УДК 003.26.09
DOI: http://dx.doi.org/10.21686/1818-4243-2018-1-59-69

Ю.И. Петренко

Московский авиационный институт
(национальный исследовательский университет), Москва, Россия

Разделение и ранжирование гласных в русских текстах для задач криптоанализа

Эта работа была выполнена при обучении студентов криптоанализу. Курс включает изучение статистики (русских зашифрованных) текстов. Цель тренинга-научиться извлекать избыточную информацию из текста и расшифровывать криптограмму без пароля. Одним из наиболее удобных для обучения является простой метод замещения и аналогичные системы шифрования, которые представлены в большинстве курсов криптографии. В статье представлен метод автоматического разделения гласных и согласных в русских текстах, который освобождает часть избыточности шифротекста. Кроме того, этот метод значительно облегчает задачу расшифровки некоторых других симметричных шифров, которые могут быть сведены к простой подстановке.

Данная работа подготовлена в процессе обучения студентов навыкам криптографического анализа. В курс обучения входит изучение статистики русских (зашифрованных) текстов. Цель обучения – научиться извлекать избыточную информацию текста и вскрывать шифрограммы при отсутствии пароля. Одним из самых удобных для изучения является метод простой замены, который излагается в большинстве курсов криптографии. В статье рассмотрен метод автоматического метода разделения гласных и согласных в русских текстах, повышающего избыточность зашифрованного текста. Кроме того, данный метод значительно облегчает задачи вскрытия некоторых других симметричных шифров, сводящихся к простой замене.

*Целью работы является разработка автоматического метода разделения гласных и согласных в русских текстах, зашифрованных простой подстановкой и сводящихся к ней систем шифрования и выявление свойств и характеристик этого метода. Согласно теории Шеннона, для однозначной расшифровки текста требуется, чтобы его избыточность была больше, чем энтропия пароля. После разделения гласных и согласных избыточность текста возрастает до одного бита на символ, что позволяет вскрывать более короткие зашифрованные тексты. Кроме того, разделение гласных и согласных значительно облегчает задачу вскрытия некоторых шифров. Например, вскрытие самого известного способа шифрования – метода подстановок, требует выбора одного из N! возможных паролей (здесь N – количество букв алфавита). Для русского языка это составляет 33! или примерно 2 в 123 степени вариантов. После разделения гласных и согласных потребуется выбор из 10!*23!, или около 2 в 96 степени вариантов. Число вариантов пароля сокращается примерно в сто миллионов раз, и это существенно упрощает криптоанализ.*

Программа, реализующая этот метод, сначала создает матрицу вероятностей биграмм текста.

Для этой матрицы выполняется расчет критерия Маркова, определенного как разность условных вероятностей биграмм типов гласная-согласная и гласная-гласная. Разработана программа, реализующая этот алгоритм. Для текста, использующего алфавит из N символов программа находит подмножество из k «гласных». Это подмножество, максимизирующее критерий, определяется полным перебором всевозможных сочетаний из N по k символов. Порядок появления новых «гласных» при k = 1, 2, 3. . . характеризует их «силу» по убыванию и может быть использован для разделения гласных и согласных, а при достаточном объеме текста – и для приближенного ранжирования гласных. Можно достигнуть более точного ранжирования, если в качестве меры «силы» символа использовать не очередность появления «гласных» в подмножестве других «гласных», а приращение величины критерия Маркова при появлении данной буквы Работу данного алгоритма можно существенно ускорить, используя методы быстрого спуска. Испытание программы выявило независимость критерия Маркова от автора текста, а также его унимодальность для достаточно длинных текстов. Используя этот критерий, алгоритм сохраняет работоспособность при разделении гласных и согласных для коротких (до 100 символов) текстов и при ранжировании гласных для текстов размером 250–500 букв и более. Впервые замечено сходство статистики биграмм гласных букв и символов «Ь» и «Ъ», которые неотделимы данным методом от обычных гласных. Результаты испытаний показали, что данный метод можно эффективно использовать при криптоанализа коротких русских текстов, а также текстов других консонантных языков.

Ключевые слова: *криптоанализ, разделение гласных и согласных, критерий Маркова, ранжирование гласных, консонантный язык.*

Yuri I. Petrenko

Moscow Aviation Institute (National Research University), Moscow, Russia

Automatic vowels selection and ranking in Russian enciphered texts

This work was developed while teaching students the cryptanalysis. The course includes the study of statistics of (Russian encrypted) texts. The purpose of training is to learn how to extract redundant information of the text and to descript the cryptogram without a password. One of the most comfortable methods for learning is a simple substitution and similar encryption systems, which are presented in most courses of cryptography. This paper presents a method of automatic separation of vowels and consonants in Russian texts, which releases some of the redundancy of the cipher text. In addition, this method greatly facilitates the task of descript some other symmetric ciphers which may be reduced to simple substitution.

The aim of this work is to develop and implement a method for the automatic selection of vowels in Russian texts, enciphered by a simple substitution and similar encryption systems.

According to the theory of Shannon, for unambiguous decoding of the text you want the redundancy of the text that exceeds the entropy of the password. After the separation of vowels and consonants redundancy of the text increases to one bit per symbol, this allows you to open shorter encrypted texts. Moreover, the separation of vowels and consonants greatly simplifies the cryptanalysis of some ciphers. For instance, cryptanalysis of the most famous encryption method - method of simple substitution-requires selection of one of N! possible passwords (where N is the number of letters in the alphabet). For the Russian language it is 33! or nearly 2 to 123rd degrees of options. After the separation of vowels and consonants

*you will need a selection of 10!*23!, or nearly 2 to 96th degrees of options. The number of combinations is reduced to one hundred million times, that makes the cryptanalysis much easier. The program that implements this method first creates a matrix of the probabilities of bigrams of the text.*
For this matrix Markov criterion calculated, defined as the difference between the conditional probabilities of vowel-consonant and vowel-vowel diagram's types. For an alphabet consisted of N characters the program defines a combination of a given number k of "vowels" by exhaustive search. This combination of k symbols maximizes Markov criterion. The order relation of the new "vowels" for k = 1, 2, 3... characterizes the descending of their "strength" and can be used to separate vowels from consonants. In texts of sufficient volume there are possible approximate ranking of the vowel's set. A more accurate ranking is possible when as a measure of "symbol power" Markov criterion's increments are used. The algorithm speed can be greatly accelerated by using some tricks of steepest descent method. The test program discovered the independence of Markov criterion from the text's author as well as its unimodality for long texts. Using this criterion, the algorithm can separate vowels from consonants for short (up to 100 characters) texts as well as the ranking of vowels for texts as small as 250-500 letters. The similarity of Markov criterion's statistics of letters "ь", "ъ" and standard vowels is discovered. These two letters are inseparable by Markov criterion method from the standard vowels. The test results showed that Markov criterion method can be used for cryptanalysis of short Russian texts as well as texts of the other consonant languages.

Keywords: *cryptanalysis, separation of vowels and consonants, Markov criterion, ranking of vowels, consonant language.*

Введение

Разделение гласных и согласных имеют большое значение в прикладной лингвистике, информатике и технике связи.

В рассмотренном далее методе разделения гласных и согласных используется теория цепей Маркова. При анализе особенностей чередования гласных и согласных Марков обнаружил [1,2], что гласные с гораздо большей вероятностью следуют за согласной, чем за предшествующей гласной. Эту закономерность мы будем далее для краткости называть критерием Маркова.

Закономерность чередования гласных и согласных используется лингвистами. Языки, в которых гласные и согласные чередуются, получили даже специальное название консонантные или vfc-языки (vfc = vowels followed consonants and vice versa, т.е. «гласные следуют за согласными и обратно»). Определение консонантного языка качественно соответствует выполнению критерия Маркова.

Разделение гласных и согласных интересно также для криптоанализа, так как значительно облегчает задачу вскрытия некоторых шифров. Например, вскрытие самого известного способа шифрования – метода подстановок, требует выбора одного из N! возможных паролей (здесь N – количество букв алфавита). Для русского языка это составляет 33! или примерно 2^{123} вариантов. После разделения гласных и согласных потребуется выбор из 10!*23!, или около 2^{96} вариантов. Число вариантов пароля сокращается примерно в 100 000 000 раз, и это существенно упрощает криптоанализ.

Согласно теории Шеннона [3], для успешного вскрытия шифра избыточность текста должна быть не меньше, чем длина энтропия ключа шифрования. С этой точки зрения разделение гласных и согласных также весьма полезно, так как при этом избыточность текста возрастает примерно на 1 бит в пересчете на 1 символ, что позволяет выполнять дешифрование более коротких зашифрованных текстов.

В известной книге Фридмана [4] рассмотрен эффективный, но достаточно трудоемкий ручной алгоритм разделения гласных и согласных.

В работе [5 Moler C.B. Morrison Singular Value Analysis] рассмотрен машинный метод разделения гласных и согласных. Этот метод достаточно сложен, и малоэффективен на коротких текстах.

Далее излагается автоматизированный метод разделения гласных и согласных на основе критерия Маркова, использующий матрицу биграмм, позволяющий не только отделить гласные буквы от согласных, но при достаточно длинных текстах также частично упорядочить гласные.

Постановка задачи

Исторически первыми открыли и описали методы криптоанализа арабы. В 9 веке арабская наука была наиболее развитой среди всех стран средиземноморья, а их медицина и математика – лучшими в мире. Религиозные запреты ислама в области живописи и скульптуры, одновременно способствовали развитию литературы. Широко распространились словесные загадки, ребусы и каламбуры. Грамматика была важным элементом культуры арабов. Она включала в себя в том числе и тайнопись. Фактически арабы для целей криптографии использовали метод простой замены, заменяя весь алфавит целиком. Один такой алфавит, называвшийся <дауди> (по имени израильского царя Давида), использовался для шифрования трактатов по черной магии.

Как указывает Д.Канн [6], знания арабов в области криптологии были изложены в 1412 г., в одном из разделов 14-томной энциклопедии Калкашанди, под названием «Относительно сокрытия в буквах тайных сообщений», часть которого посвящена криптологии. Впервые в истории шифров в энциклопедии приводился список систем шифрования, включавших перестановки и замены. Кроме всего прочего, там было и первое в истории описание криптоаналитического исследования шифрованного текста. В по-

мощь криптоаналитику впервые приводятся списки букв, которые не встречаются вместе ни в одном слове, букв, которые редко контактируют друг с другом, а также невозможные сочетания букв. Наконец, дан список букв в порядке «частоты их использования в арабском языке в свете результатов изучения священного Корана», то есть ранжирование букв алфавита.

Фактически ранжирование всех букв алфавита достаточно для расшифровки криптограммы. К сожалению, метод частотного ранжирования редко работает.

Есть две основных причины нарушения порядка ранжирования. Во-первых, при малой длине текста могут наблюдаться отклонения от закона больших чисел, по порядку величины – это примерно корень из частот их встречаемости. Во-вторых, это могут быть систематические отклонения закона распределения, вызванные характером текста. В совокупности эти причины могут вызвать отклонения в ранжировании букв до 5 позиций. При этом замена букв в соответствии со стандартными частотами ничего не дает. Текст становится неузнаваемым. В таких случаях разделение гласных и согласных сильно упрощает криптоанализ.

В качестве самого первого рецепта для разделения гласных и согласных рекомендуется использовать различия в их частотах. Суммарная частота гласных в русском тексте составляет примерно 42% [7 Алферов], но количество гласных в русском алфавите в два раза меньше числа согласных. В результате четыре самых употребительных гласных сильно выделяются по частоте встречаемости и легко выделяются среди остальных букв. Это позволяет отделить часть гласных от остальных гласных и согласных.

Лучшие результаты можно получить, используя различия в сочетаемости букв в тексте [8 FM-34-40-2].

Действительно, гласные гораздо чаще следуют за согласными, чем за гласными. Буквы каждой из этих групп чаще контактируют с буквами другой группы, чем с буквами собственной группы. Кроме того, так как гласных меньше, чем согласных, они соседствуют с большим количеством разных букв, чем согласные. Исследуя количество контактов с разными буквами, можно достаточно надежно распределить гласные и согласные зашифрованного текста по указанным группам.

Эти соображения объясняют принцип действия так называемого метода подсчета триграмм. Для иллюстрации метода произведем подсчет триграмм в тексте [9]:

УЖНЕБООСЕНЬЮДЫША
ЛОУЖРЕЖЕСОЛНЫШКОБ
ЛИСТАЛО
КОРОЧЕСТАНОВИЛСЯ
ДЕНЬЛЕСОВТАИНСТВЕН
НАЯСЕНЬ
СПЕЧАЛЬНЫМШУМОМОБ
НАЖАЛАСЬЛОЖИЛСЯНА
ПОЛЯТУМАН

Пробелы и знаки препинания опущены, как это часто делают при шифровании. Сам текст сообщения не зашифрован для того, чтобы наглядно продемонстрировать особенности метода.

Результат подсчета триграмм показан в таблице на рис. 1. Здесь в первой строке (Ч) показаны частоты букв в данном тексте, во второй (К) – количество контактов каждой буквы, помещенной в третьей строке (Б) таблицы, а еще ниже перечень всех контактов соответствующих букв (средние буквы в триграммах пропущены, так как они одинаковы в каждом столбце). Как видно из таблицы, большим количеством контактов явно выделяются гласные буквы А, О и Е, которые почти не контактируют друг с другом. Это позволяет сделать вывод о том, что эти буквы – гласные. С этими гласными часто контактируют буквы ж, л, м, с и т, предположительно – соглас-

Ч	11	3	3	2	10	5	4	2	11	4	13	14	2	2	11	5	4	2	3	3	5	1	4
К	11	4	4	4	10	7	6	2	9	5	12	16	4	3	10	5	5	3	5	4	4	2	6
Б	А	Б	В	Д	Е	Ж	И	К	Л	М	Н	О	П	Р	С	Т	У	Ч	Щ	Ы	Ь	Ю	Я
	ШЛ	ЕО	ОИ	ЮЫ	НБ	УН	ЛС	ЩО	АО	ЫШ	ЖЕ	БО	СЕ	ЖЕ	ОЕ	СА	ОЖ	ОЕ	ЫА	ДШ	НЮ	ЬД	СД
	ТЛ	ОЛ	ОТ	ЯЕ	СН	УР	ВЛ	ОО	ОН	УО	ЕЬ	ОС	АО	ОО	ЕО	СА	ШМ	ЕА	ЫК	НШ	НЛ		АС
	ТН	ОН	ТЕ		РЖ	ЕЕ	АН		БИ	ОО	ЛЫ	ЛУ			ИТ	ВА	ТМ		МУ	НМ	НС		СН
	ТИ				ЖС	АА	ЖЛ		АО	УА	АО	ЕО			ЕТ	СВ					ЛН		ЛТ
	НЯ				ЧС	ОИ			ИС		ЕЬ	СЛ			ЛЯ	ЯУ					СЛ		
	ЧЛ				ДН				ЬЕ		ИС	КР			ЕО								
	НЖ				ЛС				АЬ		ЕН	РЧ			НТ								
	ЖЛ				ВН				АА		НА	НВ			ЯЕ								
	ЛС				СН				ЬО		ЕЬ	СВ			ЬП								
	НП				ПЧ				ИС		ЬЫ	ММ			АЬ								
	МН								ОЯ		БА	МБ			ЛЯ								
											ЯА	АЬ											
												ЛЖ											
												ПЛ											

Рис. 1. Метод триграмм

ные. Конечно, после шифрования эти и другие буквы таблицы будут переназначены, но структура таблицы сохранится, что позволит определить принадлежность данных букв к группам гласных и согласных.

Анализ данной таблицы показывает и недостатки метода подсчета триграмм. Здесь нечетко выявлены гласные «и» и «у», а букву «с» можно принять и за гласную. В основном это вызвано малым объемом текста, но все же заставляет задуматься о его надежности.

Метод подсчета триграмм можно улучшить, если анализ контактов выполнять не для каждой буквы в отдельности, а целиком для выделенных групп гласных и согласных. В результате получается так называемый метод Т-диаграмм.

Для наглядности выполним методом Т-диаграмм анализ того же текст, что и в предыдущем примере. Здесь весь алгоритм состоит из четырех шагов, показанных далее на рис. 2 в Т-диаграммах 1–4. Начало работы метода Т-диаграмм предполагает наличие какого-нибудь предположения о группах гласных и согласных. Можно полагать, что среди наиболее частых букв с большой вероятностью окажутся гласные, а среди редко используемых – согласные. Данный метод обладает хорошей сходимостью, поэтому для первого шага пригодны оба варианта. Здесь можно использовать результат частотного анализа, выполненного ранее при подсчете триграмм. Для определенности выберем в качестве «гласных» шесть наиболее частых букв: «а», «е», «л», «н» «о» и «с». (Ведь по условию мы не знаем, какие из них гласные, а какие – согласные). Далее подсчитаем количество контактов всех букв нашего текста с буквами данной группы «гласных». Результаты поместим в Т-диаграмму «1. Гласные», действительно похожую на букву «Т». В верхнюю строку помещаем перечень нашего варианта группы «гласных», а в нижних столбцах буквы, контактирующие ними слева и справа соответственно. Для наглядности в каждой ячейке помещаем столько букв, сколько их обнаружено в контактах с каждой стороны.

Анализ Т-диаграммы «1. Гласные», ясно показывает, что буквы нашей группы легко контактируют между собой, что означает наличие в группе и гласных, и согласных. Наиболее подозрительны по множеству числу «родственных контактов» буквы «л», «н» и «с». Удалим их и повторим подсчет контактов. Результаты помещаем в Т-диаграмму «2. Гласные».

Анализ Т-диаграммы «2. Гласные» показывает, что мы на верном пути. Буквы «а», «е» и «о» почти не контактируют между собой, а более склонны к контактам с буквами, не принадлежащим к их собственной группе. Выделим среди последних наиболее активные «ж», «л», «м», «н» и «с». Вероятно, это согласные. Проверим эту гипотезу, повторив подсчет контактов этой группы «согласных». Результаты поместим в Т-диаграмму «3. Согласные».

Анализ Т-диаграммы «3. Согласные» позволяет выделить еще несколько гласных. Это «и», «у» «ы» «я». Добавление их в группу гласных дает возможность сделать новый подсчет контактов. Результат дан в Т-диаграмме «4. Гласные».

Сопоставление Т-диаграмм 2 и 4 позволяет сделать, хотя и

1. Гласные	
АЕЛНОС	
ААААААА	АААА
БББ	БББ
В	ВВ
Д	
ЕЕЕЕЕЕЕ	ЕЕЕЕ
ЖЖЖ	ЖЖЖ
ІІІІІІІ	ІІІ
КК	К
ЛЛЛЛЛЛЛЛ	ЛЛЛЛЛЛ
МММ	М
НННННННН	НННННННН
ОООО	ОООООООО
ПП	ПП
РР	Р
СССС	СССССССС
ТТТ	ТТТ
	У
ЧЧ	ЧЧ
Ш	
ЬЬЬЬ	ЬЬЬЬЬ
	ЫЫ
ЯЯ	ЯЯЯЯ

2. Гласные	
АЕО	
Б	БББ
В	ВВ
Д	
ЖЖ	ЖЖЖ
	ІІ
КК	К
ЛЛЛЛ	ЛЛЛЛЛЛ М
МММ	
НННННН	НННННН
О	О
ПП	П
РР	Р
СССС	СССССС
ТТТ	
	У
ЧЧ	ЧЧ
Ш	
	Я

3. Согласные	
ЖЛМНС	
ААААААА	АААААА
ББ	
ЕЕЕЕЕЕЕЕ	ЕЕЕЕЕ
Ж	
ИИИИ	ІІІІ
ЛЛЛ	
НН	НННН
ОООООО	ОООООООО
	П
	Р
	ССС
	ТТТ
УУУУ	
	Ш
Ы	ЫЫ
ЬЬЬЬ	ЬЬЬЬЬ
ЯЯ	ЯЯЯ

4. Гласные	
АЕОПУЫЯ	
АА	
Б	БББ
ВВ	ВВ
ДД	Д
ЖЖЖ	ЖЖЖЖЖ
	ІІ
КК	К
ЛЛЛЛЛЛЛ	ЛЛЛЛЛЛЛЛ
МММ	ММММ
НННННННН	НННННННН
ОО	О
ПП	П
РР	Р
СССССС	СССССССС
ТТТТ	Т
	У
ЧЧ	ЧЧ
ШШ	ШШ
	Я

Рис. 2. Метод Т-диаграмм

с несколько меньшей уверенностью, что буквы **БВДКПРТЧШ** тоже согласные.

Итак, несмотря на большие отклонения частот букв в данном тексте, мы определили, что АЕОИУЫЯ – гласные, а ЖЛМНС и БВДКПРТЧШ – согласные.

Метод Т-диаграмм гораздо более надежен, чем предыдущий, особенно для коротких текстов. Это объясняется тем, что здесь большинство букв объединены в две группы, которые целиком участвуют в сборе статистики, что значительно уменьшает разброс анализируемых данных (дисперсию).

Недостатком метода Т-диаграмм является то, что его трудно автоматизировать, а его результаты несколько субъективны.

Для устранения этих недостатков был предложен еще один метод, который рассмотрен далее.

Реализация метода

Прежде всего полагаем, что нам задан текст с конечным, но заранее неизвестным алфавитом (например, зашифрованный). Язык текста известен и относится к консонантным языкам. Для расшифровки текста нужно полностью или частично восстановить его алфавит. Предлагаемый метод автоматического разделения гласных и согласных позволяет получить дополнительную информацию (избыточность текста) до 1 бита на букву, что значительно облегчает задачу вскрытия шифротекста. При сравнительно длинных текстах метод позволяет также ранжировать гласные, что может дать дополнительно от одного до двух битов информации на символ.

Чтобы повысить наглядность, изложение иллюстрируется примерами открытых (незашифрованных) текстов. С точки зрения применимости результатов – это допустимо, так как при использовании шифров типа подстановок и сводящихся к ним матрица биграмм инвариантна с точностью до переименования символов.

Для более четкого понимания читателю следует помнить о том, что в реальных случаях априори может быть неизвестно, какие символы алфавита отображают гласные, какие – согласные, и даже их количество может быть не задано заранее (в коротких шифровках).

Как известно [7], вопрос о зависимости букв алфавита в открытом тексте от предыдущих букв исследовался в начале 20 в. русским математиком А.А. Марковым. Марков разработал теорию последовательностей событий, получивших впоследствии название цепей Маркова, и доказал, что закон больших чисел можно распространить и на случай зависимых величин. Как сообщает Марков в работе [1], для русского языка характерно чередование гласных и согласных. Подсчет этой особенности подтверждается следующей таблицей, полученной автором при подсчете частот биграмм в тексте [9 Онегин]:

Таблица 1

Чередование гласных и согласных

	Г	С
Г	4183	38560
С	38559	18697

Из этой таблицы видно, что гласные чаще следуют за согласными (примерно в девять раз), чем за гласными. Это значит, что в биграммах текста условная вероятность гласной после согласной значительно выше, чем условная вероятность гласной после гласной:

$$p(с/г) = 0{,}6730$$

$$p(г/г) = 0{,}0979.$$

Это обстоятельство известно со времен Маркова и часто упоминается в литературе [10 Яглом]. Гораздо реже обращают внимание на обратное утверждение о том, что согласные чаще следуют за гласными (примерно в два раза), чем за согласными:

$$p(г/с) = 0{,}6730$$

$$p(с/с) = 0{,}3263.$$

Здесь p(с/г) – вероятность следования согласной после гласной, p(г/г) вероятность гласной после гласной, p(г/с) вероятность следования гласной после согласной, p(с/с) – вероятность следования согласной после согласной.

Эти закономерности были подтверждены и обобщены А.А. Марковым в работе [2], на примере текста поэмы А.С. Пушкина «Е. Онегин» [9]. При исследовании вероятностей совместного появления в тексте гласных и согласных он выявил стабильность вероятностей чередования гласных и согласных в тексте. Он также указал на то, что вероятность появления гласной после предшествующей гласной значительно ниже, чем вероятность появления гласной после предшествующей согласной.

Этот закон носят фундаментальный характер и в данной статье мы используем для него краткое название «критерий Маркова» или просто критерий. Формально его можно определить так.

$$M_г = p(с/г) - p(г/г),$$

и

$$M_с = p(г/с) - p(с/с),$$

Здесь $M_г$ и $M_с$ – соответственно критерий Маркова для гласных и согласных. В этом случае множества гласных и согласных можно определять так, как это принято в русской орфографии, но иногда нам будет удобно использовать и более широкие понятия для этих множеств букв, отличающиеся от традиционных. Для выделения таких случаев мы будем использовать термины «множества гласных» и «множества согласных» в кавычках,

если это не ясно непосредственно из контекста.

Выбор этого критерия правомочен, так для каждой отдельной гласной его величина положительна (что легко проверить по таблице биграмм [7Алферов]), и чем больше гласных будет выделено из текста, тем больше будет величина этого критерия.

В данной работе предлагается метод автоматического разделения гласных и согласных с помощью критерия Маркова. Идея метода в следующем. Пусть задана матрица вероятностей биграмм заданного текста. Считая, что ни количество, ни состав гласных в тексте нам заранее неизвестны, мы ищем их согласно критериям Маркова.

Будем последовательно находить в тексте «гласные», выбирая из состава всего алфавита те сочетания букв, которые дают максимальное значение критерия Маркова $M_г$:

$$M_г = p(с/г) - p(г/г),$$

где: p(с/г) – вероятность следования гласной за предшествующей согласной, равная количеству пар (биграмм) вида согласная-гласная, деленному на количество согласных в тексте, а p(г/г) – вероятность следования гласной за предшествующей гласной, равная количеству пар вида гласная-гласная, деленному на общее количество гласных в тексте. Значение критерия $M_г$ зависит от требуемого количества выделенных гласных.

Последовательности появления «гласных» в указанных сочетаниях букв можно сопоставить «силе» или «рангу» гласных. Чем раньше появилась данная «гласная», тем она «сильнее». Это обстоятельство можно использовать для ранжирования или, другими словами, взаимного упорядочивания гласных. Данные о стабильности этого ранжирования в литературе отсутствуют, поэтому в данной работе была поставлена задача выяснения того, как она зависит, во-первых, от объема текста, и, во-вторых, от источника этого текста. В работе была также проверена зависимость ранжирования гласных от автора текста. Предполагалось, что если указанное ранжирование будет зависеть от автора, его можно использовать для подтверждения авторства, если же от автора ранжирование практически не зависит, то его можно использовать в криптографии для разделения гласных и согласных при вскрытии шифрованных сообщений.

Чтобы автоматически вычислять эти критерии, была разработана специальная программа, которая позволяет ввести файл с заданным текстом лингвистического ансамбля, посчитать для него частоты биграмм и вычислить соответствующую им матрицу вероятностей биграмм. После этого матрица вероятностей биграмм запоминается и не изменяется в течение всего времени работы программы с данным текстом.

Перед запуском программы в основном режиме ей в качестве параметра задано число k – предполагаемое количество «гласных». Далее программа производит полный перебор всевозможных C_N^k сочетаний «гласных» букв (здесь N – количество символов алфавита), и для каждого сочетания вычисляет значение критерия Маркова. Сочетание букв алфавита, которому соответствует максимальное значение критерия, принимается в качестве множества «гласных» данной мощности k. Ввиду симметрии критериев Маркова для «гласных» и «согласных» эту же программу, как показано далее, можно применить и для поиска «согласных».

Количество подмножеств букв C_N^k даже при фиксированном значении k – количества «гласных» довольно велико и при N = 32, k = 16 превосходит 600 000 000, поэтому программа в основном варианте полного перебора работает довольно долго. С целью ускорения работы в ней был предусмотрен вспомогательный режим ускоренного поиска экстремума. Для этого фиксировалась так называемая «маска гласных» в виде некоторого подмножества букв алфавита, которые априори считались гласными. В этом случае m заранее заданных букв, объявлялись «гласными» и в переборе не участвовали, а программа в процессе перебора находила k-m «новых гласных» среди N-m оставшихся букв алфавита. Благодаря этому, работу программы можно ускорить в тысячи и даже миллионы раз. Необходимым условием правильности работы в таком режиме является отсутствие локальных экстремумов (унимодальность). Одной из задач данной работы была именно проверка на унимодальность. Многократные испытания показали, что при длине исследуемого текста порядка килобайта и более в каждом цикле выделения новых гласных достаточно работать в режиме запроса одной-двух новых гласных. Если количество выделяемых новых гласных увеличить до 5–6, нарушения унимодальности вообще не наблюдаются даже при очень коротких текстах, длиной до 100–200 символов.

В программе был предусмотрен также режим отключения перебора для численного расчета значения критерия Маркова конкретной группы «гласных» или согласных.

Все перечисленные режимы работы были предусмотрены для фактического выделения гласных и согласных, и их возможного ранжирования в количествах до 10–12 «гласных» и до 20 «согласных».

Еще один режим программы предусматривает работу с неограниченным числом выделяемых «гласных» (естественно, в пределах алфавита).

Оказывается, в таком режиме сначала происходит выделение, всех настоящих гласных в порядке убывания их «силы», а затем среди выделенных букв начнут появляться согласные, но уже в порядке возрастания их «силы» (как согласных), о чем будет подробнее рассказано далее.

Рассмотрим особенности критериев Маркова на примере исследования текста «Е. Онегин» (без дополнительных глав и примечаний). Объем текста составил 110 050 букв. При подсчете биграмм вспомогательные знаки и символы удалялись и получался текст без пробелов, для которого традиционным для криптографии способом подсчитывались частоты биграмм со сдвигом на 1 букву, причем первая буква биграммы из потока данных изымалась.

На рис. 3 показаны графики, отображающие характер поведения критериев Маркова для «Е. Онегина».

Функция критерия Маркова для «гласных», показана сплошной линией, а соответствующая для «согласных» пунктирной. На оси ординат показаны значения критериев, а на оси абсцисс отображены подмножества букв, соответствующие этим значениям. Порядки появления «гласных» и «согласных» служат координатами абсцисс, причем их взаимозависимость позволила совместить две функции этих критериев на одном графике. Для определения состава множеств «гласных» и «согласных» при помощи этого графика надо в случае «гласных» выбирать последовательно буквы от нуля слева направо, до заданного значения критерия, а «согласные» – в обратном порядке. Очевидно, что на этом графике порядок расположения букв на оси абсцисс соответствует «силе» букв.

Расчет этих функций имеет некоторые особенности. При расчете критерия Маркова для гласных среди гласных появляются последовательно буквы О, А, Е, И, У, Ы. Это не

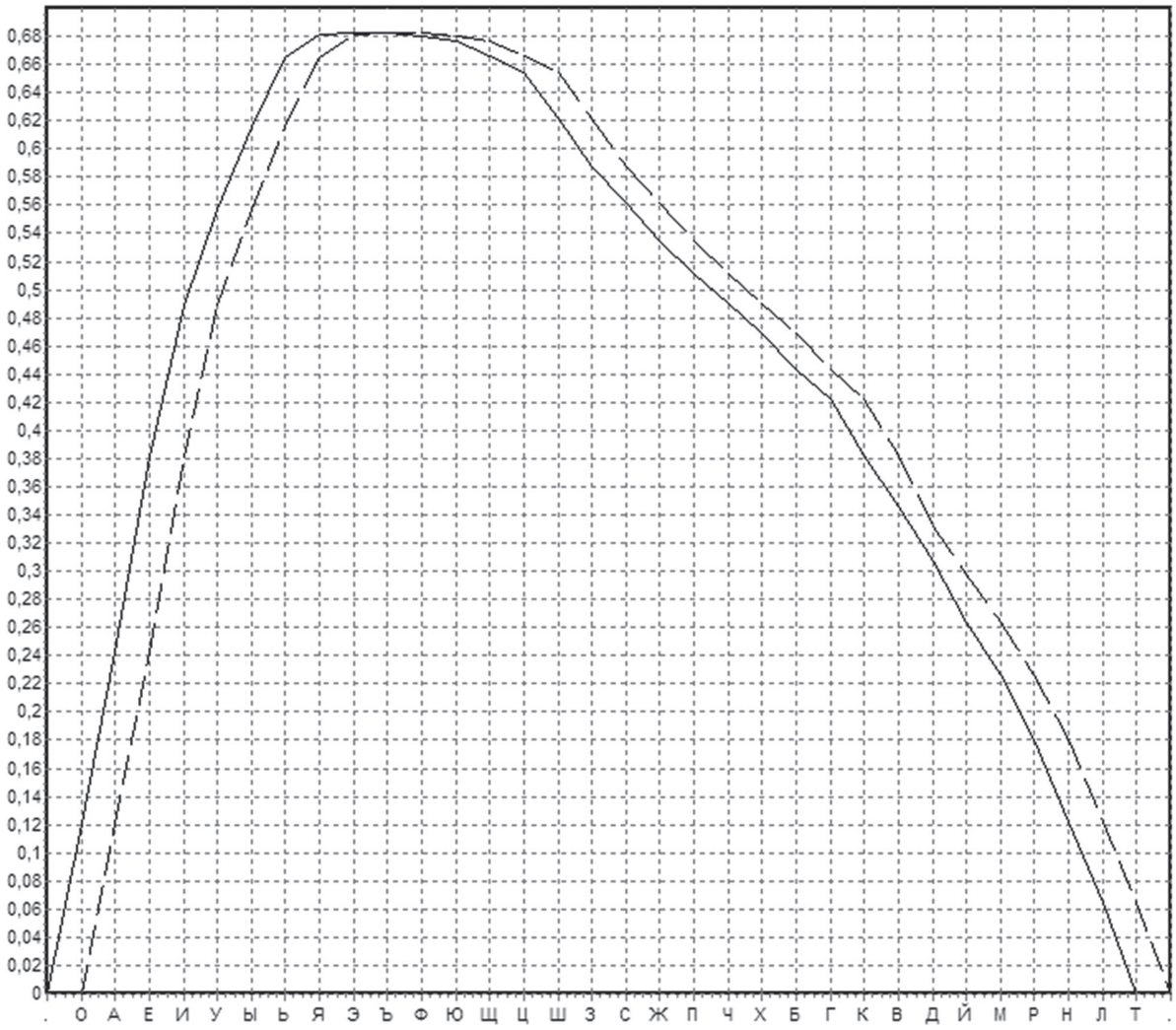

Рис. 3. Критерии

является чем-то необыкновенным и может быть объяснено, например, частотностью этих букв в тексте. Но вот появление вслед за «У» буквы «Ь», которую традиционно относят к согласным является неожиданностью. Далее появляются гласные Я и Э, а потом, снова неожиданно, «Ъ» и «Ф», а только потом уже буква «Ю».

Дальнейшее исследование полностью подтвердило тот факт, что по статистическим характеристикам буквы «Ь» и «Ъ» в русском тексте следует относить к «гласным».

Дальнейший ход расчета критерия Маркова для «гласных» проявляет еще одну неочевидную особенность. После ожидаемых букв «гласных» «ОАЕИУЫЬЯЭЪФЮЩЦШЗ» программа неожиданно выдала в качестве решения подмножество согласных «БВГДЖЗЙКЛМНПРСТХЧ». Почему? Оказывается, потому что критерий Маркова для 16 букв, больше для первого из показанных подмножеств («гласных»), чем для такого же количества «согласных», а уже для 17 букв критерий Маркова имеет большее значение, равное 0.6212 для 17 согласных, чем значение 0.5614 для 17 «гласных» букв. Так как программа заранее не знает фактически, что мы хотим, она в первом случае в результате полного перебора беспристрастно выбирает первый ансамбль букв, а во втором – второе подмножество. При дальнейшем увеличении заданного количества «гласных» программа движется по левой нижней ветви графика в направлении буквы «О».

Таким образом, в режиме полного перебора программа позволяет видеть только часть графиков, расположенных на рисунке левее буквы «З».

Правые части графика можно увидеть, если отказаться от поиска глобального экстремума. Если введем в маску гласных наиболее «сильные» гласные «О» и «А», никакое из сочетаний символов, добавленное к ним, уже не может переключить работу программы в режим «согласных», и мы сможем беспрепятственно продвигаться по нижней ветви графика (показанного сплошной линией) правее буквы «З», вдоль сплошной линии вплоть до буквы «Л» включительно. (Букву «Т», самую сильную согласную, мы программа достигнуть может, так при этом перебор отсутствует вообще.) Аналогичным образом, задавая в маске согласных буквы «Т» и «Л», мы можем пройти сначала по верхней (пунктирной) линии графика от буквы «Т» до буквы «З» и далее, вплоть до буквы «А» (хотя это и не столь интересно, так как этот участок мы уже проходили ранее).

Полученные качественные выводы подтверждаются количественными результатами расчета критериев Маркова для данного случая.

Ранее было указано что для расчета критерия Маркова следует использовать вероятности биграмм в данном тексте. На практике эти вероятности неизвестны, и вместо них используют относительные частоты биграмм, которые по закону больших чисел при достаточно длинных текстах стремятся к соответствующим вероятностям. Как определить необходимый объем текста?

В данной работе при выборе объема исследуемых текстов применялась методика, традиционная для лингвистики. Известно, что вероятность лингвистических событий можно четко определить только после выбора конкретного статистико-лингвистического ансамбля [12 Пиотровский]. При исследованиях письменной речи лингвисты обычно отождествляют его с некоторым фиксированным множеством книг или других письменных источников. Можно, например, говорить о множестве изданных в определенном году газет на русском языке («ансамбль русской газетной речи N-го года»), о конкретном издании собрания сочинений А.С. Пушкина («ансамбль языка Пушкина»). В нашей работе мы в качестве источника статистико-лингвистических ансамблей рассматривали так называемый «Машинный фонд русского языка» [13]. Фонд содержит произведения примерно пятидесяти русских прозаиков различных жанров и стилей. Эпизодически для сравнения выбирались и другие образцы русского текста.

При подсчете вероятностных характеристик текста учитывалось то, что если ансамбль фиксирован, то вероятность появления некоторого элемента текста (букв, слов или более крупных единиц текста) среди других элементов данного класса, можно найти, если пересчитать общее количество заданного элемента и число элементов всего класса, а затем разделить первое на второе. Это, по определению, считают вероятностью появления заданного элемента текста в заданном статистическом ансамбле.

Если закон распределения известен, то методами математической статистики можно оценить, каким должен быть объем выборки для того, чтобы отношение m/n с заданной точностью соответствовало истинной вероятности.

Если же статистические закономерности заранее неизвестны, прибегают к последовательным испытаниям все больших объемов выборки, полагая ее достаточно большой, если увеличение объема выборки в несколько раз не приводит к существенному изменению исследуемых вероятностей. В данной работе использован именно этот прием.

Полученные результаты

Прежде всего были исследованы ансамбли на базе произведений разных писателей

при сравнительно больших размерах тестируемых текстов (десятки и сотни килобайт). Оказалось, что отделение гласных от согласных практически не зависит от личности автора текста. Таким образом, для лингвистических испытаний на авторство текста этот метод оказался негодным. Гораздо отчетливее выгоднее выглядят преимущества его применения в криптоанализе. Приведем некоторые из полученных результатов.

В таблице 3 показаны результаты, полученные при разделения гласных и согласных указанным методом для семи выбранных статистико-лингвистических ансамблей в виде художественных произведений писателей 19–20 вв. из «Машинного фонда русского языка» [13]. Для каждого ансамбля приведен литературный источник, последовательность появления «гласных» при полном переборе и соответствующие значения критерия Маркова. Вывод букв производился группами с заданным числом «гласных». Размер ансамблей не ограничивался и статистика биграмм подсчитывалась для всего текста, вплоть до 1,5 млн. символов.

Таблица 3

Разделение гласных и согласных

1. Пушкин. «Капитанская дочка». [14]						
Буквы	О	А	Е	И	У	Ы
Критерий	0,1031	0,2187	0,3217	0,4176	0,4798	0,5179
Буквы	Ь	Я	Э	Ю	Ф	Ц
Критерий	0,5578	0,5823	0,5835	0,5833	0,5823	0,5743
2. Аксаков. «Детские годы ...». [15]						
Буквы	О	А	Е	И	У	Ь
Критерий	0,1044	0,2178	0,3242	0,4160	0,4655	0,5095
Буквы	Ы	Я	Э	Ю	Ф	Щ
Критерий	0,5547	0,5772	0,5783	0,5786	0,5786	0,5689
3. Куприн. «Поединок». [16]						
Буквы	О	А	Е	И	У	Ы
Критерий	0,1129	0,2190	0,3180	0,4089	0,4669	0,5095
Буквы	Ь	Я	Э	Ю	Ф	Ц
Критерий	0,5482	0,5708	0,5730	0,5749	0,5690	0,5592
4. Тургенев. «Дворянское гнездо». [17]						
Буквы	О	А	Е	И	У	Ы
Критерий	0,1031	0,2185	0,3220	0,4192	0,4698	0,5087
Буквы	Ь	Я	Ю	Э	Ф	Щ
Критерий	0,5495	0,5710	0,5727	0,57369	0,5684	0,5614
5. Лермонтов. «Герой нашего времени». [18]						
Буквы	О	А	Е	И	У	Ь
Критерий	0,1033	0,2123	0,3100	0,4010	0,4572	0,4986
Буквы	Ы	Я	Э	Ю	Ф	Ц
Критерий	0,5428	0,5680	0,5702	0,5708	0,5709	0,5648
6. Чехов. «Цветы запоздалые». [19]						
Буквы	О	А	Е	И	У	Ь
Критерий	0,1110	0,2228	0,3172	0,4131	0,4765	0,5174
Буквы	Ы	Я	Ю	Э	Ъ	Ф
Критерий	0,5571	0,5934	0,5942	0,5948	0,5949	0,5901
7. Василий Шукшин. Живет такой парень. [20]						
Буквы	О	А	Е	И	У	Ь
Критерий	0,1039	0,2225	0,3304	0,4246	0,4856	0,5321
Буквы	Ы	Я	Ю	Э	Ъ	Ф
Критерий	0,5716	0,6004	0,6043	0,6062	0,6062	0,6009

Целью этой части работы было определение возможной зависимости ранжирования символов от автора текста. При тестировании первых пяти произведений символы Ь и Ъ были объединены в один символ, как это часто принимают в русской криптографии (кстати, эти символы Марков исключил их из рассмотрения в своей работе, не относя их ни к гласным, ни к согласным). Можно заметить, что очередность появления первых 10 «гласных», т.е. их ранжирование, практически не зависит от автора текста. Несколько неожиданным выглядит появление среди настоящих гласных символов «Ь» и «Ъ», что и объясняет использование термина «гласные» в данной работе в кавычках. Как видно из таблицы, и как показало более широкое исследование, очередность появления гласных, т.е. ранжирование, практически не зависит от автора.

Это значит, что указанное ранжирование можно считать свойством языка и его морфологии.

«Силу» гласных можно истолковать и количественно, как приращение критерия Маркова при включении символа во множество «гласных». Аналогичное заключение можно сделать и о «силе» согласных.

Характер поведения критерия Маркова свидетельствует о том, что символам ОАЕИУЫЬЯЮЭ можно приписать положительное «свойство гласных», так как они последовательно увеличивают значение критерия, а следующие за ними символы Ф, Ц, Щ его уменьшают и по этому признаку их можно отнести к согласным. При этом самая «сильная» гласная буква О, а последующие буквы в этом ряду постепенно «слабеют», и это объясняет очередность их появления при работе алгоритма. Некоторые вариации

в ранжировании пар символов «ЬЫ» и «ЭЮ» можно объяснить тем, что их «силы» весьма близки, и поэтому небольшие изменения статистических характеристик текста влияют на их ранжирование.

Отметим еще одну неочевидную особенность критерия Маркова для гласных. При увеличении заявленного порядка множества гласных k «старые» k-1 гласные входят в него как подмножество. Это значит, что никакая пара из символов, ранее относившихся к «согласным» не может увеличить критерий больше, чем новая, k-я «гласная» в паре с одной из последних k-1 «гласных». Если бы это было не так, то на k-м шаге произошло бы вытеснение одной из предыдущих «гласных» двумя новыми. Это свойство можно назвать устойчивостью ранжирования. Эта устойчивость возможна, как легко видеть, при убывании «силы» гласных по закону, близком к экспоненциальному. Ранжирование сохраняется для гласных в широком диапазоне размеров текста, но при уменьшении его объема устойчивость упорядочивание сохраняется у все меньшего количества самых «сильных» гласных.

О ранжировании согласных, нельзя говорить столь уверенно. Дело в том, что суммарная «сила» согласных равна «силе» гласных, а согласных в два раза больше. Как следствие, сила отдельных согласных гораздо меньше силы отдельных гласных, и их ранжирование подвержено гораздо большим вариациям при изменении статистики текста, особенно при уменьшении размера текста, что подтвердилось при более подробных исследованиях.

Таблица 4

Размер ансамбля	Последовательность появления символов «гласных»
Весь текст (более 110 000 символов.)	ОАЕИУ ЫЬЯЮЭ ФЩЦ
20 000 букв.	ОАЕИУ ЫЬЯЮЭ ФЩЦ
2 000 букв.	ОАЕИУ ЫЬЯЭЮ ЩФЦ
500 букв.	ОАЕИУ ЬЯЫЮЭ –
250 букв.	ОАИЕУ ЬЯЫЮЭ –

Остается еще вопрос о том, как свойства ранжирования сохраняются при уменьшении размеров выборки.

Оценочные испытания более коротких текстов показали значительную устойчивость ранжирования при уменьшении размера ансамбля (из того же текста) вплоть до 1000 символов и ниже. Далее, в таблице 2, приводятся представительные результаты проверки стабильности ранжирования на примере текста поэмы А.С. Пушкина «Евгений Онегин». Исследовались все более уменьшающиеся в размере начальные части поэмы, начиная со слов «Мой дядя...».

Как видно из табл. 4, разделение гласных и согласных сохраняется вплоть до текстов длиной до 250 символов. Ранжирование гласных в данном случае точно вплоть до текстов размером порядка 2000 букв, далее возникают вариации. Это объясняется тем, что при уменьшении размеров текста нарушается закон больших чисел, в частности в последних двух ансамблях букв Ф, Ц и Щ в тексте совсем нет, так что о статистике здесь говорить не приходится.

Выводы

Метод критерия Маркова позволяет автоматизировать разделение гласных и согласных в русском тексте.

Гласные буквы русских текстов длиной порядка 1000 символов и более методом критерия Маркова можно ранжировать.

Ранжирование согласных нестабильно и возможно лишь для больших объемов лингвистических ансамблей.

Буквы «Ъ» и «Ь» в биграммах русского текста по статистическим свойствам сходны с гласными.

Приращение критерия Маркова для гласных, полученное данным методом, убывает приблизительно по экспоненте для «гласных» ряда «ОАЕИУЫЬЯЮЭ».

Для сравнения приводим также результат разделения и ранжирования гласных очень короткого (85 букв) и трудного для дешифровки текста [9]:

«Мой дядя самых честных правил, когда не в шутку занемог он уважать себя заставил и лучше выдумать не мог».

Здесь программа выдает такую последовательность вывода «гласных»: «аеоуыяиь». Видно, что, хотя ранжирование здесь нарушено, функция разделения гласных и согласных даже здесь работает почти безошибочно.

Литература

1. Марков А.А. Исследование замечательного случая зависимых испытаний // Известия Императорской академии наук. Серия 6. 1907. Т. 1. Вып. 3 С. 61-80.

2. Марков А.А. Пример статистическаго изследованія над текстом «Евгенія Онегина», иллюстрирующій связь испытаній в цепь» //

References

1. Markov A.A. Issledovanie zamechatel'nogo sluchaya zavisimykh ispytaniy. Izvestiya Imperatorskoy akademii nauk. Series 6. 1907. Vol. 1. Iss. 3. P. 61-80. (In Russ.)

2. Markov A.A. Primer statisticheskago izsledovaniya nad tekstom "Evgeniya Onegina", illyustriruyushchiy svyaz' ispytaniy v tsep'". Izvestiya

Известия Императорской Академии Наук. VI серія. 1913. Т. 7. Вып. 3. С. 153–162.

3. Шеннон К. Теория связи в секретных системах. В кн.: Работы по теории информации и кибернетике. М.: ИЛ, 1963.

4. Friedman W. F.,Callimahos D. Military cryptanalysis. Aegean Park Press, Laguna Hills CA, 1985. Part I. Vol. 2.

5. Moler C., Morrison D. Singular Value Analysis of Cryptograms. The American Mathematical Monthly. 1983. Vol. 90. № 2. P. 78–87.

6. Kahn D. The codebreakers. The story of secret writing. Macmillan, N.Y., 1967.

7. Алферов А. П., Зубов А. Ю., Кузьмин А. С., Черемушкин А. В. Основы криптографии. М.: Гелиос АРВ, 2002.

8. Sikora T.F. Field Manual 34-40-2, Basic Cryptanalysts. USA, Washington, DC, 13 September 1990.

9. Пушкин А.С. Евгений Онегин, М.: Эксмо, 2017.

10. Яглом А.М. Яглом И.М. Вероятность и информация. М.: Наука, 1973.

11. Конан Дойль А., Приключения Шерлока Холмса. Пляшущие человечки. М.: АСТ, 2016.

12. Пиотровский Р.Г., К.Б. Бектаев, А.А. Пиотровская. Математическая лингвистика. М.: Высшая школа. 1977. С. 383.

13. Машинный фонд русского языка. URL: http//cfrl.ruslang.ru.

14. Пушкин А.С. Капитанская дочка. М.: Эксмо, 2015. С. 1–640.

15. Аксаков С.Т. Детские годы Багрова-внука. М.: АСТ, 2017. С. 1–416.

16. Куприн А.И. Поединок. М.: Эксмо, 2013. С. 1–640.

17. Тургенев И.С. Дворянское гнездо. М.: АСТ, 2017. С. 384

18. Лермонтов М.Ю. Герой нашего времени. М.: АСТ, 2015. С. 1-285.

19. Чехов А.П. Цветы запоздалые. А. П. Чехов. Полное собрание сочинений и писем в 30-ти томах. Сочинения. М.: Наука, 1983. Том 1.

20. Шукшин В.М. Киноповести. М., 1988. С. 115.

Imperatorskoy Akademii Nauk. VI series. 1913. Vol. 7. Iss. 3. P. 153–162. (In Russ.)

3. Shennon K. Teoriya svyazi v sekretnykh sistemakh. In.: Raboty po teorii informatsii i kibernetike. Moscow: IL, 1963. (In Russ.)

4. Friedman W. F.,Callimahos D. Military cryptanalysis. Aegean Park Press, Laguna Hills CA, 1985. Part I. Vol. 2.

5. Moler C., Morrison D. Singular Value Analysis of Cryptograms. The American Mathematical Monthly. 1983. Vol. 90. No. 2. P. 78–87.

6. Kahn D. The codebreakers. The story of secret writing. Macmillan, N.Y., 1967.

7. Alferov A. P., Zubov A. Yu., Kuz'min A. S., Cheremushkin A. V. Osnovy kriptografii. Moscow: Gelios ARV, 2002. (In Russ.)

8. Sikora T.F. Field Manual 34-40-2, Basic Cryptanalysts. USA, Washington, DC, 13 September 1990.

9. Pushkin A.S. Evgeniy Onegin, Moscow: Eksmo, 2017. (In Russ.)

10. Yaglom A.M. Yaglom I.M. Veroyatnost' i informatsiya. Moscow: Nauka, 1973. (In Russ.)

11. Konan Doyl' A., Priklyucheniya Sherloka Kholmsa. Plyashushchie chelovechki. Moscow: AST, 2016. (In Russ.)

12. Piotrovskiy R.G., K.B. Bektaev, A.A. Piotrovskaya. Matematicheskaya lingvistika. Moscow: Vysshaya shkola. 1977. P. 383. (In Russ.)

13. Mashinnyy fond russkogo yazyka. URL: http//cfrl.ruslang.ru. (In Russ.)

14. Pushkin A.S. Kapitanskaya dochka. Moscow: Eksmo, 2015. P. 1-640. (In Russ.)

15. Aksakov S.T. Detskie gody Bagrova-vnuka. Moscow: AST, 2017. P. 1-416. (In Russ.)

16. Kuprin A.I. Poedinok. Moscow: Eksmo, 2013. P. 1-640. (In Russ.)

17. Turgenev I.S. Dvoryanskoe gnezdo. Moscow: AST, 2017. P. 384 (In Russ.)

18. Lermontov M.Yu. Geroy nashego vremeni. Moscow: AST, 2015. P. 1-285. (In Russ.)

19. Chekhov A.P. Tsvety zapozdalye. A. P. Chekhov. Polnoe sobranie sochineniy i pisem v 30-ti tomakh. Sochineniya. Moscow: Nauka, 1983. Vol. 1. (In Russ.)

20. Shukshin V.M. Kinopovesti. M., 1988. P. 115. (In Russ.)

Сведения об авторе

Юрий Ильич Петренко
К.т.н., с.н.с., доцент
Московский авиационный институт
(национальный исследовательский университет),
Москва, Россия
Эл. почта: tautoga@yandex.ru
Тел. 8-926-237-3501

Information about the author

Yuriy I. Petrenko
Cand. Sci. (Engineering), Senior researcher,
Associate Professor
Moscow Aviation Institute, Moscow, Russia
E-mail: tautoga@yandex.ru
Tel.: 8-926-237-3501