ЖУРНАЛ ФІЗИЧНИХ ДОСЛІДЖЕНЬ
т. 11, № 1 (2007) с. 22–33

JOURNAL OF PHYSICAL STUDIES
v. 11, No. 1 (2007) p. 22–33

# ЛИС МИКИТА І МЕРЕЖІ МОВИ

Ю. Головач[1,2], В. Пальчиков[1]
[1] *Інститут фізики конденсованих систем НАН України, Львів, 79011, Україна*
[2] *Інститут теоретичної фізики університету Йогана Кеплера, Лінц, 4040, Австрія*
(Отримано 31 січня 2007 р.)

У статті наведено результати кількісного аналізу розподілу слів у двох текстах — творах Івана Франка "Лис Микита" та "Абу-Касимові капці". Дослідження складались із двох частин: аналіз розподілу частота–ранґ та аналіз текстів із застосуванням теорії складних мереж. Результати аналізу розподілів частота–ранґ показали, що розмір текстів є достатнім для прояву в них статистичних закономірностей. Так, степеневий характер цих розподілів (закон Зіпфа) справджується в ділянці змінної ранґ $r = 20 \div 3000$ із значенням показника степеня $\alpha \simeq 1$. Тим самим обґрунтувано вибір зазначених текстів для дослідження на їхній основі характерних особливостей української мови як складної мережі. Крім того, оцінено придатність однієї з найпростіших моделей, що приводить до виникнення степеневого розподілу частоти слів у тексті — моделі Саймона — для опису неасимптотичних властивостей розподілу слів.

У другій частині статті отримано зображення мережі української мови в різних просторах (вузлами таких мереж зазвичай виступають слова, а залежно від інтерпретації зв'язків отримуємо різні зображення мереж — різні простори) і проведено їх порівняльний аналіз. Результати, наведені в статті, переконливо свідчать про те, що мережа української мови є сильно скорельованим безмасштабним тісним світом (scale-free small world) з малим значенням середньої довжини найкоротшого шляху та високим коефіцієнтом кореляції. Отримані емпіричні результати можуть бути корисними при теоретичному описі еволюції мови.

**Ключові слова:** складні системи, мережі мови, безмасштабні мережі, закон Зіпфа.

PACS number(s): 02.10.Ox, 87.75.Da, 89.75.Hc

## I. ВСТУП

*Не прожити без наук!*[1]

Список так званих екзотичних задач статистичної фізики, у яких концепції та методи цієї науки застосовуються до нефізичних об'єктів [2], недавно поповнився за рахунок ще однієї ділянки досліджень — складних мереж (complex networks) [3, 4]. На цей раз мова йде про застосування фізичних концепцій, здебільшого тих, що сформувались у теорії фазових переходів та критичних явищ [5, 6], для опису типово математичних об'єктів — графів [7]. Незважаючи на те, що тематика, пов'язана зі складними мережами, з'явилась на сторінках фізичних журналів зовсім недавно (перші статті датуються кінцем 1990-их років), ці роботи вже набули рис цілком сформованого напрямку досліджень — див., наприклад, оглядові статті [4,8] та книжки [3, 9]. Емпіричні дослідження багатьох створених людиною та природних мереж виявили низку притаманних їм незвичних властивостей. Серед таких властивостей особливе місце займають так звані ефекти тісного світу [10] та безмасштабности [11]. Перший полягає в тому, що реальні мережі з дуже великою кількістю вузлів часто виявляються надзвичайно компактними. Відповідно, другий ефект свідчить про специфічну структуру мереж, що характеризується степеневим загасанням функції розподілу ступенів вузлів $P(k)$:

$$P(k) \sim k^{-\gamma}, \quad k > 1, \tag{1.1}$$

де показник $\gamma > 0$, а $k$ — ступінь вузла (кількість зв'язків, приєднаних до нього). Приклади безмасштабних тісних світів можна знайти серед соціяльних (співпраця, співавторство, знайомства), біологічних (метаболізм, харчування), технологічних (інтернет, електропередачі, транспорт), інформаційних (цитування, www) мереж [3, 4, 8, 9]. Останнім часом з'являються роботи, у яких концепції складних мереж застосовуються для кількісного опису мови [12–21]. Саме мережі мови будуть у центрі уваги нашого дослідження.

Мабуть, найвідомішою формулою кількісної лінґвістики є співвідношення, що пов'язує частоту $f$, з якою задане слово вживається в тексті з ранґом $r$ цього слова — його місцем у впорядкованому за спаданням частоти списку всіх слів тексту [22–24] (див. таблицю 1). Це співвідношення, відоме як закон Зіпфа, має вигляд:

$$f(r) = \frac{A}{r^\alpha}, \tag{1.2}$$

[1] Цей та всі решта епіграфів узято з Франкового "Лиса Микити" [1].

де $A$ — стала нормування, а показник степеня $\alpha$ тривалий час уважався однаковим для різних людських мов ($\alpha \simeq 1$) і незалежним від таких факторів, як автор, жанр, час написання твору тощо (див., однак, детальніше обговорення цього питання на початку розділу II). На сьогодні немає єдиної думки про те, чому розподіл слів у тексті підлягає законові Зіпфа [25]. Щобільше, виникнення цього закону можна пояснити різними механізмами [26, 27]. Суттєво, однак, те, що закон Зіпфа має статистичний характер: для його прояву необхідно, щоб текст був достатньо довгим і мав достатньо великий словник (набір різних слів). Розмір аналізованих текстів сягає від декількох тисяч слів (при аналізі окремих творів) до десятків і сотень мільйонів (при аналізі великого корпусу текстів, написаних однією мовою) [29–31]. Застосування фізичних аналогій для дослідження розподілу слів у тексті дало змогу проаналізувати ентропію слів [32], їх складність (complexity) та далекосяжні кореляції [33], марківські властивості [34, 35]. Однак таким дослідженням статистики слів притаманна одна спільна риса — нехтування зв'язками між словами. А саме зв'язки та організація слів у речення дозволяють ефективно передавати інформацію [12]. Застосування теорії складних мереж до аналізу текстів якраз і покликане проаналізувати такі зв'язки, а отже і виявити нові закономірості в структурі мови.

| $r$ | $f$ | слово | $r$ | $f$ | слово |
|---|---|---|---|---|---|
| 1 | 439 | я | 1 | 165 | він |
| 2 | 323 | не | 2 | 163 | в |
| 3 | 312 | в | 3 | 143 | не |
| 4 | 272 | і | 4 | 140 | і |
| 5 | 233 | ти | 5 | 128 | той |
| 6 | 222 | що | 6 | 125 | що |
| 7 | 214 | на | 7 | 125 | на |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 16 | 140 | лис | 12 | 87 | капець |
| 21 | 109 | Микита | 18 | 69 | Абу-Касим |
| 23 | 98 | вовк | 40 | 28 | пан |
| 25 | 88 | цар | 41 | 27 | суддя |

Таблиця 1. Приклади впорядкованих за ранґом $r$ слів із творів "Лис Микита" (ліва частина таблиці) та "Абу-Касимові капці" (права частина). $f$ — кількість поя слова в тексті.

У цій статті ми подамо результати аналізу методами теорії складних мереж двох творів Івана Франка: "Лис Микита" [1] та "Абу-Касимові капці" [36]. При цьому ставимо перед собою два різні завдання: одне зумовлює вибір об'єкта вивчення — аналіз україномовного тексту методами теорії складних мереж, наскільки нам відомо, проводиться вперше. Друге завдання загальніше: дотепер різні дослідження мереж мови проводили в межах якогось одного з можливих зображень тексту як мережі (одного з "просторів", детальніше описаних у розділі III). У нашій роботі ми застосуємо різні зображення і проведемо їх порівняльний аналіз.

Подальша структура статті така: у розділі II ми розпочнемо кількісний аналіз двох згаданих текстів із дослідження розподілів частота–ранґ. Перевіривши, чи виконується для них закон Зіпфа (1.2), отримаємо в такий спосіб відповідь на питання, чи достаньою є довжина обраних текстів для прояву статистичних закономірносей. Крім цього, предметом розділу II буде опис ґенерації тексту за допомогою моделі Саймона. У розділі III розглянемо тексти як складні мережі; проаналізуємо різні зображення текстів у вигляді мереж; отримаємо типові кількісні характеристики цих мереж; дослідимо ефекти тісного світу та безмасштабности. Висновки подамо в розділі IV.

## II. ЗАКОН ЗІПФА: ЕМПІРИЧНІ РЕЗУЛЬТАТИ ТА МОДЕЛЬ САЙМОНА

*Ну-бо, хлопці, поспішімо,*
*Землю міряти ходімо!*
*Маєте кілки, дрючки?*

### A. Співвідношення частота–ранґ

Застосування методів кількісного аналізу, що використовуються в природничих науках, для дослідження мови привело, зокрема, до відкриття закону частотної залежности слів у тексті (1.2). Традиційно вважається, що це відкриття належить ґарвардському лінґвістові Дж. К. Зіпфу (1935, [23]), хоча раніше подібні міркування висловлювали Ж. Б. Есту (J. B. Estoup, 1916) та Е. У. Кондон (E. U. Condon, 1928) [28] (докладніше див. [27]). Те, що слова пов'язані *локально* у межах одного речення за допомогою граматичних правил, є очевидним. Однак те, що такі локальні послідовності організовуються *глобально* і приводять до розподілів степеневого типу, є аж ніяк не тривіяльним [34]. Сама форма розподілу (1.2) піддавалась різним модифікаціям, із яких найвідомішим є закон Зіпфа–Мандельброта (див. [30] та поклики в цій статті). Зокрема, відхилення від закону Зіпфа в його класичній формі (1.2) спостерігалось для змішаного корпусу текстів, що належали різним авторам [30], та коли апроксимація степеневою функцією проводилась на всій ділянці змінної ранґ [30, 34]. Однак спостережено, що для одиночних текстів співвідношення (1.2) з достатньою точністю описує залежність частота–ранґ у діяпазоні $r = 100 \div 2000$.

Як зазначено у Вступі, нашою метою є кількісний аналіз тексту (фактично, двох текстів) одного автора [1, 36]. Природно розпочати цей аналіз з перевірки співвідношення (1.2) як для кожного тексту окремо, так і об'єднуючи їх разом. Зокрема, це дозволить переконатися, чи довжина обраних текстів є достатньою для прояву в них статистичних закономірностей.

Для аналізу ми скористались електронною версією творів [1, 36]. Усі слова в текстах були зведені до словникової форми. Довжина (загальна кількість слів) обраних текстів становить $\mathcal{N} = 15426; 8002$, а їх словник (кількість різних слів) $\mathcal{V} = 3563; 2392$ — для "Лиса Микити" та "Абу-Касимових капців" відповідно. Об'єднуючи ці два тексти разом, отримуємо $\mathcal{N} = 23428$ і $\mathcal{V} = 4823$. Наступним кроком є підрахунок кількости появ слів у тексті — надалі ототожнюватимемо кількість появ слова в тексті з його частотою $f$ (насправді ці дві величини різняться множником нормування $\mathcal{N}$). Упорядкувавши список слів за спаданням частоти, присвоюємо словам значення змінної ранґ: $r = 1, 2, \ldots, \mathcal{V}$. Причому, якщо декілька різних слів характеризуються однаковою частотою, то їх впорядкування за змінною $r$ відбувається випадково. Слова перших ранґів — це ті, що вживаються найчастіше. Для літературної української мови — це функціональні слова (прийменники, частки, сполучники) та деякі займенники. Після них з'являються слова, пов'язані з контекстом твору. Приклади впорядкованих за ранґом слів із аналізованих текстів наведено в таблиці 1.

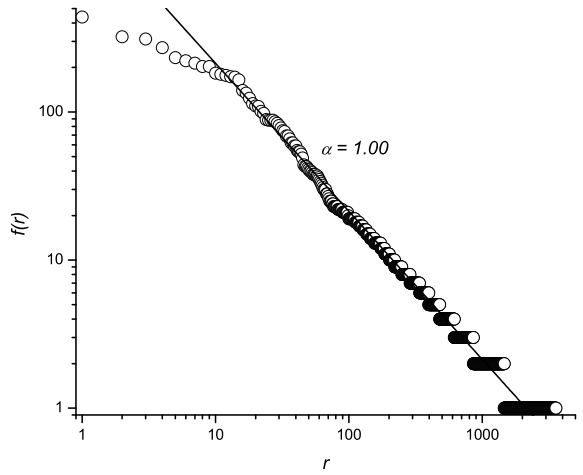


Рис. 1. Залежність частота–ранґ для "Лиса Микити". Суцільна пряма — апроксимація степеневою функцією (1.2) із показником $\alpha = 1.00$.

На рис. 1 показано залежність частота–ранґ для розподілу слів у "Лисі Микиті". Як видно з рисунка, на проміжку $r = 20 \div 3000$ функція $f(r)$ описується степеневим законом (1.2). Показник степеня, який ми визначили методом найменших квадратів за нахилом прямої $f(r)$, побудованої в подвійному логарифмічному масштабі, виявився рівним $\alpha = 1.00$. При цьому кожному значенню $f(r)$ ставилося у відповідність лише одне значення $r$ — усереднене в логарифмічному масштабі. Подібні степеневі залежності отримані також і для "Абу-Касимових капців" та для двох творів, об'єднаних разом (зі значеннями $\alpha = 0.97$ та $\alpha = 1.00$ відповідно). Типова точність визначення показника $\alpha$ при апроксимації $f(r)$ степеневою функцією становить $\chi^2/d.o.f = 0.002$.

Звернімо увагу на ще одну характерну особливість залежности $f(r)$, зображеної на рис. 1: кількість різних слів, що мають однакову частоту $f$, зростає з $r$. Щобільше, кількісний аналіз показує, що це зростання також описується степеневою залежністю [22, 23]:

$$N(f) = \frac{B}{f^\beta}, \tag{2.3}$$

де $N(f)$ — кількість різних слів, кожне з яких уживається в тексті $f$ разів, $B$ — стала нормування. Розподіл (2.3) деколи називають другим законом Зіпфа (відповідно, (1.2) — перший закон). Слід, однак, зауважити, що перший і другий закони не є незалежними: (2.3) приводить до (1.2). У цьому можна безпосередньо переконатись, задавши (2.3) і звернувши увагу на те, що довжина тексту (кількість слів у тексті) $\mathcal{N}$ та його словник (кількість різних слів) $\mathcal{V}$ однозначно визначаються через $N(f)$:

$$\mathcal{N} = \sum_{f=f_{\min}}^{f_{\max}} fN(f) = \sum_{f=f_{\min}}^{f_{\max}} \frac{B}{f^{\beta-1}}, \tag{2.4}$$

$$\mathcal{V} = \sum_{f=f_{\min}}^{f_{\max}} N(f) = \sum_{f=f_{\min}}^{f_{\max}} \frac{B}{f^\beta}. \tag{2.5}$$

У (2.4), (2.5) підсумовування ведеться за всіма частотами $f$: від мінімальної $f_{\min}$ (як правило, $f_{\min} = 1$) до максимальної $f_{\max}$. Залишається знайти зі співвідношень (2.3)–(2.5) ті значення ранґу $r$, що відповідають однаковим частотам $f$. Отримана залежність $f(r)$ і буде співвідношенням (1.2). Як відомо ще з робіт Зіпфа [23], числове значення показника $\beta \simeq 2$. Однак спостерігаються і відхилення, спричинені особливостями лексикону (див., наприклад, статтю [25] і поклики в ній). Для наших подальших досліджень важливо, однак, те, що степеневі закони (1.2) і (2.3) пов'язані один з одним і значення показника $\alpha \simeq 1$, отримане на підставі аналізу залежности частота–ранґ, рис. 1, пов'язане зі значенням $\beta \simeq 2$.

На сьогодні немає єдиного пояснення, чому розподіл слів у тексті підлягає степеневому законові. Однією з перших моделей, покликаних пояснити це явище, була модель Саймона [37]. Вона належить до класу моделей, що базуються на так званих нульових гіпотезах (null hypothesis) [25]. Нульові гіпотези не беруть до уваги певні фундаментальні аспекти того, як і чому використовуються слова (наприклад, що слова використовуються згідно з їх значенням). Як фізичну аналогію можна взяти модель середнього поля в теорії критичних явищ [5]: нехтуючи важливими для критичної поведінки факторами, ця модель, тим не менше, часто приводить до якісно правильного опису степеневих залежностей спостережуваних величин [5, 6]. Нижче ми опишемо, як відбувається ґенерація тексту за моделлю Саймона, наведемо аналітичні обчислення асимптотики розподілу слів та порівняємо відповідні характеристики, отримані емпірично і в результаті чисельного моделювання.

### В. Модель Саймона

Г. А. Саймон у 1955 році [37] запропонував таку модель для пояснення виникнення степеневих функцій розподілу, що характеризують різноманітні явища, і зокрема, розподіл слів у тексті. Нехай текст досяг розміру $n$ слів. Тоді те, яким буде $n+1$-е слово тексту, ґрунтується на двох припущеннях:

- *Припущення 1.* Нехай $N(f,n)$ — кількість різних слів, кожне з яких ужито $f$ разів серед перших $n$ слів тексту. Тоді ймовірність того, що $(n+1)$-им є слово, яке попередньо вжито $f$ разів, пропорційна до $fN(f,n)$ — загальної кількости появи всіх слів, кожне з яких ужито до цього моменту $f$ разів.
- *Припущення 2.* Існує постійна ймовірність $\delta$ того, що $(n+1)$-им словом буде нове слово — слово, яке не траплялося жодного разу серед перших $n$ слів.

Цих двох припущень виявляється достатньо, щоб показати, що функція розподілу слів у тексті, зґенерованому на їх підставі, підлягає степеневому законові [27, 37]. Покажемо це нижче, згідно з роботою [37]. Справді, з Припущення 1 виходить, що

$$N(f,n+1)-N(f,n)=K(n)[(f-1)N(f-1,n) - fN(f,n)], \qquad f=2,...,n+1, \tag{2.6}$$

де $K(n)$ — коефіцієнт пропорційности. Справді, перший доданок у правій частині дорівнює ймовірності появи слова, яке вжито серед перших $n$ слів $f-1$ разів. Він збільшує кількість слів, які вжито $f$ разів у корпусі з $n+1$ слів. Якщо ж $(n+1)$-им словом буде слово, яке серед $n$ перших слів ужито рівно $f$ разів, то серед перших $(n+1)$ слів, кількість слів, які вжито рівно $f$ разів, зменшиться на 1, за що відповідає другий доданок. Аналогічно, Припущення 2 приводить до рівняння:

$$N(1,n+1)-N(1,n)=\delta-K(n)N(1,n). \tag{2.7}$$

Враховуючи, що $K(n)fN(f,n)$ — ймовірність того, що $(n+1)$-им буде слово, яке серед $n$ перших слів вжито $f$ разів, а $\delta$ — ймовірність появи нового слова, отримуємо

$$\sum_f K(n)fN(f,n)+\delta=K(n)\sum_f fN(f,n)+\delta=1. \tag{2.8}$$

Рівність (2.8) — це ймовірність достовірної події (появи $(n+1)$-го слова). Враховуючи, що $\sum_f fN(f,n)$ — загальна кількість слів у корпусі з $n$ слів, тобто $\sum_f fN(f,n)=n$, з (2.8) отримуємо:

$$K(n)=\frac{1-\delta}{n}. \tag{2.9}$$

Щоб полегшити розрахунки, припустімо, що

$$\frac{N(f,n+1)}{N(f,n)}=\frac{n+1}{n} \qquad \text{для всіх } n \text{ та } f. \tag{2.10}$$

Співвідношення (2.10) є умовою пропорційного зростання функції $N(f,n)$ зі збільшнням розміру тексту $n$. З нього випливає, що

$$\frac{N(f,n)}{N(f-1,n)}=\frac{N(f,n+1)}{N(f-1,n+1)}=\varphi(f), \tag{2.11}$$

де $\varphi(f)$ не залежить від $n$. Підставивши в рівняння (2.6) співвідношення (2.9), (2.10) та (2.11), отримуємо:

$$\Big(\frac{n+1}{n}-1\Big)N(f,n)=\frac{1-\delta}{n}\Big(\frac{f-1}{\varphi(f)}-f\Big)N(f,n). \tag{2.12}$$

Звідки, скоротивши на спільний множник, маємо

$$\varphi(f)=\frac{(1-\delta)(f-1)}{1+(1-\delta)f}, \qquad f=2,\dots,n. \tag{2.13}$$

Щоб одержати залежність (2.3), перейдімо до функцій $N^*(f)=N(f,n)/n$, які відрізняються від $N(f)$, запроваджених у (2.3), лише множником нормування. Згідно зі співвідношенням (2.10), функція $N^*(f)$ не залежить від $n$. Тоді рівняння (2.11) можна записати у вигляді

$$\frac{N^*(f)}{N^*(f-1)}=\varphi(f). \tag{2.14}$$

Застосувавши рекурентне співвідношення (2.14) $f-1$ разів, отримуємо

$$N^*(f)=\varphi(f)N^*(f-1)=\varphi(f)\varphi(f-1)\dots\varphi(2)N^*(1). \tag{2.15}$$

Для зручности введемо позначення

$$\rho=\frac{1}{1-\delta}, \qquad 1<\rho<\infty. \tag{2.16}$$

Підставивши у (2.15) функцію $\varphi(f)$, (2.13), одержуємо

$$N^*(f)=\frac{(f-1)(f-2)\dots2\cdot1}{(f+\rho)(f+\rho-1)\dots(\rho+2)}N^*(1) = \frac{\Gamma(f)\Gamma(\rho+2)}{\Gamma(f+\rho+1)}N^*(1), \qquad f=2,\dots,n. \tag{2.17}$$

Скориставшись властивістю Г-функції $\frac{\Gamma(f)}{\Gamma(f+\rho)}\sim f^{-\rho}$, $f\to\infty$ [38], маємо асимптотичну оцінку

$$N^*(f)\sim f^{-(1+\rho)}. \tag{2.18}$$

Порівнюючи (2.18) та (2.3), знаходимо співвідношення між $\beta$ та $\rho$

$$\beta=1+\rho, \tag{2.19}$$

з якого з урахуванням (2.16) випливає оцінка $\beta\simeq2$ для малих значень імовірности появи в тексті нових слів $\delta$.

### С. Порівняльний аналіз

Перевіримо, наскільки ґенерація тексту, відповідно до припущень моделі Саймона, відповідає емпіричним результатам, що ми спостерігали в підрозділі II A. Для цього, взявши за основу перші 2000 слів тексту твору "Лис Микита", ($n = 2000$), ґенеруватимемо решту тексту, згідно з Припущеннями 1 і 2 моделі. Ймовірність появи нових слів ми визначили емпірично, і вона виявилась рівною $\delta \simeq 0,22$.

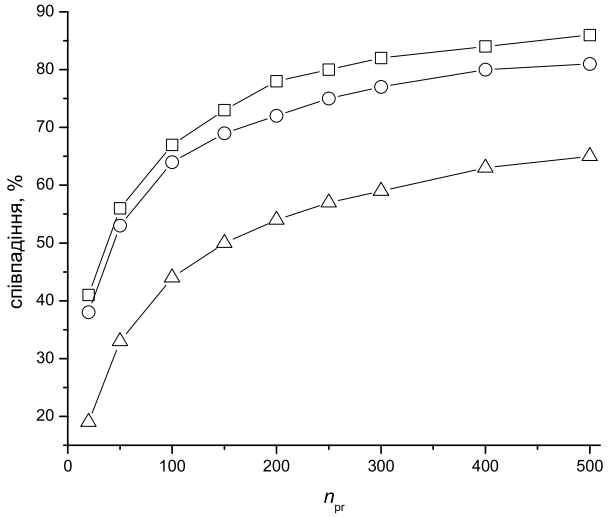


Рис. 2. Залежність відсотка збігу двох текстів як функція розміру передбачуваного блоку слів $n_{\rm pr}$. ○ — порівняння реального тексту-оригіналу та тексту, зґенерованого згідно з моделлю Саймона. △ — порівняння тексту, зґенерованого згідно з моделлю Саймона з текстом, зґенерованим випадково. □ – порівняння двох текстів, побудованих за припущеннями моделі Саймона.

Здійснюватимемо крок за кроком порівняння зґенерованого тексту з текстом-оригіналом. Порівнюватимемо блоки слів розміром $n_{\rm pr} = 20 \div 500$. Результати такого порівняння наведені на рис. 2. Зауважимо, що в моделі Саймона фіґурують не конкретні слова, а частоти їх появ $f$: одному й тому ж значенню частоти появи можуть відповідати декілька різних слів. Тому на кожному кроці ми ґенеруємо не слово, а частоту $f$. Так ґенерується множина з $n_{\rm pr}$ чисел — частот появ слів. Зіставленняння цієї зґенерованої за правилами моделі Саймона множини з аналогічною множиною, що відповідає текстові-оригіналу, і знаходження відсотка їхнього перекриття є першим кроком до оцінки придатности моделі. Залежність відсотка збігу тексту-оригіналу та зґенерованого тексту від розміру блоку слів $n_{\rm pr}$ зображена на рис. 2 кружечками. Для малих значень $n_{\rm pr}$ відсоток збігів сильно залежить від розміру передбачуваного блоку, однак цей відсоток зростає і виходить на насичення при достатньо великому розмірі $n_{\rm pr}$, коли ліпше проявляються статистичні закономірності.

Як видно з рис. 2, згадане вище насичення у збігові частот слів тексту-оригіналу та тексту, зґенерованого за моделлю Саймона, становить величину близько $\sim 80\%$. Пересвідчимось у тому, що такий збіг (отриманий для тексту порівняно невеликого розміру) свідчить про адекватність моделі. Для цього, взявши за основу перші $n$ слів реального тексту, спробуємо ґенерувати його зростання, вибираючи наступні після $n$-того слова́ незалежно від того, з якою частотою вони вживалися раніше. Оцінимо відсоток збігу такого тексту (надалі називатимемо його випадково зґенерованим) із текстом, зґенерованим за моделлю Саймона. Щоб зробити це, зґенеруємо випадково блок тексту розміром $n_{\rm pr}$ на основі тих самих 2000 перших слів твору "Лис Микита", поклавши, що ймовірність появи нових слів дорівнює $\delta = 0.22$. Імовірності появи кожного зі слів, які вжиті в блоці 2000 перших слів, покладаються однаковими, їх сума дорівнює $1-\delta$. Відсоток збігу отриманого таким чином тексту зі зґенерованим на підставі припущень Саймона зображено на рис. 2 трикутниками. Як бачимо з рисунка, максимальне значення відсотка збігу частот слів у текстах, зґенерованих випадково і за моделлю Саймона, не таке й мале ($\sim 60\%$). Проте зростання збігів частот на величину близько $\sim 20\%$ при переході до опису ґенерації тексту за моделлю Саймона (кружечки на рис. 2) свідчить про адекватність моделі Саймона вже для опису ґенерації текстів невеликих розмірів.

З іншого боку, годі сподіватися повного збігу частот слів у тексті-оригіналі та в зґенерованому тексті для блоку тексту скінченного розміру і для всіх значень частот. "Ефект скінченного розміру" спостерігається навіть при порівнянні двох текстів, зґенерованих згідно з моделлю Саймона, як це показано квадратиками на рис. 2. На основі того ж корпусу тексту, як і в попередніх випадках, ми зґенерували відповідно до моделі Саймона два блоки розміром по $n_{\rm pr}$ слів. Відсоток збігу частот слів у цих блоках можна вважати верхньою межею можливих збігів і прийняти їх за умовні 100%. Узявши за умовні 0% відсоток збігу випадкового та зґенерованого за моделлю Саймона текстів (годі очікувати меншого збігу реального тексту зі зґенерованим за припущеннями моделі Саймона), отримаємо шкалу, за якою можна оцінити придатність моделі для опису ґенерації реального тексту. Вона виявилася величиною близько $\sim 80\%$, що свідчить про реалістичність моделі щодо ґлобальної тенденції росту текстів.

## III. АНАЛІЗ ЗВ'ЯЗКІВ МІЖ СЛОВАМИ МЕТОДАМИ ТЕОРІЇ СКЛАДНИХ МЕРЕЖ

*Де лиш сіть свою закину,*
*Маю здобич повсякчас.*

### А. Текст як складна мережа

Першим кроком при застосуванні теорії складних мереж до аналізу тексту є, власне, зображення цього

тексту у вигляді сукупности вузлів і зв'язків — мережі мови (language network). Є різні способи інтерпретації вузлів і зв'язків, що приводять, відповідно, до різних зображень мережі мови. Як правило, вузлами такої мережі є слова. Проте відомо багато способів вибору зв'язків між вузлами. Так, вузли можуть бути поєднаними між собою, якщо слова, що їм відповідають, стоять у тексті поруч [12,13] чи належать до одного речення [14], поєднані синтаксично [15–18] або семантично [19–21]. У нашому дослідженні ми застосуємо декілька зображень мережі мови й покажемо зв'язок між ними. Як і в попередньому розділі, для аналізу ми скористаємось електронною версією творів [1,36], спершу звівши всі слова до словникової форми. Таким чином, ми не досліджуватимемо синтаксичних зв'язків між словами безпосередньо. Збереження синтаксичних зв'язків приводить до зображення тексту у вигляді спрямованої мережі (directed network), де напрямок зв'язку відповідає підпорядкуванню слова [15,16]. Зауважимо, що завдання, яке ми ставимо перед собою, полягає не у встановленні зв'язків між словами окремого речення. Нашим завданням є виявити закономірності між словами мови, які мали би проявитися при аналізі достатньо великого тексту.

Для подальшого аналізу скористаємось такими способами зображення тексту у вигляді мережі. Поставимо у відповідність кожному слову вузол мережі (надалі не братимемо до уваги розділові знаки всередині речення). Поєднаємо кожні два вузли зв'язком, якщо відповідні їм слова стоять у реченні поряд. Таке зображення називатимемо $L$-простором (див. рис. 3**а**). У $L$-просторі, так само як і в інших зображеннях, перелічених нижче, при виникненні кратних зв'язків залишатимемо лише один з них.

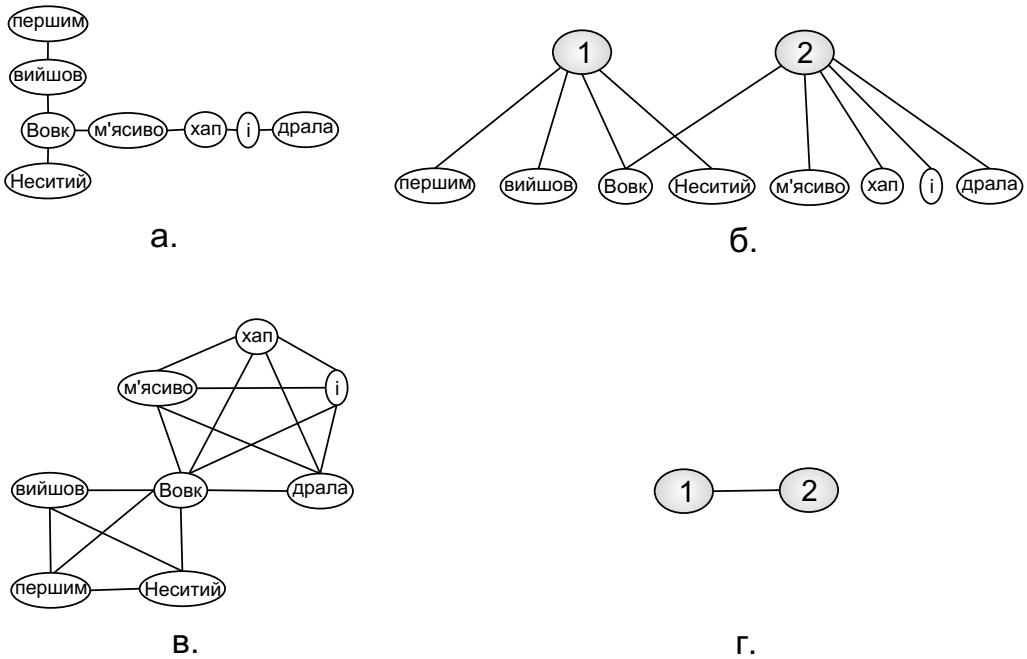


Рис. 3. Графічне зображення двох речень, 1: "Першим вийшов Вовк Неситий", 2: "Вовк м'ясиво хап — і драла!". **а.** $L$-простір. Сусідні слова, що належать до одного речення, пов'язані зв'язками. Кількість сусідів для кожного слова (вікно слова) визначається "радіусом взаємодії" $R$, в наведеному прикладі $R = 1$. **б.** $B$-простір. Наявні вузли двох сортів. Темні вузли — речення, світлі вузли — слова, що до них належать. **в.** $P$-простір. Усі слова, що належать до одного речення, пов'язані між собою. **г.** $C$-простір. Речення пов'язані між собою, якщо в них ужиті однакові слова. Зв'язок між вузлами-реченнями (1 і 2) відповідає слову "вовк", спільному для обох речень.

Не пояснюючи причини виникнення зв'язків між сусідніми словами в реченні (вони можуть бути викликані синтаксисом мови, існуванням у мові усталених виразів тощо), утворена таким чином мережа дозволяє їх графічне зображення і кількісний аналіз. Однак зв'язки можуть поєднувати не лише "найближчих сусідів", а групи з кількох слів, що перебувають на певній відстані одне від одного. Щобільше, слова, що стоять поряд у реченні, можуть часом і не мати таких зв'язків. Для часткового врахування цих фактів уведемо "радіус взаємодії" $R$: при $R = 1$ зв'язок існує лише між найближчими сусідами, при $R = 2$ — між найближчими й наступними близькими сусідами і т. д. Зрозуміло, що змінна $R$ може пробігати значення $R = 1, \ldots, R_{\max}$, де $R_{\max} + 1$ — загальна кількість слів у реченні. Для $R = 1$ уведений вище $L$-простір відповідає зображенню мережі мови, що було запропоновано в роботі [12].

Ще один спосіб зобразити текст у вигляді мережі полягає у використанні двосортних (bipartite) графів [7,40]. У такому зображенні (називатимемо його $B$-простором, рис. 3**б**) наявні вузли двох сортів. Один сорт (темні вузли на рисунку) відповідає реченням, інший (світлі вузли) — словам. Зв'язок між темним і

світлим вузлами означає, що слово належить до цього речення. Одномодові проєкції двосортного графа дозволяють отримати ще два зображення мережі мови, які називатимемо $P$- та $C$-просторами (рис. 3 **в** та **г** відповідно). У $P$-просторі всі слова, що належать до одного речення, вважаються пов'язаними між собою. Таким чином, кожне окреме речення входить у мережу як повний граф, так звана кліка (clique). У $C$-просторі вузли відповідають реченням, а зв'язок між вузлами-реченнями ставиться тоді, коли в них є спільні слова. Зображення, подібне до запровадженого вище $P$-простору, використовувалося для аналізу так званої мережі концепцій у мові [14].

Номенклатуру $L$-, $B$-, $P$-, $C$-просторів ми взяли з робіт про дослідження транспортних мереж [39, 40], що належать до технологічних мереж. Як переконаємося нижче, характеристики мереж мови і технологічних мереж суттєво різняться. Важливим є географічний чинник, що накладає обмеження на еволюцію транспорних мереж (так, транспортні мережі, як правило, реалізуються у двовимірному просторі вкладення). Однак застосування зображень, що використовувались у транспорних мережах, для опису мереж мови виявляється корисним. Відомі праці, у яких мережі мови описувалися в $P$- [14] або в $L$- [12] просторах. Запровадивши параметр $R$, що характеризує так зване вікно слів (word window), пов'язаних між собою в межах одного речення, ми тим самим указуємо на можливість отримати характеристики мережі мови в $P$-просторі як граничний випадок відповідних характеристик в $L$-просторі (при $R = R_{\max}$).

Є ряд стандартних характеристик, якими прийнято описувати мережу. До них належить розподіл ступенів вузлів $P(k)$, середній ступінь вузла $\langle k\rangle$, коефіцієнт кластерности $\langle C\rangle$ та середня довжина найкоротшого шляху між парами вузлів $\langle L\rangle$. Щоб зрозуміти зміст цих характеристик, уведемо кілька означень.

Ступенем $k_i$ вузла $i$ називається кількість зв'язків, приєднаних до нього. Знаючи ступені кожного вузла, можна побудувати розподіл ступенів вузлів $P(k)$ — ймовірність того, що випадково взятий вузол мережі має ступінь $k$. Вибираючи відповідну умову нормування, можна розглядати $P(k)$ як кількість вузлів, що мають ступінь $k$. Як зазначено у Вступі, для багатьох природних та штучно створених мереж розподіли ступенів є степеневими — описуються співвідношенням (1.1). Мережі з таким розподілом називають безмасштабними (scale-free) у зв'язку з відсутністю масштабної характеристики цього розподілу. Середнім ступенем вузла мережі $\langle k\rangle$ називають усереднений ступінь усіх її вузлів

$$\langle k\rangle = \frac{1}{\mathcal{V}}\sum_{i=1}^{\mathcal{V}} k_i, \tag{3.20}$$

де $\mathcal{V}$ — кількість вузлів мережі.

Для кількісного опису локальної "згуртованости" вузлів уводимо коефіцієнт кластерности. Коефіцієнт кластерности $C_i$ вузла $i$ дорівнює ймовірності того, що два випадково вибрані сусіди цього вузла мають зв'язок між собою:

$$C_i = \frac{2E_i}{k_i(k_i-1)}, \tag{3.21}$$

де $E_i$ — кількість наявних зв'язків між всіма парами найближчих сусідів $i$-того вузла. Відповідною ґлобальною характеристикою, що описує не окремий вузол, а всю мережу і, зокрема, характеризує її кореляційні властивості, є середнє значення коефіцієнта кластерности:

$$\langle C\rangle = \frac{1}{\mathcal{V}}\sum_{i=1}^{\mathcal{V}} C_i. \tag{3.22}$$

Довжиною найкоротшого шляху $l_{ij}$ між вузлами $i$ та $j$ називається мінімальна кількість зв'язків, яку необхідно "пройти", починаючи з вершини $i$, щоб досягнути вершини $j$. Середня довжина найкоротшого шляху

$$\langle l\rangle = \frac{2}{\mathcal{V}(\mathcal{V}-1)}\sum_{i>j} l_{ij} \tag{3.23}$$

визначає розмір мережі. Ще однією величиною, що характеризує розмір мережі, є максимальне значення довжини найкоротшого шляху між парами вершин $l_{\max}$.

Увівши основні характеристики кількісного опису мереж, обчислимо тепер їх значення для вибраних текстів [1,36]. Надалі зосередимося на зображенні цих текстів у $L$- та $P$-просторах, як пояснено вище.

### В. $L$-простір

Нагадаємо, що в цьому зображенні (див. рис. 3**а**) поєднуються зв'язками сусідні слова, які належать до одного речення. Причому кількість сусідніх слів, між якими проводяться зв'язки, визначається параметром $R$: при $R = 1$ пов'язані між собою лише найближчі сусіди, при $R = 2$ зв'язки проводяться між вузлом-словом, його найближчими та наступними близькими сусідами і т. д. У таблиці 2 наведено основні характеристики мереж, отримані при аналізі творів [1,36] при значеннях параметра $R = 1, 2, R_{\max}$. Як ми уже зауважили вище, при $R = R_{\max}$ зображення мережі в $L$- і $P$-просторах, а отже і їх кількісні характеристики, збігаються. Детальніше на властивостях мережі при $R = R_{\max}$ ми зупинимось у наступному підрозділі III C. Зміна деяких характеристик мережі зі зростанням $R$ показана на рис. 4. Зокрема, на рисунку наведено нормовані значення середнього та максимального ступенів вузлів як функції $R$: $\langle k^*(R)\rangle \equiv \langle k(R)\rangle/\langle k(R_{\max})\rangle$, $\langle k^*_{\max}(R)\rangle \equiv \langle k_{\max}(R)\rangle/\langle k_{\max}(R_{\max})\rangle$. Зростання "радіуса взаємодії" $R$ приводить до зростання кількости зв'язків. Таким чином, середнє та максимальне значення $k$ зростають, досягаючи насичення при $R = R_{\max}$.

| $R$ | $\mathcal{V}$ | $\mathcal{M}$ | $\langle k\rangle$ | $\langle k^2\rangle/\langle k\rangle$ | $k_{\max}$ | $\gamma$ | $\gamma_{\rm int}$ | $\langle C\rangle$ | $\langle C\rangle/C_r$ | $\langle l\rangle$ | $l_{\max}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2392 | 6273 | 5.24 | 48 | 228 | 1.9 | 1.15 | 0.171866 | 78 | 3.428 | 11 |
| 2 | 2392 | 11475 | 9.59 | 77 | 391 | 2.0 | 1.18 | 0.567195 | 141 | 2.897 | 7 |
| $R_{\max}$ | 2392 | 48603 | 40.64 | 208 | 1134 | 1.9 | 1.35 | 0.841215 | 50 | 2.220 | 4 |
| 1 | 3563 | 11102 | 6.23 | 76 | 419 | 1.9 | 1.12 | 0.214024 | 122 | 3.301 | 11 |
| 2 | 3563 | 20063 | 11.26 | 119 | 665 | 1.8 | 1.16 | 0.587752 | 186 | 2.852 | 7 |
| $R_{\max}$ | 3563 | 65997 | 37.05 | 269 | 1526 | 1.9 | 1.27 | 0.821895 | 79 | 2.274 | 5 |
| 1 | 4823 | 16580 | 6.88 | 102 | 537 | 1.9 | 1.13 | 0.243097 | 170 | 3.235 | 11 |
| 2 | 4823 | 29916 | 12.41 | 156 | 868 | 1.8 | 1.15 | 0.585375 | 227 | 2.826 | 7 |
| $R_{\max}$ | 4823 | 107750 | 44.68 | 360 | 2185 | 2.0 | 1.28 | 0.818495 | 88 | 2.249 | 5 |

Таблиця 2. Кількісні характеристики мереж досліджуваних текстів в $L$-просторі для кількох значень $R$. Верхня частина таблиці — мережа тексту "Абу-Касимові капці" [1], середня частина — "Лис Микита" [36], нижня частина — обидва твори разом. $\mathcal{V}$ — кількість вузлів, $\mathcal{M}$ — кількість зв'язків, $\langle k\rangle$, $k_{\max}$ — середній і максимальний ступені вузла, $\gamma$, $\gamma_{\rm int}$ — показники звичайного (1.1) та інтеґрального (3.24) розподілів ступенів вузлів, $\langle C\rangle$ — середнє значення коефіцієнта кластерности (3.22), $C_r$ — коефіцієнт кластерности відповідного випадкового графа Ердоша–Рені, $\langle l\rangle$, $l_{\max}$ — середнє та максимальне значення довжини найкоротшого шляху. Зауважимо, що при $R = R_{\max}$ зображення мережі в $L$- і $P$-просторах збігаються.

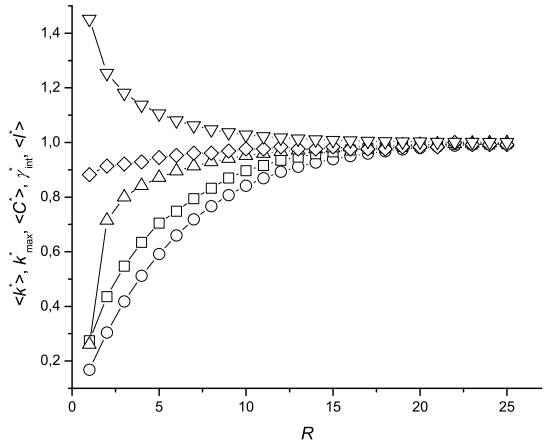


Рис. 4. "Лис Микита": залежність нормованих характеристик мережі від значення $R$: ○ — $\langle k^*\rangle$; □ — $k^*_{\max}$; △ — $\langle C^*\rangle$; ◇ — $\gamma^*_{\rm int}$; ▽ — $\langle l^*\rangle$.

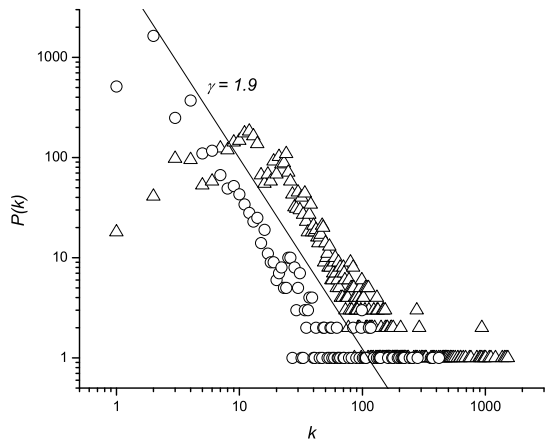


Рис. 5. "Лис Микита": розподіл ступенів вузлів. $R = 1$ (○-○-○), $R = R_{\max}$ ( △-△-△), суцільна пряма — степенева функція (1.1) із показником $\gamma = 1.9$.

На рис. 5 наведено розподіл ступенів вузлів $P(k)$ — кількість вузлів мережі, що мають однаковий ступінь $k$ — для "Лиса Микити". Видно, що в подвійному логарифмічному масштабі цей розподіл апроксимується прямою лінією, що свідчить про степеневе загасання функції $P(k)$ зі зростанням $k$. Також спостерігається характерний максимум у ділянці $k \sim 10$. Подібні залежності отримані і для тексту "Абу-Касимові капці" та для двох творів разом. Значення показника степеня, що характеризує загасання (1.1), коливається в межах $\gamma = 1.9 \div 2.0$. Однак аналізована база даних не дозволяє зробити точнішої оцінки. Для того, щоб пересвідчитися в степеневій поведінці, знайдемо так званий інтеґральний розподіл ступенів вузлів $P_{\rm int}(k)$ (cumulative node degree distribution):

$$P_{\rm int}(k) = \sum_{k'=k}^{k_{\max}} P(k'). \qquad (3.24)$$

Відповідна залежність показана на рис. 6.

Функція $P_{\rm int}(k)$ є значно гладшою, ніж $P(k)$, і дає змогу зробити однозначний висновок про степеневий характер функції розподілу. Значення показника, що характеризує загасання функції інтеґрального розподілу ступенів вузлів $P_{\rm int}(k) \sim 1/k^{\gamma_{\rm int}}$, наведено в таблиці 2. Типова точність визначення $\gamma_{\rm int}$ при апроксимації $P_{\rm int}(k)$ степеневою функцією становить $\chi^2/d.o.f = 0.001$. Залежність показника $\gamma^*_{\rm int}(R) = \gamma_{\rm int}(R)/\gamma_{\rm int}(R_{\max})$ від $R$ зображена на рис. 4.

Як видно із проведеного аналізу, висновок про степеневу поведінку (1.1) можна зробити навіть на підставі невеликого корпусу текстів. Для порівняння вкажемо, що для англійської мови подібний аналіз проводився на підставі так званого Британського національного корпусу [41]. Аналіз цих текстів із $\mathcal{N} \sim 10^7$ слів показав [12], що досліджувана мережа англійської мови безмасштабна, а степенева поведін-

ка $P(k)$ характеризується двома режимами зі значенням показника $\gamma = 1.5$ для $k \lesssim 2000$ і $\gamma = 2.7$ для $k \gtrsim 2000$ відповідно. Ми не можемо стверджувати ані того, що отримані значення показників не змінюватимуться при збільшенні досліджуваного корпусу, ані того, що таке збільшення корпусу приведе до кросоверу, як спостерігалось у роботі [12]. Однак спостережена степенева залежність розподілу ступенів вузлів свідчить про те, що мережа української мови є безмасштабною (scale-free).

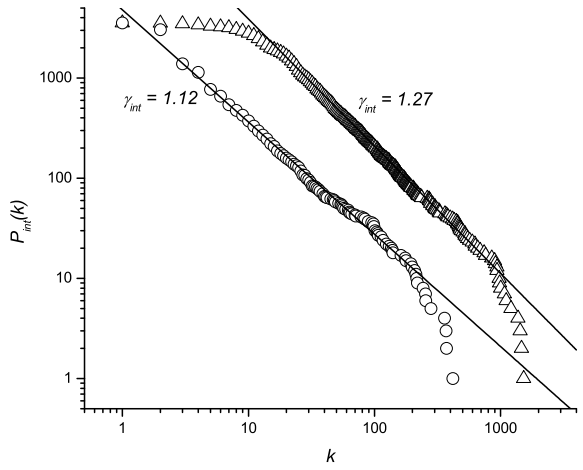


Рис. 6. "Лис Микита": інтеґральний розподіл ступенів вузлів. $R = 1$ (○-○-○), $R = R_{\max}$ (△-△-△). Суцільні прямі показують апроксимації степеневими функціями їх показниками $\gamma = 1.12$ і $\gamma = 1.27$ відповідно.

Як ми зазначили у Вступі, багато реальних мереж поряд із властивістю безмасштабности [11] володіють так званим ефектом тісного світу [10]. Згідно з означенням, якщо середня довжина найкоротшого шляху $\langle l \rangle$ зростає з розміром (кількістю вузлів) мережі $\mathcal{V}$ повільніше за будь-яку степеневу функцію, то мережа є тісним світом [3]. Зауважимо, наприклад, що для реґулярної $d$-вимірної структури $\langle l \rangle \sim \mathcal{V}^{1/d}$. Мережі тісного світу надзвичайно компактні. Незважаючи на велику кількість вузлів, кожну пару з них розділяє лише декілька зв'язків. Для прикладу, в соціології (де і виникло саме поняття тісного світу) відомо, що два випадково обрані члени суспільства перебувають на відстані шести проміжних знайомств $\langle l \rangle \sim 6$ [42]. Із таблиці 2 бачимо, що при $R = 1$ максимальна довжина найкоротшого шляху в усіх трьох мережах (із кількістю вузлів $\mathcal{V} = 2392, 3563, 4823$) становить лише $l_{\max} = 11$, а середня довжина найкоротшого шляху $\langle l \rangle \sim 3$. Таким чином, ці мережі є тісними світами. Для згаданої вище мережі англійської мови $\langle l \rangle = 2.63$ [12]. Оскільки зростання $R$ приводить лише до додавання нових зв'язків і не змінює $\mathcal{V}$, то і $l_{\max}$, і $\langle l \rangle$ зменшуються зі зростанням $R$. Залежність середньої довжини найкоротшого шляху $\langle l \rangle$ показана на рис. 7 як функція ступеня вузла $k$. Спадання $\langle l \rangle$ зі зростанням $k$ відповідає тому, що вузли з вищим ступенем, габи (hubs), розділені меншою відстанню.

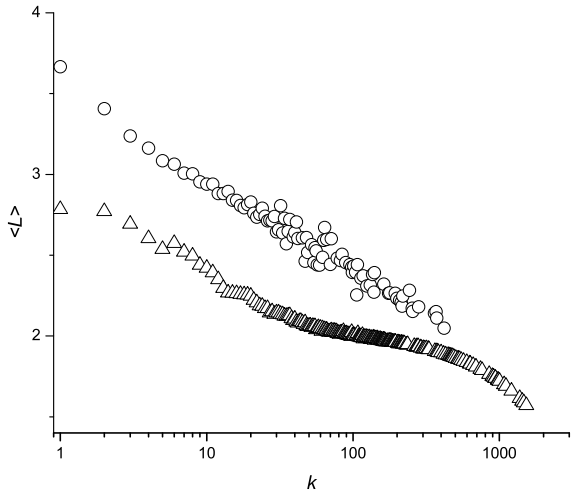


Рис. 7. "Лис Микита": середня довжина найкоротшого шляху як функція ступеня вузла. $R = 1$ (○-○-○), $R = R_{\max}$ (△-△-△).

Специфічною формою кореляції в мережах є утворення кластерів. Означений у (3.21), (3.22) коефіцієнт кластерности $C$ характеризує схильність мережі до утворення з'єднаних трійок вузлів. Так, для повного графа $C = 1$, а для мережі у формі дерева $C = 0$. Коефіцієнт кластерности нескорельованої мережі малий. Прикладом нескорельованої мережі є так званий класичний випадковий граф Ердоша–Рені: $\mathcal{M}$ зв'язків випадково розподілених між парами з $\mathcal{V}$ вузлів. Можна показати, що коефіцієнт кластерности такої мережі [3]

$$C_r = \frac{2\mathcal{M}}{\mathcal{V}^2}. \tag{3.25}$$

У таблиці 2 наведено значення середнього коефіцієнта кластерности $\langle C \rangle$ досліджуваних мереж та його відношення до коефіцієнта кластерности $C_r$ (3.25) класичного випадкового графа з такими ж значеннями $\mathcal{V}$ та $\mathcal{M}$. Значення $\langle C \rangle$ значно перевищує $C_r$, це свідчить про те, що мережі, які ми розглядаємо, є добре скорельованими структурами. Як і слід сподіватися, такі кореляції зростають зі зростанням "радіуса взаємодії" $R$, що відображено на рис. 4. На рис. 8 показано середнє значення коефіцієнта кластерности $\langle C(k) \rangle$ вузлів, що мають однаковий ступінь $k$. Його поведінка характеризується певним плато (при малих значеннях $k$), а далі $\langle C(k) \rangle$ швидко спадає.

Зауважимо, що зростання кількости речень приводить як до зростання словника (кількости різних слів) $\mathcal{V}$, так і до збільшення кількости зв'язків мережі $\mathcal{M}$. Причому значення змінної $\mathcal{V}$ обмежене, і її зростання відбувається повільно. Тому слід сподіватися на зміну $\langle C \rangle$ зі зростанням корпусу досліджуваних творів. Це можна прослідкувати вже на прикладі даних, наведених у таблиці 2. Наприклад, при $R = 1$ $\langle C \rangle$ зростає зі зростанням досліджуваного корпусу ($\langle C \rangle = 0.172, 0.214, 0.243$ для аналізованих тек-

стів). Для Британського національного корпусу [41] на підставі аналізу текстів, що містили $\sim 10^7$ слів, отримано значення $\langle C\rangle = 0.687$ [12].

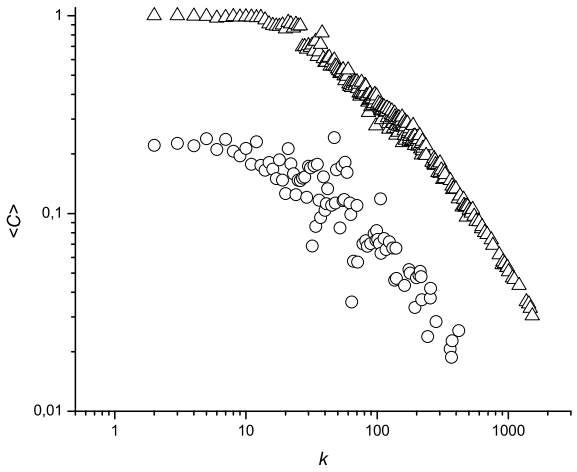


Рис. 8. "Лис Микита": середній коефіцієнт кластерности як функція ступеня вузла. $R = 1$ (○-○-○), $R = R_{\max}$ (△-△-△).

### C. $P$-простір

У $P$-просторі кожне слово-вузол пов'язане з усіма іншими словами, що належать до спільного речення. Таким чином, кожне речення тексту входить у мережу як повний граф — кліка взаємнопов'язаних вузлів. Різні речення-кліки об'єднуються в мережу завдяки спільним словам (див. рис. **3б**). У $L$-просторі слова взаємно пов'язані в межах вікна, розміри якого характеризуються змінною $R$. Коли розмір цього вікна стає рівним розмірові речення, то зображення цього речення в $L$- і в $P$-просторі збігаються. Відповідно, коли розмір вікна стає рівним розмірові найдовшого речення тексту ($R = R_{\max}$), то зображення всього тексту в $L$- і в $P$-просторах збігаються. Таким чином, характеристики досліджуваних текстів в $P$-просторі збігаються з характеристиками цих текстів у $L$-просторі при $R = R_{\max}$, як показано на рисунках 4–8 і наведено в таблиці 2. Відповідні мережі виявляються безмасштабними тісними світами зі степеневим розподілом ступенів вузлів. Їм притаманний високий ступінь кореляції ($\langle C\rangle = 0.841$, 0.821, 0.8183 відповідно для "Абу-Касимових капців", "Лиса Микити" і обох творів разом) і мала середня довжина найкоротшого шляху ($\langle l\rangle = 2.22$, 2.27, 2.25).

Цікаво порівняти результати нашого аналізу з результатами праці [14], у якій теорію складних мереж застосовано для дослідження збірки текстів, написаних португальською та англійською мовами [43] розміром від 169 до 276425 слів. Зображення мережі, яке використовували в роботі [14], за нашою класифікацією відповідає $P$-простору. Однак, на відміну від наших досліджень, у праці [14] розглянуто так звану мережу концепцій (network of concepts): до уваги не взято службових слів, займенників, скорочень. Кількісні характеристики отриманої таким чином мережі становлять: $\gamma = 1.6 \pm 0.2$, $\langle l\rangle = 2.0 \pm 0.1$, $l_{\max} = 4 \pm 1$, $\langle C\rangle = 0.82 \pm 0.03$ і добре узгоджуються з нашими даними (пор. з $\gamma = 2.0$, $\langle l\rangle = 2.25$, $l_{\max} = 5$, $\langle C\rangle = 0.818$, одержаними для мережі двох текстів [1, 36], узятих разом). Такий збіг цікавий не лише тим, що аналізовані мережі відповідають трьом різним мовам (англійська, португальська в [14] і українська в нашому дослідженні), а й тим, що обмеження певною категорією слів (так званими словами-концепціями) не приводить до суттєвих змін у наведених вище характеристиках мережі.

## IV. ВИСНОВКИ

*Тут кінчиться наша казка.*
*Бубликів солодких в'язка*
*Тим, хто слухав, не шумів...*

У цій статті ми розповіли про результати кількісного аналізу розподілу слів у двох текстах — творах Івана Франка "Лис Микита" [1] та "Абу-Касимові капці" [36]. Наші дослідження складались із двох частин: аналіз розподілу частота–ранґ (закон Зіпфа) та аналіз текстів із застосуванням теорії складних мереж. Основною метою аналізу розподілів частота–ранґ було перевірити, чи розмір текстів є достатнім для прояву в них статистичних закономірностей. Показавши, що закон Зіпфа справджується в ділянці змінної ранґ $r = 20 \div 3000$, і отримавши типове значення показника $\alpha \simeq 1$, ми тим самим обґрунтували вибір зазначених текстів для вивчення на їх основі характерних особливостей української мови як складної мережі.

Наскільки нам відомо, наше дослідження є першою спробою застосувати теорію складних мереж для аналізу україномовних текстів. Результати, наведені в розділі III, переконливо свідчать про те, що мережа української мови є сильно скорельованим безмасштабним тісним світом (scale-free small world). Це і не дивно, бо подібні ефекти виявлені раніше для інших мов [12–14]. Властивість тісного світу означає, зокрема, малу середню відстань $\langle l\rangle$ між словами в мережі мови. Незважаючи на величезну кількість слів, які становлять словниковий запас, між будь-якими двома словами в мережі мови в середньому стоять лише два проміжні слова. Високе значення коефіцієнта кластерности свідчить про сильні кореляції. Чисельні значення цих та інших характеристик мережі наведені в таблиці 2. Отримані емпіричні результати можуть бути корисними при теоретичному описі еволюції мови, наприклад, на підставі еволюційної теорії ігор [44].

Є низка праць, у яких зроблено спробу пояснити властивості мереж мови за допомогою сценарію переважного приєднання (preferential attachment [11]), розглядаючи їх як результат процесу росту, коли нові вузли-слова з більшою ймовірністю приєднуються до вузлів-габів, що мають багато зв'язків [12–14]. Цікаво

зазначити, що за таким сценарієм поява синтаксису, як певний фазовий перехід у процесі еволюції мови, є природним наслідком безмасштабної структури мережі мови [17, 18].

Корисним виявляється зображення мережі мови в різних просторах (див. рис. 3) [40] і їх порівняльний аналіз. Зокрема, характеристики мережі мови в $P$-просторі можна отримати як граничний випадок зображення в $L$-просторі. Раніше такі зображення розглядались як незалежні [12, 14]. Щобільше, проведене у $P$-просторі порівняння різних мереж показало, що обмеження певною категорією слів не спричиняє суттєвих змін у характеристиках цих мереж.

На закінчення хочемо подякувати Джефрі Вилзові, Уляні та Тарасові Головачам, Олесі Мриглод, Вікторії Ратушній, Олександрові Сабану, Волтові Стівенсону та Крістіанові фон Ферберу за допомогу при проведенні досліджень, підсумованих у цій роботі.

Подаючи цю статтю у випуск "Журналу фізичних досліджень", присвячений 60-річчю професора Івана Вакарчука, маємо честь і приємність привітати ювіляра та побажати йому щастя, здоров'я і многая літа.

[1] Іван Франко, *Лис Микита*, (Дитвидав, Київ, 1959). Електронна версія є за адресою: http://poetyka.uazone.net/franko/.

[2] G. Parisi, *Complex Systems: a Physicist's Viewpoint* (preprint cond-mat/0205297, 2002); R. N. Mantegna, H. E. Stanley, *An Introduction to Econophysics: Correlations and Complexity in Finance* (Cambridge University Press, Cambridge, 1999); B. K. Chakrabarti, A. Chakraborti, A. Chatterjee, *Econophysics and Sociophysics: Trends and Perspectives* (Wiley-VCH, Berlin, 2006).

[3] S. N. Dorogovtsev, S. N. Mendes, *Evolution of Networks* (Oxford University Press, Oxford, 2003).

[4] Ю. Головач, О. Олємской, К. фон Фербер, О. Мриглод, Т. Головач, І. Олємской, В. Пальчиков, Журн. фіз. досл. **10**, 247 (2006).

[5] H. E. Stanley, *Introduction to Phase Transitions and Critical Phenomena* (Clarendon Press, Oxford, 1971); C. Domb. *The Critical Point* (Taylor & Francis, London Bristol, 1996).

[6] Yu. Holovatch (Ed.), *Order, Disorder and Criticality. Advanced Problems of Phase Transition Theory* (World Scientific, Singapore, 2004); vol.2: World Scientific, Singapore, 2007.

[7] *Graph Theory*, (Springer-Verlag, Heidelberg, Graduate Texts in Mathematics, Volume 173, 2005); S. Bornholdt, H. Schuster (Eds.), *Handbook of Graphs and Networks* (Wiley-VCH, Weinheim, 2003).

[8] R. Albert, A.-L. Barabási, Rev. Mod. Phys. **74**, 47 (2002); S. N. Dorogovtsev, J. F. F. Mendes, Adv. Phys. **51**, 1079 (2002); M. E. J. Newman, SIAM Review **45**, 167 (2003); S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, D.-U. Hwang, Physics Reports **424**, 175 (2006); A. Lesne, Lett. Math. Phys. **78**, 235 (2006).

[9] D. J. Watts, *Small Worlds* (Princeton University Press, Princeton, NJ, 1999); R. Pastor-Satorras, A. Vespignani, *Evolution and Structure of the Internet: A Statistical Physics Approach* (Cambridge University Press, Cambridge, 2004); M. E. J. Newman, A.-L. Barabási, D. J. Watts, *The Structure and Dynamics of Networks* (Princeton University Press, Princeton, 2006).

[10] D. J. Watts, S. H. Strogatz, Nature (London) **393**, 440 (1998).

[11] R. Albert, H. Jeong, A.-L. Barabási, Nature (London) **401**, 130 (1999).

[12] R. Ferrer i Cancho, R. V. Solè, Proc. R. Soc. Lond. B **268**, 2261 (2001).

[13] S. N. Dorogovtsev, J. F. F. Mendes, Proc. R. Soc. Lond. B **268**, 2603 (2001).

[14] S. M. G. Caldeira, T. C. Petit Lobao, R. F. S. Andrade, A. Neme, J. G. V. Miranda, preprint physics/0508066 (2005).

[15] R. Ferrer i Cancho, R. V. Solé, R. Kohler, Phys. Rev. E **69**, 051915 (2004).

[16] R. Ferrer i Cancho, Phys. Rev. E **70**, 056135 (2005).

[17] R. Ferrer i Cancho, O. Riordan, B. Bollobás, Proc. R. Soc. Lond. B **272**, 561 (2005).

[18] R. Solé, Nature **434**, 289 (2005).

[19] A. E. Motter, A. P. S. de Moura, Y.-C. Lai, P. Dasgupta, Phys. Rev. E **65**, 065102(R) (2002).

[20] M. Sigman, G. A. Cecchi, Proc. Natl. Acad. Sci. USA, **99**, 1742 (2002).

[21] A. de Jesus Holanda, I. Torres Pisa, O. Kinouchi, A. Souto Martinez, E. E. Seron Ruiz, Physica A **344**, 530 (2004).

[22] G. F. Zipf, *Human Behaviour and the Principle of least Effort. An Introduction to Human Ecology*, 1st edition (Hafner reprint, New York, 1972) (Addison-Wesley, Cambridge, MA, 1949).

[23] G. F. Zipf, *The Psycho-Biology of Language*, (Houghton-Mifflin, Boston, 1935).

[24] Бібліографію робіт про закон Зіпфа можна знайти на сайті: http://www.nslij-genetics.org/wli/zipf/

[25] R. Ferrer i Cancho, Eur. Phys. J. B **44**, 249 (2005).

[26] M. Mitzenmacher, Internet Mathematics **1**, 226 (2004).

[27] M. V. Simkin, V. P. Roychowdhury, preprint physics/0601192 (2006).

[28] E. U. Condon, Science **67**, 300 (1928). Автором цієї статті про статистику розподілу слів у словнику є Едвард Улер Кондон (1902-1974), автор принципу Франка-Кондона та першої англомовної статті з квантової механіки (разом з Філіпом Морзе у 1929 р.).

[29] A. N. Pavlov, W. Ebeling, L. Molgedey, A. R. Ziganshin, V. S. Anishchenko, Physica A **300**, 310 (2001).

[30] M. A. Montemuro, Physica A **300**, 567 (2001).

[31] W. Dahui, L. Menghui, D. Zengru, Physica A **358**, 545 (2005).

[32] M. A. Montemuro, D. H. Zanette, Advances in Complex Systems **5**, 7 (2002).

[33] P. Kokol, V. Podgorelec, Complexity International **7**, 1 (2000).

[34] I. Kanter, D. A. Kessler, Phys. Rev. Lett. **74**, 4559 (1995).

[35] S. S. Melnyk, O. V. Usatenko, V. A. Yampol'skii,

V. A. Golick, Phys. Rev. E **72**, 026140 (2005).
[36] Іван Франко, *Абу-Касимові капці*, (Зібрання творів в 50 томах, т. 4, с. 295, Наукова думка, Київ, 1976). Електронна версія є за адресою: http://poetyka.uazone.net/franko/.
[37] H. A. Simon, Biometrika **42**, 425 (1955).
[38] М. Абрамовиц, И. Стиган (Ред.), *Справочник по специальным функциям* (Наука, Москва, 1979).
[39] V. Latora, M. Marchiori, Physica A **314**, 109 (2002); P. Sen, S. Dasgupta, A. Chatterjee, P. A. Sreeram, G. Mukherjee, S. S. Manna, Phys. Rev. E **67**, 036106 (2003); K. A. Seaton, L. M. Hackett, Physica A **339**, 635 (2004); J. Sienkiewicz, J. A. Holyst, Phys. Rev. E **72**, 046127 (2005)
[40] C. von Ferber, T. Holovatch, Yu. Holovatch, V. Palchykov, Physica A **380**, 585 (2007), preprint physics/0608125 (2006).
[41] Британський національний корпус — це збірка зразків (тексти, зразки усного мовлення), зібраних з різних джерел для того, щоб репрезентувати сучасну англійську мову. На сьогодні його обсяг становить $10^8$ слів. Див. http://www.natcorp.ox.ac.uk/.
[42] S. Milgram, Psychol. Today **2**, 60 (1967).
[43] Тексти, проаналізовані в роботі [14], в основному взяті з сайту проєкту Ґутенберґ: http://www.gutenberg.org
[44] M. A. Nowak, D. C. Krakauer, Proc. Natl. Acad. Sci. USA, **96**, 8028 (1999).

## FOX MYKYTA AND NETWORKS OF LANGUAGE

Yu. Holovatch[1,2], V. Palchykov[1]
[1] Insitute for Condensed Matter Physics, National Academy of Sciences of Ukraine, 79011 Lviv, Ukraine
[2]Institut für Theoretische Physik, Johannes Kepler Universität Linz, 4040 Linz, Austria

The results of quantitative analysis of word distribution in two fables in Ukrainian by Ivan Franko: "Fox Mykyta" and "Abu-Kasym's Slippers" are reported. Our study consists of two parts: the analysis of frequency-rank distributions and the application of complex networks theory. The analysis of frequency-rank distributions shows that the text sizes are sufficient to observe statistical properties. The power-law character of these distributions (Zipf's law) holds in the region of rank variable $r = 20 \div 3000$ with an exponent $\alpha \simeq 1$. This substantiates the choice of the above texts to analyse typical properties of the language complex network on their basis. Besides, an applicability of the Simon model to describe non-asymptotic properties of word distributions is evaluated.

In describing language as a complex network, usually the words are associated with nodes, whereas one may give different meanings to the network links. This results in different network representations. In the second part of the paper, we give different representations of the language network and perform comparative analysis of their characteristics. Our results demonstrate that the language network of Ukrainian is a strongly correlated scale-free small world. The empirical data obtained may be useful for a theoretical description of language evolution.