Russian Journal of Linguistics
ISSN 2687-0088 (print), ISSN 2686-8024 (online)

2022 Vol. 26 No. 2 493–514
http://journals.rudn.ru/linguistics

https://doi.org/10.22363/2687-0088-30084

Research article

Word frequency and text complexity: an eye-tracking study of young Russian readers

Antonina N. LAPOSHINA ✉, Maria Yu. LEBEDEVA and Alexandra A. BERLIN KHENIS

Pushkin State Russian Language Institute, Moscow, Russia
✉ANLaposhina@pushkin.institute

Abstract
Although word frequency is often associated with the cognitive load on the reader and is widely used for automated text complexity assessment, to date, no eye-tracking data have been obtained on the effectiveness of this parameter for text complexity prediction for the Russian primary school readers. Besides, the optimal ways for taking into account the frequency of individual words to assess an entire text complexity have not yet been precisely determined. This article aims to fill these gaps. The study was conducted on a sample of 53 children of primary school age. As a stimulus material, we used 6 texts that differ in the classical Flesch readability formula and data on the frequency of words in texts. As sources of the frequency data, we used the common frequency dictionary based on the material of the Russian National Corpus and DetCorpus – the corpus of literature addressed to children. The speed of reading the text aloud in words per minute averaged over the grades was employed as a measure of the text complexity. The best predictive results of the relative reading time were obtained using the lemma frequency data from the DetCorpus. At the text level, the highest correlation with the reading speed was shown by the text coverage with a list of 5,000 most frequent words, while both sources of the lists – Russian National Corpus and DetCorpus – showed almost the same correlation values. For a more detailed analysis, we also calculated the correlation of the frequency parameters of specific word forms and lemmas with three parameters of oculomotor activity: the dwell time, fixations count, and the average duration of fixations. At the word-by-word level, the lemma frequency by DetCorpus demonstrated the highest correlation with the relative reading time. The results we obtained confirm the feasibility of using frequency data in the text complexity assessment task for primary school children and demonstrate the optimal ways to calculate frequency data.
Keywords: *text complexity, text readability, word frequency, eye tracking*

For citation:
Laposhina, Antonina N., Maria Yu. Lebedeva & Alexandra A. Berlin Khenis. 2022. Word frequency and text complexity: An eye-tracking study of young Russian readers. *Russian Journal of Linguistics* 26 (2). 493–514. (In Russian). https://doi.org/10.22363/2687-0088-30084

Научная статья

Влияние частотности слов текста на его сложность: экспериментальное исследование читателей младшего школьного возраста методом айтрекинга

А.Н. ЛАПОШИНА ✉, М.Ю. ЛЕБЕДЕВА, А.А. БЕРЛИН ХЕНИС

Государственный институт русского языка имени А.С. Пушкина, Москва, Россия
✉ANLaposhina@pushkin.institute

Аннотация
Параметр частотности слова во многих исследовательских трудах связывается с когнитивной нагрузкой на читателя и широко используется в автоматических системах анализа сложности текста. Однако к настоящему моменту для русскоязычного материала не представлено достаточное количество экспериментальных данных о влиянии параметра частотности слов на сложность текста, собранных с помощью метода айтрекинга. Кроме того, не определены оптимальные способы учета частотности отдельных слов для характеристики целого текста. Целью данной статьи является заполнение этих лакун. Исследование проводилось на выборке 53 детей младшего школьного возраста. Материалом для эксперимента выступили 6 текстов, отличающихся по параметрам классической формулы читабельности Флеша и данным о частотности слов в текстах. В качестве источников данных о частотности слов использованы как стандартный частотный словарь на материале Национального корпуса русского языка, так и корпус литературы, адресованной детям, ДетКорпус. В качестве меры сложности текста использовался параметр скорости чтения текста вслух в словах в минуту, усредненный по классам. Для более детального анализа были произведены подсчеты корреляции параметров частотности конкретных словоформ и их лемм с тремя параметрами глазодвигательной активности: средней относительной скорости чтения слова, средней длительности фиксаций и средним количеством фиксаций. На пословном уровне анализа наивысший коэффициент корреляции с относительным временем чтения продемонстрировали данные частотности леммы по корпусу детской литературы. На уровне анализа текстов наиболее высокую корреляцию со средним временем чтения фрагмента показал параметр процента покрытия текста списком 5 000 самых частотных слов, при этом данные по разным источникам показали близкие значения. Приведенные результаты айтрекингового эксперимента подтверждают связь сложности текста и частотности входящих в него слов на материале для младших школьников, а также обозначают оптимальную методику и источники подсчета частотности для данной задачи.

Ключевые слова: *сложность текста, читабельность текста, частотность слова, айтрекинг*

Для цитирования:
Лапошина А.Н., Лебедева М.Ю., Берлин Хенис А.А. Влияние частотности слов текста на его сложность: экспериментальное исследование читателей младшего школьного возраста методом айтрекинга. *Russian Journal of Linguistics.* 2022. Т. 26. № 2. С. 493–514. https://doi.org/10.22363/2687-0088-30084

1. Введение

Тенденция к постоянному росту объема информации, характерная для современного этапа развития человечества, в полной мере касается и учебной информации. Новые учебники и пособия, образовательный контент, размещаемый на различных цифровых платформах, требуют тщательной и одновременно быстрой оценки с точки зрения их доступности читателям определенного возраста. Проблема предиктивной оценки сложности учебных материалов стоит особенно остро для категории детей младшего школьного возраста: в это время ребенок не только осваивает технику чтения, но и формирует читательскую грамотность – способность к осмыслению письменных текстов и эффективному взаимодействию с ними. Критически важно, чтобы в этот период человек не сталкивался с нецелесообразными сложностями при чтении.

За более чем столетнюю историю разработки формальных способов оценки сложности текста исследовались самые разные группы признаков текста – лексические, морфологические, синтаксические, семантические, способные указывать на его уровень сложности (DuBay 2007). В качестве базовых и наиболее распространенных рассчитываемых показателей сложности текста выступают значения средней длины слова и средней длины предложения. Логика подсчета таких признаков может варьироваться: кроме длины в знаках, есть формулы, учитывающие длину слова в слогах или процент слов длиннее заданного количества слогов. К настоящему моменту предложено огромное количество подобных формул для разных групп читателей, а также для русскоязычных учебных материалов (Солнышкина, Кисельников 2015, Криони и др. 2008, Шпаковский 2008) и сервисов по автоматизированной оценке сложности текста[1]. При этом такие формальные показатели имеют существенные ограничения.

В качестве аргумента против использования подобных формул часто указывается их низкая интерпретируемость (Мизернов, Гращенко 2015) и игнорирование лексики, составляющей текст (Graesser et al. 2011), – очевидно, что семантическая сложность слова не всегда коррелирует с его длиной (ср., например, слова *здравствуйте* и *штифт*). Чтобы снять это ограничение оценки сложности текста базовыми подсчетами, предлагается учитывать лексическую и семантическую сложность слов, входящих в текст. Одним из наиболее распространенных способов расчета сложности лексики являются подсчеты процента лексики из специально составленных списков «простых»

[1] Аналитик чтения. http://read-analytic.ru (дата обращения: 20.12.2021); Оценка читабельности текста. http://ru.readability.io (дата обращения: 20.12.2021)

слов (например, Chall & Dale 1995). В отечественных классических трудах 1970-х гг., посвященных сложности русских учебных текстов, Я.А. Микк предлагает учитывать «знакомость» слова: количество знакомых слов в тексте, которое определяется испытуемыми эмпирически по пятибалльной шкале (5 – очень хорошо знакомое слово, 0 – незнакомое слово) (Микк 1970).

С развитием корпусной лингвистики и появлением больших корпусов текстов для получения статистики стало возможным объективизировать интуитивное понятие «знакомости» слова с помощью данных о его частотности по релевантным большим коллекциям текстов. Основная идея такого подхода заключается в том, что частотные слова воспринимаются и производятся быстрее, чем редкие (Raney & Rayner 1995): возвращаясь к приведенному выше примеру, все формы слова *здравствовать* встречаются в основном корпусе НКРЯ 17 965 раз, а слова *штифт* – 129 раз. Соответственно, текст с высокой долей частотных слов должен восприниматься лучше (Chen & Meurers 2018). Частотность слова в качестве одного из признаков, оказывающих влияние на сложность, широко используется как в исследованиях для англоязычных материалов (Lexile 2007, Graesser et al. 2014), так и для текстов на русском языке (Solovyev et al. 2018, Glazkova et al. 2021, Иомдин, Морозов 2021). С другой стороны, исследователи указывают на отсутствие значимой корреляционной связи частотности и сложности текста, оцениваемой с помощью классической формулы Флеша-Кинкейда (Мартынова и др. 2020). Следовательно, экспериментальное установление связи (или ее отсутствия) частотности слова со сложностью на материале текстов на русском языке представляется актуальной исследовательской задачей.

Целью данного исследования является экспериментальное уточнение связи параметров движений глаз и частотности слова на материале русскоязычных текстов для младших школьников и определение оптимальной методики учета информации о частотности.

В ходе исследования мы отвечаем на следующие исследовательские вопросы:

1) Оказывает ли параметр частотности слова значимое влияние на сложность текста на русском языке для учеников младших классов?

2) Какой источник информации о частотности слова наиболее релевантен для этой задачи?

3) Какая из предлагаемых методик подсчета данных о частотности – среднее от логарифма нормализованной частотности всех слов текста, процент покрытия текста частотными списками объемом 5 000 слов по двум указанным источникам, процент слов в тексте со значением ipm ниже 5 – оптимальна для поставленной цели?

2. Частотность слова как параметр оценки сложности текста: теоретическое обоснование

Помимо установления связи параметра частотности слов и сложности составленного из них текста отдельную исследовательскую проблему

представляет выбор источника данных о частотности лексики, релевантного выбранной возрастной категории, так как данные о частотности слова сильно зависят от типа и наполнения корпуса, по которому ведутся подсчёты (Ляшевская, Шаров 2009, предисловие к словарю). Ряд исследователей используют для этих целей данные больших национальных корпусов текстов (Dorofeeva et al. 2019, Glazkova et al. 2021, Иомдин, Морозов 2021). Аргументами в пользу этого выбора могут служить большой размер таких корпусов, а также представленность в их составе различных жанров «официального» кодифицированного языка, с которым учащимся предстоит столкнуться в жизни: художественная литература, новости, публицистика – всё это составляет основу данных.

Однако, с другой стороны, остается открытым вопрос о правомерности сравнения материалов для детей с языком «взрослых». Например, ряд слов, высокочастотных по данным Национального корпуса русского языка (НКРЯ), полностью отсутствует в текстах учебников младших классов (*цена, проблема, государство*). Напротив, относительная встречаемость другой группы слов (*лес, птица, мороз)* значительно выше в текстах учебников, чем в коллекции текстов НКРЯ (Лапошина et al. 2019). Поэтому альтернативным решением здесь может стать подсчет частотности по специальным коллекциям текстов, предназначенным для детей (Lexile 2007).

Наконец, существуют разные способы учета частотной информации для текстовых фрагментов. Классический способ, представленный еще в ранних формулах читабельности, предлагает расчет процента слов текста, входящих в релевантный список слов, одной из разновидностей которого может стать частотный список. Этот метод расчета и сейчас используется в ряде исследований сложности текста (Glazkova et al. 2021, Sato 2014). Ещё один популярный способ учета частотности слов текста – это расчет среднего или медианного значения из частотности каждого слова текста (Francois & Fairon 2012, Reynolds 2016). Система анализа сложности англоязычных текстов Lexile предлагает средний логарифм нормализованной частотности всех слов текста как меру, демонстрирующую наивысшую корреляцию со сложностью, однако для расчетов используют не стандартную меру количества вхождений на миллион, ipm (instances per million), а на 5 миллионов слов (Lexile 2007). В работе, посвященной сравнению метрик частотности для текстов на английском языке с помощью построения предсказательных моделей, лучший результат показала модель, основанная на двух показателях: среднего логарифма с основанием 10 от нормализованной частотности слова на миллиард слов и стандартного отклонения (Chen & Meurers 2016). В работе на материале русского языка, посвященной диагностической валидности Стандартизированной методики исследования навыков чтения на русском языке, предлагается в качестве метрики средняя нормализованная частотность (ipm) только полнозначных слов текста (Dorofeeva et al. 2019). Часть исследований вообще использует не числовые значения, а систему деления текстов на группы по

частотности входящей в них лексики: «тексты с частотными словами/тексты с нечастотными словами» (Rello et al. 2013). Таким образом, представляется перспективным сравнение способов выражения данных о частотности на материале целых текстов и выбор оптимальной методики для данной задачи.

Метод бесконтактной регистрации движений глаз (айтрекинг) является одним из наиболее точных и трудозатратных способов проверки механизмов восприятия текста, который активно используется уже более 25 лет (Rayner 1998). Он позволяет отслеживать непроизвольное направление взгляда с высокой точностью (вплоть до доли градуса) и временным разрешением (вплоть до сотых долей секунды). Данная технология применима как для экспериментальной проверки гипотез о влиянии лингвистических и паралингвистических параметров текста на его сложность, так и для обнаружения различных затруднений при чтении. Чаще всего применительно к исследованиям чтения используются такие глазодвигательные параметры, как средняя продолжительность фиксаций, относительное время чтения слова и среднее количество фиксаций. Средняя продолжительность фиксаций (fixation duration) отражает время остановки взора на конкретном слове, и, чаще всего, характеризует скорость лексической активации и восприятия прочитанного (Henderson et al. 1989, Rayner 1998). Относительное время чтения слова (dwell time, %) является показателем процентного отношения суммы всех фиксаций на слове к другим слова в тексте (Griffin & Spieler 2006). Среднее количество фиксаций (fixation count) отражает количество фиксаций на слове, дополняя понимание о стратегии чтения (Clifton et al. 2007).

В современной науке накоплена большая база данных и установлены связи между определенными параметрами движений глаз и сложностью текста (Jian & Ko 2017). Сложность текста может быть представлена в различных лексических параметрах слов. Широко изучено влияние таких факторов, как длина слова (Rayner 2011), регулярность и согласованность слова (Farris-Trimble et al. 2018), орфографические и фонологические характеристики слова (Tiffin-Richards & Schroeder 2015), семантическое разнообразие значений слова (Luke et al. 2015) и прочее.

В контексте нашего исследования особый интерес представляют работы, оценивающие изменения в глазодвигательном поведении в зависимости от частотности слова (Rello et al. 2013, White et al. 2018). Так, в ряде исследований реципиенты демонстрировали значительно большую продолжительность взгляда на низкочастотных словах, чем на высокочастотных (Rau et al. 2014). Есть также свидетельства о большем влиянии фактора длины слова на параметры движения глаз у детей при чтении слов с низкой частотностью в сравнении с чтением слов с высокой частотностью (Rau et al. 2015). В работе на русскоязычном материале в записью движений глаз, техника чтения оценивалась по методикам «Чтение регулярных и нерегулярных слов» Т.В. Ахутиной и «Стандартизованная методика исследования навыка чтения» (СМИНЧ) А.Н. Корнева и О.А. Ишимовой (Корнеев и др. 2019). В описании обеих групп

этих методических материалов частотность указывается в качестве одной из характеристик, однако не поясняется источник данных о частотности слова и методика подсчета этой метрики для целого текста.

3.Материалы и методы

Участники эксперимента. Для установления зависимости частотности слов текста на русском языке с его сложностью был проведен эксперимент, в котором приняли участие 53 ученика 1–3 классов средних школ города Москвы: 26 учеников 1 класса (10 мальчиков, 16 девочек), 15 учеников 2 класса (4 мальчика, 11 девочек), 12 учеников 3 класса (2 мальчика, 10 девочек). Исследования проводились в апреле и мае, в конце учебного года, когда предполагается освоение навыков чтения, соответствующих классу обучения.

В качестве ***материала для эксперимента*** были использованы 6 текстов из современных учебников русского языка для 2–3 класса (табл. 1). В некоторых случаях текст незначительно модифицировался: слова заменялись на синонимы для получения более контрастирующих значений длины и частотности.

Методика подсчета данных о частотности лексики русского языка. Для определения оптимального источника информации о частотности слова, релевантного задаче оценки сложности текста для младшей школы, мы использовали два источника информации: Частотный словарь современного русского языка (по материалам Национального корпуса русского языка)[2] (далее – ЧС НКРЯ) и корпус детской литературы[3] (далее – ДетКорпус).

Частотный словарь основан на выборке текстов Национального корпуса русского языка объемом 100 млн словоупотреблений и включает в себя 20 тысяч наиболее употребительных слов современного русского языка (2-я половина XX – начало XXI вв.). Для получения частотных данных использовалась мера нормализованной частоты (ipm) общего частотного списка лемм. В данном словаре снята омонимия, поэтому частоты для разных значений омонима приводятся отдельно.

ДетКорпус – это аннотированный корпус русской детской литературы, включающий более 2097 прозаических произведений, написанных на русском языке в период с 1920-х по 2010-е гг. и адресованных детям и подросткам. Корпус содержит как художественные тексты различных жанров (реализм, приключения, детектив, ужастик), так и текст нон-фикшн. В данной коллекции текстов омонимия не снята. Поэтому в дальнейших пословных подсчетах мы не учитывали многозначные слова. Примеры анализируемых слов в табл. 2 дают представление о возможных различиях в частотных данных в зависимости от выбранного корпуса. Так, частотность тематически и

[2] *Ляшевская О.Н., Шаров С.А.* Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). М.: Азбуковник, 2009.

[3] ДетКорпус. http://detcorpus.ru (дата обращения: 20.12.2021).

стилистически нейтральной леммы *себя* показывает весьма близкие между собой значения по указанным коллекциям текстов, лемма *мальчик* значительно чаще встречается в корпусе детской литературы, а лемма *современный,* более характерная для документальной прозы и публицистики, намного частотнее в ЧС НКРЯ. Частотные списки 5 000 самых частотных слов по этим двум источникам, которыми мы будем оперировать далее, пересекаются на 81%.

Для учета частотных данных для целых текстов были рассчитаны следующие метрики:

1. Процент покрытия текста списком 5000 самых частотных слов;
2. Процент слов в тексте со значением ipm ниже 5;
3. Среднее от логарифма нормализованной частотности всех слов текста;
4. Средняя относительная частотность (ipm) полнозначных слов текста.

Анализ результатов показал, что метрика № 4 на текстах небольшого объема показывает слишком сильный разброс: например, значение для текста «В траве» равно 42, а для текста «Собака» – 1545. Вероятнее всего, это связано с тем, что в текстах малого объема появление хотя бы нескольких слов из 100 самых частотных для русского языка (*она, человек, такой* и т.п.*)* и отсутствие логарифмирования приводят к увеличению среднего значения частотности в разы. Поскольку такие данные трудно интерпретировать в дальнейшем, мы отказались от этой метрики.

В качестве ***традиционной метрики сложности текста***, основанной на средних значениях длины и предложения, был использован индекс читабельности Флеша со скорректированными для русского языка коэффициентами (Оборнева 2006).

$$\text{FRE(Oborneva)} = 206{,}835 - 1{,}52 \times \text{ASL} - 65{,}14 \times \text{ASW},$$

где FRE(Oborneva) – оценка читабельности текста; AWL – среднее число слогов в слове; ASL – среднее количество слов в предложении. Результатом этой формулы является число от 0 до 100, где 100 – очень легкий текст, 65 – легкий текст, 30 – немного трудно читать, а 0 – очень сложный для чтения текст.

В табл. 1 для каждого отобранного текста представлены данные о лингвистических параметрах, влияние которых на сложность исследовалось в ходе эксперимента. Первые две строки табл. 1 содержат параметры, по которым контрастируют отобранные текста. Тексты № 1 и № 2 уравновешены по проценту частотных слов, но отличаются по индексу FRE(Oborneva), тексты № 3 и № 4, наоборот, контрастируют по частотным данным, текст № 5 является примером текста, определенного как сложный и по индексу FRE(Oborneva) и по количеству частотных слов, текст № 6, напротив, пример простого текста и по индексу FRE(Oborneva), и по количеству частотных слов.

Таблица 1. **Основные лингвистические параметры используемых в эксперименте текстов**

Параметр текста	Текст 1. Трактор	Текст 2. Умка	Текст 3. В траве	Текст 4. Мышка	Текст 5. Цветы	Текст 6. Собака
FRE (Oborneva)	49	78	75	66	15	80
Покрытие текста частотным списком 5 000 (ЧС НКРЯ)	84%	85%	35%	89%	62%	92%
Средняя длина слова	6.4	4.8	5.6	5.6	6.8	4.1
Покрытие текста частотным списком 5 000 (ДетКорпус)	81%	83%	61%	93%	69%	96%
Процент слов с ipm < 5 (ЧС НКРЯ)	8%	14%	43%	4%	21%	2%
Процент слов с ipm < 5 (ДетКорпус)	11%	12%	22%	4%	24%	0%
Среднее от логарифма ipm всех слов текста (ЧС НКРЯ)	4.5	4.2	3.6	4.7	3.8	4.9
Среднее от логарифма ipm всех слов текста (ДетКорпус)	4.2	4.6	3.8	4.8	4	4.9

Table 1. **Main linguistic parameters of the texts used in the experiment (FD RNC is a frequency dictionary based on Russian National Corpus, DetCorpus is a corpus of literature addressed to children)**

Параметр текста	Text 1. Tractor	Text 2. Umka	Text 3. In the grass	Text 4. Mouse	Text 5. Flowers	Text 6. Dog
FRE (Oborneva)	49	78	75	66	15	80
Text coverage by the list 5000 (FD RNC)	84%	85%	35%	89%	62%	92%
Average word length	6.4	4.8	5.6	5.6	6.8	4.1
Text coverage by the list 5000 (DetCorpus)	81%	83%	61%	93%	69%	96%
Percent of words with ipm < 5 (FD RNC)	8%	14%	43%	4%	21%	2%
Percent of words with ipm < 5 (DetCorpus)	11%	12%	22%	4%	24%	0%
Average log word frequency (FD RNC)	4.5	4.2	3.6	4.7	3.8	4.9
Average log word frequency (DetCorpus)	4.2	4.6	3.8	4.8	4	4.9

Пословный анализ. Для более детального анализа и выбора оптимального источника данных о частотности и методике подсчета, были также произведены подсчеты корреляции лингвистических параметров конкретных слов с данными глазодвигательной активности. Каждая словоформа длиной более 3 символов была размечена по длине в символах и в слогах, частотности конкретной словоформы и частотности леммы (табл. 2). В анализе не участвовали многозначные слова. Поскольку слово в тексте может встретиться в непривычной, редкой форме, в ходе эксперимента отдельно проверялось влияние частотности конкретной словоформы на параметры глазодвигательной активности респондентов. Так, частотность всех форм существительного

волна составляет 31 662 вхождения, тогда как конкретная словоформа, представленная в тексте эксперимента, *волною*, встретилась в корпусе лишь 279 раз, поэтому все подсчеты были выполнены отдельно по совокупности частот всех словоформ этой лексемы (леммы *волна*) и отдельно по словоформе (*волною).*

Таблица 2. **Пословные параметры длины, частотности и характеристик движений глаз**

Словоформа	мальчики	гладиолусов	себе	современный
Лемма	мальчик	гладиолус	себя	современный
Длина словоформы в знаках	8	11	4	11
Длина словоформы в слогах	2	4	2	4
Частотность леммы в ЧС НКРЯ, ipm	188	0	2272	236
Частотность леммы в ДетКорпусе, ipm	597	1.1	2243	14
Частотность словоформы в ЧС НКРЯ, ipm	19	0	90	33
Частотность словоформы в ДетКорпусе, ipm	91	0.4	86	4
Относительное время чтения слова	0.026	0.089	0.019	0.032
Средняя длительность фиксаций, мс	257	288	255	250
Среднее количество фиксаций	3.22	9.15	2.46	4.43

Table 2. **Word-by-word values of word length, frequency and eye movement parameters (FD RNC is a frequency dictionary based on Russian National Corpus, DetCorpus is a corpus of literature addressed to children)**

Word form	мальчики	гладиолусов	себе	современный
Lemma	мальчик	гладиолус	себя	современный
Length of word form in characters	8	11	4	11
Length of word form in syllables	2	4	2	4
Lemma frequency by FD RNC, ipm	188	0	2272	236
Lemma frequency by DetCorpus, ipm	597	1.1	2243	14
Word form frequency by FD RNC, ipm	19	0	90	33
Word form frequency by DetCorpus, ipm	91	0.4	86	4
Dwell time, %	0.026	0.089	0.019	0.032
Fixation duration, ms	257	288	255	250
Fixation count	3.22	9.15	2.46	4.43

В ***ходе эксперимента*** испытуемых просили вслух и с максимальной скоростью прочитать предъявляемые тексты и предупреждали, что после прочтения будут заданы вопросы на понимание. Параллельно велась аудиорегистрация чтения и ответов на вопросы. Перед чтением каждого текста участник отвечал на вопрос, знаком ли ему предложенный текст. Учеников случайным образом разделили на 2 равные группы. Для уменьшения утомления каждая из групп читала только 3 из 6 отобранных текстов в случайном порядке, а также первый «тренировочный» текст, данные которого впоследствии не учитывались. Рис. 1 иллюстрирует пример результата прочтения текста одним из

испытуемых: точками обозначены фиксации, линии демонстрируют траекторию перемещения взгляда при чтении.

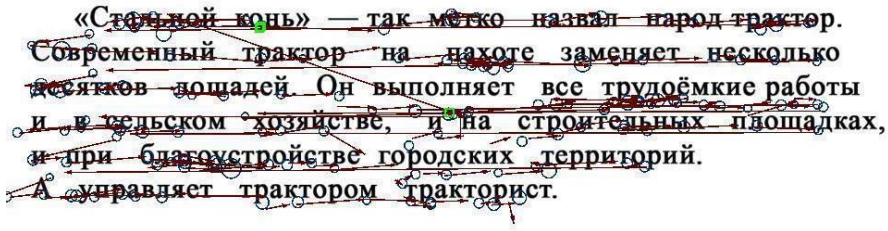

Рис. 1. Пример анализируемых данных глазодвигательной активности / Pic. 1. An example of the analyzed data of oculomotor activity

Оборудование. Исследование проводилось с применением айтрекера SR Research Eyelink 1000+, с частотой регистрации 500 Гц и 13-точечной калибровкой перед началом эксперимента. Испытуемые садились перед экраном компьютера диагональю 23 дюйма, с разрешением 1920 на 1080 точек (расстояние между глазами и экраном 940 мм), голова фиксировалась с помощью лобной опоры. В центре экрана предъявлялись отобранные тексты в той же верстке, которая использовалась в исходных учебниках, в виде изображения шириной 1400 пикселей и соответствующей тексту высотой. Это обеспечивало соответствие угловых размеров текста таковым при чтении учебника в привычном положении.

Методика подсчетов. В качестве меры сложности целого текста использовался параметр скорости чтения текста вслух в словах в минуту, усредненный по классам. При пословном анализе в качестве показателей сложности использовались значения средней относительной скорости чтения слова (dwell time %) – эта величина показывает, какую часть от времени прочтения всего текста конкретным испытуемым занимает чтение данного слова; средней длительности фиксаций (fixation duration); средним количеством фиксаций (fixation count).

4. Результаты

4.1. Анализ на уровне текстов

Для всех текстов ожидаемо наблюдалось увеличение скорости чтения от первого к третьему классу, хотя для разных текстов средние скорости были различны (Рис. 2). Следует отметить, что часть учеников 1 класса, участвовавших в исследовании, уже были знакомы с текстом «Собака», что дополнительно могло повысить скорость чтения этого фрагмента. Все остальные тексты были отмечены учениками как незнакомые.

Для оценки влияния класса обучения и конкретного текста на скорость чтения был проведен двухфакторный дисперсионный анализ. Результаты анализа показали, что оба этих фактора играют статистически значимую роль и

не зависят друг от друга, то есть «простой» или «сложный» текст оставался таковым независимо от класса обучения (см. рис. 1) (ANOVA, фактор «класс» ($F(2,135) = 28{,}55$, $p < 0{,}0001$) и «текст» ($F(5,135) = 8{,}40$, $p < 0{,}0001$).

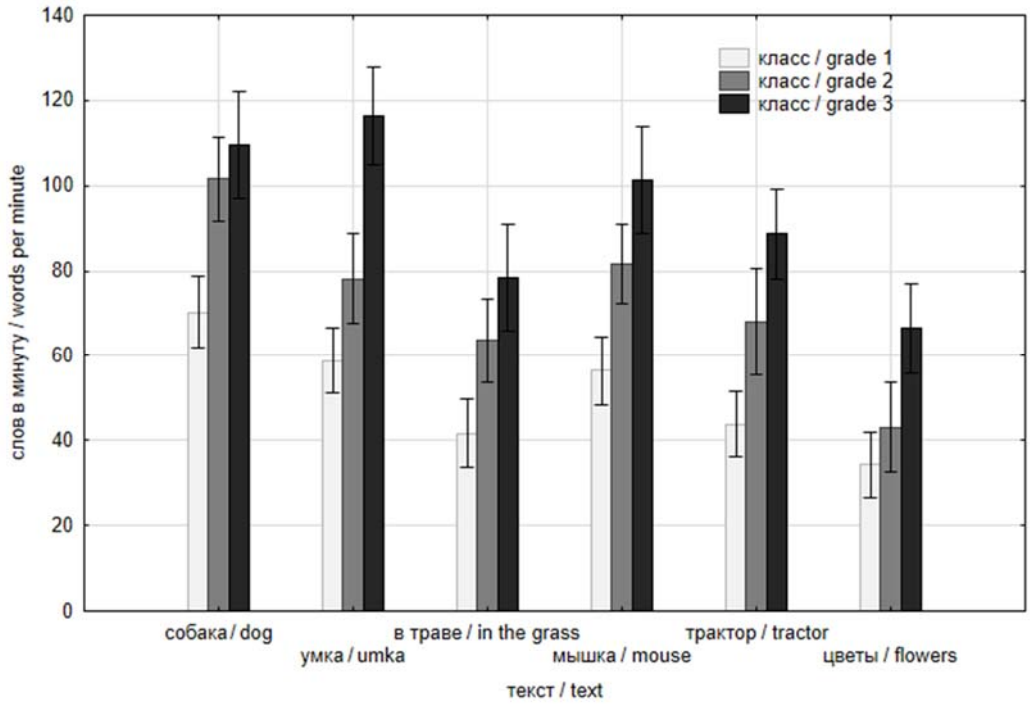

Рис. 2. Средние скорости чтения исследуемых текстов учениками 1–3 классов / Fig. 2. Average reading speed of the texts by students of grades 1–3

Анализ связи скорости чтения с характеристиками текстов иллюстрирует, что значения классической формулы Флеша не всегда способны адекватно предсказывать скорость чтения (рис. 2). Так, первая контрастирующая пара текстов с близкими показателями частотной лексики, но отличающаяся значениями формулы FRE (Oborneva), – «Умка» («простой» текст) и «Трактор» («сложный» текст), демонстрирует стабильную связь скорости чтения и показателей формулы: средняя скорость «простого» текста заметно выше по всем трем классам. Вторая пара текстов, имеющая близкие значения формулы FRE (Oborneva), соответствующие «простым» текстам, и контрастирующая по проценту покрытия текста частотным списком ЧС НКРЯ, иллюстрирует интересный случай расхождения показателей сложности по формуле и частотным данным. Так, на рис. 2 тексты расположены в порядке возрастания сложности по формуле FRE (Oborneva). Если допустить, что такой расчет точно измеряет сложность текста, мы ожидаем увидеть на графике плавное падение скорости чтения. Однако обратим внимание, что скорость чтения текста «В траве» (пример 1) заметно ниже ожидаемой.

(1) В траве трещат кузнечики, скрипит жук. Воркуют дикие голуби. Стучат по деревьям дятлы, пищат рябчики. Жужжит золотая пчёлка. Поют певчие дрозды, трещит сойка.

Этот текст действительно состоит из очень коротких слов и предложений, что и определило его значение формулы FRE (Oborneva), соответствующее самому простому тексту из всей коллекции. Однако при этом текст имеет наименьший процент покрытия списками частотной лексики. Можно предположить, что в данном случае речь идет о повышении сложности текста исключительно за счет трудности лексического состава текста. И оценка его сложности с помощью традиционных показателей (формуле FRE, основанной на длине слов и предложений) не соответствует его наблюдаемой сложности, выраженной средней относительной скоростью чтения.

Таким образом, были экспериментально зафиксированы две глобальные возможные причины повышения сложности текста, выраженные в снижении средней скорости чтения: структурная сложность, связанная с длиной слов и предложений («Трактор»), и лексическая сложность («В траве»). Быстрее всего испытуемые справлялись с текстами, отнесенными по обеим группам параметров к «простым»: «Собака», «Умка» и «Мышка». Эти тексты отличаются короткими словами и частотной лексикой. А наименьшая скорость чтения была показана на тексте «Цветы», содержащем и длинные слова, и крайне нечастотную лексику (*хризантема – 1.2 ipm; клубень – 3 ipm; гладиолус – 1.1 ipm; георгин – 1.4 ipm*)[4].

Результаты корреляционного анализа параметров частотности слов текста, полученных различными методиками, с параметром средней по всем классам скорости чтения исследуемых текстов (табл. 3), демонстрируют наивысшую корреляцию средней относительной скорости чтения с метриками текстов, представляющими процент покрытия текста частотными списками. Тип источника данных о частотности – ЧС НКРЯ или ДетКорпус – не всегда играет значимую роль в исследуемых текстах: при подсчетах среднего логарифма частотности слов текста и процента слов с ipm ниже 5 коэффициент корреляции оказывается выше у подсчетов по данным ДетКорпуса; процент же покрытия текстов частотными списками объемом 5000 слов по двум исследуемым источникам показывает одинаковый коэффициент детерминации. Интересно также, что индекс FRE (Oborneva) не показал статистически значимой корреляции со временем чтения в изучаемых текстах.

Малый объем текстов, на материале которых трудно делать общие выводы, стоит отнести к ограничениям эксперимента, связанным со спецификой процесса записи данных глазодвигательной активности. В айтрекинговых экспериментах текст выводится на экране, расположенном относительно далеко от глаз участника исследования. Для получения достаточно точных для анализа чтения результатов шрифт не может быть мелким, а текст не должен занимать значительную часть экрана, прокрутка также невозможна. Кроме

[4] Приведены расчеты частотности по леммам по данным ДетКорпуса.

того, возраст участников данного исследования также налагал ограничения: для детей 7–10 лет неподвижное сидение с фиксированным положением головы в сочетании с чтением на скорость утомительно. Поэтому весь эксперимент не мог длиться более 15 минут, включая калибровку, чтение тренировочного текста и ответы на вопросы.

Таблица 3. **Корреляционный анализ параметров глазодвигательной активности с параметрами частотности (Корреляция Спирмена, выделенные жирным значения с p-value < 0,05)**

Параметр	Средняя скорость чтения текста
Средняя длина слова	**-0.83**
FRE(Oborneva)	0.66
Покрытие текста частотным списком 5 000 (ЧС НКРЯ)	**0.89**
Покрытие текста частотным списком 5 000 (ДетКорпус)	**0.89**
Процент слов с ipm < 5 (ЧС НКРЯ)	-0.77
Процент слов с ipm < 5 (ДетКорпус)	**-0.83**
Среднее от логарифма ipm всех слов текста (ЧС НКРЯ)	**0.78**
Среднее от логарифма ipm всех слов текста (ДетКорпус)	**0.85**

Table 3. **Correlation analysis of oculomotor activity parameters with word frequency parameters (Spearman correlation, bold values have p-value <0.05)**

Parameter	Average reading speed
Average word length	**-0.83**
FRE(Oborneva)	0.66
Text coverage by the list 5000 (FD RNC)	**0.89**
Text coverage by the list 5000 (DetCorpus)	**0.89**
Percent of words with ipm < 5 (FD RNC)	-0.77
Percent of words with ipm < 5 (DetCorpus)	**-0.83**
Average log word frequency (FD RNC)	**0.78**
Average log word frequency (DetCorpus)	**0.85**

4.2. Результаты пословного анализа

Результаты корреляционного анализа (табл. 4) иллюстрируют степень связи параметров глазодвигательной активности с лингвистическими параметрами отдельных словоформ. Затемнённым фоном отмечены наиболее высокие значения корреляции в столбце. Максимально тесную связь с относительным временем чтения демонстрируют данные частотности леммы в ДетКорпусе, хотя все остальные варианты подсчета частотности слова по лемме и словоформе показывают очень близкие по значению результаты. Это говорит о том, что гипотеза о необходимости подсчета именно словоформ, а не лемм, на данных текстах не подтверждается. Возвращаясь к выбору оптимального источника данных о частотности, можно отметить, что, несмотря на то, что лучший результат показали данные по ДетКорпусу, разница не является существенной.

Таблица 4. **Корреляционный анализ параметров глазодвигательной активности с лингвистическими параметрами словоформ (Корреляция Спирмена, выделенные жирным значения имеют значение p-value < 0.05)**

Параметр	Среднее относительное время чтения	Средняя длительность фиксации	Среднее количество фиксаций
Длина словоформы в знаках	**0.53**	-0.02	**0.73**
Длина словоформы в слогах	**0.36**	-0.09	**0.55**
Частотность леммы в ЧС НКРЯ, ipm	**0.55**	**0.49**	**0.46**
Частотность леммы в ДетКорпусе, ipm	**0.59**	**0.42**	**0.54**
Частотность словоформы в ЧС НКРЯ, ipm	**0.58**	**0.47**	**0.53**
Частотность словоформы в ДетКорпусе, ipm	**0.58**	**0.42**	**0.53**

Table 4. **Correlation analysis of oculomotor activity parameters and linguistic parameters of word forms (Spearman correlation, bold values have a p-value <0.05)**

Parametr	Dwell time	Fixation duration	Fixation count
Length of word form in characters	**0.53**	-0.02	**0.73**
Length of word form in syllables	**0.36**	-0.09	**0.55**
Lemma frequency by FD RNC, ipm	**0.55**	**0.49**	**0.46**
Lemma frequency by DetCorpus, ipm	**0.59**	**0.42**	**0.54**
Word form frequency by FD RNC, ipm	**0.58**	**0.47**	**0.53**
Word form frequency by DetCorpus, ipm	**0.58**	**0.42**	**0.53**

Показательно, что длина слова в знаках сильнее всего коррелирует со средним количеством фиксаций, но не демонстрирует значимой связи со средней длительностью фиксаций, в отличие от частотных данных. Иными словами, от длины слова зависит, на какое количество «отрезков» глаз делит слово при чтении, тогда как то, сколько времени он задерживается на каждом таком отрезке, зависит именно от частотности слова. Можно предположить, что длительность фиксаций свидетельствует о когнитивных усилиях, требуемых для распознавания и обработки данного слова. Например, в выборке слов одинаковой длины в 5 знаков самые высокие значения длительности фиксаций занимают низкочастотные слова *сойка* (2.4 ipm / 307 мс.), *юркий* (4 ipm / 302 мс.), а самые низкие – частотные *разве* (240 ipm / 230 мс.) и *белые* (425 ipm / 234 мс.)[5].

Количество слогов в данном эксперименте показывает статистически значимую корреляцию со временем чтения и количеством фиксаций, но заметно меньшую, чем длина слов в знаках.

Также важно отметить, что на материале пословного анализа в совокупности отобранных для исследования текстов взаимная корреляция между параметрами частотности словоформы по НКРЯ и ее длины в знаках крайне слабая (корреляция Спирмена, $r = –0,21$, $p < 0,05$).

[5] Приведены расчеты частотности по леммам по данным ДетКорпуса.

4.3. Практическое применение результатов эксперимента

Полученные в ходе эксперимента данные, подтверждающие предиктивный потенциал информации о частотности слова, были использованы при создании пилотной версии системы автоматизированной оценки уровня сложности текста. Разработанный сервис Текстометр[6] при переключении в режим «русский язык как родной» предлагает оценку сложности текста по двум векторам: структурному и лексическому. Структурная сложность текста основывается на традиционном индексе FRE(Oborneva), приведенной для удобства интерпретации[7] к шкале возрастающей сложности от 0 до 10 по формуле:

$$\text{TEXTOMETR\_Struc} = \frac{100 - FRE(Oborneva)}{10}$$

Коэффициент лексической сложности текста основывается на результатах проведенного эксперимента и подсчитывается с помощью формулы:

$$\text{TEXTOMETR\_Lex} = 10 - \frac{(Freq - 50)}{5}$$

где Freq – процент покрытия всех слов текста списком 5 000 наиболее частотных лемм ДетКорпуса, показавший наивысшую корреляцию с относительным временем чтения текста, а остальные параметры являются константами. Результатом вычислений становится коэффициент лексической сложности текста, число от 0 до 10[8], означающее степень вероятной знакомости слов текста читателями детской аудитории. Предположительный возраст, для которого данный текст является оптимальным по сложности, рассчитывается на основании усредненной оценки двух описанных параметров. В качестве иллюстрации приведем значения параметров структурной и лексической сложности изученных текстов.

Как видно из табл. 5, почти все тексты отмечены программой как соответствующие возрасту учеников российской начальной школы. Исключение составляет лишь текст «Цветы» из-за входящих в его состав длинных и крайне нечастотных слов, которые привели к высоким показателям структурной и лексической сложности текста. Дополнительной причиной таких высоких показателей сложности может являться и небольшой объем текста, при котором появление даже нескольких длинных или низкочастотных слов может значительно повлиять на финальные оценки теста. Стоит упомянуть, однако, что данный текст действительно показал наименьшие значения усредненной скорости чтения по всем трем классам, что также свидетельствует о возникших трудностях при его чтении. Описанный выше нетипичный текст

6 Текстометр. https://textometr.ru (обращение 27.12.2021)

7 Оригинальный вариант коэффициента представляет собой число от от 0 до 100, причем коэффициент измеряет уровень простоты текста, т.е. меньшие значения означают большую сложность текста, что приводит к неудобству интерпретации.

8 Формально максимальным значением коэффициента является 20, однако подавляющее большинство русскоязычных текстов, включая новостные, научные, художественные тексты и определения сложных терминов, укладывается в шкалу от 0 до 10.

«В траве» оценивается по шкале структурной сложности в 2 балла из 10, тогда как по шкале лексической сложности – в 7 баллов из 10. Общая усредненная оценка остается при этом в границах возрастной группы младшей школы, однако такой способ оценки позволяет получить более полную и интерпретируемую информацию о сложности текста.

Таблица 5. **Пример работы сервиса Текстометр (русский язык как родной) на материале текстов из эксперимента**

Текст	Структурная сложность	Лексическая сложность	Предположительный возраст
Текст 1. Трактор	4	3	9–10 лет
Текст 2. Умка	3	3	9–10 лет
Текст 3. В траве	2	7	9–10 лет
Текст 4. Мышка	3	1	7–8 лет
Текст 5. Цветы	9	6	13–15 лет
Текст 6. Собака	2	1	7–8 лет

Table 5. **An example of the output of the Textometr tool (Russian as a native language section) for the texts from the experiment**

Text	Structural complexity	Lexical complexity	Estimated age
Text 1. Tractor	4	3	9–10 years
Text 2. Umka	3	3	9–10 years
Text 3. In the grass	2	7	9–10 years
Text 4. Mouse	3	1	7–8 years
Text 5. Flowers	9	6	13–15 years
Text 6. Dog	2	1	7–8 years

К ограничениям эксперимента, связанным со спецификой процесса записи данных глазодвигательной активности, стоит отнести малый объем текстов. В айтрекинговых экспериментах текст выводится на экране, расположенном относительно далеко от глаз участника исследования. Для получения достаточно точных для анализа чтения результатов шрифт не может быть мелким, а текст не должен занимать значительную часть экрана, прокрутка также невозможна. Кроме того, возраст участников данного исследования также налагал ограничения: для детей 7–10 лет неподвижное сидение с фиксированным положением головы в сочетании с чтением на скорость утомительно. Поэтому весь эксперимент не мог длиться более 15 минут, включая калибровку, чтение тренировочного текста и ответы на вопросы.

5. Выводы

Приведенные результаты айтрекингового эксперимента, осуществленного на материале текстов учебников русского языка для младшей школы, подтверждают связь сложности текста и частотности слов, в него входящих, и демонстрируют потенциал учета частотных данных в системах оценки сложности текста на русском языке.

На уровне текстов самую высокую корреляцию со средним временем чтения фрагмента показал параметр процента покрытия текста списком 5000

самых частотных слов, при этом данные по разным источникам – ЧС НКРЯ и ДетКорпусу – показали одинаковые значения. Также в результате эксперимента удалось выявить текст, где классическая формула Флеша, базирующаяся на средней длине слова и предложения, дает ошибку прогноза, тогда как данные о частотности слов текста верно диагностируют вероятную лексическую сложность текста. Целесообразно включение такого текста в коллекцию материалов для валидизации качества систем, оценивающих сложность текста на русском языке.

На пословном уровне самую тесную связь с относительным временем чтения продемонстрировали данные частотности леммы по корпусу детской литературы. Также анализ показал, что хотя длина слова в знаках сильнее всего коррелирует со средним количеством фиксаций, она не демонстрирует значимой связи со средней длительностью фиксаций. Напротив, частотность слова показала самую высокую корреляцию с этим параметром. Такие данные позволяют сделать предположение, что длина слова больше влияет на механическую часть процесса чтения, а именно на какое количество «отрезков» глаз делит слово для удобства прочтения, тогда как частотность слова оказывает влияние на когнитивный аспект чтения – на то, сколько времени потребуется на каждом таком отрезке для распознавания облика слова и его восприятия. При этом более трудозатратная методика подсчета частотности словоформ не дала значимого прироста качества в исследуемом материале, что позволяет сделать вывод о достаточности данных о частотности леммы в задаче оценки сложности лексики текста.

Среди направлений дальнейшей работы отметим, во-первых, проверку найденных закономерностей на большем количестве текстов, во-вторых, экспериментальную проверку влияния других формальных лингвистических показателей текста на русском языке на его сложность.

Благодарности и финансирование

Работа выполнена с использованием средств государственного бюджета по госзаданию на 2020–2024 годы (проект FZNM-2020-0005).

REFERENCES / СПИСОК ЛИТЕРАТУРЫ

Иомдин Б.Л., Морозов Д.А. Кто поймет «Незнайку»? Автоматическое определение сложности текстов для детей // *Русская речь*. 2021. № 5. С. 55–68. [Iomdin, Boris L. & Dmitry A. Morozov. 2021. Who can understand “Dunno”? Automatic assessment of text complexity in children’s literature. *Russian Speech* 5. 55–68 (In Russ.)]. https://doi.org/10.31857/S013161170017239-1

Корнеев А.А., Ахутина Т.В., Матвеева Е.Ю. Особенности чтения третьеклассников с разным уровнем развития навыка: анализ движений глаз // *Вестник Московского университета. Серия 14. Психология*. 2019. № 2. С. 64–87. [Korneev, Aleksei A., Tatiana V. Akhutina & Ekaterina Yu. Matveeva. 2019. Reading in third graders with different state of the skill: An eye-tracking study. *Vestnik Moskovskogo Universiteta. Seriya 14. Psikhologiya* 2. 64–87. (In Russ.)]. https://doi.org/10.11621/vsp.2019.02.64

Криони Н.К., Никин А.Д., Филиппова А.В. Автоматизированная система анализа сложности учебных текстов // *Вестник Уфимского государственного авиационного технического университета.* 2008. № 11 (1). С. 101–107. [Krioni, Nikolai K., Aleksei D. Nikin & Anastasia V. Filippova. 2008. Automated system for analyzing the complexity of educational texts. *Bulletin of the Ufa State Aviation Technical University* 11(1). 101–107. (In Russ.)].

Лапошина А.Н., Веселовская Т.С., Лебедева М.Ю., Купрещенко О.Ф. Лексический состав текстов учебников русского языка для младшей школы: корпусное исследование // *Компьютерная лингвистика и интеллектуальные технологии: по материалам международной конференции «Диалог 2019».* 2019. Т. 18 (25). С. 351–363. [Laposhina, Antonina N., Tatiana S. Veselovskaya, Maria U. Lebedeva & Olga F. Kupreshchenko. 2019. Lexical analysis of the Russian language textbooks for primary school: Corpus study. *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2019"*18. 351–363. (In Russ.)].

Мартынова Е.В., Солнышкина М.И., Мерзлякова А.Ф., Гизатулина Д.Ю. Лексические параметры учебного текста (на материале текстов учебного корпуса русского языка) // *Филология и культура.* 2020. № 3 (61). С. 72–80. [Martynova, Ekaterina V., Marina I. Solnyshkina, Amina F. Merzlyakova & Diana Yu. Gizatulina. 2020. Lexical parameters of the academic text (based on the texts of the academic corpus of the Russian language). *Philology and Culture* 3. 72–80. (In Russ.)]. https://doi.org/10.26907/2074-0239-2020-61-3-72-80

Мизернов И.Ю., Гращенко Л.А. Анализ методов оценки сложности текста. // *Новые информационные технологии в автоматизированных системах.* 2015. № 18. С. 572–581. [Mizernov, I. Yu. & L. A. Grashchenko. 2015. Analysis of methods for assessing text complexity. *New Information Technologies in Automated Systems* 18. 572–581. (In Russ.)].

Микк Я.А. О факторах понятности учебного текста: автореф. дис. … канд. пед. наук. Тарту, 1970. 22 с. [Mikk, Ya.A. 1970. Factors of educational text clarity. *Abstract of Pedagogy Cand. Diss. Tartu.* (In Russ.)].

Оборнева И.В. Автоматизированная оценка сложности учебных текстов на основе статистических параметров: дис... канд. пед. наук: 13.00.02. М., 2006. 165 с. [Oborneva, Irina V. 2006. Automated estimation of complexity of educational texts on the basis of statistical parameters. *Pedagogy Cand. Diss*. Moscow. (In Russ.)].

Солнышкина М.И., Кисельников А.С. Сложность текста: этапы изучения в отечественном прикладном языкознании. // *Вестник Томского государственного университета. Филология.* 2015. № 6 (38). С. 86–99. [Solnyshkina, Marina I. & Alexander S. Kiselnikov. 2015. Text complexity: Study phases in Russian linguistics. *Tomsk State University Journal of Philology* 6. 86–99. (In Russ.)]. https://doi.org/10.17223/19986645/38/7

Шпаковский Ю.Ф. Разработка количественной методики оценки трудности восприятия учебных текстов для высшей школы // *Научно-технический вестник информационных технологий, механики и оптики*. 2008. № 1 (83). С. 110–117. [Shpakovsky, Yury F. 2008. Development of a quantitative methodology for assessing the difficulty of perceiving educational texts for higher education. *Scientific and Technical Bulletin of Information Technologies, Mechanics and Optics* 1(83). 110–117. (In Russ.)].

Chall, Jeanne S. & Edgar Dale. 1995. *Readability Revisited: The New Dale-Chall Readability Formula.* Cambridge, MA: Brookline Books.

Chen, Xiaobin & Detmar Meurers. 2016. Characterizing text difficulty with word frequencies. In Joel Tetreault, Jill Burstein, Claudia Leacock & Helen Yannakoudakis (eds.),

Proceedings of the 11th workshop on innovative use of nlp for building educational applications, 84–94. San Diego: Association for Computational Linguistics.

Clifton, Jr. Charles, Adrian Staub & Keith Rayner. 2007. Eye movements in reading words and sentences. In Roger P. G. van Gompel, Martin H. Fischer, Wayne S. Murray & Robin L. Hill (eds.), *Eye movements: A window on mind and brain,* 341–371. Elsevier. https://doi.org/10.1016/B978-008044980-7/50017-3

Dorofeeva, Svetlana V., Victoria Reshetnikova, Margarita Serebryakova, Daria Goranskaya, Tatiana V. Akhutina & Olga Dragoy. 2019. Assessing the validity of the standardized assessment of reading skills in Russian and verifying the relevance of available normative data. *The Russian Journal of Cognitive Science* 6(1). 4–24.

DuBay, William H. 2007. *Smart Language: Readers, Readability, and the Grading of Text.* Costa Mesa, California: Impact Information.

Farris-Trimble, Ashley & Bob McMurray. 2018. Morpho-phonological regularities influence the dynamics of real-time word recognition: Evidence from artificial language learning. *Laboratory Phonology* 9(1). 1–34. https://doi.org/10.5334/labphon.41

Francois, Tomas & Cedrick Fairon. 2012. An 'AI readability' formula for French as a foreign language. *Proceedings of the EMNLP and CoNLL 2012, Jeju Island, Korea, 12–14 July 2012*. 466–477.

Glazkova, Anna, Yury Egorov & Maxim Glazkov. 2021. A comparative study of feature types for age-based text classification. In *Analysis of Images, Social Networks and Texts. AIST 2020. Lecture Notes in Computer Science* 12602. 120–134.

Graesser, Arthur C., Danielle S. McNamara, Zhiqang Cai, Mark Conley, Haiying Li & James Pennebaker. 2014. Coh-Metrix measures text characteristics at multiple levels of language and discourse. *The Elementary School Journal* 115. 210–229.

Griffin, Zenzi M. & Daniel H. Spieler. 2006. Observing the what and when of language production for different age groups by monitoring speakers' eye movements. *Brain and Language* 99(3). 272–288.

Henderson, John M., Aleksander Pollatsek & Keith Rayner. 1989. Covert visual attention and extrafoveal information use during object identification. *Perception & Psychophysics* 45. 196–208. https://doi.org/10.3758/BF03210697

Jian, Yu-Cin & Hwawei Ko. 2017. Influences of text difficulty and reading ability on learning illustrated science texts for children: An eye movement study. *Computers & Education* 113. 263–279.

Lexile. 2007. *The Lexile Framework for Reading: Theoretical Framework and Development. Technical Report.* MetaMetrics, Inc., Durham, NC

Luke, Steven G., John M. Henderson & Fernanda Ferreira. 2015. Children's eye-movements during reading reflect the quality of lexical representations: An individual differences approach. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 41(6). 1675–1683. https://doi.org/10.1037/xlm0000133

Raney, Gary E. & Keith Rayner. 1995. Word frequency effects and eye movements during two readings of a text. *Canadian Journal of Experimental Psychology* 49. 151–172.

Rau, Anne K., Kristina Moll & Karin Landerl. The transition from sublexical to lexical processing in a consistent orthography: An eye-tracking study. *Scientific Studies of Reading* 18. 224–233. https://doi.org/10.1080/10888438.2013.857673

Rau, Anne K., Kristina Moll, Margaret J. Snowling & Karin Landerl. 2015. Effects of orthographic consistency on eye movement behavior: German and English children and adults process the same words differently. *Journal of Experimental Child Psychology* 130. 92–105. https://doi.org/10.1016/j.jecp.2014.09.012.

Rayner, Keith. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin* 124. 372–422. https://doi.org/10.1037/0033-2909.124.3.372

Rayner, Keith, Timothy J. Slattery, Denis Drieghe & Simon P. Liversedge. 2011. Eye movements and word skipping during reading: Effects of word length and predictability. *Journal of Experimental Psychology: Human Perception and Performance* 37(2). 514–528.

Rello, Luz, Ricardo Baeza-Yates, Laura Dempere-Marco & Horacio Saggion. 2013. Frequent words improve readability and short words improve understandability for people with dyslexia. In Paula Kotzé & Gary Marsden (eds.), *Human-Computer interaction – INTERACT 2013. Lecture notes in computer science vol 8120,* 203–219. Berlin/Heidelberg: Springer. https://doi.org/10.1007/978-3-642-40498-6_15

Reynolds, Robert. 2016. Insights from Russian second language readability classification: Complexity-dependent training requirements, and feature evaluation of multiple categories. *Proceedings of the 11th Workshop on the Innovative Use of NLP for Building Educational Applications, San Diego, CA 2016*. 289–300.

Sato, Satoshi. 2014. Text Readability and Word Distribution in Japanese. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14) 2014.* 2811–2815.

Schwarm, Sarah E. & Mari Ostendorf. 2005. Reading level assessment using support vector machines and statistical language models. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL '05), USA, 2005.* 523–530.

Solovyev, Valery, Vladimir Ivanov & Marina Solnyshkina. 2018. Assessment of reading difficulty levels in Russian academic texts: Approaches and metrics. *Journal of Intelligent & Fuzzy Systems* 34. 3049–3058.

Tiffin-Richards, Simon P. & Sasha Schroeder. 2015. Children's and adults' parafoveal processes in German: Phonological and orthographic effects. *Journal of Cognitive Psychology* 27. 531–548. https://doi.org/10.1080/20445911.2014.999076

White, Sarah J., Denis Drieghe, Simon P Liversedge & Adrian Staub. 2018. The word frequency effect during sentence reading: A linear or nonlinear effect of log frequency? *Quarterly Journal of Experimental Psychology* 71(1). 46–55. https://doi.org/10.1080/17470218.2016.1240813

Словари/Dictionaries

Ляшевская О.Н., Шаров С.А. Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). М.: Азбуковник. 2009. [Lyashevskaya, Olga N. & Sergey A. Sharoff. 2009. Modern Russian Frequency Dictionary (based on the data from the Russian National Corpus). Moscow: Azbukovnik. (In Russ.)]

Article history:
Received: 19 November 2021
Accepted: 21 January 2022

Bionotes:

Antonina N. LAPOSHINA is a leading expert of the Laboratory of Cognitive and Linguistic Studies at Pushkin State Russian Language Institute, member of the research group “Teaching Russian in the Digital Age”. Her research interests include computer-assisted language learning and corpus-based language learning.

Contact information:
6 Akademika Volgina street, Moscow, 117485, Russia
e-mail: ANLaposhina@pushkin.institute
ORCID: 0000-0003-0693-7657

Maria Yu. LEBEDEVA holds a PhD in Philology and is a leading researcher of the Laboratory of Cognitive and Linguistic Studies, Associate Professor of the Department of Methods of Teaching Russian as a Foreign Language at Pushkin State Russian Language Institute. Her research interests are focused on corpus-based language learning, methods of online teaching Russian and reading strategies in the digital age.

Contact information:

6 Akademika Volgina street, Moscow, 117485, Russia
e-mail: MULebedeva@pushkin.institute
ORCID: 0000-0002-9893-9846

Alexandra A. BERLIN KHENIS is a specialist of the Laboratory of Cognitive and Linguistic Studies at Pushkin State Russian Language Institute. Her research interests include cognitive psychology, as well as psychophysiological aspects of reading and learning.

Contact information:

6 Akademika Volgina street, Moscow, 117485, Russia
e-mail: alexa.munxen@gmail.com
ORCID: 0000-0003-2034-1526

Сведения об авторах:

Антонина Николаевна ЛАПОШИНА – ведущий эксперт лаборатории когнитивных и лингвистических исследований Государственного института русского языка имени А.С. Пушкина. Сфера научных интересов: компьютерная лингвистика, цифровая лингводидактика.

Контактная информация:

Российская Федерация, 117485, Москва, ул. Академика Волгина, д. 6
e-mail: ANLaposhina@pushkin.institute
ORCID: 0000-0003-0693-7657

Мария Юрьевна ЛЕБЕДЕВА – кандидат филологических наук, ведущий научный сотрудник лаборатории когнитивных и лингвистических исследований, доцент кафедры методики преподавания РКИ Государственного института русского языка имени А.С. Пушкина. Сфера научных интересов: корпусная лингводидактика, особенности чтения в цифровую эпоху.

Контактная информация:

Российская Федерация, 117485, Москва, ул. Академика Волгина, д. 6
e-mail: MULebedeva@pushkin.institute
ORCID: 0000-0002-9893-9846

Александра Алехандровна БЕРЛИН ХЕНИС – специалист лаборатории когнитивных и лингвистических исследований Государственного института русского языка имени А.С. Пушкина. Сфера научных интересов: когнитивная психология, психофизиологические аспекты чтения и обучения.

Контактная информация:

Российская Федерация, 117485, Москва, ул. Академика Волгина, д. 6
e-mail: alexa.munxen@gmail.com
ORCID: 0000-0003-2034-1526