УДК 519.6

Ю.А. Котов

ДЕТЕРМИНИРОВАННАЯ ИДЕНТИФИКАЦИЯ БУКВЕННЫХ БИГРАММ В РУССКОЯЗЫЧНОМ ТЕКСТЕ

Котов Ю.А. **Детерминированная идентификация буквенных биграмм в русскоязычном тексте.**
Аннотация. В статье рассмотрена задача идентификации символов текстов на естественном языке по числовым характеристикам этих текстов. На основе правил языка и частот биграмм предложено решение данной задачи для русских текстов. Решение представляет собой систему идентифицирующих функций для каждого символа алфавита и детерминированную последовательность их применения. Указаны ограничения для полученного решения, область его эффективного применения и возможности расширения.
Ключевые слова: идентификация, символ, биграмма, русский язык, простая замена.

Kotov Yu.A. **Determinate Identification of Russian Text Letter Bigrams.**
Abstract. A problem of symbols identification of natural language texts on numerical characteristics of these texts is considered. The proposed solution for the Russian texts is based on the language rules and bigram frequency. The solution is a system of identifying functions for each character of the alphabet and a deterministic sequence of their application. The limitations, efficiency and extension options of the proposed solution are shown.
Keywords: identification; character; bigram; the Russian language; one-to-one substitution.

1. Введение. Как известно, алфавит любого языка представляет собой множество упорядоченных кодов символов, обозначающих буквы этого языка. Буквы языка однозначно связаны с их порядковым номером в исходном алфавите, но могут быть представлены разными — в том числе неизвестными — знаками, и в то же время одинаковые знаки в разных текстах могут обозначать одну и ту же, но, возможно, неизвестную букву. Далее под символом будем понимать букву языка, код (значение) которой в исходном алфавите нам может быть заранее неизвестен.

Идентифицировать символ в произвольном тексте на некотором языке — значит приписать ему такие числовые характеристики, получаемые из данного текста, которые позволяют определить номер символа в исходном алфавите этого языка и, соответственно, букву, которую данный символ представляет. Такая задача имеет не единственное решение, а его специфическая сложность заключается в том, что каждый текст обладает собственным упорядоченным множеством используемых в нем символов.

В общем случае текст может быть представлен не только в исходной, но и в произвольной, в том числе неизвестной, кодировке символов алфавита. Поэтому традиционно задача идентификации симво-

лов таких текстов трактуется как криптографическая, связанная с шифром простой замены [1-8]. Этот шифр не считается криптографически стойким, и специальных методов криптоанализа для него нет. В [1-8] указаны возможные подходы к решению данной задачи, основанные на методах:

1) перебора;
2) частотного упорядочивания;
3) вычисления энтропии;
4) марковских цепей;
5) статистической проверки гипотез;
6) генетических алгоритмов.

Метод перебора является универсальным методом решения дискретных задач. Однако, помимо известной комбинаторной сложности, для его практического применения необходимы начальные данные в виде образцов открытого текста [2], либо известных частотных характеристик [1].

Методами частотного упорядочивания будем называть методы вскрытия шифра простой замены путем прямого сопоставления частотных характеристик появления фиксированного символа или сочетания символов с эталонными частотами букв или их сочетаний [3]. Для успешного применения данных методов требуется значительный объем текста [3], и в основном они используются для получения некоторого начального приближения. Чаще всего используется частотное упорядочивание букв [1, 3-5], в [4] применяется частотное упорядочение буквенных биграмм — но для вскрытия шифра перестановки, а не замены.

Генетические алгоритмы, применяемые при вскрытии шифров, используют частотное упорядочивание для организации направленного ограниченного поиска решения — «ключа», например [6]. Они объединяют методы перебора и частотного упорядочивания в целях сокращения числа возможных переборов в пространстве поиска. Для вскрытия шифра простой замены предложен генетический алгоритм [7], использующий частотные распределения буквенных биграмм и триграмм. В основе [7] лежит сравнение формы частотного распределения элементов текста — эталонного и обрабатываемого, выделения и объединения совпадающих фрагментов таких распределений. Как и другие алгоритмы сравнения формы, [7] в значительной степени зависит от эталона, объема анализируемого текста, совпадающих и пропущенных значений; как метод перебора — от формального условия останова, и дополнительно — от значительного числа неформальных факторов (подбора фитнесс-функции, кроссовера, мутаций и т.д.),

влияющих на его эффективность. Основное значение [7] — в демонстрации применения подхода (6) к решению задачи простой замены.

Формальные методы анализа текста (3-5) используют вероятностную модель текста [4-5] и связаны с методами прикладного статистического анализа. В [1] и [5] приведены примеры подходов к решению задачи простой замены на основе методов (3) и (5). Рассмотрение этих примеров показывает сложность соответствующего анализа текстов даже при наличии существенных допущений и ограничений. Известные практические алгоритмы для решения задачи простой замены основаны на методе (4), например [8]. Но они предназначены для вскрытия более сложных шифров, чем простая замена, то есть обладают избыточной алгоритмической и вычислительной сложностью относительно данной задачи.

Ни один из двух существующих алгоритмов [7-8] и трех описанных походов [1-3,5] не дает удовлетворительного решения рассматриваемой задачи. При этом основной характеристикой так называемой «рабочей функции» [1] должна быть ее вычислительная простота, а числовые характеристики текста в различных кодировках (при простой замене символов алфавита) сохраняются [1]. Следовательно, решение задачи числовой идентификации символов языка по тексту является инвариантным относительно любых кодировок алфавита языка. Другими словами, такое решение позволяет приводить множество текстов на данном языке, но использующих различные кодировки символов алфавита, к одной исходной кодировке букв языка. Постулируем следующий подход к решению рассматриваемой задачи.

В первую очередь необходимо определить идентифицирующие числовые свойства символов для произвольного текста заданного языка. Например, для выделения символов (О, Е, А, И) из множества всех символов текста можно определить для них свойство: «разность сумм значений определенных биграмм меньше нуля», или другое свойство: «суммарное количество биграмм больше, чем у других символов», или то и другое вместе, и т.д.

Числовые свойства, идентифицирующие некоторый символ, для конкретного естественного языка могут быть выделены в результате прямого анализа (на основе правил языка) известных данных о частотных характеристиках текстов этого языка, например [5,10]. Далее будем называть такие данные образцами или эталонами.

Тогда, в отличие от вероятностной модели, текст может рассматриваться как закономерная последовательность упорядоченных по правилам языка символов его алфавита, а эталоны частотных характеристик текстов — как числовая аппроксимация этих закономерностей. Такой подход позволяет не только выделить множество идентифици-

рующих числовых признаков для каждого символа языка, но и частично решить проблему индивидуального упорядочения символов для каждого текста, так как сама идентификация проходит в фиксированной системе координат идентифицирующих признаков.

На втором шаге для множества идентифицирующих признаков каждого символа необходимо определить идентифицирующую функцию, формализующую некоторое свойство, отделяющее данный символ от множества других символов текста.

Качество полученного решения определяется его безошибочностью и инвариантностью относительно содержания и объема текстов и должно оцениваться экспериментально.

2. Определения и обозначения. Рассмотрим решение задачи идентификации символов русского текста на основе частот встречаемости пар символов — буквенных биграмм [5,10] и предложенного выше подхода. Идентификацию биграмм и символов текста будем понимать как один процесс, так как для идентификации символов нам требуется выделить определенные биграммы, а уже идентифицированные символы в свою очередь идентифицируют соответствующие им биграммы.

Над таблицей биграмм **B**, полученной для некоторого текста *t,* сформируем поле идентифицирующих признаков **W(B)**. На **B** и **W(B)** определим множество функций, имеющих экстремум при соответствии идентифицируемого символа эталонному. Значения признаков в поле **W(B)** зависят не только от значений исходных биграмм, но и от их взаимного расположения. С учетом этого определим фиксированную последовательность применения идентифицирующих функций —вектор **F**, переупорядочения **B** и пересчета **W(B)**, которые образуют алгоритм детерминированной последовательной идентификации символов текста *t*.

В соответствии с данным алгоритмом применение функций осуществляется в порядке их определения. Идентификация отдельного символа проводится однократно, т.е. множество не идентифицированных символов последовательно сокращается на один элемент после каждого применения идентифицирующей функции. Ранее идентифицированные символы в явном виде могут использоваться при идентификации последующих символов. При существовании одинаковых экстремумов для идентифицирующих функций выбирается только один из них, как правило, первый по порядку.

По завершении выделения или идентификации некоторой последовательной группы символов из **F** выполняется переупорядочение таблицы биграмм **B** в соответствии с эталоном, и затем пересчет поля признаков**W(B)**.

Основой для определения множеств **W** и **F** и последовательности идентификации послужили правила русского языка и таблицы биграмм [5,10]. Идентифицирующие функции **F** фактически являются алгебраической интерпретацией логических функций, что позволяет распространить логическую меру сравнения на некоторую область колебаний значений идентифицирующих признаков и упорядоченности символов текстов. Вероятностная мера и ее статистические аналоги в системе **F** не применяются. В силу этого рассматриваемая в данной работе идентификация не является статистической, и соответствующая терминология [9] не используется.

Входной информацией для алгоритма идентификации является таблица биграмм **B** некоторого текста *t* без учета пробела. Считается, что текст *t* представляет собой обычный, цельный и связный последовательный русский текст, в котором используются все символы (без учета пробела) русского алфавита или любая их простая замена.

Нижняя граница по объему входного текста в таком случае может составить примерно 200 символов (без учета пробелов), но более вероятно: 1000-1200 символов. На верхнюю границу устанавливается ограничение в 10000 символов (также без учета пробелов).

При превышении текстом верхней границы из него может быть выделен любым образом (в том числе и случайным) последовательный фрагмент, отвечающий ограничениям, либо проведена его нормировка или логарифмирование системы функций **F.**

Считается, что русский алфавит содержит 31 символ («Ь» — «Ь,Ъ», «Е» — «Е,Ё»). Вводится эталонная упорядоченность символов **Z** = (О, Е, А, И, Н, Т, С, В, Р, Л, П, К, Д, М, З, Г, Б, Ч, Х, Ш, Ж, Ц, Щ, Ф, У, Я, Ю, Э, Й, Ы, Ь). Номер символа в данной последовательности является его идентификатором.

Перед началом идентификации входная таблица биграмм **B** симметрично упорядочивается по частоте появления символов в тексте. Из **B** выделяются в отдельный вектор **D** диагональные значения, а в **B** они обнуляются. Осуществляется первоначальный расчет значений множества признаков **W(B)**. Они вычисляются для каждого символа $i \in$ **B** и обозначаются следующим образом:

$$q_{1,i}{}^{L} = \sum_{j=1}^{31} b_j^L, \quad b_j^L = b_{i,j}, \text{если } L = C,\ b_j^L = b_{j,i},\ \text{если } L = T;\ \ b_{i,j} \in \mathbf{B}$$

$$q_{2,i}{}^{L} = \sum_{j=1}^{31} 1(\text{если}\ \ b_j^L > 0) \quad q_{3,i}{}^{L} = \sum_{j=1}^{4} 1(b_j^L > 0) \quad q_{4,i}{}^{L} = \sum_{j=5}^{14} 1(b_j^L > 0)$$

$$q_{5,i}{}^{L} - \max(b_j^L), \quad i,j = \overline{1,4}\ \ q_{6,i}{}^{L} - \max(b_j^L), \quad i,j = \overline{5,14}$$

$$q_{7,i}{}^{L} - \max(b_j^L), \quad i, j = \overline{1,31}$$

$$r_{1,i}{}^{L} = \sum_{j=1}^{4} b_j^L \quad r_{2,i}{}^{L} = \sum_{j=5}^{14} b_j^L \quad r_{3,i}{}^{L} = r_{1,i}{}^{L} - \sum_{j=5}^{7} b_j^L$$

$$r_{4,i}{}^{L} = 2 \cdot r_{1,i}{}^{L} - q_{1,i}{}^{L} \quad r_{5,i}{}^{L} = 2 \times (r_{1,i}{}^{L} + \sum_{j=5}^{7} b_j^L) - q_{1,i}{}^{L}.$$

Признаки $q_{1,i}{}^{L}$- $q_{7,i}{}^{L}$, $r_{1,i}{}^{L} - r_{5,i}{}^{L}$образуют векторы $\mathbf{V_i^C}$, $\mathbf{V_i^T}$: $\mathbf{V_i^C}=\mathbf{Q_i^C}||\mathbf{R_i^C}$ и $\mathbf{V_i^T}=\mathbf{Q_i^T}||\mathbf{R_i^T}$,где**Q**||**R** — конкатенация **Q** и **R**.

Следующие признаки являются общими для символа *i*:

$$u_{1,i} = q_{2,i}{}^{C} + q_{2,i}{}^{T} \quad u_{2,i} = q_{2,i}{}^{C} - q_{2,i}{}^{T} \quad u_{3,i} = r_{2,i}{}^{C} + r_{2,i}{}^{T}$$

$$u_{4,i} = \sum_{j=5}^{16} b_{i,j} - \sum_{j=5}^{16} b_{j,i}$$

$$u_{5,i} = r_{1,i}{}^{C} - r_{1,i}{}^{T} \quad d_i = b_{i,i}.$$

За исключением d_iони образуют вектор $\mathbf{U_i}$. При переупорядочении **B** поле векторов **W(B)** =<$\mathbf{V^C}$,$\mathbf{V^T}$,**U**> не переупорядочивается, а только пересчитывается по завершении перестановки.

Вектор диагональных элементов **D** является структурным расширением **B** и всегда переупорядочивается одновременно с ней.

При описании идентифицирующих функций используются следующие обозначения:

$S = \max\limits_{G} \{f_s(\mathbf{X}_i^S)\}$ - определение идентифицирующей функции S,

$$S \in \mathbf{Z}, \quad \mathbf{X}_i^S \in \mathbf{B} \cup \mathbf{D} \cup \mathbf{W(B)}, \quad i \in G, G = \overline{m,n}.$$

Значением S является значение индекса $i=k$, при котором $f_S(\mathbf{X}_i^S)$ достигает максимума (или минимума, если определено «min») на множестве идентифицируемых символов G текущей таблицы биграмм: $i \in G$, $G = \overline{m,n}$.

До идентификации $i=0$, после нее устанавливается однозначное соответствие между идентифицированным символом k и идентификатором символа S.

Использование функции S может осуществляться только после ее определения как указанием биграмм ранее идентифицированного символа S, связывающих его (по строке или столбцу) с текущим сим-

волом $i{:}S_i^C$ или S_i^T, так и указанием значения идентификатора $i{=}S$ для признака из **W(B)**.

Все вычисления $f_S(\mathbf{X}_i^S)$ выполняются на множестве действительных чисел.

3. Система идентифицирующих функций. Определение идентифицирующих функций (1.1)-(7.10) и их нумерация, а также последовательность применения следуют вектору **F**:

$$\mathbf{F}=((О, Е, А, И)_1, (Й, Ы, Ь)_2, (Т, С, Н, Р, У, Я, Ю, Э)_3, (Л, П, В)_4, (О, Е, А, И)_5, (К, Д, М)_6, (З, Б, Ч, Х, Ж, Щ, Ц, Ш, Г, Ф)_7),$$

где каждой группе соответствует свой вектор Fj , $j = \overline{1,7}$.

Перед началом идентификации значение идентифицирующих функций, кроме (6.3), устанавливается равным нулю для всех не идентифицированных символов. Начальное значение для функции (6.3) равно 1000.

По завершении идентификации (выделения) каждой группы символов из **F** производится переупорядочение **B** в соответствии с **Z** и пересчет **W(B)**. При любом переупорядочивании **B** идентифицированные символы упорядочиваются в соответствии с эталоном, не идентифицированные символы сохраняют первоначальное взаимное упорядочение по частоте.

Для выборки не идентифицированных символов из текущей таблицы **B** используются следующие последовательные поля таблицы.

$$G1 = \overline{1,4}\quad G2 = \overline{5,31}\quad G3 = \overline{5,14}\quad G4 = \overline{12,31}\quad G5 = \overline{15,31}\quad G6 = \overline{31,5}\ .$$

Если все функции группы используют одно и то же поле G, то оно указывается только в заголовке группы.

Выделение групп $\mathbf{F}j$, $j = \overline{1,7}$ имеет не случайный характер. Каждую группу образуют символы, имеющие некоторые общие свойства. Например, в $\mathbf{F_1}$ входят наиболее часто встречающиеся гласные языка, проявляющие его слоговую структуру и представляющие примерно 35% объема текста и 40% биграмм, в $\mathbf{F_2}$ — символы *(Й,Ы,Ь)*, два последних из которых никогда (для «Ы» — практически никогда) не встречаются после гласных, а первый — «Й» — после согласных букв. Группу $\mathbf{F_3}$ образуют наиболее часто встречающиеся согласные и остальные гласные, которые вместе с группами $\mathbf{F_1}$и $\mathbf{F_2}$ охватывают уже 70% объема и 82% биграмм текста.

В группу $\mathbf{F_4}$ включены согласные *(Л, П, В)*, имеющие следующие особенности. Символ «Л» устойчиво образует биграмму «ЛЬ» с ранее идентифицированным символом «Ь»; символ «П» имеет ярко выраженную асимметрию по числу биграмм в сторону начала слов и

биграммы «ПР» (эффект частого использования приставок, начинающихся с «П»); символ «В» часто встречается в сочетаниях с ранее идентифицированными символами *(Т, С)*.

Группы $\mathbf{F_5}$ и $\mathbf{F_6}$ образуют символы, начинающие *(О, Е, А, И)* и завершающие *(К, Д, М)* первую половину наиболее часто используемых символов языка. В силу этого они обладают наиболее средними (во всех смыслах) характеристиками, и для их идентификации требуются многие ранее идентифицированные символы. Особенно сложно идентифицируются символы *(К, Д, М)* **—** группа функций $\mathbf{F_6}$, с привлечением большого числа ранее идентифицированных символов. И ранее сделанные ошибки идентификации могут накапливаться в этой группе.

Последнюю группу $\mathbf{F_7}$ образуют относительно редко встречающиеся согласные, имеющие в основном индивидуальные особенности.

F1. Выделение символов **«О, Е, А, И».** Из первых 10 символов **В** выделяются четыре символа (вектор **S4**) по первым четырем наибольшим значениям признака $u_{1,i}$, если для них $r_{5,i}{}^{C}<0$ или $r_{5,i}{}^{T}<0$. При переупорядочении выделенные символы переносятся в начало **В**. Взаимное расположение этих символов сохраняется.

$$\mathbf{S4} = \max\{\, u_{1,i} (\textit{если}\ \ r_{5,i}^{C} < 0 \vee r_{5,i}^{T} < 0) \,\}. \tag{1.1}$$

F2. Идентификация символов **«Й, Ы, Ь»** *на множестве* **G2.**

$$\textit{Й} = \max\{\, \textit{если}\ u_{2,i} > 0\ \textit{то} - \frac{u_{2,i} \times u_{5,i}}{q_{1,i}{}^{C} \times u_{1,i}} \,\}. \tag{2.1}$$

Затем выбираются первые два не идентифицированных элемента **В**, для которых $r_{1,i}{}^{C}=0$.

$$\textit{Ь} = \max\{\, \frac{q_{7,i}{}^{T}}{(q_{5,i}{}^{C}+1)\times(q_{2,i}{}^{T}+1)^{2}} \,\}\ \textit{из выбранных двух элементов}, \tag{2.2}$$

$$\textit{Ы} = \textit{оставшийся символ}. \tag{2.3}$$

F3. Идентификация символов **«Т, С, Н, Р, У, Я, Ю, Э».** Из **В**, начиная с пятого диагонального элемента, выделяется подматрица **SB** 15x15. Формируется матрица $\mathbf{DB}=\mathbf{SB}+\mathbf{SB}^{\mathrm{T}}$, где $\mathbf{SB}^{\mathrm{T}}$ **—** транспонированная **SB**. Для симметричной матрицы **DB**определяется первый элемент, для которого:

$$CT = max\ \left\{\ db_{i,j} \times \left(r_{2,j}{}^{C} - q_{6,i}{}^{C} + 1\right)\ \right\},\ db_{i,j} \in \mathbf{DB}, \quad i,j = 1\ldots 15$$

$$C = \text{первый индекс большего слагаемого выбранной суммы} \quad db_{CT} \tag{3.1}$$

$$T = \text{первый индекс меньшего слагаемого выбранной суммы} \quad db_{CT} \tag{3.2}$$

$$H = \max_{G2} \{ (q_{1,i}{}^{C} - u_{3,i} - Ь_i^{T} + d_i) \times u_{1,i} \} \tag{3.3}$$

$$P = \min_{G3} \{ \text{если } PP_i < 0 \text{ то } PP_i \times abs(PP_i) \times PPP_i$$
$$\text{иначе если } PP_i = 0 \text{ то } -(C_i^{T}+1)^2 \times PPP_i \text{ иначе } 0 \}, \tag{3.4}$$
$$\text{где } PP_i = u_{4,i} + Ь_i^{T},\ PPP_i = q_{4,i}{}^{C} \times q_{4,i}{}^{T} \times r_{4,i}{}^{C}$$

$$У = \max_{G2} \{ \text{если } r_{4,i}{}^{C} < 0 \wedge r_{4,i}{}^{T} < 0 \wedge (Ы I_i^{T} + Ь_i^{T}) = 0 \wedge$$
$$\wedge\ (r_{3,i}{}^{C} \le 0 \vee r_{3,i}{}^{T} \le 0) \text{ то } (P_i^{C}+1) \times q_{2,i}{}^{T} \} \tag{3.5}$$

$$Я = \max_{G2} \{ \text{если } q_{1,i}{}^{C} \ne 0 \wedge r_{4,i}{}^{C} < 0 \wedge (Ы I_i^{T} + Ь_i^{T}) = 0 \wedge r_{1,i}{}^{C} > 0$$
$$\text{то } \frac{abs(r_{4,i}{}^{C}) \times (T_i^{T}+1) \times (C_i^{C}+1) \times q_{2,i}{}^{C}}{q_{1,i}{}^{C}} \ \}. \tag{3.6}$$

Идентификация символов *«Ю»* и *«Э»* проводится по столбцу идентифицированного символа *«Т»*. Выделяются два символа *S1* и *S2*.

$$S1 = \max_{G4} \{ \text{ если } q_{1,i}{}^{C} \ne 0 \wedge r_{4,i}{}^{C} < 0 \wedge (Ы I_i^{T} + Ь_i^{T}) = 0 \text{ то } ЮЭ_i \} \tag{3.7}$$

$$S2 = \max_{G4} \{ \text{ если } q_{1,i}{}^{C} \ne 0 \text{ то } \frac{ЮЭ_i}{(r_{1,i}{}^{C}+1) \times q_{1,i}{}^{C} \times q_{2,i}{}^{C}} \ \}$$
$$\text{где } ЮЭ_i = \frac{h_i \times abs(r_{4,i}{}^{C})}{Ы I_i^{C}+1},\ \ h_i = \text{если } r_{1,i}{}^{C} = 0,\ \text{то } T_i^{T}+1 \text{ иначе } T_i^{T}. \tag{3.8}$$

$$\text{Если } r_{1.S1}{}^{C} > r_{1,S2}{}^{C}, \text{то } \{ M_3 = 2,\ M_4 = 1 \} \text{ иначе},$$
$$\text{если } r_{1,S1}{}^{C} < r_{1,S2}{}^{C}, \text{то } \{ M_3 = 1,\ M_4 = 2 \}$$
$$\text{иначе} \{ M_3 = \max\{b_{S1,j}\},\ M_4 = \max\{b_{S2,j}\} \},\quad b_{S1,j}, b_{S2,j} \in B,\quad j = \overline{14,31},$$
$$\text{Если } M_3 \ge M_4,\ \text{то } Ю = S1,\ Э = S2 \text{ иначе } Ю = S2,\ Э = S1.$$

F4. Идентификация символов «**Л, П, В**».

$$Л = \max_{G6} \{ Ь_i^{T} \} \tag{4.1}$$

$$П = \max_{G2} \{ \text{если } u_{2,i} < 0 \wedge (r_{1,i}{}^{C} - q_{5,i}{}^{C}) \ne 0 \text{ то}$$
$$\frac{(P_i^{T}+1) \times (T_i^{C} + C_i^{C} + 1) \times q_{6,i}{}^{C} \times (q_{5,i}{}^{C} \cdot q_{3,i}{}^{C} - r_{1,i}{}^{C})}{(q_{3,i}{}^{C}-1)^2} \ \} \tag{4.2}$$

$$B = \max_{G3} \{ (T_i^C + C_i^C + C_i^T + 1) \times q_{5,i}{}^T \}. \tag{4.3}$$

F5. Идентификация символов «**О, Е, А, И**» *на множестве* **G1.**

$$O = \max\{B_i^T\} \tag{5.1}$$

$$И = \max\{ (H_i^C - П_i^C) \times q_{2,i}{}^T \} \tag{5.2}$$

$$\begin{aligned} E = \max\{ & \textit{если } EE_i > 0, \textit{ то } EE_i \\ & \textit{иначе если } P_i^C < P_i^T, \textit{то } (H_i^T - H_i^C + C_i^C) \times h_i \\ & \textit{иначе если } H_i^C < H_i^T, \textit{то } (P_i^T - P_i^C + C_i^C) \times h_i \\ & \textit{иначе } (C_i^C + 1) \times h_i \ \}, \\ & \textit{где } EE_i = (H_i^T + P_i^T + C_i^C - H_i^C - P_i^C) \times h_i, \ h_i = \frac{q_{2,i}{}^T}{q_{1,i}{}^T} \end{aligned} \tag{5.3}$$

$$A = \textit{оставшийся символ множества } \{O, E, A, И\}. \tag{5.4}$$

F6. Идентификация символов «**К, Д, М**» *на множестве* **G4.**

$$\begin{aligned} K = \max\{ & \textit{если } O_i^T - E_i^T > 0, \textit{ то} \\ & \frac{(P_i^T + 1) \times (C_i^C + 1) \times (T_i^T + 1) \times (O_i^T - E_i^T)^2 \times q_{3,i}{}^C \times u_{1,i}}{(Ы_i^T + 1) \times q_{1,i}{}^C} \ \} \end{aligned} \tag{6.1}$$

$$\begin{aligned} & Д = \max\{ Д3_i \times q_{3,i}{}^C \times q_{2,i}{}^T \times q_{2,i}{}^C \ \}, \\ & \textit{где } Д3_i = \textit{ если } Ы_i^C \le 0, \textit{ то } \frac{Д1_i \times Д2_i \times (2 \cdot H_i^T + 1)}{h + 1} \\ & \qquad \textit{иначе } \frac{Д1_i \times Д2_i \times (2 \cdot H_i^T + 1) \times (Л_i^T + 1)}{(Ы_i^C + 1) \times (Ы_i^T + 1) \times (К_i^T + 1)}, \ h = \frac{q_{1,Ы}{}^T}{q_{2,Ы}{}^T} \\ & Д1_i = \textit{ если } O_i^C + У_i^C - И_i^C > 0 \textit{то } q_{5,i}{}^T - И_i^C + 1 \ \textit{ иначе } 1, \\ & Д2_i = \textit{ если } \ E_i^T + У_i^T - И_i^T > 0, \textit{ то} \\ & \qquad (q_{6,i}{}^C + 1) \times (У_i^T + 1) \times (E_i^T + A_i^T - И_i^T + 1) \ \textit{ иначе } E_i^T + A_i^T + 1 \end{aligned} \tag{6.2}$$

$$\begin{aligned} M = \min\{ & \textit{если } q_{3,i}{}^C = 0, \textit{ то } 1000 \textit{ иначе } \frac{h_i}{u_{1,i} \times q_{3,i}{}^C \times q_{3,i}{}^T} \ \}, \\ & \textit{где } h_i = \textit{если } r_{1,i}{}^C = 0, \textit{ то } 1000 \textit{ иначе } 1 + \frac{4 \cdot q_{5,i}{}^C - r_{1,i}{}^C}{r_{1,i}{}^C}. \end{aligned} \tag{6.3}$$

F7. Идентификация символов «**З, Б, Ч, Х, Ж, Щ, Ц, Ш, Г, Ф**» *на множестве* **G5.**

$$З = \max\{\frac{A_i^C \times A_i^T \times A_i^T \times (q_{3,i}{}^T + 1) \times q_{2,i}{}^T}{(ЬI_i{}^C + 1) \times (T_i{}^T + 1)^2}\}; \quad (7.1)$$

$$Б = \max\{\frac{(h_i + 1) \times Б2_i \times (q_{5,i}{}^C + O_i^C + 1) \times q_{2,i}{}^C \times q_{2,i}{}^T}{A_i^T + 1}\};$$

где $Б2_i = если\ ЬI_i^C = 0 \wedge ЬI_i^T = 0 \wedge Ь_i^C = 0 \wedge Ь_i^T = 0,\ то$

$$если\ N = 1,\ то \frac{У_i^C + 1}{(P_i^T + 1) \times (H_i^T + 1)},\ иначе\ \frac{У_i^C + 1}{P_i^T + 1}, \quad (7.2)$$

иначе если $N = 1$ *то* $([ЬI_i{}^T]^2 + 1) \times ([Ь_i{}^T]^2 + 1) \times ([ЬI_i{}^T]^3 + 1),$

иначе $([ЬI_i{}^T]^2 + 1) \times ([Ь_i{}^T]^2 + 1).$

h_i — максимальная биграмма среди последних 10-ти (кроме«Й,Ы,Ь») биграмм строки текущего элемента,

N — количество ненулевых биграмм по столбцу символа «Ы» в диапазоне от 15 символа до конца, за исключением «Й, Ы, Ь».

$$Ч = \max\{\frac{(E_i{}^T + A_i^T + 1)^2 \times (T_i{}^T + 1)^3 \times q_{2,i}{}^C \times q_{2,i}{}^T}{(H_i{}^C + 1) \times (Д_i{}^C + 1) \times (O_i^T + H_i^C + Д_i{}^T + 1)}\}; \quad (7.3)$$

$$X = \max\{\frac{(O_i^T + 1) \times (И_i{}^C + 1) \times (ЬI_i{}^C + 1) \times q_{2,i}{}^T}{q_{1,i}{}^C}\}; \quad (7.4)$$

$$Ж = \max\{\frac{(O_i^C + 1) \times (E_i^T + A_i^T + 1) \times (H_i^T + У_i^C + 1) \times q_{2,i}{}^C}{([O_i^T]^2 + 1) \times (abs(q_{2,i}{}^C - q_{2,i}{}^T) + 1) \times ([Ю_i^C]^3 + 1) \times (ЬI_i^C + 1) \times q_{1,i}{}^C}\}; \quad (7.5)$$

$$Щ = \max\{если\ q_{2,i}{}^C = 1 \wedge СВ_i = 1,\ то\ 0,\ иначе \frac{(E_i^T + И_i^T + 1) \times (СВ_i + 1)^2}{u_{1,i} \times q_{2,i}{}^T \times q_{1,i}{}^C}\}, \quad (7.6)$$

где $СВ_i = (У_i^C + 1) \times (Я_i^C + 1) \times (Ю_i^C + 1) \times (Б_i^C + 1).$

$$Ц = \max\{если\ q_{2,i}{}^C = 1 \wedge H_i^C = 0,\ то$$

$$если\ q_{1,i}{}^C = 1,\ то\ 0,\ иначе\ \frac{1}{(q_{1,i}{}^C)^3},\ иначе\ \frac{(E_i^T + [И_i^T]^3 + 1) \times (H_i^C + 1)}{(O_i^T + 1) \times q_{2,i}{}^C \times (q_{1,i}{}^C)^2}\} \quad (7.7)$$

$$Ш = \max\{\ \textit{если}\ q_{2,i}{}^{C} \neq 1 \wedge q_{2,i}{}^{T} \neq 1,$$
$$\textit{то} \frac{(E_i^T + И_i^T + У_i^C + Я_i^C + Ю_i^C + ЬІ_i^C + 1)}{(O_i^T + 1)\times(H_i^C + 1)\times(ЬІ_i^T + 1)\times(У_i^C + 1)\times(Ю_i^C + 1)\times q_{1,i}{}^{C}}\ \}; \tag{7.8}$$

$$Г = \max\{\ \frac{([O_i^T]^2 + 1)\times(M_i{}^{T} + 1)\times q_{2,i}{}^{T} \times r_{1,i}{}^{T}}{(r_{1,i}{}^{C} - O_i^T + ЬІ_i{}^{T} + 1)\times q_{1,i}{}^{C}}\ \}; \tag{7.9}$$

$$Ф = \textit{оставшийся символ}. \tag{7.10}$$

4. Экспериментальная проверка. Экспериментальная проверка предложенной системы идентификации (1.1-7.10) была проведена на текстах двух типов: тип 1 — научно-популярные и художественные тексты; тип 2 — тексты учебных пособий для вузов. Для каждого типа были случайным образом выбраны тексты различных авторов и жанров, для типа 2 — тексты из различных областей знаний: математика, информатика, материаловедение, химия, экономика, юриспруденция и т.д. Всего в вычислительном эксперименте участвовало 200 текстов, по 100 текстов каждого типа.

Из этих текстов случайным образом были выбраны последовательные фрагменты различной длины (здесь и далее длина фрагментов определяется без учета пробелов). Они распределены по двум группам (для каждого типа): группа 1 — фрагменты длиной от 200 до 2000 символов, с шагом 200; группа 2 — фрагменты от 2000 до 10000 символов, с шагом 2000. Всего 2194 фрагмента, тип 1 — 1065 (группа 1 — 591, группа 2 — 474), тип 2 — 1129 (группа 1 — 641, группа 2 —488). При этом в каждом из фрагментов использовались все буквы русского алфавита множества **Z**.

В эксперименте определялось три вида ошибок: О1 — ошибка полной идентификации, как количество ошибочных фрагментов, содержащих хотя бы один неправильно идентифицированный символ; О2 — ошибка идентификации символов, как среднее количество неправильно идентифицированных символов в ошибочных фрагментах данной длины; О3 — средний объем ошибочного текста, связанного с неправильно идентифицированными символами. Для ошибок О2 и О3 дополнительно определялись минимальное и максимальное значение и стандартное отклонение.

Результаты эксперимента приведены в таблице 1. В ней столбец *N* содержит значения длин текстов в символах (без учета пробела), *K* — количество текстов; столбцы *O1, O2, O3* — абсолютные значения ошибок; столбцы *min, max, S.D.* — минимальные, максимальные значения и стандартное отклонение соответствующих ошибок. Для наглядного представления о динамике ошибок кусочно-

линейная аппроксимация их нормированного значения представлена на соответствующих графиках рисунок 1.

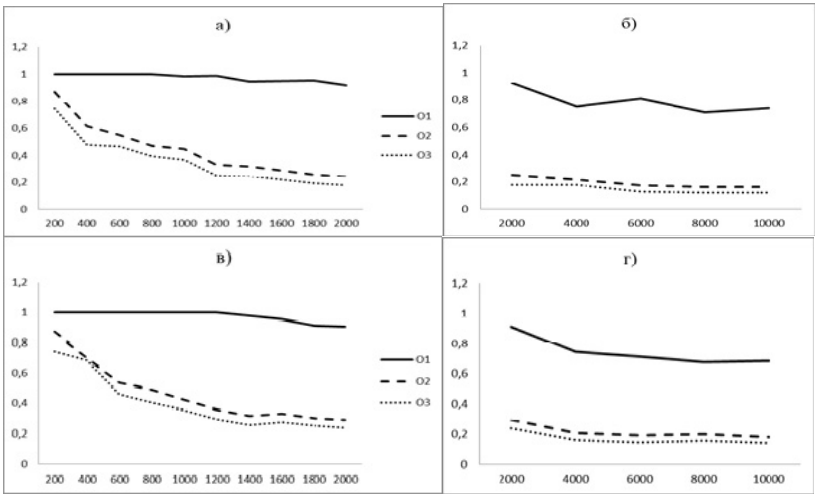

Рис. 1. Графики значений нормированных ошибок O1, O2, O3 текста типа 1: а) группа 1; б) группа 2; текста типа 2: в) группа 1; г) группа 2

Из данных эксперимента можно сделать следующие выводы:

1) в диапазоне длин фрагментов от 1000-1400 до 2000 символов точность полной идентификации меняется от 0.01 до 0.08 для текстов 1-го типа, и от 0.02 до 0.09 для текстов 2-го типа. При этом во всем диапазоне длин фрагментов от 200 до 2000 символов точность идентификации символов изменяется от 0.13 до 0.76 для текстов 1-го типа, и от 0.13 до 0.71 для текстов 2-го типа;

2) в диапазоне длин от 2000 до 10000 символов точность полной идентификации меняется от 0.08 до 0.29 для текстов 1-го типа, и от 0.09 до 0.32 для текстов 2-го типа. Точность идентификации символов изменяется от 0.76 до 0.84 для текстов 1-го типа, и от 0.71 до 0.81 для текстов 2-го типа;

3) стандартное отклонение показывает, что распределение ошибок O2 и O3 смещено в строну минимальной ошибки;

4) во всех случаях ошибка O3 коррелирует с O2 и заметно ниже нее, что указывает в первую очередь на ошибочную идентификацию символов, имеющих малую частоту появления в тексте.

Таблица 1. Результаты эксперимента

Ошибка О1			Ошибка О2				Ошибка О3			
N	*K*	*O1*	*O2*	*min*	*max*	*S.D.*	*O3*	*min*	*max*	*S.D.*
Тексты 1-го типа, группа 1										
200	1	1	27.000	27	27	0.00	149.00	149	149	0.00
400	14	14	19.071	14	27	3.75	191.86	97	364	79.39
600	34	34	17.059	8	29	4.01	279.56	90	538	100.26
800	53	53	14.660	4	30	5.85	315.75	21	777	180.34
1000	71	70	13.971	3	30	5.37	369.81	21	974	208.26
1200	79	78	10.321	2	20	4.88	303.54	8	859	213.69
1400	79	75	9.960	2	25	5.47	341.68	10	1049	264.60
1600	83	79	8.975	2	21	5.22	347.95	11	1156	295.06
1800	88	84	8.024	2	20	4.35	348.57	13	1274	289.67
2000	89	82	7.585	2	20	4.62	356.32	14	1423	321.65
Всего:	591									
Тексты 1-го типа, группа 2										
2000	89	82	7.585	2	20	4.62	356.32	14	1423	321.65
4000	93	70	6.657	2	19	6.18	707.69	18	2610	849.12
6000	99	80	5.363	2	18	4.51	755.74	33	3564	868.27
8000	97	69	4.986	2	13	4.32	951.77	37	3187	1070.40
10000	96	71	4.958	2	14	4.21	1195.97	44	5358	1336.44
Всего:	474									
Тексты 2-го типа, группа 1										
200	1	1	27.000	27	27	0.00	149.00	149	149	0.00
400	12	12	21.667	15	28	4.42	273.00	148	347	54.02
600	44	44	16.568	4	27	4.91	277.27	22	539	128.94
800	62	62	15.242	4	25	3.95	323.90	48	629	129.91
1000	77	77	13.143	4	25	4.57	354.30	45	913	200.73
1200	82	82	11.110	2	26	5.11	351.68	9	1040	232.78
1400	87	85	9.788	2	28	5.87	359.08	11	1224	302.60
1600	87	83	10.241	2	27	5.90	438.16	13	1382	340.39
1800	93	85	9.388	2	27	5.64	455.93	14	1553	352.30
2000	96	87	9.011	2	26	6.17	475.29	13	1810	425.39
Всего:	641									
Тексты 2-го типа, группа 2										
2000	96	87	9.011	2	26	6.17	475.29	13	1810	425.39
4000	98	73	6.425	2	18	5.43	650.14	29	2407	711.54
6000	98	70	5.900	2	18	5.36	890.50	16	3146	1027.14
8000	97	66	6.136	2	18	5.69	1250.33	48	5387	1431.87
10000	99	68	5.588	2	18	5.02	1425.07	31	6077	1595.10
Всего:	488									

При углубленном анализе полученных ошибок для 2-й группы текстов 1-го типа выявлено, что ошибки идентификации для различных групп символов различны. Для групп символов (О,Е,А,И), (Н,Т,С,), (В,Р,Л,П), (К,Д,М), (Й,Ы,Ь), (У,Я,Ю,Э), (З,Г,Б,Ч,Х), (Ш,Ж,Ц,Щ,Ф) ошибки идентификации появлялись в 12%, 0.8%, 22%, 32%, 9%, 7%, 63% и 89% всех случаев ошибок идентификации в текстах соответственно. Как видно из этого распределения, в наиболее значимых для идентификации группах символов ошибки встречаются реже. Поэтому даже при наличии ошибок в большинстве случаев после идентификации получается так называемый «читаемый» текст.

5. Заключение. Для предложенной системы идентификации (1.1-7.10) экспериментально установлены максимальный коэффициент полной идентификации, равный 0.3, и максимальный коэффициент идентификации символов — 0.82 при объеме анализируемого текста около 10000 символов. Важно отметить медленное нарастание ошибки идентификации символов при уменьшении объема анализируемого текста.

Система имеет потенциал прямого улучшения за счет модификации некоторых функций групп $\mathbf{F_4}$-$\mathbf{F_6}$. Но вряд ли какие либо модификации функций $\mathbf{F_7}$ позволят улучшить идентификацию символов (З, Г, Б, Ч, Х, Ш, Ж, Ц, Щ, Ф), обладающих большой информационной изменчивостью относительно частотных характеристик биграмм и дающих наибольший процент ошибок идентификации. Точность системы идентификации (1.1-7.10) и диапазон ее применения могут быть расширены применением других методов, в частности, методов прикладного статистического анализа [9].

Первичная актуальность системы (1.1-7.10) заключается в установлении функциональной связи между синтаксическими правилам русского языка и некоторыми частотными характеристиками текстов на этом языке. Прикладная актуальность, по мнению автора, не ограничивается только вскрытием шифров простой замены.

Литература

1. *Шеннон К.* Теория связи в секретных системах // Работы по теории информации и кибернетике. М.:ИЛ. 1963. С. 333–369.
2. *Бабенко Л.К., Ищукова Е.А.* Анализ симметричных криптосистем // Известия ЮФУ. Технические науки. 2012. Вып. 137. № 12. С. 136–147.
3. *Минеев М.П., Чубариков В.Н.* Лекции по арифметическим вопросам криптографии // М.: Изд-во «Попечительский совет Механико-математического факультета МГУ им. М. В. Ломоносова». 2010. 186 с.
4. *Бабаш А.В., Баранова Е.К.* Криптографические методы и средства информационной безопасности // М.:РГСУ. 2010. 65 с.
5. *Жданов О.Н., Куденкова И.А.* Криптоанализ классических шифров // Красноярск: Изд-во Сиб. гос. аэрокосм. ун-та им. акад. М.Ф. Решетнева. 2008. 107 с.

6. *Морозенко В.В., Плешкова И.Ю.* О применении генетического алгоритма для криптоанализа шифра Тритемия-Белазо-Виженера // Современные проблемы науки и образования: электронный научный журнал. 2014. №2. С. 1–11.
7. *Brownbridge J.* Decrypting Substitution Ciphers with Genetic Algorithms // Department of Computer Science. University of Cape Town. 2007. 12 p.
8. *Chen J., Rosenthal J. S.* Decrypting classical cipher text using Markov Chain Monte Carlo // Statistics and Computing. 2011. vol. 22. no. 2. pp. 397–413.
9. *Губарев В.В.* Введение в теоретическую информатику // Новосибирск: Изд-во НГТУ. 2014. 420 с.
10. *Ляшевская О.Н., Шаров С.А.* Частотный словарь современного русского языка (на материале Национального корпуса русского языка) // М.: Азбуковник. 2009. 923 с.

References

1. Shennon K. *Raboty po teorii informacii i kibernetike* [Works on the theory of information and cybernetics].M.: IL. 1963. 832 p. (In Russ.).
2. Babenko L.K., Ishhukova E.A. *Analiz simmetrichnyh kriptosistem* [Analysis of symmetric cryptosystems]. Izvestija JuFU. Tehnicheskie nauki. 2012. vol. 137. no. 12. pp. 136–147. (In Russ.).
3. Mineev M.P., Chubarikov V.N. *Lekcii po arifmeticheskim voprosam kriptografii* [Lectures on arithmetic cryptography]. M.: Izd-vo "Popechitel'skij sovet Mehaniko-matematicheskogo fakul'teta MGU im. M. V. Lomonosova". 2010. 186 p.(In Russ.).
4. Babash A.V., Baranova E.K. *Kriptograficheskie metody i sredstva informacionnoj bezopasnosti* [Cryptographic methods and means of information security]. M.: RGSU, 2010. 65 p. (In Russ.).
5. Zhdanov O.N., Kudenkova I.A. *Kriptoanaliz klassicheskih shifrov* [Cryptanalysis of classical ciphers]. Krasnojarsk: Izd-vo Sib. gos. ajerokosm. un-ta im. akad. M.F. Reshetneva. 2008. 107 p. (In Russ.).
6. Morozenko V.V., Pleshkova I.Ju. [On the application of a genetic algorithm for cryptanalysis of the cipher Triteria-Belazo-Vigenère]. *Sovremennye problemy nauki i obrazovanija: jelektronnyj nauchnyj zhurnal – Modern problems of science and education: electronic scientific journal*. 2014. no. 2. pp. 1–11. (In Russ.).
7. Brownbridge J. Decrypting Substitution Ciphers with Genetic Algorithms. Department of Computer Science. University of Cape Town. 2007. 12 p.
8. Chen J., Rosenthal J.S. Decrypting classical cipher text using Markov Chain Monte Carlo. *Statistics and Computing*. 2011. vol. 22. no. 2. pp. 397–413.
9. Gubarev V.V. *Vvedenie v teoreticheskuju informatiku* [Introduction to theoretical informatics]. Novosibirsk: Izd-vo NGTU. 2014. 420 p. (In Russ.).
10. *Ljashevskaja O.N., Sharov S.A.* Chastotnyj slovar' sovremennogo russkogo jazyka (na material Nacional'nogo korpusa russkogo jazyka) [The frequency word book of modern Russian (on material of the National case of Russian)]. M.: Azbukovnik. 2009. 923 p. (In Russ.).

Котов Юрий Алексеевич — доцент кафедры защиты информации факультета автоматики и вычислительной техники, Новосибирский государственный технический университет (НГТУ). Область научных интересов: информационная и компьютерная безопасность, криптография и криптоанализ, математическое обеспечение вычислительных систем. Число научных публикаций — 21. kotov@corp.nstu.ru; пр. К.Маркса, 20, Новосибирск, 630073; р.т.: +7(383)346-04-92, Факс: +7(383)346-04-92.

Kotov Yuri Alexeevich — associate professor of information protection department of faculty of automation and computer engineering, Novosibirsk State Technical University (NSTU). Research interests: information and computer security, cryptography, software technologies and development of information systems. The number of publications — 21. kotov@corp.nstu.ru; 20, pr. K. Marksa, Novosibirsk, 630073; office phone: +7(383)346-04-92, Fax: +7(383)346-04-92.

РЕФЕРАТ

Котов Ю.А. **Детерминированная идентификация буквенных биграмм в русскоязычном тексте.**

Задача числовой идентификации символов текста на некотором языке имеет не единственное решение. Сложность его заключается в том, что каждый текст имеет собственное упорядочение множества используемых в нем символов, если считать их неизвестными, что по существу эквивалентное простой замене символов относительно исходного алфавита.

С другой стороны, известны правила языка и некоторые совокупные данные по частотным характеристикам текстов на этом языке. Рассматривая текст как закономерную последовательность упорядоченных по правилам языка символов его алфавита, совокупные частотные характеристики — как числовую аппроксимация этих закономерностей, методом прямого анализа можно получить систему идентифицирующих признаков и функций, позволяющих решить исходную задачу.

В работе предложена такая система для решения задачи идентификации символов русского текста на основе частот встречаемости пар символов — биграмм. Функции системы фактически являются алгебраической интерпрета-цией логических функций, вероятностная мера и её статистические аналоги в рассмотрение не вводятся. Система применима для широкого круга текстов, содержащих все символы алфавита, и особенно эффективна для объемов текста от 2000 символов и выше (без учета пробела). Диапазон применения системы может быть расширен разными способами, в том числе за счет использования статистических и логико-вероятностных методов.

SUMMARY

Kotov Yu.A. **Determinate Identification of Russian Text Letter Bigrams.**

The task of numerical identification of text characters in some language have more than one unique solution. Its complexity lies in the fact that each text has its own ordering of the character set, assuming that the characters in such a set are unknown. Such assumption is equivalent to the one-to-one substitution of symbols with respect to the original alphabet .However, the rules of the language and some aggregate data on the frequency characteristics of the texts in this language are known. If we consider a text to be a logical sequence of characters of the alphabet, ordered by the rules of the language, and the aggregate frequency response — to be the numerical approximation of these laws, then by a direct analysis we can obtain an identification system of attributes and functions that allow us to solve the original problem.

In this paper, a system to solve the problem of identifying symbols of the Russian text based on occurrence frequency of characters pairs, bigrams, is proposed. The system functions are actually an algebraic interpretation of logic functions; a probability measure and its analogues are not considered. The system is applicable to a wide variety of texts, which contain all the characters of the alphabet, and is particularly effective for the text corpus starting from 2000 characters and above (excluding the gap). The application range of the system may be expanded in various ways, including using statistical and logical-probabilistic methods.