СТАТИСТИЧЕСКИЙ АНАЛИЗ ТЕКСТОВ УЧЕБНЫХ ИЗДАНИЙ ПО ИЗДАТЕЛЬСКОМУ ДЕЛУ

УДК 004.93’1

М.А. Зильберглейт, А.С. Малюкевич, БГТУ, г. Минск

Аннотация

В статье приведены результаты исследования по установлению взаимозависимости между параметрами «количество символов» – «статистические параметры» и «количество символов» – «автор» по заранее определенным статистическим показателям текстовых фрагментов. Установление взаимозависимости между параметрами проводилось методом дисперсионного анализа.

Введение

Любой текст – это связная и полная последовательность символов, однако, если рассматривать данное понятие в более широком аспекте, то можно отметить, что каждый текст имеет свои отличительные особенности, которые выделяют его среди множества текстов других авторов. К таким особенностям относятся не только особый – присущий именно данному автору – стиль написания и манера изложения материала, но и частота употребления тех или иных частей речи, оборотов и конструкций.

Целью настоящей работы является проведение статистического анализа текстов ряда учебных изданий, а также определение зависимости полученных данных от объема текстовой выборки и авторской принадлежности.

Общее количество отобранных для анализа текстовых фрагментов составило 378 единиц. В них были включены учебные издания 14 авторов, при этом все книги предназначены для обучения студентов по специальности «Издательское дело». Данные учебные издания связаны с такими дисциплинами, как книговедение, история книги, общий курс редактирования, экономика и управление в издательско-полиграфическом комплексе, обработка текстовой информации, охрана авторского права и др. На наш взгляд, такой выбор покрывает основную часть гуманитарного блока, относящегося к профессиональной деятельности специалистов данного профиля.

В каждой единице выборки были выделены три текстовые части объемом в 2000 символов, при этом каждая часть в свою очередь была разделена на дополнительные текстовые фрагменты объемом в 500, 700, 900, 1000, 1200, 1400, 1600, 1800 и 2000 символов. Выполненная работа включает в себя четыре этапа.

Этапы выполнения работы

1. Подготовительный. На данном этапе работы перед нами была поставлена задача определения статистических показателей различных по объему текстовых фрагментов. Для реализации данной операции была использована программа SuperCounter 2.1, а также автомати-ческий анализатор текстов. [1]

В качестве статистических параметров были определены 17 текстовых характеристик: средняя длина слов в слогах, средняя длина слов в буквах, средняя длина слов по Деверу, процент слов в 3-8 слогов, процент односложных слов, средняя длина предложения в словах, средняя длина предложения в слогах, процент неповторяющихся в тексте слов, процент чисел от общего количества слов, процент иностранных слов в тексте, процент ключевых слов, найденных с использованием автоматического анализатора текстов, а также процент ключевых слов, выявленных вручную.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1		Кол-во символов	Ср. длина_слов_в слогах	Ср._длина_слов_в_букв	Ср.длина_слов_по_Деверу	%слов_в_3_слога	%слов_в_4_слога	%слов_в_5_слогов	%слов_в_6_слогов	%слов_в_7_слогов	%слов_в_8_слогов	%_односложных_слов	Ср.длина_предл_в словах	Ср.длина_предл_в_слогах	%неповтор_в_тексте_слов	%чисел_от_общ_кол-ва_слов	%иностр_слов_в_тексте	%ключ_слов_интернет	%ключ_слов_сам	
2	1	500	2,61	6,48	7,64	53,3	34,7	13,3	1,33	0	0	26,7	18,8	49	92	0	2,67	53,33	73,33	1
3	2	500	3,22	7,62	8,9	66,7	44,4	23,8	9,52	1,59	0	15,9	15,8	50,8	87,3	0	0	68,25	76,19	1
4	3	500	3,52	8,03	9,18	62,3	52,5	36,1	21,3	6,56	0	14,8	15,3	53,8	90,2	0	1,64	60,66	75,41	1
5	4	500	3,22	7,17	8,63	49,2	43,1	32,3	16,9	3,08	0	18,5	32,5	105	81,5	0	0	64,06	73,44	1
6	5	500	3,44	8,12	9,67	61,4	42,1	22,8	17,5	7,02	7,02	15,8	19	65,3	87,7	0	0	64,91	75,44	1
7	6	500	3,51	8,1	9,46	61	42,4	23,7	15,3	8,47	8,47	20,3	11,8	41,4	88,1	0	0	69,49	71,19	1
8	7	500	2,91	7,16	8,25	54,4	33,8	25	5,88	2,94	0	22,1	22,7	66	89,7	0	0	58,82	69,12	1
9	8	500	2,73	6,1	7,24	49,4	31,6	11,4	3,8	0	0	16,5	9,88	27	67,1	0	0	62,03	74,68	1
10	9	500	2,36	5,82	6,99	43,4	22,9	13,3	2,41	0	0	26,5	13,8	32,7	84,3	4,82	0	58,02	69,14	1
11	10	500	2,67	6,53	7,75	50	26,4	9,72	1,39	1,39	1,39	20,8	18	48	91,7	0	0	61,11	75	1
12	11	500	2,47	6,05	7,61	44	24	13,3	8	5,33	4	30,7	6,82	16,8	78,7	5,03	0	51,32	63,16	1
13	12	500	3,18	7,15	8,6	60	41,5	27,7	10,8	3,08	0	18,5	16,3	51,8	93,8	1,54	0	57,14	74,6	1
14	13	500	3,22	7,53	8,73	62,5	37,5	20,3	9,38	4,69	1,56	12,5	16	51,5	93,8	0	0	56,25	76,56	1
15	14	500	2,84	6,68	7,86	54,8	30,1	17,8	4,11	1,37	1,37	24,7	18,3	51,8	97,8	0	0	43,84	61,64	1
16	15	500	3,09	6,9	8,09	58,6	41,4	22,9	12,9	2,86	0	17,1	14	43,2	84,3	4,29	0	50	64,29	1
17	16	500	2,26	5,58	6,72	45,9	21,2	9,41	3,53	1,18	0	31,8	14,2	32	82,4	4,71	0	46,43	70,24	1
18	17	500	2,36	5,95	7,14	44,4	17,3	8,64	2,47	1,23	0	28,4	13,5	31,8	87,7	1,23	0	41,98	67,9	1
19	18	500	2,69	6,14	7,36	56,4	33,3	11,5	5,13	2,56	0	20,5	11,1	30	82,1	5,13	0	53,85	75,64	1
20	19	500	2,76	6,4	7,55	56	34,7	14,7	2,67	1,33	0	26,7	25	69	85,3	0	0	52	65,33	1

Лист1 Лист2 Лист3

Рис. 1. Фрагмент таблицы значений статистических показателей текстовых фрагментов

По результатам проведенных в программе SuperCounter 2.1 расчетов была составлена итоговая таблица значений статистических характеристик текста. Фрагмент данной таблицы представлен на рис. 1.

Указанные в таблице (рис. 1) данные были использованы для проведения дальнейших расчетов по установлению взаимозависимости между параметрами. Все числовые значения сгруппированы по 9 блокам, соответствующим объему текстовых фрагментов.

2. Проверка значений статистических параметров на выполнение требований нормальности распределения. Одним из условий использования параметрических методов статистического исследования является нормальное распре-

деление количественных данных. Проверка на нормальность статистических данных осуществлялась с использованием пакета StatGraphics 5.1.

Анализ полученных значений критерия Шапиро-Уилкса, одного из наиболее эффективных критериев проверки на нормальность, а также критерия хи-квадрат,

Таблица 1. Пример результата работы программы StatGraphics 5.1

Параметр	Объем выборки	Критерий хи-квадрат	P-значение	Критерий Шапиро-Уилкса	P-значение	Коэффициент асимметрии	P-значение	Коэффициент эксцесса	P-значение
Средняя длина слов в буквах	500	14,67	0,40	0,97	0,39	0,66	0,51	-0,13	0,90
Средняя длина слов по Деверу	700	11,43	0,65	0,98	0,80	0,75	0,45	0,29	0,77
Процент ключевых слов (программное средство)	900	13,86	0,46	0,99	0,91	0,26	0,80	0,67	0,50
Процент слов в 3 слога и более	1000	11,43	0,65	0,99	0,95	0,06	0,95	0,06	0,95

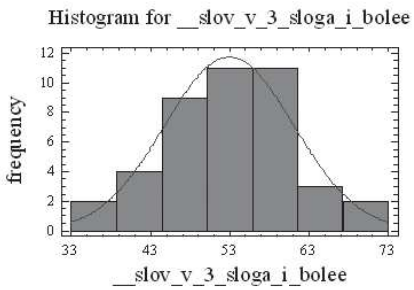

Рис. 2. Гистограмма для параметра «Процент слов в три слога и более»

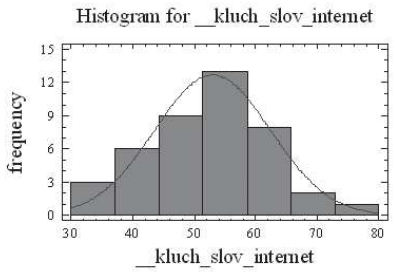

Рис. 3. Гистограмма для параметра «Процент ключевых слов (интернет)»

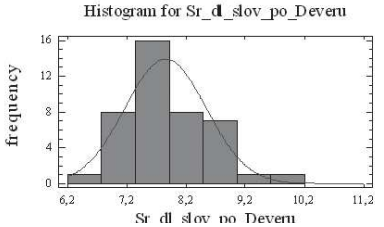

Рис. 4. Гистограмма для параметра «Средняя длина слов по Деверу»

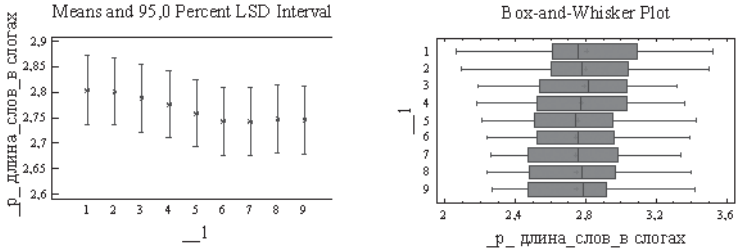

Рис. 5. Графическое представление результатов анализа данных по установлению взаимозависимости между параметрами

Таблица 2. Результаты анализа по установлению взаимозависимости между параметрами «количество символов» и «статистические показатели текста»

Параметр	Результат
Процент односложных слов	статистически значимых отличий нет
Процент иностранных слов в тексте	статистически значимых отличий нет
Процент ключевых слов (программное средство)	статистически значимых отличий нет
Процент ключевых слов (самостоятельно)	статистически значимых отличий нет
Процент неповторяющихся в тексте слов	*статистически значимых отличий есть*
Средняя длина слов в слогах	статистически значимых отличий нет
Средняя длина слов по Деверу	статистически значимых отличий нет
Средняя длина предложения в словах	статистически значимых отличий нет
Средняя длина предложения в слогах	статистически значимых отличий нет
Процент слов в 3 слога и более	статистически значимых отличий нет
Процент слов в 4 слога и более	статистически значимых отличий нет
Процент слов в 5 слогов и более	статистически значимых отличий нет
Процент слов в 6 слогов и более	статистически значимых отличий нет
Процент слов в 7 слогов и более	статистически значимых отличий нет
Процент слов в 8 слогов и более	статистически значимых отличий нет
Процент чисел от общего количества слов	статистически значимых отличий нет
Средняя длина слов в буквах	статистически значимых отличий нет

коэффициента асимметрии и коэффициента эксцесса позволили сделать вывод о том, что большая часть данных подчиняется нормальному закону распределения. Пример результата работы про-граммы представлен в таблице 1, а также на рис. 2-4.

Таким образом, данные, которые будут использованы в ходе классического дисперсионного анализа, смогут в полной мере отразить взаимосвязь исследуемых совокупностей.

3. Установление взаимозависимости между параметрами «количество символов» и «статистические показатели текста». Для установления взаимозависимости между экспериментальными данными был использован метод ANOVA. При исследовании зависимости статистических параметров от объема выборки было проведено межгрупповое и внутригрупповое сравнение полученных в ходе анализа значений, определены суммы квадратов, количество степеней свободы, средние квадраты, F-критерий Фишера, а также P-величина. Значение последнего показателя достигаемого уровня значимости позволяет отвергнуть или принять нулевую гипотезу. Пример

Таблица 3. Результатов анализа данных по установлению взаимозависимости между параметрами «количество символов» и «статистические показатели текста»

Параметр	Параметры сравнения	Сумма квадратов	Количество степеней свободы	Среднеквадратичное значение	F-критерий	P-величина
Средняя длина слов в слогах	M	0,22	8	0,03	0,29	0,97
	B	35,89	369	0,10		
Средняя длина слов в буквах	M	1,03	8	0,13	0,35	0,94
	B	134,74	369	0,37		
Средняя длина слов по Деверу	M	0,389	8	0,05	0,13	1
	B	142,26	369	0,39		
Процент слов в 3 слога и более	M	115,89	8	14,49	0,25	0,98
	B	21488,8	369	58,26		
Процент слов в 4 слога и более	M	128,95	8	16,12	0,31	0,96
	B	18889,1	369	51,19		
Процент слов в 5 слогов и более	M	59,51	8	7,44	0,22	0,99
	B	12700,6	369	34,42		
Процент слов в 6 слогов и более	M	23,19	8	2,90	0,19	0,99
	B	5764,66	369	15,62		
Процент слов в 7 слогов и более	M	3,42	8	0,43	0,13	1
	B	1248,63	369	3,38		
Процент слов в 8 слогов и более	M	0,70	8	0,09	0,05	1
	B	601,61	369	1,639		
Процент односложных слов	M	138,47	8	17,31	0,78	0,62
	B	8229,48	369	22,3021		
Средняя длина предложения в словах	M	28,211	8	3,52637	0,09	1
	B	14953,3	369	40,5239		
Средняя длина предложения в слогах	M	189,657	8	23,7071	0,05	1
	B	159651,0	369	432,659		
Процент неповторяющихся в тексте слов	M	7274,26	8	909,283	22,38	0
	B	14990,2	369	40,6239		
Процент чисел от общего количества слов	M	6118,01	8	764,751	1,0	0,44
	B	282590,0	369	765,827		
Процент иностранных слов в тексте	M	0,508407	8	0,0635508	0,03	1,0
	B	890,025	369	2,41199		

Таблица 4. Результаты установления взаимозависимости между параметрами «автор» и «количество символов»

Параметр	Параметры сравнения	Сумма квадратов	Количество степеней свободы	Среднеквадратичное значение	F-критерий	P-величина
Средняя длина слов в слогах	M	3,27	13	0,25	3,28	0,0041
	B	2,15	28	0,07		
Средняя длина слов в буквах	M	13,51	13	1,04	3,66	0,0020
	B	7,96	28	0,28		
Средняя длина слов по Деверу	M	15,49	13	1,19	4,42	0,0005
	B	7,55	28	0,27		
Процент слов в 3 слога и более	M	1426,12	13	109,70	2,06	0,0530
	B	1489,53	28	53,20		
Процент слов в 4 слога и более	M	1384,71	13	106,52	2,61	0,0164
	B	1144,48	28	40,87		
Процент слов в 5 слогов и более	M	1013,89	13	77,99	2,36	0,0278
	B	925,70	28	33,06		
Процент слов в 6 слогов и более	M	505,53	13	38,89	1,89	0,0778
	B	577,31	28	20,62		
Процент слов в 7 слогов и более	M	89,96	13	6,92	1,90	0,0747
	B	101,73	28	3,63		
Процент слов в 8 слогов и более	M	78,32	13	6,02	3,0	0,0072
	B	56,19	28	2,01		
Процент односложных слов	M	572,32	13	44,02	1,52	0,1720
	B	811,89	28	29		
Средняя длина предложения в словах	M	735,44	13	56,57	0,86	0,5952
	B	1831,48	28	65,41		
Средняя длина предложения в слогах	M	7761,9	13	597,07	1,17	0,3517
	B	14337,6	28	512,06		
Процент неповторяющихся в тексте слов	M	887,59	13	68,28	3,07	0,0063
	B	623,56	28	22,27		
Процент чисел от общего количества слов	M	116,19	13	8,94	3,50	0,0027
	B	71,60	28	2,56		
Процент иностранных слов в тексте	M	53,8351	13	4,14	1,03	0,4543
	B	112,85	28	4,03		

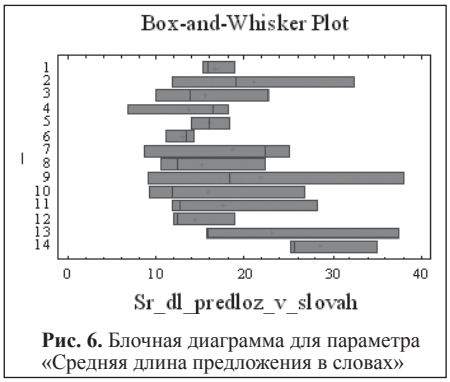

Рис. 6. Блочная диаграмма для параметра «Средняя длина предложения в словах»

Рис. 7. Блочная диаграмма для параметра «Процент слов в 6 слогов и более»

результата работы программы представлен на рис. 5, полученные в ходе анализа данные – в таблицах 2 и 3.

Как следует из табл. 2, шестнадцать из анализируемых параметров не имеют статистически значимых отличий от объема выборки в пределах 500–2000 символов и лишь один (процент неповторяющихся в тексте слов) имеет статистически значимое отличие.

4. Установление взаимозависимости между параметрами «автор» – «количество символов».

В таблице 4 представлены числовые значения результатов обработки экспериментальных данных (для количества символов n = 500) по установлению взаимозависимости между параметрами «автор» и «количество символов». Пример графического результата работы программы представлен на рис. 6 и 7. Сформулированные по результатам анализа выводы представлены в таблице 5.

В ходе анализа взаимозависимости между параметрами «автор» и «количество символов» было установлено, что при использовании выборки в интервале 1000-1600-2000 символов существуют статистически значимые отличия характеристик текста от автора издания.

Таблица 5. Выводы по установлению взаимозависимости между параметрами «автор» и «количество символов»

Параметр	Количество символов						
	500	1000	1200	1400	1600	1800	2000
Процент односложных слов	-	-	+	+	-	+	-
Процент иностранных слов в тексте	+	-	-	-	-	-	-
Процент ключевых слов (программное средство)	+	-	+	-	-	-	-
Процент ключевых слов (самостоятельно)	+	-	+	-	-	-	-
Процент неповторяющихся в тексте слов	+	+	+	+	+	+	+
Средняя длина слов в слогах	+	+	-	-	+	-	+
Средняя длина слов по Деверу	+	+	+	-	+	+	+
Средняя длина предложения в словах	-	-	-	-	-	-	-
Средняя длина предложения в слогах	-	-	+	+	-	+	+
Процент слов в 3 слога и более	-	-	+	+	-	+	-
Процент слов в 4 слога и более	+	+	+	+	+	-	+
Процент слов в 5 слогов и более	+	+	+	+	+	+	+
Процент слов в 6 слогов и более	-	-	+	+	-	+	-
Процент слов в 7 слогов и более	-	+	+	+	+	+	+
Процент слов в 8 слогов и более	+	+	-	-	+	-	+
Процент чисел от общего количества слов	+	+	-	-	+	-	+
Средняя длина слов в буквах	+	+	+	+	+	+	+

«+» – статистически значимые отличия есть; «-» – статитистически значимых отличий нет

Заключение

По результатам проведенного дисперсионного анализа можно сформулировать вывод о том, что объем текстовой выборки, начиная с количества символов, равного 500, не влияет на показатели основных статистических характеристик текста. Нами также было установлено, что статистические параметры текста являются характерными особенностями авторского стиля.

Литература

1. Анализатор текстов [Электронный ресурс]. – Апрель, 2012. – Режим доступа: http://shipbottle.ru/ir.ru

Abstract

Results of research on interdependence establishment are given in article between parameters «quantity of symbols» – «statistics» and «quantity of symbols» – «authors» on in advance established statistics of the selected text fragments. Establishment of independence is carried out by a method of the variance analysis.

Поступила в редакцию 14.09.2012 г.