Terra Linguistica. 2023, Том 14, № 1. С. 30–40.
Terra Linguistica. 2023, Vol. 14, No. 1. Pp. 30–40.

Научная статья
УДК 81.32
DOI: https://doi.org/10.18721/JHSS.14104

ИССЛЕДОВАНИЕ ВЛИЯНИЯ ПАРАМЕТРОВ МОРФОЛОГИЧЕСКОЙ СЛОЖНОСТИ НА ТРУДНОСТЬ ВОСПРИЯТИЯ МЕДИАТЕКСТА С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ СТАТИСТИЧЕСКОГО АНАЛИЗА ДАННЫХ

Т.Г. Евтушенко[1] ✉, Е.С. Клочкова[1], А.В. Лапутенко[2], Н.В. Евтушенко[3]

[1] Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург, Российская Федерация;

[2] Национальный исследовательский Томский государственный университет, г. Томск, Российская Федерация;

[3] Институт системного программирования РАН, Москва, Российская Федерация

✉ evtushenkotg@gmail.com

Аннотация. Предлагаемая работа посвящена изучению одного из аспектов сложности, влияющих на восприятие медиатекста: параметров морфологической сложности, а также их взаимодействию с поверхностными характеристиками текста, такими как средняя длина предложения, средняя длина слова и т.п. В работе исследуется вопрос о связи количественных параметров (метрик) объективной сложности текста, которая обусловлена его языковыми характеристиками, со степенью трудности восприятия текста читателем. Для определения и подсчета метрик морфологической сложности использовался корпус из 1000 размеченных новостных текстов (общим объемом 140000 словоупотреблений) с веб-сайтов российских ВУЗов. Для каждого текста были подсчитаны следующие величины: доля слов различных частей речи, доля отдельных граммем, соотношение именности-глагольности, соотношение знаменательных и служебных частей речи, средняя длина предложения, средняя длина слова и т.д. Анализ морфологической сложности был дополнен результатами опроса представителей целевой аудитории веб-сайта ВУЗа (абитуриентов, студентов и аспирантов), которые оценили трудность 255 новостных текстов по пятибалльной шкале. Далее на основе собранных данных проводился корреляционно-регрессионный анализ для определения значимости анализируемых метрик морфологической сложности и степени их влияния на трудность восприятия текста. На основе анализа используемых полученных моделей линейной регрессии было установлено, что наиболее значимыми метриками морфологической сложности являются доля полных причастий, доля словоформ в родительном падеже, доля кратких прилагательных и доля числительных. Кроме того, проведенный анализ подтвердил вывод предыдущих исследований о значимости таких поверхностных метрик как средняя длина предложения и средняя длина словоформы. В результате анализа были предложены две формулы для расчета степени понятности новостного текста: 1) формула, основанная на трех метриках, которые чаще всего встречаются в моделях; 2) формула, основанная на модели с наиболее высокой точностью и учитывающая пять морфологических и поверхностных метрик.

Ключевые слова: сложность текста, понятность, морфологические параметры, медиатекст, корреляционно-регрессионный анализ.

Финансирование: Проект выполнен при финансовой поддержке программы стратегического академического лидерства «Приоритет 2030» Российской Федерации (Договор № 075-15-2021-1333 от 30.09.2021).

Для цитирования: Евтушенко Т.Г., Клочкова Е.С., Лапутенко А.В., Евтушенко Н.В. Исследование влияния параметров морфологической сложности на трудность восприятия медиатекста с использованием методов статистического анализа данных // Terra Linguistica. 2023. Т. 14. № 1. С. 30–40. DOI: 10.18721/JHSS.14104

Research article
DOI: https://doi.org/10.18721/JHSS.14104

STUDYING THE IMPACT OF MORPHOLOGICAL PARAMETERS ON TEXT READABILITY USING STATISTICAL ANALYSIS METHODS

T.G. Evtushenko[1] ✉, Y.S. Klochkova[1], A.V. Laputenko[2], N.V. Evtushenko[3]

[1] Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russian Federation;

[2] National Research Tomsk State University, Tomsk, Russian Federation;

[3] Institute for System Programming of the Russian Academy of Sciences, Moscow, Russian Federation

✉ evtushenkotg@gmail.com

Abstract. The paper addresses one of the important aspects of text complexity, namely the dependency of text readability on a set of morphological and text surface metrics such as the average length of words, sentences, etc. The correlation between the objective text complexity which is specified by quantitative parameters of the linguistic features and the subjective text complexity, i.e. the difficulty of text comprehension as a psychological phenomenon, is analyzed. To assess the morphological text complexity we used an annotated dataset consisting of 1000 online news texts (140000 tokens) retrieved from the websites of Russian universities. For each text unit the ratio of each part-of-speech per token is measured. Online news texts of the dataset were also assessed by a target audience of the website, i.e. applicants, undergraduate and postgraduate students. As a result, the dataset was automatically annotated based on text linguistic features and human-labelled based on experts' estimates of text readability on a 5-point scale. To assess the significance of morphological metrics and their influence on text readability, the correlation and regression analysis was carried out. To automatically classify a text as 'easy-to-read' or not 'easy-to-read', both single feature and compound models including more than one metric were constructed. In agreement with the prior research the most common metrics influencing text readability appear to be text surface characteristics. However, the proposed models also made it possible to establish the significance of morphological parameters, used both in single feature and compound models, such as the use of participles, nouns in the genitive case, adjectives and numerals, which should be taken into account in analyzing news text readability. Moreover, novel formulae for assessing readability were proposed based on the studied coefficients.

Keywords: text complexity, readability, morphological features, media text, correlation and regression analysis.

Acknowledgements: The project was implemented with the financial support of the strategic academic leadership program "Priority 2030" of the Russian Federation (project No. 075-15-2021-1333 of 30.09.2021).

Citation: T.G. Evtushenko, E.S. Klochkova, A.V. Laputenko, N.V. Evtushenko, Studying the impact of morphological parameters on text readability using statistical analysis methods, Terra Linguistica, 14 (1) (2023) 30–40. DOI: 10.18721/JHSS.14104

Введение

Предлагаемое исследование посвящено выявлению тех морфологических характеристик сложности, которые оказывают существенное влияние на трудность восприятия текста читателем.

Восприятие и понимание текста зависит от его сложности как совокупности языковых средств разных языковых уровней, которые используются автором для создания текста в соответствии с определенной коммуникативной задачей и с учетом конкретной коммуникативной ситуации.

Являясь исходно математическим понятием, а следовательно, и объектом изучения в математических дисциплинах и теории информации, сложность текста трансформировалась в трансдисциплинарную область исследования [1]. В области лингвистики исследования сложности имеют длительную историю изучения (см. далее). При этом сложность как лингвистический конструкт рассматривается с различных точек зрения: как характеристика языковой системы в целом [2] и как характеристика отдельного речевого произведения [3]. В данном исследовании речь пойдет о втором аспекте сложности.

Разработка проблемы сложности текста ведется в зарубежной лингвистике, преимущественно в американской, уже с 40-х гг. 20 века. Для измерения сложности были выработаны формулы, среди которых наиболее популярными являются формула Флеша-Кинкейда [4, 5], формула читабельности SMOG [6], автоматический индекс удобочитаемости, индекс Колман-Лиау и ряд других [3]. Все перечисленные формулы основаны на учете количественных параметров текста (метрик), которые отражают сложность единиц различных языковых уровней: морфологического (длина слова в символах и слогах), синтаксического (длина предложения, длина самого текста), лексического (доля низкочастотной лексики в тексте).

В советской лингвистике в 70-х гг. 20 века проблема сложности текста изучалась в рамках направления квантитативной лингвистики [7]. В частности, были предложены формула М.С. Мацковского [8] и формула Ю.А. Тулдавы [9]. Эти формулы для расчета сложности текста так же, как и зарубежные, основаны на учете таких количественных параметров текста, как средняя длина предложения и длина слова в слогах. Популярностью пользуется также адаптированная для русского языка формула Флеша-Кинкейда [10]. Более подробный обзор работ данного направления представлен в работе [3].

В настоящее время в исследовании сложности можно отметить следующие тенденции.

Во-первых, применяется дифференцированный подход к сложности в зависимости от его функционально-стилистической принадлежности. В частности, отдельно изучаются сложность дидактического текста [3], сложность юридических документов [11–13]. Исследователи также разрабатывают подходы к автоматизированной оценке сложности учебных текстов на русском языке как иностранном [14, 15].

Во-вторых, в современных лингвистических исследованиях широко применяются методы автоматизированной обработки текстов для решения различных задач анализа речевого материала [16–18]. Такие методы, в частности модели машинного обучения, используются и при разработке моделей сложности текста.

Так, например, в исследованиях А.Н. Лапошиной сложность текста определяется с помощью регрессионной модели, обученной на корпусе из 800 текстов из пособий по РКИ [19]. В основе моделей сложности, разрабатываемых в рамках проектов РНФ 19-18-00525 «Понятность официального русского языка: юридическая и лингвистическая проблематика» [11, 12], заложен алгоритм градиентного бустинга на основе деревьев решений для задачи классификации. В работах, посвященных анализу сложности учебных текстов, предлагаются формулы, полученные на основе линейной регрессии с регуляризацией [20–22].

Таким образом, можно отметить, что за достаточно длительную историю изучения сложности текста в лингвистике разработаны различные подходы к определению этого понятия и вы-

явлению факторов, влияющих на тот или иной уровень сложности конкретного текста. Однако многие задачи, решение которых необходимо для понимания этого явления, все еще не имеют однозначного ответа. Одной из таких задач является выявление тех лингвистических параметров, которые оказывают наибольшее воздействие на понимание текста читателем, причем особый интерес представляет определение комбинаций нескольких параметров, поскольку, как показывают предыдущие исследования, измерение метрик сложности изолированно не дает адекватного представления о сложности целого текста.

Применение методов компьютерной лингвистики и больших массивов текстов позволило, наряду с уже известными формулами читабельности, поставить вопрос об учете большего количества метрик, учитывающих влияние на сложность текста языковых явлений разных уровней.

Насколько нам известно, на данный момент при определении уровня сложности текста не учитываются или ограниченно учитываются следующие факторы: 1) взаимосвязь объективной сложности текста и трудности его восприятия читателем; 2) влияние морфологических параметров, таких как соотношение слов разных частей речи в тексте, употребление отдельных граммем и т.п.

Таким образом, предлагаемая работа направлена на определение ряда морфологических параметров и их частотных комбинаций, которые оказывают влияние на трудность восприятия текста читателем. В работе мы придерживаемся дифференцированного подхода к определению сложности и фокусируемся только на медиатекстах на русском языке.

Подходы к определению и измерению сложности текста

Обсуждая сложность текста с лингвистической точки зрения, исследователи вводят различные термины для обозначения этого понятия, что отражает его определенную двойственность. Большинство исследователей так или иначе выделяют две стороны этого явления, объективную и субъективную сложность, для обозначения которых используют разные термины. Набор объективных лингвистических характеристик, в основном формальных, присущих тому или иному произведению, обозначается как собственно «сложность» (в зарубежных источниках – complexity). Субъективная сложность, которая рассматривается как психолингвистическое явление, обозначается терминами «трудность», «понятность». Субъективная сложность предполагает учет таких психолингвистических параметров, как когнитивные способности читающих, наличие у них определенных фоновых знаний, мотивации к чтению и т.п. [2, 23]. В работах зарубежных исследователей используются также такие термины, как «читабельность», «удобочитаемость» (англ. readability), которые обозначают уровень легкости восприятия текста читателем, что соотносится с его понятностью. В данной работе, вслед за Н.С. Валгиной под понятностью текста будем понимать “возможность определить смысл, доходчивость – возможность преодолеть «препятствия», возникающие при передаче информации” [24].

Как отмечалось в предыдущих исследованиях, сложность текста можно рассматривать как переменную, значение которой вычисляется на основе числовых показателей соответствующих признаков. Измерение сложности может осуществляться посредством подсчета известных индексов удобочитаемости по формулам, представленным в литературе (см.выше), учитывающим поверхностные, или базовые (surface metrics or baseline surface features), характеристики текста, такие как длина предложения, количество слов и предложений в тексте. Кроме того, в ряде работ учитываются количество языковых единиц разных уровней. На уровне морфологии выделяют такие показатели, как доля слов разных частей речи, формы родительного и творительного падежей существительных, соотношение глаголов и существительных и т.п. [25].

Методы и материал исследования

В качестве материала исследования в работе использовался датасет, сформированный в ходе реализации проекта «Цифровые технологии в лингвистике: модель автоматической оценки рече-

вого воздействия мультимодального электронного текста». Датасет включает новостные тексты на русском языке с сайтов ведущих российских ВУЗов. Все тексты были размечены с помощью pymorphy2 и синтаксического анализатора natasha.

Исследование проводилось по алгоритму:

- отбор имеющихся в литературе базисных и морфологических метрик и выбор наиболее релевантных из них для медиатекстов;
- опрос целевой аудитории для оценки понятности медиатекста;
- статистическая обработка собранных данных;
- анализ результатов с целью выявления наиболее влиятельных метрик и их частотных сочетаний.

Опрос по оценке понятности текстов проводился среди студентов и преподавателей ВУЗов: количество студентов – 90 человек, количество текстов – 250. Респондентам предоставлялись тексты с веб-сайтов ведущих высших образовательных учреждений и анкета, в которой они должны были проставить оценки читабельности/понятности предъявленных текстов. Оценка каждого текста ставилась как средняя оценка на основании результатов прочтения каждого текста тремя экспертами по 5-балльной шкале; обработка результатов опроса проводилась по методу экспертной оценки.

Существующие модели машинного обучения, в основном, решают задачи классификации или предсказания. Большинство таких моделей не позволяют выявить вес лингвистических признаков и их комбинаций, которые могут влиять в той или иной мере на понятность текста. В то же время, классические алгоритмы регрессионного анализа позволяют создавать как описательные, так и предсказательные модели. Как отмечалось выше, целью работы является определение морфологических характеристик текста, влияющих на трудность восприятия текста, и выявление частотных комбинаций этих параметров, наиболее часто встречающихся в статистических моделях, т.е. параметров, которые с высокой частотой встречаются в полученных 155 моделях.

При составлении массива данных на основе уже имеющихся списков метрик, составленных авторами других работ, с учетом жанра текста и релевантности метрик для статистического анализа, были отобраны соответствующие морфологические параметры, пронумерованные для удобства анализа следующим образом:

(0) индекс аналитичности,
(1) индекс субстантивности,
(2) индекс местоименности,
(3) доля словоформ в родительном падеже,
(4) доля словоформ в творительном падеже,
(5) доля кратких прилагательных,
(6) доля полных причастий,
(7) доля деепричастий,
(8) доля инфинитивов,
(9) доля числительных,
(10) доля частиц,
(11) соотношение имённости-глагольности.

При проведении анализа влияния метрик на понятность текста мы исключили общепринятые формулы индексов читабельности. Использование готовых формул ограничивало проводимое исследование, вследствие чего индексы читабельности были разложены на отдельные метрики, чтобы изучить сочетаемость поверхностных характеристик непосредственно с морфологическими текстовыми параметрами. Среди поверхностных характеристик текста были выделены следующие:

(12) средняя длина слова в слогах(ASW);
(13) среднее количество букв на 100 слов;

(14) среднее количество предложений на 100 слов;
(15) доля длинных слов (слова длиннее 6 букв);
(16) средняя длина предложения в словах (ASL);
(17) среднее количество букв и цифр в слове (CbyW).

Построение формулы для определения воспринимаемости текста на основе его морфологических характеристик

Для определения наиболее существенных морфологических метрик, влияющих на восприятие текста, был выбран метод корреляционно- регрессионного анализа [26]. Анализ значимости выбранных метрик выполнялся на основе полного перебора всех моделей линейной регрессии, содержащих все возможные комбинации из 18 выбранных метрик для 255 оцененных экспертами текстов.

Ниже представлен фрагмент таблицы (табл. 1) с комбинацией наиболее значимых для решения данной задачи метрик, где RMSE — значение корня из среднеквадратической ошибки, а коэффициенты могут быть положительными и отрицательными. Знак «–» указывает на то, что при увеличении количества словоупотреблений в данной форме трудность восприятия текста возрастает. Отсутствие знака «-» перед коэффициентом предполагает положительное значение количественной характеристики и указывает на то, что при увеличении количества словоупотреблений в данной форме трудность восприятия текста снижается. Так как исходные значения отдельных метрик для 255 текстов лежали в различных числовых диапазонах, все значения предварительно были стандартизованы для приведения к единой шкале. Соответственно, числовые значения коэффициентов для каждой метрики ниже приведены в единицах стандартных отклонений для этой конкретной метрики, что позволяет сравнивать метрики между собой по величине влияния на трудность текста.

Таблица 1. Фрагмент таблицы с построенными моделями
Table 1. A fragment of a table with constructed models

Комбинация метрик	RMSE	Коэффициенты
6, 14, 17	0,64	-0,1393 0,1805 -0,0921 3,8199
6, 10, 12, 16, 17	0,62	–0,1150; –0,0983; –0,2110; –0,1408; –0,1628; 3,8199
6, 9, 14, 15	0,63	–0,1201; –0,1172; 0,1659; –0,1256; 3,8199
6, 9, 13, 16	0,63	–0,1122; –0,1124; –0,1493; –0,1536; 3,8199
6, 9, 15, 16	0,63	–0,1261; –0,1042; –0,1298; –0,1626; 3,8199
10, 12, 16, 17	0,63	–0,1015; –0,2371; –0,1527; –0,1606; 3,8199
6, 9, 14, 15	0,63	–0,1201; –0,1172; 0,1659; –0,1256; 3,8199
6, 16	0,64	–0,1384; –0,1782; 3,8199
6, 14	0,64	–0,1315; 0,1759; 3,8199
14	0,66	0,2019; 3,8199
16	0,66	–0,1999; 3,8199
12	0,66	–0,1836; 3,8199
1, 6	0,66	–0,1001; –0,1667; 3,8199
6	0,67	–0,1663; 3,8199
2	0,67	0,1563; 3,8199

Из всего множества построенных моделей были выбраны 155 моделей, каждый коэффициент регрессии в которых является статистически значимым. Каждая модель имеет в своем составе не более 5 метрик. В табл. 2 приведен список метрик, упорядоченных по количеству моделей (из

числа выбранных 155 моделей), в которых соответствующая метрика присутствует. Среди наиболее часто «используемых» метрик оказалась доля полных причастий (входит в состав 65 моделей), среднее количество предложений на 100 слов (входит в состав 58 моделей) и среднее количество букв и цифр в слове (входит в состав 55 моделей). Из исходного набора метрик в этот список не попала доля словоформ в творительном падеже и индекс аналитичности/автосемантичности.

Относительно высокая частота вхождения отдельной метрики в различные модели линейной регрессии может говорить о ее важности для анализа текстов в рамках описанной задачи при использовании более сложных моделей.

Таблица 2. Распределение частотности встречаемости той или иной метрики в моделях
Table 2. Distribution of the frequency of occurrence of a particular metric in the models

№ метрики	Метрика	Частота
6	Доля полных причастий	65
14	Среднее количество предложений на 100 слов	58
17	Среднее количество букв и цифр в слове (CbyW)	55
16	Средняя длина предложения в словах (ASL)	45
13	Среднее количество букв на 100 слов	38
3	Доля словоформ в родительном падеже	35
5	Доля кратких прилагательных	34
9	Доля числительных	31
12	Средняя длина слова в слогах (ASW)	31
15	Доля длинных слов	27
10	Доля частиц	16
8	Доля инфинитивов	14
2	Индекс местоименности	11
1	Индекс субстантивности	10
11	Соотношение имённости-глагольности	2
7	Доля деепричастий	1
0	Индекс аналитичности/автосемантичности	0
4	Доля словоформ в творительном падеже	0

Например, для модели (первая строчка в табл. 1) на основе только трех самых частотных метрик была получена формула (1):

$$y = -0,1393 \cdot x_6 + 0,1805 \cdot x_{14} - 0,0921 \cdot x_{17} + 3,8199, \quad (1)$$

где y — экспертная оценка (уровень воспринимаемости текста); x_6 — доля полных причастий; x_{14} — среднее количество предложений на 100 слов; x_{17} — среднее количество букв и цифр в слове.

Согласно данной формуле трудность восприятия текста возрастает с увеличением доли полных причастий, средней длины слова и средней длины предложения.

Коэффициенты при соответствующих параметрах позволяют судить об относительной степени их положительного или отрицательного влияния на величину целевой переменной, в данном случае на оценку понятности текста. Согласно приведенной модели при увеличении в тексте доли полных причастий воспринимаемость текста уменьшится на 0,1393, при увеличении на 1 среднего числа предложений более чем со 100 словами, воспринимаемость увеличится на 0,1805, при увеличении среднего количества букв и цифр в слове, воспринимаемость уменьшится на 0,0921. Значение корня из среднеквадратической ошибки (RMSE) для представленной модели

линейной регрессии равно 0,6376, т.е., в среднем ошибка определения уровня воспринимаемости с помощью данной модели составляет 0,6376 единиц.

Наименьшим значением ошибки RMSE (0,6166) из отобранных 155 моделей обладает модель, приведенная ниже и учитывающая значения 5 метрик. Данные метрики также часто сочетаются с другими метриками в рамках отдельных моделей (например, вторая строчка в табл. 1, что отражено в формуле (2):

$$y = -0,1150 \cdot x_6 - 0,0983 \cdot x_{10} - 0,2110 \cdot x_{12} - 0,1408 \cdot x_{16} - 0,1628 \cdot x_{17} + 3,8199, \quad (2)$$

где x_6 — доля полных причастий; x_{10} – доля частиц; x_{12} – средняя длина слова в слогах (ASW); x_{16} – средняя длина предложения в словах (ASL); x_{17} — среднее количество букв и цифр в слове.

Согласно формуле (2) трудность восприятия текста возрастает с увеличением доли полных причастий, доли частиц, средней длины слова и средней длины предложения.

Заключение

Таким образом, на основе проведенных исследований были получены следующие результаты. На основе анализа полученных моделей линейной регрессии были определены наиболее часто встречающиеся метрики, позволяющие адекватно описать уровень воспринимаемости медиатекста в зависимости от его объективных характеристик.

Наиболее встречающимися метриками в построенных моделях являются поверхностные: среднее количество предложений на 100 слов, среднее количество букв и цифр в слове (CbyW), средняя длина предложения в словах (ASL), что хорошо согласуется с результатами предыдущих исследований.

Среди морфологических метрик, существенно влияющих на восприятие: доля полных причастий, доля словоформ в родительном падеже, доля кратких прилагательных, доля числительных.

На основе первых трех наиболее частотных в моделях метрик в их взаимосвязи предложена формула (1) для определения степени понятности медиатекста.

Модель с наиболее высокой точностью из рассмотренных представлена формулой (2). Эта модель учитывает комбинацию 5 метрик: доля полных причастий, доля частиц, средняя длина слова в слогах, средняя длина предложения в словах и среднее количество символов в слове.

Использованный алгоритм отбора моделей линейной регрессии является универсальным и может быть использован для текстов другой жанрово-стилистической принадлежности.

Как видно из представленных результатов, морфологические метрики тесно связаны с длиной словоформы и длиной предложения. Такая корреляция, на наш взгляд, обусловлена увеличением количества языковых единиц и иерархических и линейных связей между ними, которые читатель обрабатывает при восприятии текста.

Таким образом, наше исследование расширяет возможности автоматической оценки читабельности медиатекстов, помещенных на сайтах ведущих ВУЗов. Тексты, ориентированные на читателя (легкочитаемые), позволят увеличить читательскую аудиторию, что может способствовать повышению популярности вуза в медийном пространстве. В дальнейшем мы предполагаем изучить группы метрик других языковых уровней с целью определения параметров, в наибольшей степени влияющих на понятность текста.

СПИСОК ИСТОЧНИКОВ

1. **Bastardas-Boada A.** From language shift to language revitalization and sustainability. A complexity approach to linguistic ecology. Barcelona: Edicions de la Universitat de Barcelona, 2019, pp. 337–349.

2. **Dahl Ö.** The growth and maintenance of linguistic complexity. Amsterdam: John Benjamins, 2004. 336 p.

3. **Солнышкина С.И., Кисельников А.С.** Сложность текста: этапы изучения в отечественном прикладном языкознании // Вестник Томского государственного университета. Филология. 2015. № 6 (38). С. 86–99. DOI: 10.17223/19986645/38/7

4. **Flesch R.** The Art of Readable Writing. Harper & Row, 1949. 237 p.

5. **Kincaid J.P., Fishburne R.P., Rogers R.L., Chissom B.S.** Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel (Research Branch Report 8-75). Memphis, TN: Naval Air Station, 1975. 40 p.

6. **McLaughlin G.H.** SMOG Grading – a New Readability Formula // Journal of Reading. 1969. 12 (8). P. 639–646.

7. **Пиотровский Р.Г., Бектаев К.Б., Пиотровская А.А.** Математическая лингвистика. М.: Высшая школа, 1977. 383 с.

8. **Мацковский М.С.** Проблемы читабельности печатного материала // Смысловое восприятие речевого сообщения в условиях массовой коммуникации. М., 1976. С. 126–142.

9. **Тулдава Ю.А.** Об измерении трудности текстов // Учен. зап. Тарт. ун-та: Труды по методике преподавания иностранных языков. 1975. Вып. 345. С. 102–120.

10. **Оборнева И.В.** Автоматизированная оценка сложности учебных текстов на основе статистических параметров: дис. ... канд. пед. наук. М., 2006. 165 с.

11. **Белов С.А., Гулида В.Б.** Язык юридических документов: сложности понимания // Acta Linguistica Petropolitana. Труды Института лингвистических исследований РАН. Т. 15. Ч. 1. 2019. С. 56–103. DOI 10.30842/alp2306573715104

12. **Blinova O.V., Belov S.A.** Legal corpus «CorRIDA» and lexical complexity assessment of Russian official texts // Book of Abstracts of the 1st International Conference of the Austrian Association for Legal Linguistics "Contemporary Approaches to Legal Linguistics", University of Vienna, 2019. P. 42.

13. **Блинова О.В., Тарасов Н.А.** Сложность русских правовых текстов: методы оценки и языковые данные // Труды международной конференции «Корпусная лингвистика-2021». СПб.: Скифия-принт, 2021. С. 175–182.

14. **Лапошина А.Н.** Автоматическое определение сложности текста по РКИ // Сборник материалов международной научно-практической интернет-конференции «Актуальные вопросы описания и преподавания русского языка как иностранного/неродного». М., 2018. С.573–579

15. **Laposhina A.N., Veselovskaya T.S., Lebedeva M.U., Kupreshchenko O.F.** Automated Text Readability Assessment For Russian Second Language Learners // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2018". Issue 17 (24). 2018. P. 403–413. DOI: 10.22363/2618-8163-2021-19-3-331-345

16. **Се Линьи, Загайнов А.И.** Моделирование характеристик персонажей и их взаимосвязей в сюжете художественного произведения методами численного фрактального анализа // Terra Linguistica. 2022. Т. 13, № 3. С. 36–47. DOI: 10.18721/JHSS.13304

17. **Митрофанова О.А., Гаврилик Д.А.** Эксперименты по автоматическому выделению ключевых выражений в стилистически разнородных корпусах русскоязычных текстов // Terra Linguistica. 2022. Т. 13, № 4. С. 22–40. DOI: 10.18721/JHSS.13402

18. **Андреев В.С**. Экспоненциальное распределение частей речи в стихотворном тексте: опыт стилеметрического анализа // Общество. Коммуникация. Образование. 2021. Т. 12, № 4. С. 94–104. DOI: 10.18721/JHSS.12407

19. **Лапошина А.Н., Лебедева М.Ю.** Текстометр: онлайн-инструмент определения уровня сложности текста по русскому языку как иностранному // Русистика. 2021. Т. 19. №3. С. 331–345. DOI: 10.22363/2618-8163-2021-19-3-331-345

20. **Solovyev 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 (5). Pp. 3049–3058. DOI: 10.3233/JIFS-169489

21. **Solovyev V., Solnyshkina M., Ivanov V.** Prediction of reading difficulty in Russian academic texts // Journal of Intelligent & Fuzzy Systems. 2019. Vol. 36. Is. 5. P. 4553–4563. DOI: 10.3233/JIFS-179007

22. **Solnyshkina M., Ivanov V., Solovyev V.** Readability Formula for Russian Texts: A Modified Version // Proceedings of the 17th Mexican International Conference on Artificial Intelligence. Guadalajara. 2018. Part II. P. 132–145. DOI: 10.1007/978-3-030-04497-8_11

23. **Томина Ю.А.** Объективная оценка языковой трудности текстов (описание, повествование, рассуждение, доказательство): дис. ... канд. пед. наук. М., 1985. 226 с.

24. **Валгина Н.С.** Теория текста. Москва.: Логос, 2003. 173 с

25. **Блинова О.В., Тарасов Н.А.** Метрики сложности русских правовых текстов: отбор, использование, первичная оценка эффективности //Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог». Вып. 21, дополнительный том. 2022. С. 1017–1028

26. **James G., Witten D., Hastie T., Tibshirani R.** An Introduction to Statistical Learning. In Springer Texts in Statistics. Springer New York. 2013. 426 p. DOI: 10.1007/978-1-4614-7138-7

REFERENCES

[1] **A. Bastardas-Boada,** From language shift to language revitalization and sustainability. A complexity approach to linguistic ecology. Barcelona: Edicions de la Universitat de Barcelona, 2019, pp. 337–349.

[2] **Ö. Dahl,** The growth and maintenance of linguistic complexity. Amsterdam: John Benjamins, 2004.

[3] **S.I. Solnyshkina, A.S. Kiselnikov,** Slozhnost teksta: etapy izucheniya v otechestvennom prikladnom yazykoznanii, Vestnik Tomskogo gosudarstvennogo universiteta. Filologiya. 6 (38) (2015) 86–99. DOI: 10.17223/19986645/38/7

[4] **R. Flesch,** The Art of Readable Writing. Harper & Row, 1949.

[5] **J.P. Kincaid, R.P. Fishburne, R.L. Rogers, B.S. Chissom,** Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel (Research Branch Report 8-75). Memphis, TN: Naval Air Station, 1975.

[6] **G.H. McLaughlin,** SMOG Grading – a New Readability Formula, Journal of Reading. 12 (8) (1969) 639–646.

[7] **R.G. Piotrovskiy, K.B. Bektaev, A.A. Piotrovskaya,** Matematicheskaya lingvistika. Vysshaya shkola, Moscow, 1977.

[8] **M.S. Matskovskiy,** Problemy chitabelnosti pechatnogo materiala, Smyslovoye vospriyatiye rechevogo soobshcheniya v usloviyakh massovoy kommunikatsii. M., 1976. Pp. 126–142.

[9] **Yu.A. Tuldava,** Ob izmerenii trudnosti tekstov, Uchen. zap. Tart. un-ta: Trudy po metodike prepodavaniya inostrannykh yazykov. 345 (1975) 102–120.

[10] **I.V. Oborneva,** Avtomatizirovannaya otsenka slozhnosti uchebnykh tekstov na osnove statisticheskikh parametrov: dis. ... kand. ped. nauk. M., 2006. 165 p.

[11] **S.A. Belov, V.B. Gulida,** Yazyk yuridicheskikh dokumentov: slozhnosti ponimaniya, Acta Linguistica Petropolitana. Trudy Instituta lingvisticheskikh issledovaniy RAN. 15 (1) (2019) 56–103. DOI 10.30842/alp2306573715104

[12] **O.V. Blinova, S.A. Belov,** Legal corpus "CorRIDA" and lexical complexity assessment of Russian official texts, Book of Abstracts of the 1st International Conference of the Austrian Association for Legal Linguistics "Contemporary Approaches to Legal Linguistics", University of Vienna, 2019. P. 42.

[13] **O.V. Blinova, N.A. Tarasov,** Slozhnost russkikh pravovykh tekstov: metody otsenki i yazykovyye dannyye [The complexity of Russian legal texts: assessment methods and language data], Proceedings of the International Conference "Corpus Linguistics-2021". SPb.: Skifiya-print, 2021. Pp. 175–182.

[14] **A.N. Laposhina,** Avtomaticheskoye opredeleniye slozhnosti teksta po RKI [Automatic determination of the complexity of the text by the RCT], Collection of materials of the international scientific and practical Internet conference "Topical issues of describing and teaching Russian as a foreign/non-native language". M., 2018. Pp.573–579.

[15] **A.N. Laposhina, T.S. Veselovskaya, M.U. Lebedeva, O.F. Kupreshchenko,** Automated Text Readability Assessment For Russian Second Language Learners, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2018". Issue 17 (24) (2018) 403–413. DOI: 10.22363/2618-8163-2021-19-3-331-345

[16] **Xie Linyi, A.I. Zagaynov,** Modeling of character characteristics and their relationships in a novel plot by methods of numerical fractal analysis, Terra Linguistica, 13 (3) (2022) 36–47. DOI: 10.18721/JHSS.13304

[17] **O.A. Mitrofanova, D.A. Gavrilic,** Experiments on automatic keyphrase extraction in stylistically heterogeneous corpus of Russian texts, Terra Linguistica, 13 (4) (2022) 22–40. DOI: 10.18721/JHSS.13402

[18] **V.S. Andreev,** Exponential distribution of parts of speech in verse text: experience in stylometric analysis, Society. Communication. Education, 12 (4) (2021) 94–104. DOI: 10.18721/JHSS.12407

[19] **A.N. Laposhina, M.Yu. Lebedeva,** Tekstometr: onlayn-instrument opredeleniya urovnya slozhnosti teksta po russkomu yazyku kak inostrannomu, Rusistika. 19 (3) (2021) 331–345. DOI: 10.22363/2618-8163-2021-19-3-331-345

[20] **V. Solovyev, V. Ivanov, M. Solnyshkina,** Assessment of reading difficulty levels in Russian academic texts: Approaches and metrics, Journal of Intelligent & Fuzzy Systems. 34 (5) (2018) 3049–3058. DOI: 10.3233/JIFS-169489

[21] **V. Solovyev, M. Solnyshkina, V. Ivanov,** Prediction of reading difficulty in Russian academic texts, Journal of Intelligent & Fuzzy Systems. 36 (5) (2019) 4553–4563. DOI: 10.3233/JIFS-179007

[22] **M. Solnyshkina, V. Ivanov, V. Solovyev,** Readability Formula for Russian Texts: A Modified Version, Proceedings of the 17th Mexican International Conference on Artificial Intelligence. Guadalajara. 2018. Part II. P. 132–145. DOI: 10.1007/978-3-030-04497-8_11

[23] **Yu.A. Tomina,** Obyektivnaya otsenka yazykovoĭ trudnosti tekstov [Objective assessment of the linguistic difficulty of texts] (description, narration, reasoning, proof): dis. ... Candidate of Pedagogical Sciences. M., 1985. 226 p.

[24] **N.S. Valgina,** Teoriya teksta [Text theory]. Moskva.: Logos, 2003. 173 p.

[25] **O.V. Blinova, N.A. Tarasov,** Metriki slozhnosti russkikh pravovykh tekstov: otbor, ispolzovaniye, pervichnaya otsenka effektivnosti [Metrics of complexity of Russian legal texts: selection, use, primary evaluation of effectiveness], Computational linguistics and intelligent technologies: Based on the materials of the annual international conference "Dialogue". Vyp. 21, dopolnitelnyy tom. 2022. Pp. 1017–1028.

[26] **G. James, D. Witten, T. Hastie, R. Tibshirani,** An Introduction to Statistical Learning. In Springer Texts in Statistics. Springer New York. 2013. 426 p. DOI: 10.1007/978-1-4614-7138-7

СВЕДЕНИЯ ОБ АВТОРАХ / INFORMATION ABOUT AUTHORS

Евтушенко Татьяна Геннадьевна
Tatiana G. Evtushenko
E-mail: evtushenkotg@gmail.com
ORCID: https://orcid.org/0000-0001-5338-3656

Клочкова Елена Сергеевна
Yelena S. Klochkova
E-mail: klochkova_es@spbstu.ru
ORCID: https://orcid.org/0000-0002-6326-8392

Лапутенко Андрей Владимирович
Andrey V. Laputenko
E-mail: laputenko.av@gmail.com

Евтушенко Нина Владимировна
Nina V. Evtushenko
E-mail: evtushenko@ispras.ru
ORCID: https://orcid.org/0000-0002-4006-1161

Поступила: 27.01.2023; Одобрена: 06.03.2023; Принята: 17.03.2023.
Submitted: 27.01.2023; Approved: 06.03.2023; Accepted: 17.03.2023.