ТЕОРЕТИЧЕСКАЯ, ПРИКЛАДНАЯ И СРАВНИТЕЛЬНО-СОПОСТАВИТЕЛЬНАЯ ЛИНГВИСТИКА/THEORETICAL, APPLIED AND COMPARATIVE LINGUISTICS

DOI: https://doi.org/10.60797/IRJ.2026.163.54

ИДЕНТИФИКАЦИЯ И ВЕРИФИКАЦИЯ МНОГОФАКТОРНЫХ ПРЕДИКТОРОВ СЛОЖНОСТИ ТЕКСТА В АСПЕКТЕ ПРЕПОДАВАНИЯ РУССКОГО ЯЗЫКА КАК ИНОСТРАННОГО

Научная статья

Бай Ж.[1, *]

[1] Санкт-Петербургский государственный университет, Санкт-Петербург, Российская Федерация

* Корреспондирующий автор (261585598[at]qq.com)

Аннотация

В статье представлено корпусное исследование, направленное на выявление и верификацию объективных лексико-синтаксических предикторов сложности учебных текстов для изучающих русский язык как иностранный (РКИ). На материале текстов, стратифицированных по уровням CEFR (A1–C2), с помощью статистической среды R был проведён анализ четырёх ключевых параметров: лексического разнообразия (характеристика K Юла), синтаксической сложности (средняя длина предложения и средняя дистанция синтаксической зависимости) и стилистических особенностей (соотношение существительных и глаголов). Результаты статистического анализа (критерий Краскела-Уоллиса и корреляция Спирмена) показали, что синтаксические предикторы и лексическое разнообразие являются высокозначимыми индикаторами, демонстрирующими сильную корреляцию с уровнем владения языком. Соотношение существительных и глаголов также выявило значимую тенденцию к номинализации в текстах более высоких уровней. Исследование предлагает многомерную модель для объективной оценки текстовой сложности и закладывает эмпирическую основу для будущих психолингвистических экспериментов.

Ключевые слова: лексическая сложность, синтаксическая сложность, русский язык как иностранный (РКИ), корпусная лингвистика, вычислительная лингвистика.

IDENTIFICATION AND VERIFICATION OF MULTIFACTORIAL PREDICTORS OF TEXT COMPLEXITY IN THE CONTEXT OF TEACHING RUSSIAN AS A FOREIGN LANGUAGE

Research article

Bai R.[1, *]

[1] St. Petersburg State University, Saint-Petersburg, Russian Federation

* Corresponding author (261585598[at]qq.com)

Abstract

The article presents a corpus study aimed at identifying and verifying objective lexical and syntactic predictors of the complexity of educational texts for learners of Russian as a foreign language (RFL). Using texts stratified according to CEFR levels (A1–C2), four key parameters were analysed using the R statistical environment: lexical diversity (Yule's K characteristic), syntactic complexity (average sentence length and average syntactic dependency distance) and stylistic traits (ratio of nouns to verbs). The results of statistical analysis (Kraskel-Wallis criterion and Spearman's correlation) showed that syntactic predictors and lexical diversity are highly significant indicators demonstrating a strong correlation with language proficiency. The ratio of nouns to verbs also showed a significant tendency towards nominalisation in higher-level texts. The study proposes a multidimensional model for the objective evaluation of text complexity and lays the empirical foundation for future psycholinguistic experiments.

Keywords: lexical complexity, syntactic complexity, Russian as a foreign language (RFL), corpus linguistics, computational linguistics.

Введение

Оценка сложности текста является одной из фундаментальных задач в области преподавания русского языка как иностранного (РКИ). Адекватный подбор учебных материалов, соответствующий текущему уровню владения языком обучающегося, напрямую влияет на эффективность усвоения знаний и является залогом успешной языковой подготовки [1]. Однако традиционные методы оценки, часто опирающиеся на интуицию преподавателя или ограниченный набор лексико-грамматических тем, не всегда обеспечивают необходимую объективность и последовательность [5]. В связи с этим возникает острая необходимость в разработке и применении объективных, количественных методов анализа, способных предоставить эмпирически обоснованные критерии для классификации текстов по уровням сложности.

Современная вычислительная лингвистика предлагает широкий инструментарий для количественного анализа текста, однако многие существующие метрики, такие как простая частотность слов или классические индексы удобочитаемости, оказываются недостаточными для языков с богатой морфологией и свободным порядком слов, каким является русский. Сложность текста — это многомерное явление, которое не может быть сведено к одному или двум параметрам [9]. Она охватывает лексическое разнообразие, синтаксическую структуру, стилистические особенности и когнитивную нагрузку, возникающую при обработке текста читателем. Следовательно, для получения достоверной картины необходимо перейти от одномерных подходов к многофакторному анализу, направленному на выявление целого комплекса взаимосвязанных лингвистических характеристик [6].

Цель настоящей статьи — идентификация и верификация набора объективных корпусных предикторов, которые надёжно отражают лексико-синтаксическую сложность текстов, предназначенных для изучающих русский язык. На материале учебных текстов, предварительно размеченных по уровням Общеевропейских компетенций владения иностранным языком (CEFR), с помощью статистической среды R и инструментов компьютерной лингвистики (quanteda, udpipe) проводится анализ четырёх ключевых групп параметров: лексического разнообразия (характеристика K Юла), длины синтаксических единиц (средняя длина предложения), сложности внутренней структуры предложения (средняя дистанция синтаксической зависимости) и стилистических особенностей (соотношение существительных и глаголов). Полученные результаты призваны не только способствовать созданию более качественно градуированных учебных материалов, но и заложить эмпирическую основу для последующих экспериментальных исследований в области психолингвистики восприятия текста.

Методы и принципы исследования

2.1. Материалы

Эмпирической базой для настоящего исследования послужил корпус учебных текстов, собранных из различных пособий по русскому языку как иностранному. Ключевым преимуществом данного корпуса является то, что все тексты в нём предварительно классифицированы по уровням владения языком в соответствии с Общеевропейскими компетенциями (CEFR), охватывая диапазон от A1 до C2. Для анализа были отобраны преимущественно тексты, предназначенные для чтения (текст для чтения), так как именно они наиболее полно отражают комплексную лексико-синтаксическую структуру, с которой сталкивается обучающийся при работе с аутентичными или адаптированными материалами. Тексты упражнений, имеющие специфическую дидактическую направленность, были исключены из основного анализа для обеспечения гомогенности выборки.

2.2. Методы анализа

Анализ данных проводился в среде статистического программирования R, которая предоставляет широкие возможности для обработки и анализа текстовых данных. Для решения поставленных задач были использованы два ключевых пакета компьютерной лингвистики. Пакет quanteda применялся для быстрой и эффективной токенизации текстов и расчёта метрик лексического разнообразия. Для более глубокого морфологического и синтаксического анализа, включая частеречную разметку и построение деревьев зависимостей, был использован пакет udpipe с предварительно обученной моделью для русского языка russian-syntagrus.

2.3. Анализируемые параметры

На основе теоретических предпосылок о многомерности языковой сложности для анализа были выбраны четыре предиктора, отражающие различные аспекты текста [10]:

Характеристика K Юла: метрика лексического разнообразия, измеряющая степень повторяемости слов. В отличие от базового показателя TTR, данный индекс практически нечувствителен к длине текста, что делает его надёжным инструментом для сравнения выборок разного объёма [4].

Средняя длина предложения: классический параметр, отражающий сложность на макросинтаксическом уровне. Увеличение длины предложения, как правило, коррелирует с усложнением его структуры и повышением когнитивной нагрузки на читателя.

Средняя дистанция синтаксической зависимости: продвинутый параметр, оценивающий сложность на микросинтаксическом уровне. Он измеряет среднее расстояние (в словах) между синтаксически связанными элементами предложения. Большая дистанция требует от читателя удерживать в рабочей памяти больше информации, что напрямую связано с воспринимаемой трудностью текста.

Соотношение существительных и глаголов: стилистический параметр, указывающий на степень номинализации текста. Преобладание существительных над глаголами характерно для более абстрактного, формального и информационно плотного стиля, свойственного текстам высоких уровней сложности.

2.4. Статистическая обработка

Для верификации выбранных предикторов использовались два непараметрических статистических теста. Критерий Краскела-Уоллиса применялся для определения наличия статистически значимых различий в значениях каждого параметра между всеми уровнями CEFR. Для оценки силы и направления тенденции (т.е. последовательного роста или снижения параметра с повышением уровня) рассчитывался коэффициент ранговой корреляции Спирмена (ρ). Уровень статистической значимости был принят за $p < 0,05$.

Основные результаты

Статистический анализ данных, проведённый на материале учебного корпуса, подтвердил гипотезу о многомерности текстовой сложности и позволил верифицировать выбранные параметры в качестве надёжных предикторов уровня владения языком. Всестороннее рассмотрение каждого предиктора выявило их различный вклад и специфику, что даёт комплексное представление о том, как меняется структура текста по мере повышения его сложности. Ниже представлены результаты анализа по каждому из четырёх параметров.

3.1. Лексическое разнообразие (Характеристика K Юла)

Анализ показал, что характеристика K Юла является высокочувствительным индикатором лексической сложности. Результаты теста Краскела-Уоллиса продемонстрировали наличие статистически высокозначимых различий в значениях K между текстами разных уровней ($\chi^2(5) = 80,19$, $p < 0,001$). Корреляционный анализ по Спирмену выявил сильную, статистически значимую отрицательную связь между уровнем текста и значением K ($\rho = -0,55$, $p < 0,001$). Это означает, что с повышением уровня от A1 до C2 лексическая повторяемость текстов закономерно и значительно снижается, следовательно, их лексическое разнообразие возрастает.

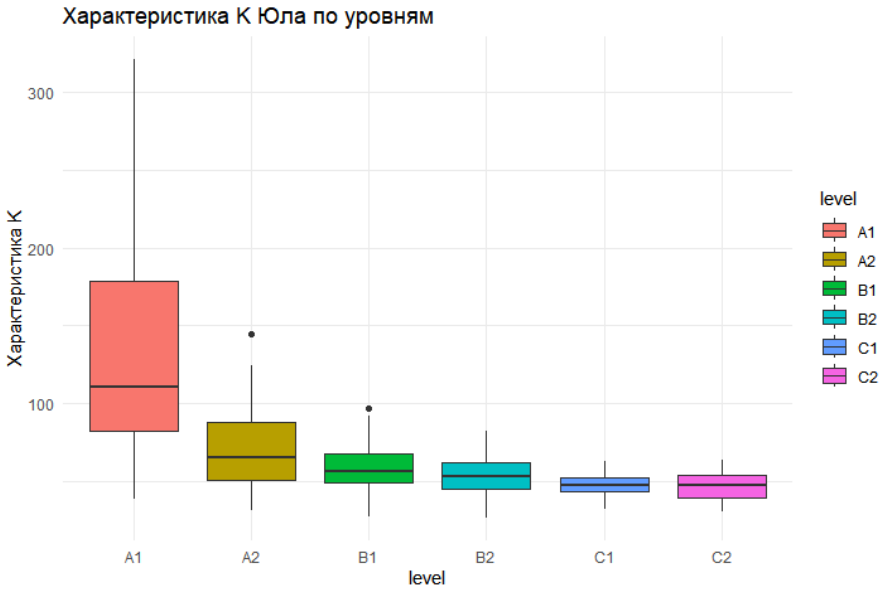

Рисунок 1 - Распределение Характеристики K Юла по уровням владения языком
DOI: https://doi.org/10.60797/IRJ.2026.163.54.1

3.2. Сложность на синтаксическом уровне

Два параметра, оценивающие синтаксическую сложность с разных сторон, показали себя как наиболее сильные предикторы [3]. Средняя длина предложения продемонстрировала положительную корреляцию с уровнем текста ($\rho = 0{,}70$, $p < 0{,}001$), что подтверждает гипотезу о том, что тексты более высоких уровней систематически используют более длинные и сложносочинённые/сложноподчинённые предложения ($\chi^2(5) = 120{,}27$, $p < 0{,}001$).

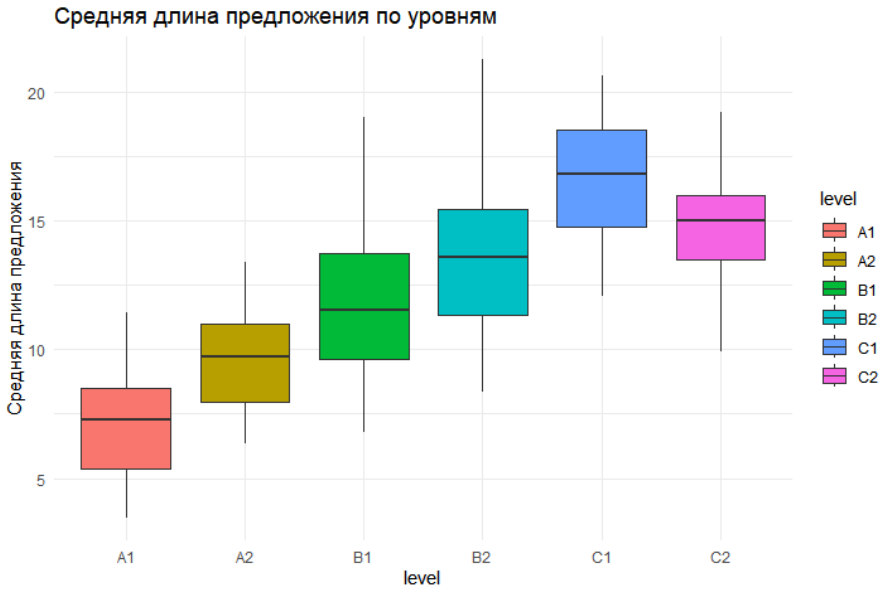

Рисунок 2 - Распределение средней длины предложения по уровням владения языком
DOI: https://doi.org/10.60797/IRJ.2026.163.54.2

Аналогичную сильную положительную тенденцию показала и средняя дистанция синтаксической зависимости ($\rho = 0,63$, $p < 0,001$), различия между уровнями также были высокозначимы ($\chi^2(5) = 95,92$, $p < 0,001$). Этот результат особенно важен, так как он свидетельствует не просто о внешнем увеличении длины предложений, но и об усложнении их внутренней структуры. Возрастающая дистанция между синтаксически связанными словами напрямую указывает на повышение когнитивной нагрузки, необходимой для успешной обработки текста читателем.

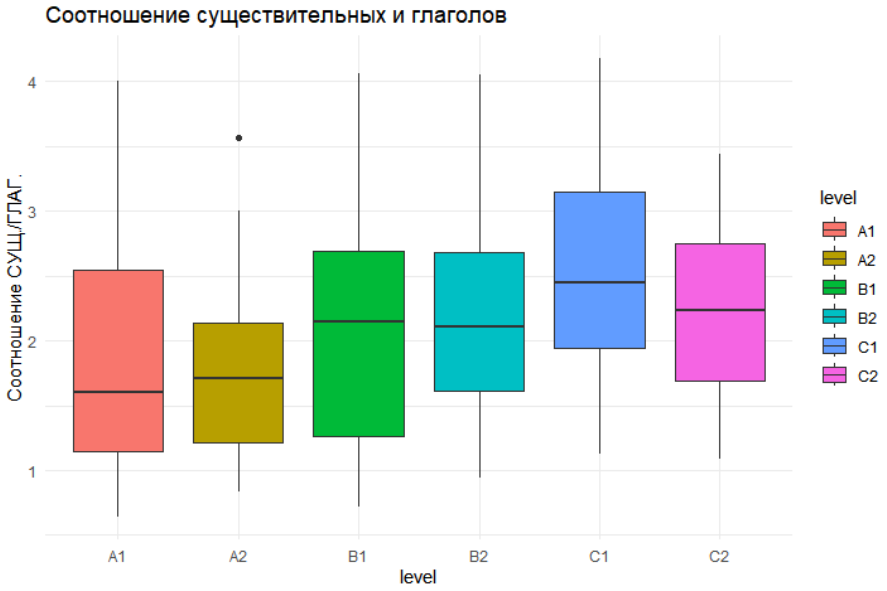

Рисунок 3 - Распределение соотношения существительных и глаголов по уровням владения языком
DOI: https://doi.org/10.60797/IRJ.2026.163.54.3

3.3. Стилистические особенности текста

Анализ соотношения существительных и глаголов выявил более тонкую, но статистически значимую закономерность. Тест Краскела-Уоллиса показал, что общие различия между группами не достигают порога статистической значимости ($\chi^2(5) = 9,94$, $p = 0,077$). Однако корреляционный анализ по Спирмену обнаружил слабую, но статистически высокозначимую положительную тенденцию ($\rho = 0,17$, $p < 0,01$). Это говорит о том, что, несмотря на незначительные колебания, по мере роста уровня сложности наблюдается системный сдвиг в сторону номинализации — увеличения доли существительных по отношению к глаголам. Данный сдвиг отражает переход к более абстрактному, формальному и информационно насыщенному стилю изложения в текстах высоких уровней.

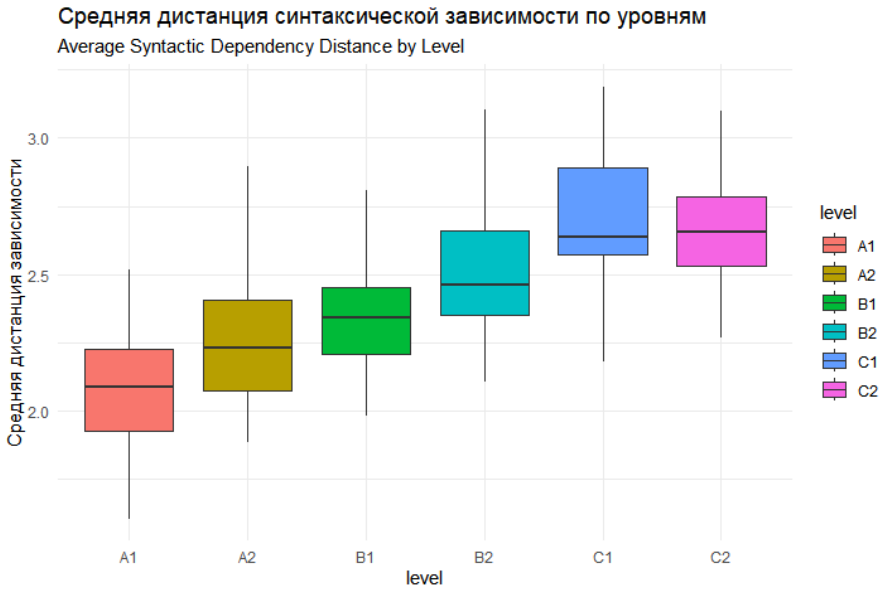

Рисунок 4 - Распределение средней дистанции синтаксической зависимости по уровням владения языком
DOI: https://doi.org/10.60797/IRJ.2026.163.54.4

Для наглядности, сводные результаты статистического анализа по всем четырём предикторам представлены в Таблице 1.

Таблица 1 - Сводные результаты статистического анализа предикторов

DOI: https://doi.org/10.60797/IRJ.2026.163.54.5

Предиктор (Predictor)	Критерий Краскела-Уоллиса (Kruskal-Wallis)		Корреляция Спирмена (Spearman's Corr.)	
Характеристика К Юла	$\chi^2(5) = 80,19$	$p < 0,001$	$\rho = -0,55$	$p < 0,001$
Средняя длина предложения	$\chi^2(5) = 120,27$	$p < 0,001$	$\rho = 0,70$	$p < 0,001$
Средняя дистанция синтаксической зависимости	$\chi^2(5) = 95,92$	$p < 0,001$	$\rho = 0,63$	$p < 0,001$
Соотношение существительных и глаголов	$\chi^2(5) = 9,94$	$p = 0,077$	$\rho = 0,17$	$p < 0,01$

Обсуждение

Проведённое исследование позволило выявить и верифицировать четыре объективных параметра, которые в совокупности дают многомерное представление о лексико-синтаксической сложности учебных текстов по РКИ. Полученные результаты не только подтверждают существующие теоретические представления, но и вносят важные уточнения, имеющие как теоретическое, так и практическое значение [7].

Ключевым выводом является то, что сложность текста для иностранного учащегося определяется не одним, а целым комплексом факторов, среди которых доминирующую роль играют предикторы синтаксического уровня. Средняя длина предложения и, что особенно важно, средняя дистанция синтаксической зависимости показали себя как наиболее сильные и надёжные индикаторы [2]. Это подчёркивает, что по мере повышения уровня владения языком основной вызов для учащегося заключается не столько в освоении новой лексики, сколько в способности обрабатывать всё более сложные синтаксические конструкции. Увеличение дистанции зависимости напрямую свидетельствует о росте когнитивной нагрузки, так как требует от читателя удерживать в рабочей памяти больше грамматических связей для успешного декодирования смысла предложения [8].

Наряду с синтаксисом, лексическое разнообразие, измеренное с помощью характеристики К Юла, также является фундаментальным параметром сложности. Сильная отрицательная корреляция данного показателя с уровнем текста доказывает, что переход на новый уровень владения языком неразрывно связан с расширением словарного запаса и уменьшением доли повторяющихся лексем. Наконец, соотношение существительных и глаголов, хоть и показало более слабую тенденцию, вскрыло важный стилистический аспект сложности. Системный сдвиг в сторону номинализации в текстах высоких уровней отражает переход от нарративного, ориентированного на действия стиля к более

абстрактному, дескриптивному и информационно плотному изложению, характерному для академической и научной речи.

Заключение

В заключение, данное исследование успешно идентифицировало и верифицировало набор из четырёх корпусных предикторов, эффективно отражающих многоаспектную природу сложности текста в РКИ. Установлено, что сложность определяется как лексическими (разнообразие), так и синтаксическими (длина и внутренняя структура предложений) и стилистическими (степень номинализации) факторами. Практическая значимость работы заключается в том, что предложенные метрики могут быть использованы авторами учебников и преподавателями для объективной оценки и градуировки учебных материалов. Теоретический вклад состоит в количественном подтверждении и детализации структуры текстовой сложности. Результаты данного корпусного анализа закладывают прочную эмпирическую основу для следующего этапа исследований — проведения психолингвистических экспериментов, направленных на изучение того, как эти объективные текстовые параметры влияют на процессы восприятия и когнитивную нагрузку у реальных читателей, изучающих русский язык.

Конфликт интересов

Не указан.

Рецензия

Все статьи проходят рецензирование. Но рецензент или автор статьи предпочли не публиковать рецензию к этой статье в открытом доступе. Рецензия может быть предоставлена компетентным органам по запросу.

Conflict of Interest

None declared.

Review

All articles are peer-reviewed. But the reviewer or the author of the article chose not to publish a review of this article in the public domain. The review can be provided to the competent authorities upon request.

Список литературы / References

1. Kupriyanov R.V. Cognitive complexity measures for educational texts: Empirical validation of linguistic parameters / R.V. Kupriyanov, O.V. Bukach, O.I. Aleksandrova // Russian Journal of Linguistics. — 2023. — Vol. 27. — № 3. — P. 641–662. — DOI: 10.22363/2687-0088-35817.

2. Галявиева Л.Ш. Сложность учебных текстов как функция терминологической плотности / Л.Ш. Галявиева, А.Т. Хусаинова, М.И. Солнышкина // Ученые записки национального общества прикладной лингвистики. — 2023. — № 2 (42). — С. 33–50.

3. Калугина Е.Н. Лингвистическая экспертиза анонимного текста: методы и подходы / Е.Н. Калугина, А.В. Волкогонова // Вестник АПК Ставрополья. — 2016. — № 2 (22). — С. 166–168.

4. Колмогорова А.В. Лексико-грамматические маркеры эмоций в качестве параметров для сентимент-анализа русскоязычных интернет-текстов / А.В. Колмогорова, Л.А. Вдовина // Вестник Пермского университета. Российская и зарубежная филология. — 2019. — Т. 11. — № 3. — С. 38–46. — DOI: 10.17072/2073-6681-2019-3-38-46.

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

6. Вахрушева А.Я. Лингвистическая сложность учебных текстов / А.Я. Вахрушева, М.И. Солнышкина, Р.В. Куприянов [и др.] // Вопросы журналистики, педагогики, языкознания. — 2021. — Т. 40. — № 1. — С. 89–99. — DOI: 10.52575/2712-7451-2021-40-1-89-99.

7. Пешкова Н.П. Лингвистические характеристики текстов как основания для их классификации (на материале научных, технических, учебных текстов): дис. ... канд. филол. наук / Пешкова Наталья Петровна. — Москва, 1987. — 272 с.

8. Резанова З.И.О выборе признаков текста, релевантных в автороведческой экспертной деятельности / З.И. Резанова, А.С. Романов, Р.В. Мещеряков // Вестник Томского государственного университета. Филология. — 2013. — № 6 (26). — С. 38–52.

9. Солнышкина М.И. Сложность текста: этапы изучения в отечественном прикладном языкознании / М.И. Солнышкина, А.С. Кисельников // Вестник Томского государственного университета. Филология. — 2015. — № 6 (38). — С. 86–99.

10. Толпегин П.В. Информационные технологии анализа русских естественно-языковых текстов. Часть I / П.В. Толпегин // Информационные технологии. — 2006. — № 8. — С. 41–50.

Список литературы на английском языке / References in English

1. Kupriyanov R.V. Cognitive complexity measures for educational texts: Empirical validation of linguistic parameters / R.V. Kupriyanov, O.V. Bukach, O.I. Aleksandrova // Russian Journal of Linguistics. — 2023. — Vol. 27. — № 3. — P. 641–662. — DOI: 10.22363/2687-0088-35817.

2. Galyavieva L.Sh. Slozhnost uchebnikh tekstov kak funktsiya terminologicheskoi plotnosti [The complexity of educational texts as a function of terminological density] / L.Sh. Galyavieva, A.T. Khusainova, M.I. Solnishkina // Uchenie zapiski natsionalnogo obshchestva prikladnoi lingvistiki [Scientific Notes of the National Society of Applied Linguistics]. — 2023. — № 2 (42). — P. 33–50. [in Russian]

3. Kalugina Ye.N. Lingvisticheskaya ekspertiza anonimnogo teksta: metodi i podkhodi [Linguistic expertise of an anonymous text: methods and approaches] / Ye.N. Kalugina, A.V. Volkogonova // Vestnik APK Stavropolya [Bulletin of the Agro-Industrial Complex of Stavropol]. — 2016. — № 2 (22). — P. 166–168. [in Russian]

4. Kolmogorova A.V. Leksiko-grammaticheskie markeri emotsii v kachestve parametrov dlya sentiment-analiza russkoyazichnikh internet-tekstov [Lexico-grammatical markers of emotions as parameters for sentiment analysis of Russian-language Internet texts] / A.V. Kolmogorova, L.A. Vdovina // Vestnik Permskogo universiteta. Rossiiskaya i zarubezhnaya filologiya [Perm University Bulletin. Russian and Foreign Philology]. — 2019. — Vol. 11. — № 3. — P. 38–46. — DOI: 10.17072/2073-6681-2019-3-38-46. [in Russian]

5. Kupriyanov R.V. Parametricheskaya taksonomiya uchebnikh tekstov [Parametric taxonomy of educational texts] / R.V. Kupriyanov, M.I. Solnishkina, P.A. Lekhnitskaya // Vestnik Volgogradskogo gosudarstvennogo universiteta. Seriya 2: Yazikoznanie [Bulletin of Volgograd State University. Series 2: Linguistics]. — 2023. — Vol. 22. — № 6. — P. 80–94. — DOI: 10.15688/jvolsu2.2023.6.6. [in Russian]

6. Vakhrusheva A.Ya. Lingvisticheskaya slozhnost uchebnikh tekstov [Linguistic complexity of educational texts] / A.Ya. Vakhrusheva, M.I. Solnishkina, R.V. Kupriyanov [et al.] // Voprosi zhurnalistiki, pedagogiki, yazikoznaniya [Issues of Journalism, Pedagogy, and Linguistics]. — 2021. — Vol. 40. — № 1. — P. 89–99. — DOI: 10.52575/2712-7451-2021-40-1-89-99. [in Russian]

7. Peshkova N.P. Lingvisticheskie kharakteristiki tekstov kak osnovaniya dlya ikh klassifikatsii (na materiale nauchnikh, tekhnicheskikh, uchebnikh tekstov) [Linguistic characteristics of texts as a basis for their classification (based on scientific, technical, and educational texts)]: dis. ... of PhD in Philological Sciences / Peshkova Natalya Petrovna. — Moscow, 1987. — 272 p. [in Russian]

8. Rezanova Z.I.O vibore priznakov teksta, relevantnikh v avtorovedcheskoi ekspertnoi deyatelnosti [On the selection of text features relevant to forensic authorship analysis] / Z.I. Rezanova, A.S. Romanov, R.V. Meshcheryakov // Vestnik Tomskogo gosudarstvennogo universiteta. Filologiya [Bulletin of Tomsk State University. Philology]. — 2013. — № 6 (26). — P. 38–52. [in Russian]

9. Solnishkina M.I. Slozhnost teksta: etapi izucheniya v otechestvennom prikladnom yazikoznanii [Text complexity: Stages of research in Russian applied linguistics] / M.I. Solnishkina, A.S. Kiselnikov // Vestnik Tomskogo gosudarstvennogo universiteta. Filologiya [Bulletin of Tomsk State University. Philology]. — 2015. — № 6 (38). — P. 86–99. [in Russian]

10. Tolpegin P.V. Informatsionnie tekhnologii analiza russkikh yestestvenno-yazikovikh tekstov. Chast I [Information technologies for the analysis of Russian natural language texts. Part I] / P.V. Tolpegin // Informatsionnie tekhnologii [Information Technologies]. — 2006. — № 8. — P. 41–50. [in Russian]