

Научная статья

УДК 81'322.2

DOI 10.25205/1818-7935-2026-24-1-102-112

Авторские инварианты в английских художественных текстах

Артем Павлович Ковалевский¹

Юлия Викторовна Павлова²

¹Институт математики СО РАН

Новосибирск, Россия

²Новосибирский государственный университет

Новосибирск, Россия

artyom.kovalevskii@gmail.com, <https://orcid.org/0000-0001-5808-3134>

pvlvaj@gmail.com, <https://orcid.org/0009-0002-1781-2630>

Аннотация

В работе предложена и протестирована методика обнаружения стилистической разладки в художественных текстах на английском языке, основанная на статистическом анализе бинарной последовательности, формируемой по признаку принадлежности слов к авторскому инварианту. В качестве авторского инварианта рассматривалось множество служебных слов, текст отображался в виде последовательности нулей и единиц, служившей для исследования стилистических особенностей текста. На основе этой последовательности был построен эмпирический мост, максимальное отклонение которого использовалось в качестве статистики нормы эмпирического моста для выявления возможной разладки. Метод был апробирован на корпусе из 100 художественных текстов британских и американских авторов XIX–XXI веков, а также на 9 000 попарных комбинаций склеенных текстов. Были выделены пороговые значения статистики, позволяющие отличать однородные тексты от текстов с возможной стилистической границей. Для оценки статистической значимости использовалась аппроксимация распределения Колмогорова, на основе которой рассчитывались *p*-значения. Результаты экспериментов показали: устойчивую способность метода выявлять стилистическую границу между произведениями разных авторов; эффективность алгоритма при различении одного текста одного автора и комбинации двух текстов одного автора; эффективность алгоритма при различении одного текста одного автора от пары текстов разных авторов; независимость метода от длины текста: высокие и низкие значения статистики и *p*-значения наблюдались при разных объемах. Анализ эмпирических мостов и поведения отдельных служебных слов показал, что стилистические особенности текста могут проявляться как в масштабах всего произведения, так и внутри отдельных его частей. Метод эффективно фиксирует подобные изменения. Разработанный подход может быть использован в рамках корпусной лингвистики, прикладной стилистики, обработки естественного языка, а также при создании интеллектуальных систем анализа текста.

Ключевые слова

служебные слова, авторский инвариант, эмпирический мост, обнаружение разладки, склейка текстов

Благодарности

Авторы искренне благодарят анонимного рецензента за тщательный анализ работы и ряд полезных замечаний, способствовавших улучшению изложения полученных результатов.

Финансирование

Работа выполнена при поддержке программы фундаментальных научных исследований СО РАН, проект FWNF-2026-0030.

© Ковалевский А. П., Павлова Ю. В., 2026

ISSN 1818-7935

Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация. 2026. Т. 24, № 1

Vestnik NSU. Series: Linguistics and Intercultural Communication, 2026, vol. 24, no. 1

Для цитирования

Ковалевский А. П., Павлова Ю. В. Авторские инварианты в английских художественных текстах // Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация. 2026. Т. 24, № 1. С. 102–112. DOI 10.25205/1818-7935-2026-24-1-102-112

Author's invariants in English literary texts

Artyom P. Kovalevskii¹, Yulia V. Pavlova²

¹Institute of Mathematics SB RAS
Novosibirsk, Russian Federation

²Novosibirsk State University
Novosibirsk, Russian Federation

artyom.kovalevskii@gmail.com, <https://orcid.org/0000-0001-5808-3134>
pvlvaj@gmail.com, <https://orcid.org/0009-0002-1781-2630>

Abstract

The paper proposes and tests a method for detecting stylistic change points in English-language fiction texts based on statistical analysis of a binary sequence taking into account the words belonging to the author's invariant. A set of function words was considered as the author's invariant; texts were displayed as sequences of zeros and ones used to study the stylistic features of texts. An empirical bridge was constructed based on this sequence. The maximum deviation of the empirical bridge served as statistical norm to detect possible change points. The method was tested on a corpus of 100 novels by British and American authors of the 19th–21st centuries, as well as on 9,000 pairwise combinations of concatenated texts. Threshold statistical values are identified that make it possible to distinguish homogeneous texts and texts with a possible stylistic change point. An approximation of the Kolmogorov distribution is used to assess statistical significance. P-values are calculated on the basis of the limiting Kolmogorov distribution. The results of the experiments show a stable ability of the method to identify a stylistic change point between the novels by different authors; the efficiency of the algorithm in distinguishing one text by one author and a combination of two texts by one author; the efficiency of the algorithm in distinguishing one text by one author and a pair of texts by different authors; independence of the method of the text length, that is, high and low values of statistics and p-values are observed for different text lengths. The analysis of empirical bridges and the behavior of individual function words show that stylistic features of the text can manifest themselves both on the scale of the entire work and within its individual parts. The method effectively records such changes. The developed approach can be used in the framework of corpus linguistics, applied stylistics, natural language processing, as well as in the creation of intelligent text analysis systems.

Keywords

function words, author's invariant, empirical bridge, change point detection, concatenation

Acknowledgments

The authors sincerely thank the anonymous referee for a thorough analysis of the paper and for a number of helpful comments that contributed to improving the presentation of the results.

Funding

This work was supported by the SB RAS Fundamental Scientific Research Program, project FWNF-2026-0030.

For citation

Kovalevskii A. P., Pavlova Y. V. Author's invariants in English literary texts. *Vestnik NSU. Series: Linguistics and Intercultural Communication*, 2026, vol. 24, no. 1, pp. 102–112. DOI 10.25205/1818-7935-2026-24-1-102-112

Введение

При анализе текстов все чаще используются методы математической статистики. В частности, склейка двух текстов разных авторов может выявляться методами обнаружения разладки случайной последовательности. В отличие от алгоритмов обнаружения разладки в реальном времени, для анализа склейки текстов используются алгоритмы апостериорного обнаружения разладки: вся изучаемая последовательность известна исследователю от начала до конца,

и надо сделать основанный на методах математической статистики вывод о том, изменились ли свойства последовательности в процессе наблюдения. Текст переводится в последовательность нулей и единиц на основании попаданий слов в авторский инвариант – множество служебных слов. Этот подход с успехом применялся к анализу текстов на русском языке [Гусарова, Ковалевский, Макаренко, 2005], в данной работе проверяется его применимость к анализу текстов на английском языке. Целью настоящего исследования является разработка и тестирование метода авторского инварианта для обнаружения склейки текстов на английском языке. Материалами исследования выступают 100 длинных текстов на английском языке. Используются романы и длинные повести десяти американских и британских авторов XIX, XX и XXI веков. Ключевой метод исследования – обнаружение точки разладки в склейке текстов.

Важной задачей прикладной лингвистики является выявление авторских особенностей в тексте. В частности, обнаружение точки разладки в тексте может отражать смену автора, темы, стиля и тона текста, а также изменения в выборе слов. Исследователи, работающие в области обработки естественного языка и области статистических языковых моделей, разработали ряд методов обнаружения изменений текста, начиная от обнаружения одной точки разладки до обнаружения множественных изменений, от одномерных данных до многомерных данных, от однопараметрических изменений до многопараметрических изменений.

Среди статей, использующих алгоритмический подход к сравнению текстов, следует выделить работу [Dembele and Lo, 2015], где рассмотрен механизм приближенного оценивания степени сходства между текстами на основе индекса Жаккара. Введены как классические реализации алгоритма RU (Rajaraman–Ullman), так и его модификации, использующие хеш-функции и методы сегментирования для повышения вычислительной эффективности. Апробация была проведена на корпусе евангельских текстов, где метод показал способность различать тексты с различной степенью семантической и лексической близости: более высокую близость метод определил между синоптическими Евангелиями, которые он смог отличить от Евангелия от Иоанна.

Дополнительную эмпирическую основу для анализа структуры текста составляет статья [Németh and Zainkó, 2002], в которой проводится статистическое сравнение корпусов на венгерском, немецком и английском языках. Особое внимание уделяется распределению служебных слов, а также их позиционным функциям в пределах предложения. Показано, что служебные слова обладают высокой частотной устойчивостью и распределяются по тексту с предсказуемыми закономерностями. Авторы подтверждают, что именно служебные слова, в отличие от лексически нагруженных единиц, лучше всего подходят для задач стилистического и структурного анализа.

Также были разработаны различные методы для анализа обнаружения точек изменения текста, например, предложенный [Hearst, 1997] метод разбиения текста на фрагменты. Новый статистический подход к разбиению текста был представлен [Beeferman et al., 1999]. Этот подход постепенно строит экспоненциальную модель для извлечения признаков, которые коррелируют с наличием границ в маркированном обучающем тексте. В этой работе используются два типа признаков: признаки соответствия определенной теме, основанные на использовании адаптивных языковых моделей новым способом для обнаружения существенных изменений темы, и признаки ключевых слов, специфичных для фрагмента текста и обычно используемых вблизи границ фрагмента. Авторы представили новую метрику ошибок, которая естественным образом сочетает точность и полноту. Ее введение мотивируется формулой Байеса. Эта метрика используется для количественной оценки относительного вклада различных типов признаков, а также для сравнения с ранее предложенными алгоритмами сегментации текста. Статья, в которой указаны ограничения предыдущей статьи, была написана [Pevzner and Hearst, 2002] и представила улучшенный метод. Согласно этому исследованию, метод, разработанный [Beeferman et al., 1999], имеет серьезное ограничение: метод слишком сильно штрафует ложные отрицательные и ложные положительные результаты и не учитывает близкие промахи.

Вероятностный алгоритм, использующий кластеризацию, был предложен [Choi, 2000]. Напротив, [Mikolov et al., 2013] разработали метод измерения сходства, используя косинус угла, соответствующего их встроенному вектору. Новый подход к сегментации текста без учителя, основанный на графе семантической связанности, путем идентификации узлов, которые имеют более высокие семантические расстояния от других узлов, был предложен [Nanni et al., 2016].

Работа по тесту однородности текста, основанному на моделях урн, использующих прямой и обратный процесс подсчета количества различных слов, была предложена [Abebe et al., 2022]. В статье [Abebe et al., 2024] авторы развивают этот подход и адаптируют его для поиска точки склейки. Был разработан алгоритм для измерения процентной ошибки для точки смены текста. Текст рассматривается в виде последовательности слов, и авторы [Abebe et al., 2024] предлагают отслеживать, как изменяется количество различных слов при движении по тексту вперед и назад, построить эмпирический мост и по отклонениям кривой эмпирического моста от условной линейности определить возможную границу между стилевыми или авторскими сегментами. Метод демонстрирует высокую точность при выделении границ в текстах, что подтверждается на примере Брауновского корпуса, русскоязычных новостей и текстах на языке тигринья.

В отличие от анализа уникальных слов, как в статье [Abebe et al., 2024], метод, представленный в настоящей работе, фокусируется на заранее фиксированном множестве. Это позволяет снизить шум и повысить чувствительность к локальным стилистическим изменениям.

Таким образом, в рамках современной статистической лингвистики наблюдается устойчивый интерес к разработке моделей, способных выявлять внутреннюю неоднородность текста и структурные границы на основе частотных характеристик. Представленные выше подходы демонстрируют как теоретическую обоснованность, так и практическую эффективность при работе с текстами различной языковой природы, длины и жанра.

1. Алгоритм анализа однородности текста

Метод построен на принципе инвариантности и использует норму эмпирического моста как критерий для выявления точки разладки в последовательности наблюдений. Предполагается, что наблюдаемая последовательность значений может быть разнородной: до момента $T \in (0,1)$ наблюдения имеют распределение F_1 , а после – F_2 , при этом математические ожидания различны, а дисперсии конечны. В качестве исходной последовательности рассматриваются индикаторы $X_1, X_2, \dots, X_n$, где каждое значение X_i определяется следующим образом:

$X_i = 1$, если i -е слово текста принадлежит словарю служебных слов;

$X_i = 0$ в противном случае.

Так, X_i отражает принадлежность i -го слова к множеству служебных слов. На основе этой бинарной последовательности вычисляются кумулятивные суммы, которые затем центрируются и нормируются, как описано в [Гусарова, Ковалевский, Макаренко, 2005]. Результатом является процесс $Z_n(t)$, называемый эмпирическим мостом. Для анализа однородности используется статистика максимума модуля эмпирического моста $M = M_n$. Для оценки p -значения (реально достигаемого уровня значимости) используется распределение Колмогорова: $p = 1 - K(M_n)$. Малые значения p (например, $p < 0,05$) свидетельствуют о наличии стилистической разладки. Точка T_n , в которой достигается максимум $|Z_n(t)|$, считается оценкой момента разладки. Эта точка интерпретируется как граница между двумя фрагментами с разной стилиевой природой. Если в экспериментах с текстами точка T_n практически совпадает с реальной границей между двумя произведениями, то можно говорить об эффективности метода.

2. Практическая апробация алгоритма

Важным этапом исследования является сбор коллекции из ста текстов художественной литературы на английском языке. Коллекция состоит из романов и длинных повестей десяти американских и британских авторов XIX, XX и XXI веков.

Agatha Christie: "Murder in Mesopotamia", "The ABC Murders", "The Big Four", "The Man in the Brown Suit", "The Murder of Roger Ackroyd", "The Murder on the Links", "The Mysterious Affair at Styles", "The Mystery of the Blue Train", "The Secret Adversary", "The Secret of Chimneys".

Louisa May Alcott: "Behind a Mask, or a Woman's Power", "Little Women", "An Old Fashioned Girl", "Little Men", "Eight Cousins", "Rose in Bloom: A Sequel to Eight Cousins", "Jack and Jill: A Village Story", "Flower Fables", "Jo's Boys and How They Turned Out: A Sequel to "Little Men", "Moods".

Charles Dickens: "A Tale of Two Cities", "Bleak House", "David Copperfield", "Great Expectations", "Hard Times", "Little Dorrit", "Old Curiosity Shop", "Oliver Twist", "The Magic Fishbone", "The Mystery of Edwin Drood".

Conan Doyle: "A Study in Scarlet", "Micah Clarke", "Rodney Stone", "Sir Nigel", "The Hound of the Baskervilles", "The Lost World", "The Refugees", "The Sign of the Four", "The Valley of Fear", "The White Company".

Ray Bradbury: "A Graveyard for Lunatics", "Dandelion Wine", "Death Is a Lonely Business", "Fahrenheit 451", "Farewell Summer", "From the Dust Returned", "Green Shadows White Whale", "Let's All Kill Constance", "Something Wicked This Way Comes", "The Martian Chronicles".

Joanne Rowling: "Harry Potter and the Chamber of Secrets", "Harry Potter and the Deathly Hallows", "Harry Potter and the Goblet of Fire", "Harry Potter and the Half-Blood Prince", "Harry Potter and the Order of the Phoenix", "Harry Potter and the Prisoner of Azkaban", "Harry Potter and the Sorcerer's Stone", "Quidditch Through the Ages", "The Casual Vacancy", "The Tales of Beedle the Bard".

Robert Louis Stevenson: "Catriona", "Kidnapped", "Prince Otto", "The Black Arrow: A Tale of the Two Roses", "The Master of Ballantrae: A Winter's Tale", "The Strange Case of Dr. Jekyll and Mr. Hyde", "The Wrecker", "The Wrong Box", "Treasure Island", "Weir of Hermiston".

Anthony Trollope: "Barchester Towers", "Can You Forgive Her", "Doctor Thorne", "Framley Parsonage", "Phineas Redux", "The Last Chronicle of Barset", "The Prime Minister", "The Small House at Allington", "The Warden", "The Way We Live Now".

Mark Twain: "A Connecticut Yankee in King Arthur's Court", "A Tramp Abroad", "Adventures of Huckleberry Finn", "Life on the Mississippi", "My Mysterious Stranger", "Personal Recollections of Joan of Arc", "Roughing It", "The Adventures of Tom Sawyer", "The American Claimant", "The Prince and the Pauper".

Roger Zelazny: "Nine Princes in Amber", "Blood of Amber", "Knight of Shadows", "Prince of Chaos", "Sign of Chaos", "Sign of the Unicorn", "The Courts of Chaos", "The Guns of Avalon", "The Hand of Oberon", "Trumps of Doom".

Словарем авторских инвариантов послужил набор из служебных слов на английском языке, состоящий:

- 1) из предлогов: in, on, from, for, to, by, at, without, about, through, with, over, under;
- 2) союзов: and, but, although, when, if, or, whether;
- 3) частиц: what, too, as, not, even, only, here, that, there, after, all, either, already.

В рамках настоящего исследования были проанализированы литературные произведения с объемом от 10 тыс. до 300 тыс. слов. Для корректной работы программы тексты были очищены от пунктуации и переведены в нижний регистр. Анализ проводился как для каждого из 100 произведений по отдельности, так и для всех 9 000 парных сочетаний, образованных последовательным объединением текстов. Для этого были рассчитаны максимумы модуля эм-

пирического моста индикаторов служебных слов для каждого текста, для всех попарных комбинаций текстов одного автора и для всех попарных комбинаций текстов разных авторов.

В ходе работы программы текстам были сопоставлены последовательности нулей и единиц на основании описанного выше алгоритма. Рассмотрим следующий отрывок романа Артура Конана Дойля “The Hound of the Baskervilles”:

“Holmes was sitting with his back to me, and I had given him no sign of my occupation.”

Этот фрагмент разбивается на последовательность слов, и каждое слово сопоставляется с инвариантом. Если слово входит в заданный список служебных слов, оно кодируется единицей (1), иначе – нулем (0):

[0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0]

Для каждого текста строились эмпирические мосты Z_n . Разработанная методика позволяет решать следующие задачи:

- 1) отличать один текст от пары текстов, написанных одним автором;
- 2) различать один текст одного автора и два текста разных авторов.

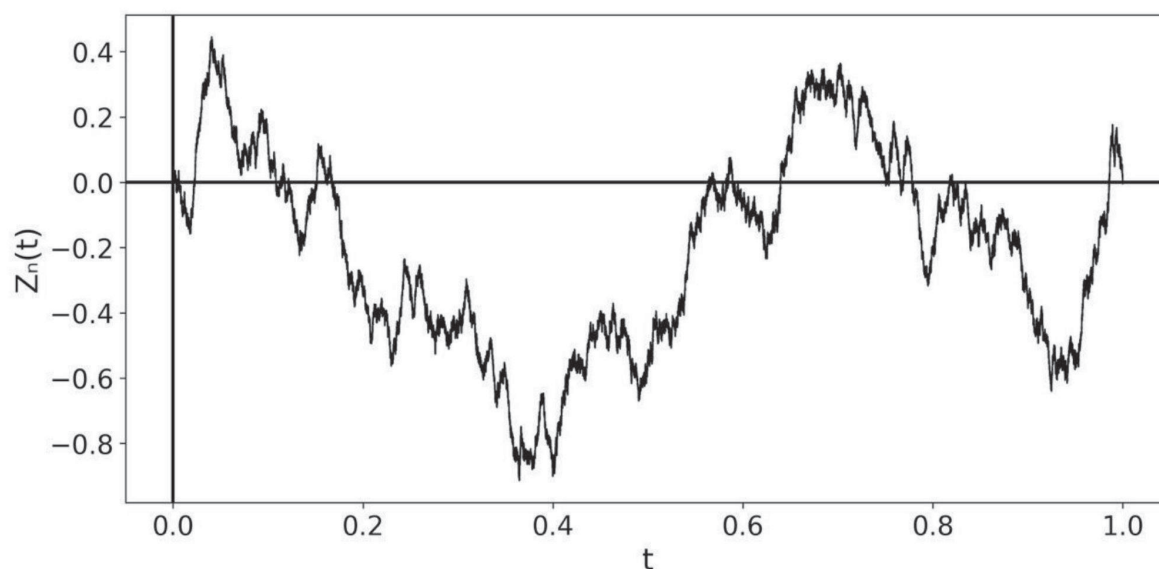


Рис. 1. График эмпирического моста описаний в повести Джоан Роулинг “Квиддич с древности до наших дней”

Fig. 1. The empirical bridge graph of descriptions in *Quidditch Through the Ages* by Joanne Rowling

Метод был апробирован на корпусе английских художественных текстов, в промежуточных результатах исследования было выявлено, что наиболее однородным текстом одного автора является произведение Джоан Кэтлинг Роулинг «Квиддич с древности до наших дней» (“*Quidditch Through the Ages*”), т. е. оно обладает минимальным значением статистики (максимума модуля эмпирического моста) среди текстов коллекции.

Для последующего анализа текста была использована разметка, отделяющая диалоги без текстов описания и описание без диалогов. На графике эмпирического моста (рис. 1) представлен текст описательной части произведения. График демонстрирует высокую степень статистической однородности, так как прослеживаются хаотические колебания вблизи горизонтальной оси, характерные для броуновского моста. В описательной части текста эмпирический мост сохраняет небольшую амплитуду (максимум модуля около 0,43). Подобным образом ведут себя и диалоги, отклонения невелики (до 0,58). Этот анализ показывает, что произве-

дение “Quidditch Through the Ages” демонстрирует однородную структуру. И в описательной части произведения, и в диалогах эмпирический мост сохраняет небольшую амплитуду.

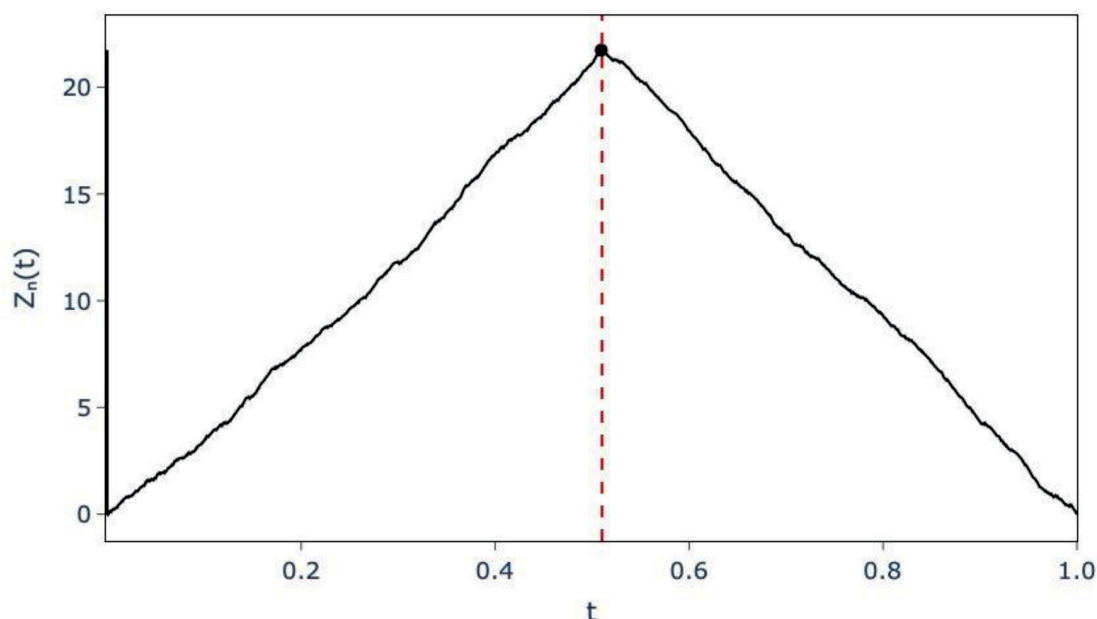


Рис. 2. График эмпирического моста для конкатенации текстов Энтони Троллопа «Фремлейский приход» (“Framley Parsonage”) и Джоан Кейтлин Роулинг «Гарри Поттер и Кубок Огня» (“Harry Potter and the Goblet of Fire”)

Fig. 2. An empirical bridge graph for the concatenation of Anthony Trollope’s *Framley Parsonage* and J. K. Rowling’s *Harry Potter and the Goblet of Fire*

Что касается максимального значения статистики двух текстов разных авторов, самой разнородной парой текстов оказались произведения Энтони Троллопа «Фремлейский приход» (“Framley Parsonage”) и Джоан Кейтлин Роулинг «Гарри Поттер и Кубок Огня» (“Harry Potter and the Goblet of Fire”). Оба текста были объединены в единую последовательность, в которой сначала идет полный текст Троллопа, затем – полный текст Роулинг. Далее была построена бинарная последовательность длиной 416 768 слов, отражающая наличие каждого слова в авторском инварианте (множестве служебных слов). Затем по этой последовательности был построен эмпирический мост.

Результаты анализа:

1. Индекс точки склейки: 212 043 слова.
2. Индекс максимального отклонения эмпирического моста: 212 043 слова.
3. Значение статистики M_n : 21,737.
4. Нормированная доля текста в точке разладки: 0,509.

Совпадение максимального отклонения с точкой склейки указывает на то, что метод успешно зафиксировал смену авторского стиля. Визуализация эмпирического моста (рис. 2) показывает характерный излом, соответствующий границе между двумя текстами. До точки склейки кривая меняется плавно, а после нее наблюдается резкий разворот в противоположную сторону – характерный признак стилистической разладки.

Для проверки устойчивости метода к длине текста были построены графики зависимости значения статистики M_n от числа слов в тексте и от автора (рис. 3). Значения M_n варьируются в диапазоне от 0,3 до 2,6 и не демонстрируют выраженной зависимости от длины текста.

Высокие значения M_n (свыше 1,3), указывающие на стилистическую разладку, наблюдаются как в текстах средней длины, так и в достаточно объемных произведениях. Однако таких текстов сравнительно немного (20 из 100, т. е. 20 %). Низкие значения M_n , характерные для однородных текстов, встречаются при любых объемах. Таким образом, метод оказывается устойчивым к изменению длины текста в диапазоне от 20 000 до 700 000 слов.

Пороговые значения статистики M_n , используемые для анализа, были определены на основе экспериментальных наблюдений и частично подтверждаются данными из предыдущих исследований: значения ниже 0,6 обычно соответствуют однородным текстам, а значения выше 1,3 – текстам со стилистической разладкой. Тексты, попадающие в промежуточный диапазон, могут содержать слабовыраженную границу между частями. Метод может фиксировать стилистическую разладку даже в пределах одного текста, если внутри произведения наблюдается резкая смена регистров речи (например, от описательного повествования к диалогам). Отметим, что распределение p -значений для текстов одного автора близко к равномерному, что и предсказывает теория для однородных последовательностей. Это подтверждает применимость элементарной модели авторского инварианта к анализу текстов на английском языке.

Критерий на основе статистики M_n и соответствующего уровня значимости позволяет выявлять разладку в текстах различных размеров. Он показывает устойчивость и применимость как к коротким, так и к длинным текстам, что важно при анализе реальных корпусных данных.

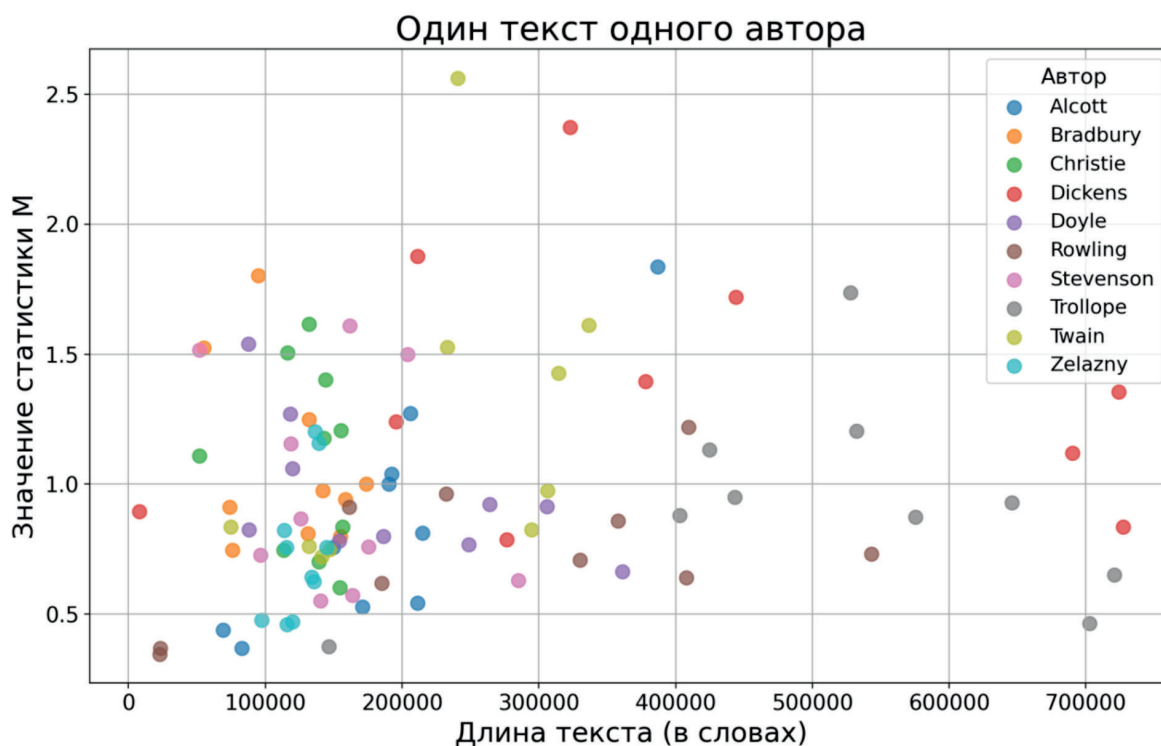


Рис. 3. Зависимость M_n от числа слов в тексте и от автора

Fig. 3. Dependence of M_n on the number of words in the text and on the author

Также рис. 3 демонстрирует, что авторов нельзя различить по степени однородности текста – у каждого автора встречаются тексты как с высокой, так и с низкой однородностью. Высокой однородности соответствуют малые значения M , а низкой – большие.

Были изучены комбинации текстов одного автора, таких комбинаций 900. Статистика M для комбинаций двух текстов одного автора меняется в пределах от 0,4 до 7,5, т. е. в целом

такие комбинации значительно менее однородны, чем тексты, взятые поодиночке. Наилучшую однородность показывают два романа Алькотт, опубликованные с интервалом в 16 лет, а наихудшую – две повести Диккенса. Если выбрать значение $C = 1,3$ в качестве границы однородности, то только для 204 комбинаций текстов одного автора из 900 (23 %) значения статистики M меньше этой границы. Итак, мы можем различать один текст одного автора и комбинацию двух текстов одного автора, выбрав $C = 1,3$ в качестве границы для M , при этом относительные частоты ошибок в ту и другую сторону составят 20 и 23 %.

Также проанализированы все попарные комбинации текстов двух разных авторов. К каждому из 10 произведений каждого из 10 авторов присоединено каждое из оставшихся 90 произведений других авторов, поэтому таких комбинаций 9 000. Здесь значения M меняются в пределах от 0,5 до 21,8. Наилучшая однородность у текстов одной эпохи (Алькотт и Диккенс, Роулинг и Желязны). Наихудшая однородность при комбинации текстов Роулинг и Троллопа.

Если использовать в качестве границы $C = 1,3$, то выясняется, что только для 740 из 9 000 комбинаций текстов двух авторов (т. е. 8 %) значение M меньше этой границы. Поэтому мы можем отличать один текст одного автора от двух текстов разных авторов с относительными частотами ошибок в 20 и 8 % в одну и другую сторону.

Заключение

Исследование было задумано как попытка выработать универсальный алгоритм, позволяющий находить склейку текстов произвольной длины. Кроме того, была надежда на то, что склейки текстов одного и того же автора будут демонстрировать гораздо более высокую однородность, чем склейки текстов любых разных авторов.

Статистический тест, выявляющий наличие неоднородности, основан на сопоставлении последовательным словам текста нулей и единиц. Для этого фиксируется множество служебных слов (предлогов, союзов и частиц). Текст переводится в нижний регистр и освобождается от пунктуации, а затем последовательные слова текста сопоставляются с множеством служебных слов. В результате формируется последовательность чисел (индикаторов появления служебных слов), в которой служебным словам соответствуют единицы, а остальным словам – нули. Длина этой последовательности равна числу слов текста.

По последовательности нулей и единиц строится эмпирический мост – центрированная и нормированная последовательность кумулятивных сумм. P -значение (реально достигаемый уровень значимости) гипотезы об однородности текста вычисляется на основании очень простой нулевой гипотезы о независимости и одинаковой распределенности индикаторов служебных слов. Точка наибольшего отклонения эмпирического моста от нуля служит оценкой точки склейки текстов, а абсолютная величина наибольшего отклонения характеризует p -значение гипотезы об однородности, т. е. вероятность того, что такое же или большее отклонение будет иметь место при верной нулевой гипотезе. В качестве критического значения отклонения выбрана величина 1,3, приблизительно соответствующая пятипроцентному размеру статистического теста.

Проведен анализ 100 текстов на английском языке (по 10 текстов десяти разных авторов), 900 попарных комбинаций текстов одного и того же автора и 9 000 попарных комбинаций текстов разных авторов. Все тексты имеют значительный объем и существенно отличаются в объеме друг от друга: от 13 000 до 750 000 слов. Предложенный статистический тест позволяет успешно (с малыми вероятностями ошибок) отличать склейку (конкатенацию) текстов двух разных авторов от одного текста одного автора. Почти настолько же успешно он отличает один текст одного автора от склейки двух текстов одного автора. Однако этот тест недостаточно уверенно различает два текста одного автора от двух текстов разных авторов. Выяснилось, что в разных произведениях один и тот же автор может совершенно по-разному использовать

служебные слова, и в этом отношении два текста одного автора часто являются существенно разнородными.

Остался открытым вопрос о том, насколько длинные тексты нужны для уверенного обнаружения склейки текстов. В проведенном исследовании использовались тексты значительной длины, позволяющей накопить необходимую информацию для принятия решения об однородности.

Возможным продолжением исследований может стать изучение последовательности векторов, ассоциированных со словами. В рамках этого подхода можно использовать существующие векторизации или изучить широкий класс возможных векторизаций и выбрать из него наиболее подходящую для обнаружения склейки текстов.

Список литературы

- Гусарова Г. В., Ковалевский А. П., Макаренко А. Г. Критерии наличия разладки // Сибирский журнал индустриальной математики. 2005. Т. 8, № 4. С. 18–33. URL: <https://www.mathnet.ru/sjim273>
- Abebe B., Chebunin M., Kovalevskii A. Text Segmentation Via Processes that Count the Number of Different Words Forward and Backward // Journal of Quantitative Linguistics. 2024. Vol. 31. No. 1. P. 1–18. DOI: <https://doi.org/10.1080/09296174.2023.2275342>
- Abebe B., Chebunin M., Kovalevskii A., Zakrevskaya N. Statistical tests for text homogeneity: Using forward and backward processes of numbers of different words // Glottometrics. 2022. Vol. 53. No. 1. P. 42–58. DOI: https://doi.org/10.53482/2022_53_401
- Beeferman D., Berger A., Lafferty J. D. Statistical models for text segmentation // Machine Learning. 1999. Vol. 34. No. 1–3. P. 177–210. DOI: <https://doi.org/10.1023/A:1007506220214>
- Choi F. Y. Y. Advances in domain independent linear text segmentation // 1st Meeting of the North American Chapter of the Association for Computational Linguistics. URL: <https://aclanthology.org/A00-2004.pdf>
- Dembele S., Lo G. S. Probabilistic, Statistical and Algorithmic Aspects of the Similarity of Texts and Application to Gospels Comparison // Journal of Data Analysis and Information Processing. 2015. Vol. 3. P. 112–127. DOI: <https://doi.org/10.4236/jdaip.2015.34012>
- Hearst M. A. Text tiling: Segmenting text into multi-paragraph subtopic passages // Computational Linguistics. 1997. Vol. 23. No. 1. P. 33–64.
- Itoh N., Kurths J. Change-point detection of climate time series by nonparametric method // Proceedings of the World Congress on Engineering and Computer Science. 2010. Vol. 1. P. 445–448.
- Mikolov T., Sutskever I., Chen K., Corrado G. S., Dean J. Distributed representations of words and phrases and their compositionality // Advances in Neural Information Processing Systems. 2013. Vol. 26. P. 1–9.
- Nanni G., Glavaš F., Ponzetto S. P. Unsupervised text segmentation using semantic relatedness graphs // In: Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics (*SEM). Berlin, Germany, 2016.
- Németh G., Zainkó C. Multilingual Statistical Text Analysis, Zipf's Law and Hungarian Speech Generation // Acta Linguistica Hungarica. 2002. Vol. 49. No. 3–4. P. 385–405.
- Pevzner L., Hearst M. A. A critique and improvement of an evaluation metric for text segmentation // Computational Linguistics. 2002. Vol. 28. No. 1. P. 19–36. DOI: <https://doi.org/10.1162/089120102317341756>

References

- Gusarova G. V., Kovalevskii A. P., Makarenko A. G.** Criteria for the existence of a change point. *Sibirskii Zhurnal Industrial'noi Matematiki*, 2005, vol. 8, no. 4, pp. 18–33. (in Russ.)
- Abebe B., Chebunin M., Kovalevskii A.** Text Segmentation Via Processes that Count the Number of Different Words Forward and Backward. *Journal of Quantitative Linguistics*, vol. 31, no. 1, 2024, pp. 1–18. DOI: <https://doi.org/10.1080/09296174.2023.2275342>
- Abebe B., Chebunin M., Kovalevskii A., Zakrevskaya N.** Statistical tests for text homogeneity: Using forward and backward processes of numbers of different words. *Glottometrics*, vol. 53, no. 1, 2022, pp. 42–58. DOI: https://doi.org/10.53482/2022_53_401
- Beeferman D., Berger A., Lafferty J. D.** Statistical models for text segmentation. *Machine Learning*, vol. 34, no. 1–3, 1999, pp. 177–210. DOI: <https://doi.org/10.1023/A:1007506220214>
- Choi F. Y. Y.** Advances in domain independent linear text segmentation. *1st Meeting of the North American Chapter of the Association for Computational Linguistics*. URL: <https://aclanthology.org/A00-2004.pdf>
- Dembele S., Lo G. S.** Probabilistic, Statistical and Algorithmic Aspects of the Similarity of Texts and Application to Gospels Comparison. *Journal of Data Analysis and Information Processing*, 2015, vol. 3, pp. 112–127. DOI: <https://doi.org/10.4236/jdaip.2015.34012>
- Hearst M. A.** Text tiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 1997, vol. 23, no. 1, pp. 33–64.
- Itoh N., Kurths J.** Change-point detection of climate time series by nonparametric method. *Proceedings of the World Congress on Engineering and Computer Science*, 2010, vol. 1, pp. 445–448.
- Mikolov T., Sutskever I., Chen K., Corrado G. S., Dean J.** Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 2013, vol. 26, pp. 1–9.
- Nanni G., Glavaš F., Ponzetto S. P.** Unsupervised text segmentation using semantic relatedness graphs. In: *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics (*SEM)*. Berlin, Germany, 2016, pp. 125–130.
- Németh G., Zainkó C.** Multilingual Statistical Text Analysis, Zipf's Law and Hungarian Speech Generation. *Acta Linguistica Hungarica*, 2002, vol. 49, no. 3–4, pp. 385–405.
- Pevzner L., Hearst M. A.** A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 2002, vol. 28, no. 1, pp. 19–36. DOI: <https://doi.org/10.1162/089120102317341756>

Информация об авторах

Ковалевский Артем Павлович, доктор физико-математических наук, доцент, ведущий научный сотрудник Института математики им. С. Л. Соболева Сибирского отделения РАН

Павлова Юлия Викторовна, студентка Новосибирского государственного университета

Information about the Authors

Artyom P. Kovalevskii, DSc., Associate Professor, Leading Researcher at the Sobolev Institute of Mathematics

Yulia V. Pavlova, Student of Novosibirsk State University

Статья поступила в редакцию 28.06.2025;

одобрена после рецензирования 02.10.2025; принята к публикации 04.07.2025

The article was submitted 07.10.2025;

approved after reviewing 02.10.2025; accepted for publication 07.10.2025