УЧЕНЫЕ ЗАПИСКИ КАЗАНСКОГО УНИВЕРСИТЕТА
Том 155, кн. 4 Физико-математические науки *2013*

УДК 81.32+519.257+519.246.2

ПРОВЕРКА ЗАКОНА ХИПСА ПО ДАННЫМ КОРПУСА GOOGLE BOOKS NGRAM

В.В. Бочкарев, Э.Ю. Лернер, А.В. Шевлякова

Аннотация

Работа посвящена проверке выполнения эмпирического закона Хипса в европейских языках на материале корпуса текстов Google Books Ngram. Показано, что закон Хипса выполняется лишь для текстов ограниченного объёма и относящихся к небольшому историческому интервалу; показатель Хипса убывает со временем, а также испытывает значительные колебания с характерными временами 60–100 лет. В рамках простой вероятностной модели порождения текста рассмотрена связь между распределением частот словоупотребления и ожидаемой зависимостью числа уникальных слов в тексте от объёма текста. Эта модель даёт объяснение наблюдаемого нисходящего тренда показателя Хипса.

Ключевые слова: закон Хипса, закон Ципфа, вероятностные модели текста, корпус Google Books Ngram.

Введение

Интерес к изучению зависимости объёма лексикона от объёма представленного текста зародился ещё в XX веке и широко обсуждался в лингвистических и нелингвистических кругах. Наиболее известные достижения, касающиеся данного вопроса, связаны с эмпирическими законами Ципфа и Хипса. Закон Хипса описывает зависимость количества уникальных слов в тексте от объёма (длины) этого текста и в изначальной формулировке утверждает, что число этих слов увеличивается как корень квадратный из числа слов в тексте. В настоящее время вопрос о зависимости размера лексикона от объёма текста не потерял своей актуальности, несмотря на то что сейчас для анализа доступны большие корпусы текстов. В настоящей работе мы рассматриваем вопрос, выполняется ли закон Хипса для диахронических корпусов, а также изучаем связь между законами Хипса и Ципфа с помощью статистического анализа большого корпуса текстов, представленных в рамках проекта Google Books Ngram.

Согласно закону Ципфа частота употребления слова p_r определяется как степенная функция его ранга:

$$p_r \sim r^{-\beta}, \tag{1}$$

где r – ранг слова, то есть его номер в упорядоченном по убыванию частот списке [1]. С законом Ципфа тесно связан закон Хипса, который утверждает, что объём лексикона N (число различных слов) в тексте или наборе текстов объёма L определяется выражением $N \sim L^{-k}$. Различные вероятностные модели порождения текста приводят (при предположении, что закон Ципфа выполняется) к простому соотношению между показателями β и k:

$$k = \beta^{-1}. \tag{2}$$

Закон Ципфа был изначально установлен на относительно небольшом материале, при этом для показателя β в формуле (1) были получены оценки, близкие к 1.

В свою очередь, закон Хипса был изначально установлен при анализе корпуса новостных сообщений, при этом в первых работах показатель k был оценен как 0.5. В последующих работах предлагались различные обобщения данных законов, в том числе предполагающие общий случай степенной зависимости. Следует отметить также, что закон Хипса был сформулирован (и проверялся в дальнейшем) на материале корпусов текстов, созданных за относительно небольшой промежуток времени. Вопрос о применимости закона Хипса к большим диахроническим корпусам требует дополнительного обоснования.

Большие возможности для исследования статистических закономерностей словоупотребления открылись с созданием большого корпуса Google Books Ngram, в котором представлены данные для 9 языков и большого временного интервала [2–4]. Объём корпуса очень велик. Так, для английского языка за период 1900–2008 гг. корпус включает $2.94 \cdot 10^6$ томов, содержащих в общей сложности $2.39 \cdot 10^{11}$ слов. Для баз русского, немецкого и французского языков эти цифры составляют соответственно $3.34 \cdot 10^5$, $3.49 \cdot 10^5$, $2.59 \cdot 10^5$ и $2.58 \cdot 10^{10}$, $2.50 \cdot 10^{10}$, $2.26 \cdot 10^{10}$. В [4] детально проанализировано выполнение закона Ципфа на материале основных европейских языков. Было показано, что закон Ципфа в классической формулировке не выполняется, однако на графике распределения частот слов можно выделить два почти степенных участка (модель, предполагающая наличие двух степенных участков, была предложена впервые в работах [5, 6]). При этом на первом участке (для часто употребляемых слов) показатель степени близок к 1, а на втором (для редких слов) он значительно выше и для различных языков и периодов изменяется от 1.7 до 2.5. Последние значения уже значительно лучше согласуются с оценками показателя Хипса и выражением (2). Это, однако, не означает, что вопрос выбора наиболее адекватной простой модели частот словоупотребления можно считать решённым. Как отмечается в [4], хотя модель с двумя степенными участками значительно лучше соответствует эмпирическим данным по сравнению с ранее предлагавшимися моделями, тем не менее она должна быть отвергнута при любом разумном уровне значимости.

Типичный вид распределения частот употребления слов представлен на рис. 1. На рисунке показано распределение частот для общей базы английского языка на 2000 год. На рисунке приведена также аппроксимация эмпирических данных степенной моделью раздельно для диапазонов рангов 3–440 и $1 \cdot 10^4 - 5 \cdot 10^5$. Подгонка осуществлялась по методу максимального правдоподобия в предположении, что вектор частот имеет мультиномиальное распределение. Можно видеть, что даже на отдельных участках степенной закон выполняется достаточно условно. Зависимость имеет достаточно замысловатую форму, и представляется маловероятным, что её можно описать какой-либо простой моделью с малым числом параметров. Следует обратить внимание на подъём кривой в области рангов $500 - 2 \cdot 10^4$. Соответственно, меняя пределы участков, по которым выполняется подгонка, можно получить значения показателя степени, изменяющиеся в достаточно широких пределах. Действительно, хорошее соответствие степенному закону наблюдается для диапазона рангов 3–300, при этом показатель значимо больше 1 (для приведённого на рисунке примера значение показателя 1.077). Отметим, что побуквенные и слоговые марковские модели словообразования дают оценки для показателя, строго большие 1 [7, 8]. Лучшее соответствие модели эмпирическим данным в области малых рангов вызывает удивление, так как естественно ожидать, что вероятностные модели лучше описывают редкую лексику и неологизмы. Для объяснения зависимости $N(L)$ большее значение имеет распределение частот в области больших рангов (то есть для редкой лексики). Следует отметить, что степенной закон для частот в этой области нельзя считать надежно установленным.

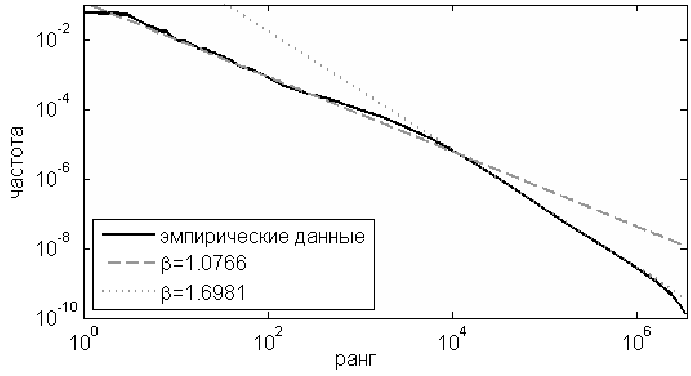

Рис. 1. Распределение частот употребления слов, общая база английского языка, 2000 г. Сплошная кривая – эмпирические данные, пунктирная и штриховая линии – степенные аппроксимации с $\beta = 1.0766$ и $\beta = 1.6981$ соответственно

Имеющиеся эмпирические данные не противоречат предположению о субстепенном законе, который также может быть обоснован в рамках марковских моделей словообразования [7].

В [4] для объяснения эмпирических данных используется модель, предполагающая повышение вероятности повторного употребления слов, а также разделение лексики на ядро и периферию.

Мы рассмотрим более простую модель порождения текста. Пусть у нас есть (конечный или бесконечный) набор возможных слов, вероятности употребления которых на очередном шаге заранее известны и равны p_i. Сколько в среднем различных слов будет хотя бы раз использовано в тексте длины L? Для ответа воспользуемся методом индикаторов. Рассмотрим случайную величину, принимающую значение 1, если i-е слово из нашего лексикона будет хотя бы раз использовано в тексте, в противном случае случайная величина принимает значение нуль. Вероятность того, что i-е слово ни разу не будет использовано, равна $(1-p_i)^L$, соответственно, наша случайная величина принимает значение, равное единице с вероятностью $1-(1-p_i)^L$, её математическое ожидание совпадает с этим же числом. Математическое ожидание суммы таких случайных величин и есть среднее количество использованных слов N. Отсюда получаем

$$N(L) = \sum_i \left(1-(1-p_i)^L\right). \tag{3}$$

Таким образом, в рамках описанной модели мы можем при заданных частотах слов оценить ожидаемую зависимость $N(L)$.

1. Сопоставление модели и эмпирических данных

На рис. 2 представлена зависимость объёма лексикона от объёма текстов по данным корпуса Google Books Ngram для общей базы английского языка. Подсчитывалось количество различных словоформ, употреблённых в данном году (без различения регистра), а также общий объём включенных в базу текстов (в словах). Учитывались только словоформы, состоящие из букв латинского алфавита. На рис. 2 приведена также аппроксимация эмпирических данных степенной зависимостью (пунктирная линия). Показатель степени находился по критерию минимума относительной квадратичной ошибки и составил 0.5503. Как видим, эмпирические данные плохо описываются степенной функцией (соответствие наблюдается

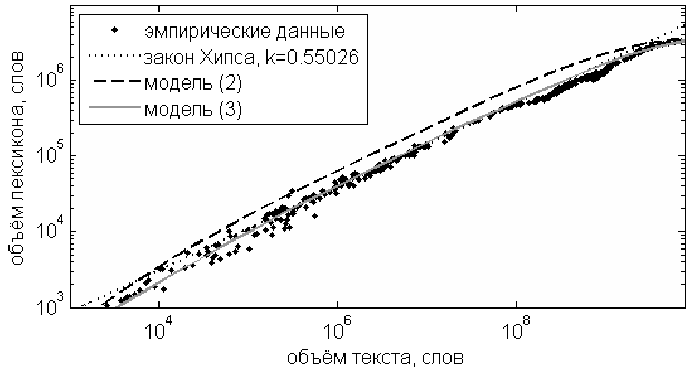

Рис. 2. Зависимость объёма лексикона от объёма текстов для общей базы английского языка. Точками показаны эмпирические данные, пунктирной линией – степенная аппроксимация. Штриховая и сплошная линии – расчет соответственно по моделям (3) и (4)

в диапазоне от 10^5 до 10^8, а вне его – значительное расхождение). Заметим, что для других представленных в базе Google Books Ngram языков соответствие степенному закону хуже, чем для английского языка.

Проводилось также моделирование ожидаемого объёма лексикона в соответствии с формулой (3). Для расчёта ожидаемого объёма лексикона в соответствии с (3) необходимо знать частоты словоупотребления. Между тем оценка частот редких слов достаточно проблематична, особенно для более раннего периода, для которого объём текстов, представленных в Google Books Ngram, относительно мал. Редко употребляемые слова могли просто не попасть в корпус, пока он не достиг достаточного годового объёма, и оценить изменения частот таких слов со временем не представляется возможным. Применение параметрических оценок частот также затруднено ввиду отсутствия на данный момент адекватной модели частот словоупотребления. Мы использовали в качестве оценки вероятностей p_r эмпирические частоты словоупотребления на 2000 год, так как для него в корпусе представлен наибольший объём текстов. Соответственно, мы таким образом можем оценить частоты наибольшего ($3.97 \cdot 10^6$) количества уникальных слов. Результаты расчётов по (3) показаны на рис. 2 штриховой линией. Можно видеть, что найденная модельная зависимость по форме близка к эмпирической, но проходит несколько выше нее. Для того чтобы добиться лучшего соответствия модели и эксперимента, модель была усовершенствована.

Традиционно слова подразделяют на содержательные и служебные. Частотность последних напрямую связана с синтаксической структурой предложения. Доля служебных слов в общем объёме словоупотребления по данным Google Books Ngram для общей базы английского языка приведена на рис. 3 (мы использовали в подсчётах список служебных слов английского языка, приведённый в [9]. Как видно из рисунка, доля употребления служебных слов может быть достаточно велика – в 1800–2000 гг. она изменялась в пределах от 0.48 до 0.57. Можно предположить, что в наборе текстов, содержащих достаточно широкий круг тем, служебные слова будут употреблены все, а число различных содержательных слов будет увеличиваться с увеличением общего объёма текста. Это приводит к следующей модифицированной модели:

$$N(L) = N_{\text{serv}} + \sum_{i \in I} \left(1 - (1 - p_i)^{\theta L}\right), \tag{4}$$

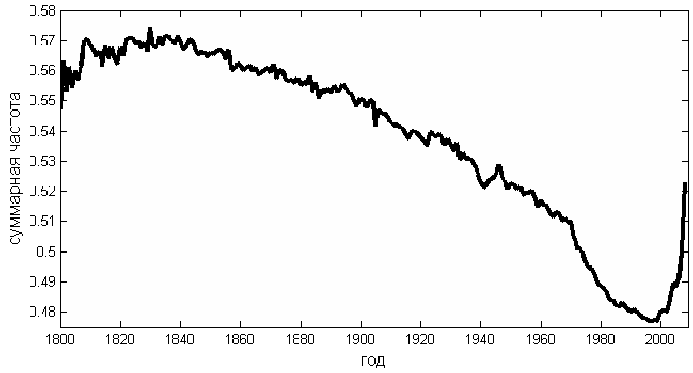

Рис. 3. Изменение доли служебных слов в английском языке со временем

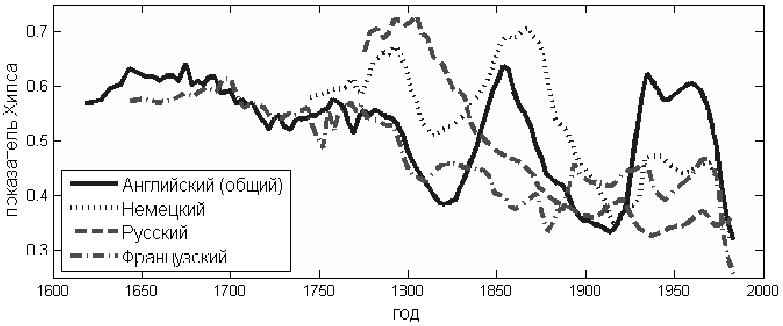

Рис. 4. Изменение показателя Хипса со временем для четырех европейских языков

где N_{serv} – число служебных слов, I – множество номеров содержательных слов в общем списке, θ – доля содержательных слов в тексте. В ходе моделирования для каждого года по данным Google Books Ngram определялся параметр θ, после чего по формуле (4) вычислялся ожидаемый объём лексикона. Результаты моделирования показаны на рис. 2 сплошной линией.

Видно, что модель (4) дает наилучшую аппроксимацию эмпирических данных. Подгонка модельной кривой степенной зависимостью (на участке от 10^3 до 10^{10}) даёт значение показателя 0.5674, что близко к указанному выше значению, полученному при подгонке эмпирических данных. В то же время наблюдаются достаточно серьёзные расхождения модельной и эмпирической зависимостей. Например, вариации сложной формы в области больших значений едва ли могут быть объяснены какой-либо простой моделью.

Соответственно, наблюдаемые расхождения могут быть как следствием несовершенства модели, так и результатом динамических процессов в языке. Для того чтобы проверить, какая из этих возможностей реализуется, проводился анализ данных раздельно для различных временных интервалов. Данные, отобранные с помощью скользящего временного окна длиной 50 лет, аппроксимировались степенной зависимостью. Полученные в результате зависимости изменения показателя Хипса со временем для английского, русского, немецкого и французского языков представлены на рис. 4. Можно отметить две особенности полученных кривых. Во-первых, для всех языков наблюдается нисходящий тренд. Во-вторых, на графиках присутствуют квазипериодические вариации с характерными временными масштабами 60–100 лет.

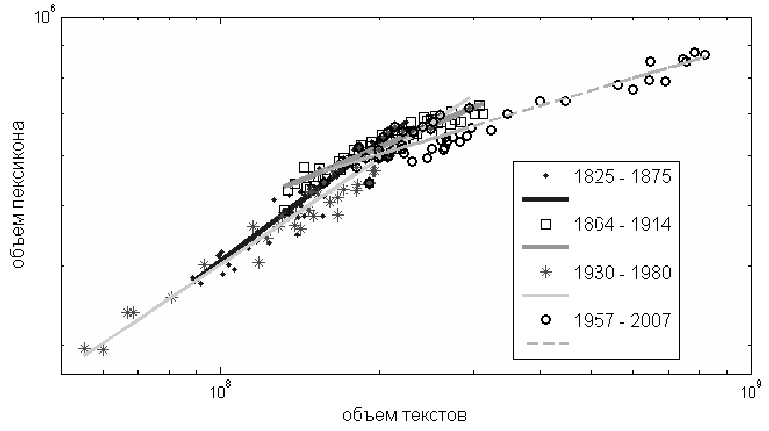

Рис. 5. Зависимость объёма лексикона от объёма текста в различные периоды для британского английского языка

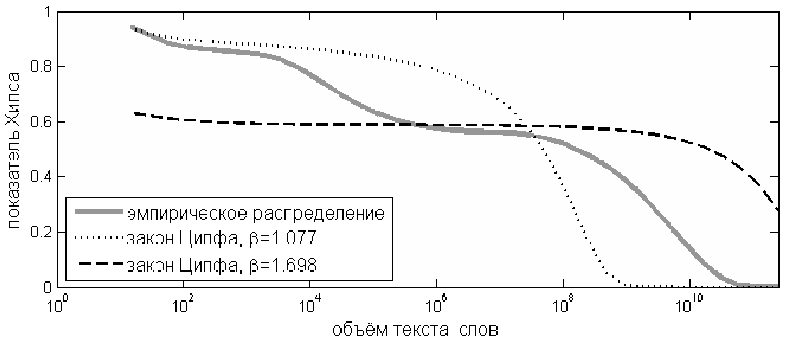

Рис. 6. Модельные значения показателя Хипса при различных объёмах корпуса. Сплошная кривая – эмпирическое распределение частот (английский язык, 2000 г.), пунктирная и штриховая линии – ципфовские распределения с показателями $\beta = 1.077$ и $\beta = 1.698$ соответственно

Для того чтобы убедиться, что вариации показателя, которые представлены на рисунке, не являются результатом ошибки при обработке, мы аппроксимировали данные степенной зависимостью раздельно для выбранных периодов. Были выбраны 4 года, соответствующие экстремумам временной зависимости показателя Хипса, и выделены интервалы длительностью 51 год, средней точкой которых являются выбранные годы. Результаты представлены на рис. 5. Различными маркерами показаны эмпирические значения, относящиеся к различным временным интервалам, а линиями – их аппроксимации степенной зависимостью. Как видно из рис. 5, изменение степенного показателя не вызывает сомнений.

В среднем для более позднего времени в базе Google Books Ngram объём представленных текстов больше. Возникает вопрос, можно ли объяснить наблюдаемые изменения показателя Хипса со временем ростом среднего объёма корпуса или они свидетельствуют о некоторых динамических процессах в языке. На рис. 6 показаны результаты моделирования ожидаемого показателя Хипса при различном объёме текстов. Мы проводили моделирование, исходя как из эмпирических частот словоупотребления, так и из модельных частот, соответствующих закону Ципфа. Для каждого малого участка зависимость $N(L)$ аппроксимируется степенным

законом $N \sim L^{-k(L)}$, причем $k(L)$ будем считать медленно изменяющейся функцией. Искомый показатель степени $k(L)$ может быть найден по формуле

$$k(L) = \frac{\partial \ln N}{\partial \ln L}. \tag{5}$$

Моделирование показателей Хипса проводилось по формулам (3), (5) с использованием эмпирических частот употребления слов для общей базы английского языка на 2000-й год (сплошная кривая). Мы наблюдаем на графике нисходящий тренд и стремление ожидаемого показателя Хипса к нулю при $L \to \infty$. Это вполне естественно, так как в нашей модели заложен конечный потенциальный лексикон, и зависимость $N(L)$ должна выходить на насыщение. На графике наблюдаются также две «ступеньки». Пунктирная и штриховая линии на рис. 6 5 соответствуют результатам моделирования в предположении, что частоты употребления слов точно соответствуют закону Ципфа $p_r = Ar^{-\beta}$ (с показателями β, выбранными в соответствии с приведенной на рис. 1 аппроксимацией эмпирических частот и равными соответственно 1.077 и 1.698), при этом число возможных слов конечно и равно $3.97 \cdot 10^6$. Так же как и в предыдущем случае, наблюдаем асимптотическое стремление кривой к нулю. При этом на графике локального показателя Хипса наблюдается почти горизонтальный участок вблизи значения $k(L) = 1/\beta$, как и следует ожидать из приведенных выше соображений о связи показателей законов Ципфа и Хипса. Таким образом, можно предположить, что два пологих участка на первой кривой соответствуют двум почти степенным участкам на графике частот словоупотребления (рис. 1).

Из сравнения рис. 4 и 6 следует, что уменьшение со временем показателя Хипса может быть, по-видимому, объяснено ростом объёма корпуса текстов. Однако как модель (3), так и модель (4) могут дать (в отличие от приведённых на рис. 3) только монотонно убывающие зависимости для показателя Хипса. Таким образом, наблюдаемые на рис. 4 квазипериодические вариации показателя Хипса, скорее всего, должны объясняться динамическими процессами в языке.

Заключение

В работе установлено, что закон Хипса выполняется для корпусов ограниченного объёма и относящихся к небольшому историческому интервалу. Для больших корпусов текстов необходимо учитывать явление насыщения зависимости $N(L)$. Для диахронических корпусов необходимо также принимать во внимание динамику лексики. Анализ эмпирических данных, представленных в корпусе Google Books Ngram, показывает, что для показателя Хипса наблюдаются вариации с характерным временным интервалом 60–100 лет, отражающие динамические процессы в языке.

Работа выполнена при финансовой поддержке РФФИ, (проект № 12-06-00404-а).

Summary

V.V. Bochkarev, E.Yu. Lerner, A.V. Shevlyakova. Verification of the Heaps Law Using the Google Books Ngram Database.

This article is devoted to the verification of the Heaps empirical law for European languages using the Google Books Ngram corpus data. It is shown that the Heaps law holds only for short texts and texts related to short historical periods. The Heaps exponent decreases in time and varies significantly within characteristic intervals of 60–100 years. The relationship between the word frequency distribution and the expected dependence of the number of individual words

on the text size is analyzed in terms of a simple probability model of text generation. This model serves as an explanation for the observed decreasing trend of the Heaps exponent.

Keywords: Heaps law, Zipf law, text probability models, Google Books Ngram corpus.

Литература

1. *Baayen R.H.* Word Frequency Distributions. – Dordrecht: Kluwer Acad. Pub., 2001. – 359 p.
2. *Michel J.B., Shen Y.K., Aiden A.P., Veres A., Gray M.K., The Google Books Team, Pickett J.P., Hoiberg D., Clancy D., Norvig P., Orwant J., Pinker S., Nowak M.A., Aiden E.L.* Quantitative analysis of culture using millions of digitized books // Science. – 2011. – V. 331. – P. 176–182.
3. *Petersen A.M., Tenenbaum J.N., Havlin S., Stanley H.E., Perc M.* Languages cool as they expand: Allometric scaling and the decreasing need for new words // Sci. Rep. – 2012. – V. 2. – Art. 943, P. 1–10. – doi: 10.1038/srep00943.
4. *Gerlach M., Altmann E.G.* Stochastic model for the vocabulary growth in natural languages // Phys. Rev. X. – 2013. – V. 3, No 2. – P. 021006-1–021006-10.
5. *Naranan S., Balasubrahmanyan V.K.* Models for power law relations in linguistics and information science // J. Quant. Linguist. – 1998. – V. 5, No 1–2. – P. 35–61.
6. *Ferrer i Cancho R., Solé R.V.* Two regimes in the frequency of words and the origins of complex lexicons: Zipf's law revisited // J. Quant. Linguist. – 2001. – V. 8, No 3. – P. 165–173.
7. *Bochkarev V.V., Lerner E.Yu.* Zipf and non-Zipf laws for homogeneous Markov chain. – 2012. – arXiv:1207.1872v2.
8. *Bochkarev V.V., Lerner E.Yu.* The Zipf law for random texts with unequal letter probabilities and the Pascal pyramid // Russ. Math. (Iz. VUZ). – 2012. – V. 56, No 12. – P. 25–27.
9. *Hughes J.M., Foti N.J., Krakauer D.C., Rockmore D.N.* Quantitative patterns of stylistic influence in the evolution of literature // Proc. Natl. Acad. Sci. USA. – 2012. – V. 109, No 20. – P. 7682–7686.

Поступила в редакцию
17.10.13

Бочкарев Владимир Владимирович – ассистент кафедры радиофизики, Казанский (Приволжский) федеральный университет, г. Казань, Россия.
E-mail: *vbochkarev@mail.ru*

Лернер Эдуард Юльевич – кандидат физико-математических наук, доцент кафедры анализа данных и исследования операций, Казанский (Приволжский) федеральный университет, г. Казань, Россия.
E-mail: *eduard.lerner@gmail.com*

Шевлякова Анна Владимировна – кандидат филологических наук, преподаватель кафедры английского языка для естественно-научных специальностей, Казанский (Приволжский) федеральный университет, г. Казань, Россия.
E-mail: *anna_ling@mail.ru*