СТАТТІ

Олексій ВАСИЛЬЄВ, Олександр ЧАЛИЙ,
Ілона ВАСИЛЬЄВА

Математичні методи й моделі в лінгвістиці[1]

Олексій ВАСИЛЬЄВ *– доктор фізико-математичних наук, професор катедри теоретичної фізики фізичного факультету Київського національного університету імені Тараса Шевченка. Сфера наукових та професійних інтересів: програмування та математичне моделювання, синергетика, біофізика, фізика рідин, математична економіка, математична лінгвістика. Електронна адреса: alex@vasilev.kiev.ua*
Олександр ЧАЛИЙ *– доктор фізико-математичних наук, професор, член-кореспондент НАПН України, завідувач катедри медичної і біологічної фізики та інформатики Національного медичного університету імені О. О. Богомольця. Сфера наукових і професійних інтересів: медична фізика, медична інформатика, критичні явища та фазові переходи, фізика рідин, синергетика, біофізика, математична лінгвістика. Електронна адреса: avchal@univ.kiev.ua*
Ілона ВАСИЛЬЄВА *– лінгвістка. Закінчила Інститут філології Київського національного університету імені Тараса Шевченка (спеціялізація – комп'ютерна лінгвістика). Сфера наукових та професійних інтересів: сучасна українська мова, автоматизовані лінгвістичні системи та комплекси, застосування математичних методів у лінгвістиці. Електронна адреса: ilona@vasilev.kiev.ua*

У статті розглянуто підходи, актуальні для побудови та використання математичних моделей у лінгвістиці. Проаналізовано типи задач лінгвістики, під час розв'язання яких видається доцільним застосовувати математичні методи. Також досліджено основні принципи створення математичних моделей. Зокрема, наголошено, що ефективним може бути підхід, який на сьогодні переважно застосовують у розв'язанні фізичних задач. Його типовою ознакою є те, що в основу моделі покладено певну теорію універсального типу, яка пояснює взаємодію функціональних елементів системи на найбільш загальному рівні. У такому разі потрібні апроксимаційні залежності не вгадуються, як це часто

[1] Автори висловлюють щиру подяку *Юрію Головачу* та всім організаторам другої міждисциплінарної майстерні "Verba et numeri" (Львів, 2017 рік).

відбувається, а їх отримують чи розраховують, спираючись на базову теорію. Визначено перспективи застосування такого підходу в лінгвістиці.

Наведено приклади моделювання лінгвістичних систем. Подано порівняльний аналіз різних способів моделювання. Значну увагу приділено універсальності математичних моделей, які застосовуються для аналізу лінгвістичного матеріялу. Відзначено: якщо використані апроксимаційні залежності базуються на загальних універсальних властивостях системи, це може значно ускладнити кількісний аналіз даних. Проілюстровано переваги й недоліки математичного моделювання у розв'язанні прикладних лінгвістичних задач.

Крім «класичних» моделей (таких, як закон Ціпфа для ранґового розподілу слів у тексті), описано й інші способи моделювання. Зокрема, увагу приділено методам побудови математичних моделей на основі нейронних мереж. У такому разі апроксимаційна залежність реалізується у вигляді послідовних суттєво нелінійних перетворень. Головний недолік цього підходу пов'язаний із технічними складностями в його реалізації та з відсутністю явної аналітичної залежности. Головна ж перевага прихована в потенційній можливості створювати апроксимаційні залежності практично будь-якого типу. Для зіставлення різних способів моделювання у статті наведено приклади розв'язання одних і тих самих лінгвістичних задач за допомогою кількох методів.

Ключові слова: модель, теорія, ранґовий розподіл, апроксимація, нейронна мережа.

Кожна формула в книзі вдвічі зменшує кількість читачів.
С. Гокінґ

1. Вступ

Застосування математичних методів у найрізноманітніших наукових дослідженнях є типовою ознакою сучасности. Причому математику все більше використовують не тільки в соціяльних науках, але й в гуманітарних. Причин для того є кілька. По-перше, надзвичайно потужний розвиток комп'ютерних технологій на якісному рівні змінив характер досліджень і значно розширив сферу застосування відповідного математичного апарату. По-друге, нові технічні можливості оброблення й аналізу великих масивів даних породжують новий клас задач. По-третє, яскраво виражений міждисциплінарний характер сучасних досліджень розширює межі застосування методів одних наук в інших. Не є винятком і математичне моделювання.

Серед гуманітарних дисциплін лінгвістика, мабуть, є однією з найбільш досліджених цариц і при цьому однією з найбільш загадкових. Людське існування не можна уявити без мови, вона не просто є супутником людини, вона в певному розумінні і є людина. Мова — це не тільки інструмент для спілкування. Це соціокультурний феномен, який супроводжує людство від

початку його існування, змінюється разом із людським соціюмом, впливає на розвиток цього соціюму і є найяскравішим відображенням процесів, які відбуваються з людством. Тому досліджуючи мову (в широкому розумінні цього поняття), ми постійно стикаємось із чинниками, які не є суто лінгвістичними. Як наслідок, у лінгвістичних дослідженнях застосовують методи різних наук – соціяльних і гуманітарних. Має застосування в лінгвістичних дослідженнях і математика. Причому способи реалізації математичних підходів є досить різноманітні: від розрахунку статистичних характеристик текстів до створення реґресійних моделей. Цікавими є не тільки самі моделі, але й спосіб їхньої побудови. Як це не дивно, але «ідеологія» побудови тієї чи тієї моделі має значення так само, як має значення конкретний «зміст» моделі. Іншими словами, шлях, яким дослідник приходить до формулювання моделі, має окрему цінність. Зокрема, останнім часом збільшується кількість наукових публікацій, у яких методи, застосовувані раніше для суто фізичних систем, використовують для моделювання в лінгвістиці (і не тільки). Такий «фізичний» спосіб моделювання має свої специфічні риси. Їхній аналіз є частиною статті.

Отже, ми обговоримо особливості застосування математичного моделювання в лінгвістиці. Тема надзвичайно широка, тому зосередимось лише на деяких її аспектах. Передусім нас цікавитимуть відповіді на такі запитання:

– Навіщо потрібне моделювання в лінгвістиці?
– Як створюються моделі в лінгвістиці і що вони описують?
– Чи має «фізичний» спосіб моделювання якісь особливості?

Також ми розглянемо кілька прикладів (переважно ілюстративних), які демонструють специфіку досліджуваних проблем та методів їхнього розв'язання.

1. Що і як ми моделюємо

Хоча на перший погляд видається, що моделювання – це певна універсальна процедура, яка має чітко означену кінцеву мету та суворо детермінований спосіб її досягнення, насправді, тут багато неоднозначного. Почати хоча би з того, *що* саме ми плануємо моделювати. Питання нетривіяльне, оскільки зовсім не кожен процес чи систему можна змоделювати математичними методами. Гарним прикладом є модель випадкової величини в теорії ймовірностей. Ідеться про те, що в певному «експерименті» отримуємо випадкове значення для певної величини. Наприклад, це може бути число, що випадає на кубику, чи номер, на який потрапляє кулька під час гри в казино. Зрозуміло, що жодне моделювання не дозволить нам передбачити результат подібних «експериментів». Однак моделювання має широке застосування навіть у таких «безнадійних» випадках. Ключовий момент полягає в тому, *що* саме ми моделюємо чи оцінюємо за допомогою математичних моделей. У наведених прикладах ми можемо математичними методами розрахувати параметри для закону розподілу випадкової величини, що дає нам змогу оцінити ймовірність настання тієї чи тієї події, пов'язаної з окремою випадко-

вою величиною. Отже, принциповим є не тільки спосіб математичного опису процесу чи системи, а вибір самого об'єкта для моделювання.

Моделювання, зокрема й у лінгвістиці, не є самоціллю. Створюючи ту чи іншу модель, дослідник намагається розв'язати конкретну задачу. Можна виділити окремі класи лінгвістичних (чи пов'язаних із лінгвістикою) задач, розв'язання яких вимагає чи передбачає залучення математичного моделювання. Прикладом є створення програмних продуктів для виконання машинного (автоматизованого) перекладу текстів чи розроблення експертних систем для аналізу текстів та живої мови. Методи математичного моделювання мають застосування у психологічній лінгвістиці, статистичному аналізі корпусів мови, криптографії та в багатьох інших ситуаціях[2].

2. Чи є «фізичний» підхід особливим

Загалом створення математичної моделі – це завдання для фахівців. Проблема в тому, що для моделювання лінгвістичних об'єктів та процесів, крім знання математики, необхідне глибинне розуміння об'єкта моделювання. Це є типовою ознакою дослідження, що виконують на межі різних наук. На практиці проблему розв'язують формуванням міждисциплінарних груп дослідників. З другого боку, важливу роль відіграє аспект, пов'язаний із кінцевою метою моделювання. Як це не дивно, але є відмінні концептуальні підходи в застосуванні математичних моделей. Різні точні науки тяжіють до різних концепцій застосування математичного апарату. Є свої особливості в економічних моделей, так само як є типові риси моделі, що їх використовують у фізиці. Оскільки методи моделювання в лінгвістиці виникли не з нічого, то, як наслідок, принцип створення моделей у лінгвістиці має «відбиток» того чи того магістрального підходу. Для нас актуальною є відповідь на питання про адекватність підходів, які застосовують фізики для опису явищ неживої природи, для використання в лінгвістичних дослідженнях. І якщо ці підходи справді адекватні, то де і як вони можуть бути застосовані. Для цього нам необхідно врахувати, які взагалі є способи використання математичних моделей. Отже, можна умовно назвати три фундаментальні підходи чи концепції, які передбачають застосування математичних моделей на практиці.

Передусім це моделі, які описують певне явище чи процес і результати яких порівнюються з емпіричними даними (моделі першого типу). Як правило, такого типу моделі використовують у точних науках, і, зокрема, у фізиці. Типовою рисою таких моделей є те, що інтерес викликає саме кінцевий результат, який може бути одержаний у вигляді числа, і це число можна порівняти зі значенням, яке вимірюється в експерименті. Назагал математичні моделі розуміють як моделі саме такого типу. Справді, у багатьох ситуаціях нас цікавить саме зазначений спосіб моделювання. Водночас є й інші можливості для ефективного застосування математичного моделювання.

[2] Юхан Тулдава, *Проблемы и методы квантитативно-системного исследования лексики* (Таллин: Валгус, 1987).

Навіть у точних науках використовують феноменологічні математичні моделі, що описують поведінку системи чи певний процес не на кількісному, а на якісному рівні (моделі другого типу)[3]. Такі моделі створюють на основі найзагальніших уявлень щодо об'єкта, який моделюється, а функціональні залежності, які використовують у моделях, мають евристичний характер. Результати, одержані на основі таких моделей, дають опис системи лише на якісному рівні. Причому не йдеться про «неповноцінність» моделей. Навіть більше, часто такі моделі дозволяють передбачити (на якісному рівні) ефекти та процеси, які не є очевидними і які складно спрогнозувати за допомогою інших методів досліджень. Застосування цих моделей вимагає використовувати математичний апарат, не менш серйозний, ніж під час роботи з кількісними моделями. Серед іншого, надійність результатів, одержаних моделюванням, перевіряється за допомогою аналізу на топологічну стійкість вихідної моделі (на основі якої одержано результат)[4].

Нарешті, є клас моделей, які дозволяють одержувати цілком конкретні кількісні результати, але головне призначення таких моделей (на відміну від моделей першого типу) полягає не в одержанні оцінок для результатів емпіричних вимірювань, а радше в класифікації систем на основі відповідних числових характеристик. Як правило, ці моделі статистичні (моделі третього типу).

У точних науках, таких як фізика, назагал використовують моделі першого чи другого типу. Моделі третього типу більш характерні для соціяльних та гуманітарних наук. Пояснення полягає не тільки в меті, яку дослідники намагаються досягти за допомогою моделювання. Важливим є також алгоритм побудови моделей та специфіка об'єктів, які описують за допомогою моделювання.

3. Теорії та моделі

«Фізичний» спосіб моделювання має одну важливу базову особливість – в основі фізичної моделі, як правило, лежить теорія. Тобто майже будь-яка сучасна фізична модель описує певну систему чи процес, і ця модель базується на певних уявленнях про характер взаємодій, які є в системі або які реалізуються в процесі. Таким моделям можна протиставити моделі «описові», які дозволяють описати систему чи навіть передбачити її динаміку. При цьому «описові» моделі не дають пояснення щодо механізмів взаємодії і не базуються на припущеннях про певний характер чи особливості цих механізмів. Фізика, до речі, дає приклади як суто «фізичних» моделей, так і «описових».

Одна з історично перших фізичних моделей – третій закон Кеплера. Закон стверджує, що квадрати періодів обертання планет відносяться як куби великих півосей їхніх еліптичних орбіт. Цю закономірність встановив екс-

[3] Владимир Арнольд, *«Жесткие» и «мягкие» математические модели* (Москва: МЦНМО, 2013).

[4] Там само.

периментально Кеплер під час аналізу статистичних даних щодо руху планет. Цей закон має нормативний характер і не пояснює причин, унаслідок яких виконується зазначене співвідношення.

На противагу третьому закону Кеплера другий закон Ньютона встановлює зв'язок між прискоренням $\vec{a}$ тіла, масою m тіла, та силою $\vec{F}$, що діє на це тіло:

$$\vec{F} = m\vec{a}\,.$$

Цей закон насправді є диференціяльним рівнянням (оскільки компоненти вектора прискорення – це другі похідні від координат тіла, а сила залежить від координат), розв'язуючи яке, ми можемо отримати закон руху тіла. Зокрема, можна і отримати закони руху небесних тіл, і довести істинність третього закону Кеплера. Тобто маємо дві моделі, одна з них є більш загальною. Однак принципова відмінність між моделями не в цьому, а в тому, що за законом Ньютона ховається теорія, яка пояснює характер взаємодій у системі й наслідки їхнього впливу на динаміку об'єктів. А закон Кеплера є лише результатом оброблення статистичних даних – тобто фактично це вдало побудована реґресійна модель.

Отже, умовно моделі можна поділити на «фізичні» й «нефізичні». «Фізична» модель пояснює механізми взаємодій у системі та дозволяє передбачити поведінку системи в майбутньому. «Нефізична» модель встановлює кількісні співвідношення між характеристиками системи та дозволяє виконувати класифікацію систем чи робити оцінки для параметрів системи. Слід зазначити, що обидва типи моделей мають широке застосування. Тобто тут не йдеться про «правильні» чи «неправильні» моделі. Інша справа, що моделі різних типів мають різне походження й різну область застосування.

4. Закон Ціпфа та реальне життя

Добре відомим у лінгвістиці і поза її межами є закон Ціпфа, який пов'язує частоту слова F із його ранґом n[5]. Відповідне співвідношення, яке багато разів було перевірено емпірично для різних текстів і різних мов, має вигляд:

$$F = \frac{A}{n^{\gamma}},$$

де γ є параметром розподілу і в багатьох випадках має близьке до одиниці значення. На практиці також часто застосовують модифікований закон Ціпфа, або закон Ціпфа з поправкою Мандельброта[6]:

[5] George Zipf, *Human Behavior and the Principle of Least Effort* (Cambrige: Addison-Wesley, 1949); Wentian Li, "Zipf's law everywhere", *Glottometrics*, 5, (2002): 14–21; Ioan-Iovitz Popescu, Gabriel Altmann, Reinhard Köhler, "Zipf's law another view", *Quality and Quantity*, 44, 4, (2010): 713–731.

[6] Юхан Тулдава, *Проблемы и методы квантитативно-системного исследования лексики* (Таллин: Валгус, 1987).

$$F = \frac{A}{(n + n_0)^{\gamma}}.$$

У цьому співвідношенні введено додатковий параметр розподілу n_0. Причина проста – виконати кращу апроксимацію статистичних даних (адже що більше параметрів у моделі, то її простіше «припасувати» до емпіричних даних).

Закон Ціпфа також може бути представлений у такому вигляді:

$$\ln(F) = \ln(A) - \gamma \ln(n).$$

Тут ми маємо лінійну залежність між натуральним логарифмом частоти появи слова $\ln(F)$ та натуральним логарифмом ранґу слова $\ln(n)$. Якщо таку залежність зобразити на графіку, на горизонтальній осі відклавши логарифм ранґу слова, а на вертикальній осі – логарифм частоти слова, то отримаємо пряму лінію. Такий графік наведено на рис. 1, на якому проілюстровано залежність логарифма частоти слова від логарифма ранґу слова для корпусу художніх текстів естонської мови[7]. Зокрема, обсяг тексту становив 99 898 слів, обсяг словника становив 30 733 слова, і відповідно обчислені значення $A = 4\,095$ та $\gamma = 0.86$ для параметрів розподілу, що входять у закон Ціпфа.

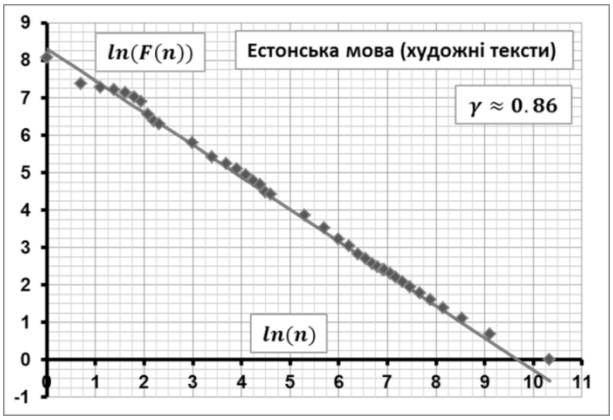

Рис. 1. *Частотний розподіл слів для естонської мови*

Закон Ціпфа емпіричний. Тобто в певному розумінні він нагадує третій закон Кеплера, оскільки обидва були одержані за допомогою оброблення й узагальнення статистичних даних. Питання про пояснення закону Ціпфа фактично залишається відкритим, хоча не можна сказати, що пояснення зовсім відсутні. Іншими словами, за законом Ціпфа, є не чітка теорія, а кілька концепцій, які в сукупності дають певні підстави для обґрунтування закону.

[7] Дані для розрахунку й побудови графіка на рис. 1 взято там само.

Також є низка моделей, які пояснюють утворення степеневої залежности між ранґом слова і його частотою в законі Ціпфа. Наприклад, модель Саймона базується на припущеннях, що додавання слів у текст є випадковим процесом, імовірність появи нового слова є сталою величиною, а ймовірність появи слова, що вже траплялось, залежить від частоти вживання цього слова[8]. У «термодинамічній» теорії Мандельброта[9] письмова мова ототожнюється з кодуванням, у якому кожен символ має «вартість». Умова мінімальної «вартости» повідомлення дає степеневий ранґовий розподіл. Ще один спосіб пояснення закону Ціпфа має радше математичний характер. Зокрема, степеневий розподіл можна отримати як розв'язок такого диференціяльного рівняння:

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

і тут через dF позначено зміну частоти появи слова в тексті, а через dn – зміну ранґу слова. Отже, закон Ціпфа можна сформулювати так: відносна зміна частоти слова (величина dF/F) пропорційна до відносної зміни ранґу слова (величина dn/n). Це так званий алометричний закон[10]. Однак і тут залишається не розв'язане питання про те, чому є саме таке співвідношення між відносною зміною ранґу й частоти слова.

Закон Ціпфа широко застосовують на практиці. Наприклад, для аналізу текстів чи живого мовлення окремих індивідів. Скажімо, відомо[11], що мова психічно здорових людей описується степеневим розподілом з показником $\gamma \approx 1$. Для психічно хворих людей так само відбувається ранґовий розподіл слів (мається на увазі залежність частоти слова від його ранґу), але вже з показником $\gamma \approx 0.7$. Цю обставину використовують у психолінгвістиці для аналізу психологічного стану індивідів (на предмет наявности психічних захворювань)[12]. Проілюструємо невеликим прикладом. Відправною точкою для реалізації відповідного дослідження є цитата з книги Ліни Костенко «Записки українського самашедшого»[13]: «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». Якщо виходи-

[8] Василь Пальчиков, *Ефекти безмасштабності та тісного світу в складних мережах* (Кандидатська дисертація, Львів: 2010).

[9] Там само.

[10] Юхан Тулдава, *Проблемы и методы квантитативно-системного исследования лексики* (Таллин:Валгус, 1987).

[11] Значення показника γ залежить від рівня аналітичности мови, близьким до одиниці воно є, наприклад, для англійської, тоді як для української – як більш синтетичної (тобто з багатшою словозміною) – воно трохи менше, а для китайської (більш аналітичної) – дещо більше. Таку залежність потрібно враховувати, інтерпретуючи результати. – *Прим. ред.*

[12] Критерій застосовний лише для повсякденного мовлення. Вузькоспеціяльні, технічні й художні тексти в цю класифікацію не потрапляють.

[13] Ліна Костенко, *Записки українського самашедшого* (Львів: А-БА-БА-ГА-ЛА-МА-ГА, 2014).

ти з того, що тут висловлена гіпотеза щодо психічного стану такого казкового героя, як Вінні-Пух, то її можна перевірити. Для цього будується ранґовий розподіл словоформ у тексті казки про Вінні-Пуха, а також ранґовий розподіл слів у прямій мові Вінні-Пуха. Для обох випадків розраховується показник розподілу γ. На рис. 2 наведено ранґовий розподіл для тексту казки, а на рис. 3 – для тексту прямої мови Вінні-Пуха[14].

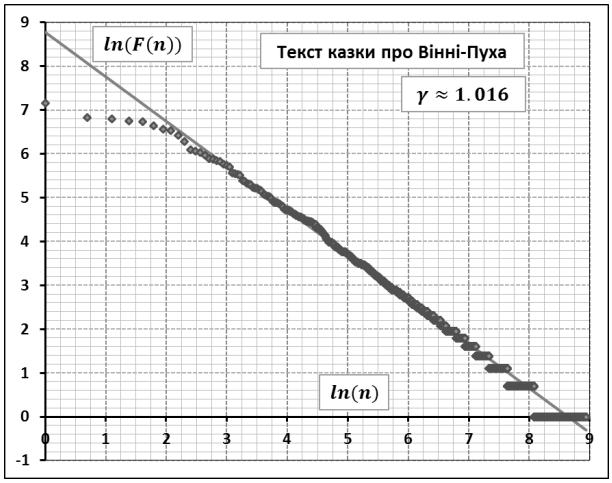

Рис. 2. *Ранґовий розподіл словоформ для тексту казки про Вінні-Пуха*

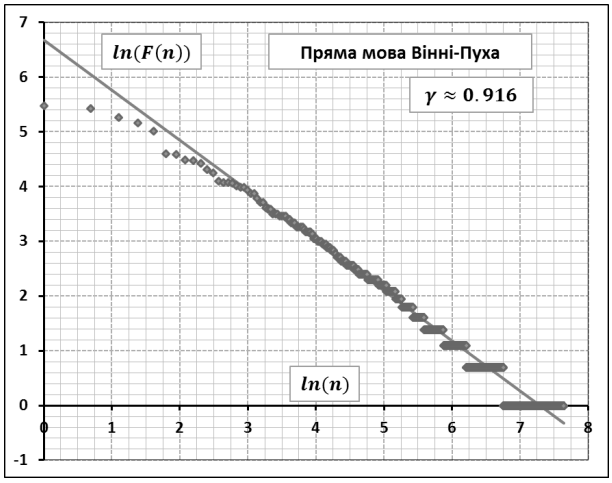

Рис. 3. *Ранґовий розподіл словоформ для прямої мови Вінні-Пуха*

[14] Дані для побудови графіків на рис. 2-3 взято з роботи: Олексій Васильєв, Олександр Чалий, Ілона Васильєва, «Про екзотичні задачі фізики, Вінні-Пуха та закон Зіпфа», *Журнал фізичних досліджень*, 17, 1 (2013): 1001 [8 pp.].

Як показують розрахунки, для тексту казки про Вінні-Пуха параметр $\gamma \approx 1.016$, тоді як для прямої мови Вінні-Пуха параметр $\gamma \approx 0.916$. Обидва значення близькі до одиничного, а отже, немає підстав хвилюватись за психічний стан Вінні-Пуха[15]. Слід також зазначити, що було проаналізовано український переклад казки. Якщо аналогічні обчислення виконати для ориґіналу англійською, то для тексту казки отримаємо значення параметра $\gamma \approx 1.26$, а для прямої мови Вінні-Пуха параметр $\gamma \approx 1.08$. Тому загальний висновок залишається незмінним.

5. Неуніверсальність моделей

Зовсім не кожна статистична модель (тобто модель, побудована на основі оброблення статистичних даних) є універсальною. Адже коли ми будуємо на основі статистичних даних модель, то передусім розраховуємо параметри розподілу (наприклад, показник γ для розподілу Ціпфа). Від чого залежать значення параметрів розподілу? Очевидно, вони залежать від фактичних статистичних даних. Однак може статись ситуація, коли результати моделювання якісно залежать ще й від того, як ми групуємо статистичні дані. Як ілюстрацію розглянемо процес (дещо спрощений) побудови моделі для розподілу слів давнього походження за частотними зонами. Алгоритм побудови моделі такий: слова в частотному словнику об'єднують у частотні групи по 100 слів, у кожній групі підраховують кількість давніх слів, що виникли в певний момент[16]. Будують апроксимаційну формулу для залежности кількости давніх слів у групі від номера групи. Використовують функціональну залежність виду:

$$F(n) = A \cdot \exp(a \cdot n^b),$$

і тут через n позначено номер групи, через $F(n)$ – кількість давніх слів у цій групі, параметр A визначається кількістю слів у групі, а параметри розподілу a та b знаходять на основі статистичних даних[17]. На рис. 4 наведено результати моделювання для розподілу давніх слів XII століття для німецької мови, а на рис. 5 показано аналогічну залежність, але для розподілу за групами давніх слів XII століття для естонської мови (залежності побудовано на основі даних роботи[18]).

Для німецької мови отримано значення для параметрів розподілу $a = 0.07$ та $b = 1.0$, а для розподілу давніх слів в естонській мові ці параметри дорівнюють $a = 0.11$ та $b = 0.96$. Водночас параметри не є універсальними і залежать, зокрема, від того, скільки слів належить до групи. Зокрема, якщо розглядати групи по 200 слів (тобто за значення параметра $A = 200$), отри-

[15] Там само.

[16] Юхан Тулдава, *Проблемы и методы квантитативно-системного исследования лексики* (Таллин: Валгус, 1987).

[17] Там само.

[18] Дані для побудови графіків на рис. 4-7 взято там само.

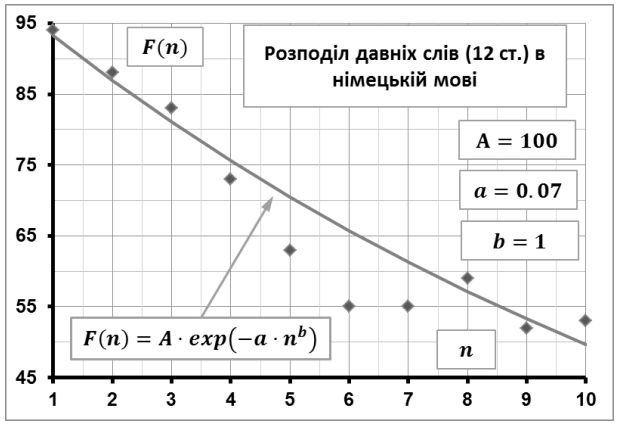

Рис. 4. *Розподіл давніх слів (XII століття) для німецької мови за умови, що група містить 100 слів*

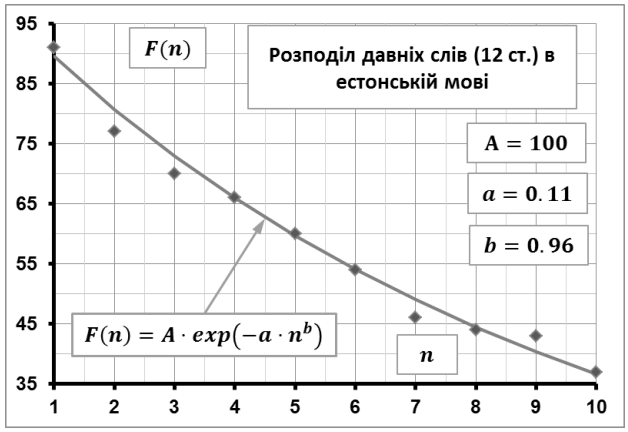

Рис. 5. *Розподіл давніх слів (XII століття) для естонської мови за умови, що група містить 100 слів*

маємо інші значення для параметрів a та b. Відповідні графіки наведено на рис. 6 та рис. 7.

Параметри розподілу змінюються. Для німецької мови маємо значення $a = 0.13$ та $b = 1.04$, а для естонської — значення $a = 0.19$ та $b = 1.02$. Однак зовсім не це найважливіше. Річ у тім, що якщо виходити із припущення, що для груп зі 100 слів розподіл давніх слів визначається залежністю виду $F(n) = A \cdot \exp(a \cdot n^b)$, то за збільшення кількости членів у групі в два рази кількість давніх слів у такій групі мала би визначатись як $F(n) + F(n + 1)$ (дві групи об'єднуються в одну). Однак є таке співвідношення:

$$F(n) + F(n + 1) = A \cdot (\exp(a \cdot n^b) + \exp(a \cdot (n + 1)^b)),$$

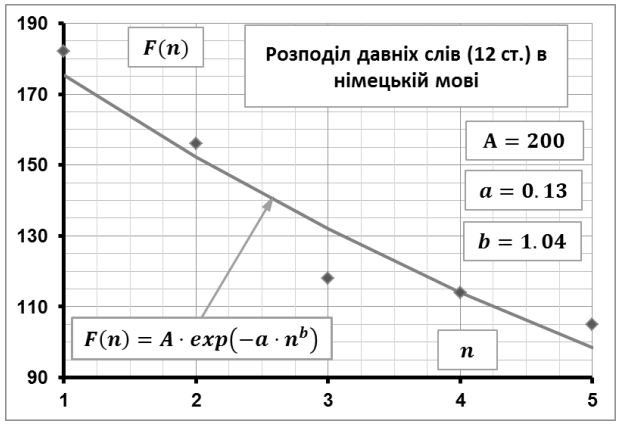

Рис. 6. *Розподіл давніх слів (XII століття) для німецької мови за умови, що група містить 200 слів*

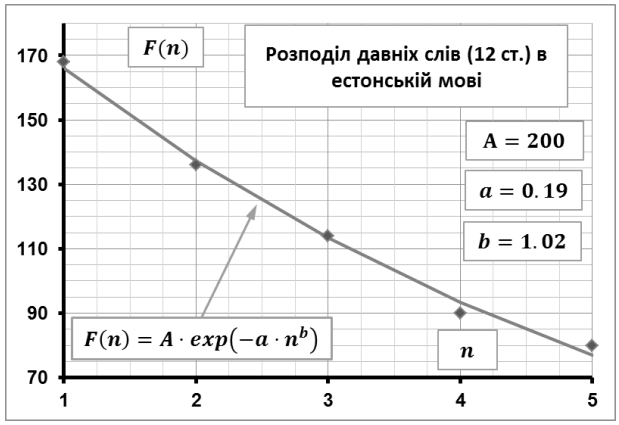

Рис. 7. *Розподіл давніх слів (XII століття) для естонської мови за умови, що група містить 200 слів*

і тут ми отримуємо зовсім іншу функціональну залежність порівняно з вихідною формулою для розподілу давніх слів у групах по 100 слів. Тому коли ми намагаємось описати «нові» статистичні дані «старою» формулою, отримуємо інші числові значення для параметрів розподілу. Іншими словами, йдеться не тільки про неуніверсальність значень параметрів розподілу, але й про неуніверсальність самої функціональної залежности.

6. Коли не працюють моделі

Спробуємо «локалізувати» проблему, із якою ми зіткнуся вище. Для цього необхідно врахувати спосіб, за яким побудовано моделі реґресійного типу (статистичні моделі, що встановлюють зв'язок між різними

наборами даних). Умовно процес побудови моделі можна розділити на кілька етапів. Передусім слід вибрати чи встановити функціональну залежність, на основі якої буде виконуватись апроксимація. Тут універсальних рецептів немає. Досить часто береться якась загальна функція, що містить низку «припасованих» параметрів. Ці параметри впливають на загальний вигляд функціональної залежности й визначаються на основі статистичних даних. Наступний етап – це розрахунок параметрів, що входять у функціональну залежність. Базова ідея полягає в тому, щоби дібрати такі значення для параметрів, за яких теоретична крива (залежність, побудована на основі апроксимуючої функції) якомога краще б описувала реальні статистичні дані. Однак щоб можна було порівнювати, який спосіб опису «кращий», потрібен критерій. Як правило, таким критерієм є метод найменших квадратів: значення для параметрів у функціональній залежності добираються так, щоби сума квадратів відхилень теоретичних (розрахованих на основі апроксимаційної функції) значень досліджуваного чинника від його безпосередньо виміряних (емпіричних) значень була мінімальною.

Якщо загальний вигляд для функціональної залежности вибрано, то розрахунок значень параметрів, що входять у функціональну залежність, – питання технічне. Насправді найбільш проблемним етапом в описаній вище схемі є вибір загального вигляду для апроксимаційної функції. Якщо модель базується на певній теорії, то часто це дозволяє отримати загальний вигляд для такої кривої. Для фізики побудова моделі на основі теорії є звичайною ситуацією. Для лінгвістики наявність теорії в підґрунті моделі є радше винятком, ніж правилом. Якщо вигляд апроксимаційної кривої апріорі встановити складно, доцільним може бути використання нейронних мереж.

7. Нейронні мережі

Спосіб моделювання на основі нейронних мереж стає дедалі популярнішим, особливо в секторі задач, пов'язаних із обробленням великих масивів даних[19]. Як і будь-який спосіб моделювання, він має свої переваги й недоліки. Головна ідея полягає в тому, що функціональна залежність, яку необхідно встановити за допомогою моделювання, інтерпретується як послідовність нелінійних перетворень. Базовими при цьому є поняття «нейрон» та «нейронна мережа».

Нейронна мережа – набір елементів, організованих у певну структуру і призначених для виконання нелінійних перетворень. Кожен елемент (нейрон) мережі отримує кілька значень «на вході», і генерує «сигнал» на виході. На рис. 8 наведено один із багатьох можливих способів організації нейронної мережі.

[19] Дирк-Эмма Бэстенс, Виллем-Макс ван ден Берг, Дуглас Вуд, *Нейронные сети и финансовые рынки: принятие решений в торговых операциях* (Москва: ТВП, 1997).

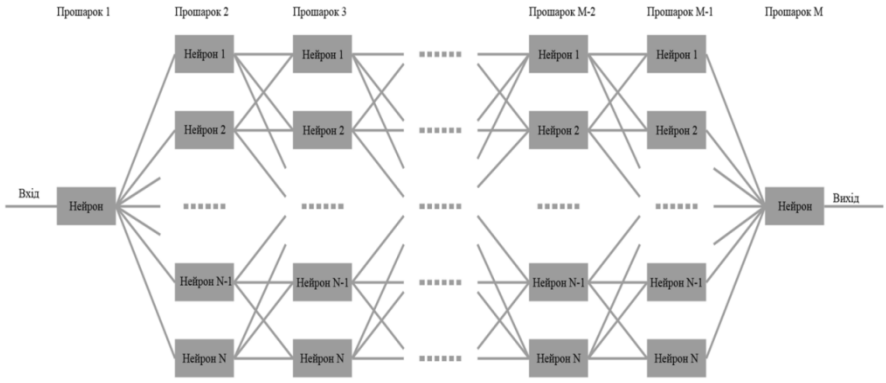

Рис. 8. *Схема організації нейронної мережі*

У цьому випадку нейронна мережа складається з кількох прошарків, і кожен прошарок містить певну кількість нейронів. Перший прошарок складається з одного нейрона. На єдиний нейрон у першому прошарку подається «сигнал» (значення арґументу у функціональній залежності, яка моделюється за допомогою нейронної мережі). Після перетворення на першому нейроні «сигнал» подається на кожен нейрон другого прошарку. Після перетворення «сигнал» з кожного нейрона другого прошарку передається на кожен із нейронів третього прошарку і так далі. «Сигнал» із нейронів передостаннього прошарку передається на єдиний нейрон останнього прошарку. «Сигнал», який отримують на виході нейрона з останнього прошарку, є результатом (значення функції). Важливим етапом у цій послідовності перетворень є спосіб, в який нейрон перетворює сигнали, що отримує на вході, на сигнал на виході. Перетворення (у спрощеному вигляді) виконується так: спочатку всі отримані на вході сигнали додаються з різними ваговими коефіцієнтами, а на основі отриманого числового значення за допомогою нелінійного перетворення отримують сигнал на виході нейрона. Функція, за допомогою якої виконується зазначене нелінійне перетворення, називається функцією активації нейрона. Це «індивідуальна» характеристика нейрона, і вибирати функцію активації нейрона можна по-різному. Популярними є так звані сигмоподібні функції активації[20]. Це функціональні залежності такого виду:

$$f(u) = \frac{1}{1 + \exp(-u)}.$$

Або такого:

$$f(u) = \frac{\exp(u) - 1}{\exp(u) + 1}.$$

[20] Там само.

Назва функцій пов'язана з тим, що графіки цих функціональних залежностей мають сигмоподібний вигляд, як це показано на рис. 9.

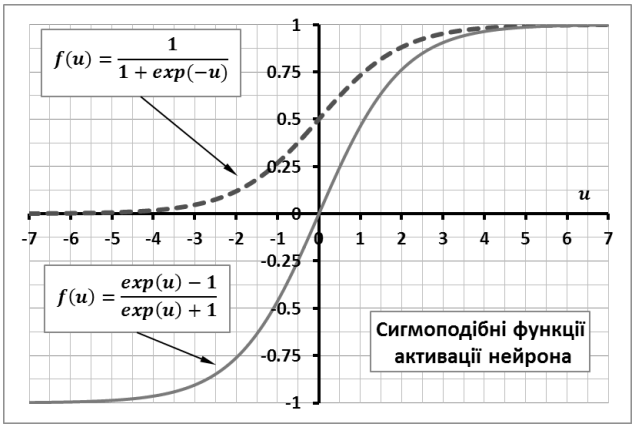

Рис. 9. *Функції активації нейрона*

Специфіка нейронних мереж передусім пов'язана з тим, що вони фактично дозволяють виконувати функціональну апроксимацію[21]. Зокрема, відомо, що за допомогою нейронної мережі з трьома внутрішніми прошарками нейронів із сигмоподібною функцією активації можна змоделювати будь-яку неперервну за арґументом функціональну залежність. Отже, навіть не маючи жодного уявлення про характер функціональної залежности, що пов'язує статистичні дані, ми можемо встановлювати між ними зв'язок. Правда, робота з нейронними мережами має свої складності й недоліки. Скажімо, немає універсальних алгоритмів та критеріїв для визначення оптимальної структури мережі. Та й сама процедура визначення параметрів мережі є складною та дещо непередбачуваною. Водночас нейронні мережі в багатьох випадках дозволяють отримувати цілком прийнятні результати. Як невелику ілюстрацію застосування нейронних мереж розглянемо вже відомі нам задачі про розподіл давніх слів і ранґовий розподіл слів у тексті казки про Вінні-Пуха.

Для моделювання розподілу давніх слів використаємо надзвичайно просту нейронну мережу, яка містить лише чотири нейрони. Схему нейронної мережі представлено на рис. 10.

За допомогою оптимізації параметрів нейронної мережі будуємо модель, що описує розподіл давніх слів у німецькій мові. Результати моделювання на основі нейронної мережі представлено на рис. 11.

[21] Там само.

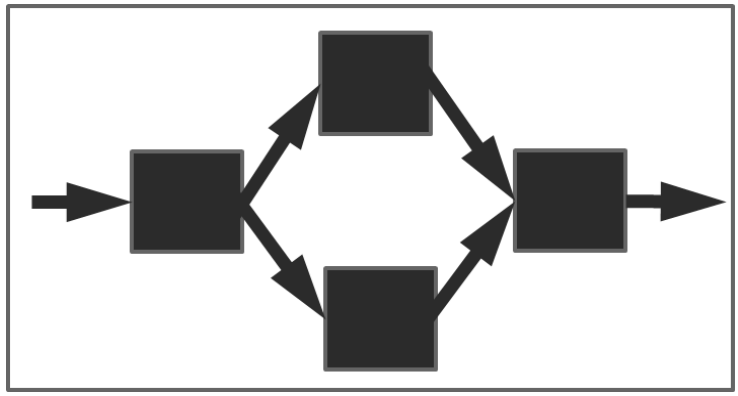

Рис. 10. *Нейронна мережа для моделювання розподілу давніх слів*

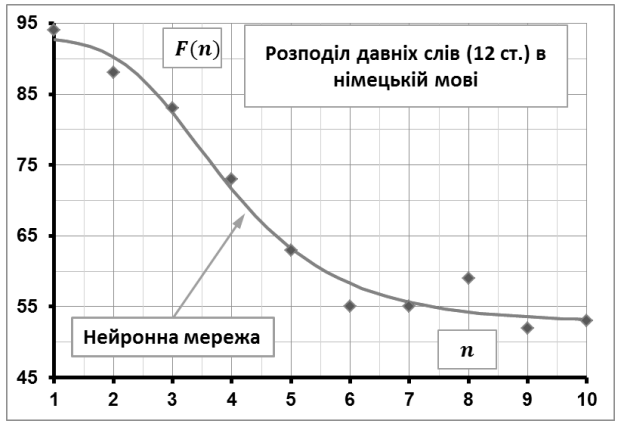

Рис. 11. *Результати моделювання на основі нейронної мережі розподілу давніх слів у німецькій мові*

Результати моделювання на основі нейронної мережі розподілу давніх слів в естонській мові представлено на рис. 12.

Для моделювання ранґового розподілу слів у тексті казки про Вінні-Пуха використовуємо дещо складнішу нейронну мережу, яка складається з восьми нейронів, що формують п'ять прошарків. Структуру нейронної мережі представлено на рис. 13.

Оптимізуючи параметри нейронної мережі, отримуємо залежність між ранґом слова в тексті та частотою його використання. Цю залежність представлено у графічному вигляді на рис. 14.

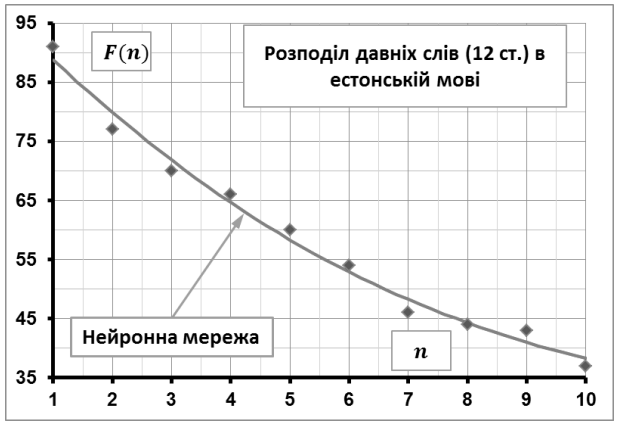

Рис. 12. *Результати моделювання на основі нейронної мережі розподілу давніх слів в естонській мові*

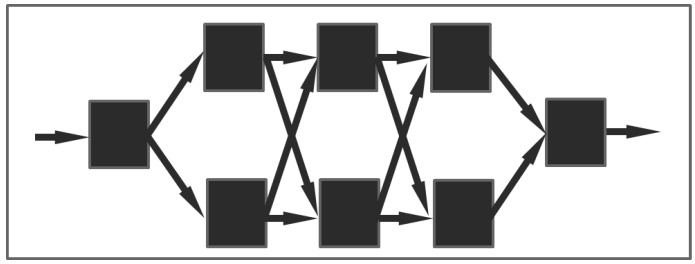

Рис. 13. *Нейронна мережа для моделювання ранґового розподілу слів у тексті казки про Вінні-Пуха*

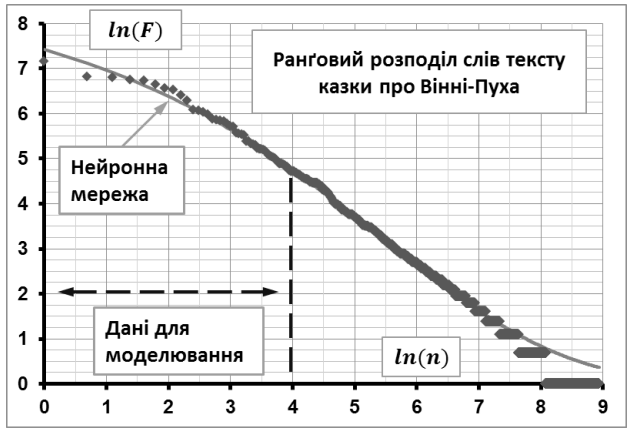

Рис. 14. *Результат моделювання на основі нейронної мережі ранґового розподілу слів у тексті казки про Вінні-Пуха*

Як бачимо, тепер теоретична крива відрізняється від прямої лінії. Ще один цікавий момент пов'язаний із тим, що процес визначення параметрів нейронної мережі (так зване навчання нейронної мережі) відбувався з використанням статистичних даних тільки для групи високочастотних словоформ. Тому перша частина кривої на рис. 14 (позначена як «дані для моделювання») – це апроксимація статистичних даних на основі нейронної мережі. А друга частина кривої — це фактично прогнозовані значення для частотного розподілу словоформ. Як бачимо, прогноз досить непогано збігається з фактичними даними. Однак найбільш важливим є той факт, що за умови використання нейронної мережі немає потреби заздалегідь визначати загальний вираз для функціональної залежности, на основі якої будується модель.

7. Замість епілогу

Сучасна наука розвивається швидко, бурхливо та різнопланово. Однією з характерних ознак цього розвитку є міждисциплінарний характер досліджень. Сучасні дослідницькі групи часто формуються з науковців, які працюють у різних наукових галузях. На рис. 15 наведено графік із даними щодо кількости спільних публікацій, що їх виконали фізики й економісти.

Рис. 15. *Кількість спільних публікацій, що їх виконали фізики й економісти*

Бачимо, що останнім часом відбувається стрімке зростання кількости таких публікацій. Є надія, що колись така ж тенденція матиме місце для публікацій, що їх виконали фізики й лінгвісти.

Oleksii VASYLIEV, Oleksandr CHALYI, Ilona VASYLIEVA

Mathematical Methods and Models in Linguistics

Oleksii VASYLIEV *– Doctor of Physical and Mathematical Sciences, Professor of Theoretical Physics Department, Physics Faculty, Taras Shevchenko National University of Kyiv. Research and professional interests: programming and mathematical modeling, synergetics, biophysics, physics of liquids, mathematical economics, mathematical linguistics. E-mail address: alex@vasilev.kiev.ua*

Oleksandr CHALYI *– Doctor of Physical and Mathematical Sciences, Professor, Corresponding Member of the National Academy of Pedagogical Sciences of Ukraine, Head of the Department of Medical and Biological Physics and Informatics, Bogomolets National Medical University. Research and professional interests: medical physics, medical informatics, critical phenomena and phase transitions, physics of liquids, synergetics, biophysics, mathematical linguistics. E-mail address: avchal@univ.kiev.ua*

Ilona VASYLIEVA *– a linguist. Graduated from the Institute of Philology of Taras Shevchenko National University of Kyiv (specialization in computer linguistics). Research and professional interests: modern Ukrainian language, automated linguistic systems and complexes, application of mathematical methods in linguistics. E-mail address: ilona@vasilev.kiev.ua*

In this article, relevant approaches to creating and using mathematical models in linguistics are examined. Analyzed are the types of mathematical methods used to solve linguistic tasks. Also investigated are the fundamental principles for creating mathematical models. Stressed in particular is that the approach usually applied in solving physical tasks can be effective. Its typical feature is that a universal theory explaining the interaction among a system's functional elements on the most general level is the basis of the model. In this case, the necessary approximating dependencies are not conjectured, as is often the case, but obtained or calculated on the basis of the fundamental theory. The prospects for applying this approach in linguistics are defined.

Examples of the modeling of linguistic systems and a comparative analysis of their various methods are provided. Much attention is paid to the universality of the mathematical models used to analyze linguistic material. It is noted that if the approximate dependencies used are based on a system's general universal properties, this can significantly complicate the quantitative analysis of the data. The advantages and shortcomings of mathematical modeling in solving applied linguistic tasks are illustrated.

Besides "classic" models (such as Zipf's law for the rank distribution of words in a text), other modeling approaches are described. In particular, attention is paid to the methods of constructing mathematical models based on neural networks. In this case, the approximation of dependence is realized in the form

of sequential essentially nonlinear transformations. The main disadvantage of this approach is connected with the technical complications of execution and with the absence of overt analytical dependence. The main advantage is hidden in the potential possibility of creating approximate dependencies of practically any type. As a comparison of various modeling approaches, the article provides examples of solving the same linguistic tasks using several methods.
Keywords: model, theory, rank distribution, approximation, neural network.

Bibliography

Arnol'd, Vladimir. "*Zhestkie" i "miagkie" matematicheskie modeli*. Moskva: MTSNMO, 2013.

Bestens, Dirk-Emma, Villem-Maks van den Berg, Duglas Vud. *Neironnye seti i finansovye rynki: priniatie reshenii v torgovykh operatsiiakh*. Moskva: TVP, 1997.

Kostenko, Lina. *Zapysky ukrains'koho samashedshoho*. L'viv: A-BA-BA-HA-LA-MA-HA, 2014.

Li, Wentian. "Zipf's law everywhere". *Glottometrics* 5 (2002): 14–21.

Pal'chykov, Vasyl'. "Efekty bezmasshtabnosti ta tisnoho svitu v skladnykh merezhakh". Kand. dys., IFKS NAN Ukrainy, L'viv, 2010.

Popescu, Ioan-Iovitz, Gabriel Altmann, and Reinhard Köhler. "Zipf's law—another view". *Quality and Quantity* 44, no. 4 (2010): 713–731.

Tuldava, Iukhan. *Problemy i metody kvantitativno-sistemnogo issledovaniia leksiki*. Tallin: Valgus, 1987.

Vasyl'iev, Oleksii, Oleksandr Chalyi, Ilona Vasyl'ieva. "Pro ekzotychni zadachi fizyky, Vinni-Pukha ta zakon Zipfa". *Zhurnal fizychnykh doslidzhen'* 17, №1 (2013): 1001(1–8).

Zipf, George. *Human Behavior and the Principle of Least Effort*. Cambridge: Addison-Wesley, 1949.