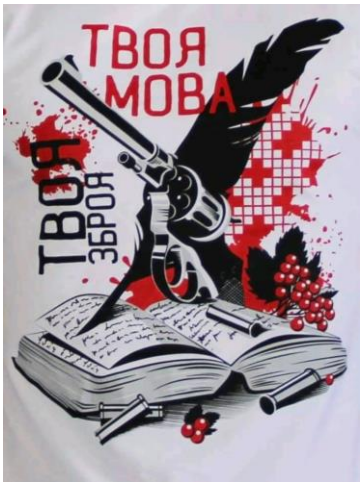
МІНІСТЕРСТВО ВНУТРІШНІХ СПРАВ УКРАЇНИ
ДЕРЖАВНА СЛУЖБА УКРАЇНИ З НАДЗВИЧАЙНИХ СИТУАЦІЙ
ЛЬВІВСЬКИЙ ДЕРЖАВНИЙ УНІВЕРСИТЕТ
БЕЗПЕКИ ЖИТТЄДІЯЛЬНОСТІ

МОВА — КОРДОН НАЦІОНАЛЬНОЇ БЕЗПЕКИ

ЗБІРНИК МАТЕРІАЛІВ
IV ВСЕУКРАЇНСЬКОЇ НАУКОВО-ПРАКТИЧНОЇ КОНФЕРЕНЦІЇ

20 ЛЮТОГО 2026 РОКУ



Львів 2026

УДК 004.912:519.21

МАТЕМАТИЧНЕ МОДЕЛЮВАННЯ ТА КОМП'ЮТЕРНИЙ АНАЛІЗ ЛЕКСИЧНОЇ СТРУКТУРИ НАУКОВИХ ТЕКСТІВ НА ОСНОВІ ЗАКОНУ ЦИПФА

Тарас ГЕМБАРА
кандидат технічних наук, доцент
Львівський державний університет безпеки життєдіяльності

Математичні закономірності природної мови давно привертають увагу як лінгвістів, так і дослідників зі сфери складних систем та аналізу даних. Однією з найвідоміших емпіричних залежностей є закон Ципфа, який описує зв'язок між рангом слова та його частотою в тексті. Пов'язаний із ним закон Хіпса описує сублінійне зростання словника V(N) зі збільшенням довжини тексту N. Такі закони використовують як компактні «підписи» тексту для порівняння жанрів, авторських стилів, доменів та якості синтетичних текстів. Наукові тексти мають специфічну лексику та структуру: поєднання термінів, загальнонаукових слів, посилань, службових елементів. Математичне моделювання частотної структури дозволяє кількісно охарактеризувати таку специфіку, а також створити комп'ютерний інструмент аналізу, наприклад у середовищі Scilab.

У праці [2, с. 2–3] показано, що навіть у природних розмовах (зокрема серед літніх людей) спостерігаються закономірності типу Ципфа та Хіпса, а параметри цих законів можуть корелювати з індивідуальними характеристиками мовлення. У прикладних задачах обробки текстів закон Ципфа використовують як інструмент розділення лексики на «часті» та «рідкісні» слова. Наприклад, у праці [6, с. 1–2] закон Ціпфа використано для керування генерацією текстів з метою боротьби з дисбалансом даних у виділенні сутностей, що демонструє практичну цінність Zipf-подібних статистик для зіставлення текстів за науковим та художнім стилем. У статті [1, с.74] запропоновано використовувати модифіковану формулу Лавалетті, яка є однією із модифікацій закону Ципфа для зіставлення текстів за науковим та художнім стилем. Теоретичні зв'язки між Zipf-подібними законами та іншими масштабними залежностями розглянуто в праці [3, с. 1–2], де подається ланцюжок «Zipf ⇒ Heaps ⇒ (ентропійні масштаби)», що додатково мотивує спільне оцінювання параметрів s

та β у прикладних дослідженнях. Питання впливу стратегій семплінгу на статистичні властивості згенерованого тексту розглядаються в [4, с. 10–13], що корисно при порівнянні Zipf-статистик людських і машинних текстів. Окрему увагу приділяють оцінюванню Zipf - Heaps законів як діагностичних інструментів (наприклад, через апроксимацію в лог-просторі та аналіз відхилень) [5, с. 1–3; 5, с. 2].

Об'єктом дослідження обрана наукова стаття англійською мовою у форматі PDF (14 стор.) [2]. Текст було автоматично витягнуто з PDF та нормалізовано (видалення переносів, заміна розділових знаків на пробіли, приведення до нижнього регістру). Далі побудовано список токенів (слів), підраховано частоти та сформовано ранговий ряд. Класична форма закону Ципфа описує спадання частоти слова зі зростанням рангу:

$$f(r) = C \cdot r^{-s} . \quad (1)$$

Тут r = 1,2,…,V — ранг слова в упорядкованому за спаданням частот списку; f(r) — частота (кількість входжень) слова рангу r; C — масштабний коефіцієнт; s — показник степеня (Zipf exponent). Для оцінювання параметрів використане логарифмування:

$$\log f(r) = \log C - s \cdot \log r . \quad (2)$$

Далі параметри (log C, s) визначали методом найменших квадратів як лінійну регресію $y=a \cdot x+b$ у просторі $x=\log r$, $y=\log f$. Діапазон апроксимації задано умовою $r \geq r_0$, щоб зменшити вплив «голови» розподілу (часто домінують службові слова та структурні елементи). Якість оцінюється коефіцієнтом детермінації R^2.

Закон Хіпса (Heaps' law) описує зростання кількості різних слів зі збільшенням довжини тексту:

$$V(N) = K \cdot N^{\beta} , \quad (3)$$

де N — кількість токенів у префіксі тексту; V(N) — кількість різних слів у перших N токенах; K — коефіцієнт масштабу; β — показник (звичайно $0<\beta<1$).

В логарифмічних координатах отримано:

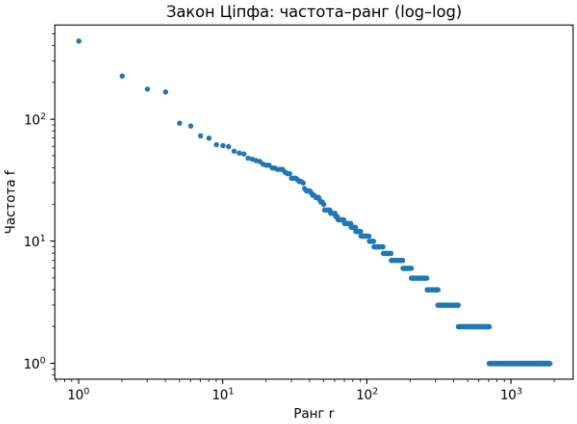
$$\log V(N) = \log K + \beta \cdot \log N , \quad (4)$$

обчислено профіль відхилення від Zipf-моделі:

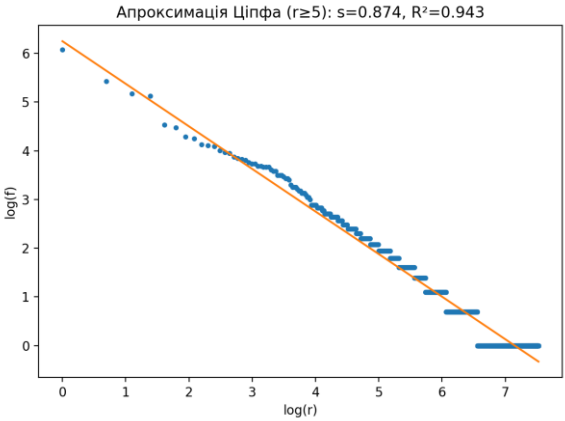
$$\Delta(r) = \log f(r) - \log f\_model(r). \quad (5)$$

Позитивні Δ(r) у середніх рангах можуть відповідати «тематичним» словам, які трапляються частіше, ніж очікує усереднена степенева модель. Комп'ютерний аналіз проведено створеним

програмним кодом в Scilab з побудовою графічних залежностей розглянутих величин на рисунках 1і 2.



а)

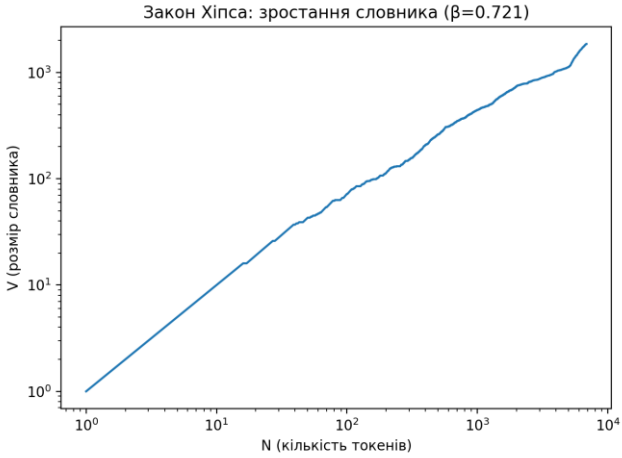


б)

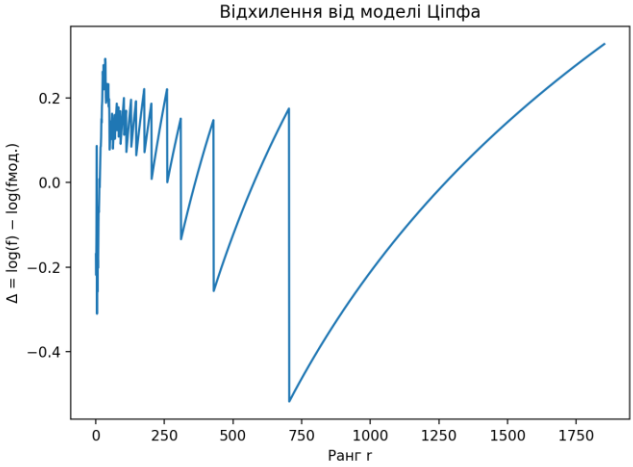
Рис. 1. Закон Ципфа: залежність частоти f від рангу r (log–log) *(а)* та апроксимація степеневою моделлю (лінійна регресія) *(б)*.

Отримане значення s≈0.874 і високе R^2≈0.943 свідчать, що рангово-частотна крива добре описується степеневою моделлю. Параметр β≈0.721 відповідає сублінійному зростанню словника, типовому для природних текстів та наукових статей [2, с. 8–9; 5, с. 1].

На прикладі PDF-статті отримано N=6869 токенів і V=1852 різних слів; оцінено параметри s≈0.874, β≈0.721 та отримано діагностичні графіки, що підтверджують степеневий характер розподілу частот.



а)



б)

Рис. 2. Закон Хіпса: зростання словника V(N) зі зростанням N (*а)* та відхилення від апроксимації Zipf: Δ(r)=log f−log f_model (*б)*.

Отримані результати можуть бути використані для порівняння корпусів жанрів, контролю якості витягу тексту з PDF, напівавтоматичного виділення тематичної лексики, швидкої діагностики статистики синтетичних текстів.

ЛІТЕРАТУРА

1. Волошиновська І. А. Зіставлення рангово-частотного розподілу слів в англо-, німецько- та україномовних наукових і художніх текстах // *Наукові записки. Серія «Філологічна»*. – Острог – 2013. – Вип. 37. – С. 74–76**.**

2. Abe M.S., Otake-Matsuura M. Scaling laws in natural conversations among elderly people. PLoS ONE 16(2): e0246884, 2021, 14 p.

3. Dębowski Ł. From Zipf’s Law to Neural Scaling through Heaps’ Law and Hilberg’s Hypothesis. arXiv:2512.13491, 2025, 33 p.

4. Finlayson M., Hewitt J., Koller A., Swayamdipta S., Sabharwal A. Closing the Curious Case of Neural Text Degeneration. ICLR 2024 (accepted), 40 p.

5. Lai J., Lu Y., Wu S. Heaps’ Law in Large Language Models: Artificial and Natural Languages in GPT-Neo. arXiv:2308.15311, 2023, 15 p.

6. Wang Z., Liu X., et al. A Zipf’s Law-based Text Generation Approach for Entity Extraction. arXiv:2205.12636, 2022, 31 p.

УДК 81’25:81’42:316.77:37.091

РЕЛІГІЙНА ТЕРМІНОЛОГІЯ ПОЛЬСЬКОЇ МОВИ: СТИЛІСТИКО-ПРАГМАТИЧНІ ТА ЕМОТИВНІ ХАРАКТЕРИСТИКИ

Тамара ГОЛІ-ОГЛУ,
кандидат філологічних наук
ДВНЗ «Приазовський державний технічний університет»

Сучасний лінгвістичний і перекладознавчий дискурси демонструють посилений інтерес до вивчення спеціалізованих