УДК 811.11

СТРАТИФИКАЦИОННЫЙ ЧАСТОТНЫЙ АНАЛИЗ НИДЕРЛАНДСКОЙ ЛЕКСИКИ (ПО ДАННЫМ НИДЕРЛАНДСКО-РУССКИХ СЛОВАРЕЙ)

Д. С. Воевудский*, В. А. Тушавин**

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

Поступила в редакцию 03.04.2014 г.

Аннотация. В статье проводится анализ распределения длины нидерландских слов в трех нидерландско-русских словарях с целью выявления его закономерностей в зависимости от объема словаря и частей речи. Исследуются взаимные пересечения анализируемых словарей и на этой основе проводится проверка существующего алгоритма выделения лексико-семантического ядра.
Ключевые слова: длина слов, нидерландский язык, лексико-семантическое ядро, распределение слов по длине, квантильная лексика.
Annotation. In the article the authors analyse the distribution of the length of words in three Dutch-Russian dictionaries in order to understand its correlations with on the dictionary size and the part of speech. Mutual intersections of the dictionaries studied are investigated, laying the basis to verify the existing algorithm of singling out the lexico-semantic nucleus.
Keywords: length of words, the Dutch language, lexico-semantic nucleus, distribution of words by length, quantile lexemes.

Целью предлагаемого исследования является выявление закономерностей в распределении частей речи и длины слов в звуках для словарей с различным количеством словарных статей.

Актуальность исследования определяется необходимостью детального изучения и верификации существующих методик выделения лексико-семантического ядра с использованием словарей различной размерности.

Для достижения поставленной цели были решены следующие задачи: 1) оцифровка словарей и создание электронных баз; 2) анализ пересечений словарных статей с помощью SQL; 3) проверка статистических гипотез с помощью критерия согласия хи-квадрат.

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 [1]. 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 [2, с. 74–75]. Для анализа были взяты три нидерландско-русских словаря различного размера [3–5]. Используя язык запросов, были построены пересечения словарей по словарным статьям. В данном случае под словарной статьей понимается уникальная совокупность описываемого слова и части речи, к которой оно принадлежит. Например, слова *aanbreken* ‘открыть’ и *aanbreken* ‘рассвет’ для целей последующего анализа являются различными. Результаты представлены в табл. 1. В отличие от [6], в данном анализе используются все части речи, поэтому количество словарных статей может отличаться.

Ниже дается таблица распределения нидерландских слов по длине, где B – множество слов в словаре Баара [4], M – множество

Таблица 1

Распределение нидерландских слов по длине (по данным двуязычных словарей)

Число звуков	B	M	D	$B \cap D \cap M$	$B \cap M \setminus D$	$B \cap D \setminus M$	$M \cap D \setminus B$	$B \setminus (M \cup D)$	$D \setminus (B \cup M)$	$M \setminus (B \cup D)$	$B \cup M \cup D$
1	3	4	2	1	1			1	1	2	6
2	124	158	65	44	46	4	10	30	7	58	199
3	1068	1522	590	464	387	29	60	188	37	611	1776
4	1864	2315	808	678	790	22	80	374	28	767	2739
5	3329	3885	1148	977	1603	29	97	720	45	1208	4679
6	4819	5520	1134	963	2550	20	88	1286	63	1919	6889
7	6940	7593	1131	954	3686	28	84	2272	65	2869	9958
8	8757	9089	1055	906	4672	17	70	3162	62	3441	12330
9	8418	8063	732	605	4211	16	66	3586	45	3181	11710
10	6896	6026	551	447	3090	11	57	3348	36	2432	9421
11	5181	4165	334	255	2115	6	35	2805	38	1760	7014
12	3639	2679	178	139	1317	7	20	2176	12	1203	4874
13	2693	1803	103	72	908	2	13	1711	16	810	3532
14	1810	1092	63	45	551	1	4	1213	13	492	2319
15	1209	664	36	22	299	1	6	887	7	337	1559
16	731	359	14	11	175		1	545	2	172	906
17	483	215	8	5	96		2	382	1	112	598
18	290	101	3	3	57			230		41	331
19	163	75	6	3	27	1		132	2	45	210
20	87	31			14			73		17	104
21	65	18			8			57		10	75
22	33	14			3			30		11	44
23	17	5	1	1	2			14		2	19
24	7	2			1			6		1	8
25	5	4	1		1			4	1	3	9
26	2							2			2
27	4	1			1			3			4
30	1							1			1
Итого	**58638**	**55403**	**7963**	**6595**	**26611**	**194**	**693**	**25238**	**481**	**21504**	**81316**
Min	1	1	1	1	1	2	2	1	1	1	1
1Qt	7	7	5	5	7	4	5	8	6	7	7
2Qt	9	8	7	7	8	6	7	10	7	8	9
3Qt	11	10	9	8	10	8	9	12	10	10	11
Max	30	27	25	23	27	19	17	30	25	25	30
Средн.	9,1	8,4	6,9	6,9	8,6	6,5	7,0	10,1	7,7	8,7	8,9

слов в словаре Миронова [3], D – множество слов в словаре Дренясовой [5], остальные колонки соответствуют различным производным множествам из этих словарей).

Распределение по частям речи по тем же множествам слов представлено в табл. 2.

Полученное в результате распределение слов по частям речи и их длине (для каждо-

Таблица 2

Распределение нидерландских слов по частям речи (по данным двуязычных словарей)

Часть речи	B	M	D	$B \cap D \cap M$	$B \cap M \setminus D$	$B \cap D \setminus M$	$M \cap D \setminus B$	$B \setminus (M \cup D)$	$D \setminus (B \cup M)$	$M \setminus (B \cup D)$	$B \cup M \cup D$
a	8485	7774	1202	945	4155	41	116	3344	100	2558	11259
adv	1544	1505	383	212	526	35	75	771	61	692	2372
art	3	3	6	3					3		6
cj	74	82	31	14	22	2	9	36	6	37	126
interj	180	184	6	2	71	3		104	1	111	292
n	39339	36551	4427	3776	16908	86	347	18569	218	15520	55424
num	134	87	35	26	39	1	4	68	4	18	160
parenth	8	3	15			2		6	13	3	24
part	2	9	4	1			1	1	2	7	12
prep	72	73	35	19	14	3	5	36	8	35	120
pron	109	135	73	52	22	3	9	32	9	52	179
v	8688	8997	1746	1545	4854	18	127	2271	56	2471	11342
Итого	**58638**	**55403**	**7963**	**6595**	**26611**	**194**	**693**	**25238**	**481**	**21504**	**81316**

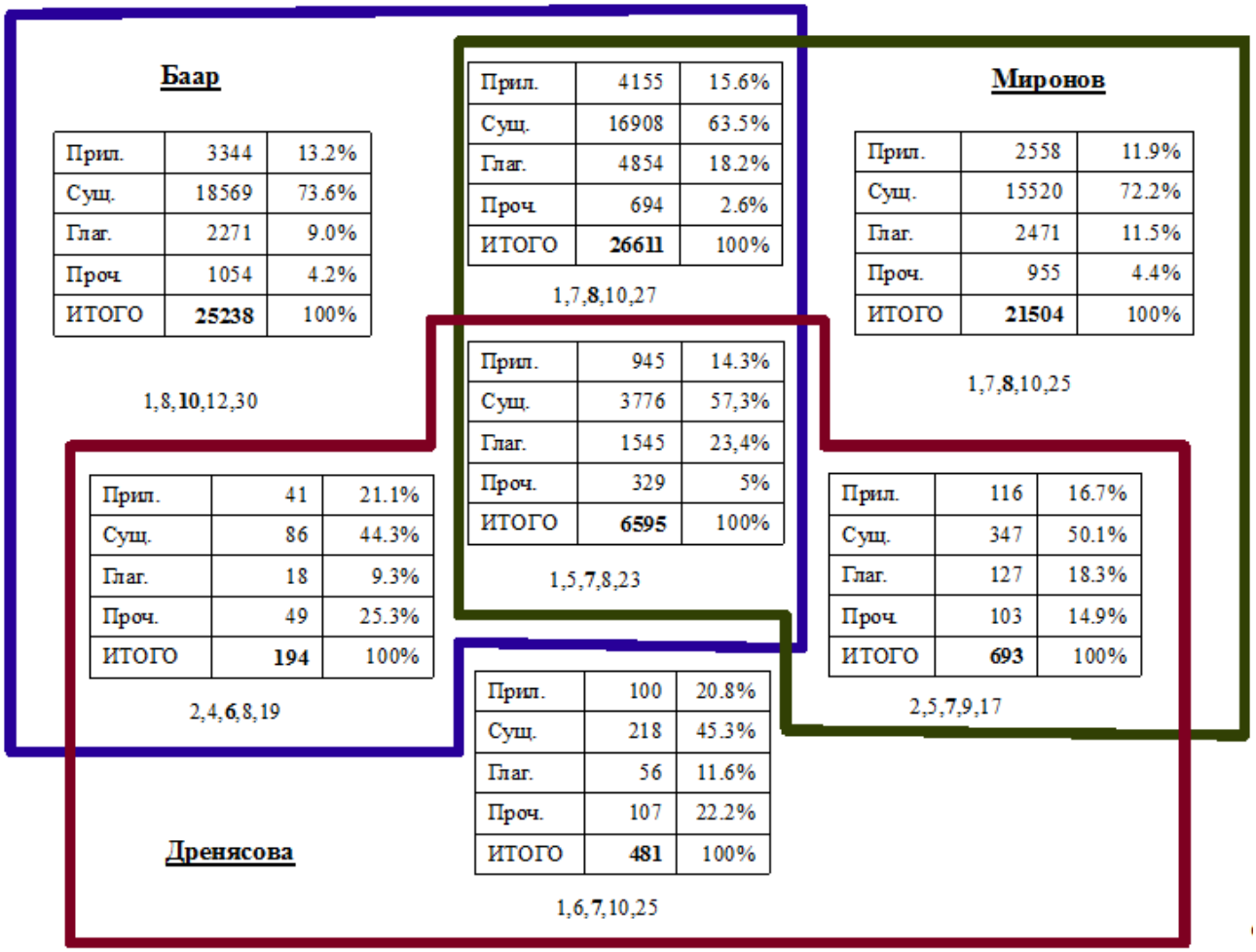

Рис. 1. Анализ пересечения трех словарей (Дренясовой, Миронова и Баара) по частям речи и длине слов с использованием диаграммы Венна

го из возможных пересечений) также можно представить в виде диаграммы Венна (рис. 1).

На рисунке для каждого пересечения в таблице приводится количественный и долевой показатели по частям речи, цифры, приведенные под таблицей, описывают распределение длины слов для каждого пересечения (слева направо: минимум, первый квартиль, медиана, третий квартиль, максимум).

Как видно из рисунка, слов, входящих во все три словаря, насчитывается не так много – около 6,5 тысяч, что может быть легко объяснено малым объемом словаря Дренясовой (около 8 тысяч слов). Если в словарях большого объема значительная доля лексики является уникальной и более не присутствует ни в одном из других словарей, то в словаре Дренясовой подобная лексика, наоборот, составляет меньшую долю словаря (около 500 слов).

Проверка гипотезы о различии пропорций частей речи для каждого из пересечений осуществлена с помощью критерия согласия Пирсона (хи-квадрат) в статистическом пакете GNU R [7, 8]. Результаты проверки показали, что гипотеза H_0 об эквивалентности пропорций отвергается (p-значение менее 2.2×10^{-16} для всех тестов).

На рис. 1 показаны медианы для каждой выборки. В данном случае выбраны медианы, а не средние значения, поскольку медианы более робастны к выбросам. Поскольку, как это было показано ранее [9], распределение длины слов в звуках в словарях не имеет нормальную природу, а выборки имеют выбросы, то для проверки статистической гипотезы о равенстве медиан был использован тест Муда (Mood's median test) [7]. Гипотеза о равенстве медиан для выборок, соответствующих уникальным словам в каждом словаре $B\setminus(M\cup D)$, $D\setminus(B\cup M)$ и $M\setminus(B\cup D)$, медиане «микрословаря» (пересечения всех трех словарей $B\cap D\cap M$) отвергается в каждом из трех тестов:

– Дренясова-Микрословарь: Z статистика = 7.127, p-значение = 1.026×10^{-12};

– Баар-Микрословарь: Z статистика = –30.8032, p-значение < 2.2×10^{-16};

– Миронов-Микрословарь: Z статистика = –11.3077, p-значение < 2.2×10^{-16}.

Проанализируем, насколько отличается распределение по длине слов в зависимости от части речи для существительных и прилагательных каждого из этих множеств, используя совмещенный график размахов и ядерной плотности Beanplot [10], как это представлено на рис. 2.

Горизонтальными линиями обозначены средние значения страт. Пунктирная линия среднюю длину слова в звуках для всего представленного множества слов.

Визуализация распределений данных по стратам показывает, что распределение существительных и глаголов различается. Проверка, проведенная с помощью критерия согласия Пирсона о соответствии распределения существительных, прилагательных и глаголов ($\chi^2 = 2426.333$, $df = 54$, p-значение < 2.2×10^{-16}), отвергает гипотезу о равенстве этих распределении. Средняя длина слов в словаре $B\cup M\cup D$ для существительных, прилагательных и глаголов, соответственно, равна 9.16, 8.62 и 8,66 звуков. Следует также отметить, что доля количество слов большой длины (более 12 звуков) среди глаголов в объединенном словаре значительно меньше по сравнению с существительными (5 и 15 процентов соответственно). Это можно объяснить как общим количественным преобладанием существительных над глаголами, так и тем, что слова большой длины представляют собой преимущественно слова-композиты, интернационализмы и терминологическую лексику, которые в основном являются именно существительными.

Полученный микрословарь, являющийся пересечением трех словарей, позволяет проверить качество работы алгоритма выбора лексико-семантического ядра языка. Для этого в каждом базовом словаре был взят первый дециль по традиционно выбираемым для выделения лексико-семантического ядра параметрам – длине слов в звуках, числу синонимов, числу значений слова и идиом [11, 12, 13]. После чего были выбраны слова, входящие как минимум в две из этих выборок. Иными словами, если обозначить эти выборки как X_1, X_2, X_3 и X_4, то результирующую

выборку для словаря i можно представить в виде

$$S_i = X_1 \cap X_2 \cup X_1 \cap X_3 \cup X_1 \cap X_4 \cup \cup X_2 \cap X_3 \cup X_2 \cap X_4 \cup X_3 \cup X_4.$$

Вернемся к табл. 2. Из нее видно, что общее количество существительных, прилагательных и глаголов для микрословаря $|(B \cap D \cap M)| = 6266$ слов. Число слов в словаре Баара не входящих в микрословарь $|(B \setminus B \cap D \cap M)| = 50246$, в словаре Миронова $|(M \setminus B \cap D \cap M)| = 47056$, в словаре Дренясовой $|(D \setminus B \cap D \cap M)| = 1109$. Очевидно, что большая часть слов словаря Дренясовой (83 %) входит в микрословарь, поэтому тестирование таким способом алгоритма будет непродуктивно.

Вероятность случайного выбора из словаря заданного количества слов, входящих в микрословарь, для данного размера выборки описывается функцией гипергеометрического распределения:

$$f(k, N, D, n) = \frac{\binom{D}{k}\binom{N-D}{d-k}}{\binom{N}{n}}, \quad (1)$$

где N – общее число слов в словаре; D – число слов входящих из него в микрословарь; n – размер выборки; k – число совпадений. Математическое ожидание рассчитывается по формуле:

$$\mu = \frac{nD}{N}. \quad (2)$$

Произведенные результаты расчетов представлены в табл. 3.

Как видно из таблицы, несмотря на незначительное число совпадений слов, полученных с использованием алгоритма выборки, с микрословарем (14 и 10 процентов соответственно), вероятность получить такой результат случайным образом пренебрежительно мала. Неслучайность в работе алгоритма видна на рисунке 3, на котором показана функ-

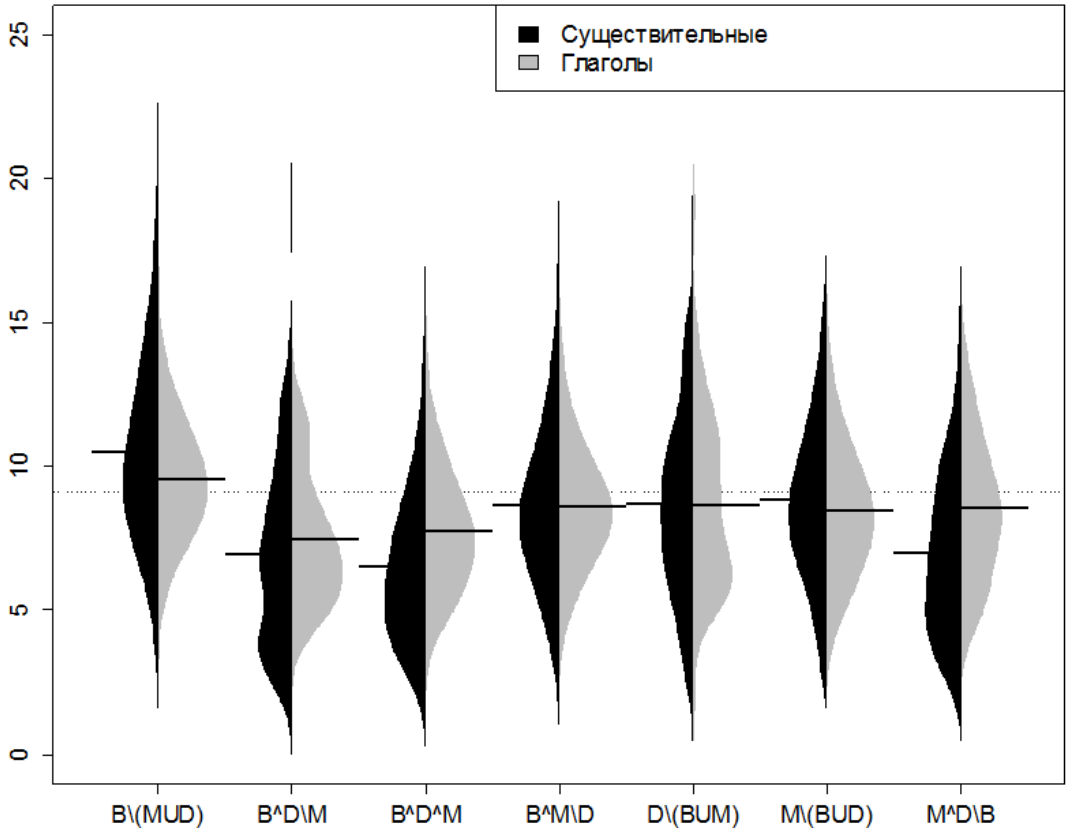

Рис. 2. Анализ плотности распределения слов по длине (существительных и глаголов) для множеств возможных пересечений трех словарей (Дренясовой (D), Миронова (M) и Баара (B)) в диапазоне длины в звуках (1–25).

Таблица 3

Анализ эффективности алгоритма выделения лексического ядра

Словарь	Размер выборки по алгоритму	Число совпадений	Математическое ожидание случайной выборки	p -значение случайного совпадения
Баар	5686	850	631	2×10^{-21}
Миронов	5084	644	597	0.016

ция гипергеометрического распределения для данной комбинации параметров (серым), а также результат, полученный с использованием описанного выше алгоритма выборки.

Штрих-пунктирной утолщенной линией показано число совпадений в полученном ядре с пересечением трех словарей, сплошной толстой черной линей – математическое ожидание случайной выборки.

Таким образом, при использовании заданного алгоритма мы получаем преимущество по сравнению со случайной выборкой методом Монте-Карло. Однако полученный результат вызывает вопрос о качестве используемого алгоритма. Оценим его с помощью каппа-статистики Флейса [14].

Пусть существует некое ядро языка Y. Тогда, очевидно, что $Y \subseteq B \cap M \cap D$, а выборки, полученные на предыдущем этапе с помощью алгоритма, в идеальном случае должны описываться формулами $S_B \subseteq Y$, $S_M \subseteq Y$, $S_D \subseteq Y$. Мы же проверим качество оценки, сравнив между собою оценку принадлежности слов микрословаря к ядру Y, используя в качестве критерия их принадлежность к множествам S_B, S_M и S_D. Анализ согласия оценивается с помощью каппа-статистики Флейса для всех трех выборок. Всего пересечение словарей содержит 6266 слов. Из них оценка принадлежности или непринадлежности к ядру полностью совпадает в 5169 случаях во всех трех выборках, что составля-

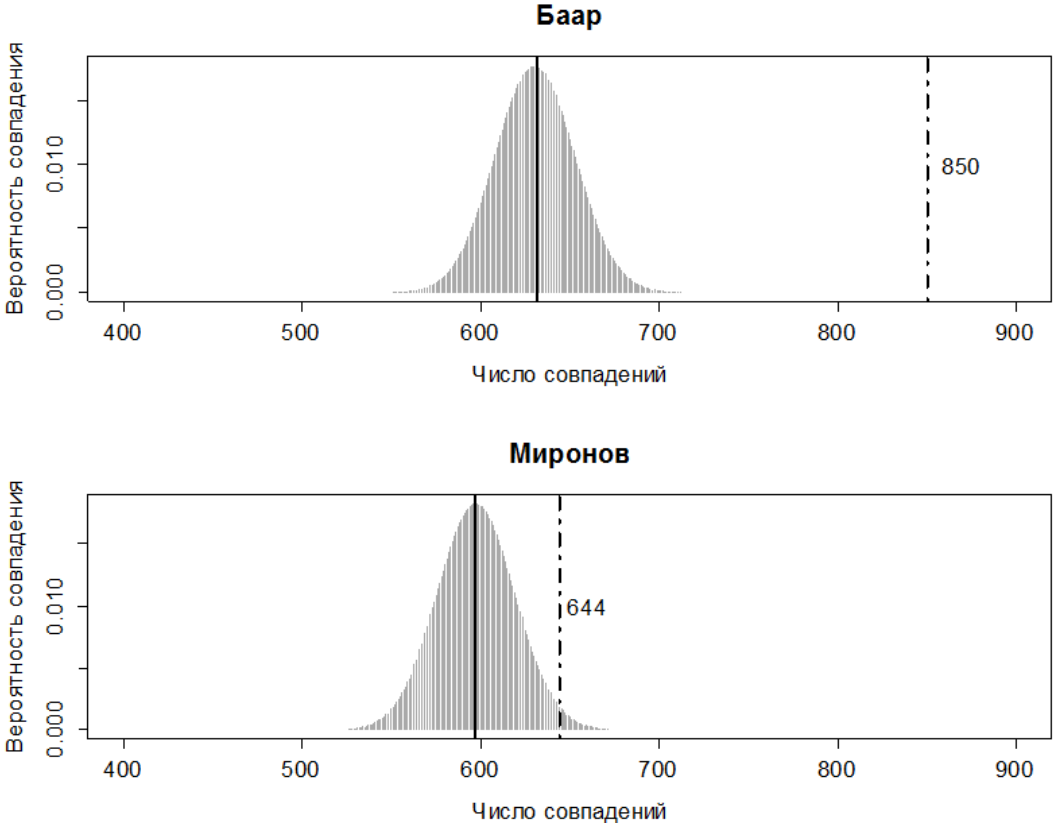

Рис. 3. Анализ работы алгоритма выделения ядра языка в сравнении со случайной выборкой

Таблица 4

Анализ согласованности выборок для трех словарей

Результат	Kappa	SE Kappa	*Z*	*P* (*vs* > 0)
Не ядро	0.401936	0.0072936	55.1078	0.0000
Ядро	0.401936	0.0072936	55.1078	0.0000

ет 82.49 % (95 % доверительный интервал: 81.53; 83.43). Каппа-статистика Флейса приведена в табл. 4.

Как видно из таблицы, коэффициент каппа незначительно превышает 0.4 при p-значении, не превосходящим уровень альфа 0.05, что свидетельствует как о применимости оценки в целом, так и необходимости её дальнейшей доработки [14].

Таким образом, на основании проведенного анализа подтверждена зависимость длины слова в звуках от частоты его использования на примере достаточно значительной выборки. Верифицирован алгоритм выделения ядра языка на словарях большого объема. Полученные результаты являются новыми. В качестве прикладного значения результатов исследования следует отметить проведенную работу по частотному анализу словарных статей трех словарей нидерландского языка, представленную в табл. 1 и 2, которые могут быть полезны другим исследователям. Помимо этого можно также отметить потенциальную возможность создания метасловаря, объединяющего в себя все словарные статьи, однако для её решения необходимо разработать механизм, позволяющие объединять словарные статьи в автоматическом режиме с учетом возможных опечаток. Развитием данного исследования станет дальнейшая оцифровка словарей, обработка аналогичных данных по остальным германским языкам и их анализ с целью разработки более совершенного алгоритма выделения ядра языка.

СПИСОК ЛИТЕРАТУРЫ

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

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

3. 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.

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

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

6. *Воевудский Д.С.* Парадигматическая стратификация лексики нидерландского языка. – Вестник Воронежского государственного университета. Серия: Лингвистика и межкультурная коммуникация. 2013. № 1. С. 111–114.

7. *Lewis N.D.* 100 Statistical Tests in R. – Heather Hills Press, 2013. – 496 p.

8. R Core Team. R: A language and environment for statistical computing / R Foundation for Statistical Computing. – Vienna, Austria. – URL http://www.R-project.org/ (дата обращения 31.03.14)

9. *Воевудский Д.С., Тушавин В.А.* Статистическая обработка лингвистических данных нидерландско-русских словарей. – Вестник Воронежского государственного университета. Серия: Системный анализ и информационные технологии. 2013. № 1. С. 169–176.

10. *Kampstra P.* Beanplot: A Boxplot Alternative for Visual Comparison of Distributions. / Journal of Statistical Software, Code Snippets 28(1). 1-9, 2008. – URL http://www.jstatsoft.org/v28/c01/

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

12. *Титов В.Т.* Частная квантитативная лексикология романских языков: Моногра-

фия / В.Т. Титов; Воронеж. гос. ун-т. – Воронеж : Изд-во Воронеж. гос. ун-та, 2004. – 552 с.

13. *Воевудская О.М., Воевудский Д.С.* Язык идиш на фоне германских языков и языков Восточной Европы. – Вестник Воронежского государственного университета. Серия: Лингвистика и межкультурная коммуникация. 2013. № 1. С. 188–195.

14. *Fleiss J.L., Cohen J.* The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. – Educational and Psychological Measurement, 1973. – Vol. 33. – pp. 613–619.

Воевудский Д.С. – кандидат филологических наук, методист Научно-методического центра компьютерной лингвистики при кафедре теоретической и прикладной лингвистики Воронежского государственного университета.
Тел.: 8-906-581-69-31. E-mail: dimavoev@mail.ru

Тушавин В.А. – кандидат технических наук, кандидат экономических наук, доцент кафедры инноватики и управления качеством Санкт-Петербургского государственного университета аэрокосмического приборостроения. Тел.: 8-911-720-43-43.
E-mail: tushavin@gmail.com

Voevudskiy D.S. – Candidate of Philology, methodologist of Scientific centre of computational linguistics of the Theoretical and Applied Linguistics Department of Voronezh State University. Phone: 8-906-581-69-31
E-mail: dimavoev@mail.ru

Tushavin V.A. – Candidate of Technology, Candidate of Economics, Associate Professor, Department of Innovation and Quality Management of Saint-Petersburg State University of Aerospace Instrumentation. Phone: 8-911-720-43-43
E-mail: tushavin@gmail.com