УДК 519.178
MSC 05C70

Вестник СПбГУ. Прикладная математика. Информатика... 2023. Т. 19. Вып. 4

Теоретико-игровая оценка сложности учебных текстов*

А. В. Хитрый[1], *В. В. Мазалов*[1,2], *Н. А. Буре*[2], *П. В. Дробная*[3]

[1] Федеральный исследовательский центр «Карельский научный центр Российской академии наук»,
Российская Федерация, 185910, Петрозаводск, ул. Пушкинская, 11
[2] Санкт-Петербургский государственный университет,
Российская Федерация, 199034, Санкт-Петербург, Университетская наб., 7–9
[3] Петрозаводский государственный университет,
Российская Федерация, 185910, Петрозаводск, пр. Ленина, 33

Для цитирования: *Хитрый А. В., Мазалов В. В., Буре Н. А., Дробная П. В.* Теоретико-игровая оценка сложности учебных текстов // Вестник Санкт-Петербургского университета. Прикладная математика. Информатика. Процессы управления. 2023. Т. 19. Вып. 4. С. 509–521. https://doi.org/10.21638/11701/spbu10.2023.407

Предлагается метод оценки сложности текстов на основе методов теории кооперативных игр. Игроками в этой игре являются длины слов в тексте. Сама игра представлена в виде игры голосования, где ценность игрока определяется числом коалиций, в которых игрок ключевой. Ранги игроков формируются путем вычисления значения Шепли — Шубика или индекса Банцафа в игре голосования с заданным порогом голосования. Таким образом, каждому тексту ставится в соответствие вектор значений Шепли — Шубика или Банцафа. После этого в пространстве векторов проводится ранжирование текстов по сложности на основе экспертных оценок, полученных в данной области.

Ключевые слова: обработка текстов, игра голосования, значение Шепли — Шубика, индекс Банцафа, кластеризация.

1. Введение. Чтение текстов — основной источник получения информации, а также средство обучения иностранному языку. В связи с этим отбор текстового материала является важнейшим этапом в практике преподавания русского языка как иностранного. Параметры отбора текста зависят от уровня владения изучаемого языка, способности восприятия учащимися содержательной стороны текста, цели коммуникативного процесса. На начальном этапе текст должен быть понятным, информативным, с четкой логикой изложения и несложной сюжетной линией. Сложность текстового материала определяется как языковыми, так и неязыковыми факторами: количеством незнакомых слов, длиной слов и предложений, сложностью повествования, смысловой нагрузкой текста, степенью понимания учащимися содержания текста. Для успешного развития речевых и коммуникативных навыков и умений при отборе текста необходимо учитывать их лексические, грамматические и культурологические характеристики, познавательную ценность. В соответствии с Государственным образовательным стандартом и Образовательной программой по русскому языку как иностранному выделяются пять сертификационных уровней общего владения русским языком как иностранным [1, 2]: уровень (включающий элементарный уровень) и четыре сертификационных.

* Исследование выполнено за счет гранта Российского научного фонда № 22-11-00051, https://rscf.ru/project/22-11-00051/

Для анализа сложности текстов применяются параметрические формулы (индекс Флеша, индекс туманности Ганнинга и др.) [3, 4], в которых числовые коэффициенты подбираются на основе существующих текстов с уже установленной сложностью. Такие формулы удобны для быстрого подсчета сложности, но могут давать не согласованные между собой результаты и не подходить для конкретных практических ситуаций.

В настоящей работе продемонстрировано, как методы теории кооперативных игр могут быть использованы для решения задачи определения сложности набора текстов, а также для кластеризации текстов из наборов. Предложен метод на основе игры голосования, позволяющий сопоставить тексту или набору текстов вектор, от которого зависит характеристика сложности данного текста на основе количества слов определенной длины.

Статья построена следующим образом. В п. 2 проведен обзор существующих методов измерения удобочитаемости текстов. В п. 3 описано построение игры голосования для текстовых документов на основе длин слов. В п. 4 рассмотрены методы определения индексов влияния игроков в заданной игре, в п. 5 приведен процесс построения векторов, характеризующих сложность текстов на основе этих индексов. Наконец, в п. 6 представлен алгоритм кластеризации текстов с использованием потенциала и произведено сравнение с экспертными оценками.

2. Индексы измерения удобочитаемости текста. Для измерения лингвистической сложности текста существуют различные индексы. Наиболее популярен индекс удобочитаемости Флеша, изначально созданный по заказу американских военных для составления текстов инструкций использования и применения оружия и оборудования в 1948 г. [5]. В настоящее время он используется в качестве оценки читаемости различных текстов:

$$\mathrm{FRE} = 206.835 - 1.015\frac{\mathrm{words}}{\mathrm{sentences}} - 84.6\frac{\mathrm{syllables}}{\mathrm{words}}, \tag{1}$$

где words — количество слов; sentences — количество предложений; syllables — количество слогов в тексте. Основными параметрами данной формулы являются среднее количество слогов в словах, мера измерения сложности слов, среднее количество слов в предложениях, мера измерения сложности предложений. Для распределения получаемых значений существует соответствующая шкала FRES (Flesch Reading Ease Scale). И. В. Оборневой была представлена адаптация формулы (1) для русского языка [6]:

$$\mathrm{FRE} = 206.835 - 1.52 \cdot \mathrm{ASL} - 65.14 \cdot \mathrm{ASW},$$

здесь ASL — средняя длина предложения в словах, ASW — средняя длина слова в слогах.

Сопоставим значение индекса удобочитаемости Флеша с группами сложности для изучения русского языка как иностранного (табл. 1).

В 1975 г. был выведен индекс Колмана — Лиау, использующийся для вычисления сложности восприятия текста путем аппроксимирования сложности текста к номеру класса в американской системе образования, ученикам которого данный текст будет понятен [7]. Формула для расчета имеет следующий вид:

$$\mathrm{CLI} = 0.0588 \cdot L - 0.296 \cdot S - 15.8,$$

Таблица 1. **Сопоставление значений индекса Флеша с группами сложности текста**

Значение индекса	Уровень читателя	Группа сложности
100.0–80.0	5–6-й класс	A1
80.0–70.0	7–8-й класс	A2
70.0–60.0	9-й класс	B1
60.0–50.0	10–11-й класс	B2
50.0–0.0	Университет	C1

где L — среднее количество букв на 100 слов; S — среднее количество предложений на 100 слов.

В случае индекса Колмана — Лиау, чем выше индекс, тем сложнее текст для чтения. Результатом является число лет обучения в американской системе образования, необходимых для понимания текста. Сопоставление значений индекса Колмана — Лиау с группами сложности представлено в табл. 2.

Таблица 2. **Сопоставление значений индекса Колмана — Лиау с группами сложности текста**

Значение индекса	Группа сложности
0.0–3.0	A1
3.0–6.0	A2
6.0–9.0	B1
9.0–12.0	B2
$>$ 12.0	C1

Рассмотрим предоставленные филологами наборы документов [8], которые используются для обучения студентов русскому языку в качестве иностранного. Документы заранее были разделены по группам сложности от наименьшей к наибольшей:

$$g = [\mathrm{A1}, \mathrm{A2}, \mathrm{B1}, \mathrm{B2}, \mathrm{C1}]. \tag{2}$$

Количество документов в группах $|\mathrm{A1}| = 31$, $|\mathrm{A2}| = 25$, $|\mathrm{B1}| = 15$, $|\mathrm{B2}| = 12$, $|\mathrm{C1}| = 36$.

Для каждого документа вычислены индексы Флеша и Колмана — Лиау и определена группа сложности на основе сопоставлений в табл. 1 и 2. Поскольку границы групп определены не явно, рассмотрим попадания в смежные по сложности группы.

Проанализировав данные табл. 3, приходим к выводу, что из всех текстов было выделено 14 со значением индекса Флеша от 100.0–80.0, только 6 из них совпали с заранее определенной группой сложности. Схожая ситуация наблюдается и для остальных групп.

Таблица 3. **Попадание текстов в корректные группы на основе индекса Флеша**

Группа	Попали в группу	Попали в смежную группу	Всего текстов
A1	6	10	31
A2	4	1	25
B1	4	3	15
B2	1	10	12
C1	8	6	36

В случае использования индекса Колмана — Лиау 41 текст имел значение индекса от 0.0 до 3.0, но только 18 из них совпали с группой A1 (табл. 4).

Таблица 4. **Попадание текстов в корректные группы на основе индекса Колмана — Лиау**

Группа	Попали в группу	Попали в смежную группу	Всего текстов
A1	18	7	31
A2	2	11	25
B1	5	6	15
B2	5	5	12
C1	0	1	36

Из полученных результатов для индексов Флеша и Колмана — Лиау можно сделать вывод, что классические индексы дают более общую оценку и могут быть неприменимы в чистом виде для оценки сложности текстов и определения их категории для изучения иностранного языка.

3. Построение игры голосования для текстовых документов. Кооперативные игры применяются во многих областях, в том числе при принятии политических решений [9]. В таких играх может быть определена сила той или иной структуры, которая отражает ее значимость при принятии общего решения. В зависимости от игры и заданной структуры индексы влияния или значимости могут показывать силу конкретного игрока при формировании коалиций. Можно применить этот подход при анализе структуры текста на сложность. Игроками в данной игре могут выступать слова, лексемы, формулы и пр. Например, в работе [10] для анализа сложности текстов на испанском языке игроками рассматривались буквы алфавита.

Определение 1. Игрой голосования называется кооперативная игра $\langle N, v\rangle$, в которой характеристическая функция принимает всего два значения: 0 и 1, $v(N) = 1$. Коалиция S, для которой $v(S) = 1$, называется выигрывающей.

Определение 2. Игрой взвешенного голосования называется кооперативная игра $\langle q, w_1, ..., w_n\rangle$, в которой характеристическая функция имеет вид

$$v(S) = \begin{cases} 1, & \text{если } w(S) \geqslant q, \\ 0, & \text{если } w(S) < q. \end{cases}$$

Построим взвешенную игру голосования $\Gamma = \langle q; w_1, w_2, ..., w_n\rangle$ для текстового документа, где w_i — количество слов длины i (веса игроков); $q = \sum w_i \cdot q_p$ — квота, необходимая для выигрыша коалиции. Значение $q_p = (0, 1)$ — это порог голосования, который положим равным 0.75. Предполагается, что q и w_i — положительные целые, $0 < w_i < q$.

Пример 1. Продемонстрируем построение игры голосования на основе одного предложения текста «Научный центр» из набора документов для обучения русскому языку в качестве иностранного категории B2. Исходный текст: «Сформулировав законы механики, он изложил учение о системе мира и установил закон всемирного тяготения». Веса слов (длины слов) в порядке их появления в предложении следующие: 13, 6, 8, 2, 7, 6, 1, 7, 4, 1, 9, 5, 10, 9.

На их основе можно определить вектор весов $\overline{w}$ и квоту q для построения игры голосования

$$\Gamma = \langle 10.5; 2, 1, 0, 1, 1, 2, 2, 1, 2, 1, 0, 0, 1\rangle.$$

Заметим, что $\sum w_i$ равно числу слов в рассматриваемом документе.

В Γ выигрывающими будут коалиции с весами 11, 12, 13, 14, в игре три игрока-болвана (w_3, w_{11}, w_{12}) и нет диктаторов, вето или мастер игроков. Минимальной выигрывающей будет коалиция размера 8, образованная игроками $w_1 = 2, w_6 = 2, w_7 =$

$2, w_9 = 2$, сумма весов которых равна 6, а также любых трех игроков в весами 1 $(w_2, w_4, w_5, w_8, w_{10}, w_{11})$.

4. Измерение индексов влияния игроков. В играх голосования можно определить количественные меры, которыми обладает каждый игрок или коалиция. Одной из таких мер является индекс влияния игрока. Для его нахождения вводится понятие «ключевой игрок». Игрок называется ключевым, если при присоединении к коалиции она становится выигрывающей. Индекс влияния игрока зависит от того, в скольких коалициях игрок ключевой.

Пример 2. Рассмотрим фрагмент из текста категории сложности А1 «Наша комната». Исходный текст: «Где стоят стол и два стула? Что стоит на столе? Что висит на стене?» На его основе построим игру голосования $\Gamma = \langle 10.5; 1, 2, 4, 1, 6\rangle$.

Для такой игры несложно вычислить все выигрывающие коалиции:

$$\begin{aligned}(w_3, w_4, w_5) &= 11,\\ (w_2, w_3, w_5) &= 12,\\ (w_2, w_3, w_4, w_5) &= 13,\\ (w_1, w_3, w_5) &= 11,\\ (w_1, w_3, w_4, w_5) &= 12,\\ (w_1, w_2, w_3, w_5) &= 13,\\ (w_1, w_2, w_3, w_4, w_5) &= 14.\end{aligned}$$

Игрок 1 будет ключевым только в коалиции $(1, 3, 5)$, игрок 2 — только в $(2, 3, 5)$, игрок 4 — только в $(3, 4, 5)$, а игроки 3 и 5 являются ключевыми во всех коалициях. Таким образом, некий вектор влияния игроков принял бы вид $\{1, 1, 7, 1, 7\}$. Существуют два классических индекса влияния (силы) игроков: индекс Пенроуза — Банцафа и индекс Шепли — Шубика.

Определение 3. Вектором Шепли — Шубика для игры голосования $\langle N, v\rangle$ будем называть вектор $\varphi(v) = (\varphi_1(v), ..., \varphi_n(v))$, где индекс i-го игрока имеет вид

$$\varphi_i(v) = \sum_{S\not\in W, S\cup\{i\}\in W} \frac{(|S|)!(n - |S| - 1)!}{n!}, \quad i = 1, ..., n,$$

где W — множество выигрывающих коалиций.

Определение 4. Вектором Банцафа в игре голосования $\langle N, v\rangle$ называется вектор $\beta(v) = (\beta_1(v), ..., \beta_n(v))$, где индекс игрока i равен

$$\beta_i(v) = \frac{\eta_i(v)}{\sum_{i\in N} \eta_i(v)}, \quad i = 1, ..., n,$$

здесь $\eta_i(v)$ — число пар коалиций $(S \cup i, S)$ таких, что коалиция $(S \cup i)$ является выигрывающей, а коалиция S нет (переключения игрока).

Для игры голосования Γ из примера 2 индексы примут следующие значения:

$$\begin{aligned}\varphi(v) &= (0.03, 0.03, 0.45, 0.03, 0.45),\\ \beta(v) &= (0.06, 0.06, 0.41, 0.06, 0.41).\end{aligned}$$

5. Использование индексов Банцафа и Шепли — Шубика для построения векторов текстов. Вычислим распределение длин слов для каждой группы (2)

Таблица 5. **Распределение длин слов для каждой группы сложности**

g_i/w_{ij}	Группы				
	A1	A2	B1	B2	C1
1	639	1013	506	770	2089
2	545	589	457	381	1961
3	570	594	502	369	2131
4	472	606	438	367	1789
5	768	933	517	569	2248
6	696	861	510	585	2050
7	516	975	418	586	1848
8	277	859	340	511	1312
9	201	559	250	386	883
10	156	492	179	388	602
11	92	256	119	232	341
12	50	143	70	201	255
13	22	125	33	106	105
14	6	53	22	66	47
15	2	25	7	35	18
16	3	10	8	17	10
17	3	9	1	13	3

сложности текстов w_{ij} для значений от 1 до 17, так как число слов большей длины для всех групп текстов ничтожно мало или равно нулю (табл. 5).

Для каждой группы документов построим взвешенную игру голосования $\Gamma_{g_i} = \langle q_i, w_{ij} \rangle : g_i \in g,\ j \in [1, ..., 17]$, где w_{ij} — количество слов длины j в группе документов i, а $q_i = \sum_j w_{ij} \cdot q_p$ — квота.

Найдем значения индексов Банцафа и Шепли — Шубика для каждой группы из g (табл. 6). Используем их как векторы, идентифицирующие группу. Так как эти векторы получены как индексы влияния игрока (длины слова) в наборе текстов, можно вычислить евклидово расстояние между группами.

Для определения принадлежности текста к группе сложности построим игру голосования Γ для конкретного текста из группы. Повторим процесс нахождения распределения длины слов и индексов влияния для каждого текста из группы. Каждому тексту поставим в соответствие группу на основе минимального евклидова расстояния. В табл. 7 приведены результаты попадания текстов на основе индекса Банцафа в корректную группу, а также в смежные по сложности группы. Для индекса Шепли — Шубика получены схожие результаты.

В сравнении с индексами Флеша (см. табл. 3) и Колмана — Лиау (см. табл. 4) представленный метод показывает лучшие результаты для заранее выявленных групп сложности. Он может быть использован для определения соответствия текста заранее заданной группе сложности. Более того, полученные векторы текстов могут быть использованы как «отпечаток» текста.

Для ранжирования текстов можно применять методы теории кооперативных игр. При этом на первом этапе каждому тексту нужно поставить в соответствие некий вектор, характеризующий данный текст. Если тексты имеют какую-то связь между собой, строится граф связей в их наборе. Затем можно провести ранжирование текстов или их кластеризацию с помощью индексов теории кооперативных игр [11–13].

6. Кластеризация текстов. Итак, каждому тексту x_i соответствует вектор Банцафа β_i в евклидовом пространстве. Разобьем это множество векторов на кластеры, используя теоретико-игровой метод кластеризации, основанный на теории гедо-

Таблица 6. **Индексы Банцафа и Шепли — Шубика для групп сложности текстов**

g_i/w_{ij}	Группы				
	A1	A2	B1	B2	C1
Индексы Банцафа					
1	0.0975	0.0803	0.0803	0.0792	0.0850
2	0.0861	0.0518	0.0733	0.0458	0.0807
3	0.0896	0.0521	0.0797	0.0442	0.0870
4	0.0755	0.0529	0.0705	0.0440	0.0729
5	0.1110	0.0754	0.0817	0.0645	0.0920
6	0.1038	0.0714	0.0808	0.0662	0.0838
7	0.0818	0.0779	0.0675	0.0662	0.0755
8	0.0361	0.0713	0.0544	0.0591	0.0496
9	0.0303	0.0499	0.0435	0.0461	0.0404
10	0.0235	0.0450	0.0346	0.0463	0.0288
11	0.0151	0.0195	0.0233	0.0281	0.0178
12	0.0087	0.0117	0.0127	0.0244	0.0125
13	0.0031	0.0105	0.0056	0.0117	0.0054
14	0.0008	0.0045	0.0039	0.0089	0.0021
15	0.0003	0.0022	0.0012	0.0043	0.0009
16	0.0004	0.0009	0.0013	0.0022	0.0004
17	0.0004	0.0009	0.0002	0.0019	0.0002
Индексы Шепли — Шубика					
1	0.1287	0.1237	0.1132	0.1407	0.1145
2	0.1068	0.0758	0.1001	0.0666	0.1052
3	0.1177	0.0763	0.1120	0.0642	0.1193
4	0.0839	0.0780	0.0973	0.0641	0.0951
5	0.1708	0.1157	0.1160	0.1028	0.1283
6	0.1476	0.1035	0.1144	0.1063	0.1125
7	0.0999	0.1204	0.0940	0.1063	0.0984
8	0.0441	0.1033	0.0751	0.0933	0.0644
9	0.0363	0.0710	0.0676	0.0669	0.0598
10	0.0284	0.0639	0.0521	0.0671	0.0464
11	0.0187	0.0268	0.0295	0.0400	0.0263
12	0.0111	0.0157	0.0155	0.0343	0.0182
13	0.0048	0.0141	0.0058	0.0202	0.0073
14	0.0005	0.0062	0.0045	0.0155	0.0031
15	0.0002	0.0031	0.0014	0.0057	0.0007
16	0.0003	0.0013	0.0014	0.0031	0.0003
17	0.0003	0.0012	0.0001	0.0027	0.0001

Таблица 7. **Попадание текстов в корректные группы на основе евклидова расстояния между векторами Банцафа**

Группа	Попали в корректную группу	Попали в смежную группу	Всего текстов
A1	25	0	31
A2	15	4	25
B1	4	8	15
B2	7	2	12
C1	11	0	36

нических игр [14], и сравним с экспертными оценками. В таком случае тексты становятся игроками и вводится отношение между ними. В зависимости от этого игроки могут предпочитать нахождение в той или иной коалиции или, наоборот, отказываться от нее. Коротко опишем теоретико-игровую модель.

Предположим, что множество игроков $N = \{1, \ldots, n\}$ разбито на K коалиций в виде $\pi = \{S_1, \ldots, S_K\}$. Пусть $S_\pi(i)$ определяет коалицию $S_k \in \pi$ такую, что $i \in S_k$. Гедоническая игра устанавливается посредством предпочтений игроков для нахождения в различных коалициях. Предпочтения игрока i выражаются через полное рефлексивное и транзитивное бинарное отношение $\succeq_i$ на множестве $\{S \subset N : i \in S\}$. Тогда задача коалиционного разбиения игроков может быть решена с помощью аддитивно сепарабельных предпочтений [14].

Предпочтения игроков являются аддитивно сепарабельными [14], если существует такая функция $v_i : N \to \mathbb{R}$, что $v_i(i) = 0$ и

$$S_1 \succeq_i S_2 \Leftrightarrow \sum_{j \in S_1} v_i(j) \geqslant \sum_{j \in S_2} v_i(j).$$

Предпочтения $\{v_i, i \in N\}$ симметричные, если $v_i(j) = v_j(i) = v_{ij} = v_{ji}$ для всех $i, j \in N$.

Коалиционное разбиение π называется устойчивым по Нэшу, если $S_\pi(i) \succeq_i S_k \cup \{i\}$ для всех $i \in N,\ S_k \in \pi \cup \{\emptyset\}$. В устойчивом разбиении никому из игроков не выгодно покидать свою коалицию. В работе [14] показано, что устойчивое коалиционное разбиение можно найти, максимизируя потенциальную функцию. Потенциал для коалиционного разбиения $\pi = \{S_1, \ldots, S_K\}$ имеет вид

$$P(\pi) = \sum_{k=1}^{K} P(S_k) = \sum_{k=1}^{K} \sum_{i,j \in S_k} v_{ij}. \tag{3}$$

Положим, что

$$v_{ij} = v_{ji} = \begin{cases} 1, & \text{если } |x_i - x_j| < \epsilon, \\ -1, & \text{иначе.} \end{cases}$$

Это значит, что если игроки близки друг другу, то это поощряется в коалиции, в противном случае они наказываются. Величину ϵ выберем в зависимости от среднего значения евклидова расстояния между векторами ρ.

Для максимизации данной функции можно применить метод отжига, на каждом этапе выбирая перестановку вектора текста в другую группу сложности и оценивая изменение потенциала. Алгоритм начинается с предоставленного экспертами разбиения на 5 кластеров $\pi = \{\mathrm{A1}, \mathrm{A2}, \mathrm{B1}, \mathrm{B2}, \mathrm{C1}\}$. Поскольку игра имеет потенциал, приведенный выше алгоритм гарантированно сходится за конечное число шагов.

Большое влияние на значение функции оказывает величина ϵ, так как на ее основе строится матрица близости $A = |v_{ij}|$. Рассмотрим результаты максимизации потенциала в зависимости от ϵ.

В качестве первого значения ϵ выберем среднее расстояние между векторами, поделенное на количество изначальных кластеров. На рис. 1 видно, что матрица является разряженной, т. е. очень малое количество пар векторов имеют меру схожести, равную 1.

Примем, что значение ϵ равно половине среднего расстояния (рис. 2).

Как видно на рис. 3, при заданном базовом количестве итераций 500 алгоритм выполняет около 2000 итераций, при этом не происходит отката в худшее положение по вероятностному критерию. В результате отжига потенциал увеличится с 298 до 698.

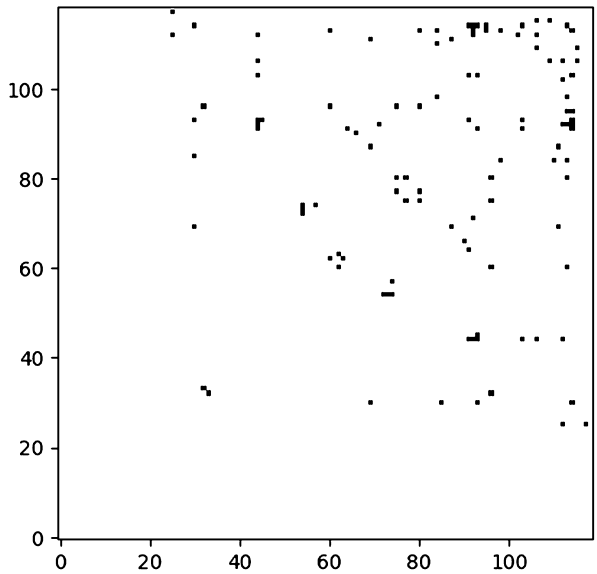

Рис. 1. Значения матрицы A при $\epsilon = \frac{\rho}{5}$

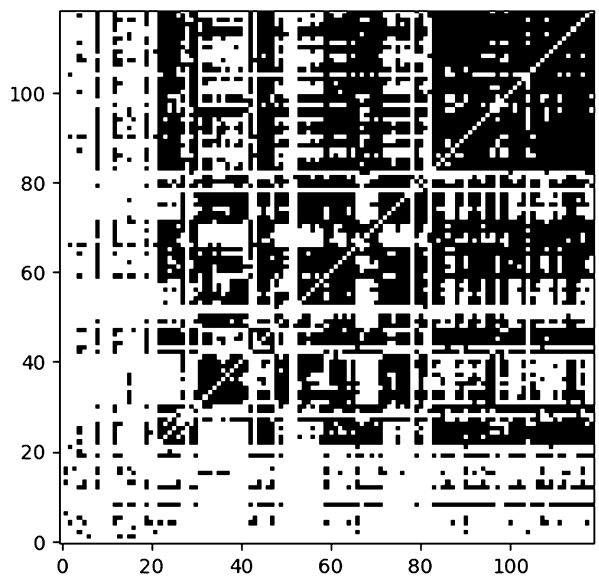

Рис. 2. Значения матрицы A при $\epsilon = 0.5 \cdot \rho$

Рассмотрим, как изменяется изначальное распределение текстов по группам сложности после проведения кластеризации:

Группа.....	A1	A2	B1	B2	C1
Исходное распределение.....	31	25	15	12	36
Распределение после отжига..	20	22	26	12	39

Представив общее количество текстов в исходных и полученных группах при $\epsilon = 0.5 \cdot \rho$, можно заметить, что часть текстов из A1, A2 и C1 были перенесены в дру-

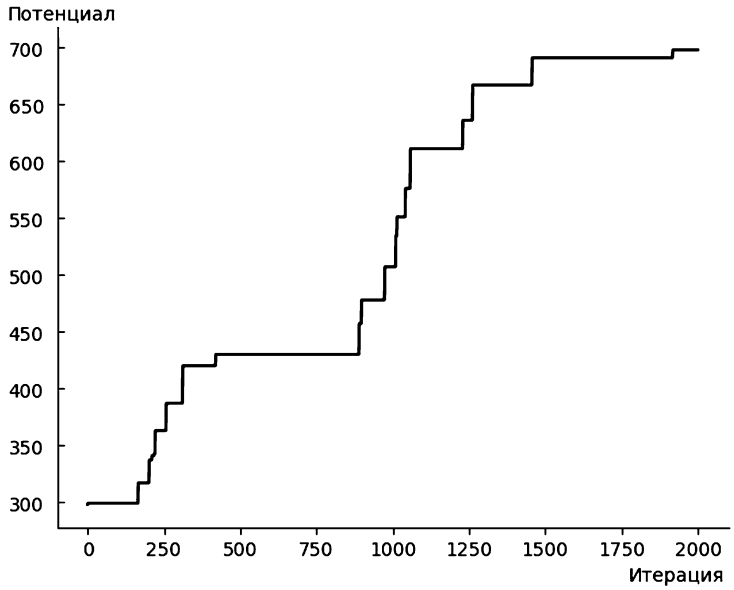

Рис. 3. Изменение потенциала при отжиге с $\epsilon = 0.5 \cdot \rho$

гие группы сложности. Опишем подробнее тексты, для которых в процессе кластеризации изменилась группа сложности. В табл. 8 представлены тексты из [8], для которых произошло данное изменение. Например, часть текстов из группы A1 были

Таблица 8. **Переопределение групп текстов на основе результата кластеризации**

Текст из набора [8]	Исходная группа	Полученная группа
A1_7_3_Паровозик.txt	A1	B1
A1_7_7_Боксер.txt	A1	B1
A1_Алмаз.txt	A1	B1
A1_Взгляд_в_окно.txt	A1	B1
A1_Жизнь_это_эхо.txt	A1	B1
A1_Зощенко.txt	A1	B1
A1_Мальчик.txt	A1	B1
A1_Олимпийские_игры.txt	A1	A2
A1_У_парикмахера.txt	A1	B1
A1_Цепочка_любви.txt	A1	B1
A2_Биология_5.txt	A2	B2
A2_Вундеркинды.txt	A2	B1
A2_Злой_мальчик.txt	A2	B1
A2_Карл_Фаберже.txt	A2	B2
A2_Наслаждайтесь_жизнью.txt	A2	B1
A2_С_женой_поссорился.txt	A2	B1
B1_13_02_Эрарта.txt	B1	A2
B1_13_11_Борис_и_Марина.txt	B1	C1
B1_13_12_Муж_и_жена.txt	B1	C1

перенесены в группу B1, а это может свидетельствовать о том, что в соответствии с построенной на основе длин текстов игре голосования они принадлежат к группе с более высоким уровнем сложности. Такие данные могут быть переданы на рассмотрение экспертам по составлению программ обучения русскому языку в качестве иностранного для уточнения категории сложности текстов.

Изменяя различные параметры алгоритма кластеризации, можно добиться уточнения распределения текстов по группам на основе изначально заданной игры голосования по длине слов.

7. Заключение. Решение задачи определения сложности текста является важным этапом при подготовке материалов для обучения иностранным языкам. Такие материалы в большинстве случаев подбираются вручную, при этом учитываются как семантические, так и лингвистические и синтаксические особенности текстов, например длина слов, сложность предложений для восприятия или смысловая нагрузка текста.

Существующие методы определения сложности текстов могут неадекватно оценивать тексты. В статье это демонстрируется на текстах, применяемых в преподавании русского языка как иностранного. Возможным подходом может быть описанный в работе метод на основе теории кооперативных игр. Метод определяет индекс влияния длин слов в тексте, которые рассматриваются как игроки в некоторой игре голосования. Таким образом, каждому тексту ставится в соответствие некоторый вектор значений. После этого в пространстве векторов проводится ранжирование текстов.

Полученные результаты показывают применимость методов теории игр при анализе текстов на сложность их восприятия и кластеризации — возможного уточнения группы сложности для конкретного текста, что в дальнейшем может быть учтено специалистами по изучению русского языка в качестве иностранного для уточнения его принадлежности к определенной категории из набора A1–C1.

Литература

1. Государственный образовательный стандарт по русскому языку как иностранному. Первый уровень. Второй уровень. Профессиональные модули / сост. Н. П. Андрюшина, Т. Е. Владимирова, Л. П. Клобукова. СПб.: Златоуст, 2000. 56 с.

2. Образовательная программа по русскому языку как иностранному. Предвузовское обучение / сост. З. И. Есина, А. С. Иванова, Н. И. Соболева и др. М.: Изд-во Российского университета Дружбы народов им. Патриса Лумумбы, 2001. 137 с.

3. *Майер Р. В.* Дидактическая сложность учебных текстов и ее оценка. Глазов: Изд-во Глазовского государственного педагогического университета, 2020. 149 с.

4. *Gunning R.* The technique of clear writing. New York: McGraw-Hill, 1952. 289 p.

5. *Flesch R.* A new readability yardstick // Journal of Applied Psychology. 1948. N 3. P. 221–233.

6. *Оборнева И. В.* Математическая модель оценки учебных текстов // Вестник Московского государственного педагогического университета. Сер. Информатика и информатизация образования. 2005. № 1 (4). С. 141–147.

7. *Coleman M., Liau T. L.* A computer readability formula designed for machine scoring // Journal of Applied Psychology. 1975. N 60. P. 283–284.

8. Тексты для обучения русскому языку в качестве иностранного. URL: https://github.com/arkty/ru_learning_data (дата обращения: 14 августа 2023 г.).

9. *Мазалов В. В.* Математическая теория игр и приложения: учеб. пособие. 2-е изд., стер. СПб.: Лань, 2016. 448 с.

10. *Molinero X., Laamiri A., Riquelme F.* Readability and power indices // The Fifteenth International Conference on Game Theory and Management (GTM 2021). St. Petersburg, 2021. P. 7.

11. *Мазалов В. В., Хитрая В. А., Хитрый А. В.* Методы теории кооперативных игр в задаче ранжирования текстов // Вестник Санкт-Петербургского университета. Прикладная математика. Информатика. Процессы управления. 2022. Т. 18. Вып. 1. С. 63–78. https://doi.org/10.21638/11701/spbu10.2022.105

12. *Kondratev A. A., Mazalov V. V.* Tournament solutions based on cooperative game theory // International Journal of Game Theory. 2020. Vol. 49. P. 119–145.

13. *Алескеров Ф. Т., Хабина Е. Л., Шварц Д. А.* Бинарные отношения, графы и коллективные решения. Примеры и задачи: учеб. пособие для вузов. М.: Юрайт, 2023. 458 с.

14. *Bogomolnaia A., Jackson M. O.* The stability of hedonic coalition structures // Games Econ. Behav. 2002. Vol. 38. N 2. P. 201–230.

Статья поступила в редакцию 14 сентября 2023 г.
Статья принята к печати 12 октября 2023 г.

Контактная информация:

Хитрый Андрей Владимирович — аспирант; andrey.khitryy@gmail.com

Мазалов Владимир Викторович — д-р физ.-мат. наук, проф.; vmazalov@krc.karelia.ru

Буре Наталья Анатольевна — канд. филол. наук, доц.; nataly.bure@gmail.com

Дробная Полина Васильевна — severnayapol@mail.ru

Cooperative game theory methods for determining text complexity*

A. V. Khitryy[1], *V. V. Mazalov*[1,2], *N. A. Bure*[2], *P. V. Drobnaya*[3]

[1] Karelian Research Center of the Russian Academy of Sciences, 11, Pushkinskaya ul., Petrozavodsk, 185910, Russian Federation
[2] St. Petersburg State University, 7–9, Universitetskaya nab., St. Petersburg, 199034, Russian Federation,
[3] Petrozavodsk State University, 33, ul. Lenina, Petrozavodsk, 185910, Russian Federation

For citation: Khitryi A. V., Mazalov V. V., Bure N. A., Drobnaya P. V. Cooperative game theory methods for determining text complexity. *Vestnik of Saint Petersburg University. Applied Mathematics. Computer Science. Control Processes*, 2023, vol. 19, iss. 4, pp. 509–521. https://doi.org/10.21638/11701/spbu10.2023.407 (In Russian)

We propose a method for estimating the complexity of texts based on the methods of cooperative game theory. The players in this game are the lengths of words in the text. The game itself is represented as a voting game in which the value of a player is determined by the number of coalitions in which the player is key. The ranks of the players are formed by computing the Shepley — Schubik value or the Banzaf index in a voting game with a given voting threshold. Thus, a vector of Shepley — Schubik or Banzaf values is assigned to each text. After that, the vector space is used to rank the texts in terms of complexity based on the expert evaluations obtained in this domain.

Keywords: text processing, voting game, Shepley — Schubik value, Banzaf power index, klasterization.

References

1. *Gosudarstvennyj obrazovatel'nyj standart po russkomu yazyku kak inostrannomu. Pervyj uroven'. Vtoroj uroven'. Professional'nye moduli* [*State educational standard for Russian as a foreign language. First level. Second level. Professional modules*]. Compilers: N. P. Andryushina, T. E. Vladimirova, L. P. Klobukova. St. Petersburg, Zlatoust Publ., 2000, 56 p. (In Russian)

2. *Obrazovatel'naya programma po russkomu yazyku kak inostrannomu. Predvuzovskoe obuchenie* [*Educational program on Russian as a foreign language. Pre-university education*]. Compilers: Z. I. Esina, A. S. Ivanova, N. I. Soboleva. Moscow, Patrice Lumumba Peoples' Friendship University of Russian Press, 2001, 137 p. (In Russian)

3. Majer R. V. *Didakticheskaya slozhnost' uchebnyh tekstov i ee ocenka* [*Didactic complexity of educational texts and its assessment*]. Glazov, Glazovskiy State Pedagogical University Press, 2020, 149 p. (In Russian)

4. Gunning R. *The technique of clear writing.* New York, McGraw-Hill Publ., 1952, 289 p.

* This work was founded by the Russian Science Foundation (project N 22-11-00051, https://rscf.ru/project/22-11-00051/).

5. Flesch R. A new readability yardstick. *Journal of Applied Psychology*, 1948, no. 3, pp. 221–233.

6. Oborneva I. V. Matematicheskaya model' ocenki uchebnyh tekstov [A mathematical model for evaluating instructional texts]. *Vestnik of Moscow State Pedagogical University. Series Information and Informatization of education*, 2005, no. 1 (4), pp. 141–147. (In Russian)

7. Coleman M., Liau T. L. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 1975, no. 60, pp. 283–284.

8. *Teksty dlya obucheniya russkomu yazyku v kachestve inostrannogo* [*Texts for teaching Russian as a foreign language*]. Available at: https://github.com/arkty/ru_learning_data (accessed: August 14, 2023). (In Russian)

9. Mazalov V. V. *Matematicheskaya teoriya igr i prilozheniya.* Uchebnoe posobie. 2-e izd. [*Mathematical game theory and applications.* Textbook]. 2nd ed. St. Petersburg, Lan' Publ., 2016, 448 p. (In Russian)

10. Molinero X., Laamiri A., Riquelme F. Readability and power indices. *The Fifteenth International Conference on Game Theory and Management (GTM 2021)*. St. Petersburg, 2021, p. 7.

11. Mazalov V. V., Khitraya V. A., Khitryj A. V. Metody teorii kooperativnyh igr v zadache ranzhirovaniya tekstov [Methods of cooperative game theory in the task of text ranking]. *Vestnik of Saint Petersburg University. Applied Mathematics. Computer Science. Control Processes*, 2022, vol. 18, iss. 1, pp. 63–78. https://doi.org/10.21638/11701/spbu10.2022.105 (In Russian)

12. Kondratev A. A., Mazalov V. V. Tournament solutions based on cooperative game theory. *International Journal of Game Theory*, 2020, vol. 49, pp. 119–145.

13. Aleskerov F. T., Habina E. L., Shvarc D. A. *Binarnye otnosheniya, grafy i kollektivnye resheniya. Primery i zadachi.* Uchebnoe posobie dlya vuzov [*Binary relations, graphs and collective solutions. Examples and problems.* Textbook for universities]. Moscow, Yurite Publ., 2023, 458 p. (In Russian)

14. Bogomolnaia A., Jackson M. O. The stability of hedonic coalition structures. *Games Econ. Behav.*, 2002, vol. 38, no. 2, pp. 201–230.

Received: September 14, 2023.
Accepted: October 12, 2023.

Authors' information:

Andrei V. Khitryi — Postgraduate Student; andrey.khitryy@gmail.com

Vladimir V. Mazalov — Dr. Sci. in Physics and Mathematics, Professor; vmazalov@krc.karelia.ru

Natalia A. Bure — PhD in Philology, Associate Professor; nataly.bure@gmail.com

Polina V. Drobnaya — severnayapol@mail.ru