# Об определении сложности мультимодальных текстов сайтов российских университетов (на материале корпуса PolyLing)

М.С. Коган, Д.А. Гаврилик, С.В. Чистякова, А.В. Рубцова, Е.Р. Никулина, А.В. Черкас, М.В. Болсуновская

***Аннотация*— Веб-сайты университетов, представляющие собой сложные мультимодальные конструкции, играют очень важную роль в современном образовательном пространстве: являясь неотъемлемой частью электронной информационно-образовательной среды университета, они осуществляют коммуникацию вуза с внешней и внутренней средой и служат инструментом формирования имиджа вуза. С момента своего появления в 1990-х гг они привлекают внимание исследователей. Однако, анализ публикаций, посвященный этой теме выявил, что лишь немногие авторы занимались оценкой сложности/удобочитаемости мультимодальных текстов, размещенных на сайтах университетов. Данное исследование, целью которого является оценка сложности сайтов ведущих российских университетов, частично восполняет этот пробел. Для реализации поставленной цели на первом этапе исследования был собран корпус из 1000+ текстов путем парсинга новостных разделов сайтов отобранных университетов. Тематический состав собранного корпуса, названного PolyLing, был определен методом кластерного анализа. Оценка сложности текстов проводилась автоматически с помощью программы, написанной на Python на основе индексов сложности текстов, и 132 респондентами, которые оценивали представительную выборку из корпуса по 10 критериям, относящимся к 3-м категориям (лингвистические, структурно-логические и вызывающие заинтересованность читателя), отвечая на вопросы специально разработанной анкеты. Корреляционный анализ показал удовлетворительное согласие между автоматическим человеческим оцениванием текстов средней сложности и отрицательную корреляцию между более легкими и более сложными текстами. В статье предлагается возможное объяснение выявленного расхождения в оценке сложности текста, и намечаются направления для дальнейшего исследования.**

***Ключевые слова*— Автоматическое определение сложности текстов, индексы читабельности, корпус PolyLing, мультимодальные сайты российских университетов, оценка сложности текстов респондентами.**

Статья получена 2 ноября 2023.

Коган Марина Самуиловна, Санкт-Петербургский политехнический университет Петра Великого, канд. техн. наук, доцент (e-mail: kogan_ms@spbstu.ru).

Гаврилик Дарья Александровна, Санкт-Петербургский политехнический университет Петра Великого, ассистент (e-mail: gavrilik_da@spbstu.ru).

Чистякова Светлана Владимировна, Санкт-Петербургский политехнический университет Петра Великого, специалист,.(e-mail: chistyakova_sv@spbstu.ru).

Рубцова Анна Владимировна, Санкт-Петербургский политехнический университет Петра Великого, докт. пед. наук, профессор (e-mail: rubtsova_av@spbstu.ru).

Никулина Елизавета Романовна, Санкт-Петербургский политехнический университет Петра Великого, специалист, магистр лингвистики (e-mail: nikulina_er@spbstu.ru).

Черкас Алина Владимировна, Санкт-Петербургский политехнический университет Петра Великого, инженер-исследователь, аспирант (e-mail: alina.cherkas@spbpu.com).

Болсуновская Марина Владимировна Санкт-Петербургский политехнический университет Петра Великого, канд. техн. наук, доцент. (e-mail: bolsun_mv@spbstu.ru).

Исследование частично поддержано Программой стратегического академического лидерства Министерства науки и высшего образования РФ «Приоритет2030» №6122042600078-2 «Цифровые технологии в лингвистике: модель автоматической оценки речевого воздействия мультимодального электронного текста». Статья подготовлена по итогам выступления на Международной объединённой конференции «Интернет и современное общество» (IMS-2023).

## I. Введение

В последние годы университеты по всему миру постоянно наращивают свое присутствие в Интернет-пространстве, активно развивая веб-сайты своих учебных заведений. Веб-сайты оказались эффективным способом коммуникации как со студентами и сотрудниками университета, так и с внешними пользователями, представляя информацию о разных сторонах жизни университета: учебной, административной, научной, культурной, спортивной, просветительской, о взаимодействии с промышленными партнерами и другими университетами и др. Официальный веб-сайт университета также служит эффективным инструментом поддержания имиджа и продвижения в национальных и международных рейтингах. Авторы [1] выделяют 3 основные функции, которые выполняют сегодня веб-сайты университетов:

Обеспечение понятной и эффективной коммуникации вуза с широкой аудиторией (абитуриенты, студенты, их родители, профессорско-преподавательский состав, специалисты по учебной работе и др.), выстраивание доверия с пользователем;

Построение, удобной информационной среды для потенциального пользователя;

Поддержание имиджа и продвижение в рейтингах:

формирование репутационной стратегии, обозначение конкурентных преимуществ и поддержка бренда вуза. [1]

Обеспечение понятной и эффективной коммуникации вуза с широкой аудиторией становится особенно важной задачей в условиях информатизации рынка образовательных услуг. Основной единицей такой коммуникации выступают мультимодальные электронные тексты. Улучшение их информативности и удобочитаемости повышает их социо-коммуникативную значимость для абитуриентов и студентов, среди которых могут быть как носители русского языка и так и те, для кого он не является родным; формирует привлекательный цифровой имидж университета и способствует повышению академической репутации вуза.

Веб-пространство современного высшего учебного заведения образуют тысячи страниц уникальных текстов, количество которых непрерывно возрастает.

Представленное на них разнообразие жанров, некоторые исследователи называют «колонии дискурса» (discourse colonies) – сложные структуры, составленные из множества текстов разных поджанров (например, тексты, напоминающие туристический справочник, эссе, блоги, фрагменты дневниковых записей, лекции, локальные нормативные акты и др.), направленные на решение единственной коммуникативной задачи: представить университет и содействовать усилению его цифрового мультимодального имиджа, построенного на взаимодействии вербальных и невербальных средств. [2].

Сайты университетов стали объектом исследования с момента их появления в 90-х годах прошлого века [3]. Краткий обзор направлений исследований веб-сайтов университетов мы дадим в следующем разделе.

## II. Обзор литературы

### *A. Направления исследований веб-сайтов университетов*

Исследования сайтов университетов многочисленны, проводились исследователями разных стран на материале разных сайтов с разным фокусом и методологией исследования. В большинстве исследований учитывалась мультимодальность университетских сайтов и исследовались такие вопросы как роль невербальных элементов в *привлечении потенциальных студентов* в университет. Автор указывает на постоянную обновляемость и качественное развитие сайтов ведущих университетов [2]. Золло (Zollo) также отмечает большую интерактивность веб-сайтов английских университетов по сравнению с итальянскими, которые используют новые коммуникативные стратегии и стратегии убеждения, рассматривающие студента как приобретателя товаров и услуг. Этот тренд отсутствует на сайтах итальянских университетов [4].

Изучению *представления на сайтах иностранных студентов* было посвящено исследование веб-сайтов трех австралийских университетов, проведенное в 2019г [5] с использованием критического мультимодального дискурс-анализа. В другой работе те же авторы изучали *способы обращения к зарубежным студентам* на сайтах австралийских и китайских университетов. Исследователи обратили внимание, что австралийские университеты обращаются к потенциальным студентам из других стран на их родном языке: китайском, вьетнамском или арабском (наряду с английским) при описании правил приема или взаимодействия университета с регионами, в которых проживают потенциальные студенты, демонстрируя таким образом тенденцию/тренд на сближение с будущими иностранными студентами [6].

Применяя метод мультимодального дискурса к анализу раздела сайта *Why Choose* одного британского и двух австралийских университетов, авторы выявили причины, по которым иностранные студенты должны отдать предпочтение тому или другому университету. Они считают, что традиционные университетские ценности такие как учеба, научная деятельность, возможность изменить свою жизнь, расширить кругозор и круг общения умело вписаны в рыночный дискурс. А иностранным студентам предлагается роль agentive doers (активных деятелей) для достижения собственных целей и максимальной прибыли от инвестирования в свое будущее [7].

Анализ раздела *Mission statements (Миссия университета)* английских версий сайтов университетов 35 российских федеральных и национальных исследовательских университетов и 12 ведущих казахских университетов позволил выявить 1) связь между научным профилем университета и его лидерством; 2) включенность университета в практику маркетизации, взаимовыгодных отношений с бизнес-структурами; 3) факт, что и российские и казахские университеты поддерживают политику своих государств и стратегии экономического развития [8].

Анализ содержания страниц *History* (*История***)** и *About the University* (*Об университете*) на веб-сайтах 45 ведущих российских университетов позволил исследователям понять, как как университеты смотрят на свое прошлое и какие коммуникативные и речевые приемы используют для выражения этого видения. В этом разделе университеты подчеркивают свое участие в жизни страны в исторической перспективе. Англоязычные версии этих разделов гораздо короче русскоязычных, что говорит об их направленности в первую очередь на русскоязычного читателя и потребителя образовательных услуг [9].

Тренд на маркетизацию заметный на сайтах российских университетов, западные исследователи называют неолиберальный подход в высшем образовании (см., например, [2, 6, 7, 10]).

Неолиберализм рассматривает высшее образование с точки зрения рыночных отношений, в рамках которых университеты конкурируют друг с другом за обладание б**о**льшей долей рынка посредством маркетизации и применения моделей корпоративного управления,

рассматривая студентов как потребителей образовательных услуг. Авторы [10] иллюстрировали эту тенденцию университетов на примере раздела сайта университета Arcadia (Пенсильвания, США), посвященного обучению за границей (*Study Abroad*), в котором текст и фотографии фокусируются на туристическом опыте, которые приобретают участники обменных программ, оставляя в тени образовательный аспект таких программ.

К. Хайлэнд провел исследование 100 *личных страниц профилей преподавателей* на сайтах разных университетов, обнаружив их сходство в том, что они главным образом содержат профессиональные биографии владельцев, отражают их научные интересы и список публикаций; по цветовой гамме и стилю дизайн личных страниц преподавателей не отличается от дизайна сайта университетов, в которых они работают; фоторяд ограничен одной фотографией паспортного формата; внешние ссылки ведут к ресурсам/сайтам/источникам, подтверждающих их компетентность и научные достижения. В целом личные страницы преподавателей на сайтах университетах преследуют цель продвижения автора и университета во внешнем мире, образуя отдельный поджанр среди представленных на сайтах университетов [11].

Среди больших проектов, касающихся исследований сайтов университетов, особое место занимает проект итальянских исследователей, поддержанный грантом Евросоюза, которые собрали англоязычный корпус объемом 90 миллионов слов с веб-сайтов европейских и английских университетов (*WaC*[1]-Eu (английский как лингва франка (ELF) и *ukWaC* ([английский] родной, NAT) соответственно), используя полуавтоматическую процедуру создания корпуса с помощью технологии «интернет как корпус». Анализ URL адресов собранных страниц показывает, что наиболее частотными в списке являются домены “news” («новости»), “courses” («курсы»), и “research” («научные исследования»). На основе этого корпуса был проведен ряд контрастивных исследований, а именно модальных и полумодальных глаголов [12], степени идиоматичности текстов, написанных носителями и не носителями английского языка [13], метафоры «путешествия» как коммуникативной стратегии, описывающей процесс обучения, которая используется на сайтах как английских так и итальянских университетов [14], способ изображения студентов с помощью визуальных и письменных компонентов университетских веб-сайтов [15]. Во всех исследованиях были выявлены различия между подкорпусами английский как лингва франка и английский родной (ELF и NAT подкорпуса соответственно). Например, в ELF было меньше английских высоко идиоматичных коллокаций, используемых носителями; с другой стороны, в англоязычных версиях сайтов обнаружены не существующие в английском языке коллокации, являющиеся заимствованиями из национальных языков (итальянского, французского, испанского) или слова, содержащие орфографические ошибки [13].

### *B. Оценка сложности текстов сайтов университетов*

Современная научная парадигма определяет сложность восприятия текста как функцию двух основных групп мегапараметров: индивидуальных характеристик читателя и объективных параметров текста. В группе последних ученые выделяют количественные и качественные параметры текста. Группа количественных параметров (метрик) включает длину текста в словах, среднюю длину предложения в словах, длину слова в слогах и др., определяющие скорость обработки информации (количество таких параметров разное в разных языках. Для русского языка этот показатель достигает 179 [16]. К качественным параметрам относятся повествовательность (наличие сюжета), синтаксическая простота текста, степень конкретности и связность. К индивидуальным характеристикам читателя относится его возраст, уровень образования, уровень владения иностранным языком, если речь идет об иноязычных текстах, общая эрудиция, специальность и др. [17].

Интуитивное понятие сложности / легкости текста для чтения и связанной с этим скорости чтения и понимания текста в лингвистике XX века было формализовано в виде индексов удобочитаемости (readability), которые, однако, не учитывают особенностей восприятия мультимодальных текстов.

Как отмечает О.Н. Ляшевская «из-за большого разнообразия ситуаций, в которых встречаются Текст и его Читатель, проводятся исследования в отдельных областях» [18: 409]:

1. оценка удобочитаемости упражнений и учебных текстов для иностранцев, изучающих язык как неродной (L2); (на материале русского языка как иностранного такие исследования проводят [16, 19];
2. экспертиза школьных учебников, экзаменационных тестов и других материалов (для носителей L1). На материале русского языка этим занимаются исследователи из Казанского университета [17, 20, 21];
3. оценка читабельности деловой документации; рекламных материалов; медицинской документации, правительственных сайтов, юридических документов. В России удобочитаемость русскоязычных юридических текстов изучают [22, 23];
4. оценка текстов веб-сайтов с точки зрения привлекательности, понятности для целевой аудитории и др..

Рассмотрим более подробно оценку удобочитаемости текстов веб-сайтов. Анализ литературы показывает, что чаще всего исследователи анализируют удобочитаемость веб-сайтов официальных организаций, включая правительственные [24, 25]. Только в 3-х работах наряду с другими характеристиками, приведенными в руководстве по доступности сетевого контента (Web Content Accessibility Guidelines (WCAG)) исследовалась удобочитаемость веб-сайтов

[1] WaC – web as corpus

университетов. Карху (Karhu) и ее коллеги анализировали удобочитаемость разделов, посвященных истории, на сайтах 7 университетов в Финляндии, используя находящийся в свободном доступе инструмент (http://flesh.sourceforge.net). По их данным три сайта имеют высокую сложность («Hard»), а четыре – очень высокую сложность («Very Hard») [26]. Патра (Patra) с коллегами изучали сайты государственных структур на предмет соответствия требованиям WCAG 2.0, включая сайты 5 индийских университетов. Они обнаружили, что 39.60% образовательных веб-сайтов не соответствуют требованиям читабельности (readability) в категории «Понятность» (Understandable category) [27, p.16]. Турецкий исследователь Akgül в сравнительном исследовании сайтов 179 государственных и частных турецких университетов оценил читабельность сайтов с помощью 6 наиболее популярных формул удобочитаемости [28, Table 12]. Он заключает, что веб-сайты и государственных и частных университетов являются трудными для чтения и указывает на недостаток, присущий измерению читабельности по формулам читабельности, требующих адаптации для разных языков. Исследователь также учитывал некоторые мультимодальные характеристики читабельности текстов на сайтах: капитализацию, подчеркивание, центрирование [28].

Авторы ряда исследований оценивали удовлетворенность целевой аудитории контентом и техническими характеристиками университетских сайтов посредством опросов. Например, бангладешские исследователи [29] провели опрос в 22 государственных и частных университетах Бангладеш с привлечением 1820 студентов для ответа на 23 вопроса анкеты. Интересно отметить, что ответах на последний свободный вопрос о факторах, требующих улучшения на веб-сайтах университетов Бангладеш, никто не упомянул высокую сложность текста. Среди недостатков наиболее часто упоминались технические (скорость загрузки сайта), дизайн интерфейса, устаревшая информация на доске объявлений, неполнота информации о преподавателях (отсутствие информации об опыте работы, интересах), недостатки онлайн сервисов регистрации и оплаты [29: 12].

Ввиду отсутствия подобных исследований на материале сайтов российских университетов, мы решили провести оценку удобочитаемости текстов на сайтах российских университетов, сравнив автоматическую оценку сложности с оценкой респондентов – студентов и аспирантов СПбПУ Петра Великого в рамках первого этапа исследования.

## III. Данные и методология исследования

### *A. Создание корпуса мультимодальных текстов сайтов российских университетов*

Для оценки удобочитаемости текстов веб-сайтов российских вузов был собран корпус. В рамках пилотной части исследования был решено собрать корпус статей с новостных разделов сайтов университетов. При выборе этого новостного раздела мы руководствовались следующими факторами: 1) частые и регулярные обновления этого раздела на университетских сайтах, 2) широкий круг освещаемых вопросов (новейшие разработки и исследования, новости науки и образования, административной и студенческой жизни, международного сотрудничества вузов и др.), 3) статьи как правило пишутся профессиональными журналистами; 4) предположительно, это наиболее посещаемый раздел сайта после домашней страницы; 5)в идеале представляет интерес для всех категорий читателей (преподавателей, студентов, сотрудников университета, потенциальными абитуриентами и их родителями). Были выбраны 20 сайтов ведущих высших учебных заведений Российской Федерации, входящих в программу Приоритет 2030, с которых было собрано более тысячи текстов для анализа с охватом новостей за период 1.01.2022 – 31.03. 2022. В таблице I представлена статистика корпуса:

Таблица I. Статистика корпуса сайтов университетов

| **Университет** | **Тексты** | **Токены** |
|---|---|---|
| ВШЭ | 84 | 156933 |
| СПбГУ | 50 | 47561 |
| СПбПУ | 215 | 158924 |
| ТГУ | 27 | 12740 |
| ТПУ | 50 | 22831 |
| БГТУ | 40 | 16286 |
| БФУ | 40 | 13824 |
| ВАС | 40 | 7817 |
| ВГУ | 40 | 12497 |
| ДВГУПС | 40 | 7185 |
| КФУ | 40 | 21533 |
| ПсковГУ | 40 | 15770 |
| СамМГУ | 40 | 13259 |
| САФУ | 40 | 20140 |
| СибСпорт | 40 | 13924 |
| СурГУ | 40 | 23255 |
| ТюмГУ | 40 | 20388 |
| Ун. Лесгафта | 40 | 11246 |
| УрФУ | 40 | 17304 |
| ЮФУ | 40 | 23460 |
| Всего | 1026 | 636877 |

Сбор данных для корпуса проводился в 2 этапа: 1) оставление списка вузов/ списка URL выбранных университетов, 2) автоматический парсинг веб-сайта с извлечением текста новостной рубрики и метаданных статьи.

Мы использовали Python 3 для автоматического парсинга выбранных российских университетов. Для сбора данных была разработана Программа для автоматического сбора текстов новостных рубрик сайтов ВУЗов [30], которая создавала два типа файлов: один – с текстом (в формате «обычный текст», txt), другой – с метаданными, включая название статьи, количество иллюстраций, университет, автор и др. Каждый текст в корпусе имеет свой уникальный ID и метаданные. Морфологическая и синтаксическая разметка корпуса осуществлялась с помощью адаптированного анализатора Pymorphy [31]. Корпус

получил название PolyLing.

### *B. Анализ содержания корпуса PolyLing с помощью методов кластеризации*

С целью изучения содержания текстов корпуса PolyLing был выполнен кластерный анализ, позволяющий группировать тексты по схожим элементам и выявлять общие темы или паттерны в наборе данных. Тексты были предобработаны, прошли процесс лемматизации и удаления стоп-слов. Для каждого текста с помощью библиотеки Rutermextract (https://github.com/igor-shevchenko/rutermextract) был выделен набор ключевых слов (150-300 элементов). Кластеризация выполнялась независимо как для полных текстов, так и для извлеченных ключевых слов. Для кластерного анализа применялись следующие методы:

- KMeans (метод k-средних)
- Spectral Clustering (спектральная кластеризация)
- Affinity Propagation (метод распространения близости)
- DBSCAN
- Agglomerative Clustering (иерархическая кластеризация)

Данные методы являются наиболее распространенными и эффективными для проведения кластерного анализа [32].

Для оценки полученных результатов была использована метрика Silhouette Score, которая позволяет определить точность группировки объектов внутри кластеров и степень их отличия от объектов в других кластерах.

Данная метрика является наиболее распространенным способом объединить коэффициенты среднего расстояния между объектами одного кластера и среднего расстояния между объектами следующего ближайшего кластера. Она не зависит от предположений о распределении данных или форме кластеров, что делает её более объективной и универсальной. Silhouette Score также полезен тем, что не требует разметки данных, и позволяет подбирать оптимальное количество кластеров [33]. Высокое значение Silhouette Score (близкое к 1) указывает на качественную кластеризацию, в результате применения которой объекты внутри кластеров располагаются плотно и отличаются от объектов в других кластерах. Низкое значение Silhouette Score (близкое к -1) указывает на низкую степень качества кластеризации, при которой объекты могут быть сгруппированы неправильно.

Для выбора оптимального числа кластеров было проведено сравнение данной метрики для разного количества кластеров, где наилучшее число соответствовало наивысшему значению Silhouette Score (Таблица II).

Таблица II. Соотношение количества кластеров и значения метрики Silhouette Score для различных методов кластеризации полного текста и ключевых слов

| **Метод кластеризации** | **Тип текста (значение метрики/кол-во кластеров)** | |
|---|---|---|
| | **полный** | **Ключевые слова** |
| KMeans | 0,37 /6 | 0,34 /6 |
| Spectral Clustering | 0,35 /3 | 0,39 /3 |
| Affinity Propagation | 0,35 /6 | 0,33 /5 |
| DBSCAN | 0,2 /5 | 0,2 /3 |
| Agglomerative Clustering | 0,32 /6 | 0,35 /3 |

Значения метрики *Silhouette Score* для полных текстов не превышали 0.37. Наилучшие результаты показали следующие методы кластеризации: метод k-средних (0.37), метод распространения близости (0.35), спектральная кластеризация (0.35), иерархическая кластеризация (0.32). Данные методы предполагают выделение 3-6 кластеров. Худший результат при оценке метрикой Silhouette Score показал метод DBSCAN (0.2). Значение близкое к нулю свидетельствует о том, что расстояние между кластерами незначительно (рис. 1).

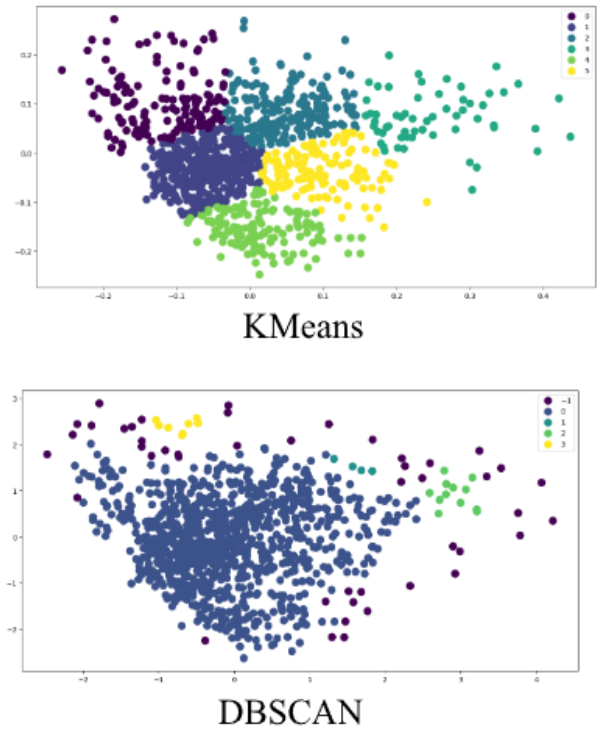
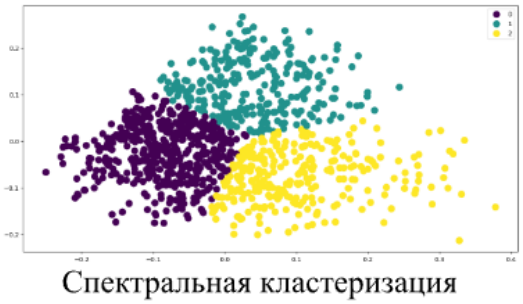



Рис. 1. Результаты кластеризации полных текстов для метода k-средних (лучший результат) и метода DBSCAN (худший результат).

Кластеризация наборов ключевых слов показала аналогичные результаты. *Silhouette Score* остался на уровне от 0.2 до 0.39, однако, количество выделенных кластеров оказалось ниже для спектральной и иерархической кластеризации и метода DBSCAN – 3 кластера (Рис. 2).



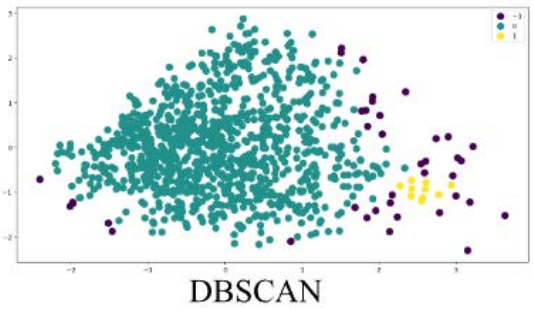



Рис. 2. Результаты кластеризации ключевых слов для метода спектральной кластеризации (лучший результат) и метода DBSCAN (худший результат).

Центроиды кластеров содержат схожие элементы, например, “наука”, “проект”, “технология”. При этом кластеры включают и уникальные элементы, позволяющие для некоторых их них примерно определить тематику. Для кластера 1 - это гуманитарная направленность, для кластера 4 - семья, для кластера 6 - спорт. Кластеры 2, 3 и 5 содержат достаточно большое количество одинаковых или синонимичных элементов, что затрудняет выделение более узкой тематики.

Новые элементы обнаруживаются только при делении корпуса на 15 и более кластеров, что позволяет сделать предположение о том, что процент текстов специальной тематики в корпусе невелик. В основном, в корпусе преобладают тексты общей направленности, посвященные учебному процессу. Для такого типа корпусов наиболее чувствительными стали метод k-средних и спектральная кластеризация.

Таким образом, анализ центроидов и результаты метрики *Silhouette Score* для проведенного кластерного анализа указывают на то, что корпус *PolyLing* однороден в разрезе представленных тематик. Несмотря на то, что источниками являются сайты разных вузов, использованные методы не позволяют найти значимые различия между ними. Тексты содержат информацию об университетских мероприятиях, программах, инициативах и академических достижениях.

### *C. Автоматическое определение сложности текста*

С помощью написанного скрипта на Python была рассчитана удобочитаемость текстов веб-сайтов университетов. Для этого использовалась специализированная библиотека ruTS для расчета показателей удобочитаемости текстов на русском языке. Библиотека ruTS представляет собой пакет Python, который содержит набор показателей удобочитаемости, включая Flesch-Kincaid Reading Ease (FRE), Gunning FOG, Simple Measure of Gobbledygook (SMOG) и индекс Coleman-Liau (CLI) [34].

В расчетах использовались формулы удобочитаемости FRE, Gunning FOG и SMOG, адаптированные ранее для русского языка. Адаптация учитывала уникальные лингвистические особенности русского языка и морфологические и синтаксические различия между английским и русским языками [35].

В результате автоматической оценки сложности все тексты были распределены по трем категориям «более легкие», «средней сложности» и «более сложные», используя терминологию [36].

### *D. Организация и проведение анкетирования*

Исследование субъективной сложности текстов новостных разделов веб-сайтов университетов проводилось путем офлайн и онлайн опроса студентов, аспирантов и сотрудников СПбПУ. В опросе приняли участие 132 респондента в возрасте от 17 до 45 лет. Для проведения исследования были составлены сеты из трех новостных текстов на русском языке, а также разработана специальная анкета. Каждый сет предлагался для оценки 5 респондентам. Для обеспечения репрезентативности выборки текстов для оценки по отношению к собранному корпусу тексты отбирались с учетом следующих факторов: 1) пропорциональности (количество текстов университета (далее подкорпус) было пропорционально количеству текстов в корпусе), 2) случайности (случайный отбор текстов внутри каждого подкорпуса); 3) соответствия объема текста среднему показателю.

Анкета включала 2 части: личную информацию участника (имя, уровень образования, специальность, возраст); вторая часть содержала 10 закрытых вопросов о разных аспектах читабельности текста. Респонденты должны были выразить свое согласие с каждым из утверждений по 5-ти балльной шкале, где 1 «совершенно не согласен», 5 – «совершенно согласен», а 3 – «затрудняюсь с ответом». Респонденты оценивали сложность текста по следующим критериям:

А. Лингвистические критерии:
- Текст легко читается, потому что он содержит короткие предложения;
- В тексте нет сложных словосочетаний и словосочетаний, затрудняющих его понимание;
- Все слова в тексте мне даются легко и понятно;
- В тексте нет непонятных слов, затрудняющих его понимание.

В. Структурно-логические критерии:
- Текст логически выстроен;
- Легко уловить основную мысль текста.

Г. Критерии заинтересованности и вовлеченности:
- Название привлекает внимание;
- Текст интересный;
- Текст мне ясен.

Д. Общая оценка: (по 5-ти балльной шкале респонденты оценивали сложность текста, где 1 получал текст, который «очень трудно понять», а 5 – текст, который «очень легко понять»).

## IV. Результаты и обсуждение

Мы не стремились к тому, чтобы наши респонденты выполняли функции экспертов. Их мнения имели такую же ценность/были настолько же ценны, как мнения участников социологических опросов, которые не могут дать «неправильных» ответов по определению. «Выбросы» в этом случае рассматриваются как выражение точки зрения (в данном случае восприятия текста сайта университета) определенных представителей целевой аудитории. Этот подход в своем

исследовании использовали, например Кольцова и ее коллеги [37, p.289]. В Таблице III представлены результаты обработки опроса и процент текстов в каждой категории сложности по каждому критерию (вопроса анкеты).

Таблица III. Результаты анкетирования

| Критерий | Более легкие тексты, % | Тексты средней сложности, % | Более сложные тексты, % |
|---|---|---|---|
| *Заинтересованность и вовлеченность* | | | |
| Название привлекает внимание | 11.73 | 48.78 | 39.49 |
| Текст интересный | 12.78 | 46.96 | 40.26 |
| Текст мне ясен | 29.75 | 35.09 | 35.16 |
| *Лингвистические* | | | |
| Текст легко читается, потому что он содержит короткие предложения | 11.75 | 47.17 | 41.08 |
| В тексте нет сложных словосочетаний и словосочетаний, затрудняющих его понимание | 21.03 | 42.51 | 36.46 |
| Все слова в тексте для меня легкие и понятные | 23.74 | 39.59 | 36.67 |
| В тексте нет непонятных слов, затрудняющих его понимание | 24.88 | 39.65 | 35.47 |
| *Структурно-логические* | | | |
| Текст логически выстроен | 30.68 | 32.13 | 37.19 |
| Легко уловить основную мысль текста | 37.85 | 30.63 | 31.52 |
| Текст легко читается | 21.36 | 40.17 | 38.47 |

Для понимания согласованности мнения между экспертами была рассчитана конвергенция, которая является показателем уровня согласия между экспертными оценками в отношении различных аспектов удобочитаемости текста. Более высокое значение сходимости указывает на более сильный консенсус среди экспертов, а более низкое значение означает более разнообразный набор мнений. Сходимость рассчитывали путем деления разницы между самым высоким и самым низким процентом оценки на сумму всех процентов оценки по каждому критерию.

Результаты показывают разные уровни согласия между респондентами по разным критериям удобочитаемости. Наиболее высокий уровень согласия показал критерий «Заголовок привлекает внимание» (сходимость = 0,49), а самый слабый — по критерий «Легко уловить основную идею» (сходимость = 0,31).

Корреляция между оценками респондентов и автоматической оценкой сложности текстов была различной для текстов, относящихся к разным категориям сложности. Для группы «более легких текстов» коэффициент корреляции составил -0,2, указывая на слабую отрицательную корреляцию. Для текстов «средней сложности» коэффициент корреляции равен 0,4, что свидетельствует о высоком уровне совпадения автоматической оценки с экспертной оценкой. Коэффициент корреляции в группе «более сложных текстов» составил -0.3, что указывает на умеренную отрицательную корреляцию.

Учитывая представительность выборки текстов, использованных в эксперименте, можно предположить, что подобная корреляция между автоматической оценкой и оценкой, сделанной респондентами, справедлива для всего корпуса *PolyLing*.

Наблюдаемое расхождение в оценках сложности можно объяснить следующим образом. Формулы читабельности разрабатывались для двух основных категорий читателей: изучающих английский как иностранный язык и англоязычных школьников, разного возраста и уровня образования. Эти формулы, возможно, не полностью применимы для оценки сложности новостных текстов с университетских веб-сайтов, ориентированных на молодого образованного читателя, читающих их на своем родном языке. Поэтому автоматическая оценка сложности таких текстов оказывается иногда завышенной по сравнению с реальной удобочитаемостью этих текстов. Также возможно, что разные лингвистические критерии имеют разное значение для разных читателей – образованных носителей языка.

Другое объяснение может заключаться в том, что для этой категории читателей качественные характеристики, такие как интерес к теме, обсуждаемой в тексте, являются более значимыми, чем объективные показатели сложности. Обе группы параметров, отвечающих за качественные характеристики текста и субъективное восприятие текста читателями, далеко не в полной мере учитывается в классических формулах определения сложности текста.

Для проверки гипотезы о важности критериев заинтересованности и вовлеченности при субъективной оценке сложности текста респондентами, мы проанализировали несколько текстов имеющих одинаковую сложность при автоматической оценке, но отнесенные к разным категориям сложности респондентами. Оказалось, что тексты с сайтов других университетов, предположительно, менее интересные для респондентов из СПбПУ, воспринимались ими как более сложные.

## V. Заключение

В результате первого этапа проекта по оценке читабельности мультимодальных текстов с сайтов 20 российских университетов путем парсинга был собран и размечен морфологически и синтаксически пилотный корпус, содержащий 1000+ новостных текстов.

Оценки респондентами трех групп критериев: лингвистических, структурно-логических и характеризующих заинтересованность и вовлеченность

позволили получить более полное представление восприятии текстов новостных рубрик сайтов российских университетов целевой аудиторией. Причины расхождения оценки сложности текстов, полученными автоматически и посредством оценки респондентами, требуют дальнейшего исследования с изменением ряда параметров: 1) на большем количестве текстов, разной тематической направленности, например, из разделов Research/Achievements, в котором как правило публикуются краткие сообщения об исследованиях, выполненных в разных подразделениях СПбПУ, получивших грантовую поддержку ведущих российских и международных фондов; 2) с более строгим контролем понимания прочитанного (например, по алгоритму, описанному в [36].

Будущие исследования также могут способствовать интеграции качественных факторов и невербальных компонентов в модели удобочитаемости и изучению их влияния на процесс оценки.

Одним из следствий, вытекающих из того факта, что респонденты находят тексты с сайта своего родного университета более легкими по сравнению с автоматической оценкой, может, на наш взгляд, стать рекомендация более активного обращения к ним на занятиях по русскому как иностранному, чего не делалось, возможно, из-за опасения высокой сложности текстов.

Логичным представляется продолжение исследования мультимодальных текстов сайтов российских университетов с целью проверки, содержат ли они тренды/характеристики, обнаруженные зарубежными исследователями университетских сайтов в разных странах.

Авторы понимают ограниченность этого пилотного исследования с точки зрения размера корпуса, размера выборки, тематического и жанрового разнообразия текстов и числа респондентов, участвовавших в эксперименте по оценке их удобочитаемости.

## Библиография


[1] Никулина Е.Р., Черкас А.В., Козина Е.Д., Бойко А.В., Дмитриева Л.А. Разработка сервиса для оценки удобочитаемости текста с применением технологий машинного обучения // Системный анализ в проектировании и управлении: сборник научных трудов / Труды XXVI Международной научно-практической конференции «Системный анализ в проектировании и управлении», Санкт-Петербург, 13-14 октября 2022 г. СПб: ПОЛИТЕХ-ПРЕСС, 2023. В 3-х частях. Часть 2. С.232–240. DOI:10.18720/SPBPU/2/id23-103.

[2] Tomášková R. A Walk through the Multimodal Landscape of University Websites // Brno Studies in English. 2015. Vol. 41. Iss. 1. P. 77–100.

[3] Middleton I., McConnell M., Davidson G. Presenting a model for the structure and content of a university World Wide Web site // Journal of Information Science. 1999. Vol. 25. Iss.3. P.219-227.

[4] Zollo S.A. Internationalization and Globalization. A Multimodal Analysis of Italian Universities' Websites // Journal of Multimodal Communication Studies. 2016. Vol.3. Iss. 1-2. P.1-17.

[5] Zhang Z.C., Tu W. Representation of international students at Australian university websites: A critical multimodal discourse analysis // Iberica. 2019. Vol. 37. P.221–243. https://revistaiberica.org/index.php/iberica/article/view/116 (дата обращения: 10.04.2023).

[6] Zhang Z., Tan S., Wignell P., O'Halloran K. Addressing international students on Australian and Chinese university webpages: A comparative study // Discourse, Context & Media. 2020. Vol.36. DOI: 10.1016/j.dcm.2020.100403.

[7] Zhang Z., Tan S., O'Halloran K.L. Managing higher education and neoliberal marketing discourses on Why Choose webpages for international students on Australian and British university websites // Discourse & Communicationю 2022. Vol.16. Iss.4. P. 462–481. DOI: 10.1177/17504813221074076.

[8] Chernyavskaya V.E., Zharkynbekova S.K. Linguistic and social construction of national university identity: Kazakh and Russian universities' mission statements //Vestnik of Saint Petersburg University. Language and Literature. 2019. Vol. 16. Iss. 2. P.304-319. DOI: 10.21638/spbu09.2019.210.

[9] Chernyavskaya V.E., Safronenkova E.L. Towards constructing identity of a National University: "Our past" at the websites of Russian universities // J. Sib. Fed. Univ. Humanit. Soc. Sci.. 2019. Vol. 12. Iss.10. P.1819–1839. DOI: 10.17516/1997-1370-0491.

[10] Michelson K., Alvarez Valencia J.A. Study Abroad: Tourism or Education? A Multimodal Social Semiotic Analysis of Institutional Discourses of a Promotional Website // Discourse & Communication. 2016. Vol.10, Iss. 3.P. 235–256. DOI: 10.1177/1750481315623893

[11] Hyland K. The presentation of self in scholarly life: Identity and marginalization in academic homepages // English for Specific Purposes. 2011. Vol. 30. Iss.4. P. 286–297. DOI: 10.1016/j.esp.2011.04.004.

[12] Bernardini S., Ferraresi A. The academic Web-as-Corpus // Proc. 8th Web as Corpus Workshop. Stroudsburg, PA: ACM. 2013. P. 53–62.

[13] Bernardini S., Ferraresi A. Institutional academic English and its phraseology: native and lingua franca perspectives // English for Academic Purposes: Approaches and Implications / G. Diani, P. Thompson (eds.). Cambridge Scholars Publishing, Newcastle, UK. 2015. ch.9, P. 225 – 244.

[14] Venuti M., Nasti C. Italian and UK university websites: comparing communicative strategies // ESP Across Cultures. 2015. Vol.12. P.127-137.

[15] Nasti C., Venuti M., Zollo S.A. UK university websites: A multimodal, corpus-based analysis // International Journal of Language Studies. 2017.Vol.11. Iss.4. P.131-152.

[16] Reynolds R.J. Russian natural language processing for computer-assisted language learning. PhD dissertation, UiT. Tromsø: The Arctic University of Norway. 2016. https://munin.uit.no/bitstream/handle/10037/9685/thesis.pdf?sequence=3&isAllowed=y (дата обращения: 10.04.2023).

[17] Солнышкина М.И., Кисельников А.С. Сложность текста: этапы изучения в отечественном прикладном языкознании // Вестник Томского государственного университета. Филология. 2015. Т. 38. №. 6. Р. 86–99. DOI: 10.17223/19986645/38/7.

[18] Ляшевская О.Н. К определению сложности русских текстов // XVII Апрельская международная научная конференция по проблемам развития экономики и общества: в 4 кн. / отв. ред. Е. Г. Ясин ; Нац. исслед. ун-т «Высшая школа экономики». М. : Изд. дом Высшей школы экономики, 2017. Кн. 4. С.408–418 https://conf.hse.ru/data/2017/04/06/1168267884/XVII%20%D0%90%D0%9C%D0%9D%D0%9A_%D0%9A%D0%BD.4-%D1%81%D0%B0%D0%B9%D1%82.pdf (дата обращения: 10.04.2023).

[19] Лапошина А.Н. Опыт экспериментального исследования сложности текстов по РКИ // Динамика языковых и культурных процессов в современной России: материалы VI Конгресса РОПРЯЛ (Уфа, 11–14 октября 2018 г.): сборник статей. 2018. Вып. 6. С. 1154–1179.

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

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

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

[23] Савельев Д.А. Исследование сложности предложений, составляющих тексты правовых актов органов власти Российской Федерации // Право. Журнал Высшей школы экономики. 2020. №.1. С. 50–74. DOI: 10.17323/2072-8166.2020.1.50.74.

[24] Akgül Y. Evaluating the performance of websites from a public value, usability, and readability perspectives: a review of Turkish national government websites // Univ Access Inf Soc. Published online Aug. 2022. DOI: 10.1007/s10209-022-00909-4.
[25] Akgül Y. The Accessibility, Usability, Quality and Readability of Turkish State and Local Government Websites: an Exploratory Study // International Journal of Electronic Government Research. 2019. Vol. 15. Iss. 1. P. 62–81. DOI:10.4018/IJEGR.2019010105.
[26] Karhu M., Hilera J.R., Fernández L., Ríos R. Accessibility and readability of university websites in Finland // J. of Accessibility and. Design for All. 2012. Vol. 2. Iss. 2. P. 178–189. DOI: 10.17411/jacces.v2i2.70.
[27] Patra M.R., Dash A.R., Mishra P.K. A quantitative analysis of WCAG 2.0 compliance for some Indian web portals // International Journal of Computer Science, Engineering and Applications (IJCSEA). 2017. Vol. 4. Iss.1. P. 9–23. DOI: 10.48550/arXiv.1710.08788.
[28] Akgül Y. Accessibility, usability, quality performance, and readability evaluation of university websites of Turkey: a comparative study of state and private universities // Univ Access Inf Soc. 2021. Vol. 20. Iss.1. P. 157–170. DOI: 10.1007/s10209-020-00715-w.
[29] Rashida M., Islam K., M. Kayes A.S., Hammoudeh M., Arefin M.S., Habib M.A. Towards developing a framework to analyze the qualities of the university websites // Computers. 2021.Vol.10. Iss. 57. P.1–16. DOI: 10.3390/computers10050057.
[30] Свидетельство № 2022669940. Программа для автоматического сбора текстов с новостных рубрик сайтов вузов: № 2022668973: заявл. 17.10.2022: опубл. 26.10.2022 / Ракова В.В., Черкас А.В., Козина Е.Д., Рубцова А.В., Коган М.С. 1с.
[31] Korobov M. Morphological analyzer and generator for Russian and Ukrainian languages // M. Khachay, N. Konstantinova, A. Panchenko, D. Ignatov, V. Labunets (eds). Analysis of Images, Social Networks and Texts. AIST 2015 / Communications in Computer and Information Science. Springer, Cham. 2015. Vol 542. P. 320–332. DOI: 10.1007/978-3-319-26123-2_31.
[32] Jain A. K. Data clustering: 50 years beyond K-means // Pattern Recognition Letters. 2010. Vol.31. Iss. 8. P. 651-666. DOI: 10.1016/j.patrec.2009.09.011.
[33] Rousseeuw P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis // Computational and Applied Mathematics. 1987. Vol. 20. P. 53-65.
[34] Gómez P.C., Sánchez-Lafuente Á.A. Readability indices for the assessment of textbooks: a feasibility study in the context of EFL // Vigo Intern. J. of Appl. Linguistics (VIAL). 2019. Vol. 16. P. 31–52.
[35] Ivanov V., Solnyshkina M., Solovyev V. Efficiency of text readability features in Russian academic texts // Computational Linguistics and Intellectual Technologies. Proc. Intern. Conf. “Dialogue 21” (May 30 – June 21, 2018). Moscow, Russia. 2018. P. 284–293. http://www.dialog-21.ru/media/4302/ivanovvv.pdf (дата обращения: 10.04.2023).
[36] Блинова О., Тарасов Н. Метрики сложности русских правовых текстов: отбор, использование, первичная оценка эффективности // Компьютерная лингвистика и интеллектуальные технологии. Ежегодная международная конференция «Диалог 2022». 15–18 июня 2022. Москва, Россия. .2022. С.1017-1028. DOI: 10.28995/2075-7182-2022-21-1017-1028.
[37] Koltsova O.Yu, Alexeeva S.V., Kolcov S.N. An Opinion word lexicon and a training data set for Russian sentiment analysis of social media computational linguistics and intellectual technologies // Proc. Intern. Conf. “Dialogue 2016” (June 1 – 4, 2016). Moscow, Russia. 2016. P. 277–287. https://www.dialog-21.ru/digest/2016/articles/ (дата обращения: 10.04.2023).

**Коган Марина Самуиловна**, канд. техн. наук, доцент, Санкт-Петербургский политехнический университет Петра Великого, ORCID 0000-0002-7519-2161 (e-mail: kogan_ms@spbstu.ru).
**Гаврилик Дарья Александровна**, ассистент, Санкт-Петербургский политехнический университет Петра Великого, ORCID 0009-0002-4410-1965 (e-mail: gavrilik_da@spbstu.ru).
**Чистякова Светлана Владимировна**, специалист, Санкт-Петербургский политехнический университет Петра Великого, ORCID 0009-0008-7465-7829.(e-mail: chistyakova_sv@spbstu.ru).
**Рубцова Анна Владимировна**, докт. пед. наук, профессор, Санкт-Петербургский политехнический университет Петра Великого, ORCID 0000-0002-0573-0980 (e-mail: rubtsova_av@spbstu.ru).
**Никулина Елизавета Романовна**, специалист, магистр лингвистики, Санкт-Петербургский политехнический университет Петра Великого, ORCID 0000-0002-6208-7637 (e-mail: nikulina_er@spbstu.ru).
**Черкас Алина Владимировна**, инженер-исследователь, аспирант Санкт-Петербургский политехнический университет Петра Великого, ORCID 0000-0001-9062-966X (e-mail: alina.cherkas@spbpu.com).
**Болсуновская Марина Владимировна**, канд. техн. наук, доцент, Санкт-Петербургский политехнический университет Петра Великого, ORCID 0000-0001-6650-6491 (e-mail: bolsun_mv@spbstu.ru).

# On readability evaluation of multimodal texts from Russian Universities websites (based on the PolyLing corpus)

M.S. Kogan, D.A. Gavrilik, S.V. Chistyakova, A.V. Rubtsova, E.R. Nikulina, A.V. Cherkas, M.V. Bolsunovskaya

***Abstract*— University websites being complex multimodal structures are very important in modern educational environment: being the central element of the University's electronic information and educational environment, they play part of intermediary between University and the outer world and a powerful University image-maker's tool. Multimodal university websites draw researchers' attention since their appearance in the 1990s. However, the analysis of studies on University websites has revealed that very few of them focused on the websites' readability. This paper focusing on evaluating readability of leading Russian university websites aims to partially bridge this gap. To achieve the goal, at the first stage of the project a corpus of 1000+ texts was built by parsing news sections of pre-selected websites. The automatic cluster analysis was used to determine the main themes of the built corpus named PolyLing. The texts' readability was evaluated both automatically with Python script based on classical readability indices and by 132 human assessors who evaluated a representative sample of the corpus according to 10 criteria belonging to three groups (linguistic, structural and logical, and appealing) through completing a specially- designed questionnaire. The correlation analysis showed a satisfactory agreement for texts of middle difficulty and a negative correlation for easier and more difficult texts between the automatic and respondents' estimates. The paper proposes possible explanation for this divergence and outlines further research.**

***Keywords*— Automatic readability evaluation, corpus PolyLing, multimodal Russian University websites, readability indices, respondents' readability evaluation.**

## REFERENCES

[1] Nikulina E.R., Cherkas A.V., Kozina E.D., Boiko A.V., Dmitrieva L.A. Development of a service for text readability assessment via machine learning technologies // Sistemnyj analiz v proektirovanii i upravlenii: sbornik nauchnyh trudov / Trudy XXVI Mezhdunarodnoj nauchno-prakticheskoj konferencii «Sistemnyj analiz v proektirovanii i upravlenii», Sankt-Peterburg, October 13-14, 2022. SPb: POLITEH-PRESS, 2023. In 3 volumes. Vol. 2. P.232–240. DOI:10.18720/SPBPU/2/id23-103.

[2] Tomášková R. A Walk through the Multimodal Landscape of University Websites // Brno Studies in English. 2015. Vol. 41. Iss. 1. P. 77–100.

[3] Middleton I., McConnell M., Davidson G. Presenting a model for the structure and content of a university World Wide Web site //Journal of Information Science. 1999. Vol. 25. Iss. 3. P. 219-227.

[4] Zollo S. A. Internationalization and Globalization. A Multimodal Analysis of Italian Universities' Websites // Journal of Multimodal Communication Studies. 2016. Vol.3. Iss. 1-2. P.1-17.

[5] Zhang Z.C., Tu W. Representation of international students at Australian university websites: A critical multimodal discourse analysis // Iberica. 2019. Vol. 37. P. 221–243. https://revistaiberica.org/index.php/iberica/article/view/116 (accessed date: 10.04.2023)

[6] Zhang S., Tan S., Wignell P., O'Halloran K. Addressing international students on Australian and Chinese university webpages: A comparative study //Discourse, Context & Media. 2020. Vol.36. DOI: 10.1016/j.dcm.2020.100403.

[7] Zhang Z., Tan S., O'Halloran K.L. Managing higher education and neoliberal marketing discourses on Why Choose webpages for international students on Australian and British university websites // Discourse & Communicationю 2022. Vol. 16. Iss. 4. P. 462–481. DOI: 10.1177/17504813221074076.

[8] Chernyavskaya V.E., Zharkynbekova S.K. Linguistic and social construction of national university identity: Kazakh and Russian universities' mission statements // Vestnik of Saint Petersburg University. Language and Literature. 2019. Vol. 16. Iss. 2. P. 304-319. DOI: 10.21638/spbu09.2019.210.

[9] Chernyavskaya V.E., Safronenkova E.L. Towards constructing identity of a National University: "Our past" at the websites of Russian universities // J. Sib. Fed. Univ. Humanit. Soc. Sci. 2019. Vol. 12. Iss. 10. P. 1819–1839. DOI: 10.17516/1997-1370-0491.

[10] Michelson K., Alvarez Valencia J.A. Study Abroad: Tourism or Education? A Multimodal Social Semiotic Analysis of Institutional Discourses of a Promotional Website // Discourse & Communication. 2016. Vol. 10, Iss. 3. P. 235–256. DOI: 10.1177/1750481315623893.

[11] Hyland K. The presentation of self in scholarly life: Identity and marginalization in academic homepages // English for Specific Purposes. 2011. Vol. 30. Iss. 4. P. 286–297. DOI: 10.1016/j.esp.2011.04.004.

[12] Bernardini S., Ferraresi A. The academic Web-as-Corpus // Proc. 8th Web as Corpus Workshop. Stroudsburg, PA: ACM. 2013. P. 53–62.

[13] Bernardini S., Ferraresi A. Institutional academic English and its phraseology: native and lingua franca perspectives // English for Academic Purposes: Approaches and Implications / G. Diani, P. Thompson (eds.). Cambridge Scholars Publishing, Newcastle, UK. 2015. ch.9, P. 225 – 244.

[14] Venuti M., Nasti C. Italian and UK university websites: comparing communicative strategies // ESP Across Cultures. 2015. Vol. 12. P. 127-137.

[15] Nasti C., Venuti M., Zollo S.A. UK university websites: A multimodal, corpus-based analysis // International Journal of Language Studies. 2017. Vol. 11. Iss. 4. P. 131–152.

[16] Reynolds R.J. Russian natural language processing for computer-assisted language learning. PhD dissertation, UiT. Tromsø: The Arctic University of Norway. 2016. https://munin.uit.no/bitstream/handle/10037/9685/thesis.pdf?sequence=3&isAllowed=y (accessed date: 10.04.2023).

[17] Solnyshkina M.I.,. Kiselnikov A.S. Text complexity: study phases in Russian linguistics // Tomsk State University Journal of Philology. 2015. Vol. 38. No. 6. P. 86–99. DOI: 10.17223/19986645/38/7.

[18] Lyashevskaya O. K opredeleniju slozhnosti russkih tekstov // XVII Aprel'skaja mezhdunarodnaja nauchnaja konferencija po problemam razvitija jekonomiki i obshhestva: v 4 kn. / E. G. Jasin (ed.) ; National research university "Higher School of Economics". Moscow : Publishing house of HSE, 2017. Kniga 4. P. 408–418. https://conf.hse.ru/data/2017/04/06/1168267884/XVII%20%D0%90%D0%9C%D0%9D%D0%9A_%D0%9A%D0%BD.4-%D1%81%D0%B0%D0%B9%D1%82.pdf (accessed date: 10.04.2023).

[19] Laposhina A.N. Insights from an experimental study on the text complexity for Russian as a foreign language // Dinamika jazykovyh i kul'turnyh processov v sovremennoj Rossii: Proceedings of VI Congress ROPRYAL (Ufa, Oct. 11-14, 2018). 2018. Iss. 6. P. 1154-1179.

[20] Solovyev V., Solnyshkina M., Ivanov V. Prediction of reading difficulty in Russian academic texts // Journal of Intelligent & Fuzzy Systems. 2019. Vol. 36. Iss. 5. P. 4553-4563. DOI: 10.3233/JIFS-179007.
[21] Solovyev V., Ivanov V., Solnyshkina M. Assessment of reading difficulty levels in Russian academic texts: Approaches and metrics // Journal of Intelligent & Fuzzy Systems. 2018. Vol. 34. Iss. 5. P. 3049–3058. DOI: 10.3233/JIFS-169489.
[22] Blinova O.V., Tarasov N.A. Complexity of Russian legal texts: assessment methods and language data // Proc. International Conference “Corpora 2021” (July 1-3, 2021). Skifija-print, Saint Petersburg. 2021. P. 175–182.
[23] Saveliev D.A. A study in complexity of sentences constituting Russian Federation legal acts // Law. Journal of HSE. 2020. No. 1. P. 50–74. DOI: 10.17323/2072-8166.2020.1.50.74.
[24] Akgül Y. Evaluating the performance of websites from a public value, usability, and readability perspectives: a review of Turkish national government websites // Univ Access Inf Soc. Published online Aug. 2022. DOI: 10.1007/s10209-022-00909-4.
[25] Akgül Y. The Accessibility, Usability, Quality and Readability of Turkish State and Local Government Websites: an Exploratory Study // International Journal of Electronic Government Research. 2019. Vol. 15. Iss. 1. P. 62–81. DOI:10.4018/IJEGR.2019010105.
[26] Karhu M., Hilera J.R., Fernández L., Ríos R. Accessibility and readability of university websites in Finland // J. of Accessibility and. Design for All. 2012. Vol. 2. Iss. 2. P. 178–189. DOI: 10.17411/jacces.v2i2.70.
[27] Patra M.R., Dash A.R., Mishra P.K. A quantitative analysis of WCAG 2.0 compliance for some Indian web portals // International Journal of Computer Science, Engineering and Applications (IJCSEA). 2017. Vol. 4. Iss. 1. P. 9–23. DOI: 10.48550/arXiv.1710.08788.
[28] Akgül Y. Accessibility, usability, quality performance, and readability evaluation of university websites of Turkey: a comparative study of state and private universities // Univ Access Inf Soc. 2021. Vol.20. Iss. 1. P. 157–170. DOI: 10.1007/s10209-020-00715-w.
[29] Rashida M., Islam K., M. Kayes A.S., Hammoudeh M., Arefin M.S., Habib M.A. Towards developing a framework to analyze the qualities of the university websites // Computers. 2021.Vol. 10. Iss. 57. P. 1–16. DOI: 10.3390/computers 10050057.
[30] Svidetel'stvo № 2022669940. Programma dlya avtomaticheskogo sbora tekstov s novostnyh rubrik sajtov vuzov: № 2022668973: zayavl. 17.10.2022: opubl. 26.10.2022 / Rakova V.V., Cherkas A.V., Kozina E.D., Rubtsova A.V., Kogan M.S. 1p.
[31] Korobov M. Morphological analyzer and generator for Russian and Ukrainian languages // M. Khachay, N. Konstantinova, A. Panchenko, D. Ignatov, V. Labunets (eds). Analysis of Images, Social Networks and Texts. AIST 2015 / Communications in Computer and Information Science. Springer, Cham. 2015. Vol. 542. P. 320–332. DOI: 10.1007/978-3-319-26123-2_31.
[32] Jain A. K. Data clustering: 50 years beyond K-means // Pattern Recognition Letters. 2010. Vol. 31. Iss. 8. P. 651–666. DOI: 10.1016/j.patrec.2009.09.011.
[33] Rousseeuw P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis // Computational and Applied Mathematics. 1987. Vol. 20. P. 53-65.
[34] Gómez P.C., Sánchez-Lafuente Á.A. Readability indices for the assessment of textbooks: a feasibility study in the context of EFL // Vigo Intern. J. of Appl. Linguistics (VIAL). 2019. Vol. 16. P. 31–52.
[35] Ivanov V., Solnyshkina M., Solovyev V. Efficiency of text readability features in Russian academic texts // Computational Linguistics and Intellectual Technologies. Proc. Intern. Conf. “Dialogue 21” (May 30 – June 21, 2018). Moscow, Russia. 2018. P. 284–293. http://www.dialog-21.ru/media/4302/ivanovvv.pdf (accessed date: 10.04.2023)
[36] Blinova O., Tarasov N. Complexity metrics of Russian legal texts: selection, use, initial efficiency evaluation // Computational Linguistics and Intellectual Technologies: Proc. International Conference “Dialogue 2022” (June 15 – 18, 2022). Moscow. 2022. P. 1017-1028. DOI: 10.28995/2075-7182-2022-21-1017-1028.
[37] Koltsova O.Yu, Alexeeva S.V., Kolcov S.N. An Opinion word lexicon and a training data set for Russian sentiment analysis of social media computational linguistics and intellectual technologies // Proc. Intern. Conf. “Dialogue 2016” (June 1 – 4, 2016). Moscow, Russia. 2016. P. 277–287. URL: https://www.dialog-21.ru/digest/2016/articles/ (accessed date: 10.04.2023).

**Kogan Marina Samuilovna**, cand. tech. sc., Associate Professor, Peter the Great Saint Petersburg Polytechnic University ORCID 0000-0002-7519-2161 (e-mail: kogan_ms@spbstu.ru).
**Gavrilik Daria Alexandrovna**, ассистент, Peter the Great Saint Petersburg Polytechnic University, ORCID 0009-0002-4410-1965 (e-mail: gavrilik_da@spbstu.ru).
**Chistyakova Svetlana Vladimirovna**, specialist, Peter the Great Saint Petersburg Polytechnic University, ORCID 0009-0008-7465-7829.(e-mail: chistyakova_sv@spbstu.ru).
**Rubtsova Anna Vladimirovna**, doctor pedagogic. sc., Peter the Great Saint Petersburg Polytechnic University, ORCID 0000-0002-0573-0980 (e-mail: rubtsova_av@spbstu.ru).
**Nikulina Elizaveta Romanovna**, specialist, Master of Linguistics, Peter the Great Saint Petersburg Polytechnic University, ORCID 0000-0002-6208-7637 (e-mail: nikulina_er@spbstu.ru).
**Cherkas Alina Vladimirovna**, research engineer, post graduate student, Peter the Great Saint Petersburg Polytechnic University, ORCID 0000-0001-9062-966X (e-mail: alina.cherkas@spbpu.com).
**Bolsunovskaya Marina Vladimirovna,** cand. tech. sc., Associate Professor, Peter the Great Saint Petersburg Polytechnic University, ORCID 0000-0001-6650-6491 (e-mail: bolsun_mv@spbstu.ru).