УДК 811.11

СТАТИСТИЧЕСКАЯ ОБРАБОТКА ЛИНГВИСТИЧЕСКИХ ДАННЫХ НИДЕРЛАНДСКО-РУССКИХ СЛОВАРЕЙ

Д. С. Воевудский*, В. А. Тушавин**

* *Воронежский государственный университет*
** *Санкт-Петербургский государственный университет аэрокосмического приборостроения*

Поступила в редакцию 04.12.2012 г.

Аннотация. В статье произведен анализ распределения длины нидерландских слов в трех нидерландско-русских словарях. На основании проведенного анализа построена математическая модель, выявляющая закономерность в распределении частот слов различной длины в зависимости от средней длины слова в анализируемых словарях посредством аппроксимации распределения Пуассона методом максимального правдоподобия. Произведена проверка адекватности полученной модели методом Монте-Карло. Теоретически обоснована и практически верифицирована математическая модель разности в длине двух случайных слов на основе распределения Скеллама.
Ключевые слова: длина слов, нидерландский язык, распределение Пуассона, метод Монте-Карло, распределение Скеллама, GNU R.
Annotation. In the article the authors analyse the distribution of the length of words in three Dutch-Russian dictionaries. Based on the analysis, a mathematical model reveals regularity in the distribution of word frequencies of different length depending on the average length of words in dictionaries analyzed by Poisson approximation method of maximum likelihood. Performed the validation of the model obtained by the Monte Carlo method. Theoretically grounded and practically verified mathematical model of the difference in the length of two random words on the basis of Skellam distribution.
Keywords: length of words, the Dutch language, Poisson distribution, Monte Carlo method, Skellam distribution, GNU R.

ВВЕДЕНИЕ

Нидерландский, или как его раньше называли голландский, язык входит в западногерманскую подгруппу германских языков. Это государственный язык Нидерландов и один из двух государственных языков Бельгии. Он также является государственным языком Суринама (бывшей нидерландской колонии) и официальным языком Нидерландских Антильских островов. Общее количество лиц, для которых он является родным, составляет примерно 21 млн.

Целью предлагаемого исследования является выявление закономерностей распределения количества слов по длине посредством построения адекватной стохастической модели.

Для достижения поставленной цели были решены следующие задачи: 1) создание электронных баз данных исследуемых словарей; 2) обработка и аппроксимация полученных данных различными видами распределений с помощью GNU R; 3) проверка полученных результатов с помощью метода Монте-Карло.

ПРАКТИЧЕСКАЯ ЧАСТЬ

Раздел квантитативной лексикологии, основанный на параметрическом анализе лексики, разработанном В. Т. Титовым [1, 2], предполагает в качестве одного из этапов исследования лексической системы языка выявление ее функционального ядра.

Показателем функциональной активности слова является его длина. Со времен Джорджа Ципфа известно, что частотность слов обратно пропорциональна их длине: чем короче слово, тем (при прочих равных условиях) чаще оно употребляется, и наоборот [3]. Поскольку именно звуковая форма является первичной реальностью языка, данные по этому параметру брались в звуках. Для этого показатели длины в буквах были обработаны по правилам чтения нидерландского языка [4, с. 74-75]. Для анализа были взяты три нидерландско-русских словаря различного размера [5-7]. По данному парамет-

ру нидерландская лексика оказалась организованной, как это представлено в таблице 1.

Вся обработка данных и их графическое представление произведена с помощью языка статистической обработки GNU R ("GNU S"), являющимся открытым и свободным программным обеспечением, насчитывающий на момент написания этой статьи 4163 пакета расширения [8] и получивший признание среди специалистов за рубежом. В 2010 году проект R вошёл в список победителей конкурса лучшего открытого программного обеспечения года «InfoWorld Bossie Awards 2010» [9]. Язык R активно применяется ведущими зарубежными компаниями, такими как Google, Pfizer, Merck, Bank of America, the InterContinental Hotels Group и Shell, а также ведущими учебными заведениями. Его поддержка включена в коммерческие

Таблица 1

Распределение нидерландских слов по длине (по данным двуязычных словарей)

	Дренясова [7]	Баар [6]	Миронов [5]
Звуков	Слов		
1	0	0	1
2	31	56	76
3	496	916	1319
4	729	1751	2143
5	979	3086	3569
6	1028	4556	5121
7	1058	6670	7185
8	1020	8468	8746
9	730	8203	7890
10	540	6713	5997
11	335	5068	4143
12	193	3576	2689
13	99	2634	1814
14	67	1775	1098
15	36	1187	682
16	13	718	372
17	10	475	213
18	3	279	110
19	6	162	77
20	0	86	32
21	0	64	20
22	0	33	13
23	1	17	7
24	0	7	1
25	1	5	3
26	0	2	1
27	0	4	1
28	0	0	0
29	0	0	0
30	0	1	0
Всего слов	7375	56512	53323
Средняя длина	7,108	9,129	8,518
1 квартиль	5	7	7
Медиана	7	9	8
3 квартиль	9	11	10

(SPSS, Statistica, Oracle Data Mining) и открытые (Gretl) пакеты программного обеспечения. В то же время в отечественной научно-практической литературе вопрос применения языка R для решения задач математической лингвистики остаётся нераскрытым. Частично эту задачу планируется решить в данной статье посредством приведением в её тексте части листингов, содержащих расчеты и выводы расчетных величин, что также должно способствовать проверке данного исследование научной общественностью.

Анализ словарей был проведен с помощью методов описательной статистики, а также посредством визуализации данных на диаграмме типа «скрипка» (violin plot), как это показано на рисунке 1. Эта графическая форма представления дает больше информации о характере распределения, чем «ящик с усами» (box-and-whisker plot), т.к. помимо данных о медиане и квартилях, отражает еще и показатели ядерной плотности распределения [10].

Листинг, содержащий данные по словарям, а также построение диаграммы и описательная статистика представлен ниже:

Из полученного распределения можно сделать вывод о том, что между словарями больших размеров наблюдается больше сходства, чем с малым словарем. Разницу в положении медианы и квартилей можно объяснить тем, что в словарях большого объема больше представлена специальная лексика и термины, которые обычно обладают большей длиной, нежели общеупотребительные слова.

Также можно заметить, что распределение не является двухмодальным и ассиметрично со смещением вправо.

Затем полученные данные были обработаны методом бутстреппинга по методике, предложенной в работе Каллена и Фрея [11], чтобы на основании расчета моментов выяснить, какое распределение следует использовать для последующей аппроксимации. Листинг с результатами расчетов моментов приводится ниже. Результат расчетов в графическом виде представлен на рисунке 2.

Как видно из рисунка, наиболее близкие результаты дают отрицательное биноминальное распределение и распределение Пуассона. Следует отметить, что пакет R использует в качес-

```
> # Ввод данных по трем словарям
> LNidM<-c (rep (1, 1), rep (2, 76), rep (3, 1319), rep (4, 2143), rep (5, 3569), rep (6,
5121), rep (7, 7185), rep (8, 8746), rep (9, 7890), rep (10, 5997), rep (11, 4143), rep
(12, 2689), rep (13, 1814), rep (14, 1098), rep (15, 682), rep (16, 372), rep (17, 213),
rep (18, 110), rep (19, 77), rep (20, 32), rep (21, 20), rep (22, 13), rep (23, 7), rep
(24, 1), rep (25, 3), rep (26, 1), rep (27, 1))
> LNidD<-c (rep (1, 0), rep (2, 31), rep (3, 496), rep (4, 729), rep (5, 979), rep (6, 1028),
rep (7, 1058), rep (8, 1020), rep (9, 730), rep (10, 540), rep (11, 335), rep (12, 193), rep
(13, 99), rep (14, 67), rep (15, 36), rep (16, 13), rep (17, 10), rep (18, 3), rep (19, 6),
rep (20, 0), rep (21, 0), rep (22, 0), rep (23, 1), rep (24, 0), rep (25, 1))
> LNidB<-c(rep (2,56), rep (3,916), rep (4,1751), rep (5,3086), rep (6,4556), rep (7,6670),
rep (8,8468), rep (9,8203), rep (10,6713), rep (11,5068), rep (12,3576), rep (13,2634),
rep (14,1775), rep (15,1187), rep (16,718), rep (17,475), rep (18,279), rep (19,162), rep
(20,86), rep (21,64), rep (22,33), rep (23,17), rep (24,7), rep (25,5), rep (26,2), rep
(27,4), rep (30,1))
> library(wvioplot)
> # рисунок 1
>wvioplot(LNidM,LNidD,LNidB,names=c("Миронов","Дренясова","Баар"), col="chartreuse1",adj
ust=2)
> summary(LNidM)
 Min. 1st Qu. Median Mean 3rd Qu. Max.
 1.000 7.000 8.000 8.518 10.000 27.000
> summary(LNidD)
 Min. 1st Qu. Median Mean 3rd Qu. Max.
 2.000 5.000 7.000 7.108 9.000 25.000
> summary(LNidB)
 Min. 1st Qu. Median Mean 3rd Qu. Max.
 2.000 7.000 9.000 9.129 11.000 30.000
```

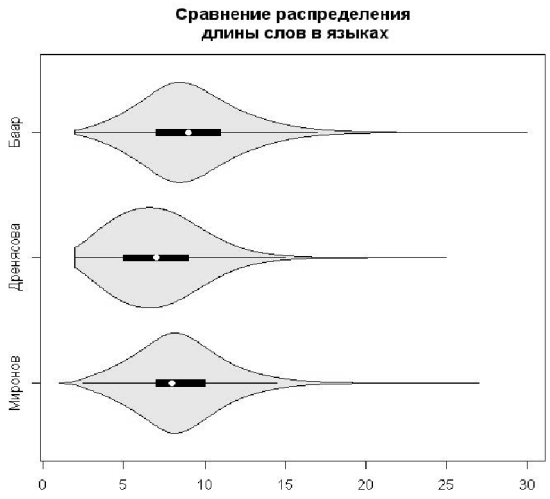

Рис. 1. Распределение в виде скрипки по исследуемым словарям

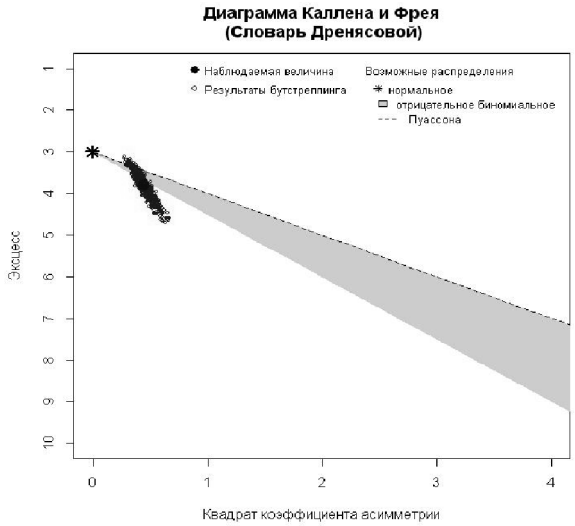

Рис. 2. Диаграмма Каллена-Фрея по словарю Дренясовой

```
> library(fitdistrplus)
> descdist(LNidD,main="Диаграмма Каллена и Фрея\n (Словарь Дренясовой)",discrete = TRUE,boot=1000)
summary statistics
------
min: 2 max: 25
median: 7
mean: 7.108475
estimated sd: 2.658719
estimated skewness: 0.6671574
estimated kurtosis: 3.806611
```

тве коэффициента эксцесса (kurtosis) именно четвертый центральный момент, без вычитания трех. Дальнейший анализ результатов аппроксимации методом максимального правдоподобия показал, что более адекватно в данном случае распределение Пуассона.

Распределение Пуассона относится к семейству дискретных распределений и задаётся следующей функцией вероятности:

$$f(k,\lambda) = \Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \qquad (1)$$

```
> library(fitdistrplus)
> summary(LNidD.fit<-fitdist(LNidD,"pois"))
Fitting of the distribution ' pois ' by maximum likelihood
Parameters :
 estimate Std. Error
lambda 7.108475 0.03104611
Loglikelihood: -17454.23 AIC: 34910.46 BIC: 34917.36
> summary(LNidD.fit1<-fitdist(LNidD,"nbinom"))
Fitting of the distribution ' nbinom ' by maximum likelihood
Parameters :
 estimate Std. Error
size 2.717821e+05 55.57441685
mu 7.109104e+00 0.03104926
Loglikelihood: -17454.23 AIC: 34912.46 BIC: 34926.27
Correlation matrix:
 size mu
size 1.000000e+00 -7.846869e-07
mu -7.846869e-07 1.000000e+00
```

где: $\lambda > 0$; k = 0,1,2,...; e – основание натурального логарифма. Проведенные вычисления показали, что при использовании при аппроксимации распределения Пуассона $\lambda \approx$ средней длине слова в анализируемом словаре. Так, для словаря ван ден Баара эта величина – 9,1, для словаря Миронова – 8,5, для словаря Дренясовой – 7,1.

Затем была проведена проверка адекватности аппроксимации распределением Пуассона – были взяты случайные выборки по каждому из словарей в количестве 2000, эти выборки были проведены 10000 раз и для каждого раза вычислялся критерий согласия Пирсона. Количество успехов, где эмпирическое и теоретическое распределения совпадают, было равно 8441 для словаря Баара, 8508 – для словаря Дренясовой и 8426 – для словаря Миронова. Таким образом, распределение Пуассона обеспечивает достаточно высокое качество аппроксимации и позволяет выявить закономерность в частотном распределении слов в словаре. Графически результаты аппроксимации и соответствующие графики квантилей представлены на рисунке 3.

Дополнительно произведен анализ параметра λ методом бутстреппинга.

Результаты, представленные на рисунке 4, позволяют говорить о сравнительно высокой точности первоначальной аппроксимации. Как видно из графика, реальные данные (медиана) и аппроксимированные отличаются незначительно.

Характерное двухмодальное распределение для словаря Баара показывает наличие страт по общей и специальной лексике, о чём было сказано выше.

Таким образом, имеющиеся эмпирические распределения слов по длине могут быть описаны распределением Пуассона. Исходя из изложенного, можно выдвинуть гипотезу, что разница в длине случайно взятых слов должна описываться распределением Скеллама, которое выражает разницу между двумя распределениями Пуассона [12]. Оно задается следующей функцией вероятности:

$$f(k,\lambda_1,\lambda_2) = e^{-(\lambda_1+\lambda_2)}\left(\frac{\lambda_1}{\lambda_2}\right)^{k/2} I_{|k|}(2\sqrt{\lambda_1\lambda_2}), \qquad (2)$$

где λ_1, λ_2 – параметры двух распределений Пуассона (1), а $I|k|$ – модифицированная функция Бесселя первого рода (функция Инфельда).

Данная гипотеза была проверена методом Монте-Карло с использованием теста Пирсона, результаты представлены в таблице 2, фрагмент листинга приведен ниже.

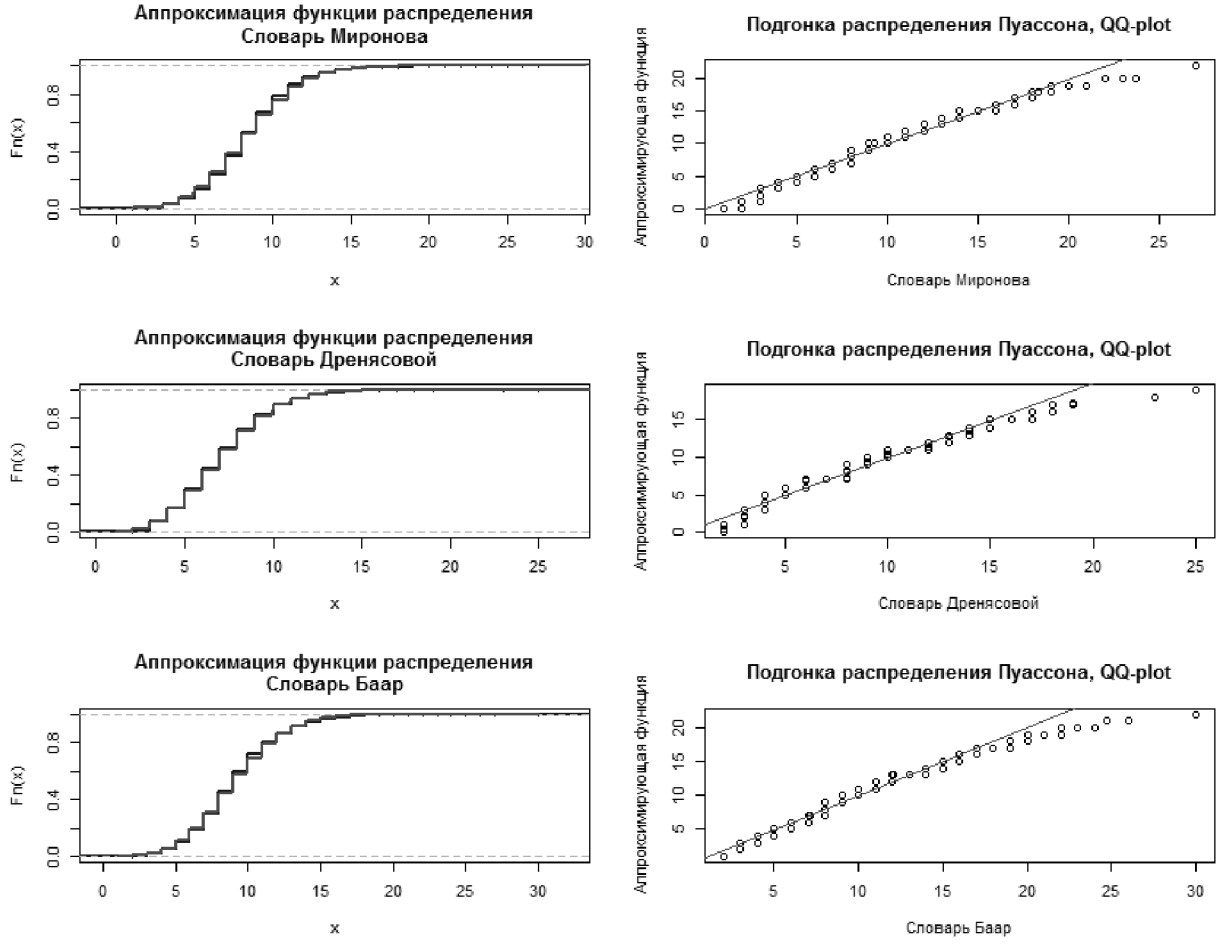

Рис. 3. Аппроксимация функций распределения
и подгонка распределений Пуассона по исследуемым словарям

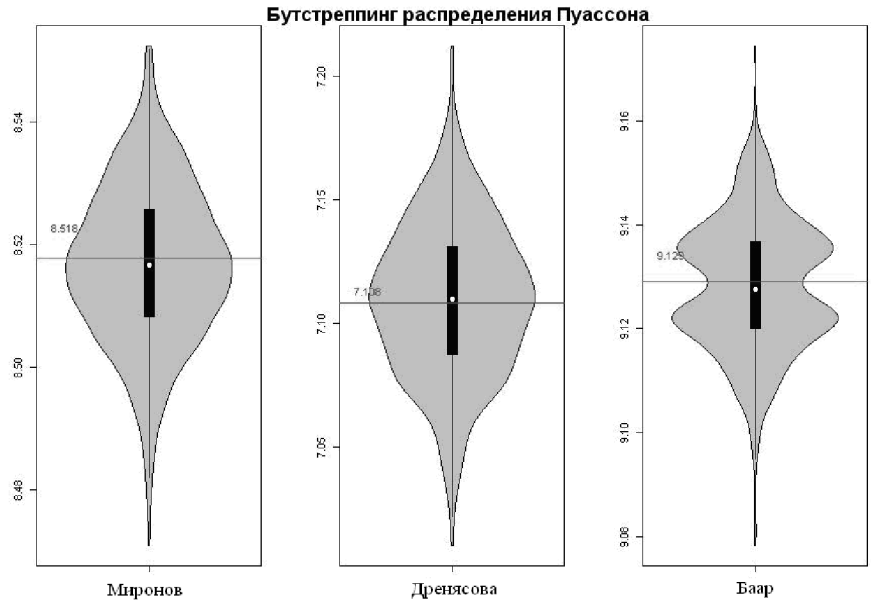

Рис. 4. Бутстреппинг величины λ в распределении Пуассона

```
> # Описание функции для проведения бутстреппинга
> DrawViol<-function(XX,names="Словарь"){
+ print(XX.fit<-fitdist(XX,"pois"))
+ XX.bw<-bootdist(XX.fit,niter=1000)
+ wvioplot(XX.bw$estim[[1]],names=names,adjust=1)
+ abline(h=XX.fit$estimate,col="red")
+ text(0.6,XX.fit$estimate+0.005,sprintf("%1.3f",XX.fit$estimate),col="red")
+ print(summary(XX.bw$estim[[1]]))
+ }
> old.par<-par(mfrow=c(1,3))
> DrawViol(LNidM,"Миронова")
Fitting of the distribution ' pois ' by maximum likelihood
Parameters:
 estimate Std. Error
lambda 8.517825 0.01263884
 Min. 1st Qu. Median Mean 3rd Qu. Max.
 8.477 8.509 8.518 8.518 8.527 8.559
> DrawViol(LNidD,"Дренясова")
Fitting of the distribution ' pois ' by maximum likelihood
Parameters:
 estimate Std. Error
lambda 7.108475 0.03104611
 Min. 1st Qu. Median Mean 3rd Qu. Max.
 7.004 7.089 7.110 7.109 7.131 7.209
> DrawViol(LNidB,"Баар")
Fitting of the distribution ' pois ' by maximum likelihood
Parameters:
 estimate Std. Error
lambda 9.129123 0.01270996
 Min. 1st Qu. Median Mean 3rd Qu. Max.
 9.089 9.121 9.128 9.129 9.137 9.176
> par(old.par)
> title("Бутстреппинг распределения Пуассона")
```

Таблица 2

Количество успешных выборок (p > 0,05) при использовании метода Монте-Карло

Успехов из 10000 выборок	Дренясова	Баар	Миронов
Дренясова	8043	8042	8039
Баар	8042	8040	7992
Миронов	8039	7992	7980

```
> library(VGAM)
> resXsq<-c()
> set.seed(500)
> # Повторяем цикл 10000 раз.
> for(k in 1:10000) {
+ xx<-sample(LNidB,2000,replace =T)-sample(LNidB,2000,replace=T)
+ yy<-rskellam(2000,9.131,9.131)
+ Xsq<-chisq.test(xx,yy)
+ resXsq<-c(resXsq,Xsq$p.value)
+ }
> length(subset(resXsq,resXsq>0.05)) library(VGAM)
[1] 8040
```

ЗАКЛЮЧЕНИЕ

Таким образом, на основании проведенного анализа построена математическая модель, выявляющая закономерность в распределении частот слов различной длины в зависимости от средней длины слова в анализируемых словарях посредством аппроксимации распределения Пуассона методом максимального правдоподобия. Верификация модели дала положительные результаты, равно как и верификация разности на основе распределения Скеллама, что говорит о правильности выдвинутой гипотезы. Полученные результаты являются новыми. Развитием данного исследования станет обработка аналогичных данных по остальным германским языкам и их сравнение с целью получения более репрезентативной выборки.

СПИСОК ЛИТЕРАТУРЫ

1. *Титов В.Т.* Общая квантитативная лексикология романских языков / В. Т. Титов. – Воронеж: Изд-во Воронеж. гос. ун-та, 2002. – 240 с.

2. *Титов В.Т.* Частная квантитативная лексикология романских языков: Монография / В. Т. Титов; Воронеж. гос. ун-т. – Воронеж: Изд-во Воронеж. гос. ун-та, 2004. – 552 с.

3. *Zipf G.K.* The Psycho-Biology of Language: an introduction to dynamic philology / Zipf G. K. – Cambridge: Mass. MIT Press, 1965. – 336 p.

4. *Берков В.П.* Современные германские языки / В. П. Берков. – М.: Астрель АСТ, 2001. – 336 с.

5. Большой нидерландско-русский словарь: Ок. 180 000 сл. и словосочетаний / С. А. Миронов, В. О. Белоусов, Л. С. Шечкова и др.; Под рук. С. А. Миронова. – 3-е изд., испр. – М.: Живой яз., 2006. – 916 с.

6. *Baar A.H., van den.* Groot Nederlands-Russisch Woordenboek / Большой голландско-русский словарь. – Amsterdam: Uitgeverij Pegasus, 2012. – 1265 p.

7. *Дренясова Т. Н., Миронов С. А.* Карманный нидерландско-русский словарь. Около 7000 слов. – М.: Русский язык, 1977. – 392 с.

8. A language and environment for statistical computing / R Development Core Team; R Foundation for Statistical Computing. – Vienna, Austria. – ISBN 3-900051-07-0. – URL http://www.R-project.org/ (дата обращения: 29.11.12)

9. Bossie Awards 2010: The best open source application development software/Andrew Binstock, Doug Dineley, Martin Heller, Rick Grehan. – URL: http://www.infoworld.com/d/open-source/bossie-awards-2010-the-best-open-source-application-development-software-140 (дата обращения: 29.11.12)

10. *Hintze J. L., Nelson R. D.* Violin Plots: A Box Plot-Density Trace Synergism / Jerry L. Hintze, Ray D. Nelson // The American Statistician. – 1998. – Vol. 52. – P. 181–184.

11. *Cullen A. C., Frey H. C.* Probabilistic Techniques in Exposure Assessment: A Handbook for Dealing with Variability and Uncertainty in Models and Inputs / Alison C. Cullen, H. Christopher Frey// Springer, 1999. – 352 p.

12. *Skellam J. G.* The frequency distribution of the difference between two Poisson variates belonging to different populations / Journal of the Royal Statistical Society, Series A, 1946. – p. 109, 296.

Воевудский Д. С. – аспирант кафедры теоретической и прикладной лингвистики. Воронежский государственный университет. E-mail: dimavoev@mail.ru

Тушавин В. А. – кандидат экономических наук, ассистент кафедры экономики и финансов. Санкт-Петербургский государственный университет аэрокосмического приборостроения. E-mail: tushavin@gmail.com

Voevudskiy D. S. – Postgraduate student of the Theoretical and Applied Linguistics Department. Voronezh State University. E-mail: dimavoev@mail.ru

Tushavin V. A. – Candidate of Economics, Teaching Assistant, Department of Economics and Finance. Saint-Petersburg State University of Aerospace Instrumentation. E-mail: tushavin@gmail.com