ЖУРНАЛ ФІЗИЧНИХ ДОСЛІДЖЕНЬ
т. 17, № 1 (2013) 1001(8 с.)

JOURNAL OF PHYSICAL STUDIES
v. 17, No. 1 (2013) 1001(8 p.)

ПРО "ЕКЗОТИЧНІ" ЗАДАЧІ ФІЗИКИ, ВІННІ-ПУХА ТА ЗАКОН ЗІПФА

О. М. Васильєв[1], О. В. Чалий[1,2], І. В. Васильєва[1]
[1] *Київський національний університет імені Тараса Шевченка, вул. Володимирська 60, Київ 01601, Україна,*
[2] *Національний медичний університет імені О. О. Богомольця, бульв. Т. Шевченка 13, Київ 01601, Україна*
(Отримано 27 листопада 2012 р.; в остаточному вигляді — 31 січня 2013 р.)

Стаття присвячена розв'язанню фізичними методами "нефізичних" задач в галузі квантитативної лінґвістики. Увагу приділено розрахунку статистичних характеристик тексту, виявленню функціональних співвідношень між різними статистичними характеристиками, перевірці законів розподілу. Емпіричні розрахунки виконано на основі україномовного тексту казки про Вінні-Пуха. Отримано статистичну залежність між частотою вживання слова та його ранґом у частотному словнику (*закон Зіпфа*, або *ранґовий розподіл*), залежність кількості слів від частоти їх уживання в тексті (*закон Хайтуна–Тулдави*, або *спектральний розподіл*), розраховано розподіл речень у тексті за довжиною. Проаналізовано моделі, що пояснюють реґресійні співвідношення між кількісними характеристиками тексту.

Ключові слова: квантитативна лінґвістика, закон Зіпфа, закон Хайтуна–Тулдави, скейлінґовий закон, ранґовий розподіл, спектральний розподіл, частотний словник.

PACS number(s): 02.30.Hq, 89.75.Da, 89.90.+n

Краще б він читав "Вінні-Пуха". Втім, психологи з якоїсь медичної установи дійшли висновку, що Вінні-Пух шизофренік, бо у нього нав'язливі ідеї та неадекватна поведінка. А в кого вона тепер адекватна? Читав, що вже сьогодні кожна третя людина у світі має порушення психіки.
Ліна Костенко, "Записки українського самашедшого"

ВСТУП

Характеризуючи ситуацію з "експансією" фізики в інші галузі досліджень, слід визнати, що цей процес об'єктивно виправданий. Ситуація дещо нагадує бурхливий розвиток інституціональної економічної теорії, яка претендує, і не безпідставно, на роль сучасної соціальної філософії [1]. Методологічний апарат інституціональної економіки дає змогу аналізувати на фундаментальному рівні більшість соціальних, політичних та економічних процесів, причому в сукупності. Методи фізики є ще різноманітнішими й універсальнішими, тому в цьому плані можна розраховувати на значні успіхи. Така думка має реальне підтвердження завдяки великій кількості публікацій, у яких розглянуто широке коло проблем у царині динаміки популяцій і соціології [2–11], політики [12–15] та економіки [16–19]. Однак навіть на тлі таких "досить екзотичних" задач застосування фізичних підходів у межах *математичної лінґвістики* [20] може видатися "занадто екзотичним". Однак і тут є серйозні успіхи (див., наприклад, [21, 22] та посилання, що містяться там).

Як правило, до математичної лінґвістики відносять коло задач та методів, які передбачають вивчення кількісних характеристик лексики та структури певної мови (чи навіть групи мов) з широким застосуванням математичного апарату. При цьому саму математичну лінґвістику поділяють на *квантитативну лінґвістику* та *комбінаторну лінґвістику* [23, 24]. Для комбінаторної лінґвістики характерним є застосування математичного апарату теорії множин, математичної логіки та теорії алгоритмів [23]. Квантитативна лінґвістика вивчає лексику методами кількісної математики — теорії ймовірностей, математичної статистики, теорії диференціальних рівнянь [24]. Крім цього, нерідко виділяють як окремий напрямок досліджень статистичне моделювання лінґвістичних систем та процесів [23, 25].

Навіть у межах одного напрямку помітна досить чітка відмінність у характері та типології різних досліджень. Наприклад, існує група "математичних" праць [26–29], у яких увагу приділено встановленню статистико-ймовірнісних характеристик текстів та відновленню математичних схем їх побудови. Ідеологічно близькими є "фізичні" дослідження в галузі квантитативної лінґвістики [30–33]. Особливість останніх — концентрація на універсальних законах та співвідношеннях, які притаманні не тільки лінґвістичним системам. Прикладом є праця [22], у якій дослідження українських текстів виконано в межах проекту вивчення ефектів безмасштабності та тісного світу для систем зі структурою складної мережі. В цьому випадку йдеться про інтерпретацію результатів кількісного аналізу лінґвістичної системи в межах концепції, яка виходить далеко за межі лінґвістики. Цей підхід є продуктивним, перспективним та універсальним. Водночас існують й інші підходи стосовно дослідження лінґвістичних систем фізико-математичними методами.

Традиційно у квантитативній лінґвістиці значну увагу приділяють інтерпретації (у термінах сучасної лінґвістичної парадигми) результатів числових розрахунків [34–36]. Така інтерпретація передбачає, крім

іншого, широке залучення моделей, що описують усю досліджувану систему чи окремі процеси в ній. Моделі створюють на основі певних загальних уявлень про лінґвістичну систему чи процес. Тобто мову загальних понять та уявлень перекладають мовою математики (у цьому разі маються на увазі диференціальні рівняння). Саме такий підхід характерний для синерґетики. Тому не дивно, що відповідний напрямок досліджень у лінґвістиці має усталену назву *лінґвістична синерґетика* [37]. Методи лінґвістичної синерґетики поєднують лінґвістичні концепції та моделювання (на основі нелінійних феноменологічних моделей). У межах таких моделей ми далі будемо пояснювати реґресійні співвідношення.

На теренах сучасної квантитативної лінґвістики вітчизняні дослідники представлені досить потужними науковими школами, які формувалася протягом останніх десятиліть і на сьогодні сконцентровані в декількох наукових центрах. Серед проблем, які висвітлювали в різні часи, можна виділити як загальні питання статистики та структури мови [38–40], так і питання організації різних лінґвістичних підсистем (див., наприклад, праці [41–45] та посилання, що містяться там). Причому характерною рисою найновіших досліджень є збільшення робіт, виконаних у межах фізичних підходів (див., наприклад, праці [46–48]). Така тенденція досить загальна. І хоча її обґрунтування виходить за межі представленої роботи, певними перевагами "фізичного" підходу ми скористаємося (правда, дещо неявно).

Перша частина статті (розділ *Емпіричні результати*) є "емпіричною" і містить результати кількісних розрахунків для різних текстів (текст казки про Вінні-Пуха та пряма мова головного героя). Фактично, там ми застосовуємо добре відому методологію до не досліджуваного раніше текстового матеріалу. Друга частина роботи (розділ *Основні математичні моделі*) присвячена аналізу перевірених та підтверджених емпірично залежностей на основі математичних моделей. На нашу думку, саме тут корисними є фізичні підходи та методи якісного аналізу, оскільки вони дають змогу не тільки отримувати реґресійні співвідношення, але й робити висновки концептуального характеру та проводити аналогії з фізичними явищами і процесами. Це дає нам підстави сподіватися, що наведені в статті міркування будуть цікавими для читачів *Журналу фізичних досліджень*.

ЗАКОН ЗІПФА

Серед об'єктів, цікавих для вивчення в межах квантитативної лінґвістики, важливе місце займають *частотні словники*, для створення яких необхідно мати базовий текст (що може складатися з різних текстових фрагментів). У тексті підраховують кількість різних слів і кількість входжень цих слів у текст. Відтак слова розміщують за порядком спадання частоти їх уживання в тексті. Порядковий номер слова в такій послідовності називається *ранґом* слова. Отже, найбільш вживане в тексті слово має ранґ 1. Ранг 2 має слово, яке є другим за частотою вживання в тексті, і так далі. Залежно від практичних потреб, частоту можна розглядати як абсолютну (кількість входження слова в текст) або відносну (кількість входження слова в текст, поділена на загальну кількість словоживань у тексті). При цьому виникає питання, що розуміти під "словом". Є два методи підрахунку слів:

- якщо всі словоживання в тексті здять до основної форми (для іменників це називний відмінок, а для дієслів — інфінітив) і потім порівнюють на предмет збігу, то в цьому разі під "словами" розуміють *лексеми*;
- якщо слова в тексті відразу порівнюють на предмет збігу (без зведення до основної форми), то в цьому разі під "словами" розуміють *словоформи*.

Ми під "словами" матимемо на увазі саме словоформи, однак відразу повинні зазначити, що ті закономірності, про які йтиметься далі (зокрема, *закон Зіпфа*), будуть наявні незалежно від того, у який спосіб (з двох названих вище) інтерпретовано "слово" [23].

Серед характеристик, які викликають найбільше зацікавлення в дослідників, можна виділити залежність частоти вживання слова в тексті від його ранґу. Відповідну залежність називають *ранґовим розподілом* [23, 24]. *Спектральний розподіл* установлює залежність між кількістю слів з однаковою частотою вживання й цією частотою [23, 24]. Результати багатьох незалежних досліджень дають підстави стверджувати, що ранґовий та спектральний розподіли мають досить універсальний характер і описуються скейлінґовими залежностями (див., наприклад, працю [23]). Це означає, що коли ми маємо побудований для певного тексту частотний словник із N словоформ, то між ранґом слова n ($n = 1, 2, \ldots, N$) та частотою (кількістю появ у тексті) цього слова $F(n)$ існує таке співвідношення:

$$F(n) = \frac{A}{n^{\gamma}}, \tag{1}$$

де A та γ є параметрами розподілу. Співвідношення (1) має назву *закону Зіпфа*, або *першого закону Зіпфа* (див. одні з перших праць у цій царині [49–52], або досить повну інформацію про застосування та модифікації закону Зіпфа в роботі [23]).

Далі, якщо через $m(F)$ позначити кількість слів, що трапляються в тексті з частотою F, то спектральний розподіл визначатиметься співвідношенням:

$$m(F) = \frac{B}{F^{1+\alpha}}, \tag{2}$$

де, як і в попередньому випадку, через B та α позначено параметри розподілу, причому маємо такий зв'язок: $\alpha = 1/\gamma$ [23, 24]. Співвідношення (2) називають *другим законом Зіпфа* [22] або *законом Хайтуна–Тулдави* [24]. Так само, як і закон Зіпфа (1), закон

Хайтуна–Тулдави (2) є емпіричним, оскільки базується передусім на результатах обробки статистичних даних (текстів). Водночас, слід мати на увазі декілька важливих обставин. По-перше, закон Хайтуна–Тулдави є наслідком закону Зіпфа [22, 23], і в цьому сенсі співвідношення (1) та (2) не є незалежними. По-друге, існує усталена думка, що скейлінґове співвідношення (1) справджується не для всього частотного словника: при малих і великих ранґах наявне значне відхилення від скейлінґового розподілу. Ця експериментально доведена обставина спричинила появу низки модифікацій закону Зіпфа.

Окрім суто теоретичного значення, закон Зіпфа широко використовують на практиці. Наприклад, частотні словники й розраховані на їхній основі ранґові розподіли дають змогу:

- виділити “ядро” мови — множину лексем, які забезпечують основний словниковий запас мови (використовують, зокрема, при вивченні іноземних мов);
- класифікувати тексти за авторською належністю (установлювати автора тексту);
- аналізувати психічний стан пацієнтів за аналізом їхньої усної та письмової мови (цими питаннями займається *психолінґвістика*);
- класифікувати мови (на аналітичні та синтетичні);
- класифікувати тексти за рівнем науковості та професійної спеціалізації;
- перевіряти моделі квантитативної лінґвістики.

Щобільше, навіть за значенням скейлінґового індексу γ можна зробити важливий висновок про характер досліджуваного тексту. Так, вважається, що для більшості неспеціалізованих текстів цей показник має бути близьким до одиниці, тобто $\gamma \approx 1$ [23]. Для тексту, який характеризує мову чи письмо людей з психічними відхиленнями, цей показник, як правило, становить величину, меншу за одиницю (див., роботи [34–36]). Тому епіграф до статті — цитата з чудового твору “Записки українського самашедшого” Ліни Костенко — окрім літературної вишуканості, має ще й цілком конкретне “наукове підґрунтя”: вважається, що для психічно хворих людей (і, зокрема, хворих на шизофренію) скейлінґовий індекс γ становить величину близько $\gamma \approx 0.7$. У цій праці ми скористаємося цією цитатою і проаналізуємо український текст казки про Вінні-Пуха та:

- по-перше, емпірично перевіримо справедливість законів Зіпфа та Хайтуна–Тулдави для тексту казки та тексту прямої мови Вінні-Пуха;
- по-друге, отримаємо розподіл речень у текстах за довжиною (вираженою через кількість слів у реченні) — ця задача розв’язується емпірично, а потім ми пропонуємо просту модель, що пояснює відповідну залежність;
- нарешті, і це по-третє, скориставшись розрахованими параметрами розподілів для тексту казки та тексту прямої мови, ми зможемо зробити певні висновки щодо психічного стану Вінні-Пуха.

Що стосується останнього пункту, то маємо зробити одне важливе зауваження. Справа в тому, що ми аналізуємо *переклад*, а не оригінальний текст. Тому аналізуватимемо не англійського Вінні-Пуха, а того, який розмовляє українською мовою (завдяки перекладачеві, звісно). Це означає, що наші висновки стосуватимуться лише україномовного Вінні-Пуха.

ЕМПІРИЧНІ РЕЗУЛЬТАТИ

Ми проаналізували український текст казки про Вінні-Пуха і на його основі виділили підтекст, який складається виключно з прямої мови Вінні-Пуха. Ці два тексти використано для побудови двох окремих частотних словників. Частотні словники дали змогу дослідити ранґовий та спектральний розподіли для кожного з текстів. Розподіл речень за довжиною розраховано безпосередньо на основі базових текстів.

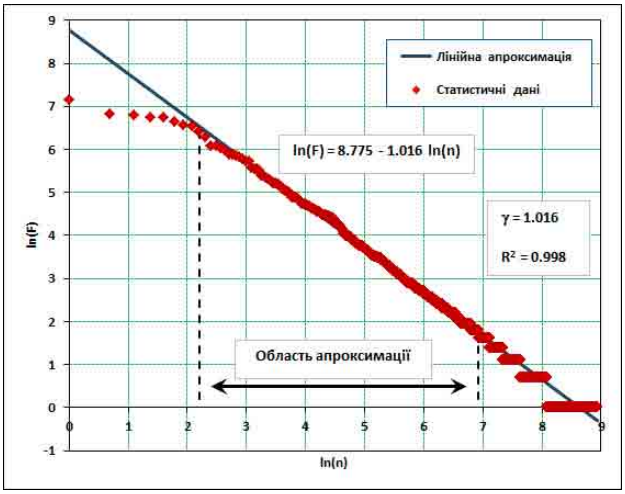

Рис. 1. Залежність $\ln(F)$ від $\ln(n)$ для всього тексту. Лінійна апроксимація виконувалась на масиві з 1000 слів (слова з 9-го ранґу по 1008-ий). Значення скейлінґового індексу дорівнює $\gamma = 1.016 \pm 0.003$, коефіцієнт детермінації $R^2 = 0.998$.

Текст казки про Вінні-Пуха (українська версія) складався з 44 783 слова (слововживання). Текст прямої мови Вінні-Пуха містив 8 160 слововживань. Частотні словники мали обсяг 7 624 словоформи для тексту всієї казки та 2 085 словоформ для тексту прямої мови Вінні-Пуха. Визначаючи параметри розподілу (ранґового та спектрального), ми вибирали центральну частину кожної залежності, відкидаючи декілька початкових точок та нехтуючи статистично малонадійними “хвостами” в розподілах. Усі побудови виконували фактично в подвійному лоґарифмічному масштабі. Точніше, ми розглядали натуральні лоґарифми для частоти $\ln(F)$, ранґу $\ln(n)$ та кількості слів

з однаковою частотою $\ln(m)$. Тому відповідні залежності, що визначають ранґовий та спектральний розподіли, апроксимувалися лінійними функціями. Отримані емпіричні результати разом з апроксимуючими лініями реґресійної залежності наведено на рис. 1 (дані для тексту казки) та рис. 2 (дані для прямої мови Вінні-Пуха).

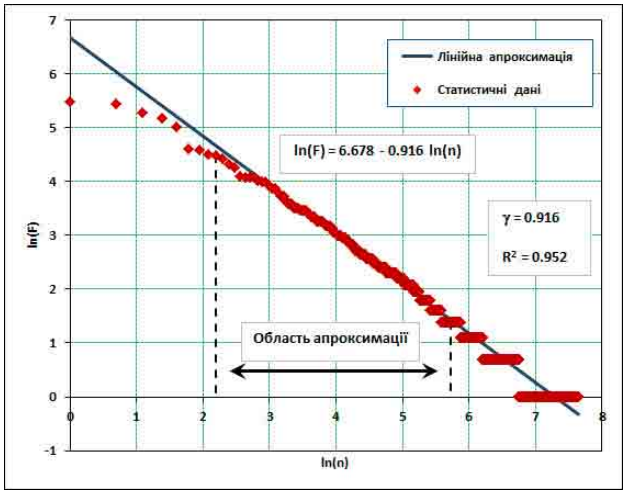

Рис. 2. Залежність $\ln(F)$ від $\ln(n)$ для тексту прямої мови. Лінійна апроксимація виконувалась на масиві з 300 слів (слова з 9-го ранґу по 308-ий). Значення скейлінґового індексу дорівнює $\gamma = 0.916 \pm 0.010$, коефіцієнт детермінації $R^2 = 0.952$.

Ми розрахували такі вирази для ранґового розподілу:

$$\ln(F) = 8.775 - 1.016 \ln(n) \qquad (3)$$

для тексту всієї казки та

$$\ln(F) = 6.678 - 0.916 \ln(n) \qquad (4)$$

для тексту прямої мови Вінні-Пуха. Отже, маємо значення скейлінґового індексу $\gamma \approx 1.016 \pm 0.003$ для тексту казки та $\gamma \approx 0.916 \pm 0.010$ для тексту прямої мови Вінні-Пуха. І хоча в останньому випадку скейлінґовий індекс менший за одиницю, відхилення можна вважати цілком прийнятним, ураховуючи незначний обсяг тексту та емпіричний характер закону Зіпфа. Щобільше, ми провели "перехресну перевірку", розрахувавши параметри спектрального розподілу. Спектральний розподіл, побудований на основі тексту казки, зображено на рис. 3. На рис. 4 показано спектральний розподіл, побудований на основі тексту прямої мови Вінні-Пуха.

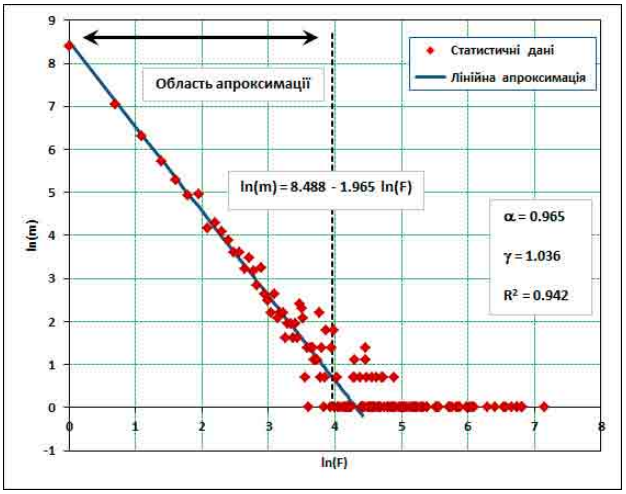

Рис. 3. Залежність $\ln(m)$ від $\ln(F)$ для всього тексту. Лінійна апроксимація виконувалась на перших 50-ти найбільш вживаних словах. Значення скейлінґових індексів $\alpha = 0.965 \pm 0.142$ та $\gamma = 1.036 \pm 0.153$, коефіцієнт детермінації $R^2 = 0.942$.

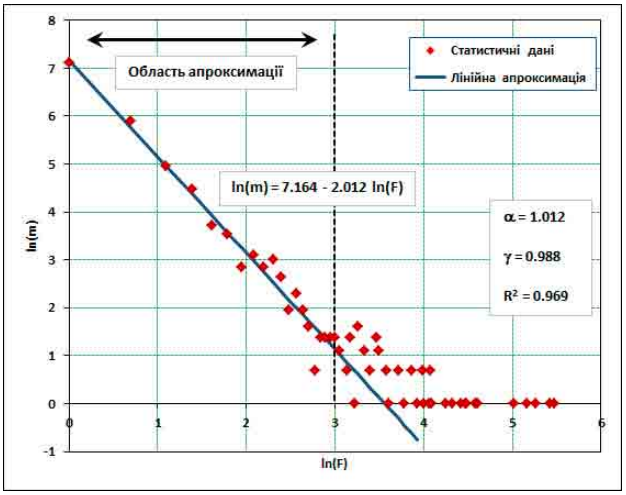

Рис. 4. Залежність $\ln(m)$ від $\ln(F)$ для тексту прямої мови. Лінійна апроксимація виконувалась на перших 20-ти найбільш вживаних словах. Значення скейлінґових індексів $\alpha = 1.012 \pm 0.178$ та $\gamma = 0.988 \pm 0.174$, коефіцієнт детермінації $R^2 = 0.969$.

Використовуючи той самий метод аналізу, як і для первірки закону Зіпфа, розрахували параметри частотного розподілу і, зокрема, значення скейлінґового індексу α. Базуючись на співвідношенні $\gamma = 1/\alpha$, розраховували значення скейлінґового індексу γ в законі Зіпфа. У результаті ми отримали такі спектральні розподіли:

$$\ln(m) = 8.488 - 1.965 \ln(F) \qquad (5)$$

для тексту казки та

$$\ln(m) = 7.167 - 2.012 \ln(F) \qquad (6)$$

для тексту прямої мови Вінні-Пуха. Значення скейлінґових індексів, відповідно, $\alpha \approx 0.965 \pm 0.142$ та $\gamma \approx$

1.036 ± 0.153 (текст казки), а також $\alpha \approx 1.012 \pm 0.178$ та $\gamma \approx 0.988 \pm 0.174$ (текст прямої мови). Так само, як і при перевірці закону Зіпфа, отримуємо значення скейлінґового індексу γ, близьке до одиниці. Отже, особливих підстав вважати мову Вінні-Пуха “неадекватною” немає. Водночас у цьому випадку доцільніше було б зауважити, що навіть на таких відносно невеликих вибірках досить спеціалізованого тексту закон Зіпфа (і похідний від нього закон Хайтуна–Тулдави) має місце — з урахуванням стандартних для практики застосування цього закону обмежень щодо ранґового діапазону, для якого закон застосовується чи перевіряється [23]. Цей висновок не є тривіальним. Адже відомо, що зі зміною обсягу тексту в ранґовій ієрархії відбувається своєрідна ротація слів частотного словника [23]. При цьому кількісні характеристики закону розподілу залишаються відносно сталими. Така властивість текстів великих обсягів добре відома й багаторазово перевірена “експериментально”. Ми отримали ще одне підтвердження — у цьому випадку на прикладі казки про Вінні-Пуха.

Менш дослідженим є питання про залежність кількості речень у тексті від довжини цих речень (*розподіл речень за довжиною*). Тут мається на увазі довжина речення, яка визначається як кількість слів у реченні. Ми провели відповідні розрахунки. А саме, отримали статистичну залежність кількості речень у тексті від кількості слів у реченні. Крім того, ми запропонували реґресійну модель, яка описує закон розподілу речень за довжиною, і розрахували його параметри. На рис. 5 проілюстровано зв’язок логарифма $\ln(S)$ кількості речень певної довжини в тексті казки з кількістю слів W у реченні. На рис. 6 аналогічна залежність показана для тексту прямої мови Вінні-Пуха. Як і раніше, для апроксимації використовувалася лінійна функція. Підґрунтя для застосування саме такої реґресійної моделі наведено в наступному розділі.

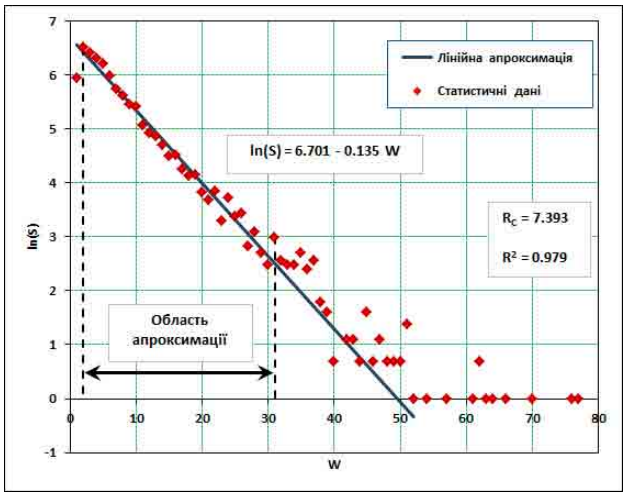

Рис. 5. Залежність $\ln(S)$ від W для всього тексту. Лінійна апроксимація виконувалась по 30-ти позиціях (ранґ із 2-го по 31-ий). Значення “радіуса кореляції” $R_C = 7.393 \pm 0.416$, коефіцієнт детермінації $R^2 = 0.979$.

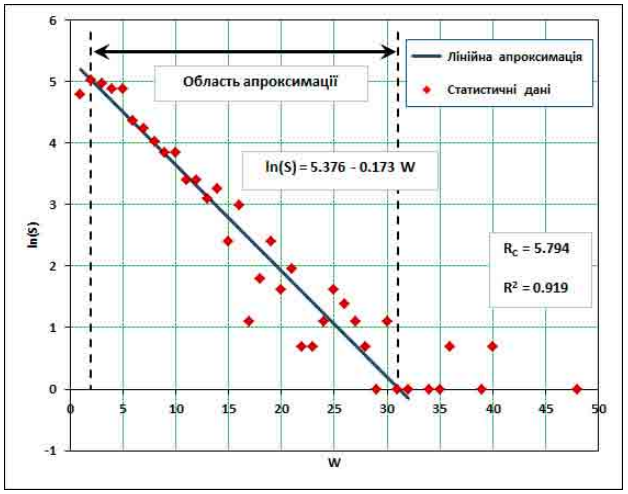

Рис. 6. Залежність $\ln(S)$ від W для всього тексту. Лінійна апроксимація виконувалась по 30-ти позиціях (ранґ з 2-го по 31-ий). Значення “радіуса кореляції” $R_C = 5.794 \pm 0.661$, коефіцієнт детермінації $R^2 = 0.919$.

Для повного тексту реґресійна модель розподілу речень за довжиною має такий вигляд:

$$\ln(S) = 6.701 - 0.135W. \tag{7}$$

Для тексту прямої мови Вінні-Пуха розподіл речень за довжиною визначаємо в межах такої реґресійної моделі:

$$\ln(S) = 5.376 - 0.173W. \tag{8}$$

Бачимо, що розподіл речень за довжиною досить непогано описується такою експоненціальною залежністю:

$$S(W) = C \exp\Big(-\frac{W}{R_C}\Big), \tag{9}$$

де параметр R_C відіграє роль характерної довжини речення (у словах). Якщо розглядати речення як блок тексту, в межах якого слова пов’язані (за смислом та граматично), то “фізичною” мовою речення можна інтерпретувати як “кореляційну область”. Тоді параметр R_C відіграє роль “радіуса кореляції” і визначає характерну відстань (у словах), на якій слова в тексті можна вважати взаємозалежними. Крім того, за умови (9) середня кількість слів у реченні з точністю до невеликої поправки визначається “радіусом кореляції”. Числові розрахунки показують, що для повного тексту маємо значення $R_C \approx 7.393 \pm 0.416$, а для тексту прямої мови $R_C \approx 5.794 \pm 0.661$. Такі результати цілком узгоджуються з даними про середню кількість слів у реченні для великої текстової вибірки.

ОСНОВНІ МАТЕМАТИЧНІ МОДЕЛІ

Існують різні способи пояснення закону Зіпфа. Наприклад, модель Саймона [22] базується на певних уявленнях про те, як формується текст великого обсягу. Модель дає досить непогані результати, а отримані

на основі цієї моделі кількісні характеристики для законів розподілу добре узгоджуються з наявними статистичними даними щодо реальних текстів. Водночас підхід, який базується на алґоритмічному способі формування тексту з повним чи частковим іґноруванням його структури, не може бути вичерпним. Досить популярними, як зазначалося у *Вступі*, є моделі синерґетичного типу, які пояснюють уже відомі функціональні залежності й дають змогу отримувати нові реґресійні співвідношення. Такого типу моделі іноді незамінні для з'ясування питання про зв'язок параметрів розподілу з процесами формування лексичних конструкцій чи просто їх інтерпретації. Наприклад, закон Зіпфа пояснюється на основі диференціального рівняння

$$\frac{dF}{F} = -\gamma\frac{dn}{n}, \tag{10}$$

яке означає, що відносна зміна частоти слова пропорційна до відносної зміни ранґу цього слова (*алометричний закон*, або *закон сталої відносної зміни*), а коефіцієнт пропорційності (зі знаком мінус) є скейлінґовим індексом у законі Зіпфа. Однак така лінійна залежність між логарифмом частоти та логарифмом ранґу слова є не для всього діапазону значень ранґу. Якщо вважати, що коефіцієнт пропорційності у виразі (10) є функцією від ранґу n (чи логарифма ранґу $\ln(n)$), то можемо отримати загальніші співвідношення. Наприклад, у першому наближенні розкладу за n маємо

$$\frac{dF}{F} = -(\gamma + \gamma_1 n)\frac{dn}{n}, \tag{11}$$

що в результаті приводить до такої залежності між частотою слова F та його ранґом n:

$$F(n) = A\frac{\exp(-\gamma_1 n)}{n^\gamma}. \tag{12}$$

Якщо як вихідну змінну розглядати $\ln(n)$, то матимемо

$$\frac{dF}{F} = -(\gamma + \gamma_1 \ln(n))\frac{dn}{n}, \tag{13}$$

і тоді для функції ранґового розподілу отримуємо

$$F(n) = \frac{A}{n^{\gamma+\frac{\gamma_1}{2}\ln(n)}}. \tag{14}$$

Також на практиці нерідко використовують поправку Мандельброта до закону Зіпфа, яка полягає в тому, що у відповідній степеневій залежності виконується "зсув ранґу" на сталу величину [23]:

$$F(n) = \frac{A}{(n+n_0)^\gamma}, \tag{15}$$

причому параметр n_0 може бути як додатним, так і від'ємним. Співвідношення (15) називають *законом Зіпфа–Мандельброта*, а параметр n_0 — *поправкою Мандельброта* [23, 53].

Вважається, що, як і для ранґового розподілу, так для спектрального розподілу має місце алометричний закон: відносна зміна кількості слів у тексті пропорційна до відносної зміни частоти входження слів у текст, тобто

$$\frac{dm}{m} = -(1+\alpha)\frac{dF}{F}. \tag{16}$$

Легко переконатися, що наслідком цього рівняння є співвідношення типу (2). Ми, як і в попередньому випадку, можемо вважати коефіцієнт пропорційності між відносною зміною частоти слів та відносною зміною кількості слів лише нульовим наближенням у розкладі за F чи $\ln(F)$. Якщо скористатися першим наближенням розкладу за F, тобто "відштовхуватися" від рівняння

$$\frac{dm}{m} = -(1+\alpha+\alpha_1 F)\frac{dF}{F}, \tag{17}$$

отримаємо співвідношення

$$m(F) = B\frac{\exp(-\alpha_1 F)}{F^{1+\alpha}}. \tag{18}$$

Якщо ж вихідним розглядати вираз

$$\frac{dm}{m} = -(1+\alpha+\alpha_1 \ln(F))\frac{dF}{F}, \tag{19}$$

у якому розклад виконується по параметру $\ln(F)$, матимемо такий спектральний розподіл:

$$m(F) = \frac{B}{F^{1+\alpha+\alpha_1\ln(F)}}. \tag{20}$$

Також при побудові реґресійних моделей для спектрального розподілу використовуємо поправку типу Мандельброта — тобто "зсув" за арґументом у степеневій залежності:

$$m(F) = \frac{B}{(F+F_0)^{1+\alpha}}, \tag{21}$$

де феноменологічний параметр "зсуву" F_0 визначається на основі статистичних даних.

Усі ці залежності використовують на практиці, і вони "експериментально" підтверджуються. Водночас, жодна з них не є універсальною і має певні межі застосовності. Узагалі, ту чи іншу модель використовують залежно від особливостей конкретної досліджуваної ситуації, хоча й тут можна виділити певні універсальні концепти (див., наприклад, роботу [54]). Перевага підходу, який базується на виведенні реґресійних співвідношень на основі диференціальних рівнянь, полягає в тому, що ми можемо проводити якісний аналіз лінґвістичної підсистеми й основних етапів її еволюції. Як ілюстрацію розгляньмо модель, що описує розподіл кількості речень за довжиною. Корисними будуть загальні міркування, які мають стосунок до лінґвістичних систем. А саме, досить універсальною є гіпотеза про наявність певного внутрішнього синерґетичного механізму, який впорядковує лінґвістичну систему [37]. Причому основна увага акцентується на

“динаміці процесу”. Мова диференціальних рівнянь у цьому плані є зручною, оскільки дає змогу пов’язати реакцію системи з основними характеристиками, що описують стан системи. Зокрема, для визначення залежності між довжиною речення та кількістю таких речень у тексті розгляньмо співвідношення між відносними змінами в кількості речень dS/S і кількості слів у реченні dW/W. Цілком очевидно, що за збільшенням кількості слів кількість речень відповідної довжини має зменшуватися (принаймні, починаючи з якогось значення). У лінійному наближенні можемо записати таке:

$$\frac{dS/S}{dW/W} = \beta_0 - \beta_1 W, \tag{22}$$

де β_0 та β_1 є феноменологічними параметрами моделі, причому параметр β_1 розглядається як невід’ємний. Після нескладних перетворень отримуємо

$$dS/S = \beta_0 \frac{dW}{W} - \beta_1 dW, \tag{23}$$

і, як наслідок, маємо формулу для залежності кількості речень у тексті від кількості слів у реченні:

$$S(W) = CW^{\beta_0} \exp(-\beta_1 W). \tag{24}$$

Якщо $\beta_0 \approx 0$ (а статистичні дані дають підстави так вважати), то формула (24), з точністю до позначення $\beta_1 = 1/R_C$, збігається з формулою (9). Останню ми використали для розрахунку параметрів закону розподілу речень за довжиною.

ВИСНОВКИ

Вище ми навели деякі міркування, що стосуються аналізу лінґвістичних систем. Зрозуміло, що запропонована методологія не є єдино можливою. Водночас, використання феноменологічних моделей на основі диференціальних рівнянь видається перспективним з декількох причин:

- це відносно простий, з методологічного погляду, підхід, який можуть адаптувати (й адаптовують) лінґвісти [23];
- у межах підходу нескладно реалізувати міркування якісного характеру стосовно особливостей лінґвістичних систем та процесів [37];
- підхід дає змогу отримувати реґресійні співвідношення, що пов’язують параметри та характеристики лінґвістичних систем (і які можна перевірити статистичними даними).

Ну, а що стосується Вінні-Пуха, то є всі підстави стверджувати, що Вінні-Пух саме такий, як треба. Принаймні той, який “розмовляє” українською.

Подяки

Автори висловлюють щиру подяку проф. Головачеві Ю. В. за корисні консультації, які, фактично, і спонукали до написання цієї статті.

[1] А. Аузан и др. *Институциональная экономика: Новая институциональная экономическая теория* (ИНФРА-М, Москва, 2011)
[2] D. Abrams, H. Yaple, Phys. Rev. Lett. **107**, 088701 (2011).
[3] Z. Zhao, J. Bohorquez, A. Dixon, N. Johnson, Phys. Rev. Lett. **103**, 148701 (2009).
[4] C. Bordogna, E. Albano, Physica A **329**, 281 (2003).
[5] F. Liu, X. Shan, Y. Ren, J. Zhan, Physica A **328**, 341 (2003).
[6] L. Sander, C. Warren, I. Sokolov, Physica A **325**, 1 (2003).
[7] P. Holme, M. Newman, Phys. Rev. E **74**, 056108 (2006).
[8] I. Benczik, S. Benczik, B. Schmittmann, R. Zia, Europhys. Lett. **82**, 48006 (2008).
[9] W. Weidlich, *Sociodynamics: A Systematic Approach to Mathematical Modelling in the Social Sciences* (Dover, London, 2006).
[10] S. Galam, Physica A **333**, 453 (2004).
[11] С. Капица, С. Курдюмов, Г. Малинецкий, *Синергетика и прогнозы будущего* (Едиториал УРСС, Москва, 2003).
[12] E. Ben-Naim, Europhys. Lett. **69** , 671 (2005).
[13] D. Stauffer, H. Meyer-Ortmanns, Int. J. Mod. Phys. B **15**, 241 (2004).
[14] E. Ben-Naim, P. Krapivsky, S. Redner, Physica D **183** 190 (2003).
[15] E. Ben-Naim, P. Krapivsky, R. Vazquez, S. Redner, Physica A **330** 99 (2003).
[16] T. Kaizoji, Physica A **326**, 256 (2003).
[17] Y. Wang, N. Ding, L. Zhang, Physica A, **324**, 665 (2003).
[18] B. Chakrabarti, A. Chakraborti, A. Chatterjee, *Econophysics and Sociophysics: Trends and Perspectives* (Wiley-VCH, Berlin, 2006).
[19] В. Владимиров, Ю. Воробьев и др., *Управление риском. Риск, устойчивое развитие, синерґетика* (Наука, Москва, 2000).
[20] D. Abrams, S. Strogatz, Nature **424**, 900 (2003).
[21] Ю. Головач *та ін.*, Журн. фіз. досл. **10**, 247 (2006).
[22] Ю. Головач, В. Пальчиков, Журн. фіз. досл. **11**, 22 (2007).
[23] Ю. Тулдава, *Проблемы и методы квантитативно-системного исследования лексики* (Валгус, Таллин, 1987).
[24] Р. Пиотровский, К. Бектаев, А. Пиотровская, *Математическая лингвистика* (Высшая школа, Москва, 1977).
[25] Н. Андреев, *Статистико-комбинаторные методы в*

теоретическом и прикладном языковедении (Наука, Ленинград, 1967).
[26] M. Montemurro, D. Zanette, Glottometrics **4**, 87 (2002).
[27] G. Altmann, R. Hammerl, *Diskrete Wahrscheinlichkeitsverteilungen I.* (Brockmeyer, Bochum, 1989).
[28] Ju. Orlov, *Ein Modell der Häufigkeitsstruktur des Vokabulars* (Brockmeyer, Bochum, 1982).
[29] Ju. Krylov, *Eine Untersuchung statistischer Gezetsmässigkeiten auf der paradigmatischen Ebene der Lexik natürlischer Sprachen* (Brockmeyer, Bochum, 1982).
[30] R. Ferrer i Cancho, Phys. Rev. E **70**, 056135 (2005).
[31] R. Ferrer i Cancho, R. V. Solé, R. Köhler, Phys. Rev. E **69**, 051915 (2004).
[32] A. Holanda, I. Pisa, O. Kinouchi, A. Martinez, E. Ruiz, Physica A **344**, 530 (2004).
[33] A. Motter, A. Moura, Y. Lai, P. Dasgupta, Phys. Rev. E **65**, 065102R (2002).
[34] П. Алексеев *Методика квантитативной типологии текста* (ЛГПИ им. А.И.Герцена, Ленинград, 1983).
[35] Ю. Гурко, Р. Пиотровский, Д. Спивак, в it Актуальные проблемы компьютерной лингвистики. Сб. научных статей (МГЛУ, Минск, 2005), с. 47.
[36] W. Piotrowska, in *Quantitative Linguistics. An International Handbook*, edited by R. Köhler, G. Altmann, R. Piotrowski (De Gruyter, Berlin, 2005), p. 988.
[37] Р. Пиотровский, *Лингвистическая синергетика: исходные положения, первые результаты, перспективы* (Санкт-Петербургский гос. ун-т, Санкт-Петербург, 2006).
[38] В. С. Перебийніс, М. П. Муравицька, Н. П. Дарчук, *Частотні словники та їх використання* (Наукова думка, Київ, 1985).
[39] В. В. Левицкий, *Квантитативные методы в лингвистике* (Рута, Черновцы, 2005).
[40] В. А. Широков, *Інформаційна теорія лексикографічних систем* (Довіра, Київ, 1998).
[41] С. Бук, Лінгвістика **19**, 169 (2010).
[42] С. Бук, Вісн. Львів. ун-ту. Сер. філолог. **55**, 230 (2011).
[43] S. N. Buk, A. A. Rovenchak, J. Quantit. Linguist. **11**, 161 (2004).
[44] S. Buk, J. Mačutek, A. Rovenchak, Glottometrics **16**, 63 (2008).
[45] J. Mačutek, A. Rovenchak, *Studies in Quantitative Linguistics 11: Issues in Qantitative Linguistics, Vol. 2*, edited by E. Kelih, V. Levickij, Yu. Matskulyak (RAM-Verlag, Lüdenscheid, 2011), p. 136.
[46] A. Rovenchak, S. Buk, J. Phys. Stud. **15**, 1005 (2011).
[47] A. Rovenchak, Glottometrics **21**, 65 (2011).
[48] A. Rovenchak, S. Buk, Physica A **390**, 1326 (2011).
[49] G. Zipf, *Human Behavior and the Principle of Least Effort* (Addison-Wesley, Cambrige, 1949).
[50] G. Zipf, *The Psycho-Biology of Language* (Addison-Wesley, Cambrige, 1935).
[51] E. Condon, Science, **67**, 300 (1928).
[52] I. Kanter, D. Kessler, Phys. Rev. Lett. **74**, 4559 (1995).
[53] B. Mandelbrot, Word. **10**, 1, (1954).
[54] G. Wimmer, G. Altmann, in *Quantitative Linguistics. An International Handbook*, edited by R. Köhler, G. Altmann, R. Piotrowski (De Gruyter, Berlin, 2005), p. 791.

ABOUT "EXOTIC" PROBLEMS OF PHYSICS, WINNIE THE POOH AND ZIPF'S LAW

A. N. Vasilev[1], A. V. Chalyi[1,2], I. V. Vasileva[1]
[1] *Taras Shevchenko National University of Kyiv, Faculty of Physics, 60, Volodymyrska St., Kyiv, UA–01601, Ukraine*
[2] *O. O. Bogomolets National Medical University, 13, Shevchenko Blvd, Kyiv, UA–01601, Ukraine*
e-mail: vasilev@univ.kiev.ua

The paper is devoted to the solution by physical methods of "nonphysical" problems in the area of quantitative linguistics. Attention is paid to the calculation of statistical properties of texts. We identify functional relations between different statistical characteristics and verify the laws of distribution. To do these empirical calculations we use Ukrainian language text of the tale about Winnie the Pooh. We receive a statistical dependence of the word frequency in text on its rank in frequency dictionary (Zipf's law, or rank distribution), and find a dependence of word number on their frequency (Haitun–Tuldava's law, or spectral distribution). Also we calculate the distribution of sentences by length in the text. Lastly, we analyze models which explain regression relations between quantitative characteristics of the text.