ISSN 2224-087X. Електроніка та інформаційні технології. 2024. Випуск 28. С. 22–37
Electronics and Information Technologies. 2024. Issue 28. P. 22–37

УДК 004.6, 004.9, 303.732, 519.1 DOI: https://doi.org/10.30970/eli.28.3

ВЛАСТИВОСТІ ЛЕКСИЧНИХ МЕРЕЖ, ПОБУДОВАНИХ НА ПРИРОДНИХ І РАНДОМНИХ ТЕКСТАХ

О. Кушнір, А. Дребот, Д. Остріков, О. Кравчук

кафедра оптоелектроніки та інформаційних технологій,
Львівський національний університет імені Івана Франка,
вул. Ген. Тарнавського, 107, 79017 м. Львів, Україна
oleh.kushnir@lnu.edu.ua

Експериментально вивчено та проаналізовано властивості лексичних мереж, побудованих на природному тексті (ПТ), його рандомізованій версії (РТ), а також мережі, одержаної з лексичної мережі ПТ шляхом рандомізації зв'язків у ній (рандомізованого графа – скорочено РГ). Ці властивості порівняно із закономірностями лексичної статистики для трьох відповідних текстів. Показано, що важкий (легкий) хвіст розподілу ступенів вузлів мережі визначається наявністю важкого (легкого) хвоста лексичного спектру відповідного тексту. Тому мережі для ПТ і РТ безмасштабні, на відміну від мережі РГ. На противагу до мережі РГ, обидві мережі для ПТ і РТ є тісними світами. Як наслідок, ефекти безмасштабності та тісного світу не слід прив'язувати до семантики або синтаксису.

Ключові слова: обробка природної мови, лінгвістичні мережі, аналіз і класифікація текстів, виявлення семантики, рандомні моделі, визначення ключових слів.

Вступ.

Ще з кінця 1990х років відомо, що природні та суспільні мережі на зразок Інтернету, WWW, транспортних або електричних мереж, харчових ланцюжків, мереж співпраці тощо не схожі на найпростіші модельні мережеві об'єкти як от регулярна ґратка або рандомний граф [1–3] (див. також огляди [4, 5] та огляд українських авторів [6]). З цього часу вивченню властивостей складних мереж надають значну увагу. Відомо, що ці мережі виявляють низку цікавих явищ, зокрема ефект тісного світу та безмасштабність (інваріантність щодо зміни масштабів мережі або скейлінг). Тісний світ означає, що усереднені найкоротші шляхи між двома будь-якими випадково обраними вузлами мережі відносно малі та порівнянні зі шляхами, притаманними для рандомних графів. Проте, на відміну від рандомних графів, складні мережі тісного світу додатково виявляють високу кластеризацію вузлів. Яскравою ілюстрацією є дані дослідження С. Мілґрема: для того, аби досягти контакту двох випадково обраних американців, достатньо всього шести проміжних осіб (див. [5]). Безмасштабність же мережі означає, що розподіл імовірності $p(k)$ для ступенів k (кількостей найближчих сусідів) її вузлів степеневий, а такому розподілові не притаманна наявність характерного масштабу. На додаток, дана функція $p(k)$ має важкий хвіст, тобто імовірність появи в мережі вузлів (хабів) із екстремально великими ступенями не є нехтовно малою, тоді як мережі з тонкими (наприклад, пуасонівськими або експоненційними) хвостами $p(k)$ практично виключають наявність хабів.

У 2001 році кілька груп дослідників [7–9] незалежно встановили, що мережі, побудовані на природних текстах (ПТ) або їхніх корпусах, теж виявляють ознаки тісного світу та безмасштабності. На думку авторів [7, 10], ці ознаки можуть відображати складну організацію природної мови і навіть еволюцію лексикону та особливості функціонування людського мозку, а їхнє дослідження тим самим робить внесок до когнітивних наук.

Властивості лінгвістичних мереж докладно вивчали в багатьох працях, зокрема для з'ясування можливостей розрізнення природних мов і типів текстових документів або «якості» текстів [11–16], а також із метою розпізнавання ПТ, які виявляють семантику та синтаксис, на тлі рандомних і семантично порожніх символьних послідовностей – т. зв. рандомних текстів (РТ) [17–21]. Більше того, було запропоновано використовувати різноманітні мережеві характеристики ПТ для визначення ключових слів або фраз, які стисло відображають семантичне поле текстового документа та дають змогу значно понизити вимірність будь-якої задачі інформаційного пошуку та класифікації текстів [9, 22–26] (див. також [27]). Наприклад, згідно з методом І. Мацуо та ін. [9], ключовими є ті терміни, при усуненні яких із тексту найпомітніше зменшується параметр, що кількісно описує тісноту світу лінгвістичної мережі даного тексту.

Мета цієї праці – подальші емпіричні дослідження складних лінгвістичних мереж і спроби з'ясування причин і походження явища тісного світу та безмасштабності лексичних мереж, побудованих на ПТ та його рандомізованих версіях.

Побудова лінгвістичних мереж і досліджувані тексти.

Почнемо з побудови лінгвістичної мережі на основі тексту (т. зв. mapping). Розроблені нами програмні засоби давали змогу створювати мережу, вузлами якої могли бути буквені, символьні або лексичні n-грами. У цьому дослідженні вузлами лінгвомережі обирали слова (елементи словника тексту), тобто лексичні n-грами при n = 1, а наявність або відсутність зв'язків між вузлами визначали за критерієм близькості появи слів у тексті. Це відповідає типовому означенню мережевих зв'язків у літературі за предметом, хоча іноді трапляються винятки (див., наприклад, працю [28], де зв'язки відповідали семантичній близькості слів). Близькість слів одне до одного означали радіусом сусідства $r = 1, 2, \ldots$. Скажімо, при $r = 1$ сусідами слова означали в попередньому реченні є одного та радіусом. Такий підхід відповідає стандартному та широко прийнятому для лінгвомереж випадку нескерованого графа, хоча послідовне врахування синтаксису, правила якого не є зворотними, потребувало би аналізу скерованого графа [13, 17]. Додатковим критерієм сусідства вузлів можна було вважати приналежність слів до того ж речення. Нижче ми обмежимося розглядом лише випадку $r = 3$.

Ми не використовували жодного частотного фільтра, так що в мережі були присутні навіть слова з найнижчою абсолютною частотою $F = 1$ у тексті (т. зв. hapax legomena). Попри відповідну технічну можливість, ми працювали тільки з т. зв. «необмеженою» мережею, зв'язки в якій не перевірялися на статистичну значимість (див., наприклад, пояснення в праці [7]). Зрештою, відомо, що результуючі мережеві параметри зазвичай виявляють значну стійкість до способів побудови лінгвомережі [7]. Нарешті, на відміну від деяких праць у галузі (зокрема [17, 19]), ми аналізували «бінарну» мережу, що відповідає незваженому графові. Іншими словами, кожен зв'язок вузлів у мережі зараховують один раз, незалежно від кратності його появи в тексті. До результатів для зважених графів, для яких вартувало би переозначити довжину шляху між вузлами, ми звернемося в окремому дослідженні.

Нижче ми зосередимося лише на аналізі трьох найголовніших мережевих параметрів: середнього коефіцієнта кластерності мережі C, середньої довжини найкоротшого шляху l і середнього ступеня N вузлів мережі. Загалом же створений нами програмний додаток давав змогу вивчати коефіцієнти кластерності, середні шляхи, ступені та кратності всіх зв'язків для кожного вузла мережі, а також розраховувати звичайний і кумулятивний розподіли ймовірності ступенів вузлів.

Основний об'єкт нашого вивчення – це англомовний ПТ художнього твору «Книга джунглів» Р. Кіплінґа, що містить близько 270.7 тис. друкованих символів і пробілів. Його довжина L на лексичному лінгвістичному рівневі складає $L \approx 51.1$ тис. слів, а словник – $V \approx 4.9$ тис. слів. Усі літери в тексті було приведено до нижнього регістру, додаткових лематизації або стемінґу не застосовували, а стоп-слова з тексту не видаляли.

Мережеві характеристики названого тексту порівнювали з характеристиками, мабуть, найпростішої рандомної моделі – ПТ, рандомізованого на лексичному лінгвістичному рівневі. Для надійності рандомізації у вихідному тексті було здійснено $M = 10^8 >> L$ циклів перестановок слів. Логіка побудови даного РТ така: слова як одиниці найнижчого лінгвістичного рівня, які володіють семантикою, поєднуються в тексті в певній послідовності згідно з правилами синтаксису та логікою тексту, будуючи та розвиваючи семантику всього тексту, а тому рандомні перестановки цих одиниць повинні повністю знищити і синтаксис, і семантику тексту. З іншого боку, такі «статичні» властивості тексту як частоти слів лишаються інваріантом рандомізації. Оскільки рандомізація докорінно змінює закономірності поєднання слів у тексті, то вона повинна радикально перебудувати лінгвомережу принаймні на локальному рівні. Постає питання: як тоді зміняться макроскопічна структура всієї мережі?

Останній об'єкт нашого дослідження – лінгвомережа, побудована на основі вихідної мережі ПТ із V вузлами і K зв'язками шляхом рандомних перестановок цих зв'язків. За принципом побудови ця мережа, яку назвемо РГ, дещо схожа до графа Ердоша–Рені [5].

Цікаво порівняти не тільки властивості всіх трьох згаданих лінгвомереж, але й статичні та найпростіші динамічні властивості відповідних текстів. Це рангова залежність частоти слів $F(r)$, лексичний спектр тексту (тобто розподіл імовірності частот $p(F)$) і відповідний доповнювальний кумулятивний розподіл $P(F)$, які для ПТ описують відповідно першим і другим законами Ціпфа та законом Парето, а також масштабна поведінка словника $V(L)$, яку описують законом Гіпса–Гердана (див., наприклад, [29–34]).

У разі штучної мережі РГ маємо деяку технічну трудність. А саме, для мереж ПТ і РТ вихідними є тексти, а от для РГ вихідним є не текст, а лінгвомережа для ПТ. Для побудови відповідного тексту на основі мережі РГ слід виконати зворотну процедуру – inverse mapping. В ідеалі, ми би мали відтворити у формі тексту таку штучну послідовність слів, яка точно та однозначно відтворює всі зв'язки вузлів у рандомній мережі, що потребувало би ускладнених алгоритмів. Наше практичне рішення було простішим: кожен зв'язок деяких двох слів, присутній у рандомній мережі РГ, позначимо реченням у відповідному синтетичному тексті, де критерієм сусідства служить виключно приналежність слів до того ж речення. Одержимо «текст» на зразок w1700 w1556. w30 w18. w4256 w2193. ... , де w1700 умовно позначає деяке слово зі словника вихідного ПТ, а кожне речення відповідає зв'язку двох слів. Такий текст ми назвемо так само, як і саму рандомну мережу, – «РГ».

Базова лінгвістична статистика текстів ПТ, РТ і РГ.

На рис. 1 проілюстровано основні лінгвостатистичні залежності для лексики РТ. Зазначимо, що залежності Рис. 1а–1в для ПТ будуть такими самими, бо зміна порядку слів у тексті не впливає на їхні частоти. Загалом залежності словника для ПТ і РТ могли би дещо відрізнятися через різну динаміку появи нових слів у цих текстах, проте ми використовували режим вимірювання словника за стандартним методом ковзного вікна (усереднення словника за локальними вікнами однакового розміру на різних ділянках тексту), а тому дані Рис. 1г одночасно стосуються й вихідного ПТ.

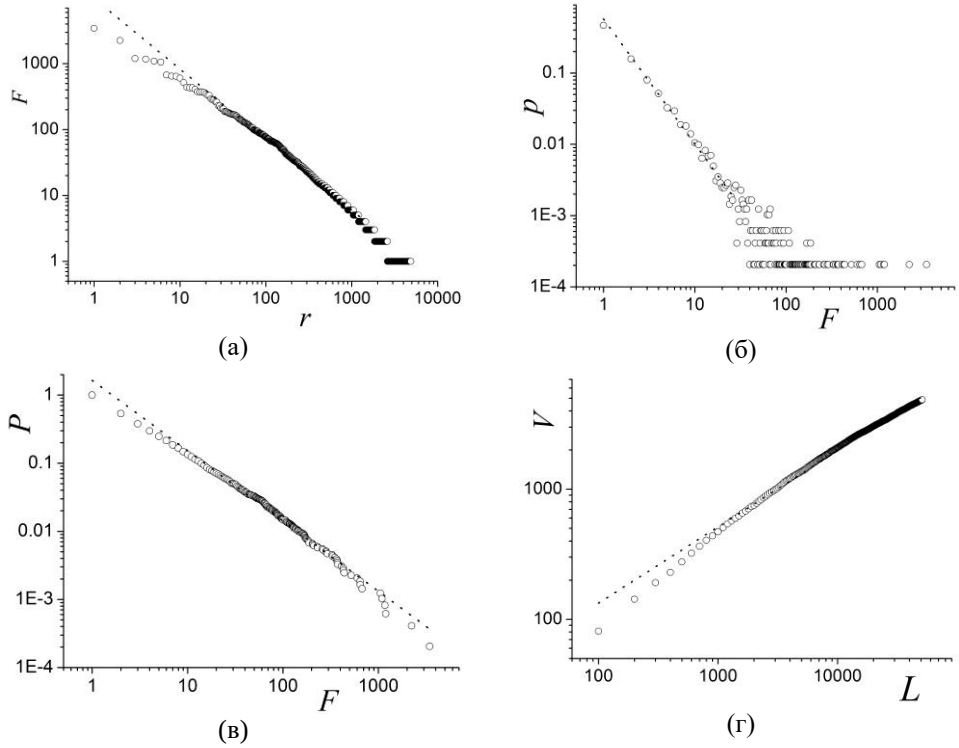

Рис. 1. Рангова залежність частоти слів $F(r)$ (а), лексичний спектр $p(F)$ (б), розподіл доповнювальної кумулятивної ймовірності $P(k)$ (в) і залежність розмірів словника від довжини тексту $V(L)$ (г) для РТ. Відповідні залежності для ПТ такі самі. Як і на всіх інших рисунках, штриховані лінії позначають лінійний фітинг емпіричних залежностей, а масштаб по обох осях підібрано так, аби лінеаризувати залежності, наскільки це можливо.

Fig. 1. Rank–frequency dependence $F(r)$ (a), lexical spectrum $p(F)$ (б), complementary cumulative probability distribution $P(k)$ (в), and text-length dependence of vocabulary $V(L)$ (г) for random text. The dependences are the same as for the natural text. Like in all the other Figures, вot lines correspond to linear fitting of empirical dependencesб while the abscissa and ordinate scales are chosen with the purpose of linearizing the dependences whenever possible.

Усі залежності на Рис. 1 наближено описуються степеневими законами [32–34]:

$$F(r) \sim r^{-\alpha},\ p(F) \sim F^{-\beta},\ P(F) \sim F^{-\pi},\ V(L) \sim L^{\theta}, \qquad (1)$$

де α, β, π і θ – степеневі показники відповідно першого та другого законів Ціпфа, закону Парето та закону Гіпса. Добре відомо, що ці параметри не є незалежними і їх можна поєднати своєрідними «співвідношеннями універсальності» (див. [32–34]):

$$\beta = 1 + 1/\alpha\,,\ \beta = 1 + \theta\,,\ \pi = 1/\alpha\,,\ \begin{cases} \theta = 1/\alpha & (\alpha > 1) \\ \theta = 1 & (\alpha \le 1) \end{cases}. \qquad (2)$$

Перевірка даних Рис. 1 для залежностей $F(r)$, $p(F)$, $P(F)$ і $V(L)$ засвідчує загалом задовільне виконання теоретичних співвідношень (1) із типовою точністю ~ 10%, за винятком останньої формули в (1). Послідовний аналіз причин такої недостатньої точності потребував би значного відхилення від предмету нашого обговорення. Ми лише перерахуємо кілька таких причин. Це недостатня точність визначення деяких зі степеневих показників α, β, π і θ (наприклад, унаслідок використання типового прийому лінійного фітингу залежностей в подвійному логарифмічному масштабі, який вносить небажане логарифмічне зважування; наявності сходинок на Рис. 1а; домінування шумів на хвості на Рис. 1б), математична нестрогість базового припущення про те, що функції $F(r)$ і $P(F)$ є оберненими, а також наближеність самого степеневого характеру функцій дискретної змінної $F(r)$, $p(F)$, $P(F)$ і $V(L)$ (див., зокрема, [35, 36]). Проте згаданої точності дотримання співвідношень (2) цілком достатньо для одержання принаймні якісних висновків про функціональний характер цих залежностей (див. нижче).

На Рис. 2 показано залежності $F(r)$, $p(F)$, $P(F)$ і $V(L)$ для тексту РГ, побудованого за умови $r = 3$. Вони мають складний характер, який навряд чи можна описати елементарними функціями на зразок степеневої, логарифмічної або експоненційної. Проте в цьому дослідженні ми використаємо саме такий спрощений, напівкількісний підхід.

Із Рис. 2а видно, що рангова залежність $F(r)$ для РГ наближено логарифмічна, хвости залежностей $p(F)$ і $P(F)$ наближено експоненційні, а залежність $V(L)$ насичується при великих L. Для якісного аналізу такої поведінки скористаємося співвідношеннями (2), узагальненими на випадок відхилення лінгвостатистичних законів від степеневого характеру. А саме, вважатимемо, що в «узагальненому» степеневому законі $f(x) \sim x^{-a}$ степінь $a \to 0$ відповідатиме логарифмічній функції ($f(x) \sim A - B\log x$) із надважким хвостом, а степінь $a \to \infty$ – експоненційній функції ($f(x) \sim \exp(-x/x_0)$, де x_0 – деяка константа) з тонким хвостом.

Отже, дані Рис. 2а схиляють до висновку $\alpha \to 0$, тобто рангова залежність логарифмічна: $F(r) \sim A - B\log r$. Тоді з першої із формул (2) одержимо експоненційну поведінку масової функції розподілу $p(F)$ ($\beta \to \infty$ і $p(F) \sim \exp(-F/F_0)$, де F_0 – константа), що якісно підтверджують дані Рис. 2б. Формули (2) так само передбачають експоненційну поведінку кумулятивної функції розподілу $P(F)$ (див. Рис. 2в). Нарешті, з формул (2) випливає, що $\theta \to \infty$. Але найпростіше експоненційне зростання у формі $V(L) \sim \exp(L/L_0)$ суперечило би й логіці, і даним експерименту. Проте даним Рис. 2г для словника задовольняє експоненційна залежність іншого вигляду:

$$V = V_0[1 - \exp(-L/L_0)], \qquad (3)$$

де L_0 – деяка характеристична довжина тексту. Дані вставки на Рис. 2г і відповідний фітинґ на основному графіку Рис. 2г засвідчують, що це припущення якісно правильне ($L_0 \approx 7384$ слів).

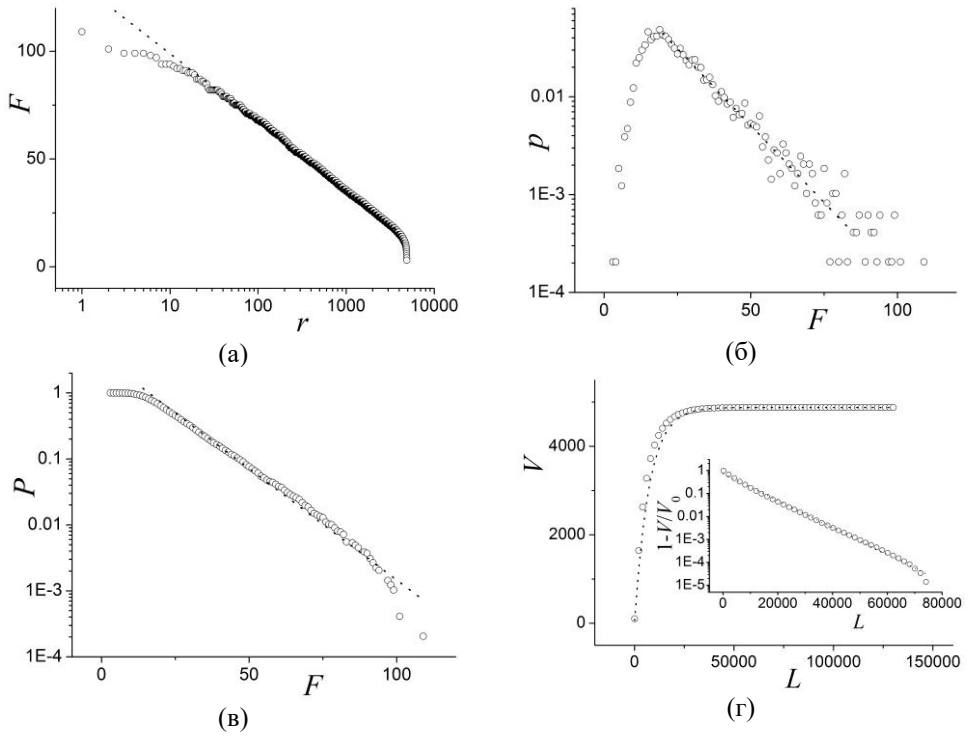

Рис. 2. Рангова залежність частоти слів $F(r)$ (а), лексичний спектр $p(F)$ (б), доповнювальний кумулятивний розподіл $P(k)$ (в) і залежність розмірів словника від довжини тексту $V(L)$ (г) для РГ, одержаного для радіуса сусідства $r = 3$. На вставці панелі (г) – масштабна залежність параметра $(1 – V/V_0)$ (див. формулу (3)). Штриховані лінії позначають лінійний фітинг емпіричних залежностей.

Fig. 2. Rank–frequency dependence $F(r)$ (а), lexical spectrum $p(F)$ (б), complementary cumulative distribution $P(k)$ (в), and text-length dependence of vocabulary $V(L)$ (г) for the random graph, as obtained at the neighbor radius $r = 3$. The insert in panel (г) shows the text-length dependence of $(1 – V/V_0)$ parameter (see formula (3)).

Зазначимо, що скінченність словника тексту РГ не є чимось незвичним: така властивість притаманна деяким узагальненим рандомним моделям природної мови (див. працю [34]). Отже, на відміну від ПТ і РТ, розподіли ймовірності частоти слів для РГ на хвості мають не степеневий, а наближено експоненційний характер, а словник наростає за експоненційним законом, що описує перехідні процеси в електротехніці.

Користуючись уведеним вище підходом, можна також успішно описати всі закони лінгвостатистики для букв або залежності $F(r)$, $p(F)$, $P(F)$ і $V(L)$ для ієрогліфічних мов (див. [37]). Відповідні результати буде висвітлено в окремому дослідженні.

Скейлінг у лінгвістичних мережах ПТ, РТ і РГ.

Перейдемо до аналізу лексичних мереж для ПТ, РТ і РГ. На Рис. 3а, б представлено рангові залежності $k(r)$, на Рис. 3в, г – функції масової ймовірності ступенів $p(k)$, а на

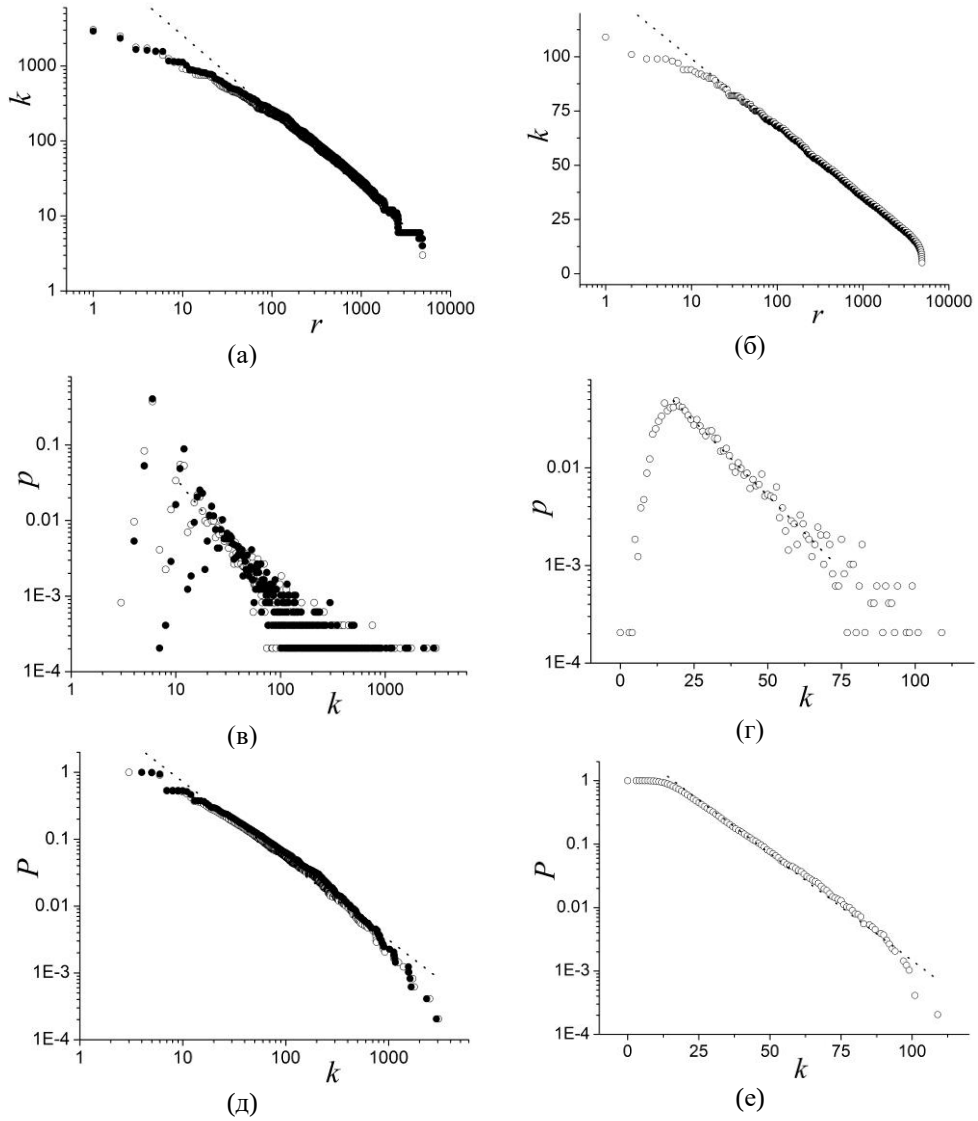

Рис. 3. Рангові залежності ступенів мережевих вузлів $k(r)$ (а, б), а також відповідні розподіли ймовірності $p(k)$ (в, г) і розподіли доповнювальної кумулятивної ймовірності $P(k)$ (д, е) для ПТ і РТ (а, в, д) і РГ (б, г, е) при $r = 3$. Незаповнені кружки на панелях (а), (в) і (д) відповідають ПТ, а заповнені – РТ. Штриховані лінії позначають лінійний фітинґ емпіричних залежностей, а фітинґ на панелях (в) і (д) стосується ПТ.

Рис. 3. Rank dependences of node degrees $k(r)$ (а, б), degree distributions $p(k)$ (в, г), and complementary cumulative distributions $P(k)$ (д, е), as obtained for natural and random texts (а, в, д) and random graph (б, г, е) at $r = 3$. Open and full circles in panels (а), (в) and (д) correspond to natural and random texts, respectively, and fitting in panels (в) and (д) to natural text.

Рис. 3д, е – функції кумулятивної ймовірності $P(k)$. Їх було одержано з радіусом сусідства $r = 3$. Заслуговує на увагу фактично повна якісна відповідність між поведінкою

функцій $F(r)$ і $k(r)$, $p(F)$ і $p(k)$, а також $P(F)$ і $P(k)$ (порівн. із відповідними графіками на Рис. 1 і Рис. 2), включно з орієнтовною функціональною формою та характером хвоста. Це емпірично доводить твердження авторів праць [17, 19] про те, що в лексичній мережі, побудованій на деякому тексті, ступінь слова-вузла в деякому сенсі еквівалентний до частоти слова в цьому тексті. Чи не єдиною відмінністю є відсутність (або наявність) початкової зростаючої ділянки та максимуму відповідно на залежностях $p(F)$ (або $p(k)$) для ПТ і РТ. Причини низької ймовірності вузлів мережі з найменшими ступенями слід вивчати додатково.

З іншого боку, порівняння даних Рис. 1 і Рис. 2 із даними Рис. 3 вказує на те, що взаємні зв'язки між функціональним характером різних мережевих характеристик $k(r)$, $p(k)$ і $P(k)$ можна аналізувати, використовуючи ті ж формули (2) для «узагальненого» степеневого закону, що були застосовані нами вище для аналізу зв'язків між залежностями $F(r)$, $p(F)$ і $P(F)$. Наприклад, наближено логарифмічний характер функції $k(r)$ означає наближено експоненційний характер $P(k)$.

Значні шуми залежностей $p(k)$ для всіх мереж стримують від категоричних висновків про характер їхніх хвостів. Проте залежності «інтегрального» розподілу $P(k)$ не залишають сумнівів у тому, що їхні хвости для ПТ і РТ важкі, тобто мережі, побудовані на цих текстах, мають безмасштабний характер. Їх можна наближено описати формулами

$$p(k) \sim k^{-\gamma}, \; P(k) \sim k^{-(\gamma-1)}. \tag{4}$$

Із точніших даних Рис. 3д одержуємо усереднений нахил для ПТ $\gamma \approx 2.2$, що корелює з параметрами γ, відомими з літератури для англійської мови (наприклад, $\gamma \approx 2.1$–2.2 [16], $\gamma \approx 1.9$ [17], $\gamma \approx 1.6$ [18], $\gamma \approx 2.0$ і 2.2 [19] і $\gamma \approx 1.8$–2.2 [21]). Крім того, всі залежності для ПТ і РТ на Рис. 3а, в, д майже нерозрізнимі, що узгоджується з висновками [18, 19, 21]. Щоправда, автори [16] стверджували, що ПТ і РТ характеризуються різними залежностями $p(k)$, але в цій праці «РТ» формували, випадково переставляють пробіли й тим самим докорінно (і непрогнозовано) змінюючи і словник тексту, і частотні співвідношення між словами.

Водночас, лінгвомережа РГ однозначно виявляє тонкий хвіст розподілу ступенів вузлів, можливо навіть тонший за експоненційний (див. Рис. 3е). За теорією, РГ Ердоша–Рені має описуватися пуасонівським розподілом імовірності $p(k)$ і неповною гамма-функцією для кумулятивного розподілу. Проте докладніший аналіз, деталі якого ми опускаємо, засвідчує, що кількісні деталі поведінки нашої рандомної мережі дещо інші. Відповідні причини потребують додаткового з'ясування.

Один із РТ, який сформували автори праці [21], базувався по-суті на однорідній ранговій залежності для слів, узятих зі словника вихідного ПТ. У дусі формалізму, представленого формулами (2), функція $F(r) = \text{const}$ мала би означати дельта-функцію для розподілу $p(F)$, хвіст якої безмежно тонкий. Водночас, для лінгвомережі відповідного РТ автори [21] одержали тонкий експоненційний хвіст $p(k)$.

Основний же висновок порівняльного аналізу даних Рис. 1–3 такий: важкий або тонкий хвіст розподілу ймовірності ступенів вузлів у лінгвомережі однозначно диктується наявністю важкого або тонкого хвоста лексичного спектру базового тексту. Іншими словами, поведінку на хвості $p(k)$ визначають характеристики слів, які мають найвищі частоти та, як наслідок, найбільше зв'язків, – навіть попри те, що мережа незважена та не враховує кратних зв'язків. Тому висока або низька ймовірність таких слів у тексті означатиме й важкий або тонкий хвіст $p(k)$. Відповідно, степеневі закони Ціпфа для лексичних частот і степеневі закони (4) еквівалентні. Оскільки рандомізація ПТ залишає закони

Ціпфа інваріантом, то лінгвомережі ПТ і РТ однаково виявляють безмасштабність. Ці результати узгоджуються з даними та висновками праць [17–19, 21]. Отже, безмасштабність мережі навряд чи можна пов'язувати із семантикою і/або синтаксисом тексту. З іншого боку, «справжні» рандомні мережі неможливо побудувати на основі рандомізованих текстів; мережі, побудовані на таких РТ, не виявляють масштабних ефектів, притаманних РГ.

До предмету нашого обговорення мають прямий стосунок наші попередні дані вивчення лінгвостатистики та властивостей мереж із буквеними n-грами як вузлами. Як і для тексту РГ, тут наближено виконуються умови $\alpha \to 0$ і $\beta \to \infty$, тобто хвіст $p(F)$ наближено експоненційний і тому тонкий. Це дає тонкий хвіст $p(k)$ і для ПТ, і для відповідних РТ. Зазначимо, що букви та буквені n-грами (принаймні при малих n) є семантично порожніми лінгвістичними об'єктами навіть у ПТ, хоча їхні послідовності в ПТ й описуються своєрідним «синтаксисом» (точніше, особливостями типового поєднання букв у комбінації, що формують слова, які частково охоплюють морфологія і фонетика). Водночас, у РТ, одержаних рандомізацією ПТ на рівні букв, системі буквених n-грам не можна приписувати ні семантики, ні «синтаксису». У дусі зроблених вище висновків, слід вважати, що відсутність ефекту безмасштабності буквених мереж є наслідком не відсутності семантики і/або «синтаксису», а відсутності важкого хвоста $p(F)$. Взагалі, за влучним висловом [17], розподіл імовірності ступенів вузлів – це не найліпше мірило самоорганізованості структури мережі.

Оскільки з наших даних і з літератури випливає прив'язка степеневого характеру функції $p(k)$ лексичних мереж до законів Ціпфа [17–19], то постає цікаве питання: чи буде лінгвомережа безмасштабною, якщо залежність $p(F)$ матиме важкий хвіст, проте не буде точно описуватися законом Ціпфа? Така ситуація матиме місце, наприклад, для РТ, побудованих за моделлю мавпи Міллера з однаковими частотами букв (див. обговорення [38, 39]). На підставі одержаних нами даних відповідь на це питання ствердна. Зрештою, ця відповідь стає ще очевиднішою, коли згадати, що закони Ціпфа не є математично точними навіть для вивчених нами ПТ і РТ.

Нарешті, було би цікаво *прямо* перевірити наведені вище висновки стосовно причин появи безмасштабності лексичних мереж шляхом дослідження нульових рандомних моделей природної мови інших типів, розподіли ймовірності частот слів $p(F)$ для яких мали би тонкий хвіст. Відповідні результати буде наведено в окремому дослідженні.

Ефект тісного світу в лінгвістичних мережах ПТ, РТ і РГ.

Перейдемо тепер до аналізу основних параметрів мереж, побудованих на досліджуваних текстах (див. Таблицю 1). У повній відповідності до висновків [21], відмінності параметрів C, l і N для ПТ і РТ ледь помітні. З нашого досвіду, навіть найменші зміни в дефініції лексики тексту (наприклад, чи вважати числа словами; чи розрізняти слова на зразок its і it's) дадуть зміни до C, l і N, порівняльні або й дещо більші за зареєстровані нами відмінності між ПТ і РТ. Тому ПТ і РТ справді важко розрізнити за мережевими параметрами [21].

Наші дані для понад 100 текстів різними мовами світу, які будуть опубліковані окремо, показують, що відмінності основних мережевих параметрів між ПТ і РТ мають швидше статистичний, аніж детермінований характер, тобто важко однозначно передбачити, чи параметр мережі РТ буде більшим або меншим за його відповідник для деякого ПТ. Зокрема, усереднені по всіх ПТ і РТ середні довжини шляху l становили

$l_{\text{ПТ}} = 1.83 \pm 0{,}13$ і $l_{\text{РТ}} = 1.77 \pm 0{,}12$. Визначальним є той факт, що середньоквадратичні відхилення $\Delta l_{\text{ПТ}}$ і $\Delta l_{\text{РТ}}$ помітно вищі за відмінність середніх значень $< l_{\text{ПТ}} >$ і $< l_{\text{РТ}} >$, так що стандартний Z-тест дає оцінку $Z \approx 2(< l_{\text{ПТ}} > - < l_{\text{РТ}} >)/(\Delta l_{\text{ПТ}} + \Delta l_{\text{РТ}}) \approx 0.5$, яка в жодному разі не спростовує нульову гіпотезу про ідентичність середніх довжин шляху для ПТ і РТ. Тому, на нашу думку, висновки праці [21] про зменшення середнього шляху в мережі при переході від ПТ до РТ слід розуміти швидше як факт, притаманний конкретним дослідженим у цій праці текстам.

Таблиця 1. Деякі мережеві параметри ПТ, РТ і РГ, визначені за умови радіуса сусідства $r = 3$.
Table 1. Some network parameters of natural and random texts and random graph obtained at $r = 3$.

Мережевий параметр	ПТ	РТ	РГ
Кількість вузлів V	4876	4876	4876
Кількість зв’язків	65512	65512	65512
Коефіцієнт кластерності C	0.6692	0.6673	0.00872
Середня довжина шляху l	2.280	2.318	2.919
Середній ступінь N	30.66	34.23	26.84
Параметр Волша μ	98.25	96.37	1

Дослідження згаданої сотні текстів також засвідчили, що мережеві параметри C і l слабко корелюють з розміром словника тексту V (тобто кількістю вузлів у мережі) та його довжиною L. Так, показник кореляції найменший (8%) для залежності $C(V)$ і найбільший (28%) для залежності $l(V)$. Звісно, ці «ансамблеві» дані не суперечать висновкам

[13] про динаміку $l(V)$ для окремих текстів із виходом на насичені значення l_c після деяких критичних значень V_c, оскільки словники всіх вивчених нами текстів $V > V_c$.

Перейдемо до найважливішого питання – порівняння мережевих параметрів для ПТ і РТ із параметрами модельного РГ. Для цього розрахуємо параметр Волша μ [40]:

$$\mu = (C/l)/(C_{\text{РГ}}/l_{\text{РГ}}), \quad (5)$$

де індекс «РГ» стосується коефіцієнта кластерності C і середньої довжини шляху l для нашого РГ. Параметр μ є кількісним мірилом ефекту тісного світу в мережі. Із Таблиці 1 видно, що мережевий світ для обох текстів ПТ і РТ тісний ($\mu_{\text{ПТ}}, \mu_{\text{РТ}} >> 1$).

Із Таблиці 1 також видно, що ефект тісного світу для РТ лише незначно ($\mu_{\text{РТ}} / \mu_{\text{ПТ}} \approx 0.98$) слабший, ніж для ПТ. Віддалені наслідки цього факту важко переоцінити, хоча, наскільки відомо авторам, вони досі не були належно усвідомлені. Нагадуємо, що РТ не притаманні ні семантика, ні синтаксис, а тому тісний світ у мережі, побудованій на РТ, означає, що цей ефект категорично не можна прив'язувати до семантики або синтаксису тексту. З цим узгоджуються й наші попередні емпіричні дані щодо наявності ефекту тісного світу в мережах буквених n-грам і в мережах для РТ інших типів, аніж висвітлено в цій праці. Тісний світ притаманний навіть лінгвомережам, побудованим на найпростіших рандомних моделях і нульових гіпотезах у галузі опрацювання природної мови на зразок моделей «мішка зі словами», мавпи Міллера або «багатий стає багатшим». Тому слід визнати, що тісний світ не має зв'язку із семантикою ні лінгвістичних одиниць, ані їхньої сукупності, а також зі складною структурою природної мови. Він не зумовлений міркуваннями легкості переходу людського мозку від однієї семантичної одиниці до іншої серед сотень тисяч таких одиниць і навряд чи пов'язаний із предметом когнітивних наук (порівн. із висновками [7, 10]).

У світлі цього твердження постають серйозні сумніви в спроможності якщо не всіх мережевих методів визначення ключових слів, то принаймні методу [9], заснованого на використанні явища тісного світу (див. Розділ 1). І справді, наш практичний аналіз параметрів якості мережевих методів пошуку ключових термінів засвідчив їхні серйозні труднощі та, зокрема, той факт, що ці методи дають очевидно недостатні результати за умови, коли текст містить стоп-слова на зразок the, of, and. Саме ці слова-хаби переважно – і невиправдано – потрапляють на перші позиції за релевантністю в рамках мережевих методів. Водночас, відносні успіхи згаданих методів за альтернативної умови очищення тексту від стоп-слів фактично можуть завдячувати не так плідності мережевих підходів, як неявному використанню, мабуть, історично першого, проте доволі недосконалого частотного методу пошуку ключових слів [41]. За цим методом, ключовими в тексті є слова, які належать до «ядра» рангової залежності $F(r)$, тобто ті слова, що йдуть першими в списку за спаданням частоти після того, як ми позбулися стоп-слів.

У світлі наведених міркувань можна піддати сумнівам і висновки [12] про можливість кореляції кількісних параметрів мережі із «якістю» відповідного тексту. Насправді тут, мабуть, знову йде мова не про самостійний мережевий метод, а про непомічену авторами [12] неявну залежність мережевих параметрів од лінгвостатистичних параметрів на зразок степеневих показників α або β. А от вже останні параметри справді володіють певним ресурсом в оцінювання «якості» текстів [42]. Зокрема, можна припустити, що тексти з крутішою ранговою залежністю $F(r)$ видаватимуться конкретнішими, більш зосередженими на деякому предметі обговорення, тоді як тексти із пологішою $F(r)$ потерпатимуть од «розпиленості уваги» автора – а чи недостатній концентрації на предметі.

Подібні критичні зауваження можна висловити й на адресу праць [14–16], автори яких пробували класифікувати мови за мережевими параметрами. Насправді, оскільки мережеві параметри ПТ і РТ майже однакові, ці параметри навряд чи можуть містити *пряму* інформацію про специфіку різних природних мов. Відповідно, автори [14–16], мабуть, досягли успіху в розрізненні та класифікації мов знову завдяки неявній залежності характеристик лінгвомереж від лінгвостатистичних параметрів. Іншими словами, тих же результатів можна було би досягти шляхом прямого аналізу даних лінгвостатистики. Отже, ми припускаємо, що всі згадані щойно дослідження мають спільну методичну проблему: неврахування суттєвого, хоча й неявного впливу лексичної лінгвостатистики на мережеві параметри.

Нарешті, масштабний статистичний аналіз [20] мережевих параметрів як маркерів розрізнення семантично наповнених ПТ від семантично порожніх РТ засвідчив неспроможність мережевого підходу, у повній згоді з нашими результатами. У цьому плані твердження [16] про успішну диференціацію ПТ і РТ за допомогою мережевих підходів, яке присутнє навіть в резюме роботи, не слід розуміти буквально. Насправді, як уже загадувалося вище, під поняттям «РТ» автори [16] мали на увазі не модель «мішка зі словами», тобто рандомізовані тексти, в яких знищено впорядкування лексичних одиниць вихідного ПТ, а текст, словник якого докорінно змінено. Більше того, автори [16] не з'ясували особливостей лексичного спектру їхнього «РТ», хоча останній має визначальний вплив і на скейлінґ лексичної мережі, й на наявність і кількісні характеристики ефекту тісного світу в ній.

Висновки.

Отже, в цій праці було експериментально досліджено та феноменологічно проаналізовано властивості лексичних мереж, заснованих на сусідстві слів у тексті. Ці лінгвістичні мережі ми будували на основі ПТ, а також на основі РТ, який представляв собою результат надійної рандомізації ПТ на лексичному рівні. Іншим предметом вивчення була мережа типу РГ, яка мала таку ж кількість вузлів, як і мережа ПТ, і рандомізовані зв'язки між цими вузлами. Ми розглядали незважені та необмежені мережі, де слова не фільтрувалися за частотою, а стоп-слова не видалялися.

Було розраховано основні параметри таких мереж і порівняно їх із даними для лексичної статистики, одержаними для ПТ, РТ і РГ. Останні дані зазвичай виражають законами Ціпфа, Парето і Гіпса. З цією метою було створено додатковий «текст» РГ на основі мережі РГ. І ПТ, і РТ виявляють розподіли частот слів, що описуються степеневим законом із важким хвостом. Навпаки, лексична статистика для тексту РГ характеризується майже логарифмічною частотно-ранговою залежністю та тонким (приблизно експоненційним) розподілом частот.

Розподіли ймовірностей для ступенів вузлів, знайдені для мереж ПТ і РТ, близькі один до одного. Крім того, розподіли ступенів для ПТ і РТ, з одного боку, та РГ, з іншого боку, мають відповідно важкий (степеневий) і тонкий (майже експоненціальний) хвости. Іншими словами, мережі ПТ і РТ безмасштабні, на відміну від мережі РГ. Крім того, це означає, що важкий (або тонкий) хвіст розподілу ступенів вузлів є наслідком важкого (або тонкого) хвоста частотного розподілу слів у тексті. Середні коефіцієнт кластерності та довжина шляху для мереж, побудованих на ПТ і РТ, дуже близькі. На відміну від мережі РГ, мережі ПТ і РТ є тісними світами, а їхні параметри Волша, що вимірюють тісноту світу, відрізняються лише на 2%.

Нарешті, нами проаналізовано низку наслідків одержаних нами емпіричних даних і деяких результатів, відомих з літератури. Зокрема, оскільки PT не має жодної семантики чи синтаксису, ми вважаємо, що було би неправильно пов'язувати властивості безмасштабності та тісного світу лексичних мереж із семантикою або синтаксисом.

На завершення зазначимо таке. Наскільки відомо авторам, розглянуті досі дослідниками складні мережі одночасно виявляли (або не виявляли) і ефект тісного світу, і відсутність характерного масштабу. Постає питання: чи ці явища жорстко прив'язані одне до одного – чи вони певною мірою незалежні (ортогональні)? Іншими словами, чи існують мережі, в яких є одне з цих явищ, проте відсутнє інше? Для відповіді слід розширити коло досліджуваних складних мереж. Відповідна робота виконується в нашій лабораторії.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

[1] *Watts D. J., Strogatz S. H.* Collective dynamics of 'small-world' networks // Nature. – 1998. – Vol. 393. – P. 440–442.

[2] *Albert R., Jeong H., Barabási A.-L.* Diameter of the world-wide web // Nature. – 1999. – Vol. 401. – P. 130–131.

[3] *Barabási A.-L., Albert R.* Emergence of scaling in random networks // Science. – 1999. – Vol. 286. – P. 509–512.

[4] A*lbert R., Barabási A.-L.* Statistical mechanics of complex networks // Rev. Mod. Phys. – 2002. – Vol. 74. – P. 47–97.

[5] *Newman M. E. J.* The structure and function of complex networks // SIAM Rev. – 2003. – Vol. 45. – P. 167–256.

[6] *Головач Ю., Олємской О., Фербер фон К., Головач Т., Мриглод О., Олємской І., Пальчиков В.* Складні мережі // Журн. фіз. дослідж. – 2006. – Т. 10, №4. – С. 247–289.

[7] *Ferrer i Cancho R., Solé R. V.* The small world of human language // Proc. Roy. Soc. Lond. B. – 2001. – Vol. 268. – P. 2261–2265.

[8] *Dorogovtsev S. N., Mendes J. F. F.* Language as an evolving word web // Proc. Roy. Soc. Lond. B. – 2001. – Vol. 268. – P. 2603–2606.

[9] *Matsuo Y., Ohsawa Y., Ishizuka M.* KeyWorld: Extracting keywords from a document as a small world / In: Jantke K. P., Shinohara A. (Eds.). Discovery Science, 2001. Lecture Notes in Computer Science, Vol. 2226. – Berlin, Heidelberg: Springer.

[10] *Solé R. V., Corominas-Murtra B., Valverde S., Steels L.* Language networks: Their structure, function, and evolution // Complexity. – 2010. – Vol. 15. – P. 20–26.

[11] *Ferrer i Cancho R., Solé R. V., Köhler R.* Patterns in syntactic dependency networks // Phys. Rev. E. – 2004. – Vol. 69. – P. 051915 (8 pp.).

[12] *Antiqueira L., Nunes M. G. V., Oliveira Jr. O. N., Costa L. da F.* Strong correlations between text quality and complex networks features // Physica A. – 2007. – Vol. 373. – P. 811–820.

[13] *Frabska-Gradzińska I., Kulig A., Kwapień J., Drożdż S.* Complex network analysis of literary and scientific texts // Int. J. Mod. Phys. C. – 2012. – Vol. 23. – P. 1250051 (15 pp.).

[14] *Liu H., Cong J.* Language clustering with word co-occurrence networks based on parallel texts // Chin. Sci. Bull. – 2013. – Vol. 58. – P. 1139–1144.

[15] *Cong J., Liu H.* Approaching human language with complex networks // Phys. Life Rev. – 2014. – Vol. 11. – P. 598–618.

[16] *Espitia D., Larralde H.* Universal and non-universal text statistics: Clustering coefficient for language identification // Physica A. – 2020. – Vol. 553. – P. 123905 (25 pp.).

[17] *Masucci A. P., Rodgers G. J.* Network properties of written human language // Phys. Rev. E. – 2006. – Vol. 74. – 026102 (8 pp.).

[18] *Caldeira S. M. G., Petit Lobão T. C., Andrade R. F. S., Neme A., Miranda J. G. V.* The network of concepts in written texts // Eur. Phys. J. B. – 2006. – Vol. 49. – P. 523–529.

[19] *Masucci A. P., Rodgers G. J.* Differences between normal and shuffed texts: Structural properties of weighted networks // Adv. Complex Syst. – 2009. – Vol. 12. – P. 113–129.

[20] *Amancio D. R., Altmann E. G., Rybski D., Oliveira Jr. O. N., Costa L. da F.* Probing the statistical properties of unknown texts: Application to the Voynich manuscript // PLOS ONE. – 2013. – Vol. 8. – e67310 (10 pp.).

[21] *Buk S., Krynytskyi Y., Rovenchak A.* Properties of autosemantic word networks in Ukrainian texts // Adv. Complex Syst. – 2019. – Vol. 22. – 1950016 (22 pp.).

[22] *Ohsawa Y., Benson N. E., Yachida M.* KeyGraph: Automatic indexing by co-occurrence graph based on building construction metaphor / In: Proc. IEEE International Forum on Research and Technology Advances in Digital Libraries ADL'98. – Santa Barbara, USA, 1998. – pp. 12–18.

[23] *Matsumura N., Ohsawa Y., Ishizuka M.* PAI: Automatic indexing for extracting asserted keywords from a document // NGCO. – 2003. – Vol. 21. – P. 37–47

[24] *Matsuo Y., Ishizuka M.* Keyword extraction from a single document using word co-occurrence statistical information // Int. J. Artif. Intell. Tools. – 2004. – Vol. 13. – P. 157–169.

[25] *Mihalcea R., Tarau P.* TextRank: bringing order into texts / In: Proc. 2004 Conference on Empirical Methods in Natural Language Processing. – pp. 404–411.

[26] *Palshikar G. K.* Keyword extraction from a single document using centrality measures / In: Pattern Recognition and Machine Intelligence Lecture Notes in Computer Science Ed. by Ghosh A., De R. K., Pal S. K. – Vol. 4815. – Berlin, Heidelberg: Springer-Verlag, 2007. – pp. 503–510.

[27] *Rossi R. G., Marcacini R. M., Rezende S. O.* Analysis of domain independent statistical keyword extraction methods for incremental clustering // Learning and Nonlinear Models. – 2014. – Vol. 12. – P. 17–37.

[28] *Motter A. E., de Moura A. P. S., Lai Ying-Cheng, Dasgupta P.* Topology of the conceptual network of language // Phys. Rev. E. – 2002. – Vol. 65. – 065102(R) (4 pp.).

[29] *Zanette D. H.* Statistical patterns in written language / Centro Atomico Bariloche, 2012. – 87 p. – URL: http://fisica.cab.cnea.gov.ar/estadistica/2te/

[30] *Altmann E. G., Gerlach M.* Statistical laws in linguistics / In: Proc. Flow Machines Workshop: Creativity and Universality in Language. – Paris, 2014. – arXiv:1502.03296 (2015).

[31] *Kushnir O. S., Buryi V. O., Grydzhan S. V., Ivanitskyi L. B., Rykhlyuk S. V.* Zipf's and Heaps' laws for the natural and some related random texts // Електроніка та інформаційні технології. – 2018. – Вип. 9. – С. 94–105.

[32] *van Leijenhorst D. C., van der Weide Th. P.* A formal derivation of Heaps' law // Inf. Sci. – 2005. – Vol. 170. – P. 263–272.

[33] *Lü L., Zhang Z.-K., Zhou T.* Zipf's law leads to Heaps' law: Analyzing their relation in finite-size systems // PLOS ONE. – 2010. – Vol. 5. – e14139 (11 pp.).

[34] *Tria F., Loreto V., Servedio V. D. P.* Zipf's, Heaps' and Taylor's laws are determined by the expansion into the adjacent possible // Entropy. – 2018. – Vol. 20. – 752 (19 pp.).

[35] *Gerlach M., Altmann E. G.* Stochastic model for the vocabulary growth in natural languages // Phys. Rev. X. – 2013. – Vol. 3. – 021006 (10 pp.).

[36] *Font-Clos F., Corral A.* Log-log convexity of type-token growth in Zipf's systems // Phys. Rev. Lett. – 2015. – Vol. 114. – 238701 (5 pp.).

[37] *Lü L., Zhang Z.-K., Zhou T.* Deviation of Zipf's and Heaps' laws in human languages with limited dictionary sizes // Sci. Rep. – 2013. – Vol. 3. – 1082 (7 pp.).

[38] *Li W.* Random texts exhibit Zipf's-law-like word frequency distribution // IEEE Trans. Inform. Theory. – 1992. – Vol. 38. – P. 1842–1845.

[39] *Ferrer-i-Cancho R., Elvevåg B.* Random texts do not exhibit the real Zipf's law-like rank distribution // PLOS ONE. – 2010. – Vol. 5. – e9411 (10 pp.).

[40] *Walsh T.* Search in a small world / In: Proc. 16th Int. Joint Conf. on Artificial Intelligence IJCAI'99. – 1999. – Vol. 2. – pp. 1172–1177.

[41] *Luhn H. P.* The automatic creation of literature abstracts // IBM J. Res. Development. – 1958. – Vol. 2. – P. 159–165.

[42] *Rovenchak A., Rovenchak O.* Quantifying comprehensibility of Christmas and Easter addresses from the Ukrainian Greek Catholic Church hierarchs // Glottometrics. – 2018. – Vol. 41. – P. 57–66.

PROPERTIES OF LEXICAL NETWORKS BUILT ON NATURAL AND RANDOM TEXTS

O. Kushnir, A. Drebot, D. Ostrikov, O. Kravchuk

Department of Optoelectronics and Information Technologies,
Ivan Franko National University of Lviv
107 Tarnavsky Street, UA-79017 Lviv, Ukraine
oleh.kushnir@lnu.edu.ua

We study experimentally and analyze phenomenologically the properties of lexical co-occurrence networks. Our linguistic networks are built on a natural text (NT) and a random text (RT), which has been obtained after randomizing the NT on the lexical level. Another subject of our interest is a random-graph-type (RG) network having the same number of nodes as in the NT network and randomized links among those nodes. We consider non-weighted non-restricted networks where words are not filtered by their frequency and stop words are not removed.

The main parameters of the above networks are calculated and compared with the data of lexical statistics obtained for the NT, RT and RG texts. The latter data is usually expressed by so-called Zipf, Pareto and Heaps laws. For this aim an additional ‘RG text’ has been built following form the RG network. Both of the NT and RT reveal well-known power-law word frequency distributions with heavy tails. On the contrary, the lexical statistics for the RG text is characterized by a nearly-logarithmic rank–frequency dependence and a thin-tail (approximately exponential) frequency distribution.

The probability distributions for the degrees of nodes found for the NT and RT networks are close to each other. Moreover, the degree distributions for the NT (RT) and RG have respectively heavy (power-law) and thin (nearly exponential) tails. In other words, the NT and RT networks are scale-free, unlike our RG network. Moreover, this implies that a heavy (thin) tail of the degree distribution is a consequence of a heavy (thin) tail of the frequency distribution.

The average clustering coefficient and path lengths for the networks built upon the NT and RT are very close to each other. Contrary to the RG network, the NT and RT networks are small worlds and their Walsh’ parameters measuring the small-worldliness are only 2 per cents different. Finally, we analyze a number of consequences of our empirical results and some data known from the literature. In particular, since the RT lacks any semantics or syntax, it would not be proper to associate the scale-free and small-world properties of the lexical networks to either semantics or syntax.

Key words: natural language processing, linguistic networks, analysis and classification of texts, semantics recognition, random models, keyword detection.

Стаття надійшла до редакції 19.08.2024.
Прийнята до друку 05.10.2024.