УДК 004.912:519.816

М. А. Зильберглейт, доктор химических наук, профессор, заведующий кафедрой (БГТУ);
А. С. Малюкевич, аспирант (БГТУ)

ОЦЕНКА ВОЗМОЖНОСТИ ИСПОЛЬЗОВАНИЯ МОРФОЛОГИЧЕСКОГО АНАЛИЗА ТЕКСТА ДЛЯ ВЫДЕЛЕНИЯ СТИЛЕЙ

В статье дается оценка возможности использования морфологического анализа текста для выделения стилей, приводятся результаты статистического анализа по установлению равноточности и неравноточности средних значений, основанного на использовании критерия Фишера и критерия Стьюдента. С помощью метода кластерного анализа определено минимальное количество морфологических характеристик, позволяющих установить стиль того или иного текста сети. Результаты проведенного исследования свидетельствуют о том, что для выделения стиля конкретного текстового фрагмента анализу могут быть подвергнуты не все морфологические характеристики текста, а лишь некоторые их сочетания (комбинации).

There is given in article the assessment of possibility of use of the morphological analysis of the text for allocation of styles, results of the statistical analysis on establishment of a equalexact and a unequalexact of average values on the basis of use of criterion of Fischer and criterion of Student. With the help of a method of the Cluster analysis is defined the minimum quantity of morphological characteristics which allow to establish style of this or that text of a network. Results of the carried-out research testify that for allocation of style of a concrete text fragment all morphological characteristics of the text, and only their some combinations (combinations) can be subjected to the analysis not.

Введение. Применение различных подходов к автоматизированной обработке текста с целью выявления его стилистических особенностей, атрибуции, анализа удобочитаемости и авторизации получило в последнее время широкое распространение. С помощью данных подходов решаются также такие прикладные задачи, как автоматическое индексирование и реферирование текстов, извлечение текстовой информации, статистическая обработка, машинный перевод, многоязыковая генерация, автоматизированная оценка трудности, извлечение знаний из больших массивов информации, интеллектуальный поиск и установление авторства текстов.

Для успешной автоматической обработки исходного материала исследователю необходимо определить параметры, с помощью которых будет проведен соответствующий анализ. Известно более 30 параметров, используемых для характеристики текста, к которым также относятся морфологические, синтаксические и лексические [1].

Эффективность научно-исследовательских работ напрямую зависит от качества информационного обеспечения, а поиск информации является ключевым этапом любого научного исследования [2]. В связи с тем, что рост научного знания предполагает и увеличение количества предлагаемого научным работникам материала, то вопрос поиска действительно ценной и полезной для исследователя информации является весьма актуальным.

Интернет стал одним из самых популярных источников информации, однако выделить среди множества результатов запроса документ, который будет соответствовать всем поставленным требованиям, достаточно тяжело. Обычно выборку на запрос составляют тысячи страниц виртуального пространства, а заданным требованиям отвечают лишь единицы, именно поэтому разработка методов повышения эффективности поиска научной информации в сети Интернет является сегодня одной из ключевых задач развития информационных технологий.

В статье «Морфологический строй функциональных стилей (на материале документов Internet)» [3] представлены результаты исследования по выявлению основных морфологических характеристик, анализ которых позволяет определить стиль текста в сети Интернет. За основу автором была взята функционально-стилевая концепция, опытный массив данных для проведения исследования составил 305 текстов, среди них к официально-деловому стилю было отнесено 50 законов, научному – 54 текста, публицистическому – 61 статья, художественному – 79 рассказов, разговорному – 61 фрагмент.

Используя возможности модуля Linguist, в качестве самостоятельных морфологических классов были выделены: существительные (+ имена собственные, отчества, фамилии, географические названия, аббревиатуры), прилагательные, местоимения, числительные, наречия, глаголы, причастия, деепричастия, предлоги, союзы, частицы, междометия и прочие (предикативные или вводные слова). Всего было обработано 239 696 слов.

Результат одновременного рассмотрения пяти стилей заключался в том, что был установлен монотонный рост средних долей существительных и прилагательных и монотонное же уменьшение

долей местоимений, наречий, глаголов и частиц от разговорного к официально-деловому стилю. На основании этого была выдвинута гипотеза о том, что морфологические параметры принадлежат к важнейшим маркерам функционального стиля и поэтому могут быть использованы для автоматической классификации текстов по стилям [3].

К сожалению, в работе не приведен статистический анализ полученных результатов, а делается лишь утверждение о возможности использования морфологических признаков в качестве основополагающих для классификации.

Основная часть. Цель работы заключается в проведении статистического анализа результатов работы [3], а также нахождении минимального количества морфологических характеристик, которые позволяют выделить тот или иной стиль речи на основании анализа исходного текста.

Статистический анализ заключался в установлении равноточности данных и сравнительном анализе средних значений.

В качестве соответствующих методов расчета применялся статистический анализ с использованием критериев Фишера, Стьюдента [4]. Для определения минимального количества морфологических характеристик — метод кластерного анализа (построение дендрограмм в программе StatGraphics Plus 5.1).

В табл. 1 представлены исходные средние значения ($x_{ср}$) и стандартное отклонение (s) [3] употребления каждой морфологической характеристики для того или иного стиля.

Для установления равноточности средних значений был проведен сравнительный анализ каждой из стилевых пар: разговорный – художественный, разговорный – публицистический, разговорный – научный, разговорный – официально-деловой, художественный – публицистический, художественный – научный, художественный – официально-деловой, публицистический – научный, публицистический – официально-деловой, научный – официально-деловой – для существительных, прилагательных, местоимений, числительных, наречий, глаголов, причастий, деепричастий, предлогов, союзов, частиц, междометий и прочих. Критерий Фишера вычислен по формуле

$$F_{эмп} = \frac{s_x^2}{s_y^2},$$

где s^2 — дисперсии сравниваемых стилей.

Количество степеней свободы df_1 и df_2 определено по формулам:

$$df_1 = n_2 - 1;\ df_2 = n_2 - 1.$$

После сравнения полученных значений с табличными данными, при условии $\alpha = 0{,}95$, было установлено, что не все средние значения неравноточны (отличаются), следовательно, гипотеза о равенстве всех дисперсий сравниваемых совокупностей отвергается.

В табл. 2 приведены результаты проведенных расчетов, где полужирным начертанием выделены значения, которые превышают $F_{эмп}$, следовательно, данные результаты можно считать неравноточными, то есть гипотеза о равенстве дисперсий сравниваемых совокупностей отвергается. Значения остальных ячеек меньше табличных, то есть в данном случае гипотеза о равенстве дисперсий принимается.

При вычислении статистики для сравнения двух средних значений при неизвестных, но равных генеральных дисперсиях была использована формула:

Таблица 1

Исходные значения средних и стандартного отклонения

Часть речи	Стиль									
	разговорный		художественный		публицистический		научный		официально-деловой	
	$X_{ср}$	S	$X_{ср}$	S	$X_{ср}$	S	$X_{ср}$	S	$X_{ср}$	S
Существительное	0,194	0,040	0,243	0,049	0,335	0,034	0,396	0,054	0,497	0,037
Прилагательное	0,000	0,000	0,063	0,020	0,107	0,024	0,130	0,028	0,184	0,048
Местоимение	0,161	0,027	0,126	0,039	0,075	0,019	0,047	0,013	0,029	0,011
Числительное	0,002	0,002	0,006	0,004	0,007	0,005	0,005	0,004	0,009	0,012
Наречие	0,068	0,017	0,065	0,017	0,049	0,012	0,029	0,016	0,008	0,007
Глагол	0,167	0,024	0,162	0,027	0,120	0,019	0,090	0,020	0,048	0,018
Причастие	0,028	0,011	0,045	0,013	0,066	0,017	0,091	0,021	0,091	0,023
Деепричастие	0,006	0,007	0,013	0,007	0,009	0,005	0,017	0,010	0,005	0,005
Предлог	0,051	0,013	0,055	0,010	0,058	0,013	0,061	0,015	0,046	0,020
Союз	0,050	0,013	0,037	0,011	0,038	0,008	0,033	0,022	0,009	0,008
Частица	0,210	0,031	0,158	0,030	0,130	0,029	0,090	0,022	0,071	0,019
Междометие	0,016	0,009	0,003	0,003	0,000	0,001	0,001	0,004	0,000	0,000
Прочие	0,013	0,008	0,006	0,003	0,007	0,004	0,008	0,006	0,002	0,004

$$\hat{t} = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\left[\frac{n_1 + n_2}{n_1 \cdot n_2}\right] \cdot \left[\frac{(n_1 - 1) \cdot s_1^2 + (n_2 - 1) \cdot s_2^2}{n_1 + n_2 - 2}\right]}},$$

где x_1 и x_2 — средние значения; n_1 и n_2 – количество выборок; s_1^2 и s_2^2 — дисперсии для первого и второго параметра соответственно; (n_1 + + n_2 – 2) — число степеней свободы.

При вычислении статистики для сравнения двух средних значений при неизвестных и неравных генеральных дисперсиях была применена формула

$$\hat{t} = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\left[\frac{s_1^2}{n_1}\right] + \left[\frac{s_2^2}{n_2}\right]}},$$

где x_1 и x_2 — средние значения; n_1 и n_2 — количество выборок; s_1^2 и s_2^2 — дисперсии для первого и второго параметра соответственно; (n_1 + + n_2) — число степеней свободы.

Значения рассчитанной статистики для равноточных и неравноточных дисперсий представлены в табл. 3.

Таблица 2

Сравнение средних по критерию Фишера

Часть речи	Стиль									
	Разговор. + + худож.	Разговор. + + публиц.	Разговор. + + науч.	Разговор. + + офиц.-дел.	Худож. + + публиц.	Худож. + + науч.	Публиц. + + науч.	Публиц. + + офиц.-дел.	Науч. + + офиц.-дел.	Худож. + + офиц.-дел.
Существительное	**1,50**	1,38	**1,82**	1,17	**2,07**	1,21	**1,75**	**2,52**	1,18	**2,13**
Прилагательное	—	—	—	—	**1,44**	**1,96**	**5,76**	1,35	**4,00**	**2,93**
Местоимение	**2,08**	**2,02**	**4,31**	**6,01**	**4,21**	**9,00**	**12,58**	**2,13**	**2,98**	1,39
Числительное	**4,00**	**6,25**	**4,00**	**36,00**	**1,56**	1,00	**9,00**	**1,56**	**5,76**	**9,00**
Наречие	1	**2,01**	1,13	**5,90**	**2,01**	1,13	1,00	**1,77**	**2,94**	**5,22**
Глагол	1,27	**1,60**	**1,44**	**1,77**	**2,02**	**1,82**	**2,25**	1,11	1,11	1,23
Причастие	1,40	**2,39**	**3,64**	**4,37**	**1,71**	**2,61**	**3,13**	**1,53**	**1,83**	1,20
Деепричастие	1,00	**1,96**	**5,90**	**1,96**	**1,96**	**5,90**	**1,96**	**11,56**	1,00	**11,56**
Предлог	**1,69**	1,00	1,33	**2,37**	**1,69**	**2,25**	**4,00**	1,33	**2,37**	**1,78**
Союз	1,40	**2,64**	**2,86**	**2,64**	**1,89**	**4,00**	**1,89**	**7,56**	1,00	**7,56**
Частица	1,07	1,14	**1,99**	**2,66**	1,07	**1,85**	**2,49**	**1,74**	**2,33**	1,34
Междометие	**9,00**	**81,00**	**5,06**	—	**9,00**	**1,77**	—	**16,00**	—	—
Прочие	**7,09**	**4,00**	**1,78**	**4,00**	**1,77**	**4,00**	**2,25**	1,00	**2,25**	**1,78**

Таблица 3

Значения *t*-статистики для равноточных и неравноточных дисперсий

Часть речи	Стиль									
	Разговор. + + худож.	Разговор. + + публиц.	Разговор. + + науч.	Разговор. + + офиц.-дел.	Худож. + + публиц.	Худож. + + науч.	Публиц. + + науч.	Публиц. + + офиц.-дел.	Науч. + + офиц.-дел.	Худож. + + офиц.-дел.
Существительное	**7,19**	0,34	**22,55**	0,82	**13,10**	0,31	**33,42**	**7,14**	0,48	**11,20**
Прилагательное	0,31	0,57	0,67	0,60	**11,55**	**17,51**	**16,92**	0,09	**10,33**	**6,94**
Местоимение	**6,27**	**20,34**	**29,36**	**34,82**	**10,87**	**16,70**	**20,84**	**9,31**	**15,93**	0,15
Числительное	**7,73**	**7,25**	**4,99**	**4,08**	1,28	0,02	1,71	**2,38**	1,10	**2,24**
Наречие	0,01	**7,13**	0,23	**25,09**	**6,52**	0,22	0,45	**7,51**	**22,43**	**8,78**
Глагол	0,01	**11,99**	**19,17**	**29,82**	**10,79**	**17,65**	**28,76**	0,15	0,41	0,22
Причастие	0,10	**14,66**	**19,77**	**17,77**	**8,01**	**14,33**	**12,90**	**6,96**	**6,39**	0
Деепричастие	0,07	**2,72**	**4,43**	0,88	**3,94**	1,64	**7,56**	**3,33**	0,08	**4,96**
Предлог	**1,99**	0,05	0,05	1,52	1,49	**2,57**	**2,96**	0,02	**3,66**	**4,30**
Союз	0,08	**6,14**	**4,96**	**20,37**	0,62	1,23	**16,70**	**9,17**	0,38	**7,50**
Частица	0,127	0,241	**24,14**	**29,00**	0,09	**15,07**	**20,17**	**8,39**	**12,87**	0,094
Междометие	**10,83**	**13,80**	**11,77**	0,25	**8,89**	**3,12**	0,14	1,79	0	0,04
Прочие	**6,49**	**5,24**	**3,82**	**9,40**	1,63	**2,26**	1,04	0,13	**6,04**	**6,07**

Полужирным начертанием выделены ячейки, значения которых меньше $t_{табл}$, следовательно, нульгипотеза для них принимается, в противном случае — отклоняется.

По результатам проведенных расчетов, основанных на использовании критерия Фишера и критерия Стьюдента для равноточных и неравноточных дисперсий, можно сделать вывод о том, что существуют средние значения стилей, различие между которыми статистически незначимо.

Цель второго этапа работы заключалась в определении с помощью методов кластерного анализа основных морфологических характеристик, которые имеют важное значение при установлении стиля отобранного материала.

На базе заданных средних значений в программа Statgraphics Plus 5.1 выполнено построение дендрограмм, на которых были отображены наиболее близкие значения стилей при задании тех или иных морфологических характеристик.

В качестве основного метода построения кластеров был выбран метод ближайшего соседа (Nearest Neighbor), мерой близости определены расстояние Евклида (Euclidean), квадрат расстояния Евклида (Square Euclidean) и мера Хэмминга (City Block), заданное минимальное количество кластеров — 2.

Для установления первоначальных данных о группировке стилей на основе исходных средних значений нами была построена дендрограмма (рис. 1). На ней четко выделено объединение художественного и разговорного стилей в один кластер, а также научного, официально-делового и публицистического в другой.

Исходная классификация требует внесения изменений — перенос публицистического стиля из кластера научный/официально-деловой в художественный/разговорный. Таким образом, целью кластерного анализа будет выделение таких морфологических характеристик, которые позволят выделить две группы кластеров: художественный/разговорный/публицистический; научный/официально-деловой.

Дальнейшим этапом анализа было определение минимального количества морфологических характеристик, необходимых для установления стиля текста. Сравнение каждой части речи с соответствующим морфологическим признаком проводилось поочередно, то есть существительное с прилагательным, местоимением, наречие, числительным и т. д. (табл. 4).

Таким образом, всего было построено 78 дендрограмм, результаты объединения стилей в соответствующие кластеры представлены в табл. 5.

На основании полученных данных можно удалить сочетания, разделение на стили которых не поддается логическому пониманию, например объединение художественного и научного стилей, разговорного и официально-делового и т. д.

Интерес представляют выборки, в которых наблюдается объединение научного, официально-делового стилей в один кластер и художественного, публицистического и разговорного стилей — в другой, при этом разговорный и художественный стили также объединены в отдельный кластер (табл. 6).

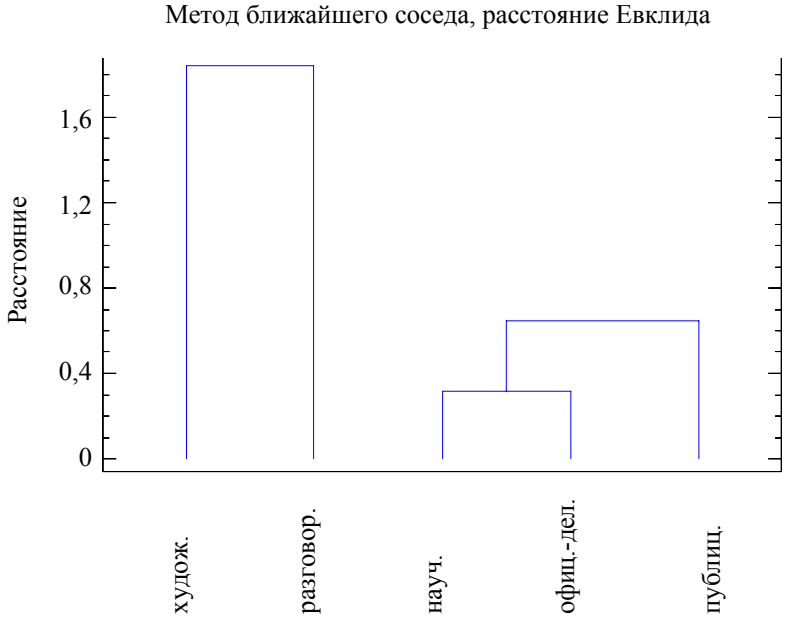

Рис. 1. Дендрограмма по исходным средним значениям

Для формулировки более точных результатов нами сопоставлены полученные после группировки морфологических характеристик сочетания, для этого построены дополнительные дендрограммы с различной мерой близости (Euclidean, Square Euclidean, City Block).

Таблица 4

Соотношение частей речи

Часть речи	Существительное	Прилагательное	Местоимение	Числительное	Наречие	Глагол	Причастие	Деепричастие	Предлог	Союз	Частица	Междометие	Прочие
Существительное	—	*	—	—	—	—	—	—	—	—	—	—	*
Прилагательное		—	*	—	*	*	—	—	—	—	*	—	—
Местоимение			—	*	—	—	*	—	*	—	—	—	—
Числительное				—	—	*	—	—	*	—	—	—	—
Наречие						—	*	—	*	*	—	—	—
Глагол						—	*	—	—	—	—	—	—
Причастие								—	*	—	*	—	—
Деепричастие								—	—	—	—	—	—
Предлог									—	—	—	—	—
Союз										—	—	—	—
Частица											—	—	—
Междометия												—	—
Прочие													—

Таблица 5

Результаты построения дендрограмм для каждой морфологической характеристики с учетом неудачных группировок

Сочетание	Полученные кластеры
Причастие + частица	(Худож. + разговор. + публиц.) + (науч. + офиц.-дел.)
Причастие + предлог	(Худож. + разговор.) + (науч.+ офиц.-дел.+ публиц.)
Глагол + причастие	(Худож. + разговор.+ публиц.) + (науч.+ офиц.-дел.)
Местоимение + причастие	(Худож. + разговор.+ публиц.) + (науч.+ офиц.-дел.)
Местоимение + предлог	(Худож. + разговор.+ публиц.) + (науч.+ офиц.-дел.)
Местоимение + числительное	(Худож. + разговор.+ публиц.) + (науч.+ офиц.-дел.)
Наречие + предлог	(Худож. + разговор.) + (науч.+ офиц.-дел.+ публиц.)
Наречие + союз	(Худож. + разговор.+ публиц.) + (науч.+ офиц.-дел.)
Наречие + причастие	(Худож. + разговор.) + (науч.+ офиц.-дел. + публиц.)
Прилагательное + частица	(Худож. + разговор.+ публиц.) + (науч. + офиц.-дел.)
Прилагательное + местоимение	(Худож. + разговор.) + (науч.+ офиц.-дел. + публиц.)
Прилагательное + наречие	(Худож. + разговор.) + (науч.+ офиц.-дел. + публиц.)
Существительное + прочее	(Худож. + разговор.+ публиц.) + (науч. + офиц.-дел.)
Числительное + предлог	(Худож. + разговор.+ публиц.) + (науч. + офиц.-дел.)
Числительное + глагол	(Худож. + разговор.) + (науч.+ офиц.-дел. + публиц.)
Прилагательное + глагол	(Худож. + разговор.+ публиц.) + (науч. + офиц.-дел.)

Таблица 6

Группировки морфологических характеристик, отобранных для дальнейшего анализа

Сочетание		Полученные кластеры
Причастие + частица	Прилагательное + частица	(Худож. + разговор. + публиц.) + + (науч. + офиц.-дел.)
Глагол + причастие	Существительное + прочее	
Местоимение + причастие	Числительное + предлог	
Местоимение + предлог	Прилагательное + глагол	
Местоимение + числительное	Наречие + союз	

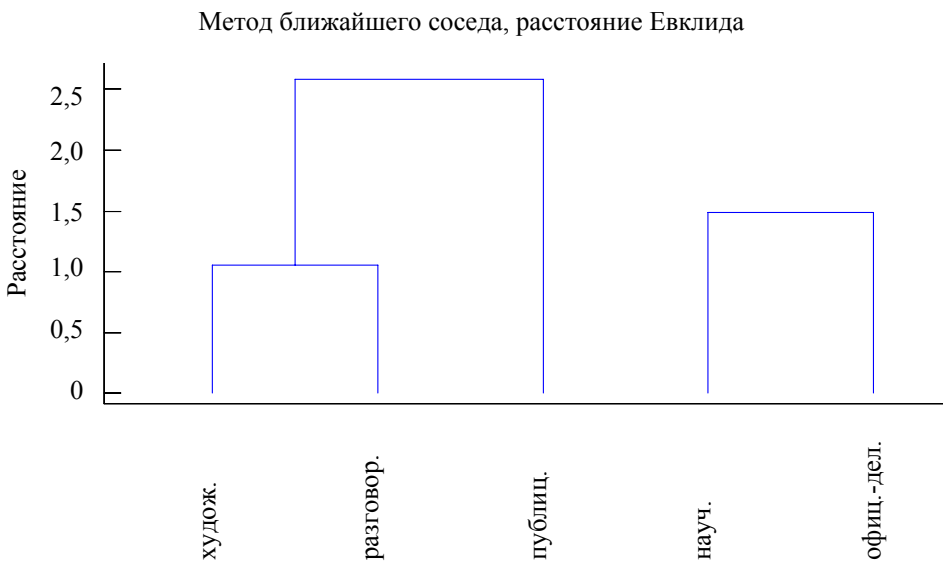

Рис. 2. Окончательный вариант дендрограммы

В результате выделены следующие сочетания морфологических характеристик: местоимение и причастие, местоимение и предлог, наречие и союз, прилагательное и частица, прилагательное и глагол — именно по ним происходит объединение стилей в заранее установленные кластеры (рис. 2).

Заключение. На основании проведенного анализа можно сделать вывод о том, что в работе П. И. Браславского представлены средние значения употребления морфологических характеристик конкретного стиля, различие между которыми статистически незначимо. Использование для сравнения средних значений критерия Фишера и критерия Стьюдента позволяет определить равноточность и неравноточность данных и тем самым подвести итог проведенного автором исследования.

Результаты второй части нашей работы показали, что установить стиль конкретного материала сети можно при использовании для анализа не всех морфологических характеристик, а лишь некоторых сочетаний: местоимение и причастие, местоимение и предлог, наречие и союз, прилагательное и частица, прилагательное и глагол. Именно они позволяют с высоким уровнем точности определить стиль текста.

В качестве заключения можно сформулировать вывод о том, что статистический анализ с использованием различных критериев позволяет более полно проанализировать полученные в результате исследования данные, а метод кластерного анализа — установить связь полученных значений.

Литература

1. Невдах, М. М. Применение информационных технологий в исследовании текстов / М. М. Невдах // Труды БГТУ. Сер. IX, Издат. дело и полиграфия. — 2009. — Вып. XVII. — С. 77–81.

2. Браславский, П. И. Методы повышения эффективности поиска научной информации (на материале Internet): автореф. дис. … канд. техн. наук: 05.13.16 / П. И. Браславский. — Екатеринбург, 2000. — 16 с.

3. Браславский, П. И. Морфологический строй функциональных стилей (на материале документов Internet) [Электронный ресурс] / П. И. Браславский // Известия Уральского государственного университета. — 2001. — № 21. — С. 9–17. — Режим доступа: http://proceedings.-usu.ru/?base=mag/0021(03_11-2001). — Дата доступа: 11.12.2011. — Екатеринбург, 2011. — 11 с.

4. Закс, Л. Статистическое оценивание / Л. Закс; под ред. Ю. П. Адлера [и др.]; пер. с нем. В. Н. Варыгина. — М., 1976. — 598 с. — (Зарубежные статистические исследования).

Поступила 13.03.2012