О.М. ВАСИЛЬЄВ, І.В. ВАСИЛЬЄВА
Київський національний університет імені Тараса Шевченка
(Вул. Володимирська, 60, Київ 01601; e-mail: vasilev@univ.kiev.ua)

УДК 53.01+81'32

ФІЗИКА ЗА МЕЖАМИ ФІЗИКИ: ФІЗИЧНІ ПІДХОДИ В КВАНТИТАТИВНІЙ ЛІНГВІСТИЦІ

В статті розглядається проблема використання фізичних методів для розв'язання задач нефізичного характеру. Зокрема, аналізуються перспективи застосування фізичних підходів в кількісній (квантитативній) лінгвістиці. Різниця між фізичними та нефізичними способами моделювання ілюструється на прикладі уже існуючих "класичних" моделей. Також пропонуються математичні моделі, котрі дозволяють встановлювати рангово-частотну залежність для слів у частотному словнику та залежність розміру словника від об'єму тексту. Показано, що підходи і принципи, котрі є характерними для фізики, можуть бути з успіхом задіяні при створенні математичних моделей у лінгвістиці.

Ключові слова: фізична теорія, модель, економфізика, соціофізика, квантитативна лінгвістика.

Тобі одна знайома путь,
А я — стою на роздоріжжі...

ГЕТЕ, "Фауст"

1. Вступ

Спектр задач, які розв'язуються фізиками, постійно розширюється, а фізичні методи все частіше залучаються для вирішення проблем, які не мають прямого стосунку до фізики. Мова йде не про поодинокі випадки, а про системний підхід, в рамках якого економічні, соціальні, політичні, лінгвістичні (а також деякі інші) задачі втілюються у вигляді моделей, подібних до тих, що широко застосовуються у фізиці. На сьогодні цілком звичними і прийнятними для фізичної спільноти стали такі напрямки досліджень, як *економфізика* [1–8] та *соціофізика* [8–13]. Більше того, відповідні дослідження отримують визнання серед економістів, соціологів, політологів – тобто фахівців нефізичного профілю. Дана обставина не є тривіальною, оскільки методологія досліджень, яка притаманна, скажімо, для економфізики, кардинально відрізняється від методик та моделей, які є звичними для професійних економістів. Разом з тим, ще зарано говорити про цілковите визнання методів фізики в нефізичних наукових царинах. Наведені вище як приклад дослідження залишаються сферою діяльності, в якій застосовують сили в основному фізики, і "споживачами" результатів також є фізики. Така ситуація видається не зовсім правильною. Особливо з урахуванням того, що мова йде про підходи, котрі, як мінімум, нічим не поступаються способу моделювання, традиційному для соціальних та гуманітарних наук. Як підтвердження можна навести низку робіт, котрі стосуються моделювання та дослідження складних систем [14–18]. Їх специфіка у тому, що іноді взагалі досить важко визначити, до якої галузі знань слід віднести відповідну систему. Але незмінно від цього застосування фізичних підходів дає чудові результати.

Одним з напрямків досліджень, окрім економфізики та соціофізики, в якому з успіхом можуть застосовуватись фізичні підходи та моделі, є *кількісна* (або *квантитативна*) *лінгвістика* [19–24]. Незважаючи на те, що на сьогодні тут існує значний доробок, задача пошуку ефективних шляхів для створення нових моделей залишається актуальною. В даній роботі аналізуються переваги фізичних методів моделювання та окреслюються напрямки для застосування цих методів у квантитативній лінгвістиці.

2. Фізичний спосіб моделювання

Отже, якою може бути мотивація для застосування фізичних підходів при розв'язанні нефізичних задач, і, зокрема, задач квантитативної лінгвістики? Аби відповісти на це запитання, слід врахува-

ти декілька важливих обставин. Так, існує велика кількість цікавих задач, які вимагають чи допускають застосування математичного апарату. Це не є дивиною, і відповідний математичний напрямок у лінгвістиці має довгу і продуктивну історію. Відповідно, раніше уже запропоновано (і постійно з'являються нові) математичні моделі, котрі з успіхом реалізуються у квантитативній лінгвістиці для опису самих різноманітних систем та процесів. Разом з тим, має принципове значення не сам факт наявності тої чи іншої моделі, а ще й спосіб, в який вона створювалась. І тут ми стикаємось з тим, що можна було би назвати особливістю фізичного способу моделювання.

Як правило моделі, що створюються для опису фізичних явищ чи процесів, ґрунтуються на певній теорії. Тобто спочатку в той чи інший спосіб формулюються якісь правила чи закони взаємодії елементів, котрі входять в досліджувану систему, і вже після цього модель отримується як наслідок цих законів. І навіть якщо історично послідовність дій інша (тобто спочатку на основі емпіричних даних створюється модель чи будується регресійна залежність), то згодом все ж з'являється теорія, яка пояснює відповідні математичні співвідношення і стає підґрунтям для їх отримання. Як приклад можна навести *третій закон Кеплера* або *закон Стефана–Больцмана*, котрі спочатку були встановлені експериментально, і лише потім для них знайшли теоретичне пояснення. На відміну від такого фізичного підходу, в процесі математичного моделювання під час вирішення лінгвістичних (та й не тільки) задач відповідна математична модель просто постулюється або конструюється, виходячи із питання зручності та загального вигляду наявної "експериментальної" залежності. Такого типу моделі далеко не завжди є ефективними та повноцінними в плані опису системи чи процесу. Чому так? Є два важливих пункти, які слід виділити. В першу чергу, відсутність теорії в основі моделі не дозволяє робити висновки щодо суті механізмів, які зумовлюють отриману залежність. Фактично, модель у такому випадку є описовою, що значно зменшує її цінність. Відразу постає питання щодо області застосовності моделі, а це, у свою чергу, може поставити під сумнів надійність результатів, отриманих на основі моделі. По-друге, слід врахувати специфіку верифікації моделей квантитативної лінгвістики на основі фактичних даних. Справа в тому, що зазвичай результати безпосередніх "вимірювань", перед використанням їх для моделювання, групуються та обробляються [21–24]. За відсутності базової теорії буває важко (а іноді й неможливо) визначити, наскільки спосіб групування даних впливає на характер кінцевої математичної залежності. Іншими словами, модель може виявитися "неуніверсальною" настільки, що зміна способу представлення даних буде на якісному рівні впливати на характер відповідної функціональної залежності. Це є серйозною проблемою, і шлях до її розв'язання проходить через використання розумних, універсальних принципів, які визначають спосіб створення математичних моделей. Це як раз ті підходи, які були розвинені фізиками і використовуються ними для успішного моделювання систем різної природи, у тому числі і лінгвістичних (див., наприклад, [25–30] та посилання, що містяться там).

3. Моделі квантитативної лінгвістики

Перед тим, як безпосередньо перейти до аналізу способів реалізації фізичних підходів для розв'язання лінгвістичних задач, розглянемо деякі "класичні" моделі, які уже існують і використовуються у квантитативній лінгвістиці. Історично одним з перших у лінгвістиці з'явився *закон Зіпфа* [31–34], котрий пов'язує частоту появи слова f у тексті з рангом цього слова n. Мова йде про те, що у певному тексті великого об'єму визначається кількість різних слів і для кожного такого слова розраховується кількість його входжень у текст (традиційно називається *частотою* появи слова у тексті). Слова упорядковуються за спаданням частоти появи у тексті. *Рангом* називається порядковий номер слова у такій послідовності (тобто слово з рангом 1 зустрічається в тексті найчастіше). Зокрема, відповідно до закону Зіпфа ця залежність має бути степеневою:

$$f(n) = \frac{A}{n^{\alpha}}. \tag{1}$$

Тут через A позначено неуніверсальну константу, а через α позначено показник степеня для степеневого розподілу. Саме його розрахунок зазвичай є головною метою дослідження, оскільки існують дані, що для багатьох мов і різних не спеціалізованих текстів значення цього показника близьке

до одиниці [21, 35]. Причини і наслідки відхилення значення показника α від 1 є предметом низки окремих досліджень. Співвідношення (1) є "емпіричним" (воно встановлене шляхом обробки великої кількості лінгвістичних даних) і виконується лише для певного діапазону значень рангу слова.

Очевидно, що в силу означення величин f та n має виконуватись співвідношення

$$V = \sum_{n=1}^{N} f(n), \tag{2}$$

де через N позначено кількість *різних* слів (лексем) у тексті, а V позначає об'єм тексту (кількість слововживань – тобто загальна кількість слів у тексті).

Також відразу слід відзначити, що якщо співвідношення (1) є справедливим, то це означає наявність лінійного зв'язку між параметрами $\ln(f)$ та $\ln(n)$:

$$\ln(f) = B - \alpha \ln(n), \tag{3}$$

де $B = \ln(A)$. Як ілюстрацію, на рис. 1 наведено рангово-частотну залежність для тексту "Записки українського самашедшого" авторства *Ліни Костенко* [36].

Загальний об'єм тексту становить 85227 слів. Кількість лексем (*різних* слів) дорівнює 14796. На основі даних, представлених на інформаційному ресурсі *www.mova.info*, вдалося розрахувати значення $\alpha \approx 0{,}948$ та $B \approx 8{,}796$ для параметрів закону (3), на основі якого виконувалась апроксимація. Коефіцієнт детермінації $R^2 \approx 0{,}964$.

В даному випадку апроксимація виконувалась для всього діапазону значень рангів слів. Якщо ж обмежитись лише тим интервалом, де залежність є суттєво лінійною (зокрема, для $4 \le \ln(n) \le 7$), то отримаємо значення $\alpha \approx 0{,}995$ при значенні коефіцієнта детермінації $R^2 \approx 0{,}997$.

Нескладно помітити, що зі зростанням рангу зростає кількість слів, котрі мають однакову частоту. Якщо через $m(f)$ позначити кількість слів, які зустрічаються з частотою f, то "спектральним" аналогом рангового розподілу є такий закон [21]:

$$m(f) = \frac{M}{f^{1+\gamma}}, \tag{4}$$

де γ та M є параметрами розподілу. Зокрема, для згаданого вище тексту маємо значення $\gamma \approx 0{,}825$ та $M \approx 9{,}015$, коефіцієнт детермінації $R^2 \approx 0{,}889$. Відповідна залежність $\ln(m)$ від $\ln(f)$ наведена на рис. 2.

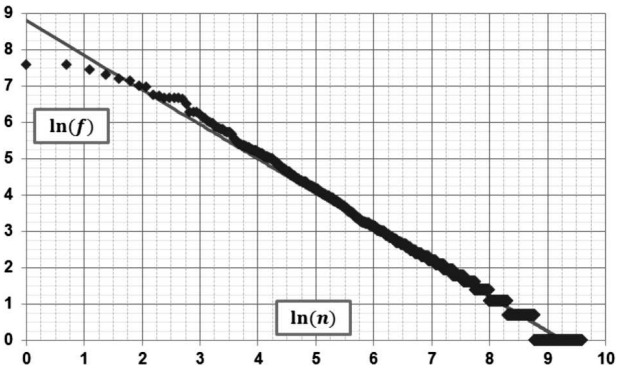

Рис. 1. Рангово-частотна залежність для тексту "Записки українського самашедшого". Маркерами позначені "експериментальні" значення. Суцільна пряма відповідає апроксимації на основі закону Зіпфа. Показник степеневого розподілу $\alpha \approx 0{,}948$

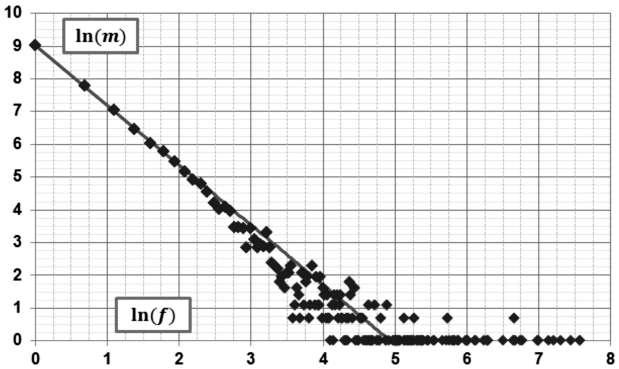

Рис. 2. "Спектральний" розподіл слів для тексту "Записки українського самашедшого". Маркерами позначені "експериментальні" значення. Суцільна пряма відповідає апроксимації на основі закону (4). Показник степеневого розподілу $\gamma \approx 0{,}825$

Зазначимо, що співвідношення (1) та (4) іноді називають відповідно першим та другим законами Зіпфа, і між цими співвідношеннями існує нетривіальний зв'язок (див., наприклад, [37]).

Ще один приклад залежності, яка досить часто використовується на практиці – це залежність розміру словника N (кількість *різних* слів у тексті) від об'єму тексту V (кількість усіх слів у тексті). Ця залежність є нелінійною при великих V і на загал універсальної формули в даному випадку немає. Існують різні підходи, в яких вигляд апроксимаційної функції вибирається апріорно [30, 38–

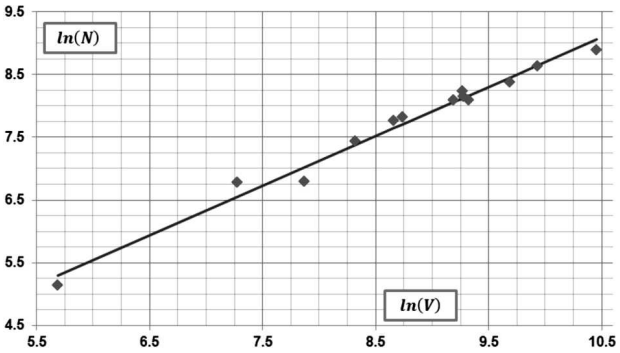

Рис. 3. Залежність між розміром словника та об'ємом тексту на основі творів Тараса Прохаська. Маркери відповідають фактичним значенням, а суцільна крива є апроксимацією на основі залежності (5). Значення параметра $\beta \approx 0{,}788$

49]. Виходячи із загальних міркувань, можна спробувати описати відповідну функціональну залежність степеневим законом вигляду

$$N(V) = kV^{\beta}, \tag{5}$$

або те саме для логарифмів:

$$\ln(N) = \beta \ln(V) + K, \tag{6}$$

де β, k та $K = \ln(k)$ є параметрами моделі. Як приклад застосування цієї залежності для опису реальних даних наведено рис. 3. Там подано залежність між логарифмом розміру словника $\ln(N)$ та логарифмом об'єму тексту $\ln(V)$ для низки творів *Тараса Прохаська* (дані отримано на основі інформаційного ресурсу *www.mova.info*). Розрахунки дають значення для параметрів розподілу $\beta \approx 0{,}788$ та $K \approx 0{,}816$, коефіцієнт детермінації $R^2 \approx 0{,}982$.

Усі наведені співвідношення і результати обробки "емпіричних" даних висвітлюють дві проблеми: одна технічного, а інша методологічного характеру. Перша пов'язана з тим, що кожне із наведених співвідношень (1), (4) чи (5) описує дані тільки в певному діапазоні. Наприклад, добре відомо, що закон Зіпфа (1) не застосовний для розподілу слів із малим та великим рангом. Аналогічно, є певні проблеми із застосуванням "спектрального" закону (4) для розподілу високочастотних слів [21]. Для залежності (5) також існують обмеження у застосуванні, оскільки, наприклад, має виконуватись очевидне співвідношення $N = 1$ при $V = 1$. Все це означає, що відповідні функціональні залежності (1), (4) чи (5) є наближеними і в принципі їх слід уточнювати. Але як? І тут ми стикаємося з методологічною проблемою. На загал, в такому випадку вибирають більш складний вираз для апроксимаційної залежності і параметри цієї залежності визначають на основі "емпіричних" даних. Однак на сьогодні не розроблені чіткі критерії того, як саме слід вибирати апроксимаційну залежність. І саме тут корисними можуть виявитися методи, які використовують фізики.

4. Реалізація фізичних підходів у лінгвістиці

Ідея надзвичайно проста і ґрунтується вона на тому, що апроксимаційна залежність може бути отримана як розв'язок диференціального рівняння. Рівняння, у свою чергу, записується відштовхуючись від загальних уявлень про характер процесів, котрі "відповідають" за наявність досліджуваної закономірності. Наприклад, залежність (1) для закону Зіпфа може бути отримана як розв'язок такого диференціального рівняння першого порядку:

$$\frac{df}{f} = -\alpha \frac{dn}{n}. \tag{7}$$

Відповідно до цього рівняння відносна зміна частоти появи слова пропорційна до відносної зміни рангу слова. Нескладно здогадатися, що аналогічне рівняння може бути вихідним для отримання законів (4) та (5) з поправкою на позначення, які використовуються у відповідному співвідношенні. Фактично це означає, що аналогічно до того, як відносна зміна частоти появи слова пропорційна до відносної зміни рангу слова, відносна зміна кількості слів, котрі зустрічаються з певною частотою, пропорційна до відносної зміни частоти, а відносна зміна кількості різних слів у тексті пропорційна до відносної зміни об'єму тексту. Ці закони можна узагальнити і зробити дещо універсальнішими. Як вихідне використаємо припущення, що зміна двох параметрів (на кшталт кількості різних слів у тексті і об'єму тексту) є такою, що шляхом нелінійного перетворення кожного з параметрів окремо можна добитися того, аби зміна одного параметра була пропорційною до зміни іншого параметра. Якщо згадані параметри позначити через x та y, то дане твердження може бути

реалізоване таким диференціальним рівнянням:

$$\phi(y)\,dy = \psi(x)\,dx, \tag{8}$$

де $\phi(y)$ та $\psi(x)$ є деякими функціями. Ці функції апріорі невідомі. Ми можемо їх оцінити шляхом розкладу в ряд Тейлора, причому по від'ємних показниках степеня аргументу (щоб у першому наближенні отримати уже відомі закони). Зокрема, якщо використати лінійне (по оберненому аргументу) наближення для функції $\phi(y)$ та кубічне наближення для функції $\psi(x)$ і врахувати довільність у виборі одного з коефіцієнтів, отримаємо таке:

$$\frac{dy}{y} = \left(\frac{a}{x} + \frac{b}{x^2} + \frac{2c}{x^3}\right) dx, \tag{9}$$

і параметри a, b, c визначаються в результаті апроксимації "емпіричних" даних. Розв'язком рівняння є вираз

$$y(x) = y_0 x^a \exp\left(-\frac{b}{x} - \frac{c}{x^2}\right). \tag{10}$$

який є основою для побудови апроксимаційної залежності (з чотирма параметрами a, b, c та y_0). Якщо увести позначення $z = \ln(y)$ та $t = \ln(x)$, то в нових змінних матимемо таку апроксимаційну залежність:

$$z(t) = at - b\exp(-t) - c\exp(-2t) + d, \tag{11}$$

де $d = \ln(y_0)$. Таким чином, у логарифмічних змінних ми отримали залежність, яку можна розглядати як таку, що містить експоненційні поправки до лінійного закону. Причому оскільки залежності, для яких виконується апроксимація, досить монотонні, а параметрів для виконання апроксимації декілька, то цілком можливо для розрахунку цих параметрів не просто використовувати метод найменших квадратів (чи інший критерій), а й накласти деякі додаткові обмеження, що є важливим при розв'язанні лінгвістичних задач. Як приклад на рис. 4 наведено результати апроксимації рангового розподілу слів у тексті *Ліни Костенко*, однак тепер уже на основі виразу вигляду (11). В даному випадку визначається залежність між логарифмом частоти слова $\ln(f)$ та логарифмом його рангу $\ln(n)$, а параметри розподілу такі: $a \approx -0{,}951$, $b \approx 1{,}531$, $c \approx -0{,}285$ та $d \approx 8{,}822$. Коефіцієнт детермінації $R^2 \approx 0{,}964$. При цьому було застосовано

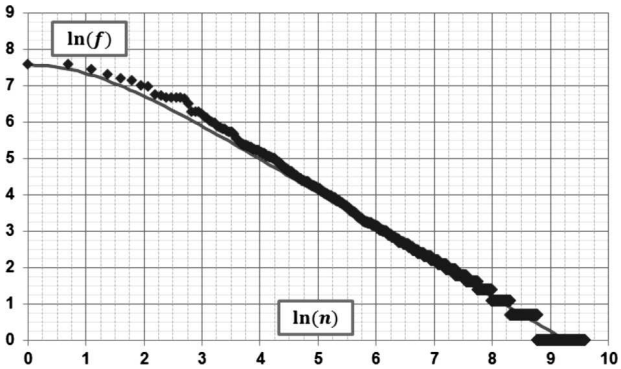

Рис. 4. Рангово-частотна залежність для тексту "Записки українського самашедшого". Маркерами позначені "експериментальні" значення. Суцільна пряма відповідає апроксимації на основі залежності вигляду (11)

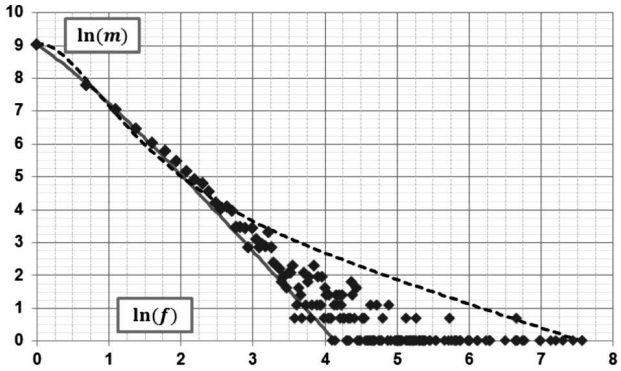

Рис. 5. Спектральний розподіл слів для тексту "Записки українського самашедшого". Маркерами позначені "експериментальні" значення. Суцільна пряма відповідає апроксимації на основі закону (11) по мінімальних значеннях, а штрихована крива відповідає апроксимації по максимальних значеннях

дві додаткових умови: (i) значення апроксимаційної функції має збігатися з "експериментальним" значенням в початковій точці; (ii) похідна не може бути більше нуля.

При моделюванні спектрального розподілу стикаємося з тією проблемою, що за збільшення частоти відповідна функція стає суттєво неоднозначною. В такому випадку, наприклад, можемо виконувати апроксимацію, фіксуючи значення першої і останньої точки в залежності. На рис. 5 наведено результати апроксимації для тексту "Записки українського самашедшого" *Ліни Костенко* [36] (дані отримано за допомогою інформаційного ресурсу *www.mova.info*). Якщо виконувати апроксимацію, зафіксувавши значення першої точки

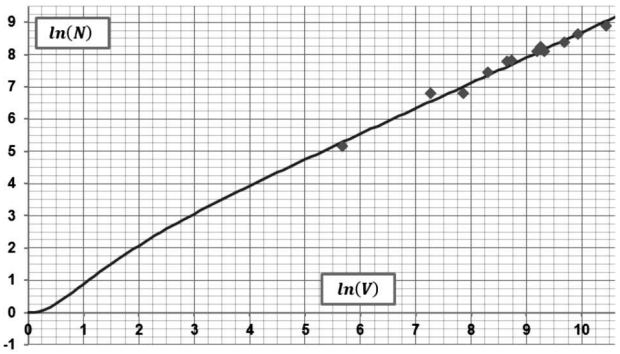

Рис. 6. Залежність розміру словника від об'єму тексту (для творів Тараса Прохаська). Маркери позначають "експериментальні" значення. Суцільна крива відповідає апроксимаційній залежності вигляду (11))

(нульове значення для логарифма частоти появи слова $\ln(f)$) і першої точки, в якій логарифм кількості слів з даною частотою $\ln(m)$ обертається на нуль, то отримуємо такі значення для параметрів апроксимації: $a \approx -2{,}424$, $b \approx 1{,}081$, $c \approx -0{,}045$ та $d \approx 10{,}061$. Коефіцієнт детермінації $R^2 \approx 0{,}898$. Якщо ж зафіксувати останню точку з нульовим значенням логарифма кількості слів з відповідною частотою, то в результаті апроксимації отримуємо значення $a \approx -0{,}708$, $b \approx -8{,}643$, $c \approx 4{,}981$ та $d \approx 5{,}362$. Коефіцієнт детермінації $R^2 \approx 0{,}290$. В останньому випадку, вочевидь, навряд чи можна говорити про якісну апроксимацію, а скоріше про криву, котра описує граничні значення для відповідної залежності.

Нарешті, на рис. 6 наведено результати апроксимації на основі виразу вигляду (11) залежності між кількістю лексем в тексті (розмір словника) від об'єму тексту (на основі текстів *Тараса Прохаська*, дані отримано за допомогою інформаційного ресурсу *www.mova.info*). В процесі розрахунків використовувалися додаткові обмеження, які полягали у тому, що: (i) якщо текст складається із одного слова, то словник також складається з одного слова; (ii) кількість слів у словнику не може бути від'ємною. Розраховані (за вказаних додаткових умов) значення параметрів апроксимації становлять $a \approx 0{,}786$, $b \approx 2{,}670$, $c \approx -1{,}835$ та $d \approx 0{,}835$. Коефіцієнт детермінації $R^2 \approx 0{,}982$. В даному випадку ефект від застосування нелінійної залежності незначний, однак він дозволяє побудувати залежність, котра дає коректні значення навіть при малих об'ємах тексту.

5. Висновки

Таким чином, в статті запропоновано підхід, в рамках якого вдається створювати апроксимаційні залежності загального вигляду, котрі можуть застосовуватися в математичних моделях квантитативної лінгвістики. Базова ідея полягає у тому, аби використати для опису процесу чи співвідношення певного диференціального рівняння. Апроксимаційна залежність будується на основі загального розв'язку такого диференціального рівняння. Крім безпосередньої переваги, пов'язаної з можливістю отримати власне апроксимаційну залежність, даний підхід дозволяє виконувати класифікацію лінгвістичних моделей та визначати області їх застосовності, що у методологічному плані може мати вирішальне значення. Продемонстровані в статті приклади застосування зазначеного підходу дають підстави для сподівання, що він може бути перспективним і у інших випадках. Також застосування запропонованої методики може надати додаткове підтвердження та обґрунтування для результатів, отриманих у рамках альтернативних теорій – як це має місце, наприклад, для різних способів пояснення закону Зіпфа [21, 35, 50].

Автори висловлюють щиру подяку проф. Наталії Дарчук та її колегам за розробку та підтримку інформаційного ресурсу www.mova.info, котрий було використано під час роботи над статтею.

Також автори надзвичайно вдячні рецензентам за їх побажання та пропозиції, завдяки чому вдалося покращити статтю.

1. D. Walker. Economics and Social Physics. *Economic J.* **101**, 615 (1991).
2. R. Mantegna, H. Stanley. *An introduction to econophysics* (Cambridge University Press, 2000).
3. J. McCauley. *Dynamics of markets: econophysics and finance* (Cambridge University Press, 2004).
4. F. Jovanovic, C. Schinckus. Econophysics: A new challenge for financial economics. *J. Hist. Econ. Thought* **35**, 319 (2012).
5. Y. Gingras, C. Schinckus. The institutionalization of econophysics in the shadow of physics. *J. Hist. Econ. Thought* **34**, 109 (2012).
6. C. Schinckus, F. Jovanovic. Towards a transdisciplinary econophysics. *J. Econ. Method.* **20**, 164 (2013).
7. D. Sornette. Physics and financial economics (1776–2014): Puzzles, Ising and agent-based models. *Rep. Progr. Phys.* **77**, 1 (2014).

8. B. Chakrabarti, A. Chakraborti, A. Chatterjee. *Econophysics and Sociophysics: Trends and Perspectives* (Wiley-VCH, 2006).
9. S. Galam, Y. Gefen, Y. Shapir. Sociophysics: A mean behavior model for the process of strike. *J. Math. Soc.* **9**, 1 (1982).
10. S. Galam. *Sociophysics: A Physicist's Modeling of Psychopolytical Phenomena* (Springer, 2012).
11. D. Stauffer. A Biased Review of Sociophysics. arXiv: 1207.6178v1.
12. C. Castellano, S. Fortunato, V. Loreto. Statistical physics of social dynamics. *Rev. Mod. Phys.* **81**, 591 (2009).
13. S. Galam. Sociophysics: A review of Galam models. arXiv: 0803.1800.
14. B. Berche, C. von Ferber, T. Holovatch, Yu. Holovatch. Transportation network stability: A case study of city transit. *Adv. Complex Syst.* **15**, 1, 1250063 (2012).
15. Y. Holovatch, V. Palchykov. *Complex Networks of Words in Fables In: Maths Meets Myths: Complexity-science approaches to folktales, myths, sagas, and histories.* Edited by R. Kenna, M. Mac Carron, P. Mac Carron (Springer, 2016).
16. Yu. Holovatch, R. Kenna, S. Thurner. Complex systems: Physics beyond physics. *Eur. Journ. Phys.* **38**, 023002 (2017).
17. Ю. Головач, М. Дудка, В. Блавацька, В. Пальчиков, М. Красницька, О. Мриглод. *Статистична фізика складних систем.* Препринт І СМР-17-06U (Львів, 2017).
18. Ю. Головач, М. Дудка, В. Блавацька, В. Пальчиков, М. Красницька, О. Мриглод. Статистична фізика складних систем у світі та у Львові. *ЖФД* **22**, 2, 2801 (2018).
19. G. Altmann, R. Köhler. "Language Forces" and synergetic modelling of language phenomena. *Glottometrika* **15**, 62 (1996).
20. R. Köhler. Synergetic linguistics. In: *Quantitative Linguistics. An International Handbook* (Walter de Gruyter, 2005).
21. Ю. Тулдава. *Проблемы и методы квантитативно-системного исследования лексики* (Валгус, 1987).
22. Р. Пиотровский, К. Бектаев, А. Пиотровская. *Математическая лингвистика* (Высшая школа, 1977).
23. Р. Пиотровский. *Лингвистическая синергетика: исходные положения, первые результаты, перспективы* (Санкт-Петербургский гос. ун-т, 2006).
24. В.В. Левицкий. *Квантитативные методы в лингвистике* (Рута, 2005).
25. Ю. Головач, В. Пальчиков. Лис Микита і мережі мови. *Журнал фізичних досліджень* **11**, 1, 22 (2007).
26. A.A. Rovenchak, S. Buk. Defining thermodynamic parameters for texts from word rank-frequency distributions. *J. Phys. Stud.* **15**, 1, 1005 (2011).
27. A.A. Rovenchak, S. Buk. Application of a quantum ensemble model to linguistic analysis. *Phys. A* **390**, 7, 1326 (2011).
28. A. Rovenchak, S. Buk. Part-of-Speech Sequences in Literary Text: Evidence From Ukrainian. *J. Quant. Ling.* (2017).
29. О.М. Васильєв, О.В. Чалий, І.В. Васильєва. Про "екзотичні" задачі фізики, Вінні-Пуха та закон Зіпфа. *ЖФД* **17**, 1, 1001 (2013).
30. A. Vasilev, I. Vasileva. Text length and vocabulary size: Case of the Ukrainian writer Ivan Franko. *Glottometrics* **43**, 1 (2018).
31. G. Zipf. *Human Behavior and the Principle of Least Effort* (Addison-Wesley, 1949).
32. G. Zipf. *The Psycho-Biology of Language* (Addison-Wesley, 1935).
33. W. Li. Zipf's law everywhere. *Glottometrics* **5**, 14 (2002).
34. I.-I. Popescu, G. Altmann, R. Köhler. Zipf's law another view. *Qual. Quant.* **44**, 4, 713 (2010).
35. В. Пальчиков. *Ефекти безмасштабності та тісного світу в складних мережах* (Кандидатська дисертація, Львів: 2010).
36. Л. Костенко. *Записки українського самашедшого* (А-БА-БА-ГА-ЛА-МА-ГА, 2014).
37. I. Moreno-Sánchez, F. Font-Clos, Á. Corral. Large-Scale Analysis of Zipf's Law in EnglishTexts. *PLoS ONE* **11(1)**: e0147073 (2016).
38. G. Thomson, J.R. Thompson. Outline of a measure for the quantitative analysis of writing vocabularies. *Brit. J. Psychol.* **8**, 52 (1915).
39. G. Herdan. *Type-token mathematics: A textbook of mathematical linguistics* (Gravenhage, 1960).
40. J. Tuldava. The statistical structure of a text and its readability. In *Quantitative text analysis* (Wissenschaftlicher Verlag Trier, 1993).
41. J. Tuldava. On the relation between text length and vocabulary size. In *Methods in Quantitative Linguistics* (Wissenschaftlicher Verlag Trier, 1995).
42. E. Panas. The Generalized Torquist: Specification and Estimation of a New Vocabulary-Text Size Function. *J. Quant. Linguist.* **8**, 233 (2001).
43. E. Panas, A.N. Yannacopoulos. Stochastic Models for the Lexical Richness of a Text: Qualitative Results. *J. Quant. Linguist.* **11**, 251 (2004).
44. G. Wimmer. The type-token relation. In *Quantitative Linguistics. An International Handbook* (Walter de Gruyter, 2005).
45. R. Köhler. Synergetic linguistics. In *Quantitative Linguistics. An International Handbook* (Walter de Gruyter, 2005).

46. F. Fan. Text length, vocabulary size and text coverage constancy. *J. Quant. Linguist.* **20**, 288 (2013).

47. M. Kubát, J. Milička. Vocabulary Richness Measure in Genres. *J. Quant. Linguist.* **20**, 339 (2013).

48. D. Mitchell. Type-token models: a comparative study. *J. Quant. Linguist.* **22**, 1 (2015).

49. F. Fan, Y. Yang, W. Yaqin. The Probability Distribution of Textual Vocabulary in the English Language. *J. Quant. Linguist.* **23**, 49 (2016).

50. S. Thurner R. Hanel, B. Liu, B. Corominas-Murtra. Understanding Zipf's law of word frequencies through sample-space collapse in sentence formation. *Journ. Royal Soc. Interface* **12** 20150330 (2015).

Одержано 15.10.19

A.N. Vasilev, I.V. Vasileva

PHYSICS BEYOND PHYSICS:
APPLICATION OF PHYSICAL APPROACHES
IN QUANTITATIVE LINGUISTICS

S u m m a r y

The application of physical methods to solve non-physical problems has been considered. In particular, the prospects of physical approaches in quantitative linguistics are analyzed. The difference between the physical and non-physical methods is illustrated by an example of already existing "classical" models. A few mathematical models which make it possible to determine the rank-frequency dependence for words in a frequency dictionary, as well as the dependence of the dictionary volume on the text length, are proposed. It is shown that the physical approaches and principles that are used in physics can also be successfully applied to create mathematical models in linguistics.