УДК 655.5

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

СНИЖЕНИЕ РАЗМЕРНОСТИ ФАКТОРНОГО ПРОСТРАНСТВА ПРИ ИЗУЧЕНИИ СТАТИСТИЧЕСКИХ ХАРАКТЕРИСТИК ТЕКСТА

В статье приводится анализ текстовых характеристик, выбранных в качестве основных параметров исследования, с помощью различных методов обработки данных, среди которых нами были выбраны: метод факторного анализа, метод кратчайшего незамкнутого пути, метод корреляционных плеяд, а также метод многомерного шкалирования. По результатам проведенных расчетов, выполненных с использованием специального программного обеспечения, были построены таблицы и графики, отражающие выявленные зависимости между исследуемыми параметрами. Дальнейший анализ данных позволит сформулировать решающие правила для определения уровня восприятия текста по специальности учащимися высших учебных заведений.

This article provides an analysis of characteristics of the text, that were selected as the main parameters of the study, through various processing of methods, were chosen: the method of factor analysis, the method of the shortest non-closed path, the method of correlation of the Pleiades and the method of multidimensional scaling. The results of the calculations made on the basis of the of special software tools, the results were illustrated in tables and graphs, that show the level of relationship between the identified parameters. Further analysis of the data allows us to formulate the decision rules for determinate of the readability of the text and its level of students' perception of higher education.

Введение. В работах, связанных с исследованием статистических характеристик текста, приводится большое число параметров, которые следует обработать и осмыслить.

Для наглядности картины и простоты интерпретации часто бывает необходимым выбрать существенно меньшее количество факторов из числа исследуемых. Такой подход необходим также в силу того, что многие параметры текста коррелированны между собой, что при последующем их совместном использовании ухудшает качество исследования. Данный подход относится к методам снижения размерности факторного пространства [1].

Целью настоящей работы является снижение размерности факторного пространства при изучении статистических характеристик текста, представленных нами ранее в работах [2].

Основная часть. В качестве исходного материала нами были использованы статистические характеристики текста, которые включали в себя следующие факторы: средняя длина слов в слогах; средняя длина слов в буквах; средняя длина слов по Деверу; процент слов в 3–7 слогов и более; процент односложных слов; средняя длина предложения в словах; средняя длина предложения в слогах; процент чисел от общего количества слов; процент опорных слов, выявленных вручную; процент опорных слов, выявленных с использованием программного средства; отношение показателя «Процент слов в 3 слога и более» к показателю «Процент слов в 6 слогов и более»; отношение показателя «Процент слов в 4 слога и более» к показателю «Процент слов в 6 слогов и более».

В качестве методов снижения размерности в настоящей работе были использованы методы факторного анализа, многомерного шкалирования, корреляционных плеяд, а также кластерного анализа и кратчайшего незамкнутого пути.

В качестве исходных данных были использованы значения, полученные в предыдущих этапах исследования [2].

В табл. 1 приведена корреляционная матрица параметров исследуемых текстов. Все значения коэффициентов корреляции после соответствующей проверки по критерию Фишера оказались значимыми для вероятности, равной 0,95.

В качестве критического значения коэффициента корреляции была принята величина $r_{крит} = 0{,}7$. Такое значение было выбрано исходя из того, что в научной литературе принято считать, что значение коэффициентов корреляции ниже 0,7 характерно для слабых связей.

Все значения коэффициентов корреляции ниже 0,7 были приравнены к нулю в результате чего были получены следующие плеяды:

1) отношение показателя «Процент слов в 3 слога и более» к показателю «Процент слов в 6 слогов и более»; отношение показателя «Процент слов в 4 слога и более» к показателю «Процент слов в 6 слогов и более»;

2) процент слов в 6 слогов и более и процент слов в 7 слогов и более;

3) процент слов в 5 слогов и более и процент слов в 6 слогов и более;

4) средняя длина слов в слогах и процент слов в 3 слога и более;

5) средняя длина слов в слогах и процент слов в 4 слога и более;

Таблица 1

Корреляционная матрица параметров исследуемых текстов

Корреляция по Пирсону	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	1	,965	,958	,931	,938	,928	,836	,693	−,756	,011	,298	−,365	,339	,621	−,534	−,452
2	,965	1	,993	,920	,928	,908	,817	,712	−,751	,022	,299	−,290	,411	,688	−,526	−,442
3	,958	,993	1	,917	,925	,898	,808	,705	−,757	,009	,284	−,280	,428	,704	−,526	−,441
4	,931	,920	,917	1	,904	,838	,696	,569	−,726	−,051	,218	−,257	,364	,614	−,426	−,356
5	,938	,928	,925	,904	1	,907	,771	,634	−,645	,033	,305	−,245	,390	,592	−,498	−,373
6	,928	,908	,898	,838	,907	1	,878	,663	−,612	,108	,374	−,228	,335	,524	−,616	−,534
7	,836	,817	,808	,696	,771	,878	1	,816	−,499	,123	,360	−,17	,282	,491	−,693	−,658
8	,693	,712	,705	,569	,634	,663	,816	1	−,367	,016	,212	−,148	,283	,500	−,552	−,513
9	−,756	−,751	−,757	−,726	−,645	−,612	−,499	−,367	1	,134	−,083	,21	−,281	−,662	,324	,270
10	,011	,022	,009	−,051	,033	,108	,123	,016	,134	1	,955	−,018	−,163	−,211	−,151	−,156
11	,298	,299	,284	,218	,305	,374	,360	,212	−,083	,955	1	−,119	−,063	−,026	−,297	−,276
12	−,365	−,290	−,280	−,257	−,245	−,22	−,170	−,148	,214	−,018	−,119	1	,052	−,150	,013	−,009
13	,339	,411	,428	,364	,390	,335	,282	,283	−,281	−,163	−,063	,052	1	,686	−,239	−,192
14	,621	,688	,704	,614	,592	,524	,491	,500	−,662	−,211	−,026	−,150	,686	1	−,314	−,260
15	−,534	−,526	−,526	−,426	−,498	−,616	−,693	−,552	,324	−,151	−,297	,013	−,239	−,314	1	,977
16	−,452	−,442	−,441	−,356	−,373	−,534	−,658	−,513	,270	−,156	−,276	−,009	−,192	−,260	,977	1

Примечание. 1. Средняя длина слов в слогах. 2. Средняя длина слов в буквах. 3. Средняя длина слов по Деверу. 4. Процент слов в 3 слога и более. 5. Процент слов в 4 слога и более. 6. Процент слов в 5 слогов и более. 7. Процент слов в 6 слогов и более. 8. Процент слов в 7 слогов и более. 9. Процент односложных слов. 10. Средняя длина предложения в словах. 11. Средняя длина предложения в слогах. 12. Процент чисел от общего количества слов. 13. Процент опорных слов, выявленных вручную. 14. Процент опорных слов, выявленных с использованием программного средства. 15. Отношение показателя «Процент слов в 3 слога и более» к показателю «Процент слов в 6 слогов и более». 16. Отношение показателя «Процент слов в 4 слога и более» к показателю «Процент слов в 6 слогов и более».

6) средняя длина слов в слогах и процент слов в 5 слогов и более;

7) процент слов в 4 слога и более и процент слов в 5 слогов и более;

8) средняя длина предложения в словах и средняя длина предложения в слогах;

9) средняя длина слов по Деверу и средняя длина слов в буквах;

10) средняя длина слов в буквах и средняя длина слов в слогах и др.

Метод кратчайшего незамкнутого пути основан на последовательном выборе пар факторов с наиболее высокими коэффициентами корреляции и последовательном достраивании графа до критического значения коэффициента корреляции.

По результатам проведенного анализа нами были получены три графа, которые представлены на рис. 1.

Первый из них включает в себя два фактора: отношение показателя «Процент слов в 3 слога и более» к показателю «Процент слов в 6 слогов и более» (1) и отношение показателя «Процент слов в 4 слога и более» к показателю «Процент слов в 6 слогов и более» (2). Второй граф содержит факторы «Средняя длина предложения в слогах» (3) и «Средняя длина предложения в словах» (4). Третий граф включает такие факторы, как: средняя длина слов в слогах (5), средняя длина слов в буквах (6), средняя длина слов по Деверу (7), процент односложных слов (8), процент слов в 3 слога и более (9), процент слов в 4 слога и более (10), процент слов в 5 слогов и более (11), процент слов в 7 слогов и более (12), процент слов в 6 слогов и более (13).

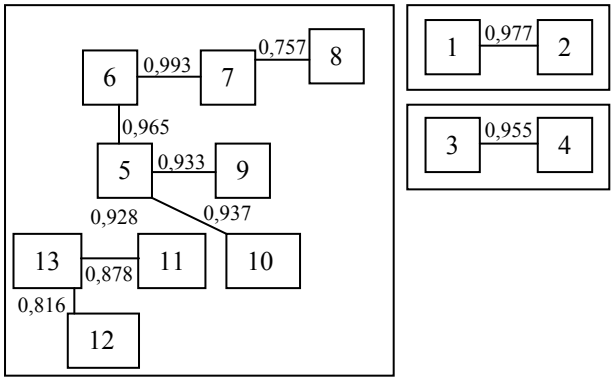

Рис. 1. Графы, полученные по результатам проведенного анализа данных

Среди методов снижения размерности метод факторного анализа является наиболее распространенным. Согласно модели данного метода свертка факторного пространства осуществляется путем уменьшения избыточности за счет введения так называемых общих факторов, включающих изучаемые, причем эти общие факторы не коррелированы, а их число меньше

исходных. Результаты факторного анализа приведены в табл. 2, табл. 3 и на рис. 2, сам метод был реализован при помощи пакета StatGraphics Plus 5.1 (метод главных компонент).

Таблица 2

Результаты факторного анализа

Номер фактора	Собственное значение	Процент дисперсии	Совокупный процент
1	8,91071	55,692	55,692
2	2,28105	14,257	69,948
3	1,4925	9,328	79,277
4	1,03124	6,445	85,722
5	0,734603	4,591	90,313
6	0,611918	3,824	94,138
7	0,406618	2,541	96,679
8	0,166383	1,04	97,719
9	0,140349	0,877	98,596
10	0,090466	0,565	99,161

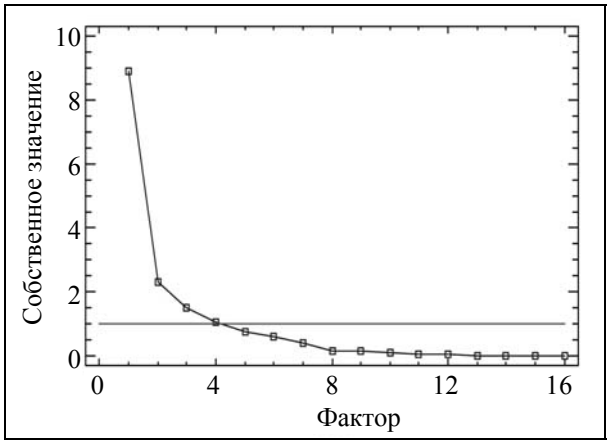

Рис. 2. Графическое представление результатов факторного анализа

Из табл. 3 видно, что около 86% дисперсии может быть объяснено при помощи четырех факторов. Разбиение результатов вычислений на четыре класса позволило выделить следующие группы факторов:

1 группа. Средняя длина предложения в словах; средняя длина предложения в слогах;

2 группа. Процент чисел от общего количества слов; отношение показателя «Процент слов в 3 слога и более» к показателю «Процент слов в 6 слогов и более»; отношение показателя «Процент слов в 4 слога и более» к показателю «Процент слов в 6 слогов и более»; процент односложных слов.

3 группа. Процент опорных слов, выявленных вручную; процент опорных слов, выявленных с использованием программного средства;

4 группа. 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; средняя длина слов в слогах.

Таблица 3

Результаты факторного анализа (4 класса)

№	Фактор			
	1	2	3	4
1	0,968332	−,0,05477	−,0,17084	−,0,10336
2	0,055969	0,891871	−,0,28124	0,332638
3	0,331142	0,837609	−,0,31869	0,285449
4	−,0,26835	0,037776	0,540636	0,466177
5	0,445007	−,0,374	0,245801	0,627647
6	0,698595	−,0,43326	0,052779	0,350266
7	−,0,6622	−,0,35621	−,0,57644	0,157726
8	−,0,58678	−,0,38438	−,0,62649	0,182989
9	0,972652	−,0,07815	−,0,13711	0,009632
10	0,970204	−,0,09593	−,0,12559	0,024992
11	0,898211	−,0,15637	−,0,19633	−,0,00902
12	0,923802	−,0,05256	−,0,16295	0,021523
13	0,935414	0,0966	−,0,04322	−,0,05343
14	0,886951	0,192016	0,161017	−,0,12848
15	0,759597	0,069061	0,200295	−,0,12336
16	−,0,73017	0,28963	0,18658	−,0,01657

Примечание. 1. Средняя длина слов в слогах. 2. Средняя длина предложения в словах. 3. Средняя длина предложения в слогах. 4. Процент чисел от общего количества слов. 5. Процент опорных слов, выявленных вручную. 6. Процент опорных слов, выявленных с использованием программного средства. 7. Отношение показателя «Процент слов в 3 слога и более» к показателю «Процент слов в 6 слогов и более». 8. Отношение показателя «Процент слов в 4 слога и более» к показателю «Процент слов в 6 слогов и более». 9. Средняя длина слов в буквах. 10. Средняя длина слов по Деверу. 11. Процент слов в 3 слога и более. 12. Процент слов в 4 слога и более. 13. Процент слов в 5 слогов и более. 14. Процент слов в 6 слогов и более. 15. Процент слов в 7 слогов и более. 16. Процент односложных слов.

В данной работе был использован также метод снижения размерности, который основан на нелинейном отображении выборочных точек в пространство меньшей размерности, наименее искажающих их геометрическую конфигурацию. В качестве алгоритма был выбран метод, предлагаемый пакетом SPSS, – ALSCAL (multidimensional scaling).

В результате расчета для стресса, равного 0,07806, при проектировании на двухмерную плоскость были выделены классы:

1) средняя длина предложения в слогах;

2) процент опорных слов, выявленных вручную; процент слов в 3 слога и более; процент слов в 4 слога и более; процент односложных слов; процент опорных слов, выявленных с использованием программного средства;

3) отношение показателя «Процент слов в 3 слога и более» к показателю «Процент слов

в 6 слогов и более»; отношение показателя «Процент слов в 4 слога и более» к показателю «Процент слов в 6 слогов и более»; средняя длина слов в слогах; процент слов в 6 слогов и более; процент чисел от общего количества слов; средняя длина слов в буквах; процент слов в 7 слогов и более; средняя длина предложения в словах; процент слов в 5 слогов и более.

Заключительный этап исследования состоял в анализе данных методом кластерного анализа.

В результате использования меры близости «Квадрат Евклидового расстояния» и алгоритмов Варда, центроидов и группового сравнения были получены кластеры, представленные в табл. 4.

Таблица 4

Результаты кластерного анализа

Переменная	Кластер*	Переменная	Кластер**	Переменная	Кластер***
Средняя длина слов в слогах	1	Средняя длина слов в слогах	1	Средняя длина слов в слогах	1
Средняя длина слов в буквах	1	Процент опорных слов, выявленных вручную	1	Процент опорных слов, выявленных вручную	1
Средняя длина слов по Деверу	1	Процент опорных слов, выявленных с использованием программного средства	1	Процент опорных слов, выявленных с использованием программного средства	1
Процент слов в 3 слога и более	1	Средняя длина слов в буквах	1	Средняя длина слов в буквах	1
Процент слов в 4 слога и более	1	Средняя длина слов по Деверу	1	Средняя длина слов по Деверу	1
Процент слов в 5 слогов и более	1	Процент слов в 3 слога и более	1	Процент слов в 3 слога и более	1
Процент слов в 6 слогов и более	1	Процент слов в 4 слога и более	1	Процент слов в 4 слога и более	1
Процент слов в 7 слогов и более	1	Процент слов в 5 слогов и более	1	Процент слов в 5 слогов и более	1
Средняя длина предложения в словах	2	Процент слов в 6 слогов и более	1	Процент слов в 6 слогов и более	1
Средняя длина предложения в слогах	2	Процент слов в 7 слогов и более	1	Процент слов в 7 слогов и более	1
Процент чисел от общего количества слов	3	Средняя длина предложения в словах	2	Средняя длина предложения в словах	2
Процент односложных слов	3	Средняя длина предложения в слогах	2	Средняя длина предложения в слогах	2
Процент опорных слов, выявленных вручную	4	Процент чисел от общего количества слов	3	Процент чисел от общего количества слов	3
Процент опорных слов, выявленных с использованием программного средства	4	Отношение показателя «Процент слов в 3 слога и более» к показателю «Процент слов в 6 слогов и более»	4	Отношение показателя «Процент слов в 3 слога и более» к показателю «Процент слов в 6 слогов и более»	4
Отношение показателя «Процент слов в 3 слога и более» к показателю «Процент слов в 6 слогов и более»	5	Отношение показателя «Процент слов в 4 слога и более» к показателю «Процент слов в 6 слогов и более»	4	Отношение показателя «Процент слов в 4 слога и более» к показателю «Процент слов в 6 слогов и более»	4
Отношение показателя «Процент слов в 4 слога и более» к показателю «Процент слов в 6 слогов и более»	5	Процент односложных слов	5	Процент односложных слов	5

* Метод Варда.

** Центроидный метод.

*** Метод группового сравнения.

Таблица 5

Результаты снижения размерности статистических характеристик текста методом кластерного анализа

Фактор	Коэффициент корреляции по Пирсону (метод Варда)	Коэффициент расстояния по Минковскому (метод Варда)	Корреляция по Пирсону (метод медианы)
Средняя длина слов в слогах	1	1	1
Средняя длина слов в буквах	1	1	1
Средняя длина слов по Деверу	1	1	1
Процент слов в 3 слога и более	1	1	1
Процент слов в 4 слога и более	1	1	1
Процент слов в 5 слогов и более	1	1	1
Процент слов в 6 слогов и более	1	1	1
Процент слов в 7 слогов и более	1	1	1
Процент односложных слов	2	2	2
Средняя длина предложения в словах	3	3	3
Средняя длина предложения в слогах	3	3	3
Процент чисел от общего количества слов	2	2	4
Процент опорных слов, выявленных вручную	4	4	1
Процент опорных слов, выявленных с использованием программного средства	4	4	1
Отношение показателя «Процент слов в 3 слога и более» к показателю «Процент слов в 6 слогов и более»	5	5	5
Отношение показателя «Процент слов в 4 слога и более» к показателю «Процент слов в 6 слогов и более»	5	5	5

Кроме того в качестве меры близости были использованы: коэффициент корреляции по Пирсону – алгоритм Варда; коэффициент расстояния по Минковскому – алгоритм Варда; коэффициент по Пирсону – метод медианы. Результаты выполненного анализа представлены в табл. 5.

Заключение. В результате использования различных методов свертки факторного пространства были выделены относительно однородные группы, что позволит в дальнейшем провести корректную интерпретацию статистических показателей текста.

Литература

1. Айвазян, С. А. Классификация многомерных наблюдений / С. А. Айвазян; З. И. Бежаева; О. В. Староверов. – М.: Статистика, 1987. – С. 134.

2. Статистический анализ текстов учебных изданий по издательскому делу / М. А. Зильберглейт, А. С. Малюкевич // Электроника ИНФО: науч.-практ. журнал для специалистов. – № 1 (2013).

Поступила 26.04.2013