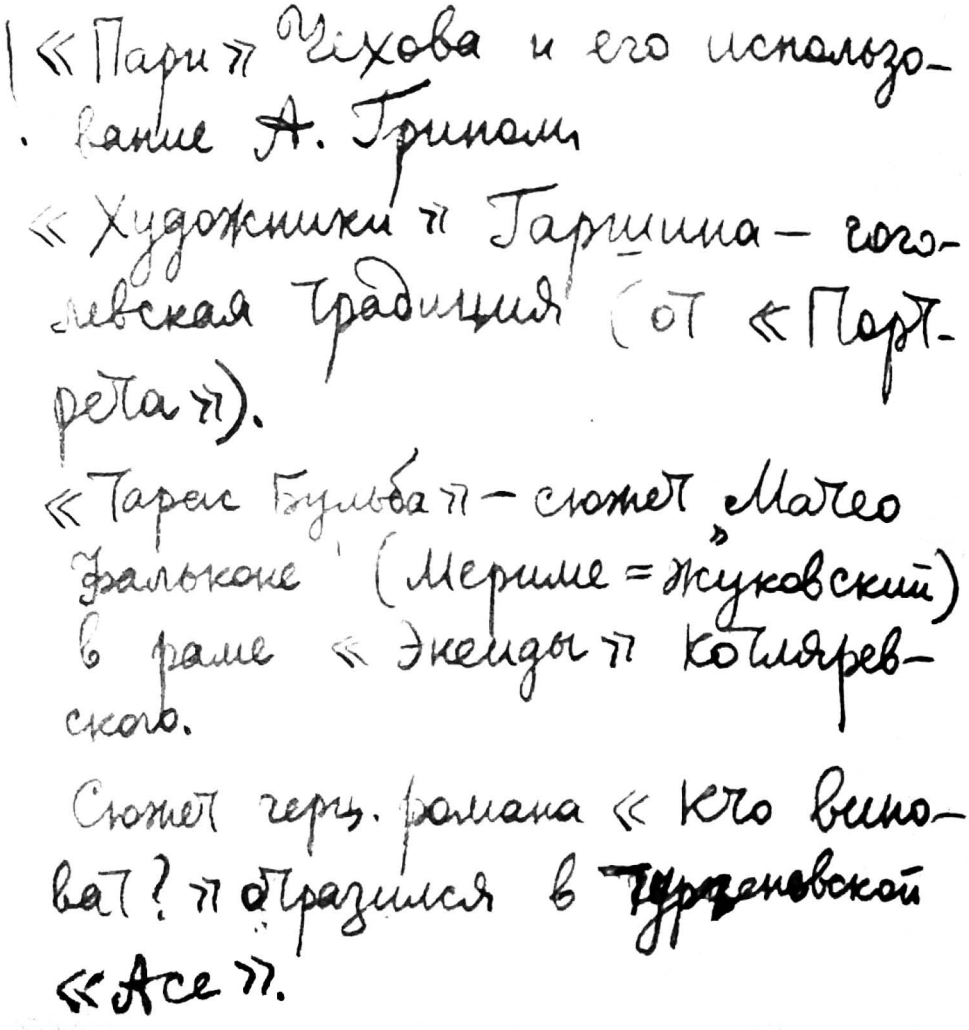
Исследования



Исследования

Формальная оценка удобочитаемости текстов Р. Г. Назирова

К. В. Самойленко

Национальный исследовательский университет «Высшая школа экономики»

1 Введение

Научные работы Р. Г. Назирова охватывают самый широкий круг тем: от русской литературы XIX века до мифологии. Его труды не только представляют интерес для современных исследователей-специалистов, но также, как считается, могут быть понятны и неподготовленному читателю. Тексты Назирова характеризуются «прозрачным, лапидарным стилем, не допускавшем ничего лишнего, в том числе, и избыточной терминологии»[1]. О Ромэне Гафановиче вспоминают как о великолепном лекторе, который мог увлекательно рассказывать о самых сложных вещах. Это нашло отражение и в его работах, которые как бы противопоставляются привычному научному, академическому стилю письма, с целью вовлечь читателя в материал не только интеллектуально, но и эмоционально[2].

Заметная часть наследия Р. Г. Назирова оцифрована. Это позволяет применять к его работам различные инструменты компьютерной лингвистики, в том числе те, которые дают возможность оценить, насколько в действительности просто устроены тексты Назирова, и сравнить их с работами других авторов. Один из таких инструментов — различные метрики удобочитаемости. Полученные результаты анализа текста посредством этих метрик можно соотнести с читательской оценкой «читабельности» текстов.

Удобочитаемость (readability) — это величина, показывающая, насколько понятным и легким для прочтения является текст. Чаще всего оценки удобочитаемости опираются на различные статистические характеристики: среднюю длину предложения, среднюю длину слов, количество «сложных» слов (то есть с большим количеством слогов), и так далее [3].

[1] Рыбина М. С. Предисловие к публикации [Спор Достоевского с Кальдероном] // Назировский сборник. Уфа, 2011. С. 43.

[2] Орехов Б. В., Шаулов С. С. Сумма мифологии // Назиров Р. Г. Становление мифов и их историческая жизнь. Уфа, 2014. С. 8.

[3] Crossley Sc. A., Skalicky St., Dascalu M., McNamara D. S., Kyle K. Predicting Text Comprehension, Processing, and Familiarity in Adult Readers: New Approaches to Readability Formulas // Discourse Processes. Vol. 54, 2017.

Первые метрики такого рода появились еще в середине XX века. Способы оценки ридабилити создавались как инструмент для подбора текстов в образовательных целях: для школьной программы и обучения иностранному языку. Большинство популярных метрик ранжирует тексты как раз согласно уровням школьного образования: 1 — понятный первокласснику, 5 — пятикласснику, 10 — ученику выпускных классов, больше 12 — студенту университета и старше. Для большинства популярных метрик эти уровни ассоциированы с классами американской школы, а при ранжировании текстов для обучения иностранному языку они соотносятся с уровнями владения языком (традиционные A1, A2, B1 и так далее)[4]. Современные исследования в этой области направлены на поиск новых характеристик, помогающих точнее определять уровень сложности текста.

Метрики удобочитаемости позволяют сравнивать тексты и по количественным признакам определять, какие из них являются более, а какие менее сложными.

Чтобы проверить гипотезу, что тексты Р. Г. Назирова действительно отличаются от других научных сочинений большей легкостью в чтении, мы решили сравнить средние значения различных метрик удобочитаемости в его сочинениях и в работах других авторов.

Корпус текстов Назирова включает его дневниковые записи, работы о мифологии, труды о творчестве Ф. М. Достоевского и общие работы об истории литературы. Мы изучили внутреннее разнообразие текстов Назирова, а также сравнили его тексты по мифологии и достоевистике с работами других исследователей по этим темам. Тексты по мифологии мы сопоставляли с работами Елеазара Моисеевича Мелетинского, специалиста по истории культуры и фольклору (1918 – 2005), и Ольги Михайловны Фрейденберг, антиковеда и историка словесности (1890 – 1955). Работы о Достоевском будут сравниваться с трудами философа и культуролога Михаила Михайловича Бахтина (1895 – 1975) и литературоведа Аркадия Семеновича Долинина (1880 – 1968) .

Тексты анализировались по следующим признакам:

- длина текста в символах
- длина текста в словах
- количество предложений в тексте
- среднее количество слов в предложении (и обратная пропорция — «количество предложений на слово»)
- среднее количество символов в словах
- среднее количество символов в предложениях
- количество слогов в тексте
- среднее количество слогов в предложениях

[4]Gallagher T., Fazio X., Ciampa K. A Comparison of Readability in Science-Based Texts: Implications for Elementary Teachers // Canadian Journal of Education = Revue canadienne de l'éducation. 40:1, 2017.

- среднее количество слогов в словах
- количество «сложных» слов (в которых более 3 слогов для английского языка и более 4 для русского)
- процент «сложных» слов [1]

Начиная с 40-х годов прошлого века учеными было разработано несколько метрик для оценки удобочитаемости. Самая популярная из них — индекс Флеша (Flesch reading Ease — FRE), которая применяется во многих сервисах для оценки сложности текстов, в частности, она используется Word от Microsoft Office. Все эти метрики (мы использовали Flesch reading Ease, Dale-Chall readability formula, Gunning Fog, Индекс Колман-Лиау, Индекс SMOG (Simple Measure of Gobbledygook)) рассчитываются на основе описанных выше признаков и специальных коэффициентов, чаще всего средней длины слова, предложения и доли общих слов. Для FRE сложность обратно пропорциональна полученному в результате применения формулы числу (т. е. 70 — простой текст, 25 — очень сложный. Формула предполагает шкалу от 100 до 0). Для остальных метрик напротив, большее число означает большую сложность.

Общей проблемой для всех метрик является то, что они разработаны только для английского языка. Это особенно заметно на индексе удобочитаемости Флеша: получаемый результат во многих случаях выбивается из принятой шкалы значений и показывает то отрицательное число, то более 100. В русском языке слова обычно длиннее, а предложения в среднем короче, чем в английском, так как служебных слов в среднем используется меньше. Поэтому для индекса Флеша, как и для других метрик, существуют адаптации под русский язык. Для FRE наиболее распространенная версия была разработана И. В. Оборневой [2]. Также существуют русскоязычные адаптации и для других метрик. Они позволяют существенно улучшить качество анализа текстов.

Кроме того, мы решили ввести несколько дополнительных признаков, которые могли бы нагляднее показать различия между текстами. Все они количественные и отражают степень присутствия того или иного вида слов в текстах.

Чтобы сделать корректные подсчеты, нам было необходимо лемматизировать тексты — то есть привести слова в них к начальной форме.

Обработанные таким образом статьи мы сравнили со списком наиболее частотных для русского языка слов (предполагается, что такие слова обычно покрывают около 80 % среднестатистического текста[3]. Подсчитывалось количество слов из списка в каждом тексте и делилось на общее количество слов в этом тексте. Кроме этого, таким же образом тексты были проанализированы на наличие разговорных слов русского языка.

[1] Saggion, Horacio, “Automatic Text Simplification”, 2017.

[2] Reynolds R. Insights from Russian second language readability classification: complexity-dependent training requirements, and feature evaluation of multiple categories // 11th Workshop on Innovative Use of NLP for Building Educational Applications, 2016.

[3] Batinic D., Birzer S. Creating an extensible, levelled study corpus of Russian // 13th Conference on Natural Language Processing (KONVENS 2016).

Помимо этого мы прибегли к частеречному анализу текстов, то есть посчитали среднюю долю каждой части речи в текстах Назирова по разным темам и в работах других авторов. Считается, что если в тексте много существительных и прилагательных, то читать его сложнее, а если в нем используется много глаголов и союзов, то такой текст напротив, воспринимается легче. Подобные наблюдения часто высказываются среди людей, профессионально работающих с текстами — редакторов, переводчиков, журналистов. В частности, известная переводчица и редактор Нора Галь в своем труде «Слово живое и мертвое» отмечает, что замена глаголов на отглагольные существительные и деепричастные обороты делает текст перегруженным и менее понятным.

2 Описание данных

Работа построена на анализе оцифрованного корпуса текстов Р. Г. Назирова[1], включающего в себя его дневники, работы о культуре, статьи о литературе, блок трудов о Достоевском. Кроме того, в исследовании использовались доступные в электронном виде монографии и статьи Е. М. Мелетинского, О. М. Фрейденберг, М. М. Бахтина и А. Г. Долинина.

3 Внутреннее разнообразие текстов Р. Г. Назирова

Корпус сочинений Р. Г. Назирова разделен на несколько больших блоков: работы о Достоевском, о мифологии, литературоведческие труды общего профиля, личные дневники. Мы решили сначала проанализировать при помощи метрик удобочитаемости внутреннее разнообразие этих произведений, чтобы понять, возможно ли сравнивать этот корпус с текстами других авторов.

Анализ текстов по некоторым статистическим признакам (рис. 1, рис. 2, рис. 3), таким как средняя длина слов и предложений, а также процент сложных слов (в которых более 4 слогов), показал, что по этим характеристикам тексты разных жанров у Назирова не слишком различаются между собой. Работы о Достоевском характеризуются чуть большей средней длиной слов и предложений, но эти различия колеблются в рамках десятых долей. Средняя длина слова составляет 6 символов, а предложений — 11 слов. Также в работах о Достоевском сосредоточено ощутимо большее количество «сложных слов»: в текстах остальных жанров их доля составляет около 17 – 18 %, а в исследованиях о Федоре Михайловиче — 23 %.

Особенности, связанные с небольшой средней длиной слова может объяснить частеречный анализ.

Оказалось, что работы Назирова о мифологии отличаются высокой долей существительных (более 50 %), а меньше всего их в дневниках — около 36 % (рис. 4). Мифологические работы также отличаются более высокой долей глаголов (17 %) по сравнению с другими

[1] https://github.com/nevmenandr/nazirov-texts-dataset

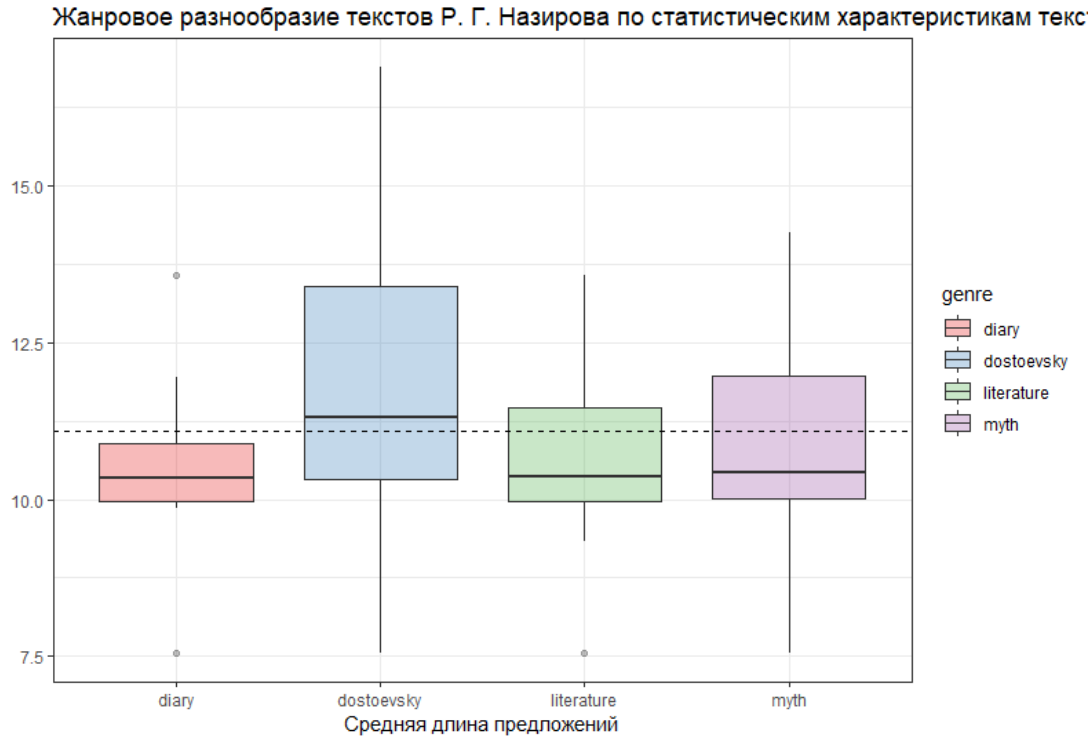


Рис. 1: Сравнение средней длины предложений в текстах Р. Г. Назирова

жанрами (12 – 13 %, рис. 5) и прилагательных (19 %, рис. 6), которые меньше всего используются в дневниках (13 %). При этом в мифологических работах значительно меньше предлогов и союзов (менее 5 %), тогда как в текстах других жанров их довольно много (10 – 12 %, рис. 7, рис. 8). Высокая доля предлогов и союзов может объяснять небольшую среднюю длину слова (хотя понятно, что в случае с мифологией это не является определяющим факторов). Мы предполагаем, что раз в текстах о мифологии значительно меньше служебных частиц, то и для читателя эти работы будут ощутимо сложнее. Работы по Достоевскому и мифологии отличаются большим разбросом значений, в отличие от текстов по литературоведению и дневников, которые, видимо, являются наиболее похожими друг на друга внутри группы.

Важным показателем понятности текста для неподготовленного читателя является наличие в нем общеупотребительных слов русского языка. Для обычного, неакадемического текста, они могут составлять до 80 % всех используемых слов. В работах Назирова (рис. 9) это в среднем 63 %, причем в дневниковых текстах и работах о Достоевском их больше (хотя у Достоевского, опять же, большой разброс). Дневники также превосходят остальные жанры по наличию разговорных слов (19 %, рис. 10), хоть и незначительно. Это неудивительно для текстов личного характера.

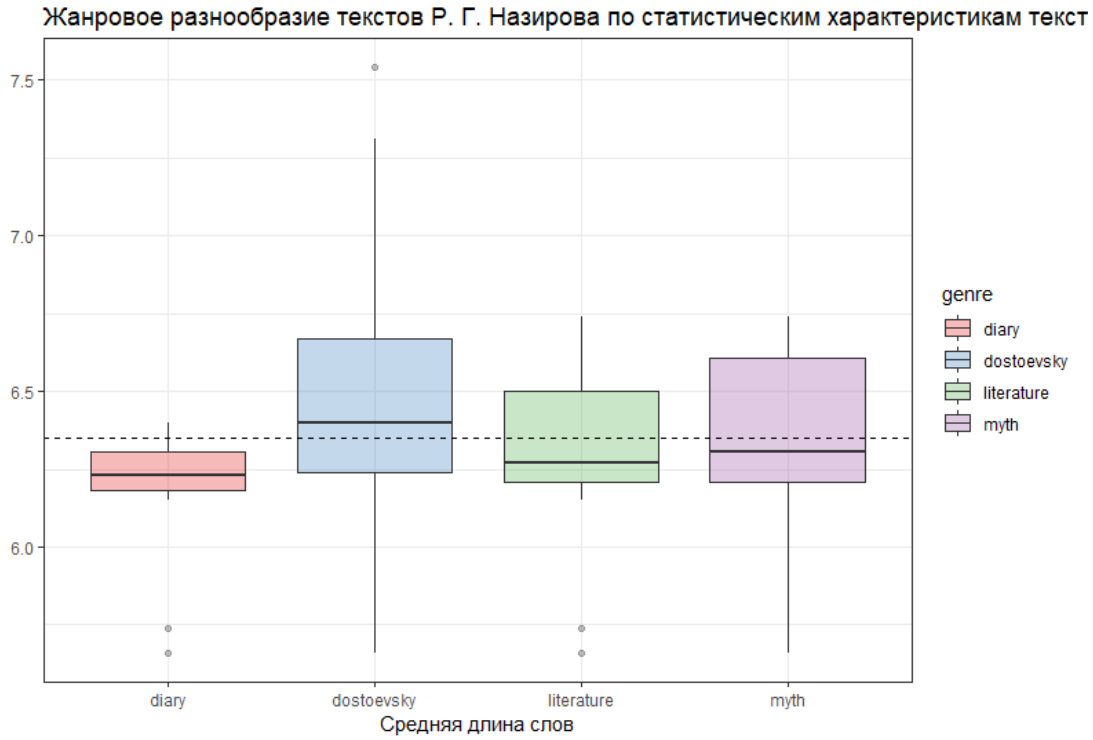


Рис. 2: Сравнение средней длины слов в текстах Р. Г. Назирова

Традиционные метрики ридабилити показывают, что уровень сложности всех текстов Назирова примерно одинаковый. Все метрики показывают, что чуть большей сложностью отличаются работы о Достоевском (рис. 11, рис. 12, рис. 13, рис. 14, рис. 15), но это отличие, опять же, небольшое, хотя в текстах о Достоевском есть большой разбросу в сторону большей сложности. По индексу Флеша среднее значение сложности текстов — 40, что является показателем средней сложности. Индексы, ориентированные на сравнение с уровнями образования, определяют тексты Назирова как понятные ученикам старшей школы.

4 Сравнение мифологических текстов Назирова с текстами О. М. Фрейденберг и Е. М. Мелетинского

Мы сравнили работы Назирова о мифологии с трудами других авторов по близкой теме. Такое сравнение позволит понять, действительно ли работы Назирова могут быть более понятны неподготовленному читателю.

На первый взгляд, нельзя сказать, что тексты Назирова однозначно выигрывают. Например, средняя длина слова оказывается меньше в работах Фрейденберг (рис. 16), соответственно, доля сложных (т.е. длинных слов) слов у нее тоже ниже, чем у Назирова

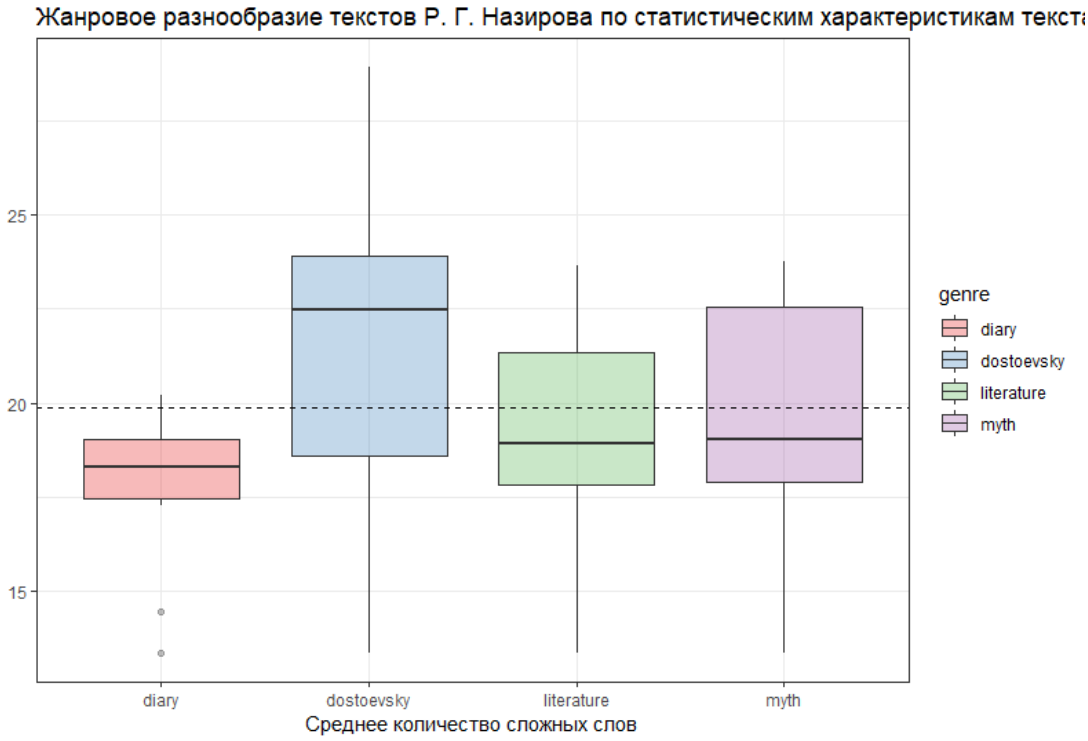


Рис. 3: Сравнение доли существительных в текстах Р. Г. Назирова

и Мелетинского (рис. 18). При этом средняя длина предложения у Назирова значительно меньше, чем у двух других авторов (рис. 17) — около 12 слов, тогда как у Фрейденберг это 22 слова, а у Мелетинского — 25. Вероятно, это является одним из признаков «лаконичности» стиля Назирова, отличного от традиционной академической строгости и избыточности.

Кроме того, в текстах Фрейденберг и Мелетинского гораздо выше процент общеупотребительных слов русского языка (рис. 19), 73 % и 71 % соответственно. В текстах Назирова таких слов на 10 % меньше. С разговорными словами все менее однозначно: медианы всех трех авторов довольно близки (рис. 20), чуть меньше разговорные слова использует Мелетинский.

Частеречный анализ объясняет невысокую среднюю длину слов у Фрейденберг (рис. 23, рис. 24): она использует очень много служебных частей речи, союзов и предлогов, обе группы занимают по 10 % в ее текстах. То же мы наблюдаем и у Мелетинского, хотя в его трудах больше существительных и прилагательных, что, видимо, уравновешивает большое количество служебных частей речи. Назиров же, как мы выяснили ранее, почти не пользуется служебными частицами в текстах этого жанра, зато в его работах о мифологии много существительных и глаголов (рис. 21, рис. 22).

На первый взгляд, это может говорить о большей сложности его текстов, но в сочетании с другими характеристиками мы можем понять, что мифологические тексты Назирова

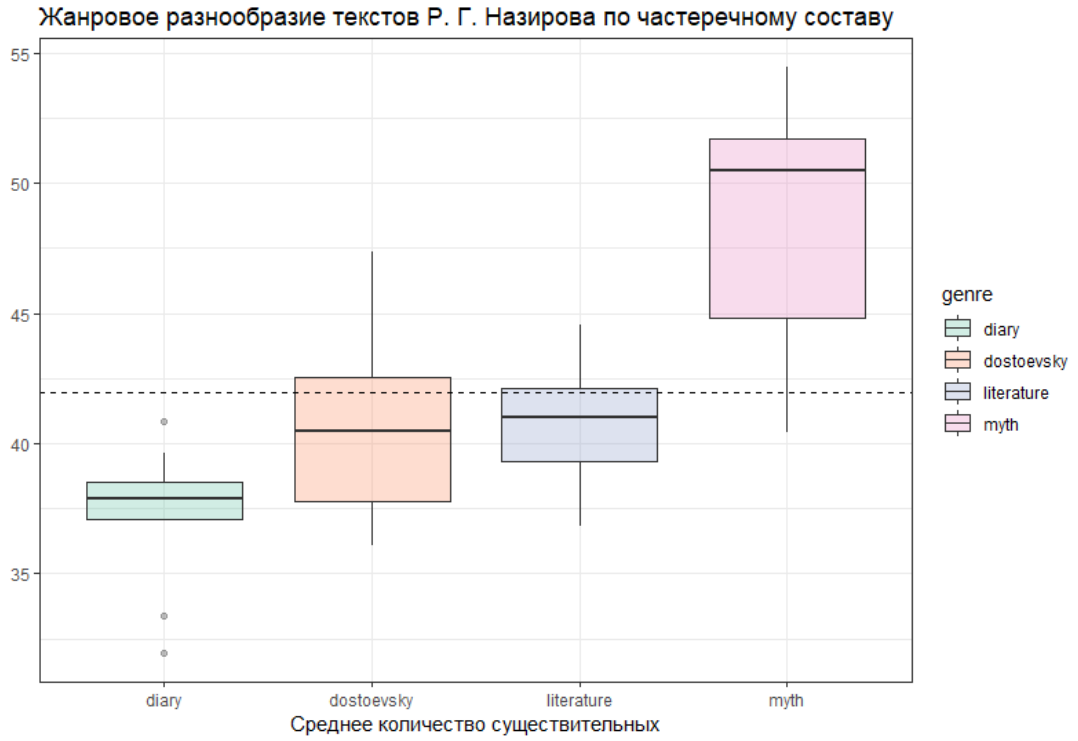


Рис. 4: Сравнение доли существительных в текстах Р. Г. Назирова

состоят из коротких предложений, не перегруженных сочинительными и подчинительными связями, зато содержащих много активных связей глагол-существительное. Мы предполагаем, что такой текст, действительно, может оказаться более легким для восприятия.

По сравнению с текстами Фрейденберг нельзя сказать, что статьи Назирова однозначно проще. Данные по индексам удобочитаемости разнятся: по некоторым из них сочинения Назирова оказываются сложнее, чем у Фрейденберг, а по другим наоборот, легче.

При этом Назиров использует в речи больше глаголов и существительных, но у Фрейденберг больше союзов и предлогов, а предложения в целом длиннее. Можно предположить, что тексты Фрейденберг для читателя кажутся сложнее именно потому, что в них длиннее предложения, и предложения эти — сложносочиненные. При этом тексты Фрейденберг проще по составу словаря — в них больше общеупотребительных слов и меньше длинных. То есть по «техническим» признакам тексты Назирова пока оказываются сложнее.

Метрики удобочитаемости тоже не дают однозначного ответа. Например, по индексу Флеша наиболее сложными оказываются тексты Мелетинского (рис. 26), которые получили оценку 20. Тексты Назирова и Фрейденберг же напротив, имеют почти одинаковые значения. Остальные индексы также подчеркивают более высокую сложность текстов Мелетинского (рис. 27, рис. 28, рис. 29, рис. 30) — для их понимания нужна если не ученая степень, то

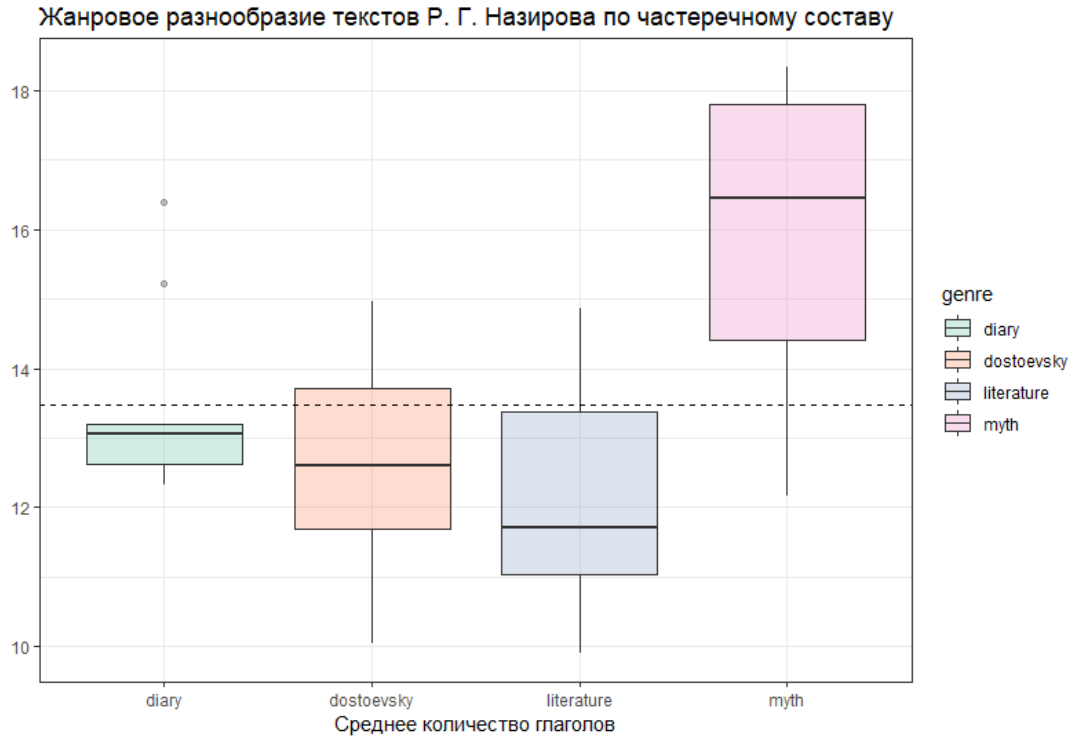


Рис. 5: Сравнение доли глаголов в текстах в текстах Р. Г. Назирова

по крайней мере несколько курсов высшего образования. Тексты Назирова и Фрейденберг определяются как подходящие для средней школы.

По некоторым из метрик (SMOG и индекс Ганнинг-Фога) тексты Назирова определяются как наиболее простые, но так как согласованности между всеми пятью метриками нет, этот результат нельзя считать однозначным и итоговым. Так как метрики удобочитаемости во многом опираются на среднюю длину слов, понятно, что тексты Фрейденберг отмечены как несложные за счет большого количества служебных частиц.

5 Сравнение текстов о Ф. М. Достоевском

Так как тексты Р. Г. Назирова довольно сильно различаются друг от друга по сложности в зависимости от их темы, работы о Достоевском мы решили сравнить с текстами авторов, которые занимались исследованиями на ту же тему.

В данном случае тексты Р. Г. Назирова также не выглядят однозначно более легкими. Средняя длина слова у него чуть выше, больше и количество общих слов. Средняя длина предложения значительно выше в текстах Долинина (30 слов, тогда как у Назирова и Бах-

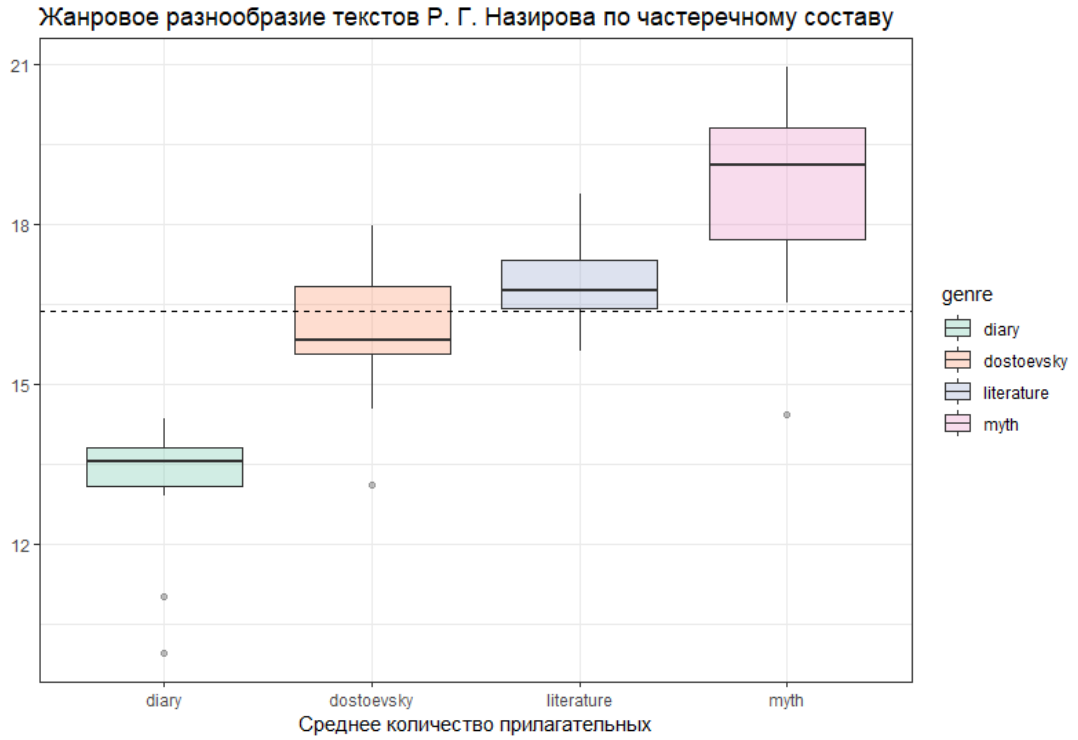


Рис. 6: Сравнение доли прилагательных в текстах Р. Г. Назирова

тина — 13 и 16). Здесь мы снова видим стилистическую особенность Назирова, связанную с тем, что в его текстах предложения довольно короткие (рис. 31, рис. 32, рис. 33).

Кроме того, в текстах Назирова, по сравнению с другими авторами, гораздо меньше общих и разговорных слов (рис. 34, рис. 35). У Бахтина и Долинина общеупотребительные слова составляют почти 80 % текста, а у Назирова в среднем меньше 65 %.

Интересные результаты дает частеречный анализ. Например, в текстах Назирова доля существительных значительно выше, чем у других авторов-достоевистов (рис. 36), но при этом заметно ниже доля союзов (рис. 38). Глаголы, предлоги и прилагательные представлены в текстах этих авторов примерно одинаково (рис. 37, рис. 38, рис. 39).

По индексам удобочитаемости результаты немного разнятся (при том что для текстов о мифологии они согласовывались между собой). Например, индекс Флеша для всех трех авторов отражает близкие значения (рис. 41), но согласно индексу SMOG и Ганнинг-Фога самыми сложными оказываются тексты Долинина — они будут понятны только выпускникам университетов и людям с научными степенями, тогда как Бахтин и Назиров подходят для старшеклассников. Схожие результаты показывает индекс Дейл-Челла, а по индексу Колимана-Лиау напротив, самыми сложными оказываются работы Назирова, хотя в целом у всех трех авторов результаты довольно похожи (рис. 42, рис. 43, рис. 44, рис. 45). Получается, что хотя в текстах о Достоевском сохраняется ряд типичных для Назирова сти-

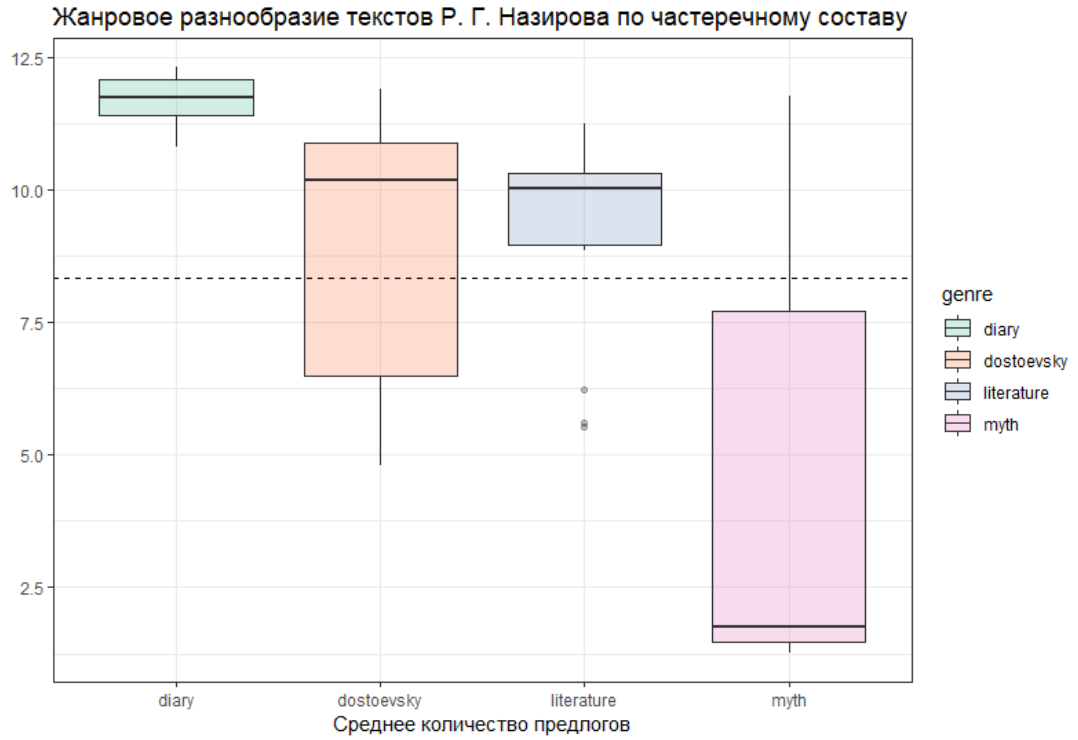


Рис. 7: Сравнение доли предлогов в текстах в текстах Р. Г. Назирова

листических черт, эти его работы едва ли можно считать нетипично легкими для текстов данной категории. Мы помним, что среди всех текстов Назирова работы о Достоевском отличались чуть более высоким уровнем сложности, возможно, это также находит отражение в результатах нашего анализа.

6 Заключение

По результатам нашего исследования нельзя однозначно сказать, что предложенная в начале гипотеза о том, что тексты Р. Г. Назирова можно определить как легкие и понятные по формальным признакам удобочитаемости, полностью подтвердилась. Анализ количественных признаков, таких как статистические характеристики текстов, доли различных частей речи и общеупотребительных слов русского языка позволил выделить некоторые отличительные особенности стиля Назирова. Они, предположительно, и формируют впечатление легкости и понятности его текстов.

Но наиболее популярные сегодня метрики удобочитаемости не позволяют сделать вывод, что его работы исключительно легкие для чтения по сравнению с исследованиями других авторов. С другой стороны, они и не определяются как невероятно сложные. Большинство

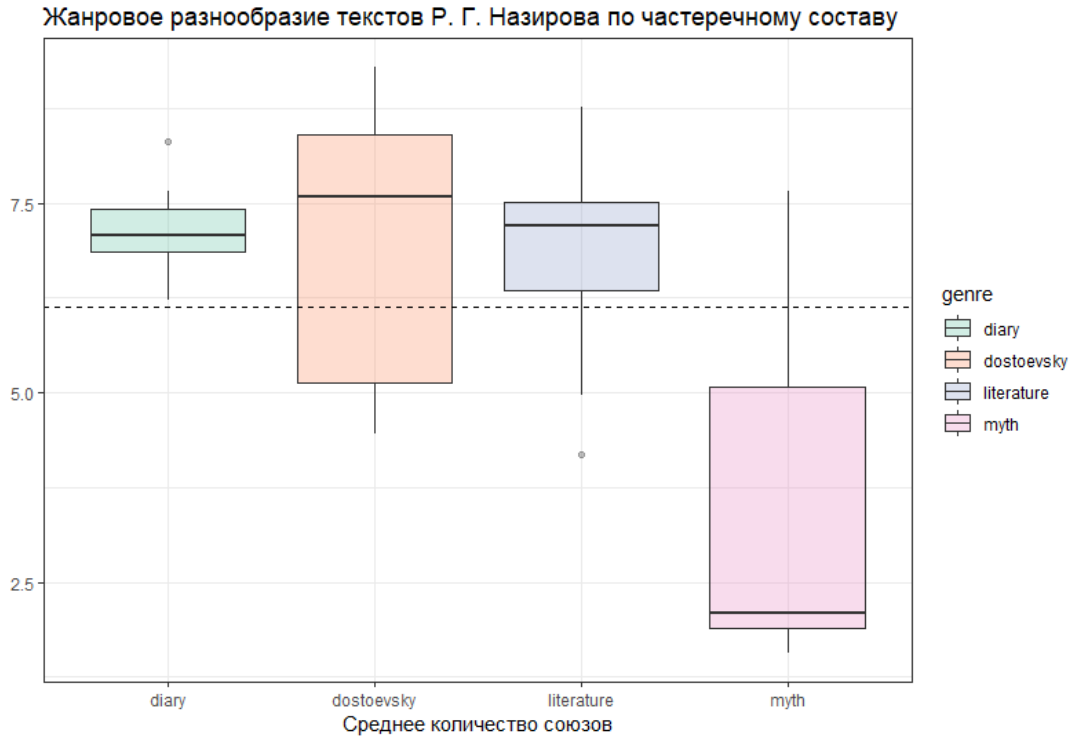


Рис. 8: Сравнение доли союзов в текстах Р. Г. Назирова

работ Назирова, согласно этим метрикам, подходят для учеников старшей школы, но, как мы увидели, тем же свойством могут обладать и тексты других ученых.

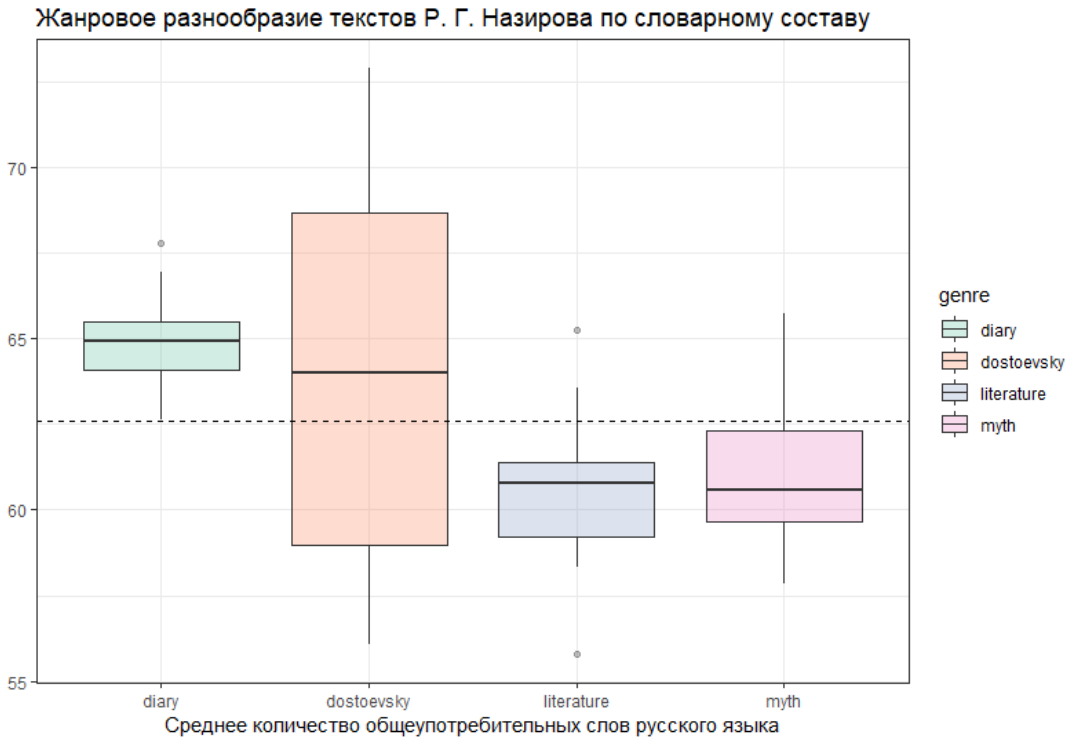


Рис. 9: Сравнение доли общеупотребительных слов в текстах в текстах Р. Г. Назирова

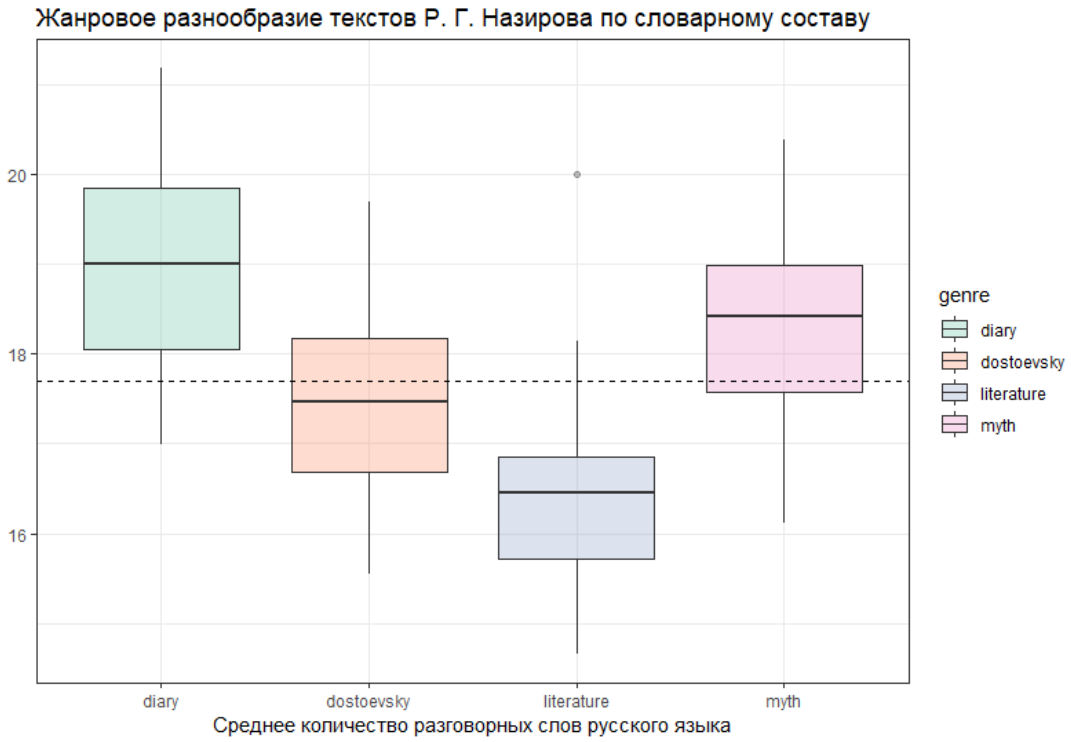


Рис. 10: Сравнение доли разговорных слов в текстах Р. Г. Назирова

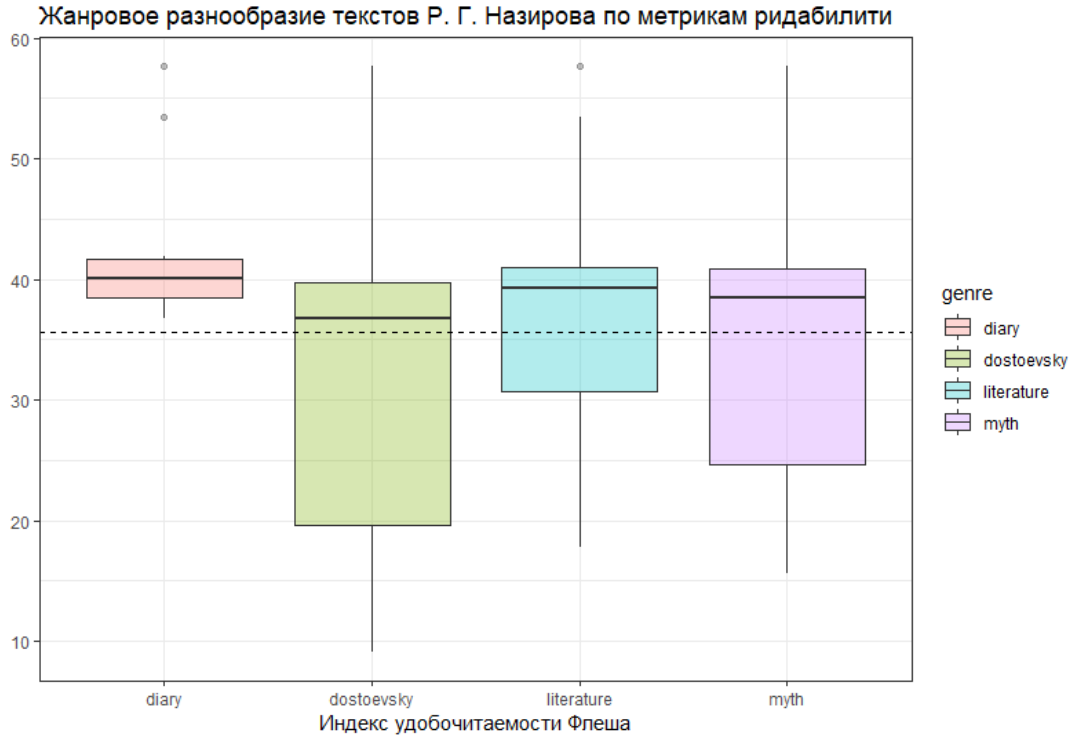


Рис. 11: Сравнение индекса удобочитаемости Флеша в текстах в текстах Р. Г. Назирова

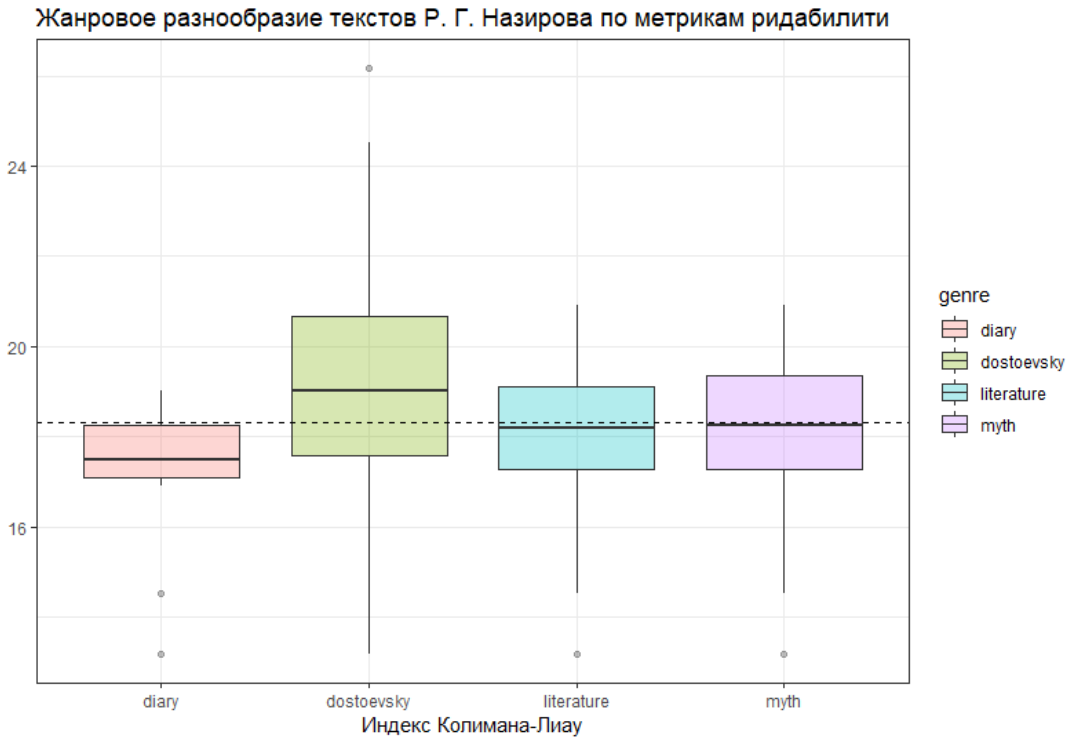


Рис. 12: Сравнение индекса Колимана-Лиау в текстах Р. Г. Назирова

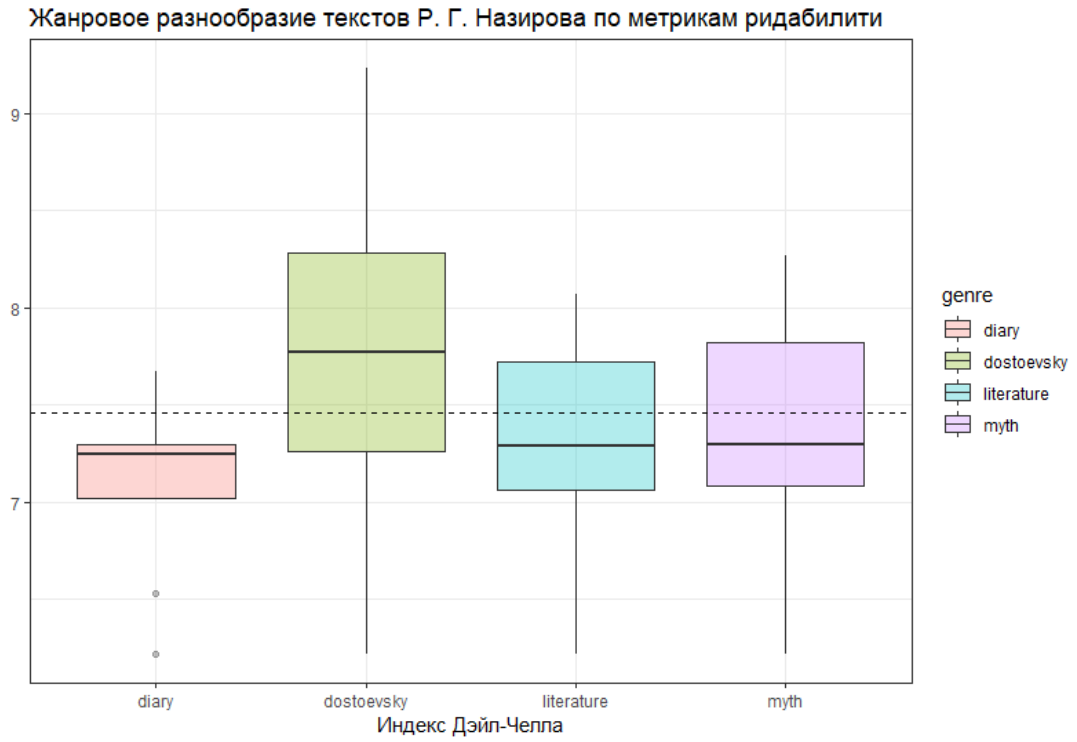


Рис. 13: Сравнение индекса Дэйл-Челла в текстах в текстах Р. Г. Назирова

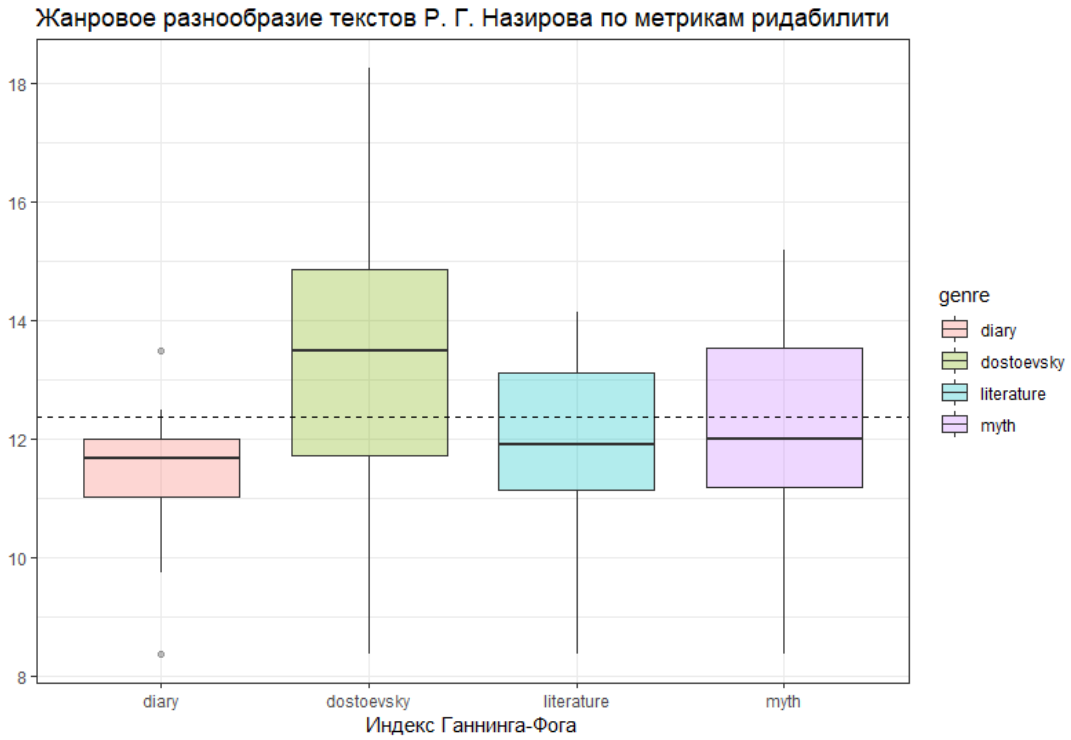


Рис. 14: Сравнение индекса Ганнинг-Фога в текстах Р. Г. Назирова

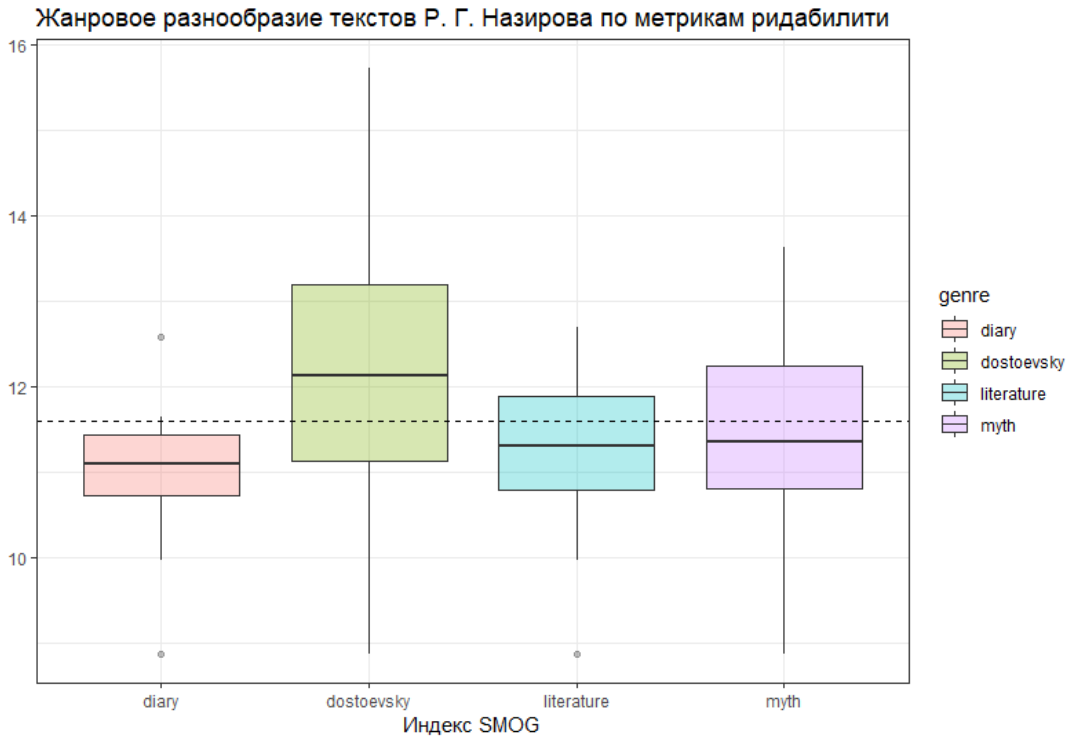


Рис. 15: Сравнение индекса SMOG в текстах в текстах Р. Г. Назирова

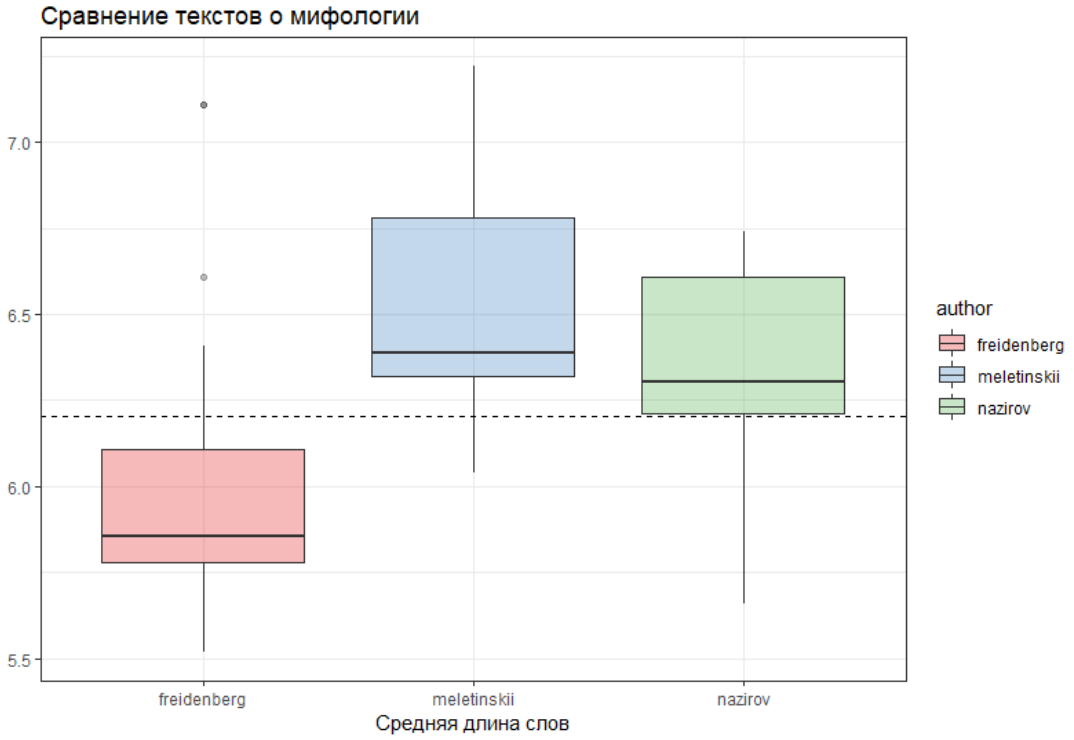


Рис. 16: Сравнение средней длины слов в текстах о мифах и культуре

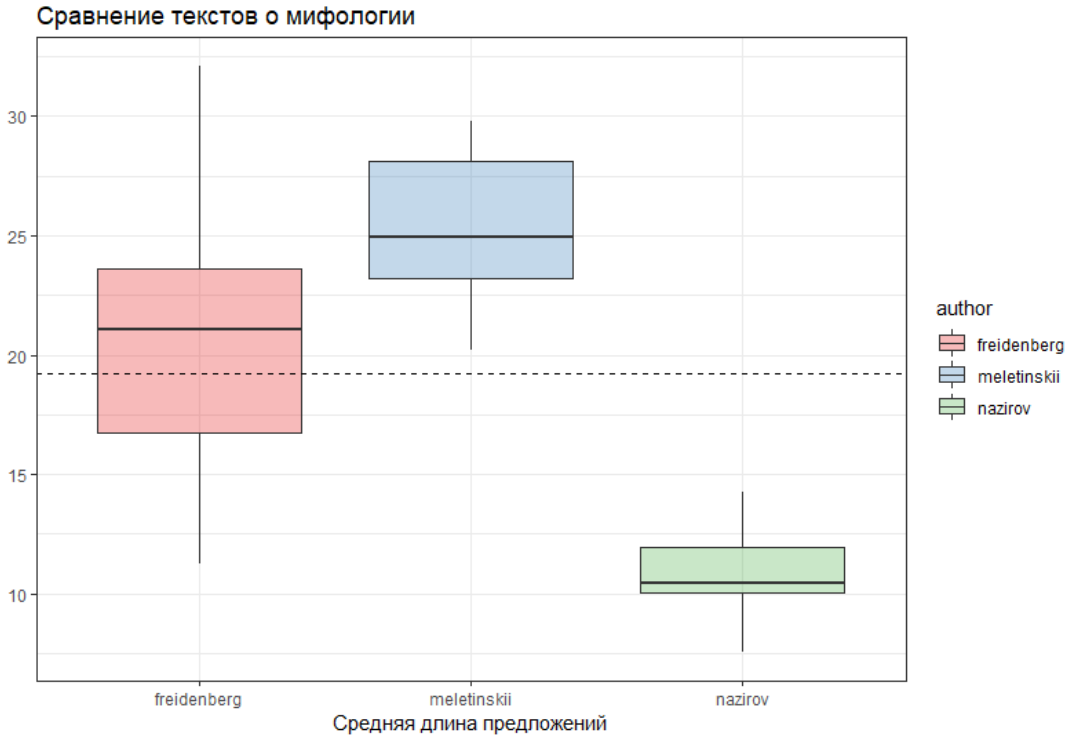


Рис. 17: Сравнение средней длины предложений в текстах о мифах и культуре

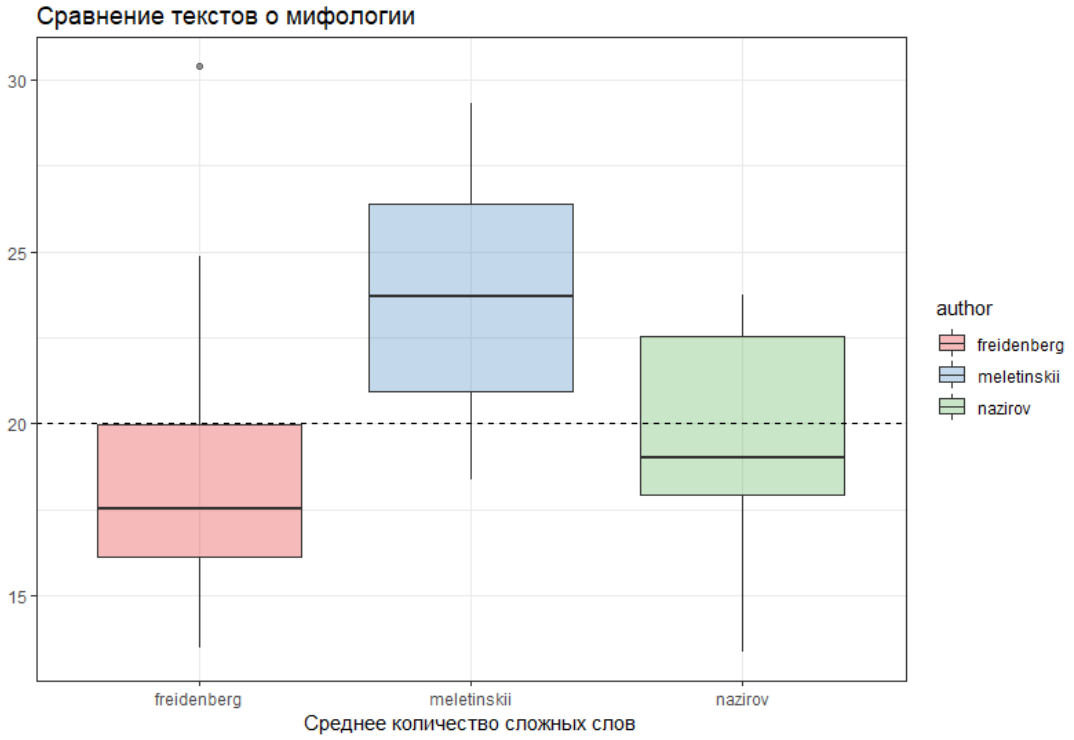


Рис. 18: Сравнение среднего количества сложных слов в текстах о мифах и культуре

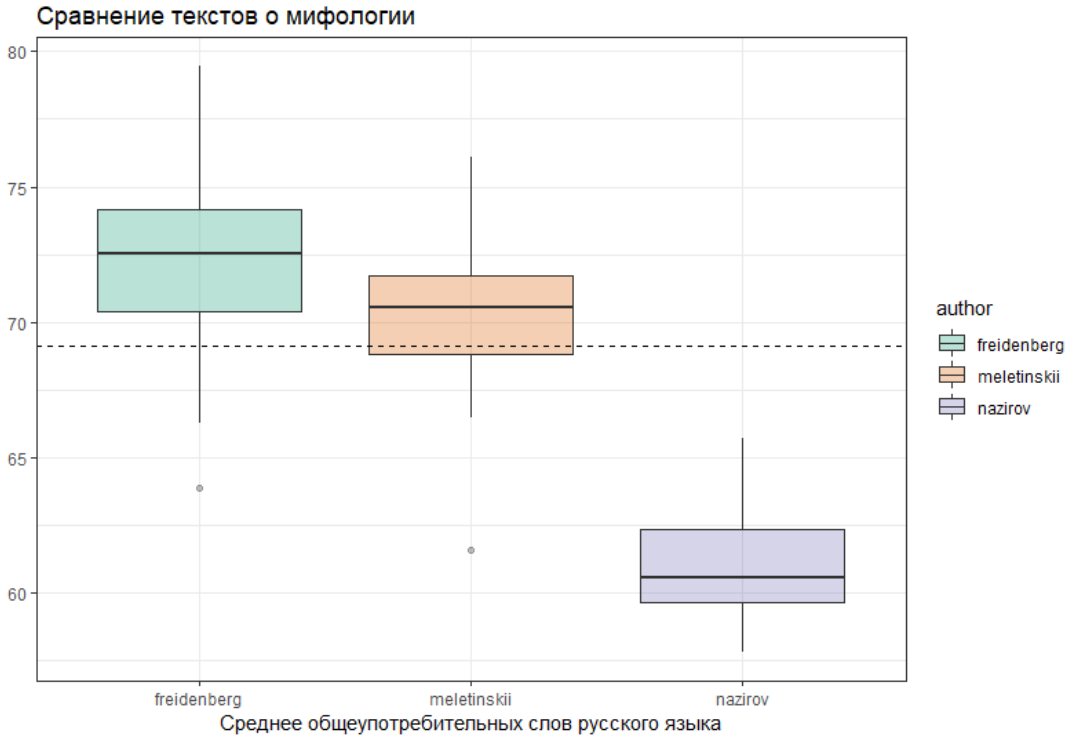


Рис. 19: Сравнение доли общеупотребительных слов в текстах о мифах и культуре

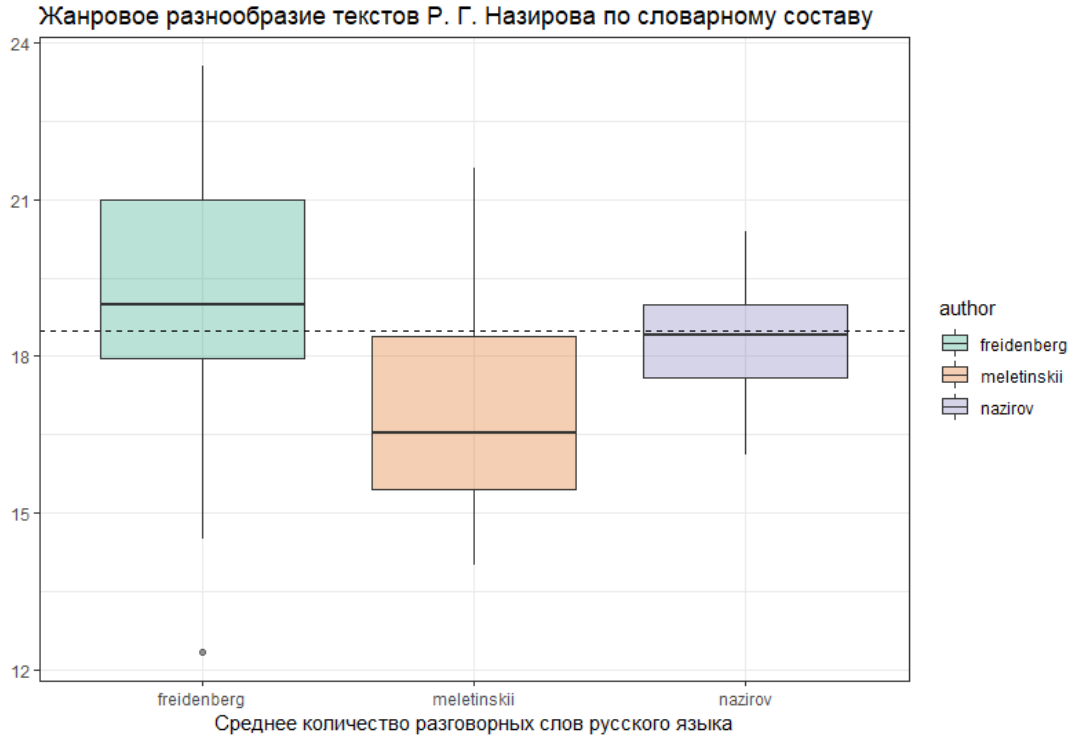


Рис. 20: Сравнение доли разговорных слов в текстах о мифах и культуре

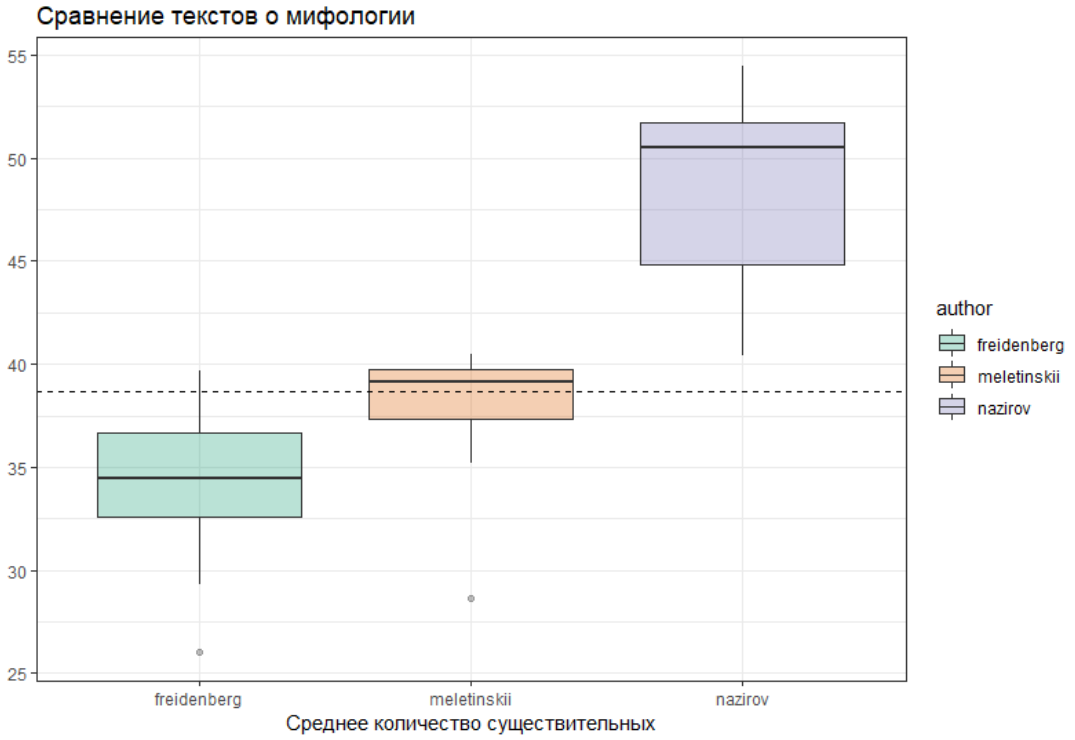


Рис. 21: Сравнение доли существительных в текстах о мифах и культуре

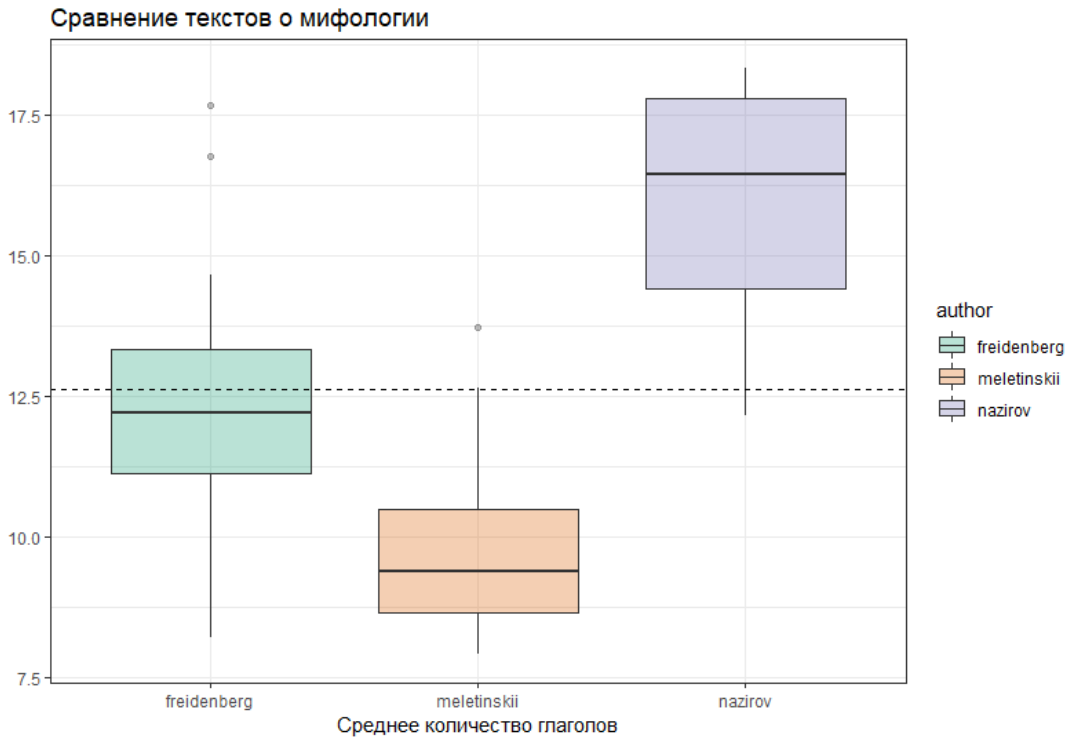


Рис. 22: Сравнение доли глаголов в текстах о мифах и культуре

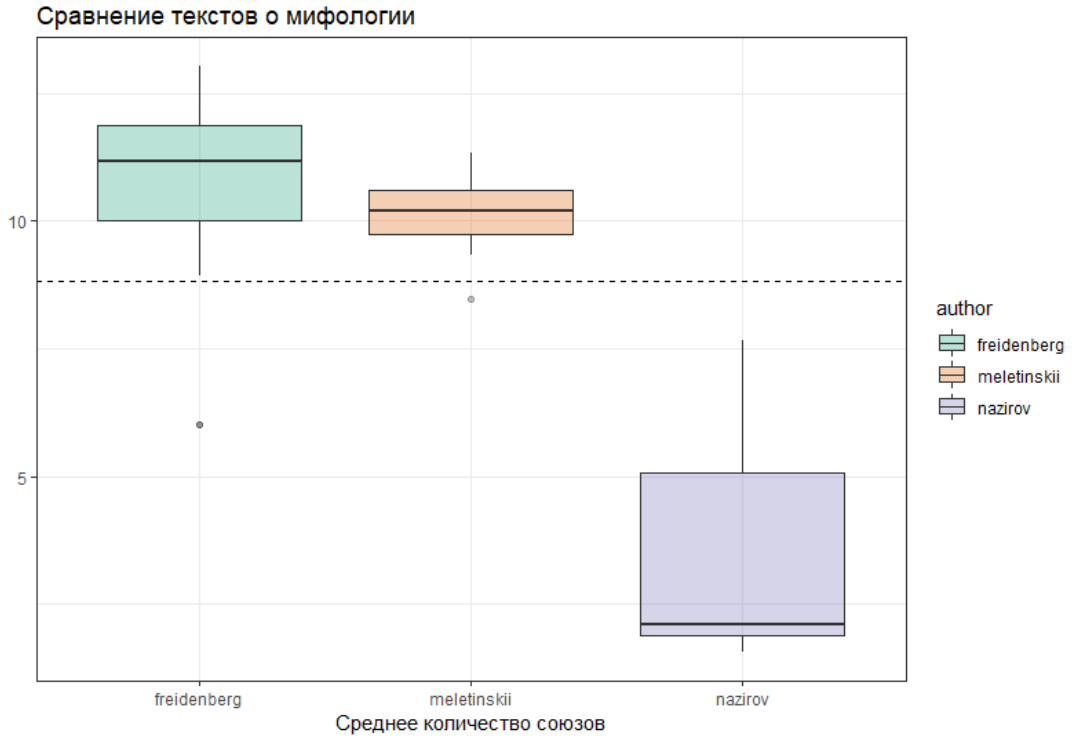


Рис. 23: Сравнение доли союзов в текстах о мифах и культуре

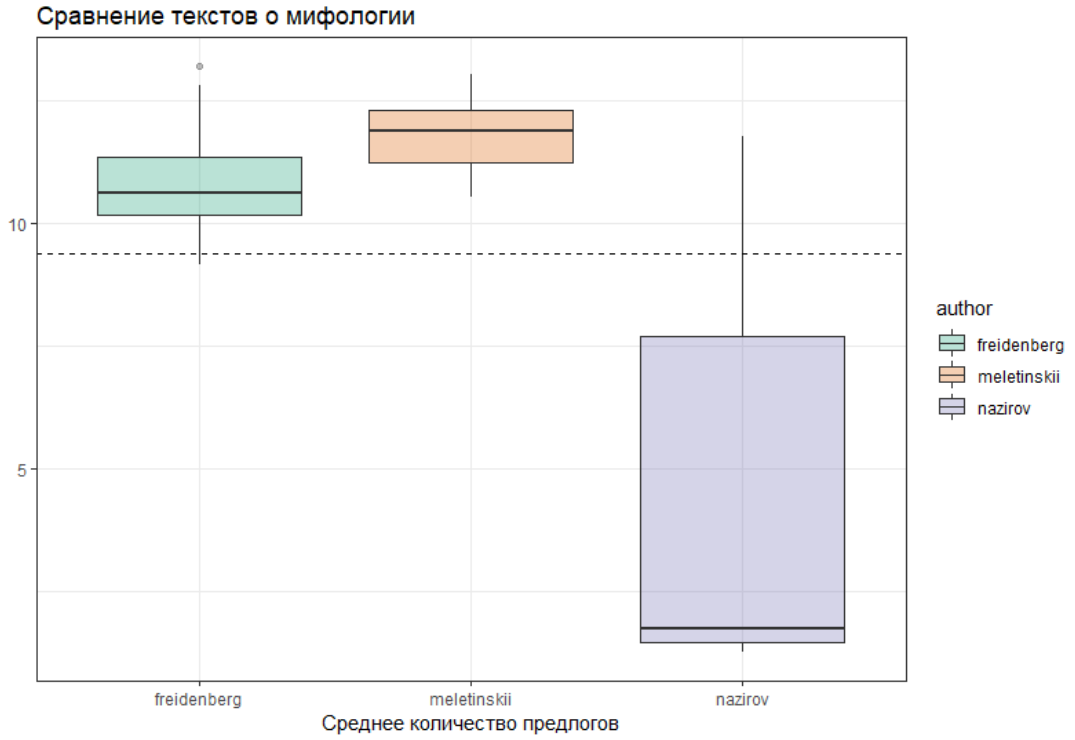


Рис. 24: Сравнение доли предлогов в текстах о мифах и культуре

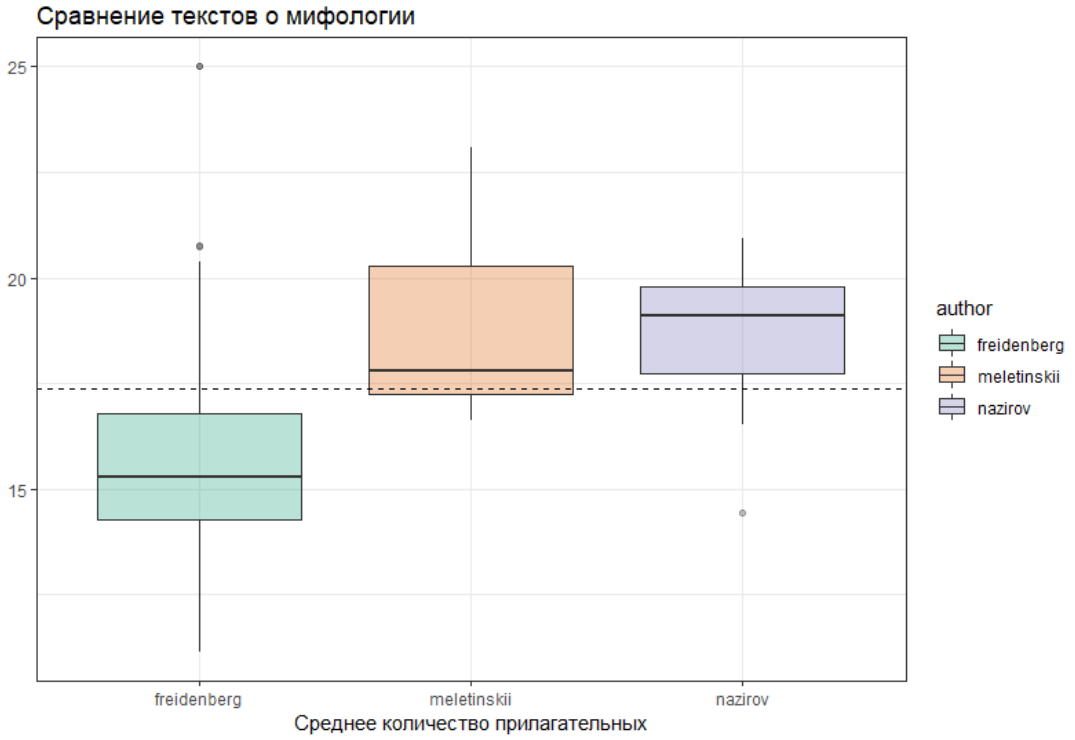


Рис. 25: Сравнение доли прилагательных в текстах о мифах и культуре

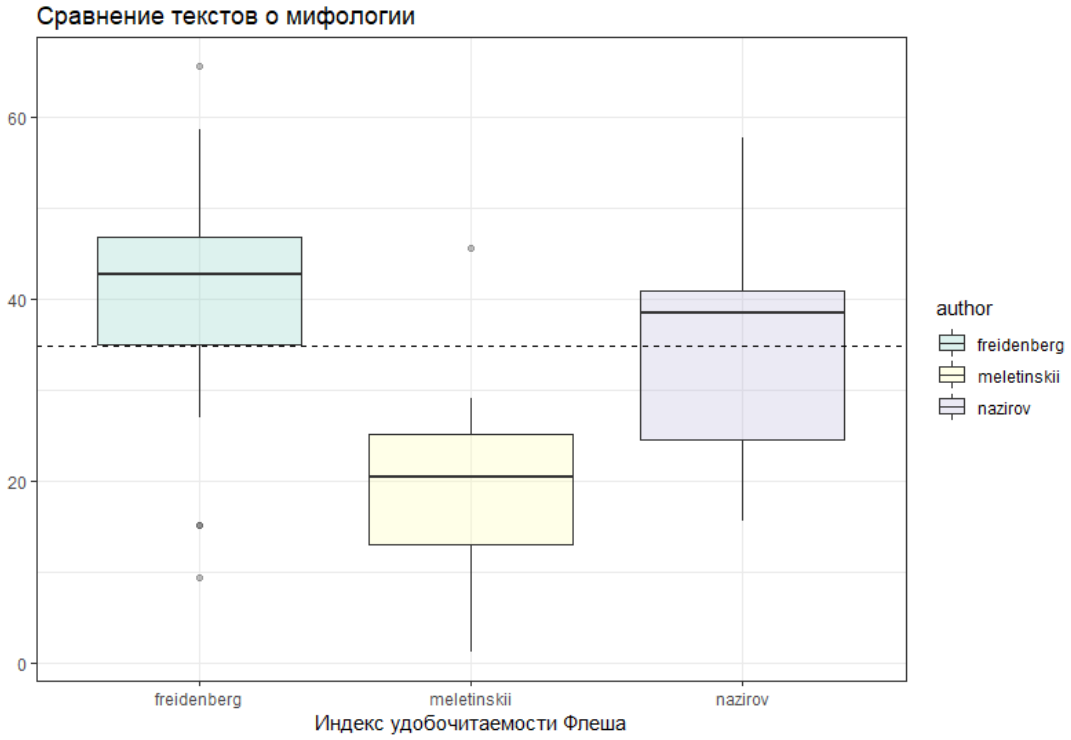


Рис. 26: Сравнение индекса удобочитаемости Флеша в текстах о мифах и культуре

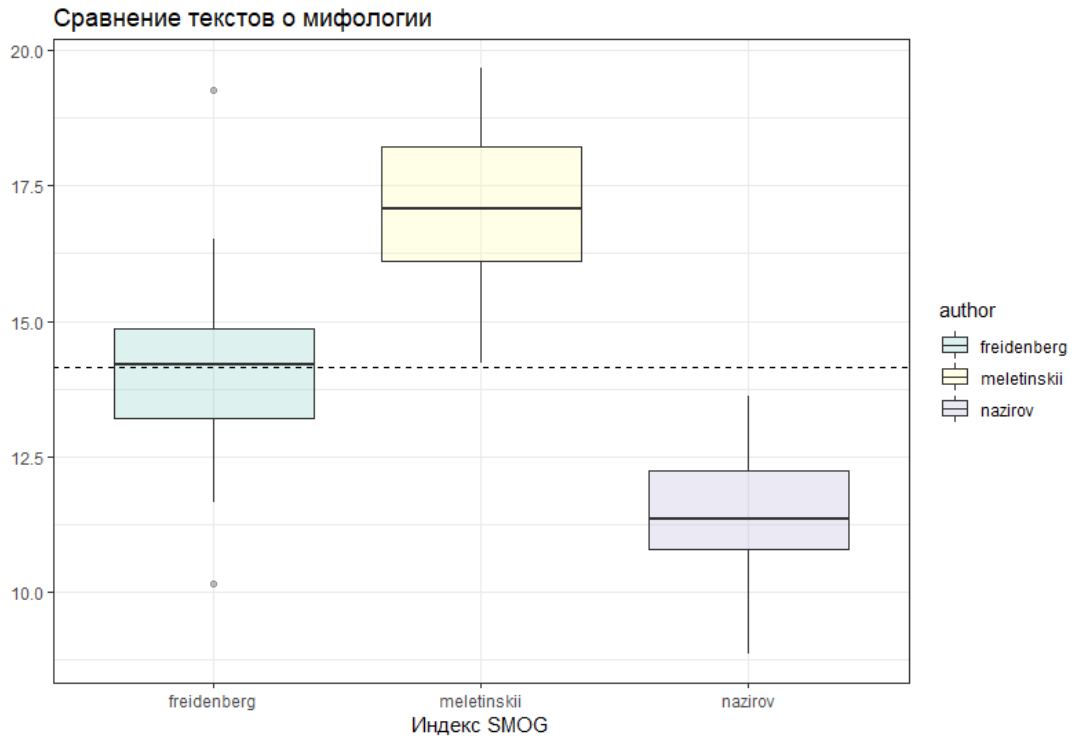


Рис. 27: Сравнение индекса SMOG в текстах о мифах и культуре

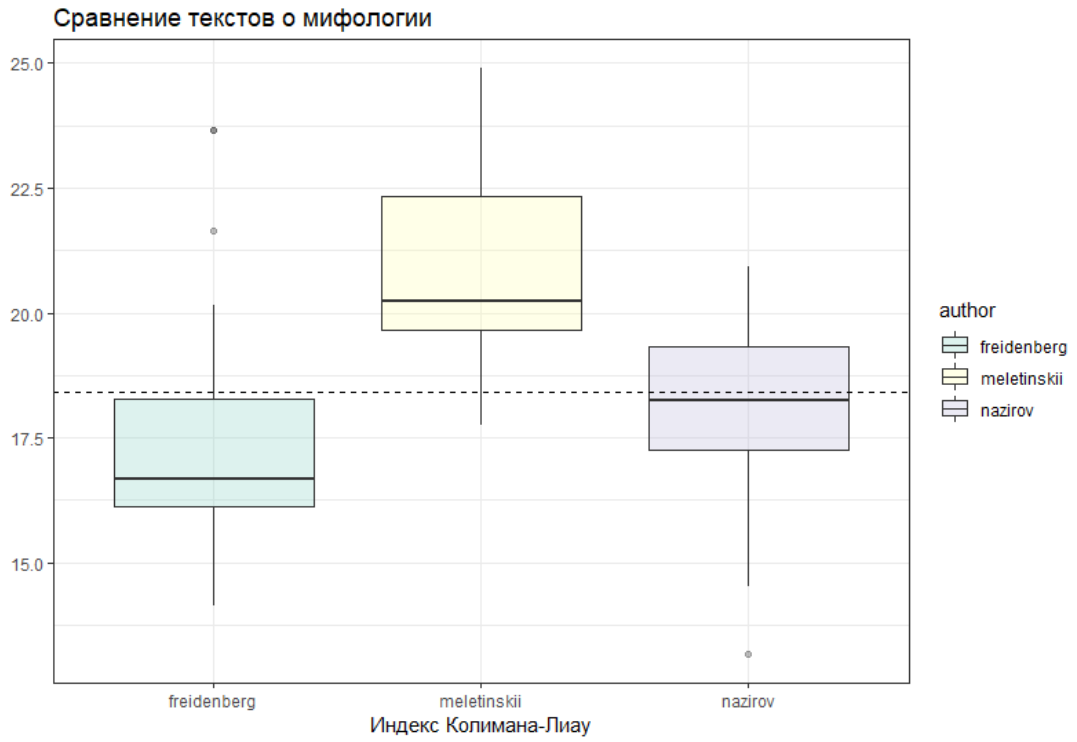


Рис. 28: Сравнение индекса Колиман-Лиау в текстах о мифах и культуре

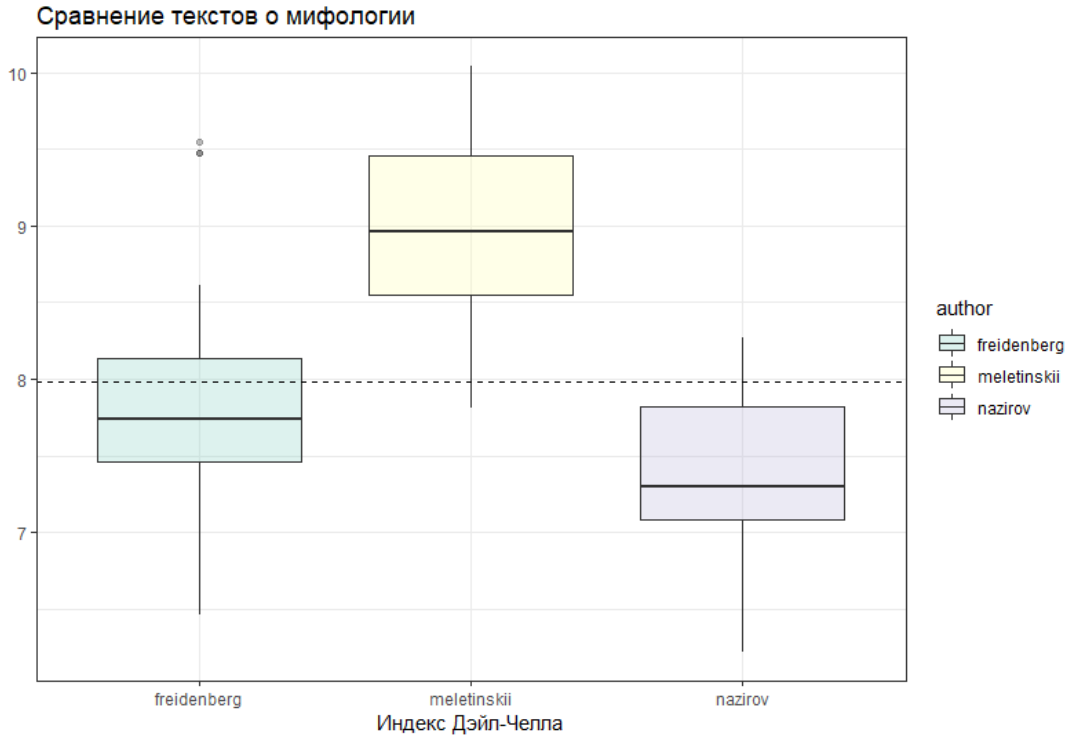


Рис. 29: Сравнение индекса Дэйл-Челла в текстах о мифах и культуре

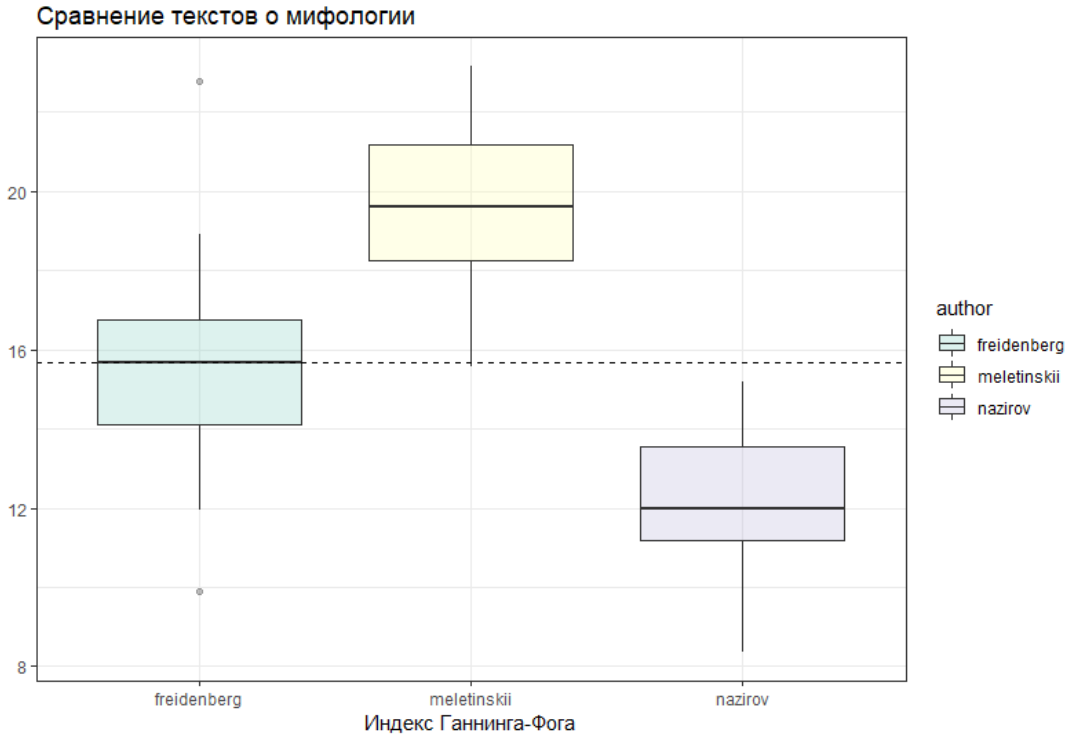


Рис. 30: Сравнение индекса Ганнинг-Фога в текстах о мифах и культуре

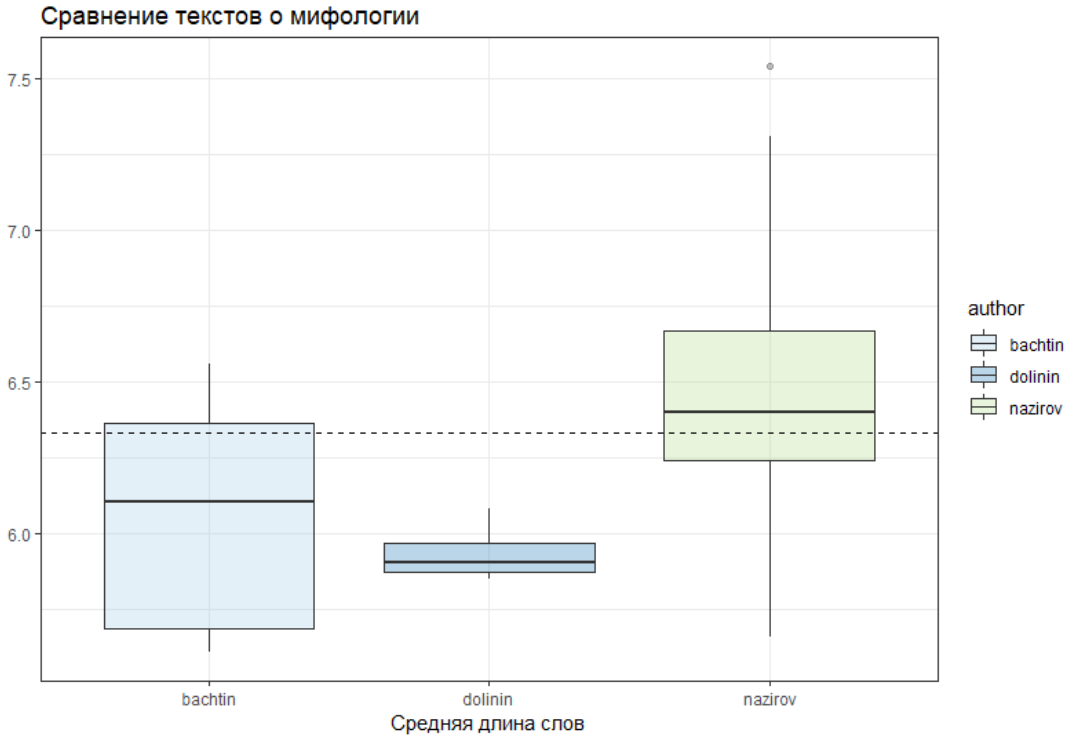


Рис. 31: Сравнение средней длины слов в текстах о Ф. М. Достоевском

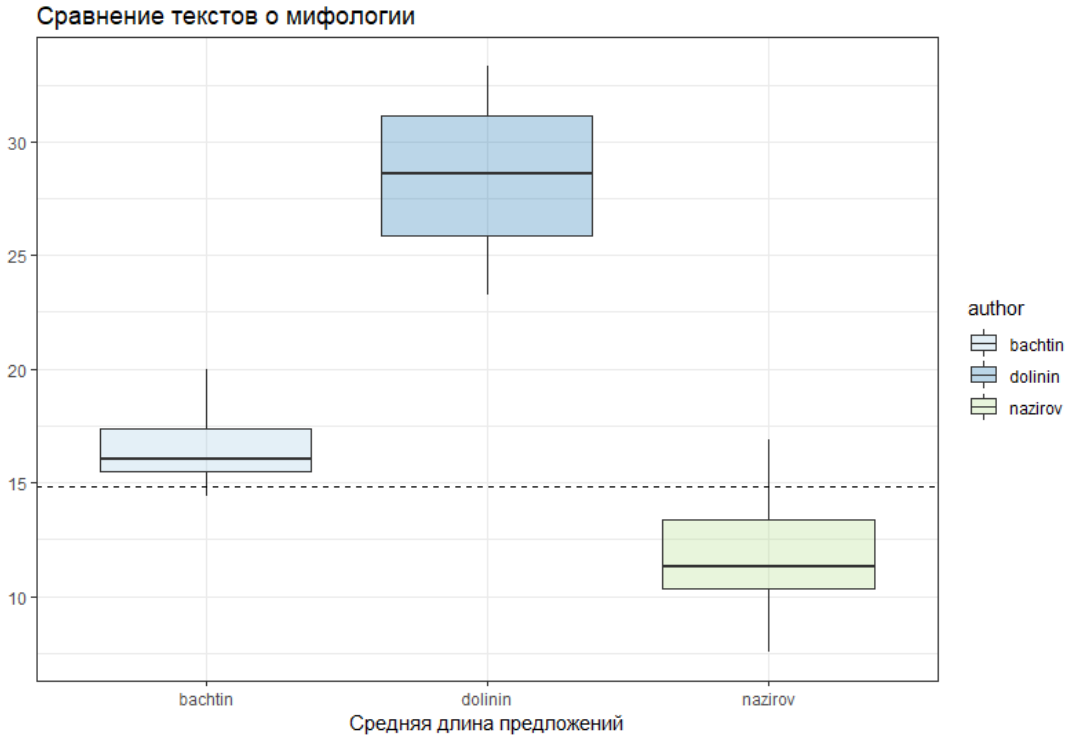


Рис. 32: Сравнение средней длины предложений в текстах о Ф. М. Достоевском

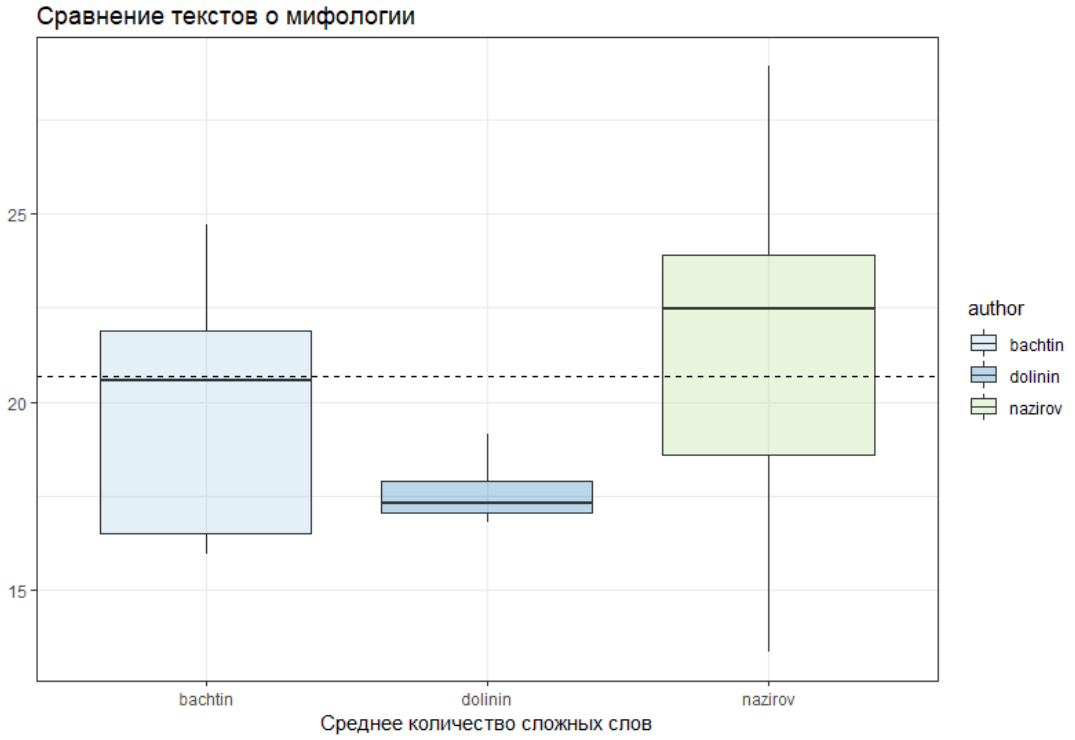


Рис. 33: Сравнение среднего количества сложных слов в текстах о Ф. М. Достоевском

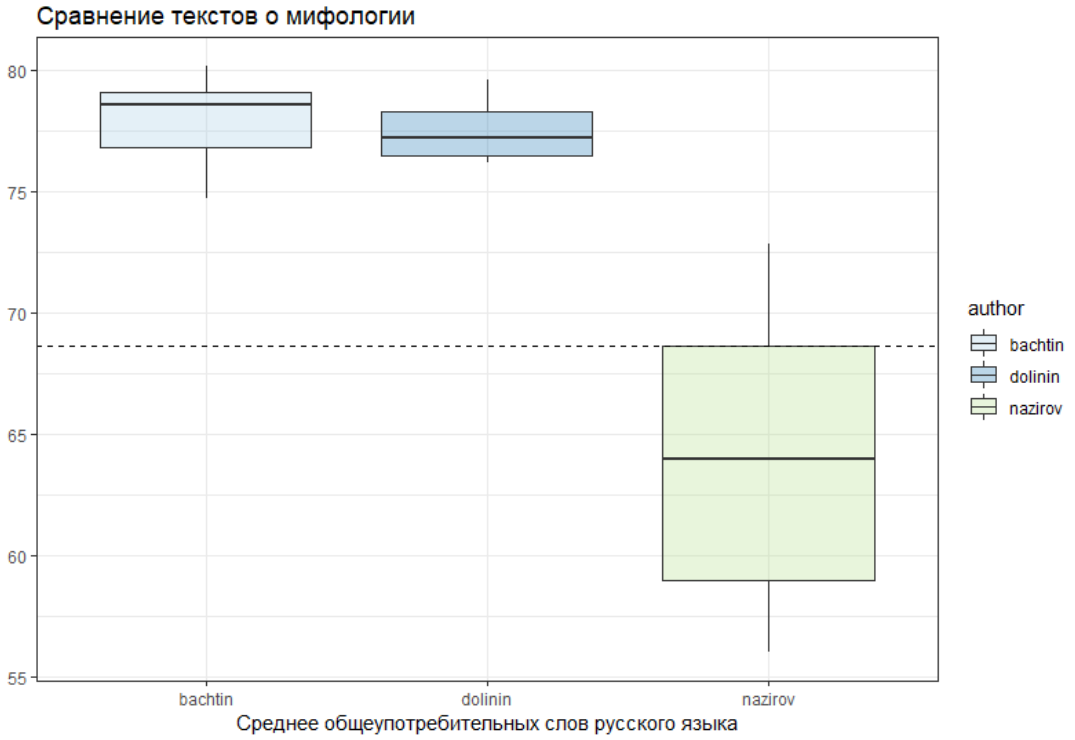


Рис. 34: Сравнение доли общеупотребительных слов в текстах о Ф. М. Достоевском

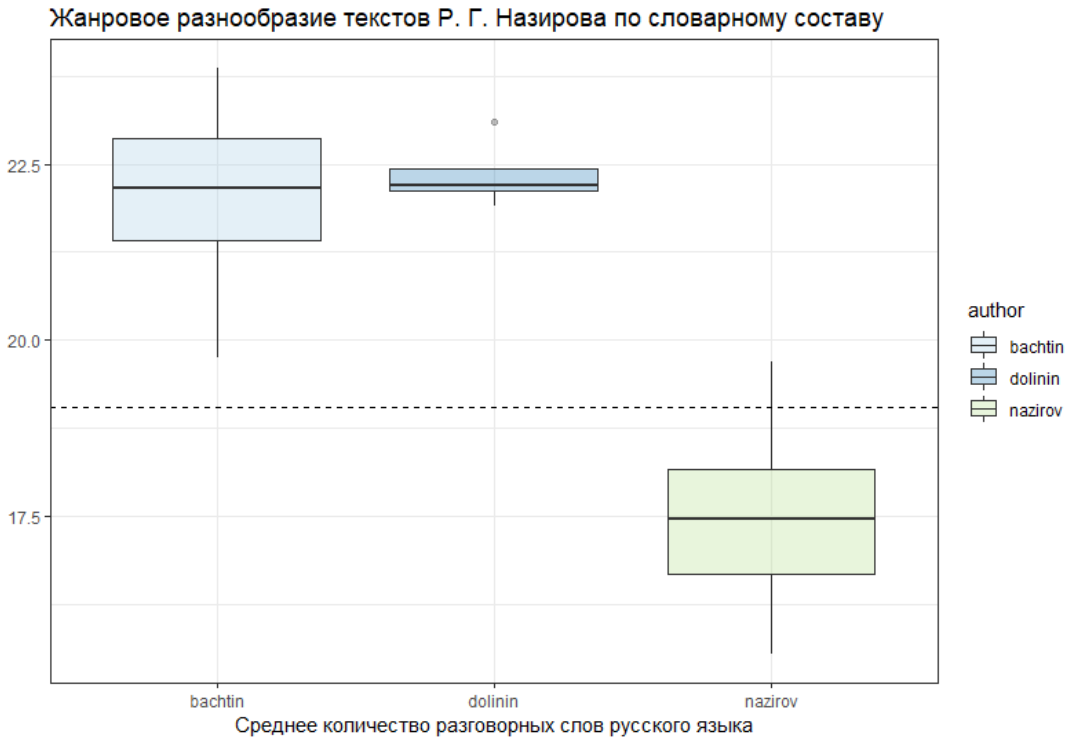


Рис. 35: Сравнение доли разговорных слов в текстах о Ф. М. Достоевском

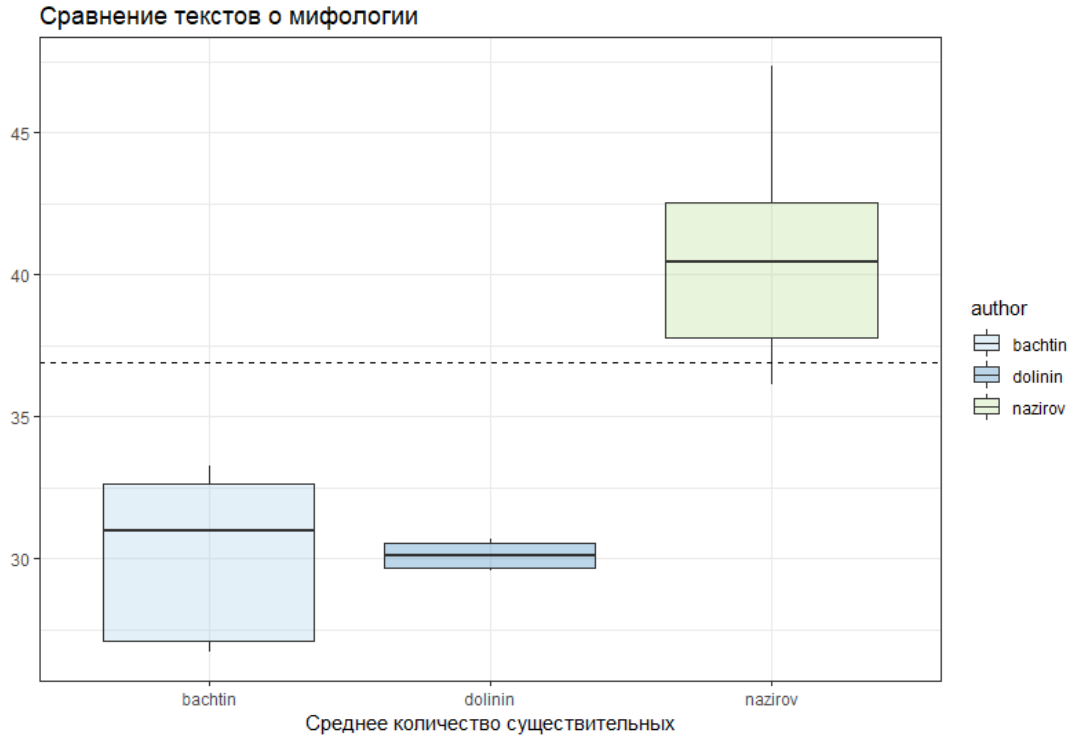


Рис. 36: Сравнение доли существительных в текстах о Ф. М. Достоевском

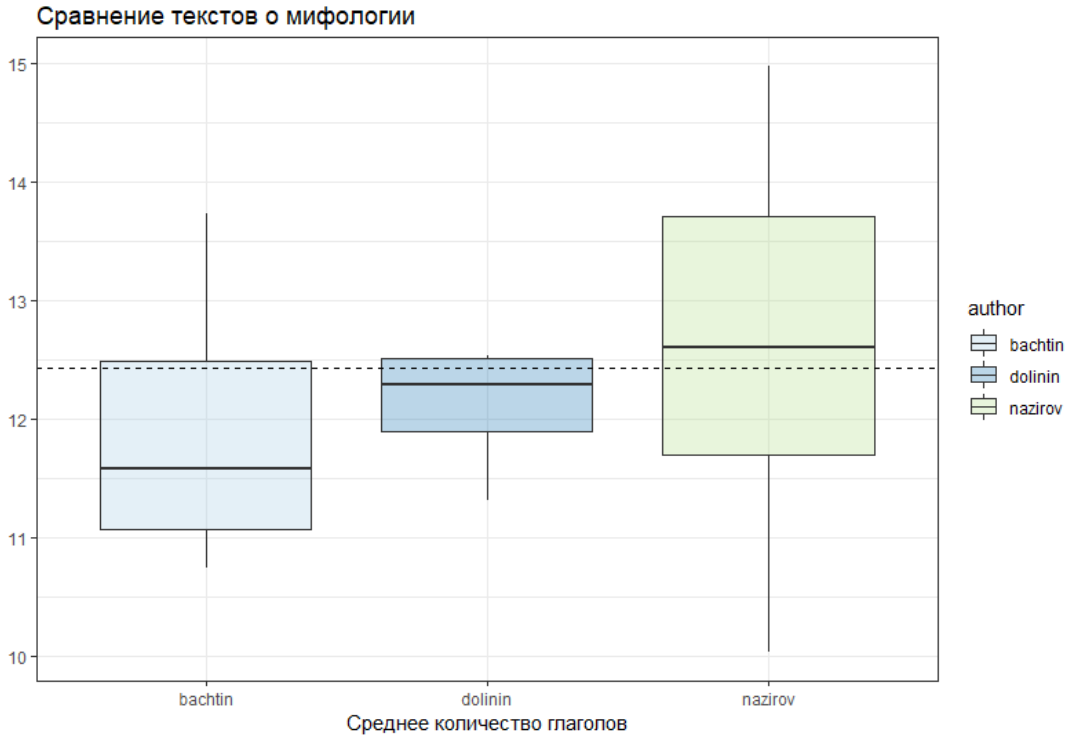


Рис. 37: Сравнение доли глаголов в текстах о Ф. М. Достоевском

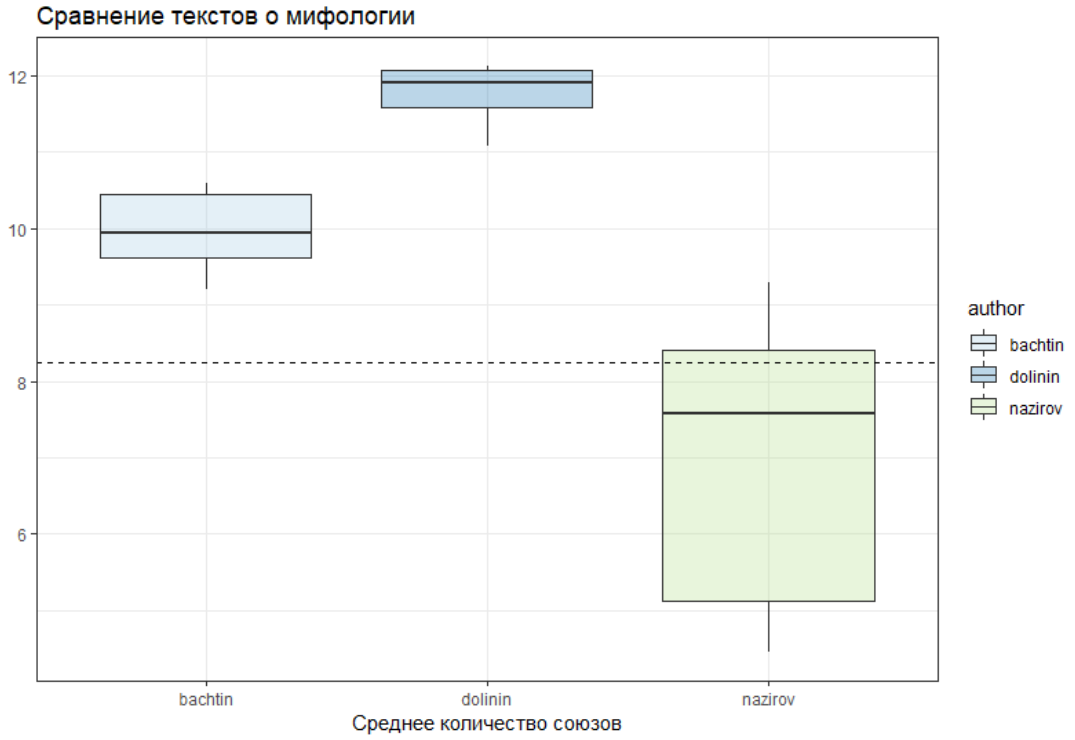


Рис. 38: Сравнение доли союзов в текстах о Ф. М. Достоевском

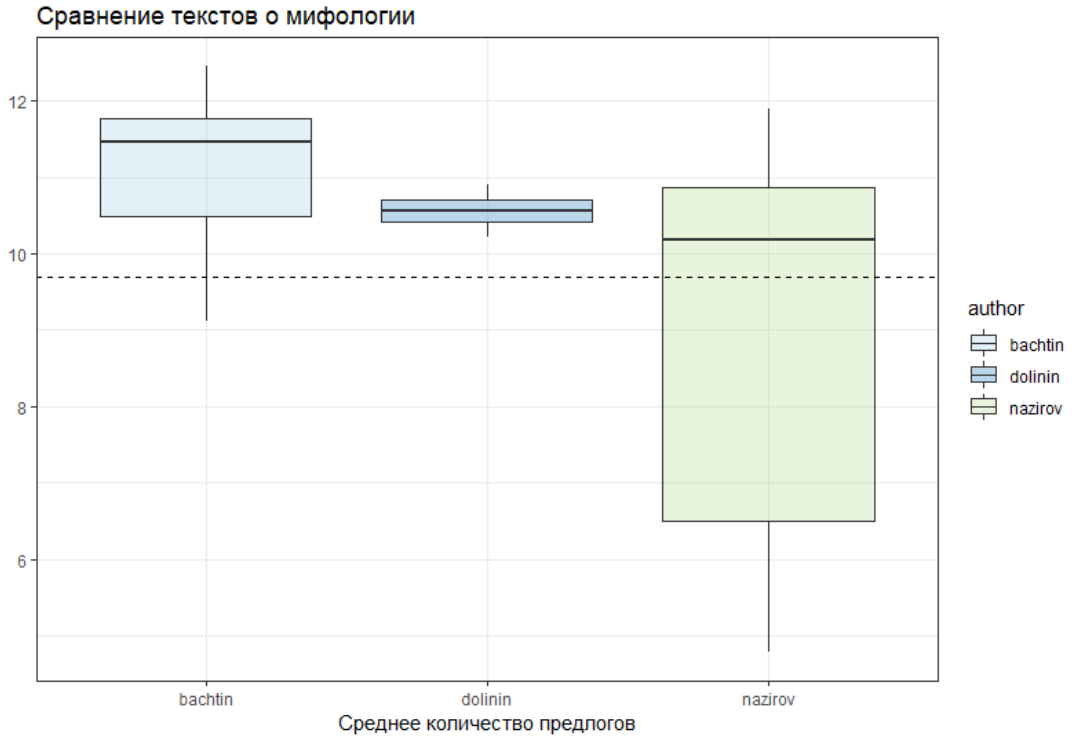


Рис. 39: Сравнение доли предлогов в текстах о Ф. М. Достоевском

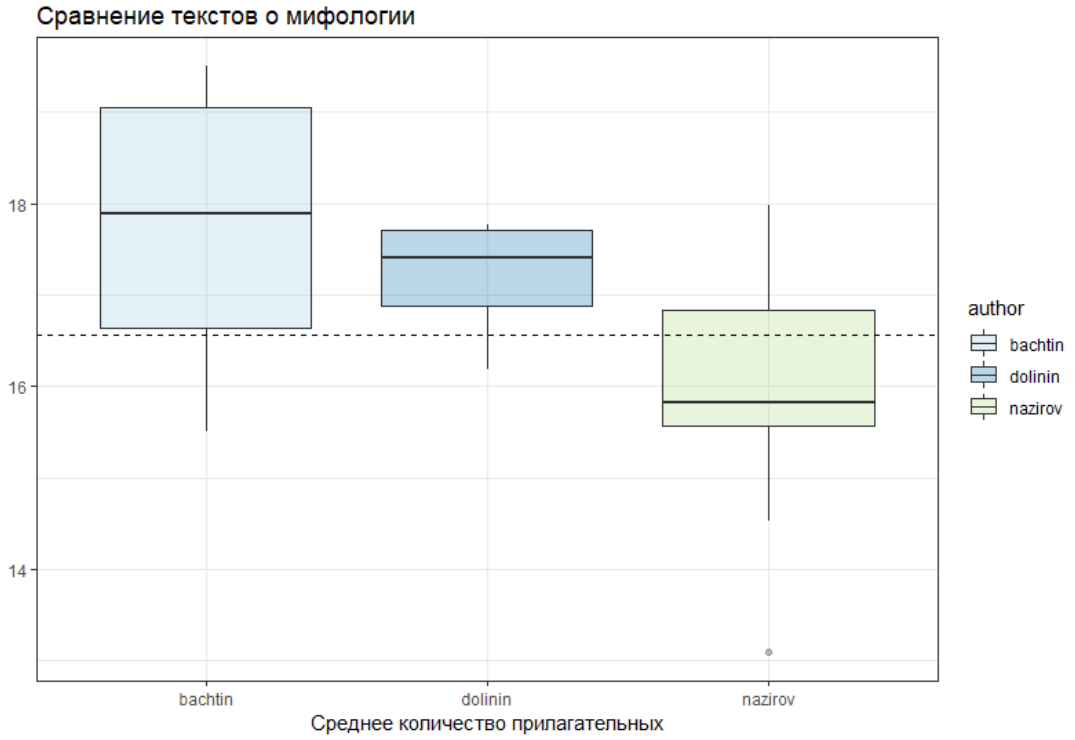


Рис. 40: Сравнение доли прилагательных в текстах о Ф. М. Достоевском

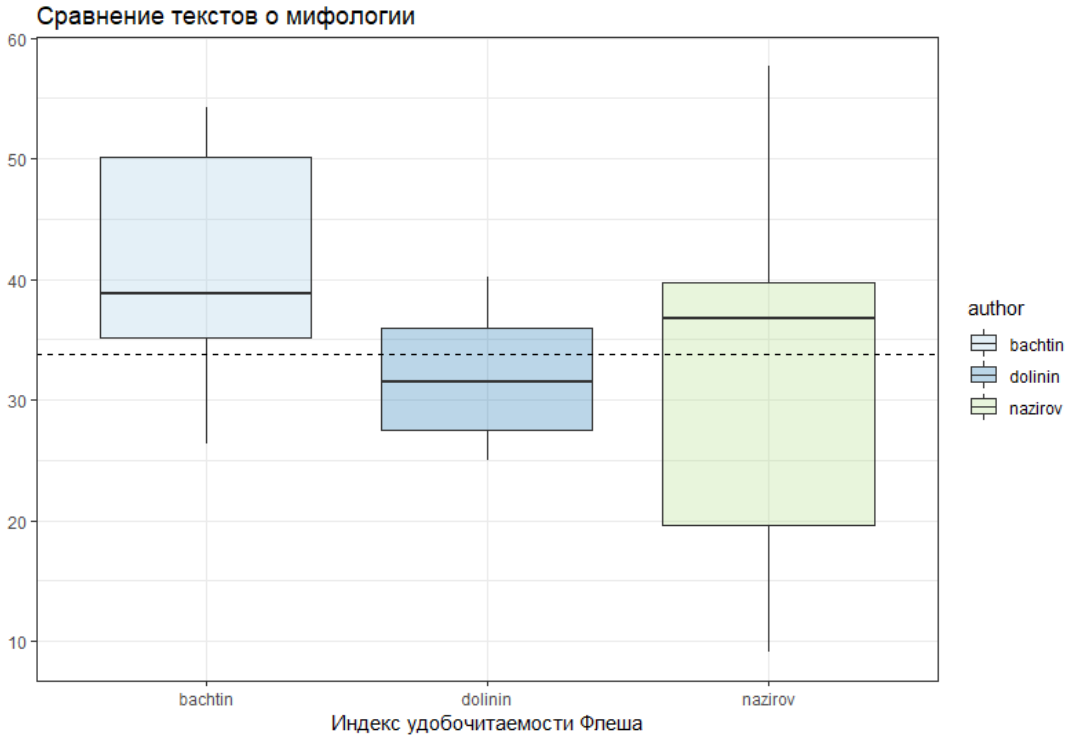


Рис. 41: Сравнение индекса удобочитаемости Флеша в текстах о Ф. М. Достоевском

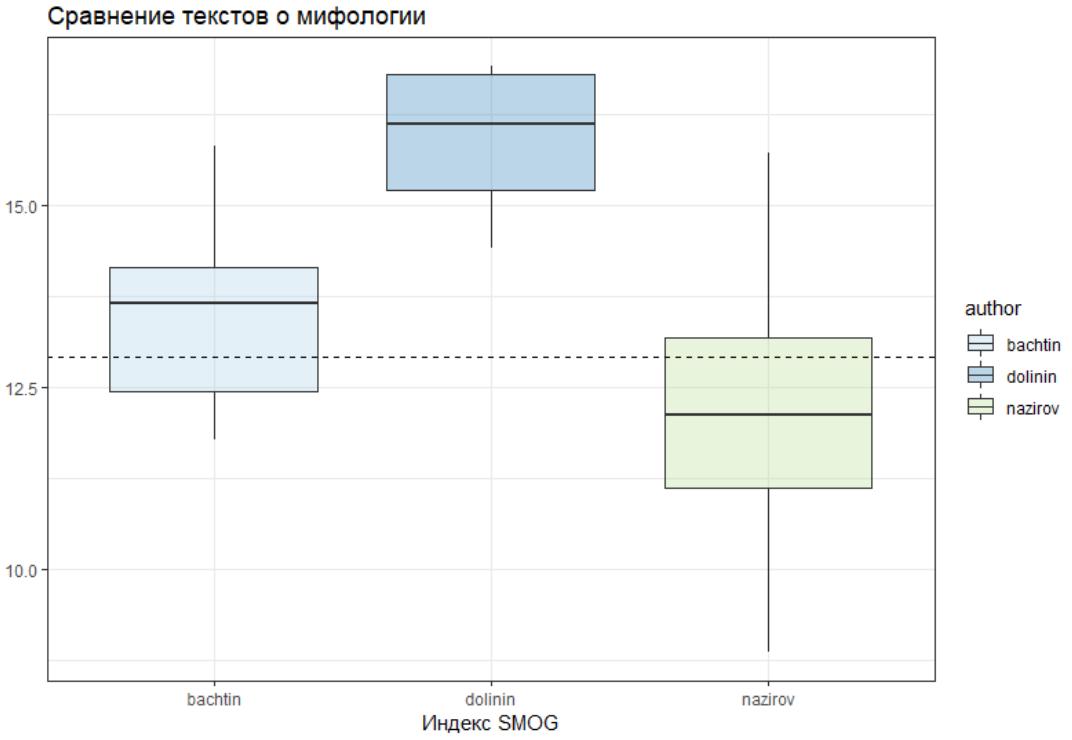


Рис. 42: Сравнение индекса SMOG в текстах о Ф. М. Достоевском

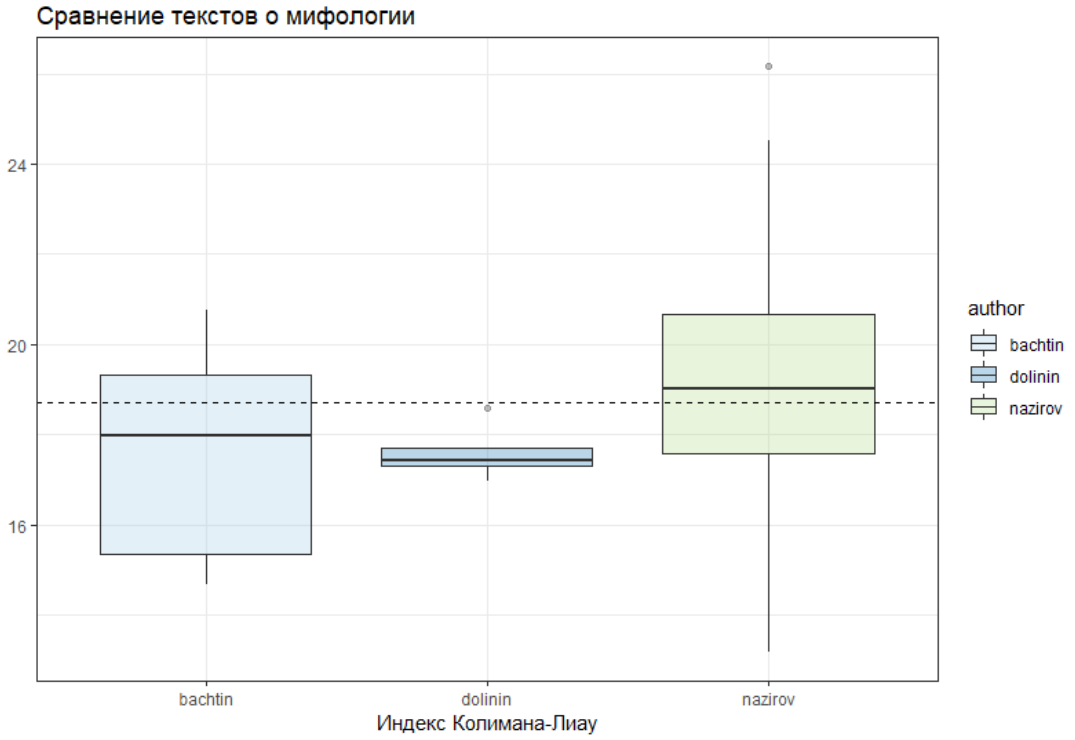


Рис. 43: Сравнение индекса Колиман-Лиау в текстах о Ф. М. Достоевском

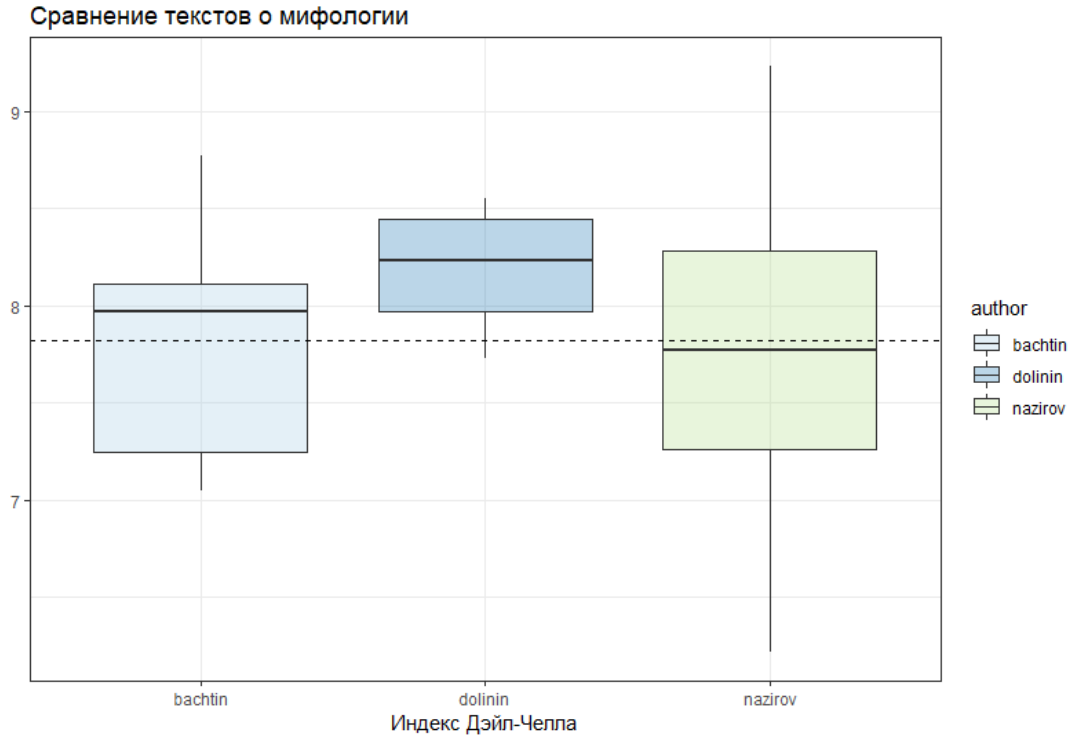


Рис. 44: Сравнение индекса Дэйл-Челла в текстах о Ф. М. Достоевском

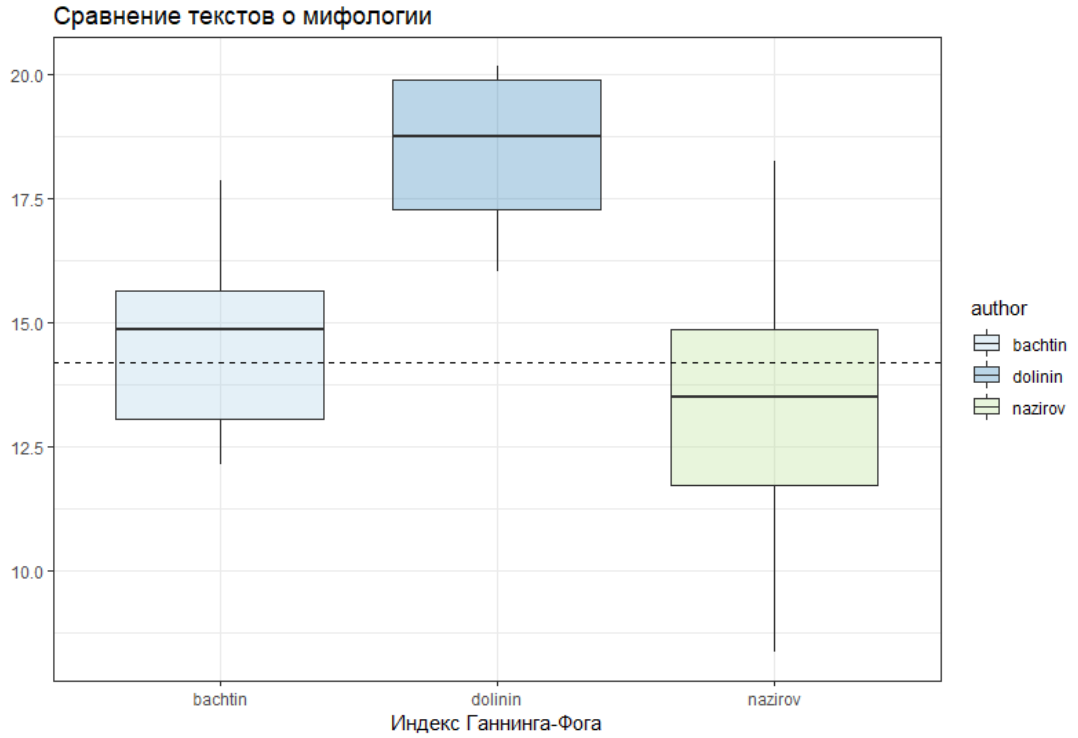


Рис. 45: Сравнение индекса Ганнинг-Фога в текстах о Ф. М. Достоевском