www.volsu.ru

DOI: https://doi.org/10.15688/jvolsu2.2023.6.6

UDC 81’322.2
LBC 81.112

Submitted: 13.05.2023
Accepted: 09.10.2023

PARAMETRIC TAXONOMY OF EDUCATIONAL TEXTS [1]

Roman V. Kupriyanov
Kazan Federal University, Kazan, Russia;
Kazan National Research Technological University, Kazan, Russia

Marina I. Solnyshkina
Kazan Federal University, Kazan, Russia

Polina A. Lekhnitskaya
Kazan Federal University, Kazan, Russia

Abstract. The article is aimed at considering the issue of the discursive text typology and developing a parametric model of the elementary school texts for the ontological domain by employing a corpus-based approach and methods of linguistic statistics. The research corpus of over 90,000 tokens comprises texts of 13 textbooks acknowledged in the 2nd grade of Russian schools. The applied multifactor discriminant analysis enabled identification and validation of typological characteristics of the texts under study, offering the formula for referring educational texts to a subject domain on Philology, Mathematics, and Natural Sciences. The discriminant analysis results confirmed the hypothesis that each type of text corresponds to a parametric model, which includes six constants: the average number of words in a sentence, the average number of nouns, the average number of verbs and the average number of adjectives per sentence, local noun overlap, global argument overlap. The assessment of linguistic parameters was performed by an automatic Russian text analyzer RuLingva. The classification accuracy of the parametric model was identified as 80%, which ensures its high reliability and allows for the data obtained to be employed in linguistic expertise, as well as for in automated linguistic profiling of texts. The prospect of the research implies installation of the model in RuLingva and development of similar models for texts of other subject domains.

Key words: discourse, subject domain, lexical parameters, syntactic parameters, mathematical model, discriminant analysis.

Citation. Kupriyanov R.V., Solnyshkina M.I., Lekhnitskaya P.A. Parametric Taxonomy of Educational Texts. *Vestnik Volgogradskogo gosudarstvennogo universiteta. Seriya 2. Yazykoznanie* [Science Journal of Volgograd State University. Linguistics], 2023, vol. 22, no. 6, pp. 80-94. (in Russian). DOI: https://doi.org/10.15688/jvolsu2.2023.6.6

УДК 81’322.2
ББК 81.112

Дата поступления статьи: 13.05.2023
Дата принятия статьи: 09.10.2023

ПАРАМЕТРИЧЕСКАЯ ТАКСОНОМИЯ УЧЕБНЫХ ТЕКСТОВ [1]

Роман Владимирович Куприянов
Казанский (Приволжский) федеральный университет, г. Казань, Россия;
Казанский национальный исследовательский технологический университет, г. Казань, Россия

Марина Ивановна Солнышкина
Казанский (Приволжский) федеральный университет, г. Казань, Россия

Полина Александровна Лехницкая
Казанский (Приволжский) федеральный университет, г. Казань, Россия

Аннотация. Представленное исследование нацелено на решение проблемы типологизации текста как единицы дискурса и выполнено в рамках корпусного подхода с применением методов лингвистической статистики. Исследовательский корпус, общий объем которого превышает 90 тыс. словоформ, включает тексты на русском языке из 13 учебников для 2-го класса российских школ. В результате многофакторного дискриминантного анализа выявлены типологические характеристики текстов учебных дискурсов трех предметных областей – филологии, математики, естествознания. Рассчитаны формулы для классификации текстов по предметным областям. На основе этих формул разработана и валидирована параметрическая модель. В нее входят шесть параметров: среднее количество слов в предложении, среднее количество существительных, среднее количество глаголов и среднее количество прилагательных на предложение, локальный повтор существительного, глобальный повтор аргумента. Расчеты значений лингвистических параметров произведены при помощи автоматического анализатора текстов на русском языке RuLingva. Высокая степень классификационной точности параметрической модели – 80 % – обеспечивает ее достаточную надежность и позволяет применять полученные данные в лингвистической экспертизе, а также для автоматизации лингвистического профилирования текстов. Перспектива исследования связана с инсталляцией модели в RuLingva и разработкой аналогичных моделей для текстов учебного дискурса других предметных областей. *Вклад авторов:* Р.В. Куприянов – проведение дискриминантного анализа, описание результатов статистической обработки данных, анализ результатов и формулировка выводов; М.И. Солнышкина – разработка концептуального подхода исследования, анализ результатов и формулировка выводов; П.А. Лехницкая – подготовка материала исследования, обработка корпуса текстов в автоматическом анализаторе текстов, описание первичных результатов.

Ключевые слова: дискурс, предметная область, лексические параметры, синтаксические параметры, математическая модель, дискриминантный анализ.

Цитирование. Куприянов Р. В., Солнышкина М. И., Лехницкая П. А. Параметрическая таксономия учебных текстов // Вестник Волгоградского государственного университета. Серия 2, Языкознание. – 2023. – Т. 22, № 6. – С. 80–94. – DOI: https://doi.org/10.15688/jvolsu2.2023.6.6

Введение

Современная прикладная лингвистика решает актуальные для общества задачи: разрабатывает переводческие алгоритмы, осуществляет автоматизированную обработку естественного языка, проводит так называемый «интеллектуальный» анализ текста (англ. text mining) и пр. Область применения ее достижений весьма обширна – от подбора текста с требуемым контентом до рекомендаций по модификации текста для определенной категории потенциальных реципиентов [Solovyev, Solnyshkina, McNamara, 2022]. В настоящее время созданы и успешно функционируют автоматические анализаторы или профайлеры текстов, рассчитывающие до 200 лингвистических параметров. Каждый из такого рода профайлеров, в том числе Coh-Metrix-Port 3.0 (fw.nilc.icmc.usp.br:23380/cohmetrixport) или RuLingva (rulingva.kpfu.ru), используется не только для расчета лингвистических параметров текстов, но и для их сравнения и изменения. Спектр лингвопрагматических целей исследователей, аналитиков и разработчиков учебных и контрольно-измерительных материалов может включать, например, упрощение текста для определенных категорий читателей, являющееся в настоящее время весьма востребованным [Ermakova et al., 2023], выявление происхождения лексики, иллюстрации использования слова в определенном регистре, подтверждение принадлежности слова к лексическому минимуму Общеевропейской шкалы (англ. CEFR) (www.coe.int/ru/web/lang-migrants/cefr-and-profiles) и др. Именно такого рода функционал имеют Lextutor (www.lextutor.ca/vp/eng), TextInspector (textinspector.com), Текстометр (textometr.ru), ReaderBench (readerbench.com), T.E.R.A (soletlab.adaptiveliteracy.com:8443). Однако ни один из существующих анализаторов не является дискурсивным профайлером, то есть не позволяет автоматически определить регистр, дискурс и тип текста по его лингвистическим параметрам. Особо актуально сегодня и определение референтного диапазона значений параметров для классификации различных типов текстов. Востребованность такого рода профайлеров велика при подборе тестов для определенных целей (учебных, мониторинговых, информативных, суггестивных и др.), а также при установлении авторства,

отборе и подготовке текстовых материалов для различных категорий пользователей.

Предлагаемое читателям исследование является частью большого проекта, конечная цель которого – составление списка типологических лингвистических параметров, определяющих предметную область учебного текста, уровень его когнитивной и лингвистической сложности. В данной статье рассматривается возможность применения дискриминантного анализа для создания математической модели, которая дифференцирует используемые в начальной школе учебные тексты трех предметных дискурсов (математического, филологического, естественно-научного). Такая модель позволит осуществлять «профилирование» текста.

Гипотеза исследования заключается в том, что академические тексты заданной сложности (например, в диапазоне одного учебного года), предназначенные для применения в различных предметных областях (математика, филология, естествознание), имеют квантитативные лексические и синтаксические различия. Эти различия носят типологический характер и позволяют идентифицировать тип, дискурс, уровень сложности и даже автора (источник) текста.

Материал и методы

Специфика академического (или учебно-научного) текста состоит в его особой коммуникативной функции и прагматике, а именно в его направленности на целевую аудиторию. Определяя учебно-научный текст как сообщение в письменной форме, характеризующееся смысловой и структурной завершенностью, связанностью и направленное на передачу и усвоение знаний, то есть на процесс обучения, Ж.И. Жеребцова особо подчеркивает его нацеленность на передачу информации [Жеребцова, 2007, с. 29]. Очевидно, что информативность текста для читателей во многом обусловлена их готовностью к восприятию и пониманию содержания текста, то есть их когнитивными способностями. Такого рода готовность трактуется в дискурсивной комплексологии как относительная сложность или трудность [Östen, 2004; Bulté, Housen, 2012; Pallotti, 2015]. При этом очевидно, что «объективная сложность» всегда манифестируется в лингвистических параметрах текста: морфологических, лексических, синтаксических и дискурсивных [Solnyshkina, Harkova, Kazachkova, 2020]. Именно они вместе с содержанием или ситуационной моделью референта текста [Solovyev, Solnyshkina, McNamara, 2022] детерминируют трудность его восприятия для различных категорий языковых личностей читателей [Солнышкина, Казачкова, Харькова, 2020].

Значимыми при восприятии устного и письменного (печатного или электронного) текста являются так называемые количественные параметры, к которым, в частности, относят длину текста, среднее количество слогов или символов в слове (длина слова) и слов в предложении (длина предложения). Именно данные параметры определяются в комплексологии как обобщенные статистические параметры, влияние которых на восприятие обусловлено относительно небольшим объемом оперативной памяти человека [Оборнева, 2006, с. 5].

Длина предложения как предиктор сложности представляет особый интерес, поскольку именно она может затруднять восприятие и понимание текста [McNamara et al., 2014, p. 2]. Аналогично длине предложения оценивается и длина слова: чем длиннее слово, тем больше времени требуется для его восприятия, понимания и удержания в кратковременной памяти [Вахрушева и др., 2021, с. 93]. Короткие слова проще читать, следовательно, они легче воспринимаются, поскольку морфологическая сложность слова создает дополнительные смыслы, влияя на его информативность [Gatiyatullina et al., 2020].

Большую роль в восприятии текста играют и другие морфологические параметры – доли различных частей речи в тексте. В корпусной лингвистике разработаны методики определения жанра на основе относительных частот отдельных частей речи [Seifart et al., 2012, p. 10]. На материале английского языка валидированы статистически значимые различия регистров и типов дискурсов [Biber, 2006]. Например, доказано, что повторяющиеся глаголы создают более связную структуру событий, которая облегчает и улучшает понимание ситуационной модели. Особенно актуален данный параметр при лингвистическом анализе повествовательных текстов [McNamara, Graesser,

Louwerse, 2012]. Аналогичные закономерности выявлены и для текстов на русском языке [Журавлев, 1988; Разговорная речь..., 2009]. Предикторами сложности, имеющими высокую степень достоверности, принято также считать среднее количество глаголов, среднее количество прилагательных, среднее количество существительных на предложение [Solnyshkina, McNamara, Zamaletdinov, 2022]. Доказанным для русского языка признано и увеличение доли имен существительных в родительном падеже по мере роста сложности текста. Например, в текстах учебников биологии в диапазоне от 5-го до 11-го класса их доля растет от 34 до 41 % , а по обществознанию – от 23 % до 38 % [Gatiyatullina et al., 2020].

К лингвистическим параметрам сложности текста относятся и так называемые относительные предикторы, то есть меры, основанные на отношении одних групп единиц к другим. К таким параметрам можно отнести, например, номинативность (отношение глаголов к существительным) и описательность (отношение прилагательных к существительным) [Мартынова и др., 2020].

Следующая группа параметров – лексические. Прежде всего это повторы отдельных лексем. Традиционно рассматриваются локальные повторы, то есть повторы внутри одного предложения или в смежных предложениях (англ. *local overlap*, букв. «локальный повтор»), а также глобальные повторы, то есть повторы внутри всего текста (англ. *global overlap*, букв. «глобальный повтор»). Например, параметр «глобальный повтор существительных» (англ. *global noun overlap*) демонстрирует количество повторов всех имен существительных в изучаемом тексте [McNamara et al., 2014, p. 2 [2]]. Аналогичными параметрами являются локальный повтор аргумента» англ. *local argument overlap*) и глобальный повтор аргумента (англ. *global argument overlap*), в которых термин «аргумент» означает существительное и/или местоимение, противопоставляемое предикату – глаголу и/или прилагательному [McNamara et al., 2014, p. 2]. Данный параметр отражает степень повторяемости аргумента в предложениях изучаемого текста [Crossley et al., 2013, p. 277].

Особое внимание ученые уделяют лексическому разнообразию (англ. *TTR*, *Type Token Ratio*, букв. «отношение слов к словоформам») [Graesser et al., 2004, p. 193]. При TTR = 1,0 ни одно из слов в тексте не повторяется. Очевидно, что такого рода тексты могут создаваться только искусственно, поскольку авторы в большинстве своем стремятся быть понятыми, а отсутствие лексических повторов во многом затрудняет восприятие текста. Низкие значения лексического разнообразия (TTR < 0,5) сигнализируют о высокой повторяемости слов, которая положительно влияет на скорость обработки текста в оперативной памяти человека. Целевая аудитория текстов такого рода – пользователи с ограниченным словарным запасом (изучающие язык как иностранный или младшие школьники) [Malvern et al., 2004].

Важной особенностью, которую необходимо учитывать при расчетах лексического разнообразия, является ограничение длины текста 1000 словоформами. Данное обстоятельство связано в первую очередь с тем, что при увеличении объема анализируемого текста растет количество служебных слов, а количество знаменательных слов сокращается. Именно поэтому расчеты лексического разнообразия в текстах более 1000 словоформ признаются недостоверными. Длинные тексты рекомендуется разбивать на отрывки по 1000 словоформ, в каждом из которых отдельно измеряется лексическое разнообразие [Вахрушева и др., 2021].

Валидированным предиктором сложности академических текстов является индекс удобочитаемости Флеша – Кинкейда (далее – ФК), первоначально рассчитанный для текстов на английском языке [Flesch, 1948] и только в начале нашего столетия адаптированный для русского языка [Солнышкина, Кисельников, 2015]. Востребованности данного индекса способствовали два фактора: легкость расчетов (и последующая успешная автоматизация для ряда языков) и корреляция с академическим возрастом читателя, то есть количеством лет формального обучения. В настоящее время данная формула успешно применяется для самых разных целей – от расчета соответствия словарного состава книги и словарного запаса читателя до прогнозирования успешности жизненного цикла сайта.

Расчеты читабельности текста на основе двух базовых метрик – средней длины пред-

ложения и средней длины слова – применяются при оценке соответствия текстов для военных, пациентов медицинских учреждений, клиентов страховых компаний и автосалонов [Corlatescu, Ruseti, Dascalu, 2022]. Для определения читабельности текстов на русском языке, принадлежащих разным дискурсам, используются две наиболее известные формулы.

1. Формула читабельности ФК (SIS) разработана и валидирована на корпусе учебных текстов и на основе психолингвистических оценок восприятия школьников [Solovyev, Ivanov, Solnyshkina, 2018]:

$$ФК(SIS) = 208{,}7 - 2{,}6 \times СДП - 39 \times СДС,$$

где СДП – это средняя длина предложения в словах; СДС – средняя длина слова в слогах.

2. Формула читабельности И.В. Оборневой [Оборнева, 2006] разработана на художественных текстах и при оценке читабельности текстов других типов дает завышенные результаты [Solnyshkina, McNamara, Zamaletdinov, 2022]:

$$ФК(О) = 206{,}835 - (1{,}3 \times СДП) - (60{,}1 \times СДС).$$

Признанным многими исследователям в качестве предиктора сложности является также индекс абстрактности [Solovyev et al., 2019; Solovyev, Ivanov, Akhtiamov, 2019], поскольку абстрактные лексические единицы всегда усложняют восприятие текста. Особенно значим данный параметр при оценивании сложности текстов для младших школьников, так как детское мышление не готово работать с абстрактными лексическими единицами, детям легче воспринимать конкретные слова [Вахрушева и др., 2021, с. 94].

Комплекс перечисленных параметров позволяет не только осуществить многофакторный анализ лингвистической сложности текста, но и составить «профиль» текста при помощи ограниченного ряда параметров, то есть отнести его к определенному типу, дискурсу и уровню сложности.

Отправной точкой представленного исследования явилось признание квантитативной типологичности – «однородности» академических текстов, предназначенных для одного года обучения в заданной предметной области. В основу типологизации как метода положена концепция «нечетких множеств» элементов типологии, при которой переход одного объекта (в нашем исследовании – текста) от принадлежности к непринадлежности заданному множеству осуществляется постепенно. При этом элементы одного множества обладают рядом типичных параметров, свойственных данному множеству, и некоторыми специфичными, индивидуальными параметрами, которые свойственны им в меньшей степени. Переход в другое множество предполагает аккумуляцию типологических параметров другого множества. Например, если сравнивать множества «Тексты по биологии для 9-го класса» и «Тексты по биологии для 10-го класса», то, очевидно, что этот переход происходит постепенно и связан с усложнением текста. Последнее должно отражаться в метриках морфологических, лексических и синтаксических параметров сопоставляемых текстов. При контрастировании текстов одной сложности, но разных предметных областей, например «Тексты по физике для 7-го класса» и «Тексты по истории для 7-го класса», естественно предположить, что они также будут отличаться рядом параметров. Причем список данных параметров может отличаться от списка параметров при сравнении текстов одного предметного блока, но разной сложности.

Исследование проводилось в три этапа и включало:

1) подготовку, очистку и предобработку корпуса исследования;

2) расчет значений (метрик) лингвистических параметров при помощи автоматического анализатора RuLingva (rulingva.kpfu.ru);

3) разработку методики профилирования (типологизации) текста на основе дискриминантного анализа значений лингвистических параметров.

Результаты

1. Подготовка, очистка и предобработка корпуса исследования.

Корпус исследования общим объемом 91 185 словоформ составили тексты 13 учебников трех предметных блоков («Русский язык», «Математика», «Окружающий мир») из Федерального перечня учебников Российской Федерации (fpu.edu.ru) [3]. Составленный на основе эк-

спертного мнения практикующих учителей начальных классов корпус был сбалансирован по объему каждого из трех составляющих подкорпусов: филологического, математического и естественно-научного. Филологический подкорпус включает четыре учебника русского языка общим объемом 34 286 словоформ, в Математический подкорпус вошло пять учебников математики, объем которых составил 28 728 словоформ, четыре учебника по предмету «Окружающий мир» составили Естественно-научный подкорпус объемом 28 171 словоформа (см. список источников). Все учебники, включенные в корпус исследования, использовались в школах РФ в 2018–2023 гг. и признаны экспертами как соответствующие когнитивным и лингвистическим способностям школьников начальных классов.

Для обеспечения единства языкового материала на этапе предобработки из текстов учебников были удалены их метаописание, предисловие, слово автора, содержание, иллюстрации, комментарии к ним, шаблонные фразы («Рисунок 1» и др.), примечания, вопросы для самоконтроля, лабораторные задания, названия параграфов, подзаголовки, тексты колонтитулов. Полученные тексты были разбиты на части длиной около 1000 слов, границы текста для анализа определялись окончанием предложения. Полный корпус исследования, содержащий 87 текстов, из которых 20 – тексты Математического подкорпуса, 30 – Филологического подкорпуса и 37 – Естественно-научного подкорпуса, был разделен на две коллекции: 77 текстов были использованы для разработки, а 10 – для тестирования параметрической модели. Тексты для тестирования были отобраны случайным образом: 3 – математических текста, 3 – филологических и 4 – естественно-научных. Эти тексты использовались не для расчетов коэффициентов в дискриминантном анализе, а для проверки математической модели.

2. Расчет значений параметров при помощи автоматического анализатора RuLingva и анализ статистически значимых параметров.

Расчеты метрик лингвистических параметров изучаемых текстов были выполнены при помощи программы автоматического анализа RuLingva (rulingva.kpfu.ru). Из 45 параметров, рассчитываемых RuLingva, в сокращенный список после первоначального отсева вошли 14 параметров: среднее количество слов в предложении, среднее количество слогов в слове, среднее количество существительных, глаголов, прилагательных на предложение, индекс ФК (SIS) или читабельность, индекс абстрактности, локальный повтор существительного, глобальный повтор существительного, локальный повтор аргумента, глобальный повтор аргумента, лексическое разнообразие (англ. TTR), номинативность, описательность. Все остальные параметры (доли существительных в разных падежах, доли временных форм глагола и прочие параметры, рассчитываемые RuLingva) были исключены из анализа на основе близости значений параметров текстов всех трех предметных подкорпусов.

Статистический анализ 14 параметров 77 текстов был проведен в программе Statistica. Для выявления типологических параметров учебного текста и расчета коэффициентов формулы применялся дискриминантный анализ. Лингвистические параметры текстов представлены в таблице 1.

3. Разработка прогностической модели, то есть методики профилирования (типологизации) текста на основе дискриминантного анализа значений лингвистических параметров.

Для создания прогностической модели был применен дискриминантный анализ, являющийся одним из наиболее апробированных многомерных методов при исследовании стиля [Андреев, 2010; Kupriyanov et al., 2022]. Данный метод используется в лингвистике также для атрибуции (выявления авторства) текста [Baayen, van Halteren, Tweedie, 1996; Holmes, Forsyth, 1995; Stamatatos, Fakotakis, Kokkinakis, 2001].

Дискриминантный анализ проведен с использованием модуля Discriminant Analysis программы Statistica, рассчитывающего значение лямбды Уилкса (λ) и *F*-критерия. Получены следующие результаты: λ Уилкса (Wilks' Lambda) = 0,02259, $F(24,126) = 29,679$, $p < 0,000$. Известно, что значения λ Уилкса, стремящиеся к 0, свидетельствуют о хорошей дискриминации сравниваемых объектов. По данным показателя λ и по значению *F*-критерия можно сделать вывод, что данная классификация корректна. Значения переменных дискриминантного анализа представлены в таблице 2.

Таблица 1. **Лингвистические параметры текстов трех предметных подкорпусов**

Table 1. **Linguistic features of texts of three sub-corpora**

Лингвистический параметр	Предметный корпус					
	Естественно-научный		Математический		Филологический	
	СЗ	СО	СЗ	СО	СЗ	СО
1. Среднее количество слов в предложении	8,94	0,66	8,76	1,51	6,21	1,04
2. Среднее количество слогов в слове	2,38	0,19	1,96	0,13	2,28	0,23
3. Среднее количество существительных на предложение	3,27	0,33	3,15	0,49	2,51	0,37
4. Среднее количество глаголов на предложение	1,40	0,16	0,96	0,16	0,93	0,17
5. Среднее количество прилагательных на предложение	0,95	0,17	0,74	0,22	0,63	0,16
6. Индекс ФК (мод) SIS	4,79	0,57	2,50	0,82	3,16	0,82
7. Индекс абстрактности	2,60	0,14	2,56	0,12	2,57	0,10
8. Локальный повтор существительного	0,15	0,06	0,38	0,07	0,10	0,04
9. Глобальный повтор существительного	0,04	0,02	0,03	0,01	0,05	0,08
10. Локальный повтор аргумента	0,45	0,13	0,67	0,10	0,28	0,09
11. Глобальный повтор аргумента	0,14	0,05	0,08	0,02	0,11	0,07
12. Лексическое разнообразие	0,63	0,06	0,46	0,05	0,60	0,04
13. Номинативность	0,43	0,07	0,31	0,03	0,37	0,06
14. Описательность	0,29	0,05	0,23	0,05	0,25	0,04

Примечание. В таблице использованы следующие обозначения: СЗ – среднее значение параметра; СО – стандартное отклонение.

Note. The following operators are used in the table: СЗ marks MEAN; CO marks STANDARD DEVIATION.

Таблица 2. **Результаты дискриминационного анализа**

Table 2. **Discriminant analysis results**

Параметры	λ Wilks'	λ Partial	*F*	*p*-value
1. Среднее количество слов в предложении	0,028	0,744	10,480	< 0,001
2. Среднее количество слогов в слове	0,021	0,993	0,227	0,798
3. Среднее количество существительных на предложение	0,023	0,925	2,456	0,094
4. Среднее количество глаголов на предложение	0,024	0,889	3,811	0,028
5. Среднее количество прилагательных на предложение	0,022	0,948	1,685	0,194
6. Индекс ФК (SIS)	0,021	0,991	0,276	0,760
7. Индекс абстрактности	0,022	0,972	0,865	0,426
8. Локальный повтор существительного	0,025	0,857	5,110	< 0,01
9. Глобальный повтор существительного	0,022	0,966	1,086	0,344
10. Локальный повтор аргумента	0,024	0,896	3,557	0,035
11. Глобальный повтор аргумента	0,025	0,845	5,580	< 0,01
12. Лексическое разнообразие	0,025	0,846	5,538	< 0,01
13. Номинативность	0,023	0,910	3,009	0,057
14. Описательность	0,022	0,974	0,819	0,446

Как видно из таблицы, наиболее сильное влияние на результаты оказывают параметры «средняя длина предложения», «глобальный повтор аргумента», «локальный повтор существительного», а также «лексическое разнообразие». Большое количество переменных в модели усложняет расчеты и интерпретацию полученных данных, связанных с «предметностью» учебного текста. Поэтому было принято решение о дальнейшей оптимизации модели с помощью метода «Backward stepwise», который позволил создать модель, состоящую из 6 параметров (см. табл. 3). После оптимизации были получены следующие

характеристики модели: λ Уилкса (Wilks' Lambda) = 0,03090, $F(12, 136) = 53,137$, $p < 0,001$.

Данные таблицы 2 показывают, большинство параметров являются синтаксическими. При этом параметры, которые обычно связывают со сложностью текста [Kupriyanov, Bukach, Aleksandrova, 2023], например «абстрактность текста», «лексическое разнообразие», «индекс удобочитаемости Флеша – Кинкейда», в модель не вошли. Это свидетельствует о том, что когнитивная сложность учебных текстов на одном образовательном уровне (в данном исследовании – начальная школа, 2-й класс) примерно одинакова и существенно не меняется в зависимости от изучаемой предметной области. Однако синтаксические параметры имеют существенные отличия в учебных текстах различных предметных областей и могут быть использованы для типологизации текстов одного уровня сложности.

Полученная модель была протестирована на 10 текстах, отобранных случайным образом (см. выше о тестовой коллекции). Проверка показала, что эффективность модели составляет 80 %: из 10 тестируемых текстов правильно были классифицированы 8 (см. рисунок).

Диаграмма рассеяния канонических значений для канонических корней демонст-

Таблица 3. **Результаты дискриминационного анализа с помощью метода «Backward stepwise»**

Table 3. **Discriminant analysis results using the "Backward stepwise" method**

Параметры	λ Wilks'	λ Partial	*F*	*p*-value
1. Среднее количество слов в предложении	0,046	0,665	17,129	<0,001
3. Среднее количество существительных на предложение	0,035	0,873	4,927	0,010
4. Среднее количество глаголов на предложение	0,063	0,491	35,182	<0,001
5. Среднее количество прилагательных на предложение	0,040	0,781	9,524	<0,001
8. Локальный повтор существительного	0,055	0,565	26,154	<0,001
9. Глобальный повтор аргумента	0,042	0,743	11,779	<0,001

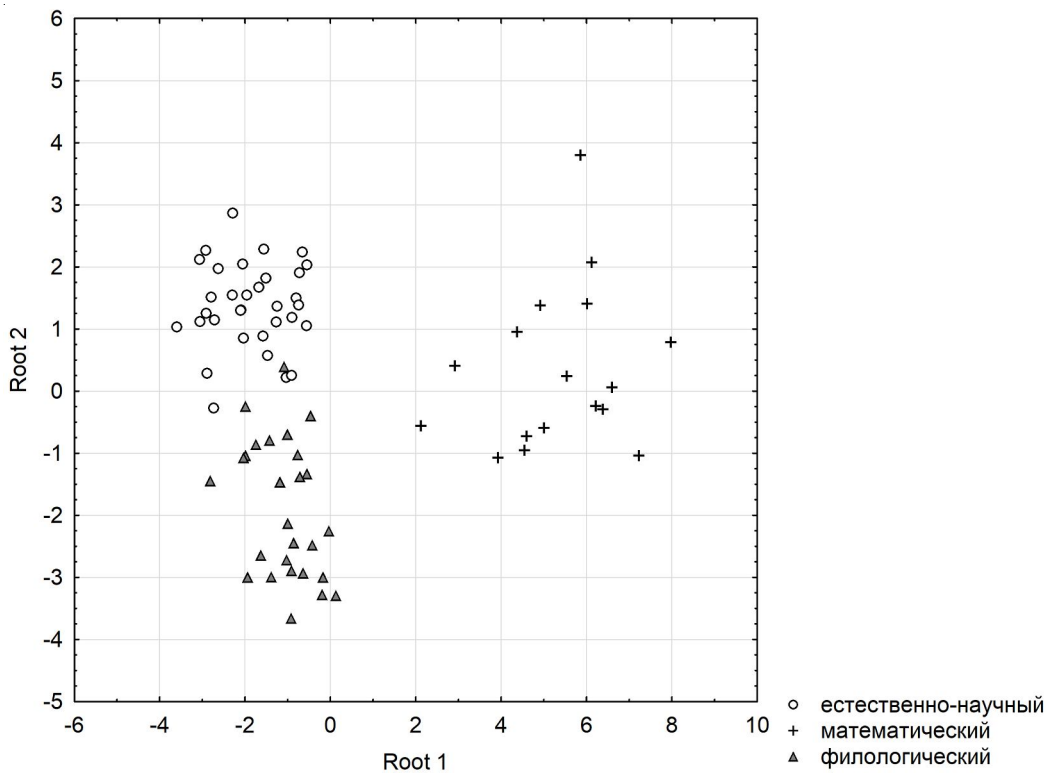

Диаграмма рассеяния канонических значений для канонических корней

Scatterplot of canonical values for canonical roots

рирует вклад, который вносит каждая дискриминантная функция в классификацию текстов предметных подкорпусов. Каноническая функция 1 (Root 1) позволяет отделить тексты по математике от текстов по филологии и естествознания: чем больше значение Root 1, тем больше вероятность того, что анализируемый текст – текст по математике. Каноническая функция 2 (Root 2) позволяет отделить тексты по филологии от текстов по математике и естествознанию: чем больше значение Root 2, тем больше вероятность того, что это текст естественно-научный. Как видно из диаграммы, области текстов по филологии и естествознанию в модели пересекаются, поэтому возможны ошибочные классификации.

Исходя из значений стандартизированных коэффициентов канонических функций (табл. 4) постулируется влияние лингвистических параметров текста на значения канонических функций 1 и 2. Судя по коэффициентам, наибольшее влияние на эти функции оказывают следующие лингвистические параметры: локальный повтор существительного, среднее количество существительных на предложение, среднее количество глаголов на предложение. В результате дискриминантного анализа выявлены коэффициенты классификационных функций (табл. 5), которые позволяют автоматически определить предметную область учебного текста по его лингвистическим параметрам.

Таким образом, формулы для классификации текстов по предметным областям выглядят следующим образом:

$$F(\text{Ест-науч}) = -62{,}88 + (-10{,}26 \times X1) + 29{,}16 \times X2 + 68{,}76 \times X3 + 16{,}90 \times X4 + 9{,}96 \times X5 + 40{,}50 \times X6;$$

$$F(\text{Матем}) = -53{,}50 + 2{,}74 \times X1 + 13{,}03 \times X2 + 15{,}21 \times X3 + (-9{,}72 \times X4) + 94{,}10 \times X5 + (-38{,}07 \times X6);$$

$$F(\text{Филол}) = -34{,}55 + (-9{,}38 \times X1) + 28{,}34 \times X2 + 49{,}05 \times X3 + 7{,}23 \times X4 + 1{,}34 \times X5 + 33{,}26 \times X6.$$

Выводы

На основе дискриминантного анализа была разработана математическая модель типологизации текстов, в которую после оп-

Таблица 4. **Стандартизированные коэффициенты канонических функций**

Table 4. **Standardized coefficients of canonical functions**

Параметры текста	Канонические функции	
	Root 1	Root 2
1. Среднее количество слов в предложении	–0,111	0,786
2. Среднее количество существительных на предложение	–0,044	0,580
3. Среднее количество глаголов на предложение	0,291	0,773
4. Среднее количество прилагательных на предложение	0,109	0,535
5. Локальный повтор существительного	–0,445	0,511
6. Глобальный повтор аргумента	0,149	0,144

Таблица 5. **Коэффициенты классификационных функций**

Table 5. **Classification function coefficients**

Переменные (X) и константы	Обозначение переменной	Предметный подкорпус		
		Естественно-научный	Математический	Филологический
X1	Среднее количество слов в предложении	–10,26	2,74	–9,38
X2	Среднее количество существительных на предложение	29,16	13,03	28,34
X3	Среднее количество глаголов на предложение	68,76	15,21	49,05
X4	Среднее количество прилагательных на предложение	16,90	–9,72	7,23
X5	Локальный повтор существительного	9,96	94,10	1,34
X6	Глобальный повтор аргумента	40,50	–38,07	33,26
Constant	Константа	–62,88	–53,50	–34,55

тимизации вошли шесть лингвистических параметров: средняя длина предложения, среднее количество существительных на предложение, среднее количество глаголов на предложение, среднее количество прилагательных на предложение, локальный повтор существительных, глобальный повтор аргумента.

Учебные тексты трех изучаемых подкорпусов характеризуются статистически значимыми различиями. Вероятностной причиной этого следует признать различия в их функционале.

Естественно-научные тексты должны формировать целостную картину мира и расширять кругозор читателя, поэтому они имеют более длинные предложения, где описываются явления живой и неживой природы, а также связи между ними. Такие учебные тексты имеют в среднем большее количество существительных, прилагательных и глаголов на одно предложение в сравнении с математическими и филологическими текстами.

Функционально учебные тексты по математике направлены на развитие математических навыков у ребенка, то есть способности применять символы и абстрактные понятия в мыслительных операциях, а также овладение правилами и способами их использования для решения задач. Спецификой учебных текстов по математике является высокая частотность терминов и абстрактных лексических единиц (традиционно обозначаемых существительными), при этом использование синонимов в ряде случаев затруднительно, а в некоторых случаях – невозможно. Этим же можно объяснить и высокие значения коэффициента у параметра «локальный повтор существительного» в функции Root 1, которая позволяет идентифицировать предметную область и отделить учебные тексты по математике от других учебных текстов.

Филологические тексты нацелены на отработку навыков письма и анализа высказывания, поэтому учебники по русскому языку для начальной школы состоят из упражнений и заданий, для текстов в этих учебниках характерна небольшая длина предложения и простая лексика по сравнению с естественно-научными текстами.

Автоматизация расчетов данных параметров на основе созданной методики типологизации текстов может служить основой лингвистической экспертизы учебных текстов, а также позволит создать профайлеры текстов, способствующие оперативному осуществлению параметрического анализа текстов. Перспектива представленного исследования видится в возможности инсталляции модели в RuLingva и разработке аналогичных моделей для текстов учебного дискурса других предметных областей, использовании его результатов для создания квантитативной лингвистической типологии текстов.

ПРИМЕЧАНИЯ

[1] Работа выполнена за счет средств Программы стратегического академического лидерства Казанского (Приволжского) федерального университета («ПРИОРИТЕТ-2030»), Стратегического проекта № 5.

This paper has been supported by the Kazan Federal University Strategic Academic Leadership Program (‘PRIORITY-2030’), Strategic Project #5.

[2] Здесь и далее перевод на русский язык осуществлен авторами статьи.

[3] Приказ Минпросвещения России № 254 от 20 мая 2020 года.

СПИСОК ЛИТЕРАТУРЫ

Андреев В. С., 2010. Методы количественного исследования стиля в лингвистике: многомерный подход // Известия Смоленского государственного университета. № 3 (11). С. 100–110.

Вахрушева А. Я., Солнышкина М. И., Куприянов Р. В., Гафиятова Э. В., Климагина И. О., 2021. Лингвистическая сложность учебных текстов // Вопросы журналистики, педагогики, языкознания. № 40 (1). С. 89–99. URL: http://jpl-journal.ru/index.php/journal/article/view/78

Жеребцова Ж. И., 2007. Использование информационной структуры предложения в обучении иностранных студентов-нефилологов чтению русских учебно-научных текстов : дис. ... канд. пед. наук. СПб. 252 с.

Журавлев А. Ф., 1988. Опыт квантитативно-типологического исследования разновидностей устной речи // Разновидности городской устной речи : сб. науч. тр. М. : Наука. С. 84–150.

Мартынова Е., Солнышкина М. И., Мерзлякова А., Гизатулина Д., 2020. Лексические параметры учебного текста (на материале текстов учебного корпуса русского языка) // Филология и куль-

тура. Philology and Culture : электрон. журн. № 3 (61). С. 72–80. URL: http://www.philology-and-culture.kpfu.ru/?q=node/2728

Оборнева И. В., 2006. Автоматизированная оценка сложности учебных текстов на основе статистических параметров : автореф. дис. ... канд. пед. наук. М. 20 с.

Разговорная речь в системе функциональных стилей современного русского литературного языка. Лексика, 2009 / [О. Б. Сиротинина и др.]. М. : Либроком. 251 с.

Солнышкина М. И., Казачкова М. Б., Харькова Е. В., 2020. Инструменты измерения сложности текстов на английском языке // Иностранные языки в школе : электрон. журн. № 3. С. 15–21. URL: https://www.elibrary.ru/item.asp?id=42609743

Солнышкина М. И., Кисельников А. С., 2015. Сложность текста: этапы изучения в отечественном прикладном языкознании // Вестник Томского государственного университета. Филология. №. 6 (38). С. 86–89.

Baayen R. H., van Halteren H., Tweedie F. J., 1996. Outside the Cave of Shadows: Using Syntactic Annotation to Enhance Authorship Attribution // Literary and Linguistic Computing. Vol. 11, № 3. P. 121–132.

Biber D., 2006. University Language: A Corpus-Based Study of Spoken and Written Registers. Amsterdam : John Benjamins. VIII, 261 p.

Bulté B., Housen A., 2012. Defining and Operationalising L2 Complexity // Dimensions of L2 Performance and Proficiency. Amsterdam : John Benjamins. P. 21–46. DOI: 10.1075/lllt.32.02bul

Corlatescu D., Ruseti Ş., Dascalu M., 2022. ReaderBench: Multilevel Analysis of Russian Text Characteristics // Russian Journal of Linguistics. Vol. 26, № 2, P. 342–370. DOI: https://doi.org/10.22363/2687-0088-30145

Crossley S. A., Varner L. K., Roscoe R. D., McNamara D. S., 2013. Using Automated Indices of Cohesion to Evaluate an Intelligent Tutoring System and an Automated Writing Evaluation System // Artificial Intelligence in Education. 16th International Conference, AIED 2013, Memphis, TN, USA, July 9–13. Berlin ; Heidelberg : Springer. P. 269–278. DOI: https://doi.org/10.1007/978-3-642-39112-5_28

Ermakova L., Solovyev V., Sidorov G., Gelbukh A., 2023. Editorial: Text Complexity and Simplification // Frontiers in Artificial Intelligence. Vol. 6. P. 01–03. DOI: https://doi.org/10.3389/frai.2023.1128446

Flesch R., 1948. A New Readability Yardstick // Journal of Applied Psychology. Vol. 32, № 3. P. 221–233. DOI: http://doi.org/ 10.1037/h0057532

Gatiyatullina G., Solnyshkina M., Solovyev V., Danilov A., Martynova E., Yarmakeev I., 2020. Computing Russian Morphological Distribution Patterns Using RusAC Online Server // 13th International Conference on Developments in eSystems Engineering (DeSE). Liverpool : IEEE. P. 393–398. DOI: http://doi.org/10.1109/DeSE51703.2020.9450753

Graesser A. C., McNamara D. S., Louwerse M. M., Cai Z., 2004. Coh-Metrix: Analysis of Text on Cohesion and Language // Behavior Research Methods, Instruments, & Computers. Vol. 36, iss. 2. P. 193–202. DOI: http://doi.org/10.3758/bf03195564

Holmes D., Forsyth R., 1995. The Federalist Revisited: New Directions in Authorship Attribution // Literary and Linguistic Computing. Vol. 10, iss. 2. P. 111–127.

Kupriyanov R. V., Solnyshkina M. I., Dascalu M., Soldatkina T. A., 2022. Lexical and Syntactic Features of Academic Russian Texts: A Discriminant Analysis // Research Result. Theoretical and Applied Linguistics. Vol. 8, № 4. P. 105–122. DOI: http://doi.org/10.18413/2313-8912-2022-8-4-0-8

Kupriyanov R. V., Bukach O. V., Aleksandrova O. I., 2023. Cognitive Complexity Measures for Educational Texts: Empirical Validation of Linguistic Parameters // Russian Journal of Linguistics. Vol. 27, № 3. P. 641–662. DOI: http://doi.org/10.22363/2687-0088-35817

Malvern D., Richards B., Chipere N., Durán P., 2004. Traditional Approaches to Measuring Lexical Diversity // Lexical Diversity and Language Development. L. : Palgrave Macmillan. P. 16–30. DOI: https://doi.org/10.1057/9780230511804

McNamara D. S., Graesser A. C., Louwerse M. M., 2012. Sources of Text Difficulty: Across Genres and Grades // Measuring Up: Advances in How We Assess Reading Ability. Lanham : R & L Education. P. 89–116.

McNamara D., Graesser A., McCarthy P., Cai Z., 2014. Automated Evaluation of Text and Discourse with Coh-Metrix. Cambridge : Cambridge University Press. XIV, 278 p. DOI: http://doi.org/10.1017/CBO9780511894664

Östen D., 2004. The Growth and Maintenance of Linguistic Complexity. Amsterdam, John Benjamins Publishing. X, 333 p.

Pallotti G., 2015. A Simple View of Linguistic Complexity // Second Language Research. Vol. 31, № 1. P. 117–134.

Seifart F., Danielsen S., Meyer R., Nordhoff S., Pakendorf B., Witzlack-Makarevich A., Zakharko T., 2012. The Relative Frequencies of Nouns, Pronouns, and Verbs Cross-Linguistically Applicant. URL: https://www.semanticscholar.org/paper/The-relative-frequencies-of-nouns-%2C-pronouns-%2C-and-Seifart-Danielsen/cd52cd7091fee4b1781c16a51fe58f87ba642c1c

Solnyshkina M. I., Harkova E. V., Kazachkova M. B., 2020. The Structure of Cross-Linguistic Differences: Meaning and Context of 'Readability' and Its Russian Equivalent 'Chitabelnost' // Journal of Language and Education. Vol. 6, iss. 1. P. 103–119.

Solnyshkina M., McNamara D., Zamaletdinov R., 2022. Natural Language Processing and Discourse Complexity Studies // Russian Journal of Linguistics. Vol. 26, № 2. P. 317–341.

Solovyev V., Andreeva M., Solnyshkina M., Zamaletdinov R., Danilov A., Gaynutdinova D., 2019. Computing Concreteness Ratings of Russian and English Most Frequent Words: Contrastive Approach // Proceedings – International Conference on Developments in eSystems Engineering, DeSE. October 2019. Kazan : IEEE. Art. № 9073272. P. 403–408.

Solovyev V. D., Ivanov V. V., Akhtiamov R. B., 2019. Dictionary of Abstract and Concrete Words of the Russian Language: A Methodology for Creation and Application // Journal of Research in Applied Linguistics. Vol. 10, № S. P. 215–227.

Solovyev V., Ivanov V., Solnyshkina M., 2018. Assessment of Reading Difficulty Levels in Russian Academic Texts: Approaches and Metrics // Journal of Intelligent & Fuzzy Systems. Vol. 34, № 5. DOI: http://doi.org/10.3233/JIFS-169489

Solovyev V., Solnyshkina M., McNamara D., 2022. Computational Linguistics and Discourse Complexology: Paradigms and Research Methods // Russian Journal of Linguistics. Vol. 26, № 2. P. 275–316.

Stamatatos E., Fakotakis N., Kokkinakis G., 2001. Computer-Based Authorship Attribution Without Lexical Measures // Computers and the Humanities. Vol. 35, № 2. P. 193–214.

ИСТОЧНИКИ

Дмитриева Н. Я., Казаков А. Н. Окружающий мир. В 2 ч. Ч. 1 : учеб. для 2 кл. 8-е изд. Самара : Учеб. лит. : Федоров, 2012. 112 с.

Дмитриева Н. Я., Казаков А. Н. Окружающий мир. В 2 ч. Ч. 2 : учеб. для 2 кл. 7-е изд. Самара : Учеб. лит. : Федоров, 2011. 112 с.

Ивченкова Г. Г., Потапов И. В. Окружающий мир. В 2 ч. Ч. 1 : учеб. для 2 кл. М. : АСТ : Астрель, 2012. 78, [2] с.

Ивченкова Г. Г., Потапов И. В. Окружающий мир. В 2 ч. Ч. 2 : учеб. для 2 кл. М. : АСТ : Астрель, 2012. 93, [3] с.

Моро М. И., Бантова М. А., Бельтюкова Г. В. и др. Математика. 2 класс. В 2 ч. Ч. 1 : учеб. для общеобразоват. орг. 6-е изд. М. : Просвещение, 2015. 96 с.

Моро М. И., Бантова М. А., Бельтюкова Г. В. и др. Математика. 2 класс. В 2 ч. Ч. 2 : учеб. для общеобразоват. орг. 6-е изд. М. : Просвещение, 2015. 112 с.

Петерсон Л. Г. Математика «Учусь учиться». 2 класс. Ч. 1 : учеб. комплекта «Учебник + рабочие тетради». Изд. 5-е, перераб. М. : Ювента, 2013. 80 с.

Петерсон Л. Г. Математика «Учусь учиться». 2 класс. Ч. 2 : учеб. комплекта «Учебник + рабочие тетради». Изд. 5-е, перераб. М. : Ювента, 2013. 112 с.

Петерсон Л. Г. Математика «Учусь учиться». 2 класс. Ч. 3 : учеб. комплекта «Учебник + рабочие тетради». Изд. 5-е, перераб. М. : Ювента, 2013. 112 с.

Рамзаева Т. Г. Русский язык. 2 кл. В 2 ч. Ч. 1. 12-е изд., дораб. М. : Дрофа, 2011. 126, [2] с.

Рамзаева Т. Г. Русский язык. 2 кл. В 2 ч. Ч. 2. 12-е изд., дораб. М. : Дрофа, 2011. 94, [2] с.

Соловейчик М. С., Кузьменко Н. С. Русский язык : К тайнам нашего языка. В 2 ч. Ч. 1 : учеб. для 2 кл. общеобразоват. учреждений. 8-е изд. Смоленск : Ассоциация XXI век, 2013. 160 с.

Соловейчик М. С., Кузьменко Н. С. Русский язык : К тайнам нашего языка. В 2 ч. Ч. 2 : учеб. для 2 кл. общеобразоват. учреждений. 8-е изд. Смоленск : Ассоциация XXI век, 2013. 160 с.

REFERENCES

Andreev V.S., 2010. Metody kolichestvennogo issledovaniya stilya v lingvistike: mnogomernyy podkhod [Methods of Quantitative Style Research in Linguistics: A Multidimensional Approach]. *Izvestiya Smolenskogo gosudarstvennogo universiteta*, no. 3 (11), pp. 100-110.

Vakhrusheva A.Ya., Solnyshkina M.I., Kupriyanov R.V., Gafiyatova E.V., Klimagina I.O., 2021. Lingvisticheskaya slozhnost uchebnykh tekstov [Linguistic Complexity of Academic Texts]. *Voprosy zhurnalistiki, pedagogiki, yazykoznaniya* [Issues in Journalism, Education, Linguistics], no. 40 (1), pp. 89-99. URL: http://jpl-journal.ru/index.php/journal/article/view/78

Zherebtsova Zh.I., 2007. *Ispolzovanie informatsionnoy struktury predlozheniya v obuchenii inostrannykh studentov-nefilologov chteniyu russkikh uchebno-nauchnykh tekstov: dis. ... kand. ped. nauk* [The Use of Information Structure of the Sentence in Teaching Foreign Non-Philological Students to Read Russian

Academic and Research Texts. Cand. pedagog. sci. diss.]. Saint Petersburg. 252 p.

Zhuravlev A.F., 1988. Opyt kvantitativno tipologicheskogo issledovaniya raznovidnostey ustnoy rechi [An Experience of Quantitative and Typological Investigation of Spoken Registers]. *Raznovidnosti gorodskoy ustnoy rechi: sb. nauch. tr.* [Varieties of Urban Oral Speech. Collection of Scientific Papers]. Moscow, Nauka Publ., pp. 84-150.

Martynova E., Solnyshkina M.I., Merzlyakova A., Gizatulina D., 2020. Leksicheskie parametry uchebnogo teksta (na materiale tekstov uchebnogo korpusa russkogo yazyka) [Lexical Parameters of Academic Text (Based on the Texts of Academic Corpus of the Russian Language)]. *Filologiya i kultura. Philology and Culture: electron. zhurn.*, no. 3 (61), pp. 72-80. URL: http://www.philology-and-culture.kpfu.ru/?q=node/2728

Oborneva I.V., 2006. *Avtomatizirovannaya otsenka slozhnosti uchebnykh tekstov na osnove statisticheskikh parametrov: avtoref. dis. ... kand. ped. nauk* [Automated Assessment of the Complexity of Educational Texts Based on Statistics. Cand. pedagog. sci. abs. diss.]. Moscow. 20 p.

Sirotinina O.B. et al., 2009. *Razgovornaya rech v sisteme funktsionalnykh stiley sovremennogo russkogo literaturnogo yazyka. Leksika* [Colloquial Speech in the System of Functional Styles of the Modern Russian Literary Language. Vocabulary]. Moscow, Librokom Publ. 251 p.

Solnyshkina M.I., Kazachkova M.B., Kharkova E.V., 2020. Instrumenty izmereniya slozhnosti tekstov na angliyskom yazyke [Tools for Measuring English Texts Complexity]. *Inostrannye yazyki v shkole: electron. zhurn.* [Foreign Languages in School. Electronic Journal], no. 3, pp. 15-21. URL: https://www.elibrary.ru/item.asp?id=42609743

Solnyshkina M.I., Kiselnikov A.S., 2015. Slozhnost teksta: etapy izucheniya v otechestvennom prikladnom yazykoznanii [Text Complexity: Chronology of Russian Applied Linguistics Studies]. *Vestnik Tomskogo gosudarstvennogo universiteta. Filologiya*, no. 6 (38), pp. 86-89.

Baayen R.H., van Halteren H., Tweedie F.J., 1996. Outside the Cave of Shadows: Using Syntactic Annotation to Enhance Authorship Attribution. *Literary and Linguistic Computing*, vol. 11, no. 3, pp. 121-132.

Biber D., 2006. *University Language: A Corpus-Based Study of Spoken and Written Registers*. Amsterdam, John Benjamins. VIII, 261 p.

Bulté B., Housen A., 2012. Defining and Operationalising L2 Complexity. *Dimensions of L2 Performance and Proficiency*. Amsterdam, John Benjamins, pp. 21-46. DOI: https://doi.org/10.1075/lllt.32.02bul

Corlatescu D., Ruseti Ş., Dascalu M., 2022. ReaderBench: Multilevel Analysis of Russian Text Characteristics. *Russian Journal of Linguistics*, vol. 26, no. 2, pp. 342-370. DOI: https://doi.org/10.22363/2687-0088-30145

Crossley S.A., Varner L.K., Roscoe R.D., McNamara D.S., 2013. Using Automated Indices of Cohesion to Evaluate an Intelligent Tutoring System and an Automated Writing Evaluation System. *Artificial Intelligence in Education. 16th International Conference, AIED 2013, Memphis, TN, USA, July 9–13*. Berlin, Heidelberg, Springer, pp. 269-278. DOI: https://doi.org/10.1007/978-3-642-39112-5_28

Ermakova L., Solovyev V., Sidorov G., Gelbukh A., 2023. Editorial: Text Complexity and Simplification. *Frontiers in Artificial Intelligence*, vol. 6, pp. 01-03. DOI: https://doi.org/10.3389/frai.2023.1128446

Flesch R., 1948. A New Readability Yardstick. *Journal of Applied Psychology*, vol. 32, no. 3, pp. 221-233. DOI: http://doi.org/ 10.1037/h0057532

Gatiyatullina G., Solnyshkina M., Solovyev V., Danilov A., Martynova E., Yarmakeev I., 2020. Computing Russian Morphological Distribution Patterns Using RusAC Online Server. *13th International Conference on Developments in eSystems Engineering (DeSE)*. Liverpool, IEEE, pp. 393-398. DOI: http://doi.org/10.1109/DeSE51703.2020.9450753

Graesser A.C., McNamara D.S., Louwerse M.M., Cai Z., 2004. Coh-Metrix: Analysis of Text on Cohesion and Language. *Behavior Research Methods, Instruments, & Computers*, vol. 36, iss. 2, pp. 193-202. DOI: http://doi.org/10.3758/bf03195564

Holmes D., Forsyth R., 1995. The Federalist Revisited: New Directions in Authorship Attribution. *Literary and Linguistic Computing*, vol. 10, iss. 2, pp. 111-127.

Kupriyanov R.V., Solnyshkina M.I., Dascalu M., Soldatkina T.A., 2022. Lexical and Syntactic Features of Academic Russian Texts: A Discriminant Analysis. *Research Result. Theoretical and Applied Linguistics*, vol. 8, no. 4, pp. 105-122. DOI: http://doi.org/10.18413/2313-8912-2022-8-4-0-8

Kupriyanov R.V., Bukach O.V., Aleksandrova O.I., 2023. Cognitive Complexity Measures for Educational Texts: Empirical Validation of Linguistic Parameters. *Russian Journal of Linguistics*, vol. 27, no. 3, pp. 641-662. DOI: http://doi.org/10.22363/2687-0088-35817

Malvern D., Richards B., Chipere N., Durán P., 2004. Traditional Approaches to Measuring Lexical Diversity. *Lexical Diversity and Language*

Development. London, Palgrave Macmillan, pp. 16-30. DOI: https://doi.org/10.1057/9780230511804

McNamara D.S., Graesser A.C., Louwerse M.M., 2012. Sources of Text Difficulty: Across Genres and Grades. *Measuring Up: Advances in How We Assess Reading Ability*. Lanham, R & L Education, pp. 89-116.

McNamara D., Graesser A., McCarthy P., Cai Z., 2014. *Automated Evaluation of Text and Discourse with Coh-Metrix*. Cambridge, Cambridge University Press. XIV, 278 p. DOI: http://doi.org/10.1017/CBO9780511894664

Östen D., 2004. *The Growth and Maintenance of Linguistic Complexity*. Amsterdam, John Benjamins Publishing. X, 333 p.

Pallotti G., 2015. A Simple View of Linguistic Complexity. *Second Language Research*, vol. 31, no. 1, pp. 117-134.

Seifart F., Danielsen S., Meyer R., Nordhoff S., Pakendorf B., Witzlack-Makarevich A., Zakharko T., 2012. *The Relative Frequencies of Nouns, Pronouns, and Verbs Cross-Linguistically Applicant*. URL: https://www.semanticscholar.org/paper/The-relative-frequencies-of-nouns-%2C-pronouns-%2C-and-Seifart-Danielsen/cd52cd7091fee4b1781c16a51fe58f87ba642c1c

Solnyshkina M.I., Harkova E.V., Kazachkova M.B., 2020. The Structure of Cross-Linguistic Differences: Meaning and Context of Readability and Its Russian Equivalent "Chitabelnost". *Journal of Language and Education*, vol. 6, iss. 1, pp. 103-119.

Solnyshkina M., McNamara D., Zamaletdinov R., 2022. Natural Language Processing and Discourse Complexity Studies. *Russian Journal of Linguistics*, vol. 26, no. 2, pp. 317-341.

Solovyev V., Andreeva M., Solnyshkina M., Zamaletdinov R., Danilov A., Gaynutdinova D., 2019. Computing Concreteness Ratings of Russian and English Most Frequent Words: Contrastive Approach. *Proceedings – International Conference on Developments in eSystems Engineering, DeSE. October 2019*. Kazan, IEEE. Art. no. 9073272, pp. 403-408.

Solovyev V.D., Ivanov V.V., Akhtiamov R.B., 2019. Dictionary of Abstract and Concrete Words of the Russian Language: A Methodology for Creation and Application. *Journal of Research in Applied Linguistics*, vol. 10, no. S, pp. 215-227.

Solovyev V., Ivanov V., Solnyshkina M., 2018. Assessment of Reading Difficulty Levels in Russian Academic Texts: Approaches and Metrics. *Journal of Intelligent & Fuzzy Systems*, vol. 34, no. 5. DOI: http://doi.org/10.3233/JIFS-169489

Solovyev V., Solnyshkina M., McNamara D., 2022. Computational Linguistics and Discourse Complexology: Paradigms and Research Methods. *Russian Journal of Linguistics*, vol. 26, no. 2, pp. 275-316.

Stamatatos E., Fakotakis N., Kokkinakis G., 2001. Computer-Based Authorship Attribution Without Lexical Measures. *Computers and the Humanities*, vol. 35, no. 2, pp. 193-214.

SOURCES

Dmitrieva N.Ya., Kazakov A.N. *Okruzhayushchiy mir. V 2 ch. Ch. 1: ucheb. dlya 2 kl.* [The World Around Us. In 2 Parts. Pt. 1. Textbook for 2nd Grade]. Samara, Ucheb. lit. Publ., Fedorov Publ., 2012. 112 p.

Dmitrieva N.Ja., Kazakov A.N., 2011. *Okruzhayushchiy mir. V 2 ch. Ch. 2: ucheb. dlya 2 kl.* [The World Around Us. In 2 Parts. Pt. 2. Textbook for 2nd Grade]. Samara, Ucheb. lit. Publ., Fedorov Publ., 2011. 112 p.

Ivchenkova G.G., Potapov I.V. *Okruzhayushchiy mir. V 2 ch. Ch. 1: ucheb. dlya 2 kl.* [The World Around Us. In 2 Parts. Pt. 1. Textbook for 2nd Grade]. Moscow, AST Publ., Astrel Publ., 2012. 78, 2 p.

Ivchenkova G.G., Potapov I.V. *Okruzhayushchiy mir. V 2 ch. Ch. 2: ucheb. dlya 2 kl.* [The World Around Us. In 2 Parts. Pt. 2. Textbook for 2nd Grade]. Moscow, AST Publ., Astrel Publ., 2012. 93, 3 p.

Moro M.I., Bantova M.A., Beltyukova G.V. et al. *Matematika. 2 klass. V 2 ch. Ch. 1: ucheb. dlya obshcheobrazovat. org.* [Mathematics. 2nd Grade. In 2 Parts. Part 1. Textbook for General Education Organizations]. Moscow, Prosveshchenie Publ., 2015. 96 p.

Moro M.I., Bantova M.A., Beltyukova G.V. et al. *Matematika. 2 klass. V 2 ch. Ch. 2: ucheb. dlya obshcheobrazovat. org.* [Mathematics. 2nd Grade. In 2 Parts. Part 2. Textbook for General Education Organizations]. Moscow, Prosveshchenie Publ., 2015. 112 p.

Peterson L.G. *Matematika «Uchus uchitsya». 2 klass. Ch. 1: ucheb. komplekta «Uchebnik + rabochie tetradi»* [Mathematics "Learning to Learn." 2nd Grade. Part 1. Textbook of the "Textbook + Workbooks" Kit]. Moscow, Yuventa Publ., 2013. 80 p.

Peterson L.G. *Matematika «Uchus uchitsya». 2 klass. Ch. 2: ucheb. komplekta «Uchebnik + rabochie tetradi»* [Mathematics "Learning to

Learn." 2nd Grade. Part 2. Textbook of the "Textbook + Workbooks" Kit]. Moscow, Yuventa Publ., 2013. 112 p.

Peterson L.G. *Matematika «Uchus uchitsya». 2 klass. Ch. 3: ucheb. komplekta «Uchebnik + rabochie tetradi»* [Mathematics "Learning to Learn." 2nd Grade. Part 3. Textbook of the "Textbook + Workbooks" Kit.]. Moscow, Yuventa Publ., 2013. 112 p.

Ramzaeva T.G. *Russkiy yazyk. 2 kl. V 2 ch. Ch. 1* [Russian Language. 2nd Grade. In 2 Parts. Part 1]. Moscow, Drofa Publ., 2011. 126, 2 p.

Ramzaeva T.G. *Russkiy yazyk. 2 kl. V 2 ch. Ch. 2* [Russian Language. 2nd Grade. In 2 Parts. Part 2]. Moscow, Drofa Publ., 2011. 94, 2 p.

Soloveychik M.S., Kuzmenko N.S. *Russkiy yazyk: K taynam nashego yazyka. V 2 ch. Ch. 1: ucheb. dlya 2 kl. obshcheobrazovat. uchrezhdeniy* [Russian Language: To the Secrets of Our Language. In 2 Parts. Part 1. Textbook for 2nd Grade of General Education Institutions]. Smolensk, Assotsiatsiya XXI vek Publ., 2013.160 p.

Soloveychik M.S., Kuzmenko N.S. *Russkiy yazyk: K taynam nashego yazyka. V 2 ch. Ch. 2: ucheb. dlya 2 kl. obshcheobrazovat. uchrezhdeniy* [Russian Language: To the Secrets of Our Language. In 2 Parts. Part 2. Textbook for 2nd Grade of General Education Institutions]. Smolensk, Assotsiatsiya XXI vek Publ., 2013. 160 p.

Information About the Authors

Roman V. Kupriyanov, Candidate of Sciences (Psychology), Senior Researcher, Text Analytics Laboratory, Kazan Federal University, Kremlevskaya St, 18, 420008 Kazan, Russia; Associate Professor, Department of Social Work, Pedagogy and Psychology, Kazan National Research Technological University, Karla Marxa St, 68, 420015 Kazan, Russia, kroman1@mail.ru, https://orcid.org/0000-0001-9794-9607

Marina I. Solnyshkina, Doctor of Sciences (Philology), Head and Chief Researcher, Text Analytics Laboratory, Professor, Department of Theory and Practice of Teaching Foreign Languages, Kazan Federal University, Kremlevskaya St, 18, 420008 Kazan, Russia, mesoln@yandex.ru, https://orcid.org/0000-0003-1885-3039

Polina A. Lekhnitskaya, Research Laboratory Assistant, Neurocognitive Research Laboratory, Kazan Federal University, Kremlevskaya St, 18, 420008 Kazan, Russia, lekhnitskaya.polina@gmail.com, https://orcid.org/0000-0002-3689-3213

Информация об авторах

Роман Владимирович Куприянов, кандидат психологических наук, старший научный сотрудник НИЛ «Текстовая аналитика», Казанский (Приволжский) федеральный университет, ул. Кремлевская, 18, 420008 г. Казань, Россия; доцент кафедры социальной работы, педагогики и психологии, Казанский национальный исследовательский технологический университет, ул. Карла Маркса, 68, 420015 г. Казань, Россия, kroman1@mail.ru, https://orcid.org/0000-0001-9794-9607

Марина Ивановна Солнышкина, доктор филологических наук, руководитель и главный научный сотрудник НИЛ «Текстовая аналитика», профессор кафедры теории и практики преподавания иностранных языков, Казанский (Приволжский) федеральный университет, ул. Кремлевская, 18, 420008 г. Казань, Россия, mesoln@yandex.ru, https://orcid.org/0000-0003-1885-3039

Полина Александровна Лехницкая, лаборант-исследователь НИЛ «Нейрокогнитивные исследования», Казанский (Приволжский) федеральный университет, ул. Кремлевская, 18, 420008 г. Казань, Россия, lekhnitskaya.polina@gmail.com, https://orcid.org/0000-0002-3689-3213