Федеральное государственное бюджетное образовательное учреждение высшего образования
«Государственный институт русского языка им. А.С. Пушкина»

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)

Диссертация на соискание ученой степени
кандидата педагогических наук

Научный руководитель:
кандидат филологических наук,
Лебедева Мария Юрьевна

Москва – 2023

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ 5-12

ГЛАВА 1. ТЕОРЕТИЧЕСКИЕ ОСНОВЫ АВТОМАТИЗАЦИИ ПРОЦЕССА АНАЛИЗА СЛОЖНОСТИ ТЕКСТА В ПРАКТИКЕ ПРЕПОДАВАНИЯ РКИ 13-54

1.1. Учебный текст в преподавании РКИ: определение понятия 13-14

1.2. Сложность и трудность как свойства текста 14-18

1.3. Отбор текстов как лингводидактическая проблема 18-21

1.4. Сложность текста в системе уровней владения русским языком как иностранным 21-29

1.5. Методы оценки языковой сложности текста 29-52

1.5.1. История развития методов оценки сложности текста в российской и зарубежной науке 29-36

1.5.2. Анализ сложности текста РКИ как задача машинного обучения 36-48

1.5.3. Существующие ресурсы и сервисы по анализу сложности текста 48-52

Выводы по главе 1 53-54

ГЛАВА 2. РАЗРАБОТКА И АПРОБАЦИЯ МАТЕМАТИЧЕСКОЙ МОДЕЛИ ДЛЯ АВТОМАТИЧЕСКОГО ОПРЕДЕЛЕНИЯ СЛОЖНОСТИ ТЕКСТА ПО ШКАЛЕ CEFR 55-100

2.1. Сбор и описание корпуса текстов пособий по РКИ 56-59

2.2. Сбор лингвистических признаков для обучения модели 60-79

2.2.1. Лексические признаки 63-70

2.2.2. Грамматические признаки 70-73

2.2.3. Синтаксические признаки 74-76

2.2.4. Дискурсивные признаки 76-77

2.2.5. Общий обзор полученных признаков 77-79

2.3. Построение и оценка качества предсказательной модели 79-83

2.4. Апробация машинной модели по определению сложности текста 83-98

2.4.1. Экспериментальная верификация качества работы модели 83-92

2.4.2. Сравнение с экспертной оценкой сложности текстов 92-98

Выводы по главе 2 98-100

ГЛАВА 3. ПРАКТИЧЕСКОЕ ИСПОЛЬЗОВАНИЕ РЕЗУЛЬТАТОВ РАБОТЫ ПРЕДСКАЗАТЕЛЬНОЙ МОДЕЛИ В ПРЕПОДАВАНИИ РКИ: СЕРВИС «ТЕКСТОМЕТР» 101-132

3.1. Интерфейс и основные возможности сервиса «Текстометр» 101-110

3.2. Варианты интерпретации результатов работы сервиса при конструировании учебных текстов (уровень А1) 110-115

3.3. Варианты интерпретации результатов работы сервиса при создании и анализе контрольно-измерительных материалов (уровень А2) 116-118

3.4. Возможности самостоятельной работы студентов с сервисом «Текстометр» на примере нехудожественных текстов для экстенсивного чтения (уровни В1 и В2) 118-120

3.5. Варианты интерпретации результатов работы сервиса при подборе фрагментов аутентичных художественных текстов (уровень В2) 121-125

3.6. Возможности работы с сервисом для специалистов по адаптации и интеграции иностранных граждан 126-129

3.7. Ограничения работы сервиса 129-131

Выводы по главе 3 131-132

ЗАКЛЮЧЕНИЕ 133-136

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ 137-152

СПИСОК УЧЕБНИКОВ И УЧЕБНЫХ ПОСОБИЙ 153-156

ПРИЛОЖЕНИЕ А. Материалы эксперимента по проверке качества работы модели: тексты и анкеты 157-164

ПРИЛОЖЕНИЕ Б. Обобщенные ответы участников эксперимента 165-171

ПРИЛОЖЕНИЕ В. Пример результата анализа текста в разработанном сервисе «Текстометр».....172-188

ПРИЛОЖЕНИЕ Г. Свидетельство о государственной регистрации программы для ЭВМ..... 189

ВВЕДЕНИЕ

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 – значить иметь оптимальное соотношение знакомой и новой информации. Однако современные исследования показывают, что представления об уровне сложности текста у разных преподавателей могут значительно отличаться друг друга. Это значительно затрудняет сравнимость сложности текстов, оцененных разными экспертами, а также маркировку учебных пособий по уровню. Таким образом, необходимость 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.

Теоретико-методологическую 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 иностранной аудитории (В.Г. Сибирцевой, Н.В. Карпова, R. Reynolds, S. Sharoff и др.).

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. При этом **объектом исследования** выступает сложность учебного текста в аспекте обучения РКИ как его объективное измеряемое свойство, выражаемое набором лингвистических характеристик, а **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.

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»;

– формировании комплекта рекомендаций по работе с сервисом для широкого круга специалистов (преподавателей, методистов, авторов пособий, представителям 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 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 как иностранному Государственного института русского языка им. А.С. Пушкина и их преподаватели (78 студентов и 7 преподавателей);

– российские и зарубежные преподаватели РКИ (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 международной филологической конференции (СПбГУ, 26 марта 2019), международном форуме «РКИ-перезагрузка 2021: уроки пандемии» (Москва, 18 – 19 июня 2021);

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

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

По теме диссертации опубликовано 8 работ, в том числе 5 опубликовано в рецензируемых научных изданиях, включенных в перечень ВАК при Минобрнауки России (из них 3 – в научных рецензируемых журналах, рекомендованных ВАК РФ, 1 – в изданиях, приравненных к перечню ВАК при Минобрнауки России (индексируемых в МБД «Scopus») и 1 свидетельство о государственной регистрации программы для ЭВМ в Федеральной службе по интеллектуальной собственности, приравниваемое к перечню ВАК при Минобрнауки России.

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.

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.

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.

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. – 2018. – Issue 17. – P. 396-406.

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.

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.

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.

По теме диссертации получено свидетельство о государственной регистрации программы для ЭВМ 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.

ГЛАВА 1. ТЕОРЕТИЧЕСКИЕ ОСНОВЫ АВТОМАТИЗАЦИИ ПРОЦЕССА АНАЛИЗА СЛОЖНОСТИ ТЕКСТА В ПРАКТИКЕ ПРЕПОДАВАНИЯ РКИ

1.1. Учебный текст в преподавании РКИ: определение понятия

В методике преподавания РКИ термин *учебный текст* имеет множество толкований, отличающихся по ширине взгляда на объект, вслед за широким или узким пониманием лингвистического термина *текст*. Текст в более широком смысле понимается как «продукт (реализация, конечный результат) речевой деятельности» [Акишина, Каган 2002: 34]. В рамках данного определения текстом может считаться и одно слово (надпись «Вход» на дверях), и несколько взаимосвязанных предложений, начиная от диалогового единства и заканчивая литературным произведением – законченным связным целым [Тёрёчик 2012: 13]. Однако для поставленной задачи исследования более релевантным является понимание текста в узком смысле – как продукта речевой определенной деятельности, характеризующегося целенаправленностью и наличием прагматической замысла, установкой, относительной завершенностью, связностью и целостностью [Щукин 2003: 148].

Кроме того, для предстоящей задачи автоматической обработки учебных текстов важным является представленность текста в письменной форме: это понимание понятия соотносится с определением И.Р. Гальперина в работе «Текст как объект лингвистического исследования», в которой автор определяет текст как «письменное сообщение, объективированное в виде письменного документа, состоящее из ряда высказываний, объединённых разными типами лексической, грамматической и логической связи, имеющее определённый моральный характер, прагматическую установку и соответственно литературно обработанное» [Гальперин 2006: 20].

Наконец, термин *учебный текст* с точки зрения концепции его создания в узком смысле может означать текст, специально созданный для инофонов в учебных целях: противоположностью учебного текста в данном случае выступает неучебный,

аутентичный текст [Щукин 2003]. С другой стороны, ряд исследователей отмечает, что любой текст, предлагаемый иностранным студентам в процессе обучения, в том числе текст, изначально созданный для носителей языка, «все равно переходит в статус учебных, ибо приобретает определенную функцию в учебном процессе» [Григоренко 1991: 76]. Поскольку уровень лингвистической сложности текста является его объективной характеристикой, не зависящей от его изначальной цели создания, мы в дальнейшей работе будем оперировать широким пониманием учебного текста как любого текстового произведения, предъявляемого студенту-инофону в учебных целях.

Таким образом, базируясь на предложенных в методической литературе определениях учебного текста [Азимов, Щукин 2009: 303; Щукин 2003: 148; Гальперин 2006: 20; Тёрёчик 2012: 13] и дополнив их в соответствии с особенностями задачи предстоящего исследования, определим 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 прагматической замысла, установкой, относительной завершенностью, связностью и целостностью.

Кроме того, лингвистические характеристики текста во многом зависят от его формы. В связи с этим обозначим, что объектом нашего исследования будет выступать конкретная разновидность учебных текстов: прозаический учебный текст монологического характера.

1.2. Сложность и трудность как свойства текста

В исследованиях, посвященных проблеме отбора и оценки доступности текстового материала, используется ряд близких по значению терминов: *сложность, трудность, понятность, читабельность*. Представляется полезным определить понятия, которыми мы будем оперировать далее в работе. Поскольку

лингвистические характеристики текста очень сильно зависят от его формы, обозначим, что в дальнейшем под текстом мы будем подразумевать только прозаические письменные произведения недиалогического характера.

Так, один из первых системных исследователей данного вопроса на русскоязычном материале Я.А. Микк определяет *сложность текста* как его объективное свойство, не зависящее от человека, читающего этот текст [Микк 1980]. Сложность текста определяется при помощи анализа этого текста, например, по проценту незнакомых слов, по длине предложений, по сложности логической структуры и другим компонентам сложности.

Трудность текста, согласно Я.А. Микку, является «свойством текста препятствовать пониманию» и зависит не только от его сложности, но и от подготовленности читателя [Микк 1980: 11]. Один и тот же текст может быть легким для подготовленного читателя и трудным для неподготовленного. Трудность текста устанавливается по результатам понимания данного текста, т. е. экспериментально.

При этом *понимание текста* интерпретируется как осознание связей между элементами текста и объектами реального мира, которые обозначают эти элементы текста, а *понятность текста* – как «свойство текста содействовать пониманию» [Микк 1980: 11].

В.С. Цетлин, ссылаясь на И.Я. Лернера, указывает, что "следует различать *сложность* учебного материала как его объективную характеристику и *трудность* как субъективный фактор подготовленности учащихся к преодолению сложности" [Цетлин 1980: 30].

М.Н. Вятютнев дает ценные с позиций методики преподавания русского языка как иностранного характеристики понятий *сложность* и *трудность* как лингвистическим и психолингвистическим категориям. Он пишет: "Легкость/трудность – это понятие психолингвистическое, субъективное, поскольку обусловливается интеллектуальной деятельностью человека, простота/сложность – понятие лингвистическое, объективное, присущее только языку" [Вятютнев 1978:

с. 46]. Теоретически разграничивая эти понятия, М.Н. Вятютнев приходит к выводу, что как только учащиеся приступают к изучению языка, между легкостью/трудностью и простотою/сложностью устанавливается своя зависимость, свои изменчивые связи.

Аналогичное понимание оппозиции сложности и трудности как категорий, в разной степени поддающихся объективному измерению, предлагается в работе Ю.А. Томиной, посвященной разработке критериев трудности текстов РКИ. *Сложность* исследовательница понимает как категорию, поддающуюся количественной оценке и в силу этого могущей быть объективированной. *Трудность*, вслед за К.М. Ушаковым, она считает частично объективной характеристикой, зависящей не только от сложности изучаемого материала, но и от состояния учащихся, объема и качества их навыков, предшествующего опыта и других индивидуальных особенностей [Томина 1985: 47].

Уточненное толкование оппозиции сложности vs. трудности текста предлагается в работе А.С. Кисельникова, согласно которому при *анализе сложности текста* наряду с широким спектром количественных параметров предполагается учет и качественных параметров, анализ лексических единиц текста (многозначные слова, национально-маркированные лексические единицы, словарь лексического минимума и частотные словари) и его «абстрактных» единиц (формулы, графики, схемы и др.), анализ связей на уровне предложения и текста в целом (анафора и антецедент, синонимия, референция и др.). *Трудность текста* определяется на основе анализа параметров сложности текста применительно к конкретной целевой аудитории (или отдельному читателю), т.к. текст одной сложности может иметь различную трудность [Кисельников 2015].

Таким образом, большая часть вышеприведенных исследований, посвященным поиску формальных вычисляемых лингвистических признаков, оказывающих влияние на восприятие текста, оперируют термином *сложности.* Этот же термин

широко используется в современных исследованиях автоматизации процесса определения уровня сложности текста [Оборнева 2006; Карпов 2015; Криони 2008].

Параллельно в работах данной тематики существуют термины *читабельность* или *удобочитаемость* как варианты перевода английского термина readability. По мнению А.С. Кисельникова, отличительной чертой определения *читабельности текста* является учет исключительно количественных параметров: количество слов в тексте, количество предложений в тексте, средняя длина предложения, среднее количество слогов в слове, среднее количество знаков в слове и ряд других, что относит нас к многочисленным формулам читабельности. Однако анализ современных исследований показывает, что этот термин может использоваться как в значении, полностью синонимичном понятию сложности текста [Мацковский 1973, Heilman 2007], так и в значении удобства, скорости чтения [Rello et al. 2013]. Ю.А. Томина связывает такое недифференцированное понимание термина *читабельность* с тем, что он сочетает в себе оценку объективных параметров текста, т.е. его сложность, осуществляемую, как правило, исследователями этой проблемы с учетом субъективных факторов подготовленности определенного контингента к преодолению сложности, т.е. с учетом трудности [Томина 1985: 46].

Кроме того, в некоторых исследованиях понятие *читабельности* трактуется значительно шире, включая в себя не только лингвистическую, но и типографскую сторону доступности текста для чтения: шрифт, размер, расположение текста и т.п. [Collins-Thompson 2014].

Исходя из проведенного анализа литературы, мы приняли решение оперировать в работе описанным выше термином *сложность текста,* который 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. При этом мы будем избегать

термина *читабельность* как не имеющего однозначного толкования и пользоваться им только в составе устойчивого словосочетания *формулы читабельности*.

1.3. Отбор текстов как лингводидактическая проблема

Поскольку доминирующий в современной лингводидактике коммуникативный подход к обучению подразумевает «уподобление процесса обучения процессу реальной коммуникации» [Щукин 2003: 167], именно текст, рассматриваемый как основная коммуникативная единица, которой человек пользуется в речевой деятельности [Фоломкина 1987; Каменская 1990; Тёрёчик 2012], считается «исходной и конечной единицей обучения» [Акишина, Каган 2002: 35], и целью, и средством обучения языку. Закономерно, что текстоцентрическая концепция обучения русскому языку признается исследователями как одна из наиболее перспективных в рамках реализации коммуникативно-деятельностного подхода [Басова и др. 2014].

Учебный текст признается главной единицей представления учебного материала: он, с одной стороны, передаёт студентам социокультурную информацию о стране изучаемого языка, а с другой стороны, является и моделью для учащегося, по которой он строит собственное высказывание-информацию о реалиях своего социума или участвует в общении на новом для себя языке [Шустикова, Кулакова 2011: 7].

Центральная роль текста в обучении РКИ обуславливает особое значение, которое придается отбору текстового материала: «успешное достижение целей обучения во многом зависит от того, насколько оптимально подобран учебный материал» [Кулибина 2015: 13]. Чрезвычайно важным и принципиальным для любого учебника признается проблема отбора коллекции текстов [Шустикова, Кулакова 2011: 7].

Соответствие текста уровню владения русским языком, его оптимальная сложность, является одним из центральных критериев отбора текстов: 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 соответствие текста уровню не является единственным критерием. Остановимся на релевантных нашему исследованию аспектах отбора текстового материала. Во-первых, методистами отмечается необходимость включения в обучающий процесс **аутентичных** текстов. Их объем и соотношение со специально сконструированными или адаптированными учебными текстами может разниться в зависимости от уровня владения учащимися русским языком, однако ориентация исключительно на тексты из учебника и невключение «живого потока текстов» (информации газет, радио, телевидения) отмечается специалистами как методическая ошибка преподавателя [Акишина, Каган 2002: 44]. Не менее важным аспектом отмечается **релевантность тематики и содержания текста** целям изучения языка конкретного учащегося. Так, специалисты отмечают влияние чтения на запоминание языкового материала и формирование словарного запаса [Крючкова, Мощинская 2009; Чеснокова 2015]: часто встречающиеся при чтении лексические единицы переходят в активное использование в речи. Следовательно, в том числе и от подбора текстов зависит результат обучения, выраженный в овладении лексикой, отвечающей коммуникативным нуждам учащегося. Наконец, одним из требований к тексту для предъявления в учебных целях является его **увлекательность**. Влияние степени интересности информации текста конкретному учащемуся на качество понимания и готовности преодолевать языковые трудности отмечается большим количеством исследователей [Вятютнев 1984; Alexander 1996]. Напротив, чтение же без заинтересованности приводит либо к утрате части информации, либо к искажению понимания, даже полному непониманию [Горелов, Седов 2001: 44].

Принимая во внимание три вышеназванных аспекта отбора текстов, можно сделать вывод, что поскольку сферы интересов, цели и задачи студентов, изучающих

русский язык как иностранный, могут быть самыми разными, перед современным преподавателем РКИ стоит непростая и трудозатратная задача отбора и подготовки большого количества материалов, а также регулярное пополнение и обновление личной текстотеки.

При этом в современном информационном пространстве вряд ли возникает проблема недостатка самих текстовых материалов: новостные сайты, интернет-издания, блоги, тексты социальных сетей являются источниками огромного количества аутентичных текстов. Проблема скорее состоит в отборе материалов, подходящих студентам по уровню языковой сложности. 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]. Однако, как отмечает Ю.А. Томина в диссертации 1985 года, не существует единой четкой процедуры установления уровня сложности текста: «преподаватели и авторы учебников нередко устанавливают меру языковой трудности текстов интуитивно, "на глазок", ограничиваются подсчетом в них знакомых и незнакомых лексических единиц, следуя невольно упрощенной логике – чем больше незнакомых слов в тексте, тем труднее текст – или руководствуются общей протяженностью текста и/или длиной включенных в него предложений. В результате в действующих учебниках русского языка для иностранцев иногда начальные тексты отличаются от срединных и завершающих именно размерами, а нередко оказываются стабильными по размерам на протяжении всего учебника» [Томина 1985].

С одной стороны, за прошедшие годы произошло множество изменений в методике РКИ в сторону объективизации этого процесса: описаны и регулярно обновляются Общеевропейские компетенции владения иностранным языком [Common European Framework 2018], появился официальный комплекс материалов Российской государственной системы тестирования граждан зарубежных стран по

русскому языку [Государственный стандарт 1999, 2001a, 2001b; Требования 2015]. Эти документы играют важную роль в описании уровневой системы РКИ и установлении связи конкретных лингвистических категорий и степенью сложности текста. Однако, с другой стороны, необходимо признать, что на практике основным методом определения уровня доступности текста до сих пор зачастую остается описанная Ю.А. Томиной индивидуальная экспертная оценка, несущая множество рисков: субъективности, непрозрачности критериев, непоследовательности и несравнимости текстов, оцененных разными экспертами.

Объективная единая оценка уровня сложности текста также является необходимым первым шагом в разработке систем автоматизированной адаптации или упрощения текстов: такая работа предполагает знание конкретных грамматических, лексических, синтаксических и др. признаков текста, способных оказывать влияние на его сложность [Сибирцева, Карпов 2014].

Всё это доказывает необходимость дальнейшей активной разработки надежных и достаточно объективных критериев языковой сложности текстов для облегчения процедуры отбора текстов по критерию языковой сложности.

1.4. Сложность текста в системе уровней владения русским языком как иностранным

Поскольку в дальнейших параграфах работы нам предстоит ранжировать тексты по сложности, необходимым этапом является выбор шкалы сложности текста. В отличие от сложности текста для носителей языка, где нет единого решения для выбора единиц измерения сложности: школьный класс [Chall and Dale 1995], возраст, абстрактные условные единицы [Orphee De Clercq 2017; Pitler, Nenkova 2008] – в случае с иностранными языками подавляющее большинство исследователей делает выбор в пользу Общеевропейской шкалы уровней владения иностранным языком (Common European Framework of Reference, CEFR) [Reynolds 2016; Karpov, Baranova, Vitugin 2014; Schwarm and Ostendorf 2005].

CEFR – общеевропейская система оценки уровня владения иностранным языком – устанавливает единые стандарты, которые применяются для определения языковой компетенции во всем мире и служит для взаимного признания квалификаций, полученных в разных системах образования. В актуальной редакции спецификаций 2018 года [Common European Framework 2018] шкала CEFR состоит из 7 уровней: от pre-A1 до C2 (см. Таблицу 1). Уровни владения русским языком как иностранным в системе РКИ соотносятся с уровнями CEFR [Лексический минимум 2012], соответствия также приведены в Таблице 1. К её плюсам стоит отнести наличие подробных спецификаций и международное признание, которое делает возможным сравнивать исследования для разных языков. Мы не будем исключением и принимаем систему уровней CEFR в качестве шкалы измерения сложности текста.

Таблица 1 – Соответствие уровней систем CEFR и ТРКИ

pre-A1	A1	A2	B1	B2	C1	C2
–	ТЭУ	ТБУ	ТРКИ-1	ТРКИ-2	ТРКИ-3	ТРКИ-4
–	Элементарный уровень	Базовый уровень	Первый уровень	Второй уровень	Третий уровень	Четвертый уровень
–	Basic user		Independent user		Proficient user	

Для того, чтобы оценка компетенции могла применяться для любого языка, не привязываясь к конкретным реалиям, Ассоциацией ALTE (The Association of Language Testers of Europe) был разработан и описан набор ключевых навыков и умений, соответствующих каждому уровню. В частности, разделы спецификаций, посвященные чтению как виду речевой деятельности, предлагают логику распределения комплекса навыков по шкале CEFR, представленную в Таблице 2 [Common European Framework 2018].

Таблица 2 – Спецификация навыков, связанных с чтением, по разным уровням CEFR

Уровень	**Расшифровка компетенций**
Pre-A1	Могу распознавать знакомые слова, сопровождаемые изображениями, например, простое иллюстрированное меню или иллюстрированная книга с использованием знакомой лексики.
A1	Могу понять знакомые имена, слова, а также базовые фразы.
A2	Могу понять короткие простые тексты. Я могу найти конкретную, **легко предсказуемую** информацию в простых текстах повседневного общения: в рекламах, проспектах, меню, расписаниях. Я понимаю простые письма личного характера.
B1	Понимаю тексты, построенные **на частотном языковом материале** повседневного и профессионального общения. Я понимаю описания событий, чувств, намерений в письмах личного характера.
B2	Понимаю статьи и сообщения по современной проблематике, авторы которых занимают особую позицию или высказывают особую точку зрения. Я понимаю **современную художественную прозу**.
C1	Понимаю **большие сложные нехудожественные и художественные тексты**, их стилистические особенности. Я понимаю также **специальные статьи и технические инструкции** большого объема, даже если они не касаются сферы моей деятельности.
C2	**Свободно понимаю все типы текстов**, включая тексты абстрактного характера, сложные в композиционном или языковом отношении: инструкции, специальные статьи и художественные произведения.

В числе параметров текста в данном описании можно заметить указания на жанры (например, книга комиксов, художественная проза, инструкции и т.п.), частотность слов, стили речи, однако нельзя говорить о каком-то конкретном наборе

характеристик текста, присущем тому или иному уровню. Это связано с тем, что система Общеевропейских компетенций разрабатывалась как универсальная, не учитывающая специфику конкретного языка. В дальнейшем на основе общей рамки CEFR разрабатываются уровневые системы для конкретных языков.

Отметим также конвенциональный характер предложенной шкалы, используя метафору авторов: «Подобно цветам настоящей радуги, овладение языком является континуумом. Тем не менее, как и в случае с радугой, несмотря на размытость границ между цветами, мы склонны видеть некоторые цвета больше, чем другие. Тем не менее, для удобства коммуникации мы упрощаем систему и фокусируемся на основных цветах» [Common European Framework 2018: 34, перевод наш].

Еще одной особенностью, которую отмечают сами авторы спецификаций, является возможность при необходимости более дробного деления на подуровни для того, чтобы более точно обозначить место ученика или предназначенного ему учебного контента внутри того или иного уровня. Например, Рисунок 1 демонстрирует систему подуровней на материале основной шкалы CEFR, используемую в сети институтов Polyskills[1].

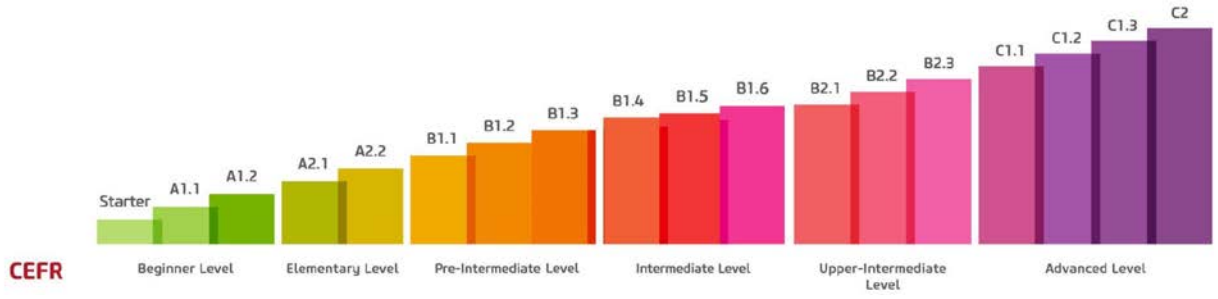


Рисунок 1 – Дробное представление уровней CEFR для удобства использования в практике преподавания

Таким образом, неоспоримыми плюсами выбора шкалы CEFR в качестве шкалы сложности текстов можно назвать её универсальность, наличие указания на

[1] Polyskills – International House Frankfurt: https://ihfrankfurt.de/cefr-levels/

уровень в большинстве пособий РКИ, возможность сравнивать результаты исследований для других языков, свободу от таких субъективных категорий, как класс/возраст/количество лет обучения, соотнесенность с жизненными ситуациями (в описании уровней содержится информация о связи того или иного уровня владения с возможностями и нуждами реальной жизни: поступление в российский вуз, разрешение на работу преподавателем русского языка и т.п.) и возможность дальнейшего разбиения на подуровни. Из возможных ограничений отметим отсутствие описания конкретных лингвистических характеристик текстов по уровням CEFR, что мы надеемся компенсировать с помощью информации из Требований к владению русским языком в системе ТРКИ.

Если описание системы уровней CEFR намеренно является абстрактным и применимым к любому языку, официальный комплекс материалов Российской государственной системы тестирования граждан зарубежных стран по русскому языку содержит набор более конкретных и подлежащих измерению параметров текста в контексте его предъявления в иностранной аудитории. Данный комплекс состоит из набора Требований по русскому языку как иностранному (далее – Требования), Лексических минимумов – списков лексики, обязательной к усвоению на каждом уровне, а также системы Типовых тестов по русскому языку как иностранному.

Система Требований содержит информацию о минимальных обязательных требованиях, определяющих цели и содержание обучения на каждом конкретном уровне. Данные материалы очень ценны в подготовке автоматизации оценки сложности текста, поскольку в них зафиксированы формальные измеряемые критерии текстов (количество слов, процент незнакомой лексики), их тематика, уровень знания морфологии, грамматики и синтаксиса на этом уровне. Таблица 3 обобщает информацию из Требований различных уровней к текстам, используемым в учебном процессе [Государственный стандарт 1999a, 1999b, 2001a, 2001b; Глазунова и др. 2017].

Таблица 3 – Требования к текстам на разных уровнях владения русским языком как иностранным в системе ТРКИ

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	актуальна для социально-культурной,	300–600	до 10%

	рассуждения и эксплицитно выраженной авторской оценкой; **художественный текст** повествовательного характера.	официально-деловой сфер общения		
C1	полилог, дискуссия с элементами описания и повествования в качестве аргументирующих элементов, содержащий эксплицитно и имплицитно выраженную оценку; интервью, содержащее элементы устной разговорной речи; текст информационно – описательного и информационно-регламентирующего характера (законы, постановления, информационные сообщения); художественный текст (рассказ, законченный фрагмент повести, романа и т.д.)	актуальна для социально-культурной, официально-деловой сфер общения	400–750	до 10%

Несмотря на некоторую противоречивость представленных данных (например, требования к объему текста уменьшаются от B1 к B2: возможно, это вызвано разным темпом обновления Требований), становится видна общая канва постепенного усложнения текстов и их тематики. Переход от специально составленных текстов к аутентичным происходит на третьем уровне (B1). Тогда же добавляется текст с элементами рассуждения как более сложный тип по сравнению с повествованием.

Меняется и тематика текстов: от бытовой и учебной, связанной с необходимостью быстрого освоения лексики для повседневных нужд (так называемый survival Russian), к появлению на 4 уровне официально-деловой сферы. Также мы видим, что полноценное чтение художественной литературы больших объемов (роман, повесть) появляется лишь к 5 уровню. Тогда же впервые выдвигаются требования к пониманию различных государственных законов и документов. Еще одним формальным критерием, описанным в Требованиях, является скорость чтения текста: она зависит не только от уровня владения языком, но и от типа чтения.

Вторым важнейшим источником информации, регламентирующим уровни владения иностранным языком, является **лексический минимум**, т.е. список слов, знание которых необходимо для успешной сдачи сертификационного тестирования. Кроме того, эти документы взаимосвязаны: в Требованиях к текстам фигурирует фраза «с учетом лексико-грамматического материала данного уровня». В настоящий момент линейка лексических минимумов разработана для 5 уровней CEFR, от A1 до C1. Объем лексических минимумов растет вместе с уровнем владения (см. Таблица 4).

Таблица 4 – Объем лексического минимума в зависимости от уровня владения русским языком

Уровень	**A1**	**A2**	**B1**	**B2**	**C1**
Количество словарных единиц	730	1 300	2 565	5 500	11 000

Отбор лексических единиц для включения в минимум производится по следующим общим критериям:

1. стилистическая немаркированность;

2. способность слова входить в различные словосочетания;

3. семантическая ценность (способность слова обозначать часто встречающиеся предметы и явления);

4. высокая словообразовательная способность слова;

5. частотность (при этом учитывались показатели частотности по частотным словарям, по использованию в учебниках РКИ, а также «тематическая» частотность) [Андрюшина 2011].

Именно лексические минимумы позволяют рассчитывать процент незнакомой лексики в тексте, что является одной из важнейших метрик, указывающих на сложность текста.

Наконец, третьим и последним типом документов, входящих в комплекс материалов Российской системы тестирования граждан зарубежных стран по русскому языку, являются **типовые тесты по русскому языку как иностранному**. В них содержатся примеры заданий и текстов, которые необходимо будет выполнить для сдачи экзамена на определенный уровень. Для нашего исследования такие тесты также являются ценным источником пополнения корпуса для обучения математической модели, т.к. именно эти тексты являются наиболее авторитетными источниками информации о соответствии текстового материала уровню ТРКИ.

1.5. Методы оценки языковой сложности текста

1.5.1. История развития методов оценки сложности текста в российской и зарубежной науке

История развития научной проблемы ранжирования текстов по сложности на основании их лингвистических характеристик насчитывает более чем сто лет [DuBay 2006] и, по мнению Томаса Франсуа, условно делится на три этапа в зависимости от превалирующих методик подсчетов. Так, первые попытки изучения сложности текста связаны с экспериментальным установлением связи сложности текста с простыми для вычисления характеристиками текста, такими как длина слов и предложений, т.н. формулы читабельности. Второй этап характеризуется поиском более глубоких лингвистических признаков сложности, внимание к синтаксическим и дискурсивным

параметрам текста. Наконец, превалирующий в современной науке взгляд на измерение сложности текста как на стандартную задачу машинного обучения исследователь называет третьей парадигмой, AI readability [Francois, Fairon 2012]. Следуя предложенной Т. Франсуа логике, рассмотрим наиболее важные исследования предшествующих парадигм оценки сложности текста в зарубежной и отечественной науке.

Активно тема оценки уровня сложности текста начинает разрабатываться с 20 годов прошлого века в США в аспекте школьного образования и подбора подходящих материалов для чтения. Ранние работы посвящены в основном поиску простых признаков сложности текста, основанных на легко вычисляемых метриках текста, таких как средняя длина слова, средняя длина предложения, количество длинных слов и т.д. Среди наиболее популярных можно назвать формулы Флеша-Кинсайда [Kincaid et al. 1975], индекс Gunning Fog [Gunning 1968], SMOG (Simple Measure of Gobbledygook). В качестве примера приведем формулу удобочитаемости Флеша (Flesch Reading Ease, FRE), расчет которой производится с помощью двух параметров: средней длине предложения в словах и средним числом слогов в слове:

(1) FRE = 206.835 – (1.015 × средняя длина предложения) – (84.6 × среднее число слогов)

Расчет сложности по этой формуле строится на достаточно очевидном предположении, что более короткие предложения и слова в тексте говорят о его простоте. При этом константные коэффициенты в формуле необходимы, чтобы приблизить значение формулы к какой-либо интерпретируемой шкале. Так, формула Флеша оценивает текст по 100-балльной шкале, где 0 – очень сложный текст, а 100 – напротив, чрезвычайно простой текст.

Подобные формулы часто критикуются в последующих исследованиях. Так, среди причин отмечается, во-первых, внимание к одномерным метрикам сложности на уровне слов и предложений и игнорирование тем самым важности более глубокого уровня понимания текста, дискурсивного. Во-вторых, при подобных измерениях не

учитываются такие важнейшие характеристики как связность текста и жанр. В-третьих, исследователи отмечают неинформативность подобных мер для преподавателей, поскольку они не содержат в себе информации о том, что конкретно делает тот или иной текст сложным или легким [Graesser et al. 2014]. Кроме того, понимая принцип работы формулы, преподаватель невольно или специально может её «сломать», создавая короткие предложения, которые ещё уменьшают связность текста и затрудняют его понимание.

Однако, несмотря на частую критику, формулы читабельности завоевали популярность благодаря своей простоте и по сей день широко используются как в контексте оценки школьных учебных текстов [Оборнева 2006], учебников для иностранных учащихся [Browne 2011; Kismarianto 2019], так и проверки на доступность государственных документов (http://readability.io), SEO оптимизации текстов (https://readable.com) и мн др. Достаточно упомянуть, что расчет сложности текста по формуле Флеша включен в редактор Microsoft Word.

Ранние работы по объективной оценке сложности текстов применительно к преподаванию иностранного языка также тесно связаны с разработкой формул читабельности и представлены именами Дж. Царп, С. Шпаулдинг, Ж. Витковская, Г. Гринюк, Э. Стебдянко, однако они выполнены на материале французского, испанского, английского и немецкого языков [Томина 1985].

Тема оценки сложности русскоязычных текстов с помощью статистических параметров текста появляется в отечественном научном ландшафте позднее, в 70-е года XX века. Так, В 1973 г. М.С. Мацковский предложил формулу читабельности на материале русского языка [Мацковский 1973]. Для выведения формулы было выбрано 50 текстов из СМИ, различающихся как по тематике, так и по типу изданий. Каждый текст оценивался по трудности 60 учениками седьмых классов одной из московских школ. Среди исследуемых параметров текста были как факторы, связанные со структурой предложения (средняя длина предложения), так и факторы, связанные со словарным составом (число слогов на 100 слов текста). В результате

применения стандартных процедур вывода уравнения регрессии была получена следующая формула, основанная на значениях средней длины предложения в словах, проценте слов больше 3 слогов и наборе постоянных корректирующих коэффициентов. Однако, к сожалению, в научной литературе нет информации, свидетельствующей о практическом применении формулы Мацковского для исследования параметров удобочитаемости широкого класса текстов на русском языке [Оборнева, 2006], исходя из которой можно было бы оценить качество работы данной формулы.

Отдельную группу современных исследований сложности текстов на русском языке представляют работы по адаптации уже названных выше классических формул читабельности под русскоязычный материал, поскольку коэффициенты этих формул нуждаются в корректировке под каждый конкретных язык [Оборнева 2006; Дружкин 2016]. Так, И.В. Оборнева предлагает вариант адаптации индексов Флеша, Флеша-Кинсайда и Фога на основании сравнения средней длины слова и количества многосложных слов на базе "Толкового словаря русского языка" Ожегова и "Англо-русского словаря" В.К. Мюллера: на основании полученных данных исследовательница делает вывод, что слова в тексте на английском языке содержат слогов меньше чем русские в среднем в 0,71 раза. Количество слов в предложениях на английском языке при этом больше примерно в 1,25 раза. Помимо скорректированных формул читабельности, И.А. Оборнева проводит данные их апробации на материале литературных произведений из программы средней школы.

Среди работ второй парадигмы, характеризующейся поиском более глубоких лингвистических признаков сложности, отдельно стоит отметить серию трудов Я.А. Микка, посвященных сложности и трудности текстовых материалов в контексте оптимизации школьного учебника [Микк 1975, 1981]. Исследователь произвел детальную разработку терминологии области и ввел термин "понятность текста», который определил как «свойство текста содействовать пониманию». В качестве измеряемых параметров сложности текста он выделяет: 1) количество слов в

предложении; 2) «знакомость» слов; 3) абстрактность слов (соотношение абстрактных и конкретных слов в тексте), а также предлагает методику экспериментальной проверки качества понимания текста учащимися [Микк 1981].

Одной из первых и важнейших работой по оценке сложности текста в контексте преподавания РКИ является диссертация Ю.А. Томиной, посвященная объективной языковой трудности текстов в зависимости от их типа. Исследовательница предлагает в качестве параметров не только классические метрики текста, но и такие показатели, как знакомость слова (процент слов, входящих в лексический минимум подготовительных факультетов СССР); абстрактность слова; сложность синтаксической структуры предложения. В частности, результаты расчетов показывают, что наибольшую сложность с точки зрения частного лексического показателя имеют тексты типа доказательство и описание. Наименьшая величина среднего числа незнакомых слов, и, следовательно, меньшая сложность имеет место в текстах типа повествование и рассуждение. Также в этой работе были поставлены вопросы методики проведения экспериментальной проверки трудности текста с помощью методов компетентных судей, анализа ответов на вопросы по тексту (открытые и закрытые); суждением информантов о трудности текста [Томина 1985].

С ростом возможностей автоматической обработки больших коллекций текстов на естественном языке (автоматическое морфологическая и синтаксическая разметка), а также производительности компьютеров, работы по определению сложности текста оказываются вариантом классической задачи машинного обучения, основанной на обучении предсказательной модели на тренировочном корпусе текстов и наборе признаков. В работах Франсуа этот этап изучения вопросов сложности текста называется AI readability (читабельность на основе искусственного интеллекта, перевод мой) или третьей парадигмой.

Среди особенностей этого этапа изучения сложности можно отметить возможность использования значительно большего количества признаков, возможность легко переобучиться на обновленной коллекции данных, а благодаря

готовым библиотекам для обработки данных не приходится говорить о сложности вычислений. Ключевыми вопросами таких исследований при этом становятся подбор оптимальной математической модели, а также поиск адекватного задаче обучающего корпуса.

Цель всех подобных систем состоит в том, чтобы научиться определять сложность текста так, как это сделал бы человек. *Научиться* здесь значит проанализировать предложенные образцы текстов и их лингвистические характеристики, и на основании такого же набора лингвистических признаков нового, незнакомого системе текста, сделать вывод о его уровне сложности. *Человек* же здесь значит носитель экспертного знания, что подчеркивает крайнюю важность унификации и формализации представлений об уровне сложности между экспертами и специалистами области, а также создания обучающей коллекции данных, размеченных по этим установленных принципам. В противном случае алгоритм не сможет обучиться корректно.

Примерами работ, где сложность текста для иностранных учащихся измеряется с помощью статистических моделей, могут служить исследования на материале немецкого [Hancke et al. 2012], французского [Francois, Fairon 2012], шведского [Pilan et al. 2016], португальского [Curto et al. 2016] и других языков. Ключевыми вопросами работ, выполненных в русле машинного подхода к определению сложности текста, становятся поиск оптимального источника текстовых образцов, извлечение и оценка влияния разных типов признаков на сложность и выбор оптимальной математической модели.

Рассмотрим подробнее существующие решения для русского языка как иностранного. Одной из пионерских здесь можно назвать работу С. Шарова по построению модели автоматического определения сложности текстов на материале коллекции неадаптированных статей на одну и ту же тему в обычных новостных изданиях и на сайте BBC, которые, по экспертной оценке, являются более простыми [Sharoff et al. 2008]. В качестве признаков были предложены как классические

признаки, основанные на длинах (средняя длина слов и предложений текста), так и частеречные признаки (среднее число полнозначных глаголов на предложение, среднее число глаголов в пассивной форме на предложение и др.) Далее к полученным признакам был применен метод главных компонент (principal component analysis), в результате чего стало возможным обозначить два главных "измерения" сложности исследуемых текстов, которые условно можно обозначить как грамматический и лексический векторы сложности. При этом в большинстве случаев простые тексты имеют ожидаемо более низкие показатели по обоим измерениям, однако есть интересные примеры, когда грамматически более простой текст может оказаться сложнее лексически.

Первой работой на материале русского языка, использующей в качестве обучающих данных специализированные тексты для иностранных учащихся, становится исследование Н. В. Карпова, Ю. Н. Барановой и Ф. М. Витюгина [Karpov et al. 2014]. Исследование описывает серию опытов по определению как сложности текстов, так и отдельных предложений. На материале сравнительно небольшого корпуса текстов и 25 базовых текстовых характеристик, включая длину документа, длину предложения, длину слова, сложность лексики и наличие каждой части речи, было исследовано качество работы различных типов моделей машинного обучения для автоматической классификации сложности русского текста, а также сложности одного предложения. В частности, описано применение таких методов, как логистическая регрессия, метод опорных векторов и построение деревьев классификации для бинарной классификации (A1-C2, A2-C2 и B1-C2) и достижении точности, близкой к 100%. Однако следует отметить, что при этом не сообщалось о результатах с менее очевидными и наиболее практически применимыми парами текстов, таких как A1-A2, A2-B1 и B1-B2.

Отдельно стоит упомянуть масштабную работу Р. Рейнолдса. Автор проводит серию экспериментов по обучению модели классификации русских текстов на материале коллекции более 4 000 текстов из различных источников, размеченных по

шкале CEFR и нескольких группах признаков, включая лексические, морфологические и синтаксические [Reynolds 2016]. Наилучшее значение F-меры в мультиклассовой классификации равняется 0,67 при точности 0,92. Бинарные классификации смежных уровней (A1-A2, A2-B1 и т.д.) показали значения F-меры от 0.8 до 0.89, что говорит о высоком уровне качества полученной модели.

Среди новейших подходов к оценке читабельности в рамках задач машинного обучения обозначим работы по созданию систем оценки «без учителя» (unsupervised approach) [Martinc 2021]. Такие системы больше не нуждаются в обучающей размеченной коллекции текстов и извлечении набора признаков: определение сложности основано на работе нейронных систем, обученных на объемных коллекциях данных. Подобный взгляд на оценку сложности текстов меняет и тип исследователя: от него требуется все больше навыков программирования и обработки данных, и практически не требуется собственно лингвистических знаний. С одной стороны, эта тенденция абстрагирования от собственно лингвистической информации является закономерной и параллельно происходит и в других прикладных задачах: машинном переводе, автоматическом упрощении текстов и др. С другой стороны, Collins-Thompson отмечает, что важной отличительной чертой задачи автоматической оценки сложности текста, особенно применительно к учебным задачам, является интерпретируемость результатов работы модели, возможность посмотреть и продемонстрировать пользователю, какие конкретно факторы позволили модели сделать вывод о простоте или сложности текста, а это невозможно сделать в случае построения систем без извлечения осмысленных лингвистических признаков [Collins-Thompson 2014].

1.5.2. Анализ сложности текста РКИ как задача машинного обучения

Задача автоматического определения сложности текста, как и любая другая задача построения системы машинного обучения с учителем, включает в себя три базовых шага [Francois, Fairon 2012; Reynolds 2016]: подготовку обучающего набора

данных, автоматическое извлечение их них признаков и, наконец, построение на основании этих данных модели машинного обучения. Первым шагом является создание эталонной коллекции текстов, на которой модель будет обучаться, т.н. золотого стандарта. В частном случае оценки сложности текста эта коллекция должна содержать тексты и информацию о их сложности. Для получения такой коллекции могут использоваться как экспертная оценка, так и автоматический сбор текстов из интернет-источников, или сбор текстов из учебников или книг для чтения.

Самым очевидным и продуктивным решением здесь является обращение к текстам учебных пособий для иностранных учащихся. В частности, одной из первых работ, использовавшей корпус учеников в качестве тренировочных данных по сложности, является работа по созданию автоматической системы оценки французских текстов для иностранных учащихся, для которой исследователи отобрали 2160 текстов из разделов, посвященных чтению, из 28 учебников французского как иностранного, изданных после 2001 г. [Francois, Fairon 2012]. В этой же работе ставится проблема субъективности суждения об уровне пособия в зависимости от коллектива авторов или издания, которая может привести к негомогенности коллекции с точки зрения сложности текстов внутри одного уровня CEFR.

Исследования сложности текстов для изучающих шведский язык как иностранный проводились на материале части корпуса учебников шведского языка COCTAILL. Этот корпус состоит из двенадцати книг (от четырех разных издательств) и снабжен подробной метаразметкой, включая уровень учебника по шкале CEFR, тип упражнения, жанр и стиль текстов. Для задачи определения сложности текстов были использован подкорпус из 867 текстов (отрывков для чтения) из этого корпуса. Кроме того, исследователи изучали возможности определения уровня сложности отдельных предложений, для этого из того же корпуса были использованы 1874 предложений из языковых примеров, иллюстрирующих использование определенных

грамматических шаблонов или лексических элементов в учебниках иностранного языка [Pilan et al. 2015].

Для создания подобной системы для изучающих китайский язык как иностранный, CRIE-CFL был собран корпус из 28 серий учебников китайского как иностранного, опубликованных в 23 странах, где используется стандарт CEFR. Корпус насчитывает 1578 текстов. Чтобы избежать проблемы неоднозначности в маркировке уровня учебников, исследователи дополнительно использовали экспертную оценку собранных текстов опытными преподавателями [Sung et al. 2015].

Обратимся к примерам тренировочных данных для моделей машинного обучения по оценке сложности русских текстов для иностранных учащихся. Так, в одной из первых работ, посвященных автоматической оценке сложности на материале корпуса эталонных текстов, в качестве обучающей коллекции для английского языка выступили пары параллельных статей стандартной Википедии и их упрощенных версий, написанных по специальным принципам (Simple English Wikipedia). Для русского же языка, за неимением подобного материала, исследование базировалось на материале сравнения неадаптированных статей на одну и ту же тему в обычных новостных изданиях и на сайте BBC, которые, по экспертной оценке, являются более простыми. Несмотря на очевидные плюсы такого подхода, заключающиеся в доступности такого рода данных, к минусам стоит отнести ограничение всего двумя уровнями сложности текста (простой-сложный), риск необъективности данных о сложности конкретных текстов [Sharoff et al. 2008].

Первой работой на материале русского языка, обратившейся в качестве тренировочных данных к текстам, оцененным по шкале CEFR стала работа группы исследователей, занимающихся автоматическим определением сложности предложений на русском языке [Karpov et al. 2014]. В качестве материала для исследования они использовали 169 текстов из Текстотеки ЦМО МГУ[2] для

[2] https://www.irlc.msu.ru/irlc_projects/texts

иллюстрации текстов элементарного (52 текста), базового (57 текстов) и первого сертификационного уровней (60 текстов). В качестве примеров текстов уровня C2 были использованы 50 аутентичных новостных текстов. Однако объем такого корпуса для задачи определение сложности целого текста представляется недостаточным. Кроме того, отнесение всех аутентичных новостных текстов к уровню C2 рискует не отражать их реальный уровень сложности по шкале CEFR, ведь такие тексты в зависимости от темы и используемых конструкций могут широко использоваться и на менее продвинутых уровнях.

Наконец, самый объемный корпус русскоязычных текстов, размеченных по уровням CEFR, представляет в своей работе Р. Рейнольдс [Reynolds 2016]. Общий объем корпуса составляет 4 689 текстов. Самый большой сегмент корпуса составляет 3 481 текстов портала для изучения языков LinqQ[3]: на этом сайте желающие могут сами добавлять понравившиеся им тексты и обмениваться ими с другими студентами. Материалом для корпуса в этом случае служит оценка уровня текста, которую поставил человек, загрузивший материал. Помимо этого, исследователь использует коллекцию Н.В. Карпова, описанную выше, адаптированные тексты из «Библиотеки Златоуста»[4], типовых тестов ТРКИ и серии книг для адаптированного чтения.

Таким образом, базируясь на анализе опыта зарубежных и российских ученых, мы принимаем решение использовать в качестве источника обучающих данных для модели машинного обучения корпус текстов из пособий по РКИ, Текстотеки ЦМО МГУ, типовых тестов ТРКИ и Библиотеки Златоуста, используя в качестве значения уровня сложности текста уровень, указанный авторами в предисловии или методическом сопровождении пособия. Несмотря на то, что идея краудсорсинга очень перспективна и нам близка, на данном этапе мы отказались от использования данных краудсорсинговой разметки текстов пользователями портала LinqQ, во-первых, из-за неполного соответствия со шкалой уровней CEFR, а во-вторых, из-за

[3] https://www.lingq.com
[4] https://www.zlat.spb.ru

риска субъективности суждений о сложности текстов (уровень присваивает не профессиональный преподаватель, а сам студент).

Среди плюсов использования текстов из пособий в качестве обучающих данных системы по определению сложности текста отметим понятность и прозрачность методики сбора коллекции; наличие информации об уровне текста по шкале CEFR; связь с информацией Требований и Лексических минимумов.

К минусам такого подхода можно отнести малый выбор пособий по общему курсу русского языка для высоких уровней (B2, C1); отсутствие образцов текстов уровня C2; размытость указания на уровень; трудоемкий процесс перевода текстов в цифровой формат.

Вторым шагом построения математической модели оценки сложности после создания коллекции образцов текстов является определение набора признаков, которые будут использованы в качестве характеристик текста для предсказания его уровня сложности. В самом широком смысле читабельность текста можно обозначить как сумму всех элементов текста, влияющих на понимание темы, скорость чтения и уровень интереса к прочитанному. К. Коллинз-Томпсон в обзоре исследований, посвященных оценке читабельности текстов [Collins-Thompson 2014] предлагает иерархию компонентов трудности текстов, представленную на Рисунке 2.

Рисунок 2 – Возможные компоненты сложности текста, предложенные К. Коллинз-Томпсоном

Эта пирамида описывает компоненты читабельности (здесь мы намеренно используем более широкий термин *читабельность*, которым оперирует Коллинз-Томпсон) от низших к высшим уровням организации текста. Очевидно, что они имеют прямую связь с признаками текста, на которых необходимо тренировать предсказательную модель. Так, первой ступенью пирамиды являются признаки текста, связанные с *разборчивостью, визуальной читабельностью*: текст, шрифт, интервалы. Некоторые исследователи добавляют в этот тип присутствие иллюстраций и графиков. Второй пласт занимают *лексико-семантические* (знакомость и частотность слов) и *морфологические* признаки (частотность/редкость морфологических форм). Далее следуют *синтаксические* признаки текста (сложность предложений, особенности грамматической структуры), *дискурсивные* (связность предложений между собой). Под прагматическим уровнем К. Коллинз-Томпсон понимает различную игру слов, идиомы, культурный контекст, необходимый для понимания текста. Сюда же он относит такое явление, как сарказм. И, наконец, самый высокий пласт признаков представляет собой некую *персональную информацию* о читателе – его интерес к теме, мотивация к чтению текста, фоновые знания и опыт.

Подавляющее большинство исследований читабельности и сложности текстов используют в своих расчетах признаки из середины пирамиды: чаще всего исследователи обращаются к лексическому и синтаксическому уровням, реже к дискурсивному. Визуальная доступность и экстралингвистические факторы трудности текста обычно являются объектами отдельных исследований. Фокус нашего исследования также будут составлять эти группы признаков.

Лексические признаки являются одной из самых обширных и широко используемых групп в задаче определения сложности текста. Внутри этой группы можно выделить несколько подгрупп. Первой разумно назвать самые простейшие признаки лексической сложности текста, основанные на длинах слов в знаках и слогах. Выбор этого параметра связан с предположением, что наличие длинных слов в тексте увеличивает вероятность того, что перед нами сложный текст, и наоборот.

Этот параметр широко используется как в ранних формулах читабельности, так и подавляющим большинством современных исследователей лингвистических признаков сложности текста. Методика подсчета при этом может варьироваться от средней длины слова к медианной [Селегей 2015], от подсчета длины слова по знакам [Coleman, Liau 1975] или слогам [Kincaid et al. 1975].

Отдельной большой группой признаков являются признаки, связанные со знакомостью лексики читателю. Так, начиная с ранних формул читабельности появляется стремление оценить степень сложности лексики текста с помощью подсчетов её вхождения в специально составленные списки слов (‘vocabulary-based formulas’). Среди наиболее известных формул такого типа назовем формулу Дэйла-Чалл, где используется средняя длина предложения и процент "сложных" слов, который авторы предлагают считать с помощью списка 3 000 слов английского языка, знакомых 80 процентам английских школьников 4 класса [Chall, Dale 1995] и формулу оценки сложности текстов Spache, использующую 1 000 самых простых слов английского языка, знакомых большинству учеников начальной школы [Spache 1953].

В более поздних работах одним из решений оценки знакомости слов становятся подсчеты по вхождениям лексики текста в частотные списки, составленные на материале больших корпусов текстов на этом языке. Связь частотности слова с когнитивной нагрузкой, требуемой для его обработки называется в литературе *эффектом частотности слова* (the word frequency effect) [Monsell et al. 1989] и используется в большом количестве прикладных задач, одной из которых является ранжирование текстов по сложности или читабельности [Francois, Fairon 2012; Rello et al. 2013; Reynolds 2016].

Еще одной важнейшей группой признаков, уникальной для задачи определения сложности материалов для иностранной аудитории, является вхождение лексики текста в специализированные учебные списки слов – **лексические минимумы**. Подобные списки лексики содержат наиболее ценные с лингводидактической точки

зрения лексические единицы и мотивированы в первую очередь прагматическими целями, такими как предоставление преподавателям и авторам пособий рекомендаций, какие слова стоит включать в материалы на данном этапе владения языком. При этом стоить отметить, что при создании подобных списков частотность слова выступает лишь одним из критериев оценки методической ценности слова, таким образом, эта информация не дублирует предыдущую группу признаков. Включение процента покрытия текста специальными списками для изучающих язык как иностранный является распространенной практикой в подобных исследованиях [Francois, Fairon 2012; Reynolds 2016; Pilan et al. 2016], а в исследовании для текстов РКИ отмечаются как признаки, показавшие наивысшую информативность [Karpov et al. 2014].

Следующую группу составляют признаки, связанные с лексическим разнообразием текста. **Коэффициент лексического разнообразия** (англ. lexical diversity) показывает степень воспроизводимости, повторяемости лексики в тексте, что, по данным исследований, может оказывать влияние на процесс чтения [Bowers 2000; Francois, Fairon 2012]. Одним из базовых способов подсчета этой меры является мера TTR, которая вычисляется как отношение уникальных лексем текста к общему числу лексем текста. Однако поскольку эта мера очень чувствительна к размеру текста, существуют более сложные подсчеты лексического разнообразия, например, мера MTLD (measure of textual lexical diversity) [McCarthy 2005]. Для расчета этой меры текст делится на сегменты равной длины, а затем для каждого сегмента рассчитывается значение TTR. Размер сегментов увеличивается, пока значение TTR не достигнет 0.72. Тогда применяется расчет значения MLTD, как отношения длины текста в словах к получившемуся числу сегментов.

Значение коэффициента лексической плотности текста (lexical density) также может быть связано с когнитивной нагрузкой, затрачиваемой на восприятие текста, а также влиять на запоминаемость сообщения [To, Le 2013]. Он рассчитывается как доля слов знаменательных частей речи в тексте. Более лексически

плотными, таким образом, признаются тексты, в которых используется больше знаменательной и меньше служебной лексики.

Появление автоматических морфологических анализаторов текста сделало возможным включать грамматическую информацию в набор признаков текста и оценить роль этих признаков в оценке сложности текста. Так, например, М. Хайлмен с коллегами сравнивает вклад группы **грамматических признаков** в качество определения сложности текстов с позиции носителей языка и иностранных учащихся. Эксперимент показал, что добавление грамматических признаков принесло бóльший прирост точности в коллекции текстов как иностранных – 22% против 7% как родных [Heilman et al. 2007].

Роль морфологических признаков в определении сложности текстов РКИ обсуждается в работе Р. Рейнолдса [Reynolds 2016]. Исследователь выдвигает гипотезу, что роль морфологии в формировании сложности текста недооценена из-за того, что огромная часть исследований проводилось для английского языка с относительно бедной морфологией. Действительно, анализируя лучшие признаки, автор сравнивает результаты индивидуального вклада признака и вклада признака в наборе признаков. В полученных результатах в наборе из 32 лучших признаков 14 оказываются из группы морфологических признаков 14. Основываясь на этих результатах, автор делает предположение, что хотя морфологические признаки не так информативны, как другие, но они предоставляют уникальную информацию о тексте, чем оказываются ценны при создании набора признаков.

При этом может отличаться способ подсчетов такого типа признаков от учета присутствия конкретных единичных категорий – причастий и деепричастий, форм пассива, форм пассива или императива – до использования абсолютной или относительной частотности всех возможных при автоматическом анализе грамматических тэгов, отражающих, в частности, для русского языка, число, род, падеж, время, лицо, вид, краткость, степень сравнения слов текста [Дружкин 2016; Reynolds 2016].

Семантические признаки, связанные с уровнем абстрактности слов текста, стабильно появляются в работах, посвященных сложности текста на русском языке. При этом для расчетов этих показателей чаще всего используется список суффиксов, выражающих в русском языке абстрактность: *-ость, ение, -ание*, *-изм, -изна* и мн.др. [Микк 1981; Томина 1985; Криони и др. 2008].

Безусловно, важную роль в понимании текста играет его структурная сложность, выраженная рядом **синтаксических признаков**. Для этой группы также наблюдается разная степень детализированности признаков. Так, самый простой способ оценки сложности структурной – средняя или медианная длина предложения – опирается на предположение, что длинные конструкции чаще оказываются более сложными. Этот признак используется, наряду со средней длиной слова, как практически во всех традиционных формулах, так и современных исследованиях. Более сложной группой признаков является подсчет количества частеречных тэгов в предложении. Например, количество союзов, предлогов, знаменательных частей речи в предложении может говорить о сложности его синтаксической структуры и используется в качестве лингвистических признаков для обучения моделей [Sharoff 2008; Francois, Fairon 2012].

Наконец, автоматизированные системы построения синтаксических разборов предложения в виде деревьев зависимостей (синтаксические парсеры) делают возможными и более детализированные синтаксические признаки текста, например, такие как количество глагольных и именных групп на предложение, средняя глубина синтаксического дерева, среднее количество придаточных предложений в сложноподчиненных конструкциях [Schwarm, Ostendorf 2005; Pitler, Nenkova 2008]. Вклад синтаксических признаков в оценку сложности текстов на русском языке доказывается на материале сразу ряда исследований [Криони и др. 2008; Ivanov et al. 2018]. С другой стороны, в единственной работе на материале текстов РКИ, использующей синтаксические признаки, основанные на работе синтаксических парсеров, вклад данной группы признаков оказывается немногим больше, чем

использование простейших метрик, основанных на длине предложения [Reynolds 2016].

Поднимаясь на уровень выше, мы сталкиваемся с необходимостью формального представления **дискурсивных признаков**, связности текста с помощью количественных параметров. Так, одним из базовых признаков связности текста, вклад которых отмечается в исследованиях, является лексический повтор [Rashotte, Torgesen 1985]. Технически параметр высчитывается с помощью поиска в двух соседних предложениях повторяющихся знаменательных слов, или только существительных. Данный параметр представляет особенный интерес в задаче определения сложности текстов для иноязычной аудитории, поскольку исследования письменной речи людей, изучающих английский язык как иностранный, говорят о том, что частый лексический повтор говорит о низком уровне владения языком, тогда как на более высоких уровнях используются такие приемы как референция и более сложные лексические связки [Ferris 1994].

Coh-Metrix – вычислительный инструмент для оценки сложности текста, разработанный на основе теории психолингвистических и когнитивных моделей чтения [Crossley et al. 2011], помимо лексического повтора использует такие параметры связности текста как плотность употребления лексических связок (союзов, частиц, вводных слов) определенных групп значений, а также наличие логических операторов (и, или, если–то), употребление которых доказало связь с повышением абстрактности текста и более высокой нагрузкой на оперативную память при восприятии текста [Costerman, Fayol 1997].

Помимо собственно связности, еще одним из ключевых показателей текста, оказывающих влияние на простоту или трудность его восприятия, данная группа исследователей называет нарративность, повествовательность: «Повествовательный текст рассказывает историю с персонажами, событиями, местами и предметами, которые знакомы читателю. Повествование тесно связано с повседневной устной речью» [Graesser et al. 2014, перевод наш].

Вклад описанных групп признаков в определение сложности текста может изучаться по-отдельности [Reynolds 2016], однако чаще всего используется набор из нескольких описанных групп признаков. Так, основу Coh-Metrix L2 Readability Index – индекса для автоматической оценки сложности англоязычных текстов для неносителей составляет лексический повтор знаменательных частей речи, синтаксическое сходство смежных предложений и лексические признаки на основе частотности [Crossley et al. 2008; Crossley et al. 2011]. Предложенный индекс применяется к анализу уровней сложности учебных материалов по английскому языку как иностранному [Cárcamo Morales 2020; Kiselnikov et al. 2020]. Однако примеров адаптации индекса к русскоязычным материалам нами обнаружено не было.

Наконец, третий шаг в построении математической модели оценки сложности текстов заключается в выборе оптимальной для данной задачи модели машинного обучения и построении предсказательной модели на основании эталонного корпуса текстов и набора лингвистических характеристик этого текста и оценке качества полученной модели. В зависимости от выбора шкалы измерения задача оценки сложности текста может решаться как в рамках классификации (тогда результатом работы модели будет предполагаемый класс, т.е. один из возможных уровней), так и регрессии (тогда результатом работы модели является любое десятичное число на заданной шкале). При этом на вход модель получает набор признаков текста, который преобразуется в вектор, и «правильное» значение уровня сложности текста-носителя этого набора признаков. В задачах классификации исследователи часто прибегают к SVM и random forest модели [Karpov et al. 2014; Francois, Fairon 2012; Reynolds 2016]. В качестве регрессионных моделей используются как классическая линейная регрессия [McCullagh 1980], так и регрессия гауссовского процесса, SVM регрессия и др. [Kate et al. 2010].

На вопрос о наиболее оптимальном выборе модели нет единственно правильного ответа. Так, сравнительное исследование влияния различных моделей

классификации и регрессии на качество прогноза уровня читабельности текста показало, что наилучшие результаты демонстрируют алгоритмы с порядковой шкалой измерения [Heilman et al. 2008]. Другими словами, лингвистические параметры постепенно изменяются в зависимости от уровня текста, однако эта связь не является линейной. Оптимальным решением в таком случае К. Коллинз-Томпсон называет порядковую регрессию или регрессию гауссовского процесса [Collins-Thompson 2014].

Выбор модели обуславливает и методику проверки качества полученной модели. Так, в случае построения регрессионной модели, наиболее релевантным и информативным способом оценки её качества являются среднеквадратическая ошибка (RMSE). Она представляет собой среднее значение квадратов всех расстояний предсказания от правильного ответа: чем ближе к нулю значение метрики, тем лучше признается модель. Наряду со среднеквадратической ошибкой часто используется мера средней абсолютной ошибки (MAE), которая рассчитывается как среднее расстояние модуля предсказания от правильного ответа [Collins-Thompson 2014; Sowmya, Meurers 2012].

При этом оптимальным решением поиска «честных» данных для тестирования модели является разделение обучающего корпуса текстов случайным образом на сегмент для тренировки модели (обычно используется 80% коллекции) и сегмент для тестирования (20% соответственно). Для повышения устойчивости модели к конкретным получившимся сегментам обычно используется механизм кросс-валидации, при котором результатом работы модели является среднее число от нескольких циклов разделения данных на тренировочные и обучающие, тренировки и тестирования полученной модели.

1.5.3. Существующие ресурсы и сервисы по анализу сложности текста

Сервисы по автоматическому анализу сложности текстов призваны освободить преподавателей от рутинных и трудозатратных действий по отбору текстов и

оставить тем самым возможность сфокусироваться на более творческих педагогических задачах. Кроме того, они способствуют стандартизации и объективизации процесса оценки текста, что представляется особенно важным моментом в ситуации коллективного авторства учебника или при создании контрольно-измерительных материалов. Однако примеров таких сервисов, особенно подходящих для русскоязычных текстовых материалов, немного. Перечислим несколько удачных на наш взгляд примеров сервисов для английского языка, а затем проведем обзор существующих ресурсов для анализа сложности русскоязычных текстов.

Один из самых детальных инструментов анализа текста на основании лингвистических параметров можно назвать сервис Coh-Metrix[5] [Graesser et al. 2011]. Сервис предлагает оценку текста по 108 параметрам, среди которых присутствуют как количественные (формулы Флеша и формулы Флеша-Кинкейда), так и качественные показатели текста: повествовательность (narrativity), синтаксическая простота (syntactic simplicity), конкретность слов (word concreteness), референциальная когезия текста (referential cohesion), глубокая когезия (deep cohesion). Среди прочего авторы предлагают оценку сложности текста на английском языке для иностранных читателей, рассчитывающиеся на основании трех показателей: уровня «переиспользования» слова в других частях текста, частотности слов текста и близостью синтаксических структур внутри текста. К минусам этого инструмента можно отнести чрезмерную нагруженность терминами и отсутствие интерпретации результатов анализа, что затрудняет её использование в повседневной преподавательской практике.

Примером сервиса, в большей степени ориентированного на нужды практикующих преподавателей английского языка как иностранного, является сервис TextInspector[6], пример работы которого представлен на Рисунке 3.

---

[5] Coh-Metrix tool 3.0. Доступна на: http://tool.cohmetrix.com/ [дата обращения: 02.11.2020]
[6] TextInspector. Доступна на: https://textinspector.com/ [дата обращения: 30.11.2021]

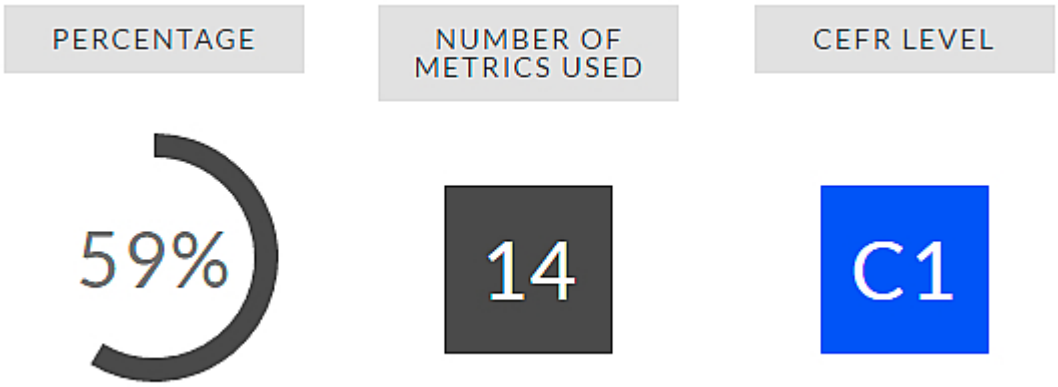


Рисунок 3 – Пример работы сервиса TextInspector

Для любого введенного текста он предлагает расчет как стандартных показателей сложности текста: средней длины слова и предложения, лексического разнообразия, процента длинных слов, популярных формул читабельности, так и метрики, специфические для использования текста в иностранной аудитории: процент лексики из лексических списков, стратифицированных по уровням шкалы CEFR, вхождение лексики текста в частотные списки крупнейших корпусов английского языка – BNC и COCA, дискурсивные маркеры текста, позволяющие оценить его связность. Наконец, сервис выдает оценку предполагаемого уровня сложности текста по шкале CEFR, однако использует только лексическую информацию для оценки.

Среди существующих сервисов по анализу русскоязычных текстов необходимо отметить проект «Простой русский»[7], предлагающий статистику по 5 популярным формулам читабельности, изначально разработанным для английского языка и адаптированным для русскоязычных текстов: формуле Flesch-Kincaid, индексу Колман-Лиау, Automatic Readability Index, SMOG (Simple Measure of Gobbledygook), и формуле Дэйла-Чейла. Результат обработки текста представляет собой оценку текста по каждой из формул, а также совокупную оценка сложности текста по этим 5

[7] Оценка читабельности текста. Доступна на: http://ru.readability.io [дата обращения: 02.11.2021]

параметрам. Кроме того, сервис предлагает значения базовых расчетных показателей, помогающих оценить уровень сложности текста: число знаков, букв, слов, предложений, процент использования сложных слов и проч.

В отечественной лингвистике существует также информация о разработке алгоритма «Анализ читабельности» (Readability analysis), нацеленного на оценку трудности учебных текстов для студентов высших учебных заведений [Зильберглейт и др. 2012], компьютерной программы «Оценка сложности параметров текста» [Криони и др. 2008], исследовательского программного стенда, рассчитывающего 12 формул читабельности по каждому абзацу текста [Мизернов, Гращенко 2015], однако стоит подчеркнуть, что данные сервисы, насколько нам известно, не сопровождаются открытым веб-интерфейсом, и, следовательно, недоступны широкому кругу пользователей.

Стоит отдельно отметить ресурс «Лексикатор», оценивающий текст с точки зрения его соответствия уровню знания русского языка как иностранного по лексическим и структурным параметрам по заданным методистами правилам [Баранова, Елипашева 2014; Сибирцева, Карпов 2014]. Лексический уровень сложности сервис оценивает с помощью вхождения слов в лексические минимумы уровней A1, A2 или B1. Структурный уровень обработки заключается в выделении структурных осложнителей текста, заданных сводом разработанных правил, например, количества сочинительных и подчинительных союзов, и т.п. В связи с тем, что не существует чётких границ соответствия сложности различных структурных осложнителей уровням изучения языка, все правила созданы для одного общего уровня. Результат лексической обработки текста представлен на Рисунке 4.

Данный ресурс предполагался стать частью более масштабного проекта по созданию рекомендательной системы по адаптации текстов для занятий РКИ, однако на момент обращения к сервису он перестал функционировать по данному адресу.

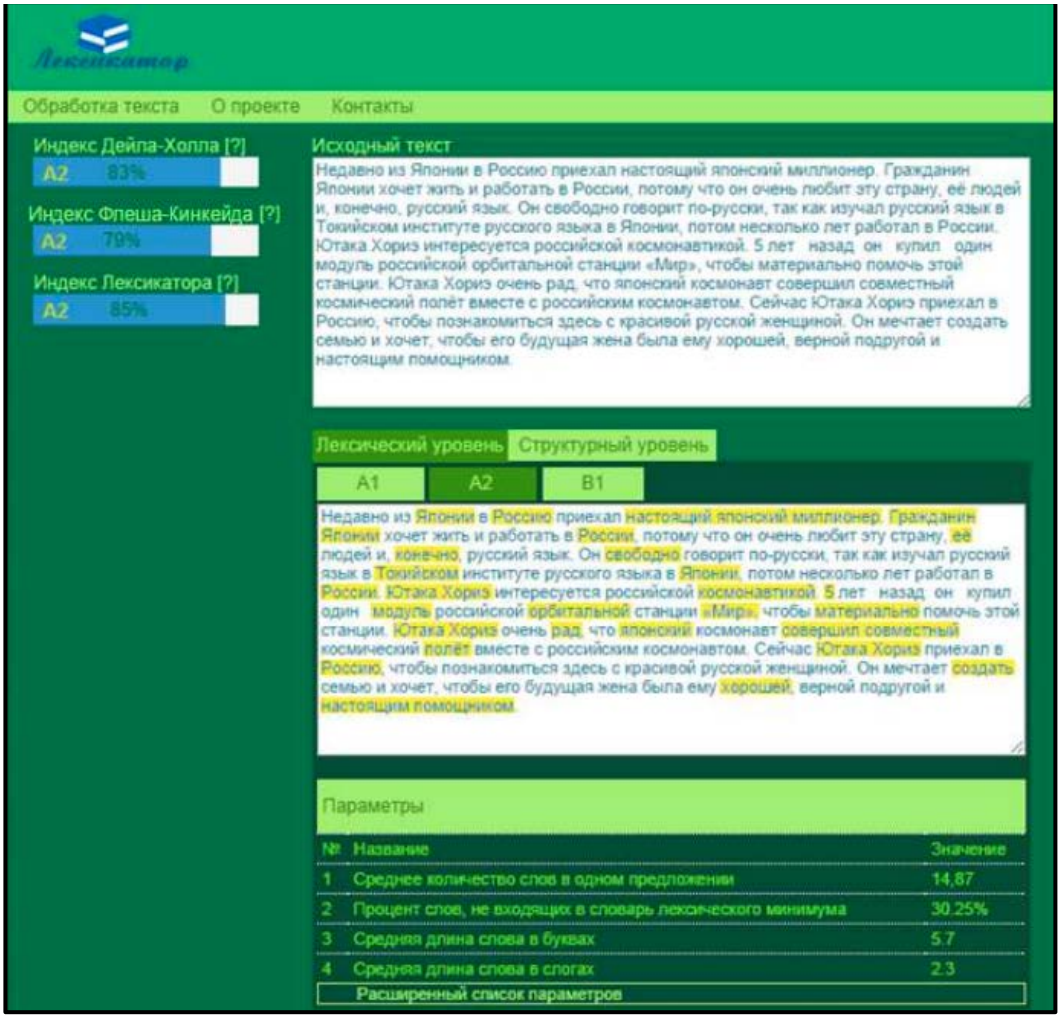


Рисунок 4 – Пример выделения лексических показателей сложности текста сервисом «Лексикатор»

Таким образом, при активной разработке методов и алгоритмов оценки лингвистических параметров текста, наблюдается, во-первых, недостаток сервисов для анализа русскоязычных текстов, открытых для использования широкому кругу пользователей, а во-вторых, явная лакуна в специализированном сервисе, содержащем информацию о параметрах текста с точки зрения преподавания русского как иностранного. Это в свою очередь является причиной недостаточной разработанности методики использования подобных систем в практике преподавания РКИ и процедур проверки качества автоматических систем анализа текста применительно к иностранной аудитории. Эти вопросы, насколько нам известно, впервые ставятся в данной работе.

Выводы по главе 1

Отбор текстов для учебников или занятий является актуальной задачей в методике преподавания РКИ, при этом к тексту предъявляется целый ряд требований, среди которых одним из центральных признается языковая доступность текста.

Доступность текста может быть описана в понятиях сложности и трудности текста, при этом следует различать сложность учебного материала как его объективную характеристику, выраженную в его лингвистических параметрах, и трудность текста как более широкую и частично субъективную характеристику, включающую себя, помимо сложности материала, факторы подготовленности учащихся к преодолению этой сложности.

История разработки темы автоматической оценки сложности текста для изучающих язык как иностранный тесно связана с исследованием сложности текста вообще 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, более развернутого, чем простые формулы читабельности. При этом обзор аналогичных продуктов для других языков демонстрирует возможные направления для деятельности.

ГЛАВА 2. РАЗРАБОТКА И АПРОБАЦИЯ МАТЕМАТИЧЕСКОЙ МОДЕЛИ ДЛЯ АВТОМАТИЧЕСКОГО ОПРЕДЕЛЕНИЯ СЛОЖНОСТИ ТЕКСТА ПО ШКАЛЕ CEFR

Вторая глава диссертационного исследования описывает опыт создания математической модели автоматического определения сложности текста для занятий РКИ. Работа включала в себя 3 основных этапа, схематично представленных на Рисунке 5.

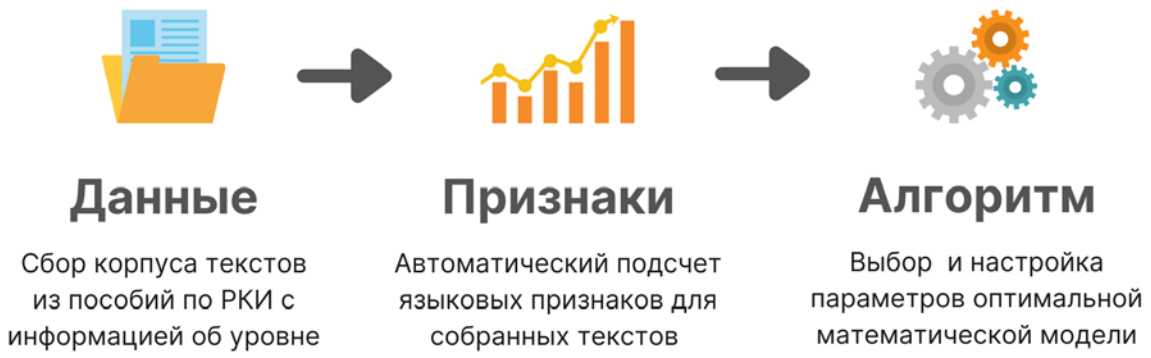


Рисунок 5 – Схема работы по построению модели автоматического определения сложности текста

Первым этапом является создание эталонного корпуса текстов, на которых модель будет обучаться и тестироваться. Этот шаг является ключевым, поскольку качество модели напрямую зависит от качества данных, на которых она обучается. Работа над таким корпусом описана в параграфе 2.1. Второй шаг заключается в подсчете лингвистических признаков для каждого текста обучающего корпуса. Этот этап работы не менее важен, поскольку с его помощью получается выделить формальные, рассчитываемые показатели текста, коррелирующие с уровнем его сложности. Методика получения лингвистических признаков с помощью систем автоматической обработки языка описана в параграфе 2.2., а изучение их информативности в задаче определения сложности текста – в параграфе 2.3. И, наконец, финальным этапом является обучение математической модели линейной

регрессии на основе полученных признаков. Эта часть работы описана в параграфе 2.4.

2.1. Сбор и описание корпуса текстов пособий по РКИ

В качестве эталонной коллекции текстов для обучения математической модели, был собран корпус из 802 текстов из пособий и электронных ресурсов по РКИ. Поскольку лингвистические характеристики текста очень сильно зависят от его формы, в корпус отбирались только фрагменты прозаических произведений не диалогического характера. В качестве уровня сложности текстов принималась информация, отображённая 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)*»). На данном этапе исследования мы принимаем допущение, что указанные уровни адекватно оценивают уровень сложности входящих в пособие текстов. Отбор учебников и текстов электронных ресурсов проходил по следующим критериям:

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

В результате отбора сформировалась коллекция из 30 учебников и пособий по чтению с общим количеством текстов 802. Среди них есть как учебные комплексы, уже ставшие классикой («Дорога в Россию», «Жили-были», «Поехали!»), так и сравнительно новые пособия («Русский сувенир», «Точка Ру»). Таблица 5 иллюстрирует распределение отобранных учебников по годам. Полный список источников корпуса находится в разделе «Список учебников и учебных пособий» библиографии.

Таблица 5 – Распределение отобранных печатных учебных пособий по годам издания

Период издания (г.)	**Количество пособий в корпусе**
2005 – 2010	10
2010 – 2015	11
2015 – 2020	9
Итого	**30**

Помимо пособий, в коллекцию вошли тексты ресурса Learn Russian With Interest (продолжающим проект текстотеки ЦМО МГУ)[8], текстовые материалы раздела «Учу русский» портала «Образование на русском»[9] и тексты из раздела Чтение тренировочных тестов ТРКИ. Пособия, направленные на изучение языка специальности (русский для медиков, русский язык в сфере туризма и т.п.), намеренно не включались в коллекцию из-за наличия в них специфической лексики и грамматики.

Общий объем корпуса, а также распределение текстов по числу примеров на каждый уровень представлены в таблице 6.

Таблица 6 – Объем корпуса RuFoLa

Параметр	**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

[8] http://lrwi.ru
[9] https://pushkininstitute.ru

Стоит отметить и ряд проблем, с которыми мы столкнулись на этапе сбора эталонной коллекции текстов. Туманное авторское описание уровня (например, *для продвинутых, для второго семестра первого года обучения*) затрудняет отнесение текста к уровням по шкале CEFR и сравнение текстов разных авторских коллективов. Размытость границ уровней (например, *ориентирован на B1-C1*) также затрудняет что затрудняет отнесение текста к одному определенному классу сложности. Наконец, существует риск субъективности информации об уровне: насколько нам известно, в настоящий момент не существует единой формальной процедуры маркировки пособия по уровням CEFR и принятие этого решения ложится на авторов пособия и редколлегию издательства (а иногда и только авторов, поскольку на рынке существуют пособия, изданные самостоятельно), что может приводить к необъективности данных и несопоставимости между собой разных пособий.

Отдельной проблемой явилось практически полное отсутствие образцов текстов для уровня C2. Поскольку C2 считается наивысшим уровнем владения иностранным языком, предполагается, что человек может читать и понимать практически любые оригинальные тексты общей тематики. Чтобы получить образцы текстов этого уровня, мы наполнили корпус для этого уровня текстами из различных научно-популярных и новостных изданий: Русский репортер, Вокруг Света, The Village и др. Однако дальнейшие эксперименты показали, что отобранные аутентичные тексты часто оказывались по своим лингвистическим признакам проще, чем представленные в пособиях высоких уровней (B2 и C1), поэтому мы приняли решение исключить эти тексты из коллекции и завершить шкалу сложности текстов корпуса на уровне C1. Таблица 7 содержит примеры текстов корпуса для каждого уровня владения русским языком как иностранным.

Таблица 7 – Образцы разметки текстов корпуса по уровням CEFR

Уровень текста	**Пример (фрагмент текста)**
A1	Сегодня воскресенье. Мы отдыхаем. Папа читает журнал «Спорт». Он очень любит спорт, особенно футбол. Мама в свободное время любит готовить. Сегодня она готовит пельмени.
A2	Все события фильма происходят 31 декабря. Герои фильма – молодые люди Настя и её жених Коля – готовятся встречать Новый год. У Насти никого нет, кроме Коли. Мама Насти умерла, а своего отца Настя никогда не знала. <...>
B1	Георгий Гречко, лётчик-космонавт: «Моя мать работала главным инженером завода. Помню, как на следующий день после того, как она ушла на пенсию, она мне сказала: "Первый раз я спала спокойно". До этого она каждую ночь беспокоилась, не случилось ли что-нибудь на заводе, но если бы кто-нибудь предложил моей матери не работать, а только заниматься домашним хозяйством, она бы не согласилась – она любила свой завод, свою работу. <...>
B2	Всем присутствующим было предложено стать конструкторами придуманной организаторами «мастерской будущего», в которой бы прорабатывались основные проблемы, интересующие молодежь накануне XXI века. Выяснилось, что волнующие россиян и немцев темы во многом схожи. Это бюрократизм и закоснелость мышления, безработица, преувеличение роли денег, равнодушие людей, языковые и культурные барьеры между странами. <...>
C1	Независимая оценка знаний заключается в том, что студенты сдают экзамены независимым экспертам – преподавателям других вузов, которые не проводили занятия в этих группах. Процедура сдачи экзаменов традиционна – в форме собеседования и письменных ответов на вопросы билетов. Идею независимой оценки качества образования поддержал Уполномоченный по правам студентов в РФ Артем Хромов, о чем сообщается на его официальном сайте. <...>

2.2. Сбор лингвистических признаков для обучения модели

Следующей задачей исследования является анализ полученного корпуса и сбор лингвистических признаков текстов. Для решения этой задачи был написан программный код на языке Python. Каждый текст коллекции проходил цикл автоматической обработки, включающий следующие шаги:

1. Предобработка текста: чистка от знаков ударений, спецсимволов, ссылок и т.п.
2. Деление текста на предложения с помощью модуля NLTK sent_tokenize[10], дополненный собственными правилами.
3. Токенизация текста (деление на слова) и лемматизация (приведение слов к их начальным, словарным формам) осуществлялось с помощью модуля Mystem[11].
4. Далее каждое слово получало набор грамматических характеристик с помощью модуля Mystem. Например, существительное СТОЛЕ получал такой набор тэгов: *существительное, единственное число, мужской род, предложный падеж*, начальная форма – стол.
5. В случае грамматической омонимии выбирался вариант, статистически наиболее вероятный в данном контексте: например, *печь* в сочетании с существительным в винительном падеже получал глагольные грамматические характеристики.
6. Далее вся полученная информация обобщалась до набора признаков данного текста, описанных ниже. Например, мера «доля слов в родительном падеже» вычислялась как отношение количества слов с тэгом «существительное, родительной падеж» к количеству слов с тэгом «существительное».

В результате анализа релевантных работ был сформирован набор из нескольких групп признаков, потенциально способных оказывать влияние на уровень сложности текста. Полный список использованных в работе признаков представлен в Таблице 8.

[10] https://www.nltk.org
[11] https://github.com/nlpub/pymystem3

Таблица 8 – Лингвистические признаки для обучения модели

Группа	**Признак**
Лексические признаки	средняя длина слова в знаках
	медианная длина слова в знаках
	средняя длина слова в слогах
	медианная длина слова в слогах
	процент слов длиннее 4 слогов
	лексическое разнообразие (type-token ratio, TTR)
	лексическое разнообразие MLTD (MLTD TTR)
	лексическая плотность (lexical density)
	покрытие текста частотным списком 1 000
	покрытие текста частотным списком 5 000
	покрытие текста частотным списком 10 000
	покрытие текста списком ЛМ A1
	покрытие текста списком ЛМ A2
	покрытие текста списком ЛМ B1
	покрытие текста списком ЛМ B2
	покрытие текста списком ЛМ C1
	покрытие текста списком KELLY A1
	покрытие текста списком KELLY A2
	покрытие текста списком KELLY B1

	покрытие текста списком KELLY B2
	покрытие текста списком KELLY C1
	покрытие текста списком KELLY C2
	процент слов из списка абстрактной лексики
	процент слов с абстрактными суффиксами
Грамматические признаки[12]	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
	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

[12] Для группы грамматических признаков в таблице приведены лишь несколько примеров, т.к. общее количество грамматических признаков составляет 49

	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

В следующих параграфах представим более подробный обзор техники извлечения отобранных групп признаков и изучение их корреляционной связи с уровнем сложности текста по шкале CEFR.

2.2.1. Лексические признаки

Первую группу составляют лексические признаки текста, которые свою очередь можно разделить на несколько подгрупп: признаки, основанные на длинах; признаки лексического разнообразия; признаки, основанные на частотных данных; признаки, основанные на специфических списках, ориентированных на обучение РКИ; лексические признаки, связанные со значением слова. Средняя и медианная длина слова в тексте являются, пожалуй, наиболее часто используемыми показателями в контексте определения сложности текста и опираются на предположение, что короткие слова чаще оказываются проще для чтения и восприятия. Рисунок 6 подтверждает плавный рост этих показателей на коллекции текстов РКИ.

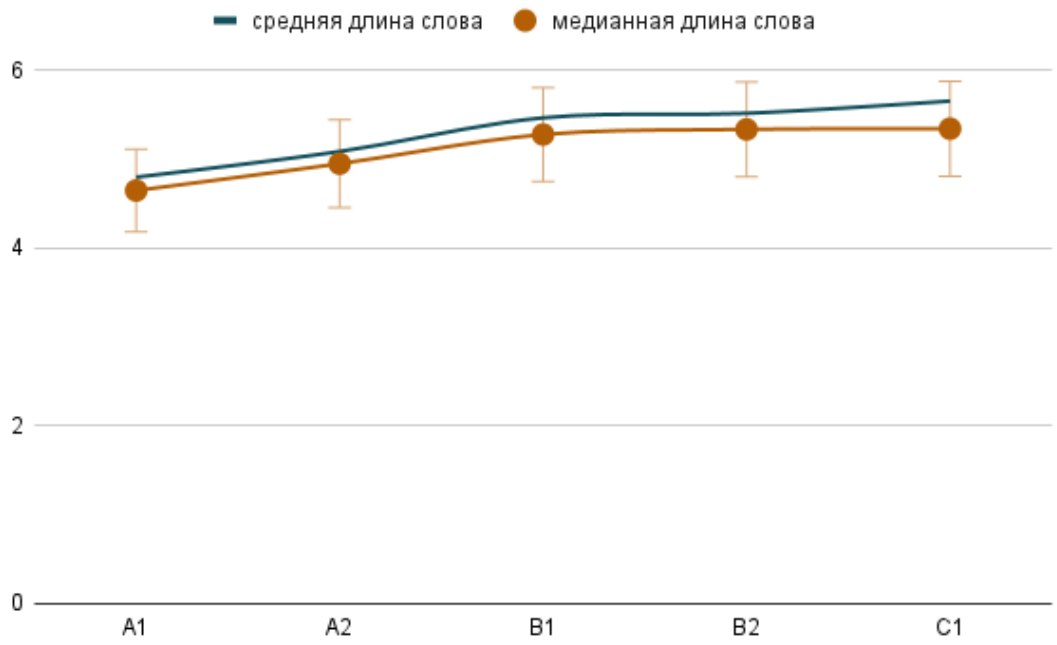


Рисунок 6 – Распределение средней и медианной длины слова в текстах по уровням CEFR

Вслед за большинством исследователей, мы также используем рассчитываем коэффициент лексического разнообразия (англ. lexical diversity) с помощью мер TTR и MLTD, и коэффициент лексической плотности текста (lexical density).

В качестве источника данных о частотности слов для расчета лексических признаков, основанных на частотности слов, мы использовали данные Нового частотного словарь современного русского языка [Ляшевская, Шаров 2009], созданного на материалах художественных и публицистических текстов 1950–2007 гг. Национального корпуса русского языка. Мы использовали список лемм словаря, отсортированный по убыванию частотности для создания нескольких списков: 1 000 самых частотных слов, 3 000, 5 000 и 10 000. Признаком текста при этом является процент слов текста, покрываемых тем или иным частотным списком.

Одним из важнейших показателей знакомости лексики в контексте преподавания РКИ является покрытие текста специальными списками лексики, стратифицированными по уровням владения языком. Для расчета этой меры была

использована линейка лексических минимумов ТРКИ (далее ЛМ ТРКИ) от A1 до C1 [Лексический минимум 2013; 2015; 2017a; 2017b; 2018). Здесь признаком выступает процент лексики текста, входящей в список того или иного уровня, т.е. потенциально знакомой студенту. Соответственно, чем выше этот показатель, тем предположительно проще лексика текста. Отдельной проблемой, которую было необходимо решить при этом типе подсчетов, явилось отсутствие общепринятой методики учета дериватов от слов, присутствующих в лексическом минимуме и образованных по словообразовательным моделям, доступным на данном уровне владения русским языком. Например, слова *небольшой, фотограф, узнать, прилететь, красиво* отсутствуют в ЛМ уровня А2, однако почти все из них имеют однокоренные слова, включенные в ЛМ (*большой, фотографировать, знать, лететь, красивый*). В результате экспериментальной работы было принято решение о создании расширенных версий списков на базе ЛМ ТРКИ с добавлением дериватов, образование которых должно быть доступно студентам данного уровня. Подробнее работа по созданию расширенных списков описана в [Лапошина 2021]. Таким образом, здесь в качестве ЛМ ТРКИ выступают их расширенные версии с учетом дериватов. Интересно отметить, что даже расширенные версии списков значительно превышают описанные в требованиях. Так, в Таблице 9 приведены усредненные значения процента покрытия всех образцов текстов уровня линейкой лексических минимумов. Затемненным цветом помечены ключевые ячейки таблицы, демонстрирующие расчеты по лексическому минимуму, целевому для данного уровня. Например, из Таблицы 9 видно, что средний процент покрытия текстов уровня A1 лексическим списком уровня A1 составляет 88 процентов, что говорит о значительном превышении количества незнакомой лексики, указанном в требованиях: 12% вместо 2–3%. Превышение значений, описанных в требованиях, наблюдается на всех уровнях, кроме самого высокого, C1: там предложенный усредненный показатель соответствует норме. Данное несоответствие не является критичным для дальнейшего построения модели машинного обучения, однако, как

нам кажется, оставляет задел для дальнейшей методической работы в области уточнения нормативных документов или методики автоматизированного подсчета незнакомой лексики текста.

Таблица 9 – Усредненный процент покрытия текстов разных уровней лексическими минимумами

Уровень подкорпуса текстов	**Процент покрытия текстов подкорпуса лексическими минимумами**				
	A1	**A2**	**B1**	**B2**	**C1**
A1	88	92	95	97	99
A2	72.5	82	89	94	98
B1	61	71	83	90	96
B2	59	69	80	89	95
C1	55	66	76	86	93

В качестве альтернативного источника оценки степени знакомости лексики текста мы использовали линейку лексических списков проекта Kelly, собранных на основе подсчета частотности слов на материале большого интернет-корпуса текстов [Kilgariff и др. 2014].

Необходимо уточнить, что при подсчете признаков на основе вхождения в лексические списки мы использовали дополнительную предобработку текстов: из них были исключены имена собственные, географические названия и слова, неизвестные морфологическому анализатору (скорее всего они написаны с опечаткой).

Для расчета признаков, связанных с абстрактностью слов текста мы использовали два метода: стандартный подсчет слов, содержащих абстрактные суффиксы *(-ость, -ение* и мн. др.) [Томина 1985], а также долю слов из семантических списков абстрактной лексики, полученных с помощью лингвистического модуля ABBYY COMPRENO, где каждой лексеме в полуручном режиме приписываются специальные семантические метки – семантемы. Список для настоящего

исследования был составлен на основе перечня семантем абстрактности [Анисимович и др. 2012].

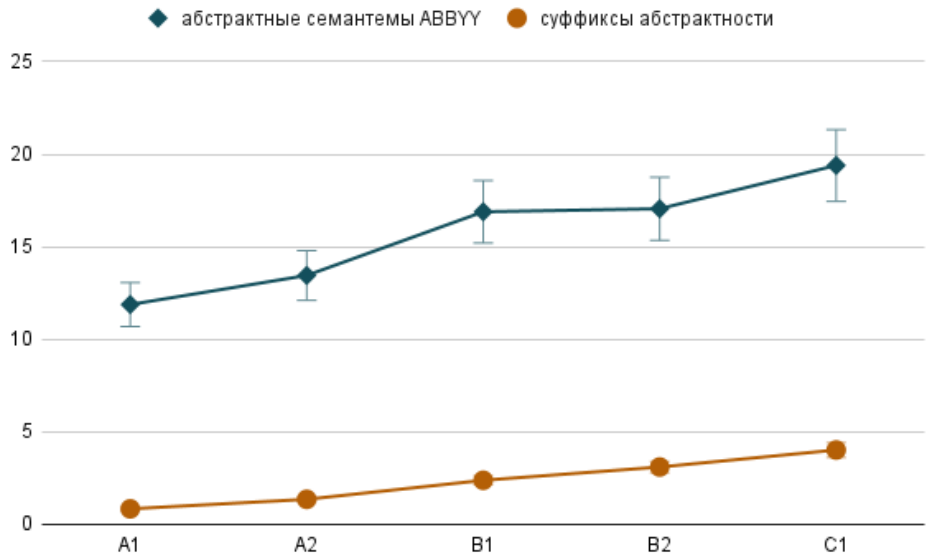


Рисунок 7 – Процент абстрактных существительных в тексте в зависимости от уровня CEFR

График 7 демонстрирует плавный рост процента абстрактных существительных с ростом уровня сложности текстов: это связано вероятнее всего и с изменением круга обсуждаемых тем по мере роста уровня владения языком. Становится также заметно, что список с использование семантем ABBYY COMPRENO находит в несколько раз больше абстрактных существительных в тексте: если поиск по суффиксам позволяет найти «стандартные» случаи абстрактных слов: такие лексемы как *строительство, ситуация, государство, требование* и т.п., то список лексем с семантемами абстрактности находит абстрактные слова и термины простой морфологической формы: *ритм, закон, бемоль, блеск, успех* и т. п.

Таблица 10 содержит результаты корреляционного анализа данной группы признаков, расположенные в порядке убывания коэффициента корреляции. Здесь и далее для изучения величины корреляции использован коэффициент ранговой корреляции Спирмена. При этом значения выше 0.7 оцениваются как высокая

корреляционная связь, от 0.5 до 0.7 – заметная связь, от 0.3 до 0.5 – умеренная связь и ниже 0.3 – слабая связь [Лагутин 2021]. Для оценки значимости полученных коэффициентов использовалась мера p-value. При этом поля таблицы белого цвета означают параметр p-value менее 0.05, т.е. значимую корреляцию; поля с p-value больше 0.05 обозначены серым цветом.

Таблица 10 – Коэффициенты корреляции лексических признаков текстов с уровнями CEFR

Параметр	**Коэффициент корреляции Спирмена**
покрытие текста списком ЛМ A1	-0.79
покрытие текста списком ЛМ A2	-0.79
покрытие текста списком ЛМ B1	-0.75
покрытие текста списком KELLY A1	-0.74
покрытие текста списком KELLY A2	-0.74
покрытие текста списком KELLY B1	-0.74
покрытие текста списком ЛМ B2	-0.71
покрытие текста списком KELLY B2	-0.64
покрытие текста списком ЛМ C1	-0.58
покрытие текста частотным списком 5 000	-0.55
покрытие текста списком KELLY C1	-0.55
процент слов длиннее 4 слогов	0.54
покрытие текста частотным списком 10 000	-0.52

покрытие текста списком KELLY C2	-0.52
процент слов с абстрактными суффиксами	0.52
средняя длина слова в знаках	0.51
средняя длина слова в слогах	0.49
покрытие текста частотным списком 1 000	-0.48
процент слов из списка абстрактной лексики	0.38
медианная длина слова в знаках	0.33
лексическое разнообразие MLTD (MLTD TTR)	0.23
лексическое разнообразие (type-token ratio, TTR)	0.11
лексическая плотность (lexical density)	0.09

Анализ коэффициентов корреляции показывает наиважнейшую роль учета специализированных списков лексики для изучающих РКИ: эти признаки показали самую тесную связь с уровнем текста по шкале CEFR. При этом коэффициент корреляции уменьшается по мере роста уровня и, соответственно, объема списка. Списки проекта KELLY также показывают одни из наивысших коэффициентов, следуя похожей логике постепенного уменьшения коэффициента с ростом уровня. Процент слов текста, входящих в частотные списки русского языка, также показывает статистически значимую связь с его сложностью: лучший результат здесь показывает список 5 000 слов. Среди простейших метрик, основанных на длинах, наилучший результат показывает процент слов текста длиннее 4 слогов и средняя длина слова в знаках, при этом результаты подсчетов по средним или медианным значениям не показывают заметной разницы. Интересно, что процент слов с абстрактными суффиксами, содержащих значительно меньше лексем, получает коэффициент

корреляции выше, чем процент слов из специализированных списков с использованием семантем. Наконец, меры лексического разнообразия и лексической плотности показывают самые низкие результаты корреляции в данной группе признаков.

2.2.2. Грамматические признаки

Грамматические признаки рассчитываются как доля того или иного грамматического тэга. Морфологическая информация слова рассчитывается автоматически с помощью модуля pymystem3. Использованы расчеты количества тэгов на общее количество слов (например, доля существительных от всех слов текста, доля слов в родительном падеже от всех слов текста. Значимым для уровня сложности может оказаться как увеличение, так и уменьшение доли определенной грамматической категории. Например, на Рисунке 8 представлено изменение процента слов в именительном, родительном, предложном и творительном падежах с возрастанием уровня сложности текстов.

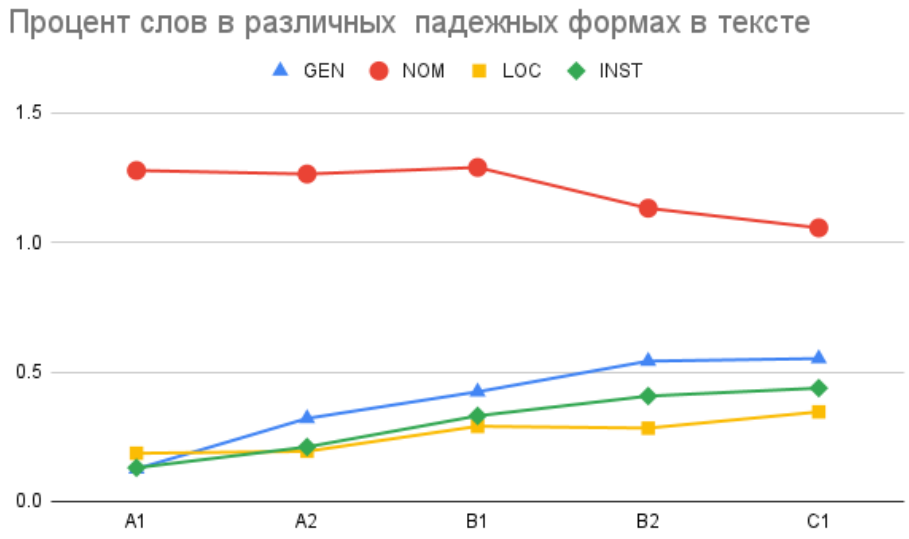


Рисунок 8 – Процент слов в различных падежных формах в корпусе текстов из пособий по РКИ

Можно заметить, что увеличение процента косвенных падежей имеет положительную корреляцию со сложностью, т.е. возрастает вместе с ростом уровня сложности текстов, имеет прямую зависимость. Увеличение же процента именительного падежа, наоборот, может свидетельствовать о более низком уровне сложности текста, имеет обратную зависимость. С помощью графика можно обнаружить еще одну методическую особенность, связанную со спецификой преподавания РКИ: процент предложного падежа, в целом постепенно возрастающий, не показывает такого роста от уровней A1 к A2. Скорее всего, это обусловлено тем, что конструкции с предложным падежом вводятся одними из первых и активно отрабатываются в учебниках элементарного уровня. Таблица 11 содержит данные корреляционного анализа группы грамматических признаков на уровень сложности текста. Для оценки значимости полученных коэффициентов использовалась мера p-value. При этом поля таблицы белого цвета означают параметр p-value менее 0.05, т.е. значимую корреляцию; поля с p-value больше 0.05 обозначены серым цветом.

Таблица 11 – Коэффициенты корреляции грамматических признаков текстов с уровнями CEFR

Процент слов с грамматическим тэгом в тексте	**Коэффициент корреляции Спирмена**
средний род	0.66
родительный падеж	0.57
причастие	0.57
дательный падеж	0.56
творительный падеж	0.54
страдательный залог	0.54

действительный залог	0.51
краткость	0.48
совершенный вид	0.48
деепричастие	0.47
полннота	0.47
неодушевленность	0.47
множественное число	0.46
женский род	0.44
единственное число	0.42
винительный падеж	0.41
инфинитив	0.34
сравнительная степень	0.34
предложный падеж	0.33
мужской род	0.32
прошедшее время	0.3
превосходная степень	0.26
изъявительность	0.24
несовершенный вид	0.2
1 лицо	-0.18

именительный падеж	-0.17
одушевленность	0.15
3 лицо	0.11
повелительное наклонение	0.09
притяжательность	0.08
непрошедшее время	0.06
2 лицо	-0.06

Анализ данных коэффициентов корреляции показывает несколько интересных закономерностей. Так, самая тесная связь между грамматической категорией и сложностью текста наблюдается у признака процента слов среднего рода в тексте. Категория рода нечасто отмечается в качестве показателя сложности в работах, посвященных близких задачам на других языках и имеет свои специфические черты в русскоязычном материале. Позволим себе сделать предположение, что это может быть связано с долей абстрактной лексики и нематериальных понятий в тексте, однако подтверждение такого предположения требует дополнительного анализа. Категория падежа наблюдается на верхних позициях таблицы корреляций. При этом самую тесную связь со сложностью обнаруживает процент слов в родительном, дательном и творительном падежах. Это соотносится с данными, полученными на материале русских текстов для носителей языка [Дружкин 2016]. Ожидаемо влияют на сложность текста процент причастий, деепричастий и связанных с ними категорий залога. Категория лица не показывает значимой связи с уровнем текста. Однако на графике, представленном на рисунке видно, как сильно на начальных уровнях превышен уровень употребления форм 1 лица.

2.2.3. Синтаксические признаки

Самым базовым признаком возможной сложности синтаксической структуры предложения является его длина: чем длиннее предложение, тем больше вероятность, что в нем присутствуют сложные синтаксические конструкции. Другие признаки, способные оказывать влияние на уровень сложности синтаксической структуры текста, были получены, во-первых, с помощью подсчетов специальных морфологических тэгов слов (союз, союзное слово, знак пунктуации), а также с помощью более детального исследования структуры текста автоматическим синтаксическим парсером Spasy[13]. Этот инструмент позволяет строить синтаксические деревья зависимостей (dependency trees) для каждого предложения коллекции (см. пример на Рисунке 9).

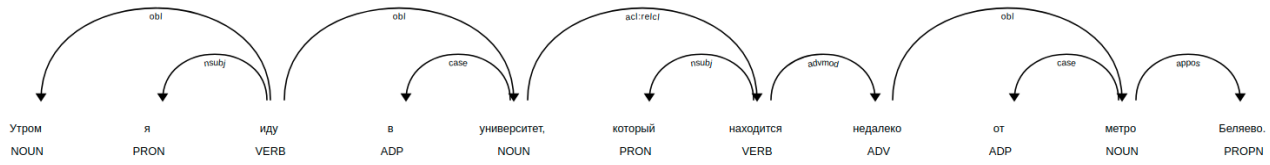


Рисунок 9 – Схема синтаксического разбора предложения парсером Spacy

В результате автоматического анализа построенных парсером синтаксических деревьев мы произвели расчеты параметров, встреченных для аналогичных задач: средней глубины дерева (расстояния от главного слова, корня, до самого дальнего зависимого), среднее количество слов, стоящих в предложении до главного слова.

[13] https://spacy.io/models/ru

Таблица 12 – Коэффициенты корреляции синтаксических признаков текстов с уровнями CEFR

Признак	**Коэффициент корреляции Спирмена**
средняя длина предложения	0.59
средняя максимальная глубина синтаксического дерева	0.58
медианная длина предложения	0.55
процент местоимений-прилагательных[14] в тексте	0.54
процент предлогов в тексте	0.54
среднее количество пунктуаторов в предложении	0.47
процент глаголов в тексте	0.47
процент существительных в тексте	0.46
процент прилагательных в тексте	0.46
среднее количество слов до корневого слова	0.44
процент союзов в тексте	0.39
процент частиц в тексте	0.38
покрытие текста списков 500 самых частотных pos-триграмм	-0.34
процент слов в форме пассива в тексте	0.31

[14] Сохранена расшифровка частеречных тэгов модуля mystem3: https://yandex.ru/dev/mystem/doc/grammemes-values.html

процент числительных-прилагательных в тексте	0.23
процент наречий в тексте	0.2
процент междометий в тексте	0.12
процент местоимений-существительных в тексте	-0.06
процент местоименных наречий в тексте	0.04

Из полученных коэффициентов становится очевидно, что самую тесную связь с уровнем сложности показывает самая базовая метрика средней длины предложения в словах. Средняя максимальная глубина синтаксического дерева находится на второй позиции, однако следует понимать, что для её расчета требуется гораздо больше вычислительных ресурсов. Среди признаков на основе частеречных тэгов самую высокую корреляционную связь показывают процент местоимений-прилагательных (*какой, чей, этот, наш* и т.п.) и предлогов в тексте. Количество пунктуаторов в тексте и покрытие текста списком 500 самых частотных pos-триграмм показывают умеренную, но статистически значимую связь с уровнем текста.

2.2.4. Дискурсивные признаки

Анализ научной литературы по смежным темам показал, что дискурсивные признаки способны отразить более широкое видение сложности и связности текста как единого продукта речевой деятельности. Так, показателем связности признаков является уровень лексического повтора существительных двух соседних предложений (noun overlap) и уровень пересечения любых лемм двух соседних предложений. Кроме того, мы использовали подсчеты встречаемости в тексте двух групп маркеров связности: временные связки (*до, после, затем, сначала, вначале, далее, потом* и др.) и причинные связки *(потому что, поэтому, следовательно,*

соответственно, так как и др.). Наконец, мы добавили подсчет количества местоимения *который*, зачастую выступающего в роли союзного слова.

Таблица 13 – Коэффициенты корреляции дискурсивных признаков текстов с уровнями CEFR

Признак	**Коэффициент корреляции Спирмена**
процент местоимений *который* на текст	0.41
лексический повтор (леммы)	0.33
дискурсивные маркеры времени	0.26
лексический повтор (существительные)	0.22
нарративность	-0.11
дискурсивные маркеры причины	0.02

2.2.5. Общий обзор полученных признаков

Анализ лингвистических признаков, извлеченных из корпуса текстов RuFoLa, дает представление о характеристиках текстов, наиболее релевантных задаче ранжирования текстов по уровням CEFR. Так, очевидным лидером являются лексические признаки, особенно основанные на вхождении слов текста в лексические минимумы. Этот результат соотносится и с данными предыдущих работ на материале РКИ: в работе Ю. Н. Барановой и Ф. М. Витюгина признаки, связанные с вхождением в лексические минимумы, показали значительно более сильную связь с уровнем текстов по РКИ, чем стандартные метрики как средняя длина слова или предложения [Karpov et al. 2014]. Р. Рейнолдс также сообщает о том, что объединенная группа всех лексических признаков показывает наилучший результат в индивидуальных зачетах [Reynolds 2016].

Среди грамматических признаков достаточно высокая степень корреляции показывает форма среднего рода, некоторые падежи и особые формы глагола. Среди синтаксических признаков наивысшую корреляцию показывает мера средней длины предложения. Это ставит под вопрос необходимость расчета куда более сложных метрик, основанных на построении синтаксических деревьев для этой задачи. С другой стороны, интерпретируемость результатов анализа уже не раз отмечалась как важнейшая особенность задачи автоматической оценки сложности учебных текстов. С этой позиции результат работы синтаксического парсера представляется более полезным, чем простая мера длины предложения.

Дискурсивные признаки не показывают ожидаемой сильной связи с ростом уровня сложности текстов. В ходе анализа признаков были также отмечены особенности, которые необходимо учесть в дальнейшей работе. Во-первых, становится очевидно, что полученные признаки во многом коррелируют между собой. Например, рост грамматического признака ‘действительное’ будет линейно связан с ростом доли причастий в тексте; рост количества косвенных падежей в тексте связан с падением количества слов в именительном падеже. Учитывая этот факт, мы добавили в эксперимент модель гребневой регрессии (Ridge Regression), которая хорошо справляется с проблемой мультиколлинеарности признаков.

Второй особенностью является возможная нелинейность признаков. Так, на Рисунке 10 показан график распределения процента форм 1, 2 и 3 лица в текстах различных уровней CEFR.

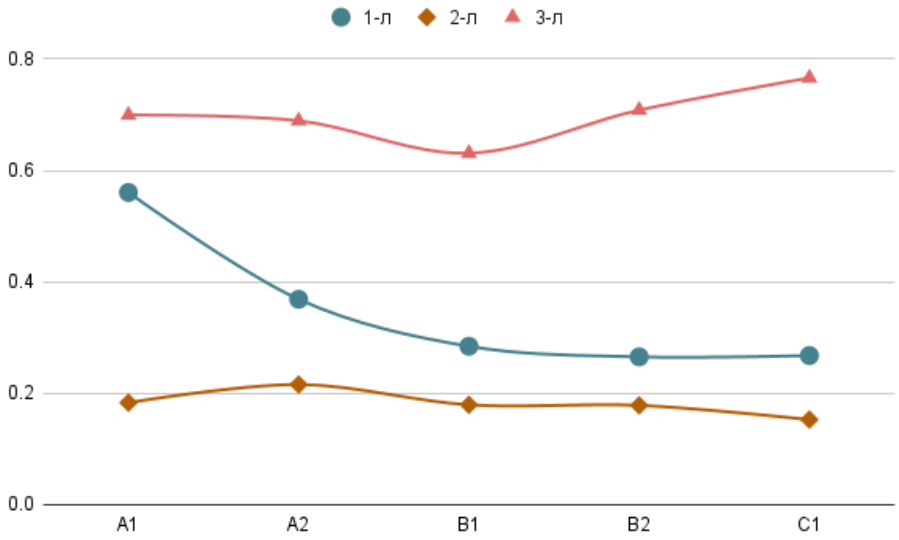


Рисунок 10 – Распределение процента форм 1, 2 и 3 лица в текстах различных уровней CEFR

На графике видно, что процент форм 2 и 3 лица меняется в зависимости от уровня незначительно. Процент же форм 1 лица стремительно падает по мере прохождения начальных уровней, и к уровню B1 выравнивается. Вероятнее всего, это связано с особенностями методики преподавания РКИ и активной работой по удовлетворению самых базовых коммуникативных потребностей, связанных с «я»-сообщениями. Подобные факты заставляют нас также осторожнее относиться к выбору модели машинного обучения в дальнейшем.

2.3. Построение и оценка качества предсказательной модели

Завершающим этапом нашей работы является построение модели машинного обучения, которая получает на вход признаки текста и делает предположение о его уровне сложности. Код выполнен на языке Python, все основные расчеты производятся с помощью библиотеки sklearn[15].

Проблема ранжирования текстов на определенное количество классов рассматривается в нашем исследовании как задача регрессии. Это вызвано взглядом на набор уровней сложности текстов, изложенный в том числе в Общеевропейских

[15] https://scikit-learn.org

компетенциях: не как на набор закрытых классов, а как постепенно возрастающую величину, где сами границы уровней являются условными. Кроме того, выбор регрессии в качестве метода позволяет оценить качество работы модели по метрикам, которые, по нашему мнению, более релевантным задаче: так, если при классификации любой ответ модели, отличающийся от эталонного, оценивается как ошибочный, то такие метрики оценки регрессии, как средний квадрат ошибки или средняя абсолютная ошибка позволяют увидеть масштаб, значительность ошибки. Наконец, результат работы регрессионной модели представляется более информативным и удобным для интерпретации: уровень сложности текста представляется при этом в виде дробного числа, давая возможность тем самым не только увидеть предполагаемый уровень сложности, но и ранжировать тексты по сложности внутри одного уровня: например, тексты с результатом 3.7. и 3.1. будут относится к уровню B1, при этом первый текст оценен как более сложный, чем второй.

Для построения регрессионных моделей были выбраны модели линейной (linear regression) и гребневой регрессии (ridge regression). Гребневая регрессия представляет собой одну из техник регуляризации, которые применяются в линейных моделях классификация и регрессии для того, чтобы решить проблемы зависимости признаков друг от друга и переобучения. Для этого вводится дополнительное штрафное слагаемое к основному функционалу регрессии, которое штрафует за избыточное увеличение нормы вектора коэффициентов. Т.е. помимо традиционной оценки по методу наименьших квадратов гребневая регрессия опирается еще и на квадрат нормы весов, который должен быть минимизирован. Все указанные результаты получены с дефолтными параметрами модели. Деление на обучающие и тестовые данные выполнялось автоматически с помощью модуля StratifiedShuffleSplit: случайным образом корпус делился из расчета 80% для обучения модели, 20% – для тестирования. Поскольку при каждом запуске прогона генерируется новое соотношение обучающих и тестовых данных, результат может отличаться, поэтому в таблице результатов мы используем десятикратную кросс-

валидацию (ten-fold cross validation). Для оценки качества модели были использованы классические метрики оценки регрессионных моделей из модуля sklearn.metrics:

1. Уровень дисперсии (variance score) является метрикой успешности регрессионной модели и представляет собой отношение квадрата стандартного отклонения ошибки к квадрату стандартного отклонения правильного ответа. Наивысшее значение – единица.

2. Коэффициент детерминации (coefficient of determination, r2) демонстрирует долю дисперсии зависимой переменной, объясняемой рассматриваемой моделью зависимости. Наивысшее значение – единица.

3. Средний квадрат ошибки (mean squared error) представляет собой среднее значение квадратов всех расстояний предсказания от правильного ответа. Следовательно, чем ближе к нулю значение метрики, тем лучше модель.

4. Средняя абсолютная ошибка (mean absolute error) высчитывается как среднее расстояние модуля предсказания от правильного ответа. Чем число ближе к нулю, тем лучше оценивается качество модели. В Таблице 14 представлены результаты сравнения качества моделей линейной и гребневой регрессии для полного набора лингвистических признаков.

Таблица 14 – Качество построенных регрессионных моделей

Тип регрессионной модели	**Параметр**	**Значение**
Линейная регрессия	Уровень дисперсии	0.82
	Коэффициент детерминации	0.75
	Средний квадрат ошибки	0.49
	Средняя абсолютная ошибка	0.56
Гребневая регрессия	Уровень дисперсии	0.84
	Коэффициент детерминации	0.76
	Средний квадрат ошибки	0.46
	Средняя абсолютная ошибка	0.55

Как видно из таблицы, лучший результат – высшее значение уровня дисперсии и коэффициента детерминации при меньших значениях абсолютной и средней ошибки – показала модель гребневой регрессии. Это может быть связано с тем, что она более устойчива к мультиколлинеарности признаков. Значение средней абсолютной ошибки, равное 0.55 говорит о том, что в среднем модель ошибается на 0.5 уровня. Для более детального анализа ошибок модели визуализируем их с помощью матрицы неточностей (confusion matrix), которая показывает количество правильных и неправильных прогнозов, а также дает представление о масштабе ошибок. Более темный цвет ячейки говорит о большем количестве совпадений предсказанных и истинных значений. Для этого приведем результаты работы регрессионной модели к целым числам для совпадения со шкалой обучающих данных.

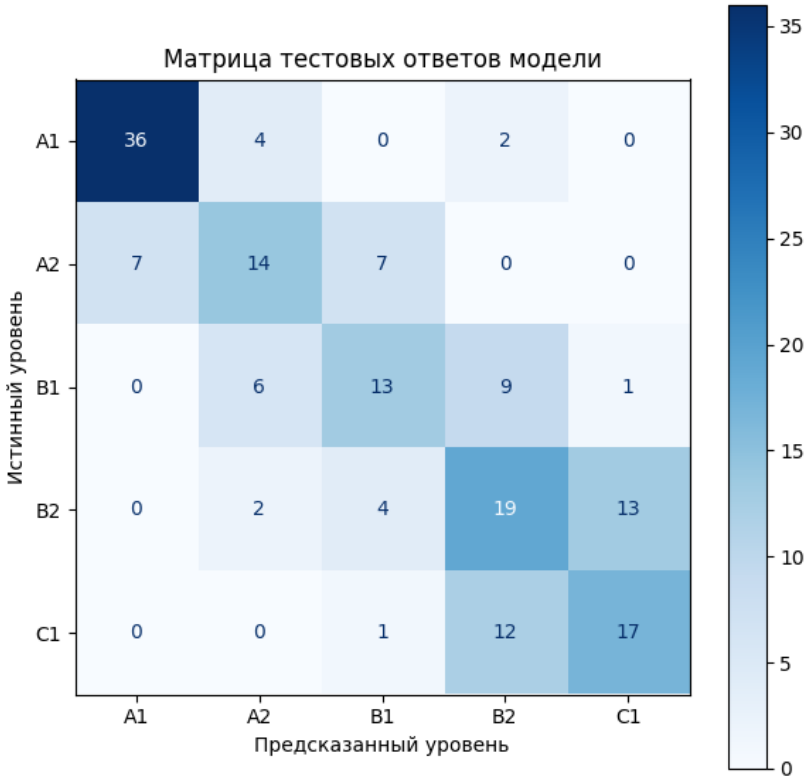


Рисунок 11 – Матрица тестовых предсказаний модели

Матрица ошибок на Рисунке 11 показывает, что лучше всего модель отличает тексты уровня A1. На остальных уровнях модель чаще всего присваивает верный уровень, а в случае неправильного прогноза, ошибка чаще всего составляет один уровень, при этом случаев завышения уровня моделью наблюдается чуть больше. Показательна ситуация с высокими уровнями, B2 и C1. Там качество работы модели нельзя назвать удовлетворительным: практически в половине случаев она ошибается, правда чаще всего ошибка составляет один уровень. Такая ситуация наиболее вероятно вызвана неоднозначностью экспертной оценки сложности текстов таких уровней и, как следствие, качеством эталонной коллекции таких текстов. Эта гипотеза выдвигалась нами и в предыдущем параграфе, где лингвистические параметры текстов высоких уровней становятся размытыми. Альтернативным объяснением может служить гипотеза, что на высоких уровнях выбранные нами лингвистические признаки оказываются нерелевантными.

2.4. Апробация машинной модели по определению сложности текста

Данный раздел диссертационного исследования посвящён описанию апробации работы предсказательной модели и состоит из двух этапов. 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 иностранных студентов и их преподавателей. Параграф 2.4.2 описывает алгоритм сравнения результатов работы предсказательной модели с оценками экспертов.

2.4.1. Экспериментальная верификация качества работы модели

Помимо стандартных метрик и методов проверки качества работы математической модели, представляется важным верификация её применимости в реальных педагогических условиях. Одним из центральных методов получения эмпирических данных в образовательной среде является эксперимент, как «особая

процедура, с помощью которой исходное предположение – гипотеза – приобретает иной статус и становится либо более правдоподобной, либо менее правдоподобной» [Фрумкина 1999].

В качестве гипотезы эксперимента было выдвинуто утверждение о наличии статистически значимой корреляции между оценкой текста математической моделью и наблюдаемыми в ходе эксперимента количественными показателями сложности текста, а именно: относительной скоростью чтения текста (слов в минуту), количеством слов, отмеченных респондентами незнакомыми, количеством правильных ответов на 3 закрытых и 1 открытый вопрос, ответом респондентов на закрытый вопрос о их собственном суждении о сложности предъявляемого текста и, наконец, экспертной оценкой уровня сложности текста преподавателями респондентов.

Эксперимент проводился на базе Государственного института русского языка им. А.С. Пушкина в феврале 2018 года. В нем приняли участие 78 студентов-иностранцев, учащихся в интернациональных группах уровня B1 (Вьетнам, Китай, Куба, Сербия, Франция и др.) и 7 преподавателей этих групп. Преподаватели и студенты заполняли анкеты одновременно, не имея возможности общаться. 67% студентов отметили, что изучают русский язык уже более 2 лет, 24% – 1-2 года. В качестве опыта чтения на русском языке помимо текстов из учебника были указаны новости на русском языке (62%), русские блоги и социальные сети (39%), специальные книги для чтения (37%), неадаптированная русская литература (28%), тексты по специальности (24%). Эти цифры говорят о том, что студенты уже достаточно активно пользуются аутентичными текстовыми материалами на русском языке.

Материалы для эксперимента. Для проведения эксперимента были отобраны 3 аутентичных текста интернет-издания TheVillage[16], построенные по схожему

[16] http://www.the-village.ru/

образцу: рассказ о необычной профессии. Тексты были предложены модели с минимальной степенью адаптации и оценены ею как A2, B1 и B2. Мы остановили свой выбор на этих трех уровнях сознательно, т.к. именно на них происходит усвоение основных грамматических тем, переход от учебных текстов к аутентичным и текстам с минимальной степенью адаптации. Тема профессий активно изучается на всех уровнях РКИ и актуальна для студентов. Для удобства проведения эксперимента выбран одинаковый объем текстов около 200 слов. В Таблице 15 представлены отрывки этих текстов. Полные тексты представлены в образце анкеты респондента-студента в Приложении А.

Таблица 15 – Фрагменты текстов для эксперимента

Текст	**Уровень (мат. модель)**	**Фрагмент текста**
Текст №1. Фотограф дикой природы	A2	Надо понимать, что фотографы National Geographic не живут в дорогих отелях и не ходят смотреть достопримечательности. Это настоящая работа. Многие люди сидят в офисах и думают, что у фотографа простая и приятная работа: путешествовать, плавать с дельфинами.
Текст №2. Блогер в аквапарке	B1	В аквапарк я приходила в любое время, вход для меня был бесплатный. Самое сложное было уговорить людей участвовать в видео и говорить на камеру. Конечно, это логично: люди пришли отдохнуть, заплатили деньги, может, у них тариф «2 часа 40 минут», и они не хотят это время тратить на тебя.
Текст №3. Технолог на шоколадной фабрике	B2	За это время все пробуют минимум 20 видов разной продукции. Какое количество шоколада в килограммах, я не смогу сказать точно, тем более мы пробуем не только шоколадные изделия, но и вафельные, мармеладные. После каждого кусочка шоколада рот ополаскивается тёплой водой – чтобы нейтрализовать вкус, оставшийся во рту.

Методика эксперимента. Эксперимент включал следующие стадии:

1. Установочная речь. Участникам сообщалось, что предстоит работать с текстами и объяснялась методика работы. Стоит отметить отдельно, что миссия эксперимента сообщалась туманно, «чтобы модель могла подбирать оптимальные для вас тексты», чтобы не акцентировать внимание участников на уровне сложности текстов.

2. Предъявление текстов студентам. Каждому участнику предлагалось прочитать 3 текста в произвольном порядке (тексты в раздаточном материале также располагались в случайном порядке) в комфортном для них темпе, отметить время начала и конца чтения, отметить незнакомые слова, ответить на 3 закрытых и 1 открытый вопрос по содержанию каждого текста. После чтения было предложено выбрать наиболее подходящее из 3 вариантов суждение:

    1. Это простой текст для меня, я знаю почти все слова.
    2. Я знаю не все слова, но общий смысл текста я понимаю.
    3. Это трудный текст для меня, я почти ничего не понял.

3. Предъявление текстов преподавателям этих студентов. Анкета преподавателя включала в себя просьбу отметить лексику, по их мнению, незнакомую студентам и требующую объяснения, отнести текст к одному из уровней владения языком, а также отметить, подходит ли текст этой конкретной группе.

4. Анализ полученных результатов. Качественный и количественный анализ результатов письменных ответов учащихся и преподавателей, статистическая обработка данных.

Анкеты для участников и преподавателей представлены в Приложении А.

Результаты обработки полученных данных представлены в сводной Таблице 16. Совпадение порядка текстов на шкале возрастания трудности подтверждается многими признаками: ростом количества незнакомой лексики, уменьшением скорости чтения и процента успешно выполненных послетекстовых заданий.

Сперва сравним полученные данные с информацией из официальных Требований: норма скорости чтения для B1 при изучающем чтении – 40-50 слов в минуту, при чтении с общим охватом содержания – 80-100 слов в минуту, фактическая скорость чтения соответствует скорее изучающему чтению. Это может быть связано с отсутствием временных границ, студенты читали в комфортном для них темпе. Норма процента лексики, отсутствующей в лексическом минимуме уровня B1, 5-7%, соблюдена в тексте №1 (6%) и превышена для текстов №2 (10%) и №3 (15%). Однако, по мнению студентов, процент незнакомой лексики в предложенных текстах был значительно меньше: от 0% в тексте №1 до 3% в тексте №3. Преподаватели указали чуть больше предположительно незнакомой студентам лексики (от 0% в тексте №1 до 4% в тексте №3), однако в среднем и эти показатели ниже нормативных. Это может связано с рядом причин: словообразовательные навыки помогают студентам понимать значение слов, отсутствующих в лексическом минимуме (слова *фотограф* нет в лексическом минимуме, но *фотография* и *фотографировать* есть); предыдущий языковой опыт, владение английским или другими европейскими языками также помогают учащимся догадаться о значении достаточно трудных слов – *дельфин, нейтрализовать, блогер*; кроме того, всегда остается вероятность несоответствия анкеты реальному положению дел (непонятно задание, забыл/а отметить, постеснялся/лась отметить).

Интересные результаты даёт сравнение незнакомых слов по мнению студентов с данными лексических минимумов. Рассмотрим подробнее данные для лексики текста №3 о технологе на шоколадной фабрике. Всего студентами отмечено незнакомыми 47 слов, однако чтобы отфильтровать возможные частные случаи, мы отберем слова, которые не знакомы более 5% выборки (более 4 человек), получится 30 слов. В лексический уровень B1 не входит 29 уникальных слов этого текста. Однако эти списки не идентичны. Лексический минимум верно идентифицирует 19 слов (65%), отмеченных студентами незнакомыми (например, *вафельный, следить, рецептор*). Следовательно, остальные 35% включены в минимум, но опыт показал,

что студенты часто сталкиваются с проблемами понимания данной лексемы (*обучать, желудок, период и др.*). И наоборот, почти в половине случаев лексический минимум «перестраховывается», считает незнакомыми слова, понимание которых у студентов не вызвало проблем (*молочный, фигура, десерт, меню и др*). Полная статистика результатов ответов участников представлена в Приложении Б.

Таблица 16 – 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%

Количество верных ответов на послетекстовые вопросы также указывает на успешность понимания текстов студентами и выстраивает их в последовательности от простого к сложному. Так, для текста №1 81% студентов дает правильные ответы на 3 из 4 послетекстовых вопросов, у текста №2 этот показатель снижается до 60%, у текста №3 – до 45%.

Рисунок 12 иллюстрирует мнение самих студентов о сложности предложенных текстов. Текст №1 большинством студентов отмечен как простой, это соотносится и с другими параметрами: процентом верных ответов на послетекстовые вопросы, мнением преподавателей. Ситуация с текстами №2 и №3 не так однозначна: в неформальной беседе студенты легко расставляют тексты в порядке усложнения, отдельно отмечая, что текст №3 был очень сложный. Однако в анкете они избегают ответа «Это сложный текст для меня, много незнакомых слов, мне будет трудно пересказать его», что привело к росту второго, менее категоричного варианта ответа. Это явление может быть связано как с объективной оценкой третьего текста как подходящего по сложности, так и психологической нежелательностью третьего ответа, подразумевая неудачу, неуспех студента.

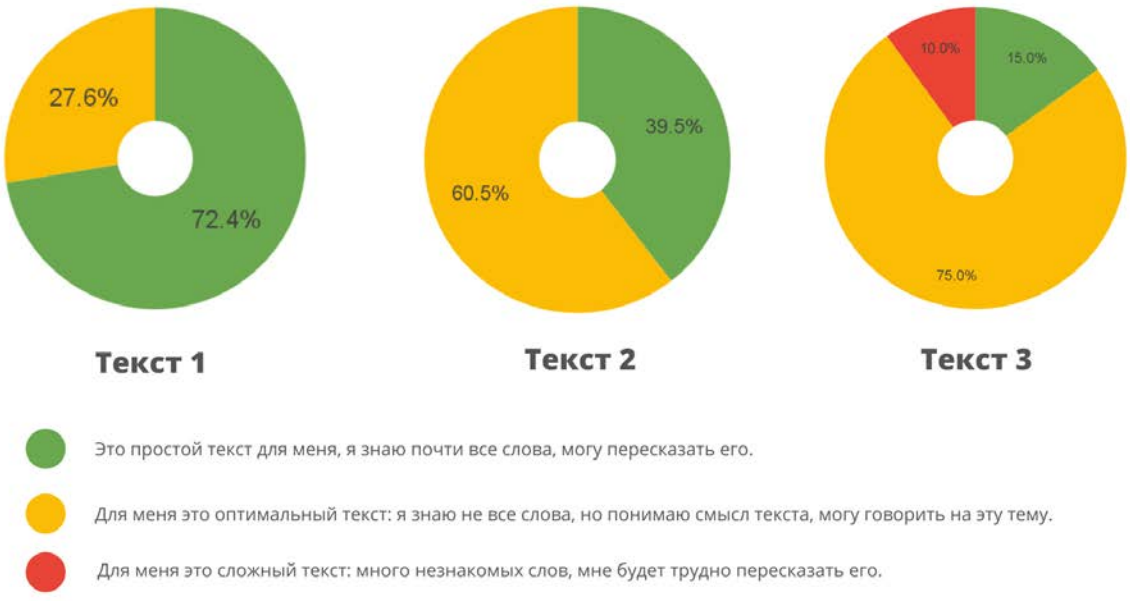


Рисунок 12 – Ответы испытуемых на вопрос об их восприятии сложности предложенных текстов

Далее обратимся к экспертной оценке данных текстов, её результаты также представлены в Таблице 16. Стоит отметить, что мы указываем в таблице усредненную оценку, т.к. мнения преподавателей о сложности одних и тех же текстов порой может сильно разниться. Например, текст №1 получил следующие оценки: A1+(1 оценка), A2(4 оценки), A2+(1 оценка), B1(1 оценка). Вопрос об уровне текста для преподавателей был открытого типа (без возможных вариантов ответа), что привело к нескольким случаям уточнения уровня сложности текста с помощью знаков + («сильный A2, конец курса A2»). Это позволило подтвердить наше субъективное предположение о том, что на практике преподаватели используют более дробную, нежели 6 уровней, шкалу, уточняя таким образом, в какой момент прохождения курса стоит предложить данный текст.

Текст №1 получил идентичные оценки от математической модели и усредненным мнением экспертов. При сравнении оценок текстов №2 и №3 видна тенденция к отнесению преподавателя текстов к более простым, чем определила система. Так, текст № 2 остался в рамках A2, но ближе к началу курса, а текст №3 оценен как B1+, тогда как модель отнесла его к началу B2.

Показателен ещё один комментарий преподавателя: текст №3 был отмечен уровнем B1 (т.е. подходящего уровня сложности для этих студентов), однако трудным для данной группы в силу различий кулинарных традиций: студенты из Вьетнама могут столкнуться с трудностями понимания текста об изготовлении шоколада, вафель и мармелада, так как им с большой вероятностью не знакомы не только лексические единицы, но и сами сладости. Этот факт хорошо иллюстрирует терминологическую разницу сложности и трудности текста: такие параметры труднее формализуются и не учитываются в текущей версии программы, однако являются возможным направлением для дальнейших исследований.

Результаты корреляционного анализа полученных количественных параметров представлены в Таблице 17. Поскольку нас интересовало прежде всего верное расположение выбранных текстов на шкале постепенного усложнения, параметры

скорости чтения, количества отмеченных незнакомыми слов и количества правильных ответов на вопросы были нормализованы для каждого студента. Например, если студент А отмечал незнакомыми в тексте №1 2 слова, в тексте №2 – 4 слова и в тексте №3 – 10 слов, а студент Б – 1, 2 и 5 слов соответственно, нормализованные значения для этих студентов будут одинаковы: 12.5, 25 и 62.5.

Таблица 17 – Коэффициент корреляции наблюдаемых параметров и оценки сложности текста математической модели (жирным выделены значения с p-value < 0.5)

Параметр	**Коэффициент корреляции Спирмена**
Экспертная оценка	**0.72**
Мнение студентов о сложности текста (закрытый вопрос)	**0.49**
Нормализованная скорость чтения текста	-0.56
Нормализованное количество отмеченных незнакомыми студентом слов	**0.89**
Нормализованное количество правильных ответов на 4 вопроса на понимание	**0.77**

Полученные результаты подтверждают гипотезу о связи оценки текста математической моделью и наблюдаемыми в ходе эксперимента параметрами, традиционно применяемыми для оценки понимания текста в иноязычной аудитории. Наивысшие значения корреляции показали параметры количества отмеченных студентами незнакомых слов и относительное количество правильных ответов на вопросы. Скорость чтения текста, напротив, не показала значимой связи с уровнем сложности текста. Это может быть связано с неоптимальной методикой замера скорости чтения с помощью самонаблюдения (студентам самим предлагалось

отметить точное время начала и окончания чтения, что могло привести к погрешностям).

2.4.2. Сравнение с экспертной оценкой сложности текстов

Вторым этапом проверки качества работы модели стало сравнение результатов работы модели с мнением экспертов на более представительном количестве текстов. Для такого сравнения были специально отобраны и размечены 100 текстов из пособий по РКИ, разделов «Чтение» вариантов тестов ТРКИ и аутентичных источников – новостных сайтов, блогов и других СМИ – стандартных источников аутентичного текстового материала. Эти тексты не участвовали в обучении модели. Полученная коллекция текстов позволяет детально проанализировать качество работы модели и масштаб её ошибок.

Отобранные тексты необходимо было оценить с точки зрения сложности вручную несколькими экспертами РКИ. Самым прямым способом получить такую оценку является прямая просьба отнести текст к одному из 6 возможных уровней нескольких экспертов. Однако против этого решения выступали несколько факторов. Во-первых, риск субъективности мнения эксперта заставил бы нас использовать консолидированную оценку по крайней мере 3 экспертов для каждого текста, что представляет собой очень трудоемкую задачу, в первую очередь для экспертов. Во-вторых, полученная нами регрессионная модель представляет сложность как постоянно возрастающее дробное число, поэтому для максимально точного сравнения результатов её работы с мнением эксперта нам необходимо было получить экспертную оценку в таком же формате. Эти факторы обусловили решение использовать методику попарного оценивания текстов на основе рейтингов Эло. Идея рейтингов венгерского математика Арпада Эло состоит в том, что ценность победы того или иного игрока стоит рассчитывать исходя из предсказуемости (т.е. математического ожидания) его победы [Jue Hou et al. 2019]. Эта мера широко применяется для расчёта относительной силы игроков в шахматы. В настоящее время

различные вариации рейтинга Эло используются для самых разных задач ранжирования каких-либо данных: ранжирования уровня сложности заданий в системе адаптивного обучения [Mangaroska et al. 2019], создания шкалы трудности лексических и грамматических тем по РКИ, а также для тестирования учащегося на уровень владения данными темами [Jue Hou et al. 2019], что позволило нам принять решение использовать данную методику для разметки текстов РКИ по уровню сложности с помощью попарного сравнения. Стартовый уровень текстов из учебников или вариантов тестов ТРКИ был равен указанному в их методической справке (A1 = 1, A2 = 2, B1 = 3 и т.д.), все тексты из аутентичных источников получили стартовый уровень 6 (эквивалент уровня C2). Продолжая аналогию с шахматами, стартовый уровень – это «рейтинг игрока в предварительной турнирной таблице», которая показывает, кто тут «гроссмейстер», кто – «новичок», а кто – примерно равные игроки.

Далее происходила сама «партия»: эксперту демонстрировались на экране два текста и предлагалось выбрать, какой из текстов является сложнее. В результате «партии» каждый текст получал баллы:

1 – если текст оказался сложнее;

0 – если текст оказался легче;

0.5 – если эксперт затруднился ответить.

Пример интерфейса для оценивания текстов приведен на Рисунке 13.

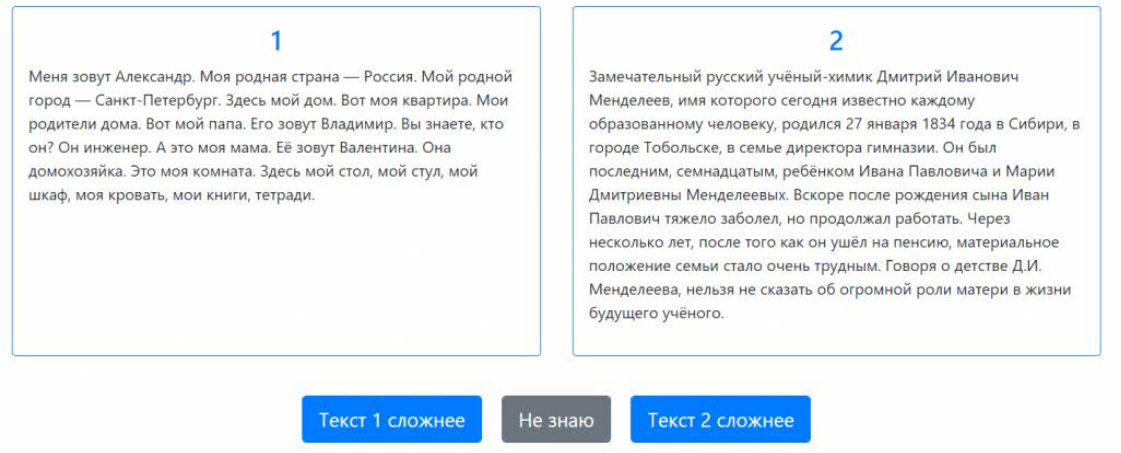


Рисунок 13 – Интерфейс программы для экспертной попарной разметки текстов

Подход получения информации с помощью разметки экспертами зачастую требует проверки валидности ответов: необходимо удостовериться, что аннотаторы правильно поняли задачу и интерфейс страницы разметки, и отвечали внимательно.

В качестве такой проверки нами был разработан алгоритм, который с определенной периодичностью просил оценить пару текстов, где исход сравнения максимально очевиден: с уровнем 1 VS уровнями 5 или 6. Примером таких текстов могут служить отрывки (3) и (4).

(3) Наш город – Петербург. Он не очень старый. Здесь есть новые широкие проспекты и старые узкие улицы. Главная улица – Невский проспект. Он длинный и широкий. Все знают музеи Петербурга: Эрмитаж, Русский музей, Исаакиевский собор. Петербург не очень зелёный город, но здесь есть известные сады и парки: Летний сад, Марсово поле. Российские и иностранные туристы часто гуляют здесь.

(4) Карелия – страна озер. Карелия. Я знаю несколько языков, и, поверьте, ни в одном из них нет слов, чтобы описать удивительную красоту этого места. Нетронутый северный край сочных зеленых лесов, тихих рассветов и белых ночей. Куда не посмотри: везде реки перетекают в озера, а те – снова впадают в реки. Мы с друзьями ходили в Карелии на байдарках. И нам казалось, что есть в этом месте что-то необъяснимое, тайное, может быть даже мистическое.

Аннотатор проходил две такие проверки на каждые 10 ответов, если обе проверки были пройдены, результаты его разметки считались нами корректными и участвовали в оценке уровня сложности текста. Из 102 сессий аннотации 2 ошиблись в обеих проверочных задачах, 17 ошиблись в одной из задач, и, наконец, 83 сессии прошли оба проверочные задания и результаты их разметки были приняты.

Помимо такой проверки на общую корректность ответов, нами также был рассчитан уровень согласия между экспертами во время оценивания одной и той же пары текстов. Процент согласия между аннотаторами на одинаковых парах текстов составил 79%, что позволяет принять полученные ответы в качестве валидных.

В результате попарной экспертной оценки и подсчетов, описанных выше, мы получили коллекцию текстов, плавно распределенных по шкале сложности. Например, на Рисунке 14 представлено, как распределились тексты с одинаковым стартовым уровнем 3 (B1) после оценок экспертами-аннотаторами.

Рисунок 14 – Тексты со стартовым уровнем 3 (B1), расположенные в порядке возрастания своей новой оценки после экспертной разметки

Как видно из Рисунка 14, изменения в уровнях текста нельзя назвать революционными: тексты уровня 3 плавно распределились на шкале сложности от 2.5 до 3.7. На наш взгляд, такая оценка текста более информативна, поскольку позволяет ранжировать тексты по сложности в том числе и внутри одного уровня. Кроме того, подобный взгляд на уровень сложности текста более натуралистично отражает идею освоения языка как постепенно движение от простого к сложному.

Изменились минимальные и максимальные значения уровня сложности. Если раньше коллекция была размечена по шкале от 1 (A1) до 6 (C2), теперь минимальным значением уровня стало 0.9 (см. пример 5), а максимальным – 6.8 (см. пример 6). Таким образом, мы получили образцы текстов, сложных даже для носителей (иронично, но самым сложным текстом в коллекции стал отрывок из закона об образовании).

(5) Стив из Америки. Он хорошо знает английский язык. Это его родной язык. Сейчас он изучает русский язык, но ещё плохо понимает по-русски.

(6) При реализации образовательных программ с применением исключительно электронного обучения, дистанционных образовательных технологий в организации, осуществляющей образовательную деятельность, должны быть созданы условия для функционирования электронной информационно-образовательной среды, включающей в себя электронные информационные ресурсы, электронные образовательные ресурсы, совокупность информационных технологий, телекоммуникационных технологий, соответствующих технологических средств и обеспечивающей освоение обучающимися образовательных программ в полном объеме независимо от места нахождения обучающихся.

Обладая коллекцией текстов, размеченных экспертами, мы можем протестировать качество работы предсказательной модели, сравнив результаты её работы с мнением экспертов. Для оценки качества регрессионной модели, т.е. разницы между фактическими и предсказанными величинами, широко используется расчет корреляции между этими двумя коллекциями данных. Метод вычисления коэффициента корреляции зависит от вида шкалы, к которой относятся переменные. Так, для измерения переменных с количественной шкалой необходимо использовать *коэффициент корреляции Пирсона*. Полученный коэффициент – 0.86 (при p-value < 0.05) позволяет утверждать, что между оценками математической модели и оценками экспертов наблюдается сильная связь, что говорит о достаточной успешности модели.

Для того, чтобы проанализировать результаты сравнения более предметно, необходимо детально посмотреть на случаи несовпадения мнения экспертов с результатом работы математической модели. Очевидно, что критичность ошибки будет зависеть от величины разницы между мнением экспертов и математической модели. Например, ошибка величиной в 0.5 говорит о том, что модель ошиблась на

половину уровня, что мы считаем приемлемым: такая разница возможна и при оценке экспертами (см. пункт про сравнение мнения экспертов). Ошибка же в 1 уровень и более говорит о более серьезных недоработках, на которые стоит обратить внимание.

Для оценки общей величины ошибки прогнозирования была использована метрика *средней абсолютной ошибки*. Для данной коллекции текстов величина средней абсолютной ошибки составила 0.77, что говорит о том, что в среднем модель ошибается в пределах одного уровня. При этом интересно, что 30% случаев модель указывает уровень сложнее, чем назвали эксперты, а в большинстве случаев – в 70% модель указывает уровень сложности ниже, чем оценили эксперты.

Таблица 18 иллюстрирует распределение случаев верных и ошибочных прогнозов в коллекции «золотого стандарта». Разница менее 0.5 рассматривается нами как верный прогноз, таких случаев больше всего, 38%. Разница больше 0.5, но в пределах одного уровня считается нами как прогноз приемлемого качества, и таких случаев 32%. Таким образом, модель дает верный прогноз в 70% случаев, и, соответственно, ошибается больше чем на уровень в 30% оставшихся случаях.

Таблица 18 – Распределение величины расхождения разметки модели и экспертов

Разница между оценкой машинной модели и оценкой экспертов	**Процент подобных случаев на прогоне текстов «золотого стандарта»**
0-0.5 (хороший прогноз)	38%
0.5-1 (приемлемый прогноз)	32%
> 1 (неправильный прогноз)	28%
> 2 (критически неправильный прогноз)	2%

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.

Выводы по главе 2

Данная глава диссертационного исследования содержала поэтапное описание работ по созданию машинной модели для оценки уровня сложности текста и осветила такие этапы, как создание эталонного корпуса текстов из пособий РКИ, размеченных по уровню сложности, извлечение лингвистических признаков, способных оказывать влияние на уровень сложности текста, построение и тестирование модели машинного обучения на полученных данных.

В качестве эталонной коллекции текстов для обучения математической модели, был собран корпус RuFoLa из 802 текстов из пособий и электронных ресурсов по РКИ общим объемом 13 720 слов.

Анализ лингвистических признаков, извлеченных из корпуса текстов RuFoLa, позволил выявить характеристики текстов, наиболее релевантные задаче ранжирования текстов по уровням CEFR. Наивысший коэффициент корреляции показали лексические признаки, основанные на вхождении слов текста в лексические минимумы и некоторые группы грамматических признаков. Среди синтаксических признаков наивысшую корреляцию показывает мера средней длины предложения. Это ставит под вопрос необходимость расчета куда более сложных метрик, основанных на построении синтаксических деревьев для этой задачи. Дискурсивные признаки не показали ожидаемой сильной связи с ростом уровня сложности текстов.

Анализ признаков также выявил особенности, которые мы учитывали при выборе регрессионной модели, такие как мультиколлинеарность и нелинейность признаков.

В ходе тестирования выбранных регрессионных моделей 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.

Разработанная технология оценки сложности текстов была протестирована в ходе эксперимента с группой из 78 студентов. Эксперимент, показал, что модель верно выстроила текстовых 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 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 полученной модели было подтверждено экспериментально. Представляется, что разработанная технология оценки сложности текстов может способствовать повышению объективности процесса оценки уровня текста и сравнимости результатов оценок нескольких текстов между собой.

ГЛАВА 3. ПРАКТИЧЕСКОЕ ИСПОЛЬЗОВАНИЕ РЕЗУЛЬТАТОВ РАБОТЫ ПРЕДСКАЗАТЕЛЬНОЙ МОДЕЛИ В ПРЕПОДАВАНИИ РКИ: СЕРВИС «ТЕКСТОМЕТР»

Для того, чтобы апробировать результаты работы полученной математической модели на широкой аудитории специалистов в области РКИ, а также дать возможность пользоваться ей в своей профессиональной деятельности преподавателям РКИ, был создан открытый веб-сервис «Текстометр»[17].

3.1. Интерфейс и основные возможности сервиса «Текстометр»

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 программный язык Python, для создания пользовательского интерфейса использован программный язык JavaScript. Схема работы сервиса представлена на Рисунке 15.

С точки зрения пользователя сервиса, интерфейс представляет собой окно ввода текста, куда можно вставить любой текст на русском языке от 5 до 10 000 слов, 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 16). Пример полного результата анализа текста в веб-версии сервиса доступен в Приложении В. Кроме того, результат анализа текста доступен для скачивания в текстовом формате TXT для удобства сохранения результата, составления лексического списка по тексту и пр. Пример результат анализа текста в сохраняемой текстовой версии также доступен в Приложении В.

[17] https://textometr.ru

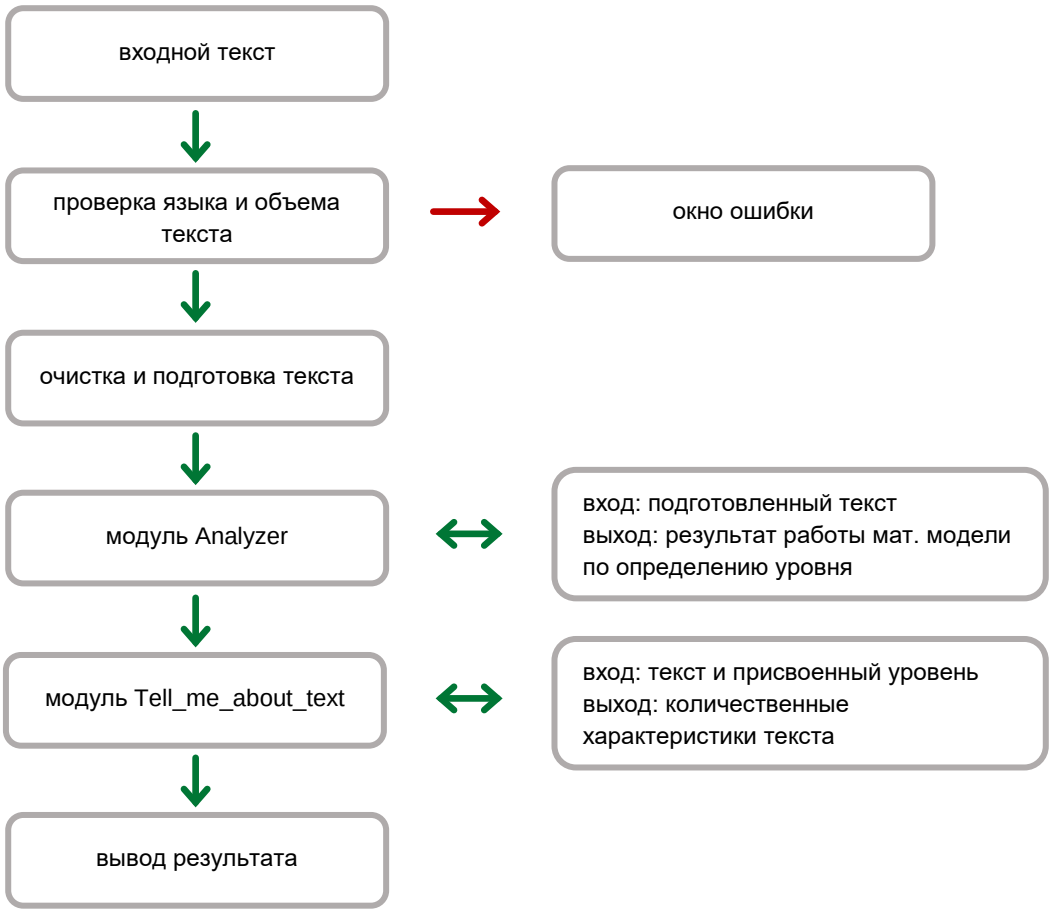


Рисунок 15 – Схема работы модуля по обработке текстовой информации сервиса «Текстометр»

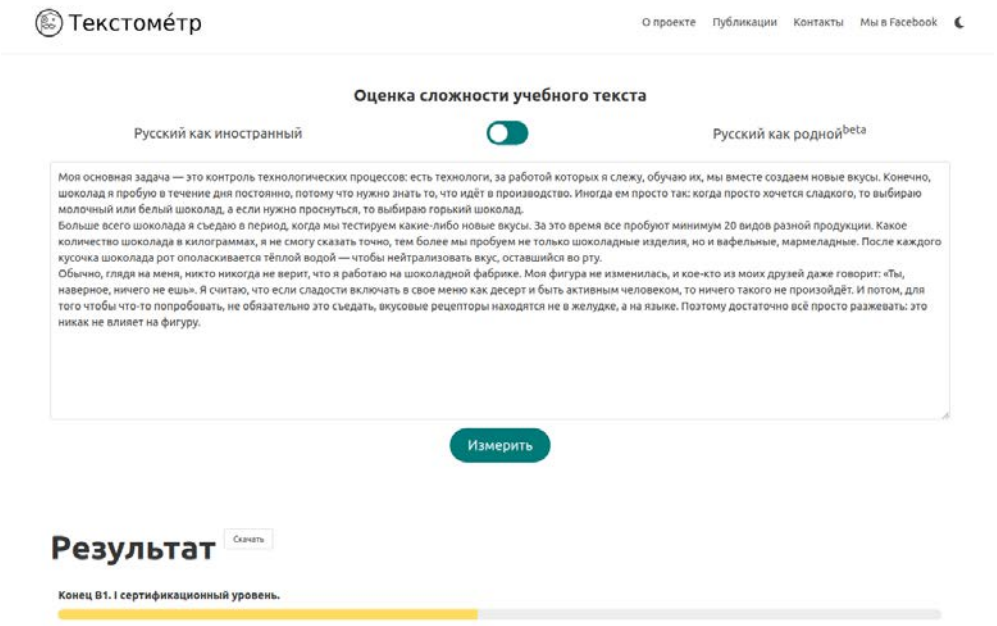


Рисунок 16 – Интерфейс сервиса «Текстометр»

Для корректного автоматического анализа введенный текст проходит несколько этапов предобработки:

1. очистку от всех символов и букв, отличных от русского алфавита (например, чисел, названий компаний на английском языке, элементов верстки);

2. лемматизацию, т.е. приведение каждого слова текста к начальной, словарной форме слова (для подсчетов лексической информации, например, количества уникальных слов);

3. автоматический морфологический анализ (для подсчета грамматических форм, оказывающих влияние на сложность текста, например, форм пассива, причастий, цепочек родительного падежа и мн.др.);

4. проверку полученных списков лексики и фильтрацию слов, отсутствующих в словаре (например, слов с опечатками, сокращений типа *вин.п.*).

Таким образом, 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.

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

Поскольку и авторы спецификаций общеевропейских компетенций отмечают возможность выделения подуровней внутри принятой шкалы, и анализ результатов экспертной оценки в ходе эксперимента (параграф 2.5.1) подтверждает, что в практике преподавания РКИ наблюдается потребность более дробного деления уровней сложности, мы приняли решение интерпретировать результаты работы машинной модели с учетом этой потребности. Поскольку результатом работы регрессионной модели является дробное число, значения десятых долей от 0 до 33 считаются началом уровня, от 34 до 66 – серединой уровня и, наконец, значения выше

67 описываются как конец освоения уровня. Например, результат работы модели 2.2 интерпретируется как начало A2, 3.8 – конец B1.

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. Помимо уровня сложности текста, сервис «Текстометр» предлагает информацию о тексте, представляющую ценность для его подготовки к занятию РКИ: списки ключевых слов, и слов-наилучших кандидатов в словарик 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.

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 показывает значимую корреляцию с уровнем текста, а значит, может свидетельствовать о сложности текста или его отдельных фрагментов. Так, например, большое количество формул читабельности используют эти показатели в качестве основных [DuBay, 2004].

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). Эта мера полезна для оценки повторяемости, воспроизводимости лексики текста и также способна сигнализировать о трудности

текста [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 словами текста называется совокупность нескольких слов или сочетаний, способных дать высокоуровневое описание содержания текстового документа, выявить его тематику. Одним из наиболее часто используемых автоматических методов выделения ключевых слов из текста является вычисление меры TF/IDF [Kaur, Gupta 2010]. Эта мера предполагает подсчет для каждого слова текста его рейтинга: количество раз, которое слово встречается в этом тексте / частота слова по Национальному корпусу русского языка[18] (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.

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

[18] https://ruscorpora.ru/new

уровня, а ниже указывается список слов, не вошедших в лексический минимум данного уровня. Пример такой информации представлен на Рисунке 17.

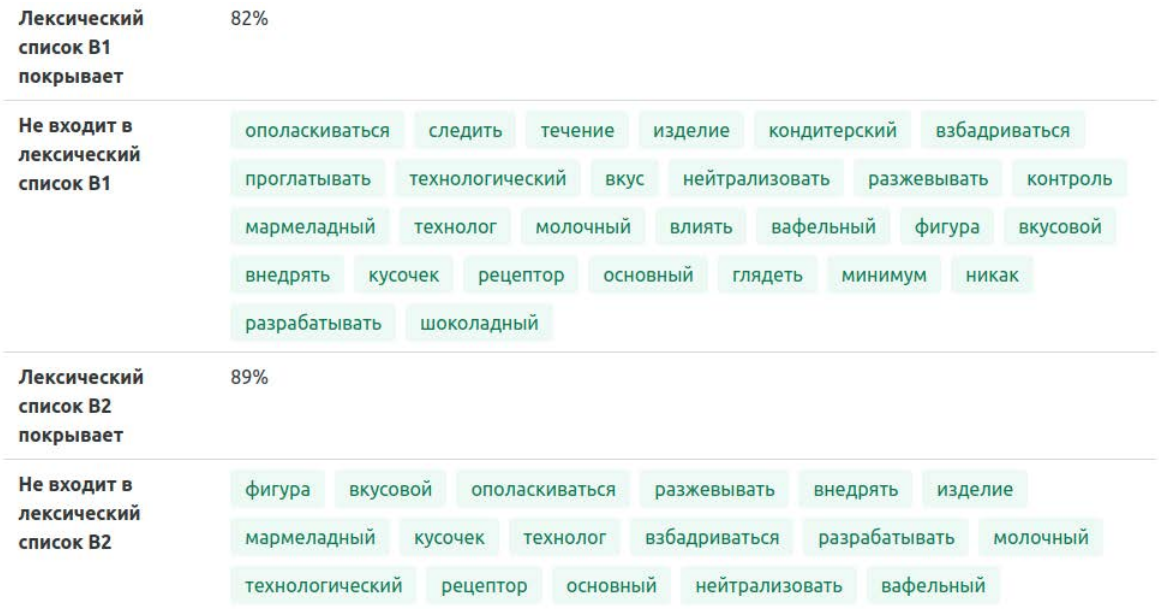


Рисунок 17 – Пример статистики вхождения лексики в лексические минимумы в сервисе «Текстометр»

Количество незнакомой лексики является важнейшим методическим показателем языковой доступности текста: многочисленные исследования говорят о самой тесной связи знакомости лексики текста и успешности его понимания [Nation 2006; Qian 2002]. Официальные Требования по РКИ также содержат информацию о рекомендуемом количестве незнакомой лексики, который постепенно растет от 2-3% для уровня А1 до 10% уровня С1.

Однако лексические минимумы не всегда оказываются информативны для оценки знакомости лексики: во-первых, они ориентированы в первую очередь 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 современным студентам знакома (*смартфон, офис, туалет* и мн.др.). В качестве дополнительного показателя

вероятной знакомости лексики может быть использована частотность слова. Этот параметр широко используется для составления списков и словарей для изучающих русский язык [Sharoff et al. 2013; Лапошина, Лебедева 2019; Система лексических минимумов 2003] и оценки связи знакомости лексики текста и его пониманием [Keskisärkkä 2012; Chen, Meurers 2016].

Для расчета статистики по частотности слов мы использовали Новый частотный словарь современного русского языка (далее ЧС)[19], 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, ещё 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 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 линейки Требований 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

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

ниже B1 нормы скорости чтения не обозначены в Требованиях, поэтому для них используются значения средней скорости чтения, полученные на основе педагогических наблюдений автора на базе Гос. ИРЯ им. А.С. Пушкина. 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, и, наоборот, которая подлежит удалению или упрощению.

Нарративность текста рассчитывается как отношение количества глаголов к количеству существительных и способна показать уровень динамики текста, что также может быть связано с простотой или трудностью его восприятия [Graesser et al. 2011]. Описательность текста демонстрирует количество прилагательных и описательных конструкций в тексте. Для удобства интерпретации эти параметры приведены к шкале от 0 до 10. Высокая степень нарративности в сочетании с низкой описательностью зачастую говорит о сюжетности текста, которая в свою очередь отмечается специалистами как один из важных критериев отбора учебных текстов: ясно выраженная сюжетная линия делает текст удобным для изложения в других формах (замена повествования диалогом и наоборот, пересказ от лица персонажа и т.п.) [Акишина, Каган 2011: 49]. Напротив, высокая степень описательности в сочетании с низкой нарративностью может указывать на потенциально трудный для восприятия текст или фрагмент текста. Значения этих параметров часто являются взаимно обратными (как, например, в примерах 1 и 2) и могут указывать в том числе и на тип или жанр текста: художественные сюжетные тексты отличаются большей нарративностью (пример 1), информационные и рекламные тексты характеризуются низкой нарративностью и высокой степенью описательности (пример 2), проблемные

тексты, тексты-рассуждения и репортажи характеризуются средними показателями обоих параметров (пример 3).

(1) Зимний день... Сильный мороз. Наденька и я стоим на высокой горе. Около нас стоят санки. – Поехали вниз, Надежда Петровна! Только один раз! С нами ничего не случится. Но Наденька боится. Она с ужасом смотрит вниз, и ей кажется, что она умрёт, если съедет с горы. – Я прошу вас! Не бойтесь! Наконец Наденька согласилась, и мы быстро летим вниз. Ветер шумит так сильно, что мы почти ничего не слышим.

Нарративность: 9

Описательность: 1

(2) 9 апреля 2005 года известному русскому художнику, скульптору, теоретику искусства, члену Европейской академии искусств, наук и гуманитарных знаний Эрнсту Неизвестному исполнилось 80 лет. Скульптурные работы Эрнста Неизвестного можно увидеть во многих городах России и за рубежом. Монументальные памятники скульптора стоят на его родине, в городе Екатеринбурге, а также в Магадане, Воркуте, Волгограде... Монумент «Память шахтёрам Кузбасса» стоит в городе Кемерове, где добывают уголь. На одной из московских улиц находится пятиметровая скульптурная композиция «Возрождение» – символ возрождения российской промышленности.

Нарративность: 0

Описательность: 10

(3) Москва всегда считалась очень красивым городом. Она была известна своими белокаменными церквями, старинными особняками. Но в советское время архитектура города сильно изменилась. Нужно было решить жилищную проблему, т. е. обеспечить людей жильём, поэтому строили быстро, часто не уделяя внимания внешнему виду здания. В результате блочно-панельная безликость засорила даже

центр. Перестройка также внесла свой вклад в архитектуру города. В последние годы появилось много зданий с башенками, шпилями, колоннами, которые часто диссонируют с московским стилем. Но это последствия запретов, существовавших в советское время. Тогда бал правил строитель – подешевле, побыстрее и подоступнее. А сегодня бал правит архитектор.

Нарративность: 7

Описательность: 6

После выполнения анализа текста протокол анализа текста со всеми описанными параметрами становится доступен для скачивания в целях удобства сохранения результатов анализа.

3.2. Варианты интерпретации результатов работы сервиса при конструировании учебных текстов (уровень А1)

Конструирование учебных текстов для обучения языку на начальных этапах представляется весьма трудоемким процессом, поскольку автор должен соблюсти требование доступности языкового материала, при этом по возможности предложив студентам связный и интересный материал для чтения и обсуждения. Лексический запас студентов элементарного уровня составляет примерно 730 лексическими единицами (ЛМ). Арсенал грамматических, словообразовательных и синтаксических знаний также весьма ограничен. В этих условиях наиболее важными количественными характеристиками текста можно назвать количество незнакомой лексики, её полезность (т.е. релевантность поставленным целям обучения конкретных учащихся), а также посильность грамматического материала. Рассмотрим подробнее возможные количественные характеристики текста с позиции его подготовки к занятиям РКИ на элементарном уровне на примере анализа реального текста из пособия:

(4) ВЕГЕТАРИАНКА

Меня зовут Миша, у меня есть подруга Маша. Все говорят, что мы хорошая пара. Да, но в последнее время у нас есть небольшая проблема. Уже год моя подруга Маша – вегетарианка. И она не просто не ест мясо, она строгая вегетарианка. Это значит, что она также не ест рыбу. Мясо она и раньше ела редко, не любила она его. А вот рыбу она очень любила! Всегда, когда я приглашал Машу в ресторан, она ела рыбу. Но это еще не все. Она еще не ест яйца! Раньше, когда я приглашал её в гости, мы всегда готовили омлет и вместе его ели. А теперь мы готовим только салаты. Она не ест даже торт, если в нем есть яйца. И теперь, когда мы покупаем торт, она долго смотрит на этикетку – читает, какие в нем продукты. Если там есть яйца или молоко, которое она тоже не пьет, я ем торт один. А я ненавижу есть один! Поэтому обычно мы покупаем еду, которую можем есть оба. Не понимаю! Маша раньше так любила все эти продукты! Как она сейчас может их не есть? Я понимаю, когда женщины не едят некоторые продукты, потому что у них плохая фигура. Но у неё фигура идеальная! Почему она делает это? Это все так глупо! Она говорит, что сейчас, когда она не ест мясо, она меньше устает. Но это абсурд! Конечно, она может жить, как она хочет. Я даже могу приглашать её только в вегетарианские рестораны. Правда, где обедать или ужинать – это не проблема. В ресторане она обычно ест салат, а я ем мясо или рыбу. К счастью, она не говорит, что я тоже, как и она, должен есть только овощи. Да, конечно, я тоже ем фрукты и овощи. Но если я не ем мясо, то я злой! Я не могу есть только овощи! Почему она может?! [20]

[20] Русский язык: 5 элементов: уровень А1 (элементарный) / Эсмантова, Т.Л. – СПб. : Златоуст, 2016. – 320 с.

Таблица 19 – Основные показатели текста из УМК «Русский язык. 5 элементов: уровень А1», полученные с помощью сервиса «Текстометр»

Параметр	**Значение**
Уровень, заявленный в пособии	А1
Уровень, предсказанный моделью	конец А1 (1)
Слов в тексте	296
Уникальных слов	116
Лексическое разнообразие	0.39
Предложений в тексте:	37
Средняя длина предложения:	8
Нарративность текста	8 из 10
Описательность текста	1 из 10
Ключевые слова текста	мясо, вегетарианка, рыба, овощ, торт, яйцо, есть, приглашать, ресторан, подруга
Самые полезные слова	небольшой, последний, злой, просто, если, пара, значить, оба, поэтому, строгий, правда, фигура, то, также, торт, вегетарианка, некоторый, ненавидеть
Лексический список А1 покрывает	83% текста
Не входит в лексический список А1	небольшой, последний, злой, просто, если, пара, вегетарианский, идеальный, значить, этикетка, оба, поэтому, строгий, правда, фигура, то, омлет, также, торт, глупо, вегетарианка, абсурд, некоторый, ненавидеть
Лексический список А2 покрывает	94% текста
Не входит в лексический список А2	фигура, вегетарианка, абсурд, омлет, некоторый, ненавидеть, идеальный, небольшой, этикетка, оба, вегетарианский, глупо, строгий, пара

Лексический список B1 покрывает	95% текста
Не входит в лексический список B1	фигура, вегетарианка, абсурд, омлет, ненавидеть, идеальный, небольшой, этикетка, оба, вегетарианский, глупо, строгий, пара
Лексический список B2 покрывает	98% текста
Не входит в лексический список B2	фигура, вегетарианка, вегетарианский
Лексический список C1 покрывает	99% текста
Не входит в лексический список C1	вегетарианка, вегетарианский
Полезные слова, которых нет в лексическом минимуме	оба, ненавидеть, небольшой
Редкие слова	вегетарианский, омлет
Изучающее чтение займет	10 минут
Просмотровое чтение займет	6 минут
Возможные грамматические темы	местоимения, краткие формы прилагательных и причастий

Таблица 19 демонстрирует показатели текста из УМК «Русский язык. 5 элементов» уровня A1, полученные с помощью сервиса «Текстометр». Видно, что оценка математической модели совпадает с информацией в методической справке пособия. Текст соответствует обозначенному уровню по длине предложений и количеству слов, однако находится в верхних границах нормы, что может говорить в пользу его размещения ближе к концу освоения уровня. Показатель лексического разнообразия показывает крайне низкое значение, причиной чему является многократная повторяемость лексики в тексте. Этот показатель особенно важен при составлении текстов для изучающего чтения на начальных уровнях владения русским языком, т.к. является одним из приемов эффективного усвоения и запоминания новых лексических единиц.

Высокий уровень нарративности текста и, напротив, низкая описательность, характерны для простых текстов бытовой направленности и показывают нормальные значения этих параметров для текстов начальных уровней владения русским языком.

Ключевые слова текста, принцип отбора которых связан с поиском в тексте лексем, частотность которых в Национальном корпусе русского языка значительно ниже, чем в предлагаемом тексте, на учебных материалах элементарных уровней часто иллюстрируют «проблему апельсинов и бананов» [Kilgariff 2010], которая заключается в том, что повседневная лексика, или «лексика выживания», не всегда часто появляется в письменных текстах и может оказаться не частотной. Так, в список ключевых слов рассматриваемого текста попали практически все упоминаемые в нем продукты питания (*мясо, рыба, овощ*).

В связи с этим при работе с текстами начальных уровней представляется более эффективным обращаться к списку т.н. «полезных» слов, которые составляют слова, не вошедшие в лексических минимум данного уровня (т.е. предположительно незнакомые студентам при чтении), но либо входящие в лексические минимумы следующих уровней, либо получившие высокую оценку частотности в корпусе русского языка. Таким образом, эти слова являются кандидатами для дальнейшей лексической работы отработки и запоминания. Имеет смысл оценить, насколько представленный список лексики совпадает с поставленными учебными целями.

Особенно полезным представляется анализ лексики, попавшей в разряд редкой. Попадание в этот список не всегда является сигналом к удалению слова: например, в разряд редких попало слово *вегетарианка*, являющееся названием текста и одним из ключевых понятий текста, кроме того, толкование слова объясняется в тексте. Однако, это поле демонстрирует пути возможного упрощения неключевой лексики текста через замену на более простые синонимы, толкование слова или все-таки его удаление (например, слова *этикетка* или *абсурд*).

Покрытие текста списками лексических минимумов по мере возрастания уровня ожидаемо растет, однако для заявленного уровня А1 оно составляет 83 процента, что

подразумевает 17 процентов незнакомой лексики, тогда как требования к уровню А1 говорят о 2-3 процентах незнакомой лексики. Такое серьезное несовпадение требует более детального анализа. Самой первой причиной такого несовпадения может будет нерегулярная представленность в лексических минимумах ТРКИ словообразовательных моделей. Это приводит к тому, что лексемы *небольшой, просто, глупо* помечаются незнакомыми, хотя их дериваты (*большой, простой, глупый*) присутствуют в лексическом минимуме указанного уровня. Вторая причина может крыться в особенностях самого пособия: поскольку автор адресует его аудитории автор намеренно включает слова, имеющие схожие аналоги в европейских языках: *вегетарианский, пара, фигура, идеальный, омлет, абсурд, этикетка.* Однако для остальных групп студентов эти слова могут представлять серьезную трудность, поскольку некоторые из них отсутствуют даже в лексическом минимуме С1.

Вычет двух указанных групп лексики из разряда незнакомой дает нам 6% процентов незнакомой лексики, что все еще сильно выше норм, заявленных в Требованиях. Так, среди незнакомой лексики содержатся высокочастотные слова, которые, однако, появляются в лексических минимумах позже, на уровнях А2 (*если, поэтому, также, то, строгий*) или даже В2 (*ненавидеть, оба*).

Таким образом, приведенный детальный анализ лексики позволяет, во-первых, обозначить целевую аудиторию текста с точки зрения языковой сложности: учащиеся, завершающие прохождение уровня А1, имеющие в запасе знание какого-либо европейского языка, позволяющего догадаться о значении интернациональных слов, а также способные производить базовый словообразовательный анализ слова. Остальная лексика, попавшая в разряд незнакомой, входит в лексические минимумы следующих уровней, т.е. данный текст предлагает лексический материал на опережение. Ключевыми лексическими единицами, которые студент отработает в предложенном тексте, будут продукты питания, обозначение различных пищевых привычек и глагол *есть*.

3.3. Варианты интерпретации результатов работы сервиса при создании и анализе контрольно-измерительных материалов (уровень А2)

При создании и валидации контрольно-измерительных материалов особое внимание уделяется строгому соответствия материала уровню сложности, а также сравнимости лингвистических характеристик материалов разных вариантов. Эта проблема особенно актуальна для материалов начальных уровней, где необходимо обеспечить вариативность материала задействуя весьма ограниченный лексико-грамматический арсенал.

Одним из этапов проверки текстовых материалов здесь может стать сравнительный анализ лингвистических показателей отобранных текстов-кандидатов. Так, субтест Чтение типовых тестов ТРКИ уровня А2 [Антонова и др. 2019] содержит 3 типа текстов в зависимости от установки. Так, первое задание связано с пониманием основной идеи объявлений и кратких анонсов и соответствует ознакомительному чтению с общим охватом содержания. Второе задание представляет собой более объемные фрагменты текстов одной тематики (в первом варианте это рецензии на фильмы, во втором – несколько мнений граждан по одному вопросу социологов), понимание которых проверяется достаточно детально с помощью 15 вопросов, что может соответствовать изучающему чтению. Наконец, третье задание содержит объемный текст с описанием биографии или хронологии событий с установкой понять основную информацию текста и значимые детали, что также соответствует ознакомительному чтению с более детальным анализом содержания. Таблица 20 содержит лингвистические характеристики текстов трех типов, включенных в первый (светлые ячейки) и третий (затемненные ячейки) варианты теста.

Таблица 20 – Сравнительные характеристики текстов разных типов из двух вариантов тестов ТРКИ базового уровня А2

Тип чтения	**ознакомительное (задания 6-10)**		**изучающее (задания 11-25)**		**ознакомительное + значимые детали (задания 25-30)**	
Вариант	**1**	**3**	**1**	**3**	**1**	**3**
Оценка модели	1.6	1.4	1.3	1.4	1.5	1.1
Средняя длина предложения	9.8	5.8	9.1	11.8	14.5	10.8
Лексическое разнообразие	0.69	0.9	0.46	0.55	0.46	0.48
Процент лексики не входящей в А2	24	13	6	8	19	9
Нарративность текста	3	2	8	6	6	6

Общая оценки сложности текстов моделью соответствует уровню А2, что согласуется с мнением экспертов-составителей. Интересно проследить изменение сложности текстов в пределах этого уровня, в зависимости от типа текста и установки на чтение. Так, можно ожидать, что тексты для изучающего чтения должны содержать меньше незнакомой лексики и желательно иметь высокую степень нарративности для удобства чтения, тексты же первого типа с установкой понять основную идею, могут иметь чуть более сложную структуру и лексический состав. Действительно, тексты объявлений (первый тип заданий) характеризуются сравнительно короткими конструкциями, высокой степенью лексического разнообразия (скорее всего, это связано с маленьким объемом текстов), значительным количеством незнакомой лексики и низкой нарративностью. Тексты для изучающего чтения, напротив, содержат минимальный из наблюдаемых процент незнакомой

лексики и обладают высокой нарративностью. Тексты третьей группы занимают срединную позицию по этим показателям.

Вторым важным моментом при анализе контрольно-измерительных материалов может стать равномерность и сравнимость сложности текстов разных вариантов. С этой точки зрения выделяются тексты разных вариантов третьего типа: судя по формальным показателям, можно предположить, что текст для первого варианта (Фабрика «Эйнем») превосходит по сложности текст третьего варианта (биография Д. Хворостовского). Это выражается как в общей оценке текстов моделью, так и в ряде отличий лингвистических характеристик текста, таких, например, как средняя длина предложения и процент незнакомой лексики.

Таким образом, предложенный анализ тестовых материалов подтверждает мнение экспертов об их уровне сложности, демонстрирует небольшую вариативность в показателях и аспектах сложности текстов в зависимости от установки на чтения, и, наконец, указывает на разницу в сложности текстов третьего типа в разных вариантах теста, что может послужить поводом обращения дополнительного внимания методистов на выявленную пару текстов.

3.4. Возможности самостоятельной работы студентов с сервисом «Текстометр» на примере нехудожественных текстов для экстенсивного чтения (уровни B1 и B2)

Экстенсивное домашнее чтение на изучаемом языке признается специалистами мощнейшим фактором прогресса обучения, источников обогащения словарного запаса и арсенала лексико-грамматических средств [Krashen 2011]. Кроме того, напомним, что в ходе анкетирования участников эксперимента, описанного в параграфе 2.5.1, 62% испытуемых в качестве источника читательского опыта помимо учебника отметили новости на русском языке, 39% – русские блоги и социальные сети. Эти варианты опередили специальные книги для чтения, которые оказались лишь на третьем месте с 37% участников.

При этом самостоятельный поиск и выбор материала для домашнего чтения способствует активизации роли учащегося в образовательном процессе и отвечает концепции персонификации обучения. Наиболее релевантными данной задаче самостоятельного отбора материалов для экстенсивного, досугового чтения видятся возможности сервиса «Текстометр», связанные с выделением ключевых слов текста и указание ориентировочного уровня сложности текста.

В частности, изучение списка ключевых слов помогает учащемуся оценить, во-первых, соответствует ли текст области его интереса, во-вторых, оценить, насколько он знаком с ключевой лексикой для понимания данного текста и, при желании, посмотреть заранее их толкование, т.к. экстенсивное чтение не предполагает детальной проработки лексики по ходу чтения. При этом выделение ключевых слов может оказаться наиболее полезным при выборе оптимального источника информации из нескольких текстов на определенную тему. Например, в Таблице 21 представлены ключевые слова и предполагаемый уровень 4 различных текстов крупных интернет-изданий, предложенных по ключевым словам «изменение климата». Как видно из таблицы, они отличаются не только по уровню сложности, но и по специфике раскрытия этой темы: в материале Ленты.Ру на первое место выходит экономическая составляющая, Интерфакс говорит о политических процессах, тогда как BBC News и N+1 фокусируются на собственно экологической и биологической проблематике.

Вторым источником информации, на который стоит обращать внимание учащимся при самостоятельной работе с сервисом «Текстометр», является предполагаемый уровень сложности текстов. На основании этой информации и предыдущего опыта оценивания и чтения текстов учащийся сможет подобрать материал оптимальной сложности, а также отслеживать свой личный прогресс и выстроить траекторию постепенного усложнения материалов для чтения.

Таблица 21 – Сравнительный анализ уровня сложности и наборов ключевых слов текстов российских СМИ

Заголовок и источник текста	**Уровень**	**Ключевые слова**
Фестивали в разных странах, Chip Trip Travel Blog	конец B1	фестиваль, гуляние, массовый, житель, парад, костюм, пиво, проходить, открытие, праздновать, радость, зрелище, регата, томат, друг, начинаться
«Норникель» сообщил об «унаследованных» со времен СССР свалках в Арктике, РБК Бизнес	середина B2	вице-президент, мера, сооружение, мусор, очистка, переработка, ликвидация, забросить, отход
Археологи нашли в Южном Туркменистане поселение цивилизации Окса, N+1	конец B2	археолог, эра, ученый, поселение, артефакт, гектар, бронзовый, южный, древний, обнаруживать, изделие, участок, сосуд
Кремль в ответ на критику Байдена отметил последовательность РФ в вопросах климата, Интерфакс	конец B1	климат, мера, энергобаланс, мероприятие, параметр, вклад, постановка, лесной, конференция, климатический, советник
Глобальное потепление в 100 словах, BBC News	начало B2	климат, океан, ледник, климатический, глобальный, кислотность, потепление, среднее, таяние
Компании обязали отчитываться о борьбе с изменением климата, Лента ру	середина B2	выброс, альянс, отчитываться, угроза, сектор, триллион, достигать, инфраструктура, глобальный, предстоять, последствие
Изменение климата повысило риск красных приливов в Чукотском море, N+1	конец B2	море, цист, клетка, концентрация, вегетативный, размножение, расчет, массовый, температура

3.5. Варианты интерпретации результатов работы сервиса при подборе фрагментов аутентичных художественных текстов (уровень B2)

Чтение художественных произведений на языке оригинала может быть обусловлено как желанием студентов познакомиться с российским культурным контекстом, так и закономерным стремлением изучать русский язык на актуальном аутентичном материале. Кроме того, освоение художественных текстов повествовательного характера отдельно указывается в Требованиях начиная со второго сертификационного уровня владения русским языком, B2 [Требования 2015: 12]. Безусловно, аутентичный художественный текст представляет собой особый целостный продукт, трудность восприятия которого не ограничивается лингвистической сложностью материала. Более того, она в более значительной степени, чем в случае с информационными или новостными текстами, зависит от читательской подготовки учащегося и проделанной работы с текстом: детальной беседы по тексту, толкования лексики, составления схем персонажей и т.п. [Кулибина 2001; 2002]. Однако, с другой стороны, посильность языкового материала художественного текста и его релевантность целям учащегося все же остается одним из важнейших критериев отбора текстовых материалов [Акишина, Каган 2011: 56]. В качестве иллюстрации возможностей сервиса по предварительному отбору авторов и фрагментов художественных текстов, подходящих студентам по уровню лингвистической сложности, приведем сравнительный анализ количественных характеристик текстов современной российской прозы (конца XX в. – начала XXI в.). Отбор произведений для анализа проводился по нескольким основаниям: место в рейтингах издательств и книжных магазинов[21], наличие литературных премий, востребованность авторов за рубежом, а также оценка профессионального сообщества (опрос среди преподавателей). На основе данных критериев был

[21] Список топ-100 российской современной прозы от портала livelib.ru: https://www.livelib.ru/genre/prose/top [дата обращения: 10.05.2020]; списки бестселлеров магазина litres.ru: https://www.litres.ru/kollekcii-knig/bestsellery-litres/ [дата обращения: 10.05.2020]

сформирован список из 12 произведений. Список всех проанализированных произведений представлен в Таблице 21.

Художественный текст зачастую оказывается неоднороден по сложности: диалогические фрагменты обычно содержат более простую лексику и синтаксис, по сравнению с описательными частями. Для того, чтобы получить более объективную общую оценку сложности произведений, мы разделили их на одинаковые фрагменты по 5 000 знаков (около 700 слов). В дальнейшем все расчеты приведены как среднее значение по всем фрагментам данного произведения. При этом внутри одного произведения сложность может варьироваться: например, в сборнике повестей Наринэ Абгарян со средним значением 3.9 фрагмент с минимальной сложностью составляет 3.1 (начало B1), а максимальной – 5.7 (C1). Последовательный анализ фрагментов художественного произведения сервисом может помочь отобрать «кандидатов» посильной языковой сложности.

В Таблице 21 представлены некоторые усредненные характеристики текста, способные оказывать влияние на уровень его сложности. Средняя длина слова и предложения традиционно используются для подсчетов сложности текста. В качестве примера грамматических признаков, указывающих на усложнение текста, в таблице приводится количество причастий и деепричастий. Доля слов в тексте, присутствующая в лексических минимумах по ТРКИ показала одну и самых сильных предсказательных способностей в задаче определения уровня текста по РКИ, в качестве примера расчетов такого типа в Таблице 22 приведены значения доли лексики, присутствующей в ЛМ уровня B2 [Лексический минимум 2017b].

Во-первых, отметим, что поскольку математическая модель обучалась на корпусе, где только 15% текстов были художественными, нельзя гарантировать корректный результат её работы в определении уровня художественного текста. Однако его можно рассматривать в качестве ориентира и аргумента в процессе выбора автора, произведения, конкретного фрагмента текста или порядка предъявления фрагментов учащимся.

Таблица 22 – Усредненные лингвистические показатели сложности текстов российской современной прозы

Автор	**Название произведения**	**Уровень сложности (усредненный)**	**Уровень по системе CEFR**	**Средняя длина слова**	**Средняя длина предложения**	**Доля лексики из ЛМ B2**	**Среднее кол-во прич. и дееприч. на фрагмент**
Павел Санаев	Похороните меня за плинтусом	3.4	B1	5	8.4	84	4
Дмитрий Глуховский	Текст	3.7	B1	5	7.1	82	6
Евгений Водолазкин	Авиатор	3.9	B1	5.2	9	84	10
Сергей Лукьяненко	Ночной дозор	3.9	B1	5.2	7.5	80	9
Наринэ Абгарян	Люди, которые всегда со мной	3.9	B1	5.2	9.6	78	6
Виктор Пелевин	Рассказы	4	B2	5.4	12.7	81	11
Алексей Иванов	Географ глобус пропил	4	B2	5.1	8.3	76	6
Дина Рубина	На солнечной стороне улицы	4.1	B2	5.3	12.3	78	9
Гузель Яхина	Зулейха открывает глаза	4.3	B2	5.4	9.7	74	10
Татьяна Толстая	Сборник рассказов Река	4.4	B2	5.3	16.5	78	12
Борис Акунин	Азазель	4.5	B2	5.4	9.8	75	5
Людмила Улицкая	Казус Кукоцкого	4.6	B2	5.7	16.8	79	12

Данные Таблицы 22 также иллюстрируют, что источниками сложности текста могут стать разные его аспекты: так, например, в произведениях Т. Толстой и Л. Улицкой сложность представляют прежде всего длинные предложения, которые свидетельствуют о синтаксической, структурной сложности текста. Эта гипотеза подкрепляется и сравнительно большим количеством причастных и деепричастных форм глагола в данных произведениях. Тексты же Г. Яхиной и Б. Акунина, напротив, написаны короткими предложениями, но с бóльшим количеством незнакомых студентам слов, то есть здесь речь в первую очередь идет о лексической сложности. Лексическая сложность текста представляется нам более неоднозначным параметром, нежели синтаксическая, структурная сложность, так как она во многом зависит от страны и родного языка учащихся, близости тематики произведения интересам читателя и читательского опыта учащегося.

Таблица 23 – Анализ лексики первых фрагментов выбранных произведений, не вошедшей в лексический минимум уровня B2

Автор	**Не входит в ЛМ B2 (слов)**	**Из них редких слов (количество: примеры)**	**Из них полезных слов (количество: примеры)**
Павел Санаев	89	37: отсыхать, *стопроцентный*, лязгнуть, прислоняться, рейтузы, гайморит, щиколотка, журчать, *проинструктировать*, смердячий	19: довольно, значительный, пожалуй, процедура, увы, остальное, вновь, понадобиться, уверять, насчет, проклятый
Гузель Яхина	140	57: войлок, подмерзать, *молочно-белый*, прошмыгивать, половица, пиала, палас, *полжизни*, исподний, *январский*, хлев	78: чуть, дух, главное, спустя, многочисленный, едва, опираться, едва, порог, биться, полагаться, слышно, конь, ледяной, тронуть
Татьяна Толстая	129	51: сдвигаться, *сливочный*, зятек, *крашеный*, крем-брюле,	71: зато, пространство, специально, едва, делаться,

		центрифуга, несравненный, акриловый, *ягодка*, зашевелиться, зазор, ороговеть	справедливость, подробность, заодно, ленинградский

Подобный анализ лексики, предположительно незнакомой студентам (не вошедшей в ЛМ уровня В2), представлен в Таблице 23 на материале первых фрагментов отобранных произведений (отрывки около 700 слов). Он позволяет увидеть не только количество потенциально сложной лексики, но и ее коммуникативный потенциал, вероятность, что она пригодится студенту в дальнейшей учебной деятельности или общении. Так, в произведении П. Санаева меньше всего незнакомых слов, однако и самое низкое число потенциально полезных слов. Среди редких слов присутствуют жаргонизмы и сниженная бытовая лексика, причем частично уже устаревшая. Роман Г. Яхиной изобилует специфической лексикой, связанной с бытом татарской деревни (*войлок, хлев, палас, пиала*). Бóльшая же часть редкой лексики рассказов Т. Толстой стилистически нейтральна. Мысль о том, что предшествующий читательский опыт является важным фактором расчета сложности текста для конкретного учащегося подкрепляется и расчетами: около 500 лексических единиц, характерных для художественного повествования (*вновь, едва, броситься, застыть, судя*), отсутствуют в ЛМ В2 и С1, однако встречаются почти во всех выбранных произведениях.

Представленный анализ произведений современной российской прозы с помощью веб-сервиса «Текстометр» позволяет ранжировать тексты в зависимости от сложности и подобрать оптимальный материал для чтения в конкретной аудитории. Предлагаемая система оценивания языкового материала может быть использована не только для выбора произведения, но и для помощи в поиске подходящего фрагмента выбранного текста.

3.6. Возможности работы с сервисом для специалистов по адаптации и интеграции иностранных граждан

Проблема языковой сложности тех или иных текстов может рассматриваться не только с позиций учебной деятельности, но и с точки зрения доступности информации на русском языке для иностранных граждан. Так, например, в работах отмечается, что успешная интеграция мигрантов невозможна без доступа к информации: возможности получить корректную информацию о законодательстве страны пребывания, механизмах правовой поддержки, культурных особенностях и нормах поведения, о рабочих местах и вакансиях и т.д. [Якимов 2018]. В соответствии с основными потребностями мигранта в научной литературе [Мукомель 2016] выделяют следующие направления адаптации и интеграции:

1. правовая адаптация/интеграция (обеспечение доступа к легальному статусу);
2. экономическая адаптация/интеграция (доступ на рынок труда, обеспечение занятости, доступ к жилью);
3. социальная адаптация/интеграция (доступ к жилью, здравоохранению, образованию и др.);
4. культурная адаптация/интеграция (доступ к изучению языка, других элементов культуры при возможности сохранения своей культуры, религии, обычаев).

Каждый из этих аспектов интеграции предполагает понимание определенного набора текстов: правовых текстов и инструкций на сайте государственных услуг для иностранных граждан, навигационной информации в общественном транспорте, вакансий, трудового договора, объявлений об аренде жилья, договора найма жилья, рекламных проспектов кафе и ресторанов, развлекательных и культурных мероприятий и мн.др. Однако на практике такие тексты зачастую оказываются для иностранных граждан недоступными из-за своей языковой сложности. Эффективный инструмент оценки таких текстов может помочь специалистам по работе с иностранными гражданами проверить адресованные им тексты на языковую доступность. В качестве иллюстрации в Таблице 24 приведены значения трудности

некоторых текстов, от понимания которых зависит доступность легализации нахождения иностранного гражданина в России, соблюдение местных норм и законов, уровень комфорта и включенности в жизнь города.

Образцы инструкции для получения вида на жительство были получены со специального раздела сайта Госуслуги для иностранных граждан[22]. Расчеты для коротких жанров, отмеченных в таблице знаком *, были рассчитаны как усредненное значение по 20 случайно отобранным образцам жанра в сети Интернет.

Таблица 24 – Сравнительные характеристики языковой сложности бытовых и правовых текстов для обеспечения интеграции иностранных граждан

Источник текста	**Уровень**	**Средняя длина предложения**	**Средняя длина слова**	**Процент лексики из ЛМ B2**
Текст из теста на РВП	начало B1	5.9	6	83%
Отзывы на рестораны*	начало B1	6.2	5.2	87%
Инструкция по аренде электросамоката	середина B1	5.1	5.4	77%
Объявления о сдаче квартиры*	середина B1	9.3	5.9	79%
Объявления в московском транспорте*	конец B1	10.7	7	88%
Договор медицинского страхования	конец B2	13.6	7	69%
Инструкция для получения вида на жительство	конец B2	30	7.1	81%
Типовой брачный договор	начало C1	16.9	6.7	71%
Типовой договор на аренду квартиры	середина C1	9.4	6.3	63%
Правила пользования московским метро	начало C2	10.2	6.8	67%

[22] Госуслуги: https://www.gosuslugi.ru/foreign-citizen [дата обращения 01.11.2021]

Приведенный сравнительный анализ иллюстрирует проблему нехватки ожидаемого уровня владения русским языком иностранных граждан (в таблице представлены характеристики текста из тренировочных тестов на получение разрешения на временное пребывание) для решения задач интеграции и легализации в России. Так, становится очевидной острая необходимость адаптации инструктивных текстов на странице Госуслуг для иностранных граждан: написанные сложными синтаксическими конструкциями, с обилием терминов и абстрактных слов, они имеют уровень сложности значительно выше требуемого от претендентов на легализацию в России во время сдачи теста по русскому языку. Такие тексты рискуют остаться непонятыми, а рекомендации, изложенные в них – невыполненными.

Среди самых сложных правовых текстов достаточно закономерно оказываются несколько типов договоров (договор медицинского страхования, брачный договор, договор аренды квартиры). Поскольку подобные тексты не всегда могут быть изменены из-за обязательных юридических формулировок, представляется крайне важным создание адаптированных версий подобных документов с толкованием основных формулировок и наиболее важных пунктов договора для того, чтобы иностранный гражданин мог заранее подготовиться к чтению подобных документов.

Самым сложным текстом из рассмотренных оказались правила пользования московским метрополитеном[23]. Обилие сложной абстрактной лексики, длинные конструкции, сложность структуры документа значительно затрудняют его понимание. Однако стоит отметить положительный опыт адаптации отдельных пунктов правил и их оформления в виде информационных плакатов с иллюстративным материалом, что значительно повышает их доступность. Для иллюстрации подобного опыта приведем пункт правил пользования московским

[23] Правила пользования московским метрополитеном: https://mosmetro.ru/passengers/information/rules/ [дата обращения: 01.11.2021]

метрополитеном (пример 5) и соответствующий ему информационный плакат (Рисунок 18).

(5) Багаж, сумма измерений которого по длине, ширине и высоте находится в пределах от 121 см до 150 см, длинномерные предметы, длина которых от 151 см до 220 см, оплачиваются отдельно за каждое место. Количество мест багажа, разрешенного к провозу, не должно превышать двух мест на одного пассажира.

Рисунок 18 – Информационный плакат московского метрополитена

3.7. Ограничения работы сервиса

Не менее важным, чем описание вариантов и методов использования сервиса, нам представляется обозначение границ его возможностей. Первое и самое важное ограничение работы алгоритма заключается в том, что он ориентирован на определение сложности законченного фрагмента прозаического текста. При анализе поэтических произведений, лексико-грамматических упражнений, списков слов адекватная оценка уровня сложности алгоритмом не может быть гарантирована. Это связано с тем, что в таких текстах представлены принципиально другие

синтаксические конструкции (или их полное отсутствие, как в случае со списком лексики), поэтому адекватная оценка уровня сложности таких материалов алгоритмом не может быть гарантирована. Впрочем, это не мешает использовать для данных материалов остальные статистические характеристики, связанные с расчетами вхождения лексики в лексические минимумы и частотные списки.

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.

Автоматизация подсчета процента незнакомой лексики текста выявила несколько вопросов, связанных недостаточной методической разработанностью этой темы. В частности, мы не смогли найти однозначного ответа на вопрос о методике учета дериватов от слов, входящих в лексический минимум и образованных по словообразовательным моделям, доступным на данном уровне владения языком (например, *город-городок, обычно-необычный* и т.д.). Отсутствие таких слов в лексических списках приводило к большому количеству лексики, помеченной незнакомой, которая при ручном подсчете скорее всего была бы отнесена экспертом к разряду знакомой. Частично это проблема была решена с помощью ручного расширения лексических списков и включения в них указанных групп дериватов.

Также сервис не застрахован от единичных ошибок на этапе автоматического морфологического анализа текста модулем Mystem: так, иногда морфологический анализатор ошибается в приведении к начальной форме слова (*воскресенье* приводит к форме *воскресение*); также он приводит все глаголы к форме инфинитива

несовершенного вида: начальная форма от *узнали* помечается как *узнавать*, а не *узнать*. В большинстве случаев это не оказывает серьезного влияния на результат, т.к. в лексическом минимуме указаны обе видовые формы, однако в ряде случаев это оказывается критичным, если в минимуме присутствует только одна форма. Анализатор может распознать географические названия и имена собственные и не учитывает их при подсчете незнакомых слов, однако не всегда способен отфильтровать названия компаний (*Ореанда, Уют*).

Выводы по главе 3

Результат работы математической модели по определению уровня сложности текста были представлены в виде открытого веб-сервиса «Текстометр». В зависимости от уровня текста и целей обращения к сервису меняется информация, на которую пользователю сервиса рекомендуется обратить свое внимание. 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 т.з. лексической **и синтаксической сложности.**

Самостоятельная работа студентов с сервисом может заключаться в поиске источников материалов для экстенсивного домашнего чтения, сообразного интересам конкретного учащегося. Наконец, сервис может быть использован и вне контекста учебной деятельности, например, для оценки доступности информации инструктивного, правового или рекламного характера **для иностранных граждан.**

Сервис также имеет ряд ограничений работы, первое и самое важное из которых заключается в том, что он ориентирован на оценку целостного фрагмента

прозаического текста. При анализе поэтических произведений, лексико-грамматических упражнений или списков слов адекватная оценка уровня сложности алгоритмом не может быть гарантирована.

ЗАКЛЮЧЕНИЕ

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.

Отбор текстов для учебных пособий, поиск и оценка сложности актуальных неадаптированных текстов для аудиторной работы или самостоятельного чтения отмечается многими исследователями как насущная задача современной методики преподавания русского языка как иностранного. Вместе с тем, одним из важнейших критериев отбора текста признается его доступность и посильность, оптимальное соответствие уровню владения русским языком учащимся. Практическая необходимость оперативной и объективной оценки сложности текстов по этому критерию широкого круга специалистов – преподавателей, методистов, редакторов – составляет актуальность выбранной темы.

Анализ профильной литературы, приведенный в Главе 1, позволил обозначить терминологический аппарат исследования, определить оптимальную шкалу уровней сложности текста для данной задачи: такой шкалой большинство исследователей справедливо выбирают международную систему уровней CEFR. Анализ истории и современного состояния науки в области автоматического определения сложности текста показал, что на современном этапе эта задача 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 показал отсутствие инструментов детального анализа русскоязычных учебных текстов, при этом обзор аналогичных

продуктов для других языков продемонстрировал возможные направления для деятельности.

Основные работы по созданию математической модели оценки сложности текста для русскоязычных текстов описаны в Главе 2 настоящего исследования. Эти работы включали в себя сбор представительного эталонного корпуса текстов, содержащего образцы текстов разных уровней сложности из пособий по РКИ; выделение лингвистических признаков собранных текстов средствами автоматической обработки естественного языка и оценка их корреляции с уровнем сложности текста и, наконец, обучение и тестирование модели машинного обучения на материале подготовленного корпуса текста и их признаков. В качестве эталонной коллекции текстов для обучения математической модели, был собран корпус RuFoLa из 802 текстов из пособий и электронных ресурсов по РКИ общим объемом 13 720 слов. Анализ лингвистических признаков, извлеченных из корпуса текстов RuFoLa, позволил выявить характеристики текстов, наиболее релевантные задаче ранжирования текстов по уровням CEFR. Наивысший коэффициент корреляции показали лексические признаки, основанные на вхождении слов текста в лексические минимумы, и некоторые группы грамматических признаков. Анализ признаков также выявил особенности, которые мы учитывали при выборе регрессионной модели, такие как мультиколлинеарность и нелинейность признаков.

Применимость полученного алгоритма в практической деятельности преподавателей и изучающих РКИ было подтверждено экспериментально в ходе сравнения оценок текстов моделью с оценками экспертов, временем чтения и качеством понимания текстов студентами и, наконец, мнением самих студентов. Эксперимент показал, что математическая 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. Вторая серия оценки качества работы модели состояла в сравнении предсказания сложности текста моделью и экспертной оценкой

преподавателей, полученной в результате попарной оценки текстов и применения 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. На разработанную технологию оценки сложности текстов был получен сертификат РИД[24] (см. Приложение Г).

Пример практического применения разработанной математической модели оценки сложности текста в области преподавания РКИ описан в Главе 3 настоящего исследования. 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. Описаны варианты работы и интерпретации результатов анализа текста при создании собственных учебных текстов, конструировании и анализе контрольно-измерительных материалов, подборе фрагментов художественных текстов, варианты самостоятельной работы студентов с сервисов, а также возможности сервиса для оценки доступности текстов для специалистов по адаптации иностранных граждан. В ходе массового тестирования веб-сервиса были получены комментарии и отчеты о неточностях от пользователей, которые позволили усовершенствовать алгоритм выдачи статистической информации по текстам. Кроме того, получены положительные отзывы пользователей, сигнализирующие об актуальности и практической значимости сервиса. Таким образом, все поставленные задачи исследования были поэтапно решены и изложены в настоящем исследовании.

В качестве основных направлений дальнейшей работы отметим расширение эталонной коллекции текстов, которое позволит более детально оценивать уровень

[24] Свидетельство о государственной регистрации программы для ЭВМ № 2021661785 от 15.07.2021

текста в зависимости от его типа (художественный, информационный и т.д.) и вида чтения (изучающее, просмотровое, с охватом общего содержания). Еще одним вектором развития можно назвать 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. Данное исследование также может стать первым шагом к созданию модели автоматической системы по упрощению текстов для изучающих русский язык как иностранный.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Азимов Э. Г. Новый словарь методических терминов и понятий (теория и практика обучения языкам) / Э. Г. Азимов, А. Н. Щукин. – М. : ИКАР, 2009. – 448 с.

2. Акишина, А.А. Учимся учить. Для преподавателя русского языка как иностранного / А.А. Акишина, О.Е. Каган. – М.: Русский язык, 2012. – 256 с.

3. Андрюшина, Н.П. Лексические минимумы по русскому языку как иностранному: проблема отбора лексических и фразеологических единиц / Н.П. Андрюшина // Проблемы истории, филологии, культуры. – 2011. – № 3 (33). – С. 648-652.

4. Анисимович, К.В. Синтаксический и семантический парсер, основанный на лингвистических технологиях Abbyy Compreno / К.В. Анисимович, К.Ю. Дружкин, К.А. Зуев [и др.] // Международная конференция по компьютерной лингвистике «Диалог». – М.: РГГУ, 2012. – URL: http://www.vestnik.vsu.ru/pdf/analiz/2013/02/2013-02-31.pdf (дата обращения: 04.10.2021).

5. Антонова, В.Е. Типовые тесты. Базовый уровень. Общее владение / В.Е. Антонова, И.В. Курлова, М.М. Нахабина [и др.]. – СПб.: Златоуст, 2019. – 124 с.

6. Арутюнов, А.Р. Многофакторный количественный анализ учебников ин. Языков / А.Р. Арутюнов, Л.Б. Трушина, П.Г. Чеботарев // Содержание и структура учебника РКИ: сб. ст. / сост. Л.Б. Трушина. – М.: Русский язык, 1981. – С. 58-76.

7. Бабайлова А.Э. Текст как продукт, средство и объект коммуникации при обучении неродному языку: социопсихолингвистические аспекты / под общ. ред. А.А. Леонтьева. – Саратов: Изд-во Сарат. ун-та, 1987. – 151 с.

8. Баранова, Ю.Н. Создание вспомогательного информационного ресурса для анализа учебных текстов на русском языке / Ю.Н. Баранова, Т.С. Елипашева // Человек в информационном пространстве. – Ярославль: ЯГПУ, 2014. – URL: http://yspu.org/images/c/cf/BaranovaYuN_ElipashevaTS.pdf (дата обращения: 04.10.2021).

9. Бурвикова, Н.Д. К проблеме экспрессивности учебного текста // Лингвострановедение и текст: сб.ст. / сост. Е.М. Верещагин, В.Г. Костомаров. – М.: Русский язык, 1987. – С. 113-117.

10. Волкова, Т.Г. Пороговый уровень. Русский язык. Том I. Повседневное общение / Т.Г. Волкова, Е.Л. Корчагина, А.Л. Кузнецов. – Страсбург: Совет Европы Пресс Г, 1996. – 236 с.

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

12. Вятютнев, М.Н. Легкость/трудность в стратегии усвоения русского языка как иностранного / М.Н. Вятютнев // Русский язык за рубежом. – 1978. – № 3. – С. 46-49.

13. Гальперин, И.Р. Текст как объект лингвистического исследования. – М.: КомКнига, 2006. – 144 с.

14. Глазунова, О.И. Программа по русскому языку как иностранному. Уровни А1 – С2. Основной курс. Фонетика. Лексика, Грамматика. Аудирование. Чтение. Говорение. Письмо / О.И. Глазунова, Д.В. Колесова, Т.И. Попова. – М.: Русский язык. Курсы, 2017. – 216 с.

15. Горелов И.Н. Основы психолингвистики. Учебное пособие. Третье, переработанное и дополненное издание / И.Н. Горелов, К.Ф. Седов. – М.: "Лабиринт", 2001. – 304с.

16. Государственный образовательный стандарт по русскому языку как иностранному. Второй уровень. Общее владение / Иванова Т.А. [и др.] – М. – СПб: Златоуст, 1999.– 40 с.

17. Государственный образовательный стандарт по русскому языку как иностранному. Третий уровень. Общее владение / Иванова Т.А. и др. – М. – СПб: Златоуст, 1999. – 44 с.

18. Государственный стандарт по русскому языку как иностранному. Базовый уровень / Нахабина М.М. [и др.]– 2-е изд., испр. и доп. – М. – СПб.: Златоуст, 2001. – 32 с.

19. Государственный стандарт по русскому языку как иностранному. Элементарный уровень / Владимирова Т.Е. [и др.] – 2е изд., испр. и доп. – М. – СПб.: Златоуст, 2001. – 28 с.

20. Григоренко, В.А. Об учебно-коммуникативной значимости текста / В.А. Григоренко // Функционально-коммуникативный подход к описанию и преподаванию русского языка: межвуз. сб. науч. тр. – Воронеж, 1991. – С. 212-217.

21. Дергачева, Г.И. Методика преподавания русского языка как иностранного на начальном этапе / Г.И. Дергачева, О.С. Кузина, Н.М. Малашенко [и др.]. – М.: Русский язык, 1986. – 239 с.

22. Дружкин, К.Ю. Метрики удобочитаемости для русского языка. Выпускная квалификационная работа / К.Ю. Дружкин. – М.: Национальный исследовательский университет «Высшая школа экономики», 2016. – 65 с. – URL: https://www.hse.ru/edu/vkr/184791276 (дата обращения: 04.10.2021)

23. Зарубина, Н.Д. Текст: лингвистические и методические аспекты / Н.Д. Зарубина. – М.: Русский язык, 1981. – 112 с.

24. Зильберглейт, М.А. Повышение качества учебной литературы / М.А. Зильберглейт, Ю.Ф. Шпаковский, М.М. Невдах // Труды БГТУ. Серия 4: Принт- и медиатехнологии. – 2012. – № 9. – URL: https://cyberleninka.ru/article/n/povyshenie-kachestva-uchebnoy-literatury (дата обращения: 04.10.2021).

25. Каменская, О.Л. Текст и коммуникация / О.Л. Каменская. – М.: Высшая школа, 1990. – 152 с.

26. Карпов, Н.В. Идентификация уровня сложности текста и его адаптация / Н.В. Карпов. – URL: http://www.slideshare. net/karpnv/ss-31225145№14356960593761&fbinitialized (дата обращения: 13.08.2021).

27. Кисельников, А.С. К проблеме характеристик текста: читабельность, понятность, сложность, трудность / А.С. Кисельников // Филологические науки. Вопросы теории и практики. – 2015. – № 11 (53), Ч. II. – С. 79-84.

28. Клычникова, З.И. К вопросу о показателях понимания содержания иноязычного текста / З.И. Клычникова // Психология в обучении иностранному языку. – М.: Просвещение, 1967. – С. 89-108.

29. Кожетьева, Т.А. Параметры оптимизации текста // Когнитивные аспекты языковой категоризации: сб. науч. тр. / отв. ред. Л.А. Манерко. – Рязань: РГПУ им. С.А. Есенина, 2000. – С. 179-183.

30. Костомаров В.Г. Принципы отбора лексического минимума / В.Г. Костомаров // Русский язык в национальной школе. – 1963. – № 1. – С. 29–35.

31. Криони, Н.К. Автоматизированная система анализа параметров сложности учебного текста / Н.К. Криони, А.Д. Никин, А.В. Филиппова // Технология и организация обучения: науч. издание. – Уфа: УГАТУ, 2008. – С. 155-161

32. Крючкова Л.С. Практическая методика обучения русскому языку как иностранному. Учебное пособие для начинающего преподавателя, для студентов-филологов и лингвистов, специализирующихся по РКИ / Л.С Крючкова, Н.В. Мощинская. – М.: Флинта: Наука, 2009. – 480 с.

33. Кулибина, Н.В. Зачем, что и как читать на уроке. Художественный текст на уроке русского языка как иностранного / Н.В. Кулибина. – СПб.: Златоуст, 2001. – 269 с.

34. Кулибина, Н.В. О том, как текст превращается в урок русского языка / Н.В. Кулибина // Мир русского слова. – 2002. – № 1. – С. 85-93.

35. Лагутин М.Б. Наглядная математическая статистика / М.Б. Лугутин. – М: Лаборатория знаний, 2021. – 472 с.

36. Лапидус, Б.А. Некоторые теоретические вопросы методики обучения неродному языку / Б.А. Лапидус // Общая методика обучения иностранным языкам: хрестоматия / сост. А.А. Леонтьев. – М.: Русский язык, 1991. – С.61-69.

37. Лапошина А.Н. Что значит «не входит в лексический минимум»? Подсчет процента незнакомой лексики в тексте по РКИ с учетом доступных словообразовательных моделей /А.Н.Лапошина // Преподаватель XXI век. – 2021. – № 4. Часть 2. – С. 473–483.

38. Лексический минимум по русскому языку как иностранному. Базовый уровень. Общее владение / Н.П. Андрюшина, Т.В. Козлова. – 5е изд. – СПб. : Златоуст, 2015. – 116 с.

39. Лексический минимум по русскому языку как иностранному. Второй сертификационный уровень. Общее владение / под редакцией Н.П. Андрюшиной. – 7-е изд. – СПб. : Златоуст, 2017. – 164 с.

40. Лексический минимум по русскому языку как иностранному. Первый сертификационный уровень. Общее владение / Н.П. Андрюшина и др. – 9е изд. – СПб. : Златоуст, 2017. – 200 с.

41. Лексический минимум по русскому языку как иностранному. Третий сертификационный уровень. Общее владение / под ред. Н.П. Андрюшиной. – СПб. : Златоуст, 2018. – 201 с.

42. Лексический минимум по русскому языку как иностранному. Элементарный уровень. Общее владение / под ред. Н.П. Андрюшиной, Т.В. Козловой. – 4е изд., испр. и доп. – СПб. : Златоуст, 2012. – 80 с.

43. Ляховицкий, М.В. Методика преподавания иностранных языков / М.В. Ляховицкий. – М.: Высшая школа, 1981. – 159 с.

44. Ляшевская, О.Н. Частотный словарь современного русского языка (на материалах Национального корпуса русского языка) / О.Н. Ляшевская, С.А. Шаров. – М.: Азбуковник, 2009. – 1087 с.

45. Маркина, Е.И. Лингводидактические основы разработки лексических минимумов по русскому языку как иностранному (для разных уровней и профилей обучения): дис. … канд. пед. наук: 13.00.02 / Маркина Елена Игоревна. – М., 2011. – 235 с.

46. Маркина, Е.И. Основные подходы к минимизации лексики в российской и европейской учебной лексикографии / Е.И. Маркина, К.М. Руис-Соррилья // Полилингвиальность и транскультурные практики. – 2011. – № 3. – С. 77-84.

47. Мацковский, М.С. Проблемы читабельности печатного материала / М.С. Мацковский // Смысловое восприятие речевого сообщения в условиях массовой коммуникации / отв. ред. Т.М. Дридзе, А.А. Леонтьев. – М.: Наука, 1976. – С. 126-142.

48. Мизернов, И.Ю. Анализ методов оценки сложности текста / И.Ю. Мизернов, Л.А. Гращенко // Новые информационные технологии в автоматизированных системах. – 2015. – № 18. – URL: https://cyberleninka.ru/article/n/analiz-metodov-otsenki-slozhnosti-teksta (дата обращения: 29.09.2021).

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

50. Мирошникова, Е.А. Адаптация текстового учебного материала при дифференцированном обучении иностранному языку / Е.А. Мирошникова // Вестник БГУ. – 2016. – № 3 (29). – С. 229-234.

51. Мукомель, В.И. Адаптация и интеграция мигрантов: методологические подходы к оценке результативности и роль принимающего общества / В.И. Мукомель // Россия реформирующаяся: ежегодник: сборник научных статей / отв. ред. М.К. Горшков. – Москва: Новый хронограф, 2016. – Вып. 14. – С. 411-467.

52. Муравьев, Н.А. Подходы к составлению лексических минимумов в России и за рубежом: проблемы и перспективы / Н.А. Муравьев, М.Ю. Ольшевская // Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация. – 2019. – Т. 17, № 1. С. 78-89.

53. Нахабина, М. Типовые тесты по русскому языку как иностранному. Элементарный уровень. Общее владение / М. Нахабина, В. Антонова, А. Толстых. – М.: Златоуст, 2014. – 43 с.

54. Невдах, М.М. Исследование информационных характеристик учебного текста методами многомерного статистического анализа / М.М. Невдах // Прикладная информатика. – 2008. – № 4 (16). – С. 117-130.

55. Норейко, Л. Лексический минимум по русскому языку как иностранному. Первый сертификационный уровень. Общее владение / Л. Норейко. – М.: Златоуст, 2014. – 200 с.

56. Оборнева, И.В. Автоматизированная оценка сложности учебных текстов на основе статистических параметров: дис. ... канд. пед. наук: 13.00.02 / Оборнева Ирина Владимировна. – М., 2006. – 165 с.

57. Пиневич, Е.В. Методика обучения чтению иностранных учащихся инженерного профиля с использованием компьютерных технологий: этап предвузовской подготовки: дис. ... канд. пед. наук: 13.00.02 / Пиневич Елена Валентиновна. – М., 2006. – 229 с.

58. Пиотровский, Р.Г. Математическая лингвистика: учеб. пособие для пед. ин-тов / Р.Г. Пиотровский, К.Б. Бектаев, А.А. Пиотровская. – М.: Высшая школа, 1977. – 383 с.

59. Пушкина, Е.С. Теоретико-экспериментальное исследование структурно-семантических параметров текста : дис. ... канд. филол. наук: 10.02.19 / Пушкина Елена Сергеевна. – Кемерово, 2004. – 155 с.

60. Русский язык – мой друг. Базовый уровень: учебник русского языка для студентов-иностранцев / под ред. Т.В. Шустиковой, В.А. Кулаковой. – М.: РУДН, 2011. – 851 с.

61. Русский язык как иностранный (с электронным приложением): учеб.-метод. пособие / под ред. А.И. Басовой. – Минск: БГУ, 2014. – 119 с.

62. Селегей, В.П. Отчет о научно-исследовательской работе по договору № 081-R приложению А1 от 01.10.2013. Тема исследования «Центр по исследованию разума и машин» / В.П. Селегей. – М., 2015.

63. Сибирцева В.Г. Автоматизация процесса адаптации текстов для электронных учебников. Проблемы и перспективы (на примере русского языка) / В.Г. Сибирцева, Н.В. Карпов // Nová rusistika. – 2014. – № 1. – С. 19-35.

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

65. Содержание и структура учебника русского языка как иностранного: сб. ст. / сост. Л.Б. Трушина. – М.: Русский язык, 1981. – 288 с.

66. Солнышкина, М.И. Сложность текста: этапы изучения в отечественном прикладном языкознании / М.И. Солнышкина, А.С. Кисельников // Вестник Томского государственного университета. Филология. – 2015. – № 6 (38). – URL: https://cyberleninka.ru/article/n/slozhnost-teksta-etapy-izucheniya-v-otechestvennom-prikladnom-yazykoznanii (дата обращения: 29.09.2021).

67. Текст: проблемы и перспективы: аспекты изучения в целях преподавания русского языка как иностранного: материалы III Междунар. науч.-практ. конф. – М.: Изд-во Моск. ун-та, 2004. – 360 с.

68. Тёрёчик, Л.Б. Текст как средство обучения русскому языку как иностранному: фреймовый подход: начальный этап обучения: дис. ... канд. пед. наук: 13.00.02 / Тёрёчик Людмила Беловна. – М., 2012. – 294 с.

69. Томина, Ю.А. Объективная оценка языковой трудности текстов (описание, повествование, рассуждение, доказательство): дисс. ... канд. пед. наук: 13.00.02 / Томина Юлия Алексеевна. – Москва, 1985. – 226 с.

70. Требования к Первому сертификационному уровню владения русским языком как иностранным. Общее владение. Профессиональный модуль / Н. П. Андрюшина [и др.]. 3-е изд. – СПб. : Златоуст, 2015. – 64 с.

71. Тулдава, Ю.А. О некоторых квантитативно-системных характеристиках полисемии / Ю.А. Тулдава // Ученые записки Тартуского университета. – 1979. – Вып. 502. – С. 107-124.

72. Тулдава, Ю.А. Об измерении трудности текста / Ю.А. Тулдава // Научные труды ЛГУ. – Тарту, 1975. – Вып. 345. – № 4. – С. 102-115.

73. Филиппова, А.В. Управление качеством учебных материалов на основе анализа трудности понимания учебных текстов: автореф. дис. ... канд. техн. наук: 05.13.10 / Филиппова Анастасия Владимировна. – Уфа, 2010. – 16 с.

74. Фоломкина, С.К. Обучение чтению на иностранном языке в неязыковом вузе: учеб.-метод. пособие для вузов / С.К. Фоломкина. – М.: Высшая школа, 1987. – 207 с.

75. Фрумкина, Р.М. Степень понимания текста как критерий оценки объёма словаря-минимума / Р.М. Фрумкина // Психология и методика обучения второму языку. – М.: Институт языкознания АНСССР, 1965. – 71 с.

76. Фрумкина, Р. М. Самосознание лингвистики – вчера и завтра / Р.М. Фрумкина // Известия Академии наук. Сер. лит. и яз. – 1999. – Т. 58. – № 4. – С. 28–38.

77. Цетлин, В.С. Дидактические требования к критериям сложности учебного материала / В.С. Цетлин // Новые исследования в педагогических науках. – 1980. – № 1 (35). – С. 30-31.

78. Чеснокова М.П. Методика преподавания русского языка как иностранного: учеб. пособие. / М.П. Чеснокова. – М.: МАДИ, 2015. – 132 с.

79. Шпаковский, Ю.Ф. Оценка трудности восприятия и оптимизация сложности учебного текста (на материале текстов по химии): автореф. дисс. … канд. филол. наук: 10.02.19 / Шпаковский Юрий Францевич. – Минск, 2007. – 21 с.

80. Щукин, А.Н. Методика преподавания русского языка как иностранного: учеб. пособие для вузов / А.Н. Щукин. – М.: Высшая школа, 2003. – 334 с.

81. Эсмантова, Т.Л. Русский язык: 5 элементов: уровень А1 (элементарный). – СПб.: Златоуст, 2016. – 320 с.

82. Якимов, А.Н. Адаптация и интеграция мигрантов: сборник эффективных практик / А.Н. Якимов. – СПб.: БФ «ПСП-фонд», 2018. – 69 с.

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

84. Bowers J.S. In defense of abstractionist theories of repetition priming and word identification / J.S. Bowers // Psychonomic bulletin and review. – 2000. – №7(1). – P. 83–99.

85. Cárcamo Morales, B. Readability and types of questions in Chilean EFL high school textbooks / B. Cárcamo Morales // TESOL Journal. – 2019. – P. e498. https://doi.org/10.1002/tesj.498

86. Chall, J.S. Manual for the new Dale-ChallReadability Formula / J.S. Chall, E. Dale. – Cambridge, MA: Brookline Books, 1995. – 149 p. книга

87. Chen X. Characterizing Text Difficulty with Word Frequencies / X. Chen, D. Meurers // Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications, San Diego, 2016. – Vol. 11. – P. 84–94.

88. Coleman, M. A computer readability formula designed for machine scoring / M. Coleman, T.L. Liau // Journal of Applied Psychology. – 1975. – Т. 60, № 2. – P. 283-284.

89. Collins-Thompson, K. A language modeling approach to predicting reading difficulty / K. Collins-Thompson, J. Callan // Proceedings of HLT//NAACL. – 2004. – Vol. 4. – P. 193-200.

90. Collins-Thompson, K. Computational assessment of text readability: a survey of current and future research. Recent Advances in Automatic Readability Assessment and Text Simplification / eds. F.T. Bernhard, F.D. Bernhard // Special issue of International Journal of Applied Linguistics. – 2014. – P. 97-135.

91. Collins-Thompson, K. Predicting reading difficulty with statistical language models / K. Collins-Thompson, J. Callan // Journal of the American Society for Information Science and Technology. 2005. – Vol. 56 (13). – P. 1448-1462.

92. Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume with New Descriptors / B. North, E. Piccardo, T. Goodier. – Strasbourg: Council of Europe Publishing, 2018. – 227 p.

93. Costerman, J. Processing interclausal relationships: Studies in production and comprehension of text / J. Costerman, M. Fayol. –Hillsdale, NJ: Lawrence Erlbaum. – 1997. 302 p.

94. Crossley, S. Assessing text readability using cognitively based indices / S. Crossley, S., J. Greenfield, D. McNamara // TESOL Quarterly. – 2008. – №42, P. 475–493.

95. Crossley, S. Text readability and intuitive simplification: A comparison of readability formulas / S. Crossley, D. Allen, D. McNamara // Reading in a Foreign Language. – 2011. – №23(1), P. 84–101.

96. Curto P., Mamede N., Baptista J. Assisting European Portuguese teaching: linguistic features extraction and automatic readability classifier / P. Curto, N. Mamede, J. Baptista // Computer Supported Education. – 2016. – Vol. 583. – P. 81-96.

97. DuBay, W. The principles of readability / W.H. DuBay // Impact information. – 2004. – P. 1-76.

98. DuBay, W.H. Smart Language: readers, readability, and the grading of text / W.H. DuBay. – California: ERIC, 2007. – 160 p.

99. Ferris, D. Lexical and syntactic features of ESL writing by students at different levels of L2 proficiency / D. Ferris // TESOL Quarterly. – 1994. – №28, – P. 414–420.

100. Flesch, R. A new readability yardstick / R. Flesch // Journal of applied psychology. – 1948. – Vol. 32 (3). – P. 221.

101. Francois, T. An ’AI readability’ formula for French as a foreign language / T. Francois, C. Fairon // Proceedings of the 2012 Joint Conference on Empirical methods in natural language processing and computational natural language learning, Korea, 2012. – P. 466-477

102. Graesser A.C. Coh-Metrix: Providing Multilevel Analyses of Text Characteristics / A.C. Graesser, D.S. McNamara, J.M. Kulikowich // Educational Researcher. – 2011. – №40(5). –P. 223-234.

103. Graesser, A. Coh-Metrix Measures Text Characteristics at Multiple Levels of Language and Discourse / A.C. Graesser, D.S. McNamara, Z. Cai, M. Conley, H. Li, J. Pennebaker // The Elementary School Journal. – 2014. – №115. – P. 210-229.

104. Gunning, R. The technique of clear writing / R. Gunning. – NewYork: McGraw-Hill, 1968. – 289 p.

105. Hancke, J. Readability classification for German using lexical, syntactic, and morphological features / J. Hancke, S. Vajjala, D. Meurers // Proceedings of COLING. – 2012. – P. 1063-1080.

106. Heilman, M. An analysis of statistical models and features for reading difficulty prediction / M. Heilman, K. Collins-Thompson, M. Eskenazi // Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications. – 2008. – P. 71-79.

107. Heilman, M. Combining lexical and grammatical features to improve readability measures for first and second language texts / M. Heilman, K. Collins- Thompson, J. Callan [et al.] // Proceedings of HLT-NAACL'07. – 2007. – P. 460-467.

108. Hou, J. Modeling language learning using specialized Elo ratings / J. Hou, M.W. Koppatz, J.M.H. Quecedo [et al.] // Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications. – 2019. – P. 494-506.

109. Ivanov, V.V. Efficiency of Text Readability Features in Russian Academic Texts / V.V. Ivanov, M.I. Solnyshkina, V.D. Solovyev // Proceedings of the International Conference «Dialogue 2018» Moscow, 2018. – P.284-293.

110. Joachims, T. Text categorization with support vector machines: learning with many relevant features / T. Joachims // Technical Report 23, Universitat Dortmund, LS VIII, 1997. – P. 138-142.

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

112. Kaur J. Effective approaches for extraction of keywords / J. Kaur, V. Gupta // International Journal of Computer Science Issues (IJCSI). – 2010. – №6. – P. 144.

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

114. Kilgarriff, A. Comparable corpora within and across languages, word frequency lists and the Kelly project / A. Kilgarriff // Proceedings of the 3rd Workshop on Building and Using Comparable Corpora at LREC 2010. – Malta, 2010. – P. 1-5.

115. Kilgarriff, A. Corpus-based vocabulary lists for language learners for nine languages / A. Kilgarriff, F. Charalabopoulou, M. Gavrilidou [et al.] // Lang Resources & Evaluation. – 2014. – Vol. 48. – P. 121-163.

116. Kincaid, J.P. Derivation of new readability formulas (automated readability index, fog count and Flesch reading ease formula) for navy enlisted personnel / J.P. Kincaid, R.P. Fishburne jr., R.L. Rogers. – Florida: Institute for Simulation and Training, 1975. – URL: https://stars.library.ucf.edu/cgi/viewcontent.cgi?article=1055&context=istlibrary (date of application: 15.10.2021).

117. Kiselnikov, A. Coh-metrix readability formulas for an academic text analysis / A. Kiselnikov, D. Vakhitova, T. Kazymova // IOP Conference Series: Materials Science and Engineering. – 2020. –Vol. 890. No. 1. – P. 1-6. doi:10.1088/1757-899X/890/1/012207

118. Kismarianto, T. The readability of reading materials in English textbook for the eleventh grade students of SMK negeri 1 Beringin. Register / T. Kismarianto, M. Siregar, Y. Erlita // Journal of English Language Teaching of FBS-Unimed. – 2019. – Vol. 7. – P. 94-102.

119. Krashen, S.D. Free Voluntary Reading: Still a Very Good Idea / S.D. Krashen. – Santa Barbara, 2011. – 90 p.

120. Lehmann, E.L. Theory of Point Estimation / E.L. Lehmann, G. Casella. – New York: Springer, 1998. – 617 p.

121. Mangaroska, K. Elo-rating method: towards adaptive assessment / K. Mangaroska, B. Vesin, M. Giannakos // E-Learning. – Vol. 1. – 2019. – P. 380-382.

122. Martinc, M. Supervised and unsupervised neural approaches to text readability / M. Martinc, S. Pollak, M. Robnik-Šikonja // Computational Linguistics 2021. – Vol. 47 (1). – P. 141-179.

123. Mc Laughlin, G.H. SMOG grading a new readability formula / G.H. Mc Laughlin // Journal of reading. – 1969. -Vol. 12. – № 8. – P. 639-646.

124. Monsell, S. Effects of frequency on visual word recognition tasks: Where are they? / S. Monsell, M.C. Doyle, P.N. Haggard // Journal of Experimental Psychology: General. –1989. – №118. – P. 43–71.

125. Nation, P. How large a vocabulary is needed for reading and listening? / P. Nation // Canadian modern language review-revue Canadienne des Langues Vivantes – CAN MOD LANG REV. – 2006. – Vol. 63. – P. 59-81.

126. Pilan, I. A readable read: automatic assessment of language ´ learning materials based on linguistic complexity / I. Pilan, S. Vajjala, E. Volodina // Proceedings of CICLing 2015, to appear in International Journal of Computational Linguistics and Applications. – 2015. – Vol. 6, № 2.

127. Qian D.D. Investigating the Relationship Between Vocabulary Knowledge and Academic Reading Performance: An Assessment Perspective / D.D.Qian // Language Learning. – 2002. – № 52(3). P. 513–536.

128. Rello, L. Frequent Words Improve Readability and Short Words Improve Understandability for People with Dyslexia / L. Rello, R. Baeza-Yates, L. Dempere-Marco [et al.] // Human-Computer Interaction – INTERACT 2013. INTERACT 2013. Lecture Notes in Computer Science / eds. P. Kotzé, G. Marsden, G. Lindgaard [et al.]. – Springer, Berlin, Heidelberg, 2013. – Vol. 8120. – P. 203-219.

129. Reynolds R. Insights from Russian second language readability classification: complexity-dependent training requirements, and feature evaluation of multiple categories, San Diego, 2016. – P. 289–300.

130. Schwarm, S.E. Reading level assessment using support vector machines and statistical language models / S.E. Schwarm, M. Ostendorf // Proceedings of the 43rd annual meeting on association for computational linguistics (ACL ’05). – USA, 2005. – P. 523-530.

131. Segalovich, I. A Fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine / I. Segalovich // MLMTA. – Citeseer. – 2003. – P. 273-280.

132. Shannon, C.E. A Mathematical theory of communication / C.E. Shannon // Bell System Technical Journal. – 1948. – Vol. 27. – P. 379-423, 623-656.

133. Sharoff S. A Frequency Dictionary of Russian: Core vocabulary for learners / S. Sharoff, E. Umanskaya, J. Wilson. – New York: Routedge, 2013. – 400 p.

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

135. Smith, E.A. Automated readability index / E.A Smith, R.J. Senter // AMRL TR. – Ohio, 1967.

136. Sowmya V. On Improving the Accuracy of Readability Classification using Insights from Second Language Acquisition / V. Sowmya, D. Meurers // Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications (BEA7). – 2012. – P. 163–173.

137. Spache, G. A new readability formula for primary-grade reading materials / G. Spache // The Elementary School Journal. – 1953. – Vol. 55. – P. 410-413.

138. Stenner, A.J. The Lexile Framework for Reading Technical Report / A.J. Stenner, H. Burdick, E. E. Sanford [et al.]. – Durham, NC: MetaMetrics, Inc, 2007.

139. Sung, Y.T. Leveling L2 Texts Through Readability: Combining Multilevel Linguistic Features with the CEFR / Y.T. Sung, W.C. Lin, S. Dyson [et al.] // Modern Language Journal. – 2015. – Vol. 99. – P. 371-391.

140. To V. Lexical density and Readability: A case study of English Textbooks / V.To, T. Le // Proceedings of the Australian Systemic Functional Linguistics Association Conference, Melbourne, 2013. – P.61–71.

141. Vinh, T. Lexical density and readability: a case study of English textbooks / T. Vinh, F. Si, T. Damon // The international journal of language, society and culture. – 2013. – Is. 37. – P. 61-71.

142. West, R. Keeping it in the family / R. West // English teaching professional. – 2015. – Vol. 97. – P. 60-63.

143. Wright, B.D. Readability and reading ability / B.D. Wright, A.J. Stenner // Paper presented to the Australian Council on Education Research. – Washington, D.C.: Distributed by ERIC Clearinghouse, 1998. – 28 p.

144. Xia, M. Text readability assessment for second language learners. in proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications / M. Xia, E. Kochmar, T. Briscoe. – San Diego, CA: Association for Computational Linguistics, 2016. – P. 12-22.

145. Zalmout, N. Analysis of foreign language teaching methods: an automatic readability approach / N. Zalmout, H. Saddiki, N. Habash // Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA 2016). – Osaka, 2016. – P. 122-130.

СПИСОК УЧЕБНИКОВ И УЧЕБНЫХ ПОСОБИЙ

1. Андрюшина Н.П. Тренировочные тесты по русскому языку как иностранному: I сертификационный уровень: общее владение./ Н.П. Андрюшина, Н.И. Пращук, М.Н. Макова

2. Андрюшина Н.П. Тренировочные тесты по русскому языку как иностранному. II Сертификационный уровень/ Андрюшина, Н.П., Макова, М.Н. СПб.: “Златоуст”, 2019, 140 с.

3. Андрюшина Н.П. Тренировочные тесты по русскому языку как иностранному. III Сертификационный уровень/ Андрюшина Н.П., Жорова А.П., Макова М.Н., Норейко Л.Н. СПб.: “Златоуст”, 2019, 152 с.

4. Андрюшина Н.П. Тренировочные тесты по русскому языку как иностранному. Базовый уровень / В.Е. Антонова, И.В. Курлова, М.М. Нахабина, А.А. Толстых СПб.: “Златоуст”, 2019, 124 с.

5. Антонова, В.Е. Дорога в Россию : учебник русского языка (первый уровень): в 2 т. т. I / В.Е. Антонова, М.М.Нахабина, А.А. Толстых. — М.: ЦМО МГУ им. М.В. Ломоносова ; СПб. : Златоуст, 2013. — 256 с.

6. Антонова, В.Е. Дорога в Россию : учебник русского языка (первый уровень): в 2 т. т. II / В.Е. Антонова, М.М.Нахабина, А.А. Толстых. — СПб. : Златоуст, 2012. — 184 с.

7. Антонова, В.Е. Дорога в Россию: учебник русского языка (базовый уровень) / В.Е. Антонова, М.М.Нахабина, А.А. Толстых.— 4-е изд. — М.: ЦМО МГУ им. М.В. Ломоносова ; СПб. : Златоуст, 2009. — 256 с.

8. Антонова, В.Е. Дорога в Россию: учебник русского языка (элементарный уровень) / В.Е. Антонова, М.М.Нахабина, М.В.Сафронова, А.А. Толстых. — СПб. : Златоуст, 2013. — 344 с.

9. Барсукова-Сергеева О.М. Читая сказки...: учеб, пособие / О.М. Барсукова-Сергеева. — М. : Флинта : Наука, 2021. — 200 с.

10. Беликова, Л.Г. Русский язык: первые шаги : учебное пособие : В 3 ч. Ч. 3 / / Л.Г. Беликова, Т.А. Шутова, И.Н. Ерофеева — СПб. : Златоуст, 2017. — 296 с.

11. Беликова, Л.Г. Русский язык: первые шаги : учебное пособие : В 3 ч. Ч. 2. / Л.Г. Беликова, Т.А. Шутова, И.Н. Ерофеева — СПб. : Златоуст, 2017. — 296 с.

12. Беликова, Л.Г. Русский язык: первые шаги: учебное пособие : В 3 ч. Ч. 1. / Л.Г. Беликова, Т.А. Шутова, И.Н. Ерофеева — СПб. : Златоуст, 2018. — 264 с.

13. Головко, О.В. Вперёд! Пособие по русской разговорной речи / О.В. Головко. — М.: Русский язык. Курсы, 2009. — 184 с.

14. Долматова О. Точка Ру. Tochka Ru. «Точка ру В1 часть 1». Textbook+Workbook. В 2 книгах. / О. Долматова, Е. Новачац — М. : Перо, 2021 — 128 с.

15. Долматова О. Точка Ру. Tochka Ru. Russian course A1. Textbook+Workbook. В 2 книгах. / О. Долматова, Е. Новачац — М. : Перо, 2017 — 294 с.

16. Костина, И. Русский класс: учебное пособие по русскому языку как иностранному/ И. Костина, Т. Александрова-Сканлан, Е. Богословская, Н. Александрова — Ростов н/Д: Феникс, 2006. — 336 с.

17. Кумбашева Ю.А. Человек в современном мире : учеб. пособие по разговорной практике / Ю.А. Кумбашева. — М. : Флинта : Наука, 2006. — 200 с.

18. Курлова И.В. Начинаем читать по-русски. Пособие по чтению для студентов, начинающих изучать русский язык. / И.В. Курлова — М.: Русский язык. Курсы, 2018 —112 с.

19. Куцерева-Жаме А.М. Спасибо! Начальный курс русского языка / А.М. Куцерева-Жаме А.М., М. Китадзё — СПб .: Златоуст, 2014. — 192 с.

20. Миллер Л.В. Жили-были...28 уроков русского языка для начинающих: учебник / Л.В. Миллер, Л.В. Политова, И.Я. Рыбакова — 14-е изд., испр. — СПб: Златоуст, 2016 — 112 С.

21. Новикова Н.С. Удивительные истории. 116 текстов для чтения, изучения и развлечения : учеб, пособие / Н.С. Новикова , О.М. Щербакова — 15-е изд. — М. : Флинта : Наука, 2019. — 368 с.

22. Плюснина Т. Д . Читаю, говорю и пишу по-русски: Учебное пособие для иностранных учащихся (элементарный, базовый уровни владения русским языком). / Т.Д. Плюснина, Г.А. Виноградова — СПб.:, 2009. — 77 с.

23. Родимкина А. М. Россия: экономика и общество. Тексты и упражнения. / А.М. Родимкина А., Н. Ландсман — СПб.: Златоуст, 2007. — 160 с.

24. Родимкина, А.М. Россия день за днем : Выпуск 1. Тексты и упражнения. / А.М. Родимкина А., Н. Ландсман — СПб .: Златоуст, 2009. — 136 с.

25. Скороходов, Л.Ю. Окно в Россию: учебное пособие по русскому языку как иностранному для продвинутого этапа. В двух частях. Часть первая. / Л.Ю. Скороходов, О.В. Хорохордина — 4-е изд. — СПб. : Златоуст, 2015. — 192 с.

26. Тектстотека проекта "Learn Russian with interest" [электронный ресурс]. — Режим доступа: http://lrwi.ru

27. Чернышов, С.И. Поехали! Русский язы к для взрослых. Начальный курс : учебник. Часть 1.1. / С.И. Чернышов, А.В. Чернышова. — 2-е изд. — СПб. : Златоуст, 2019. — 176 с.

28. Чернышов, С.И. Поехали! Русский язы к для взрослых. Начальный курс : учебник. Часть 1.2. / С.И. Чернышов, А.В. Чернышова. — 2-е изд. — СПб. : Златоуст, 2019. — 176 с.

29. Чернышов, С.И. Поехали!-2. Русский язык для взрослых. Базовый курс: в 2 т. Т. I. / С.И. Чернышов, А.В. Чернышова. — 6-е изд., — СПб. : Златоуст, 2014. — 168 с.

30. Чернышов, С.И. Поехали!-2. Русский язык для взрослых. Базовый курс: в 2 т. Т. II. / С.И. Чернышов, А.В. Чернышова. — 4-е изд. — СПб. : Златоуст, 2014. — 200 с.

31. Шустикова Т.В. Русский язык для вас. Первый сертификационный уровень: Учебник русского языка для иностранных учащихся / Т.В. Шустикова, В.А. Кулакова. — 5-е изд., доп. — М.: РУДН., 2016. 322 с.

32. Электронный учебник "Между нами" [электронный ресурс] / Lynne deBenedette, William J. Comer, Alla Smyslova,Jonathan Perkins. 2019. — Режим доступа: https://mezhdunami.org

33. Эсмантова, Т.Л. Русский язык: 5 элементов : уровень А1 (элементарный). / Т.Л. Эсмантова — СПб. : Златоуст, 2021. — 320 с.

34. Эсмантова, Т.Л. Русский язык: 5 элементов : уровень А2 (базовый). / Т.Л. Эсмантова — СПб. : Златоуст, 2013. — 328 с.

35. Эсмантова, Т.Л. Русский язык: 5 элементов : уровень В1 (базовый — первый сертификационный) / Т.Л. Эсмантова — СПб. : Златоуст, 2019. — 340 с.

ПРИЛОЖЕНИЕ А.

МАТЕРИАЛЫ ЭКСПЕРИМЕНТА ПО ПРОВЕРКЕ КАЧЕСТВА РАБОТЫ МОДЕЛИ: ТЕКСТЫ И АНКЕТЫ

Лист №1:

Время начала чтения _____ Время окончания чтения _____

Текст 1. Мария, профессиональный отдыхающий в аквапарке.

Самая важная моя задача – снимать видео о «Ква-ква парке», они попадают на сайт аквапарка в «Блог профессионального отдыхающего». Такое видео я должна была снимать каждую неделю на абсолютно разные темы. Например, для первого видео мне нужно было протестировать все горки аквапарка. Я неуверенно себя чувствую в воде, поэтому просила это сделать за меня друзей, но потом прокатилась и сама. На одной неделе я тестировала меню ресторана аквапарка «Троя»: мне бесплатно приносили блюда, я их пробовала и на камеру рассказывала, понравились ли они мне. Кстати, со стороны аквапарка мне никто не запрещал говорить, если мне что-то не понравилось.

В аквапарк я приходила в любое время, вход для меня был бесплатный. Самое сложное было уговорить людей участвовать в видео и говорить на камеру. Конечно, это логично: люди пришли отдохнуть, заплатили деньги, может, у них тариф «2 часа 40 минут» и они не хотят это время тратить на тебя. К тому же не у всех людей хорошие фигуры, а многие просто не любят, когда их снимают, тем более в купальнике.

Друзьям нравилась моя работа, потому что я их проводила в аквапарк бесплатно, а родители сначала относились к моей работе с сомнением – почему выбрали меня? Почему я буду ходить в купальнике? А потом, когда стали видео мои смотреть, поняли, что всё хорошо.

Вопрос 1. Что было самое сложное в этой работе?

1) Тестировать горки
2) Снимать видео о меню ресторана «Троя»
3) Снимать видео с посетителями аквапарка
4) Объяснить родителям, что это нормальная работа

Вопрос 2. Как Мария тестировала меню ресторана «Троя»?

1) Пробовала еду и говорила повару, понравилась ли она ей
2) Пробовала еду и снимала видео об этом
3) Просила других людей рассказать на камеру, какие блюда им нравятся
4) Приглашала друзей и предлагала попробовать еду из меню

Вопрос 3. Какой у Марии был график работы?

1) Она работала по стандартному графику с 9 до 18 часов
2) Она работала по выходным, когда больше всего людей
3) Она работала ночью, когда в аквапарке мало людей
4) Она могла работать когда хочет

Вопрос 4. Какая проблема была у Марии, когда она делала первое видео?

Спасибо! А теперь выберите наиболее подходящий пункт:

☐Для меня это сложный текст: много незнакомых слов, мне будет трудно пересказать его.

☐Для меня это оптимальный текст: я знаю не все слова, но понимаю смысл текста, могу говорить на эту тему.

☐Для меня это легкий текст: я знаю почти все слова, могу пересказать его.

Лист №2:

Время начала чтения _____ Время окончания чтения _____

Текст 2. Олег, фотограф в National Geographic

Надо понимать, что фотографы National Geographic не живут в дорогих отелях и не ходят смотреть достопримечательности. Это настоящая работа. Многие люди сидят в офисах и думают, что у фотографа простая и приятная работа: путешествовать, плавать с дельфинами. Но когда тебе нужно каждый день вставать в восемь утра, собирать фототехнику, идти к дельфинам, полтора часа с ними плавать, потом завтракать, ещё полтора часа плавать с дельфинами, потом обедать, потом ещё полтора часа плавать с дельфинами – и так семь месяцев, – ты начинаешь ненавидеть этих дельфинов.

Еще это и опасная работа. Например, в 2011 году мы делали проект в Кабардино-Балкарии, и во время экспедиции у нас умер человек, еще один стал инвалидом.

Что нужно делать, чтобы получить работу мечты? Для этого нужно стать лучшим. Например, чтобы фотографировать для National Geographic, для этого тоже нужно стать лучшим, и это очень просто. Ты выбираешь тему или стиль фотографии, работаешь пять-десять лет, становишься лучшим в этом стиле, и рано или поздно начинаешь работать с National Geographic.

Вопрос 1. Что Олег советует делать, чтобы работать с National Geographic?

1) Выбрать стиль фотографии и работать 15 лет
2) Выбрать стиль фотографии и стать в нем лучшим
3) Делать много фотографий в разных стилях и на разные темы
4) Работать с дельфинами

Вопрос 2. Что Олег говорит о работе с дельфинами?

1) Это сложная работа, надо рано вставать
2) Это тяжелая работа, надо жить в плохих отелях
3) Это прекрасная работа, можно увидеть много интересного
4) Это тяжелая работа, надо много времени делать одно и то же

Вопрос 3. Что из этого Олег НЕ говорил о своей профессии?

1) Это опасная работа
2) Это офисная работа
3) Это настоящая работа
4) Это работа мечты

Вопрос 4. Сколько времени Олег делал проект с дельфинами?

Спасибо! А теперь выберите наиболее подходящий пункт:

☐Для меня это сложный текст: много незнакомых слов, мне будет трудно пересказать его.

☐Для меня это оптимальный текст: я знаю не все слова, но понимаю смысл текста, могу говорить на эту тему.

☐Для меня это легкий текст: я знаю почти все слова, могу пересказать его.

Лист №3:

Время начала чтения _____ Время окончания чтения _____

Текст 3. Наталья, технолог на шоколадной фабрике.

Моя основная задача – это контроль технологических процессов: есть технологи, за работой которых я слежу, обучаю их, мы вместе создаем новые вкусы. Конечно, шоколад я пробую в течение дня постоянно, потому что нужно знать то, что идёт в производство. Иногда ем просто так: когда просто хочется сладкого, то выбираю молочный или белый шоколад, а если нужно проснуться, то выбираю горький шоколад.

Больше всего шоколада я съедаю в период, когда мы тестируем какие-либо новые вкусы. За это время все пробуют минимум 20 видов разной продукции. Какое количество шоколада в килограммах, я не смогу сказать точно, тем более мы пробуем не только шоколадные изделия, но и вафельные, мармеладные. После каждого кусочка шоколада рот ополаскивается тёплой водой – чтобы нейтрализовать вкус, оставшийся во рту.

Обычно, глядя на меня, никто никогда не верит, что я работаю на шоколадной фабрике. Моя фигура не изменилась, и кое-кто из моих друзей даже говорит: «Ты, наверное, ничего не ешь». Я считаю, что если сладости включать в свое меню как десерт и быть активным человеком, то ничего такого не произойдёт. И потом, для того чтобы что-то попробовать, не обязательно это съедать, вкусовые рецепторы находятся не в желудке, а на языке. Поэтому достаточно всё просто разжевать: это никак не влияет на фигуру.

Вопрос 1. Какая основная работа Натальи?

1) Разрабатывать и тестировать новые вкусы конфет
2) Контролировать, чтобы работники фабрики не ели шоколад
3) Готовить конфеты, мармеладную и вафельную продукцию
4) Разрабатывать новый дизайн для конфет

Вопрос 2. Какой шоколад она ест, когда хочется взбодриться?

1) Молочный или белый
2) Шоколад, вафельную и мармеладную продукцию
3) Горький шоколад
4) Она не ест шоколад вне работы, чтобы не портить фигуру

Вопрос 3. Что из этого НЕ говорила Наталья?

1) После дегустации надо прополоскать рот
2) В период выбора новых вкусов она тестирует около 20 видов кондитерских изделий
3) Наталья ест шоколад на производстве каждый день
4) Если включить сладости в свое меню, вы будете более активным человеком

Вопрос 4. Какой рецепт даёт Наталья, чтобы сохранить хорошую фигуру?

Спасибо! А теперь выберите наиболее подходящий пункт:

☐Для меня это сложный текст: много незнакомых слов, мне будет трудно пересказать его.

☐Для меня это оптимальный текст: я знаю не все слова, но понимаю смысл текста, могу говорить на эту тему.

☐Для меня это легкий текст: я знаю почти все слова, могу пересказать его.

АНКЕТА ДЛЯ СТУДЕНТА

Здравствуйте!

Расскажите, пожалуйста, немного о себе:

1. Ваш родной язык — _____
2. Русский для вас:
   - ☐ Первый иностранный язык
   - ☐ Второй иностранный язык
   - ☐ Третий и более иностранный язык
   - ☐ Семейный язык, вы билингв
3. Сколько времени вы изучаете русский язык?
   - ☐ Несколько месяцев
   - ☐ Около 1 года
   - ☐ 1-2 года
   - ☐ Более 2 лет
4. Сколько времени вы живете в России?
   - ☐ Несколько месяцев
   - ☐ Около 1 года
   - ☐ 1-2 года
   - ☐ Более 2 лет
5. Вы читаете по-русски (можно выбрать несколько вариантов):
   - ☐ Тексты из учебника
   - ☐ Специальные книги для чтения
   - ☐ Русская литература в оригинале
   - ☐ Новости на русском языке
   - ☐ Русские журналы, блоги, социальные сети
   - ☐ Профессиональные тексты (статьи, книги по специальности)
   - ☐ Другое: _____

АНКЕТА ДЛЯ УЧАСТНИКА-ПРЕПОДАВАТЕЛЯ

1. Перед текстом:

Прочитайте текст. Отметьте, пожалуйста, слова и конструкции, которые, скорее всего, студенты данной группы не знают, их нужно будет объяснить.

2. После текста:

Спасибо! Оцените, пожалуйста, уровень сложности этого текста:

☐ A1 ☐ A2

☐ B1 ☐ B2

☐ C1 ☐ C2

В качестве ознакомительного чтения (понять общий смысл, пересказать, ответить на вопросы, поддержать дискуссию) для этой группы он:

☐ Слишком сложный

☐ Оптимальной сложности

☐ Слишком простой

Комментарий (доп. информация о группе, текстах, замечения):

_____

ПРИЛОЖЕНИЕ Б.

ОБОБЩЕННЫЕ ОТВЕТЫ УЧАСТНИКОВ ЭКСПЕРИМЕНТА

	Текст №1. Фотограф дикой природы	**Текст №2. Блогер в аквапарке**	**Текст №3. Технолог на шоколадной фабрике**
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
Медианное время чтения (мин.)	4	5	6
Процент анкет с правильными ответами на послетекстовые вопросы	все правильные - 42% 3,5 правильных- 44% 3 правильных - 81% 2 правильных- 96%	4 - 20% 3.5 - 28% 3 - 60% 2 - 88%	4 - 15% 3.5 - 18% 3 - 45% 2 - 76%
Количество слов не вошедших в минимум B1	11 (6% от текста)	21 (10% от текста)	31 (15% от текста)
Среднее	0.9/1.1	3.11/4.5	5.9/7.2

количество незнакомых слов по мнению студентов/преподавателей			
Всего отмечено незнакомых слов у студентов/преподавателей	18/6	32/16	47/26
Слова, не вошедшие в лексические минимумы B1	дельфин Достопримечатель-ность инвалид многий ненавидеть отель офис проект стиль фотограф фототехника	аквапарк блог камера ква-ква кстати купальник логичный меню неуверенно попадать прокатиться просто протестировать профессиональный сайт сомнение тариф тестировать троя уговаривать фигура	вафельный взбодриться вкус вкусовой влиять глядеть дегустация десерт есть изделие какой-либо кое-кто контроль кусочек мармеладный меню минимум молочный нейтрализовать никак ополаскиваться основный

			просто разжевывать рецептор сладость следить тестировать технолог технологический течение фигура шоколадный
Незнакомые слова (мнение студентов)	Пустое поле : 42 дельфин : 12 экспедиция : 9 Кабардино-Балкария : 9 инвалид : 7 отель : 6 опасный : 5 становиться : 4 достопримечательность : 3 полтора : 3 фототехника : 3 офис : 2 ненавидеть : 2 стиль : 2 National Geographic: 1 фотограф : 1 плавать : 1	прокатиться : 38 купальник : 23 протестировать : 23 сомнение : 19 логично : 16 тариф : 16 запрещать : 15 абсолютно : 15 Пустое поле : 13 аквапарк : 10 блог : 8 горка : 7 тестировать : 7 камера : 7 уговаривать : 6 кстати : 4 тратить : 4 сторона : 3 попадать : 3 график : 2	вафельный : 43 разжевывать : 42 ополаскиваться : 42 мармеладный : 41 взбодриться: 30 нейтрализовать : 29 следить : 28 желудок : 27 обучать : 20 кусочек : 15 изделие : 14 просыпаться : 12 съедать : 12 глядеть : 9 дегустация : 9 происходить : 7 тестировать : 7 производство : 7 рецептор : 7 прополаскивать : 6

	офисный : 1 мечта : 1	Троя : 2 снимать : 2 Ква-Ква : 2 относиться : 2 повар : 1 пробовать : 1 сложный : 1 отдыхающий : 1 логичный : 1 неуверенно : 1 видео : 1	период : 6 Пустое поле : 5 сладость : 4 десерт : 4 минимум : 3 технологический : 3 горький : 3 технолог : 3 влиять : 3 либо : 3 продукция : 2 ополаскивать : 2 оставаться : 2 рот : 2 постоянно : 2 течение : 2 дизайн : 1 никак : 1 кое-кто : 1 фигура : 1 молочный : 1 основной : 1 сладкий : 1 портить : 1 вкусовой : 1 какой-либо : 1
Незнакомые слова (мнение преподавателей)	Пустое поле : 4 инвалид : 2 дельфин : 2 полтора : 2 Кабардино-	уговаривать : 4 тариф : 3 аквапарк : 3 блог : 3 ква-ква : 3	нейтрализовать : 5 разжевывать : 6 тестировать : 4 рецептор : 4 вафельный : 4

	Балкария : 1 экспедиция : 1	протестировать : 3 сомнение : 2 прокатиться : 2 абсолютно : 2 кстати : 1 горка : 1 тестировать : 1 купальник : 1 запрещать : 1 что-то : 1 камера : 1	ополаскиваться : 4 мармеладный : 3 следить : 2 кое-кто : 2 изделие : 2 технолог : 2 вкусовой : 2 горький : 1 десерт : 1 съедать : 1 ополаскивать : 1 молочный : 1 Пустое поле : 1 какой-либо : 1 ополаскиваться : 1 желудок : 1 продукция : 1 вкус : 1 белый : 1 количество : 1 технологический : 1
Слова, которых нет в минимуме, но их отметило незнакомыми меньше 5% студентов	фотограф проект многий	фигура попадать неуверенно просто сайт профессиональный троя ква-ква меню	молочный просто вкус технолог никак течение минимум контроль шоколадный технологический

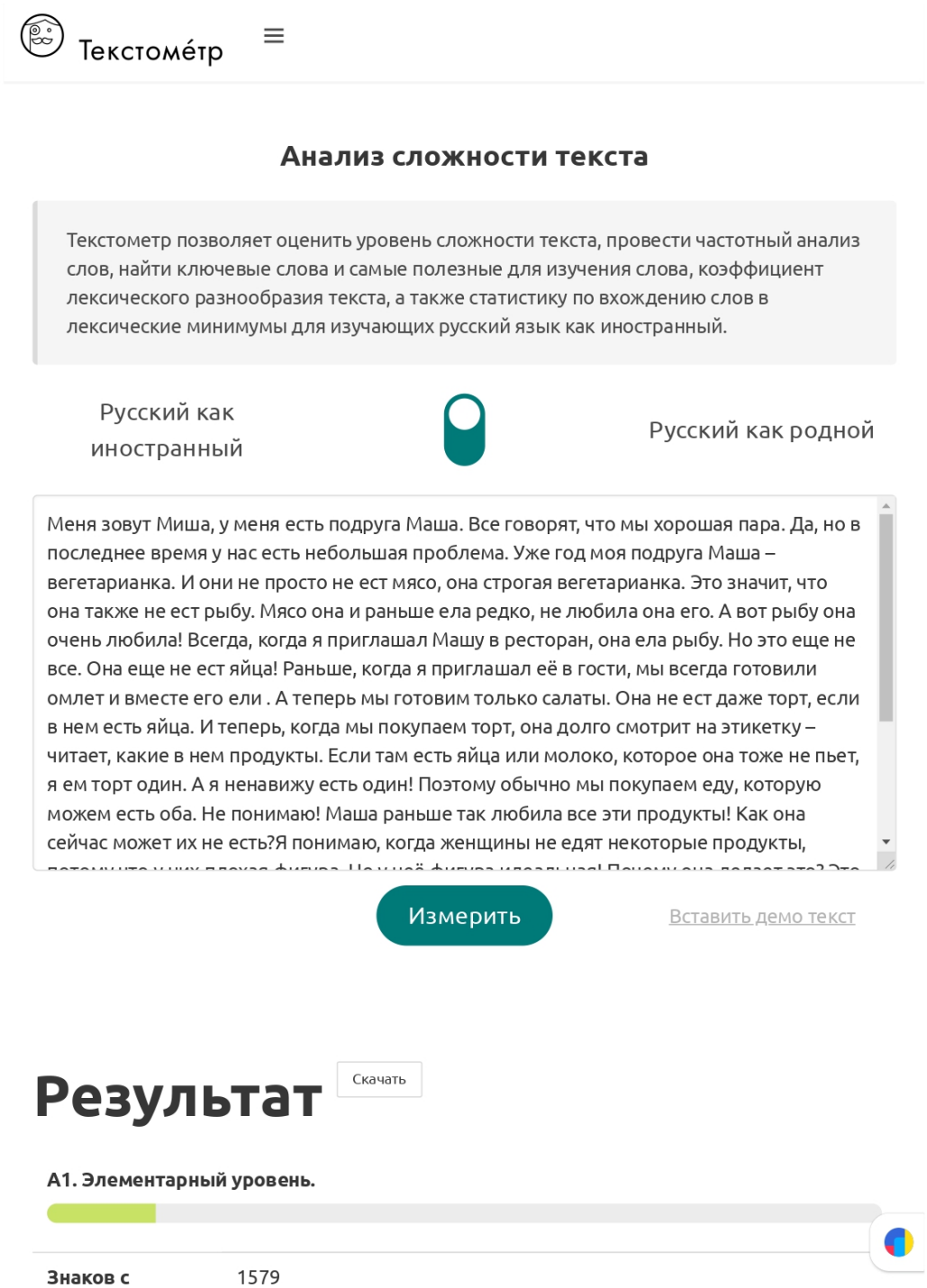
			фигура меню какой-либо влиять кое-кто вкусовой
Слова, которые есть в ЛМ уровня B1, но > 5% студентов отметили незнакомыми)	экспедиция становиться опасный	горка абсолютно тратить запрещать	съедать происходить обучать желудок производство период просыпаться желудок
Среднее количество незнакомых конструкций по мнению студентов/преподавателей	0.07	0.37	0.64
Незнакомые конструкции (мнение преподавателей)	Пустое поле:5 пять-десять:1 собирать фототехнику:1	Пустое поле: 3 с сомнением:3 на камеру рассказывать:3 просила это сделать за меня: 2 со стороны аквапарка:2 время тратить на	Пустое поле: 3 технологические процессы: 3 вкусовые рецепторы находятся не в желудке, а на языке:2 создаем новые вкусы:1

		тебя: 2 проводить бесплатно:1 попадать на сайт:1 у них тариф:1 блог профессионального отдыхающего:1	то, что идет в производство: 1 количество шоколада к килограммах:1 глядя на меня:1 ничего такого не произойдет:1 оставшийся во рту: 1 никак не влияет на фигуру:1 просто так:1

ПРИЛОЖЕНИЕ В

Пример результата анализа текста в разработанном сервисе «Текстометр»

ПРИМЕР РЕЗУЛЬТАТА А



Предложений	36
Слов	295
Уникальных слов	115
Лексическое разнообразие	0.39
Ключевые слова	есть мясо рыба овощ торт яйцо вегетарианка приглашать ресторан салат
Самые полезные слова	торт значить продукт ненавидеть правда поэтому если просто оба злой небольшой также строгий некоторый пара фигура последний
Лексический список A1 покрывает	89%
Не входит в лексический список A1	торт продукт строгий поэтому ненавидеть вегетарианка если просто небольшой вегетарианский некоторый пара фигура этикетка идеальный значить последний омлет оба глупо злой также немой правда абсурд
Лексический список A2 покрывает	94%
Не входит в лексический список A2	продукт ненавидеть вегетарианка омлет оба этикетка строгий идеальный немой некоторый вегетарианский пара фигура абсурд
Лексический	95%

список B1 покрывает	
Не входит в лексический список B1	ненавидеть вегетарианка омлет оба этикетка строгий идеальный немой вегетарианский пара фигура абсурд
Лексический список B2 покрывает	98%
Не входит в лексический список B2	немой вегетарианский вегетарианка фигура
Лексический список C1 покрывает	99%
Не входит в лексический список C1	вегетарианский вегетарианка
Частотный список 5000 покрывает	93%
Полезные слова, которых нет в лексическом минимуме	оба продукт ненавидеть
Редкие слова	вегетарианский омлет
Лексический список РКИ-дети 1000 покрывает	90 %
Не входит в список РКИ-дети 1000	проблема строгий ненавидеть ужинать вегетарианка вегетарианский пара фигура счастие этикетка идеальный значить редко омлет оба

	глупо злой уставать немой правда абсурд
Лексический список РКИ-дети 2000 покрывает	94 %
Не входит в список РКИ-дети 2000	ненавидеть вегетарианка омлет оба глупо этикетка идеальный немой вегетарианский фигура абсурд
Лексический список РКИ-дети 5000 покрывает	95 %
Не входит в список РКИ-дети 5000	вегетарианка оба глупо этикетка идеальный немой вегетарианский абсурд
Изучающее чтение текста займет	10 мин.
Просмотровое чтение текста займет	6 мин.
Возможные грамматические темы	Местоимения Причастия Краткие формы прилагательных и причастий
Частотный словарь по тексту	она 24 есть 16 не 16 я 14 быть 7 в 7 и 6 мочь 6 мы 6 это 6

когда	5
мясо	5
но	5
что	5
а	4
маша	4
рыба	4
только	4
у	4
все	3
говорить	3
если	3
или	3
как	3
любить	3
овощ	3
они	3
приглашать	3
продукт	3
рано	3
ресторан	3
тоже	3
торт	3
яйцо	3
вегетарианка	2
всегда	2
готовить	2
да	2
даже	2
еще	2
конечно	2
который	2
обычно	2

один	2
он	2
подруга	2
покупать	2
понимать	2
почему	2
проблема	2
салат	2
сейчас	2
так	2
теперь	2
фигура	2
абсурд	1
вегетарианский	1
весь	1
вместе	1
вот	1
время	1
где	1
глупо	1
год	1
гость	1
делать	1
долго	1
должный	1
его	1
еда	1
женщина	1
жить	1
звать	1
злой	1
значить	1
идеальный	1
к	1

какой	1
мало	1
миша	1
мой	1
молоко	1
на	1
небольшой	1
некоторый	1
немой	1
ненавидеть	1
оба	1
обедать	1
омлет	1
очень	1
пара	1
пить	1
плохой	1
последний	1
потому	1
поэтому	1
правда	1
просто	1
редко	1
смотреть	1
строгий	1
счастие	1
также	1
там	1
то	1
уже	1
ужинать	1
уставать	1
фрукт	1

Пример скачиваемого результата анализа текста в формате TXT в веб-сервисе «Текстометр»

Text

Меня зовут Миша, у меня есть подруга Маша. Все говорят, что мы хорошая пара. Да, но в последнее время у нас есть небольшая проблема. Уже год моя подруга Маша – вегетарианка. И они не просто не ест мясо, она строгая вегетарианка. Это значит, что она также не ест рыбу. Мясо она и раньше ела редко, не любила она его. А вот рыбу она очень любила! Всегда, когда я приглашал Машу в ресторан, она ела рыбу. Но это еще не все. Она еще не ест яйца! Раньше, когда я приглашал её в гости, мы всегда готовили омлет и вместе его ели. А теперь мы готовим только салаты. Она не ест даже торт, если в нем есть яйца. И теперь, когда мы покупаем торт, она долго смотрит на этикетку – читает, какие в нем продукты. Если там есть яйца или молоко, которое она тоже не пьет, я ем торт один. А я ненавижу есть один! Поэтому обычно мы покупаем еду, которую можем есть оба. Не понимаю! Маша раньше так любила все эти продукты! Как она сейчас может их не есть? Я понимаю, когда женщины не едят некоторые продукты, потому что у них плохая фигура. Но у неё фигура идеальная! Почему она делает это? Это все так глупо! Она говорит, что сейчас, когда она не ест мясо, она меньше устает. Но это абсурд! Конечно, она может жить, как она хочет. Я даже могу приглашать её только в вегетарианские рестораны. Правда, где обедать или ужинать – это не проблема. В ресторане она обычно ест салат, а я ем мясо или рыбу. К счастью, она не говорит, что я тоже, как и она, должен есть только овощи. Да, конечно, я тоже ем фрукты и овощи. Но если я не ем мясо, то я злой! Я не могу есть только овощи! Почему она может?!

Уровень текста

0.9

Уровень текста

A1. Элементарный уровень.

Слов

295

Знаков с пробелами

1579

Предложений

36

Уникальных слов

115

Лексическое разнообразие

0.39

Изучающее чтение текста займет

10 мин.

Просмотровое чтение текста займет

6 мин.

Частотный словарь по тексту

она,24

есть,16

не,16

я,14

быть,7

в,7

и,6

мочь,6

мы,6

это,6

когда,5

мясо,5

но,5

что,5

а,4

маша,4

рыба,4

только,4

у,4

все,3

говорить,3

если,3

или,3

как,3

любить,3

овощ,3
они,3
приглашать,3
продукт,3
рано,3
ресторан,3
тоже,3
торт,3
яйцо,3
вегетарианка,2
всегда,2
готовить,2
да,2
даже,2
еще,2
конечно,2
который,2
обычно,2
один,2
он,2
подруга,2
покупать,2
понимать,2
почему,2
проблема,2
салат,2
сейчас,2
так,2
теперь,2
фигура,2
абсурд,1
вегетарианский,1
весь,1

вместе,1

вот,1

время,1

где,1

глупо,1

год,1

гость,1

делать,1

долго,1

должный,1

его,1

еда,1

женщина,1

жить,1

звать,1

злой,1

значить,1

идеальный,1

к,1

какой,1

мало,1

миша,1

мой,1

молоко,1

на,1

небольшой,1

некоторый,1

немой,1

ненавидеть,1

оба,1

обедать,1

омлет,1

очень,1

пара,1

пить,1

плохой,1

последний,1

потому,1

поэтому,1

правда,1

просто,1

редко,1

смотреть,1

строгий,1

счастие,1

также,1

там,1

то,1

уже,1

ужинать,1

уставать,1

фрукт,1

хороший,1

хотеть,1

читать,1

этикетка,1

этот,1

Ключевые слова

есть

мясо

рыба

овощ

торт

яйцо

вегетарианка

приглашать

ресторан

салат

Лексический список А1 покрывает

89

Не входит в лексический список А1

торт

продукт

строгий

поэтому

ненавидеть

вегетарианка

если

просто

небольшой

вегетарианский

некоторый

пара

фигура

этикетка

идеальный

значить

последний

омлет

оба

глупо

злой

также

немой

правда

абсурд

Лексический список А2 покрывает

94

Не входит в лексический список А2

продукт

ненавидеть

вегетарианка

омлет

оба

этикетка

строгий

идеальный

немой

некоторый

вегетарианский

пара

фигура

абсурд

Лексический список B1 покрывает

95

Не входит в лексический список B1

ненавидеть

вегетарианка

омлет

оба

этикетка

строгий

идеальный

немой

вегетарианский

пара

фигура

абсурд

Лексический список B2 покрывает

98

Не входит в лексический список B2

немой

вегетарианский

вегетарианка

фигура

Лексический список С1 покрывает

99

Не входит в лексический список С1

вегетарианский

вегетарианка

Частотный список 5000 покрывает

93

Лексический список РКИ-дети 1000 покрывает

90%

Не входит в список РКИ-дети 1000

проблема

строгий

ненавидеть

ужинать

вегетарианка

вегетарианский

пара

фигура

счастие

этикетка

идеальный

значить

редко

омлет

оба

глупо

злой

уставать

немой

правда

абсурд

Лексический список РКИ-дети 2000 покрывает

94%

Не входит в список РКИ-дети 2000

ненавидеть

вегетарианка

омлет

оба

глупо

этикетка

идеальный

немой

вегетарианский

фигура

абсурд

Лексический список РКИ-дети 5000 покрывает

95%

Не входит в список РКИ-дети 5000

вегетарианка

оба

глупо

этикетка

идеальный

немой

вегетарианский

абсурд

Самые полезные слова

торт

значить

продукт

ненавидеть

правда

поэтому

если

просто

оба

злой

небольшой

также

строгий

некоторый

пара

фигура

последний

Редкие слова

вегетарианский

омлет

Полезные слова, которых нет в лексическом минимуме

оба

продукт

ненавидеть

Возможные грамматические темы

Местоимения

Причастия

Краткие формы прилагательных и причастий

ПРИЛОЖЕНИЕ Г.
СВИДЕТЕЛЬСТВО О ГОСУДАРСТВЕННОЙ РЕГИСТРАЦИИ ПРОГРАММЫ ДЛЯ ЭВМ

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2021661785

Текстометр

Правообладатель: ***Федеральное государственное бюджетное образовательное учреждение высшего образования «Государственный институт русского языка им. А.С. Пушкина» (RU)***

Авторы: ***Лапошина Антонина Николаевна (RU), Лапошин Алексей Александрович (RU), Лебедева Мария Юрьевна (RU)***

Заявка № **2021660920**
Дата поступления **09 июля 2021 г.**
Дата государственной регистрации
в Реестре программ для ЭВМ ***15 июля 2021 г.***

Руководитель Федеральной службы
по интеллектуальной собственности

Г.П. Ивлиев