

DOI: 10.22363/2618-8163-2024-22-4-555-578

EDN: AWRLUL

Научная статья

Лингвистическое профилирование учебных и художественных текстов

К.В. Воронин , Ф.Х. Исмаева , А.В. Данилов

Казанский (Приволжский) федеральный университет, Казань, Российская Федерация

voronin.konstantin@outlook.com

Аннотация. Исследование проведено в рамках текстовой аналитики как одного из стратегических направлений современной российской лингвистики и посвящено лингвистическому профилированию учебных и художественных текстов. Определение жанровой специфики текстов востребовано для разработки автоматических программных комплексов и обработки больших языковых данных. Цель исследования состоит в выявлении диапазонов значений лингвистических параметров, позволяющих дифференцировать учебные (вторичные) и художественные (первичные) тексты. Исследование выполнено на материале 72 текстов биографий из учебников по русскому языку как иностранному и 90 отрывков из приключенческих рассказов. Выбор для изучения оппозиции жанров обусловлен (1) высокой степенью нарративности и динамичности обоих; (2) функциональными различиями: нацеленностью биографии на информативность и приключений — на развлечения. Общий объем исследовательского корпуса составил 120932 словоформ. В исследовании использованы два инструмента обработки данных: (1) с помощью платформы RuLingva рассчитывали значения лингвистических параметров; (2) программу STATISTIKA использовали для выявления статистически значимых различий изучаемых текстов. Определено, что список параметров, дифференцирующих тексты биографий и приключений, включает глобальные и локальные повторы существительных и аргумента, дистрибуцию имен существительных в предложном и родительном падежах, дистрибуцию глаголов прошедшего и настоящего времени. Выявлено, что повторы существительных и аргумента способствуют высокой связности текстов биографий, что реализуется в номинации каждого следующего события жизни человека в связи с предыдущим. Родительный падеж превалирует в биографиях по причине использования номинативных словосочетаний существительных. Перспектива исследования видится в типологии жанров, основанной на лингвистическом профилировании официальных и личных биографий, с одной стороны, и приключенческих романов, с другой. Отдельный аспект дальнейших исследований в изучаемой области составляет инсталляция матриц значений в текстовый профайлер RuLingva с целью автоматизации лингвистического профилирования текстов.

Ключевые слова: русский язык как иностранный, РКИ, биография, приключенческий рассказ, корпус исследования, лингвистическая экспертиза, языковой анализ, жанр, связность

Вклад авторов: *Воронин К.В.* — концепция исследования; *Исмаева Ф.Х.* — сбор и обработка материалов, написание текста; *Данилов А.В.* — анализ полученных данных.

Финансирование. Исследование выполнено при финансовой поддержке РФФ, грант 24-78-10129 «Когнитивная и лингвистическая сложность текста».

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

История статьи: поступила в редакцию 25.03.2024; принята к печати: 11.06.2024.

Для цитирования: *Воронин К.В., Исмаева Ф.Х., Данилов А.В.* Лингвистическое профилирование учебных и художественных текстов // *Русистика*. 2024. Т. 22. № 4. С. 555–578. <http://doi.org/10.22363/2618-8163-2024-22-4-555-578>

Введение

Автоматическая дифференциация жанров рассматривается в современной лингвистике в целом и в текстовой аналитике в частности как многоплановая и сложная задача, поскольку параметры, по которым можно разграничивать письменные жанры, весьма многочисленны. Жанр, определяемый как «тематически, композиционно и стилистически устойчивый тип текста» (Федосюк, 1997: 107), характеризуется на основе ряда признаков: функция, авторская интенция, форма, структура, тематика, объем, проблематика, способы передачи информации, выражение субъективности, образ адресата и адресанта, ситуация и пр. (Стаценко, 2016). Актуальным в рамках проблематики жанровой дифференциации является лингвистическое профилирование, которое как лингвистическая методология имеет целью не только создание лингвистического «профиля» текста определенного типа/жанра, но и выявление значений параметров-дискриминантов, на основе которых исследователь вручную или автоматически может дифференцировать тексты разных жанров и типов, а также тексты, принадлежащие разным авторам. Лингвистическое профилирование включает ряд исследовательских процедур: (1) сначала «подсчитывается количество вхождений в текст большого количества языковых признаков, объективируемых либо в отдельных элементах, либо в комбинациях элементов; (2) затем эти подсчеты нормализуются для того, чтобы (3) обнаружить и количественно оценить различия и сходства в текстах различных жанров и типов (Van Halteren, 2004¹).

Следует признать, что при обширном опыте такого рода исследований на материале европейских языков (Swales, 2004; Mendhakar, 2022), работы в данной области на материале русского языка весьма немногочисленны (Литвинова и др., 2013; Чурунина и др., 2023; Solnyshkina et al., 2024). Основы лингвистической параметризации текстов были положены в 1986 г. Д. Байбером, который впервые (1) выдвинул предположение о возможности профилирования типа (или жанра) текста как функции параметризации текста (Biber, 1986), а затем (2) разработал основы многомерного (многофакторного) метода жанрового варьирования (Biber, 1988). Аналогичные исследования осуществляют как зарубежные, так и российские ученые, стремясь

¹ Перевод на русский язык здесь и далее осуществлен авторами статьи.

параметризовать письменные тексты в оппозиции к устным, учебные в оппозиции к научным (Mendhakar, 2022; Сиротинина и др., 1983; Кожина, 1999). Многомерный (или многофакторный) анализ используется в компьютерной лингвистике для выявления жанро-специфических характеристик, позволяющих классифицировать типы текстов и жанры с помощью ограниченного числа лингвистических параметров (Литвинова и др., 2013; Ян, 2023).

Ученые признают, что лингвистическое профилирование и параметризация текста весьма востребованы в ряде областей (Manning, Schütze, 1999; Mendhakar, 2022), а современные алгоритмы обработки естественного языка успешно используются как для идентификации лингвистических переменных, специфичных для определенного типа текста, так и для разработки классификационных языковых моделей (Brunato et al., 2020). Особенностью современного этапа развития данной области является и то, что профилирование текста / жанра, т.е. определение текстовых шаблонов, которые могут быть использованы для классификации текстовых жанров или типов, обязательно включает использование статистических методов (Paltridge, 1994; Van Halteren, 2004).

Современная традиция жанрометрических подходов заключается в выявлении как наиболее заметных, так и наиболее редких признаков в тексте (Dell'Orletta et al., 2013), при этом особое внимание и интерес ученых сосредоточены на выявлении закономерностей в минимально возможных сегментах текста (Montemagni et al., 2013). Отдельное направление лингвистического профилирования — выявление набора грамматических и синтаксических параметров текста, диагностирующих личность автора (Литвинова и др., 2013).

В рамках современной парадигмы лингвистического профилирования параметризация рассматривается как ключ не только к жанровому профилированию, но и к объективной оценке лингвистической сложности текста и когнитивных трудностей для респондента (Andreeva et al., 2020; Ivanov, Solnyshkina, 2020; Sakhovsky, Solovyev, Solnyshkina, 2020; Ismaeva, Tomin, Sharifullina, 2023).

В представленном исследовании мы тестируем **гипотезу** о том, что лингвистические параметры, присущие жанру биографии, могут быть сведены к конкретному списку и имеют строго определенные диапазоны значений, отличающие их от текстов приключений. Очевидно, что стилистическое варьирование биографий различного типа, таких как биографическое эссе, биографическая справка и документально художественная биография (Сафонова, 2018), детерминирует широкие диапазоны метрик лингвистических параметров текстов данного жанра, именно поэтому задача идентификации самого списка параметров и диапазонов их метрик видится как трудновыполнимая. Однако, если речь идет об «официальной биографии» (Ян, 2023: 7), используемой в преподавании языков вообще и русского как иностранного (РКИ) в частности, то логично предположить, что биографии как вторичному, сконструированному для учебных целей жанру свойственен более узкий диапазон метрик. Дополнительным аргументом в пользу едино-

образия и стабильности значений лингвистических параметров биографий, используемых в преподавании, может служить учитываемая разработчиками сложность, соответствующая установленной в РКИ шкале (элементарный, базовый, первый и т.д.²).

Во втором десятилетии XXI в. биографии выдающихся личностей — один из наиболее востребованных жанров. А.А. Холиков определяет биографию как «один из способов познания и реконструкции личности в ее становлении и развитии», а в качестве ведущей жанровой особенности выделяет «стремление третьего лица воссоздать словесными средствами целостный процесс становления, развития и деятельности исторической личности» (Холиков, 2016: 50). В соответствии с ситуацией и сферой общения биографические тексты классифицируются на три типа: официальные, приватные и художественные биографии (Ян, 2023). Официальные биографии адресуются широкой аудитории, описывают жизнь и достижения известных людей или предназначены для официального документального свидетельства. Приватные биографии представлены в дневниках, письмах, блогах и т.д. Художественные биографии имеют субъективную окраску автора, вымысел, оценки и эмоциональное выражение (Ян, 2023). Многообразие официальной биографии как жанра представлено *биографией-справкой*, *биографическим очерком* и *биографическим эссе* (Ефремова, 2016). Среди основных особенностей текстов биографий ученые выделяют линейную структуру, шаблонность и клишированность (Терпугова, 2011).

Определяя приключение как жанр, М.М. Бахтин утверждал, что «оно изменяет реальное течение времени, сжимает его: время движется быстрее, почти не меняя персонажей; оно разделено на ряд коротких отрезков (приключений)» (Бахтин, 1975: 157). Характерной особенностью приключенческих рассказов является быстрая смена действий, обстоятельств, событий, в которых герой не думает о будущей судьбе и живет настоящим. Все это требует определенной формы художественного произведения. В приключенческих рассказах понятия «времени» и «места» часто условны (Савирова, 2017). В силу этого жанр приключенческого рассказа рассматривается как составная часть жанра «травелог», характеризующего путешествия героев. Элементы приключений в данном жанре изобилуют действиями и сменами локаций (Абдулвахиду, 2014).

Конкретные *исследовательские вопросы*, поставленные в исследовании:

- 1) Каковы лингвистические параметры-дискриминанты, позволяющие установить различия между (а) учебными текстами биографий, с одной стороны, и (б) художественными текстами (приключениями)?
- 2) Каковы «жанровые» диапазоны значений параметров-дискриминантов для указанных жанров?

Цель исследования — выявить значения жанро-различительных параметров учебных текстов биографий и приключенческих рассказов.

² Институт русского языка и культуры МГУ им. М.В. Ломоносова. URL : <https://test.irlc.msu.ru/trki/> (дата обращения: 21.03.2024).

Методы и материалы

Алгоритм исследования включал пять этапов: (1) создание и предобработка корпуса исследования — текстов биографий из учебников по РКИ и текстов приключенческих рассказов; (2) отбор текстов с соответствующими значениями читабельности; (3) расчеты значений лингвистических параметров текстов; (4) нормализация отдельных параметров для (5) идентификации статистически значимых различий лингвистических параметров текстов двух разных жанров и (6) определение диапазонов значений параметров-дискриминантов жанров.

На *этапе I* файлы в формате pdf были конвертированы в txt формат. Каждому тексту присваивался код, содержащий указание на авторов (AnNa — В.Е. Антонова, М.М. Нахабина), уровень владения языком по CEFR³ (A1–C2), жанр (Bio), объект биографического описания, год и количество словоформ. В код файлов приключенческих рассказов вошли зоны авторов, жанра (Adv), года, количества словоформ. Корпус приключенческих рассказов российских авторов 1920–1990 гг. составлялся в ходе реализации проекта по изучению жанровой специфики текста (Solnyshkina et al., 2024). Поскольку достоверность расчетов значений параметров достигается только в тексте, длина которого не превышает 1000 словоформ (Biber, 2006; Вахрушева и др., 2021; Соловьев и др., 2022), расчеты метрик лингвистических параметров текста приключенческого рассказа были сегментированы на фрагменты по 1000 словоформ⁴.

На *этапе II* каждый текст исследовательского корпуса (биографии $n = 72$, приключения $n = 90$) был проанализирован на платформе RuLingva⁵ с целью определения индекса читабельности. Выявлено, что преимущественное большинство текстов в коллекции имеют читабельность 6–8, т.е. соответствуют когнитивному возрасту читателей с уровнем образования 6–8 классов общероссийской школы. Именно поэтому материал данного исследования составили тексты со средним индексом FKGL = $7 \pm 1,9$. Индекс FKGL рассчитывался по формуле

$$0,36 \times \text{СДП} + 5,76 \times \text{СДС} - 11,97,$$

где СДП — это средняя длина предложения в словах; СДС — это средняя длина слова в слогах (Solnyshkina et al., 2018).

Поскольку индекс читабельности позволяет осуществить отбор текстов, имеющих приблизительно одинаковую трудность восприятия, т.е. предназначенных для одной читательской аудитории, все остальные тексты были исключены из исследования. В финальный корпус исследования были включены 72 биографических текста (табл. 1), извлеченных из 17 учебников РКИ, и 90 фрагментов из текстов 7 приключенческих рассказов (табл. 2).

³ CEFR, Common European Framework of Reference — Европейская шкала уровней владения языком. URL : clck.ru/3Cqwgj (дата обращения : 21.03.2024)

⁴ RuLingva. Segmentator. URL : <https://rulingva.kpfu.ru/chunkizer> (дата обращения : 21.03.2024).

⁵ RuLingva text profiler. URL : <https://rulingva.kpfu.ru/> (дата обращения : 21.03.2024).

Таблица 1

Подкорпус биографий РКИ

Код текста	Количество текстов	Количество словоформ
AnNa_A2 ⁶	3	1170
AnNa_B ⁷	6	2718
Ag_A2 ⁸	1	234
Ar_B2 ⁹	6	2588
BeLu_A2 ¹⁰	1	380
BoAg_C1 ¹¹	3	1595
FiDm_A2 ¹²	1	142
Gr_A2 ¹³	11	5015
JaSu_C2 ¹⁴	5	1196
KaFr_B2 ¹⁵	1	291
Li_C1 ¹⁶	7	5750
Mo_B1 ¹⁷	1	215
MoSi_A1 ¹⁸	3	1187
OdNo_A1B2 ¹⁹	8	2785
Sa_C1 ²⁰	6	1512
ShKu_B1 ²¹	4	2366
TiKo_B1 ²²	5	1696
ИТОГО	72	30840

⁶ Здесь и далее в работе используется авторская маркировка учебников корпуса буквенно-числовым кодом, соответствующим году обучения и начальным буквам фамилий двух первых авторов. Антонова В.Е., Нахабина М.М., Толстых А.А. Дорога в Россию : учебник рус. яз. : базовый уровень. 4-е изд. М. : ЦМО МГУ им. М.В. Ломоносова ; СПб. : Златоуст, 2009. 256 с.

⁷ Антонова В.Е., Нахабина М.М., Толстых А.А. Дорога в Россию : учебник рус. яз. : первый уровень : в 2 томах. 4-е изд. СПб. : Златоуст, 2012. Т. 2. — 184 с.

⁸ Агеева А.Ю., Касарова В.Г. Жизнь великих врачей. Ч. 1. М. : МАДИ, 2011.

⁹ Архипова Л.В. Мы живем в России : учеб.-метод. пособие. Тамбов : Изд-во Тамб. гос. техн. ун-та, 2007. 88 с.

¹⁰ Беляева Г.В., Луцкая Н.Э. Я пишу по-русски : пособие по письму : кн. для студентов базовый уровень. М. : ред. Изд. Совет МОЦ МГ, 2008. 150 с.

¹¹ Бочина Т.Г., Агеева Ю.В. Обсуждаем, разговариваем, спорим : учеб. пособие по рус. яз. для иностр. студентов. Казань : Изд-во Казань ун-та, 2014. 144 с.

¹² Финагина Ю.В. Русский язык как иностранный : пособие по чтению : учеб. пособие. СПб. : НИУ ИТМО ; ИХиБТ, 2014. 81 с.

¹³ Гречиго Т.А. и др. Русский язык как иностранный: великие ученые. Тексты для чтения и обсуждения : методические рекомендации для иностранных студентов. Витебск : ВГУ имени П.М. Машерова, 2021. 52 с.

¹⁴ Жаумитова В.К., Суюнбаева А.Ж. Русский язык : учебное пособие. Актобе, 2020. 186 с.

¹⁵ Казакова О.А., Фрик Т.Б. Практикум по культуре речевого общения на русском языке : учеб. пособие. Томск : Изд-во Том. политех. ун-та, 2009. 136 с.

¹⁶ Литвинова Л.Б. Биографии ученых-медиков : учеб. пособие для иностр. студентов мед. вузов. М. : Ай Пи Ар Медиа, 2021. 94 с.

¹⁷ Мозелова И. Новый сувенир : учебник : русский язык для иностранцев. М. : Mozi-House, 2019. 180 с.

¹⁸ Московкин Л.В., Сильвина Л.В. Русский язык : учебник для иностр. студентов подгот. фак. СПб. : СММО Пресс, 2006. 528 с.

¹⁹ Одицова Р.И. История русской литературы X–XXI вв. : для иностр. уч-ся подгот. фак. СПб. : Златоуст, 2022. 343 с.

²⁰ Санникова А.В. Русские ученые-медики : учеб.-метод. пособие. Минск : БГМУ, 2014. 48 с.

²¹ Шустикова В.Т. Русский язык для вас. Первый сертификационный уровень : учебник русского языка для иностранных учащихся / под ред. Т.В. Шустиковой, В.А. Кулаковой. 2-е изд., доп. М. : РУДН, 2009. 319 с.

²² Тимофеева И.М., Кожевникова М.Н., Алёшичева Н.Н., Рычагова Л.В. Учимся пересказывать тексты : учеб. пособие по рус. яз. для иностр. уч-ся. М. : МАДИ, 2017. 96 с.

Table 1

Subcorpus of RFL biographies

Text code ²³	Number of texts	Number of tokens
AnNa_A2	3	1170
AnNa_B1	6	2718
Ag_A2	1	234
Ar_B2	6	2588
BeLu_A2	1	380
BoAg_C1	3	1595
FiDm_A2	1	142
Gr_A2	11	5015
JaSu_C2	5	1196
KaFr_B2	1	291
Li_C1	7	5750
Mo_B1	1	215
MoSi_A1	3	1187
OdNo_A1B2	8	2785
Sa_C1	6	1512
ShKu_B1	4	2366
TiKo_B1	5	1696
TOTAL	72	30840

Таблица 2

Подкорпус приключенческих рассказов

Код текста	Количество сегментов	Количество словоформ
Pl_1973 ²⁴	21	21007
St_1974 ²⁵	2	2004
Vu_1975 ²⁶	12	12023
Bo_1984 ²⁷	22	21995
Ga_1990 ²⁸	23	23054
Kn_1990 ²⁹	4	4022
Ch_1991 ³⁰	6	5987
ИТОГО	90	90092

Table 2

Subcorpus of adventure stories

Text code ³¹	Number of texts	Number of tokens
Pl_1973	21	21007
St_1974	2	2004
Vu_1975	12	12023
Bo_1984	22	21995
Ga_1990	23	23054
Kn_1990	4	4022
Ch_1991	6	5987
TOTAL	90	90092

²³ Hereinafter in the research the authors K.V. Voronin, F.Kh. Ismaeva, A.V. Danilov developed a code for the textbooks comprising a letter and a number to mark the family names of the textbook authors and the grades.

²⁴ Плотников А.Н. Молчаливое море. Калининград : Калининград. кн. изд-во, 1973.

²⁵ Степанов В.А. Венок на волне. М. : Воениздат, 1974.

²⁶ Вулис А.З. Хрустальный ключ. Ташкент : Джош Гвардия, 1975.

²⁷ Борщаговский А.М. Тревожные облака. М. : Физкультура и спорт, 1984.

²⁸ Гагарин С.С. Дело о Бермудском треугольнике. М. : Интерпринт, 1990.

²⁹ Князев Л.Н. Сатанинский рейс. М. : Интерпринт, 1990.

³⁰ Черкашин Н.А. Сын «Святого Петра». М. : Молодая гвардия, 1991.

³¹ Hereinafter in the research the authors K.V. Voronin, F.Kh. Ismaeva, A.V. Danilov developed a code for the textbooks comprising a letter and a number to mark the family names of the textbook authors and the grades.

В исследовании использовали тексты официальных биографий выдающихся личностей, поэтов, писателей (А.С. Пушкин, И.И. Шишкин), ученых (И.П. Павлов, М.В. Ломоносов). В текстах данных биографий содержатся краткая личная информация, анкетные данные и отсутствует авторская оценка (рис. 1).

Михаил Васильевич Ломоносов родился в 1711 году в маленьком северном городе Холмогоры в крестьянской семье. Он не мог учиться в школе, потому что в то время на севере России не было школ. Его первыми учителями были священник С.М. Сабельников и крестьянин И.А. Шубной.

В 1730 году он пришёл пешком в Москву и поступил в Славяно-греко-латинскую академию. В течение пяти лет он изучил латинский и греческий языки, математику, географию, латинскую и русскую поэзию, историю, богословские науки. В 1735 году его послали продолжать обучение в Петербургскую академию наук.

В 1741 году Ломоносов вернулся в Петербургскую академию наук. Он занимался многими науками: математикой, физикой, географией, астрономией, геологией, историей, русским языком и литературой, но самой любимой наукой Ломоносова была химия.

В 1745 году он стал профессором химии, а в 1748 году основал первую в России химическую лабораторию.

Рис. 1. Фрагмент текста биографии М.В. Ломоносова из учебника РКИ
Л.В. Московкина, Л.В. Сильвиной³²

Fig. 1. Fragment of M.V. Lomonosov biography from RFL textbook by L.V. Moskovkin, L.V. Silvina

Выбор жанров учебного текста биографии и художественного текста приключений как материалов для сравнения продиктован как сходствами, так и различиями этих жанров. Оба жанра повествовательные и имеют письменную форму. Различия жанров: (1) биография выполняет преимущественно информативную функцию, в то время как приключения — развлекательную; (2) биография имеет линейное изложение, структурные компоненты приключенческих рассказов гибки и переменчивы; (3) хронологическое жизнеописание одного лица как основа биографии противопоставляется экспрессивному изложению жизни героев приключений; (4) разработчик учебного материала и обучающийся как основная аудитория жанра учебной биографии представляется более узкой, чем широкий спектр читателей приключенческих рассказов; (5) биография используется в учебной ситуации, в то время как приключенческий рассказ — во внеучебной.

На *этапе III* был осуществлен расчет значений параметров, на основании которых позднее сравнивались тексты биографий и приключений. Список рассчитываемых параметров включал следующие: глобальный и локальный повторы существительных и аргумента, предложный и родительный падежи, среднее количество прилагательных, наречий и существительных в предложении, количество существительных, наречий, прилагательных, глаголов, время глаголов, соотношение количества глаголов к существительным, читабельность. Параметр «повтор существительных» рассчитывает количество (1) повторов существительных в смежных предложениях (локальный) и (2) повтор существительных во всем тексте (глобальный). Индекс «повтор аргумента» отражает количество повторов существитель-

³² Московкин Л.В., Сильвина Л.В. Русский язык : учебник для иностр. студентов подгот. фак. СПб. : СММО Пресс, 2006. 528 с., ил.

ных и местоимений в смежных (локальный) и во всем тексте (глобальный) (McNamara, 2014).

На этапе IV метрики количества вхождений частей речи, падежей существительных и времен глаголов были нормализованы на 1000 словоформ по формуле

$\text{количество вхождений в изучаемом тексте параметров} / \text{количество словоформ в отрывке} \times 1000.$

Табл. 3 содержит рассчитанных при помощи RuLingva только 19 из 49 статистически значимых параметров текстов биографий РКИ и приключений, объединенных в четыре группы, и объективирует основные квантитативные различия в значениях параметров сравниваемых жанров. Предиктивная сила указанных параметров подтверждена статистическими расчетами ($p < 0,05$) (табл. 3 и рис. 2).

Таблица 3

Различия лингвистических параметров текстов биографий и приключений

Группы параметров	Параметр		Среднее статистическое значение для приключения (n = 90)	Среднее статистическое значение для биографии (n = 72)	p^*	Различие, %
Дискурсивные	1	Глобальный повтор сущ.**	0,03	0,14	< 0,01	314,7
	2	Локальный повтор сущ.	0,06	0,18	< 0,01	171,7
	3	Глобальный повтор аргумента	0,16	0,32	< 0,01	100,6
	4	Локальный повтор аргумента	0,32	0,46	< 0,01	45,9
Морфологические	5	Предложный падеж (сущ.)	27,28	55,37	< 0,01	103,0
	6	Родительный падеж (сущ.)	75,03	122,38	< 0,01	63,1
	7	Ср. кол-во прил. в предлож.	1,15	1,81	< 0,01	58,3
	8	Ср. кол-во сущ. в предлож.	3,76	5,4	< 0,01	43,6
	9	Наречие	68,46	39,75	< 0,01	41,9
	10	Прилагательное	98,73	138,79	< 0,01	40,6
	11	Соотношение глаг. / сущ.	0,54	0,34	< 0,01	36,8
	12	Ср. кол-во нареч. в предлож.	0,8	0,52	< 0,01	35,7
	13	Существительное	321,25	412,6	< 0,01	28,4
	14	Глагол	172,49	137,44	< 0,01	20,3
	15	Соотношение прил. / сущ.	0,3	0,34	< 0,01	9,0
Времена глагола	16	Будущее время (глагол.)	6,48	0,83	< 0,01	87,2
	17	Настоящее время (глагол.)	41,17	19,45	< 0,01	52,7
	18	Прошедшее время (глагол.)	98,78	101,26	< 0,01	2,5
Читабельность	19	FKGL (SIS)	5,82	6,86	< 0,01	17,9

Примечание. * $p < 0,05$ — статистически значимые различия. ** Заливкой маркированы параметры с наибольшей выявленной долей различий.

Table 3

Differences of linguistic parameters of the texts of two genres

Groups of parameters	Parameter		Mean adventures (n = 90)	Mean biography (n = 72)	p-value*	Difference, %
Discourse	1	Global noun overlap	0.03	0.14	< 0.01	314.7
	2	Local noun overlap	0.06	0.18	< 0.01	171.7
	3	Global argument overlap	0.16	0.32	< 0.01	100.6
	4	Local argument overlap	0.32	0.46	< 0.01	45.9
Morphological	5	Prepositional case (Noun)	27.28	55.37	< 0.01	103.0
	6	Genitive case (Noun)	75.03	122.38	< 0.01	63.1
Parts of speech	7	Average number of adjectives per sentence	1.15	1.81	< 0.01	58.3
	8	Average number of nouns per sentence	3.76	5.4	< 0.01	43.6
	9	Adverbs	68.46	39.75	< 0.01	41.9
	10	Adjectives	98.73	138.79	< 0.01	40.6
	11	Verb/Noun ratio	0.54	0.34	< 0.01	36.8
	12	Average number of adverbs per sentence	0.8	0.52	< 0.01	35.7
	13	Nouns	321.25	412.6	< 0.01	28.4
	14	Verbs	172.49	137.44	< 0.01	20.3
	15	Adjective/Noun ratio	0.3	0.34	< 0.01	9.0
Verb tense	16	Future tense (Verb)	6.48	0.83	< 0.01	87.2
	17	Present tense (Verb)	41.17	19.45	< 0.01	52.7
	18	Past tense (Verb)	98.78	101.26	< 0.01	2.5
Readability	19	FKGL (SIS)	5.82	6.86	< 0.01	17.9

Note. * $p < 0.05$ — statistically significant differences. ** The parameters with the largest detected proportion of differences are highlighted.

Для реализации следующей исследовательской задачи были определены диапазоны параметров-дискриминантов (табл. 4–7). Идентификация статистически значимых различий лингвистических параметров текстов двух разных жанров (*эман V*) и определение диапазона жанровых различительных параметров (*эман VI*) рассчитывались в программе STATISTIKA³³.

³³ StatSoft. Statistika. URL : <https://statsoftstatistica.ru/> (дата обращения : 21.03.2024).

Результаты

Сравнительный анализ выявил ряд параметров, метрики которых в жанрах текстов биографии РКИ и приключений имеют статистически значимые различия, позволяющие дифференцировать тексты данных жанров. Выявлено, что динамичность повествования в приключенческих рассказах объективируется в более высокой частотности глаголов и наречий по сравнению с текстами биографий (см. строки 9 и 14 в табл. 3). Высокая номинативность и сопутствующая ей более высокая трудность восприятия текстов биографий по сравнению с текстами приключений эксплицирована в высокой частотности имен существительных (см. строку 10 в табл. 3). Биографиям также свойственна бо́льшая по сравнению с приключениями описательность, отражение которой находим в частотности прилагательных. Жанровая специфика биографий отражается также в дистрибуции падежных форм: доля имен существительных в предложном и родительном падежах значительно превышает другие падежи (см. строки 5 и 6 в табл. 3).

Наибольшее различие (от 100 до 314 %) выявлено для четырех параметров, а именно: глобальный и локальный повторы существительных, глобальный повтор аргумента, количество имен существительных в предложном и родительном падежах, дистрибуция глаголов прошедшего и настоящего времени. В текстах биографий выявлено почти в два раза больше повторов существительных и аргументов. Объясняются данные результаты тем, что учебные тексты биографий изобилуют описаниями этапов и достижений жизни выдающихся личностей.

Обсуждение

Проведенный статистический анализ показал значительные различия в метриках ряда параметров в текстах биографии и приключений (рис. 2).

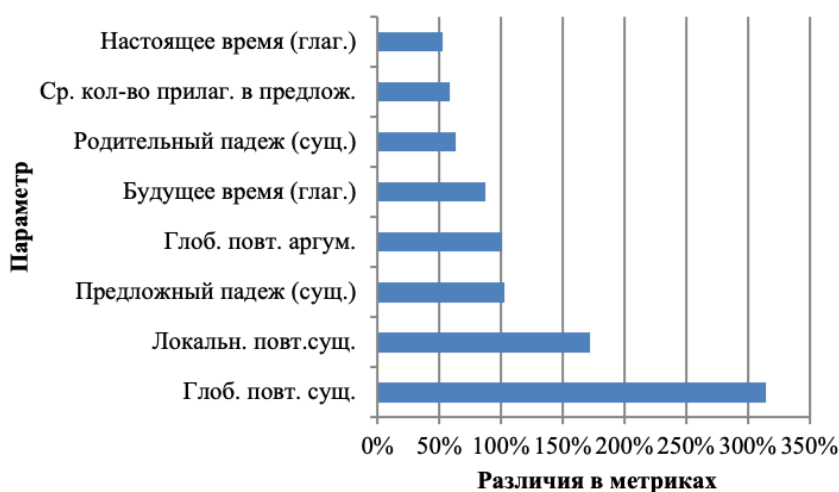


Рис. 2. Различия в лингвистических параметрах текстов биографий и приключений
Источники: Расчеты выполнены К.В. Ворониным, Ф.Х. Исмаевой, А.В. Даниловым на материале авторского корпуса исследования.

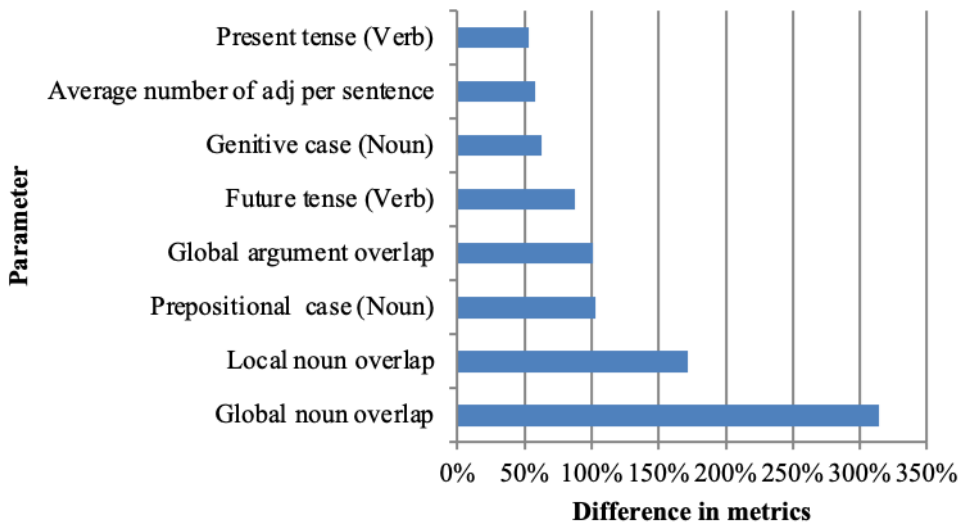


Figure 2. Differences in linguistic parameters of biographies and adventure stories
S o u r c e : Calculated by K.V. Voronin, F.Kh. Ismaeva, A.V. Danilov based on the Research corpus.

Дистрибуция частей речи показывает (а) преобладание глаголов (с диапазоном 136–212) и наречий (с диапазоном 50–98) в приключенческих рассказах, (б) преобладание существительных (с диапазоном 290,91–503,21) и прилагательных (81,82–226,87) в биографиях РКИ (табл. 4).

Таблица 4

Диапазон метрик частеречных параметров биографий и приключений

Параметр	Диапазон метрик		Стандартное отклонение	
	Приключения	Биография	Приключения	Биография
Глагол	136–212	89,74–201,47	16,29	22,91
Существительное	270–384	290,91–503,21	25,00	47,11
Прилагательное	53–134	81,82–226,87	16,77	26,47
Наречие	50–98	6,80–100	9,11	17,40
Соотношение глаг. / сущ.	0,38–0,70	0,18–0,68	0,07	0,09
Соотношение прил. / сущ.	0,16–0,44	0,24–0,46	0,05	0,05
Среднее количество сущ. в предл.	2,69–5,72	3,48–8,62	0,66	1,10
Среднее количество прил. в предл.	0,76–1,77	0,93–3,30	0,23	0,45
Среднее количество нареч. в предл.	0,52–1,42	0,08–1,50	0,17	0,24

Table 4

Metrics of parts-of-speech in biographies and adventure stories

Parameter	Metrics scope		Standard deviation	
	Adventure	Biography	Adventure	Biography
Verbs	136–212	89.74–201.47	16.29	22.91
Nouns	270–384	290.91–503.21	25.00	47.11
Adjectives	53–134	81.82–226.87	16.77	26.47
Adverbs	50–98	6.80–100	9.11	17.40
Verb/Noun ratio	0.38–0.70	0.18–0.68	0.07	0.09
Adjective/Noun ratio	0.16–0.44	0.24–0.46	0.05	0.05
Average number of nouns per sentence	2.69–5.72	3.48–8.62	0.66	1.10
Average number of adjectives per sentence	0.76–1.77	0.93–3.30	0.23	0.45
Average number of adverbs per sentence	0.52–1.42	0.08–1.50	0.17	0.24

Рассмотрим параметры, в которых выявлены наибольшие различия.

В целом диапазоны метрик локального повтора аргумента довольно широки как для приключений (0,04–0,57), так и для биографий РКИ (0,10–1,14). Однако в текстах биографий содержится почти в два раза больше повторов существительных и аргументов (табл. 5, рис. 3).

Таблица 5

Диапазон значений дискурсивных параметров биографий и приключений

Параметр	Диапазон метрик		Стандартное отклонение	
	Приключения	Биография	Приключения	Биография
Глобальный повтор сущ.	0,01–0,09	0,04–0,39	0,02	0,07
Локальный повтор сущ.	0,00–0,24	0,03–0,62	0,05	0,13
Глобальный повтор аргумента	0,06–0,29	0,13–0,57	0,06	0,10
Локальный повтор аргумента	0,04–0,57	0,10–1,14	0,13	0,21

Table 5

Discourse parameters of biographies and adventure stories

Parameter	Metrics scope		Standard deviation	
	Adventure	Biography	Adventure	Biography
Global noun overlap	0.01–0.09	0.04–0.39	0.02	0.07
Local noun overlap	0.00–0.24	0.03–0.62	0.05	0.13
Global argument overlap	0.06–0.29	0.13–0.57	0.06	0.10
Local argument overlap	0.04–0.57	0.10–1.14	0.13	0.21

Выявленное превалирование повторов обеспечивает большую степень связности текста биографии. Показательно, что установленные нами диапазоны метрик биографий подтверждают более ранние наблюдения (Терпугова, 2011; Ефремова, 2016), которые, правда, не были подтверждены количественно. Связность текста, реализуемая при помощи повторов, является ключевым параметром для жанра биографии, поскольку изложение жизненного пути человека предполагает последовательную номинацию во взаимосвязи со всеми ступенями его достижений. Например, начало события А, завершение события А, начало события Б, завершение события Б и т.д.

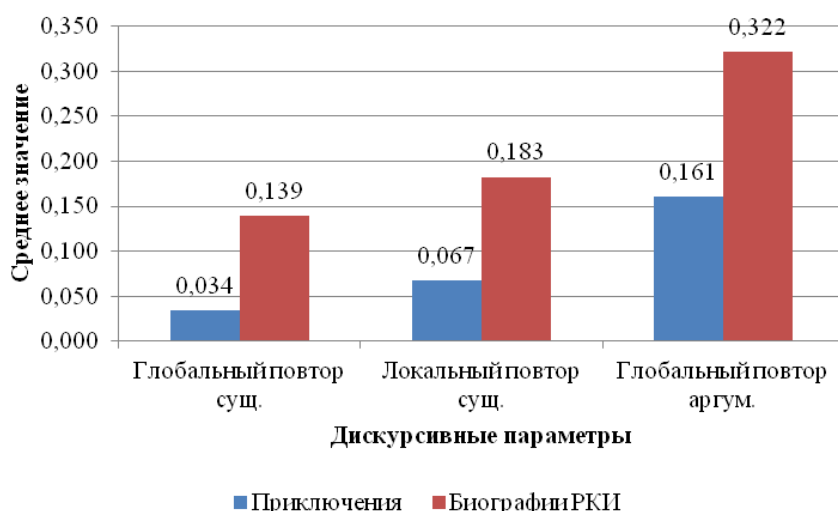


Рис. 3. Различия в дискурсивных параметрах текстов жанров «биография» и «приключение»

И с т о ч н и к : Расчеты выполнены К.В. Ворониным, Ф.Х. Исмаевой, А.В. Даниловым на материале авторского корпуса исследования.

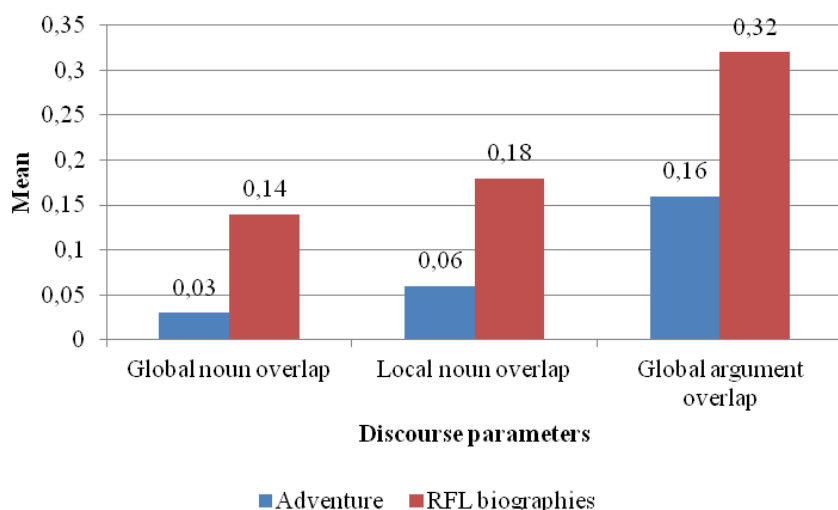


Figure 3. Differences in discourse parameters of biographies and adventure stories

S o u r c e : Calculated by K.V. Voronin, F.Kh. Ismaeva, A.V. Danilov based on the Research corpus.

Проиллюстрируем глобальные и локальные повторы существительных фрагментом из биографии И.И. Мечникова из учебника РКИ уровня С1 (код файла *Sa_C1_Bio_Mechnikov_2014_195*). Отметим повторы существительных курсивом.

«И.И. Мечников — известный русский *микробиолог* и патолог, основоположник эволюционной эмбриологии, сравнительной патологии, иммунологии, геронтологии. Свою научную деятельность знаменитый русский *микробиолог* И.И. Мечников начал с изучения жизни простейших животных. В 22 *года* Мечников защитил магистерскую *диссертацию*, а через *год* — докторскую *диссертацию*. В 1883 *году* в Одессе состоялся *съезд* врачей и естествоиспытателей. Председателем *съезда* был И.И. Мечников. По приглашению Луи Пастера И.И. Мечников согласился приехать для продолжения своих исследований в организованный в Париже *институт*, где работал с 1888 *года*, а с 1904 *года* руководил этим *институтом* до конца своих дней³⁴».

Повторы в учебном тексте не только обеспечивают высокую связность текста, но и выполняют весьма важную роль для понимания и запоминания текста обучающимися (McNamara, 2014).

Отдельного внимания заслуживает широкий диапазон метрик родительного падежа в текстах обоих жанров (табл. 6, рис. 4). Как видим, метрики дистрибуции существительных в родительном и предложном падежах в текстах приключений значительно ниже: 13–42 вхождения существительных в предложном падеже на 1000 словоупотреблений в приключениях против 22–93 в биографиях. 50–135 — диапазон значений родительного падежа существительных в приключениях против 40–229 в биографиях.

Таблица 6

Количество вхождений категории «Падеж имен существительных» в биографиях и приключениях

Параметр	Диапазон метрик		Стандартное отклонение	
	Приключения	Биография	Приключения	Биография
Предложный падеж (сущ.)	13–42	22–93	6,31	17,7
Родительный падеж (сущ.)	50–135	40–229	12,3	35,7

Table 6

Occurrences of category “Case of nouns” in biographies and adventures stories

Parameter	Metrics scope		Standard deviation	
	Adventure	Biography	Adventure	Biography
Prepositional case (Noun)	13–42	22–93	6.31	17.7
Genitive case (Noun)	50–135	40–229	12.3	35.7

³⁴ Санникова А.В. Русские ученые-медики : учеб.-метод. пособие. Минск : БГМУ, 2014. 48 с.

Большее количество существительных в предложном и родительном падежах в текстах биографий РКИ также является жанровой спецификой и свидетельствует о более высокой степени номинативности текстов биографий, в которых авторы отдают предпочтения конструкциям *сущ. + сущ. (род)* вместо придаточных предложений. Например, «основоположник эволюционной эмбриологии» вместо **были заложены основы эмбриологии*, «с изучения жизни простейших животных» вместо **изучал жизнь простейших* и т.д. (Gatiyatullina et al., 2020).

Что касается установленных нами различий в метриках предложного падежа, то они, насколько нам известно, выявлены впервые, а их верификацию в корпусе большого объема мы рассматриваем как одну из перспектив исследования.

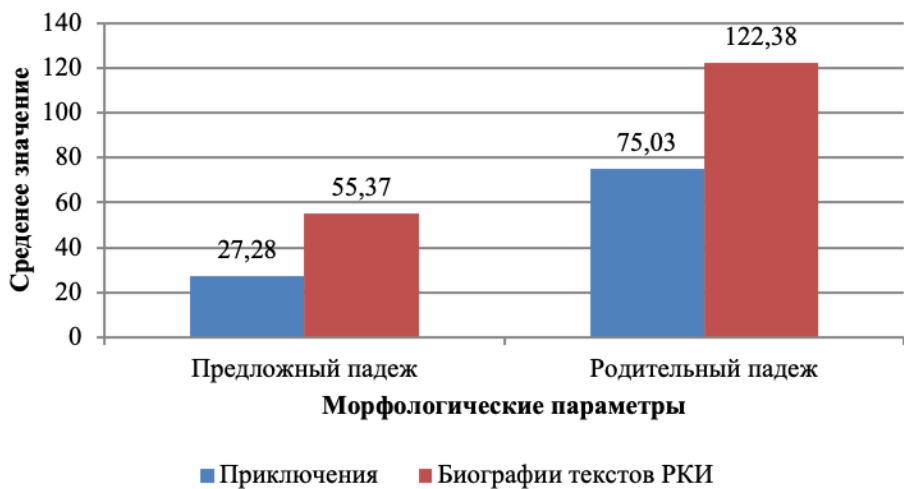


Рис. 4. Различия в количестве вхождений параметров «падеж» в текстах жанров «биография» и «приключение»

И с т о ч н и к : Расчеты выполнены К.В. Ворониным, Ф.Х. Исмаевой, А.В. Даниловым на материале авторского корпуса исследования.

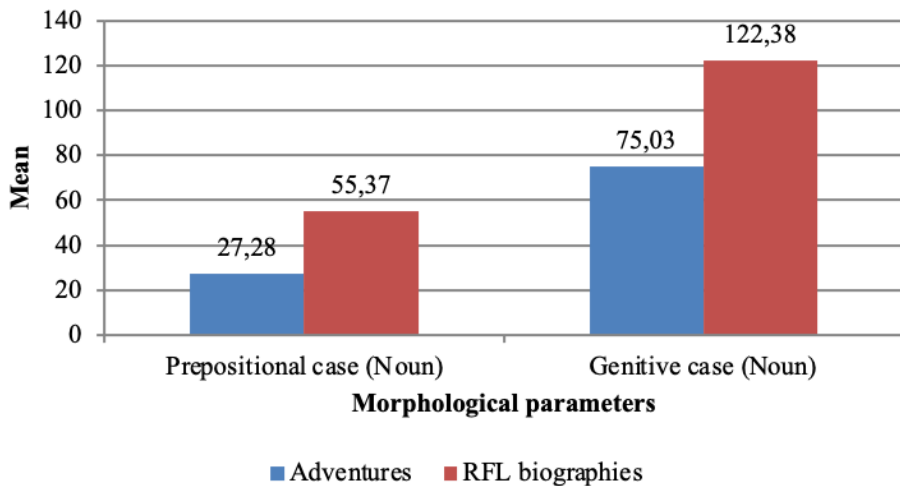


Figure 4. Differences in Noun cases parameters in biographies and adventure stories
S o u r c e : Calculated by K.V. Voronin, F.Kh. Ismaeva, A.V. Danilov based on the Research corpus.

Определено, что количество глаголов в прошедшем времени в текстах обоих жанров не является параметром-дискриминантом в силу отсутствия статистически значимых различий между значениями: 98,78 vs 101,26 (табл. 6, рис. 5).

Таблица 6

Диапазон метрик времен глаголов биографий текстов РКИ и приключений

Параметр	Диапазон метрик		Стандартное отклонение	
	Приключения	Биография	Приключения	Биография
Будущее время	0–21	0–9,09	3,89	1,94
Настоящее время	21–91	0–54,05	14,25	13,18
Прошедшее время	45–147	44,84–142,86	21,86	22,09

Table 6

Verb tenses in biographies and adventure stories

Parameter	Metrics scope		Standard deviation	
	Adventure	Biography	Adventure	Biography
Future tense (Verb)	0–21	0–9.09	3.89	1.94
Present tense (Verb)	21–91	0–54.05	14.25	13.18
Past tense (Verb)	45–147	44.84–142.86	21.86	22.09

Очевидным представляются и низкие значения дистрибуции глаголов в будущем (0,83) и настоящем (19,45) времени в биографических текстах. Изложение материала в прошедшем времени является одной из важнейших характеристик биографического текста (Ян, 2023).

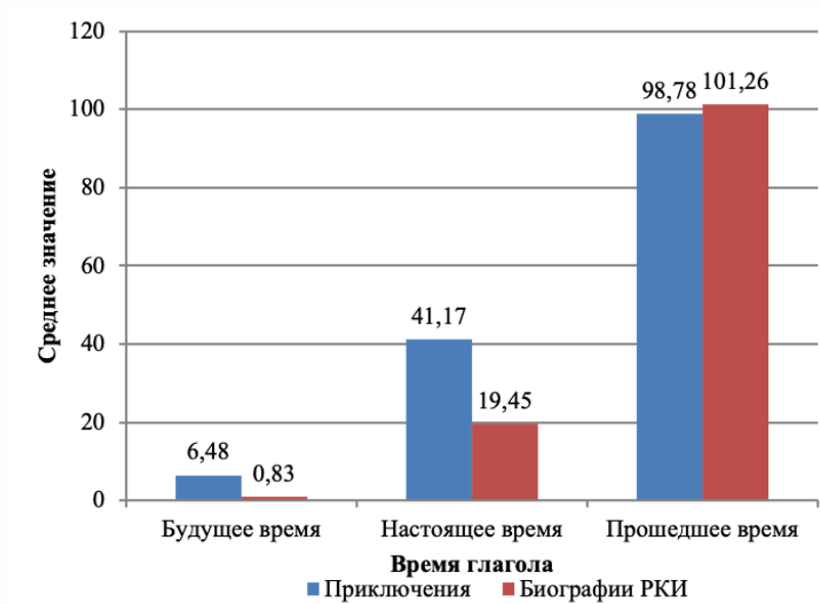


Рис. 5. Различия в количестве вхождений параметров «времени глагола» в текстах приключений и биографий

И с т о ч н и к : Расчеты выполнены К.В. Ворониным, Ф.Х. Исмаевой, А.В. Даниловым на материале авторского корпуса исследования.

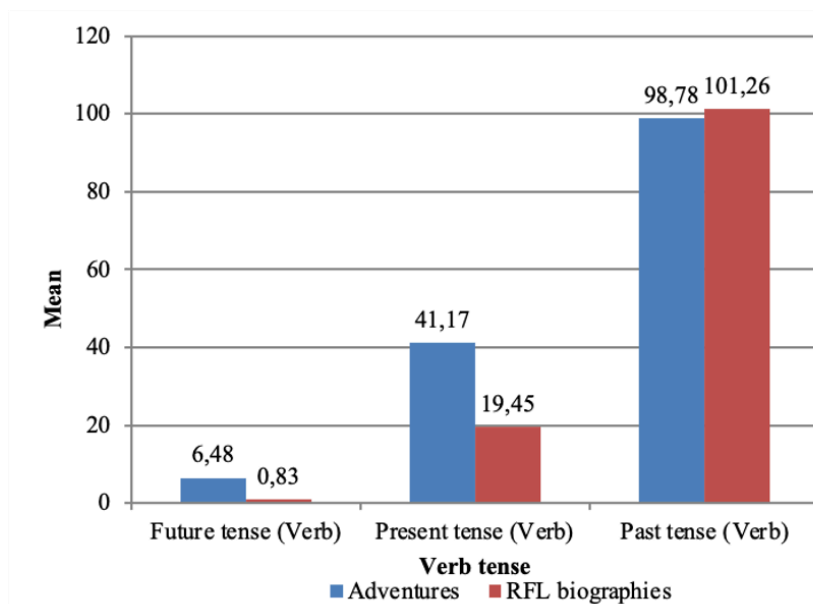


Figure 5. Differences in Verb tense parameters in biographies and adventure stories
Source : Calculated by K.V. Voronin, F.Kh. Ismaeva, A.V. Danilov based on the Research corpus.

Полученные результаты указывают на то, что диапазоны значений глобальных и локальных повторов существительных и аргумента, вхождения предложного и родительного падежей существительных, а также глаголов в будущем и настоящем времени являются жанро-различительными для текстов биографий и приключений.

Заклучение

Исследование позволило определить список и диапазоны дискурсивных и морфологических параметров, характерных для жанров биографии и приключений. Список параметров-дискриминантов включает следующие: глобальный и локальный повторы существительных, глобальный повтор аргумента, предложный и родительный падежи имен существительных, количество вхождений глаголов в будущем и настоящем времени.

Перспективы дальнейших исследований видятся в расширении корпуса исследования, например, в изучении специфики энциклопедических и словарных биографий. Особое внимание при изучении первичных и вторичных текстов биографий требуется при параметризации морфологических категорий в целом и категории падежа имен существительных в частности. Интересным представляется сравнение полученных метрик для жанра биографий текстов РКИ и академических текстов, например, учебных текстов по истории. Отдельного внимания заслуживает определение диапазона жанродифференцирующих параметров текстов других жанров, содержащихся в учебниках РКИ, например, диалогов, фрагментов энциклопедических текстов и т.д. Полученные данные о диапазоне метрик текстов РКИ могут быть использованы для расширения функций текстовых профайлеров русского языка.

Список литературы

- Абдулвахиду М.А. Литературный травелог: специфика жанра // Филология и культура. 2014. № 3 (37). С. 254–259.
- Бахтин М.М. Вопросы литературы и эстетики. Исследования разных лет. М. : Худож. лит., 1975. 504 с.
- Вахрушева А.Я., Солнышкина М.И., Курьянов Р.В., Гафиятова Э.В., Климагина И.О. Лингвистическая сложность учебных текстов // Вопросы журналистики, педагогики, языкознания. 2021. Т. 40. № 1. С. 88–99. <https://doi.org/10.18413/2712-7451-2021-40-1-89-99>
- Ефремова Д.А. Лингвостилистические средства выражения модальности в тексте биографии (на материале англоязычных текстов) : автореф. дис. ... канд. филол. наук. М., 2016. 22 с.
- Кожина М.Н. Некоторые аспекты изучения речевых жанров в нехудожественных текстах // Стереотипность и творчество в тексте. Пермь : ПГНИУ, 1999. С. 22–39.
- Литвинова Т.А., Лантюхова Н.Н., Рыжкова Е.С., Шевченко И.С. Профилирование автора текста как одно из стратегических направлений исследований // Вестник Воронежского института ГПС МЧС России. 2013. Т. 1. № 6. С. 38–41.
- Савирова М.П. Сравнительно-типологические особенности приключенческих жанров в литературоведении Урало-Поволжья // Национальные языки и литературы в поликультурных условиях. Ч 2. Чебоксары : Чувашский государственный педагогический университет им. И.Я. Яковлева, 2017. С. 129–131.
- Сиротинина О.Б. Разговорная речь в системе функциональных стилей современного русского литературного языка: Лексика. Саратов : Изд-во Саратов. ун-та, 1983. 256 с.
- Соловьев В.Д., Вольская Ю.А., Андреева М.И., Заикин А.А. Словарь русского языка с индексами конкретности / абстрактности // Russian journal of linguistics. 2022. Т. 26. № 2. С. 515–549. <https://doi.org/10.22363/2687-0088-29475>
- Стаценко А.С. Структура и критерии дифференциации языкового жанра // Филологические науки. Вопросы теории и практики. 2016. № 5–2 (59). С. 32–34.
- Терпугова А.В. Биографический текст как объект лингвистического исследования : автореф. дис. ... канд. филол. наук. М., 2011. 26 с.
- Федосюк М.Ю. Нерешенные вопросы теории речевых жанров // Вопросы языкознания. 1997. № 5. С. 102–121.
- Холиков А.А. Жанровый потенциал биографии литературоведа // Новый филологический вестник. 2016. № 4 (39). С. 46–51.
- Чурунина А.А., Солнышкина М.И., Ярмачев И.Э. Лексическое разнообразие как предиктор сложности учебников по русскому языку // Русистика. 2023. Т. 21. № 2. С. 212–227. <https://doi.org/10.22363/2618-8163-2023-21-2-212-227>
- Ян Ж.Ж. Жанрово-стилистические характеристики биографического текста в русской лингвокультуре : автореф. дис. ... канд. филол. наук. М., 2023. 23 с.
- Andreeva M., Solnyshkina M., Bukach O., Zaikin A., Zamaletdinov R. Assessment of comparative abstractness : Quantitative approach // CEUR Workshop Proceedings. Kazan, 2020. Pp. 132–144.
- Biber D. Spoken and written textual dimensions in English : Resolving the contradictory findings // Language. 1986. Vol. 62. No. 2. Pp. 384–414. <https://doi.org/10.2307/414678>
- Biber D. University language : a corpus-based study of spoken and written registers. Amsterdam : John Benjamins, 2006. <https://doi.org/10.1075/sci.23>
- Biber D. Variation across speech and writing. Cambridge : Cambridge University Press, 1988. <https://doi.org/10.1017/CBO9780511621024>
- Brunato D., Cimino A., Dell'Orletta F., Venturi G., Montemagni S. Profiling-ud : a tool for linguistic profiling of texts // Proceedings of the twelfth language resources and evaluation

- conference. Marseille : European Language Resources Association, 2020. Pp. 7145–7151.
- Dell’Orletta F., Montemagn S., Ventur G.* Linguistic profiling of texts across textual genre and readability level. An exploratory study on Italian fictional prose // Proceedings of the recent advances in natural language processing conference. 2013. Pp. 189–197.
- 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>
- Ismaeva F., Tomin E., Sharifullina E.* Comparison of algorithms for automatic terminology extraction on material of educational texts on biology // Proceeding of the 33rd conference of FRUCT associationistics. Helsinki : FRUCT Oy, 2023. Pp. 95–100. <http://doi.org/10.23919/FRUCT58615.2023.10143073>
- Ivanov V., Solnyshkina M.* A method for assessment of text complexity based on knowledge graphs // CEUR Workshop Proceedings. 2020. Vol. 2852.
- Manning C., Schütze H.* Foundations of statistical natural language processing. Cambridge, MA ; London : MIT Press, 1999. 720 p.
- McNamara D.S.* Automated evaluation of text and discourse with Coh-Metrix. Cambridge : Cambridge University Press, 2014. 289 p. <https://doi.org/10.1017/CBO9780511894664>
- Mendhakar A.* Linguistic profiling of text genres: Linguistic profiling of text genres: An exploration of fictional vs. non-fictional texts // Information. 2022. Vol. 13. No. 8. P. 357. <https://doi.org/10.3390/info13080357>
- Montemagni S., Wieling M., de Jonge B., Nerbonne J.* Synchronic patterns of Tuscan phonetic variation and diachronic change. Evidence from a dialectometric study // Literary and linguistic computing. 2013. Vol. 28. No. 1. Pp. 157–172. <https://doi.org/10.1093/llc/fqs057>
- Paltridge B.* Genre analysis and the identification of textual boundaries // Applied linguistics. 1994. Vol. 15. No. 3. Pp. 288–299.
- 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., Ivanov V., Solovyev V.* Readability formula for Russian texts: a modified version // Advances in Computational Intelligence. MICAI 2018. Lecture Notes in Computer Science. 2018. Vol 11289. Pp. 132–145. https://doi.org/10.1007/978-3-030-04497-8_11
- Solnyshkina M.I., Kupriyanov R.V., Shoeva G.N.* Linguistic profiling of text genres : adventure stories vs. textbooks // Research result. Theoretical and applied linguistics. 2024. Vol. 10. No. 1. Pp. 115–132. <https://doi.org/10.18413/2313-8912-2024-10-1-0-7>
- Swales J.M.* Research genres : Explorations and applications. Cambridge : Cambridge University Press, 2004. 314 p. <https://doi.org/10.1017/CBO9781139524827>
- Van Halteren H.* Linguistic profiling for author recognition and verification // Proceedings of the association for computational linguistics. 2004. Pp. 199–206. <http://doi.org/10.3115/1218955.1218981>

Сведения об авторах:

Воронин Константин Валерьевич, ассистент кафедры теории и практики преподавания иностранных языков, инженер НИЛ «Мультидисциплинарные исследования текста», Казанский (Приволжский) федеральный университет, Российская Федерация, 420008, Республика Татарстан, г. Казань, ул. Кремлевская, д. 18, корп. 1. *Сфера научных интересов:* сложность текста, корпусная лингвистика, учебный текст. E-mail: voronin.konstantin@outlook.com

Исмаева Фарида Хамисовна, кандидат филологических наук, доцент, доцент кафедры теории и практики преподавания иностранных языков, Казанский (Приволжский) федеральный университет, Российская Федерация, 420008, Республика Татарстан, г. Казань, ул. Кремлевская, д. 18, корп. 1. *Сфера научных интересов*: сложность текста, жанры текста, профессиональная терминология, лингвистическое профилирование. ORCID: 0000-0003-4496-0700. SPIN-код: 4728-3163. Researcher ID: B-5420-2016. Scopus ID: 57191851333. E-mail: fismaeva@yandex.ru

Данилов Андрей Владимирович, кандидат педагогических наук, доцент кафедры билингвального и цифрового образования, старший научный сотрудник НИЛ «Мультидисциплинарные исследования текста», Казанский (Приволжский) федеральный университет, Российская Федерация, 420008, Республика Татарстан, г. Казань, ул. Кремлевская, д. 18, корп. 1. *Сфера научных интересов*: параметризация текста, сложность текста, русский язык как иностранный, алгоритмы автоматизированного анализа текста. ORCID: 0000-0002-2358-1157. SPIN-код: 8525-5480. Researcher ID: L-8745-2013. Scopus ID: 57008755500. E-mail: tukai@yandex.ru

DOI: 10.22363/2618-8163-2024-22-4-555-578

EDN: AWRLUL

Research article

Linguistic profiling of educational and artistic texts

Konstantin V. Voronin^{ID}✉, Farida H. Ismaeva^{ID}, Andrew V. Danilov^{ID}

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

✉ voronin.konstantin@outlook.com

Abstract. Implemented within text analytics as one of the strategic directions of modern Russian linguistics, the research focuses on the linguistic profiling of educational and artistic texts. The identified genre features contribute to software systems development and big language data processing. The study is aimed at revealing the ranges of linguistic parameters which differentiate educational (secondary) and artistic (primary) texts. The study was based on 72 biographies from textbooks on Russian as a foreign language and 90 excerpts from adventure stories. The genres were chosen for the contrastive study for the following reasons: (1) a high degree of narrativity and actions; (2) functional differences, namely, informative function of biography and entertaining function of adventure stories. The research corpus comprised 120932 words. Two data processing tools were used in the study: (1) the linguistic parameters were calculated using RuLingva platform, (2) STATISTIKA program was used to identify statistically significant differences between the two genres. The revealed genre-specific parameters include global and local overlaps of nouns and arguments, nouns in prepositional and genitive cases, past and present verb forms. Global and local overlaps of nouns and arguments were found to contribute to high cohesion of biographies, because the nomination of each next event in the person's life is connected with the previous one. The genitive case prevails in biographies due to the use of nominative word combinations. The research perspective is seen in the typology of genres based on linguistic profiling of official and personal biographies, on the one hand, and adventure novels, on the other hand. An important aspect of further research in the studied area is the installation of meaning matrices in the RuLingva text profiler to automate text linguistic profiling.

Keywords: Russian as a foreign language, biography, adventure story, research corpus, text complexity, genre, text cohesion

Contribution: *Voronin K.V.* — research concept; *Ismaeva F.H.* — collection and processing of materials, writing the text; *Danilov A.V.* — analysis of the data obtained.

Funding. The research was supported by the RSF, grant 24-78-10129 ‘Cognitive and linguistic text complexity’.

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

Article history: received 25.03.2024; accepted 11.06.2024.

For citation: Voronin, K.V., Ismaeva, F.H., & Danilov, A.V. (2024). Linguistic profiling of educational and artistic texts. *Russian Language Studies*, 22(4), 555–578. (In Russ.). <http://doi.org/10.22363/2618-8163-2024-22-4-555-578>

References

- Abdulahidu, M.A. (2014). Literary travelogue: genre particularities. *Philology and Culture*, (3), 254–259. (In Russ.).
- Andreeva, M., Solnyshkina, M., Bukach, O., Zaikin, A., & Zamaletdinov, R. (2020). Assessment of comparative abstractness: Quantitative approach. In *CEUR Workshop Proceedings* (pp. 132–144). Kazan.
- Bakhtin, M.M. (1975). Questions of literature and aesthetics. Studies of different years. Moscow: Khudozhestvennaya Literatura Publ. (In Russ.).
- Biber, D. (1986). Spoken and written textual dimensions in English: Resolving the contradictory findings. *Language*, 62(2), 384–414. <https://doi.org/10.2307/414678>
- Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511621024>
- Biber, D. (2006). *University language: a corpus-based study of spoken and written registers*. Amsterdam: John Benjamins Publ. <https://doi.org/10.1075/scl.23>
- Brunato, D., Cimino, A., Dell’Orletta, F., Venturi, G., & Montemagni, S. (2020). Profiling-ud: a tool for linguistic profiling of texts. In *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 7145–7151). Marseille: European Language Resources Association.
- Churunina, A.A., Solnyshkina, M.I., & Yarmakeev, I.E. (2023). Lexical diversity as a predictor of the complexity of textbooks on the Russian language. *Russian Language Studies*, 21(2), 212–227. (In Russ.). <https://doi.org/10.22363/2618-8163-2023-21-2-212-227>
- Dell’Orletta, F., Montemagni, S., & Venturi, G. (2013). Linguistic profiling of texts across textual genre and readability level. An exploratory study on Italian fictional prose. In *Proceedings of the Recent Advances in Natural Language Processing Conference (RANLP-2013)* (pp. 189–197). RANLP 2013 Organising Committee Publ.
- Efremova, D.A. (2016). Linguistic stylistic means of expressing modality in the text of a biography (based on the material of English-language texts). [Author’s abstr. cand. philol. diss.]. Moscow. (In Russ.).
- Fedosyuk, M.Y. (1997). Unresolved issues of the theory of speech genres. *Voprosy Jazykoznanania*, (5), 102–121. (In Russ.).
- Gatiyatullina, G., Solnyshkina, M., Solovyev, V., Danilov, A., Martynova, E., & Yarmakeev, I. (2020). Computing Russian morphological distribution patterns using RusAC Online Server. In *Proceedings of the International Conference on Developments in eSystems Engineering (DeSE 2020)* (pp. 393–398). <https://doi.org/10.1109/DeSE.2020.9450753>
- Holikov, A.A. (2016). Genre potential of the biography of a literary critic. *The New Philological Bulletin*, (4), 46–51. (In Russ.).

- Ismaeva, F., Tomin, E., & Sharifullina, E. (2023). Comparison of algorithms for automatic terminology extraction on material of educational texts on biology. In *Proceeding of the 33rd Conference of FRUCT Associationistics* (pp. 95–100). Helsinki: FRUCT Oy. <http://doi.org/10.23919/FRUCT58615.2023.10143073>
- Ivanov, V., & Solnyshkina, M. (2020). A method for assessment of text complexity based on knowledge graphs. In *CEUR Workshop Proceedings*. CEUR-WS.
- Jan, J.J. (2023). *Genre and stylistic characteristics of a biographical text in Russian linguoculture*. [Author's abstr. cand. philol. diss.]. Moscow. (Russ.).
- Kozhina, M.N. (1999). Some aspects of speech genres study in non-fiction texts. In *Stereotypes and art in text* (pp. 22–39). Perm: PSU Publ. (In Russ.).
- Litvinova, T.A., Lantukhova, N.N., Ryzhkova, E.S., & Shevchenko, I.S. (2013). Profiling the author of the text as one of the strategic directions of research. *Vestnik Voronezhskogo instituta GPS MChS Rossii*, 1(6), 38–41. (In Russ.).
- Manning, C., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge (MA): MIT Press.
- McNamara, D.S. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511894664>
- Mendhakar, A. (2022). Linguistic profiling of text genres: Linguistic profiling of text genres: An exploration of fictional vs. non-fictional texts. *Information*, 13(8), 357. <https://doi.org/10.3390/info13080357>
- Montemagni, S., Wieling, M., de Jonge B., & Nerbonne, J. (2013). Synchronic patterns of Tuscan phonetic variation and diachronic change. Evidence from a dialectometric study. *Literary and Linguistic Computing*, 28(1), 157–172. <https://doi.org/10.1093/lilc/fqs057>
- Paltridge, B. (1994). Genre analysis and the identification of textual boundaries. *Applied Linguistics*, 15(3), 288–299.
- Sakhovskiy, A., Solovyev, V., & Solnyshkina, M. (2020). Topic modeling for assessment of text complexity in Russian textbooks. In *2020 Ivannikov Ispras Open Conference (ISPRAS)* (pp. 102–108). IEEE. <https://doi.org/10.1109/ISPRAS51486.2020.00022>
- Savirova, M.P. (2017). Comparative typological features of adventure genres in literary studies of the Ural-Volga region. In *National languages and literatures in multicultural conditions* (pp. 129–131). Cheboksary: Yakovlev Chuvash State Pedagogical University Publ. (In Russ.).
- Sirotnina, O.B. (Ed.). (1983). *Colloquial speech in the system of functional styles of the modern Russian literary language: Vocabulary*. Saratov: Saratov University Press. (In Russ.).
- Solnyshkina, M., Ivanov, V., & Solovyev, V. (2018). Readability formula for Russian texts: a modified version. In *Advances in Computational Intelligence. MICAI 2018. Lecture Notes in Computer Science* (pp. 132–145), 11289. Springer, Cham. https://doi.org/10.1007/978-3-030-04497-8_11
- Solnyshkina, M.I., Kupriyanov, R.V., & Shoeva, G.N. (2024). Linguistic profiling of text genres: adventure stories vs. textbooks. Research Result. *Theoretical and Applied Linguistics*, 10(1), 115–132. <https://doi.org/10.18413/2313-8912-2024-10-1-0-7>
- Solovyev, V., Volskaya, Y., Andreeva, M., & Zaikin, A. (2022). Russian dictionary with concreteness/abstractness indices. *Russian Journal of Linguistics*, 26(2), 515–549. (In Russ.). <https://doi.org/10.22363/2687-0088-29475>
- Statsenko, A.S. (2016). Structure and criteria of differentiation of the language genre. *Philology. Theory and practice*, (5-2), 32–34. (In Russ.).
- Swales, J.M. (2004). *Research genres: Explorations and applications*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9781139524827>
- Terpugova, A.V. (2011). Biographical text as an object of linguistic research. [Author's abstr. cand. philol. diss.]. Moscow. (In Russ.).

- Vakhrusheva, A.Y, Solnyshkina, M.I., Kupriyanov, R.V., Gafiyatova, E.V., & Klimagina, I.O. (2021). Linguistic complexity of educational texts. *Issues of journalism, education, linguistics*, 40(1), 88–99. <http://doi.org/10.52575/2712-7451-2021-40-1-89-99>
- Van Halteren, H. (2004). Linguistic profiling for author recognition and verification. In *Proceedings of the Association for Computational Linguistics* (pp. 199–206). <http://doi.org/10.3115/1218955.1218981>

Bio notes:

Konstantin V. Voronin, assistant at the Department of theory and practice of teaching foreign languages, engineer of the research laboratory ‘Multidisciplinary Text Investigation’, Kazan (Volga region) Federal University, 18 Kremlevskaya St, Kazan, 420008, Russian Federation. *Research interests*: text complexity, corpus linguistics, academic text. E-mail: voronin.konstantin@outlook.com

Farida H. Ismaeva, PhD in Philology, Associate Professor, Associate Professor of the Department of theory and practice of teaching foreign languages, Kazan (Volga region) Federal University, 18 Kremlevskaya St, Kazan, 420008, Russian Federation. *Research interests*: text complexity, text genres, professional terminology, linguistic profiling. ORCID: 0000-0003-4496-0700. SPIN-code: 4728-3163. Researcher ID: B-5420-2016. Scopus ID: 57191851333. E-mail: fismaeva@yandex.ru

Andrew V. Danilov, PhD in Philology, Associate Professor of the Department of bilingual and digital education, Senior researcher of the research laboratory ‘Multidisciplinary Text Investigation’, Kazan (Volga region) Federal University, 18 Kremlevskaya St, Kazan, 420008, Russian Federation. *Research interests*: text parametrization, text complexity, Russian as a foreign language, algorithms of automated text analysis. ORCID: 0000-0002-2358-1157. SPIN-code: 8525-5480. Researcher ID: L-8745-2013. Scopus ID: 57008755500. E-mail: tukai@yandex.ru