УДК 519.24
DOI: 10.15827/0236-235X.123.448-454
Дата подачи статьи: 23.04.18
2018. Т. 31. № 3. С. 448–454

Моделирование восприятия мозгом анаграммно искаженного текста

З.Д. Усманов [1], *д.ф.-м.н., профессор кафедры информатики и информационных систем, zafar-usmanov@rambler.ru*

[1] *Российско-Таджикский (Славянский) университет, г. Душанбе, 734025, Таджикистан*

Объектом исследования являются тексты естественных языков, слова которых обессмыслены случайными перестановками букв. Рассматривается способность человеческого мозга безошибочно распознавать смысл непривычной продукции.

В статье предлагается математическая модель объяснения того, каким образом мозг справляется с решением этой задачи в случаях, когда а) первая, б) последняя, в) первая и последняя буквы слова остаются на своих местах, а все прочие переставляются произвольным образом, и, наконец, в самом общем случае г), когда ни одна буква слова не фиксируется и все они в пределах слова могут располагаться в любом порядке. Объяснение основывается на понятии в широком смысле анаграммы слова как совокупности его букв, расставленных в какой-либо последовательности, а также на понятии прообраза анаграммы, в роли которой выступает само слово.

В упрощенной математической модели предполагается, что мозг воспринимает каждую анаграмму изолированно; распознает ее правильно, если ей соответствует единственный прообраз, а если таких прообразов несколько, то автоматически останавливает свой выбор на том из них, который имеет наибольшую частоту встречаемости в текстах. Приемлемость такой модели проверялась на английском, литовском, русском и таджикском языках, а также на искусственном языке эсперанто. Для всех языков эффективность безошибочного распознавания искаженного текста оказывалась приблизительно одинаковой, на уровне 97–98 %. При необходимости достижения более высоких показателей можно обратиться к расширенной модели, в которой мозг учитывает пары, а возможно, и тройки соседствующих буквенных совокупностей.

Ключевые слова: *текст, анаграмма, мозг, восприятие, математическая модель.*

Поводом для данного исследования послужило привлекшее внимание сообщение в Интернете. Прочтите его не по буквам и не по слогам, а как обычно читаете тексты.

> *По рзелульаттам илссеовадний одонго анлигйсокго унвиертисета, не иеемт занчнеия, в кокам пряоддке рсапожплены бкувы в солве. Осатьлыне бкувы мгоут селдовтаь в плоонм бсепордяке, все рвано ткест чтаитсея без побрелм. Пичрионй эгото ялвятеся то, что мы не чиатем кдаужю бкуву по отдльенотси, а все солво цликеом.*

Составив на английском, литовском, русском, таджикском языках и на языке эсперанто разнообразные тексты, делая во всех словах случайные перестановки букв и выставив такое творчество на обозрение, можно убедиться в том, что респонденты без затруднений понимают содержание полученного «произведения».

Постановка задачи

Примем, что текст какого-либо естественного языка является *анаграммно искаженным*, если каждому его слову сопоставлен набор из тех же самых букв, но расставленных в произвольном порядке. Фрагмент из Интернета как раз является примером такого текста. Каким же образом человеческий мозг легко справляется с его прочтением, восприятием и пониманием?

Предлагаемая далее математическая модель распознавания мозгом содержания искаженного текста основывается на гипотезе (упрощенной), состоящей из трех пунктов:

– мозг воспринимает текст пословно;

– мозг распознает слово правильно, если из заданного набора букв может быть составлено одно и только одно слово;

– мозг допускает ошибку, если из заданного набора букв может быть составлено несколько слов.

В двух последних пунктах в неявном виде просматривается понятие анаграммы. Поскольку в настоящей статье оно будет использоваться в несколько нетрадиционном смысле, приведем вначале два общеизвестных определения.

В энциклопедии *анаграмма* означает слово или словосочетание, образованное перестановкой букв другого слова или словосочетания.

В толковом словаре русского языка *анаграмма* – это перестановка букв, посредством которой из одного слова составляется другое.

Из первого определения следует, что анаграмма – *слово*, а из второго – *процедура* (перестановка букв). В дальнейшем будем пользоваться обоими определениями, а также и следующим: *анаграмма* – это конечное множество (по крайней мере, пара) слов естественного языка, составленных из одного и того же набора букв.

Анаграмму естественно называть *тривиальной*, если она состоит из одного элемента-слова. Рассмотрим, как обстоит дело с нетривиальными анаграммами в некоторых языках.

Инструмент для формирования подмножеств словоформных анаграмм

Пусть L – естественный язык с алфавитом A и $W = "\alpha_1\alpha_2 \ldots \alpha_n"$ – некоторое слово длины n, состоящее из букв $\alpha_k \in A$. Рассмотрим цепочку $CW = "\alpha_{s1}\alpha_{s2} \ldots \alpha_{sn}"$, составленную из тех же самых букв, что и слово W, но упорядоченных по алфавиту.

Определение 1. *Отображение* F: $W \rightarrow CW$ *назовем упорядоченным алфавитным кодированием (αβ-кодированием) слова* W*, а цепочку букв* CW *– его αβ-кодом* [1].

Для пояснения определения укажем, что αβ-кодирование, например, слова W = *"реферат"* приводит к цепочке CW = "*аеерртф*", а слова W = = *"агент"* – к той же самой цепочке CW = "*агент*", поскольку в этом слове буквы уже расположены в алфавитном порядке.

Кодирующее отображение F: $W \rightarrow CW$ является однозначным, декодирующее F^{-1}: $CW \rightarrow W$, вообще говоря, является многозначным, поскольку коду могут соответствовать несколько словоформ. Нарушение однозначности происходит на образах *анаграмм.* Например, цепочке букв CW = = "иикопрт" соответствует анаграмма из двух прообразов (тропики − киприот).

Всякой анаграмме соответствует единственный образ, соответствующий αβ-код. Этот факт подсказывает простой способ извлечения множества словоформных анаграмм из какого-либо текста. Вначале следует построить частотный словарь словоформ, затем каждой словоформе сопоставить ее αβ-код и, наконец, скомпоновать в подмножества словоформы с одинаковыми αβ-кодами.

Исследование декодирующего отображения F^{-1}

Обратимся к таблице 1. В ней в 1-м столбце представлен список языков, во 2-м – размеры (в словах) коллекций текстов. Результаты обработки исходных данных приведены в последующих столбцах: в 3-м указывается число различных словоформ, обнаруженных в коллекциях; в 4-м – число различных αβ-кодов словоформ; в 5-м и 6-м из общего числа кодов выделяются однозначные и многозначные, а в 7-м и 8-м столбцах данные 5-го и 6-го столбцов показаны в процентах.

Из столбцов 5–6 и 7–8 таблицы 1 видно, что для рассматриваемых языков *количество однозначно декодируемых кодов на порядок больше суммарного количества кодов анаграмм*. Полученный результат, отдавая приоритет однозначно декодируемым кодам, создает искаженную картину об обратном отображении F^{-1}, ибо не принимает в расчет частотности кодов и тех, и других.

Уточненные свойства F^{-1} выявляются в столбцах 4 и 6 таблицы 2. Действительно, в этих столбцах приводятся данные о суммарных частотах встречаемости однозначно декодируемых и двусмысленных (многозначных) кодов для соответствующих коллекций. Эти данные, пересчитанные в столбцах 4 и 6 в процентах, показывают недостаточно приемлемый для практических целей уровень декодирования слов, реализуемый отображением F^{-1}: относительные частоты встречаемости элементов анаграмм для отмеченных языков группируются вокруг значения 0,5. Иначе говоря, почти каждое второе слово из корпуса текстов принадлежит множеству анаграмм.

Определение модифицированного отображения $F^{(*)}$

Итак, αβ-кодирование, будучи удобным средством для выявления всевозможных элементов анаграмм, оказывается неэффективным для декодирования. В связи с этим обратимся к отображению $F^{(*)}$, которое наделим следующими свойствами [1]:

– как и F, оно определено на множестве $\{W\}$ слов естественного языка L;

– как и F, оно ставит в соответствие слову W его αβ-код, то есть $F^{(*)}$: $W \rightarrow CW$;

– обратное отображение $F^{(*)-1}$ на множестве однозначно декодируемых кодов совпадает с F^{-1}, а на множестве многозначных кодов (образов анаграмм) каждому образу CW ставит в соответствие единственное слово W^*, которое имеет максимальную частоту встречаемости в текстах в сравнении с другими словами из набора слов рассматриваемой анаграммы.

Последнее свойство назовем $F^{(*)}$*-схемой декодирования.* Очевидно, что принятие решений по этой схеме имеет вероятностный характер и предполагает возможность допущения ошибки в случаях, когда при правильном декодировании на выходе должно появиться слово не с максимальной частотой.

Практическое использование отображения $F^{(*)}$ предполагает наличие развитой БД $\{W \Leftrightarrow CW\}$, реализующей взаимно однозначное соответствие между элементами множеств $\{W\}$ и $\{CW\}$. Поскольку установление такого соответствия основывается на отмеченных свойствах функции $F^{(*)}$, ее априорную эффективность естественно определять суммарной частотой слов, αβ-коды которых декодируются по $F^{(*)}$*-схеме.* Соответствующие данные, приведенные в столбцах 7 и 8 таблицы 2, указывают на высокий уровень такого декодирования (не менее 94 % для эсперанто и 97 % для других языков).

Итак, $F^{(*)}$*-отображение* можно трактовать как *математическую модель* работы распознающего

Таблица 1

Статистические свойства отображения F: $W \to CW$

Table 1

The statistical properties of displaying F: $W \to CW$

Язык	Размер коллекции текстов	Число словоформ	Число различных αβ-кодов	Число однозначных кодов	Число многозначных кодов	(5)/(4), %	(6)/(4), %
1	2	3	4	5	6	7	8
Английский	11 252 496	137 732	119 055	106 841	12 214	89,7	10,3
Литовский	34 165 084	693 995	605 039	546 254	58 785	90,3	9,7
Русский	19 175 074	509 031	462 886	430 517	32 369	93,0	7,0
Таджикский	2 323 965	87 181	80 080	74 512	5 568	93,0	7,0
Эсперанто	5 080 195	165 570	147 220	133 851	13 369	90,9	9,1

Таблица 2

Частотности кодов

Table 2

Code frequencies

Язык	Размер коллекции	Частоты однозначных кодов	Отношение (3)/(2), %	Частоты многозначных кодов	Отношение (5)/(2), %	Частоты слов, декодируемых по F^*-схеме	Отношение (7)/(2), %
1	2	3	4	5	6	7	8
En	11 252 496	4 738 825	42,1	6513671	57,9	10 961792	97,4
Lt	34 165 084	15637702	45,8	18527382	54,2	33 199 335	97,2
Ru	19 175 074	10413280	54,3	8761794	45,7	18 724 384	97,7
Tj	2 323 965	1152420	49,6	1171545	50,4	2 280 334	98,1
Eo	5080195	1788681	35,2	3291514	64,8	4 782 380	94,1

мозга. Действительно, всякий раз, когда приходится иметь дело с очередным «сумбурным» словом, мозг интуитивно сопоставляет ему то единственное слово (с той же самой совокупностью букв), которое встречалось в его практике чаще других слов из их общей анаграммы. Именно такое свойство мозга заключено в $F^{(*)}$-схеме.

Обратим внимание еще на одну важную сторону $F^{(*)}$-отображения. Оно указывает, что исследование (а фактически заключение) «…одного английского университета», представленное во введении настоящей статьи, является *неточным*: для понимания анаграммно испорченного текста нет никакой необходимости фиксировать неподвижными первую и последнюю буквы слова (остальные перемешивать произвольно).

Далее рассмотрим другие модели понимания текста, которые в сравнении с $F^{(*)}$-моделью подчиняются дополнительным ограничениям. Априори очевидно, что такие модели будут иметь более высокий процент безошибочного восприятия текстов, однако, как будет показано далее, повышение точности не будет столь принципиальным.

***$F^{(*)}$*-модель распознавания текста** (фиксируются неизменными первые буквы слов).

Функция $F^{(f)}$ задается на множестве $\{W\}$ слов естественного языка L.

Определение 2. *Отображение $F^{(f)}$ слову W ставит в соответствие цепочку $\alpha_1 C(W/\alpha_1)$, в которой α_1 – первая буква в слове W и $C(W/\alpha_1)$ – αβ-код цепочки W/α_1, то есть слова W без первой буквы.*

В отличие от F это отображение оставляет в слове W неизменной первую букву, то есть α_1, и упорядочивает по алфавиту прочие буквы. Из общих соображений ясно, что декодирование $\alpha_1 C(W/\alpha_1) \to W$ в определенном смысле обладает лучшими свойствами, чем $CW \to W$.

Пример. Обратимся к анаграмме $\{W$: $W =$ *автор, втора, отвар, рвота, тавро, товар*$\}$. Отображение $F^{(f)}$ первые четыре элемента кодирует следующим образом: ***а**ворт, **в**аорт, **о**аврт, **р**авот*, оставляя неизменными первые буквы элементов анаграммы (отмечены жирным шрифтом) и располагая в алфавитном порядке прочие буквы. Этим кодам однозначно соответствуют первые четыре элемента анаграммы. Пятый и шестой элементы анаграммы кодируются одинаково – ***т**авро*.

***$F^{(f,l)}$*-модель распознавания текста** (не изменены первые и последние буквы слов).

Как и $F^{(f)}$, функция $F^{(f,l)}$ задается на множестве $\{W\}$ слов естественного языка L.

Определение 3. $F^{(f,l)} : W \to \alpha_1 C(W/\{\alpha_1, \alpha_n\})\alpha_n$.

В нем α_1 – первая и α_n – последняя буквы слова W остаются неподвижными, а цепочка букв между ними, то есть $W/\{\alpha_1, \alpha_n\}$, подвергается αβ-кодированию.

Теперь рассмотрим применение отображения $F^{(f,l)}$ к той же анаграмме, что и в примере для $F^{(f)}$. В этом случае первая и последняя буквы (далее показаны жирным шрифтом) элементов анаграмм должны оставаться неизменными, а все другие буквы упорядочиваются по алфавиту. Результаты

кодирования записываются в виде: ***автор, ворта, оавтр, рвота, тавро, тавор***, то есть все шесть слов рассматриваемой анаграммы получили собственные коды. Декодирование с помощью обратной функции также однозначно.

Итак, в сравнении с F отображения $F^{(f)}$ и $F^{(f,l)}$ несколько сложнее, зато наверняка успешнее в вопросах декодирования (см. табл. 3). Отметим также, что $F^{(f,l)}$ – именно то отображение, о котором шла речь в начале статьи.

Отметим, что для отображения F речь идет об анаграммах в смысле приведенного определения, а для отображений $F^{(f)}$ и $F^{(f,l)}$ – о соответствующим образом модернизированных анаграммах.

Замечание относительно отображений $F^{(f)}$ и $F^{(f,l)}$

Из данных столбцов 8 и 9 видно, что для рассматриваемых языков количество различных однозначно декодируемых кодов на порядок больше суммарного количества различных кодов анаграмм. Отметим, что здесь имеется в виду список слов частотных словарей, причем без учета частот их встречаемости. Полученный результат, отмечающий высокий процент однозначно декодируемых кодов, создает, однако, искаженную картину мощности множества слов, входящих в состав анаграмм. Сделанный вывод подтверждает дальнейшее исследование, результаты которого представлены в таблице 4.

Прежде чем переходить к рассмотрению этой таблицы, объясним смысл обозначений $\overline{F} = F^{(*)}$, $\overline{F}^{(f)}$ и $\overline{F}^{(f,l)}$. Как отмечалось ранее, отображения F, $F^{(f)}$ и $F^{(f,l)}$ каждому слову приписывают единственный код, однако обратные отображения в общем случае не обеспечивают однозначного декодирования. Использование отображений $\overline{F} = F^{(*)}$, $\overline{F}^{(f)}$ и $\overline{F}^{(f,l)}$ – это по существу попытка устранения неоднозначности при декодировании анаграмм и распознавании порождающих их прообразов за счет использования дополнительных атрибутов, присоединяемых к αβ-кодированию.

Определения отображений $\overline{F}$, $\overline{F}^{(f)}$ и $\overline{F}^{(f,l)}$

Определение 4. *Отображения $\overline{F}$, $\overline{F}^{(f)}$ и $\overline{F}^{(f,l)}$ обладают следующими свойствами:*

– *задаются на множестве слов $\{W\}$ языка L;*

– *совпадают, соответственно, с F, $F^{(f)}$ и $F^{(f,l)}$ при кодировании слов;*

– *их обратные отображения $(\overline{F})^{-1}$, $(\overline{F}^{(f)})^{-1}$ и $(\overline{F}^{(f,l)})^{-1}$ на кодах, однозначно декодируемых, совпадают, соответственно, с F^{-1}, $(F^{(f)})^{-1}$, $(F^{(f,l)})^{-1}$, а на кодах анаграмм каждому из них ставят в соответствие единственное слово W^*, которое имеет максимальную частоту встречаемости в текстах в сравнении с другими словами с одинаковым кодом.*

Таблица 3

Свойства отображений $F^{(f)}$ и $F^{(f,l)}$

Table 3

The properties of displaying $F^{(f)}$ and $F^{(f,l)}$

Язык	Размер коллекции	Число словоформ	Тип кода	Число различных кодов	Число однозначных кодов	Число многозначных кодов	Отношение (6)/(5), %	Отношение (7)/(5), %
1	2	3	4	5	6	7	8	9
En	11252496	137732	F	119055	106841	12214	89.74	10.26
			$F^{(f)}$	130644	124366	6278	95.19	4.81
			$F^{(f,l)}$	**135618**	**133570**	**2048**	**98.49**	**1.51**
Lt	34165084	693995	F	605039	546254	58785	90.28	9.72
			$F^{(f)}$	654475	621487	32988	94.96	5.04
			$F^{(f,l)}$	**675208**	**657925**	**17283**	**97.44**	**2.56**
Ru	19175074	509031	F	462886	430517	32369	93.01	6.99
			$F^{(f)}$	488286	470336	17950	96.32	3.68
			$F^{(f,l)}$	**500433**	**492360**	**8073**	**98.39**	**1.61**
Tj	2323965	87181	F	80080	74512	5568	93.05	6.95
			$F^{(f)}$	84220	81501	2719	96.77	3.23
			$F^{(f,l)}$	**85805**	**84499**	**1306**	**98.48**	**1.52**
Eo	5080195	165570	F	147220	133851	13369	90.92	9.08
			$F^{(f)}$	158310	151885	6425	95.94	4.06
			$F^{(f,l)}$	**162940**	**160407**	**2533**	**98.45**	**1.55**

Пример. Положим, что в анаграмме {*W*: *W* = *казан, казна, наказ*} наибольшую частоту в корпусе текстов имеют слово *наказ*, затем *казна*. При отображении *F* (= $\overline{F}$) рассматриваемой анаграмме будет соответствовать код *аазкн*, которому отображение $(\overline{F})^{-1}$ поставит в соответствие слово *наказ*.

Если же применить отображение $\overline{F}^{(f)}$ (= $F^{(f)}$), то первые два слова анаграммы получат одинаковый код ***к**аазн*, а третье слово – код ***н**аазк* (напомним, что при кодировании первые буквы в словах анаграмм фиксируются). Коду *каазн* будет сопоставляться слово *казна*, у которого частота больше, чем у слова *казан*.

Замечание. Предлагаемый в определении 4 метод выбора единственного прообраза того или иного кода анаграммы носит вероятностный характер. Он не исключает возможности принятия ошибочного решения в случаях, когда при правильном декодировании на выходе должно появиться слово не с максимальной частотой.

В таблице 4 приводятся результаты статистической обработки данных коллекций текстов с учетом определений трех типов кодирования словоформ.

Поясним эту таблицу. В ней первые два столбца – те же, что и в таблице 3. Столбец 3 отмечает три типа используемых способов кодирования слов. Столбцы 4–7 по существу продолжают таблицу 3. С учетом того, что, согласно определению 4, при кодировании слов отображения *F* и $\overline{F}$, $F^{(f)}$ и $\overline{F}^{(f)}$, $F^{(f,l)}$ и $\overline{F}^{(f,l)}$ совпадают, эти столбцы представляют информацию о частотах встречаемости однозначно декодируемых кодов и элементов анаграмм (многозначных кодов), причем столбцы 4 и 6 информируют об абсолютных, а 5 и 7 – об относительных значениях частот, выраженных в процентах по отношению к общему количеству слов (словоупотреблений) текстовых коллекций.

Из данных столбцов 4–7 и пяти строк, привязанных к отображению $\overline{F}$, видно, что для всех языков, за исключением русского, слова, являющиеся элементами анаграмм, составляют более половины общего числа слов (словоупотреблений) текстовых коллекций (для русского языка – 45,69 %). По этой причине при перемешивании всех букв в слове читающий мозг будет допускать много ошибок.

Таблица 4

Свойства отображений $\overline{F}$, $\overline{F}^{(f)}$ и $\overline{F}^{(f,l)}$

Table 4

The properties of displaying $\overline{F}$, $\overline{F}^{(f)}$ and $\overline{F}^{(f,l)}$

Язык	Размер коллекции	Тип кода	Частота однозначных кодов	Отношение (4)/(2), %	Частота многозначных кодов	Отношение (6)/(2), %	Частота слов, декодируемых по опр. 4	Отношение (8)/(2), %
1	2	3	4	5	6	7	8	9
Eng	11252496	$\overline{F}$	4738825	42.11	6513671	57.89	10961792	97.42
		$\overline{F}^{(f)}$	8282101	73.60	2970395	26.40	11179579	99.35
		$\overline{F}^{(f,l)}$	**10830821**	**96.25**	**421675**	**3.75**	**11224361**	**99.75**
Lt	34165084	$\overline{F}$	15637702	45.77	18527382	54.23	33199335	97.17
		$\overline{F}^{(f)}$	23737620	69.48	10427464	30.52	33834328	99.03
		$\overline{F}^{(f,l)}$	**28999545**	**84.88**	**5165539**	**15.12**	**34028900**	**99.60**
Ru	19175074	$\overline{F}$	10413280	54.31	8761794	45.69	18724384	97.65
		$\overline{F}^{(f)}$	14490319	75.57	4684755	24.43	19063618	99.42
		$\overline{F}^{(f,l)}$	**16449834**	**85.79**	**2725240**	**14.21**	**19145587**	**99.85**
Tj	2323965	$\overline{F}$	1152420	49.59	1171545	50.41	2280334	98.12
		$\overline{F}^{(f)}$	1759262	75.70	564703	24.30	2309336	99.37
		$\overline{F}^{(f,l)}$	**2021532**	**86.99**	**302433**	**13.01**	**2316348**	**99.67**
Eo	5080195	$\overline{F}$	1788681	35.21	3291514	64.79	4782380	94.14
		$\overline{F}^{(f)}$	4185715	82.39	894480	17.61	5033967	99.09
		$\overline{F}^{(f,l)}$	**4834326**	**95.16**	**245869**	**4.84**	**5068691**	**99.77**

Из пяти ячеек, стоящих на пересечении столбца 7, и пяти строк, привязанных к отображению $\overline{F}^{(f)}$, видно, что относительная частотность многозначных кодов располагается между значениями 17,61 % (эсперанто) и 30,52 % (литовский язык). Это и предопределяет невозможность достаточно успешного восприятия смысла слов с фиксированной первой буквой.

Из пяти ячеек, стоящих на пересечении столбца 7, и пяти строк, привязанных к отображению $\overline{F}^{(f,l)}$, видно, что частотности многозначных кодов среди общего числа словоупотреблений равны 4,84 % и 3,75 % соответственно для эсперанто и английского языка, то есть являются относительно малыми величинами, вследствие чего читающий мозг в большинстве случаев правильно справляется с пониманием слов искаженного текста. Наряду с этим в трех других языках частотность многозначных кодов примерно на 10 % выше, поэтому ошибочность в понимании искаженных слов будет недопустимой для практического применения $\overline{F}^{(f,l)}$-модели.

Итак, несмотря на то, что количество различных однозначно декодируемых слов (словоформ) оказалось на порядок больше количества различных слов, входящих в состав анаграмм (см. данные столбцов 8 и 9 для строк *F*-отображения), частоты встречаемости рассматриваемых элементов в тестовых коллекциях пяти языков оказались одного порядка.

Последние два столбца, 8-й и 9-й, таблицы 4 выдают количественные показатели эффективности декодирования в соответствии с определением 4. Для всех пяти языков ошибки принятия неверных решений заключаются в пределах 1 % при декодировании посредством $(F^{(f)})^{-1}$ и $(\overline{F}^{(f,l)})^{-1}$ и не превосходят 3 % при декодировании с помощью $(\overline{F})^{-1}$ (для эсперанто – не более 6 %).

Обсуждение результатов

Итак, отображения $\overline{F} = F^{(*)}$, $\overline{F}^{(f)}$ и $\overline{F}^{(f,l)}$ можно рассматривать в качестве математических моделей функционирования мозга, распознающего смысл анаграммно искаженного текста. В согласии с этим процесс понимания мозгом произвольного набора букв происходит следующим образом:

– если набору букв соответствует однозначный αβ-код, мозг сопоставляет ему единственную словоформу;

– если набору букв соответствует многозначный αβ-код, мозг сопоставляет ему ту словоформу из анаграммы, которая имеет максимальную частоту встречаемости.

Математическая модель, описываемая отображением $\overline{F}$, охватывает более общую ситуацию в сравнении с $\overline{F}^{(f)}$ и $\overline{F}^{(f,l)}$. В ней нет по существу каких-либо ограничений на характер перестановок букв в слове. Вместе с тем она обеспечивает достаточно высокий уровень (не менее 97 %) распознавания смысла искаженных слов, уступая двум другим моделям в точности результата не более чем на 2 % (см. столбец 9). Если сравнивать эффективность моделей по этим двум факторам, предпочтение следовало бы отдать $\overline{F}$-отображению.

Гипотеза (расширенная) распознавания искаженного текста

Предложенные в настоящей статье математические модели основывались на гипотезе восприятия каждого искаженного слова в отдельности. Даже в этом случае удалось добиться почти безошибочного понимания их смысла в искаженном тексте. При необходимости достижения более точных результатов можно обратиться к цепям Маркова [2, 3], привлекая для анализа пары соседствующих слов (так называемые словоформные биграммы), которые привнесут дополнительную информацию о семантической связи слов. Последние должны быть предварительно распознаны с помощью, например, $\overline{F}$-модели.

Заключение

Настоящая статья по содержанию созвучна работам [4] и [5], в которых обсуждается утверждение Г. Роулинсона о том, что случайное расположение букв в середине слов либо слабо влияет, либо совсем не влияет на способность квалифицированного читателя понимать текст. При этом авторы проходят мимо того факта, что если искаженному слову соответствует многозначный код, то для его правильного восприятия мозг интуитивно использует накопленную в регулярных чтениях информацию о частотности слов, входящих в анаграмму, и извлекает из памяти то слово, которое имеет максимальную частоту.

Литература

1. Усманов З.Д. Об упорядоченном алфавитном кодировании слов естественных языков // ДАН РТ. 2012. Т. 55. № 7. С. 545–548.

2. Марков А.А. Исчисление вероятностей. М.: ГИЗ, 1924. 589 с.

3. Гнеденко Б.В. Курс теории вероятностей. М.: Физматгиз, 1961. 408 с.

4. Rawlinson G.E. The significance of letter position in word recognition: PhD Thesis. Univ. of Nottingham Publ., UK, 1976.

5. MRC Cognition and Brain Sciences Unit. URL: http://www.mrc-cbu.cam.ac.uk/people/matt.davis/Cmabrigde (дата обращения: 20.02.2018).

Software & Systems
DOI: 10.15827/0236-235X.123.448-454

Received 23.04.18
2018, vol. 31, no. 3, pp. 448–454

Modeling brain activity recognizing anagrammatically distorted words

Z.D. Usmanov [1], *Dr.Sc. (Physics and Mathematics), Professor, zafar-usmanov@rambler.ru*

[1] *The Russian-Tajik (Slavonic) University (RTSU), Dushanbe, 734025, Tajikistan*

Abstract. The object of research are natural language texts the words in which were corrupted by random letter transpositions. The authors analyze the ability of a human brain to accurately recognize the meaning of distorted texts. offer mathematical models how the brain decides the problem.

The paper describes a mathematical model that explains how the brain solves the problem in cases when a) the first, b) the last, c) the first and last letters of words remain in their places, and all others are reset arbitrarily and, finally, in the most general case, d) when no letter is fixed and all letters within a word can be placed in any order. The explanation is based on the concept of a word anagram (in the broad sense, the set of its letters arranged in any sequence) as well as on the concept of an anagram prototype.

A simplified mathematical model assumes that the brain perceives each anagram separately; recognizes it correctly if it has a single prototype. In the case when there are several such prototypes, the brain automatically selects the one that has the highest frequency of occurrence in texts. The acceptability of this model was tested in English, Lithuanian, Russian and Tajik, as well as in the artificial language such as Esperanto. For all languages, efficiency of the correct recognition of distorted text was at the level of 97–98%. If it is necessary to achieve higher indicators, one can refer to an extended idea in which the brain takes into account couples, and maybe triples of neighboring letter sets.

Keywords: text, anagram, brain, perception, mathematical model.

References

1. Usmanov Z.D. On the Ordered Alphabetic Coding of Natural Language Words. *Reports of the Academy of Sciences of the Republic of Tajikistan*. 2012, vol. 55, no 7, pp. 545–548 (in Russ.).
2. Markov A.A. *Calculus of Probabilities*. Moscow, GIZ, 1924, 589 p.
3. Gnedenko B.V. *Course of the Probability Theory*. Moscow, Fizmatgiz Publ., 1961, 408 p.
4. Rawlinson G.E. *The Significance of Letter Position in the Word Recognition: Unpublished*. Univ. of Nottingham Publ., UK, 1976.
5. *MRC Cognition and Brain Sciences Unit*. Available at: http://www.mrc-cbu.cam.ac.uk/people/matt.davis/Cmabrigde (accessed February 20, 2018).

Примеры библиографического описания статьи

1. Усманов З.Д. Моделирование восприятия мозгом анаграммно искаженного текста // Программные продукты и системы. 2018. Т. 31. № 3. С. 448–454. DOI: 10.15827/0236-235X.123.448-454.

2. Usmanov Z.D. Modeling brain activity recognizing anagrammatically distorted words. *Software & Systems*. 2018, vol. 31, no. 3, pp. 448–454 (in Russ.). DOI: 10.15827/0236-235X.123.448-454.