Russian Language Studies
Русистика
ISSN 2618-8163 (Print); ISSN 2618-8171 (Online)
2021 Vol. 19 No. 3 331–345
http://journals.rudn.ru/russian-language-studies

МЕДИАДИДАКТИКА И ЭЛЕКТРОННЫЕ СРЕДСТВА ОБУЧЕНИЯ
MEDIADIDACTICS AND ELECTRONIC MEANS OF INSTRUCTION

DOI 10.22363/2618-8163-2021-19-3-331-345

Научная статья

Текстометр: онлайн-инструмент определения уровня сложности текста по русскому языку как иностранному

А.Н. Лапошина✉, М.Ю. Лебедева

Государственный институт русского языка имени А.С. Пушкина,
Российская Федерация, 117485, Москва, ул. Академика Волгина, д. 6
✉ ANLaposhina@pushkin.institute

Аннотация. Оценка текста с точки зрения его языковой доступности представляется крайне актуальной и трудозатратной задачей в процессе его подготовки к занятию по русскому языку как иностранному. С другой стороны, процесс отнесения текста к одному из уровней по шкале CEFR (от A1 до C2) является достаточно формализованным и описанным в методической литературе, что открывает возможности по его автоматизации. Цель исследования – описать возможности и методику использования нового онлайн-инструмента «Текстометр» для автоматического анализа уровня сложности текста по шкале CEFR и его подготовки к уроку русского языка в иностранной аудитории. Материалом для построения математической модели по определению уровня текста послужили более чем 800 текстов из современных учебников по русскому языку как иностранному. В процессе разработки концепции и создания сервиса применялись методы теоретического анализа научно-методической литературы и регламентирующих документов в области русского языка как иностранного, анкетирования и тестирования учащихся и преподавателей, машинного обучения и автоматической обработки текстов на естественном языке. В результате установлены и описаны основные возможности сервиса: определение уровня текста по шкале CEFR, предоставление информации, полезной для адаптации текста к учебным задачам, такой как списки ключевых слов и слов – оптимальных кандидатов в словарь к данному тексту, статистика по покрытию текста лексическими минимумами ТРКИ и списками частотных слов русского языка, меры лексического разнообразия текста, прогноз времени, необходимого для разных видов чтения текста. Выявлены недостатки работы сервиса на данном этапе разработки и предложены пути их решения. Приведены результаты экспериментальной проверки качества работы инструмента и намечены векторы дальнейшего развития сервиса. Сервис может быть полезен преподавателям, методистам, а также авторам пособий и представителям издательств для проверки соответствия текстового материала заявленному уровню и учебным целям.

Ключевые слова: русский язык как иностранный, учебный текст, сложность текста, обучение чтению, адаптация текстов, компьютерная лингводидактика, компьютерные технологии, преподавание русского языка, интернет-ресурсы, обучение русскому языку

История статьи: поступила в редакцию 16.02.2021; принята к печати 18.05.2021.

Для цитирования: *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. http://dx.doi.org/10.22363/2618-8163-2021-19-3-331-345

Введение

Качество текстовых материалов, используемых на занятиях по иностранному языку, способно оказывать значительное влияние на результат обучения. Так, например, положительный эффект текстовых материалов, интересных учащимся, на качество понимания текста и уровень мотивации студентов не только согласуется с интуитивными представлениями преподавателей, но и подтверждается научными работами (Alexander, Jetton, 1996). Поскольку сферы интересов, цели и задачи студентов, изучающих русский язык как иностранный, могут быть самыми разными, перед преподавателем стоит непростая задача подбора и подготовки большого количества материалов, а также регулярное пополнение и обновление личной текстотеки. При этом в современном информационном пространстве вряд ли возникает проблема недостатка текстовых материалов: новостные сайты, интернет-издания, блоги, тексты социальных сетей являются источниками огромного количества аутентичных текстов. Проблема скорее состоит в отборе материалов, подходящих студентам как по тематике, так и по уровню языковой сложности. При этом языковая доступность материалов представляется чрезвычайно важным критерием: исследования показывают, что подходящие по уровню материалы для чтения способствуют развитию языковых навыков, тогда как слишком простые тексты могут вызвать скуку, а чересчур сложные – снизить мотивацию (Graesser et al., 2014; Микк, 1981).

Вопросы языковой доступности русского учебного текста тесно связаны с теорией создания и оценки учебника и поднимаются в работах И.Л. Бим (Бим, 1977), А.Р. Арутюнова (Арутюнов, 1990), Я.А. Микка (Микк, 1981), М.Н. Вятютнева (Вятютнев, 1984), Ю.А. Томиной (Томина, 1985).

С развитием технологий автоматической обработки естественного языка в мировой научной практике появляются работы, посвященные возможностям автоматизации процесса оценки доступности текста (DuBay, 2004). История разработки автоматизированной оценки сложности русских текстов для преподавания иностранной аудитории пока не столь богата, как, например, для англоязычных текстов, однако все же содержит несколько пионерских работ (Karpov et al., 2014; Reynolds, 2016; Sharoff et al., 2008), на которые опирается данное исследование.

Среди существующих сервисов по анализу русскоязычных текстов можно отметить ресурс «Лексикатор»[1], оценивающий текст с точки зрения его соответствия уровню знания русского языка как иностранного по лексическим и структурным параметрам по заданным методистами правилам, а также проект «Простой русский»[2], предлагающий статистику по пяти популярным формулам

[1] Лексикатор. URL : https://corplings.pythonanywhere.com (дата обращения: 15.02.2021).

[2] Проверка на читабельность текстов. URL : https://ru.readability.io (дата обращения: 15.02.2021).

читабельности, изначально разработанным для английского языка и адаптированным для русскоязычных текстов. Однако вопросы оценки уровня сложности текста с помощью модели машинного обучения и представления результатов ее работы широкому кругу пользователей, а также проверки качества работы модели в реальной практике преподавания РКИ, насколько нам известно, впервые ставятся в данной работе. Таким образом, **цель исследования** – описать возможности и методику использования онлайн-инструмента «Текстометр» для оценки сложности русского текста как иностранного.

Методы и материалы

Для разработки научной концепции сервиса был проведен анализ релевантной научной и методической литературы по вопросам автоматического определения сложности текста для задач лингводидактики, подготовки текста к занятию по РКИ, нормативных документов в области РКИ, описания характеристик текстов в системе уровней CEFR.

В процессе проверки качества работы модели и ее настройки были применены методы анкетирования и тестирования учащихся и преподавателей.

Для обучения математической модели по определению уровня сложности текста был собран корпус из 800 текстов из пособий по РКИ, информация об уровне сложности которых отображена в методической справке пособия. Каждый текст был автоматически размечен по более чем 100 лингвистическим признакам (средняя длина слова и предложения, количество различных частей речи и грамматических форм, покрытие текста списками из лексических минимумов, списками частотных слов русского языка, списками абстрактной лексики и пр.) (Laposhina et al., 2018). Автоматическая обработка текста и подсчет лингвистических характеристик проводилась с помощью программного кода на языке Python. Построение математической модели осуществлено с помощью библиотеки Scikit-learn[3].

Результаты

В 2020 году на базе предшествующих методических разработок (Laposhina et al., 2018; Лапошина, 2018) был создан интерфейс веб-сервиса «Текстометр»[4], позволяющий любому пользователю получить результаты работы анализатора.

Интерфейс представляет собой окно ввода, куда можно вставить любой текст на русском языке до 10 000 слов и получить значение уровня сложности для введенного текста по шкале CEFR, а также информацию о тексте, представляющую ценность для его подготовки к занятию по РКИ (рис. 1).

Для корректного автоматического анализа введенный текст проходит несколько этапов предобработки: очистку от всех символов и букв, отличных от русского алфавита (например, чисел, названий компаний на английском языке, элементов верстки); лемматизацию, то есть приведение каждого слова текста к начальной, словарной форме слова (для подсчетов лексической информации,

[3] Scikit-learn. Machine learning in Python. URL : https://scikit-learn.org (accessed: 15.02.2021).
[4] Текстометр. URL : https://textometr.ru/ (дата обращения: 15.02.2021).

например, количества уникальных слов); автоматический морфологический анализ (для подсчета грамматических форм, оказывающих влияние на сложность текста, например, форм пассива, причастий, цепочек родительного падежа и мн. др.); проверку полученных списков лексики и фильтрацию слов, отсутствующих в словаре (например, слов с опечатками, сокращений типа *вин. п.*).

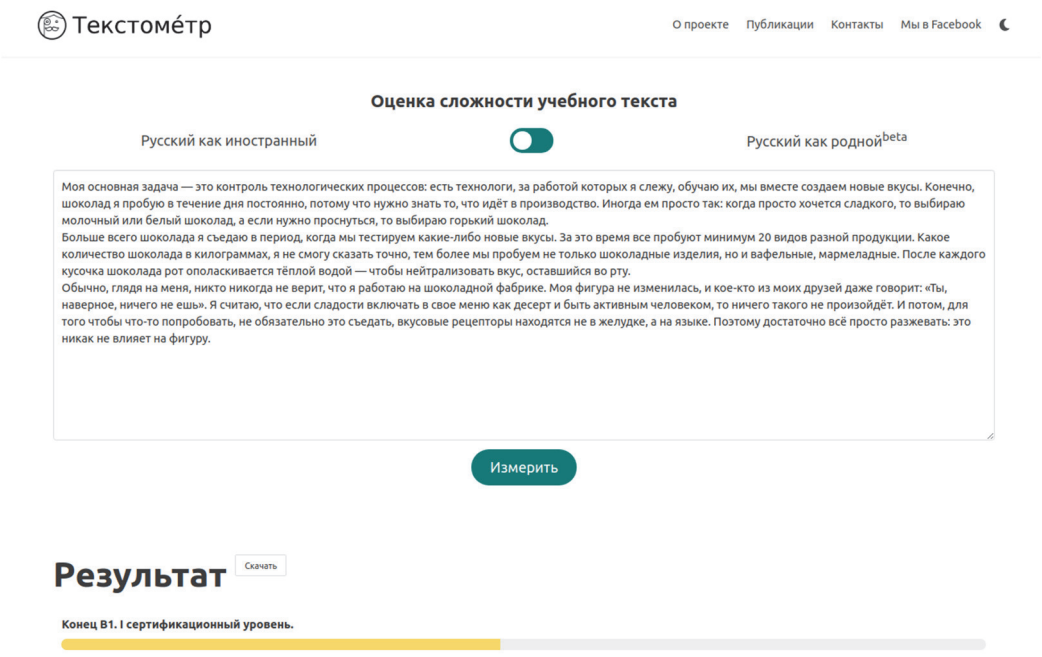

Рис. 1. Интерфейс сервиса «Текстометр»
Figure 1. Interface of Textometr

Во всех дальнейших подсчетах вхождения слов текста в лексические списки учитываются только слова, написанные буквами русского алфавита, присутствующие в словаре морфологического анализатора, приведенные к начальной форме.

Уровень сложности текста по шкале CEFR (от A1 до C2) определяется автоматически на основании результатов работы математической модели, обученной на коллекции из 800 текстов из пособий по РКИ, информация об уровне сложности которых нам уже известна, и более чем 100 лингвистических признаков текста. Таким образом, модель делает предположение об уровне сложности текста, опираясь не на заданные разработчиками правила, а на практический опыт большого количества авторов пособий и методистов, реальный текстовый материал, с которым сталкиваются студенты, изучая русский язык по данным пособиям.

В ходе экспериментов по определению сложности текста практикующими преподавателями мы заметили, что они зачастую пользуются более дробной шкалой уровней, вводя, например, такие обозначения как *«A2+»*, *«B1 продвинутый»* и т. п. (Лапошина, 2018). Поэтому было принято решение о представлении информации об уровне в более дробном формате: так, каждый уровень CEFR получает дополнительную маркировку *начало, середина* или *конец*. Эта информация особенно полезна при сравнении несколь-

ких текстов: например, в ситуации сравнения текстов *«конец B1»* и *«начало B2»* становится очевидна их близость по уровню сложности, а *«начало B1»* и *«конец B2»* – наоборот.

Обсуждение

Информация об уровне языковой сложности текста является важнейшей, но не единственной характеристикой текста, влияющей на выбор текста преподавателем. Например, важным критерием может стать информация о том, насколько хорошо данный текст подходит для целей контроля, какая лексика может быть изучена на его материале и насколько полезна эта лексика для данной аудитории и ситуации обучения. Поэтому, помимо уровня сложности текста, сервис «Текстометр» предлагает информацию о тексте, представляющую ценность для его подготовки к занятию по РКИ: списки ключевых слов и слов – наилучших кандидатов в словарик к данному тексту, статистику по покрытию текста лексическими минимумами ТРКИ, частотный словарь текста, прогноз времени, необходимого для разных видов чтения текста, а также грамматические темы, которые можно отработать на данном тексте.

Длины текста в словах и предложениях являются базовыми характеристиками текста, особенно полезными для расчета времени, которое потребуется на его освоение, или при подготовке проверочных материалов, где объем текста обычно строго определен государственным стандартом по РКИ. Например, рекомендуемая длина текста для чтения уровня A1 составляет 250–300 слов, A2 – 600–700 слов и т. д.

Средняя длина слова и предложения также может свидетельствовать о сложности текста или его отдельных фрагментов. Так, большое количество формул читабельности используют данные показатели в качестве основных (DuBay, 2004). На рис. 2 представлена иллюстрация этой достаточно простой, но «работающей» характеристики на материале корпуса текстов из учебных пособий по РКИ: чем предложения в тексте длиннее, тем вероятнее, что перед нами текст высокого уровня. На графике видно, как среднее значение длины предложения плавно растет от приблизительно 7 слов на уровне A1 до 16 слов на уровне C2.

Лексическое разнообразие (англ. lexical diversity) представляет собой отношение количества уникальных слов текста к количеству всех слов текста и обозначается величиной от 0 до 1 (когда все слова в тексте уникальны и встретились только по одному разу). Под словом здесь понимается *лексема*, то есть совокупность словоформ данной лексической единицы (Зализняк, 1967). Эта мера полезна для оценки повторяемости, воспроизводимости лексики текста и также способна сигнализировать о его трудности (To, Le, 2013). Например, коэффициент лексического разнообразия отрывка аутентичного публицистического текста в среднем составляет 0,8, а учебного текста уровня B1 – 0,5. Однако этот коэффициент стоит с осторожностью использовать на коротких текстах: в одном абзаце, скорее всего, почти все знаменательные слова будут уникальны, тогда как в целом тексте более вероятно повторяются основные имена, локации, понятия и действия.

Ключевыми являются слова, составляющие уникальность данного текста. Они рассчитываются с помощью специального рейтинга: количество раз, которое слово встречается в этом тексте/частота слова по Национальному

корпусу русского языка[5] (мера TF/IDF с корректирующим коэффициентом). Таким образом, наивысший рейтинг получают слова, которые часто встречаются в данном тексте, но редко – во всех других текстах корпуса, то есть максимально характерные именно для этого текста. Например, в тексте интервью с музыкантом слова *музыка* и *рэп* встречаются по три раза. Но при этом *музыка* встречается в НКРЯ 45 000 раз, а *рэп – 270*. С этой точки зрения, слово *рэп* является более характерным и необходимым для понимания данного текста.

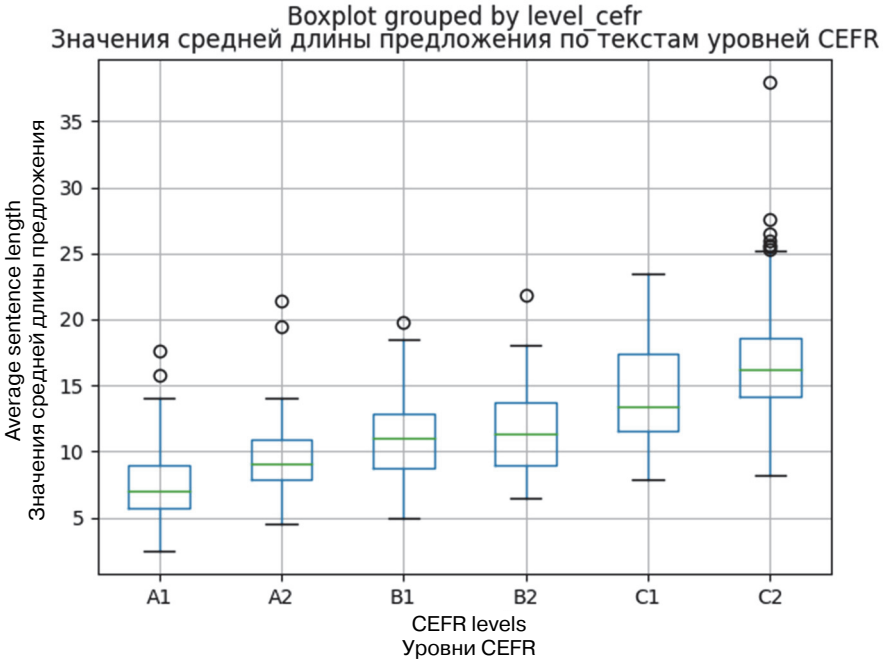

Рис. 2. Значения средней длины предложения по уровням CEFR
Figure 2. Average sentence length values by CEFR Level

При этом появление слова в списке ключевых слов вовсе не означает, что оно должно остаться в тексте при адаптации: слово может быть заменено на синоним или снабжено толкованием. Его присутствие в списке говорит лишь о том, что оно играет важную роль для понимания данного текста и на него стоит обратить особое внимание при переработке текста.

Статистика по лексическим минимумам включает в себя информацию о том, сколько процентов текста покрывается лексическими минимумами того или иного уровня, а ниже указывается список слов, не вошедших в лексический минимум данного уровня. Количество незнакомой лексики является важнейшим показателем языковой доступности текста: многочисленные исследования говорят о самой тесной связи знакомости лексики текста и успешности его понимания (Nation, 2006; Qian, 2002). Государственный стандарт по РКИ также содержит информацию о рекомендуемом количестве незнакомой лексики, который постепенно растет от 2–3 % для уровня А1 до 10 % для уровня С1.

Однако лексические минимумы не всегда оказываются информативны для оценки знакомости лексики: во-первых, они ориентированы прежде все-

[5] Национальный корпус русского языка. URL : https://ruscorpora.ru/new

го на иностранных студентов, поступающих в российские вузы, что приводит к присутствию в них специфической учебной лексики (*деканат, факультет, общежитие*), во-вторых, по разным причинам могут не содержать актуальную лексику, которая с большой вероятностью студентам знакома (*смартфон, офис, туалет* и мн. др.). Поэтому в качестве еще одного показателя вероятной знакомости лексики мы используем частотность слова. Этот параметр широко используется для составления списков и словарей для изучающих русский язык (Sharoff et al., 2013; Лапошина, Лебедева, 2019; Лапошина, 2020; Система.., 2003) и оценки связи между знакомостью лексики текста и его пониманием (Keskisärkkä, Jönsson, 2012; Chen, Meurers, 2016).

Для расчета статистики по частотности слов мы использовали Новый частотный словарь современного русского языка (далее – ЧС)[6], который был составлен на материале коллекции художественных и публицистических текстов 1950–2007 гг. На основании информации о частотности слова сервис «Текстометр» предлагает статистику по ***доле в тексте слов из списка 5 000 самых частотных слов русского языка***, предположительно ***полезным*** и ***редким*** словам, а также отдельно отмечает ***частотные слова, отсутствующие в лексических минимумах***. ***Полезными*** мы обозначили слова, которые, вероятнее всего, еще не знакомы студентам (их нет в лексических минимумах предыдущих уровней), но они есть в минимуме данного уровня или в списке 3 000 самых частотных слов русского языка, согласно Новому частотному словарю. Этот список может использоваться для составления словаря к тексту и заданий на отработку лексики. ***Редкими*** помечаются слова, которые не входят в лексические минимумы, частотный словарь для изучающих РКИ (Sharoff et al., 2013) и список 5 000 самых частотных слов русского языка по Новому частотному словарю. Данный список можно использовать как ориентир при удалении или замене слова.

Примерное время чтения текста рассчитывается с опорой на информацию из государственного стандарта по РКИ и включает информацию по ориентировочному времени чтения в зависимости от вида чтения – изучающего или просмотрового. Такая информация появляется начиная с уровня B1 и составляет для этого уровня 50 слов в минуту для изучающего чтения и 100 слов в минуту для просмотрового. Для уровней ниже B1 мы взяли на себя смелость продолжить эту шкалу расчетной скорости чтения исходя из педагогического опыта. Однако стоит понимать ориентировочный характер подобной информации: скорость чтения, помимо уровня владения языком, может зависеть от таких факторов, как родной язык, читательский опыт студента, фонетическое и синтаксическое удобство текста и пр.

Частотный словарь текста представляет собой список всех лексем текста, отсортированный по количеству их упоминаний в тексте. Он может быть полезен для объективизации процесса выбора лексики, которая будет в фокусе изучения или, наоборот, подлежит удалению или упрощению.

Для ***демонстрации работы сервиса «Текстометр»***, приведем показатели текста из УМК «Жили-были» (Миллер и др., 2016) элементарного уровня (таблица, фрагмент исследованного текста представлен перед таблицей).

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

Скоро мы поедем на экскурсию в Москву. Мы много говорили об этом на уроке, и вчера Иван Петрович попросил нас сделать необычное домашнее задание – написать о нашем самом интересном путешествии. Вот что написала Ирена. В этом году я и мои друзья ездили в Крым, в Ялту. Эта поездка мне очень понравилась. В Симферополь, столицу Крыма, мы ехали на поезде. [...]

Показатели сложности текста из УМК «Жили-были», полученные с помощью сервиса «Текстометр»

Параметр	Значение
Уровень, заявленный в пособии	A1
Уровень, предсказанный моделью	A1. Элементарный уровень
Слов в тексте	200
Уникальных слов	121
Лексическое разнообразие	0,6
Предложений в тексте	22
Средняя длина предложения	6,57
Ключевые слова	Крым, Ялта, поезд, Симферополь, час, автобус, интересный, поездка
Самые полезные слова	Берег, деревня, во-первых, купе, выбирать, пешком, задание, через, во-вторых, домашний, есть, самый, необычный, узнавать
Лексический список A1 покрывает	87 % текста
Не входит в лексический список A1	Во-первых, выбирать, ботанический, уютный, экспресс, задание, купе, через, необычный, современность, чудесный, деревня, пешком, во-вторых, ореанда, самый, берег, домашний, городок, узнавать
Лексический список A2 покрывает	92 % текста
Не входит в лексический список A2	Современность, чудесный, во-первых, купе, ботанический, уютный, экспресс, во-вторых, ореанда, ну, необычный, городок, узнавать
Лексический список B1 покрывает	95 % текста
Не входит в лексический список B1	Современность, чудесный, купе, ботанический, уютный, экспресс, ореанда, необычный, городок
Лексический список B2 покрывает	98 % текста
Не входит в лексический список B2	Современность, городок, ореанда, ботанический
Лексический список C1 покрывает	98 % текста
Не входит в лексический список C1	Городок, ореанда, ботанический
Частотный список 5000 покрывает	96 % текста
Полезные слова, которых нет в лексическом минимуме	Узнавать
Редкие слова	Экспресс, ботанический, ореанда
Изучающее чтение займет	7 мин
Просмотровое чтение займет	4 мин
Возможные грамматические темы	Предложный падеж
Частотный список текста	В 13; мы 13; и 9; быть 6; на 6; я 5; час 4; интересный 3; Крым 3; …

Parameter values of the text from "Zhili-byli" textbook obtained by Textometr

Parameter	Value
Text level declared in the textbook	A1
Predicted by Textometr level	A1. Elementary level
Words	200
Unique words	121
Lexical diversity	0.6
Sentences	22
Average sentence length	6.57
Keywords	Крым, Ялта, поезд, Симферополь, час, автобус, интересный, поездка
Most useful words	Берег, деревня, во-первых, купе, выбирать, пешком, задание, через, во-вторых, домашний, есть, самый, необычный, узнавать
Text coverage by A1 vocabulary list	87% of text
Words out of A1 vocabulary list	Во-первых, выбирать, ботанический, уютный, экспресс, задание, купе, через, необычный, современность, чудесный, деревня, пешком, во-вторых, ореанда, самый, берег, домашний, городок, узнавать
Text coverage by A2 vocabulary list	92% of text
Words out of A2 vocabulary list	Современность, чудесный, во-первых, купе, ботанический, уютный, экспресс, во-вторых, ореанда, ну, необычный, городок, узнавать
Text coverage by B1 vocabulary list	95% of text
Words out of B1 vocabulary list	Современность, чудесный, купе, ботанический, уютный, экспресс, ореанда, необычный, городок
Text coverage by B2 vocabulary list	98% of text
Words out of B2 vocabulary list	Современность, городок, ореанда, ботанический
Text coverage by C1 vocabulary list	98% of text
Words out of C1 vocabulary list	Городок, ореанда, ботанический
Text coverage by frequency list 5 000	96% of text
Useful words that are out of lexical minima	Узнавать
Rare words	Экспресс, ботанический, ореанда
Detail reading for details will take	7 min
Skimming reading will take	4 min
Possible grammar topics	Prepositional case
Frequency list of the text	В 13; мы 13; и 9; быть 6; на 6; я 5; час 4; интересный 3; Крым 3; …

Таблица демонстрирует лингвистические характеристики текста, полученные с помощью сервиса «Текстометр». Оценка математической модели,

«A1», совпадает с информацией в методической справке пособия. Текст соответствует норме элементарного уровня по длине предложений и количеству слов. Количество незнакомых слов в лексических минимумах по мере возрастания уровня ожидаемо падает, однако для заявленного уровня A1 оно составляет более 13 %, тогда как в государственном стандарте рекомендовано количество незнакомых слов 2–3 %. При детальном рассмотрении становятся видны особенности входных данных и нашей системы, которые объясняют такую разницу показателей незнакомых слов:

1. В некоторых случаях слово может отсутствовать в лексическом минимуме, однако расцениваться преподавателями и авторами пособий как потенциально знакомое студентам. Например, слово *городок* отсутствует в лексическом минимуме и, следовательно, определяется сервисом «Текстометр» как незнакомое, тогда как *город* есть в лексическом минимуме уровня A1 и, по информации государственного стандарта по РКИ, учащийся уровня A1 должен владеть начальными знаниями об образовании уменьшительных форм существительных. В результате формально слово *городок* оказывается незнакомым, отсутствующим в лексическом списке даже на уровне C1. Сюда же можно отнести случаи со словами *необычный* (появляется в списке B2) – *обычно* (есть в A1), *современность* (появляется в списке C1) – *современный* (есть в A1).

2. В текстах встречаются интернациональные слова, о значении которых студенты, владеющие европейскими языками, могут догадаться: *экспресс, купе, ботанический*. Однако для остальных групп студентов эти слова незнакомы и могут представлять трудность, некоторые из них отсутствуют даже в лексическом минимуме C1.

3. Некоторых слов нет в лексическом минимуме данного уровня, однако есть их синонимы, например: *задание* vs *упражнение*.

4. Наконец, ошибка может произойти на этапе автоматического грамматического анализа текста модулем Mystem: так, автоматический анализатор предлагает начальную форма от *узнали – узнавать*, а не *узнать*, тогда как в лексическом минимуме указана форма совершенного вида.

Проверка качества работы сервиса. С целью проверки качества работы модели в реальных педагогических условиях, был также разработан и проведен эксперимент, позволяющий сравнить результаты работы модели по автоматическому определению уровня сложности текста с нормами государственного стандарта по РКИ, экспертной оценкой практикующих преподавателей РКИ, суждением самих студентов об уровне сложности предложенных текстов и результатами выполнения студентами послетекстовых заданий. Эксперимент был проведен на базе Государственного института имени А.С. Пушкина и охватил 78 студентов интергрупп уровня B1. Для эксперимента были отобраны три аутентичных текста интернет-издания The Village, оцененные моделью как A2, B1 и B2 (Лапошина, 2018). Верное выстраивание текстов математической моделью на шкале сложности было подтверждено и экспертной оценкой, и мнением студентов. Однако была обнаружена тенденция автоматической системы завышать сложность. Возможно, это связано с тем, что нормы из государственного стандарта по РКИ, откуда частично черпаются сведения о тех или иных показателях уровня текста, оказываются несколько ниже, чем реальные возможности студентов и мнение

преподавателей о сложности текста, полученные в ходе эксперимента. По результатам эксперимента мы учли полученные данные и провели дополнительную настройку системы определения уровня.

Заключение

Данные статистики использования сервиса «Текстометр» за первые полгода его работы – более 3200 уникальных пользователей из 76 стран мира – ярко иллюстрируют актуальность задачи самостоятельной подготовки текстов преподавателями и востребованность информации такого рода. Она необходима не только преподавателям и методистам, но также авторам пособий и представителям издательств для проверки соответствия текстового материала заявленному уровню и учебным целям.

С другой стороны, очевидно, что любая автоматизация является долгим процессом и требует тщательного тестирования и доработки. Перечислим основные недостатки сервиса на данный момент и связанные с ними векторы дальнейшей работы.

Во-первых, это описанное выше отсутствие в лексических минимумах слов, которые авторами и преподавателями оцениваются как знакомые студентам, исходя из наличия в минимуме однокоренных слов и знаний студентов о словообразовании. Эта проблема стала очевидна именно при автоматизации подсчетов вхождения лексики в списки, поскольку то, что для преподавателя кажется очевидным (если студент знает слово *автобус*, то, скорее всего, он поймет и *автобусный*), для машины таковым не является. Данная проблема будет решена расширением лексических минимумов с учетом доступного для студентов данного уровня словообразования.

Во-вторых, встает вопрос маркировки пособий по РКИ по уровням сложности. Качество работы математической модели напрямую зависит от качества и единообразия материалов, на которых она обучается. Для задач данного исследования это информация об уровне в методической справке пособий по РКИ, из которых была сформирована обучающая коллекция текстов. Стоит отметить ряд проблем, с которыми мы столкнулись на этом этапе: туманность описания уровня (*для продвинутых, для второго семестра первого года обучения*), размытость границ уровней (*B1–C1*), а также риск субъективности этой информации. Насколько нам известно, в настоящий момент не существует единой формальной процедуры маркировки пособия по уровням CEFR. Получается, принятие этого решения ложится на авторов пособия и редколлегию издательства (а иногда только на авторов, поскольку на рынке существуют пособия, изданные самостоятельно), что может приводить к необъективности данных и несопоставимости между собой разных пособий. Поэтому одним из важнейших этапов дальнейшего развития проекта мы считаем дополнительную разметку текстов для обучения модели несколькими экспертами.

Наконец, машинная модель, в отличие от опытного преподавателя, пока не способна учесть всю совокупность внетекстовых факторов, влияющих на восприятие текста для конкретного учащегося: родной язык, используемый учебный комплекс, личностные особенности, интересы, опыт изучения языков, способность к догадке, общая начитанность и многое другое. В связи с этим перспективным направлением нам кажется совершенствование системы

обработки лексики: подключение учета интернационализмов и общеславянской лексики для изменения списка полезных и неизвестных слов в зависимости от контингента, разработка собственных лексических списков с учетом информации о частотности слов в русском языке и в пособиях по РКИ.

Среди более смелых и масштабных векторов развития обозначим работы в области создания рекомендательной системы по симплификации, адаптации текстов для изучающих русский как иностранный.

Список литературы

Арутюнов А.Р. Теория и практика создания учебника русского языка для иностранцев. М. : Русский язык, 1990. 167 с.

Бим И.Л. Методика обучения иностранным языкам как наука и проблемы школьного учебника. М. : Русский язык, 1977. 288 с.

Вятютнев М.Н. Теория учебника русского языка как иностранного (методические основы). М. : Русский язык, 1984. 144 с.

Зализняк А.А. Русское именное словоизменение. М. : Наука, 1967. 373 с.

Лапошина А.Н. Корпус текстов учебников РКИ как инструмент анализа учебных материалов // Русский язык за рубежом. 2020. № 6 (283). С. 22–28.

Лапошина А.Н. Опыт экспериментального исследования сложности текстов по РКИ // Динамика языковых и культурных процессов в современной России : материалы VI Конгресса РОПРЯЛ (Уфа, 11–14 октября 2018 г.) : сборник статей. 2018. Вып. 6. С. 1544–1549.

Лапошина А.Н., Лебедева М.Ю. Корпусный подход к решению проблемы отбора лексики в обучении РКИ // Slavica Helsingiensia. 2019. № 52. С. 359–368.

Микк Я.А. Оптимизация сложности учебного текста : в помощь авторам и редакторам. М. : Просвещение, 1981. 119 с.

Миллер Л.В., Политова Л.В. Рыбакова И.Я. Жили-были... 28 уроков русского языка для начинающих : учебник. СПб. : Златоуст, 2016. 112 с.

Система лексических минимумов современного русского языка : 10 лексических списков : от 500 до 5000 самых важных русских слов / под ред. В.В. Морковкина. М. : Астрель, 2003. 768 с.

Томина Ю.А . Объективная оценка языковой трудности текстов (описание, повествование, рассуждение, доказательство) : дис. ... канд. пед. наук. М., 1985. 225 с.

Alexander P.A., Jetton T.L. The role of importance and interest in the processing of text // Educational Psychology Review. 1996. No 8 (1). Pp. 89–121.

Chen X., Meurers D. Characterizing text difficulty with word frequencies // Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications. 2016. Vol. 11. Pp. 84–94.

DuBay W. The principles of readability. Costa Mesa, CA : Impact Information, 2004. 76 p.

Graesser A.C., McNamara D.S., Cai Z., Conley M., Li H., Pennebaker J. Coh-Metrix measures text characteristics at multiple levels of language and discourse // The Elementary School Journal. 2014. No. 15 (2). Pp. 210–229.

Karpov N., Baranova J., Vitugin F. Single-sentence readability prediction in Russian // Proceedings of Analysis of Images, Social Networks, and Texts Conference (AIST). 2014. Vol. 3. Pp. 91–100.

Keskisärkkä R., Jönsson A. Investigations of synonym replacement for Swedish // Northern European Journal of Language Technology. 2013. No 3. Pp. 41–59.

Laposhina A.N., Veselovskaya T.S., Lebedeva M.U., Kupreshchenko O.F. Automated text readability assessment for Russian second language learners // Dialogue 2018 : Proceedings of the International Conference. 2018. Vol. 17. Issue 24. Pp. 396–406.

Nation P. How large a vocabulary is needed for reading and listening? // Canadian Modern Language Review. 2006. No 63. Pp. 59–81.

Qian D.D. Investigating the relationship between vocabulary knowledge and academic reading performance: an assessment perspective // Language Learning. 2002. No 52 (3). Pp. 513–536.

Reynolds R. 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. 2016. Vol. 11. Pp. 289–300.

Sharoff S., Kurella S., Hartley A. Seeking needles in the web's haystack : finding texts suitable for language learners // Proceedings of the 8th Teaching and Language Corpora Conference (TaLC-8). Lisbon, 2008. Pp. 365–370.

Sharoff S., Umanskaya E., Wilson J. A frequency dictionary of Russian: core vocabulary for learners. New York: Routledge, 2013. 400 p.

To V., Le T. Lexical density and readability : a case study of English textbooks // Proceedings of the Australian Systemic Functional Linguistics Association Conference. Melbourne, 2013. Pp. 61–71.

Сведения об авторах:

Лапошина Антонина Николаевна, ведущий эксперт лаборатории когнитивных и лингвистических исследований, Государственный институт русского языка имени А.С. Пушкина, участник исследовательской группы «Обучение русскому языку в цифровую эпоху», автор и разработчик проекта «Текстометр». *Сфера научных интересов:* компьютерные технологии в РКИ, корпусная лингводидактика. E-mail: ANLaposhina@pushkin.institute

Лебедева Мария Юрьевна, кандидат филологических наук, ведущий научный сотрудник лаборатории когнитивных и лингвистических исследований, доцент кафедры методики преподавания РКИ, Государственный институт русского языка имени А.С. Пушкина. Руководитель грантов РФФИ и РНФ. *Сфера научных интересов:* корпусная лигводидактика, онлайн-обучение РКИ, особенности чтения в цифровую эпоху. E-mail: MULebedeva@pushkin.institute

DOI 10.22363/2618-8163-2021-19-3-331-345

Research article

Textometr: an online tool for automated complexity level assessment of texts for Russian language learners

Antonina N. Laposhina✉, Maria Yu. Lebedeva

Pushkin State Russian Language Institute,
6 Akademika Volgina St, Moscow, 117485, Russian Federation
✉ ANLaposhina@pushkin.institute

Abstract. Evaluation of text accessibility seems to be an extremely urgent and labor-consuming task in the process of preparing texts for teaching Russian as a foreign language. On the other hand, the procedure of assigning a text to one of the levels on the CEFR scale (from A1 to C2) is well-formalized and described in the professional literature, which opens opportuni-

ties for its automation. This paper presents Textometr – a new free web-based tool for estimating CEFR level and other key statistics from any given text in Russian that can be relevant for adapting it for foreign students. The automated assessment of the text level here is based on a regression model, trained on the dataset of more than 800 texts from Russian textbooks for foreigners, applying several machine learning and natural language processing methods. In addition to the CEFR level, the tool provides information relevant for adapting the text to educational tasks: lists of keywords and words for a potential vocabulary list, statistics on the text coverage by frequency lists and CEFR-graded vocabulary lists (lexical minima), a frequency list of the text, a forecast of the time needed for reading. The tool shortages at the current stage of development and suggested ways to solve them are also discussed. Finally, the results of the test on the tool quality and the vectors for its further development are reported. Textometr can provide helpful information not only to teachers and guidance teachers, but to authors of textbooks and publishers to check the compliance of the text content with the declared level and educational goals.

Keywords: Russian as a foreign language, educational text, text complexity, reading, text adapting, computational linguodidactics, computer assisted language learning, Russian language learning, web tools

Article history: received 16.02.2021; accepted 18.05.2021.

For citation: Laposhina, A.N., & Lebedeva, M.Yu. (2021). Textometr: An online tool for automated complexity level assessment of texts for Russian language learners. *Russian Language Studies, 19*(3), 331–345. (In Russ.) http://dx.doi.org/10.22363/2618-8163-2021-19-3-331-345

References

Alexander, P.A., & Jetton, T.L. (1996). The role of importance and interest in the processing of text. *Educational Psychology Review*, 8(1), 89–121.

Arutyunov, A.R. (1990). *Theory and practice of creating a textbook of the Russian language for foreigners*. Moscow: Russkii Yazyk Publ. (In Russ.)

Bim, I.L. (1977). *Methods of teaching foreign languages as a science and problems of a school textbook.* Moscow: Russkii Yazyk Publ. (In Russ.)

Chen, X., & Meurers, D. (2016). Characterizing text difficulty with word frequencies. *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications (June 16, 2016), 11*, 84–94. San Diego, CA, USA.

DuBay, W. (2004). *The principles of readability.* Costa Mesa, CA: Impact Information.

Graesser, A.C., McNamara, D.S., Cai, Z., Conley, M., Li, H., & Pennebaker, J. (2014). Coh-Metrix measures text characteristics at multiple levels of language and discourse. *The Elementary School Journal*, 15(2), 210–229.

Karpov, N., Baranova, J., & Vitugin, F. (2014). Single-sentence readability prediction in Russian. *Proceedings of Analysis of Images, Social Networks, and Texts conference (AIST)*, (3), 91–100.

Keskisärkkä, R., & Jönsson, A. (2013). Investigations of synonym replacement for Swedish. *Northern European Journal of Language Technology*, (3), 41–59.

Laposhina, A.N. (2018). Insights from an experimental study on the text complexity for Russian as a foreign language. *The Dynamics of Linguistic and Cultural Processes in Modern Russia: Proceedings of the VI Congress of ROPRYAL*, (6), 1544–1549. (In Russ.)

Laposhina, A.N. (2020). A corpus of Russian textbook materials for foreign students as an instrument of an educational content analysis. *Russian Language Abroad*, (6(283)), 22–28. (In Russ.)

Laposhina, A.N., & Lebedeva, M.U. (2019). Corpus approach to vocabulary selection for learning Russian as a foreign language. *Slavica Helsingiensia*, (52), 359–368. (In Russ.)

Laposhina, A.N., Veselovskaya, T.S., Lebedeva, M.U., & Kupreshchenko, O.F. (2018). Automated text readability assessment for Russian second language learners. *Dialogue 2018: Proceedings of the International Conference, 17*(24), 396–406.

Mikk, Ya.A. (1981). Optimizing the complexity of educational text: A help for authors and editors. Moscow: Prosveshchenie Publ. (In Russ.)

Miller, L.V., Politova, L.V., & Rybakova, I.A. (2016). *Once upon a time... 28 Russian lessons for beginners: Textbook*. Saint Petersburg: Zlatoust Publ. (In Russ.)

Morkovkin, V.V. (Ed.). (2003). *The system of lexical minima of the modern Russian language: 10 lexical lists: From 500 to 5000 of the most important Russian words.* Moscow: Astrel Publ. (In Russ.)

Nation, P. (2006). How Large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, (63), 59–81.

Qian, D.D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning*, 52(3), 513–536.

Reynolds, R. (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, 11*, 289–300.

Sharoff, S., Kurella, S., & Hartley, A. (2008). Seeking needles in the web's haystack: Finding texts suitable for language learners. *Proceedings of the 8th Teaching and Language Corpora Conference (TaLC-8)* (pp. 365–370). Lisbon.

Sharoff, S., Umanskaya, E., & Wilson, J. (2013). *A frequency dictionary of Russian: Core vocabulary for learners.* New York: Routledge.

To, V., & Le, T. (2013). Lexical density and readability: A case study of English textbooks. *Proceedings of the Australian Systemic Functional Linguistics Association Conference* (October 1–3, 2013) (pp. 61–71). Melbourne.

Tomina, Yu.A. (1985). *Objective assessment of the language difficulty of texts (description, narration, reasoning, argumentation)* (Candidate dissertation, Moscow). (In Russ.)

Vyatyutnev, M.N. (1984). *Textbook theory of Russian as a foreign language (methodological foundations).* Moscow: Russkii Yazyk Publ. (In Russ.)

Zaliznak, A.A. (1967). *Russian nominal inflection.* Moscow: Nauka Publ. (In Russ.)

Bio notes:

Antonina N. Laposhina, leading expert, Laboratory of Cognitive and Linguistic Studies, Pushkin State Russian Language Institute, member of the research group "Teaching Russian in the Digital Age", the author and developer of the "Textometr" project. *Research interests:* computer assisted language learning, corpus-based language learning, methods of teaching Russian as a foreign language. E-mail: ANLaposhina@pushkin.institute

Maria Yu. Lebedeva, Candidate of Philology, leading researcher of the Laboratory of Cognitive and Linguistic Research, Associate Professor of the Department of Methods of Teaching Russian as a Foreign Language, Pushkin State Russian Language Institute. Head of grants from the Russian Foundation for Basic Research and the Russian Science Foundation. *Research interests:* corpus based language learning, methods of online teaching Russian, reading strategies in the digital age. E-mail: MULebedeva@pushkin.institute