На правах рукописи

Лапошина Антонина Николаевна

Лингводидактическое обоснование применения автоматической оценки сложности учебного текста в преподавании РКИ

5.8.2 – Теория и методика обучения и воспитания (русский язык как иностранный, уровень общего, профессионального, дополнительного образования, профессионального обучения)

Автореферат
диссертации на соискание ученой степени
кандидата педагогических наук

Москва – 2023

Работа выполнена в федеральном государственном бюджетном образовательном учреждении высшего образования «Государственный институт русского языка им. А.С. Пушкина» на кафедре методики преподавания русского языка как иностранного

Научный руководитель: кандидат филологических наук **Лебедева Мария Юрьевна**

Официальные оппоненты:

Попова Татьяна Игоревна, доктор филологических наук, профессор, федеральное государственное бюджетное образовательное учреждение высшего образования «Санкт-Петербургский государственный университет», кафедра русского языка как иностранного и методики его преподавания, профессор, заведующий кафедрой

Дьякова Татьяна Александровна, кандидат педагогических наук, доцент, федеральное государственное бюджетное образовательное учреждение высшего образования «Тамбовский государственный университет имени Г.Р. Державина», кафедра русского языка, заведующий кафедрой

Ведущая организация: федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет “Высшая школа экономики”» (г. Москва)

Защита состоится 18 мая 2023 г. в 10.00 на заседании диссертационного совета 24.2.292.02, созданного на базе федерального государственного бюджетного образовательного учреждения высшего образования «Государственный институт русского языка им. А.С. Пушкина», по адресу: 117485, г. Москва, ул. Академика Волгина, 6.

С диссертацией можно ознакомиться в научной библиотеке федерального государственного бюджетного образовательного учреждения высшего образования «Государственный институт русского языка им. А.С. Пушкина» и на официальном сайте: https://www.pushkin.institute

Материалы по защите диссертации размещены на официальном сайте ФГБОУ ВО «Государственный институт русского языка им. А.С. Пушкина»: https://www.pushkin.institute/science/dissovet/detail.php?ELEMENT_ID=32370

Автореферат разослан «___»_____ 2023 г.

Ученый секретарь
диссертационного совета Филиппова Варвара Михайловна

Общая характеристика диссертационной работы

Диссертация посвящена разработке системы автоматической оценки уровня сложности текста, основанной на его количественных характеристиках, и возможностям применения такой системы в практике обучения русскому языку как иностранному.

Актуальность исследования. Обучение русскому языку как иностранному характеризуется текстоцентричностью, а отбор и подготовка текстов признается исследователями крайне актуальной задачей современной методики преподавания русского языка как иностранного. Для достижения поставленных учебных задач текстовый материал должен соответствовать учащимся по множеству параметров, среди которых одним из важнейших считается уровень языковой сложности текста. Соответствовать – значит иметь оптимальное соотношение знакомой и новой информации. Однако современные исследования показывают, что представления об уровне сложности текста у разных преподавателей могут не совпадать. Это значительно затрудняет сравнимость сложности текстов, оцененных разными экспертами, а также маркировку учебных пособий по уровню. Таким образом, потребность в единой формальной объективной системе оценки сложности текста на фоне постоянной потребности в обновлении коллекций учебных текстов для занятий РКИ и составляет актуальность данного исследования.

Степень разработанности научной проблемы

Проблема отбора текстов на русском языке для учебников или занятий в иноязычной аудитории признается крайне актуальной задачей в методике преподавания РКИ, при этом к тексту предъявляется целый ряд требований, среди которых одним из центральных выделяется языковая доступность текста (см. [Акишина, Каган 2002]; [Кулибина 2015]; [Шустикова, Кулакова 2011]; [Щукин 2003] и мн. др.).

Большинство исследователей доступности текста разграничивают понятия *сложности* текста как его объективной характеристики, выраженной в его лингвистических параметрах, и *трудности* текста как более широкой и частично субъективной характеристики, включающей в себя, помимо сложности материала, факторы подготовленности читателя (см. [Микк 1980]; [Томина 1985]; [Солнышкина, Кисельников 2015]; [Оборнева 2006]; [Филиппова 2010], [Collins-Thompson 2014] и др.).

Современный подход к созданию системы оценки сложности текстов в большинстве случаев подразумевает использование методов машинного обучения на коллекции текстовых образцов (см. [Francois, Fairon 2012], [DuBay 2006], [Graesser et al. 2014]).

Отмечается ряд особенностей, связанных с автоматизацией процесса оценки сложности текста для его предъявления в иноязычной аудитории (см. [Pilan et al. 2015], [Sung et al. 2015]; [Zalmout 2016] и др.).

Опыт создания подобных систем для нужд РКИ описан в нескольких пионерских работах этой области (см. [Karpov et al. 2014]; [Reynolds 2016]; [Sharoff 2006]).

В то же время следует отметить, что вопросы репрезентативности корпуса текстов по РКИ для обучения модели, экспериментальной проверки качества и применимости полученной модели в практической деятельности широкого круга преподавателей РКИ, а также методики интерпретации результатов работы автоматических систем оценки сложности текста, до сих пор не были системно описаны в научной литературе.

Целью данной работы мы ставим разработку системы автоматической оценки сложности текста для изучающих РКИ на основании статистических параметров текста и обоснование её применимости в практике преподавания РКИ.

Объектом исследования выступает сложность учебного текста в аспекте обучения РКИ как его объективное измеряемое свойство, выражаемое набором лингвистических характеристик.

Предметом исследования является система автоматической оценки сложности учебного текста в практике преподавания РКИ.

Основная **гипотеза исследования** состоит в том, что тексты, предъявляемые в качестве единиц обучения РКИ, могут быть объективно дифференцированы по мере их языковой сложности автоматической системой на основании ряда вычисляемых лингвистических параметров текста.

Цель исследования обусловила необходимость постановки и решения следующих **задач**:

1. Анализ научных работ зарубежных и отечественных ученых, посвященных проблеме оценки сложности текстов, в том числе в парадигме уровневой системы русского языка как иностранного.

2. Отбор и систематизация представительного эталонного корпуса текстов, содержащего образцы текстовых единиц разных уровней сложности.

3. Выделение и оценка эффективности лингвистических признаков для текстов данного корпуса.

4. Создание модели машинного обучения на материале подготовленного корпуса текстов и их признаков.

5. Экспериментальное исследование качества работы модели.

6. Разработка и тестирование сервиса по автоматическому анализу текстов, основанного на созданной модели машинного обучения.

7. Разработка комплекта рекомендаций по работе с сервисом по анализу текстов в зависимости от практической задачи обучения РКИ.

Междисциплинарный характер обусловливает сочетание в работе **методов исследования** компьютерной лингвистики и лингводидактики:

1. Теоретический анализ, который проводился с целью всестороннего изучения разработанности рассматриваемой проблемы, возможных подходов к её решению, а также определения шкалы сложности текста и нормоустанавливающих документов.

2. Корпусные методы сбора и анализа коллекции текстов для создания корпуса текстов из учебных пособий по РКИ для корректного обучения математической модели.

3. Методы компьютерной лингвистики для автоматической обработки текста на естественном языке.

4. Статистические методы и методы машинного обучения для создания предсказательной модели по определению уровня текста на основании лингвистических признаков.

5. Методы анкетирования и тестирования учащихся и преподавателей, методы статистического анализа результатов эксперимента.

Теоретическую основу исследования составляют труды по:

– методике обучения чтению и отбору текстов в иноязычной аудитории (А.А. Акишиной, Н.В. Кулибиной, Т.В. Шустиковой, К.А. Роговой, А.Н. Щукина и др.);

– уровневой системе ТРКИ (Н.П. Андрюшиной, Т.Е. Владимировой, Т.В. Козловой, М.М. Нахабиной, Н.И. Соболевой, Л.П. Клобуковой и др.);

– теории учебника русского языка (И.Л. Бим, А.Р. Арутюнова, М.Н. Вятютнева);

– проблеме оценки доступности учебных текстов (Я.А. Микка, Ю.А. Тулдавы, Ю.А. Томиной, М.И. Солнышкиной, А.С. Кисельникова, О.В. Филипповой и др.);

– методам автоматизации процесса оценки сложности текста (W. DuBay, A. Graesser, D. McNamara, Y.T. Sung, N. Zalmout и др.);

– разработке автоматической оценки сложности русских текстов для преподавания в иностранной аудитории (В.Г. Сибирцевой, Н.В. Карпова, R. Reynolds, S. Sharoff и др.).

Материалом исследования послужили:

– зарубежная и российская научная, научно-методическая и учебная литература по проблемам теории и методики преподавания русского языка, теории текста, прикладной и корпусной лингвистики;

– нормоустанавливающие и рекомендательные документы, такие как «Общеевропейские компетенции владения иностранным языком: изучение, преподавание, оценка», система требований к освоению русского языка как иностранного, включая лексические минимумы по РКИ;

– разработанный корпус текстов из печатных и электронных учебных пособий для изучающих русский язык как иностранный;

– данные анкетирования преподавателей русского языка и иностранных студентов.

Научная новизна настоящего исследования заключается в комплексном анализе учебного текста с точки зрения формальных показателей его сложности для иностранных учащихся и определяется следующими результатами:

– обобщена и формализована система лингвистических признаков текста, оказывающих влияние на уровень его сложности в системе обучения русскому языку как иностранному;

– разработана и реализована методика сбора и разметки сбалансированного корпуса образцов текстов из различных пособий по русскому языку как иностранному;

– создана и внедрена в практику математическая модель автоматической оценки сложности текста на русском языке для изучающих русский язык как иностранный;

– разработана и апробирована методика верификации применимости автоматической системы оценки сложности текстов в практике обучения РКИ

путем сравнения результатов работы системы с экспертной оценкой сложности текстов и оценкой текстов иностранными учащимися;

– предложен комплект методических материалов по вариантам интерпретации формальных лингвистических характеристик текста в зависимости от уровня владения русским языком и/или типа методической задачи.

Теоретическую значимость работы составляют:

– систематизация опыта исследователей в области определения сложности текста с позиций обучения иностранным языкам;

– разработка и реализация концепции эталонного корпуса текстов из учебных пособий по РКИ с информацией об их уровне;

– обобщение и проверка эффективности формальных признаков текста при оценке его сложности в преподавании РКИ.

Практическая значимость проведенного исследования состоит в:

– разработке системы автоматической оценки сложности текста для иностранных студентов, изучающих русский язык и создании на её основе веб-сервиса «Текстометр»[1];

– формировании комплекта рекомендаций по работе с сервисом для широкого круга специалистов (преподавателей, методистов, авторов пособий, представителей издательств и др.) для отбора учебных текстов оптимального уровня языковой сложности.

Основные положения, выносимые на защиту:

1. Сложность учебного текста является объективной характеристикой, детерминированной совокупностью признаков, оказывающих влияние на трудность его восприятия.

2. Описание уровневой системы РКИ может служить источником информации о базовых значениях измеряемых лингвистических характеристик учебных текстов для учащихся различных уровней владения русским языком и стать основой системы автоматической оценки сложности текста для иностранных учащихся.

3. Репрезентативный корпус текстов из пособий по РКИ отражает совокупность авторских интерпретаций уровневых описаний CEFR и ТРКИ и

[1] Ресурс доступен по веб-адресу: https://textometr.ru

представляет обобщенный коллективный опыт экспертного сообщества в ранжировании учебных текстов по шкале уровней владения русским языком.

4. Оценка уровня языковой сложности текста может быть получена в результате работы машинной модели, обученной на корпусе текстов из пособий по РКИ и их лингвистических характеристиках.

5. Применимость технологии автоматической системы оценки сложности текстов может быть верифицирована путем сравнения результатов работы системы с экспертной оценкой сложности текстов и оценкой текстов иностранными учащимися.

6. Создание веб-сервиса, основанного на разработанной технологии оценки сложности текстов, способствует повышению объективности оценки уровня текста и оптимизирует процесс подготовки текста к занятию РКИ.

Апробация исследования осуществлялась среди иностранных студентов и преподавателей русского языка как иностранного. В эмпирических исследованиях приняли участие:

– студенты подготовительного факультета и факультета обучения русскому языку как иностранному Государственного института русского языка им. А.С. Пушкина и их преподаватели (78 студентов и 7 преподавателей);

– российские и зарубежные преподаватели РКИ, привлеченные посредством краудсорсинга в специализированных профессиональных интернет-сообществах (41 человек).

Основные положения диссертации были изложены автором на следующих научно-практических конференциях и вебинарах:

1. Ежегодная международная конференция по компьютерной лингвистике и интеллектуальным технологиям «Диалог 2017» (РГГУ, 31 мая – 3 июня 2017 г.);

2. Международная научно-практическая интернет-конференция «Актуальные вопросы описания и преподавания русского языка как иностранного/неродного» (Москва, 27 ноября – 1 декабря 2017 г.);

3. Ежегодная международная конференция по компьютерной лингвистике и интеллектуальным технологиям «Диалог 2018» (РГГУ, 30 мая – 2 июня 2018 г.);

4. Международная научно-практическая конференции «Корпусные и компьютерные технологии и лингвистические проблемы» (Нижний Новгород, 12–14 октября 2018 г.);

5. XLVIII международная филологическая конференция (СПбГУ, 26 марта 2019);

6. Международный форум «РКИ-перезагрузка 2021: уроки пандемии» (Москва, 18 – 19 июня 2021);

7. VII конгресс РОПРЯЛ «Динамика языковых и культурных процессов в современной России» (УрФУ, 7 – 8 октября 2021 г.);

8. Вебинар «Текстометр: новый инструмент для подготовки текста к занятию по РКИ» на портале «Образование на русском» (27 марта 2019).

Личный вклад автора. Автором проведен всесторонний анализ профильной литературы в области оценки сложности текстов для различных аудиторий читателей, методики отбора текстов в иноязычной аудитории и методов автоматизации процесса оценки сложности текста.

Автором подробно изучены и систематизированы описания уровневых систем в области преподавания русского языка как иностранного на предмет возможности формализации лингвистических показателей, указывающих на уровень сложности текста для иностранных учащихся.

Для создания предсказательной модели автором был собран репрезентативный корпус образцов текстов из учебников и пособий по РКИ уровней A1-C1 общим объемом 802 текста, 266 000 слов.

Автором была создана предсказательная модель по оценке сложности текстов на русском языке по шкале уровней CEFR, а также создан веб-сервис для доступа к результатам работы модели широкого круга пользователей.

Для экспериментальной верификации применимости созданной модели автором были сконструированы экспериментальные тексты и контрольно-измерительные материалы.

Наконец, автором был разработан комплект рекомендаций по работе с сервисом по анализу текстов в зависимости от практической задачи обучения РКИ.

Объем и структура исследования определяются поставленными в нем целями и задачами. Общий объем диссертации – 189 страниц. Работа содержит введение, три главы, заключение, библиографию, 4 приложения.

Основное содержание диссертационной работы

Во **Введении** обосновывается актуальность темы исследования, формулируются цель и задачи работы, определены объект и предмет, методы исследования. Устанавливаются основные теоретические предпосылки исследования, формируется его рабочая гипотеза, излагаются положения, выносимые на защиту. Характеризуются теоретическая и практическая значимость, научная новизна исследования.

В первой главе – «Теоретические основы автоматизации процесса анализа сложности текста в практике преподавания РКИ» – представляется терминологический аппарат исследования, определяется шкала сложности текстов применительно к задачам преподавания русского языка как иностранного и рассматривается предыдущий исследовательский опыт отбора и оценки сложности учебных текстов.

Базируясь на предложенных в методической литературе определениях учебного текста (см. [Азимов, Щукин 2009]; [Щукин 2003]; [Гальперин 2006]; [Тёрёчик 2012]) и дополнив их в соответствии с задачами исследования, в **параграфе 1.1.** мы определяем в качестве объекта исследования *учебный текст* как речевое произведение, предъявляемое студентам-инофонам в учебных целях, независимо от того, создано оно для носителя языка или специально для учащихся-иностранцев, объективированное в виде письменного документа, и характеризующееся целенаправленностью, наличием прагматического замысла, установки, относительной завершенностью, связностью и целостностью. Кроме того, количественные лингвистические характеристики текста во многом зависят от его формы. Поэтому обозначается, что объектом исследования выступает конкретная разновидность учебных текстов: прозаический учебный текст монологического характера.

Исходя из проведенного в **параграфе 1.2.** анализа литературы, мы приняли решение оперировать в работе изложенным выше термином *сложность текста,* которая понимается нами как объективная характеристика текста, набор вычисляемых признаков текста, оказывающих влияние на трудность его восприятия, и термином *трудность текста,* понимаемая как более комплексное понятие, значение которого зависит от ряда субъективных, в том числе неязыковых факторов [Кисельников 2015].

Параграф 1.3. освещает проблему отбора текстов в лингводидактических целях. Соответствие текста уровню владения русским языком – его *оптимальная сложность* – признается одним из центральных критериев отбора текстов в обучении языку: 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] и стать причиной неприязни к чтению, а иногда и к изучаемому языку [Акишина, Каган 2002: 44]. Однако необходимо признать, что на практике основным методом определения уровня сложности текста до сих пор зачастую остается индивидуальная экспертная оценка, несущая множество рисков: субъективность, непрозрачность критериев, непоследовательность и несравнимость текстов, оцененных разными экспертами. Это доказывает необходимость дальнейшей активной разработки доступных и достаточно объективных критериев оценки языковой сложности учебных текстов для облегчения процедуры отбора текстов.

Параграф 1.4. посвящен обоснованию выбора в качестве шкалы сложности текстов системы уровней общеевропейских компетенций владения иностранным языком [Common European Framework 2018], включающей 6 основных уровней от A1 до C2. Важнейшими нормоустанавливающими документами, уточняющими связь конкретных лингвистических категорий и уровневой системы CEFR, являются официальный комплекс материалов Российской государственной системы тестирования граждан зарубежных стран по русскому языку. Таблица 1 обобщает информацию из Требований различных уровней к текстам, используемым в учебном процессе [Государственный стандарт 1999a, 1999b, 2001a, 2001b; Требования 2015].

Таблица 1 – Требования к текстам на разных уровнях владения русским языком как иностранным

Уровень CEFR	**Тип текста**	**Тематика текста**	**Объем текста**	**Допустимый процент незнакомой лексики**
A1	Специально составленные или адаптированные сюжетные тексты (на основе лексико-грамматического материала, соответствующего элементарному уровню).	актуальна для бытовой, социально-культурной и учебной сфер общения	250–300	1–2%
A2	Сообщение, повествование, описание, а также тексты смешанного типа. Специально составленные или адаптированные сюжетные тексты, построенные на основе лексико-грамматического материала, соответствующего базовому уровню.	актуальна для сферы повседневного общения, социально-культурной и учебной сфер	600–700	3–4 %
B1	Сообщение, повествование, описание, а также тексты смешанного типа с элементами рассуждения. Тексты аутентичные (допустима минимальная степень адаптации) с учетом лексико-грамматического материала данного уровня.	актуальна для социально-культурной сферы общения	900–1000	5–7%
B2	Тексты описательного и повествовательного характера с элементами рассуждения и эксплицитно выраженной авторской оценкой; художественный текст повествовательного характера.	актуальна для социально-культурной, официально-деловой сфер общения	300–600	до 10%
C1	Полилог, дискуссия с элементами описания и повествования в качестве аргументирующих элементов, содержащий эксплицитно и имплицитно выраженную оценку; интервью, содержащее элементы устной разговорной	актуальна для социально-культурной, официально-деловой сфер общения	400–750	до 10%

	речи; текст информационно — описательного и информационно-регламентирующего характера (законы, постановления, информационные сообщения); художественный текст (рассказ, законченный фрагмент повести, романа и т.д.).			

Параграф 1.5. содержит обзор истории и современных подходов к автоматизации оценки сложности текстов для изучающих иностранный язык. История разработки этой темы тесно связана с исследованием сложности текста в целом и повторяет путь от простейших вычисляемых формул читабельности к поиску более сложных лингвистических признаков и, наконец, созданию предсказательных моделей на основе искусственного интеллекта.

Среди уникальных черт задачи оценки сложности текстов в контексте преподавания иностранного языка на основании анализа научной литературы отмечается наличие понятной единой шкалы уровней сложности материалов, совпадающей с уровнями владения языком по системе CEFR, наличие регламентирующих документов, частично описывающих набор лексических и грамматических тем, доступных на каждом уровне, а также больший вес лексических и грамматических признаков при анализе их вклада в качество работы модели.

В современной науке задача автоматического определения сложности текста чаще всего решается с помощью обучения математической модели и включает в себя три базовых шага: подготовку обучающего набора данных (сбор коллекции образцов текстов с присвоенной им информацией о сложности), автоматическое извлечение их них лингвистических признаков и, наконец, построение на основании этих данных модели машинного обучения.

Анализ существующих сервисов показывает отсутствие инструментов детального анализа русскоязычных учебных текстов. При этом обзор аналогичных продуктов для других языков демонстрирует возможные направления деятельности.

Вторая глава «Разработка и апробация математической модели для автоматического определения сложности текста по шкале CEFR» посвящена описанию процесса создания и проверки качества математической модели автоматического определения сложности текста для занятий РКИ на основе количественных показателей. Работа включала в себя 3 основных этапа, схематично представленных на Рисунке 1.

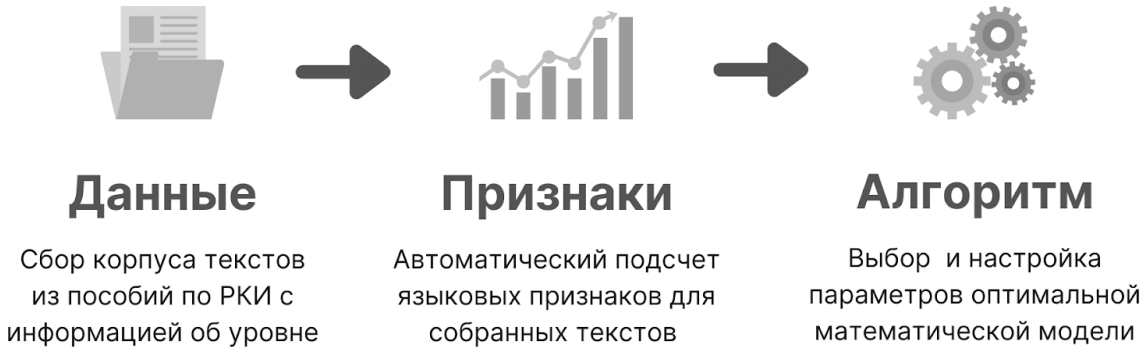

Рисунок 1 – Схема работы по построению модели автоматического определения сложности текста

В параграфе 2.1. представлены результаты работы по созданию эталонного корпуса текстов, на котором модель должна обучаться и тестироваться. В результате работы был собран корпус RuFoLa из 802 текстов из пособий, электронных ресурсов по РКИ и тренировочных тестов ТРКИ. В качестве уровня сложности текстов принималась информация, указанная в аннотации пособия (например, «*настоящий учебный комплекс по русскому языку как иностранному* предназначен *для* взрослых *учащихся и обеспечивает подготовку в объеме I сертификационного уровня*»; «*предназначен для уровня A1 (элементарного)*»). Отбор учебников и текстов электронных ресурсов проходил по следующим критериям:

- изданы/созданы после 2000 года;
- содержат указание на уровень владения русским языком;
- адресованы студентам общего курса русского языка.

Объем корпуса и сбалансированность по уровням владения языком представлены в Таблице 2.

Таблица 2 – Объем корпуса RuFoLa

	A1	**A2**	**B1**	**B2**	**C1**	**Всего**
Количество текстов	220	137	144	158	143	802
Количество слов	29 422	42 529	53 619	89 076	51 444	266 090
Объем словника (кол-во уникальных слов)	1 690	4 352	6 685	11 381	8 715	13 720

Второй этап работы заключался в подсчете лингвистических признаков для каждого текста эталонного корпуса и описан в **параграфе 2.2.** Этот этап работы позволил выделить формальные, рассчитываемые показатели текста, коррелирующие с уровнем его сложности. В результате анализа релевантных работ был сформирован набор из нескольких групп признаков, потенциально способных оказывать влияние на уровень сложности текста и доступных для формализации. Для решения этой задачи был написан код на языке программирования Python. В Таблице 3 приведены основные группы и примеры признаков, полный список количественных характеристик для каждого текста составляет 92 признака.

Таблица 3 – Группы и примеры лингвистических признаков текста для обучения модели

Группа	**Пример признака**
Лексические признаки	средняя длина слова в знаках
	медианная длина слова в знаках
	средняя длина слова в слогах
	медианная длина слова в слогах

Группа	**Пример признака**
Лексические признаки	процент слов длиннее 4 слогов
	лексическое разнообразие (type-token ratio, TTR)
	лексическое разнообразие MLTD (MLTD TTR)
	лексическая плотность (lexical density)
	покрытие текста частотным списком 1, 5 и 10 тысяч самых употребительных слов русского языка
	покрытие текста списками официальной линейки лексических минимумов ТРКИ (от A1 до C1)
	покрытие текста списками альтернативной линейки лексических списков проекта KELLY (от A1 до C2)
	процент слов из списка с семантемами абстрактности
	процент слов с абстрактными суффиксами
Грамматические признаки	процент слов в родительном падеже в тексте
	процент глаголов в финитных формах в тексте
	процент слов в форме 1-го лица в тексте
	процент глаголов в финитных формах в тексте
Синтаксические признаки	средняя длина предложения
	количество противительных союзов на текст
	процент существительных в тексте
	количество сочинительных союзов

Группа	**Пример признака**
	среднее количество пунктуаторов на предложение
	средняя глубина синтаксического дерева
	среднее количество слов, стоящих до главного слова предложения
	покрытие текста списком 500 самых частотных POS-триграмм
Дискурсивные признаки	лексический повтор лемм (lemma overlap)
	количество причинных связок в тексте
	количество временных связок в тексте
	количество аддитивных связок в тексте
	количество негативных связок в тексте
	количество уточняющих связок в тексте
Нарративность текста	отношение количества глаголов на количество существительных в предложении
Описательность текста	количество прилагательных и причастий на предложение

Параграф 2.3 содержит результаты работ по обучению математической модели линейной регрессии на основе полученных признаков. Для построения регрессионных моделей были выбраны модели классической линейной (linear regression) и гребневой регрессии (ridge regression), которая представляет собой одну из техник регуляризации, которые применяются в линейных моделях для решения проблемы зависимости признаков друг от друга и переобучения. Лучший результат показала модель гребневой регрессии, при этом средняя абсолютная ошибка составила 0.7, что говорит о том, что чаще всего модель ошибается в пределах одного уровня.

В **параграфе 2.4** представлены результаты двух этапов оценки применимости полученной математической модели в практике преподавания РКИ. **Параграф 2.4.1** описывает процедуру экспериментальной верификации оценки уровня текста предсказательной моделью путём её соотнесения со временем чтения, качеством ответов на вопросы и субъективными оценками сложности материалов на выборке из 78 иностранных студентов и их преподавателей. Основные результаты оценки представлены в таблице 4.

Таблица 4 – Результаты экспериментальной оценки текстов

Параметр	**Текст №1. Фотограф**	**Текст №2. Блогер**	**Текст №3. Технолог**
Количество слов	179	206	199
Оценка автоматической системы	1.6 (A2)	2.7 (B1)	3.1 (B2 начало)
Средняя оценка текстов преподавателями	1. 6 (A2)	2.1 (B1 начало)	2.8 (B1+)
Средняя скорость чтения текста студентами (слов в минуту)	45	41	33
Процент слов текста, не вошедших в лексический минимум B1	6%	10%	15%
Процент незнакомых слов текста по мнению студентов	0%	1%	3%
Процент незнакомых слов текста по мнению преподавателей	0%	2%	4%
Всего отмечено незнакомых слов у студентов	18	32	47
Всего отмечено незнакомых слов у преподавателей	6	16	26
Процент анкет, где дан корректный ответ на все 4 вопроса	42%	20%	15%
Процент анкет, где дан корректный ответ на 3 вопроса из 4	81%	60%	45%

Этот этап оценки программы, во-первых, показал, что модель верно выстроила текстовый материал по шкале постепенного усложнения на основании целого ряда параметров: скорости чтения, качества ответов на вопросы по тексту, а также анкеты самонаблюдения студентов. Однако обнаружена тенденция модели завышать уровень сложности текстов продвинутых уровней, которая была учтена в дальнейшей настройке системы.

Усредненные результаты оценки текстов преподавателями также совпадают с уровнем, предсказанным моделью, однако оценки одного текста несколькими преподавателями могут достаточно сильно варьироваться, что подтверждает предположение о возможной субъективности суждения об уровне отдельного эксперта-преподавателя.

В ходе второй части опытного исследования, представленного в **параграфе 2.4.2**, было произведено сравнение результатов работы математической модели с выборкой из 100 текстов, размеченных с помощью попарной экспертной оценки и системы рейтингов Эло. Полученный в результате сравнения оценок текстов экспертами и моделью коэффициент корреляции Пирсона 0.86 (при p-value <0.05) позволяет утверждать, что между оценками математической модели и оценками экспертов наблюдается сильная связь. Величина средней абсолютной ошибки составила 0.77, что говорит о том, что в среднем модель ошибается в пределах одного уровня. Анализ ошибок прогноза показал, что модель хорошо справляется с текстами со стандартными характеристиками, сравнимыми с текстами из пособий РКИ, на которых она обучалась. Однако она не способна делать более тонкие выводы, например, учитывать для незнакомых слов шансы на догадку студентов. С другой стороны, при анализе целого текста, а не отрывка, его метрики становятся более стандартизированными, и модель дает верный прогноз.

Таким образом, качество работы и применимость полученной модели были подтверждены экспериментально.

Третья глава «Практическое использование результатов работы предсказательной модели в преподавании РКИ: сервис «Текстометр» содержит описание процесса создания веб-интерфейса сервиса формальной единообразной оценки сложности текста «Текстометр». Сервис рассматривается как площадка для массовой апробации результатов работы полученной математической модели; предлагается его применение широкой

аудиторией специалистов в области РКИ для решения ряда учебно-методических задач.

Параграф 3.1. содержит описание интерфейса сервиса, его основных технических и методических функций. Сервис в техническом плане представляет собой веб-приложение. Для создания части программы, связанной с извлечением лингвистических характеристик текста и построением математической модели определения уровня текста использован язык программирования Python, для создания пользовательского интерфейса использован язык программирования JavaScript. Схема работы сервиса представлена на Рисунке 2.

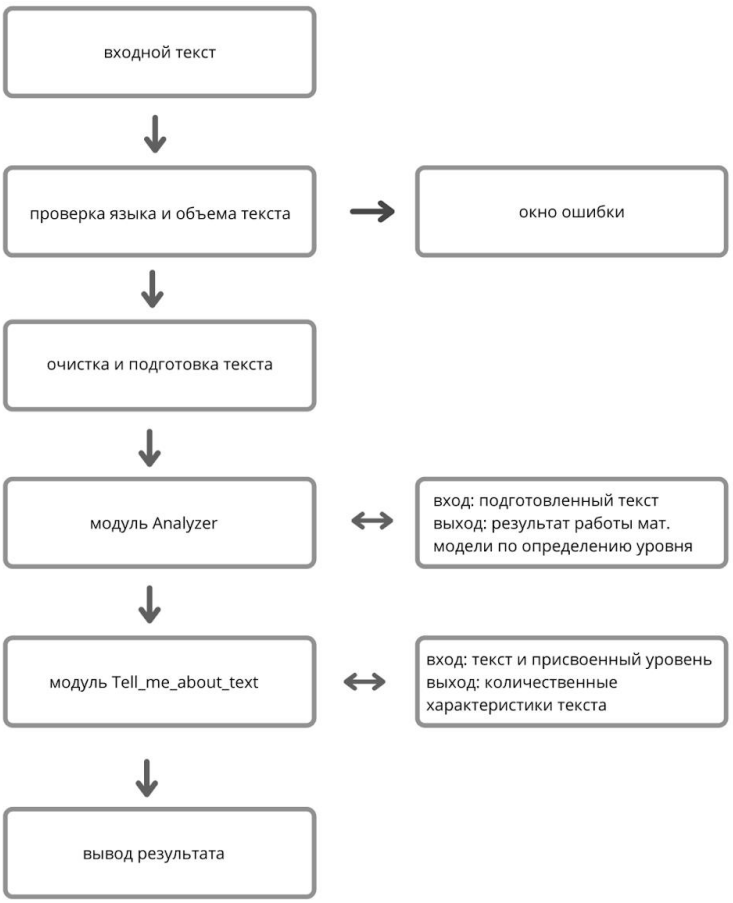

Рисунок 2 – Схема работы модуля по обработке текстовой информации сервиса «Текстометр»

С точки зрения пользователя, интерфейс представляет собой окно ввода текста, куда можно вставить любой текст на русском языке длиной от 5 слов до 10 000 знаков и получить предполагаемое значение уровня сложности текста по методике, описанной во второй главе настоящего исследования, а также статистические параметры введенного текста, релевантные с точки зрения его подготовки к использованию в обучающем процессе (Рисунок 3).

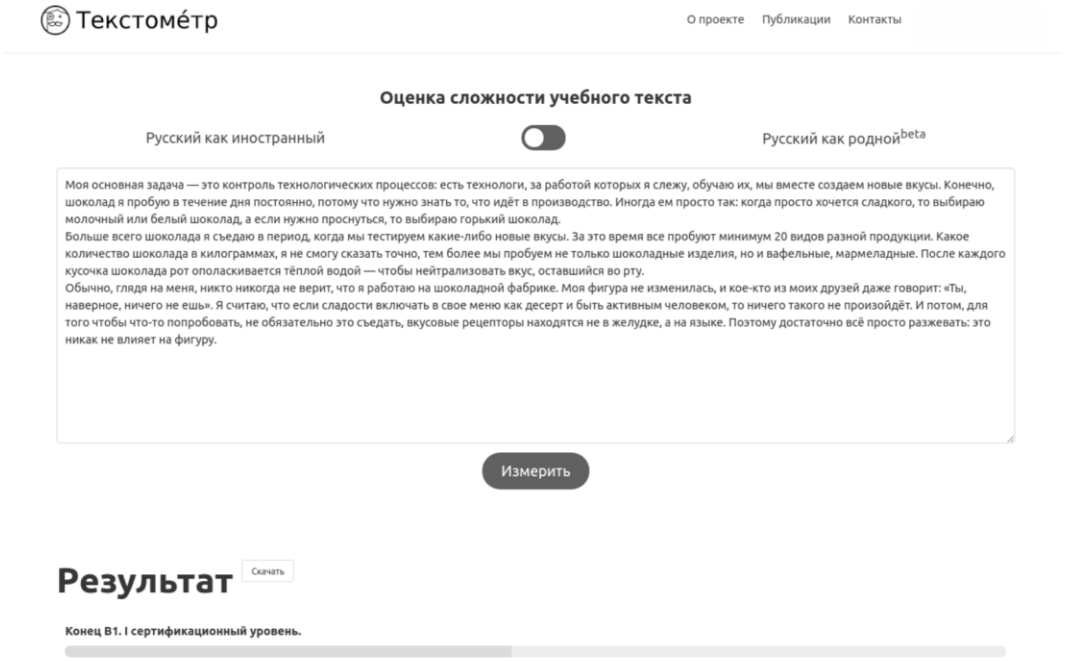

Рисунок 3 – Интерфейс сервиса «Текстометр»

Информация об уровне сложности текста, определенном моделью, для удобства демонстрируется в терминах уровневых систем CEFR и ТРКИ, а также дополнительно визуализируется с помощью цветовой шкалы.

Информация об уровне языковой сложности текста является важнейшей, но далеко не единственной характеристикой текста, влияющей на выбор текста преподавателем. Например, важным критерием может стать информация, 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 РКИ: списки ключевых слов; слов – наилучших кандидатов в словник к данному тексту;

статистика 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.

В зависимости от уровня текста и целей обращения к сервису меняется информация, на которую пользователю сервиса рекомендуется обратить свое внимание. **Параграфы 3.2-3.5** содержат варианты интерпретации количественных характеристик текста, полученных с помощью сервиса, в зависимости от уровня студента и учебных задач. Так, при подготовке учебных текстов начальных уровней большое внимание уделяется проработке списков незнакомой и нечастотной лексики для оптимизации ее объема, а также уровню лексического разнообразия текста. При работе с сервисом для создания контрольно-измерительных материалов наиболее важными критериями должны стать соответствие уровню, а также равномерность уровня сложности материалов среди нескольких тестовых вариантов. Сервис может быть использован для поиска фрагментов аутентичных художественных произведений, оптимальных с точки зрения лексической и синтаксической сложности. Самостоятельная работа студентов с сервисом может заключаться в поиске источников материалов для экстенсивного домашнего чтения, сообразного интересам конкретного учащегося. Наконец, **параграф 3.6.** иллюстрирует дополнительный потенциал использования сервиса вне контекста учебной деятельности, например, для оценки доступности информации инструктивного, правового или рекламного характера для иностранных граждан.

В **параграфе 3.7.** приведена информация об ограничениях работы сервиса, первое и самое важное из которых заключается в том, что он ориентирован на оценку целостного фрагмента прозаического текста. Поэтому при анализе поэтических произведений, лексико-грамматических упражнений или списков слов адекватная оценка уровня сложности алгоритмом не может быть гарантирована. Вторым важным ограничением является ориентация модели в первую очередь на определение уровня сложности нехудожественного текста: информационных, проблемных, сюжетных текстов, репортажей и т.п. Это связано в первую очередь с малой представленностью художественных текстов в обучающей коллекции и принципиально другими критериями сложности и трудности художественных

текстов для работы в иностранной аудитории. Однако и в этом случае представляется полезной практикой проверка незнакомой и ключевой лексики, а также сравнительная характеристика различных фрагментов художественного произведения между собой.

В **Заключении** подводятся итоги исследования и намечаются векторы его дальнейшего развития. В результате проведенного исследования была разработана система автоматической оценки уровня сложности прозаического текста на русском языке по шкале уровней CEFR, проведена экспериментальная работа по оценке её применимости в области преподавания РКИ, а также разработаны методические рекомендации по использованию открытого сервиса на основе разработанной системы.

Применимость полученного алгоритма в практической деятельности преподавателей и изучающих РКИ была подтверждена экспериментально в ходе сравнения оценок текстов моделью с оценками экспертов, временем чтения и качеством понимания текстов студентами и, наконец, мнением самих студентов.

На основе разработанной технологии оценки сложности текстов был создан открытый веб-сервис «Текстометр», а также комплект материалов по работе с ним в зависимости от уровня владения языком и цели обращения.

В качестве основных направлений дальнейшей работы отмечаются расширение эталонной коллекции текстов, которое позволит более детально оценивать уровень текста в зависимости от его типа/жанра (художественный, информационный и т.д.) и вида чтения (изучающее, просмотровое, ознакомительное, просмотровое и поисковое) и работы по уточнению автоматического подсчета незнакомой лексики: возможность учета предшествующего языкового опыта с помощью дополнительных списков интернациональных и общеславянских слов, корректировка лексических списков на основании востребованности лексики в современных пособиях РКИ.

Библиографический список включает используемые и цитируемые в диссертации научные труды (145 наименований) и список используемых учебников и учебных пособий (35 наименований). В **Приложениях** приводятся материалы для проведения экспериментальной верификации качества модели (анкеты для студентов и преподавателей), обобщенные данные по результатам проведения эксперимента, пример результата анализа

текста в разработанном сервисе «Текстометр» и свидетельство о государственной регистрации программы для ЭВМ.

Основное содержание диссертации отражено в следующих публикациях автора (общим объемом 4,18 п. л.):

Статьи в ведущих рецензируемых изданиях, рекомендованных ВАК Министерства науки и высшего образования Российской Федерации:

1. **Лапошина А. Н.** Корпус текстов учебников РКИ как инструмент анализа учебных материалов / А. Н. Лапошина // Русский язык за рубежом. – 2020. – № 6. – С. 22–28. (0, 65 п.л.)
2. **Лапошина А. Н.** Текстометр: онлайн-инструмент определения уровня сложности текста по русскому языку как иностранному / А. Н. Лапошина, М. Ю. Лебедева // Русистика. – 2021. – Т. 19, № 3. – С. 331–345. DOI 10.22363/2618-8163-2021-19-3-331-345. (0,9 п.л.)
3. **Лапошина А. Н.** Что значит «не входит в лексический минимум?» Подсчет процента незнакомой лексики в тексте РКИ с учетом доступных словообразовательных моделей / А. Н. Лапошина // Преподаватель XXI век. – 2021. – №4. Часть 2. – С. 473–483. DOI: 10.31862/2073-9613-2021-4-473-483 (0,78 п.л.)

Публикации в научных изданиях, индексируемых в Scopus:

4. **Laposhina A. N.** Automated Text Readability Assessment For Russian Second Language Learners / A.N. Laposhina, T.S. Veselovskaya, M.Y. Lebedeva, O.F. Kupreshchenko // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialogue 2018». Moscow, Russia. – 2018, – Issue 17. – P. 396–406. (0,62 п.л.)

Публикации в других научных изданиях:

5. **Лапошина А. Н.** Автоматическое определение сложности текста по РКИ / А. Н. Лапошина // Международная научно-практическая интернет-конференция «Актуальные вопросы описания и преподавания русского языка как иностранного/неродного»: Сборник материалов, Москва, 27 ноября – 01

декабря 2017 года. Москва: Государственный институт русского языка им. А.С. Пушкина, 2018. – С. 573–579. (0,3 п.л.)

6. **Лапошина А. Н.** Опыт экспериментального исследования сложности текстов по РКИ / А. Н. Лапошина // Динамика языковых и культурных процессов в современной России. Материалы VI Конгресса РОПРЯЛ. Уфа, 11–14 октября 2018 года. – Уфа, 2018. – С. 1544–1549. (0,5 п.л.)

7. **Лапошина А. Н.** Анализ релевантных признаков для автоматического определения сложности русского текста как иностранного / А. Н. Лапошина // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог». Москва, 31 мая — 3 июня 2017 г. – Москва, 2017. – С. 1–7. (0,43 п.л.)

Программы, зарегистрированные в Федеральной службе по интеллектуальной собственности:

1. Свидетельство о государственной регистрации программы для ЭВМ №2021661785. Текстометр / Лапошина А.Н. (RU), Лапошин А.А. (RU), Лебедева М.Ю. (RU); правообладатель ФГБОУ ВО Государственный институт русского языка им. А.С. Пушкина (RU). Заявка № 2021660920; дата поступления – 09.07.2021; дата государственной регистрации в Реестре программ для ЭВМ – 15.07.2021.