УДК 655.512

М. М. Косова, аспирант; М. А. Зильберглейт, профессор

ОПИСАТЕЛЬНАЯ СТАТИСТИКА УЧЕБНЫХ ТЕКСТОВ ПО ФИЗИКЕ

In this article the descriptive statistical analysis of school supplies on physics is representative..

Ранее нами были рассмотрены проблемы и достижения читабельности текстов в историческом аспекте [1]. Было показано, что научные исследования по читабельности текста начались в конце 19 – начале 20 в. в США. Число работ, связанных с этой проблемой, на сегодняшний день составляет более 200 для более чем 13 языков мира. В тоже время проблемы русскоязычных текстов в нашей стране практически не затрагивались. Нам известны исследования Мацковского и Микка, проведенные в 70-х гг. 20 в. на территории бывшего СССР. При этом работы Микка были опубликованы на русском, но касались текстов на эстонском языке. Работы же Мацковского проводились на небольшом экспериментальном материале, что, конечно, связано с проблемами в анализе статистических данных больших объемов.

В настоящее время нами разработан алгоритм и программное обеспечение (SuperCounter 2.0), позволяющее проводить статистическую обработку любых текстов на русском языке. Блок-схема данной программы представлена на рис. 1.

Рис. 1

Не исключено, что данная программа окажется полезной для решения проблем, связанных с анализом читабельности.

В представленной статье сделана попытка провести статистический анализ учебных текстов по физике для учащихся средних школ 7, 8, 9, 10, 11 классов. Всего было исследовано 13 учебников, из которых отобрано по 5 отрывков приблизительно одинаковой длины. Определение и обоснование объема выборки будет показано ниже.

В качестве исследуемых параметров изучили показатели текста, представленные в табл. 1.

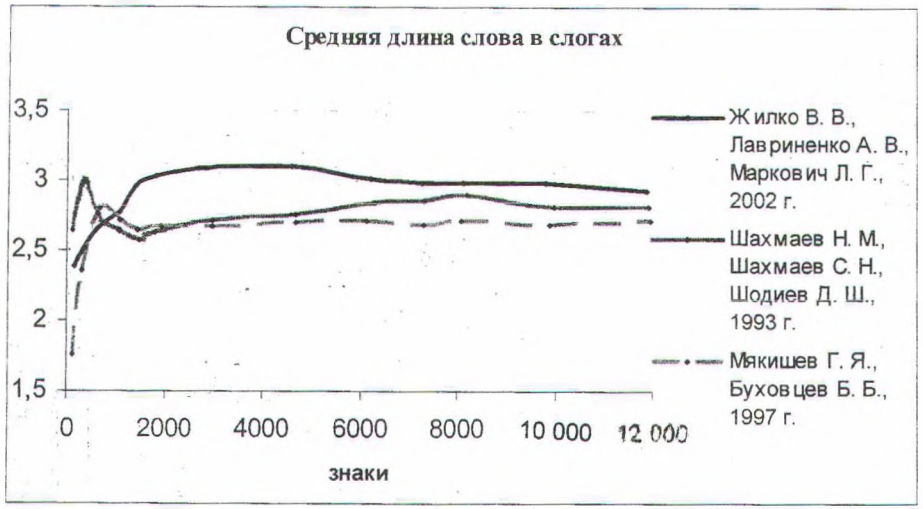
Рассмотрим полученные данные для 11 класса[1]. Они были получены при помощи следующих пакетов: SuperCounter 2.0 (параметры 3–21, 40, 41), BYKBA141 (параметры 22–39, 46), а также некоторых функций пакета Microsoft Word XP (параметры 1–2, 47, 48). Остальные параметры определили вручную.

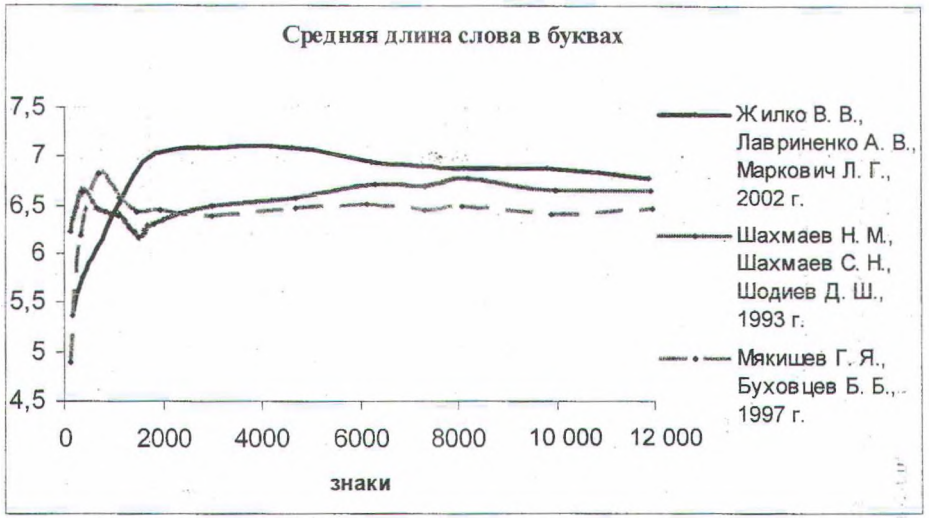
Данный материал в первом приближении может рассматриваться как «отпечатки пальцев» изучаемых объектов, так как нельзя исключить, что каждый автор имеет свою индивидуальность, которая может быть запечатлена статистическим методом. Кроме того, полученные результаты могут быть полезны для установления связи между восприятием текста и его статистическими характеристиками, а также последующей авторизации текста.

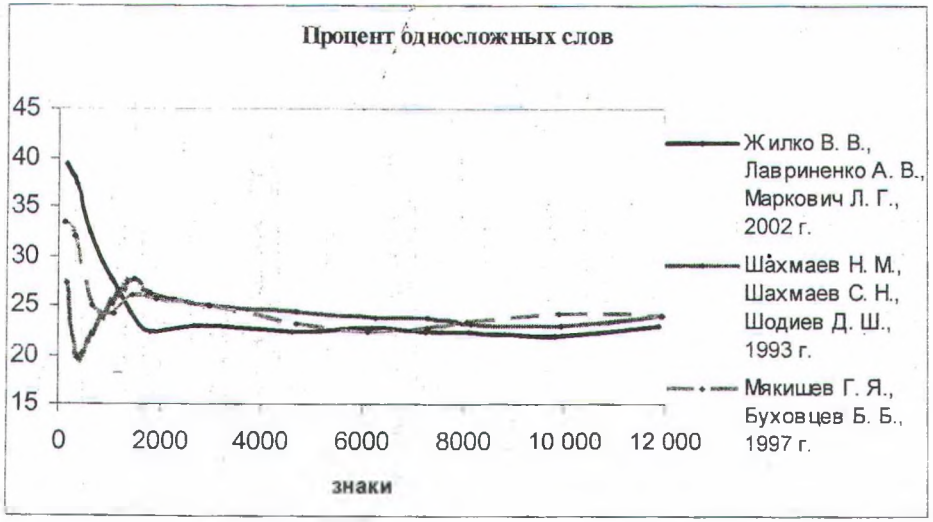
Очевидно, что при данной постановке задачи наиболее важным представляется определение оптимального объема выборки, при котором статистические характеристики текста становятся относительно постоянными. С этой целью нами была исследована зависимость между объемом анализируемой выборки, выраженной в числе символов, и остальными характеристиками текста. Минимальный и максимальный объем для данного эксперимента составил 100 и 12 000 знаков соответственно (рис. 2).

Как следует из полученных графиков, оптимальный объем выборки, при которой исследуемые характеристики текста практически не меняются, начинается с 1800–2000 знаков. Поэтому в дальнейшем объем анализируемых отрывков не превышал 2000 знаков. Число параллельных выборок равнялось 5.

[1] В представленной статье весь иллюстрационный материал будет рассмотрен для 11 класса


Средняя длина слова в слогах
3,5
3
2,5
2
1,5
0
2000
4000
6000
8000
10 000
12 000
знаки
Жилко В. В., Лавриненко А. В., Маркович Л. Г., 2002 г.
Шахмаев Н. М., Шахмаев С. Н., Шодиев Д. Ш., 1993 г.
Мякишев Г. Я., Буховцев Б. Б., 1997 г.


Средняя длина слова в буквах
7,5
7
6,5
6
5,5
5
4,5
0
2000
4000
6000
8000
10 000
12 000
знаки
Жилко В. В., Лавриненко А. В., Маркович Л. Г., 2002 г.
Шахмаев Н. М., Шахмаев С. Н., Шодиев Д. Ш., 1993 г.
Мякишев Г. Я., Буховцев Б. Б., 1997 г.


Процент односложных слов
45
40
35
30
25
20
15
0
2000
4000
6000
8000
10 000
12 000
знаки
Жилко В. В., Лавриненко А. В., Маркович Л. Г., 2002 г.
Шахмаев Н. М., Шахмаев С. Н., Шодиев Д. Ш., 1993 г.
Мякишев Г. Я., Буховцев Б. Б., 1997 г.

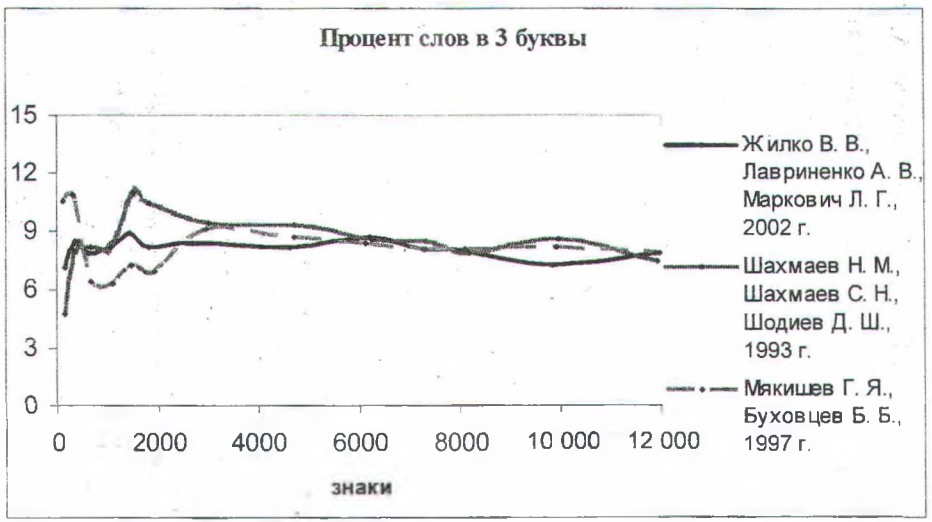

Процент слов в 3 буквы
15
12
9
6
3
0
2000
4000
6000
8000
10 000
12 000
знаки
Жилко В. В., Лавриненко А. В., Маркович Л. Г., 2002 г.
Шахмаев Н. М., Шахмаев С. Н., Шодиев Д. Ш., 1993 г.
Мякишев Г. Я., Буховцев Б. Б., 1997 г.

Таблица 1

№	Параметр	R	$X_{ср}$	S^2	A	E	V	$S(X_{ср})$
1.	Кол-во символов (без пробелов)	335	1886,8	11 275,74	0,07	–1,19	5,63	2,44
2.	Ср. длина абзаца в предложениях	3,82	3,06	1,07	0,43	–0,19	33,66	0,59
3.	Ср. длина слов в слогах	0,88	2,84	0,06	0,61	–0,26	8,45	0,14
4.	Ср. длина слов в буквах	1,62	6,69	0,2	0,8	0,01	6,73	0,17
5.	Процент односложных слов	10,9	23,99	10,7	0,65	–0,4	13,63	0,67
6.	Процент слов в 2 слога и больше	13,2	75,29	14	–0,97	0,43	4,97	0,43
7.	Процент слов в 3 слога и больше	22,7	54,9	44,23	–0,03	–0,71	12,11	0,9
8.	Процент слов в 4 слога и больше	28,6	34,36	51,56	0,43	0,51	20,9	1,22
9.	Процент слов в 5 слогов и больше	25,25	18,32	46,45	1	1,09	37,23	1,59
10.	Процент слов в 6 слогов и больше	12,71	7	10,25	1,22	2,46	45,71	1,21
11.	Процент слов в 7 слогов и больше	4,72	1,63	1,68	0,88	0,58	79,75	1,02
12.	Процент слов в 8 слогов и больше	0,83	0,37	0,12	0,04	–1,89	94,59	0,58
13.	Процент слов в 9 слогов и больше	0,42	0,05	0,02	2,41	4,38	280	0,63
14.	Процент слов в 2 слога	19,9	20,39	35,79	0,7	0,11	29,33	1,32
15.	Процент слов в 3 слога	10,4	20,54	11,93	–0,02	–1,47	16,8	0,76
16.	Процент слов в 4 слога	8,8	16,04	8,64	–0,39	–1,19	18,33	0,73
17.	Процент слов в 5 слогов	18,91	11,32	26	1,78	3,41	45,05	1,52
18.	Процент слов в 6 слогов	8,89	5,37	5,68	0,69	0,6	44,32	1,03
19.	Процент слов в 7 слогов	4,72	1,27	1,75	1,28	1,95	103,94	1,17
20.	Процент слов в 8 слогов	0,74	0,31	0,09	0,14	–1,62	96,77	0,54
21.	Процент слов в 9 слогов	0,42	0,03	0,01	3,87	15	333,33	0,58
22.	Процент слов в 1 букву	9,49	10,18	6,53	0,6	0,49	25,15	0,8
23.	Процент слов в 2 буквы	5,93	6,04	2,93	0,2	–0,37	28,31	0,7
24.	Процент слов в 3 буквы	7,37	7,37	4,29	–0,05	–0,24	28,09	0,76
25.	Процент слов в 4 буквы	7,45	6,32	4,17	1,06	1,1	32,28	0,81
26.	Процент слов в 5 букв	16,22	10,6	18,66	1,27	1,8	40,75	1,33
27.	Процент слов в 6 букв	11,75	8,2	10,15	1,06	1,22	38,9	1,11
28.	Процент слов в 7 букв	12,24	10,39	9,36	0,72	0,91	29,45	0,95
29.	Процент слов в 8 букв	8,45	9,55	4,67	0,55	0,75	22,62	0,7
30.	Процент слов в 9 букв	8,75	8,66	5,9	0,86	0,37	28,06	0,83
31.	Процент слов в 10 букв	8,45	6,57	6,47	0,1	–1,08	38,66	0,99
32.	Процент слов в 11 букв	6,18	4,77	2,52	–0,99	1,16	33,33	0,73
33.	Процент слов в 12 букв	4,98	3,95	2,39	–0,34	–0,78	39,24	0,78
34.	Процент слов в 13 букв	9,31	2,97	6,22	2,48	6,25	83,84	1,44
35.	Процент слов в 14 букв	6,57	2,23	3,09	1,32	2,37	78,92	1,18
36.	Процент слов в 15 букв	3,02	0,7	0,74	1,68	2,83	122,86	1,03
37.	Процент слов в 16 букв	2,5	0,83	0,52	0,63	0,35	86,75	0,79
38.	Процент слов в 17 букв	1,72	0,52	0,35	0,92	–0,41	113,46	0,82
39.	Процент слов в 18 букв	0,66	0,14	0,05	1,24	0,36	157,14	0,59
40.	Ср. длина предложения в словах	6	14,91	2,37	–0,43	0,74	10,33	0,4
41.	Ср. длина предложения в слогах	24,3	42,46	42,42	0,32	0,32	15,33	1
42.	Процент им. существительных	14,71	38,78	19,78	–0,1	–1,03	11,47	0,71
43.	Процент им. прил.+прич.+нар.	15,52	16,35	24,54	0,52	–0,93	30,28	1,22
44.	Процент гл. и деепричастий	18,33	14,76	27,2	1,04	0,58	35,37	1,36
45.	Процент сложных предложений	46,32	35,76	197,51	–0,09	–1,15	39,29	2,35
46.	Процент согласных букв	3,46	56,84	0,86	0,01	–0,38	1,64	0,12
47.	Индекс Флеша[2]	19,2	67,11	42,68	–0,17	–1,1	9,73	0,8
48.	Уровень образования[3]	6,4	11,59	4,75	0,25	–1,26	18,81	0,64

[2] Показатель, который определяет удобочитаемость текста в единицах от 0 до 100 (более подробно см [1]).
[3] Показатель уровня образования по Флешу – Кинсайду ([1]).

В табл. 1 приведены результаты описательной статистики. Применены следующие буквенные обозначения:

R –размах вариаций;

$X_{ср}$ – среднее арифметическое значение;

S^2 – дисперсия;

A – коэффициент асимметрии;

E – эксцесс;

V – коэффициент вариаций;

$S(X_{ср})$ – ошибка среднего.

Анализ полученных результатов показал, что с увеличением средней длины слова в слогах наблюдается непрерывное уменьшение процента их использования. Эти данные полностью согласуются с первым законом Зипфа. В то же время увеличение процента слов в буквах с длинной от 1 буквы до 10 примерно одинаково. Очевидно, что данный показатель является мало информативным.

Сравнивая полученные результаты для учебников 7–11 классов (см табл. 2), можно прийти к следующим выводам. У старших классов наблюдается увеличение средней длины слов в слогах, буквах и увеличение слов по Деверу. Первый закон Зипфа подтверждается для выборки всех классов. Со слов в 5 слогов до слов в 7 слогов наблюдается явное увеличение их количества для старших классов. Процент слов в 2, 3, 4 слога не связан с классом. Изучение влияния процента слов в буквах от номера класса показало отсутствие какой-либо четкой зависимости.

Таблица 2

№	Параметр	11 класс	10 класс	8 класс	7 класс
1.	Ср. длина слов в слогах	2,84	2,85	2,67	2,6
2.	Ср. длина слов в буквах	6,69	6,69	6,18	6,04
3.	Ср. длина слов по Деверу	7,91	7,92	7,42	7,27
4.	Процент односложных слов	23,99	22,81	25,03	26,42
5.	Процент слов в 2 слога	20,39	22,97	24,17	24,35
6.	Процент слов в 3 слога	20,54	19,39	20,21	21,38
7.	Процент слов в 4 слога	16,04	15,65	15,89	15,78
8.	Процент слов в 5 слогов	11,32	10,86	9,59	7,55
9.	Процент слов в 6 слогов	5,37	5,99	3,65	2,99
10.	Процент слов в 7 слогов	1,27	1,19	0,63	0,86
11.	Процент слов в 8 слогов	0,31	0,32	0,09	0,03
12.	Процент слов в 9 слогов	0,03	0,04	0	0,03
13.	Ср. длина предложения в словах	14,91	16,83	14,99	14,76
14.	Ср. длина предложения в слогах	42,46	48,11	39,79	38,43
15.	Процент простых предложений	64,24	58,43	55	66,39
16.	Процент сложных предложений	35,76	41,58	45	33,62
17.	Индекс Флеша	67,11	65,92	70,55	71,64
18.	Уровень образования	11,59	11,76	10,21	9,69

Анализ средней длины предложения в словах показал, что учебники для 10 класса имеют более длинную среднюю длину по сравнению с учебниками остальных классов. В то же время средняя длина предложений в слогах возрастает с 7 класса по 11 класс.

Сравнивая количество простых и сложных предложений, можно ответить, что наибольший процент сложных предложений характерен для 8 и 10 классов, а наименьший для 7 и 11 – учебники написаны разными авторами, обладающими неодинаковыми стилями.

После проведения однофакторного дисперсионного анализа было показано, что не существует принципиальной разницы между распределением длины слов в слогах и буквах между классами.

Литература

1. Косова М. М., Зильберглейт М. А. Проблема читабельности текстов в историческом контексте и на современном этапе // Труды БГТУ. Сер. IX. Издат. дело и полиграфия. – 2005. – Вып. XIII. – С. 3–9.

2. Мацковский М. С. Проблемы читабельности печатного материала: В кн. Смысловое восприятие речевого сообщения (в условиях массовой коммуникации). – М., 1976. – С. 126–142.

3. Рокицкий П. Ф. Биологическая статистика. – 3-е изд., испр. – Мн., 1973. – С. 320.