УДК 004.6, 004.9, 538.9

О. С. Кушнір, М. А. Альфавіцький, В. Є. Дзіковський [*], Л. Б. Іваніцький, С. В. Рихлюк, В. І. Сокульський
Львівський національний університет імені Івана Франка,
кафедра оптоелектроніки та інформаційних технологій,
[*] Природничий коледж Львівського національного університету імені Івана Франка

СТАТИСТИКА ПОЯВИ СЛІВ У ПРИРОДНИХ І РАНДОМНИХ ТЕКСТАХ

Експериментально досліджено статистичні розподіли, що описують появу слів у кількох природних текстах, а також похідних від них рандомних текстах. Показано, що масова функція ймовірності відповідних інтервалів між словами є практично однаковою для природних і рандомних текстів і виявляє важкий вейбулівський хвіст, що не узгоджується із суто стохастичним характером цих інтервалів. Помітні відхилення динаміки зростання словника природних і рандомних текстів від динаміки, передбаченої степеневим законом Гіпса, а також кросовер у словнику одного з природних текстів підтверджують потребу в узагальненні цього закону.

Ключові слова: природні тексти, рандомні тексти, статистичні закони лінгвістики, словник, розподіли з важким хвостом.

We study experimentally statistical distributions that describe the appearance of words in a number of natural texts, as well as in the random texts derived on their basis. It is shown that the probability mass function of the respective intervals between words is practically the same for the natural and random texts and manifests a fat tail, which is inconsistent with purely stochastic character of those intervals. Significant deviations of the vocabulary growth dynamics found for the natural and random texts from the dynamics predicted by the power Heaps' law, together with a crossover found in the dictionary of one of the natural texts, confirm a need in generalization of that law.

Key words: natural texts, random texts, statistical laws of linguistics, vocabulary, fat-tailed distributions.

Вступ і постановка проблеми

Однією з важливих проблем теорії складних систем і мереж є співвідношення між кількістю елементів у системі (розміром системи) та кількістю V різних класів, до яких належать ці елементи (різноманітністю класів або "словником") [1, 2]. Прикладами є кількість міст залежно від населення країни або залежність кількості словоформ (словника) від загальної кількості слововживань (довжини тексту). Досліджуючи міста, переважно спостерігають за деяким їхнім ансамблем, тобто за статичним часовим зрізом ($t = \text{const}$) складної системи, а не за розвитком у часі одного міста. Тоді висновки про динаміку системи роблять за припущення її ергодичності. З іншого боку, тексти є впорядкованими в "часі" системами, в яких явно є часовий вимір: текст розвивається за деяким динамічним законом, а кожному слову відповідає певний дискретний "час" появи. Тому співвідношення між V і t можна розглядати залежно від розміру системи, що зростає, аж до повних розміру тексту $t = t_{\max}$ і словника $V = V_{\max}$ ($V_{\max} = V(t_{\max})$). Функціональну залежність словника $V(t)$ переважно описують степеневим законом Гіпса (або Гердана – див., наприклад, [3–5]):

$$V(t) \propto t^q , \qquad (1)$$

за яким часова динаміка словника сублінійна ($q = \text{const}$, $0 < q \le 1$). Оскільки вираз (1) лише наближено описує емпіричні дані, існують інші моделі опису зростання словника. Наприклад, у працях [6–9] припускають зміни параметра $q = q(t)$, а автори [10–13] виражають залежність $V(t)$ складнішими за (1) формулами. Так, за результатами [13] для достатньо великих текстів маємо

$$V(t) \approx V_{\max}\left(1 - \frac{\mathrm{Li}_g(1 - t/t_{\max})}{x(g)}\right). \qquad (2)$$

У (2) входять спеціальні функції – полілогарифмічна ($\mathrm{Li}_g(x) = \sum_{n=1}^{\infty} x^n / n^g$, де $|x| < 1$) і дзета-функція Рімана ($x(g) = \mathrm{Li}_g(1)$), а g – це постійний показник степеня в другому законі Ціпфа (в разі виконання закону Гіпса і нерівності $g \le 2$ припускають асимптотичну рівність $g = q + 1$ – див., наприклад, [8, 14]).

Незалежно від конкретної моделі опису словника, універсальною нульовою гіпотезою в описі природних текстів (NT) є припущення про їхній рандомний характер, за яким поява того чи іншого слова в тексті є випадковою подією, що не залежить від появи інших слів. Тривіальним практичним утіленням такого підходу є рандомізовані тексти (RT) – тексти, побудовані на основі NT так, що черговість слів у них змінена довільно. Хоча RT принципово не описують окремих явищ в NT, пов'язаних із масштабними флуктуаціями та довгосяжними кореляціями (див. [15–20]), ці явища, хай інтуїтивно добре зрозумілі, порівняно слабкі. Мабуть, саме тому успіхи нульової гіпотези виявилися несподівано переконливими. Зокрема, всі статичні властивості NT і RT збігаються, оскільки перестановки лінгвістичних елементів не впливають ні на залежність ранг–частота для них (перший закон Ціпфа), ні на розподіл імовірності слів залежно від їхньої частоти (другий закон Ціпфа). В літературі також переважно припускають ідентичність більшості динамічних властивостей NT і RT, окрім кореляційних, зокрема незмінність закону Гіпса для них, хоча ці питання потребують уважного аналізу. У будь-якому разі вважається, що степеневий закон Гіпса (1) або складнішу формулу (2) для $V(t)$ можна безпосередньо одержати зі степеневих законів Ціпфа та припущення про стохастичний розподіл слів у тексті; наслідками стохастичності вважають і асимптотичні співвідношення поміж різними степенями лінгвістичних законів на зразок q і g [1–7, 9, 13, 21]. Це доводить важливість вивчення стохастичних складних систем. Чи не єдині винятки – це висновок [8] про відсутність потреби в стохастичній гіпотезі для обґрунтування зв'язків між законами Гіпса та Ціпфа і висновок [14] про непринциповість такої потреби.

Широковідомою стохастичною моделлю текстів, до якої часто вдаються дослідники, є динамічна модель зростання Саймона [22]. Рандомний текст Саймона (ST) будують на основі "резервуару слів" – словника. Динаміка генерування ST така, що кожне наступне слово в тексті буде або "новим" (його вибирають із резервуару з деякою ймовірністю p_0), або "старим" (його вибирають з імовірністю $(1 - p_0)$ з наявних у тексті слів так, що ймовірність вибору цього слова пропорційна до кількості таких слів у тексті). Такий алгоритм узгоджується із загальновідомими для складних систем механізмом переважного приєднання і принципом "багатий стає багатшим" [23]. Модель Саймона передбачає степеневі лінгвістичні закони на рівні слів і масштабну інваріантність властивостей ST, що узгоджується з властивостями NT. Хоч і нечасто, в літературі висловлювали окремі критичні зауваження на адресу ST, зокрема через неприродну властивість $q = 1$ [24, 25] і розподіл низько- та високочастотних слів у тексті, не притаманний NT [1, 7].

Предметом вивчення в цій праці є часова динаміка появи нових слів у текстах різних типів (у загальніших термінах – появи нових класів елементів у складній системі), яка врешті-решт і визначає закон зростання словника, а також порівняння цієї динаміки для низки NT, RT і ST. Найпершим предметом вивчення вибрано статистичний розподіл часів очікування нових слів, який є чутливим зондом кореляцій у тексті [18] і дає змогу розпізнати ключові слова [26–29], а також імовірність появи нових слів у тексті. За нашою інформацією, питання часової динаміки появи

нових слів у літературі досі майже не торкались, за винятком недавньої праці [13] у чи не найповажнішому міжнародному періодичному виданні з фізики. Водночас, дані й висновки цієї праці, яка безпосередньо стимулювала наше дослідження, на нашу думку, суперечливі та заслуговують на докладніший аналіз.

Об'єкти та методи досліджень

Об'єктами досліджень були чотири природні тексти NT1÷NT4, їхні рандомізовані версії RT1÷RT4, а також чотири саймонівські тексти ST1÷ST4 ($p_0 = 0{,}1$), створені на основі словників відповідних NT. Для коректного порівняння результатів [13] із нашими даними і унеможливлення впливу мови на них ми вивчали англомовні тексти. Текстами NT1÷NT4 були художні тексти "Joseph Andrews", том 1 (автор Г. Філдінг), "Villa Rubein and Other Stories" (Дж. Голсуорсі), "Redgauntlet" (В. Скотт) і "The Lord of the Rings" (Дж. Р. Р. Толкін). Їхні розміри відповідно 64,5; 95,9; 188,0 і 479,2 тис. слів. Тексти в кодуванні UTF8 здійснено стандартне попереднє оброблення для досліджень на лексичному лінгвістичному рівні, зокрема уніфіковано великі й малі літери, а також вилучено всі знаки, крім літер і цифр (див., наприклад, [20, 29]). Відповідно до алгоритму рандомізації, розміри текстів RT1÷RT4 були ідентичними до розмірів NT1÷NT4, а розміри текстів ST1÷ST4 становили 50, 100, 150 і 170 тис. слів.

Для рандомізації текстів NT у середовищі Visual Studio було створено програму-рандомізатор на лінгвістичних рівнях знаків, слів і речень. Програма передбачала режими глобального "перемішування" (на масштабах усього тексту) та локального "перемішування" (на визначеному користувачем масштабі $t_1 \div t_2$, де $t_2 - t_1 < t_{\max}$, із подальшим переміщенням поточного вікна "перемішування"). Для надійного забезпечення стохастичного характеру RT кількість циклів рандомізації становила 10^9; було здійснено перевірку роботи базового рандомного генератора на предмет однорідності статистичного розподілу випадкової змінної. Нижче висвітлено лише дані, одержані внаслідок глобальної рандомізації на лексичному рівні.

Також створено програму для формування текстів ST. Нарешті, наша основна програма опрацьовувала статистичні розподіли ймовірності часів очікування $p(\tau)$ перших появ слів і ймовірності $p(t)$ появи нових слів на тій чи іншій ділянці тексту.

Табл. 1 пояснює основні поняття, потрібні для кількісного опису появи нових слів у тексті та їхньої часової динаміки, на прикладі короткого синтетичного тексту із восьми слів. Тут нові слова з'являються на позиціях $\{t_i\} = 0, 1, 3, 6, \ldots$. Попри дискретність послідовності, поява нових слів – порівняно рідкісна подія (для текстів досліджуваних розмірів маємо $V_{\max} / t_{\max} \sim 0{,}03 \div 0{,}08$), а тому ймовірність цієї появи нульова за багатьох значень t (наприклад, $p(t) = 0$ за $t = 2, 4, 5, 7, \ldots$) і поводиться як квазінеперервна функція. Відповідно, можна формально визначити функцію густини ймовірності першої появи $p(t)$, розраховуючи її гістограму як відношення кількості нових слів (кількості одиничок у четвертому рядку табл. 1) на деякому малому інтервалі тексту (біні) до повного словника $V_{\max}$ на всьому тексті. Так, у тексті-зразку з табл. 1 за ширини бінів, що дорівнює чотири слова, для першого та другого бінів маємо відповідно $p(1) = 3/4$ і $p(2) = 1/4$. Зауважимо, що тут порядковий номер біну ($n = 1, 2$) вжито як незалежну змінну замість абсолютної позиції t слів у тексті.

Зазвичай термін "часи очікування" вживають у значенні часових інтервалів t' між двома сусідніми появами того самого слова на позиціях t'_{i-1} і t'_i, тобто йдеться про часи очікування повторень деякого слов. Випадкові величини t' формують дискретну послідовність; їх можна розрахувати за формулою $t'_i = t'_i - t'_{i-1} - 1$ (див., наприклад, [26, 28–30]). Наприклад, для слова "he" у тексті з табл. 1 маємо набір позицій $\{t'_i\} = 0, 2, 5, 7, \ldots$ і відповідний набір інтервалів $\{t'_i\} = 1, 2, 1, \ldots$. У цій праці ми узагальнюємо поняття часів очікування на випадок послідовності перших появ слів: $t_i = t_i - t_{i-1} - 1$. Із табл. 1 для часів очікування першої появи слів маємо послідовність $\{t_i\} = 0, 2, 1, \ldots$. Масову функцію ймовірності $p(\tau)$ розраховують за визначенням:

$p(t) = N_t / N$ ($\tau = 0, 1, \ldots$), де N_t – кількість часів очікування, значення яких дорівнює t, $N = \sum_0^\infty N_t$ – кількість усіх часів очікування. Наприклад, за даними останнього рядка табл. 1 одержимо $p(0) = p(1) = p(2) = 1/3$.

Таблиця 1

Ілюстрація кількісних параметрів перших появ слів у текстах на прикладі синтетичного тексту

Послідовність слів у тексті	he	she	he	she	it	he	they	he	…
Абсолютна позиція слова	0	1	2	3	4	5	6	7	…
Позиція першої появи слова t_i	0	1	–	–	4	–	6	–	…
Наявність/відсутність факту першої появи слова на позиції t_i	1	1	0	1	0	0	1	0	…
Час очікування нового слова (час між двома сусідніми появами нових слів) t_i	–	0	–	–	2	–	1	–	…

На додаток до функцій розподілів $p(\tau)$ і $p(t)$, визначають відповідні комплементарні кумулятивні функції ймовірності $P(\tau)$ і $P(t)$, які зазвичай виявляють слабші шуми, ніж $p(\tau)$ і $p(t)$: наприклад, $P(t) = \sum_{h=t}^\infty p(h)$. На завершення зазначимо, що серед часів t'_i автори праці [13] окремо виділяють "часи очікування" для випадку $i = 1$, трактуючи їх як проміжки від початку тексту до першої появи ($i = 1$) кожного слова. За такого визначення параметри t'_1 із роботи [13] не є часами очікування в прямому розумінні – це швидше позиції першої появи слів, тобто в наших позначеннях маємо $t'_1 \equiv t$.

Емпіричні дані та їхнє обговорення

Природні тексти

За браком місця нижче наведено лише типові приклади одержаних емпіричних розподілів імовірності $p(\tau)$ і $P(\tau)$ часів очікування першої появи слів і масові функції розподілу $p(t)$ позицій першої появи слів для окремих текстів. На рис. 1, *а, б* показано функції розподілу $p(\tau)$ і $P(\tau)$ для текстів NT3 і NT4. Зазначимо, що за найменших τ (типово за $\tau < 4$) на залежностях $p(\tau)$ є ділянка з пологим нахилом, не помітна в масштабі рис. 1, *а*. Її наявність, мабуть, пов'язана із впливом синтаксису, який обмежує надто часту появу нових слів (порівн. із даними праці [17]).

За стандартною нульовою гіпотезою, появи нових слів у природному тексті – це рідкісні незалежні події, ймовірність сукупної появи яких повинна описуватися розподілом Пуассона. Тоді часи очікування τ повинен характеризувати експоненційний розподіл

$$p(t) = \bar{t}^{-1} \exp(-t / \bar{t}), \qquad (3)$$

де $\bar{t}$ – це середній час очікування ($\bar{t} \approx t_{\max} / V_{\max}$). Оскільки формула (3), які і всі наступні формули, означає континуальне наближення, строго кажучи, вона описує дискретні емпіричні дані лише для порівняно великих аргументів (на "хвості" розподілу), але не в області його "голови" (принаймні, не в діапазоні $\tau << \bar{t}$). Універсальна особливість усіх досліджених нами текстів NT – це помітна увігнутість функції $p(\tau)$ – тобто її відхилення від експоненційного розподілу (3), який у масштабі рис. 1, а передбачав би лінійну залежність. Це означає, передусім, що ймовірність малих і великих τ є вищою, аніж передбачає формула (3) (див. також [31, 32]). Дані рис. 1, *б* засвідчують, що причина цього не пов'язана з істотними шумами на хвості залежностей $p(\tau)$. А саме: графік кумулятивної функції $P(\tau)$ розподілу (3), яка визначається виразом

$$P(t) = \exp(-t / \bar{t}), \qquad (4)$$

повинен бути прямою лінією у масштабі $\lg P(t)$, що суперечить емпіричним даним рис. 1, б уже за умови фактичної відсутності шумів. Хоча в кількісному плані відхилення даних $\lg P(t)$ від лінійної залежності незначні, вони регулярні, а тому їх не можна звести лише до впливу похибок розрахунків або ефекту скінченних розмірів статистичної вибірки.

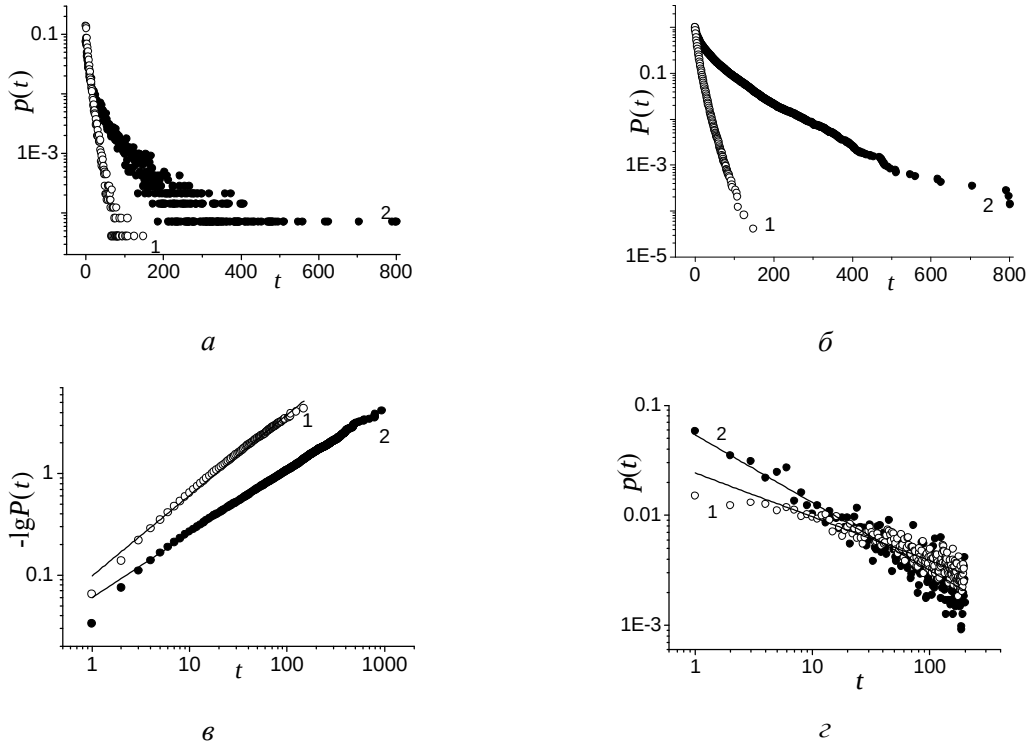

Рис. 1. Масові функції розподілу ймовірності p(τ) (а) і комплементарні кумулятивні функції розподілу ймовірності P(τ) (б) часів очікування першої появи слів для текстів NT3 (світлі кружки, 1) і NT4 (темні кружки, 2) у напівлогарифмічному масштабі; залежності –lgP(τ) (в) і гістограми функції густини розподілу ймовірності p(t) позицій першої появи слів (г) для тих самих текстів у подвійному логарифмічному масштабі. Прямі відповідають лінійній апроксимації емпіричних залежностей, а абсцисою на рис. 1, г є номер біна (n = 1÷200 – див. розділ 2)

Відхилення від експоненційного розподілу часів очікування означає порушення гіпотези про незалежний характер появи нових слів у тексті та скорельований (а не суто стохастичний) характер таких подій – т. зв. явище "спалахів" або "пульсацій" ("burstiness" – див., наприклад, праці [17, 26, 29, 30, 33]). За цих умов функцію $p(\tau)$ часто описують розподілом Вейбула

$$p(t) = a_b\, b\, t^{\,b-1} \exp[-(a_b t)^b], \tag{5}$$

а кумулятивну функцію $P(\tau)$ – частковим випадком "розтягнутого" експоненційного розподілу

$$P(t) = \exp[-(a_b t)^b], \tag{6}$$

де $0 < b \le 1$, $a_b = \bar{t}^{-1}\Gamma[(b+1)/b]^b$, $\Gamma(x)$ – гамма-функція. Зокрема, якщо $b = 1$, формули (5) і (6) зводяться до (3) і (4) (відсутність пульсацій; експоненційний розподіл), а якщо $b \to 0$, у формулі (5) маємо очевидний граничний перехід до степеневого розподілу зі степенем –1 (максимально можливі пульсації). Зазначимо, що в літературі розтягнутим експоненційним іноді називають розподіл, який описується функцією (6) для густини ймовірності $p(\tau)$ (див., наприклад, [17]). Щоб уникнути непорозумінь, користуємося термінологією [31, 34, 35], але не [17, 33].

Слабкість шумів емпіричної залежності $P(\tau)$ сприяє простій перевірці коректності її опису функцією Вейбула. Для цього в формулі (6) необхідно прологарифмувати $P(\tau)$ і побудувати графік $\lg(-\lg P)$ від $\lg\tau$ (див. [17, 34, 35]). Рис. 1, в засвідчує, що для обох текстів NT3 і NT4 ці графіки з високою точністю ($R^2 \approx 0{,}997$) є лінійними, тобто емпіричні розподіли справді описуються

вейбулівською функцією. Відхилення від лінійності на сильно розтягнутій у логарифмічному масштабі ділянці найменших τ ($\tau < 4$) на рис. 1, *в* природні з міркувань обмеженої застосовності континуального наближення і впливу синтаксису за цих умов (див. вище); зрозумілими є й причини деяких залишкових шумів на рис. 1, в за найбільших τ (порівн. із рис. 1, *б*). Графічна лінійна апроксимація, проілюстрована на рис. 1, в, дає змогу одержати показники степеня β для текстів NT. Порядок одержаних величин β (0,60÷0,85) добре узгоджується з даними [17] для часів очікування t' слів із низьким і середнім семантичним навантаженням. Загалом висновок про нерівність $b < 1$ для NT не викликає сумнівів принаймні з урахуванням такого самого результату, відомого для часів очікування t' між сусідніми появами того самого слова [17]. Серед можливих причин порушення рандомності часів очікування першої появи слів та появи кореляцій у тексті на думку спадають локальні зміни тематики тексту, розгортання його сюжетних ліній тощо. Нарешті, з рис. 2 видно, що параметр β майже лінійно зменшується зі зростанням довжини тексту $t_{\max}$. Відповідні причини розглянуто в підрозділі 3.5.

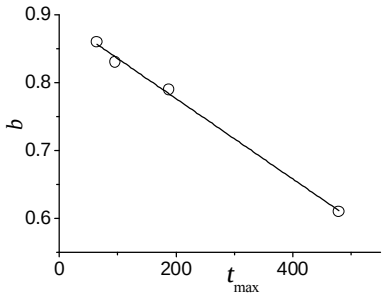

Рис. 2. Залежність показника степеня β, який описує емпіричні комплементарні кумулятивні функції розподілу ймовірності P(τ) часів очікування першої появи слів функцією Вейбула (див. формули (5) і (6)), від довжини t_{max} досліджених текстів NT, вираженої в тисячах слів. Пряма лінія відповідає лінійній апроксимації $\beta = A + Bt_{max}$ із $A \approx 0{,}9$ і $B \approx -6\times10^{-4}$

Попри дуже серйозні шуми, залежності $p(t)$ найближчі до прямої лінії в подвійному логарифмічному масштабі (див. рис. 1, *г*). Це означає, що серед усіх найпростіших альтернативних функцій (експоненційної, степеневої та логарифмічної) саме степенева функція найкраще описує емпіричну густину ймовірності $p(t)$ появи нових слів на тій чи іншій ділянці текстів NT3 і NT4:

$$p(t) = Ct^{-e}, \qquad (7)$$

де C і ε – сталі (наприклад, $\varepsilon \approx 0{,}4$ і 0,6 для NT3 і NT4, відповідно). Наслідки цього факту докладніше проаналізовано в підрозділі 3.4.

Отже, збіг теорії та експериментальних залежностей $P(\tau)$ достатній, аби стверджувати, що часи очікування появи нових слів у природних текстах описуються не експоненційним, а вейбулівським розподілом з "важчим" хвостом [*]. Це перший основний емпіричний результат цієї роботи. Він істотно уточнює, якщо не заперечує, висновок праці [13], згідно з яким формула (2) для словника природних текстів застосовна завдяки стохастичному розподілу нових слів у текстах, тобто експоненційному характерові залежності $p(t)$ (або $p(\tau)$). Проте жодна з них не є експоненційною; ці статистичні розподіли виявляють важкий (степеневий або вейбулівський) хвіст. Через принциповість цього результату ми виконали його перевірку на прикладі рандомізованих текстів і текстів Саймона. Ці додаткові дослідження потрібні також для встановлення причин нетривіальної поведінки набору часів очікування першої появи слів у текстах.

Рандомізовані тексти

Як засвідчують дані рис. 3 для тексту RT3, емпіричні залежності $p(\tau)$, $P(\tau)$ і $p(t)$ для текстів RT виявляються і кількісно, і якісно схожими до відповідних залежностей для NT. Незначною відмінністю є відсутність на залежності $p(\tau)$ для RT3 ділянки з пологим нахилом за найменших τ, оскільки будь-який вплив синтаксису в RT відсутній. З цієї ж причини відхилення емпіричної

[*] Термін "хвіст статистичного розподілу $p(x)$ випадкової змінної x" стандартно використовують на позначення ділянки функції розподілу за великих значень x, переважно для монотонно спадних функцій $p(x)$. Термін "важкий хвіст" означає, зокрема, аномально високу ймовірність екстремально великих значень x.

залежності lg(–lg$P$) від lgτ від прямої (див. рис. 3, *в*) на ділянці малих τ виявляються істотно меншими за відповідні відхилення для NT. В RT вони спричинені, мабуть, надзвичайно слабким ефектом дискретизації, а загалом якість лінійної апроксимації на рис. 3, в дуже висока ($R^2 > 0{,}999$).

Оцінені за методом графічної апроксимації степеня b для текстів RT1÷RT3 займають діапазон 0,79÷0,83, який з урахуванням похибок розрахунку майже збігається з відповідним діапазоном для текстів NT1÷NT3 (0,79÷0,86 – див. рис. 2). Наприклад, параметри b для текстів NT3 і RT3 практично однакові (0,79). Інакше кажучи, рандомізація практично не впливає на характер непуасонівської статистики часів очікування появи нових слів. Якісно схожою до даних для NT є й густина ймовірності $p(t)$ нових слів у текстах RT, яка теж наближено описується степеневою функцією (див. рис. 3, *г*). З урахуванням значних експериментальних шумів, близькими є й кількісні параметри залежностей $p(t)$ для NT і RT: наприклад, нахил ε для RT3 становить 0,5, що не надто відрізняється від значення 0,4 для NT3 (див. підрозділ 3.1).

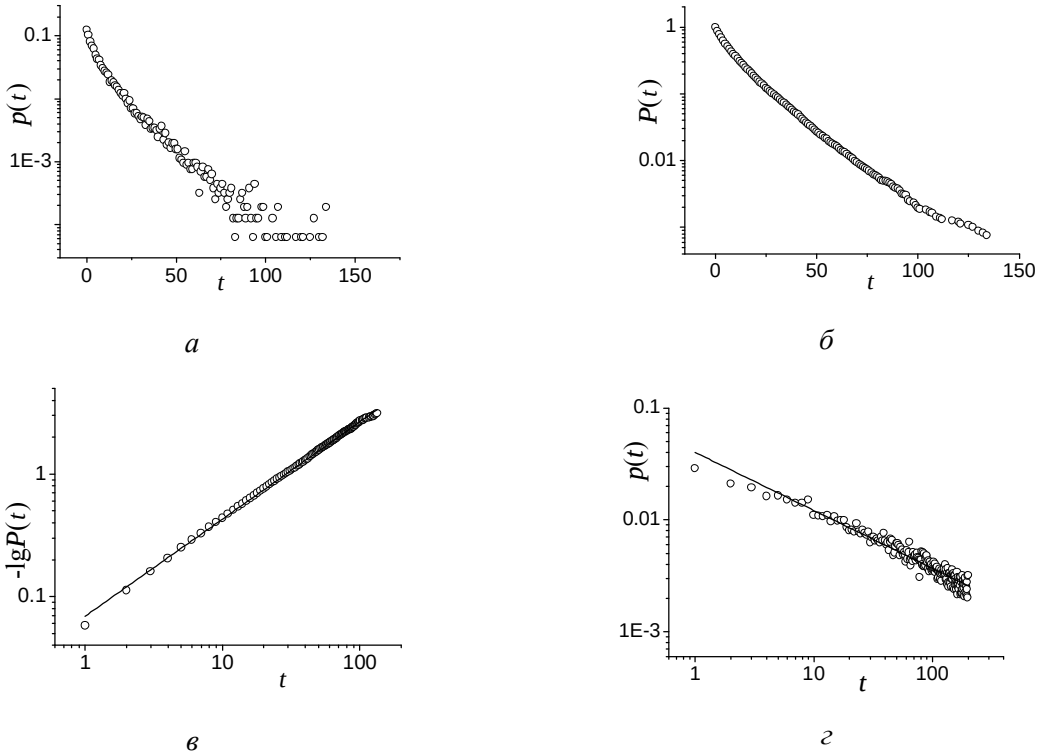

Рис. 3. Масова функція розподілу ймовірності p(τ) (а) і комплементарна кумулятивна функція розподілу ймовірності P(τ) (б) часів очікування першої появи слів для тексту RT3 у напівлогарифмічному масштабі; залежність –lgP(τ) (в) і гістограма функції густини розподілу ймовірності p(t) позицій першої появи слів (г) для того ж тексту в подвійному логарифмічному масштабі. Прямі відповідають лінійній апроксимації емпіричних залежностей, а абсцисою на рис. 3, г є номер біна (n = 1÷200)

Отже, поява нових слів у текстах RT також скорельована і виявляє пульсації. Це другий важливий висновок цієї праці. З огляду на парадоксальний наслідок вейбулівської статистики інтервалів t – наявність у рандомних текстах RT деяких кореляцій – ми виконали додаткову перевірку роботи нашого рандомізатора вихідних текстів NT. Метод перевірки ґрунтувався на порівнянні для текстів NT і RT параметра $R = \Delta t'/\bar{t}'$, де $\Delta t'$ – середньоквадратичне відхилення часів очікування t' від середнього значення $\bar{t}'$ для низки слів із різною семантикою. Відомо, що для т. зв. функціональних слів у природних текстах маємо $R \approx 1$ (тобто пуасонівську статистику інтервалів t'), а для важливих змістових або ключових слів – випадки $R > 1$ або й $R >> 1$ (тобто істотні відхилення від стохастичного розподілу слів – їхню кластеризацію або явище пульсацій) [17, 18, 26–30].

Перевірка засвідчила, що процес рандомізації NT надійно знищує пульсації у розподілах $p(t')$ фактично для всіх слів, так що в RT на статистично значущому рівні вони відсутні.

Наприклад, параметр R для змістового слова “her” (частота появи в тексті $F = 1077$) і ключового слова “Swithin” (частота $F = 180$) у тексті NT2 набуває дуже великих значень (відповідно $R \approx 2{,}97$ і 11,07), проте в тексті RT2 вони фактично одиничні (відповідно $R \approx 1{,}00$ і 0,97). Щоправда, в тексті RT2 є низка слів із $R = 1{,}3 \div 1{,}8$, але частоти цих сліва вкрай низькі ($F \le 10$), що знецінює надійність відповідної статистики. Отже, вжиті нами практичні алгоритми “перемішування” слів у RT таки коректні – жодних кореляцій (довгосяжних чи короткосяжних) у розподілі окремих слів у цих текстах немає. Водночас питання про причини кореляцій у розподілі часів очікування перших появ слів залишається відкритим (див. підрозділ 3.4).

Тексти Саймона

Приклади статистики першої появи слів для тексту Саймона ST4 наведено на рис. 4. Найголовніша риса текстів ST – це практична незалежність густини ймовірності $p(t)$ нових слів від позиції t у тексті. Так, для тексту ST4 кутовий нахил лінійної залежності $p(t)$ менший за $4 \cdot 10^{-3}$ (рис. 4, в). У термінах формули (7) для ST маємо $e \approx 0$. Причиною, очевидно, є сам алгоритм генерування ST, за яким імовірність появи нового слова, що визначається параметром p_0, є незмінною.

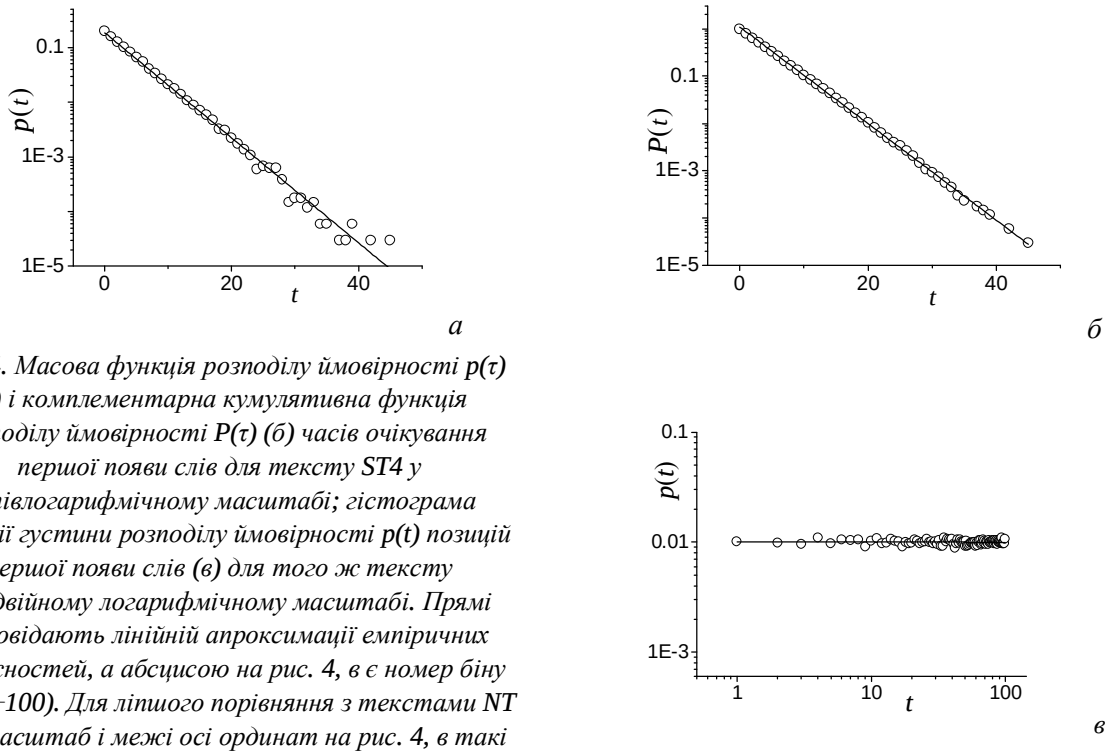

Рис. 4. Масова функція розподілу ймовірності p(τ) (а) і комплементарна кумулятивна функція розподілу ймовірності P(τ) (б) часів очікування першої появи слів для тексту ST4 у напівлогарифмічному масштабі; гістограма функції густини розподілу ймовірності p(t) позицій першої появи слів (в) для того ж тексту в подвійному логарифмічному масштабі. Прямі відповідають лінійній апроксимації емпіричних залежностей, а абсцисою на рис. 4, в є номер біну (n = 1÷100). Для ліпшого порівняння з текстами NT і RT масштаб і межі осі ординат на рис. 4, в такі самі, як на рис. 1, г і рис. 3, г

Зазначимо, що спільною особливістю всіх згенерованих саймонівських текстів є порівняно малі найбільші значення τ (наприклад, для ST4 маємо $t_{\max} = 45$). Це пов’язано з вищою, порівняно з текстами NT, імовірністю появи нових слів. Незважаючи на слабкі шуми в емпіричній залежності $p(\tau)$ (див. рис. 4, *а*), регулярних відхилень у ній від пуасонівської статистики (3) не спостерігаємо. Це підтверджує й кумулятивний розподіл $P(\tau)$, який з надзвичайно високою точністю ($R^2 > 0{,}999$) описується експоненційною функцією ($b = 1$). Відповідно, для ST нема потреби в додатковій графічній апроксимації $P(\tau)$ функцією Вейбула загального виду ($b \ne 1$), за аналогією до рис. 1, *в* і рис. 3, *в*. Зазначимо, що статистика часів очікування t' повторень однакових слів для ST також пуасонівська [24].

Отже, на відміну від текстів NT і RT, статистика часів очікування першої появи слів у текстах ST строго пуасонівська та не виявляє явища пульсацій. Це третій основний емпіричний результат цієї праці. Сам факт відсутності кореляцій у суто рандомних за способом побудови текстах ST певною мірою природний, навіть тривіальний, проте незрозумілими залишаються причини

докорінно різної поведінки статистики випадкової змінної t для текстів RT (а також NT), з одного боку, і ST, з іншого боку. Це питання пояснено в наступному підрозділі.

Математична модель і аналіз даних

Факт відхилення емпіричних розподілів імовірності $p(\tau)$ для текстів NT і RT від експоненційної функції можна було би пояснювати порушенням припущення про зникомо малу ймовірність появи нових слів для вивчених текстів. Саме це припущення обґрунтовує розподіл Пуасона для частоти появи нових слів і розподіл (3) для часів їхнього очікування. Справді, усереднена по тексту ймовірність $\overline{p}_0$ появи нових слів, яку визначають за середнім часом очікування $\overline{t} \approx t_{\max}/V_{\max}$ як $\overline{p}_0 = \overline{t}^{-1}$, помітно відхиляється від нуля (~0,03 ÷ 0,08 – див. підрозділ 3.2) для досліджених текстів NT і RT. За цих умов і за припущення про стохастичний характер появи нових слів частота випадання цих слів у тексті повинна би описуватися процесом Бернуллі, а емпіричні дані $p(\tau)$ – геометричним розподілом (див. [27, 28]), дещо відмінним від експоненційного. Хоча відповідні уточнення опису формулою (3) справді актуальні, таке трактування неекспоненційності наших даних $p(\tau)$ для текстів NT і RT усе ж не є правильним. Ключовими тут є дані для текстів ST, параметри $\overline{t}$ для яких навіть менші, ніж для NT і RT, а відповідна імовірність $\overline{p}_0$ – вища ($\overline{p}_0 = p_0 = 0{,}1$ – див. рис. 4, в і підрозділи 1 і 2). Проте відмінності геометричного та експоненційного розподілів для ST, очевидно, настільки малі, що навіть не помітні в масштабі рис. 4, б. Отже, причини вейбулівської залежності $p(\tau)$ для NT і RT не пов'язані з порушенням континуального наближення для дискретної випадкової величини.

Встановимо зв'язок ймовірності $p(t)$ появи нових слів у тексті й залежності словника $V(t)$ від розмірів тексту. Із загальних міркувань очевидно, що

$$V(t) = V_{\max} \int_0^t p(t')\,\mathrm{d}\,t' = V_{\max}[1 - P(t)]\,. \tag{8}$$

Припускаючи в першому наближенні найпростішу степеневу форму (5) для функції $p(t)$, із формули (8) одержимо залежність словника, а із формули (1) – тривіальний зв'язок параметрів ε і θ:

$$V(t) = C_0 t^{1-e}\,, \quad \theta = 1 - \varepsilon, \tag{9}$$

де C_0 – константа. На рис. 5 подано залежності словника $V(t)$ у відносних одиницях $V/V_{\max}$ від $t/t_{\max}$, розраховані за формулою (8) й емпіричними даними функції $P(t)$. Із рис. 5, в видно, що для тексту ST4 з високою точністю виконуються рівності $\theta = 1$ ($\theta \approx 1{,}00$, $R^2 > 0{,}9999$) і $e = 0$, тобто словник зростає лінійно з t; останній факт збігається з висновками теорії [25, 36] (із літератури відомі також спроби модифікувати модель Саймона з метою відтворення "природнішої" поведінки текстів NT, для яких $\theta < 1$ [24, 25]). Водночас, для текстів NT і RT нахил θ менший за одиницю (див. рис. 5, *а, б*); помітною є й опуклість кривих $V(t)$ у подвійному логарифмічному масштабі (див. також [13]) [**].

Для найбільшого тексту NT4 за $t_0 \approx 0{,}04 t_{\max}$ і $V_0 \approx 0{,}20 V_{\max}$ (або за $V_0 \approx 2{,}9$ тис. слів, повний словник $V_{\max} \approx 14$ тис. слів) додатково спостерігаємо явище "кросоверу" – перехід між двома степеневими режимами з різними θ, тобто злам на прямій $\lg V(\lg t)$ (див. рис. 5, *а*). Докладне обговорення кросоверу виходить за межі цього дослідження. Зазначимо лише, що схема його пояснення ґрунтується на понятті "ядра лексикону" V_0. На основі аналізу великих корпусів текстів (а не єдиного тексту, як у нас) автори праць [5, 37] також спостерігали явище кросоверу, проте за дещо більших V_0 ($V_0 \sim 5 \div 6$ тис. слів [37] або $V_0 \sim 8$ тис. слів [5]). Зокрема, в праці [5] виявлено лінійне зростання ($\theta \approx 1$) словника за малих t, яке змінюється степеневою ділянкою із $\theta \approx 0{,}56$; для нашого тексту NT4 на початковій ділянці t одержано $\theta \approx 0{,}96$, а за великих t – $\theta \approx 0{,}48$. Для

[**] Явища опуклості функцій $V(t)$ для NT і RT, а також "кросоверу" $V(t)$ для NT (див. нижче) повинні би виявлятися на рис. 1, в і рис. 3, в як відхилення функції густини ймовірності $p(t)$ від степеневого закону, проте через значні шуми про це важко робити висновки.

коротшого тексту NT3 явище кросоверу не виражене – нахил залежності lg$V$(lg$t$) тут більш-менш плавно зменшується з переходом до більших t. Якщо грубо взяти як координату кросоверу ту саму точку $V_0 \sim 0{,}2V_{\max}$, то одержуємо приблизно таке саме "ядро лексикону" $V_0 \sim 2{,}9$ тис. слів, що і для тексту NT4 (словники NT3 і NT4 майже однакові). Для тексту NT3 маємо $\theta \approx 0{,}98$ і 0,66 відповідно для малих і великих t. Для тексту RT3 за малих t знаходимо $\theta \approx 0{,}97$, а за великих – $\theta \approx 0{,}58$, тобто вплив рандомізації на характеристики опуклості функції lg$V$(lg$t$) і явище кросоверу сумнівний. Нарешті, з урахуванням нашого аналізу природно припустити, що саме перераховані вище відмінності поведінки словника текстів ST ($\theta = 1$ і відсутність опуклості або кросоверу) і NT, RT ($\theta < 1$ й опуклість або й кросовер функції $V(t)$) зумовлюють принципово різну поведінку статистики появи нових слів у них.

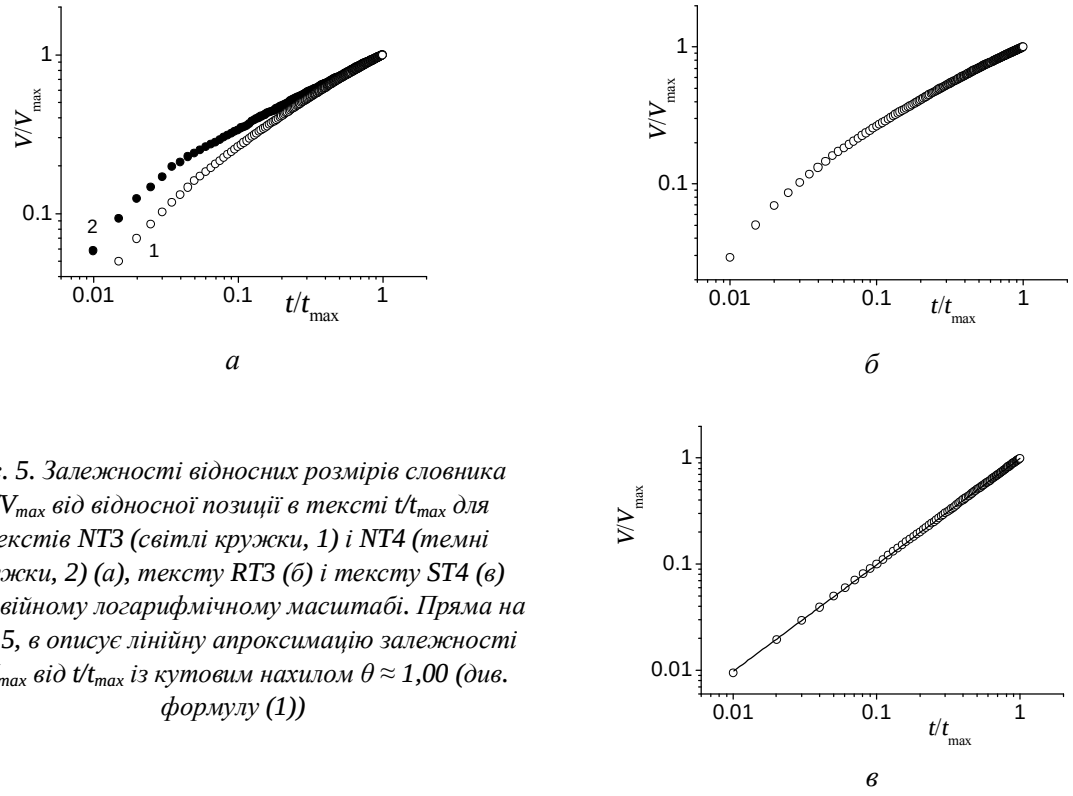

Рис. 5. Залежності відносних розмірів словника V/V_{max} від відносної позиції в тексті t/t_{max} для текстів NT3 (світлі кружки, 1) і NT4 (темні кружки, 2) (а), тексту RT3 (б) і тексту ST4 (в) у подвійному логарифмічному масштабі. Пряма на рис. 5, в описує лінійну апроксимацію залежності V/V_{max} від t/t_{max} із кутовим нахилом $\theta \approx 1{,}00$ (див. формулу (1))

Для пояснення емпіричних даних використаємо генеративну модель стаціонарного стохастичного процесу з пам'яттю [38], застосовану раніше до пояснення вейбулівської статистики часів очікування t' повторень однакових слів [17]. Припустимо, що ймовірність $p(t)$ появи нового слова в тексті не постійна, але залежить від часового проміжку після вживання попереднього нового слова t за степеневим законом:

$$p_b(t) = a_b b t^{b-1} \quad (0 < \beta \le 1), \tag{10}$$

де, згідно з формулою (7), маємо $C = a_b b$ і $\varepsilon = 1 - \beta$. Це припущення означає модель резервуару слів із пам'яттю, за якою поява деякого нового слова в тексті стимулює вживання наступних нових слів, а ймовірність цього процесу зменшується з віддаленням від нового слова. Тоді можна довести, що ймовірність $p(\tau)$ і кумулятивна ймовірність $P(\tau)$ часу очікування τ визначаються формулами, аналогічними до відповідних формул [17, 38]:

$$p(t) \approx p_b(t) \exp\left(-\int_0^t p_b(t)dt\right), \; P(t) \approx \exp\left(-\int_0^t p_b(t)dt\right). \tag{11}$$

Нарешті, інтегрування формули (10) (або (7)) згідно з (8) дає степеневу залежність (1) або (9) для словника $V(t)$, так що показник степеня θ закону Гіпса і показник β вейбулівської функції із формул (5) і (6) мають бути однаковими:

$$\theta = 1 - \varepsilon = \beta. \tag{12}$$

Якщо тепер $p_b(t) = t^{-1} = \text{const}$ ($\beta = 1$), то з (11) одержуємо статистику $\{\tau\}$, яка описується формулами (3) і (4), а згідно з (8) або (12) словник $V(t)$ тоді зростає лінійно. Інакше, за відсутності пам'яті про попередні вживання нових слів ($p(t) = \text{const}$), статистика $p(t)$ першої появи цих слів пуасонівська і виконується закон Гіпса із $\theta = 1$. Якщо ж система "пам'ятає" про попередні перші появи слів за степеневим законом ($p(t) \neq \text{const}$, $\beta < 1$), то статистика часів очікування нових слів вейбулівська [17, 38], а степінь у законі Гіпса $\theta < 1$. Отже, рівносильними є такі твердження:

{імовірність $p(\tau)$ часів очікування першої появи слів експоненційна / вейбулівська};

{густина ймовірності $p(t)$ появи нових слів постійна / степенева}; (13)

{словник $V(t)$ зростає за степеневим законом Гіпса із $\theta = 1$ / $\theta < 1$}.

Наведена вище математична модель вичерпно пояснює всі емпіричні дані цієї праці для текстів ST (див. рис. 3, 4 і 5, *в*) за умови $\beta = \theta = 1$. Нагадаємо, що ці тексти генерують так, що "швидкість" появи нових слів постійна (p_0), а зростання словника лінійне:

$$V(t) = p_0 t, \quad \mathrm{d}V/\mathrm{d}t = p_0. \tag{14}$$

З літератури відома модифікована модель Саймона [24, 25], в якій загалом маємо $\theta \leq 1$ і

$$V(t) = p_0 t^q, \quad \mathrm{d}V/\mathrm{d}t = p(t) = p_0 q t^{q-1} \quad (0 < \theta \leq 1), \tag{15}$$

де порівняння з формулою (10) дає рівність $p_0 = a_b$. Тут "швидкість" появи нових слів у тексті сповільнюється зі зростанням часу t, а степінь у законі Гіпса стає меншим ($\theta < 1$). Фактично автори [25] вводили модифіковану модель Саймона з доволі скромною метою – забезпечити характерну ознаку багатьох природних текстів $\gamma < 2$. Із нашого ж аналізу випливає, що така модифікація зі степеневим зменшенням функції $p(t)$ набагато принциповіша. Оскільки закон Гіпса із сублінійним зростанням словника (15) еквівалентний розподілу часів очікування $p(\tau)$ з важким хвостом, ми передбачаємо, що за динамічними властивостями модифікована модель Саймона повинна виявитися істотно ближчою до природних текстів і, зокрема, зніме критичні зауваження, які висловили автори [1, 7] на адресу ST. Цікаво також, що були спроби [25, 36] ввести в модель Саймона локальну пам'ять, що загасає, про попередньо вжиті в тексті "старі" слова. На нашу думку, вплив цих модифікацій на відповідність природним текстам буде слабшим, аніж базова модифікація (15).

Таблиця 2

Параметри генеративної моделі для деяких досліджених текстів, розраховані на підставі емпіричних даних

Текст	Значення параметра β, знайдене із залежності lg[–lgP(τ)] від lgτ	Значення параметра ε = 1 – β = 1 – θ, знайдене із залежності p(t)	Значення параметра β, знайдене із залежності p(t)	Значення параметра θ = β, знайдені із залежності V(t) *
NT3	0,79	0,4	0,6	0,98; 0,66; 0,82; 0,73
RT3	0,79	0,5	0,5	0,97; 0,58; 0,78; 0,64
NT4	0,61	0,6	0,4	0,96; 0,48; 0,72; 0,51

** Чотири цифри $\theta_1 \div \theta_4$ в останній колонці відповідають значенням θ_1 і θ_2, знайденим лінійною апроксимацією даних рис. 5, а, б відповідно на ділянках малих і великих t, їхньому незваженому середньому θ_3, а також значенню θ_4, знайденому лінійною апроксимацією в усьому діапазоні t.*

Запропонована математична модель якісно описує і тексти NT і RT, хоча її кількісна відповідність цим текстам дещо неповна. Загалом дані рис. 1, *в* і рис. 3, *в* для NT і RT не заперечують можливості опису залежностей $p(\tau)$ і $P(\tau)$ виразами (5) і (6) із $\beta < 1$. З іншого боку, степенева форма (10) залежності $p(t)$ для цих текстів є хіба наближенням (див. рис. 1, *г* і рис. 3, *г*), надійність якого важко перевірити через значні шуми. Це підтверджує порівняння значень параметрів β, ε і θ, розрахованих відповідно до даних рис. 1, *в*, 3, *в*, рис. 1, *г*, 3, *г* і рис. 5, *а, б*. Як видно з табл. 2, збіг значень швидше якісний, аніж кількісний. На додаток, очевидними є

відхилення даних рис. 5, *а, б* від степеневих формул (1) або (9) для словника $V(t)$, що добре підтверджує обмеженість математичної моделі в описі текстів NT і RT (див. твердження (13)). Строго кажучи, опуклість функції $V(t)$ для текстів NT і RT свідчить на користь того, що статистика $p(\tau)$ часів очікування першої появи може описуватися функцією, складнішою за вейбулівську.

Отже, формули (5), (6), (8), (10) і (11) вичерпно пояснюють усі наші результати для текстів ST і наближено, якісно пояснюють дані для текстів NT і RT.

Обговорення результатів

З урахуванням одержаних результатів насамперед критично проаналізуємо методичну та емпіричну бази висновку [13] щодо рандомного (експоненційного) характеру появи нових слів у природних текстах. Наведемо такі основні зауваження:

1) викладення матеріалу в праці [13] залишає сумніви стосовно того, статистику якого саме параметра досліджували автори: $\{\tau\}$ чи $\{t'_1 \equiv t\}$ (за нашими даними, перша з них наближено вейбулівська, а друга – наближено степенева);

2) експоненційну ($b = 1$) і вейбулівську ($b < 1$) функції $p(\tau)$ важко розрізнити на практиці, особливо за наявності шумів (див. рис. 1, а і рис. 2, а);

3) автори [13] залишили поза аналізом слова з найнижчою ймовірністю (найвищою частотою F) на зразок слів “the” або “of” (вони мають найбагатшу статистику інтервалів $\{t'\}$, але дають найменший внесок до словника і статистики $\{\tau\}$ і $\{t\}$); що важливіше, не аналізувалися й найчисленніші слова з найнижчою частотою $F = 1$, хоча саме вони домінують у статистиці $\{\tau\}$ і $\{t\}$ (слова із $F = 1$, які ще називають “hapax legomena” – це до 50 % усього словника!);

4) автори [13] не описали методів розрахунку статистики $\{t'_1\}$ (зокрема бінування), називаючи залежність $p(t'_1)$ “масовою функцією ймовірності”, хоча значний діапазон величини $t'_1 = 1 \div t_{\max}$ робить її фактично неперервною;

5) графіки $p(\tau)$ і $p(t)$ подано в подвійному логарифмічному масштабі, що утруднює наочне порівняння з експоненційною функцією;

6) ні густина (або масова функція) ймовірності, ні кумулятивна ймовірність випадкової змінної $\{t\}$ (або $\{\tau\}$) не можуть сягати величин $10^1 \div 10^2$ при $t \to 0$ (або $t \to 0$), що спостерігаємо на рис. 3 із праці [13]; якщо ж ідеться про абсолютні частоти спостереження N_t (або N_t – див. пояснення в розділі 2), то формулу (3) у разі $\bar{t} = 1$ [13] заміняє вираз $N_t = N\exp(-t)$, який теж не дорівнює одиниці за умови $t \to 0$ (порівн. з даними рис. 3 у праці [13]);

7) усереднення статистики часів очікування для багатьох природних текстах, ужите в [13], має низку недоліків, оскільки за умови залежності параметра b від тексту (зокрема, від його розмірів $t_{\max}$) воно може призвести до додаткових шумів, на фоні яких експоненційну та вейбулівську статистику важче відрізнити.

Як висновок, ми схиляємося до думки про ненадійність даних [13] щодо експоненційної статистики часів очікування першої появи слів у природних текстах.

Незалежно від того, чи одержана у нашій праці експериментально статистика часів очікування нових слів у текстах NT і RT вейбулівська, чи ще складніша, вона в жодному разі не є пуасонівською і відповідає описаному вище процесові генерування тексту з пам’яттю про попередні слова. Для статистики часів очікування t' повторень однакових слів у природних текстах автори [17] припускали, що вживання деякого слова стимулює його повторні вживання, а далі ймовірність цих вживань поступово зменшується – функція $p(t')$ локально загасає. Хоча якісно схожу картину можна уявити і для появи нових слів у тексті, фактичне припущення [17] про роль “локального контексту” й аргументи математичної психології та науки про людську пам’ять стають сумнівними, беручи до уваги одержані нами результати про практично однакову статистику часів очікування нових слів для текстів NT і RT. Фактично, інваріантність явища степеневого загасання ймовірності $p(t)$ появи нових слів щодо рандомізації доводить, що це явище не зумовлене лінгвістичними причинами. По-друге, на відміну від припущення [17] про локальне загасання

імовірності вживання того чи іншого слова в тексті (кусково-заданої функції $p(t')$), у нас імовірність $p(t)$ появи нових слів загасає глобально.

З одного боку, практично однакові статистика появи нових слів і динаміка словника NT і RT добре зрозумілі зі встановлених нами взаємозв'язків поміж залежностями $p(t)$, $p(t)$ і $V(t)$: оскільки повні словники NT і його рандомізованої версії RT тотожні, то природно припустити й однакову динаміку зростання словника $V(t)$ і, відповідно, однакові функції $p(t)$ і $p(t)$. Наші емпіричні дані підтримують цю гіпотезу. Отже, ми фактично довели один з основних результатів авторів [13] – їхнє припущення про близькість або й тотожність динаміки словника $V(t)$ для NT і RT, яке обґрунтовує застосовність формули (2) до опису словника NT. Зокрема, це означає інваріантність функцій $p(t)$ і $p(t)$ стосовно процедури рандомізації тексту. Проте, на відміну від даних [13], ця спільна динаміка виявляється не пуасонівською, а має важкий хвіст.

З іншого боку, за нашими даними, в літературі досі не було даних про порушення пуасонівської статистики для лінгвістичних елементів у рандомних чи рандомізованих текстах будь-яких типів. Відомо, що рандомізація завжди знищує будь-які кореляції в символьних часових рядах, а вихідні розподіли відповідних часів очікування з важким хвостом стають експоненційними (див., наприклад, [24, 39]). Так, у дослідженні [24] було виявлено експоненційний розподіл часів очікування t' повторення однакових слів у рандомних ST, навіть у модифікованій моделі Саймона зі спадною ймовірністю $p(t)$ появи нових слів (15). Із теорії, наведеної в підрозділі 3.4, випливає, що за цих умов для перших поява слів слід все-таки очікувати вейбулівської статистики $\{\tau\}$, як і для RT. Наведені міркування визначають наукову новизну наших результатів для часів очікування першої появи слів у текстах RT. Чи не єдиним відомим нам результатом стосовно нетривіальної поведінки RT є по суті не обговорений і не пояснений результат праці [40] про одиничний показник степеня в законі Тейлора, який притаманний текстам NT.

Як уже згадано вище, нетривіальність статистичного розподілу із важким хвостом для часів очікування деяких елементів полягає в тому, що він еквівалентний скорельованому характерові часової послідовності цих елементів (у нашому випадку – нових слів у тексті) [17, 18, 31, 32, 34, 35, 39, 41]. Наприклад, якщо статистичний розподіл $p(t)$ описується якісно схожою до вейбулівської функції розтягнутою експонентою (5), то відповідний параметр $\beta < 1$ безпосередньо дорівнює показнику δ степеневої залежності вихідної кореляційної функції [31, 34, 35, 41]; у разі степеневої залежності $p(t)$ зв'язок β і δ теж існує, але він дещо інший [32]. Такі кореляції повинні бути додатними – інакше хвіст функції $p(t)$ був би легким ($\beta > 1$) [41]. Тісно пов'язане з кореляціями і явище пульсацій. Інтуїтивно визначене в праці [33] як відхилення часової послідовності елементів від суто стохастичної (рандомної), воно має два незалежні вияви (див. також [18]): 1) кореляції в послідовності цих елементів, яка описується функцією $p(t)$ (тобто пам'ять) і 2) важкий хвіст розподілу $p(t)$. У нас наявні обидва вияви пульсацій – і пам'ять, що полягає в "згущенні" нових слів на початкових ділянках тексту, і важкий вейбулівський хвіст. Тому послідовність часових позицій нових слів у текстах NT і RT відповідає суперпозиції символьних послідовностей на рис. 1, b і рис. 1, e із праці [33].

Звісно, що поняття "пам'яті" в послідовності нових слів у тексті доволі специфічне, як і саме поняття "нове слово", дефініція якого означає врахування "передісторії" тексту. У цьому плані корисна така фізична аналогія: хоча швидкість світла у вакуумі гранична, це все ж не забороняє рух об'єктів, які не переносять інформації, з надсвітловою швидкістю. "Нові слова" можна вважати аналогами цих об'єктів і приписувати їм нульове інформаційне навантаження. Хоча для послідовності будь-яких конкретних слів, як "матеріальних об'єктів", кореляції та пам'ять у RT справді відсутні, це не унеможливлює їхню наявність для послідовності нових слів. Так вирішується парадокс "кореляцій" і "пам'яті" в загалом рандомній послідовності елементів, якою є RT. Насправді йдеться лише про те, що на початку будь-якого тексту (NT або RT) нових слів більше, ніж в кінці [***]. Згідно з (15), це по суті переформулювання сублінійного зростання словника.

[***] Це явище, хоча й видається "інтуїтивно" зрозумілим, насправді потребує додаткових пояснень.

Повністю узгоджено з теорією, кореляції між першими появами слів у тексті справді додатні: зменшення часу очікування t зі зростанням t приводить до подальшого зменшення t. Нарешті, наші результати ставлять під сумнів універсальність висновку [18, 33] про "ортогональність" пам'яті в символьній послідовності та важкого хвоста відповідних часів очікування: принаймні в межах механізму пульсацій, підсумованого виразами (8), (10), (11) і (13), кореляції відстаней між новими словами, які визначаються функцією $p(t)$, і важкий хвіст $p(t)$ зникають одночасно за умови $\beta \to 1$.

Коротко обговоримо залежність степеня β від довжини тексту $t_{\max}$ (див. рис. 2). Оскільки $\beta = \theta$, це означає залежність сталої Гіпса $\theta(t_{\max})$. У недавній праці [42] на велетенському корпусі текстів було виявлено, що степінь γ другого закону Ціпфа незначно зменшується зі зростанням розмірів текстів. За умови $g \le 2$ і зв'язку $g = q + 1$ (див. розділ 1) це зводиться до зменшення θ, що якісно узгоджується з нашими даними $\theta(t_{\max})$. На нашу думку, причиною явища може бути внутрішньотекстова залежність "ефективного" степеня $\theta(t)$, знайденого як локальна похідна функції $\lg V(\lg t)$. За нашими спостереженнями для багатьох природних текстів, зміни $\theta(t)$ є слабкими для коротших текстів, а для довших текстів значення θ істотно зменшуються на ділянках великих t (див. останню колонку табл. 2 і дані [1, 6, 7]). Ці фактори можуть привести до видимого зменшення функції $\theta(t_{\max})$. Оскільки ми спостерігаємо явище і для RT, таке пояснення видається нам надійнішим, аніж аргументація [42] стосовно комбінованого впливу залежності θ від жанру тексту і жанру від його довжини $t_{\max}$. Щоправда, за даними [42] залежність $\gamma(t_{\max})$ дуже слабка і відсутня для синтетичних текстів (порівн. із даними підрозділу 3.2), а автори іншої праці [8] роблять протилежний висновок про слабке асимптотичне зростання функції $q(t_{\max})$. Тому явище потребує додаткового вивчення.

Оскільки степеневий закон Гіпса (1) для NT і RT правильний принаймні в деякому наближенні, залежність $p(t)$ для позицій першої появи слів принципово не може бути експоненційною – це повинна бути функція з істотно товстішим хвостом. І навіть більше, якби $p(t)$ була експоненційною, то словник, який визначається кумулятивною функцією ймовірності $P(t)$, також описувався би експонентою:

$$V = V_{\max}[1 - P(t)] = V_{\max}[1 - \exp(-t/\bar{t}\,)] \qquad (\bar{t} = \text{const}). \tag{16}$$

Хоча формула (16) прямо суперечить усім емпіричним результатам для слів у текстах, зростання словника згідно з (16), що є аналогом швидкого "перехідного процесу" в електричному колі, таки описує реальні дані. Це закон зростання обмеженого "словника" букв у природних текстах [43].

Наші дані не заперечують, що динаміка словника $V(t)$ для NT описується виразом (2) [13] або може виявитися навіть складнішою за формулу (10) із праці [13]. Нагадаємо, що автори [13] виводили формулу (2) у припущенні строго степеневого (ціпфівського) характеру частотної залежності для всіх слів в NT (див. підрозділ 3.1). Порушення цього припущення, яке спостерігають для багатьох реальних NT, може ускладнити залежність $V(t)$, порівняно з формулою (2). Швидше за все, коректна залежність $V(t)$ таки повинна мати загальну форму $V/V_{\max} = 1 - P(t)$ на зразок формули (8) (див. [8, 12, 13]). Тоді статистика $p(\tau)$ і $p(t)$ набуде ще складнішої форми, аніж формули (5) і (10). Якщо формула (2) таки є достатнім наближенням, то уточнений вираз $p(t)$ можна знайти із (8) диференціюванням функції дискретної змінної (2), а відповідний вираз $p(\tau)$ – інтегруванням у формулі (11).

Наприкінці торкнемося питання трактування законів Дж. К. Ціпфа та Г. С. Гіпса. Протягом більш ніж півстоліття в літературі триває дискусія щодо походження законів статистичної лінгвістики (див., наприклад, [1, 25, 44]). Історично першими альтернативами були гіпотези Б. Мандельброта і Г. А. Саймона про закони Ціпфа відповідно як тривіальний наслідок статистики рандомних послідовностей символів або як вияв деякого лінгвістичного змісту. Наші результати, згідно з якими природні та рандомізовані тексти NT і RT не відрізняються одні від одних за всіма статичними закономірностями статистичної лінгвістики, а також схожі за динамікою появи нових слів і законом зростання словника, схиляють до висновку про те, що закони Ціпфа та Гіпса (разом з його модифікацією (2)) навряд чи є наслідками якихось глибоких лінгвістичних і семантичних

закономірностей або особливостей людської мови [44], а швидше випливають із (можливо, ускладнених) нульової стохастичної гіпотези для тексту або моделі резервуару слів [1].

Висновки

Отже, в цій праці досліджено статистичні розподіли $p(t)$ і $p(\tau)$ позицій і часів очікування першої появи слів у природних, рандомізованих і саймонівських текстах. Показано, що статистика $\{\tau\}$ і $\{t\}$ для текстів NT і RT має відповідно вейбулівський і наближено степеневий характер, тоді як для текстів ST функція $p(\tau)$ експоненційна, а $p(t) = \text{const}$. Наявність важкого хвоста в розподілі $p(\tau)$ для NT і, особливо, для RT є принципово новим емпіричним фактом, який засвідчує наявність явища пульсацій і довгосяжних кореляцій, пов'язаних із появами нових слів у тексті. Це чи не перша експериментальна демонстрація скорельованої поведінки, виявленої в рандомній часовій послідовності лінгвістичних знаків. Докладно проаналізовано методичні недоліки дослідження [13] і показано, що відповідний висновок [13] про пуасоніську статистику часів очікування першої появи слів ненадійний. Доведено також, що відхилення від пуасонівської статистики інтервалів τ не можна звести до наслідків некоректного використання континуального наближення в описі дискретної випадкової змінної. Виявлено і пояснено наближено лінійну залежність параметра β вейбулівської функції від розмірів вивчених текстів t_{max}.

Фактична інваріантність залежностей $p(\tau)$ і $p(t)$ під дією рандомізації доводить застосовність раніше запропонованої в літературі [13] формули (2) для зростання словника текстів навіть без дотримання додаткової гіпотези [13] про пуасонівську статистику $\{\tau\}$ в NT. Порівняння розрахованої за емпіричними залежностями $p(t)$ динаміки зростання словника $V(t)$ для текстів різних типів показує, що для ST з високою точністю виконується степеневий закон Гіпса, а для NT і RT спостерігаємо помітні відхилення від нього – опуклість у координатах $\lg V(\lg t)$. Уперше виявлено кросовер на залежності $V(t)$, одержаній не для корпусу текстів, а для єдиного порівняно великого за розмірами природного тексту NT; це явище потребує докладнішого вивчення. На додаток показано, що рандомізація тексту майже не впливає на динаміку словника.

Для кількісного пояснення емпіричних результатів застосовано математичну модель стаціонарного стохастичного процесу з пам'яттю, запропоновану раніше для опису статистики часів очікування повторень однакових слів [17]. Модель дає змогу встановити співвідношення між усіма параметрами розподілів $p(\tau)$, $p(t)$ і $V(t)$. Зокрема, параметр β вейбулівської залежності $p(\tau)$ дорівнює степеню θ у формулі Гіпса для словника $V(t)$. Запропонований підхід вичерпно пояснює всі закономірності динаміки появи нових слів у найпростіших за принципом побудови текстах ST.

Згадана модель лише частково пояснює властивості складніших текстів NT і RT. Найбільша проблема полягає в помітному відхиленні емпіричної функції $V(t)$ від сублінійної степеневої, яку передбачає модель. Відповідно, не виключено, що статистика $\{\tau\}$ описується виразом, складнішим за функцію Вейбула (5). Можливо, це пов'язано з відхиленням від степеневих ціпфівських законів для реальних NT. Оскільки всі вивчені нами властивості ST принципово відмінні від властивостей NT і RT, генеративна модель Саймона, принаймні в її найпростішій формі, не описує останніх текстів та належить до іншого класу стохастичності. Хоча в літературі вже висловлювали схожу думку, наші емпіричні аргументи набагато серйозніші, ніж наведені в працях [1, 7]. З іншого боку, передбачаємо, що проста модифікація моделі Саймона з урахуванням загасання ймовірності $p(t)$ появи нових слів дасть динамічні властивості ST, близькі до NT і RT.

Показано, що текстам NT і RT притаманні обидва відомі з літератури [17, 33] вияви явища пульсацій – довгосяжні додатні кореляції в послідовності часових позицій $\{t\}$ нових слів (ефект пам'яті) й важкий хвіст статистичного розподілу $p(\tau)$. Запропоновано пояснення парадоксального явища кореляцій, виявленого в рандомних текстах. Його природа пов'язана зі специфічною сутністю поняття "нове слово" і відповідною "пам'яттю", яка зводиться до (загалом нетривіального) поступового зменшення часової густини появи цих слів у будь-яких (природних чи рандомізованих) текстах. Вперше показано, що пам'ять у послідовності лінгвістичних знаків і важкий хвіст розподілу їхніх часів очікування не завжди можна вважати "оргогональними"

ефектами: в межах запропонованого механізму пульсацій перших поява слів ці ефекти взаємопов'язані та одночасно зникають у границі, якщо $\beta = \theta \to 1$.

Тотожність статичних властивостей NT і RT на зразок законів Ціпфа і доведена у нашій роботі схожість статистики появи нових слів і динаміки словника для цих текстів схиляють до думки про суто статистичну природу законів Гіпса та Ціпфа, а також відсутність потреби в лінгвістичному обґрунтуванні цих законів. Це узгоджується з успішною якісною інтерпретацією функцій $p(\tau)$, $p(t)$ і $V(t)$ для NT і RT у межах модифікованої нульової стохастичної гіпотези.

Серед перспектив цього дослідження – аналіз явища флуктуацій у часових рядах $\{t_i\}$ позицій першої появи слів для текстів різних типів, які повинні бути пов'язані з явищем пульсацій для часів очікування. Скажімо, на підставі даних табл. 1 маємо ряд $\{t_i\} = \{110010\}$, який було би цікаво вивчити за стандартним методом рандомних прогулянок [45].

1. Baek S. K. Zipf's law unzipped / S. K. Baek, S. Bernhardsson, P. Minnhagen // New J. Phys. – 2011. – Vol. 13. – 043004 (21 pp.). 2. Adamic L. Unzipping Zipf's law / L. Adamic // Nature. – 2011. – Vol. 474. – P. 164–165. 3. Kornai A. How many words are there? / A. Kornai // Glottometrics. – 2002. – Vol. 4. – P. 60–85. 4. van Leijenhorst D. C. A formal derivation of Heaps' law / D. C. van Leijenhorst, Th. P. van der Weide // Inf. Sci. – 2005. – Vol. 170. – P. 263–272. 5. Gerlach M. Stochastic model for the vocabulary growth in natural languages / M. Gerlach, E. G. Altmann // Phys. Rev. X. – 2013. – Vol. 3. – 021006 (10 pp.). 6. Bernhardsson S. The meta book and size-dependent properties of written language / S. Bernhardsson, L. E. Correa da Rocha, P. Minnhagen // New J. Phys. – 2009. – Vol. 11. – 203015 (15 pp.). 7. Bernhardsson S. Size-dependent word frequencies and translational invariance of books / S. Bernhardsson, L. E. Correa da Rocha, P. Minnhagen // Physica A. – 2010. – Vol. 389. – P. 330–341. 8. Lü L. Zipf's law leads to Heaps' law: Analyzing their relation in finite-size systems / L. Lü, Z.-K. Zhang, T. Zhou // PLOS ONE. – 2010. – Vol. 5. – e14139 (11 pp.). 9. Yan X.-Y. Comment on 'A scaling law beyond Zipf's law and its relation to Heaps' law' [Electronic resource] / X.-Y. Yan, P. Minnhagen. – 2014. – Access mode: http://arxiv.org/abs/1404.1461. – Title from the screen. 10. Lü L. Deviation of Zipf's and Heaps' laws in human languages with limited dictionary sizes / L. Lü, Z.-K. Zhang, T. Zhou // Sci. Rep. – 2013. – Vol. 3. – 1082 (7 pp.). 11. Font-Clos F. A scaling law beyond Zipf's law and its relation to Heaps' law / F. Font-Clos, G. Boleda, A. Corral // New J. Phys. – 2013. – Vol. 15. – 093033 (16 pp.). 12. Bochkarev V. V. Deviations in the Zipf and Heaps laws in natural languages / V. V. Bochkarev, E. Yu. Lerner, A. V. Shevlyakova // J. Phys.: Conf. Ser. – 2014. – Vol. 490. – 012009 (4 pp.). 13. Font-Clos F. Log-log convexity of type-token growth in Zipf's systems / F. Font-Clos, A. Corral // Phys. Rev. Lett. – 2015. – Vol. 114. – 238701 (5 pp.). 14. Egghe L. Untangling Herdan's law and Heaps' law: Mathematical and informetric arguments / L. Egghe // J. Amer. Soc. Inf. Sci. Technol. – 2007. – Vol. 58. – P. 702–709. 15. Ebeling W. Long-range correlations between letters and sentences in texts / W. Ebeling, A. Neiman // Physica A. – 1995. – Vol. 215. – P. 233–241. 16. Hierarchical structures induce long-range dynamical correlations in written texts / E. Alvarez-Lacalle, B. Dorow, J.-P. Eckmann, E. Moses // Proc. Nat. Acad. Sci. (USA). – 2006. – Vol. 103. – P. 7956–7961. 17. Altmann E. G. Beyond word frequency: Bursts, lulls, and scaling in the temporal distributions of words / E. G. Altmann, J. B. Pierrehumbert, A. E. Motter // PLOS ONE. – 2009. – Vol. 4. – e7678 (7 pp.). 18. Altmann E. G. On the origin of long-range correlations in texts / E. G. Altmann, G. Cristadoro, M. D. Esposti // Proc. Nat. Acad. Sci. (USA). – 2012. – Vol. 109. – P. 11582–11587. 19. Флуктуації частоти літер і знаків в українських і російських текстах / О. С. Кушнір, А. М. Байовський, Л. Б. Іваніцький, С. В. Рихлюк // Матер. VII Укр.-польськ. наук.-практ. конф. "Електрон. та інф. технол.". – Львів : ЛНУ, 2015. – С. 76–79. 20. Статистичний розподіл і флуктуації довжин речень в українському, російському і англійському корпусах / О. С. Кушнір, О. С. Брик, В. Є. Дзіковський, Л. Б. Іваніцький, І. М. Катеринчук, Я. П. Кісь // Вісн. нац. ун-ту "Львівська політехніка". Сер. "Інф. сист. та мережі". – 2016. – № 854. – С. 228–239. 21. Eliazar I. The growth statistics of Zipfian ensembles: Beyond Heaps' law / I. Eliazar // Physica A. – 2011. – Vol. 390. – P. 3189–3203. 22. Simon H. On a class of skew distribution functions / H. Simon // Biometrika. – 1955. – Vol. 42. – P. 425–440. 23. Barabási A.-L. The origin of bursts and heavy

tails in human dynamics / A.-L. Barabási // Nature. – 2005. – Vol. 435. – P. 207–211. 24. Chen Y. S. Exponential recurrence distribution in the Simon-Yule model of text / Y. S. Chen // Cybernetics and Systems. – 1988. – Vol. 19. – P. 521–545. 25. Zanette D. H. Dynamics of text generation with realistic Zipf distribution / D. H. Zanette, M. A. Montemurro // J. Quant. Linguist. – 2005. – Vol. 12. – P. 29–40. 26. Keyword detection in natural languages and DNA / M. Ortuño, P. Carpena, P. Bernaola-Galván, E. Muñoz, A. M. Somoza // Europhys. Lett. – 2002. – Vol. 57. – P. 759–764. 27. Herrera J. P. Statistical keyword detection in literary corpora / J. P. Herrera, P. A. Pury // Eur. Phys. J. – 2008. – Vol. 63. – P. 135–146. 28. Level statistics of words: Finding keywords in literary texts and symbolic sequences / P. Carpena, P. Bernaola-Galván, M. Hackenberg, A. V. Coronado, J. L. Oliver // Phys. Rev. E. – 2009. – Vol. 79. – 035102(R) (4 pp.). 29. Про статистику відстаней між словами в тексті та проблему розпізнавання змістових слів / О. С. Кушнір, А. В. Волоско, Л. Б. Іваніцький, С. В. Рихлюк // Електроніка та інф. технол. – 2016. – Вип. 6. – С. 155–164. 30. До пояснення механізму явища "спалахів" у статистиці лінгвістичних елементів: часи очікування буквених n-грам / О. С. Кушнір, М. А. Альфавіцький, В. Є. Дзіковський, Л. Б. Іваніцький, І. М. Катеринчук, О. І. Шарга // Матер. VIII Укр.-польськ. наук.-практ. конф. "Електрон. та інф. технол.". – Львів : ЛНУ, 2016. – С. 84–89. 31. The effect of long-term correlations on the return periods of rare events / A. Bunde, J. F. Eichner, S. Havlin, J. W. Kantelhardt // Physica A. – 2003. – Vol. 330. – P. 1–7. 32. Vajna S. Modelling bursty time series / S. Vajna, B. Tóth, J. Kertész // New J. Phys. – 2013. – Vol. 15. – 103023 (17 pp.). 33. Goh K.-I. Burstiness and memory in complex systems / K.-I. Goh, A.-L. Barabási // Europhys. Lett. – 2008. – Vol. 81. – 48002 (5 pp.). 34. Altmann E. G. Recurrence time analysis, long-term correlations, and extreme events / E. G. Altmann, H. Kantz // Phys. Rev. E. – 2005. – Vol. 71. – 056106 (9 pp.). 35. Statistics of return intervals in long-term correlated records / J. F. Eichner, J. W. Kantelhardt, A. Bunde, S. Havlin // Phys. Rev. E. – 2007. – Vol. 75. – 011128 (9 pp.). 36. Cattuto C. A Yule-Simon process with memory / C. Cattuto, V. Loreto, V. D. P. Servedio // Europhys. Lett. – 2006. – Vol. 76. – P. 208–214. 37. Ferrer i Cancho R. Two regimes in the frequency of words and the origins of complex lexicons: Zipf's law revisited / R. Ferrer i Cancho, R. V. Solé // J. Quant. Linguist. – 2001. – Vol. 8. – P. 165–173. 38. Santhanam M. S. Return interval distribution of extreme events and long-term memory / M. S. Santhanam, H. Kantz // Phys. Rev. E. – 2008. – Vol. 78. – 051113 (9 pp.). 39. Long-term memory: A natural mechanism for the clustering of extreme events and anomalous residual times in climate records / A. Bunde, J. F. Eichner, J. W. Kantelhardt, S. Havlin // Phys. Rev. Lett. – 2005. – Vol. 94. – 048701 (4 pp.). 40. Gerlach M. Scaling laws and fluctuations in the statistics of word frequencies / M. Gerlach, E. G. Altmann // New J. Phys. – 2014. – Vol. 16. – 113010 (19 pp.). 41. Improving statistical keyword detection in short texts: Entropic and clustering approaches / C. Carretero-Campos, P. Bernaola-Galván, P. Ch. Ivanov, P. Carpena // Phys. Rev. E. – 2012. – Vol. 85. – 011139 (6 pp.). 42. Moreno-Sánchez I. Large-scale analysis of Zipf's law in English texts / I. Moreno-Sánchez, F. Font-Clos, A. Corral // PLOS ONE. – 2016. – Vol. 11. – e0147073 (19 pp.). 43. Kushnir O. S. New text-length scaling effects in statistics of natural texts / O. S. Kushnir, L. B. Ivanitskyi, S. V. Rykhlyuk // Матер. VII Укр.-польськ. наук.-практ. конф. "Електрон. та інф. технол.". – Львів : ЛНУ, 2015. – P. 80–83. 44. Ferrer i Cancho R. Zipf's law from a communicative phase transition / R. Ferrer i Cancho // Eur. Phys. J.: B. – 2005. – Vol. 47. – P. 449–457. 45. Long-range correlations in nucleotide sequences / C.-K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, H. E. Stanley // Nature. – 1992. – Vol. 356. – P. 168–170.