

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

DOI: <https://doi.org/10.60797/IRJ.2026.172.28> EDN: IEYDYC

РАСХОЖДЕНИЕ ПОКАЗАТЕЛЕЙ ЯЗЫКОВОЙ СЛОЖНОСТИ В ТЕКСТАХ ПЕСЕН: КОЛИЧЕСТВЕННЫЙ АНАЛИЗ АНГЛОЯЗЫЧНЫХ ХИТОВ 1955–2025 ГОДОВ

Научная статья

Кагиров Р.Д.¹, Гибадуллин Р.Ф.^{2*}, Мухамедьярова Л.Р.³²ORCID : 0000-0001-9359-911X;¹ Казанский национальный исследовательский технологический университет, Казань, Российская Федерация² Казанский национальный исследовательский технический университет им. А.Н. Туполева – КАИ, Казань, Российская Федерация³ Гимназия №18 с татарским языком обучения – «Галэм», Казань, Российская Федерация

* Корреспондирующий автор (landwatersun[at]mail.ru)

Предложена: 27.08.2026; Принята: 01.10.2026; Опубликовано: 16.10.2026

Аннотация

Статья посвящена количественному анализу изменений формальных лингвистических характеристик текстов англоязычных популярных песен за 1955–2025 годы. Материалом послужил корпус из 1 695 текстов из годовых перечней популярных композиций; все показатели рассчитывались по полным текстам с равным весом каждой песни в годовом среднем. Оценивались структурные индексы читаемости (уровень Флеша–Кинкейда, индекс Колмана–Лиау), доля повторяющихся предложений, показатели лексического разнообразия (скорректированный TTR и Yule's K), частеречный профиль по 34 лексическим тегам Penn Treebank и доля форм первого лица единственного числа. Тренды оценивались наклоном Тейла–Сена и OLS-регрессией со стандартными ошибками Ньюи–Уэста, р-значения корректировались по процедуре Бенджамини–Хохберга; устойчивость выводов дополнительно проверялась на наборе уникальных текстов. Установлено расхождение измерений сложности: структурные индексы и средняя длина предложения снижаются, доля дословных повторов растёт, тогда как лексическое разнообразие по обоим индексам увеличивается. Среди частеречных сдвигов устойчивы рост личных местоимений, сравнительных прилагательных, глагольных форм настоящего времени и частиц при снижении притяжательных местоимений, экзистенциального there, глаголов прошедшего времени, маркера to, детерминативов, сочинительных союзов и модальных глаголов; разведочный спектральный анализ устойчивой периодичности не выявил. Сделан вывод, что обобщение о линейном «упрощении языка песен» слишком грубо: лирика стала структурно компактнее и более повторяющейся, но не менее разнообразной на уровне словоформ.

Ключевые слова: тексты песен, языковая сложность, индексы читаемости, лексическое разнообразие, повторяемость, частеречная разметка, корпусная лингвистика, диахронический анализ, англоязычная популярная музыка.

VARIATIONS IN LANGUAGE COMPLEXITY METRICS IN SONG LYRICS: A QUANTITATIVE ANALYSIS OF ENGLISH-LANGUAGE HITS FROM 1955 TO 2025

Research article

Kagirov R.D.¹, Gibadullin R.F.^{2*}, Mukhamedyarova L.R.³²ORCID : 0000-0001-9359-911X;¹ Kazan National Research Technological University, Kazan, Russian Federation² Kazan National Research Technical University named after A.N. Tupolev – KAI, Kazan, Russian Federation³ Gymnasium No. 18 with Tatar as the Language of Instruction – "Galem", Kazan, Russian Federation

* Corresponding author (landwatersun[at]mail.ru)

Suggested: 27.08.2026; Accepted: 01.10.2026; Published: 16.10.2026

Abstract

The article is devoted to a quantitative analysis of changes in the formal linguistic traits of the lyrics of English-language popular songs from 1955 to 2025. The source material consisted of a corpus of 1,695 lyrics from annual charts of popular songs; all metrics were calculated using the full lyrics, with each song given equal weight in the annual average. The following were assessed: structural readability indices (the Flesch–Kincaid Grade Level, the Coleman–Liau index), the proportion of repeated sentences, measures of lexical diversity (adjusted TTR and Yule's K), a frequency profile based on 34 Penn Treebank lexical tags, and the proportion of first-person singular forms. Trends were evaluated using the Theil–Sen estimator and OLS regression with Newey–West standard errors; p-values were adjusted using the Benjamini–Hochberg procedure; the reliability of the findings was further tested on a set of unique texts. A divergence in measures of complexity was observed: structural indices and average sentence length are decreasing, the proportion of verbatim repetitions is increasing, whilst lexical diversity is increasing according to both indices. Among the shifts in part-of-speech distribution, there is a consistent increase in personal pronouns, comparative adjectives, present-tense verb forms and particles, alongside a decrease in possessive pronouns, the existential 'there', past-tense verbs, the marker 'to', determiners, coordinating conjunctions and modal verbs;

exploratory spectral analysis did not reveal any stable periodicity. It has been concluded that the generalisation regarding a linear ‘simplification of song lyrics’ is too broad: the lyrics have become structurally more concise and repetitive, but no less varied in terms of word forms.

Keywords: song lyrics, linguistic complexity, readability indices, lexical diversity, repetition, part-of-speech tagging, corpus linguistics, diachronic analysis, English-language popular music.

Введение

Тексты популярных песен является удобным материалом для наблюдения за тем, как культурные продукты меняются на протяжении десятилетий: они массовы, точно датированы и сопоставимы между собой. Varnum с соавторами проанализировали 14 661 композицию из Billboard Hot 100 за 1958–2016 годы, приняв сжимаемость текста за меру лирической простоты, и обнаружили устойчивое движение в сторону более простой лирики [1]. Parada-Cabaleiro с соавторами, охватив пять жанров англоязычной популярной музыки, пришли к сходному выводу: лексическая и структурная сложность снижается, а повторяемость растёт [2]. Схожая динамика прослеживается и в аудиохарактеристиках записей [3], [4], а на уровне содержания ранее отмечались рост самореференции [5] и усиление негативной эмоциональной окраски [6]. Вместе с тем корпуса, правила сегментации, набор метрик и принципы отбора песен в этих работах заметно расходятся. Поэтому новые результаты разумнее читать как самостоятельное описание конкретной выборки, а не как прямую репликацию предшествующих исследований.

Цель настоящей работы — воспроизводимо описать, как менялись формальные лингвистические характеристики песенных текстов в собранном корпусе за 1955–2025 годы. Анализ строится на четырёх принципах. Во-первых, все показатели рассчитаны непосредственно по полным текстам, а не по предварительно агрегированным данным. Во-вторых, каждое наблюдение снабжено устойчивым идентификатором, складывающимся из года и имени файла. В-третьих, наряду с характеристиками длины учитываются лексическое разнообразие и местоименная перспектива. В-четвёртых, каждая оценка тренда сопровождается мерой неопределённости, поправкой на множественность сравнений и проверкой устойчивости к точным дубликатам текстов.

Методы и принципы исследования

2.1. Корпус и контроль качества

Материалом послужили 1 699 текстов песен из годовых перечней популярных англоязычных композиций за 1955–2025 годы. Каждый текст хранился отдельным файлом в каталоге соответствующего года; итоговый состав выборки приведён в таблице 1.

Точные текстовые совпадения не удалялись автоматически: один и тот же текст может относиться к разным чартовым позициям или к разным версиям песни. Такие группы выявлялись и учитывались на этапе контроля качества. Чтобы проверить устойчивость выводов, дополнительно был сформирован набор из 1 598 уникальных текстов: для каждого SHA-256-хэша в нём оставлена самая ранняя запись.

Таблица 1 - Состав аналитической выборки и контроль качества

DOI: <https://doi.org/10.60797/IRJ.2026.172.28.1>

Показатель	Значение	Решение
Файлов в исходном корпусе	1 699	Все проверены
Исключено	4	Тексты короче 20 токенов или без единиц, сопоставимых с предложением
Включено в основной анализ	1 695	Равный вес каждой песни в годовом среднем
Годы	1955–2025	Представлены все годы периода
Групп точных текстовых дубликатов	97	Сохранены; проведена проверка устойчивости
Набор уникальных текстов	1 598	Одна самая ранняя запись на текст

2.2. Лингвистические показатели

Текст сегментировался по терминальным знакам (. ! ?) и по исходным строкам после нормализации. Предложение считалось повторным, если нормализованная последовательность его токенов уже встречалась в той же песне; доля повторов вычислялась как $100 \times (\text{число предложений} - \text{число уникальных нормализованных предложений}) / \text{число предложений}$. Это узкая мера дословного повтора строки: она не описывает ни повторяемость припева, ни рифму, ни мелодию, ни смысловую близость фрагментов.

Во избежание терминологической путаницы оговорим: основной мерой читаемости здесь служит уровень Флеша–Кинкейда (Flesch–Kincaid Grade Level), а не шкала Flesch Reading Ease, которая рассчитывалась лишь дополнительно. Уровень Флеша–Кинкейда вычислялся по формуле $0,39 \times (\text{слов/предложений}) + 11,8 \times (\text{слогов/слов}) - 15,59$ [7], [8], индекс Колмана–Лиау — по формуле $0,0588L - 0,296S - 15,8$, где L — число букв на 100 слов, S — число предложений на 100 слов [9]. Слоговой состав определялся по CMU Pronouncing Dictionary [10]; для форм, в словаре

отсутствующих, применялось простое эвристическое правило подсчёта групп гласных. В качестве дополнительных показателей рассчитывались ARI и Gunning Fog, а также средняя длина предложения в словах и средняя длина слова в слогах.

Лексическое разнообразие описывалось двумя взаимодополняющими индексами: скорректированным $TTR = \frac{V}{\sqrt{2N}}$, где V — число типов, а N — число токенов [11], и Yule's $K = 10000 \cdot \frac{(M_2 - N)}{N^2}$, где M_2 — сумма квадратов частот типов [12]. Рост скорректированного TTR и снижение Yule's K одинаково указывают на более разнообразное словоупотребление. Оба индекса зависят от длины текста и не заменяют друг друга, поэтому используются здесь как парные описательные характеристики, а средний годовой объём лирики приводится отдельно [13], [14]. Частеречная разметка выполнена англоязычным averaged perceptron tagger из пакета NLTK [15]; доли категорий рассчитывались от числа словесных токенов. Местоименная перспектива оценивалась по доле форм первого лица единственного числа (I, me, my, mine, myself) среди словесных токенов.

Индексы читаемости опираются главным образом на длину слов и предложений и ничего не говорят о том, насколько слушателю знакома лексика, насколько связан текст и как он воспринимается на слух. Поэтому далее мы называем их структурными текстовыми индексами, а не прямыми показателями понятности песни [16].

2.3. Агрегация и статистический анализ

Для каждого года рассчитывалось равновзвешенное среднее по песням, чтобы длинные тексты не получали преимущества перед короткими. На рисунках приведены годовые средние и 7-летние центрированные скользящие средние; последние служат исключительно описательной визуализации. Для каждой из 13 основных метрик оценивались наклон Тейла–Сена [17] и линейный наклон OLS со стандартными ошибками Ньюи–Уэста (НАС, максимальный лаг 3 года) [18] — последнее снижает риск занижить ошибки при краткосрочной автокорреляции годовых рядов. Р-значения скорректированы по процедуре Бенджамини–Хохберга внутри семейства метрик [19]; в тексте и таблицах далее приводятся именно скорректированные значения. Все проверки носят разведочный характер: статистический план анализа заранее не фиксировался. Связи между годовыми средними оценивались коэффициентом Спирмена с той же FDR-коррекцией и причинной интерпретации не допускают.

Отдельные частеречные теги рассматривались как самостоятельное разведочное семейство из 34 лексических тегов Penn Treebank [20]. Для каждого тега вычислялись годовые средние, наклон Тейла–Сена и OLS-оценка с НАС-стандартными ошибками, а FDR-коррекция Бенджамини–Хохберга применялась в пределах этих 34 проверок. На каждой временной панели дополнительно нанесено 7-летнее центрированное среднее: оно доступно с 1958 по 2022 год, описывает локальную форму ряда и самостоятельной проверкой гипотезы не является. Пунктирная линия OLS также остаётся визуальным ориентиром — статистическая аннотация панели опирается на НАС и FDR. Чтобы частые категории не заслоняли редкие, каждая малая панель снабжена собственной шкалой y . Диаграммы размаха сопоставляют распределения на уровне отдельных песен в первом и последнем десятилетиях; это описательное сравнение, новых проверок оно не добавляет.

Дополнительный отбор тегов проводился уже после просмотра всех результатов и опирался на два независимых разведочных правила. По первому из ещё не представленных тегов отбирались те, для которых в основной годовой панели $FDR \leq 0,05$, направление наклона Тейла–Сена совпадает с полученным на уникальных текстах, а средняя распространённость составляет не менее 0,10%; отобранные теги ранжировались по абсолютной величине стандартизированного наклона. По второму среди остальных тегов с той же минимальной распространённостью выбирались ряды с наибольшим размахом 7-летней центрированной средней, а затем — с наибольшей суммарной вариацией сглаженной кривой. Второе правило представляет собой описательный апостериорный отбор и новых статистически подтверждённых выводов не даёт.

Годовые ряды дополнительно проверялись на периодичность разведочным спектральным анализом (FFT). Перед преобразованием из рядов удалялся линейный тренд; рассматривались только периоды от 2,0 до 23,7 года, то есть укладывающиеся в 71-летний ряд не менее трёх раз, а максимальные пики сопоставлялись с 5 000 AR(1)-суррогатами с последующей FDR-коррекцией. Ни один пик коррекцию не пережил, поэтому результаты FFT не вошли в основные и не рассматриваются как свидетельство устойчивой периодичности.

Основные результаты

3.1. Структурная сложность, повторяемость и лексическое разнообразие

В основной выборке оба структурных индекса снижаются (табл. 2, рис. 1). Средний уровень Флеша–Кинкейда упал с 2,73 в 1955–1964 годах до 2,30 в 2016–2025 годах: наклон Тейла–Сена $-0,082$ класса за десятилетие (95% ДИ $[-0,125; -0,033]$), $p = 0,016$. Индекс Колмана–Лиану снизился заметнее — с 3,36 до 2,23, наклон $-0,154$ класса за десятилетие, $p < 0,001$. За первой динамикой стоит прежде всего сокращение средней длины предложения с 9,29 до 7,62 слова, тогда как средняя слоговая длина слова достоверного тренда не показала ($p = 0,636$).

Доля повторяющихся предложений выросла с 27,28% до 30,24%, что отвечает наклону Тейла–Сена $+0,688$ процентного пункта за десятилетие ($p = 0,029$). При этом данные о словаре не подтверждают привычный тезис об «обеднении» лексики: скорректированный TTR увеличился с 3,76 до 4,56, а Yule's K снизился с 293,55 до 220,05 (рис. 2). Оба сдвига указывают в одну сторону — на более разнообразное словоупотребление, хотя оговорка о зависимости этих индексов от длины текста сохраняет силу. Иными словами, структурная компактность и рост дословных повторов уживаются здесь с растущим лексическим разнообразием.

Таблица 2 - Годовые тренды основных метрик

DOI: <https://doi.org/10.60797/IRJ.2026.172.28.2>

Показатель	1955–1964 → 2016–2025	Наклон/10 лет [95% ДИ]	p (FDR)
Флеш–Кинкейд, классы	2,73 → 2,30	–0,082 [–0,125; –0,033]	0,016
Колман–Лиау, классы	3,36 → 2,23	–0,154 [–0,211; –0,102]	< 0,001
Повторяющиеся предложения, %	27,28 → 30,24	+0,688 [+0,128; +1,216]	0,029
Скорректированный TTR	3,76 → 4,56	+0,125 [+0,090; +0,162]	< 0,001
Yule's K	293,55 → 220,05	–14,570 [–20,425; –9,384]	< 0,001
Слов в предложении	9,29 → 7,62	–0,190 [–0,279; –0,110]	< 0,001
Первое лицо ед. ч. (I, me, my...), %	7,04 → 8,16	+0,228 [+0,116; +0,369]	0,002

Примечание: N = 1 695 песен; 71 год; наклон — оценка Тейла–Сена за десятилетие с 95% ДИ; p (FDR) — коррекция Бенджамини–Хохберга среди 13 метрик

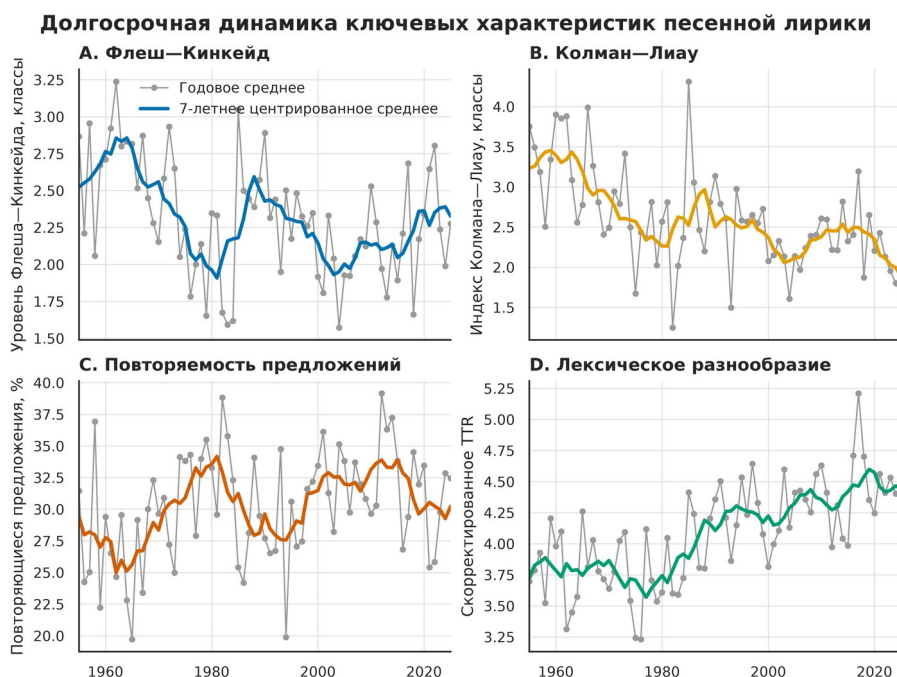


Рисунок 1 - Структурные индексы, повторяемость предложений и скорректированный TTR: годовые средние и 7-летние центрированные средние
DOI: <https://doi.org/10.60797/IRJ.2026.172.28.3>

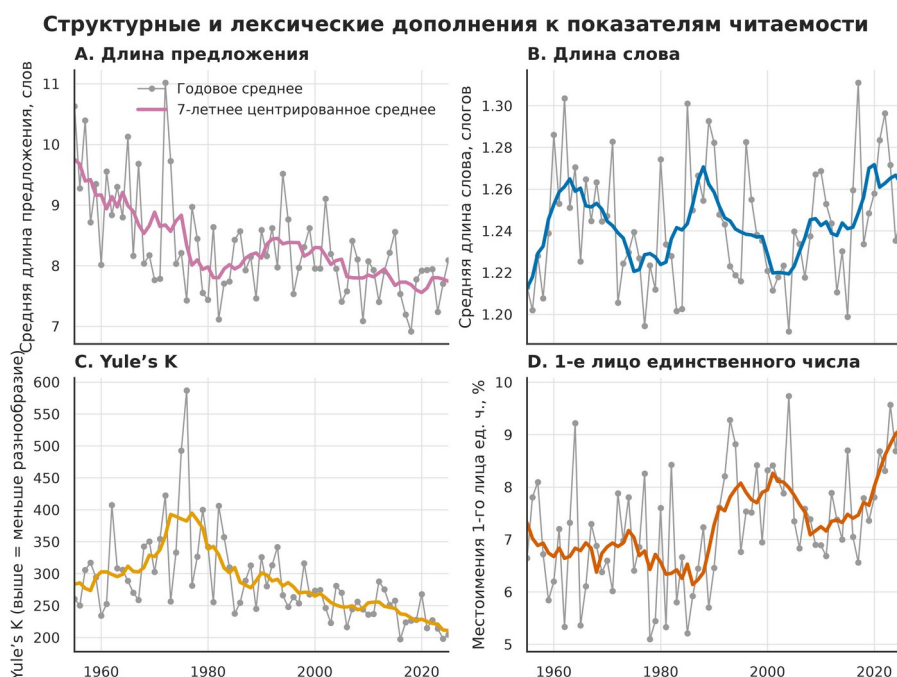


Рисунок 2 - Длина предложения и слова, Yule's K и доля первого лица
DOI: <https://doi.org/10.60797/IRJ.2026.172.28.4>

Примечание: более высокий Yule's K означает меньшее разнообразие

3.2. Индивидуальные POS-теги и местоименная перспектива

Разбор отдельных тегов делает частеречный профиль нагляднее, чем сводный график с общей шкалой. После НАС-оценки и FDR-коррекции направление тренда сохранилось при проверке на уникальных текстах у 13 из 34 лексических тегов. Девять наиболее содержательных представлены в таблице 3 и на рисунке 3, ещё три разбираются в разделе 3.3; один редкий тег со средней распространённостью ниже 0,10% отдельно не рассматривается. Доля личных местоимений (PRP) выросла с 7,12% до 8,93% (наклон Тейла—Сена +0,301 п.п. за десятилетие, $p = 0,004$), тогда как притяжательные местоимения (PRP\$) сократились с 3,69% до 2,42% ($p < 0,001$). Растёт доля сравнительных прилагательных (JJR), снижаются экзистенциальное there (EX) и глаголы прошедшего времени (VBD) — эти

направления также устойчивы к FDR-коррекции. Одновременно увеличивается доля VBP, то есть глагольных форм настоящего времени не третьего лица единственного числа; без ручной проверки разметки и учёта контекста этот результат требует осторожного толкования.

Теги JJS и UN вынесены на отдельный описательный рисунок 4: их наклоны выглядят содержательно, но после коррекции по 34 тегам значимости не сохраняют ($p = 0,813$ в обоих случаях). В основной перечень подтверждённых направлений они не входят, и трактовать их как самостоятельные долговременные изменения нельзя.

Таблица 3 - POS-теги с FDR-устойчивым трендом, подтверждённым на уникальных текстах

DOI: <https://doi.org/10.60797/IRJ.2026.172.28.5>

Тег Penn Treebank	1955–1964 → 2016–2025, %	Наклон/10 лет, п.п.	p (FDR)
Экзистенциальное there (EX)	0,12 → 0,07	–0,008	0,017
Сравнительные прилагательные (JJR)	0,26 → 0,41	+0,022	0,017
Личные местоимения (PRP)	7,12 → 8,93	+0,301	0,004
Притяжательные местоимения (PRP\$)	3,69 → 2,42	–0,139	< 0,001
Глаголы, прошедшее время (VBD)	3,27 → 2,69	–0,068	0,041
Глаголы наст. времени, не 3-е л. ед. ч. (VBP)	5,95 → 7,79	+0,296	< 0,001
Маркер to (TO)	2,13 → 1,70	–0,105	< 0,001
Частицы (RP)	0,56 → 1,06	+0,079	0,013
Детерминативы (DT)	7,58 → 6,76	–0,187	0,002

Примечание: наклон — оценка Тейла–Сена, п.п. за десятилетие; p (FDR) — HAC + коррекция Бенджамини–Хохберга среди 34 тегов

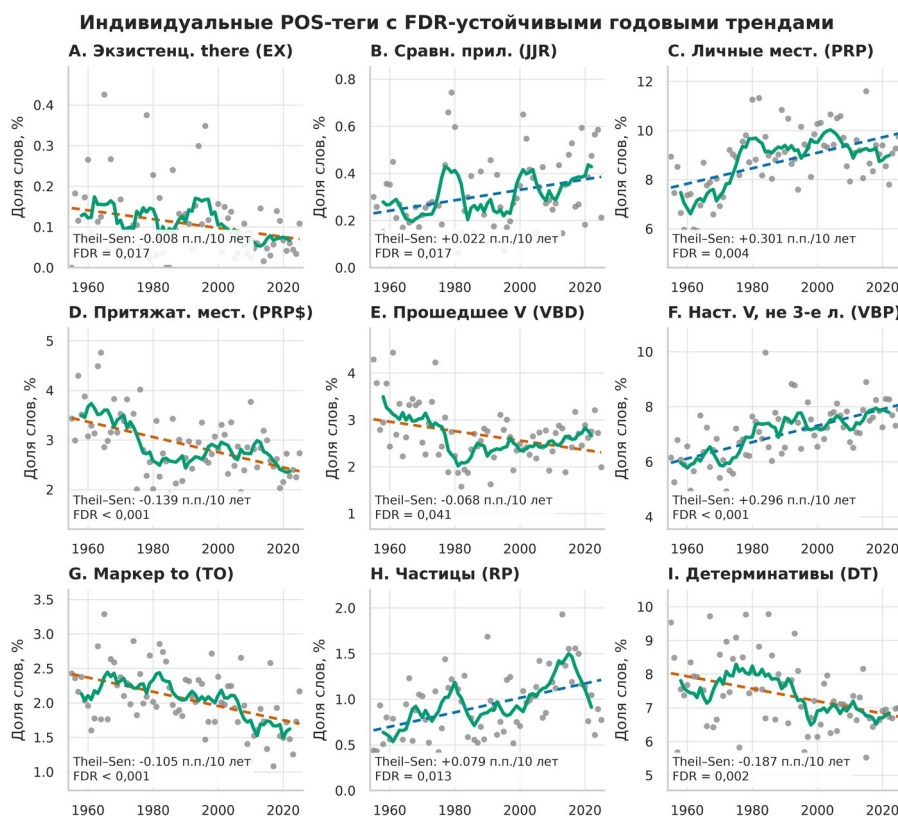


Рисунок 3 - Девять FDR-устойчивых POS-тегов:

точки – годовые средние; сплошная линия – 7-летние центрированные средние; пунктир – OLS-тренд

DOI: <https://doi.org/10.60797/IRJ.2026.172.28.6>

Примечание: шкалы панелей независимы

Направления 7-летней средней, линии OLS и оценки Тейла–Сена на рисунке 3 совпадают и выдерживают FDR-коррекцию, поэтому все девять панелей можно описывать как устойчивые многолетние сдвиги — но не как монотонные процессы. EX и TO в целом снижаются: у EX заметны ранние локальные колебания, у TO спад усиливается ближе к концу периода. JJR растёт, но с выраженными локальными пиками, так что речь идёт о повышении среднего уровня, а не о равномерном приросте год за годом.

Личные местоимения PRP заметно поднимаются к концу 1970-х – 1980-м годам и далее держатся выше исходного уровня; притяжательные PRP\$, напротив, после 1970-х снижаются, задержавшись на промежуточном плато. Местоименный профиль, следовательно, перестраивается неравномерно и не сводится к росту всех местоименных форм сразу. VBD падает к концу 1970-х – 1980-м, затем частично восстанавливается, но прежнего уровня не достигает, тогда как VBP особенно прибавляет после 1980-х. RP растёт с поздним локальным максимумом, DT в целом снижается, временами отыгрывая назад. Всё сказанное относится к частотам тегов и само по себе не обосновывает ни тематических, ни психологических, ни причинных объяснений.

Запрошенные индивидуальные POS-теги: отдельные шкалы и линейные тренды

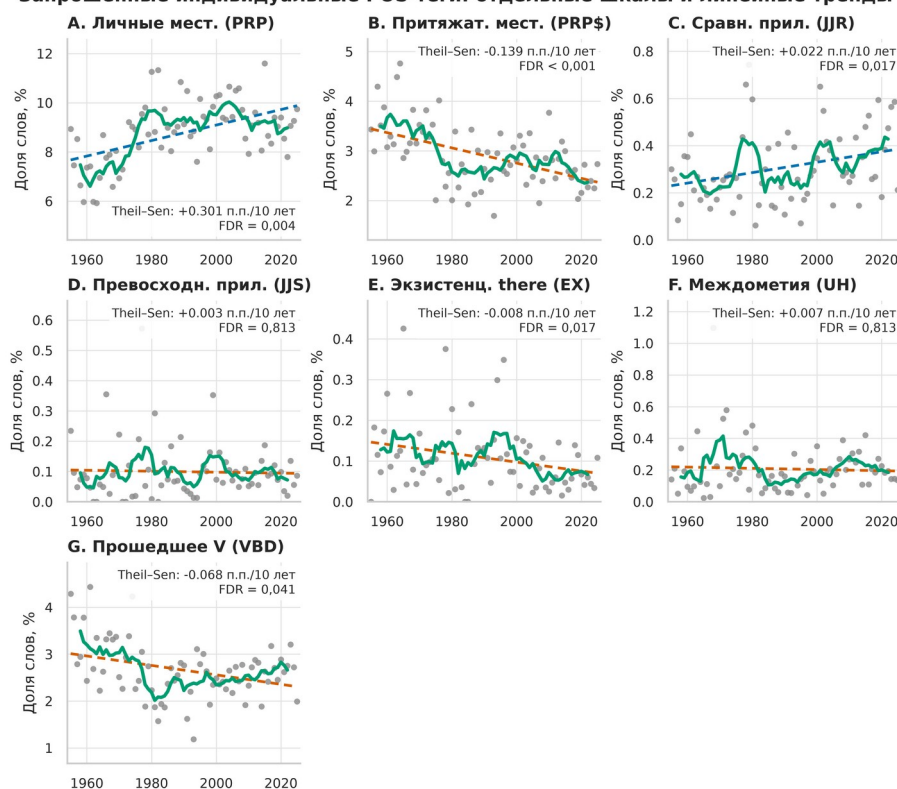


Рисунок 4 - Теги PRP, PRP\$, JJR, JJS, EX, UH и VBD на отдельных шкалах:
 точки – годовые средние; сплошная линия – 7-летние центрированные средние; пунктир – OLS-тренд
 DOI: <https://doi.org/10.60797/IRJ.2026.172.28.7>

Примечание: JJS и UH приведены описательно и не являются FDR-устойчивыми

Для PRP, PRP\$, JJR, EX и VBD 7-летние средние на рисунке 4 воспроизводят те же общие направления, что и на рисунке 3, но яснее показывают локальные отклонения от линии регрессии. Сглаженная кривая JJS складывается из кратковременных подъёмов и спадов без устойчивого однонаправленного движения, у UH виден ранний локальный подъём и последующие колебания. Поскольку обе категории не выдерживают FDR-коррекции, форма их рядов не может служить доказательством долгосрочного изменения. Распределения песенных значений девяти устойчивых тегов в первом и последнем десятилетиях сопоставлены на рисунке 5.

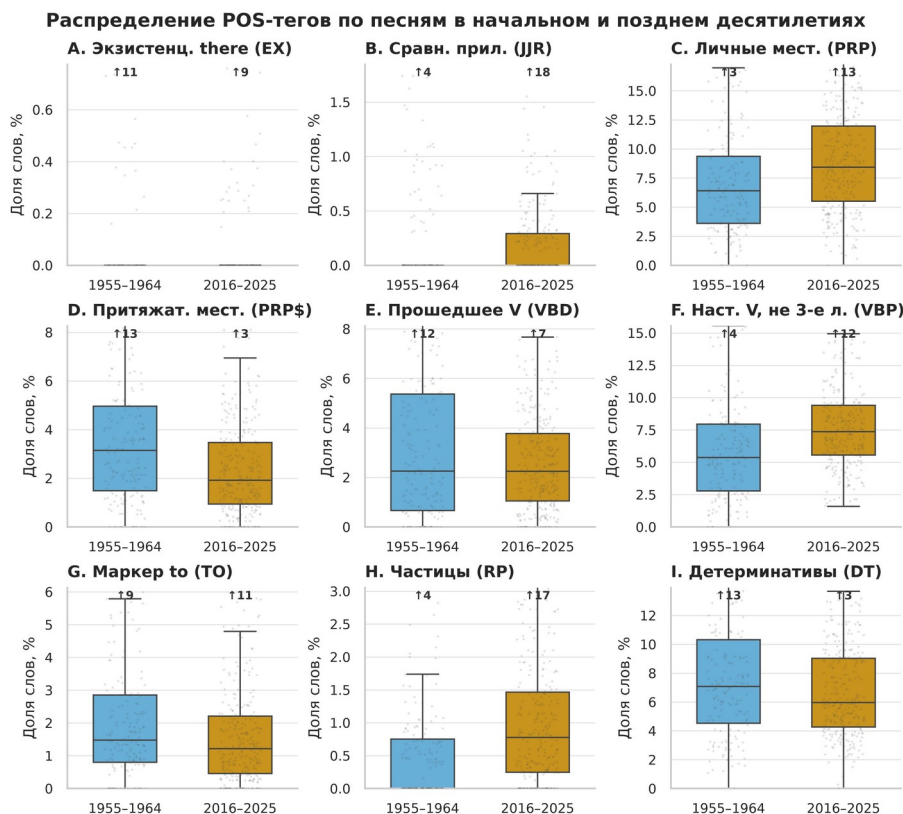


Рисунок 5 - Распределения девяти FDR-устойчивых тегов по песням в 1955–1964 и 2016–2025 годах:

↑n — число значений выше предела отображения (учтены, но не отрисованы)

DOI: <https://doi.org/10.60797/IRJ.2026.172.28.8>

Примечание: медиана, межквартильный размах и точки до 95-го перцентиля; сравнение описательное, шкалы независимы

3.3. Дополнительный отбор тегов: выраженные наклоны и вариативные средние

Из тегов, не попавших на предыдущие рисунки, критерию устойчивого тренда отвечают ещё три (табл. 4, рис. 6): сочинительные союзы CC (−0,077 п.п. за десятилетие, $p = 0,045$), модальные глаголы MD (−0,050 п.п., $p = 0,045$) и сравнительные наречия RBR (−0,007 п.п., $p = 0,050$, то есть ровно на границе порога). У всех трёх направление наклона Тейла–Сена сохраняется и на уникальных текстах. Их отрицательные наклоны дополняют описанную выше картину, однако RBR встречается настолько редко, что требует особой осторожности.

Теги NN, JJ и VB отобраны по другому основанию — высокой изменчивости 7-летней средней: размах сглаженной траектории составил 5,34, 2,54 и 2,47 п.п. соответственно. В этих рядах чередуются локальные подъёмы, спады и частичные восстановления, которые единственная линия OLS неизбежно упрощает. Однако значения p после FDR-коррекции (0,906, 0,813 и 0,813) общего долгосрочного тренда не подтверждают, и приводятся эти ряды как описательные примеры формы сглаженных кривых, а не как новые результаты.

Таблица 4 - Дополнительный отбор тегов: устойчивые к FDR-коррекции наклоны (CC, MD, RBR) и описательная изменчивость (NN, JJ, VB)

DOI: <https://doi.org/10.60797/IRJ.2026.172.28.9>

Тег Penn Treebank	Основание отбора	1955–1964 → 2016–2025, %	Наклон/10 лет, п.п.	p (FDR)	Размах 7-летней средней, п.п.
Сочинительные союзы (CC)	Устойчивый наклон	2,88 → 2,35	−0,077	0,045	1,18
Модальные глаголы (MD)	Устойчивый наклон	1,37 → 0,98	−0,050	0,045	0,73
Сравнительные наречия (RBR)	Устойчивый наклон	0,14 → 0,07	−0,007	0,050	0,15
Нарицательные сущ., ед. ч. (NN)	Высокая изменчивость	23,52 → 24,45	+0,037	0,906	5,34
Прилагательны	Высокая	9,18 → 8,78	+0,015	0,813	2,54

Ter Penn Treebank	Основание отбора	1955–1964 → 2016–2025, %	Наклон/10 лет, п.п.	p (FDR)	Размах 7-летней средней, п.п.
е (JJ)	изменчивость				
Глаголы, базовая форма (VB)	Высокая изменчивость	7,01 → 6,52	–0,011	0,813	2,47

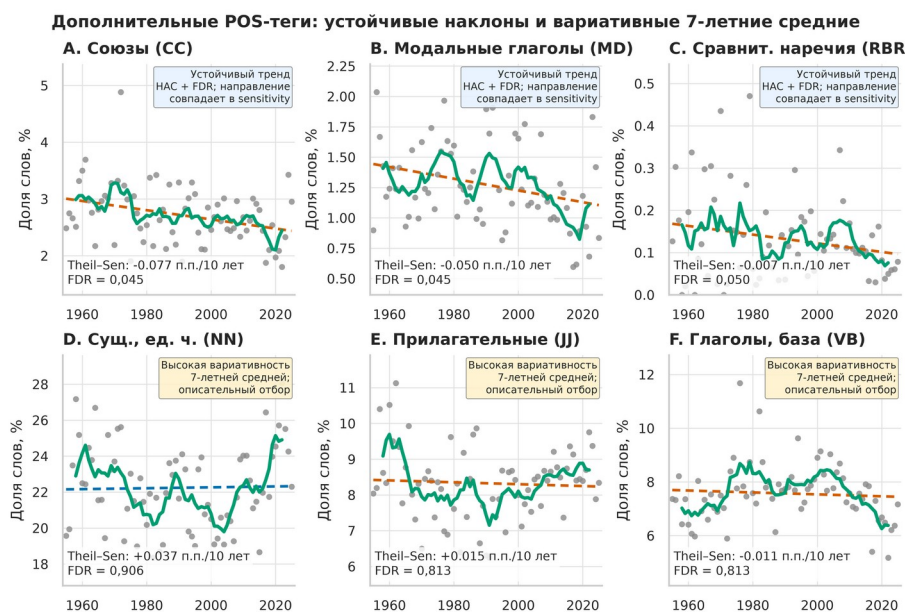


Рисунок 6 - Дополнительные теги: FDR-устойчивые CC, MD и RBR (панели А–С) и описательные NN, JJ и VB (панели D–F):

точки – годовые средние; сплошная линия – 7-летние центрированные средние; пунктир – OLS-тренд

DOI: <https://doi.org/10.60797/IRJ.2026.172.28.10>

На рисунке 6 у CC, MD и RBR 7-летние средние снижаются, но форма спада неровная: у CC и MD видны многолетние колебания, у RBR после ранних пиков идёт долгое сползание вниз. Кривые NN, JJ и VB прежде всего немонотонны: позднее повышение NN контрастирует с провалом в середине периода, JJ проходит через раннее снижение и последующее восстановление, VB даёт несколько локальных максимумов и поздний спад. Так как NN, JJ и VB отобраны по изменчивости уже после просмотра данных и FDR-коррекции не выдерживают, доказательством направленного языкового изменения эти наблюдения служить не могут.

3.4. Связи с повторяемостью и проверка устойчивости

На уровне годовых средних (табл. 5) повторяемость отрицательно связана с уровнем Флеша–Кинкейда ($\rho = -0,757$), индексом Колмана–Лиану ($\rho = -0,620$) и средней длиной предложения ($\rho = -0,476$). Связь со скорректированным TTR слабее ($\rho = -0,268$), но выдерживает FDR-коррекцию. Все эти ряды разделяют общую временную структуру, поэтому обнаруженные соответствия не доказывают, что изменение одного признака вызывает изменение другого.

Таблица 5 - Корреляции годовой средней повторяемости с основными метриками

DOI: <https://doi.org/10.60797/IRJ.2026.172.28.11>

Признак	ρ с повторяемостью	p (FDR)
Уровень Флеша–Кинкейда, классы	–0,757	< 0,001
Индекс Колмана–Лиану, классы	–0,620	< 0,001
Средняя длина предложения, слов	–0,476	< 0,001
Скорректированный TTR	–0,268	0,041
Все местоимения, %	0,305	0,020
Глаголы, %	0,358	0,005

Примечание: ρ — коэффициент Спирмена; p (FDR) — коррекция Бенджамини–Хохберга среди 12 ассоциаций

Проверка на наборе уникальных текстов, где на каждый текст оставлена самая ранняя запись, подтвердила направление 12 наклонов из 13. Единственное расхождение пришлось на долю существительных, но устойчивого тренда для неё не было ни в одном из вариантов анализа. Ключевые выводы — снижение обоих структурных индексов, рост повторяемости, рост скорректированного TTR, снижение Yule's K, сокращение длины предложения и увеличение доли первого лица — на уникальных текстах воспроизводятся. Наполненность выборки по годам и средний годовой объём лирики показаны на рисунке 7.

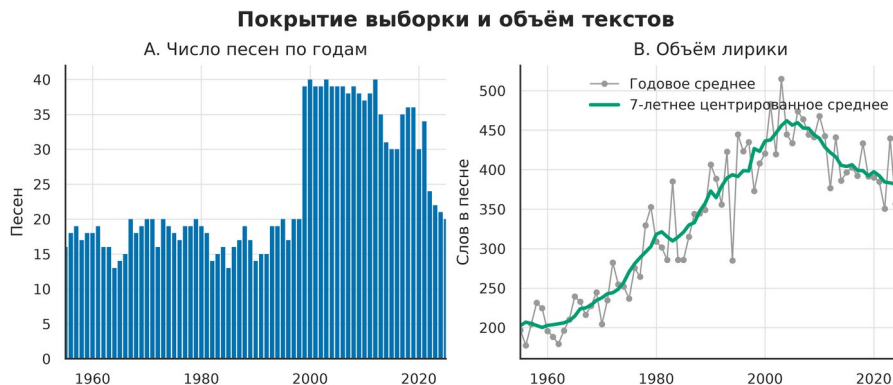


Рисунок 7 - Число песен и средний годовой объём текста — ограничения для сопоставления периодов

DOI: <https://doi.org/10.60797/IRJ.2026.172.28.12>

Обсуждение

Полученные результаты уточняют расхожий тезис об упрощении песенной лирики. Тенденция к более коротким предложениям, более низким значениям структурных индексов и большей доле дословных повторов данными подтверждается и по направлению согласуется с предшествующими работами, опиравшимися на другие корпуса и другие операционализации простоты [1], [2]. Но стоит добавить лексические метрики, и картина меняется: одновременного сокращения лексического разнообразия в этой выборке нет — скорректированный TTR растёт, а Yule's K падает. Точнее поэтому говорить о расхождении измерений: структурная компактность и повторяемость смещались в сторону простоты, тогда как разнообразие словоупотребления по обоим выбранным индексам увеличивалось.

Раздельная визуализация тегов объясняет, почему сводного графика с общей шкалой недостаточно: редкие, но систематически меняющиеся категории различимы только на собственных панелях. А 7-летнее центрированное среднее обнаруживает локальные подъёмы, плато и частичные восстановления, которые прямая OLS скрывает. К перечню устойчивых результатов — росту PRP, JJR, VBP и RP и снижению PRP\$, EX, VBD, TO и DT — дополнительный отбор прибавляет снижение CC, MD и редкого RBR, причём направление сохраняется и при проверке на уникальных текстах. Ряды NN, JJ и VB, напротив, показывают лишь изменчивость сглаженных траекторий: они отобраны апостериорно, FDR-коррекции не выдерживают и основной перечень трендов подкреплять не могут. Линейные регрессии, центрированные средние и диаграммы размаха ценны здесь как прозрачные описательные инструменты, а не как доказательство причинных связей.

Разведочный спектральный анализ не обнаружил устойчивой периодичности: после удаления линейного тренда, сопоставления с AR(1)-суррогатами и FDR-коррекции ни один пик не сохранился. Объяснения, апеллирующие к циклам, этот 71-летний ряд наблюдений не поддерживает, и в основные выводы они не вошли.

Наконец, различия между периодами могут отражать не только сам язык песен, но и меняющийся состав чартов, практики публикации текстов, доступность источников, пропуски пунктуации и жанровую структуру выборки.

Ограничения

Во-первых, корпус составлен из конкретных годовых перечней популярных композиций и не представляет генеральную совокупность англоязычной музыки; число наблюдений по годам к тому же заметно колеблется (рис. 7). Во-вторых, происхождение текстов, их полнота и редакторские особенности ограничивают прямое сопоставление с корпусами других исследований. В-третьих, сегментация предложений и подсчёт слогов особенно уязвимы к песенной орфографии, стяжениям, повторам строк и произвольной пунктуации, а это затрагивает все показатели, опирающиеся на границы предложения. В-четвёртых, автоматическая частеречная разметка не заменяет ручной лингвистической аннотации и ошибается на неполных фразах, именах собственных, диалектных формах и нестандартной грамматике. В-пятых, читаемость, вычисленная по формуле, — не то же самое, что понимание или слуховое восприятие [16]. Наконец, автокорреляция и множественность проверок учтены лишь отчасти, поэтому приведённые *p*-значения следует читать как разведочные, а не как подтверждение причинных гипотез.

Заключение

На очищенном и заново пересчитанном корпусе из 1 695 песенных текстов за 1955–2025 годы структурные характеристики меняются согласованно: оба индекса, опирающиеся на длину предложений и слов, снижаются, а доля



повторяющихся предложений растёт. Показатели лексического разнообразия дают противоположную картину: в поздней части выборки скорректированный TTR выше, а Yule's K ниже. Анализ отдельных частеречных тегов эту картину уточняет — рост личных местоимений и форм VBP соседствует со снижением притяжательных местоимений, экзистенциального there, маркера to, VBD, а также сочинительных союзов CC и модальных глаголов MD; снижаются и редкие сравнительные наречия RBR, хотя их низкая частотность требует особой осторожности в выводах. Семилетние центрированные средние показывают, что эти сдвиги идут через локальные колебания, а не по ровной годовой траектории, а JJS и UH устойчивого долгосрочного тренда не обнаруживают. Обобщение о линейном «упрощении языка песен» оказывается, таким образом, слишком грубым. Точнее сказать, что в этой чартовой выборке лирика стала структурно компактнее и более повторяющейся, но не менее разнообразной на уровне словоформ. Дальнейшая работа должна сопоставить жанры, страны и рынки, источники текстов и альтернативные алгоритмы сегментации; особенно полезны здесь открытая публикация списка песен и версий текстов, а также заранее зафиксированный статистический план.

Конфликт интересов

Не указан.

Рецензия

Сообщество рецензентов Международного научно-исследовательского журнала

DOI: <https://doi.org/10.60797/IRJ.2026.172.28.13>

Conflict of Interest

None declared.

Review

Community of Reviewers of the International Research Journal

DOI: <https://doi.org/10.60797/IRJ.2026.172.28.13>

Список литературы на английском языке / References in English

1. Varnum M.E.W. Why are song lyrics becoming simpler? A time series analysis of lyrical complexity in six decades of American popular music / M.E.W. Varnum, J.A. Krems, C. Morris [et al.] // PLOS ONE. — 2021. — № 16 (1). — Art. e0244576. — DOI: [10.1371/journal.pone.0244576](https://doi.org/10.1371/journal.pone.0244576).
2. Parada-Cabaleiro E. Song lyrics have become simpler and more repetitive over the last five decades / E. Parada-Cabaleiro, M. Mayerl, S. Brandl [et al.] // Scientific Reports. — 2024. — № 14. — Art. 5531. — DOI: [10.1038/s41598-024-55742-x](https://doi.org/10.1038/s41598-024-55742-x).
3. Serrà J. Measuring the evolution of contemporary Western popular music / J. Serrà, Á. Corral, M. Boguñá [et al.] // Scientific Reports. — 2012. — № 2. — Art. 521. — DOI: [10.1038/srep00521](https://doi.org/10.1038/srep00521).
4. Mauch M. The evolution of popular music: USA 1960–2010 / M. Mauch, R.M. MacCallum, M. Levy [et al.] // Royal Society Open Science. — 2015. — № 2 (5). — Art. 150081. — DOI: [10.1098/rsos.150081](https://doi.org/10.1098/rsos.150081).
5. DeWall C.N. Tuning in to psychological change: Linguistic markers of psychological traits and emotions over time in popular U.S. song lyrics / C.N. DeWall, R.S. Pond Jr., W.K. Campbell [et al.] // Psychology of Aesthetics, Creativity, and the Arts. — 2011. — № 5 (3). — P. 200–207. — DOI: [10.1037/a0023195](https://doi.org/10.1037/a0023195).
6. Brand C.O. Cultural evolution of emotional expression in 50 years of song lyrics / C.O. Brand, A. Acerbi, A. Mesoudi // Evolutionary Human Sciences. — 2019. — № 1. — Art. e11. — DOI: [10.1017/ehs.2019.11](https://doi.org/10.1017/ehs.2019.11).
7. Flesch R. A new readability yardstick / R. Flesch // Journal of Applied Psychology. — 1948. — № 32 (3). — P. 221–233. — DOI: [10.1037/h0057532](https://doi.org/10.1037/h0057532).
8. Kincaid J.P. Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel (Research Branch Report 8-75) / J.P. Kincaid, R.P. Fishburne Jr., R.L. Rogers [et al.]. — Naval Technical Training Command, 1975.
9. Coleman M. A computer readability formula designed for machine scoring / M. Coleman, T.L. Liao // Journal of Applied Psychology. — 1975. — № 60 (2). — P. 283–284. — DOI: [10.1037/h0076540](https://doi.org/10.1037/h0076540).
10. The CMU Pronouncing Dictionary / Carnegie Mellon University. — <http://www.speech.cs.cmu.edu/cgi-bin/cmudict> (accessed: 16.06.2026).
11. Carroll J.B. Language and thought / J.B. Carroll. — Prentice-Hall, 1964.
12. Yule G.U. The statistical study of literary vocabulary / G.U. Yule. — Cambridge University Press, 1944.
13. McCarthy P.M. MTL, D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment / P.M. McCarthy, S. Jarvis // Behavior Research Methods. — 2010. — № 42 (2). — P. 381–392. — DOI: [10.3758/BRM.42.2.381](https://doi.org/10.3758/BRM.42.2.381).
14. Tweedie F.J. How variable may a constant be? Measures of lexical richness in perspective / F.J. Tweedie, R.H. Baayen // Computers and the Humanities. — 1998. — № 32 (5). — P. 323–352. — DOI: [10.1023/A:1001749303137](https://doi.org/10.1023/A:1001749303137).
15. Loper E. NLTK: The Natural Language Toolkit [Preprint] / E. Loper, S. Bird // arXiv. — 2002. — DOI: [10.48550/arXiv.cs/0205028](https://doi.org/10.48550/arXiv.cs/0205028).
16. Tip 6: Use caution with readability formulas for quality reports // Agency for Healthcare Research and Quality. — 2015. — URL: <https://www.ahrq.gov/talkingquality/resources/writing/tip6.html> (accessed: 16.06.2026).
17. Sen P.K. Estimates of the regression coefficient based on Kendall's tau / P.K. Sen // Journal of the American Statistical Association. — 1968. — № 63 (324). — P. 1379–1389. — DOI: [10.1080/01621459.1968.10480934](https://doi.org/10.1080/01621459.1968.10480934).
18. Newey W.K. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix / W.K. Newey, K.D. West // Econometrica. — 1987. — № 55 (3). — P. 703–708. — DOI: [10.2307/1913610](https://doi.org/10.2307/1913610).
19. Benjamini Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing / Y. Benjamini, Y. Hochberg // Journal of the Royal Statistical Society: Series B (Methodological). — 1995. — № 57 (1). — P. 289–300. — DOI: [10.1111/j.2517-6161.1995.tb02031.x](https://doi.org/10.1111/j.2517-6161.1995.tb02031.x).



20. Marcus M.P. Building a large annotated corpus of English: The Penn Treebank / M.P. Marcus, B. Santorini, M.A. Marcinkiewicz // Computational Linguistics. — 1993. — № 19 (2). — P. 313–330. — URL: <https://aclanthology.org/J93-2004/> (accessed: 16.06.2026).