ISSN 2707-1898 (print)
ISSN 2786-9172 (online)

Correspondence author
O. S. Kushnir
oleh.kushnir@lnu.edu.ua

Український журнал інформаційних технологій
Ukrainian Journal of Information Technology
http://science.lpnu.ua/uk/ujit
https://doi.org/10.23939/ujit2025.02.082
Article received 10.10.2025 p.
Article accepted 30.10.2025 p.
UDC 004.6, 004.9, 303.732, 519.1

О. С. Кушнір, В. О. Футей

Львівський національний університет імені Івана Франка, м. Львів, Україна

ДОСЛІДЖЕННЯ ПОВТОРЮВАНОСТІ ДЛЯ НАЙПРОСТІШИХ РАНДОМНИХ МОДЕЛЕЙ ПРИРОДНОЇ МОВИ

У статті вирішується актуальна проблема опрацювання природної мови – розроблення методів оцінювання повторюваності в текстових документах і емпіричне з'ясування ресурсів цих методів для аналізу наявності семантичного навантаження текстів. Досі для цього переважно залучали підходи, основані на закономірностях статистичної лінгвістики на зразок законів Ціпфа, Парето та Гіпса, а також аналіз явищ кластеризації слововживань і довгосяжних кореляцій лексики. Ми розробили програмне забезпечення для кількісного дослідження повторюваності в текстах за алгоритмом дерев суфіксів Укконена. Запропоновано аналізувати усереднений параметр повторюваності v_0 та відповідне стандартне відхилення Δv. Емпірично доведено, що реалізований за допомогою нашого програмного забезпечення алгоритм виявляє наближено лінійну часову складність O ($L\log L$) (де L – довжина символьного ряду), яка відповідає теоретичним передбаченням. На предмет повторюваності проаналізовано природні тексти англійською мовою та низкою інших алфавітних (неієрогліфічних) мов, а також рандомні тексти мавпи Міллера та природні тексти, дозовано рандомізовані на лінгвістичних рівнях символів, слів і речень. Підтверджено, що поза початковою ділянкою природних текстів, на якій функція повторюваності флуктуює, ця функція виявляє явище насичення, де насичене значення $v_0 \approx 0{,}5$. Докладно досліджено тексти, що ґрунтуються на рандомній моделі Міллера. З'ясовано, що на поведінку функції повторюваності й на параметри v_0 і Δv цих текстів впливають насамперед розміри алфавіту та розподіл заданих відносних частот символів алфавіту. Доведено, що усереднена повторюваність v_0 цих текстів пов'язана із інформаційною ентропією Шеннона в її найпростішому поданні. Встановлено, що повторюваність у моделі "мішка зі словами" корелює із параметром усередненого семантичного навантаження тексту. Запропоновано використовувати параметри повторюваності v_0 і Δv для розпізнавання семантично наповнених природних текстів і семантично порожніх стохастичних символьних часових рядів.

Ключові слова: опрацювання природної мови, моделі природної мови, аналіз семантики, опрацювання текстових даних, символьні часові ряди, повторюваність.

Вступ / Introduction

В останні десятиліття спостерігаємо стрімкий розвиток галузей комп'ютерної лінгвістики та опрацювання природної мови, які привели, зокрема, до значного прогресу в розумінні структури та законів функціонування природної мови та появи великих мовних моделей. Однією з відкритих проблем в опрацюванні природної мови є автоматичний аналіз текстів, пов'язаний із розпізнаванням їхньої семантики. Такий аналіз повинен надійно відрізняти семантично наповнені, інформативні природні тексти від рандомних послідовностей символів, які фактично є стохастичними часовими рядами без жодної семантики. Зазначимо, що універсальність алгоритмів вирішення таких завдань означає, що ми не повинні покладатися на апріорні знання "мови" досліджуваного тексту, а користуватися лише деякими статистично надійними властивостями взаємного розміщення лінгвістичних символів у текстах.

Для аналізу будови та характеристик природних текстів дослідники розробили низку простих моделей цих текстів, незмінно основаних на рандомній появі символів або слів у них. Зокрема, це відома рандомна модель "мішка зі словами" ("bag of words"), якій відповідають випадкові перестановки лінгвістичних елементів у тексті (символів, слів, речень тощо), а також моделі мавпи Міллера, Саймона (або модель "багатий стає багатшим" – "rich gets richer"), урни Поля, ланцюжків Маркова тощо (див., наприклад, [1], [2], [3], [4], [5]). Для вирішення проблеми розрізнення семантично наповнених і семантично порожніх послідовностей лінгвістичних елементів дослідники неодноразово залучали такі властивості текстів, як лінгвістичні закони Ціпфа, Парето та Гіпса, що наближено описують кількісні закономірності розподілу слів або букв за частотами [1], [3], [6], [7], динамічні явища на зразок кластеризації позицій окремих слів у тексті або довгосяжних автокореляцій цих позицій [6], характеристики лінгвістичних мереж [4] тощо.

Однією із перспективних можливостей є вивчення кількісних закономірностей повторюваності лінгвістичних елементів у текстах [8], [9], [10], [11], [12], [13], [14], [15]. Інтуїтивно зрозуміло, що природний текст міститиме більше повторень, аніж рандомний текст, оскільки, окрім випадкових повторень унаслідок законів комбінаторики, в ньому будуть також повторення, які відповідатимуть типовим комбінаціям літер у словах або широковживаним лексичним колокаціям.

У цій роботі ми продовжуємо розвиток напряму, започаткованого працями [8], [11], [13], [14], і хочемо з'ясувати можливості кількісних характеристик повторюваності для відмежування інформативних, семантично наповнених текстів від суто "шумових" стохастичних символьних послідовностей. Особливого значення набуває дослідження базових рандомних моделей природних текстів як таких, що принципово допускають аналітичні інтерпретацію та розрахунки повторюваності, а також відповідні порівняння з емпіричними даними для повторюваності.

Об'єкт дослідження – явища повторюваності в природних текстах і їхніх базових рандомних моделях.

Предмет дослідження – реалізація алгоритму визначення повторюваності та порівняльний аналіз основних кількісних характеристик повторюваності для природних і рандомних текстів.

Мета дослідження – розроблення технології визначення повторюваності в текстах і вивчення практичних ресурсів цього підходу, насамперед порівняння параметрів повторюваності природних текстів і їхніх рандомних моделей із метою оцінювання семантичного навантаження текстів.

Для досягнення зазначеної мети виконано такі *завдання дослідження*:

- практична реалізація кількісного аналізу повторюваності та дослідження його швидкодії;
- вивчення впливу розмірів алфавіту, відносних частот лінгвістичних символів у тексті та інших факторів на характеристики повторюваності в рандомних текстах мавпи Міллера та рандомізованих текстах;
- з'ясування взаємозв'язків повторюваності із системою письма, семантичним навантаженням текстів та інформаційною ентропією Шеннона;
- порівняння поведінки характеристики повторень у природних і рандомних текстах і встановлення кількісних критеріїв їх розрізнення.

Аналіз останніх досліджень і публікацій. Математичну характеристику повторюваності вперше ввів Ф. Гольхер у праці [8]. Згідно з [8], повторюваність *v(t)* як функцію поточної позиції *t* символу в тексті та кількості *V* внутрішніх вузлів дерева суфіксів означають як

$$v(t) = V(T_t)/t, \qquad (1)$$

де $t = 1, 2, \ldots, L$ ($L = t_{\max}$ – кількість усіх лінгвістичних символів у тексті або довжина цього тексту); T_t – це неповне суфіксне дерево, побудоване на субтексті S_t повної послідовності символів (тобто всього тексту) S, довжина якої $|S| = L$. Інакше кажучи, *v(t)* – кількість усіх завершених символьних *n*-грам ($n = 1, 2,\ldots,$), що повторилися в тексті принаймні один раз аж до позиції *t*, поділена на номер цієї позиції. Зауважимо, що в межах такого алгоритму розрахунку всі наступні повторення після першого не враховуються, тобто фактично маємо справу зі "словником" повторень. Термін "завершеність повторення" деякої *n*-грами означає, що в тексті відсутні ($n + 1$)-грами або ще довші повторювані послідовності, які продовжують цю *n*-граму.

У праці [8] зроблено висновок, що для природних текстів вже за помірно коротких їх довжин функція $v(t)$ "насичується" (тобто майже перестає залежати від дискретного часу *t*) на величині $v_0 \sim 1/2$. З іншого боку, точне значення насиченої величини v_0 дещо залежить від мови, а для ієрогліфічних мов на зразок китайської явище насичення функції *v(t)* сумнівне. Водночас у текстах комп'ютерних кодів спостерігали немонотонні зміни *v(t)* за великих *t*, а для єдиного проаналізованого прикладу тексту мавпи Міллера із "алфавітом" $M = 100$ букв – значні коливання *v(t)* за цих самих умов [8]. Хоча коло вивчених у [8] текстів і мов було доволі обмеженим, ці дані наводять на думку про деякі ресурси характеристики *v(t)* для відрізняння текстів природними мовами від рандомних символьних рядів.

Повторюваність $v(t)$ надалі вивчали для набагато довших текстів (фактично корпусів текстів) кількома мовами [9], [11] і дійшли висновку про деякі залишкові флуктуації $v(t)$ для цих текстів. На додаток зроблено припущення про можливий зв'язок параметра v_0 із надлишковістю мови та системою письма [9]. У цій праці ми продовжуємо та узагальнюємо перші спроби вивчення поведінки повторюваності для кількох типів рандомних послідовностей символів [11], [13], [14]. Актуальність цього напряму пов'язана із багатьма факторами: недостатнім вивченням базових алгоритмів розрахунку, зокрема їхньої швидкодії, потребою у масштабному дослідженні та порівнянні даних для різних рандомних моделей природної мови і природних текстів, а також з'ясуванні ресурсів характеристики повторюваності для розрізняння семантично наповнених і семантично порожніх текстів.

Результати дослідження та їх обговорення / Research results and their discussion

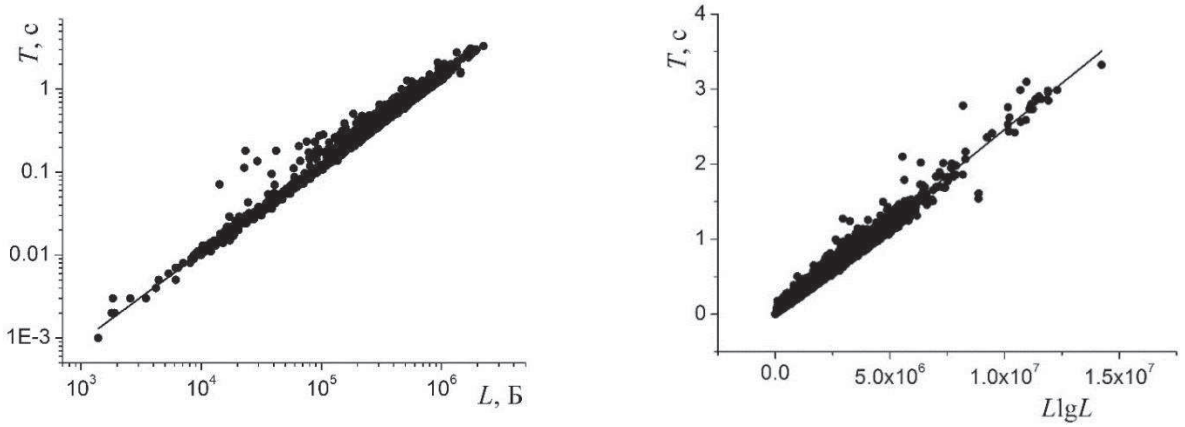
Швидкодія алгоритму розрахунку повторюваності в символьних послідовностях. Проілюструємо схему розрахунку залежності *v(t)* на короткому "тексті", що складається із єдиного слова, – "національний". За означенням, у такому разі зараховують завершені повторення символьних *n*-грам НА, А на позиції 8 і повторення Н на позиції 11. Такий основний режим алгоритму назвемо "символи". В альтернативному режимі "слова" текст вважають часовою послідовністю слів, а не символів (лексичним часовим рядом). Наприклад, у тексті "дощик, дощик, кап-кап-кап" у такому режимі маємо одне завершене повторення "дощик" на позиції 2.

Оскільки залежність *v(t)* для багатьох текстів може досягати рівноважного значення, то функцію *v(t)* зручно характеризувати двома параметрами – її середнім значенням v_0 і стандартним відхиленням Δv на деякій прикінцевій ділянці графіка *v(t)*. Тобто значення v_0 і Δv обчислюємо за допомогою статистичного опрацювання всіх значень *v(t)*, починаючи з деякого часу $t = t_0$ (типово $t_0 = (0{,}8 \div 0{,}9)\, L$).

Технічні труднощі розрахунку повторюваності за наведеним вище алгоритмом очевидні: для встановлення факту повторення потрібно пам'ятати всі можливі символьні (або лексичні) *n*-грами довільної довжини *n*, що трапилися в тексті, аж до деякої поточної позиції *t* у ньому. Проте, за законами комбінаторики, кількість цих *n*-грам дуже швидко зростає (наближено за формулою t^a, де *a* – деяке число, помітно більше за одиницю). Інакше кажучи, обчислення повторюваності для достатньо довгих символьних рядів потребує дуже великих ресурсів. Одним із виходів є застосування відомого алгоритму суфіксних дерев Укконена (див. [16]).

Через потребу масштабних досліджень характеристики повторюваності для багатьох порівняно довгих текстів важливу роль відіграє швидкодія відповідного програмного забезпечення. Науковці стверджують, що час роботи *T* алгоритму Укконена наближено лінійно залежить від довжини *L* досліджуваного часового ряду символів ($t_{\max} = L$) для незмінного розміру алфавіту символів, що означають як лінійну часову складність алгоритму *O*(*L*). Ми експериментально перевіряли різні практичні аспекти реалізації алгоритму Укконена, а також часову складність задачі вивчення функцій *v(t)* у базовому режимі повторюваності символів. Це було виконано для природних текстів англійською мовою з різними довжинами *L*. Корпус текстів для досліджень містив 2000 текстів, із загальним обсягом файлів 700 МБ. У кодуванні UTF-8 це приблизно відповідає 700 млн символів. Розрахунки виконано на комп'ютері із такими характеристиками: Intel Core i7-9750H @ 2.60 GHz / 16 GB RAM / 512 GB SSD.

Виявилося, що повний час *T* опрацювання текстів і одержання даних повторюваності *v(t)* справді наближено лінійно залежить від розміру тексту. Маємо також деякі флуктуації часу *T* – насамперед через труднощі контрольованого уникнення запуску додаткових фонових системних процесів. Нахил залежності *T*(*L*) залежить від потужності процесора комп'ютера. Додатковий аналіз засвідчив (див. рис. 1, *а*), що час роботи алгоритму описується степеневою функцією $T \sim L^{1{,}06}$, що дещо повільніше за *O*(*L*). Водночас, точніший підхід (див. рис. 1, *б*) підтвердив відому теоретичну залежність *O*(*L*log*L*) (див., наприклад, [16]) для нашої реалізації алгоритму. Тут точність лінійної апроксимації (коефіцієнт кореляції *r* за Пірсоном) перевищує $r \geq 0{,}99$.



а *б*

Рис. 1. Залежність часу розрахунку *T* повторюваності від довжини *L* тексту в байтах для корпусу 2000 англомовних природних текстів. Пряма лінія позначає лінійну апроксимацію у подвійному логарифмічному масштабі та відповідає степеневій функції $T \sim L^{1{,}06}$(а); та сама залежність у координатах $T = f$(LlgL) (б) / Dependence of the calculation time *T* of the repetitiveness on the length *L* of text in bytes for a corpus of 2000 English-language natural texts. The straight line denotes a linear fitting on a double logarithmic scale and corresponds to the power-law function $T \sim L^{1.06}$(a); The same dependence in the coordinates $T = f$ (LlgL) (b)

Максимальний розмір файлу, з яким може упоратися алгоритм дерев суфіксів, залежить від виділеної оперативної пам'яті: кожен її 1 ГБ дає змогу опрацювати ~ 2,0–2,5 МБ тексту в кодуванні UTF-8. Отже, типові сучасні комп'ютери дають змогу працювати з дуже великими текстами, проте не великими корпусами. Тому для розрахунків *v(t)* на корпусах доцільно розглядати родину альтернативних алгоритмів масивів суфіксів (suffix array). Застосування цього алгоритму буде одним із предметів аналізу в наших наступних працях.

Результати для рандомних текстів мавпи Міллера. У цьому підрозділі зосередимося суто на основному режимі обчислення повторюваності на лінгвістичному рівні символів. Для аналізу проблеми розрізнення природних і рандомних текстів на основі повторюваності передусім з'ясуємо поведінку функції *v(t)* для найпростішої рандомної моделі – текстів мавпи

Міллера. Їхня класична версія передбачає рандомну послідовність букв або, загальніше, символів, із деяким заданим розміром алфавіту M, до якого ми – з міркувань зручності застосування підходу до питань повторюваності – зарахуємо й пробіли. Відносні частоти всіх символів f_A, f_B,… задають однаковими ($\sum_i f_i = 1$), а в узагальненій версії моделі ці частоти можуть бути й різними [13], [14]. Переважно достатньо розглянути найпростішу лінійну частотно-рангову залежність цих частот, де єдиним параметром є "градієнтний параметр" $b = f_{max}/f_{min}$ (f_{max} і f_{min} – відповідно максимальна та мінімальна частоти символів) або логарифмічну залежність, яка грубо моделює співвідношення між частотами букв у природних текстах. Водночас, класичним текстам мавпи Міллера відповідає границя $f_{max} = f_{min}$ або $b = 1$.

На рис. 2 подано кілька прикладів результатів вивчення класичних текстів мавпи Міллера ($b = 1$) із розмірами алфавіту $M = 2 \div 14$ (разом з пробілом – порівн. із некоректними означеннями [13], [14]) і довжинами $L = 8$ млн символів. Для підвищення надійності даних $v(t)$ кожен розмір алфавіту представлено 5 випадковими реалізаціями текстів. Вивчено також кілька текстів із $M \approx 27 \div 34$, що відповідає алфавітам англійської, української та деяких інших європейських мов.

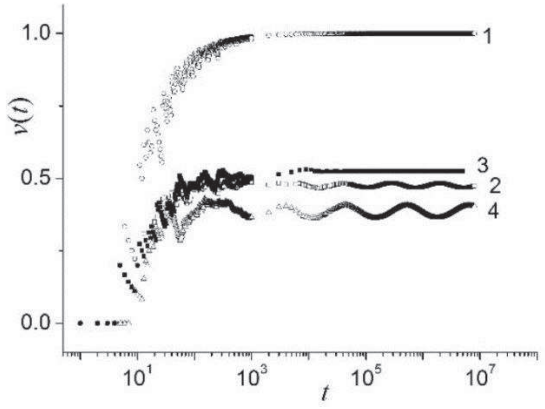


Рис. 2. Функція повторюваності, розрахована для текстів мавпи Міллера: крива 1 відповідає алфавітові $M = 2$ і параметрові $b = 1$, крива 2 – $M = 7$ і $b = 1$, крива 3 – $M = 7$ і $b = 50$, а крива 4 – $M = 12$ і $b = 1$ / The repetitiveness function calculated for the Miller's monkey texts: curve 1 corresponds to the alphabet $M = 2$ and the parameter $b = 1$, curve 2 to $M = 7$ and b = 1, curve 3 to $M = 7$ and $b = 50$, and curve 4 to $M = 12$ and $b = 1$

Зазначмо, що у разі $M = 1$ повторюваність $v(t)$ невизначена. З подальшим зростанням алфавіту рівноважне значення v_0 плавно зменшується від 1 за $M = 2$ (див. рис. 2 і рис. 3, *а*), а за $M = 27 \div 34$ становить орієнтовно ~ 0,3. Цю закономірність легко пояснити нижчою ймовірністю стохастичних повторень з-за більшого алфавіту. Зі зростанням M на графіках $v(t)$ з'являються та посилюються осциляції (див. рис. 2 і значення стандартного відхилення Δv на рис. 3, *б*). Якщо $M \geq 14$, вони домінують у функції $v(t)$, а досягнення рівноважного значення v_0 істотно уповільнюється, тобто параметр t_0 зростає (див. також [9]). Це потребує аналізу надзвичайно довгих текстів і помітно ускладнює високоточне визначення параметрів v_0 і Δv. З цих причин ми не вводили даних для $M \geq 14$ у подальший кількісний аналіз, а розрахунки v_0 і Δv для $M < 14$ виконували за допомогою усереднення з урахуванням фази згаданих осциляцій.

Із даних лінійної апроксимації в подвійному логарифмічному масштабі на рис. 3, *а, б* видно, що найближчим до емпіричних даних виявляється припущення про степеневу залежність параметрів v_0 і Δv від розмірів алфавіту M:

$$v_0(M) \propto M^{-B_1}, \Delta v(M) \propto M^{B_2}, \quad (2)$$

де $B_1 \approx 0.52$ (порівн. із набагато менш точними та докладними даними $B_1 \approx 0.37$ і $B_1 \approx 0.36$ [13], [14]), $B_2 \approx 0.37$, а коефіцієнти кореляції Пірсона дорівнюють відповідно $r \approx 0{,}99$ і 0,98.

Природно припустити, що параметр повторюваності v_0 залежить від розмірів алфавіту лінгвістичних одиниць, притаманних тій чи іншій системі письма, та інформаційної надлишковості "мови". Ці величини найпростіше виразити через односимвольну ($n = 1$) інформаційну ентропію Шеннона H, яка залежить від частотно-рангової залежності символів алфавіту (див. [17]). H описує суто формальний "потік текстової інформації" в одиницях біт / символ, який принципово можливий за умови такого розподілу ймовірностей p_i символів [17]:

$$H = -\sum_{i=1}^{M} p_i \log_2 p_i, \quad (3)$$

де ймовірності можна прирівняти до відносних частот ($p_i = f_i$). Дані рис. 3, *в* підтверджують припущення про зв'язок рівноважної повторюваності v_0 і ентропії H, який є степеневим: $v_0(H) \propto H^{-B_3}$, де $B_3 \approx 1.09$ і $r \approx 0{,}99$.

Ми вивчали також залежності параметрів повторюваності v_0 і Δv від параметра b (по 21 точці в діапазоні $b = 1 \div 10^5$) для трьох фіксованих розмірів алфавіту $M = 5$, 7, 12. Для цього було згенеровано 63 тексти за узагальненою моделлю мавпи Міллера із довжиною $L = 5$ млн символів. На рис. 3, *г*, *е* подано окремі приклади згаданих залежностей – відповідно дані $v_0(b)$ і Δv (b).

Із рис. 3, *г* видно, що зі збільшенням параметра b середнє значення повторюваності монотонно зростає, досягаючи деякого граничного значення $v_0^{(0)}$. Його існування зрозуміле інтуїтивно: достатньо великі параметри b означають дуже низькі частоти деяких букв алфавіту, тобто "ефективний" розмір M_{eff} останнього стає меншим ($M_{eff} < M$) і прямує до деякого фіксованого значення. Статистично це дає деяке зростання повторюваності, порівняно із випадком $b = 1$ (порівн. криві 2 і 3 на рис. 2 і див. пояснення, наведені вище).

Залежність $\Delta v(b)$ стандартного відхилення Δv функції $v(t)$ за найбільших часів t (див. рис. 3, *е*) фактично описує явище загасання осциляцій $v(t)$ зі зростанням параметра b. Хоча точність даних $\Delta v(b)$ виявляється набагато нижчою, вона інтуїтивно зрозуміла: зростання b

приводить до зменшення ефективного розміру алфавіту $M_{eff}(b)$, а функція $\Delta v(M_{eff})$ зростає (див. рис. 2 і рис. 3, *б*). Хоча через брак місця ми не відображали на рисунках відповідних даних $\Delta v(b)$ для всіх досліджених алфавітів M, зауважимо: що більший алфавіт текстів мавпи Міллера, то більші параметри b потрібні, аби компенсувати осциляції $v(t)$. Для прикладу, для алфавіту $M = 5$ осциляції стають майже невидимими вже за $b > 5$ (див. рис. 2, крива 3), а в разі алфавітів, розміри яких відповідають алфавітам європейських мов ($M \approx 27 \div 34$) для того, щоб осциляції $v(t)$ стали непомітними, потрібні вже параметри $b \sim 200 \div 300$.

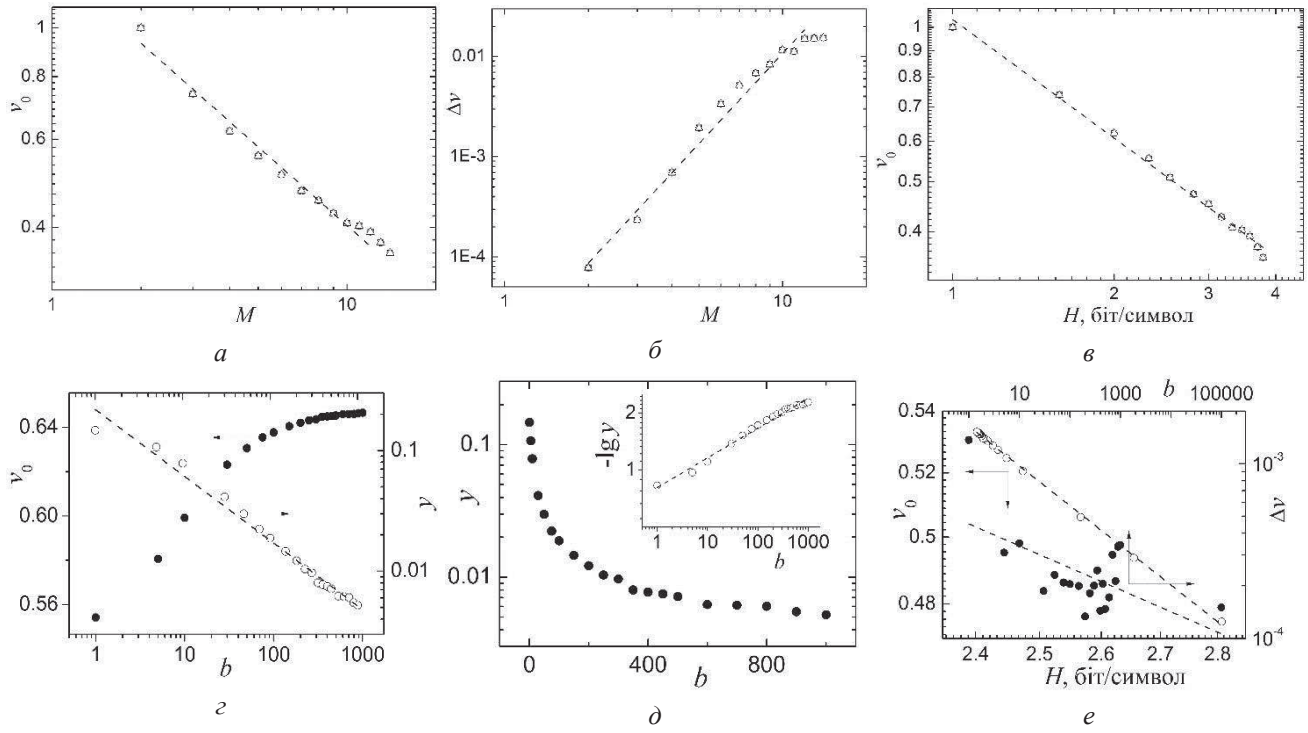


Рис. 3. Залежності параметрів повторюваності $v_0(M)$ (а) і $\Delta v(M)$ (б) від розмірів алфавіту M і залежність $v_0(H)$ від інформаційної ентропії H (в) для текстів мавпи Міллера із параметром алфавіту $b = 1$. Наведено дані лише для трьох статистичних реалізацій рандомних текстів. Залежності $v_0(b)$ (г – заповнені кружки), $\Delta v(b)$ (е – заповнені кружки), залежності допоміжного параметра $y = 1 - (v_0/v_0^{(0)})$ від b у різних масштабах (г, д) і залежність $v_0(H)$ від інформаційної ентропії H (е) для текстів мавпи Міллера за умов фіксованих M і змінного b: дані (г) і (д) відповідають $M = 5$, а дані (е) – M = 7. Штриховані лінії – лінійна апроксимація залежностей (див. пояснення у тексті) / Dependences of repetitiveness parameters $v_0(M)$ (a) and $\Delta v(M)$ (b) on the alphabet size M, and the dependence $v_0(H)$ on the information entropy H (c) for the Miller monkey texts with the alphabet parameter $b = 1$. Data are shown only for three statistical realizations of the random texts. Dependences v_0(b) (panel d; filled circles) and $\Delta v(b)$ (panel f; filled circles), dependences of the auxiliary parameter $y = 1 - (v_0/v_0^{(0)})$ on b plotted on different abscissa and ordinate scales (panels d and e), and dependence $v_0(H)$ on the information entropy H for the Miller's monkey texts with fixed M's and variable b: the data (d) and (e) corresponds to $M = 5$ and the data (f) to $M = 7$. Dashed lines correspond to linear fittings (see explanation in the text)

Нагадуємо, що практична відсутність осциляцій є однією із ознак функції $v(t)$ для природних текстів. До речі, саме значення $b \approx 300$ описує типовий перепад частот різних літер в алфавітних природних мовах, хоча частотно-рангові залежності тоді не лінійні, як у нас для простоти, а наближено логарифмічні. Тому з наведених вище результатів, а також не наведених через брак місця даних вивчення текстів мавпи Міллера з логарифмічними ранговими залежностями частот букв випливає, що точна форма функції рангової залежності частоти букв не відіграє особливої ролі в явищі загасання осциляцій для текстів мавпи Міллера. Природний наслідок цих висновків такий: на відміну від текстів мавпи Міллера, функція повторюваності $v(t)$ природних текстів не виявляє осциляцій саме через великий перепад частот різних літер алфавіту природних мов. Нарешті, залежність $\Delta v(b)$ на рис. 3, *е* можна грубо описати степеневою функцією $\Delta v(b) \propto b^{-B_4}$ із $B_4 \approx 0.126$ і кореляцією $r \approx 0{,}59$.

Залежність $v_0(b)$ на рис. 3, *г* можна спробувати описати степеневою, експоненційною або розтягнутою експоненційною функцією із насиченням:

$$v_0(b) = v_0^{(0)}[1 - b^{-c}],\ v_0(b) = v_0^{(0)}[1 - exp(-b/b_0)],$$
$$v_0(b) = v_0^{(0)}\left[1 - exp\left((-b/b_0)^{\beta}\right)\right], \qquad (4)$$

де c, b_0 і β – константи, а $v_0^{(0)} \approx 0{,}65$ – насичене значення. Зауважимо, що наведені вирази, насамперед другий вираз у формулах (4), формально відповідають своєрідним "перехідним процесам", добре відомим із електротехніки. Зазначмо також, що в нашому кількісному аналізі точки за найменших значень b менш надійні, оскільки повторюваність $v(t)$ тоді виявляє сильніші осциляції, що знижує точність визначення v_0 і Δv.

На рис. 3, *г*, д подано залежності допоміжного параметра $y = 1 - (v_0/v_0^{(0)})$ у різних масштабах, які ілюструють різні вирази в (4). Так, якість лінійної апроксимації y (b) у подвійному логарифмічному масштабі на рис. 3, *г* описує можливість застосування першого (степеневого) виразу в формулах (4): $c \approx 0{,}54$, $r \approx 0{,}992$. З іншого боку, правильність другого (експоненційного) виразу в (4) мала би означати лінійність залежності y (b) у напівлогарифмічному масштабі. Проте дані рис. 3, *д* засвідчують, що це не так: згадана залежність виражено нелінійна, з характерною увігнутістю, яка притаманна розтягнутій експоненті. Справді, високоякісна лінійна апроксимація залежності $-\lg y$ (b) у подвійному логарифмічному масштабі на вставці рис. 3, *д* підтверджує, що остання альтернатива з формул (4) теж із достатньою точністю ($r \approx 0{,}996$) описує характер залежності $v_0(b)$. Як висновок, однозначний вибір між першим і третім виразами в (4) на основі наших експериментальних даних утруднений.

Останній значущий експериментальний факт, що випливає з наших досліджень повторюваності для текстів мавпи Міллера, – це однозначна негативна кореляція рівноважної повторюваності v_0, знайденої за умови різних параметрів b, із односимвольною інформаційною ентропією H "системи письма" цих текстів ($r \geq 0{,}999$ – див. рис. 3, *е*). Як і дані рис. 3, *в*, цей факт, який поєднує повторюваність у текстах із особливостями графічної системи їхнього написання, підтверджує припущення роботи [9], проте спростовує висновки [13].

Результати для рандомізованих природних текстів. Наступною після текстів мавпи Міллера важливою теоретичною моделлю природних текстів є рандомізовані тексти. Рандомізація означає багатократні (кількість циклів $N >> 1$) випадкові перестановки лінгвістичних елементів (символів, слів, речень тощо) вихідного природного тексту. Зокрема, в разі перестановок слів маємо класичну модель "мішка зі словами". Зазначмо, що рандомізація не змінює розподілу відносних частот символів або слів, а лише їхню впорядкованість, знищуючи зміст (семантику) тексту.

Перш ніж перейти до аналізу повторюваності рандомізованих текстів порівняно з природними текстами, коротко пояснимо параметр кількісного оцінювання семантичного навантаження тексту (див. [6], [13]). Відомо, що семантично вбогі слова (т. зв. стоп-слова) "стохастично однорідно" розташовані в природному тексті, тоді як семантично багатші слова (найперше т. зв. ключові слова) виявляють явища "згущення" (кластеризації) або "розрідження" своїх появ у тексті [6]. Просторову неоднорідність розподілу кожного слова w легко оцінити за множиною $\{\tau_i\}$ т. зв. часів очікування τ_i усіх його слововживань, тобто відстаней $\tau_i = t_{i+1} - t_i$ між часовими позиціями t_i і t_{i+1} сусідніх слововживань цього слова w у тексті.

Розглянемо параметр $R_w = \Delta\tau/\langle\tau\rangle$ слова w, у якому $\langle\tau\rangle$ і $\Delta\tau$ – відповідно середнє значення та стандартне відхилення множини $\{\tau_i\}$ часів очікування цього слова. Увесь масив експериментальних даних засвідчує, що для більшості слів у природних текстах маємо $R_w \sim 1$, що відповідає випадкові стохастичного розподілу слововживань (фактично моделі рандомного тексту), і лише для меншої частки слів (передусім ключових) маємо $R_w > 1$ або $R_w >> 1$, тоді як ситуація $R_w < 1$ практично відсутня [6], [13]. Тепер введемо кумулятивний параметр тексту $R = \langle R_w \rangle$, усереднений за всіма словами w, статистика яких допускає розрахунки R_w. Легко переконатися, що для семантично наповнених природних текстів матимемо нерівність $R > 1$, тоді як $\Delta R \approx 1$ для всіх семантично порожніх рандомних текстів, у яких немає змісту, а тому й ключових слів. Отже, параметр R слугує мірилом семантичного навантаження тексту.

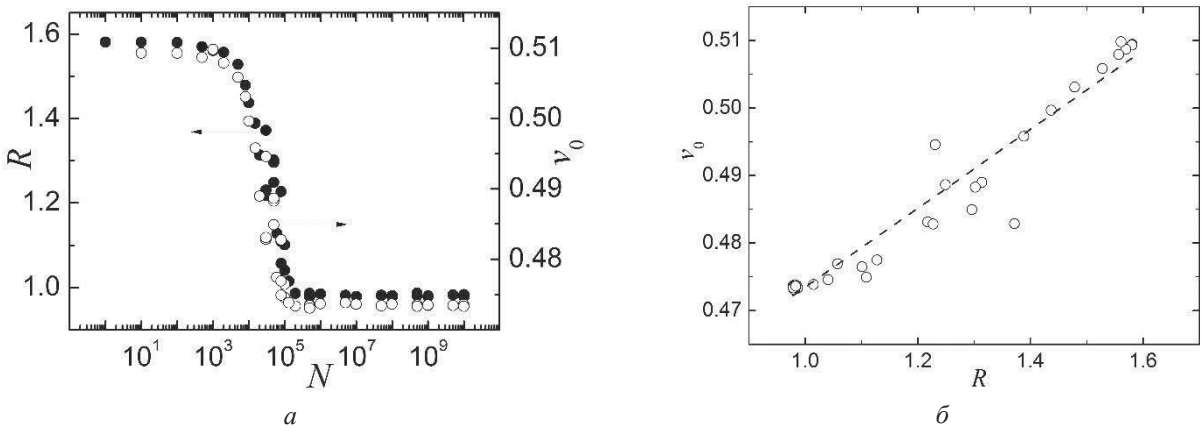
На рис. 4, *а* подано залежності рівноважного параметра повторюваності v_0 й параметра семантики R від кількості N елементарних циклів рандомізації тексту "The Lord of the Rings" за авторством J. R. R. Tolkien. Тут рандомізація відбувалася на лінгвістичному рівневі слів. Відповідно до наведених вище пояснень, нарощування кількості циклів рандомізації поступово знищує семантику, пов'язану із упорядкуванням слів у вихідному тексті, а тому зменшує значення R. За умов, коли N приблизно досягає довжини тексту ($L \approx 2{,}5$ млн символів $\approx$ 480 тис. слів), маємо різкий перехід до теоретичної границі рандомного тексту $R \to 1$. Спостерігаємо і різке зменшення параметра повторюваності v_0: очевидно, що тоді складова повторюваності, пов'язана із внутрішньою будовою окремих слів, залишається незмінною, проте зникає складова, зумовлена типовими комбінаціями та впорядкуванням слів у тексті. Кореляція параметрів R і v_0 доволі висока ($r = 0{,}97$ – див. рис. 4, *б*), а основний внесок до неузгодження дає лише порівняно нестабільна ділянка крутих змін графіків, якщо $N \sim 10^4 \div 10^5$ на рис. 4, *а*. Отже, повторюваність і семантичне навантаження текстів непогано корелюють, що дає підстави сподіватись на розрізнення природних і рандомних текстів за параметром v_0.

Хоча ми не відображали відповідні результати графічним матеріалом, зазначимо також, що рандомізація природних текстів приводить до інтенсивнішого насичення функції $v(t)$, тобто вона дещо зменшує стандартне відхилення Δv.

Одночасні залежність $v_0(N)$ повторюваності від ступеня рандомізації природних текстів, з одного боку, та взаємозв'язок $v_0(H)$ між повторюваністю та інформаційною ентропією для текстів мавпи Міллера, з іншого боку, видаються несумісними явищами та породжують парадокс. Справді, рандомізація принципово не впливає на односимвольну ентропію Шеннона, оскільки перестановка лінгвістичних елементів у моделі мішка зі словами не може змінити базових частотно-рангових співвідношень для символів алфавіту та ймовірностей p_i у формулі (3). А це мало б означати, що повторюваність рандомізованих текстів,

на відміну від повторюваності текстів мавпи Міллера, не залежить від ентропії, хоча, звісно, важко припустити, що різні рандомні моделі природної мови керуються різними закономірностями для ентропії.

Описану суперечність можна пояснити тим, що для коректного врахування надлишковості природної мови насправді не можна обмежуватися односимвольним підходом – натомість необхідно розраховувати ентропію, яка ґрунтується на багатосимвольному (n-грамному) підході, тобто на умовних імовірностях $p_{i|jkl\ldots}$ елементів алфавіту в формулі (3) [17]. Інакше: за рандомізації залишається незмінною лише односимвольна ентропія, проте змінюється багатосимвольна ентропія, яку, щоправда, вкрай складно розрахувати. Так формально не порушується можлива залежність $v_0(H)$ для рандомізованих текстів. Відповідно, односимвольна ентропія дає коректні результати для повторюваності тільки для рандомних (наприклад, повністю рандомізованих) текстів, у яких відсутні будь-які закономірності переважного комбінування букв.



а *б*

Рис. 4. Залежності параметра повторюваності v_0 (шкала справа) і параметра семантики R (шкала зліва) від кількості N циклів рандомізації тексту (див. деталі в тексті статті) (а); кореляція параметрів R і v_0 (б) / Dependences of the repetitiveness parameter v_0 (right scale) and the semantics parameter R (left scale) on the number N of text randomization cycles (see details in the text of the article) (a); correlation of the parameters R and v_0 (b)

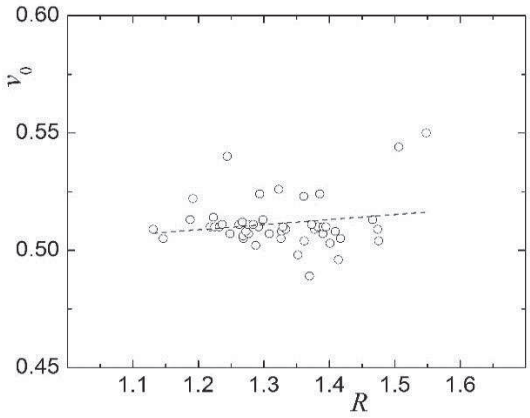


Рис. 5. Кореляція параметра повторюваності v_0 і параметра семантики R, розрахована для 50 природних текстів англійською мовою / Correlation of the repetitiveness parameter v_0 and the semantics parameter R calculated for 50 natural texts in English

Наявність складніших закономірностей для природних текстів підтверджує й залежність рівноважної повторюваності v_0 від параметра семантичного наповнення R, одержана для 50 англомовних текстів (рис. 5): кореляція між названими параметрами слабка і становить лише $r \approx 0{,}18$. Інакше кажучи, хоча повторюваність можна використовувати для відмежування семантично наповнених природних текстів від семантично порожніх рандомних текстів, на практиці параметр v_0 непридатний для розв'язання тоншої задачі – розрізнення семантичного навантаження та структурування різних природних текстів. Нарешті, стандартне відхилення Δv_0 рівноважної повторюваності для доволі різних за змістом вивчених нами англомовних текстів виявляється не набагато меншим, аніж відповідний параметр для природних текстів, написаних 30 різними алфавітними мовами (див. дані [13]). Отже, міжмовні та внутрішньомовні відмінності повторюваності Δv_0 мають порівнянні величини.

Цікаво також виключити вплив на повторюваність у природних текстах такої змінної, як семантика, досліджуючи зміни величини v_0 внаслідок впливу лише системи письма. Це найпростіше зробити для текстів із однаковим змістом, написаних різними мовами, – скажімо для Біблії. З цією метою ми вивчили 88 текстів Біблії, написаними 47 мовами (разом з різними варіантами та перекладами різних авторів). Для порівняння також вивчено надійно рандомізовані за словами тексти Біблії ($N = 5 \cdot 10^7$ циклів рандомізації). Як і обговорені вище результати, дані табл. 2 теж підтверджують вагомий внесок лінгвістичної системи подання тексту до параметра v_0, розрахованого в обох режимах "символи" та "слова". Справді, діапазон змін повторюваності v_0 між різними текстами, який ми вимірювали стандартним відхиленням δv_0, за порядком величини такий самий, як для текстів із різною семантикою (див. дані, наведені вище), а тому навряд чи семантика визначально впливає на повторюваність,

допоки ми порівнюємо між собою різні природні тексти, а не рандомні тексти з природними.

Тепер порівняємо внески до параметра повторень v_0, спричинені такими можливими факторами: (1) повтореннями, не зумовленими лінгвістичними закономірностями, які є суто стохастичними і визначаються законами комбінаторики для заданих розміру алфавіту та частот символів, незалежно від часового масштабу цих повторень, (2) внутрішньою структурою слів, (3) типовими поєднаннями або повтореннями слів у межах окремих речень і (4) типовими повтореннями, спостережуваними на масштабах різних, зокрема віддалених одне від одного, речень. Для цього треба оцінити зміни повторюваності із переходом від природних текстів до текстів, рандомізованих на лінгвістичних рівнях символів (абревіатура c-RT), слів (w-RT) і речень (s-RT).

Табл. 2. Параметри повторюваності v_0, розраховані в режимах "символи" та "слова" для 88 природних текстів (позначення "NT"; "Біблія"), написаних 47 алфавітними мовами світу, та їхніх версій, рандомізованих за словами (*w-RT*) / Repetitiveness parameters v_0 calculated in the "symbols" and "words" modes for 88 natural texts (notation "NT"; "Bible") written in 47 alphabetic languages of the world, and their word-randomized versions (*w-RT*)

Параметри	v0 (режим "символи", NT)	v0 (режим "слова", NT)	v0 (режим "символи", w-RT)	v0 (режим "слова", w-RT)
Середнє значення <v0>	0,54	0,27	0,45	0,14
Стандартне відхилення δv0	0,02	0,03	0,02	0,01

У табл. 3 наведено результати масштабного порівняння параметрів v_0 рандомізованих текстів із даними для природних текстів (абревіатура NT). Загалом вивчено 150 текстів, написаних 40 алфавітними мовами, середня довжина яких дорівнювала $\langle L \rangle \approx 190$тис. символів. Загалом дані табл. 3 для рандомізованих за словами текстів (w-RT) добре корелюють із даними табл. 2 для режиму "символи". Наголосимо, що параметри v_0 природних текстів визначаються всіма механізмами повторюваності (1)–(4), а і зниженням лінгвістичного рівня рандомізації відповідні параметри почергово позбуваються впливу повторень типу (4), (3) і (2), так що тексти c-RT можуть містити лише внесок механізму (1).

Із табл. 3 бачимо, що вплив повторень на великих масштабах (4) на v_0 незначний або й взагалі відсутній, оскільки статистичні межі ефекту ($\pm 0{,}03$) істотно перевищують середнє значення зниження повторюваності із переходом від природних текстів до текстів, рандомізованих за реченнями ($-0{,}01$). Отже, довгосяжні (на часових відстанях, близьких до довжини L самого тексту) кореляції символьних n-грам навряд чи роблять істотний внесок до рівноважної повторюваності v_0. Оскільки саме такі кореляції найбільше відрізняють природні тексти від їхніх рандомних моделей і фактично відповідають за семантику тексту як цілого та за наявність у ньому ключових слів, цей експериментальний факт добре узгоджується з тим, що кореляція повторюваності v_0 та параметра R природних текстів слабка (див. вище).

Водночас внесок (3) до повторюваності від типових колокацій слів у межах речень помітніший і є статистично надійним фактом (зміни $-0{,}04 \pm 0{,}02$). Нарешті, внесок (2) до повторюваності, зумовлений повтореннями літер завдяки типовій внутрішній будові слів, домінує, тому рандомізація на рівні символів, яка "руйнує" слова, найбільше знижує повторюваність ($-0{,}17 \pm 0{,}02$).

Табл. 3. Параметри повторюваності v_0, розраховані в режимі "символи" для 150 природних текстів (позначення "NT"), написаних 40 алфавітними мовами, а також їхніх версій, рандомізованих на рівнях речень (*s-RT*), слів (*w-RT*) і символів (*c-RT*) / Repetitiveness parameters v_0 calculated in the "symbols" mode for 150 natural texts (notation "NT") written in 40 alphabetic languages, and their versions randomized on the sentence (*s-RT*), word (*w-RT*) and character (*c-RT*) levels

Параметри	v_0 (NT)	v_0 (*s-RT*)	v_0 (*w-RT*)	v_0 (c-RT)
Середнє значення $\langle v_0 \rangle$	0,51	0,50	0,47	0,34
Стандартне відхилення δv_0	0,03	0,03	0,03	0,02
Середнє значення відмінності від NT	–	–0,01	–0,04	–0,17
Стандартне відхилення відмінності від NT	–	0,03	0,02	0,02

Зауважмо, що, як і для природних текстів, для жодного із досліджених нами 150 рандомізованих за символами текстів не виявлено осциляцій повторюваності $v(t)$, які притаманні класичним текстам мавпи Міллера із параметром $b = 1$ перепаду частот літер. Водночас за фактичною будовою і цілковитою відсутністю семантики повністю рандомізовані тексти еквівалентні текстам мавпи Міллера. Вище ми вже припускали, що причиною такої разючої відмінності поведінки функцій $v(t)$ для природних текстів і текстів мавпи Міллера з $b = 1$ є значний перепад ($b \sim 300$) частот літер у природних текстах, який зумовлює швидке загасання осциляцій $v(t)$. Для прямої перевірки цієї гіпотези потрібні порівняння залежностей $v(t)$ для (еквівалентних

у всіх інших аспектах) рандомізованих за символами текстів і узагальнених текстів мавпи Міллера ($b \neq 1$).

Для відповідних експериментів вибрано 9 вихідних англомовних текстів. З метою коректного порівняння з текстами мавпи Міллера, які типово генерують без розділових знаків, цифр тощо, ми очистили природні тексти від цих знаків (тобто результуючий "алфавіт" містив $M = 27 \div 36$ символів, разом з пробілом), а також надійно ($N = 20$ млн $>> L$) рандомізували їх за символами. Для порівняння було згенеровано аналогічні за параметрами тексти мавпи Міллера. Вони містили той самий "алфавіт" M, разом з пробілом, а рандомні частоти всіх символів ми задали на основі розподілів частот f_i, притаманних кожному з природних текстів. За браком місця не наводимо графічних ілюстрацій залежностей характеристик повторюваності $v(t)$ для порівнюваних рандомізованих текстів і аналогічних текстів мавпи. Зазначимо лише, що для першої групи текстів одержано усереднене для всіх текстів значення $\langle v_0 \rangle \approx 0{,}3390$, а для другої – $\langle v_0 \rangle \approx 0{,}3394$, де стандартне відхилення відмінностей параметрів δv_0, знайдене усередненням для всіх пар рандомізований текст – текст мавпи Міллера, виявилося екстремально малим: $(3 \div 5) \cdot 10^{-4}$. Тобто, принаймні в термінах повторюваності, ці дві рандомні моделі природної мови та відповідні рандомні тексти є ідентичними.

Отже, можна вважати повністю доведеним той факт, що відсутність явища насичення величини v_0 для класичних текстів мавпи Міллера із $b = 1$ не є якимось характерним маркером цих текстів; це пов'язано лише із відсутністю перепаду відносних частот символів у них. Якщо ввести згаданий перепад у рандомну модель Міллера і зробити його достатньо великим (наприклад, $b \sim 300$ для алфавіту $M = 27$), то явище осциляцій для цих узагальнених текстів мавпи Міллера зникає, так само, як в рандомізованих за символами природних текстах або у вихідних природних текстах.

Обговорення отриманих результатів. Отже, ми здійснили практичну реалізацію розрахунків і подальший статистичний аналіз повторюваності у текстах на основі алгоритму дерев суфіксів Укконена, яка виявилася життєздатною, а її швидкодія – достатньою для досліджень навіть великих за розмірами окремих текстів. Значення повторюваності рандомних текстів мавпи Міллера залежить насамперед від розміру алфавіту та перепаду частот різних символів у тексті, причому вплив останніх факторів можна підсумувати в термінах односимвольної інформаційної ентропії Шеннона. Зі зростанням розмірів алфавіту та за умови однакових частот усіх символів посилюються осциляції на залежностях $v(t)$, які можна послабити, збільшивши відмінності цих частот. Водночас, повторюваність рандомізованих текстів залежить найперше від ступеня рандомізації, зі зростанням якого зменшуються семантичне навантаження та повторюваність у тексті.

Попри загалом доволі різну поведінку функцій повторюваності $v(t)$ і підсумкових параметрів v_0 і Δv для природних текстів і рандомних текстів різних типів, на практиці все ж важко уникнути ситуації, коли ці тексти можна сплутати, зважаючи тільки на дані повторюваності v_0 і Δv. Прикладом є тексти мавпи Міллера з малими алфавітами, для яких параметр v_0 може становити ~ 0,5, що близько до параметрів природних текстів. Тому для підвищення надійності розпізнавання природних і рандомних символьних послідовностей можна порекомендувати, окрім повторюваності, розглядати додатково й сам розмір алфавіту. Так приходимо до ідеї про двовимірне подання текстів на шкалі $M - v_0$ (порівн. із дещо схожою ідеєю [11]).

На рис. 6 цей підхід проілюстровано даними для понад 170 вивчених вище текстів. Очевидно, що кількісна поведінка природної мови та її простих рандомних моделей статистично надійно розрізняється: "крива", сформована "темними" точками для природних текстів, помітно зміщена вверх на шкалі повторюваності, порівняно із "кривою", сформованою "світлими" точками. Як результат, повторюваність слугує статистично надійним критерієм відрізняння семантично наповнених природних текстів від рандомних текстів, породжених найпростішими моделями природної мови.

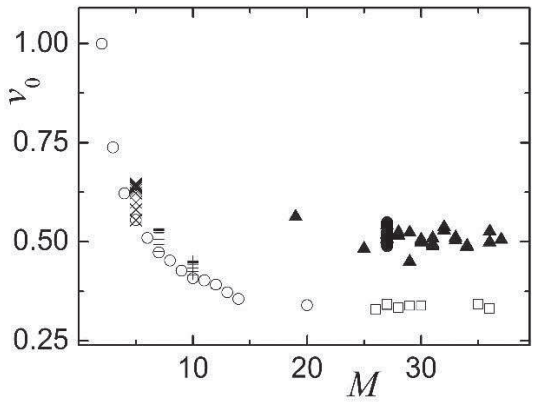


Рис. 6. Порівняння залежностей параметра повторюваності v_0 у режимі "символи" для природних текстів і різних типів рандомних текстів: "○" – тексти мавпи Міллера для $b = 1$, "×", " – " і "+" – ті самі тексти для $b \neq 1$ і $M = 5$, 7 і 10, відповідно, "□" – природні тексти, рандомізовані за символами, "●" – 50 англомовних природних текстів і "▲" – 30 природних текстів іншими алфавітними мовами / Comparison of dependences of the repetitiveness parameter v_0 calculated in the "symbol" mode for the natural texts and the random texts of different types: "○" corresponds to the Miller's monkey texts at b = 1, "×"," – " and "+" to the same texts at $b \neq 1$ and $M = 5$, 7 and 10, respectively, "□" to the natural texts randomized by symbols, "●" to 50 English-language natural texts, and "▲" to 30 natural texts in different alphabetic languages

Наукова новизна дослідження полягає у тому, що в цій роботі вивчено характеристики повторюваності для найпростіших рандомних моделей природної мови і на цій основі розроблено та апробовано метод розпізнавання природних текстів і текстів, породжених рандомними нульовими моделями природної мови.

Практична значущість отриманих результатів роботи полягає в можливості застосування розробленого підходу до оцінювання семантичного навантаження текстів.

Висновки / Conclusions

Запропоновано практичну реалізацію кількісного статистичного аналізу параметрів повторюваності за алгоритмом суфіксних дерев Укконена. Емпірично доведено, що алгоритм має наближено лінійну часову складність $O(L\log L)$, що дає змогу успішно розраховувати параметри повторюваності для великих текстів, хоча не для текстових корпусів.

Виконано масштабні емпіричні дослідження текстів, що ґрунтуються на моделі мавпи Міллера. Встановлено основні фактори, які впливають на поведінку функції повторюваності $v(t)$ і параметри v_0 і Δv відповідних текстів – розміри алфавіту та перепад частот символів алфавіту. Доведено однозначний зв'язок повторюваності із односимвольною інформаційною ентропією Шеннона.

Вивчено повторюваність природних текстів, рандомізованих за символами, словами та реченнями, та виявлено її зв'язок із параметрами семантичного навантаження цих текстів. Доведено еквівалентність моделей мавпи Міллера та "мішка зі словами" з погляду поведінки повторюваності.

Показано, що параметри повторюваності, доповнені інформацією про алфавіт, можна використовувати для розрізняння семантично наповнених природних текстів і семантично порожніх стохастичних послідовностей символів.

References

1. Aletti, G., & Crimaldi, I. (2021). Twitter as an innovation process with damping effect. *Sci. Rep*., 11, 21243 (15 p.). https://doi.org/10.1038/s41598-021-00378-4
2. Deng, W., Xie, R., Deng, S., & Allahverdyan, A. E. (2021). Two halves of a meaningful text are statistically different. *J. Statistical Mechanics: Theory and Experiment*, 3, 033413 (28 p.). https://doi.org/10.1088/1742-5468/abe947
3. Lai, U., Randhawa, G. S., & Sheridan, P. (2023). Heaps' law in GPT-Neo large language model emulated corpora. Proceedings of the Tenth International Workshop on Evaluating Information Access (EVIA 2023), a Satellite Workshop of the NTCIR-17 Conference, 20-23. https://doi.org/10.20736/0002001352
4. Kushnir, O., Drebot, A., Ostrikov, D., & Kravchuk, O. (2024). Vlastyvosti leksychnyh merezh, pobudovanyh na pryrodnyh i randomnyh tekstah [Properties of lexical networks built on natural and random texts]. *Electronics and Information Technologies*, 28, 22-37 (in Ukrainian). https://doi.org/10.30970/ eli.28.3
5. Zhenhan Qi (2025). An analysis of Markov model's applications. *Theoretical and Natural Science*, 92, 82-87. https://doi.org/ 10.54254/2753-8818/2025.21613
6. Amancio, D. R., Altmann. E. G., Rybski. D., Oliveira Jr., O. N., & da F. Costa, L. (2013). Probing the statistical properties: application to the Voynich manuscript. *PLoS ONE*, 8 e67310 (10 p.). https://doi.org/10.1371/journal.pone.0067310
7. Kim Chol-jun (2025). Proper interpretation of Heaps' and Zipf's laws. arXiv:2305.15413v3, 1-18. Retrieved from: https://arxiv. org/abs/2305.15413
8. Golcher, F. (2007). A stable statistical constant specific for human language texts, 1-6. Retrieved from: https://www. academia.edu/5986557/A_Stable_Statistical_Constant_Specific _for_Human_Language_Texts
9. Kimura, D., & Tanaka-Ishii, K. (2014). Study on constants of natural language texts. *J. Language Processin*g, 21, 877-895. https://doi.org/10.5715/jnlp.21.877
10. Smyth, W. F. (2014). Large-scale detection of repetitions. *Phil. Trans. R. Soc. A*, 372, 20130138 (11 p.). https://doi.org/ 10.1098/rsta.2013.0138
11. Tanaka-Ishii, K., & Aihara, S. (2015). Computational constancy measures of texts – Yule's K and Renyi's entropy. *Computational Linguistics*, 41, 481-502. https://doi.org/10.1162/ COLI_a_00228
12. Fu, Z., Lam, W., So, A. M.-C., & Shi, B. (2021). A theoretical analysis of the repetition problem in text generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14), 12848-12856. https://doi.org/10.1609/aaai.v35i14.17520
13. Kushnir, O. S., Ivanitskyi, L. B, Kashuba, A. I., Mostova, M. R., & Mykhaylyk, V. B. (2021). Repetition characteristic for single texts. *CEUR Workshop Proceedings*, 2870, 629-641. https://ceur-ws.org/Vol-2870/paper47.pdf
14. Kushnir, O. S., Ivanitskyi, L. B., Kashuba, A. I., Mostova, M. R., & Mykhaylyk, V. B. (2021). Large-scale studies of the repetition characteristic for different models of symbolic sequences. Proceedings of 12th IEEE International Conference on Electronics and Information Technologies, 61-66. https://doi.org/ 10.1109/ELIT53502.2021.9501102
15. Salkar, N., Trikalinos, T., Wallace, B., & Nenkova, A. (2022). Self-repetition in abstractive neural summarizers. Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing, 2, 341-350. https://doi.org/10.18653/v1/2022.aacl-short.42
16. Rabea, Z., El-Metwally, S., Elmougy, S., & Zakaria, M. (2022). A fast algorithm for constructing suffix arrays for DNA alphabets: a review. *Journal of King Saud University – Computer and Information Sciences*, 34(7), 4659-4668. https://doi.org/10.1016/ j.jksuci.2022.04.015
17. Sepúlveda-Fontaine, S. A., & Amigó, J. M. (2024). Applications of entropy in data analysis and machine learning. *Entropy*, 26, 1126 (42 p.). https://doi.org/10.3390/e26121126

O. S. Kushnir, V. O. Futey
Ivan Franko National University of Lviv, Lviv, Ukraine

STUDIES OF REPETITIVENESS FOR THE SIMPLEST RANDOM NATURAL LANGUAGE MODELS

The article addresses a currently important problem of natural language processing, the development of methods for assessing repetitiveness in textual documents and the empirical clarification of the resources of these methods for analyzing the presence of semantic load in texts. So far, the approaches based on the laws of statistical linguistics such as Zipf's, Pareto's and Heaps' laws have been mainly used for this aim, as well as the analysis of word-clustering phenomena and long-term correlations of word tokens. We have developed software for quantitative analysis of the repetitiveness in texts, using the Ukkonen's suffix-tree algorithm. We suggest analyzing the averaged repetitiveness parameter v_0 and the corresponding standard deviation Δv. It is empirically proven that the algorithm implemented with our software exhibits an approximately linear time complexity, O (L log L) (with L denoting the symbolic-series length), which corresponds to theoretical predictions. The texts analyzed for the repetitiveness have benn natural texts in English and a number of other alphabetic (non-hieroglyphic) languages, as well as random Miller's monkey texts and natural texts randomized gradually at the linguistic levels of symbols, words and sentences. It is confirmed that, outside the initial section of natural texts, where the repetition function fluctuates, this function exhibits a saturation phenomenon, with a saturated value $v_0 \approx 0.5$. The texts based on the Miller's random model are studied in detail. It is found that the behavior of the repetitiveness function and the parameters v_0 and Δv of these texts are influenced primarily by a size of the alphabet and a preset distribution of relative frequencies of the alphabet symbols. It is proved that the average repetition v_0 for these texts is related to the Shannon information entropy in its simplest representation. We have ascertained that the repetitiveness for the "bag of words" model correlates with the parameter of the average semantic load of a text. Finally, we suggest employing the repetition parameters v_0 and Δv for recognizing semantically filled natural texts against semantically empty, stochastic symbolic time series.

Keywords: natural language processing, natural language models, semantic analysis, text data processing, symbolic time series, repetitiveness.

Інформація про авторів:

Кушнір Олег Степанович, д-р фіз.-мат. наук, професор, завідувач кафедри оптоелектроніки та інформаційних технологій.
Email: oleh.kushnir@lnu.edu.ua; https://orcid.org/0000-0002-1545-7666

Футей Володимир Олександрович, аспірант кафедри оптоелектроніки та інформаційних технологій.
Email: volodymyr.futei@lnu.edu.ua; https://orcid.org/0009-0003-1053-9835

Цитування за ДСТУ: Кушнір О. С., Футей В. О. Дослідження повторюваності для найпростіших рандомних моделей природної мови. *Український журнал інформаційних технологій*. 2025, т. 7, № 2. С. 82-92.

Citation APA: Kushnir, O. S., & Futey, V. O. (2025). Studies of repetitiveness for the simplest random natural language models. *Ukrainian Journal of Information Technology*, 7(2), 82-92. https://doi.org/10.23939/ujit2025.02.82