УДК 81’32 DOI: 10.17673/vsgtu-pps.2019.3.13

АВТОМАТИЧЕСКАЯ ОБРАБОТКА ТЕКСТОВ; ТЕМАТИЧЕСКАЯ СЕГМЕНТАЦИЯ УЧЕБНЫХ ТЕКСТОВ

М.И. Солнышкина[1], И.Э. Ярмакеев[2], Э.В. Гафиятова[3], Ф.Х. Исмаева[4]
[1,2,3,4]Казанский (Приволжский) федеральный университет
Россия, 420008, г. Казань, ул. Кремлевская, 18
[1]E-mail: mesoln@yandex.ru
[2]E-mail: ermakeev@mail.ru
[3]E-mail: rg-777@yandex.ru
[4]E-mail: fismaeva@yandex.ru

Аннотация
Статья посвящена проблеме автоматического количественного определения сложности и тематической сегментации текстов. Дана краткая характеристика положения дел в данной области, показано, что существующие формулы расчета индекса читабельности являются жанрозависимыми и утрачивают достоверность при их использовании для текстов другого жанра. На основе корпуса учебных текстов и анализа количественных параметров авторы предлагают новый способ анализа соответствия текста лингвистическим способностям школьников. Исследование осуществлено на материале УМК Spotlight 11, общий объем корпуса составил 38 текстов с суммарным объемом 12891 словоупотреблений. В качестве методов использованы тематическая сегментация, компонент-анализ, метод статистического анализа, в работе применялись формулы читабельности Флеш – Кинкейда для англоязычных текстов, программы автоматизированной обработки текстов Coh-Metrix, WebFX, MonkeyLearn. Оценка сложности текстов показала, что динамика текстов с заданиями такова: на ознакомительное чтение – от более сложных к менее сложным (–0.2); на просмотровое чтение – от менее сложных к более сложным (+0.4); в текстах на полное понимание прочитанного индекс сложности поднялся на 5.2. Тематическая сегментация, осуществленная при помощи программы Monkey Learn, выявила лексику 15 тем, которые в течение учебного года предлагаются учащимся для изучения в среднем 3–5 раз. Наиболее частотной является тема Humanities (гуманитарные науки), обращение к которой выявлено в 9 модулях. Показательно, что к темам Gardening (садоводство), Computers & Internet (компьютер и интернет), Science & Mathematics (наука и математика), Entertainment & Recreation (развлечения) авторы учебника обращаются только в одном модуле.

Ключевые слова: сложность текстов, читабельность, математические модели, английский язык, тематическая сегментация.

[1] *Солнышкина Марина Ивановна*, доктор филологических наук, профессор кафедры «Теория и практика преподавания иностранных языков».

[2] *Ярмакеев Искандер Энгелевич*, доктор педагогических наук, профессор кафедры «Билингвальное и цифровое образование»

[3] *Гафиятова Эльзара Василовна*, кандидат филологических наук, доцент кафедры «Теория и практика преподавания иностранных языков».

[4] *Исмаева Фарида Хамисовна*, кандидат филологических наук, доцент кафедры «Иностранные языки».

Введение

Современные наукометрические исследования показывают, что объем генерируемой информации в настоящее время настолько велик, что она почти не достигает адресата [1]. В качестве одной из причин ученые указывают на несоответствие сложности текста сообщения когнитивным и лингвистическим способностям целевой аудитории [2]. Именно поэтому весьма актуальной в современном мире является проблема оценки сложности текста, определение его «трудности» или «понятности» для конкретной аудитории, его читабельности. Особую значимость данная проблема обретает в системе образования: завышение уровня читабельности текста ведет к снижению объема воспринимаемого текста, занижение – замедляет развитие школьников [3, 4, 5]. До недавнего времени в качестве основного инструмента при оценке соответствия текста читателю использовалась экспертная оценка квалифицированных специалистов, учителей, авторов учебников [6].

В современной системе образования актуальной является задача осуществления автоматической экспертизы учебных текстов [7, 8], а проблемы автоматической оценки сложности и тематической сегментации текстов в последние годы привлекают внимание широкого круга ученых [9, 10]. Сфера образования рассматривается наиболее важной областью применения данных технологий, поскольку именно в образовании особо значимо соответствие учебных текстов когнитивным и лингвистическим способностям школьников. Возможность автоматически определять сложность текстов позволит авторам учебников точнее ориентироваться на целевую аудиторию [11].

В нашей стране автоматическая оценка сложности текстов постепенно становится важным инструментом не только индивидуализации и персонификации образовательного процесса [12], но и оценивания работ школьников и студентов [13]. Сложно переоценить роль автоматизированных инструментов анализа сложности текста для экспертизы учебников [14]. Особую значимость проблема создания инструментов автоматической оценки сложности учебников обрела в настоящее время: проводимые международной организацией PISA исследования способности старшеклассников к усвоению прочитанного материала выявили неудовлетворительную подготовку российских школьников. В 2015 г. в числе 72 обследованных стран по оценке читательской грамотности Россия заняла 26-е место [15]. Правительством России поставлена задача в течение 5 лет войти в десятку лучших стран по этому показателю [16].

1. Обзор литературы

Для английского языка проблема сложности текстов имеет более длительную историю и обсуждалась в ряде работ [17, 18, 19]. К настоящему времени предложено более 200 формул читабельности, при помощи которых рассчитываются «индексы читабельности» англоязычного текста [9]. Наибольшую известность приобрела формула Флеша – Кинкейда: FKG = 0.39 ASL + 11.8 ASW − 15.59, где FKG (Flesch-KincaidGradeLevel) – сложность текста, ASL – средняя длина предложения (в словах), ASW – средняя длина слов (в слогах) [17]. Считается, что индекс Флеша (FKG/FKGL) академического текста должен соответствовать году обучения в школе или университете США. Инструменты автоматической оценки сложности текстов традиционно используются для создания корпусов текстов для чтения лиц с определенным уровнем владения языком [9, 21].

Проблема оценки сложности текста изучалась и для ряда европейских языков. В последнее десятилетие инструменты автоматической оценки сложности текста созданы и для восточных языков: корейского, китайского, японского [20].

Однако при всей своей значимости, общедоступности и экономичности [22] формулы читабельности опираются только на ограниченное число количественных параметров (количество знаков тексте; число слов, предложений; количество слов с более чем 4 слогами; среднее число слов в предложении; среднее число слогов в предложении; процент сложных слов) и не могут представить весь спектр параметров сложности [23].

Общество нуждается в инструментах, которые осуществляют более глубокий лингвостатистический анализ текста, а также его ранжирование с возрастом или количеством лет формального образования. Для английского языка в настоящее время созданы и успешно используются программные комплексы с возможностями анализа широкого спектра параметров текстов: Coh-Metrix, TAACO, WebFX, MonkeyLearn, iSTART и др. В англоязычном обществе эти программные средства нашли применение не только в сфере образования [24]; при их помощи, например, оценивается уровень сложности речей президентов США [25], произведений художественной литературы [26] и даже различия в уровне сложности британских и американских юридических документов [27]. Современные автоматизированные средства оценки качественных и количественных параметров текстов хорошо описаны в современной литературе [17, 18, 28].

Coh-Metrix, разработанная коллективом американских ученых под руководством профессора А. Грейссера, ориентирована на прогнозирование удобочитаемости текста и анализ объема информации в тексте и опирается на комплекс математических формул [28]. Основу программы Coh-Metrix составили результаты исследований Ассоциации по критериям в прикладных науках (Touchstone Applied Science Associates Inc., TASA), занимающейся анализом академических текстов и создавшей корпус, включающий 11 миллионов слов, 119 627 фрагментов 37 651 текст. В корпусе представлены тексты из различных сфер: филологии (language arts), общественных (social studies) и естественных (science) наук, истории (history), здравоохранения (health), бизнеса (business), домоводства (home economics), прикладного искусства (industrial arts) [29]. Тексты в корпусе TASA распределены по 13 уровням, каждый из которых соответствует этапу академической подготовленности читателя – от детского сада до высшего учебного заведения. Определение этапа академической подготовленности читателя (Degrees of Reading Power [30]) производится с учетом количественных параметров, а результат выводится на основе индекса читабельности по одной из двух формул: удобочитаемость по Флешу или уровень Флеша – Кинкейда [20].

Наиболее значимыми в ракурсе представленного исследования являются два типа инструментов: программы оценки сложности или читабельности текста [9] и программы тематического анализа или сегментации текста. Именно тематическая сегментация текста в настоящее время рассматривается как одна из ведущих технологий обработки естественного языка (Natural Language Processing, NLP), позволяющая автоматически извлекать смысл из текстов и выявлять темы и подтемы внутри текста [9].

Тематическое моделирование как одна из форм статистического анализа текстов разрабатывается с конца 90-х гг. прошлого века и представляет собой «группирование

документов» на основе общих тем. При этом одно и то же слово в зависимости от его контекста(ов) может быть определено как принадлежащее к одной или нескольким темам, образуя таким образом «мягкую кластеризацию» (soft clustering). Именно поэтому тематические модели также именуются моделями би- или поликластеризации. Например, значение слова nucleus (ядро) может быть определено только на основе доминирующей темы: математика, физика, биология, военная история и проч. [9].

2. Материалы и методы

Представленное исследование осуществлено на материале одного из первых российских УМК, созданных издательством «Просвещение» (Россия) совместно с издательством Express Publishing (Великобритания) [31, 32], – УМК Spotlight 11. Отличительной особенностью учебника является наличие аутентичного материала о России. Учебник получил положительные заключения Российской академии наук и Российской академии образования на соответствие федеральному компоненту Государственного образовательного стандарта среднего (полного) общего образования [32]. Общий объем рассматриваемого корпуса составил 38 текстов суммарным объемом 12891 словоупотребление. Исследование осуществлено с использованием методики тематической сегментации, метода статистического анализа, в работе применялись формулы читабельности Флеш – Кинкейда для англоязычных текстов, программы автоматизированной обработки текстов Coh-Metrix, WebFX, MonkeyLearn.

3. Результаты исследования

На первом этапе исследования было осуществлено распознавание всех текстов УМК Spotlight (11-й класс) в формате txt и последующее определение их уровней сложности с помощью программы Coh-Metrix (рис. 1).

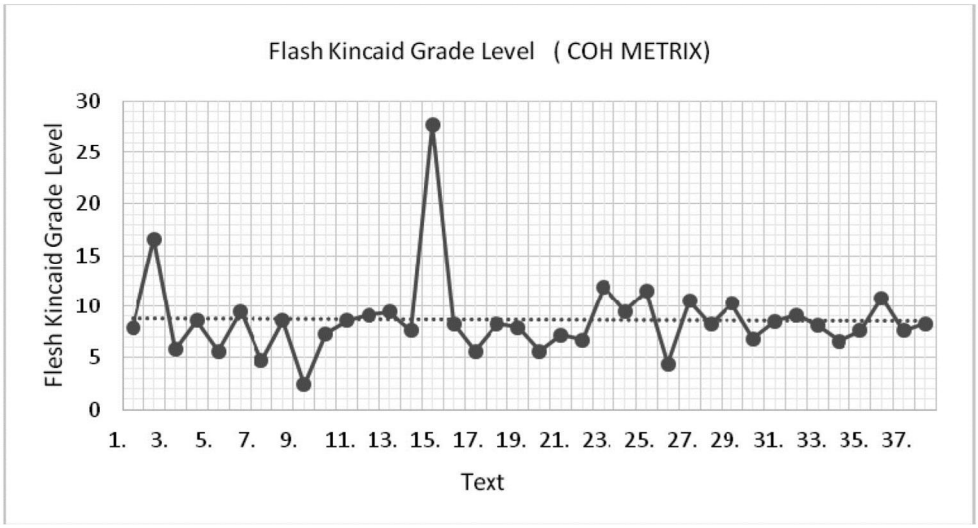

Рис. 1. Сложность текстов УМК Spotlight 11

Как показало исследование, индекс сложности изученных текстов изменяется в пределах между сложностью 2-го уровня (минимальный уровень, текст 10) и 28-го уровня (максимальный уровень, текст 16). Средний уровень сложности изученных текстов составил около 7,5, что соответствует 7–8-му классу американской школы.

На втором этапе исследования все тексты УМК Spotlight 11 были классифицированы в три группы в зависимости от типа заданий, выполняемых учащимися при чтении конкретного текста: 1) ознакомительные тексты (skimming), для которых достаточно понимания 70 % текста; 2) тексты для просмотрового чтения (scanning), нацеленные на извлечение определенной информации (дата, время, главные герои и др.); 3) тексты, нацеленные на полное понимание прочитанного (reading for detailed comprehension). В УМК Spotlight 11 для первого типа заданий (ознакомительное чтение) используются 13 текстов (3, 6, 8, 10, 14, 18, 21, 24, 25, 28, 32, 34, 35). Для второго типа (просмотровое чтение) – 12 текстов (1, 4, 13, 15, 16, 20, 23, 27, 29, 33, 37, 38). Для третьего типа (полное понимание прочитанного) – 13 текстов (2, 5, 7, 9, 11, 12, 17, 19, 22, 26, 30, 31, 36).

На третьем этапе исследования уровень сложности определялся для текстов каждой группы. Данные для каждой группы представлены на рис. 2, 3, 4.

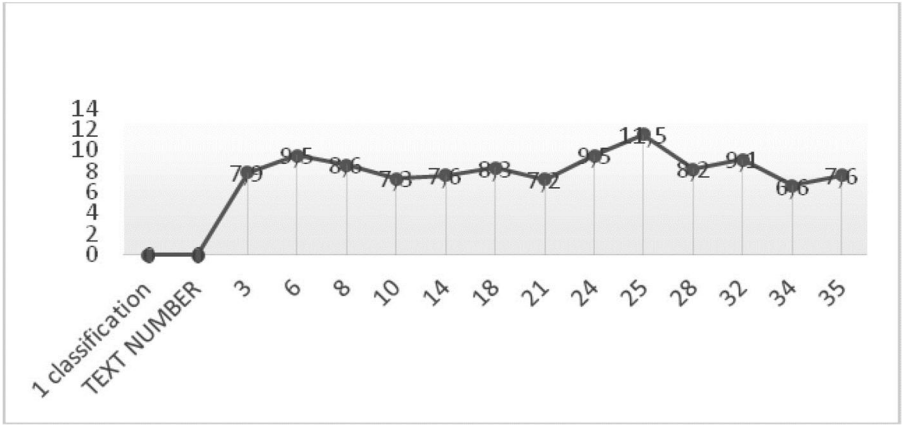

Рис. 2. Уровень сложности текстов для заданий на ознакомительное чтение

Как видим, в целом уровень сложности текстов в заданиях на ознакомительное чтение (см. рис. 2) не поднялся, а наоборот снизился на три десятых: от 7,9 в тексте 3 до 7,6 в 35 тексте.

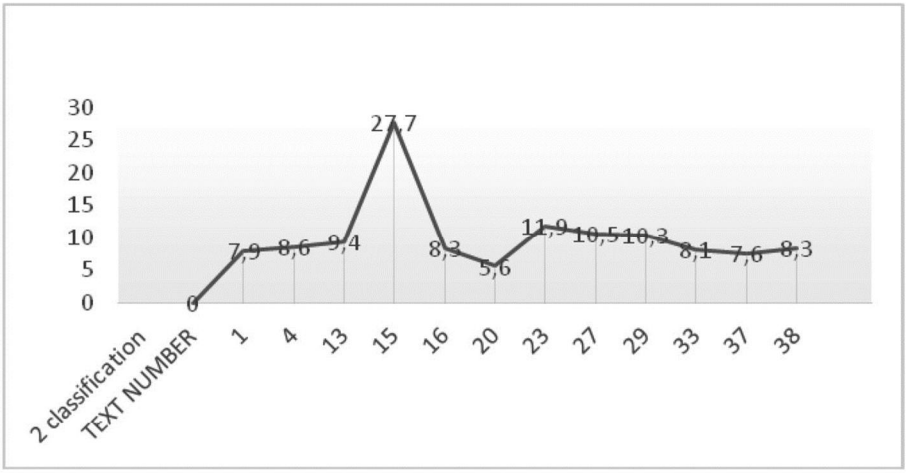

Рис. 3. Уровень сложности текстов для заданий на просмотровое чтение

В заданиях на просмотровое чтение (см. рис. 3) уровень сложности поднялся на четыре десятых (от 7,9 в тексте 1 до 8,3 в тексте 8.3). При этом весьма сложным является текст 15 (FKGL 27.7), соответствующий уровню студентов – носителей языка 23–24 лет [33, 34].

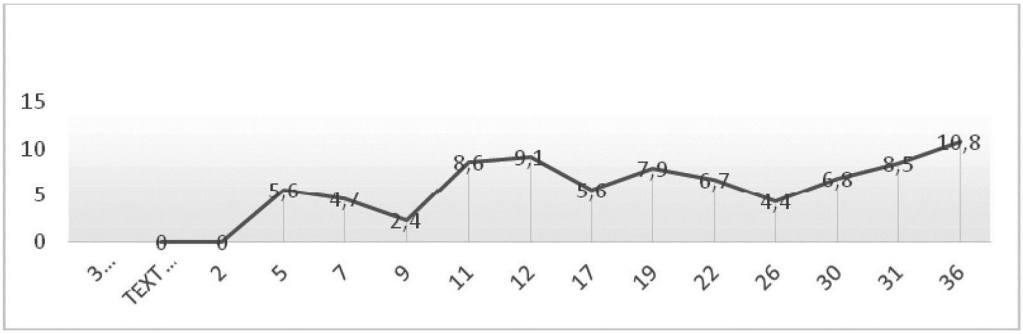

Рис. 4. Уровень сложности текста для заданий на полное понимание прочитанного

В заданиях на полное понимание прочитанного (см. рис. 4) уровень сложности поднялся с 5,6 в тексте 5 до 10,8 в тексте 36. Показательно, что в течение года наблюдается значительное повышение уровня сложности учебных текстов.

На четвертом этапе исследования при помощи программы Coh-Metrix был осуществлен анализ качественных параметров текстов: нарративности/повествовательности (Narrativity), синтаксической простоты (Syntactic Simplicty), конкретности слова (Word Concreteness), референциальной связности слов (Referential Cohesion) и глубинной связности слов (Deep Cohesion) (табл. 1).

Таблица 1

Качественные параметры сложности текста УМК Spotlight 11, %

Текст	Повествовательность (Narrativity)	Синтаксическая простота (Syntactic simplicity)	Конкретность (Concreteness)	Референциальная связность (Referential Cohesion)	Глубинная связность (Deep Cohesion)
1	80	40	33	27	79
2	73	2	99	96	35
3	90	37	80	26	29
4	67	17	29	16	79
5	94	53	78	21	86
6	17	38	77	53	75
7	52	91	59	2	98
8	64	54	18	30	62
9	81	69	99	16	79
10	61	58	59	22	91
11	26	50	95	7	37
12	33	58	63	11	85
13	50	32	25	9	24
14	70	62	72	27	94
15	96	2	98	81	87
16	26	39	94	16	4
17	57	71	87	5	79
18	26	39	94	16	4

Окончание табл. 1

Текст	Повествовательность (Narrativity)	Синтаксическая простота (Syntactic simplicity)	Конкретность (Concreteness)	Референциальная связность (Referential Cohesion)	Глубинная связность (Deep Cohesion)
19	59	66	90	12	99
20	85	61	73	20	92
21	71	38	75	20	70
22	40	53	79	2	70
23	5	45	85	7	81
24	35	31	82	11	85
25	64	21	51	17	93
26	89	77	96	23	70
27	53	58	35	44	86
28	45	79	71	10	95
29	15	70	10	8	81
30	85	67	31	9	95
31	65	33	77	14	59
32	32	35	64	2	91
33	39	52	80	23	40
34	26	63	74	2	22
35	9	46	79	28	21
36	15	40	75	59	85
37	21	41	76	4	35
38	9	74	57	9	85

Как видим, тексты 35 и 38 отличаются особо низким уровнем повествовательности, что говорит о том, что данные тексты содержат крайне низкий уровень сюжетообразующих элементов, таких как имена героев, названия действий, мест и обстоятельств действий. Несмотря на низкий уровень повествовательности, индекс сложности (FKGL) в текстах равен 7,6 и 8,3 соответственно.

Определив сложность, количественные и качественные параметры каждого текста в отдельности и объединив их в группы, выделяем ряд данных.

4 текста (6, 23, 35 и 38) имеют низкий уровень повествовательности (Narrativity) – менее 20 %, что свидетельствует о потенциальных трудностях понимания их учениками. Повествовательность текста обуславливается количеством присутствующих в нем сюжетообразующих элементов – персонажей и событий, а также его лексическим составом [35].

4 текста (2, 4, 15, 25) имеют низкий уровень синтаксической простоты. Синтаксическая простота выявляется с помощью трех переменных: 1) число грамматических основ в предложении: чем больше предложений с несколькими грамматическими основами, тем сложнее текст; 2) количество слов в предложении: чем длиннее предложение, тем выше его сложность; 3) количество слов в предложении перед главным сказуемым: чем дальше такое сказуемое удалено от начала предложения, тем текст сложнее [36].

2 текста (8, 29) имеют низкий уровень конкретности слов. Для определения конкретности слов использована база данных MRC Psycholinguistic Database [37].

36 текстов из 38 (94,7 %) имеют низкий уровень референциальной связности слов. В текстах 4, 7, 9, 11, 12, 13, 16, 17, 18, 19, 22, 23, 24, 25, 28, 29, 30, 31, 32, 34, 37, 38 процент референциальной связности менее 20 %. Референциальная связность текста в программе Coh-Metrix определяется с помощью повторов слов и синонимических замен знаков, относящихся к одному и тому же референту.

11 текстов имеют низкий уровень глубинной связности слов. В текстах 16, 18 процент связности слов менее 20 %.

Как видим, динамика увеличения параметров сложности текстов в УМК Spotlight 11 – небольшая, не превышает единицы.

На пятом этапе при помощи программы MonkeyLearn [38] была осуществлена тематическая сегментация текстов. MonkeyLearn – платформа искусственного интеллекта, которая позволяет анализировать текст с помощью машинного обучения, чтобы автоматизировать рабочий процесс. Программа позволяет классифицировать и извлекать данные из необработанных текстов. В данной программе возможно определение тональности текста, тематической сегментации, извлечения ключевых слов и фраз.

Как видно из табл. 2, программа MonkeyLearn выделила 16 тем: животные (Animals), красота и стиль (Beauty & Style), бизнес и финансы (Business & Finance), компьютер и Интернет (Computers & Internet Consumer), электроника (Electronics), образование (Education), развлечения (Entertainment & Recreation), окружающая среда (Environment), еда и напитки (Food & Drink), садоводство (Gardening), здоровье и медицина (Health & Medicine), дом (Home), гуманитарные науки (Humanities), наука и математика (Science & Mathematics), общество (Society), путешествия (Travel). При автоматической обработке текстов из УМК Spotlight были выявлены 4 основные тематические группы, в которых темы повторялись три и более раз (включительно): 1) общество (Society) – 6 текстов (тексты 1, 2, 4, 11, 12; 36) 2) дом (Home) – 7 текстов; 3) гуманитарные науки (Humanities) – 9 текстов; 4) развлечения (Entertainment & Recreation) – 3 текста (см. табл. 2).

Таблица 2

Тематическая сегментация текстов из УМК Spotlight

Текст	Тема	%
1	Society (общество)	99,3
2	Society (общество)	18,8
3	Beauty & Style (красота и стиль)	68,2
4	Society (общество)	52,1
5	Home (дом)	39,1
6	Health & Medicine (здоровье и медицина)	64,8
7	Environment (окружающая среда)	31,9
8	Education (образование)	66,4
9	Humanities (гуманитарные науки)	49,1
10	Education (образование)	37,7
11	Society (общество)	29
12	Society (общество)	84,5
13	Entertainment & Recreation (развлечения)	35
14	Entertainment & Recreation (развлечения)	41,8

Окончание табл .2

Текст	Тема	%
15	Home (дом)	27,2
16	Travel (путешествия)	72,8
17	Health & Medicine (здоровье и медицина)	89,5
18	Home (дом)	70,2
19	Home (дом)	73,2
20	Home (дом)	57
21	Humanities (гуманитарные науки)	79,4
22	Gardening (садоводство)	30,6
23	Home (дом)	37,2
24	Home (дом)	98,4
25	Science & Mathematics (наука и математика)	27,8
26	Animals (животные)	83,9
27	Humanities (гуманитарные науки)	86,3
28	Animals (животные)	100
29	Computers & Internet (компьютер и Интернет)	69,9
30	Humanities (гуманитарные науки)	72,7
31	Humanities (гуманитарные науки)	70,5
32	Entertainment & Recreation (развлечения)	31,4
33	Humanities (гуманитарные науки)	64,8
34	Humanities (гуманитарные науки)	37,2
35	Humanities (гуманитарные науки)	31,9
36	Society (общество)	96,85
37	Science & Mathematics (наука и математика)	45,4
38	Humanities (гуманитарные науки)	66,6

На шестом этапе исследования был осуществлен анализ сложности текстов одной тематики. В качестве иллюстрации приведем пример анализа темы «Общество» (рис. 5).

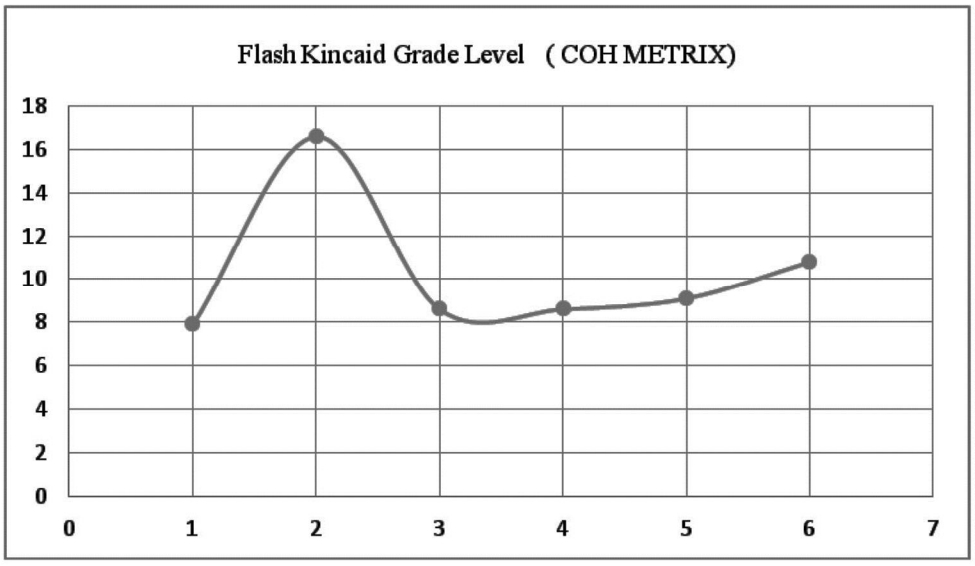

Рис 5. Уровень сложности текстов на тему «Общество»

График на рис. 5 свидетельствует о том, что к теме «Общество» учащиеся обращаются в течение учебного года шесть раз. Немаловажным фактором является то, что индекс сложности текстов в большинстве своем увеличивается, наиболее резкий подъем уровня сложности наблюдается в тексте 2.

4.Обсуждение и заключение

Таким образом, проблемы сложности текста и тематического моделирования широко изучаются в последние несколько десятилетий. Исследования в данной области базируются преимущественно на данных, собранных на основе текстов онлайн-сообщений, микроблогов, новостных обзоров и т. д. Значительно меньше исследований проведено на материале текстов академического дискурса. При этом очевидно, что именно тематическое моделирование и определение сложности академических текстов имеют широкие возможности применения в сфере образования при определении соответствия учебных материалов лингвистическим и когнитивным способностям целевой аудитории. Тематическое сегментирование и оценка сложности текста суть технологии, применение которых в образовании имеет потенциал способствовать улучшению качества преподавания, упростить подбор учебных материалов, осуществлять индивидуализацию обучения. Лингвостатистическая информация, полученная при помощи программ Coh-Metrix и MonkeyLearn, весьма полезна при изучении учебного материала, тематическом планировании, а также при подготовке учащихся к тестированию.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Русский язык как иностранный и методика его преподавания: сб. научн. тр. Вып. 28 / Редкол.: Е.И. Зиновьева, Н.А. Любимова (отв. ред.), Л.В. Московкин и др. – СПб.: РОПРЯЛ, 2017. – 160 с. [Электронный ресурс]. – Режим доступа: http://rki.spbu.ru/documents/sbornik2017.pdf (дата обращения 11.04.2017).
2. *Милованов К.* Методы интердискурсивной адаптации текста в СМИ с использованием его формальных характеристик // Культурологический журнал. – 2015. – № 2(20) [Электронный ресурс]. – Режим доступа: https://goo.gl/CBofsL (дата обращения 11.04.2017).
3. *Микк Я.А.* Методика измерения трудности текста // Вопросы психологии. – 1975. – № 3. – С. 147–155.
4. *Микк Я.А.* Факторы, определяющие время прочтения слова в связанном тексте // Вопросы психологии. – 1979. – № 3. – С. 125–128.
5. *Микк Я.А.* Оптимизация сложности учебного текста. – М.: Просвещение, 1981. – 119 с.
6. *Сидорова М.Ю.* Лингвистическая экспертиза школьных учебников // Метапредметный подход в образовании: русский язык в школьном и вузовском обучении разным предметам: сб. статей Межрегион. науч-практ. конф. (М., 19 апреля 2018). – М.: Российский учебник, 2018. – С. 49–64 [Электронный ресурс]. – Режим доступа: https://elibrary.ru/item.asp?id=36672498 (дата обращения 11.06.2018).
7. *Оборнева И.В.* Автоматизированная оценка сложности учебных текстов на основе статистических параметров: автореф. дис. … канд. пед. наук. – М., 2006. – 19 с. [Электронный ресурс]. – Режим доступа: https://www.dissercat.com/content/avtomatizirovannaya-otsenka-slozhnosti-uchebnykh-tekstov-na-osnove-statisticheskikh-parametr (дата обращения 11.04.2017).

8. *Глушань В.М.* Компьютерный анализ сложности текстов учебно-методических разработок как средство повышения качества обучения [Электронный ресурс]. – Режим доступа: https://elibrary.ru/item.asp?id=26028726 (дата обращения 25.04.2017).
9. *Солнышкина С.И., Кисельников А.С.* Сложность текста: этапы изучения в отечественном прикладном языкознании // Вестник ТГУ. Филология. – № 6(38). – 2015. – С. 86–100.
10. *Solov'ev V., Ivanov V., Solnyshkina M.* Assessment of reading difficulty levels in Russian academic texts: Approaches and metrics // Journal of Intelligent & Fuzzy Systems. 2018. Vol. 34. Is. 5. Pp. 3049–3058.
11. *Майер Р.В.* Определение уровня абстрактности, сложности и информативности различных тем школьного учебника физики // Психология, социология и педагогика. – 2013. – № 2 [Электронный ресурс]. – Режим доступа: http://psychology.snauka.ru/2013/02/1813 (дата обращения: 08.02.2018).
12. *Уша Т.Ю.* Язык школьного учебника: проблема понимания учащимся-инофоном учебного текста, терминологической лексики, формулировок заданий // Теория и практика общественного развития. – 2015. – № 15 [Электронный ресурс]. – Режим доступа: http://teoria-practica.ru/rus/files/arhiv_zhurnala/2015/15/pedagogics/usha.pdf (дата обращения: 08.02.2019).
13. *Устинова Л.В., Адекенова А.Н., Литвинова О.В.* Проверка сложности выпускных работ учащихся и студентов на основе статистических параметров // Молодой ученый. – 2015. – № 8. – С. 148–152 [Электронный ресурс]. – Режим доступа: https://moluch.ru/archive/88/16986/ (дата обращения: 28.02.2018).
14. Webcache [Электронный ресурс]. – Режим доступа: http://webcache.googleusercontent.com/search?q=cache:46AZDFGrSJoJ:www.ras.ru/FStorage/Download.aspx%3Fid%3D17d4378e-749c-45f1-84c8-812282c9b24d+&cd=15&hl=ru&ct=clnk&gl=ru
15. ФИОКО [Электронный ресурс]. – Режим доступа: https://fioco.ru/results_PISA_2015 (дата обращения: 20.02.2018).
16. ТАСС [Электронный ресурс]. – Режим доступа: https://tass.ru/obschestvo/5301919 (дата обращения: 20.02.2018).
17. Автоматическая обработка текстов на естественном языке и анализ данных: Учеб. пособие / Е.И. Большакова, К.В. Воронцов, Н.Э. Ефремова, Э.С. Клышинский, Н.В. Лукашевич, А.С. Сапин. – М.: Изд-во НИУ ВШЭ, 2017. – 269 с.
18. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: Учеб. пособие / Е.И. Большакова, Э.С. Клышинский, Д.В. Ландэ, А.А. Носков, О.В. Пескова, Е.В. Ягунова. – М.: МИЭМ, 2011. – 272 с.
19. *Аношин П.И.* Автоматический анализ текстов. Синтаксический и семантический анализ // Евразийский научный журнал. – 2017. – № 6. – С. 15.
20. Comparative Analysis about the Degree of text Complexity of Korean and Chinese Intermediate Korean textbooks – based on Internal Factors of texts [Электронный ресурс]. – Режим доступа: https://www.researchgate.net/publication/322205569_Comparative_Analysis_about_the_Degree_of_Text_Complexity_of_Korean_and_Chinese_Intermediate_Korean_Textbooks_-_based_on_Internal_Factors_of_Texts-, https://www.researchgate.net/publication/220746039_Automatic_Assessment_of_Japanese_Text_Readability_Based_on_a_Textbook_Corpus, http://wordsandmonsters.com/research/pdf/Japanese_high_school_textbook.pdf (дата обращения: 20.02.2018)
21. *Al-Khalil M., Saddiki H., Habash N., Alfalasi L.* A Leveled Reading Corpus of Modern Standard Arabic Muhamed [Электронный ресурс]. – Режим доступа: https://www.aclweb.org/anthology/L18-1366 (дата обращения: 20.06.2018).

22. *Solnyshkina M.I., Zamaletdinov R.R., Gorodetskaya L.A.* Evaluating text complexity and Flesch-Kincaid grade level // Journal of Social Studies Education Research. 2017. Vol. 8. Is. 3. Pp. 238–248.
23. *Fisher D., Lapp D., Frey N.* Homework in Secondary Classrooms: Making It Relevant and Respectful [Электронный ресурс]. – Режим доступа: https://s3-us-west-1.amazonaws.com/fisher-and-frey/documents/homework_jaal.pdf (дата обращения: 15.05.2018).
24. Using Coh-Metrix to Assess Cohesion and Difficulty in High School Textbooks [Электронный ресурс]. – Режим доступа: https://www.researchgate.net/publication/248260617_Using_Coh-Metrix_to_Assess_Cohesion_and_Difficulty_in_High-School_Textbooks (дата обращения: 20.02.2018).
25. "STABLE GENIUS" – Let's Go to the Data [Электронный ресурс]. – Режим доступа: https://factba.se/blog/2018/01/08/stable-genius-lets-go-to-the-data (дата обращения: 20.02.2018).
26. *Philip M. McCarthy, Gwyneth A. Lewis, David F. Dufty, Danielle S. McNamara.* Analyzing Writing Styles with Coh-Metrix [Электронный ресурс]. – Режим доступа: https://aaai.org/Papers/FLAIRS/2006/Flairs06-151.pdf (дата обращения: 20.02.2018).
27. Language in Law: Using Coh-Metrix to assess differences between American and English/Welsh language varieties [Электронный ресурс]. – Режим доступа: https://www.researchgate.net/publication/303288858_Language_in_law_Using_Coh-Metrix_to_assess_differences_between_American_and_EnglishWelsh_language_varieties (дата обращения: 17.04.2017).
28. *Gabitov A.I., Solnyshkina M.I., Shayakhmetova L.Kh., Ilyasova L.G.* Text Complexity In Russian Textbooks On Social Studies // Revista Publicando. 2017. Vol. 4. Is. 13. Pp. 597–606.
29. CohMetrix [Электронный ресурс]. – Режим доступа: http://cohmetrix.com (дата обращения: 20.04.2017).
30. *Вычегжанин С.В.* Анализ тональности текстов на основе ДСМ-метода. – Киров, 2013. – С. 16.
31. *Солнышкина М.И., Кисельников А.С.* Параметры сложности экзаменационных текстов // Вестник Волгоградского государственного университета. Сер. 2: Языкознание. – 2015. – № 1(25). – С. 99–107.
32. Интегративный подход в обучении младших школьников [Электронный ресурс]. – Режим доступа: integrativnyy-podhod-v-obuchenii-mladshih-shkolnikov (дата обращения: 20.02.2018).
33. Английский язык, 11 класс: Учебник для общеобраз. учреждений / О.В. Афанасьева, Дж Дули, И.В. Михеева и др. – М.: Просвещение, 2009. – 244 с.
34. *Бахтин М.М.* Литературно-критические статьи. – М.: Художественная литература, 1986. – 428 с.
35. *Леонтьева Н.Н.* Автоматическое понимание текстов: системы, модели, ресурсы: Учеб. пособие. – М.: Академия, 2006. – 304 с.
36. *Dowell N.* Analyzing Language and Discourse With Coh-Metrix. Workshop Presented at 2 nd Learning Analytics Summer Institutes (LASI 2014) / N. Dowell, Z. Cai & A.C. Graesser. Cambridge (MA), 2014. 84 p. Electronic text data. Mode of access: https://drive.google.com/file/d/0B-xloTsxGxlGcEw1RmNGTUtnSnc/edit (дата обращения: 25.04.2017).
37. *Graesser A.C., McNamara D.S., Louwerse M.M.* What do readers need to learn in order to process coherence relations in narrative and expository text. In A.P. Sweet and C.E. Snow (Eds.), Rethinking reading comprehension: New York: Guilford Publications, 2003. Pp. 82–98.
38. Coltheart. The MRC Psycholinguistic Database. Quarterly Journal of Experimental Psychology. 1981. 33A. Pp. 497–505.

39. MonkeyLearn [Электронный ресурс]. – Режим доступа: https://monkeylearn.com/topic-analysis (дата обращения: 25.04.2017).

Поступила в редакцию 30.07.2019
В окончательном варианте 15.09.2019

UDC 81’32

AUTOMATED TEXT PROCESSING: TOPIC SEGMENTATION OF EDUCATIONAL TEXTS

M.I. Solnyshkina[1], I.E. YArmakeev[2], E.V. Gafiyatova[3], F.KH. Ismaeva[4]
[1,2,3,4]Kazan (Volga) Federal University
18, Kremlevskaya st., Kazan, 420008, Russia
[1]E-mail: mesoln@yandex.ru
[2]E-mail: ermakeev@mail.ru
[3]E-mail: rg-777@yandex.ru
[4]E-mail: fismaeva@yandex.ru

Abstract
The article explores the problem of automatic quantitative assessment of text complexity and thematic segmentation of texts. The authors offer a brief description of the state of affairs in this area including the fact that the existing formulas for text readability index are genre-dependent and lose their reliability when used for texts of other genres. Based on the corpus of educational texts and analysis of quantitative text parameters, we suggest a new way of text ranking so that they correspond linguistic abilities of pupils. The study was carried out on the material of UMK Spotlight 11, the corpus used in the study comprises 38 texts of 12891 tokens in total. The methods used were topic segmentation, component analysis, statistical analysis, Flash-Kincaid readability Texts complexity assessment showed that the dynamics of texts with tasks (1) testing skimming abilities is from more complex to less complex (–0.2); (2) testing scanning abilities is from less complex to more (+0.4); (3) in the texts for intensive reading, text readability rose by 5.2. The thematic segmentation performed based on Monkey Learn revealed the vocabulary of 15 topics that, during the school year, are offered to students on average 3–5 times. The most frequent theme is "Humanities", the reference to which is revealed in 9 modules. It is significant that textbook authors offer the following topics Gardening, Computers & Internet, Science & Mathematics, Entertainment& Recreation only once during the school year.

Key words: text complexity, readability, mathematical models, English Language, thematic division.

[1] *Marina I. Solnyshkina,* Dr. Phil. Sci., Professor of Theory and Practice of Teaching Foreign Languages Department.

[2] *Iskander E. YArmakeev,* Dr. Ped. Sci., Professor of Bilingual and Digital Education Department.

[3] *Elzara V. Gafiyatova,* Cand. Phil. Sci., Associate Professor of Theory and Practice of Teaching Foreign Languages Department.

[4] *Farida KH. Ismaeva,* Cand. Phil. Sci., Associate Professor of Foreign Languages Department.

REFERENCES

1. Russkiy yazyk kak inostrannyy i metodika ego prepodavaniya [Russian language as foreign language and methods of teaching sb. nauchn. tr. Vyp. 28 / Redkol.: E.I. Zinov'yeva, N.A. Lyubimova (otv. red.), L.V. Moskovkin i dr. Saint Petersburg: ROPRYaL, 2017. 160 p. http://rki.spbu.ru/documents/sbornik2017.pdf (accessed April 11, 2017).
2. *Milovanov K.* Metody interdiskursivnoy adaptatsii teksta v SMI s ispol'zovaniyem ego formal'nykh kharakteristik [Methods of interdiscursive adaptation of the text in media with using its formal characteristics]. *Kul'turologicheskiy zhurnal.* 2015. Vol. 2(20). https://goo.gl/CBofsL (accessed April 11, 2017).
3. *Mikk YA.A.* Metodika izmereniya trudnosti teksta [Method of measuring text complexity]. *Voprosy psikhologii.* 1975. No. 3. Pp. 147–155.
4. *Mikk YA.A.* Faktory, opredelyayushchiye vremya prochteniya slova v svyazannom tekste [Factors of measuring the reading time of a word in the text]. *Voprosy psikhologii.* 1979. No. 3. Pp. 125–128.
5. *Mikk YA.A.* Optimizatsiya slozhnosti uchebnogo teksta [Optimization of teaching text complexity]. Moscow: Prosveshcheniye, 1981. 119 p.
6. *Sidorova M.YU.* Lingvisticheskaya ekspertiza shkol'nykh uchebnikov [Linguistics expertise of complex academic books]. *Metapredmetnyy podkhod v obrazovanii: russkiy yazyk v shkol'nom i vuzovskom obuchenii raznym predmetam: sb. statey Mezhregion. nauch-prakt. konf. (M., 19 aprelya 2018).* Moscow: Rossiyskiy uchebnik, 2018. Pp. 49–64. https://elibrary.ru/item.asp?id=36672498 (accessed April 11, 2017).
7. *Oborneva I.V.* Avtomatizirovannaya otsenka slozhnosti uchebnykh tekstov na osnove statisticheskikh parametrov: Avtoref. dis. kand. ped. nauk [Automatically evaluating of academic text complexity based on statistics parameters. Abstract of thesis cand. of ped. sci.]. Moscow, 2006. 19 p.
   https://www.dissercat.com/content/avtomatizirovannaya-otsenka-slozhnosti-uchebnykh-tekstov-na-osnove-statisticheskikh-parametr (accessed April 11, 2017).
8. *Glushan' V.M.* Komp'yuternyy analiz slozhnosti tekstov uchebno-metodicheskikh razrabotok kak sredstvo povysheniya kachestva obucheniya [Computer analyze of text complexity of teaching materials as improvement in the quality of teaching]. https://elibrary.ru/item.asp?id=26028726 (accessed April 25, 2017).
9. *Solnyshkina S.I., Kisel'nikov A.S.* Slozhnost' teksta: etapy izucheniya v otechestvennom prikladnom yazykoznanii [Text complexity: investigation stages in domestic applied linguistics]. *Vestnik TGU. Filologiya.* 2015. Vol. 6(38). Pp. 86–100.
10. *Solov'yev V., Ivanov V. Solnyshkina M.* Assessment of reading difficulty levels in Russian academic texts: Approaches and metrics. *Journal of Intelligent & Fuzzy systems*. 2018. Vol. 34. Is. 5. Pp. 3049–3058.
11. *Mayyer R.V.* Opredeleniye urovnya abstraktnosti, slozhnosti i informativnosti razlichnykh tem shkol'nogo uchebnika fiziki [Determination of the level of abstraction, complexity and information contents of different themes of Physics school books]. *Psikhologiya, sotsiologiya i pedagogika.* 2013. Vol. 2. http://psychology.snauka.ru/2013/02/1813 (accessed February 8, 2018).
12. *Usha T.YU.* YAzyk shkol'nogo uchebnika: problema ponimaniya uchashchimsya-inofonom uchebnogo teksta, terminologicheskoy leksiki, formulirovok zadaniy [Language of school textbooks: problem of text understanding, terminology language, formulation of tasks]. *Teoriya i praktika obshchestvennogo razvitiya.* 2015. No. 15. http://teoria-practica.ru/rus/files/arhiv_zhurnala/2015/15/pedagogics/usha.pdf (accessed February 8, 2018).
13. *Ustinova L.V., Adekenova A.N., Litvinova O.V.* Proverka slozhnosti vypusknykh rabot uchashchikhsya i studentov na osnove statisticheskikh parametrov [Verification of graduate work

complexity on based of statistical parameters]. *Molodoy uchenyy*. 2015. Vol. 8. Pp. 148–152. https://moluch.ru/archive/88/16986/ (accessed February 28, 2018).
14. Webcache.http://webcache.googleusercontent.com/search?q=cache: 6AZDFGrSJoJ:www.ras.ru/ FStorage/Download.aspx%3Fid%3D17d4378e-749c-45f1-84c8-812282c9b24d+&cd=15&hl=ru&ct=clnk&gl=ru
15. FIOKO. https://fioco.ru/results_PISA_2015 (accessed February 20, 2018).
16. TASS. https://tass.ru/obschestvo/5301919 (accessed February 20, 2018).
17. Avtomaticheskaya obrabotka tekstov na estestvennom yazyke i analiz dannykh [Automatic natural language text processing and data analysis]. E.I. Bol'shakova, K.V. Vorontsov, N.E. Efremova, E.S. Klyshinskiy, N.V. Lukashevich, A.S. Sapin. Moscow. Izd-vo NIU VSHE, 2017. 269 p.
18. Avtomaticheskaya obrabotka tekstov na estestvennom yazyke i komp'yuternaya lingvistika [Automatic processing of natural language texts and computer linguistics]. E.I. Bol'shakova, E.S. Klyshinskiy, D.V. Lande, A.A. Noskov, O.V. Peskova, E.V. YAgunova. Moscow: MIEM, 2011. 272 p.
19. *Anoshin P.I.* Avtomaticheskiy analiz tekstov. Sintaksicheskiy i semanticheskiy analiz [Automatic text analysis. Syntactic and semantic analysis]. *Evraziyskiy nauchnyy zhurnal.* 2017. Vol. 6. P. 15.
20. Comparative Analysis about the Degree of text Complexity of Korean and Chinese Intermediate Korean textbooks based on Internal Factors of texts. https://www.researchgate.net/ publication/322205569_Comparative_Analysis_about_the_Degree_of_Text_Complexity_of_Korean_and_Chinese_Intermediate_Korean_Textbooks_-based_on_Internal_Factors_of_Texts-, https://www.researchgate.net/publication/220746039_Automatic_Assessment_of_Japanese_Text_Readability_Based_on_a_Textbook_Corpus, http://wordsandmonsters.com/research/pdf/Japanese_high_school_textbook.pdf (accessed February 20, 2018).
21. *Al-Khalil M., Saddiki H., Habash N., Alfalasi L.* A Leveled Reading Corpus of Modern Standard Arabic Muhamed. https://www.aclweb.org/anthology/L18-1366 (accessed 20 February 2018).
22. *Solnyshkina M.I., Zamaletdinov R.R., Gorodetskaya L.A.* Evaluating text complexity and Flesch-Kincaid grade level. *Journal of Social Studies Education Research*. 2017. Vol. 8. Is. 3. Pp. 238–248.
23. *Fisher D., Lapp D., Frey N.* Homework in Secondary Classrooms: Making It Relevant and Respectful. https://s3-us-west-1.amazonaws.com/fisher-and-frey/documents/homework_jaal.pdf (accessed 15 May 2017).
24. Using Coh-Metrix to Assess Cohesion and Difficulty in High School Textbooks. https://www.researchgate.net/publication/248260617_Using_Coh-Metrix_to_Assess_Cohesion_and_Difficulty_in_High-School_Textbooks (accessed February 20, 2018).
25. "STABLE GENIUS" – Let's Go to the Data. https://factba.se/blog/2018/01/08/stable-genius-lets-go-to-the-data (accessed February 20, 2018).
26. *Philip M. McCarthy, Gwyneth A. Lewis, David F. Dufty, Danielle S. McNamara.* Analyzing Writing Styles with Coh-Metrix. https://aaai.org/Papers/FLAIRS/2006/Flairs06-151.pdf (accessed February 20, 2018).
27. Language in Law: Using Coh-Metrix to assess differences between American and English/Welsh language varieties. https://www.researchgate.net/publication/303288858_Language_in_law_Using_Coh-Metrix_to_assess_differences_between_American_and_EnglishWelsh_language_varieties (accessed April 11, 2017).

28. *Gabitov A.I., Solnyshkina M.I., Shayakhmetova L.Kh., Ilyasova L.G.* Text Complexity In Russian Textbooks On Social Studies. *Revista Publicando*. 2017. Vol. 4. Is. 13. Pp. 597–606.
29. CohMetrix. http://cohmetrix.com (accessed 20 April 2017).
30. *Vychegzhanin S.V.* Analiz tonal'nosti tekstov na osnove DSM-metoda [Text tone analysis on the bases of DSM-method]. Kirov, 2013. P. 16.
31. *Solnyshkina M.I., Kisel'nikov A.S.* Parametry slozhnosti ekzamenatsionnykh tekstov [Examination Text Difficulty Options]. *Vestnik Volgogradskogo gosudarstvennogo universiteta. Ser. 2: YAzykoznaniye.* Vol. 1(25). 2015. Pp. 99–107.
32. Integrativnyy podkhod v obuchenii mladshikh shkol'nikov [Integrative approach in the education of primary school students]. http://integrativnyy-podhod-v-obuchenii-mladshih-shkolnikov (accessed February 20, 2018).
33. *Afanas'yeva O.V., Duli Dzh., Mikheeva I.V. i dr.* Angliyskiy yazyk, 11 klass [Spotlight 11]. Moscow: Prosveshcheniye, 2009. 244 p.
34. *Bakhtin M.M.* Literaturno-kriticheskiye stat'i [Literary critical articles]. Moscow: KHudozhestvennaya literatura, 1986. 428 p.
35. *Leont'yeva N.N.* Avtomaticheskoye ponimaniye tekstov: sistemy, modeli, resursy [Automatic understanding of texts: Systems, models, resources]. Moscow: Akademiya, 2006. 304 p.
36. *Dowell N.* Analyzing Language and Discourse With Coh-Metrix. Workshop Presented at 2 nd Learning Analytics Summer Institutes (LASI 2014) Cambridge (MA), 2014. 84 p. https://drive.google.com/file/d/0B-xloTsxGxlGcEw1RmNGTUtnSnc/edit (accessed April 25, 2017).
37. *Graesser A.C., McNamara D.S., Louwerse M.M.* What do readers need to learn in order to process coherence relations in narrative and expository text. In A.P. Sweet and C.E. Snow (Eds.), Rethinking reading comprehension: New York: Guilford Publications, 2003. Pp. 82–98.
38. *Coltheart.* The MRC Psycholinguistic Database. *Quarterly Journal of Experimental Psychology*. 1981. No. 33A. Pp. 97–505.
39. MonkeyLearn. https://monkeylearn.com/topic-analysis (accessed April 25, 2017).

Original article submitted 30.07.2019
Revision submitted 15.09.2019