразложимыми КС-грамматиками, и построен алгоритм асимптотически оптимального кодирования, имеющий полиномиальную временную сложность.

Представляет теоретический и практический интерес изучение также и разложимых грамматик. Поскольку разложимые грамматики общего вида

мало изучены и их исследование чрезвычайно сложно, исследование разложимых грамматик целесообразно начать со случая грамматики с двумя классами нетерминальных символов.

Цель работы. Основной целью данной диссертационной работы является изучение свойств сообщений, являющихся словами стохастического КС-языка, порожденного разложимой КС-грамматикой с двумя классами нетерминальных символов, а также изучение вопросов оптимального кодирования таких сообщений. Рассматриваемые вопросы включают:
1) получение связи между средней длиной кода и энтропией для произвольного стохастического языка;
2) исследование свойств длинных (т.е. имеющих деревья вывода большой высоты) слов, порожденных разложимой грамматикой с двумя классами нетерминальных символов;
3) нахождение стоимости оптимального кодирования длинных слов;
4) построение эффективного алгоритма асимптотически оптимального кодирования слов языка.

Научная новизна. Результаты, полученные в диссертации, являются новыми. Они имеют теоретическое и познавательное значение для теории кодирования и изучения свойств формальных языков. Данная работа является продолжением исследований, проведенных Л.П. Жильцовой для неразложимых грамматик, на разложимые. Наиболее важными являются следующие результаты.
1) Получены нижние оценки средней длины кода для произвольного стохастического языка в зависимости от энтропии, которые являются более точными при больших значениях энтропии, чем полученные Л.П. Жильцовой[1].
2) Изучены свойства деревьев вывода слов стохастического КС-языка, порожденного разложимой грамматикой с двумя классами нетерминальных символов, в случае, когда перронов корень матрицы первых моментов не превосходит единицы.
3) С помощью результатов 1) и 2) найдена нижняя оценка стоимости оптимального двоичного кодирования слов языка, порожденного рассмотренной грамматикой, доказано, что эта оценка является точной.
4) Доказано, что схема асимптотически оптимального блочного кодирования выводов слов, предложенного Л.П.Жильцовой для неразложимо-

[1]Zhiltsova L. On Entropy and Optimal Coding Cost for Stochastic Language//Fundamenta Informaticae.-1998.-V.36.-P.285-305.

го случая, является асимптотически оптимальной и для рассматриваемой разложимой грамматики.

Методы исследования. В работе использованы методы дискретной математики и математической кибернетики, относящиеся к теории кодирования, теории формальных языков, теории ветвящихся процессов, линейной алгебре, математическому анализу и теории рекуррентных соотношений.

Теоретическая и практическая значимость. Результаты, полученные в диссертационной работе, имеют теоретический характер. Они могут быть использованы для дальнейшего изучения свойств разложимых грамматик. Алгоритм экономного кодирования, учитывающий вероятностные и структурные свойства сообщений, может применяться в прикладных задачах экономного кодирования.

Апробация работы. Результаты диссертации докладывались на V и VII Международной научной конференции "Дискретные модели в теории управляющих систем"(Дубна, 2003 и Покровское, 2006), Межгосударственной школе-семинаре "Синтез и сложность управляющих систем"(Н. Новгород, 2003), VIII Международном семинаре "Дискретная математика и ее приложения"(Москва, 2004), VI Международной конференции по математическому моделированию (Н. Новгород, 2004), а также на городском семинаре по дискретной математике в Нижегородском государственном университете им. Лобачевского.

Публикации. Основные результаты диссертации опубликованы в работах, список которых приведен в конце автореферата.

Структура диссертации. Диссертация состоит из введения (первая глава), трех глав, заключения и списка литературы. Объем диссертации составляет 103 страницы машинописного текста. Список литературы состоит из 33 наименований.

Содержание диссертации

I. Во введении обсуждается актуальность темы диссертации, новизна полученных результатов, теоретическая ценность работы, изложено краткое содержание и результаты диссертации.

Здесь изложена упрощенная система определений и понятий, необходимая для формулировки основных результатов.

Языком над алфавитом $B = \{b_1, \ldots, b_n\}$ называется подмножество $L \subseteq B^*$. Стохастический язык $\mathcal{L}$ определяется как пара (L, P), где L - язык, а P - распределение вероятностей на множестве его слов, причем вероятность каждого слова языка больше нуля.

Стохастическая порождающая контекстно-свободная грамматика - это система $G = \langle V_T, V_N, R, s\rangle$, где V_T и V_N - множества терминальных и нетерминальных символов соответственно (в дальнейшем называемых *терминалами* и *нетерминалами*), $|V_N| = k$, $s \in V_N$ - *аксиома*, R - конечное множество правил, $R = \bigcup\limits_{i=1}^{k} R_i$, где $R_i = \{r_{i1}, \ldots, r_{in_i}\}$, и каждое правило в R_i имеет вид

$$r_{ij} : A_i \xrightarrow{p_{ij}} \beta_{ij},\ j = 1, \ldots, n_i, \text{ где } A_i \in V_N, \beta_{ij} \in (V_T \cup V_N)^*,$$

а p_{ij} - вероятность применения правила r_{ij}, удовлетворяющая условиям $0 < p_{ij} \leq 1$, $\sum\limits_{j=1}^{n_i} p_{ij} = 1$.

Применение правила к слову состоит в замене в этом слове нетерминала, стоящего в левой части правила, на правую часть применяемого правила.

Будем говорить, что слово α *непосредственно выводимо* из слова β, если $\alpha = \alpha_1 A_i \alpha_2$, $\beta = \alpha_1 \beta_{ij} \alpha_2$ для некоторых слов $\alpha_1, \alpha_2 \in (V_T \cup V_N)^*$, и грамматика содержит правило $r_{ij} : A_i \xrightarrow{p_{ij}} \beta_{ij}$. Рефлексивное транзитивное замыкание отношения непосредственной выводимости назовем *отношением выводимости*. За L_G обозначим множество слов в терминальном алфавите, выводимых из аксиомы s. Под *выводом слова* будем понимать последовательность правил грамматики, с помощью которой данное слово выводится из аксиомы. *Левый вывод* - это вывод, в котором каждое правило применяется к самому левому по порядку нетерминалу в слове. Вероятность вывода определяется как произведение вероятностей правил, образующих вывод. Вероятность $p(\alpha)$ для слова α определяется как сумма вероятностей его различных левых выводов.

КС-грамматика называется грамматикой с однозначным выводом, если любому слову соответствует единственный левый вывод. При рассмотрении вопросов кодирования слов языка всегда будем рассматривать грамматики с однозначным выводом.

Пусть $\alpha \in L_G$. Каждому левому выводу слова α соответствует *дерево вывода*, корень которого помечен аксиомой, а вершины - терминальными и

нетерминальными символами[2]. При применении правила $A_i \stackrel{p_{ij}}{\to} h_1 h_2 \dots h_m$ в вершине a, помеченной нетерминалом A_i, добавляются m дуг от a к вершинам следующего яруса, которые помечаются слева направо символами $h_1, h_2, \dots, h_m,\ h_i \in V_T \cup V_N$. Все листья дерева помечены терминальными символами, при этом само слово α получается обходом листьев дерева слева направо. Высотой дерева вывода называется наибольшая длина пути от корня до листа. Ярусы в дереве мы будем нумеровать с единицы, т.е. ярус, на котором располагается корень дерева, имеет номер один, следующий - два и т.д.

Проиллюстрируем понятие дерева вывода на примере. Рассмотрим грамматику с двумя нетерминалами A_1, A_2, тремя терминалами y, x_1, x_2 и следующими правилами :

$$r_{11} : A_1 \stackrel{p}{\to} A_1 y A_2,$$

$$r_{12} : A_1 \stackrel{1-p}{\to} \lambda,$$

$$r_{21} : A_2 \stackrel{p/2}{\to} A_2 x_1 A_2 x_2,$$

$$r_{22} : A_2 \stackrel{1-p/2}{\to} \lambda,$$

где λ - пустое слово, $0 < p < 1$. Такая грамматика порождает слова вида $yV_1 \dots yV_n$, где V_i - правильные скобочные выражения от x_1, x_2, если отождествить x_1 с открывающей скобкой, а x_2 с закрывающей. Тогда левый вывод $r_{11} r_{11} r_{12} r_{21} r_{22} r_{22} r_{21} r_{22} r_{22}$ порождает слово $yx_1x_2yx_1x_2$ (см. рис. 1).

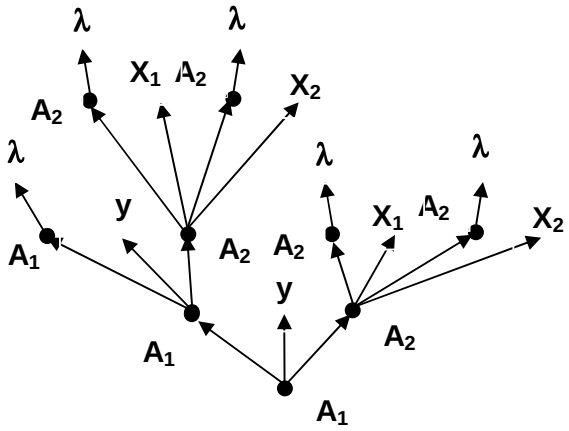

Рис. 1: Дерево вывода.

[2]Ахо А., Ульман Дж. Теория синтаксического анализа, перевода и компиляции. Том 1. М.: Мир, 1978.

Стохастическая КС-грамматика называется *согласованной*, если

$$\lim_{N\to\infty} \sum_{\alpha\in L, |\alpha|<N} p(\alpha) = 1,$$

где через $|\alpha|$ обозначается длина слова α. Согласованная грамматика порождает распределение вероятностей P на множестве слов и определяет стохастический КС-язык $\mathcal{L} = (L, P)$. В дальнейшем будем рассматривать только согласованные грамматики.

Любой стохастической КС-грамматике можно очевидным образом поставить в соответствие дискретный ветвящийся процесс[3], причем нетерминалу грамматики будет соответствовать тип частицы в ветвящемся процессе. Классу нетерминалов грамматики будет соответствовать класс типов частиц в ветвящемся процессе.

Пусть k - число нетерминалов в рассматриваемой грамматике. *Первые моменты* $a_j^i,\ i,j = 1,\ldots,k$ грамматики определяются как $a_j^i = \sum\limits_{l=1}^{n_i} p_{il} s_j^{il}$, где s_j^{il} - число нетерминалов A_j в правой части правила r_{il}. Элемент a_j^i представляет собой математическое ожидание количества нетерминалов A_j в правых частях правил, примененных к нетерминалу A_i. Матрица $A = (a_j^i)$ называется *матрицей первых моментов*. Важной характеристикой матрицы первых моментов является ее максимальный по модулю собственный корень (перронов корень). Будем обозначать его через r. Известно, что для согласованной грамматики $r \le 1$.

Вторые моменты определяются как

$$b_{jm}^i = \sum_{l=1}^{n_i} p_{il} s_j^{il} (s_m^{il} - \delta_j^m),$$

где δ_j^m - символ Кронекера ($\delta_i^i = 1,\ \delta_j^i = 0$ при $i \ne j$).

Будем говорить, что нетерминал A_j следует за нетерминалом A_i, или что A_i предшествует A_j (и обозначать $A_i \to A_j$), если из A_i выводимо хотя бы одно слово, содержащее нетерминал A_j. Грамматика называется *неразложимой*, если для любых двух различных нетерминалов A_i и A_j верно $A_i \to A_j$. В противном случае она называется *разложимой*. *Классом* нетерминалов назовем максимальное по включению подмножество $K \in V_N$, такое, что $A_i \to A_j$ для любых $A_i, A_j \in K$. Будем говорить, что класс K_1 предшествует классу K_2 (и обозначать $K_1 \prec K_2$), если для любых $A_1 \in K_1,\ A_2 \in K_2$ верно $A_1 \to A_2$. Очевидно, что множество нетерминалов V_N является объединением конечного числа непересекающихся классов.

[3] Севастьянов Б.А. Ветвящиеся процессы. М.: Наука, 1971.

Пусть G - стохастическая КС-грамматика. Через L_i обозначим язык, порожденный грамматикой, которая получена заменой аксиомы исходной грамматики на нетерминал A_i. Через D_i обозначим множество деревьев вывода слов из L_i. Обозначим через $Q_i(t)$ вероятность множества деревьев из D_i, имеющих высоту больше чем t. Эту величину назовем вероятностью продолжения по аналогии с теорией ветвящихся процессов. Пусть $P_i(t)$ - вероятность множества деревьев из D_i, имеющих высоту ровно t. Очевидно, что $P(D_i^t) = Q_i(t-1) - Q_i(t)$.

В работе рассматривается разложимая согласованная грамматика с двумя классами нетерминалов K_1, K_2, причем K_1 предшествует K_2, и $s \in K_1$. Обозначим $k_1 = |K_1|,\ k_2 = |K_2|,\ k_1 + k_2 = k$. Будем считать, что нетерминалы из K_1 имеют номера $1, \ldots, k_1$, из K_2 - соответственно $k_1 + 1, \ldots, k$. Для такой грамматики матрица первых моментов A имеет блочный вид:

$$A = \begin{pmatrix} A^{(1)} & A^{(2)} \\ 0 & A^{(3)} \end{pmatrix}, \tag{1}$$

где $A^{(1)}$ - матрица размером $k_1 \times k_1$, $A^{(3)}$ - матрица размером $k_2 \times k_2$, а $A^{(2)}$ - матрица размером $k_1 \times k_2$. Дополнительно предположим (без ограничения общности), что матрицы $A^{(1)},\ A^{(2)},\ A^{(3)}$ положительны.

Рассмотрим стохастический КС-язык $\mathcal{L} = (L, P)$, порожденный грамматикой L с однозначным выводом. Под *кодированием* языка $\mathcal{L}$ будем понимать инъективное отображение $f : L \to \{0,1\}^+$. Образ слова при отображении f назовем его кодом. Множество всех кодирований языка $\mathcal{L}$ обозначим за $F(\mathcal{L})$.

Под *энтропией языка* $\mathcal{L}$ будем понимать величину

$$H(\mathcal{L}) = -\lim_{N\to\infty} \sum_{\alpha\in L, |\alpha|\le N} p(\alpha)\cdot \log p(\alpha),$$

где через $p(\alpha)$ обозначается вероятность слова α, а $|\alpha|$ - длина слова α. Здесь и далее под log будем понимать логарифм по основанию 2. Если энтропия определена и конечна, то будем писать для краткости $H(\mathcal{L}) = -\sum\limits_{\alpha\in L} p(\alpha)\cdot\log p(\alpha)$.

Этот подход к определению стоимости кодирования был рассмотрен Л.П. Жильцовой в неразложимом случае. В некотором смысле такая постановка задачи эквивилентна задаче кодирования длинных слов, порожденных эргодическим источником, которая была рассмотрена К.Шенноном.

Для произвольного стохастического языка $\mathcal{L}$ (не обязательно контекстно-свободного), предел математического ожидания длины кодового слова

$$M(\mathcal{L}, f) = \lim_{N\to\infty} \sum_{\alpha\in L, |\alpha|\le N} p(\alpha)\cdot |f(\alpha)|$$

назовем *средней длиной кода* для кодирования f. Значение $M^*(\mathcal{L}) = \inf\limits_{f\in F(\mathcal{L})} M(\mathcal{L}, f)$ будем называть *оптимальной средней длиной кода.*

II. Во второй главе исследуется связь между энтропией и средней длиной кода для произвольного стохастического языка. Л.П.Жильцовой доказано, что если энтропия $H(\mathcal{L})$ конечна, то оптимальная средняя длина кода $M^*(\mathcal{L})$ тоже конечна и удовлетворяет неравенствам

$$H(\mathcal{L})/2 \le M^*(\mathcal{L}) < H(\mathcal{L}) + 1.$$

Во второй главе получена более точная нижняя оценка для $M^*(\mathcal{L})$, которая выполняется для больших значений энтропии, а именно:
1) если энтропия стохастического языка $\mathcal{L} = (L, P)$ удовлетворяет неравенству $1 < H(\mathcal{L}) < \infty$, то выполняется неравенство:

$$H(\mathcal{L}) \le \left(1 + \log\left(\frac{M^*(\mathcal{L})}{M^*(\mathcal{L}) - 1}\right)\right)\cdot M^*(\mathcal{L}) + \log(M^*(\mathcal{L}) - 1);$$

2) для любого $H_* > 1$ существует такой стохастический язык $\mathcal{L}$, для которого $H(\mathcal{L}) = H_*$, а неравенство пункта 1) превращается в равенство (теорема 2.2);
3) отношение $M^*(\mathcal{L})/H(\mathcal{L})$ достигает нижней оценки $1/2$ только при $H(\mathcal{L}) = H_* = 4$. При $H(\mathcal{L}) \to \infty$ и при $H(\mathcal{L}) \to 1$ отношение $M^*(\mathcal{L})/H(\mathcal{L})$ возрастает и стремится к верхней оценке 1 (теорема 2.3).

III. В третьей главе рассматривается разложимая грамматика с двумя классами нетерминалов в докритическом случае (перронов корень $r < 1$). Сначала для соответствующего грамматике ветвящегося процесса устанавливается асимптотика вероятностей продолжения.

Обозначим за v', v'' левые, за u', u'' правые собственные вектора матриц $A^{(1)}$ и $A^{(3)}$, соответствующие их перроновым корням r' и r'' при нормировке $v'u' = v''u'' = 1$, $\sum_i v'_i = \sum_i v''_i = 1$ (левый собственный вектор матрицы всегда считаем вектором-строкой).

Установлено, что для случая перронова корня кратности один вероятности продолжения имеют такую же асимптотику, как и в неразложимом

случае, а для перронова корня кратности два их асимптотика имеет другой вид, а именно:

$$Q_i(t) \sim bc_0 u'_i t r^t,\ i \le k_1,$$
$$Q_i(t) \sim c_0 u''_{i-k_1} r^t,\ i > k_1,$$

где r-перронов корень, $c_0 > 0$ - некоторая константа, и b определено формулой

$$b = v' A^{(2)} u''/r \tag{2}$$

(теорема 3.2.1). В доказательствах использовалась техника из теории ветвящихся процессов, аналогичная примененной Б.А.Севастьяновым для неразложимых процессов.

С использованием этих результатов показано, что в случае простого перронова корня вероятности продолжения, вероятность деревьев вывода высоты t при $t \to \infty$ и математические ожидания числа применений правил на фиксированном ярусе дерева вывода и во всем дереве имеют такую же асимптотику, как и в неразложимом случае.

Обозначим через $M(\xi), D(\xi)$ соответственно математическое ожидание и дисперсию случайной величины ξ. В случае кратного перронова корня математическое ожидание числа применений $q_{ij}(t,\tau)$ правила r_{ij} в дереве вывода высоты t на ярусе τ линейно зависит от τ/t при $t, \tau \to \infty$, и при этом ограничено сверху константой:

$$M(q_{ij}(t,\tau)) \sim \frac{p_{ij} \cdot (t-\tau)}{t} \cdot \left(G'_i + \frac{v'_i S'_{ij}}{r} \right)$$

для $i \le k_1$,

$$M(q_{ij}(t,\tau)) \sim \frac{p_{ij}}{t} \cdot \left(\tau \cdot (v''_{i-k_1} S''_{ij}/r + G''_i) + (t-\tau) \cdot G'_i \right)$$

для $k_1 < i \le k$. Здесь $S'_{ij},\ S''_{ij}$ - константы, определяемые по грамматике и учитывающие число нетерминалов в правых частях правил, $G'_i,\ G''_i$ - константы, определяемые вторыми моментами грамматики (теорема 3.5.1).

Доказано, что для правил, применяемых к нетерминалам из K_1, величина $M(q_{ij}(t,\tau))$ убывает, а для правил, применяемых к нетерминалам из K_2, может и убывать, и возрастать (см. пример параграфа 3.7). При этом, как и в неразложимом случае, имеет место перераспределение частот правил, т.е. при $t \to \infty$ правила, порождающие больше нетерминальных символов, используются в выводе слов чаще, что, по-видимому, нужно для достижения большей высоты дерева вывода. Величины $M(q_{ij}(t,\tau))$ не меняются при замене аксиомы на любой нетерминал из класса K_1.

Введем константы $\omega_{ij}^{(m)},\ m=1,2,\ i,j=1,\ldots,k$ равенствами

$$\begin{aligned}
&\omega_{ij}^{(1)} = p_{ij}\cdot\left(G'_i + v'_i S'_{ij}/r\right) \text{ при } i\le k_1,\\
&\omega_{ij}^{(1)} = p_{ij}G'_i \text{ при } i > k_1,\\
&\omega_{ij}^{(2)} = p_{ij}\cdot\left(G''_i + v''_{i-k_1} S''_{ij}/r\right) \text{ при } i > k_1,
\end{aligned} \tag{3}$$

$$\begin{aligned}
&\omega_{ij} = \omega_{ij}^{(1)}/2 \text{ при } i\le k_1,\\
&\omega_{ij} = (\omega_{ij}^{(1)} + \omega_{ij}^{(2)})/2 \text{ при } i > k_1.
\end{aligned} \tag{4}$$

Пусть $S_{ij}(t)$ - число правил r_{ij} в дереве вывода высоты t. В работе доказано, что математическое ожидание величины $S_{ij}(t)/t$, т.е. среднее число правил r_{ij}, приходящихся на один ярус дерева вывода высоты t, стремится к константе при $t\to\infty$:

$$M(S_{ij}(t)/t)\to\omega_{ij},\ i=1,\ldots,k,$$

где величины ω_{ij} определены формулами (4) (теорема 3.5.2).

Дисперсия этой величины для случая кратного перронова корня $r<1$ не стремится к 0 при $t\to\infty$ в отличие от неразложимого случая, а именно, справедливы соотношения

$$D(S_{ij}(t)/t)\to\left(\omega_{ij}^{(1)}\right)^2/12 \text{ при } i\le k_1,$$

$$D(S_{ij}(t)/t)\to\left(\omega_{ij}^{(2)}-\omega_{ij}^{(1)}\right)^2/12 \text{ при } i > k_1,$$

где величины $\omega_{ij}^{(l)}$ определены формулами (3) (теорема 3.6.1).

IV. За L^t обозначим множество слов языка L, деревья вывода которых имеют высоту t. Через $p_t(\alpha),\ \alpha\in L^t$, обозначим условную вероятность слова α в множестве слов L^t, она равна $p(\alpha)/P(L^t)$. Через $\mathcal{L}^t=(L^t,P_t)$ обозначим множество слов L^t с определенным на нем выше распределением вероятностей P_t. *Стоимостью кодирования* $f\in F(\mathcal{L})$ назовем величину

$$C(\mathcal{L},f)=\lim_{t\to\infty}\frac{\sum_{\alpha\in L^t}p_t(\alpha)\cdot|f(\alpha)|}{\sum_{\alpha\in L^t}p_t(\alpha)\cdot|\alpha|},$$

если этот предел существует. Через $F_*(\mathcal{L})$ обозначим множество всех кодирований, для которых величина $C(\mathcal{L},f)$ определена. Стоимостью оптимального кодирования назовем величину $C^*(\mathcal{L})=\inf\limits_{f\in F_*(\mathcal{L})}C(\mathcal{L},f)$.

С помощью найденных асимптотик для числа применений правил на ярусе дерева вывода и во всем дереве вывода в докритическом случае установлено, что энтропия $H(\mathcal{L}^t)$ множества слов, имеющих деревья вывода

высоты t, асимптотически линейно зависит от t, как и в неразложимом случае:

$$H(\mathcal{L}^t) = t \cdot \left(\log r - \sum_{i=1}^{k} \sum_{j=1}^{n_i} \omega_{ij} \log p_{ij} \right) + o(t).$$

Здесь величины ω_{ij} определены формулами (4), а p_{ij} - вероятность применения правила r_{ij} (теорема 3.8.1).

Найденная асимптотика для $H(\mathcal{L}^t)$ позволила получить нижнюю оценку $C^*(\mathcal{L})$ для стоимости оптимального (двоичного) кодирования слов языка, порождаемого рассматриваемой грамматикой:

$$C^*(\mathcal{L}) = \frac{\log r - \sum_{i,j} \omega_{ij} \log p_{ij}}{\sum_{i,j} l_{ij} \omega_{ij}},$$

где ω_{ij} определены формулами (4), p_{ij} - вероятность применения правила r_{ij}, а l_{ij} - число терминальных символов в правой части правила r_{ij} (теорема 3.8.2).

Кроме того, показано, что эта оценка является асимптотически точной, то есть существует кодирование, стоимость которого сколь угодно близка к полученной оценке $C^*(\mathcal{L})$. Для этого использовалась схема кодирования, предложенная Л.П. Жильцовой для неразложимого случая, которая состоит в кодировании последовательности правил грамматики, образующих левый вывод слова. При этом для множества правил R_i с одинаковой левой частью A_i применяется схема префиксного кодирования Шеннона с учетом найденных математических ожиданий числа применений правил в деревьях вывода. Код слова получается конкатенацией кодов правил, примененных при его выводе. Доказано, что стоимость такого кодирования стремится к $C^*(\mathcal{L})$ при укрупнении правил грамматики. Такое кодирование существенно использует и вероятностные и структурные свойства языка.

В случае простого перронова корня $r < 1$ энтропия $H(\mathcal{L}^t)$ множества слов, имеющих деревья вывода высоты t, и стоимость оптимального кодирования имеют тот же вид, что и в неразложимом случае.

V. В четвертой главе рассмотрена разложимая грамматика с двумя классами нетерминалов в критическом случае (перронов корень $r = 1$).

В первой части главы рассматривается случай кратного перронова корня. Для вероятностей продолжения соответствующего грамматике ветвящегося процесса аналогично неразложимому случаю построены рекуррентные соотношения. В результате их решения установлено, что асимп-

тотика вероятностей продолжения для нетерминалов первого класса и вероятность деревьев вывода высоты t в случае кратного перронова корня имеют другой вид по сравнению с неразложимым случаем. Доказан следующий результат.

При условии $B_1B_2 > 0$, где B_1, B_2 определяются вторыми моментами грамматики и даны формулами

$$\begin{array}{l} B_1 = \sum_{i,j,l \le k_1} v'_i b^i_{jl} u'_j u'_l, \\ B_2 = \sum_{i,j,l > k_1} v''_{i-k_1} b^i_{jl} u''_{j-k_1} u''_{l-k_1}, \end{array} \tag{5}$$

при $t \to \infty$ верны следующие асимптотические равенства:

$$Q_i(t) \sim u'_i k_0 t^{-1/2},\ P_i(t) \sim \frac{u'_i k_0}{2t^{3/2}},\ \text{при } i \le k_1,$$

$$Q_i(t) \sim \frac{2u''_{i-k_1}}{B_2 t},\ P_i(t) \sim \frac{2u''_{i-k_1}}{B_2 t^2},\ \text{при } k_1 < i \le k,$$

где $k_0 = \sqrt{\frac{4b}{B_1 B_2}}$, а b определено формулой (2) (теорема 4.1.1.2).

Таким образом, вероятности продолжения $Q_i(t)$ при $i \le k_1$ больше, чем $Q_i(t)$ при $i > k_1$.

В случае кратного перронова корня асимптотика математического ожидания $M(S_{ij}(t))$ числа применений правила $r_{ij},\ i = 1, \ldots, k,\ j = 1, \ldots, n_i$ в деревьях вывода высоты t при $t \to \infty$ была найдена в результате решения полученных рекуррентных уравнений. Доказано, что при $r' = r'' = 1$ и $B_1B_2 > 0$ справедливы следующие асимптотические равенства:

$$M(S_{ij}(t)) \sim \frac{p_{ij} B_2 t}{4b},\ i \le k_1,$$

$$M(S_{ij}(t)) \sim p_{ij} B_2 t^2/4,\ i > k_1,$$

где b введено формулой (2), B_1, B_2 определяются формулами (5), а p_{ij} - вероятность применения правила r_{ij} (теорема 4.1.2.1).

Из теоремы следует, что в случае кратного перронова корня $r = 1$ для разложимой грамматики с двумя классами нетерминалов количество правил, примененных к нетерминалам второго класса, имеет такую же ($O(t^2)$) асимптотику, как и в неразложимом случае, а количество правил, примененных к нетерминалам первого класса, значительно меньше ($O(t)$). Как и в неразложимом критическом случае, величины $M(S_{ij}(t))$ пропорциональны первоначальным вероятностям применения правил p_{ij}.

VI. С помощью установленных в случае кратного перронова корня $r = 1$ свойств деревьев вывода найдена асимптотика величины $H(\mathcal{L}^t)$ при $t \to \infty$:

$$H(\mathcal{L}^t) = \sum_{i>k_1} v''_{i-k_1} H(R_i) \cdot B_2 t^2 \cdot (1 + o(1))/4,$$

где $H(R_i) = -\sum_{j=1}^{n_j} p_{ij} \log p_{ij}$ - энтропия множества правил R_i, и $B_1 B_2 > 0$ (теорема 4.1.3.1).

С использованием этой теоремы и результатов главы 2 найдена нижняя оценка стоимости кодирования $C^*(\mathcal{L})$ для $r = r' = r'' = 1,\ B_1 B_2 > 0$:

$$C^*(\mathcal{L}) = \sum_{i>k_1} v''_{i-k_1} H(R_i) / \sum_{i>k_1} v''_{i-k_1} L(R_i),$$

где $H(R_i)$ - энтропия множества правил R_i, а $L(R_i)$ - среднее число терминалов в правой части правил из R_i (теорема 4.1.3.2).

Таким образом, в случае кратного перронова корня нижняя оценка стоимости кодирования определяется только свойствами правил, применяемых к нетерминалам из класса K_2, и совпадает со стоимостью оптимального кодирования для неразложимой грамматики, порожденной нетерминалами класса K_2. Это вызвано тем, что пропорция правил, применяемых к нетерминалам из K_1 в деревьях вывода высоты t, стремится к нулю при $t \to \infty$.

Как и в докритическом случае, доказано, что полученная оценка стоимости кодирования является точной. Установлено, что кодирование всех слов языка по невозрастанию вероятностей является оптимальным и с точки зрения кодирования слов с деревьями вывода большой высоты. Кроме того, показано, что блочное кодирование, использованное в главе 3 для докритического случая, также является асимптотически оптимальным.

VII. Во второй части главы 4 рассматривается случай простого перронова корня $r = 1$. Оказывается, что в этом случае, как и при $r < 1$, асимптотики для величин $Q_i(t), P_i(t)$, а также математических ожиданий числа применений правил на фиксированном ярусе τ дерева вывода из D_i^t и во всем дереве имеют такой же вид, как и в неразложимом случае.

Пусть $M(x_i(t,\tau))$ - математическое ожидание числа нетерминалов A_i на ярусе τ в деревьях вывода высоты t, а $M(q_{ij}(t,\tau))$ - математическое ожидание числа применений правил r_{ij} на ярусе τ в деревьях вывода высоты t. Доказано, что при $r' \neq r''$ величины $M(x_i(t,\tau)),\ M(q_{ij}(t,\tau)),\ M(S_{ij}(t))$ имеют тот же вид, что и в неразложимом случае.

В случае $r' < r'' = 1$ для величин $M(x_i(t,\tau))$, $M(q_{ij}(t,\tau))$, $M(S_{ij}(t))$ получены следующие уточненные оценки при $i \le k_1$ и $t, \tau \to \infty$:

$$M(x_i(t,\tau)) < C_1 \cdot \left(t^2 (r')^\tau / (t-\tau)\right),$$

$$M(q_{ij}(t,\tau)) < C_2 \cdot \left(t^2 (r')^\tau / (t-\tau)\right).$$

для некоторых $C_1, C_2 > 0$, и

$$M(S_{ij}(t)) \to p_{ij} \cdot (S_{ij} + p_{ij} \cdot \sum_m G_m z_m^i) z^i / u_1,$$

где S_{ij}, G_m, z^i - вычисляемые по грамматике константы (теорема 4.2.3.2).

Таким образом, в случае $0 < r' < r'' = 1$ в дереве вывода на каждом ярусе пропорция нетерминалов из класса K_1 стремится к нулю. Оценки для $H(\mathcal{L}^t)$ и $C^*(\mathcal{L})$ в этом случае имеют тот же вид, что и для неразложимой грамматики, порожденной только нетерминалами из K_2.

Выражения для энтропии $H(\mathcal{L}^t)$ множества слов $\mathcal{L}^t$ и нижней оценки стоимости кодирования $C^*(\mathcal{L})$ при $0 < r' \ne r'' = 1$ имеют такой же вид, как и в неразложимом критическом случае, причем схема кодирования, использованная для случая кратного перронова корня, является асимптотически оптимальной и в этом случае.

Список публикаций по теме диссертации

1. Борисов А.Е. О свойствах стохастического КС-языка, порожденного грамматикой с двумя классами нетерминальных символов// Дискретный анализ и исследование операций. Серия 1, т.12, N3. Новосибирск: Издательство Института математики СО РАН, 2005. С.3-31.

2. Борисов А.Е. Кодирование слов стохастического КС-языка, порожденного разложимой грамматикой с двумя нетерминалами// Вестник Нижегородского университета им. Н.И. Лобачевского. Серия Математика вып. 1(2), 2004. С. 18-28.

3. Борисов А.Е. Закономерности в деревьях вывода для стохастической разложимой КС - грамматики// Труды V Международной конференции "Дискретные модели в теории управляющих систем". М.: Изд. отдел ВМиК МГУ, 2003. С. 15-17.

4. Борисов А.Е. О свойствах стохастического КС-языка, порожденного разложимой грамматикой// Материалы XIV Международной школы-семинара "Синтез и сложность управляющих систем". Н. Новгород, 2003. С. 15-18.

5. Борисов А.Е. О свойствах слов языка, порожденного разложимой стохастической КС-грамматикой с двумя нетерминалами. Критический случай// Материалы VIII Международного семинара "Дискретная математика и ее приложения". М.: Изд. мех-мат. ф-та МГУ, 2004. С. 408-410.

6. Борисов А.Е. О числе применений правил стохастической КС-грамматики// Проблемы теоретической кибернетики. Тезисы докладов XIV Международной конференции. Изд. мех-мат. ф-та МГУ, 2005. С. 22.

7. Борисов А.Е, Жильцова Л.П. О закономерностях в словах стохастического КС - языка, порождаемого разложимой грамматикой// Труды VII Международной конференции "Дискретные модели в теории управляющих систем". М.: Изд. отдел ВМиК МГУ, 2006. С. 36-39.

8. Borisov A.E. Optimal Coding Cost for Stochastic CF-Language Induced by Decomposable Grammar// VI International Conference on Mathematical Modeling/Book of abstracts, N.Novgorod, 2004. pp. 72.