Известия Гомельского государственного университета
имени Ф. Скорины, № 6(81), 2013

УДК 004.93’1

Сравнительный анализ методов опроса и компьютерного анализа данных для изучения восприятия текстов студентами высших учебных заведений

А.С. Малюкевич, М.А. Зильберглейт

Статья посвящена анализу восприятия студентами высших учебных заведений текстовых фрагментов учебных изданий по специальности «Издательское дело». В качестве основных методик были выбраны метод опроса и метод многомерного статистического анализа. Полученные в ходе исследования данные были преобразованы для дальнейшего анализа и получения дискриминантной функции, а также установления процента корректно определенных объектов по заранее установленным признакам. Сформулированные выводы позволили установить взаимозависимость между двумя исследуемыми методами.
Ключевые слова: опрос, дискриминантный анализ, удобочитаемость, статистика, классификация объектов, восприятие, текстовые характеристики.

The article is dedicated to the analysis of the perception of textbooks fragments for specialty “Publishing” by the students of higher educational institutions. The main chosen methods were the survey method and the method of multivariate statistical analysis. The findings of the study data were transformed for further analysis and obtaining the discriminate functions as well as for the establishment of well-defined objects of interest on pre-established criteria. The conclusions made it possible to establish the relationship between the two investigated methods.
Keywords: survey, discriminant analysis, readability statistics, classification of objects, perception, text features.

Текстовая составляющая любого учебного издания имеет сложную структуру и обладает иерархическим принципом построения. Сложность структуры заключается в строго установленном расположении тематических блоков, таких как: введение, основная тема, заключение, вывод; иерархический принцип построения основан на последовательном введении в учебный текст терминологии − замене текстовых формулировок соответствующими терминами, а также увеличении частоты используемых определений в основном тексте.

На этапе подготовки рукописи к изданию автору приходится также руководствоваться правилами лексического и технического построения текста. К ним относят длину строки, количество слов в предложении, количество предложений в абзаце, объем иллюстративного материала на один учетно-издательский лист, объем текстового материала для каждого из разделов и др.

Восприятие учебного материала обучающимися находится в тесной взаимосвязи со всеми указанными признаками. Текст, содержащий в себе большое количество формульного материала, высокий процент использования иностранной лексики, сложные синтаксические конструкции, большой массив числовых данных, а также причастных и деепричастных оборотов вызовет не только определенные затруднения при его восприятии, но и будет иметь низкий показатель эффективности обучающего средства.

После автора первую оценку подготовленному изданию дает титульный редактор либо рецензент. Профессиональный и теоретический уровень подготовки этих специалистов намного выше уровня тех обучающихся, которым адресовано издание. Именно поэтому на данном этапе выпуска книги необходимо дать адекватную и правильную оценку уровню разработки материала, спрогнозировать, как оно будет воспринято читателями, а главное, готовы ли студенты воспринять материал именно в такой форме и в таком объеме.

Цель нашей работы состоит в том, чтобы установить зависимость между двумя методами исследования восприятия текста студентами высших учебных заведений.

Основная задача – получить репрезентативные результаты, отражающие взаимосвязь используемых в ходе работы методик.

В качестве объектов исследования были определены 82 текстовых фрагмента учебных изданий по издательскому делу для студентов высшего учебного заведения объемом 1800−2000 символов.

Основные методы исследования – метод статистической обработки текстовых фрагментов, опрос (методика дополнения и метод балльных оценок), дискриминантный анализ данных.

Актуальность работы заключается в том, что использование компьютерных технологий на этапе подготовки рукописи к публикации позволит не только оценить ее по основным техническим параметрам, но и определить оценку, которую получит издание после выхода в свет.

1. Определение статистических показателей текста. На первом этапе работы для каждого из отобранных текстовых фрагментов были получены основные статистические показатели, среди них: средняя длина слов в слогах, средняя длина слов в буквах, средняя длина слов по Деверу, 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 односложных слов, средняя длина предложения в слогах, средняя длина предложения в словах, процент чисел от общего количества слов. Дополнительно были введены также два параметра: отношение показателя «Процент слов в 3 слога и более» к параметру «Процент слов в 6 слогов и более», а также отношение показателя «Процент слов в 4 слога и более» к показателю «Процент слов в 6 слогов и более». Основные расчеты по статистике текста были выполнены в программе SuperCounter 2.1. Сгруппированные для последующего анализа данные представлены в таблице 1.

Таблица 1 – Основные статистические параметры текста

№	Средняя длина слов в слогах	Средняя длина слов в буквах	Средняя длина слов по Деверу	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 односложных слов	Средняя длина предложения в словах	Средняя длина предложения в слогах	Процент чисел от общего количества слов	3/6	4/6
1	2,38	5,95	7,38	43,10	23,20	11,10	3,03	0,34	29,30	21,20	50,50	1,01	14,22	7,66
2	2,77	6,48	7,77	56,70	33,80	13,50	2,55	0,73	24,40	13,10	36,30	0,00	22,24	13,25
3	2,65	6,17	7,55	48,80	29,90	16,30	3,65	0,66	26,90	13,10	34,70	0,00	13,37	8,19
4	2,70	6,38	7,53	51,90	25,40	14,40	5,15	2,06	25,10	22,40	60,50	0,00	10,08	4,93
5	2,65	6,41	7,64	52,30	26,90	13,80	3,18	0,00	23,30	16,60	44,10	0,00	16,45	8,46
6	2,88	6,71	7,96	51,90	32,30	19,60	8,42	2,46	21,80	14,30	41,10	0,35	6,16	3,84
7	2,76	6,63	7,88	53,70	33,00	15,10	6,32	2,46	26,70	19,00	52,50	0,00	8,50	5,22
8	2,36	5,95	7,29	46,10	21,10	9,15	3,52	0,35	29,20	18,90	44,60	2,11	13,10	5,99
9	3,24	7,58	8,75	64,70	47,70	27,00	7,88	2,49	20,70	21,90	70,90	0,00	8,21	6,05
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
10	2,64	6,39	7,60	52,30	33,60	15,20	4,24	1,77	29,70	18,90	49,70	1,77	12,33	7,92
11	2,97	6,89	8,23	60,50	37,60	20,30	6,27	0,74	22,90	18,10	53,60	0,37	9,65	6,00
12	2,63	6,20	7,43	45,60	30,90	15,10	6,38	1,68	26,20	21,30	55,90	1,68	7,15	4,84
13	2,71	6,48	7,78	52,80	31,50	15,70	4,12	0,75	26,60	15,70	42,60	0,75	12,82	7,65
14	2,63	6,85	8,05	49,10	30,30	18,00	4,87	1,50	22,10	24,30	63,90	0,00	10,08	6,22
15	2,69	6,41	7,68	51,50	32,60	15,90	5,56	1,85	26,30	24,50	66,00	1,85	9,26	5,86
16	2,40	5,83	6,99	44,20	24,90	9,63	1,00	0,00	29,20	20,10	48,20	1,33	44,33	24,97

2. Анализ текстов по методу дополнений

Для выполнения данного этапа нами работы была проведена предварительная обработка отобранных 82 текстовых фрагментов, цель которой заключалась в удалении каждого пятого слова (классический вариант методики) и замене его пробельным материалом для последующего восполнения респондентами. Таким образом, в ходе опроса студент получал текстовый фрагмент объемом 1800−2000 символов, в котором вместо каждого пятого слова была пропечатана полоса. Основная задача метода состояла в том, чтобы респондент восполнил как можно больше пропущенных лексических единиц.

Среднее количество удаленных слов в каждом из текстов находилось в пределах 50−60 единиц. Количество бланков, представленных к анализу, составляет – 764 единицы, количество представленных к заполнению и обработанных ячеек – 40 535 единиц. Количество респондентов – 300 человек.

Обработка полученных результатов была сведена к преобразованию данных в количественные показатели. При условии полного восполнения пропуска, то есть вставке слова соответствующего оригинальному тексту, ячейке присваивался 1 балл, при условии частичного восполнения (лексический или стилистический синоним, изменение грамматической формы) – 0,7 балла, близкого по смыслу понятия или содержания – 0,5 балла, полное несовпадение либо пропуск слова – 0 баллов. Дополнительно была определена также доля (процентное соотношение) каждой оценки в общем количество пропущенных текстовых единиц. Данные по выполненным этапам работы сведены в итоговой таблице, фрагмент которой представлен в таблице 2.

Таблица 2 – Результаты опроса по методу дополнений

	Количество ответов/Балл						Процентное соотношение			
							%			
	1 (совпадение)	0,7	0,5	0	Всего слов	Средний балл	1	0,7	0,5	0
1	198	17	8	299	522	0,410	37,93	3,26	1,53	57,28
2	230	103	31	122	486	0,653	47,33	21,19	6,38	25,10
3	121	86	10	131	348	0,535	34,77	24,71	2,87	37,64
4	158	38	19	201	416	0,467	37,98	9,13	4,57	48,32
5	147	60	25	200	432	0,466	34,03	13,89	5,79	46,30
6	135	18	52	173	378	0,459	35,71	4,76	13,76	45,77
7	103	18	19	166	306	0,409	33,66	5,88	6,21	54,25
8	100	35	16	119	270	0,491	37,04	12,96	5,93	44,07
9	137	24	24	223	408	0,406	33,58	5,88	5,88	54,66
10	140	26	11	208	385	0,425	36,36	6,75	2,86	54,03
…	…	…	…	…	…	…	…	…	…	…
78	107	13	11	191	322	0,378	33,23	4,04	3,42	59,32
79	143	72	13	164	392	0,510	36,48	18,37	3,32	41,84
80	112	38	34	248	432	0,360	25,93	8,80	7,87	57,41
81	151	34	34	222	441	0,435	34,24	7,71	7,71	50,34
82	90	29	18	128	265	0,450	33,96	10,94	6,79	48,30

Для проведения дальнейших этапов работы необходимо было получить значение, отражающее уровень воспроизведения текста респондентами. В качестве такого значения была определена интегральная сумма, рассчитанная по формуле:

$$S = 1 \times n_1 + 0{,}7 \times n_2 + 0{,}5 \times n_3,$$

где n_1 – значение процентного соотношения слов, получивших 1 балл, к общему объему всех пропущенных слов;

n_2 – значение процентного соотношения слов, получивших 0,7 балла, к общему объему всех пропущенных слов;

n_3 – значение процентного соотношения слов, получивших 0,5 балла, к общему объему пропусков.

Таким образом, для каждого из текстов для проведения дальнейшего анализа было получено значение интегральной суммы. Чем оно выше, тем лучше воспроизводимость текста, соответственно, чем оно меньше, тем хуже восприятие учебного материала.

Для проведения следующего этапа работы необходимо было определить точку отсчета, то есть выбрать значение, на основе которого можно было бы отнести текст к группе воспроизводимых либо невоспроизводимых текстов. Анализ результатов опроса по методу дополнений, сравнение полученных данных с результатами опроса по методу балльных оценок [1], а также общая оценка полученных данных позволили определить пороговое значение для данного метода в точке 35.

Для подготовки итоговой матрицы и получения дискриминантной функции, все тексты были поделены на две группы. К первой группе со значением интегральной суммы более 35 относились тексты с низким уровнем сложности, ко второй − со значением менее 35 соответственно − тексты с высоким уровнем сложности, то есть тяжелые по восприятию и воспроизведению.

Третий этап работы заключался в том, чтобы проанализировать полученные в предыдущих этапах работы данные в программе статистики StatGraphics 5.1, сравнить результаты программного анализа и опроса, определить точность классификации объектов по заранее установленным признакам, а также получить дискриминантную функцию, по которой для любого из текстов можно было бы определить оценку и уровень восприятия студентами еще на стадии допечатной подготовки.

Для реализации данного этапа работы была подготовлена матрица, включающая в себя значения статистических параметров текста, полученных с помощью расчетов в программе SuperCounter 2.1, а также указание, к какому из классов по результатам опроса, был отнесен каждый из исследуемых текстовых фрагментов.

Для реализации дискриминантного анализа была использована матрица, состоящая из 15 столбцов и 82 строк.

Результаты проведенного дискриминантного анализа: Eigenvalue (собственное значении) – 1,10566, Relative Percentage (процентное содержание) – 100, Canonical Correlation (каноническая корреляция) – 0,72463, Wilks Lambda (коэффициент Шапиро-Уилкса) − 0,474911, Chi-Square (хи-квадрат) – 54,3578, DF (количество степеней свободы) – 14, P-value (P-значение) – 0. Графическая форма результата представлена на рис. 1.

Коэффициенты классификационной функции двух групп объектов представлены в таблице 3.

Таблица 3 – Коэффициенты дискриминантных функций

№	Параметр	Значение параметра	Коэффициент (k)	
			1 группа (k_1)	2 группа (k_2)
1	Средняя длина слов в слогах	p_1	2122,76	2144,62
2	Средняя длина слов в буквах	p_2	–54,9459	–42,8671
3	Средняя длина слов по Деверу	p_3	483,726	460,195
4	Процент слов в 3 слога и более	p_4	–19,5526	–19,549
5	Процент слов в 4 слога и более	p_5	–25,0052	–24,1717
6	Процент слов в 5 слогов и более	p_6	–19,0833	–20,1734
7	Процент слов в 6 слогов и более	p_7	–13,5695	–12,2767
8	Процент слов в 7 слогов и более	p_8	–66,8416	–65,6394
9	Процент односложных слов	p_9	39,5872	39,5923
10	Средняя длина предложения в словах	p_{10}	87,3949	90,664
11	Средняя длина предложения в слогах	p_{11}	–31,0564	–32,1531
12	Процент чисел от общего количества слов	p_{12}	56,3843	57,7088
13	Отношение показателя «Процент слов в 3 слога и более» к показателю «Процент слов в 6 слогов и более»	p_{13}	–3,182627	–3,08938
14	Отношение показателя «Процент слов в 4 слога и более» к показателю «Процент слов в 6 слогов и более»	p_{14}	8,01303	7,60715
15	Постоянное значение		–4000,96	–3985,5

Дискриминантная функция для первой группы объектов имеет вид: $-4000{,}96 + 2122{,}76 \times p_1 - 54{,}9459 \times p_2 + 483{,}726 \times p_3 - 19{,}5529 \times p_4 - 25{,}0052 \times p_5 - 19{,}0833 \times p_6 - 13{,}5695 \times p_7 - 66{,}8416 \times p_8 + 39{,}5872 \times p_9 + 87{,}3949 \times p_{10} - 31{,}0564 \times p_{11} + 56{,}3843 \times p_{12} - 3{,}182627 \times p_{13} + 8{,}01303 \times p_{14}$, где p_n – значение каждого из статистических параметров (таблица 3).

Дискриминантная функция для второй группы объектов можно представить: $-3985{,}5 + 2144{,}62 \times p_1 - 42{,}8671 \times p_2 + 460{,}195 \times p_3 - 19{,}549 \times p_4 - 24{,}1717 \times p_5 - 20{,}1734 \times p_6 - 12{,}2767 \times p_7 - 65{,}6394 \times p_8 - 39{,}5923 \times p_9 + 90{,}664 \times p_{10} - 32{,}1531 \times p_{11} + 57{,}9088 \times p_{12} - 3{,}08938 \times p_{13} + 7{,}60715 \times p_{14}$, где p_n – значение каждого из статистических параметров (таблица 3).

Для определения, к какой из групп, отражающих уровень восприятия учебного материала, будет отнесен анализируемый объект, в представленные выше дискриминантные

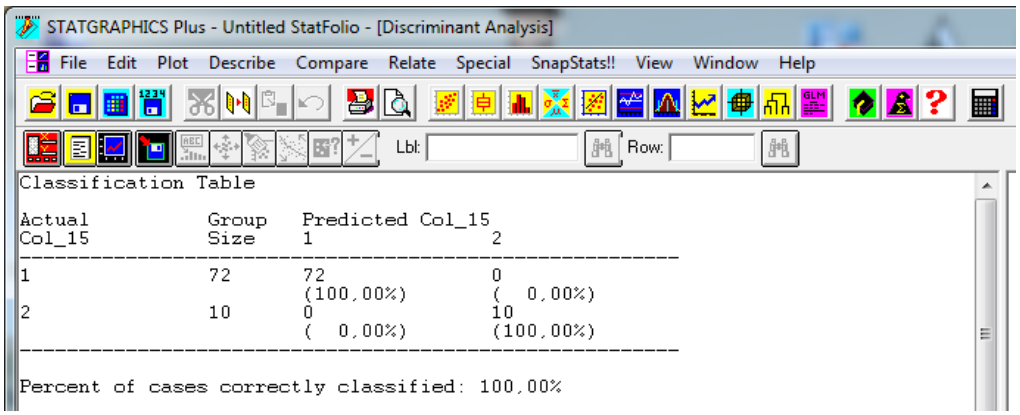

Рисунок 1 – Результаты дискриминантного анализа

функции подставляют значения, полученные на основе статистического анализа текста в программе SuperCounter 2.1 либо другом программном средстве, а затем сравнивают результаты и устанавливают принадлежность к той или иной группе.

Результатом проведенного исследования можно считать полученные дискриминантные функции для каждой группы объектов, которые служат для классификации текстовых блоков, а именно – относится исследуемый фрагмент к группе «Легкий уровень восприятия текстовой информации» либо к группе «Тяжелый уровень восприятия текстовой информации».

Результаты проведенного исследования подтверждают выполнение поставленных задач и целей работы. Процент корректно определенных объектов составляет 100 %, то есть результаты исследования уровня восприятия текстов учебных изданий по издательскому делу, полученные методом опроса, полностью соответствуют результатам, полученным в ходе автоматизированной обработки и анализа данных.

Результаты исследования имеют научную значимость, актуальны и репрезентативны. Сформулированные выводы будут использованы для формулировки решающих правил оценки уровня восприятия текстов учебных изданий студентами высших учебных заведений, а также лягут в основу разработки специального программного средства для оценки рукописи на этапе допечатной подготовки.

Литература

1. Зильберглейт, М.А. Применение метода распознавания образов для оценки качества учебных текстов / М.А. Зильберглейт, А.С. Малюкевич // Электроника-инфо : науч.-практ. журнал для специалистов. – 2013. – № 10. – С. 51−55.

Белорусский государственный
технологический университет

Поступила в редакцию 31.10.2013