

МНОГОФАКТОРНЫЙ АНАЛИЗ СЛОЖНОСТИ ТЕКСТА



Аннотация

В статье рассматривается проблема типологии текстов ОГЭ, разрабатываемая на основе семи лингвостатистических параметров: длины предложения, длины слова, индекса читабельности по Флешу – Кинкейду, индекса читабельности текста как иностранного, лексического многообразия, референциальной связности, индекса конкретности/абстрактности. Многофакторный анализ 118 текстов, осуществленный при помощи онлайн-сервисов Text Inspector и Coh-Metrix, выявил средние индексы текстов ОГЭ, позволяющие определить соответствие параметров любого текста типовым параметрам ОГЭ. Результаты исследования могут быть полезны учителям английского языка, авторам пособий и учебников, а также разработчикам тестовых заданий.



Abstract

MULTIPLE FACTOR ANALYSIS OF TEXT COMPLEXITY

The article examines the problem of typology of General State Exam texts, developed on the basis of seven linguistic-statistical parameters, i.e. sentence length, word length, Flash-Kincaid readability index, Coh-Metrix L2 readability index, lexical diversity, referential cohesion, concreteness / abstractness index. Multi-factor analysis of 118 texts, carried out using online services Text Inspector and Coh-Metrix, revealed the average indexes of General State Exam texts, which enable to determine the correspondence of parameters of any text to the standard parameters of General State Exam. The research results can be useful for English teachers, authors of manuals and textbooks as well as test items developers.



Ключевые слова

ОГЭ, учебный текст, читабельность, параметры текста



Keywords

General State Exam, educational text, readability, text parameters

Введение. Сложность текста как лингводидактическая проблема обсуждается в настоящее время в рамках междисциплинарных исследований, поскольку для ее решения необходимо привлечение лингвистов, педагогов, психологов, специалистов по информационным технологиям и статистике. Практическая значимость исследований в данной области определяется необходимостью обеспечить соответствующими возрасту и лингвистическим способностям текстами не только обучающихся и экзаменуемых, но также пользователей вебсайтов и других категорий читателей. Особую значимость данная проблема приобретает для разработчиков тестов и учителей, поскольку несоответствие сложности текста способностям читателей ведет к демотивации обучающихся и утрате валидности результатов тестов [12].

Учебный текст как «модель образцового использования языка» [25] во всех случаях должен соответствовать когнитивному и лингвистическому возрастам читателя,

т.е. находиться в таком диапазоне сложности, который способен мотивировать и расширять «зону ближайшего развития» [3]. При этом понимание учебного текста предполагает, что читатель, помимо знаний лексики и синтаксиса, обладает способностью «реконструировать» все имеющиеся в тексте пропозиции, т.е. обладает навыком чтения «между строк», понимания метафорических значений, подтекстов, ситуационных моделей [34]. Таким образом, отбор текста (-ов) для определенного «читательского возраста» предполагает решение двух равнозначных задач: определение сложности или читабельности текста, с одной стороны, и выявление «читательского возраста» потенциального читателя, с другой стороны. При этом соблюдение лингводидактических принципов предполагает размещать тексты в учебных пособиях в порядке следования «от простого к сложному», постепенно расширяя «зону ближайшего развития» читателя. В этом случае не только у читателя сохраняется мотивация, но и у педагога и экзаменатора появляется больше возможностей прогнозировать результаты тестов.

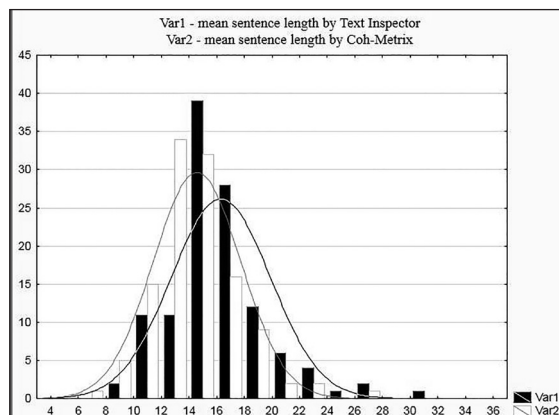


Рис. 1
Флуктуации
длины
предложения
и длины слова

Анализ современных учебных пособий и тестовых материалов, используемых в Российской Федерации, дает возможность утверждать, что в ряде случаев текст учебника и теста не всегда соответствует читательскому адресу [13]. К сожалению, экспертиза учебных материалов, осуществляемая в настоящее время в нашей стране, предъявляет крайне ограниченные требования к текстам: «Текстовый материал соответствует нормам современного русского языка, государственных языков республик Российской Федерации, языков народов Российской Федерации. Текст предоставляет возможность «расширения» информационного поля» [10]. Что касается оценки сложности текста на иностранных языках, то основополагающие документы устанавливают жесткий регламент относительно языковой сложности текстов, которая «должна соответствовать заявленному уровню сложности задания (базовый, повышенный или высокий)» [6].

Представленное исследование имеет целью выявление типичных дескриптивных и лексических метрик, а также определение диапазона читабельности текстов, предлагаемых школьникам для подготовки и сдачи ОГЭ по английскому языку. Гипотезой, которая доказывается в данном исследовании, является утверждение, что тексты ОГЭ должны соответствовать одному типу сложности и при помощи ограниченного списка параметров могут быть квалифицированы как (не) подходящие для ОГЭ. Для оценивания текстов ОГЭ были выбраны следующие параметры: средняя длина предложения, средняя длина слова, читабельность (индекс Флеша – Кинкейда и индекс читабельности текстов на иностранном языке), индекс референциальной связности, индекс лексического многообразия, конкретность.

Материал и методы. Корпус исследования включает 118 текстов для подготовки и сдачи

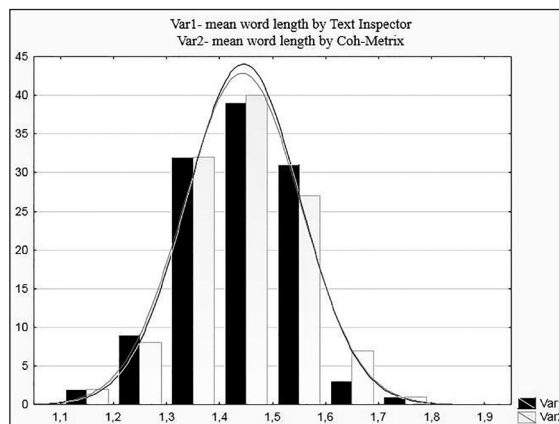


Рис. 2
Диапазон
средней длины
слова

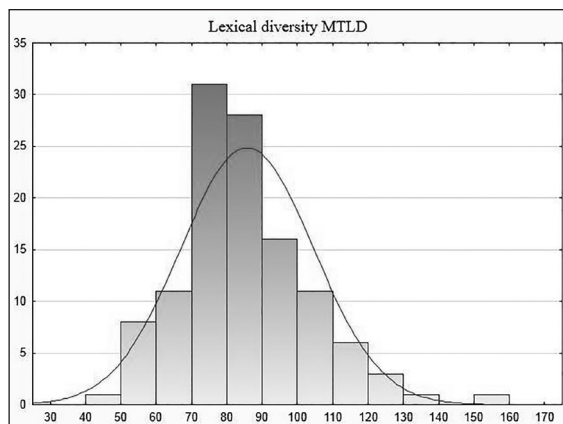


Рис. 3

Читабельность
по Флешу –
Кинкейду

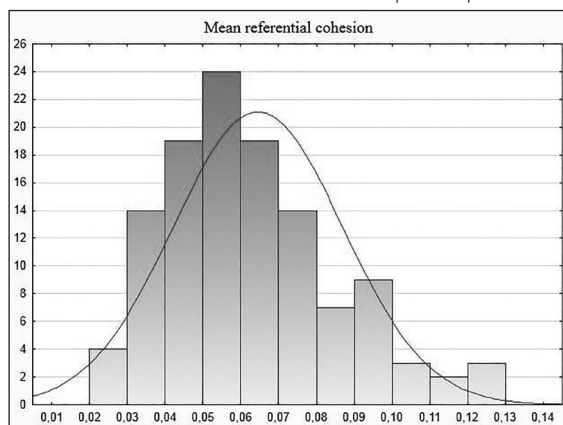
[7], 10 текстов для подготовки к ОГЭ сайта «Английский в полном порядке» [1], 19 текстов для подготовки к ОГЭ сайта «Сдам ГИА: Решу ОГЭ» [11], 1 текст для подготовки к ОГЭ сайта «Подготовка к ОГЭ (ОГЭ Английский)». Данные тексты использованы во втором разделе «Задания по чтению» в ОГЭ (задания 10–17) [16], в котором экзаменуемые должны продемонстрировать умение осуществлять изучающее чтение, предполагающее полное и точное понимание «основных и второстепенных фактов, содержащихся в тексте» [17].

Оценка и измерение параметров текстов осуществлялись при помощи онлайн-сервисов Text Inspector (далее TI) [32], Coh-Metrix (далее CM) [21], которые были дополнены сравнительным лингвистическим анализом, осуществленным независимыми экспертами. Анализ семи параметров текста, а именно количества слов, глаголов, существительных; средней длины предложения, слова; лексического многообразия, конкретности, читабельности по формуле Флеш – Кинкейда был осуществлен на платформе Text Inspector. На основе платформы Coh-Metrix были рассчитаны индекс читабельности для изучающих английский язык как иностранный (Coh-Metrix L2 Readability) и индекс референциальной связности. Во всех случаях при загрузке на сайт вышеуказанных платформ объем анализируемого текста был выше 75% оригинального текста, что позволило сохранить репрезентативность выборки [18].

Рис. 4

Читабельность
CML2 текстов ОГЭ

Анализ. Анализ текстов был осуществлен поэтапно в трех группах изучаемых параметров: дескриптивные характеристики, индексы читабельности, лексические параметры.



Дескриптивные параметры, средняя длина предложения и средняя длина слова оценивались с помощью сервисов TI и CM. Результаты представлены на рис. 1.

Как видим (см. рис. 1), инструменты выявили аналогичные метрики длины предложения и длины слов. Средняя длина предложения в изученных текстах составляет 12 (CM) – 14 (TI) слов, что несколько ниже среднестатистической длины английского предложения в 15–20 слов, рекомендуемого разработчиками так называемого «простого английского» [23]. Оба инструмента выявили одинаковую среднюю длину слова:

1,44 слога (см. рис. 2), которая также несколько ниже среднестатистического слова в «простом» английском языке, длина которого оценивается специалистами в 1,66 слога [24]. Данные результаты позволяют охарактеризовать тексты ОГЭ с точки зрения их дескриптивных параметров как тексты с индексом сложности несколько ниже среднего в современном английском языке.

В современной парадигме читабельность текстов измеряется по-разному для текстов, предназначенных для двух категорий читателей: (1) для носителей языка — при помощи индекса Флеша — Кинкейда (Flesch-Kincaid Grade Level, FKGL) [29]; (2) для неносителей английского языка — при помощи индекса читабельности Coh-Metrix L2 Reading Index (CML 2x) [27].

Онлайн-сервисы TI и CM оценивают читабельность текстов при помощи формулы Флеша — Кинкейда: $FKGL = (0,39 \times ASL) + (11,8 \times ASW) - 15,59$, где ASL — средняя длина предложения; ASW — среднее число слогов в слове [27]. Читабельность по Флешу ранжирует тексты на основе их соответствия читателям определенного года обучения: тексты с FKGL в диапазоне 1-10 соответствуют читателям средней школы, тексты с FKGL 11-15 предназначены для студентов высшего образования, а 16-20 соответствуют сложным научным текстам, читателями которых являются научные сотрудники. Индекс читабельности для текстов на английском языке для читателей-неносителей языка, Coh-Metrix L2 Reading Index (CML 2x), рассчитывается по формуле $CML2 = -45,032 + (52,230 \times CWO) + (61,306 \times SSS) + (22,205 \times CELEX)$, где CWO (Content word overlap) — повтор знаменательных слов, SSS (Sentence syntactic similarity) — синтаксическое сходство предложений, CELEX — средняя частота встречаемости всех слов в тексте. Данные параметры в значительной степени отличают формулу CML2 от формулы Флеша — Кинкейда, принимающей во внимание исключительно количественные параметры без учета содержания текста.

Индексы Флеша — Кинкейда и CML2 рассчитывались для всех изучаемых текстов ОГЭ. Средний уровень читабельности текстов ОГЭ по Флешу — Кинкейду составил 5-9, что соответствует диапазону осведомленности ученика 5-9-го года обучения в американской средней школе (см. рис. 3).

Выборка текстов ОГЭ также имеет нормальное распределение индекса читабельности текстов на иностранном языке CML2 в диапазоне 14-20 (см. рис. 4).

Тексты данного уровня соответствуют «зоне ближайшего развития» обучающегося, находясь в которой читатель способен «самостоятельно решать проблемы», а «под руководством взрослых или в сотрудничестве с более развитыми сверстниками» сможет обеспечить себе «потенциальное развитие, определяемое способностью решать проблемы» [33].

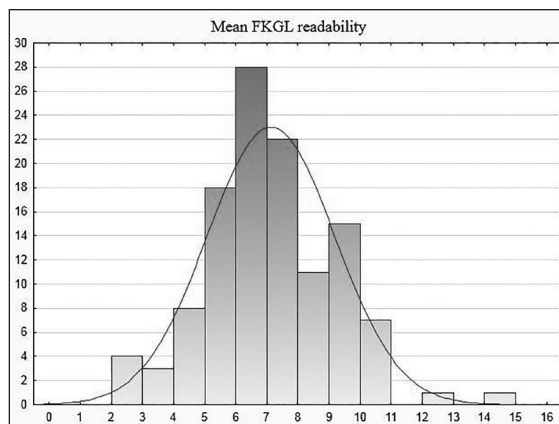


Рис. 5

Диапазон индекса
MTLD в текстах
ОГЭ

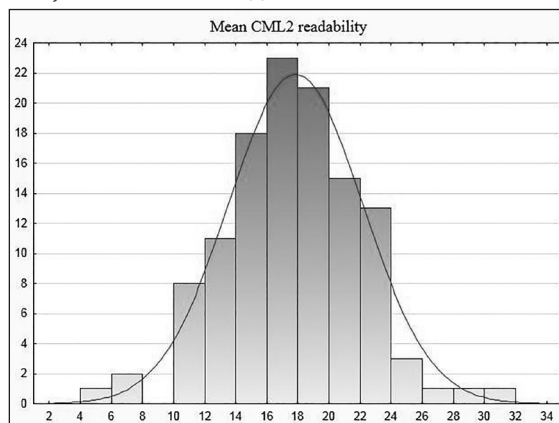


Рис. 6

Индекс
референциальной
связности

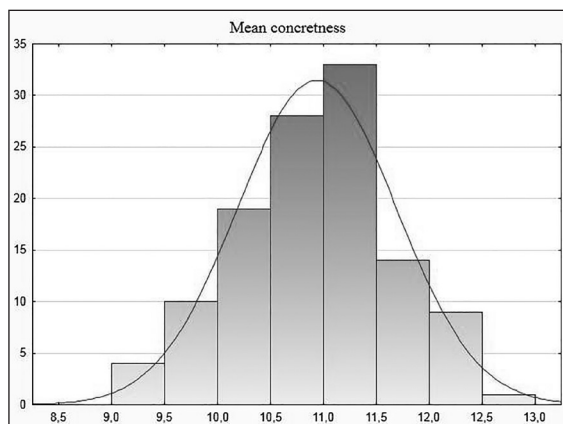


Рис. 7
Диапазон
конкретности

Лексические параметры. В представленном исследовании осуществлен анализ трех лексических параметров текстов ОГЭ: лексическое многообразие, референциальная связность и конкретность/абстрактность.

Индекс лексического многообразия характеризует лексическое богатство текста и рассчитывается как отношение неповторяющихся слов ко всем словам в тексте. Индекс лексического разнообразия текста (MTLD) вычисляется по формуле $MTLD = L/n$, где L – число словоформ, n – число строк [27].

Диапазон колебаний индекса MTLD находится в пределах от 40 до 170, а среднее значение индекса – от 70 до 90. Данные метрики характеризуют изучаемые тексты как лексически богатые.

На основе количества повторов, т.е. языковых средств замен наименований, представленных в изучаемых текстах, программа Coh-Metrix рассчитывает также *референциальную связность* текста [27]. Например, в текстах A1nf и M5f использован широкий спектр средств повторной номинации: личные и притяжательные местоимения (he, his, their), а также относительные местоимения (which) и др.

Метрики данного параметра в текстах ОГЭ колеблются в пределах от 0,05 до 0,07, свидетельствуя о невысокой референциальной связности текстов, используемых при оценивании знания иностранного языка школьников.

Индекс конкретности/абстрактности слов в тексте в значительной степени влияет на трудность его восприятия. Coh-Metrix измеряет абстрактность текста на основе соотношения количества конкретных и абстрактных имен существительных. Каждому существительному в тексте также присваивается ранг конкретности/абстрактности на основе базы данных MRC Psycholinguistic Database [23]. Анализ показал крайне низкий индекс абстрактности/конкретности изучаемых текстов (см. рис. 7).

Как видим, метрика колеблется от 8,5 до 13,5, в то время как средние значения варьируются от 10 до 12.

Заключение. Таким образом, многофакторный анализ 118 текстов позволил сформировать лингвостатистическую типологию текстов ОГЭ. Типичный текст ОГЭ имеет следующие параметры: средняя длина предложения колеблется в пределах 14,4-17,0 слов, длина слова – приблизительно 1,44 слога, читабельность по Флешу – Кинкейду находится в диапазоне от 5 до 9; индекс читабельности текстов на иностранных языках ограничен диапазоном 14-20. Флуктуации метрик лексических параметров типичного текста ОГЭ незначительны и находятся в следующих пределах: референциальная связность – от 0,05 до 0,07, лексическое разнообразие (MTLD) – от 70 до 90, конкретность – от 10 до 12. Дескриптивные и лексические параметры, а также индексы читабельности позволяют отнести каждый из изученных текстов к ядру или периферии типологии текстов ОГЭ. Измеряемые при помощи онлайн-сервисов метрики восьми параметров (длина предложения, длина слова, читабельность по Флешу – Кинкейду, читабельность для текстов

на иностранном языке (CML2), лексическое многообразие, референциальная связность и конкретность) позволяют не только классифицировать тексты по уровням сложности, но и осуществлять отбор текстов, валидных для тестирования на данном уровне.

Благодарность. Исследование выполнено при финансовой поддержке гранта 19-07-00807 Российского фонда фундаментальных исследований. Составление корпуса выполнено при финансовой поддержке Российского научного фонда, грант № 18-18-00436.



Литература

1. Английский в полном порядке [Электронный ресурс] // ОГЭ английский: сайт URL: <http://tonail.com> (дата обращения: 10.09.20)
2. **Веселова Ю. С.** ОГЭ. Готовимся к итоговой аттестации: учебное пособие. – М.: Интеллект-центр, 2020. – 128 с.
3. **Выготский Л. С.** Динамика умственного развития школьника в связи с обучением // Л. С. Выготский. Умственное развитие детей в процессе обучения. – М.-Л., 1935. – С. 33-52.
4. **Выготский Л. С.** Проблема обучения и умственного развития в школьном возрасте // Л. С. Выготский. Избранные психологические исследования. – М.: Изд-во АПН РСФСР, 1956. – С. 338-452.
5. **Гудкова Л. М., Терентьева О. В.** ОГЭ-2020: Английский язык: 30 тренировочных вариантов экзаменационных работ для подготовки к общему государственному экзамену: учебное пособие. – М., 2019. – 106 с.
6. Единый государственный экзамен по иностранным языкам [Электронный ресурс] // Иностранные языки, 11 класс: сайт URL: <https://egenglish.ru/wp-content/uploads/2019/08/INIAZ-11-EGE-2020-SPETS.pdf>, 2019. (дата обращения: 16.09.20)
7. **Махмурян К. С., Маркова Е. С., Соловова Е. Н.** ГИА. Английский язык. Практикум.: учебное пособие. – М., 2013. – 86 с.
8. Подготовка к ОГЭ. [Электронный ресурс] // ОГЭ Английский: сайт URL: <http://onlyege.ru/realnyj-variant-oge-po-anglijskomu-yazyku-2017-variant-006/> (дата обращения: 10.09.20)
9. **Попова М. А.** ГИА 13. Английский язык. 9 класс. Тематические тестовые задания для подготовки к ГИА. – Я.: 2012. – 96 с.
10. Регламент [Электронный ресурс] // Проведение педагогической экспертизы учебных пособий: сайт URL: http://rusacademedu.ru/wp-content/uploads/2015/11/reglament_ped_eksp_zch_pos_prikaz-80.pdf, 2015. (дата обращения: 10.09.20)
11. Сдам ГИА: Решу ЕГЭ [Электронный ресурс] // Каталог заданий. ОГЭ английский: сайт URL: <https://en-oge.sdangia.ru/test?theme=10> (дата обращения: 13.11.20)
12. **Солнышкина М. И., Казачкова М. Б., Харькова Е. В.** Инструменты измерения сложности текстов на английском языке // Иностр. языки в школе. – 2020. – № 3. – С. 15-21.
13. **Солнышкина М. И., Кисельников А. С.** Параметры сложности экзаменационных текстов // Вестник Волгоградского государственного университета. Серия 2: Языкознание. – 2015. – № 1. – С. 99-107.
14. **Соловова Е. Н., Маркова Е. С.** ОГЭ 2016. Английский язык. Типовые тестовые задания: учебное пособие. – М., 2016. – 87 с.
15. **Сунгатулина Д. Д.** Специфика экзаменационного текста: вторичность, функции, сложность (на материале русского и английского языков). – 2018, 263 с.
16. ФГБНУ «ФИПИ» [Электронный ресурс] // Открытый банк заданий: сайт. М., [2004-2020]. URL: <http://www.fipi.ru/> (дата обращения: 15.04.2020).
17. **Фоломкина С. К.** Обучение чтению на иностранном языке в неязыковом вузе. – М.: Высшая школа. – 2005. – 255 с.

18. **Biber D.** Representativeness in Corpus Design. Oxford, Literary and Linguistic Computing, 8, 1993.
19. **Biber D.** Register as a predictor of linguistic variation // Corpus linguistics and linguistic theory. – 2012. – Т. 8. – №. 1. – p. 9-37.
20. **Campaign P.E.** How to write in plain English // Plain English Campaign. – М., 2004.
21. **Campaign P.E.** Plain English Campaign. – М., 2011. Coh-Metrix: Analysis of text on cohesion and language. URL: <https://link.springer.com/article/10.3758/BF03195564> (accessed: 20.04.2020).
22. **Coltheart M.** The MRC Psycholinguistic Database. Quarterly Journal of Experimental Psychology, 1981, p. 497-505.
23. **Cutts M.**, 2009, Newell, C., 2014, American Journal Experts, 2015, How to write in plain English, 2015.
24. **Flesch R.** Chapter 2: Let's start with the formula // How to write plain English. URL: <http://www.appstate.edu/~steelekm/classes/psy2664/Flesch.htm> (accessed: 20.04.2020).
25. **Lukasik M.** Lexical Minimum Re (Defined). Slupsk, Lingwistyka Stosowana, 23, 2017, p. 47-63.
26. **Martin Cutts.** Oxford Guide to Plain English, Oxford: Oxford University Press, 3rd edition. – М, 2009.
27. **McNamara D. S., Graesser A. C., McCarthy P. M., Zh. Cai.** Automated evaluation of Text and Discourse with Coh-Metrix. New York, Cambridge University Press, 2014. 285 p.
28. **Newell C.** Editing tip: Sentence length. – М., 2014.
29. **Solnyshkina M. I., Gabitov A. I., Shayakhmetova L. Kh., Ilyasova L. G., Akbarova S. A.** Text Complexity In Russian Textbooks On Social Studies // Revista Publicando, 4, 2017, p. 597-606.
30. **Solnyshkina M. I., Harkova E. V., Kazachkova M. B.** The Structure of Cross-Linguistic Differences: Meaning and Context of 'Readability' and its Russian Equivalent 'Chitabelnost' // Journal of Language and Education, 6 (1) – 2020, p. 103-119.
31. **Solovyev V. D., Ivanov V. V., Akhtiamov R. B.** Dictionary of Abstract and Concrete Words of the Russian Language: A Methodology for Creation and Application // Journal of Research in Applied Linguistics, 10, 2019, p. 215-227.
32. **Text Inspector:** The professional web tool for analyzing texts. URL: <https://textinspector.com/workflow> (accessed: 20.04.2020).
33. **Vygotsky L. S.** Mind in society: The development of higher psychological processes (Cambridge, MA: Harvard University Press, 1978).
34. **Weir C. J. and Khalifa H. A.** cognitive processing approach towards defining reading comprehension, Research Notes, University of Cambridge Local Examinations Syndicate 2008 (31): p. 2-10.



**Сведения
об авторах**

Солнышкина Марина Ивановна, профессор ИФМК КФУ, руководитель НИЛ «Интеллектуальные технологии управления текстами».

E-mail: mesoln@yandex.ru

Замалетдинов Радиф Рифкатович, профессор ИФМК КФУ, заведующий кафедрой общего языкознания и тюркологии.

E-mail: director.ifmk@gmail.com

Гафиятова Эльзара Васильевна, профессор ИФМК КФУ заведующий кафедрой теории и практики преподавания иностранных языков.

E-mail: rg-777@yandex.ru

Гизатулина Диана Юрьевна, аспирант ИФМК КФУ, учитель ОШИ IT-лицей КФУ.

E-mail: giz-diana@mail.ru

Бегаева Мария Николаевна, аспирант КФУ ИФМК НИЛ «Интеллектуальные технологии управления текстами».

E-mail: Masha.begaeva95@gmail.com