


DOI: 10.22363/2618-8163-2024-22-4-501-517

EDN: AMYSNF

Вступительная статья

Подходы и инструменты лингвистического профилирования текста на русском языке

М.И. Солнышкина¹, В.Д. Соловьев¹, Ю.Н. Эбзеева²¹Казанский (Приволжский) федеральный университет, Казань, Российская Федерация²Российский университет дружбы народов, Москва, Российская Федерация mesoln@yandex.ru

Аннотация. Развитие подходов и усовершенствование инструментов оценки лингвистической и когнитивной сложности учебного текста востребовано как в науке, так и практике обучения. Особую значимость прогнозирование трудностей восприятия и понимания, а также ранжирование текстов по классам, т.е. количеству лет формального обучения, или уровням владения языком (A1–C2) имеет в системе образования. Цель исследования — продемонстрировать, каким образом современные методологии, алгоритмы и инструменты аналитики текстов на русском языке реализованы в автоматическом анализаторе RuLingva, а также представить статьи тематического выпуска, посвященного комплексному анализу учебников по русскому языку для российских и белорусских школ. Показано, что современная парадигма дискурсивной комплексологии опирается на разработанные в российском языкознании методы стилостатистики, позволяющие выявлять функциональные характеристики языковых единиц и осуществлять их верификацию на материале больших языковых данных. Функционирующие на портале RuLingva сервисы предназначены для преподавателей и исследователей и позволяют в автоматическом режиме не только осуществлять аналитику учебного текста, но и прогнозировать его целевую аудиторию на основании данных о читабельности, лексическом разнообразии, абстрактности, частотности, терминологической плотности. В режиме «Русский как иностранный» RuLingva выгружает из текста списки слов, соответствующие каждому из уровней владения языком, и оценивает долю каждого из них, предоставляя таким образом материал для пред- и посттекстовой работы преподавателя. Алгоритм функционирования RuLingva разработан на основе типологии учебных текстов и имеет в качестве перспективы создание функционала оценки вербального интеллекта и читательской грамотности обучающегося. Перспектива развития RuLingva связана с расширением спектра предикторов сложности и внедрением функции автоматического определения предметной области учебного текста. Оба направления планируется реализовать при помощи нейронных сетей и созданных на их основе классификационных моделей, а также на базе «типологических паспортов» учебных текстов различной сложности и тематической направленности.

Ключевые слова: лингвистический анализ, текстовый профайлер, RuLingva, сложность текста, учебный текст, типологический паспорт текста, предикторы сложности

Вклад авторов: Солнышкина М.И. — разработка концепции, проведение исследования, подготовка и редактирование текста; Соловьев В.Д. — разработка методологии, проведение исследования; Эбзеева Ю.Н. — проведение исследования, утверждение окончательного варианта статьи.

© Солнышкина М.И., Соловьев В.Д., Эбзеева Ю.Н., 2024

This work is licensed under a Creative Commons Attribution 4.0 International License
<https://creativecommons.org/licenses/by-nc/4.0/legalcode>

Финансирование. Работа выполнена за счет средств Программы стратегического академического лидерства Казанского (Приволжского) федерального университета (ПРИОРИТЕТ–2030). Работа выполнена в рамках проекта № 050738-0-000 системы грантовой поддержки научных проектов РУДН.

Конфликт интересов: Авторы заявляют об отсутствии конфликта интересов.

История статьи: поступила в редакцию 02.07.2024; принята к печати 18.08.2024.

Для цитирования: Солнышкина М.И., Соловьев В.Д., Эбзеева Ю.Н. Подходы и инструменты лингвистического профилирования текста на русском языке // Русистика. 2024. Т. 22. № 4. С. 501–517. <http://doi.org/10.22363/2618-8163-2024-22-4-501-517>

Введение

Смена научных парадигм современности и активные интегративные процессы поставили перед лингвистами новые задачи, решение которых предполагает, с одной стороны, включение текста в широкий исторический и дискурсивный контексты, а с другой — изучение процессов восприятия, понимания, воспроизведения и генерирования текстов. Сам факт обращения к дискурсивным аспектам текста и когнитивным характеристикам носителя языка в значительной степени способствовал расширению границ лингвистики, привлечению данных других наук, использованию не одного, а нескольких подходов для анализа данных.

В списке наиболее актуальных задач аналитики текста ученые справедливо выделяют классификацию текстов, анализ тональности, извлечение ключевых слов, «диагностику» типов отношений между единицами текста, определение семантических ролей, анализ аргументов и дискурсивных структур, структурирование больших языковых данных и т.д. (Kuznetsova, 2015; Young et al., 2018). Особую сложность имеют задачи разрешения (снятия) омонимии/полисемии, а также тематическое моделирование текста (Sakhovskiy et al., 2020). Отдельную, весьма интересную проблему представляет создание авторских лингвопрофилей, включающих совокупность количественных характеристик, свойственных конкретному автору (Михеев, Эрлих, 2018). Решение столь многоаспектных задач предполагает доступ к большим коллекциям текстов различных форм, регистров, типов и жанров, а также применение инструментов автоматического анализа.

Во все времена, ставя перед собой исследовательские цели, ученый выбирает подход и соответствующие ему методы, собирает данные и подбирает соответствующий инструментарий. Отличие современного периода состоит в возможности выбрать не один, а несколько подходов, включая междисциплинарный, а также использовать большие репрезентативные электронные корпуса, в т.ч. созданные в предыдущие десятилетия. При этом предполагается, что корпус языковых данных содержит не только метаразметку, но и детализированное описание каждого текста в коллекции, его «типологический паспорт», представляющий собой определенным образом упорядоченные данные о количественных характеристиках текста, его «лингвостатистический профиль» (Virk et al., 2020). «Профиль» текста содержит данные о частотности, дистрибуции лингвистических параметров текста и взаимосвязи лингвистических характеристик текстов. Последнее

означает, что тексты разных типов, жанров и регистров «профилируются» на основании списков характерных для них признаков и диапазонов референсных значений этих признаков. Именно последние выполняют предиктивную и дискриминантную функции, т.е. способны, с одной стороны, определять принадлежность, а с другой стороны, дифференцировать тексты как элементы определенных типов, жанров и регистров. Профилирование текстов и создание матриц языковых данных завершают этап сбора и организации корпуса.

Особую значимость для подготовки профилей текста имеют, с одной стороны, общенаучные подходы текстовой аналитики и конкретные алгоритмы, обеспечивающие основу автоматизации лингвистического анализа, а с другой стороны, инструменты — текстовые анализаторы, способные осуществлять автоматическую оценку значений параметров текста (Лукашевич, Добров, 2015; Наместников, Пирогова, Филиппов, 2021; Соловьев, Солнышкина, Макнамара, 2022; Колмогорова, Колмогорова, Куликова, 2024). Эффективная автоматизация «профилирования» текстов, а также общих, весьма трудоемких механических задач лингвистического анализа текстов на русском языке является одним из условий перехода русистики в частности и российской лингвистики в целом на качественно новый уровень.

Современные ученые относят «профили» текста к так называемым «ресурсам» языка, а сами языки делят на высокоресурсные и низкоресурсные в зависимости от достаточности данных, которые можно использовать для машинного обучения или других типов обработки (Chang et al., 2023). Аналогом в лингвистической типологии является противопоставление хорошо и недостаточно описанных языков. К первым относят, например, английский и немецкий языки. Русский язык в этом отношении квалифицируется учеными по-разному: низкоресурсный (Valeev et al., 2019), «относительно высокоресурсный» (Karakanta, Dehdari, van Genabith, 2018). Однако ученые по-прежнему пишут о необходимости создания электронных баз данных и инструментов, разработки и совершенствования подходов текстовой аналитики для русского языка (Toldova et al., 2015).

Все вышесказанное определяет актуальность обращения к проблематике принципов лингвистического профилирования и автоматизации анализа языковых данных. **Цель исследования** — описание теоретических подходов и инструментов лингвистического профилирования текстов на русском языке. Вторая часть работы содержит краткую характеристику статей представляемого тематического выпуска.

Лингвистическое профилирование как объект теоретической и прикладной лингвистики

Применение методов точных наук и классификационных математических моделей для описания текста далеко не ново. За работами Ф. де Соссюра в начале XX в. (1922, первое издание вышло в 1916) (Соссюр, 1977) последовали междисциплинарные исследования К. Шеннона и У. Уивера (1949), заложившие основы современных методов количественной лингвистики. Важным в данном контексте является подход к языковым явлениям

как стереотипным, т.е. способным маркировать определенные явления и характеристики как свойственные или чуждые какому-либо типу объектов (Lipmann, 1922). При таком подходе установление стереотипов позволяет типизировать тексты и определить свойственные каждому типу параметры. Одна из первых гипотез относительно статистических различий дискурсов принадлежит В.В. Виноградову, который еще в 1938 г. писал: «Повидимому, в разных стилях книжной и разговорной речи, а также в разных стилях и жанрах художественной литературы частота употребления разных типов слов различна. Но, к сожалению, этот вопрос пока находится лишь в подготовительной стадии обследования материала» (Виноградов, 1938: 356). За период 1930–1960 гг. российская и зарубежная лингвистика достигла больших успехов, так что лингвистика 1960-х была названа «самой точной из всех гуманитарных наук», в первую очередь, благодаря четко сформулированной и высоко формализованной теории Н. Хомского, применимой не только к естественным языкам, но даже и к языкам программирования. Причиной этому стал выбор в качестве основных объектов исследования всеобщих принципов и моделей построения смысловых констант (Кронгауз, 2009) и синтаксических конструкций. Построение формальных моделей основывалось, с одной стороны, на идентификации языковых единиц как компонентов языковой системы, организованной в соответствии с универсальными когнитивными принципами, а с другой — на вероятности появления языковой единицы в тексте определенного типа (Зиндер, Строева, 1968).

Важный вклад в развитие текстовой аналитики для русского языка внесли Б.Н. Головин (Головин, 1971) и последователи его научной школы, широко использовавшие количественные методы для описания и анализа функциональных стилей. «Основная заслуга горьковского центра, — пишут М.А. Кормилицина и О.Б. Сиротина, — состояла в разработке статистической методики исследования речевых фактов. Эти методы основаны на наблюдении, что существует сильная корреляция между семантическими и дистрибутивными свойствами языковых единиц» (Кормилицина, Сиротина, 2013: 103).

Особую значимость в данной связи имеет стилостатистический (качественно-количественный) метод, разработанный в отечественной школе в конце прошлого века и предполагающий (1) «семанτικο-стилистическую квалификацию» языковых единиц, т.е. выявление наличия у них специфических функциональных характеристик, а также (2) верификацию данных характеристик при помощи методов математической статистики (см. Кожина, 1989). Метод, получивший особую популярность в 1980-х гг. благодаря разработкам формального языка, используется и в современной квантитативной лингвистике для оценки роли фактора на конструктор (Сердобольская, Толдова, 2005).

Рубеж нового тысячелетия ознаменовался не только становлением компьютерной и корпусной лингвистики (Solnyshkina et al., 2022), но и появлением многочисленных подходов к формализованной обработке больших языковых данных. А сам анализ текста претерпел существенную эволюцию благодаря революции в области вычислительных технологий и баз данных. Количественные методологии, включая машинное обучение, сделали извле-

чение информации из массивов текстовых данных доступным, а следовательно, позволили подойти к подтверждению/опровержению сделанных ранее гипотез о системности языковых фактов в текстах отдельных жанров, регистров, типов.

Сами изменения, происходящие в науке, можно описать как взаимодействующие этапы, нацеленные на разработку моделей трех типов: на основе признаков, обучения репрезентации и генеративные модели. Современные автоматические анализаторы текстов на русском языке продолжают завоевывать популярность. Открытые платформенные решения предлагают текстовый профайлер Текстометр, активно используемый отечественными словесниками (Лапошина, Лебедева, 2021), и анализатор И. Бегтина «Оценка читаемости текста»¹, на котором установлены 5 формул читабельности. Проект И. Бегтина стал первым онлайн-сервером со встроенными формулами читабельности, при этом, к сожалению, за изучением алгоритма расчетов и формулами разработчики предлагают обратиться к англоязычным сайтам Википедии², что в целом не дает возможности оценить валидность используемых формул.

Текстовый профайлер и анализатор сложности текстов на русском языке **RuLingva** был разработан в рамках проекта Российского научного фонда «Сложность текстов на русском языке»³, направленного на достижение двух основных целей: выявление и описание типологических параметров академических текстов и разработка методов ранжирования текстов по уровням сложности. Оценка и идентификация уровня сложности текстов на RuLingva основаны на выявленных корреляциях параметров текстов и критериев читателей (возраст, образование, словарный запас и т.д.).

Следуя современной традиции в текстовой аналитике, мы используем термины «характеристики текста», «параметры текста», «значения» или «метрики», а также кластеры или группы параметров (которые Д. Макнамара (McNamara et al., 2014) именует brands). Термин «характеристика» обозначает наименование языковой категории (например, лексическое разнообразие), в то время как термин «параметр» предоставляет информацию о способе(ах), при помощи которого(ых) RuLingva оценивает соответствующие характеристики текста. Например, лексическое разнообразие (характеристика текста) оценивается на основе параметра TTR (Type Token Ratio, TTR), т.е. отношения количества словоформ к леммам. Термины «метрики» и «значения» взаимозаменяемы и используются для наименований количественных значений параметра. Например, значение/метрика 166 в строке 1 (рис. 1) указывает на количество словоформ (параметр), при помощи которых оценивается длина текста (характеристика). Группы концептуально

¹ Удобный инструмент проверки текстов. URL: <https://plainrussian.ru/#about> (дата обращения: 02.06.2024).

² Flesch – Kincaid readability tests. URL: https://en.wikipedia.org/wiki/Flesch%E2%80%93Kincaid_readability_tests (дата обращения: 16.06.2024.). https://en.wikipedia.org/wiki/Coleman%E2%80%93Liau_index (дата обращения: 16.06.2024).

³ Карточка проекта фундаментальных и поисковых научных исследований, поддержанного Российским научным фондом. URL: https://rscf.ru/prjcard_int?18-18-00436 (дата обращения: 18.06.2024).

аналогичных параметров текста объединяются в кластеры. Например, дескриптивный или описательный кластер параметров текста, включает длину текста, измеряемую в количестве словоформ, лемм, слогов или предложений.

Ru, Lingva

РКИ

О нас

Исследования

Результаты анализа

№	Параметр	Документ	Абзац	Предложение	Нормализация на диапазон токенов <div>1000</div>
Описательные параметры					
1	Количество словоформ	166	83	15.09	1000
2	Количество лемм	89	44.50	8.09	536.14
3	Количество слогов	408	204	37.09	2457.83
4	Количество предложений	11	5.50	1	66.27
5	Среднее количество слов в предложении	15.09			
6	Среднее количество слогов в слове	2.46			
7	Среднее количество букв в слове	5.38			
8	Односложные слова	36	18	3.27	216.87
9	Двусложные слова	44	22	4	265.06
10	Трехсложные слова	32	16	2.91	192.77
11	Четырехсложные слова	38	19	3.45	228.92

Рис. 1. Интерфейс RuLingva

И с т о ч н и к : RuLingva. URL: <https://rulingva.kpfu.ru/> (дата обращения: 18.06.2024).

Профайлер RuLingva⁴, поддерживаемый исследовательской группой Казанского федерального университета, выполняет автоматические расчеты значений 73 параметров учебных текстов на русском языке. В соответствии с современной традицией в компьютерной лингвистике измерение значений лингвистических параметров осуществляется с различной степенью «грануляции» (см. термин в Paraschiv et al., 2023), т.е. расчетом метрик в составе предложения, абзаца, фрагмента текста определенной длины и всего документа. Пользователь имеет возможность устанавливать нормализацию расчета параметров в зависимости от задач исследования на 100, 200, 1000 словоформ (токенов) (см. рис. 1).

Разработке функционала RuLingva предшествовало создание двух независимых корпусов: Учебного корпуса русского языка (УКРЯ) и Корпуса текстов РКИ (КТРКИ). Объем УКРЯ⁵ составляет 14 млн словоформ, объем КТРКИ – чуть более 500 тыс. словоформ. Оба корпуса созданы на основе принципа достоверности данных, в соответствии с которым в корпус включались только «эталонные» тексты, т.е. тексты, прошедшие профессиональную экспертизу и признанные лучшими текстами в своей области.

Источниками материалов для УКРЯ послужили тексты Федерального государственного образовательного стандарта⁶, а для КТРКИ — тексты, рекомендованные Комиссией по экспертизе тестовых материалов по русскому

⁴ RuLingva. URL: <https://rulingva.kpfu.ru/> (дата обращения: 18.06.2024).

⁵ Свидетельство о государственной регистрации базы данных № 2020622254.

⁶ Федеральный перечень учебников <https://fpu.edu.ru/> (дата обращения: 18.06.2024).

языку как иностранному⁷, Экспертной комиссией Государственной системы тестирования граждан зарубежных стран по русскому языку⁸, а также тексты из Открытого банка заданий ФИПИ⁹. Дифференциальная полнота, а также сбалансированность и репрезентативность УКРЯ, использовавшегося в качестве источника в ряде российских и зарубежных исследований, не вызывает сомнений (Corlatescu et al., 2022, Kupriyanov et al., 2022, Paraschiv et al., 2023). Этот факт позволяет рассматривать RuLingva как весьма ценный инструмент для изучения современного состояния научно-учебного дискурса и профилирования текстов на русском языке.

Оба корпуса — УКРЯ и КТРКИ — являются закрытыми и используются исключительно в исследовательских целях. В открытом доступе находится демонстрационный образец небольшого размера — фрагмент подкорпуса учебных текстов социо-гуманитарного предметного блока, в который вошли случайным образом отобранные тексты российских учебников по обществознанию (CORAT, Corpus of Russian Academic Texts, Корпус учебных текстов на русском языке)¹⁰. В целях сохранения авторских прав последовательность абзацев и предложений в текстах CORAT изменена.

На основании Учебного корпуса русского языка была разработана первая формула читабельности учебных текстов на русском языке:

Индекс Флеша — Кинкейда (SIS) = $208,7 - 2,6 \times \text{ASL} - 39,2 \times \text{ASW}$, где ASL (average sentence length) — средняя длина предложения; ASW (average word length in syllables) — средняя длина слова в слогах (Solovyev, Ivanov, Solnyshkina, 2018). После успешной валидации на текстах гуманитарного, филологического и естественно-научного предметного блоков средней и старшей школы (Gatiyatullina et al., 2020), формула была установлена на сайт RuLingva и применяется для оценки читабельности учебных текстов на русском языке¹¹. Удобство использования формулы состоит в том, что она позволяет ранжировать читабельность учебного текста по годам обучения в школе, т.е. классам. Например, текст, имеющий читабельность 7,62 (Индекс Флеша — Кинкейда (SIS)), ориентирован на школьников 7–8 классов (рис. 2).

Для оценки читабельности текстов художественной прозы на сайте RuLingva также установлена формула читабельности Флеша — Кинкейда, модифицированная И.В. Оборновой для русского языка:

Индекс Флеша — Кинкейда (O) = $206,835 - 1,3 \times \text{ASL} - 60,1 \times \text{ASW}$.

⁷ Приказ «Об утверждении Положения о Комиссии по экспертизе тестовых материалов по русскому языку как иностранному и ее состава» <https://docs.cntd.ru/document/901860364> (дата обращения: 18.06.2024).

⁸ Приказ от 16 февраля 2005 г. № 69 «О создании экспертной комиссии государственной системы тестирования граждан зарубежных стран по русскому языку» <https://normativ.kontur.ru/document?moduleId=1&documentId=85661> (дата обращения: 18.06.2024).

⁹ Экзамен для иностранных граждан и лиц без гражданства. <https://fipi.ru/inostr-exam> (дата обращения: 18.06.2024).

¹⁰ Научно-исследовательская лаборатория «Мультидисциплинарные исследования текста». URL: <https://ifmk.kpfu.ru/laboratory/tekstovaya-analitika/> (дата обращения: 18.06.2024).

¹¹ RuLingva. URL: <https://rulingva.kpfu.ru/> URL: RuLex <https://rulex.kpfu.ru/nlp> (дата обращения: 18.06.2024).

Данная формула разрабатывалась И.В. Оборневой на материалах авторского англо-русского корпуса параллельных текстов художественной литературы, поэтому рекомендована исключительно для оценки читабельности текстов литературной прозы (Оборнева, 2006). Использование формулы И.В. Оборневой для оценки читабельности учебных текстов дает заведомо завышенные результаты (рис. 2) (Kupriyanov et al., 2022).

Параметры читабельности		
12	Индекс Флеша-Кинкейда (SIS)	7.62
13	Индекс Флеша-Кинкейда (O)	12.60

Рис. 2. Параметры читабельности текста на RuLingva

И с т о ч н и к : RuLingva. URL: <https://rulingva.kpfu.ru/> (дата обращения: 18.06.2024).

В дополнение к индексам читабельности RuLingva рассчитывает значения четырех групп параметров: 1) дескриптивных (количество слов, предложений, слогов, лемм и словоформ); 2) морфологических (количество разных частей речи и их категорий); 3) лексических (частотность, абстрактность, количество терминов семи предметных областей, включая филологию, математику, информатику, естествознание, физику, изобразительное искусство, музыку, а также количество уникальных, т.е. неповторяющихся слов); 4) дискурсивных (локальные и глобальные повторы слов).

RuLingva оценивает уровень лексического разнообразия TTR текста, измеряет степень конкретности/абстрактности, частотности и лексической плотности. Автоматизация оценки значения лексического разнообразия, несмотря на кажущуюся простоту, требует особого подхода. Достоверность расчетов данного параметра достигается только на фрагментах от 200 до 1000 слов (Cvrček, Chlumská, 2015), поскольку высокая доля служебных частей речи в текстах большей длины в значительной степени снижает метрики этого параметра. Именно поэтому RuLingva автоматически сегментирует текст на отрывки по 1000 словоформ и расчет среднего значения лексического разнообразия всего документа производит на данных каждого из отрывков.

Индекс абстрактности/конкретности текста, установленный на RuLingva, рассчитывается на основе данных, полученных в ходе выполнения научного проекта РФФИ (Solovyev et al., 2022)¹². Базы данных о степени абстрактности/конкретности сформировали на экспериментальных данных, полученных при помощи краудсорсинга в Интернете от участников-носителей языка, а позднее на их основе создали три версии словаря абстрактной лексики: (1) словарь, содержащий 22 тыс. слов, построенный на технологии глубокого обучения по модели BERT; (2) словарь: 64 тыс. слов, построен по технологии word2vec; (3) словарь: 88 тыс. словоформ, создан на базе корпуса Google Books Ngram (Solovyev et al., 2022).

Для текстов РКИ на сайте Rulingva рассчитываются доли лексики от A1 до C2, а также доля слов, отсутствующих в лексических минимумах (рис. 3).

¹² OpenLab «Квантитативная лингвистика». URL: <https://kpfu.ru/tehnologiya-sozdaniya-semanticeskikh-elektronnyh.html> (дата обращения: 17.06.2024).

53	Доля слов уровня А1	96	48	8.73	57.83
54	Доля слов уровня А2	21	10.50	1.91	12.65
55	Доля слов уровня В1	14	7	1.27	8.43
56	Доля слов уровня В2	27	13.50	2.45	16.27
57	Доля слов уровня С1	4	2	0.36	2.41
58	Доля слов уровня С2	0	0	0	0

Рис. 3. Лексический анализ текста РКИ на RuLingva
И с т о ч н и к : RuLingva. URL: <https://rulingva.kpfu.ru/> (дата обращения: 18.06.2024).

RuLingva предлагает результаты анализа частотности лексики (рис. 4), классифицируя все слова в исследуемом документе по группам от А1 до С2 на основе данных об их частотности в Национальном корпусе русского языка (Ляшевская, Шаров, 2009). Сервис определяет долю слов каждого из уровней, а также слов, отсутствующих в лексических минимумах, и позволяет выгружать списки слов, давая преподавателю материал для пред- и посттекстовой работы.

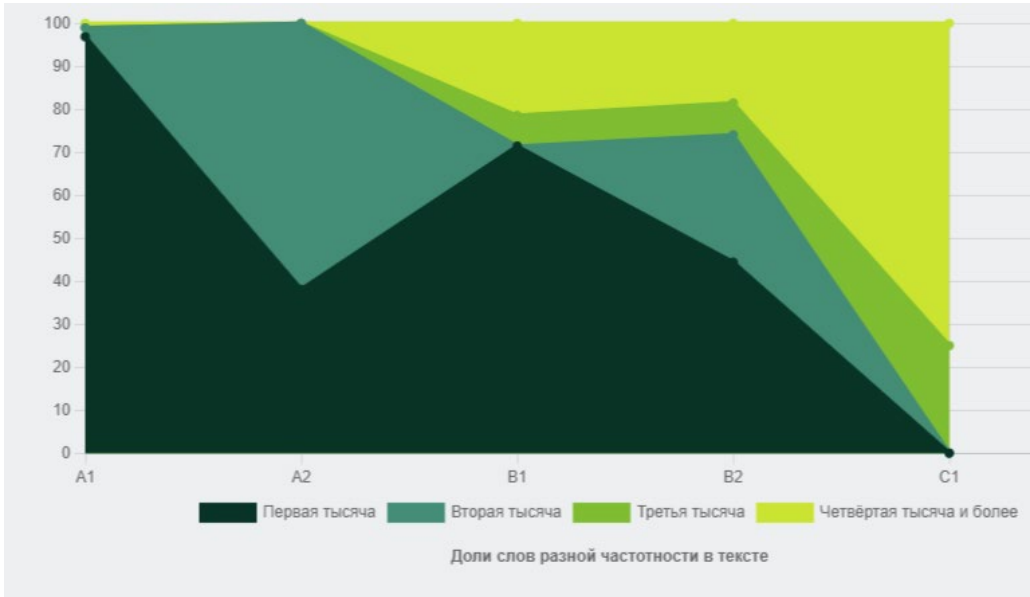


Рис. 4. Частотный анализ лексики в тексте РКИ на RuLingva
И с т о ч н и к : RuLingva. URL: <https://rulingva.kpfu.ru/> (дата обращения: 18.06.2024).

Исследователям, нацеленным на анализ больших объемов данных, RuLingva предлагает пакетную обработку, позволяющую загружать несколько файлов для параллельного анализа. Отчет выгружается с подробным описанием результатов аналитического процесса в формате таблицы Excel.

Предиктивная сила представленных параметров в качестве уровня предикторов сложности и дискриминантов предметных областей доказана в ряде исследований (Лапошина и др., 2019; Blinova, Tarasov, 2022; Dmitrieva, Laposhina, Lebedeva, 2021; Morozov, Glazkova, Iomdin, 2022; Lyashevskaya, Panteleva, Vinogradova, 2021).

В полном соответствии с парадигмой современной количественной лингвистики алгоритм исследования на платформе включает следующие этапы:

1) предобработка корпуса, предполагающая стандартизированные процедуры удаления из текста знаков других семиотических систем в целях сохранения целостности и чистоты входного текста;

2) создание матрицы значений параметров анализируемых, т.е. загружаемых в RuLingva, текстов и ее последующая выгрузка в виде Excel таблицы;

3) расчеты усредненных показателей каждого из параметров и выявление референсных диапазонов, т.е. переменных исследуемых характеристик текста;

4) обобщение и выявление универсальных статистически значимых закономерностей.

Краткий обзор статей выпуска

Данный тематический выпуск включает блок работ, посвященных учебникам по русскому языку и литературе и рассматривающих их в различных аспектах. Две статьи посвящены сопоставлению современных учебников с учебниками советской эпохи.

В открывающей выпуск статье М.И. Андреевой, Р.Р. Замалетдинова, А.С. Борисовой «Предикативная сила лексических параметров: оценка сложности текста в учебниках по русскому языку для 5–7 классов» рассматриваются лингвистические параметры, отражающие сложность текста. В первой части статьи подробно описана методика создания необходимого корпуса учебников. Важно, что удалось подобрать линейки учебников разных классов одного автора. Существенным этапом создания корпуса является предобработка текстов — лемматизация, сегментация и др. Эта часть статьи может быть полезна всем исследователям, которые создают корпуса текстов для изучения вопросов сложности. Далее авторы используют два текстовых профайлера: RuLingva — для оценки значений 49 языковых параметров, RuLex — для извлечения терминов из текстов учебников. Выделено 9 параметров, имеющих статистически значимую корреляцию со сложностью. Интересно, что в их числе не оказался TTR — параметр, характеризующий лексическое разнообразие в тексте. К основным результатам статьи следует отнести установление связи сложности текста с лексической плотностью (долей знаменательных частей речи) и связностью текста (числом лексических повторов). Пионерскими являются исследования числа терминов в учебниках разных классов. Несколько неожиданным оказалось то, что терминов больше в учебниках для 5-го класса. Этот результат требует дальнейших исследований и обсуждения. Подобное детальное изучение языковых параметров в связи со сложностью учебников по русскому языку разных классов проводится впервые.

В статье Е.Н. Булиной, М.И. Солнышкиной, Ю.Н. Эбзеевой «Учебник русского языка как проводник перемен: от СССР до нового века» сопоставлены структура и типографика учебников двух разных исторических периодов: 1935–1974 гг. и 2012–2015 гг. Авторы показывают, что основные структурные элементы учебников разных периодов времени совпадают и примерно одинаково располагаются — это тексты по теории, тексты инструкций и заданий и тексты упражнений. При этом оказалось, что доля этих трех «формантов» существенно различается. Объем заданий в современных учебниках вырос более чем в 2 раза. Сделан важный вывод об изменении характера инструкций. Директивы советских учебников имеют традиционную форму побуждения, выражаемых глагольными императивами. В тоже время в современных учебниках преобладают мотивационные вопросы, реа-

лизирующие тенденцию к диалогичности. Авторы статьи внимательно рассматривают типографику учебников. Типографика современных учебников ожидаемо разнообразнее, качественнее, что способствует лучшему восприятию текста. Статья характеризуется тщательным отбором учебников. Собственно лингвистический анализ текстов осуществляется с применением современных средств компьютерной лингвистики, в т.ч. с помощью разработанного в Казанском федеральном университете программного комплекса RuLingva. Значительное внимание уделяется общепедагогическим вопросам в контексте меняющейся социально-политической обстановки в стране. Данная статья задает некие рамки для серии последующих исследований в этом направлении, часть из которых помещена в этом номере журнала.

Авторы статьи «Лингвистическое профилирование учебных и художественных текстов» К.В. Воронин, Ф.Х. Исмаева, А.В. Данилов осуществляют детализированное профилирование художественных текстов, в качестве которых избраны приключенческие рассказы, и противопоставляют их текстам учебных биографий, используемых в учебниках РКИ. К параметрам-дискриминантам, дифференцирующим биографии из учебников и приключенческие рассказы, относятся: глобальные и локальные повторы существительных и личных местоимений, дистрибуция имен существительных в предложном и родительном падежах, а также глаголов прошедшего и настоящего времени. Жанровая специфика биографии проявляется в более широких референсных диапазонах предложного и родительного падежей имен существительных, а также большей связанности. Как и предыдущие статьи данного выпуска, исследование выполнено на весьма репрезентативном объеме материала и включает детальный анализ 15 языковых параметров, рассчитанных при помощи RuLingva, а также последовательное описание методологии исследований. Статья может рассматриваться как образец межжанрового профилирования.

Статья «Лексическое обогащение в учебниках филологического блока: корпусный и статистический подходы» Х.Н. Галимовой, Е.В. Мартыновой, С.А. Москвичевой посвящена лексическому наполнению учебников по русскому языку и литературе. Как и в предыдущих статьях этого блока, рассматриваются 66 учебников 5–7 классов российской средней школы общим объемом более 1,5 млн слов. Корпус является максимально репрезентативным, так как содержит все учебники, входящие в Федеральный государственный образовательный стандарт. Авторы изучают словарный состав учебников в разрезе объема, частотности, динамики от класса к классу. Один из заслуживающих внимания результатов — то, что наибольший словарный состав оказался в учебниках для 5-го класса. Представляется, что эти данные должны быть еще осмыслены в свете общей концепции среднего образования в России. Авторы описывают частотные словари, созданные для каждого класса. Отдельно обсуждается проблема анализа редких слов. Полученные частотные словари и словари «обогащения» детально анализируются по тематическим группам. Динамика словарного состава учебников представляет особый интерес. Оказалось, что в 6-м классе по сравнению с 5-м 25 % новых слов составляют устаревшие слова (историзмы и архаизмы), что способствует сохранению культурного кода России. В итоге авторы делают вывод, что лексический состав учебных предметов «Русский язык»

и «Литература» представляет собой важный материал для воспитания современного человека и сохранения культурных традиций России.

В статье Е.Е. Иванова и В.И. Куликовича «Теория русской орфографии в учебной литературе для студентов Республики Беларусь» представлена практика преподавания орфографии русского языка в вузах Беларуси. В качестве принципиально важной заявлена следующая оппозиция: введение орфографии исключительно на примерах и простых правилах или формирование теоретических основ и фундаментальной методологической базы. Авторы выделяют три группы учебников на основании представленности орфографии как теоретической дисциплины. Крайне важным представляется вывод авторов о том, что студенты, изучающие орфографию в рамках теоретического подхода, делают меньше ошибок в письме, чем те, которые использовали учебники исключительно с примерами. При этом обе группы студентов неплохо справляются со стандартными тестами.

В статье также уделяется много внимания вопросу единства трактовки или разночтениям в определениях орфографических понятий в различных учебниках. Приводится показательный пример, когда в одном учебнике термину «орфография» приписывается одно значение, а в другом — целых четыре (!). Авторы оценивают сложившуюся ситуацию следующим образом: «терминологическая база орфографии русского языка в белорусских учебных изданиях <...> во многих случаях носит ненаучный характер». Результаты исследования обосновывают изложение современной теории орфографии русского языка на базе четырех принципов: *системности, антропологии, семантической целостности и целесообразности*.

В целом следует отметить важность намеченного авторами подхода к преподаванию русского языка за рубежом, особенно в странах с большой долей русскоговорящего населения, — учета вариативности не только орфографии, но и других разделов языкознания. Представляется, что этому вопросу уделяется недостаточно внимания. Изложенные в статье идеи могут быть применены и в преподавании русского как иностранного в самой России.

Статья «Linguistic parameters of formants in textbooks of the Russian language: a comparative study» by R.V. Kupriyanov, G.N. Shoeva, O.I. Aleksandrova посвящена ранее не отраженному в литературе вопросу систематического количественного сравнения по многим лингвистическим параметрам текстов школьных учебников, использовавшихся в СССР/России в разные годы. Авторы выбрали значительный временной интервал с 1937 по 2015 гг. и рассмотрели учебники по русскому языку для 5 класса. Изучены 24 языковых параметра текстов. Для количественного анализа использован профайлер RuLingva, при помощи которого выявлен ряд весьма интересных закономерностей изменения учебных текстов со временем. В частности, установлено, что тексты современных учебников более простые (в смысле использования более коротких предложений и слов), что кажется несколько неожиданным. Выделен еще целый ряд параметров, по которым имеются статистически значимые различия советских и российских учебников. Авторы также обращают внимание на то, что учебники имеют смысловые фрагменты, различающиеся по своим лингвистическим параметрам: изложение теоретического материала, упражнения и задания. Детально проанализированы значения 24 языковых параметров в этих фрагментах. Обозначено направление дальнейших исследований: в первую очередь это

увеличение количества рассматриваемых учебников как по классам, так и по предметам. Статья может служить образцом для исследований в заданном направлении.

Особняком в тематическом выпуске стоит статья И.В. Приваловой и А.А. Петровой «Методы мониторинга англицизмов в русскоязычном молодежном дискурсе», в которой представлена авторская методика исследования англицизмов в современном российском молодежном дискурсе. Авторы приводят результаты трех опросов, проведенных в Саратовском, Волгоградском и Казанском университетах за последние 7 лет. В опросах приняли участие несколько сотен респондентов, изучалась частота использования более 1300 слов молодежного социолекта. Авторы пришли к следующим выводам: англицизмы значительно превосходят иные типы слов в молодежном социолекте; они в первую очередь укореняются при общении в кругу друзей и членов семьи, а также в Интернет-коммуникации. При сопоставлении частотности лексем в реальном употреблении в молодежной среде с частотностью в корпусе НКРЯ установлено, что частотность в корпусе значительно ниже. Это естественно объяснить тем, что состав НКРЯ отстает от реального употребления и требуется время для закрепления в языке новых единиц. Также в работе показано влияние зарубежных и российских телесериалов на формирование молодежного сленга.

Безусловно, изучение столь нового и динамично меняющегося явления, как молодежный сленг, — сложная исследовательская задача, требующая постоянного мониторинга ситуации. Выполненная работа, являясь одним из немногих систематических исследований в этой области, все же имеет ряд ограничений. Прежде всего следует отметить, что используется единственный корпус текстов — НКРЯ, в дальнейшем необходимо перепроверить полученные результаты на других корпусах, в первую очередь, на Google Books Ngram. Далее, в качестве источника слов молодежного социолекта использован словарь Н.Л. Шамне, Л.Н. Ребриной¹³, причем, как отмечено в статье И.В. Приваловой и А.А. Петровой, «метод сплошной выборки в алфавитном порядке оказался наиболее эффективным в плане выбора лексем». Представляется, что в дальнейшем следует расширить исследуемый словарный состав, и, с другой стороны, пересмотреть сам словарь. Ряд слов — флешка, фанат, контент и др. — выходят далеко за рамки только молодежного употребления, они давно закрепились в языке, что, в свою очередь, ставит сложную задачу разграничения молодежного сленга и слов общеупотребительного языка. Наконец, особой является проблема омонимии. Например, слово *бомбить* имеет разные значения в молодежном и медийном дискурсах. Важность заявленного направления исследований определяется актуальной задачей сохранения русского языка.

Заключение

Анализ состояния текстовой аналитики показывает, что для решения стоящих перед ней задач современная лингвистика успешно обращается к междисциплинарным подходам. Все большую актуальность приобретают инструменты лингвистического профилирования, опирающиеся на достиже-

¹³ Шамне Н.Л., Ребрина Л.Н. Словарь молодежного сленга. Волгоград : Изд-во ВолГУ, 2017.

ния лингвистической статистики, компьютерной лингвистики и искусственного интеллекта. Методологическим основанием формализованных методов анализа текста служат открытия, сделанные в области теории текста, функциональной стилистики, стилостатистики, а также компьютерной лингвистики.

Ближайшая перспектива развития RuLingva связана со следующими основными направлениями: расширением спектра предикторов сложности и внедрением на платформе функции автоматического определения предметной области учебного текста, а также создание функционала оценки вербального интеллекта и читательской грамотности пользователя. Эти направления могут быть реализованы с использованием одного из двух подходов: (1) при помощи нейронных сетей и созданных на их основе классификационных моделей; (2) на базе «типологических паспортов» учебных текстов различной сложности и тематической направленности. Первый подход весьма привлекателен и может быть осуществлен на базе репрезентативной коллекции текстов с применением больших языковых моделей. Второй подход — не менее трудоемкий, поскольку для обеспечения точности идентификации уровня сложности и предметной области профайлер должен быть обеспечен информацией о референсных диапазонах значений широкого спектра параметров, которые в дальнейшем послужат в качестве предикторов сложности и дискриминантов предметных областей. И только второй подход обеспечит «лингвистический взгляд» для понимания того, внутренних процессов при обработке текстов нейронными сетями. Будущее RuLingva, как и всей современной текстовой аналитики, во многом зависит от технологического обеспечения профайлеров, качество работы и достоверность оценок которых обеспечены достижениями междисциплинарных исследований в области лингво-когнитивных наук, лингвистической онтологии, дискурс-анализа, искусственного интеллекта, а также в широком спектре других наук, все чаще привлекаемых для решения практических задач.

Список литературы

- Виноградов В.В. Современный русский язык. Грамматическое учение о слове. М. ; Л. : Гос. учеб.-пед. изд-во Наркомпроса РСФСР, 1938. 591 с.
- Головин Б.Н. Язык и статистика. М. : Просвещение, 1971. 189 с.
- Зиндер Л.Р., Строева Т.В. Историческая морфология немецкого языка. Л. : Просвещение, Ленингр. отд-ние, 1968. 262.
- Колмогорова А.В., Колмогорова П.А., Куликова Е.Р. О прошлом, но в разное время: компьютерный анализ текстов учебников по истории СССР/России для шести поколений студентов // Вестник Томского государственного университета. Филология. 2024. № 89. С. 73–103. <http://doi.org/10.17223/19986645/89/4>
- Кожина М.Н. О функциональных семантико-стилистических категориях в аспекте коммуникативной теории языка // Разновидности и жанры научной прозы. Лингвостилистические особенности. М. : Наука, 1989. С. 3–27.
- Кормилицына М.А., Сиротинина О.Б. Функциональная стилистика и ее место в современной лингвистике // Славянская стилистика. Век XXI : сб. статей / под ред. Л.Р. Дускаевой. СПб. : С.-Петербург. гос. ун-т, Высш. шк. журн. и мас. коммуникаций, 2013. С. 101–111.
- Кронгауз М.А. Русский язык на грани нервного срыва. М. : Языки славянских культур, 2009. 232 с.
- Лапошина А.Н., Веселовская Т.С., Лебедева М.Ю., Купрещенко О.Ф. Лексический состав текстов учебников русского языка для младшей школы: корпусное исследование // Компьютерная лингвистика и интеллектуальные технологии : по материалам международной конференции «Диалог 2019». 2019. Т. 18 (25). С. 351–363.
- Лапошина А.Н., Лебедева М.Ю. Текстометр: онлайн-инструмент определения уровня сложности текста по русскому языку как иностранному // Русистика. 2021. Т. 19. № 3. С. 331–345. <http://doi.org/10.22363/2618-8163-2021-19-3-331-345>

- Лукашевич Н.В., Добров Б.В. Проектирование лингвистических онтологий для информационных систем в широких предметных областях // *Онтология проектирования*. 2015. № 1 (15). С. 47–69.
- Ляшевская О.Н., Шаров С.А. Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). М. : Азбуковник, 2009.
- Михеев М.Ю., Эрлих Л.И. Идиостилевой профиль и определение авторства текста по частотам служебных слов // *Научно-техническая информация. Сер. 2 : Информационные процессы и системы*. 2018. № 2. С. 25–34.
- Наместников А.М., Пирогова Н.Д., Филиппов А.А. Подход к автоматическому построению лингвистической онтологии для определения интересов пользователей социальных сетей // *Онтология проектирования*. 2021. Т. 11. № 3. С. 351–363. <http://doi.org/10.18287/2223-9537-2021-11-3-351-36>
- Оборнева И.В. Автоматизированная оценка сложности учебных текстов на основе статистических параметров: дис. ... канд. пед. наук. М., 2006. 165 с.
- Сердобольская Н.В., Толдова С.Ю. Оценочные предикаты: тип оценки и синтаксис конструкции // *Компьютерная лингвистика и интеллектуальные технологии: труды Междунар. конф. «Диалог» 2005*. М. : Наука, 2005. С. 436–443.
- Соловьев В.Д., Солнышкина М.И., Макнамара Д.С. Компьютерная лингвистика и дискурсивная комплексология: парадигмы и методы исследований // *Russian Journal of Linguistics*. 2022. Т. 26. № 2. С. 275–316. doi: 10.22363/2687-0088-31326
- Соссюр Ф. де. Труды по языкознанию. М. : Прогресс, 1977. 695 с.
- Blinova O., Tarasov N. A hybrid model of complexity estimation: Evidence from Russian legal texts // *Frontiers in Artificial Intelligence*. 2022. Vol. 5. <http://doi.org/10.3389/frai.2022.1008530>
- Chang T.A., Arnett C., Tu Z., Bergen B.K. When is multilinguality a curse? language modeling for 250 high-and low-resource languages // *arXiv preprint*. 2023. <https://doi.org/10.48550/arXiv.2311.09205>
- Corlatescu D., Ruseti Ș., Dascalu M. ReaderBench : Multilevel analysis of Russian text characteristics // *Russian Journal of Linguistics*. 2022. Vol. 26. No. 2. Pp. 342–370. <https://doi.org/10.22363/2687-0088-30145>
- Cvrček V., Chlumská L. Simplification in translated Czech : a new approach to type-token ratio // *Russian Linguistics*. 2015. Vol. 39. Pp. 309–325. <https://doi.org/10.1007/s11185-015-9151-8>
- Dmitrieva A., Laposhina A., Lebedeva M. A comparative study of educational texts for native, foreign, and bilingual young speakers of russian: are simplified texts equally simple? // *Frontiers in Psychology*. 2021. Vol. 12. 703690. <https://doi.org/10.3389/fpsyg.2021.703690>
- Gatiyatullina G., Solnyshkina M., Solovyev V., Danilov A., Martynova E., Yarmakeev I. Computing Russian morphological distribution patterns using RusAC online server // 2020 13th International Conference on Developments in eSystems Engineering (DeSE). IEEE, 2020. Pp. 393–398. <https://doi.org/10.1109/DeSE51703.2020.9450753>
- Karakanta A., Dehdari J., van Genabith J. Neural machine translation for low-resource languages without parallel corpora // *Machine Translation*. 2017. Vol. 32. Pp. 167–189. <https://doi.org/10.1007/s10590-017-9203-5>
- Kupriyanov R.V., Solnyshkina M.I., Dascalu M., Soldatkina T.A. Lexical and syntactic features of academic Russian texts : a discriminant analysis // *Research Result. Theoretical and Applied Linguistics*. 2022. Vol. 8 No. 4. Pp. 105–122. <http://doi.org/10.18413/2313-8912-2022-8-4-0-8>
- Kuznetsova I. Linguistic profiles : going from form to meaning via statistics. Berlin, München, Boston : De Gruyter Mouton, 2015. <http://doi.org/10.1515/9783110361858>
- Lipmann W. Public opinion. New York : Macmillan, 1922.
- Lyashevskaya O., Panteleeva I., Vinogradova O. Automated assessment of learner text complexity // *Assessing Writing*. 2021. Vol. 49. 100529. <https://doi.org/10.1016/j.asw.2021.100529>
- McNamara D.S., Graesser A.C., McCarthy P.M., Cai Z. Automated Evaluation of Text and Discourse with Coh-Metrix. Cambridge University Press, 2014.
- Morozov D.A., Glazkova A.V., Iomdin B.L. Text complexity and linguistic features: Their correlation in English and Russian // *Russian Journal of Linguistics*. 2022. Vol. 26. No. 2. Pp. 426–448. <https://doi.org/10.22363/2687-0088-30132>

- Paraschiv A., Dascalu M., Solnyshkina M.I.* Classification of Russian textbooks by grade level and topic using ReaderBench // Research Result. Theoretical and Applied Linguistics. 2023. Vol. 9. No. 1. Pp. 50–63. <https://doi.org/10.18413/2313-8912-2023-9-1-0-4>
- Sakhovskiy A., Solovyev V., Solnyshkina M.* Topic modeling for assessment of text complexity in Russian textbooks // 2020 Ivannikov Ispras Open Conference (ISPRAS). IEEE, 2020. Pp. 102–108. <https://doi.org/10.1109/ISPRAS51486.2020.00022>
- Solnyshkina M.I., Solovyev V.D., Gafiyatova E.V., Martynova E.V.* Text complexity as an interdisciplinary problem // Issues of Cognitive Linguistics. 2022. No. 1. Pp. 18–39. <https://doi.org/10.20916/1812-3228-2022-1-18-39>
- 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. No. 5. Pp. 3049–3058 <http://doi.org/10.3233/JIFS-169489>
- Toldova S., Anastasiya A.B., Lyashevskaya O., Ionov M.* Evaluation for morphologically rich language: Russian NLP // Int'l Conf. Artificial Intelligence. ICAI'15. 2015. Pp. 300–306.
- Valeev A., Gibadullin I., Khusainova A., Khan A.* Application of Low-resource Machine Translation Techniques to Russian-Tatar Language Pair // arXiv preprint. 2019. <http://doi.org/10.48550/arXiv.1910.00368>
- Virk S.M., Hammarström H., Borin L., Forsberg M., Wichmann S.* From Linguistic Descriptions to Language Profiles // Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020). Marseille : European Language Resources Association, 2020. Pp. 23–27.
- Young T., Hazarika D., Poria S., Cambria E.* Recent Trends In Deep Learning Based Natural Language Processing // IEEE Computational intelligence magazine. 2018. Vol. 13. No. 3. Pp. 55–75. <http://doi.org/10.1109/MCI.2018.2840738>

Сведения об авторах:

Солнышкина Марина Ивановна, доктор филологических наук, профессор, профессор кафедры теории и практики преподавания иностранных языков, руководитель НИЛ «Мультидисциплинарные исследования текста», Казанский (Приволжский) федеральный университет, Российская Федерация, 420008, г. Казань, ул. Кремлевская, д. 18. *Сфера научных интересов*: текстовая аналитика, сложность текста, дискурсивная комплексология, лексикография, языковая личность. ORCID: 0000-0003-1885-3039. SPIN-код: 6480-1830. Researcher ID: E-3863-2015. Scopus ID: 56429529500. E-mail: mesoln@yandex.ru

Соловьев Валерий Дмитриевич, доктор физико-математических наук, профессор, главный научный сотрудник НИЛ «Мультидисциплинарные исследования текста» Института филологии и межкультурной коммуникации, Казанский (Приволжский) федеральный университет, Российская Федерация, 420008, г. Казань, ул. Кремлевская, д. 18. Член президиума Междисциплинарной ассоциации когнитивных исследований. Автор четырех монографий и более 60 публикаций по сложности текста. *Сфера научных интересов*: когнитивная наука, компьютерная лингвистика, искусственный интеллект. ORCID: 0000-0003-4692-2564. SPIN-код: 5791-3820. Researcher ID: C-8023-2015. Scopus ID: 26665013000. E-mail: maki.solovyev@mail.ru

Эбзеева Юлия Николаевна, доктор социологических наук, кандидат филологических наук, первый проректор — проректор по образовательной деятельности, заведующая кафедрой иностранных языков, филологический факультет, Российский университет дружбы народов, Российская Федерация, 117198, г. Москва, ул. Миклухо-Маклая, д. 6. *Сфера научных интересов*: лексикология и стилистика французского языка, теория перевода, межкультурная коммуникация, социолингвистика, миграциология, образовательная политика. ORCID: 0000-0002-0043-7590. SPIN-код: 3316-4356. E-mail: ebzeeva-jn@rudn.ru

DOI: 10.22363/2618-8163-2024-22-4-501-517

EDN: AMYSNF


Introductory article

Approaches and tools for Russian text linguistic profiling

Marina I. Solnyshkina¹, Valery D. Solovyev¹, Yulia N. Ebzeeva²

¹Kazan (Volga Region) Federal University, *Kazan, Russian Federation*

²RUDN University, *Moscow, Russian Federation*

 mesoln@yandex.ru

Abstract. Approaches and tools for assessing linguistic and cognitive complexity of educational texts are in demand both in science and teaching. Predicting difficulties of perception and understanding and ranking texts by classes, i.e. the number of years of learning or levels of language proficiency (A1–C2), are of particular importance for education. The study is aimed at demonstrating modern methodologies, algorithms, and tools for analyzing Russian texts in text profiler and automatic analyzer RuLingva and at presenting articles from the thematic issue on comprehensive analysis of Russian language textbooks for Russian and Belarusian schools. The research demonstrates that the modern paradigm of discourse complexity is based on the methods of stylistic statistics, which identifies functional characteristics of language units and verifies them using big data. The services on RuLingva are designed for teachers and researchers; they automatically analyze educational texts and predict their target audience based on readability, lexical diversity, abstractness, frequency, and terminological density. In “Russian as a Foreign Language” mode, RuLingva downloads lists of words from the text according to each level of language proficiency and estimates their proportion. This provides material for pre- and post-text work. RuLingva algorithm is based on the typology of educational texts and is to be supplied with tools for assessing a person’s verbal intelligence and reading literacy. The nearest prospect of RuLingva lies in widening the range of complexity predictors and installing automatic subject area discriminator. Both directions are planned to be implemented using neural networks, classification models, “typological passports” of educational texts with different complexity, and thematic orientation.

Keywords: text profiler RuLingva, text complexity, educational text, typological passport of the text, complexity predictors

Contribution: Solnyshkina M.I. — idea, research, text preparation and editing; Solovyev V.D. — methodology, research; Ebzeeva Yu.N. — research, approval of the final version of the article.

Funding. This article has been supported by the Kazan Federal University Strategic Academic Leadership Program (PRIORITY–2030). This publication has been supported by the RUDN University Scientific Projects Grant System, project no. 050738-0-000.

Conflict of interests. The authors declare that they have no conflict of interests.

Article history: received 02.07.2024; accepted 18.08.2024.

For citation: Solnyshkina, M.I., Solovyev, V.D., & Ebzeeva, Y.N. (2024). Approaches and tools for Russian text linguistic profiling. *Russian Language Studies*, 22(4), 501–517. <http://doi.org/10.22363/2618-8163-2024-22-4-501-517>