УДК 004.056.5 (045) **Быстрова Б.В., Сарсадских О.В. (НАУ)**

СРАВНЕНИЕ ЧАСТОТ ВСТРЕЧАЕМОСТИ БУКВ В АНГЛИЙСКИХ ТЕКСТАХ И ЛИПОГРАММАХ

В данной работе рассматриваются вопросы, связанные с определением частот встречаемости букв в липограмматичсеких работах и их сравнение обычными английскими текстами. Исследования показали, что свойства липограмм существенно влияют на общепринятые подходы частотного криптоанализа, основанного на среднестатистических данных.

Ключевые слова: липограмма, криптоанализ, частотный анализ, частота встречаемости букв, Гедсби.

Частотный анализ является одним из основных инструментов криптоаналитиков. Техника, известная как частотный анализ может эффективно использоваться для вскрытия моноалфавитной подстановки, так как значение каждой буквы остается неизменным как в шифрограмме, так и в открытом тексте. Анализ относительно небольшого текстового произведения показывает, что эмпирические частоты несколько отклоняются от средних показателей. Но также известно, например, что частота встречаемости английской буквы "e" (составляющей приблизительно 13 % от всех букв [6]) в конкретных текстах несколько отличается от известных среднестатистических значений [8] и тогда, соответственно отличается частота других букв, которую используют для распознавания шифралфавита. Современные компьютеры позволяют легко дешифровать такие шифротексты, но даже во времена рукописного текста, анализ частоты мог использоваться, чтобы взламывать шифры простой замены.

Известно, что буква "e" наиболее часто встречается в текстах на английском языке, следующей по частоте использования является буква "t", а наименее используемая – "z" [6]. Эти показатели являются результатом подсчета миллионов букв из тысяч текстов и абсолютно надежны, но не обязательно относятся к каждому отдельно взятому тексту. Третьей наиболее часто употребляемой буквой в массе английских текстов является "a", но, например, в "Путешествиях Гулливера" частота её употребления составляет всего лишь 0,8% [5], а в романе Жоржа Перека "A Void" [4] частота использования "e" равна абсолютному нулю, поскольку книга была написана без использования этой буквы. Это та ситуация, когда частота является результатом поведения человека и сложно утверждать вероятность чего-либо происходящего, если человек может выбирать способ поведения отличный от общепринятых в прошлом норм. У Перека был выбор, и он сделал его в пользу неиспользования буквы "e". Но еще чаще частота, наблюдаемая в прошлом, может обоснованно быть использована для прогнозирования ожиданий о будущем. Это сравнимо с подбрасыванием монетки или кубика, падение которых на одну из сторон происходит приблизительно с одинаковой частотой.

Существует возможность изменить известные частотные свойства шифрованных текстов, что существенно усложнит соответствующий анализ. Этого можно достичь при помощи исключения использования определенной буквы (или нескольких) алфавита. Усложнение реализации подобных криптоаналитических атак на шифротексты возможно при использовании липограмм. Но на сегодня неизвестно, например, как отсутствие одной буквы в тексте влияет на частоту других, что и будет основным аргументом использования липограмм при подготовке к шифрованию исходных текстов. В этой связи, целью данной работы является определение частот встречаемости букв в липограммах и сравнение их с известными частотами.

Можем предположить, что создание липограммы является достаточно интересным заданием для автора. Липограмматическое произведение отличается своими частотными свойствами от традиционных текстов, но это не использовалось в шифровальных целях.

Напомним кратко о понятии липограммы и его истории. Липограмма – содержательный текст, состоящий из слов, в которых не встречается одна или несколько букв используемого алфавита [9]. Название происходит от греческих слов λειπω – пренебрегаю, отказываюсь и γράμ – буква. Липограмма – одно из самых незаметных для читателя формальных ограничений, особенно если речь идет о небольшом тексте. Термин был введен в начале XVII в. французским поэтом Соломоном Сертоном, активно использовавшем липограмму в своем творчестве. Первая известная нам липограмма была написана в VI веке до н.э., когда Лазос Гермионский (возможно, входивший в состав Семи мудрецов Древней Греции) сочинил без буквы сигма "Оду о кентаврах", от которой не сохранилось ничего, и "Гимн Деметре", от которой сохранилась первая строчка: Dhmhara melpw Koran te Klnmenoi alocon. В III веке н.э. римский поэт Нестор Ларандский переписал "Илиаду", избавившись в первой песне от буквы "альфа", во второй – от "бета", в третьей от "гамма", и так до конца алфавита и поэмы. Два века спустя Трифиодор Сицилийский (Tryphiodorus), грек из Египта, завершил тенденцию, применив ту же процедуру к Одиссее. Однако от этих модифиций поэм также ничего не сохранилось. Несмотря на это, Нестор и Трифиодор по сей день являются наиболее известными липограммистами. Первая удостоверенная липограмма составлена латинским грамматиком Фабиусом Планиадесом Фулгентиусом (Fabius Planiades Fulgentius, VI век н.э.), который написал «De Aetatibus Mundi & Hominus», работу из 23 глав, в которой первая глава не содержит буквы "a", вторая – "b", и т.д. Позже, в середине XI века, Пьер де Рига выполнил стихотворный перевод Библии, снабдив каждую песнь небольшим резюме, в которых также отсутствовали буквы по алфавиту: первое резюме было написано без "a", второе – без "b", и т.д. Работа имела значительный успех, и сегодня существует 250 манускриптов с этим текстом. Другая липограммическая традиция писать без буквы "r" (с XVII века до наших дней), была особенно распространена в Германии и Италии. С XVIII века липограмма переходит в поэзию (Брокс, Готтлиб Вильгельм Бурманн, Кемпнер (1803)), а также в прозу (Франц Риттлер (1813), Леопольд Колбе (1816), Кристиан Вейс (1878), Пауль фон Шонтан (1883)). В Италии в рамках той же традиции были написаны: сказка Риккобони, речи Луиджи Казолини и поэма в 1700 строк Орацио Фиделе (заменяющего Амура на Купидона и Венеру на Синтию). К этому далеко не полному списку можно добавить примеры из русской поэзии: стихотворения Г.Р. Державина "Соловей во сне", "Бабочка", "Свобода", "Тишина" сознательно написаны без буквы "р", а небольшое стихотворение Д. Бурлюка так и озаглавлено: Без "р" и "с" [10]. Одна из самых крупных липограмм на английском языке – опубликованный в 1939 году роман американского моряка Эрнеста Винсента Райта "Гэдсби": история в 50 000 слов без буквы "e" (Ernest Vincent Wright, "Gadsby") [1]. По словам автора, роман "был написан, когда из пишущей машинки вынули букву "e", чтобы ее нельзя было случайно напечатать. Многие пытались найти эту букву в тексте, но безуспешно!". В английском языке это значит отказ от "he", "she", "her", "they", "them", а также часто используемом определенном артикле "the"; глаголов в прошедшем времени, оканчивающихся на " ed": "replied", "answered", "asked"; а также чисел между шестью и тридцатью, не говоря уже о часто встречаемом "_of_ the_". Райт не использовал их даже в численной форме, равно как и сокращений Mr и Mrs, в которых "e" слышна при прочтении. "Моя книга может пригодиться английским школьникам, – писал Райт, – обучающимся основам сочинений" [1].

Взяв одновременно две буквы, можно рассмотреть частотную биграмму (взятые одновременно три буквы называются триграммой). Сравнивая частотную биграмму (или триграмму) потенциально рассматриваемого текста или произведения с обычной частотностью английского языка, легко распознать английский язык благодаря сильной корреляции.

Известно, что наиболее распространенная буква в английском языке – это буква "e". И, несмотря на этот факт, в истории литературы известны авторы, написавшие свои работы (названные липограммами) и даже целые романы без ее употребления.

Для исследования была взята одна из наиболее известных липограмм – произведение Э. Райта "Гэдсби". Здесь вычислялись количества использованных букв для определения и сравнения их частоты с обычными английскими текстами. В табл. 1 представлена частота и количество использования букв в английских текстах и в произведение "Гэдсби". Следует отметить, что данные о частоте букв в английских текстах были исследованы учеными Корне́лльского университета [8], а данные о количестве и частоте букв в романе "Гэдсби" получены путем обработки текста произведения, состоящие из 50,000 знаков не включая пробел, с помощью специально разработанной для этой цели программы [7]. В табл. 2 представлены те же данные, но в порядке убывания для наглядной демонстрации влияния отсутствия одной буквы на изменение частоты использования остальных букв алфавита.

Данные, приведенные в табл. 1 и в табл. 2 можно использовать для построения гистограмм.

Таблица 1

Частота (в процентном отношении) и количество используемых букв.

Частота букв в английских текстах		Количество и частота букв в произведении "Гэдсби"	
"a"	**8,167%**	**23218**	**10,960%**
"b"	**1,492%**	**4541**	**2,144%**
"c"	**2,782%**	**5645**	**2,664%**
"d"	**4,253%**	**8728**	**4,120%**
"e"	**12,702%**	**0**	**0**
"f"	**2,228%**	**4562**	**2,154%**
"g"	**2,015%**	**7644**	**3,609%**
"h"	**6,094%**	**10398**	**4,909%**
"i"	**6,966%**	**18663**	**8,810%**
"j"	**0,153%**	**487**	**0,230%**
"k"	**0,772%**	**2498**	**1,179%**
"l"	**4,025%**	**11261**	**5,316%**
"m"	**2,406%**	**4399**	**2,077%**
"n"	**6,749%**	**18227**	**8,604%**
"o"	**7,507%**	**22081**	**10,424%**
"p"	**1,929%**	**4038**	**1,906%**
"q"	**0,095%**	**109**	**0,052%**
"r"	**5,987%**	**10094**	**4,765%**
"s"	**6,327%**	**14771**	**6,973%**
"t"	**9,056%**	**18002**	**8,498%**
"u"	**2,758%**	**8813**	**4,160%**
"v"	**0,978%**	**665**	**0,314%**
"w"	**2,360%**	**5930**	**2,799%**
"x"	**0,150%**	**100**	**0,047%**
"y"	**1,974%**	**6731**	**3,178%**
"z"	**0,074%**	**228**	**0,108%**

Таблица 2

Частота (в процентном отношении) использования букв в порядке убывания.

Частота букв в английских текстах		Частота букв в произведении "Гэдсби"	
"e"	**12,702%**	**"a"**	**10,960%**
"t"	**9,056%**	**"o"**	**10,424%**
"a"	**8,167%**	**"i"**	**8,810%**
"o"	**7,507%**	**"n"**	**8,604%**
"i"	**6,966%**	**"t"**	**8,498%**
"n"	**6,749%**	**"s"**	**6,973%**
"s"	**6,327%**	**"l"**	**5,316%**
"h"	**6,094%**	**"h"**	**4,909%**
"r"	**5,987%**	**"r"**	**4,765%**
"d"	**4,253%**	**"u"**	**4,160%**
"l"	**4,025%**	**"d"**	**4,120%**
"c"	**2,782%**	**"g"**	**3,609%**
"u"	**2,758%**	**"y"**	**3,178%**
"m"	**2,406%**	**"w"**	**2,799%**
"w"	**2,360%**	**"c"**	**2,664%**
"f"	**2,228%**	**"f"**	**2,154%**
"g"	**2,015%**	**"b"**	**2,144%**
"y"	**1,974%**	**"m"**	**2,077%**
"p"	**1,929%**	**"p"**	**1,906%**
"b"	**1,492%**	**"k"**	**1,179%**
"v"	**0,978%**	**"v"**	**0,314%**
"k"	**0,772%**	**"j"**	**0,230%**
"j"	**0,153%**	**"z"**	**0,108%**
"x"	**0,150%**	**"q"**	**0,052%**
"q"	**0,095%**	**"x"**	**0,047%**
"z"	**0,074%**	**"e"**	**0**

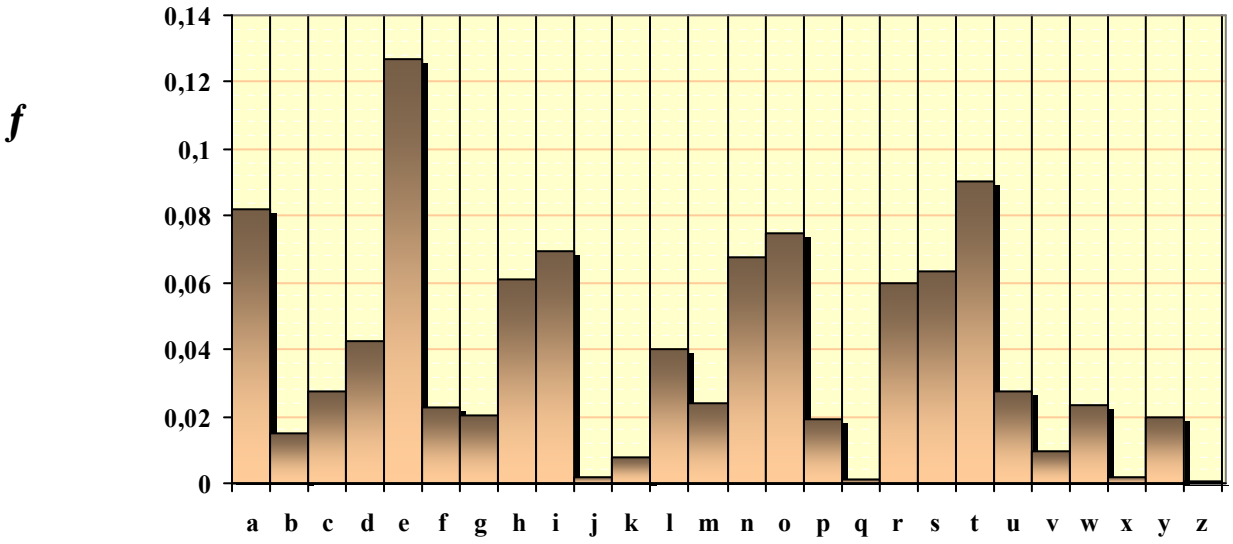

Рис. 1 Гистограмма частот встречаемости букв английского языка

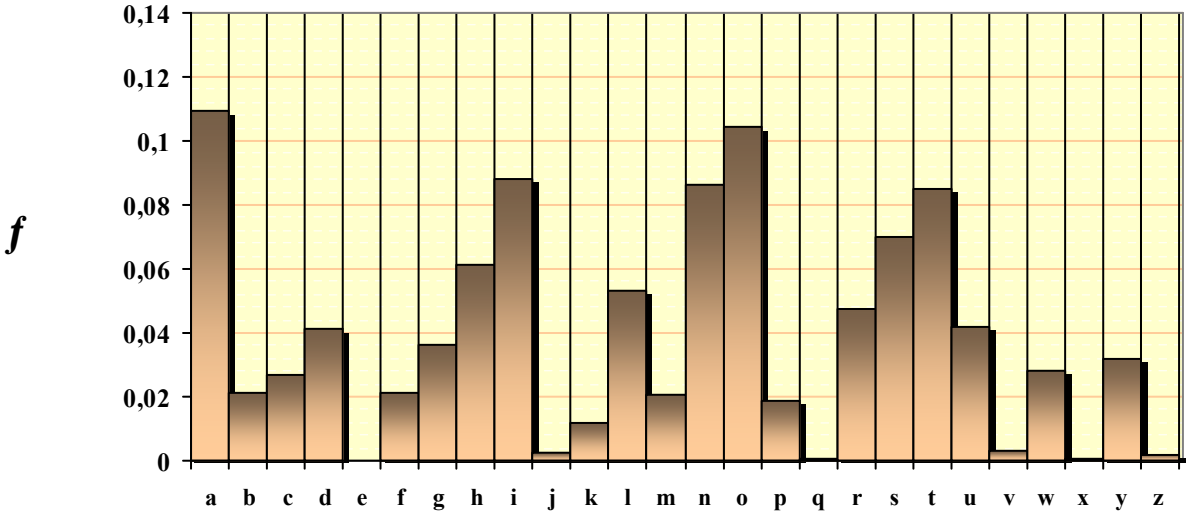

Рис. 2 Гистограмма частот встречаемости букв липограмматического произведения "Гэдсби"

Приведенные выше гистограммы дают наглядное представление о частоте встречаемости букв английского алфавита [8] и липограмматического произведения "Гэдсби". Здесь наглядно видно, что отсутствие буквы "e" повлияло на распределение частот других символов. Как видно отношение этих распределений носит нелинейный характер, а использование максимальных и минимальных показателей позволяет условно разделить буквы на категории высокой ("В"), средней ("С"), низкой ("Н") и очень низкой ("ОН") частоты встречаемости ("f"). Ее можно определить по формуле:

$$f = \begin{cases} "B" при f \geq 0{,}08 \\ "C" при\ 0{,}08 > f \geq 0{,}04 \\ "H" при\ 0{,}04 > f \geq 0{,}02 \\ "OH" при\ 0{,}02 > f. \end{cases} \quad (1)$$

Исходя из формулы (1) и рис. 1 установим:
e, t, a, – "B"; o, i, n, s, h, r, d, l, – "C"; c, u, m, w, f, g – "H"; y, p, b, v, k, j, x, q, z – "OH".
Аналогичным образом на основе рис. 2 можно распределить буквы в произведении «Гэдсби»: a, o, i, n, t – "B"; s, l, h, r, u, d, – "C"; g, y, w, c, f, b, m – "H"; p, k, v, j, z, q, x, e – "OH".

Как видно осуществлено не только значительное перераспределение частот, но и даже количество букв относящихся к определенной категории. Так для усредненных частот к категориям "B", "C", "H" и "OH" относятся соответственно 3, 8, 6 и 9, а для "Гэдсби" – 5, 6, 7 и 8.

Отсюда, можно предположить, что частотный анализ шифрограмм, основанный на знании частот встречаемости той или иной буквы, может быть неприемлем при работе с липограммами.

Отметим, что отчасти существует определенное взаимоотношение между традиционной криптографией и романом Жоржа Перека "A Void". Как было уже упомянуто, роман является примером липограммы на букву "e" – использование этой буквы сознательно избегается автором на протяжении всего романа. В то время как Жорж Перек трудился над созданием своего произведения, главным образом для достижения признания в определенных творческих кругах, связь с криптографией оказалась более чем очевидной. Отбрасывая букву "e", Перек возился с принципами теории вероятности и относительности, которые лежат в основе теории информации. Никто не предполагал, что отсутствие лишь одной буквы "e" на протяжении всего 300-страничного романа существенно влияет на общепринятые подходы частотного анализа, основанные на среднестатистических данных. Исходя из этого если бы все липограммы были зашифрованы с помощью моноалфавитной подстановки, то попытка использовать частотный анализ для их дешифрования была бы загнана в тупик из-за существенного перераспределения частот встречаемости букв.

Попытка использования липограмм в литературных текстах носит не единичный разрозненный характер. В 60-х годах XX века было создано целое объединение писателей и математиков (УЛИПО (фр. *OULIPO*, сокращение от ***Ou****vroir de* ***li****ttérature* ***po****tentielle*) – Цех потенциальной литературы), поставившее своей целью научное исследование потенциальных возможностей языка путём изучения известных и создания новых искусственных литературных ограничений, под которыми понимаются любые формальные требования к художественному тексту (например, определённый стихотворный размер или отказ от использования некоторых букв). Объединение, основано в Париже в 1960 году математиком Франсуа Ле Лионне и писателем Раймоном Кено включало наиболее известных участников-литераторов Жоржа Перека, Итало Кальвино, Жака Рубо и художника Марселя Дюшан. Приверженцы движения УЛИПО обращали особое внимание на отдельные символы – буквы, слова, предложения – и на способы их соединения формальным и математическим путем. Таким образом, применяя подходы, изложенные сторонниками УЛИПО, можно проанализировать язык с точки зрения кода и сконструировать своеобразный мост между этими двумя системами [7].

Проведенные исследования и анализ литературных произведений, написанных в виде липограмм, ставят под сомнение эффективность использования частотного анализа для реализации криптоаналитических атак на липограмматические шифротексты. Для получения полной картины о частотных свойствах липограмм необходимо дополнительно провести исследования биграмм и других часто используемых соединений букв.

Литература

1. Ernest Vincent Wright. Gadsby. Wetzel Publishing Co., 1939 – 260 pp.
2. Fletcher Pratt, Secret and Urgent: the Story of Codes and Ciphers. Blue Ribbon Books, 1939 – pp. 254-255.

3. Friedman W. F., Callimahos D., Military cryptanalysis, Part I, Vol 2, Aegean Park Press, Laguna Hills CA, 1920 – 342 p.

4. Georges Perec. A Void. The Harvill Press., 1995 – 290 pp.

5. Jonathan Swift. Travels into Several Remote Nations of the World, in Four Parts. By Lemuel Gulliver, First a Surgeon, and then a Captain of several Ships. London, Printed for C. Bathurst, 1765. – 320 pp.

6. Lewand, Robert. Cryptological Mathematics. The Mathematical Association of America, 2000 – p.36.

7. Rainbow Arch - Scripts and Web Tools (electronic source)/Word Counter and Frequency Data. Access: http://rainbow.arch.scriptmania.com/tools/word_counter.html, free.

8. Wikipedia, the free encyclopedia (electronic source)/ Letter frequency. Access: http://en.wikipedia.org/wiki/Letter_frequency, free.

9. Бабак В.П., Корченко О.Г. Інформаційна безпека та сучасні мережеві технології: Англо – українсько – російський словник термінів. – К.: НАУ, 2003. – 670 с.

10. Бонч-Осмоловская Татьяна. Литературные эксперименты группы «УЛИПО» // «НЛО» –2002. – №57. –С. 57-61

Надійшла: 05.03.11

Рецензент: д.т.н., проф. Шелест М. Є.