Андрій РОВЕНЧАК, Соломія БУК

Квантові розподіли і дослідження текстів: температура та література[1]

Андрій РОВЕНЧАК *– доктор фізико-математичних наук, професор катедри теоретичної фізики Львівського національного університету імені Івана Франка. Наукові інтереси: статистична фізика; конденсація Бозе–Айнштайна; системи з дробовою статистикою; кількісні методи в суспільних і гуманітарних науках; вивчення систем письма; історія науки. Електронна адреса: andrij.rovenchak@gmail.com*

Соломія БУК *– кандидатка філологічних наук, доцентка катедри загального мовознавства Львівського національного університету імені Івана Франка. Наукові інтереси: статистична лінгвістика, лексикографія, українська для іноземців, комп'ютерна та корпусна лінгвістика, лінгвістична семантика та прагматика, лінгвістична антропологія, психологія щастя. Електронна адреса: solomija@gmail.com*

Ранґово-частотні розподіли слів у текстах мають низку спільних рис із розподілами частинок за енерґіями, які відомі у статистичній фізиці. Це дає можливість на підставі аналогії з фізичними системами запропонувати новий набір параметрів, за допомогою якого можна здійснювати атрибуцію текстів, що є прикладами складних систем. Зокрема, вийшло показати зв'язок цих параметрів із типологічною класифікацією мов за рівнем аналітичности та проілюструвати еволюцію в межах кількох мовних «родоводів».

Запропоновані параметри розраховано за частотними даними слів, які рідко трапляються в текстах. Виявляється, що цю частину ранґово-частотного розподілу характеризує стабільніша поведінка, на відміну від високочастотної лексики, яку використовували деякі інші автори в подібних дослідженнях.

Один із параметрів, використаних у класифікації, є аналогом температури у фізиці. Його менші значення відповідають мовам із вищим рівнем аналітичности (менш розвиненою словозміною, яку фактично заміняє

[1] Автори висловлюють подяку Юрієві Головачу за обговорення результатів та корисні зауваження під час роботи над текстом статті.

більша кількість допоміжних слів та фіксованість порядку слів у реченнях). Частка рідковживаної лексики в таких мовах є іншою порівняно з мовами, в яких добре розвинена словозміна.

Наш підхід продемонстровано на прикладі перекладів новели-казки Антуана де Сент-Екзюпері «Маленький принц» та Євангелія від Івана. Перший твір, який належить до текстів секулярного (світського) характеру, перекладено понад 200 мовами, із яких ми аналізуємо близько 40. Євангеліє взято для вивчення розвитку мови в історичному розрізі, оскільки саме релігійні тексти можна знайти в перекладах, віддалених у часі на кілька століть.

Одержані результати показують нові виміри раніше відомих понять. Їх розглянуто в ширшому контексті лігвостатистичних та лінгвофілософських досягнень Вільгельма фон Гумбольдта, Морріса Сводеша, Джозефа Ґрінберґа, Ґабріеля Альтмана, Райнгарда Кьолера.

Ключові слова: закон Ціпфа, ранґово-частотний розподіл, атрибуція текстів, «температура» тексту, еволюція мов.

> Ich behaupte aber, daß in jeder beſonderen Naturlehre nur ſo viel eigentliche Wiſſenſchaft angetroffen werden könne, als darin Mathematik anzutreffen iſt.
>
> Immanuel Kant,
> Metaphyſiſche Anfangsgründe der Naturwiſſenſchaft (1786)[2]

1. Вступ

Винесена в епіграф цитата Іммануїла Канта інколи трапляється в дискусіях про співвідношення методів гуманітарних і природничих наук. Її трактування загалом потребує докладного вивчення контексту, як пише британський біолог, логік і філософ Джозеф Генрі Вуджер[3], і до сучасних наукових теорій навряд чи застосовне беззастережно. У цій статті ми спробуємо поєднати досить далекі, на перший погляд, галузі науки – мовознавство й фізику, продемонструвавши в такий спосіб єдність наукової картини світу.

Насамперед метамовою фізики (майже виключно) та лінгвістики (все ще спорадично, але вже більш та більш упевнено) є математика: фізика послуговується, і мовознавство може послуговуватися математико-статистичним апаратом для опису певних явищ, виявлення їхньої закономірности тощо.

[2] «Я стверджую, однак, що в кожному конкретному вченні про природу можна знайти лише стільки власне науки, скільки в ньому є математики». Цит. за: Immanuel Kant, *Metaphysische Anfangsgründe der Naturwissenschaft. Zweyte Auflage* (Riga: bey Johann Friedrich Hartknoch, 1787), S. VIII.

[3] J. H. Woodger, *Biological Principles: A Critical Study* (Abingdon–New York: Routledge, 2014), p. 234.

І якщо для природничих дисциплін це твердження очевидне, то для гуманітарних може вимагати пояснення.

Певний час точилася дискусія про сприйняття лінгвістичної статистики як спроби дегуманізації мовознавства та літературознавства, «знедушення» досліджень художнього твору й тексту. Проте, за словами Ґабріеля Альтмана, «кожна достатньо розвинута наукова дисципліна рано чи пізно, принаймні на певному етапі свого розвитку, може опинитися на порозі математизації. Фізику не можна уявити без математики, в гуманітарних науках намагаються загальмувати цей прогрес непереконливими обґрунтуваннями. Усі ці обґрунтування зберігаються ж більшою чи меншою мірою лише з ідеологічних міркувань. Про дегуманізацію науки математикою говорять зазвичай лише ті, хто в ній нічого не розуміє і хто не готовий прийняти кращі методи»[4].

Застосування кількісних і статистичних методів у дослідженні мови та мовлення має тисячолітню історію: наприклад, іще в античному світі (III ст. до н. е). александрійські граматики підраховували слова в різних творах Гомера. Метою такого підрахунку було визначення тих слів, які трапляються лише один раз у творі. У середньовіччі, відомому особливою увагою до сакральних текстів, ерудити укладали частотні списки слів Святого Письма, у XIX ст. стенографія запропонувала скорочення для найчастотніших елементів тексту, далі під час воєн у XX ст. посилилася увага до криптографії – науки про шифрування й дешифрування повідомлень.

Зараз складно уявити будь-яку галузь мовознавства без кількісних і статистичних методів: фонетику (фонетичні закони), морфологію (продуктивність морфем), лексикологію та семантику (семантична відстань між словами[5]), стилістику[6] й лінгводидактику (для ефективного вивчення мови треба спочатку вивчати найчастотніші її слова, оскільки саме вони дають основне розуміння будь-якого тексту), генеалогічну та типологічну класифікацію мов (про це див. детальніше далі у статті).

Погляди на мову як на простий набір звуків, морфем та слів еволюціонували через структуралізм до розуміння її як семіотичної та синергетичної системи. «Оскільки мова – це ймовірнісна, а не жорстко детермінована система, то для її пізнання квантитативні методи, пов'язані з дослідженням частотних, ймовірнісних, градуальних та інших нелогічних характеристик, не тільки бажані, але й необхідні»[7].

---

[4] Ґабріель Альтман, «Мода та істина в лінгвістиці: Особисте звернення до багатьох», у кн. *Проблеми квантитативної лінгвістики* (Чернівці: Рута, 2005), 5.

[5] Михайло Емільович Білинський, *Синоніміка англійського дієслова: Словник семантичних відстаней* (Львів: ЛДУ імені Івана Франка, 1999).

[6] Валентина Сидорівна Перебийніс (ред)., *Статистичні параметри стилів* (Київ: Наукова думка, 1967).

[7] Михайло Петрович Кочерган, *Загальне мовознавство* (Київ: Академія, 2006), 191.

2. Квантові розподіли

Сатьєндранат Бозе з Дакки (тоді це була Індія) 1924 року надіслав до німецького "Zeitschrift für Physik", одного з тодішніх провідних фізичних видань, статтю про виведення закону Планка для випромінювання. Переклад цього матеріялу зробив Альберт Айнштайн, додавши примітку, що праця Бозе є значним досягненням, а відповідну ідею буде застосовано для вивчення квантових ідеальних газів[8]. Згодом, протягом 1924–25 рр., Айнштайн опублікував дві статті[9] в журналі Прусської академії наук, "Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften zu Berlin", які й стали основою того, що ми називаємо статистикою Бозе–Айнштайна, або статистикою Бозе. Вона описує частинки, що характеризуються певною «колективною» поведінкою, оскільки в одному квантовому стані їх може бути як завгодно багато, на відміну від «індивідуалістів» – частинок, які описує так звана статистика Фермі–Дірака (або просто статистика Фермі), відома нам ще зі шкільної лави через принцип заборони Паулі, що дозволяє зрозуміти правило заповнення електронами атомних орбіталей; в одному квантовому стані таких частинок не може бути більше однієї.

У широкому розумінні аналогію між фізичними багаточастинковими системами та текстами можна вбачати в тому, що між реальними частинками існує фізична взаємодія, тоді як текст є не просто сукупністю слів – зв'язки між ними визначають «взаємодію» на різних рівнях: граматичному, семантичному, синтаксичному тощо. Саме ця «взаємодія» опосередковано проявляється, зокрема, через сполучуваність слів та частоту їхнього вживання.

Ідея застосувати квантові розподіли[10] в лінгвістичному аналізі виникла через математичну подібність між розподілом Бозе–Айнштайна та ранґово-частотними характеристиками текстів. Для того, щоб це продемонструвати, коротко опишемо спосіб укладання ранґово-частотних списків, які далі становитимуть основу описаних досліджень.

Аналіз текстів ми проводили на рівні так званих ортографічних слів – буквено-цифрових послідовностей між двома пропусками чи розділовими знаками. Це означає, що різні форми, наприклад, українські 'рука' і 'рукою',

---

8 Bose, "Plancks Gesetz und Lichtquantenhypothese", *Zeitschrift für Physik* 26, 1 (1924): 178–181.

9 Albert Einstein, "Quantentheorie des einatomigen idealen Gases", *Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften zu Berlin: physikalisch-mathematischen Klasse* (1924): 261–267; Idem, "Quantentheorie des einatomigen idealen Gases. Zweite Abhandlung", *Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften zu Berlin: physikalisch-mathematischen Klasse* (1925): 3–14.

10 Цікаво, що неможливість описати явища суспільних наук детерміністичними методами спонукала у 1930-х рр. Етторе Майорану на філософські розважання про аналогію між квантовою механікою та суспільними науками: Rosario Nunzio Mantegna, "Presentation of the English translation of Ettore Majorana's paper: The value of statistical laws in physics and social sciences", *Quantitative Finance* 5, 2 (2005): 133–140. Оригінальна стаття: Ettore Majorana, "Il valore delle leggi statistiche nella fisica e nelle scienze sociali", *Scientia* 36 (1942): 58–66.

'говорити' і 'говорила', англійські 'go', 'goes' і 'went', французькі 'petit', 'petite' і 'petites' тощо вважали різними словами. Таке рішення пов'язане з відсутністю єдиного загальноприйнятого визначення слова навіть у межах однієї мови, вже не кажучи про таке, яке би можна було застосувати до різних мов[11].

На підставі заданого тексту можна укласти список слів за спаданням їхніх абсолютних частот, тобто кількости вживань кожного слова в цьому тексті. Перше, найчастотніше, слово отримує ранґ 1, наступне за частотою – ранґ 2, третє за частотою – ранґ 3 і так далі. Якщо слова мають однакові частоти, то їм присвоюють послідовні ранґи в довільному порядку – цікаво, що в цьому підході вже можна помітити аналогію з так званим квантовомеханічним принципом нерозрізнювальности частинок. Є також трохи інший підхід, коли словам із однаковими частотами присвоюють середнє значення з відповідного діяпазону ранґів, однак такий спосіб трохи ускладнить інтерпретацію, з якої виникає фізична аналогія. Типовий ранґово-частотний розподіл зображено на рис. 1. Наведена залежність між ранґом і частотою, відома як закон Ціпфа (інколи його називають першим законом Ціпфа).

У правій нижній частині рис. 1, тобто за високих значень ранґів, які відповідають низьким абсолютним частотам (1, 2, 3, …), легко помітити горизонтальні плато. Вони свідчать про те, що в текстах є багато слів, що вживаються дуже мало, 1–2 рази. Найдовше плато – а отже, і найбільша кількість слів – мають частоту 1, тобто в конкретному тексті є унікальними. Такі слова називають *гапакс леґомена* (множина від д.-грецьк. *ἅπαξ λεγόμενον* '[щось] сказане [лише] один раз'). Цей термін походить із вивчення Біблії, а найвідомішими прикладами є לִילִית 'Ліліт' (слово незрозумілого значення, відповідає персонажеві єврейської мітології) та גֹּפֶר '[дерево] ґофер' (із якого було збудовано Ноїв ковчег)[12]. Як ми побачимо згодом, кількість гапаксів N_{hapax} буде одним із параметрів, за допомогою яких відбуватиметься кількісний опис.

На підставі ранґово-частотних списків будують так званий частотний спектр[13]. Для цього підраховують кількість слів N_j, які мають абсолютну частоту j (отже, кількість гапаксів $N_{\text{hapax}} = N_1$). Цю залежність, яку інколи називають другим законом Ціпфа, і будемо намагатися моделювати за допомогою аналогії з квантовим розподілом Бозе[14], як описано в наступному розділі.

---

[11] Ioan-Iovitz Popescu, Gabriel Altmann, Peter Grzybek, Bijapur Dayaloo Jayaram, Reinhard Köhler, Viktor Krupa, Ján Macutek, Ján Mačutek, Regina Pustet, Ludmila Uhlířová, and Matummal N. Vidya, *Word frequency studies* (Berlin–New York: Mouton de Gruyter, 2009).

[12] E. G. Hirsch, I. M. Casanowicz, J. Jacobs, and M. Schloessinger, "Hapax legomena", in *The Jewish Encyclopedia, Vol. VI* (New York: Funk and Wagnalls, 1904), 226–229. Available online at: http://www.jewishencyclopedia.com/ articles/7236-hapax-legomena.

[13] Juhan Tuldava, "The frequency spectrum of text and vocabulary", *Journal of Quantitative Linguistics*, 3 (1996): 38–50.

[14] Andrij Rovenchak and Solomija Buk, "Application of a quantum ensemble model to linguistic analysis", *Physica A* 390 (2011): 1326–1331.

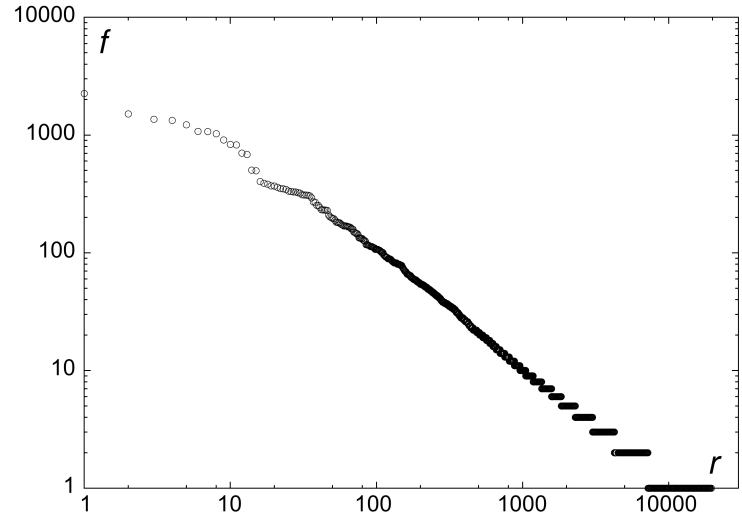

Рис. 1. *Типовий ранґово-частотний розподіл. Дані відповідають абсолютним частотам f залежно від ранґу r для ортографічних слів у романі Івана Франка «Перехресні стежки». Автори одержали результати, працюючи над укладанням частотного словника цього твору*[15].

2. Фізична аналогія

Практика дослідження текстів показує, що кількість слів із частотою 1 приблизно вдвічі більша за кількість слів із частотою 2 і далі поступово зменшується зі зростанням абсолютної частоти. Якісно така сама поведінка типова для фізичних систем: частинки воліють займати рівні з якомога меншою енерґією. Причому тут стає важливою саме статистика Бозе, оскільки лише вона дає змогу зосередитися на одному рівні багатьом частинкам, на відміну від статистики Фермі.

Математичний опис передбачає досягнення не лише якісного, але й задовільного кількісного узгодження спостережуваних значень із розрахованими. Тому для моделювання частотного спектра ми скористаємося розподілом Бозе у вигляді:

$$N_j = \frac{1}{z^{-1}\exp\left(\dfrac{(j-1)^{\alpha}}{T}\right)-1},$$

[15] Соломія Бук та Андрій Ровенчак, «Частотний словник роману «Перехресні стежки»», у кн. *Стежками Франкового тексту (комунікативні, стилістичні та лексикографічні виміри роману «Перехресні стежки»)*, Ф. С. Бацевич (наук. ред), С. Н. Бук, Л. М. Процак, А. А. Ровенчак, Л. Ю. Сваричевська, І. Л. Ціхоцький (Львів: Видавничий центр ЛНУ імені Івана Франка, 2007), 138–369.

У фізиці ця формула описує розподіл частинок на енергетичних рівнях. У цій праці *j* означатиме абсолютну частоту (а у фізичній моделі – це номер рівня), величину *z* розраховуватимемо за кількістю гапаксів N_1:

$$N_{\text{hapax}} \equiv N_1 = \frac{1}{z^{-1} - 1},$$

а параметри α і T знаходитимемо, зіставляючи розраховані N_j з отриманими для кожного з досліджуваних текстів. Приклад результатів застосування такої процедури наведено на рис. 2.

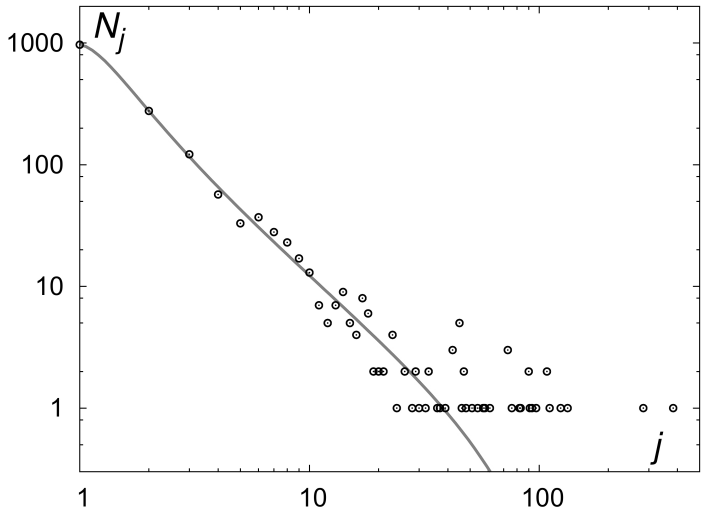

Рис. 2. *Частотний спектр перших дев'яти розділів Євангелія від Івана давньогрецькою мовою.*

Варто звернути увагу на те, що запропонована модель погано «ловить» великі значення *j*. Це можна поліпшити, використовуючи певні модифікації розподілу Бозе[16], що, однак, значно не вплине на загальні висновки.

У відповідних фізичних задачах величина T має зміст температури. Проте, пам'ятаючи застереження Ґабріеля Альтмана та Петера Маєра[17], ми не будемо намагатися надати якогось подібного значення цьому параметру в текстах попри спокуси, пов'язані з ототожненням фізичних висновків і результатів розрахунків у їхніх аналогах з інших галузей науки.

[16] Andrij Rovenchak and Solomija Buk, "Part-of-speech sequences in literary text: Evidence from Ukrainian", *Journal of Quantitative Linguistics*, 25, 1, (2018): 1–21.

[17] Gabriel Altmann and Peter Meyer, "Physicist's look at language", *Problems of Quantitative Linguistics* (Černivci: Ruta, 2005), 42–59.

Зауважимо тут, що ідея застосування методів статистичної фізики в дослідженні так званих складних систем, тобто таких, чиї властивості як цілого не визначаються сумою властивостей окремих складників, має вже досить тривалу історію[18]. Відповідні методи застосовують і в лінгвістиці, а тому не дивно, що різний зміст поняття «температури тексту» можна знайти в різних авторів. Близьким до підходу, що ми запропонували, є підхід Сасуке Міяджіми і Кейдзо Ямамото[19], які використовували класичний розподіл Больцмана для моделювання високочастотних слів, проводячи калібрування «температури» за допомогою певного еталонного тексту. Зазвичай саме на високочастотній лексиці зосереджують основну увагу в подібних дослідженнях. Ми ж розраховуємо параметри на підставі даних про низькочастотні слова, поведінка яких відзначається набагато більшою стабільністю, а тому наш підхід повинен бути релевантним у порівняльних і контрастивних мовознавчих дослідженнях.

Ранґово-частотні розподіли різних одиниць відомі в документознавстві, в соціології, географії, математиці і т. д. Відповідні зв'язки справджуються на мовному та багатьох інших матеріялах, як-от закон Ціпфа–Мандельброта, що уточнює залежність частоти слова та його ранґу у словнику (тобто закон Ціпфа) для слів із високими частотами, виявляється справедливим для музичних текстів (одиницю, що відповідає слову, називають F-мотивом) і для розподілу площ, які займають різні кольори на картинах[20].

3. Деякі результати

Описаний спосіб аналізу ми застосували до низки текстів різними мовами. Вже під час першого дослідження[21] було помічено, що параметри α і T, які було застосовано в попередньому розділі, у певний спосіб пов'язані з рівнем аналітичности мови, а саме: менші значення відповідали аналітичним мовам, а більші – синтетичним. Тут доречно пригадати, що в синтетичних мовах основним способом утворення граматичних форм є словозміна (як, наприклад, у слов'янських), тоді як в аналітичних мовах переважно використовують допоміжні слова (зокрема в англійській). Наше спостереження підтвердилося й надалі. Показовим прикладом можна вважати новелу-казку Антуана де Сент-Екзюпері «Маленький принц». Такий вибір пов'язаний із тим, що цей твір є одним із найбільше перекладених нерелігійних текстів: понад 240 різними мовами із різних родин, зокрема і штучними – есперанто та ложбан.

---

[18] Yurij Holovatch, Ralph Kenna, and Stefan Thurner, "Complex systems: physics beyond physics", *European Journal of Physics* 38, 2 (2017): 023002 [19 p.]

[19] Sasuke Miyazima and Keizo Yamamoto, "Measuring the temperature of texts", *Fractals*, 16 (2008): 25–32.

[20] Reinhard Köhler, *Zur linguistischen Synergetik. Struktur und Dynamik der Lexik* (Bochum: Brockmeyer, 1985).

[21] Andrij Rovenchak and Solomija Buk, "Application of a quantum ensemble model to linguistic analysis", *Physica A* 390 (2011): 1326–1331.

Результати наших обчислень продемонстровано на рис. 3 для сорока двох перекладів «Маленького принца»[22]. Звернемо увагу, що замість температури T ми використали логарифмічно масштабований параметр $\tau = \ln T / \ln N$, який враховує залежність T від обсягу тексту (загальної кількости слів) N.

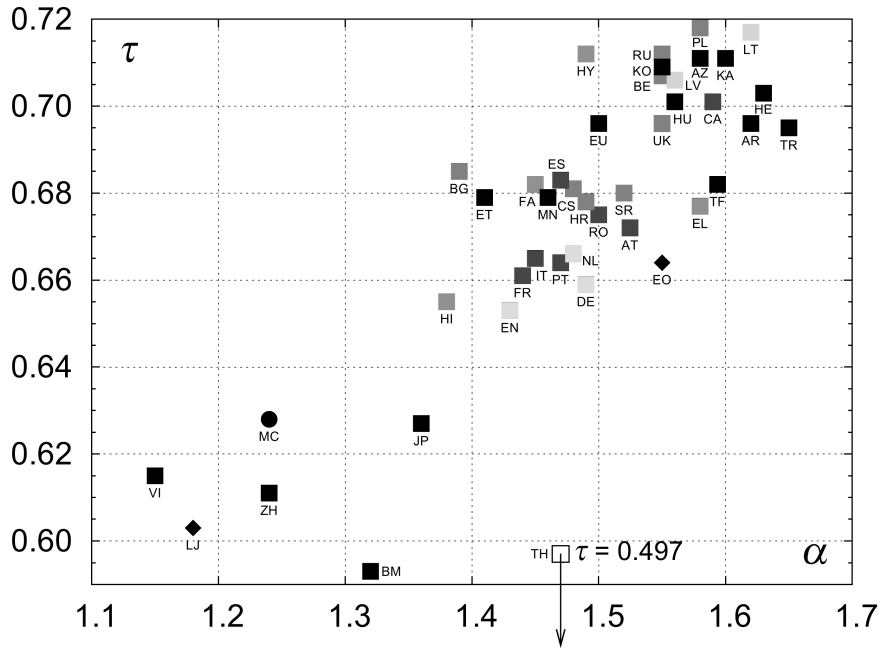

Рис. 3. *Положення перекладів «Маленького принца» різними мовами на площині (α; τ). Мови позначено кодами ISO (якщо двобуквений код не визначено, то використано найближче за звучанням позначення): AR – арабська, AT – астурійська, AZ – азербайджанська, BM – бамана, BE – білоруська, BG – болгарська, CA – каталонська, CS – чеська, DE – німецька, EN – англійська, ES – іспанська, EU – баскська (еускара), FA – фарсі, FR – французька, EL – грецька, EO – есперанто, ET – естонська, HE – іврит, HI – гінді, HR – хорватська, HU – угорська, HY – вірменська, IT – італійська, JP – японська, KA – грузинська, KO – корейська, LJ – ложбан, LV – латвійська, LT – литовська, MC – маврикійська креольська (морісьєн), MN – монгольська, PL – польська, PT – португальська, RO – румунська, RU – російська, SR – сербська, TF – амазіг (берберська, письмом тіфінаг), TH – тайська, TR – турецька, UK – українська, VI – в'єтнамська, ZH – китайська.*

[22] Більшість цих результатів взято зі статті: Andrij Rovenchak and Solomija Buk, "Defining thermodynamic parameters for texts from word rank-frequency distributions", *Journal of Physical Studies* 15, 1 (2011): 1005 [6 pp.].

У трьох мовах з-поміж тих, переклади якими ми досліджували, а саме в японській, китайській і тайській, немає звичного поділу на слова. Тут було застосовано три різні підходи: словоподіл у японському тексті забезпечено спеціяльними програмними засобами[23], в китайському тексті зроблено частотний аналіз окремих ієрогліфів замість слів, у тайському ж тексті пропусками відділяють речення або їхні частини, тому відповідні результати можуть стати корисними в майбутньому для зіставлення з полісинтетичними мовами.

На рис. 3 можна простежити групування мов на площині, яку визначають параметри (α; τ). До цих груп належать відповідно такі мови:

– бамана, в'єтнамська, китайська, ложбан, морісьєн і японська;
– англійська, гінді, італійська, німецька, португальська та французька;
– іспанська, монгольська, румунська, сербська, фарсі, хорватська, чеська;
– білоруська, польська, російська, українська, латвійська, литовська, каталонська, арабська, азербайджанська, грузинська, іврит, угорська, корейська й турецька (у межах цієї більшої групи є також власний поділ).

У першу розпорошену групу потрапили мови з високим рівнем аналітичности; у групах від другої до четвертої, які значно компактніші, рівень аналітичности зменшується (відповідно в таких мовах стає більшою частка синтетичних конструкцій).

Серед мов, які не потрапляють у ці компактні групи, опинилися баскська, естонська, вірменська, грецька, есперанто і болгарська. Також серед слов'янських мов спостерігаємо поділ на дві групи: у першій – хорватська, сербська й чеська, у другій – білоруська, польська, російська й українська. Знову ж таки зліва направо і знизу догори спадає рівень аналітичности, що найяскравіше відображено в позиції болгарської мови, у якій словозміна істотно збіднена.

Твердження про зв'язок ступеня аналітичности мови зі значеннями параметрів α й τ підтверджує також розміщення штучних мов, які позначено ромбиками, – есперанто й ложбан. Істотна відмінність між ними пов'язана з різними підходами до створення цих мов: ложбан[24] – близька до машинної високоаналітична мова, що ґрунтується на предикативній логіці, а есперанто – переважно аґлютинативна мова на кшталт природних.

Відзначимо, що близьке положення мов на рис. 3 зовсім не означає близької генетичної споріднености, а просто засвідчує подібність частотної структури текстів на рівні слів, яка випливає з характеру побудови граматичних конструкцій, типових для аналітичних чи синтетичних мов.

Згодом такі висновки про зв'язок пари параметрів α й τ зі ступенем аналітичности мови підтвердилися на інших текстах. Крім того, було показано, що перекладач не впливає істотно на ці характеристики[25].

[23] За порадою Гаруко Санади (Haruko Sanada) ми використали програмні засоби UniDic, MeCab та ChaSen.

[24] *Lojban*, available at: https://mw.lojban.org/papri/Lojban

[25] Andrij Rovenchak, "Where Alice meets Little Prince: Another approach to study language relationships", in *Sequences in Language and Text*, edited by George K. Mikros and Ján Mačutek (Berlin–Boston: Mouton de Gruyter, 2015), 217–230.

За допомогою такого самого «температурного» підходу ми проаналізували українські тексти великої прози Івана Франка[26], підтвердивши деякі виявлені раніше залежності між параметрами. Також на підставі фізичної аналогії щодо рівноваги між підсистемами ми запропонували нову величину для аналізу зв'язку між різними підсистемами в текстах на прикладі прямої й авторської мови. З цього погляду найбільше збалансованим (тобто з близькими значеннями введеної величини для прямої й авторської мови) виявився текст роману «Перехресні стежки», а найменше – друге видання роману «Петрії й Довбущуки». Надалі було би цікаво встановити, якою мірою невимірювані з першого погляду властивості тексту, як-от цілісність та зв'язність[27], емоційна забарвленість[28], напруженість[29] тощо, можна описати кількісними параметрами. Це поки що залишається відкритим питанням і вимагає ширшого аналізу за участи фахівців у галузі літературознавства.

Інші цікаві результати можна одержати, вивчаючи зміну параметрів α і τ в часі. Для цього необхідно знайти тексти, переклади яких робили однією мовою в різний час, бажано з відстанню в кілька століть. Тут не дуже великий вибір, а найбільш реалістичним видається дослідження релігійних книг. Після того, як вийшло знайти в оцифрованому вигляді перші дев'ять розділів Євангелія від Івана англосаксонською мовою[30], саме цей текст і став предметом для аналізу[31].

Початкова гіпотеза полягала в тому, що з плином часу в мові відбуваються спрощення, тобто вона набуває щораз більше аналітичних рис. Для перевірки такого твердження можна простежити зміну параметрів між такими мовами:

– латинська [lat] → романські (італійська [ita], французька [fra]) → іспанська [spa];

– старогрецька [grc] → новогрецька [ell];

– церковнослов'янська [chu] → сучасні слов'янські (російська [rus], українська [ukr]) → болгарська [bul].

---

[26] Solomija Buk and Andrij Rovenchak, "Probing the "temperature" approach on Ukrainian texts: Long-prose fiction by Ivan Franko", in *Studies in Quantitative linguistics 23: Issues in Quantitative Linguistics 4*, edited by E. Kelih, R. Knight, J. Mačutek, A. Wilson (Lüdenscheid: RAM-Verlag, 2016), 160–175.

[27] Анатолій Панасович Загнітко, *Лінгвістика тексту: Теорія і практикум* (Донецьк: ДонНУ, 2006).

[28] Anders Pettersson, *Verbal Art: A Philosophy of Literature and Literary Experience* (Montreal & Kingston–London–Ithaca: McGill-Queen's Press, 2001), chaps. 5–7.

[29] Владимир Григорьевич Адмони, «Синтагматическое напряжение в стихе и прозе», в кн. *Инвариантные синтаксические значения и структура предложения* (Москва: Наука, 1969), 16–26.

[30] https://web.archive.org/web/20150507015011/http://wordhord.org/nasb/john.html.

[31] Andrij Rovenchak, "Trends in language evolution found from the frequency structure of texts mapped against the Bose-distribution", *Journal of Quantitative Linguistics*, 21 (2014): 281–294.

На рис. 4 показано різні мови на площині (α; τ), а стрілками зображено еволюційну зміну параметрів. Усі напрямки червоних стрілок – від верхнього кута до нижнього, тобто відповідають зростанню ступеня аналітичности. Цікаво, що від англосаксонської мови [ang] до ділянки між середньоанглійською [eme] та сучасною англійською [eng] синя стрілка має трохи інший напрямок – це можна пояснити складними історичними процесами становлення англійської мови[32].

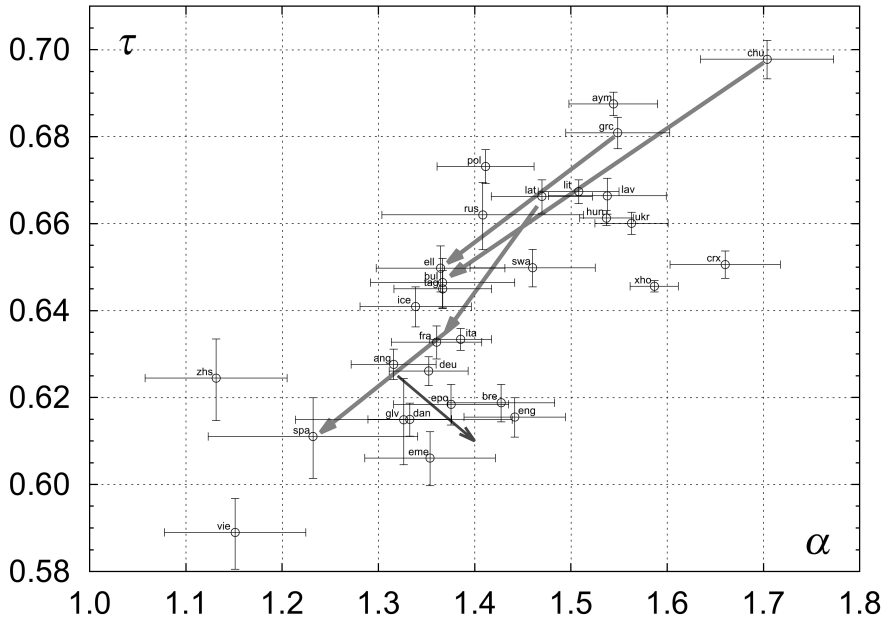

Рис. 4. *Еволюція мов відображена на площині (α; τ). Стрілки вказують напрямок еволюції. Мови позначено кодами ISO 639-2 (див. https://www.loc.gov/standards/iso639-2/php/code_list.php). Кольоровий онлайн.*

Доречно вказати тут на квантитативну типологію мов дослідника зі США Дж. Ґрінберґа, який запропонував 10 індексів, що дають об'єктивну характеристику будови мови: індекс синтетичности, префіксації, деривації, аґлютинації мов тощо[33]. Зокрема, індекс синтетичности він обчислює як відношення кількости морфів у певному тексті до кількости слів у ньому. Наші

[32] Albert C. Baugh and Thomas Cable, *The History of the English Language, 6th ed.* (London–New York: Routledge, 2013).

[33] Joseph Harold Greenberg, "A quantitative approach to the morphological typology of language", *International Journal of American Linguistics* 26 (1960): 178–194; Idem, *Language Typology: A Historical and Analytic Overview* (The Hague: Mouton, 1974).

результати суголосні такому підходу, демонструючи ще й інший – часовий – вимір явища аналітичности й синтетичности мов.

З хронологічного погляду варто згадати також факт, що в порівняльно-історичному мовознавстві вже відомі теорії, що пояснюють зміну мов із застосуванням математичних методів, як-от глотохронологія, яку застосовують для визначення ймовірного часу поділу споріднених мов. Обґрунтовуючи власну гіпотезу, американський лінгвоантрополог Морріс Сводеш робив аналогію з поняттям радіовуглецевого датування зміни віку органічних речовин. Фізичними та хімічними методами його можна визначити за кількістю відносного вмісту ізотопу ^{14}C в природних об'єктах (найчастіше – вугіллі, деревині, залишках кісток тощо), оскільки цей ізотоп має чітко передбачуваний час радіоактивного розпаду. До слова, такий метод широко застосовують в археології, за його винахід Віллард Ліббі здобув Нобелівську премію з хімії 1960 р. Так само для мовознавства Сводеш запропонував базовий словник (ядро лексики, спільне для всіх мов світу), швидкість зміни якого в усіх мовах залишається приблизно однаковою[34].

Також ще у XIX ст. визначний німецький лінгвофілософ та дипломат Вільгельм фон Гумбольдт спробував розмістити на осі часу чотири морфологічні типи мов (кореневі, аґлютинативні, інкорпоруючі, флективні) як відображення хронологічно послідовних етапів світового мовотворчого процесу, як перехід від нижчої до вищої, досконалішої форми і мови, і народу. В такий спосіб він прагнув виявити загальні закономірності історичного розвитку мов світу, вважаючи, що народ, який більше від інших обдарований природою і який перебуває у сприятливіших умовах, порівняно з іншими, повинен отримати й найдосконалішу мову[35]. І хоча зараз уважають, що за ступенем розвитку мов не варто робити висновки щодо ступеня інтелектуального розвитку народу, не можна заперечити глобальність мислення та узагальнення дослідника в галузі лінгвістичної типології. Наші результати демонструють зміну типологічної класифікації мови в часовому зрізі.

5. Висновки

Відомо, що в англо-американській науковій традиції є поділ наук на Sciences (природничі та математичні науки) та Humanities (гуманітарні: літературознавство, мовознавство, історія) та Social Sciences (економіка, політологія, соціологія), проте застосування кількісних методів для дослідження мовного матеріялу ставить мовознавство поряд із фізикою та іншими природничими науками. І це логічно, оскільки на зміну домінантного в середині XX ст. структуралізму (що розглядав мову як систему саму в собі і для самої

[34] Morris Swadesh, "Lexicostatistic dating of prehistoric ethnic contacts", *Proceedings of the American Philosophical Society* 96 (1952): 452–463; Idem, "Towards greater accuracy in lexicostatistic dating", *International Journal of American Linguistics* 21, 2 (1955): 121–137.

[35] Вильгельм фон Гумбольдт, «О различии строения человеческих языков и его влиянии на духовное развитие человеческого рода», в кн. В. А. Звегинцев, *История языкознания XIX и XX веков в очерках и извлечениях* (Москва: Просвещение, 1964), ч. 1, 85–105.

себе) прийшов погляд на мову як на самореґульовану систему, що взаємодіє з іншими філософськими прагматичними категоріями, як-от контекст, ситуація, культура, тощо.

Ми запропонували підхід до дослідження текстів, який ґрунтується на математичній аналогії між ранґово-частотними розподілами слів та заповненням енерґетичних рівнів у квантовому розподілі Бозе–Айнштайна. Завдяки аналізу низки творів, що їх написано різними мовами, розраховані параметри вийшло пов'язати зі ступенем аналітичности мови, а також одержати деякі інші характеристики тексту.

Той факт, що в явищах мови та мовлення можна виявити математичні закономірності, подібні до законів природничих наук, як такий є фундаментальний. Іншими словами, функціонування мови та мовлення підпорядковані певним законам, як у фізиці, хімії, біології. Їх вивчення допоможе пізнати глибинні закономірності світу.

Як було показано в широкому контексті попередніх лінгвостатистичних та лінгвофілософських досягнень фон Гумбольдта, Сводеша, Ґрінберґа, Альтмана, Кьолера та інших авторів, наші результати можна використати як у типологічній, так і в генеалогічній класифікації мов, як такі, що показують нові виміри раніше відомих понять. Кьолер[36] ще 1985 року підкреслював, що з'ясування законів побудови мови й тексту має бути центральною задачею мовознавства. Навіть більше, за словами Альтмана, «жодна інша лінгвістична дисципліна не мала такого впливу на інші науки як квантитативна лінгвістика. Закон Ціпфа є предметом щонайменше двадцяти інших дисциплін, які його аналізують і розвивають. Заінтриговані цією обставиною все більше фізиків, математиків і біологів підключаються до дослідження мови»[37].

Andrij ROVENCHAK and Solomija BUK

Quantum Distributions and Text Studies: Temperature and Literature

Andrij ROVENCHAK *– Doctor of Sciences in Physics and Mathematics, Professor of the Department for Theoretical Physics, Ivan Franko National University of Lviv. Scientific interests: statistical physics; Bose–Einstein condensation; systems obeying fractional statistics; quantitative methods in social sciences and humanities; studies of writing systems; history of science. Email: andrij.rovenchak@gmail.com*

[36] Reinhard Köhler, *Zur linguistischen Synergetik. Struktur und Dynamik der Lexik* (Bochum: Brockmeyer, 1985).

[37] Ґабріель Альтман, «Мода та істина в лінгвістиці: Особисте звернення до багатьох», у кн. *Проблеми квантитативної лінгвістики* (Чернівці: Рута, 2005), 11.

***Solomija BUK** – PhD in Linguistics, Associate Professor of the Department for General Linguistics, Ivan Franko National University of Lviv. Scientific interests: statistical linguistics, lexicography, Ukrainian for foreigners, computer and corpus linguistics, linguistic semantics and pragmatics, linguistic anthropology, psychology of happiness. E-mail: solomija@gmail.com*

Rank-frequency distributions of words in texts have a number of similarities to particle distributions in statistical physics. This makes it possible to propose a new set of parameters by which texts that are examples of complex systems can be attributed. In particular, it is possible to show the relation of these parameters to the typological classification of languages by their level of analyticity and to illustrate the evolution within several language "lineages".

The proposed parameters are calculated from the frequency data of words rarely occurring in texts. It turns out that this part of a rank-frequency distribution is characterized by more stable behaviour, in contrast to the high-frequency lexicon certain other authors used in similar studies. One of the parameters used in classification is an analogue of temperature in physics. Its smaller values correspond to languages with a higher level of analyticity (less developed word inflection, replaced by a greater number of auxiliary words and a fixed word order in sentences). The proportion of rarely used vocabulary in such languages is different from languages in which inflection is well developed.

Our approach is demonstrated using the example of translations of Antoine de Saint-Exupéry's novella *The Little Prince* into nearly forty languages and various translations of the Gospel of John. The latter are used to study the historical development of language given that this religious text was translated in several different centuries. The results indicate new dimensions of previously known concepts. They are considered in the broader context of the linguo-statistical and linguo-philosophical achievements of Wilhelm von Humboldt, Morris Swadesh, Joseph Greenberg, Gabriel Altmann, Reinhard Köhler, and other linguists.

Keywords: Zipf's law, rank–frequency distribution, text attribution, text "temperature", evolution of languages.

Bibliography

Admoni, Vladimir Grigor'evich. "Sintagmaticheskoe napriazhenie v stikhe i proze". V kn. *Invariantnye sintaksicheskie znacheniia i struktura predlozheniia*, 16–26. Moskva: Nauka, 1969.

Altmann, Gabriel. "Moda ta istyna v linhvistytsi: Osobyste zvernennia do bahat'okh". U kn. *Problemy kvantytatyvnoi linhvistyky*, 3–11. Chernivtsi: Ruta, 2005.

Altmann, Gabriel and Peter Meyer. "Physicist's look at language". In *Problems of Quantitative Linguistics*, 42–59. Černivci: Ruta, 2005.

Baugh, Albert C. and Thomas Cable. *The History of the English Language. 6th ed.* London–New York: Routledge, 2013.

Bilynsky, Mykhaylo. *Synonimika anhliis'koho diieslova: Slovnyk semantychnykh vidstanei*. L'viv: LDU imeni Ivana Franka, 1999.

Bilynsky, Mykhaylo. *English Verbal Synonyms: A Dictionary of Semantic Distances*. Lviv: Lviv University Press, 1999.

Bose. "Plancks Gesetz und Lichtquantenhypothese". *Zeitschrift für Physik* 26, no. 1 (1924): 178–181.

Buk, Solomija ta Andrij Rovenchak. "Chastotnyi slovnyk romanu "Perekhresni stezhky".» U kn. *Stezhkamy Frankovoho tekstu (komunikatyvni, stylistychni ta leksykohrafichni vymiry romanu "Perekhresni stezhky")*, F. S. Batsevych (nauk. red), S. N. Buk, L. M. Protsak, A. A. Rovenchak, L. Iu. Svarychevs'ka, I. L. Tsikhots'kyi, 138–369. L'viv: Vydavnychyi tsentr LNU imeni Ivana Franka, 2007.

– "Probing the "temperature" approach on Ukrainian texts: Long-prose fiction by Ivan Franko". In *Studies in Quantitative linguistics 23: Issues in Quantitative Linguistics 4*, edited by E. Kelih, R. Knight, J. Mačutek, A. Wilson, 160–175 Lüdenscheid: RAM-Verlag, 2016.

Einstein, Albert. "Quantentheorie des einatomigen idealen Gases. Zweite Abhandlung". *Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften zu Berlin: physikalisch-mathematischen Klasse* (1925): 3–14.

– "Quantentheorie des einatomigen idealen Gases". *Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften zu Berlin: physikalisch-mathematischen Klasse* (1924): 261–267.

Greenberg, Joseph Harold. "A quantitative approach to the morphological typology of language". *International Journal of American Linguistics* 26, no. 3 (1960): 178–194.

– *Language Typology: A Historical and Analytic Overview*. The Hague: Mouton, 1974.

Gumbol'dt, Vil'gel'm fon. "O razlichii stroeniia chelovecheskikh iazykov i ego vliianii na dukhovnoe razvitie chelovecheskogo roda". V kn. Zvegintsev, V. A. *Istoriia iazykoznaniia XIX i XX vekov v ocherkakh i izvlecheniiakh*, Ch. 1, 85–105. Moskva: Prosveshchenie, 1964.

Hirsch, E. G., I. M. Casanowicz, J. Jacobs, and M. Schloessinger. "Hapax legomena". In *The Jewish Encyclopedia, Vol. VI*, 226–229. New York: Funk and Wagnalls, 1904. Available online at: http://www.jewishencyclopedia.com/articles/7236-hapax-legomena.

Kant, Immanuel. *Metaphysische Anfangsgründe der Naturwissenschaft. Zweyte Auflage*. Riga: bey Johann Friedrich Hartknoch, 1787.

Kocherhan, Mykhailo Petrovych. *Zahal'ne movoznavstvo*. Kyiv: Akademiia, 2006.

Köhler, Reinhard. *Zur linguistischen Synergetik. Struktur und Dynamik der Lexik*. Bochum: Brockmeyer, 1985.

"Lojban". Accessed March 6, 2019. https://mw.lojban.org/papri/Lojban

Majorana, Ettore. "Il valore delle leggi statistiche nella fisica e nelle scienze sociali". *Scientia* 36 (1942): 58–66.

Mantegna, Rosario Nunzio. "Presentation of the English translation of Ettore Majorana's paper: The value of statistical laws in physics and social sciences". *Quantitative Finance* 5, no. 2 (2005): 133–140.

Miyazima, Sasuke and Keizo Yamamoto. "Measuring the temperature of texts". *Fractals* 16 (2008): 25–32.

Perebyinis, Valentyna Sydorivna, red. *Statystychni parametry styliv*. Kyiv: Naukova dumka, 1967.

Pettersson, Anders. *Verbal Art: A Philosophy of Literature and Literary Experience*, Chaps. 5–7. Montreal & Kingston–London–Ithaca: McGill-Queen's Press, 2001.

Popescu, Ioan-Iovitz, Gabriel Altmann, Peter Grzybek, Bijapur Dayaloo Jayaram, Reinhard Köhler, Viktor Krupa, Ján Macutek, Ján Mačutek, Regina Pustet, Ludmila Uhlířová, and Matummal N. Vidya. *Word frequency studies*. Berlin–New York: Mouton de Gruyter, 2009.

Rovenchak, Andrij and Solomija Buk. "Application of a quantum ensemble model to linguistic analysis". *Physica A* 390, no. 7 (2011): 1326–1331.

– "Defining thermodynamic parameters for texts from word rank-frequency distributions". *Journal of Physical Studies* 15, no. 1 (2011): 1005(1–6).

– "Part-of-speech sequences in literary text: Evidence from Ukrainian". *Journal of Quantitative Linguistics* 25, no. 1, (2018): 1–21.

Rovenchak, Andrij. "Trends in language evolution found from the frequency structure of texts mapped against the Bose-distribution". *Journal of Quantitative Linguistics* 21, no. 3 (2014): 281–294.

– "Where Alice meets Little Prince: Another approach to study language relationships". In *Sequences in Language and Text*, edited by George K. Mikros and Ján Mačutek, 217–230. Berlin–Boston: Mouton de Gruyter, 2015.

Swadesh, Morris. "Lexicostatistic dating of prehistoric ethnic contacts". *Proceedings of the American Philosophical Society* 96 (1952): 452–463.

– "Towards greater accuracy in lexicostatistic dating". *International Journal of American Linguistics* 21, no. 2 (1955): 121–137.

Tuldava, Juhan. "The frequency spectrum of text and vocabulary". *Journal of Quantitative Linguistics* 3, no. 1 (1996): 38–50.

Woodger, J. H. *Biological Principles: A Critical Study*. Abingdon–New York: Routledge, 2014.

Zahnitko, Anatolii Panasovych. *Linhvistyka tekstu: Teoriia i praktykum*. Donets'k: DonNU, 2006.