

APPLICATION ZIPF'S LAW IN LINGUISTIC RESEARCH

N. Kuzminska, G. Vasyutunska

National University of Food Technologies

This article describes the results of the check Zipf's law. To test selected and the narrow literary texts. The study was conducted on a frequency dictionary, which are produced using the program TextAnalyzer, set up jointly with the International Scientific and Training Center of Information Technologies and Systems, National Academy of Sciences of Ukraine and MES of Ukraine.

ПРИМЕНЕНИЕ ЗАКОНА ЦИПФА В ЛИНГВИСТИЧЕСКИХ ИССЛЕДОВАНИЯХ

Н.Л. Кузьминская, Ю.О. Васютинская

Национальный университет пищевых технологий

В статье описаны результаты проверки закона Ципфа. Для проверки выбраны узкопрофильные и литературные тексты. Исследование проводилось по частотным словарям, которые составлены с помощью программы TextAnalyzer, созданной совместно с Международным научно-учебным центром информационных технологий и систем НАН Украины и МОН Украины.

ЗАСТОСУВАННЯ ЗАКОНА ЦИПФА В ЛІНГВІСТИЧНИХ ДОСЛІДЖЕННЯХ

Н.Л. Кузьмінська, Ю.О. Васютинська

Національний університет харчових технологій

У статті описані результати перевірки закону Ципфа. Для перевірки вибрані вузькопрофільні та літературні тексти. Дослідження проводилось по частотним словникам, які складені за допомогою програми TextAnalyzer, створеної спільно з Міжнародним науково-навчальним центром інформаційних технологій та систем НАН України та МОН України.

Вступ.

У різні часи у багатьох науковців виникало бажання пов'язати алгебру чисел із законами природи, довівши цим досконалість усього навколишнього.

Будь-який витвір мистецтва – це складно організоване ціле, яке складається із багатьох порівняно простих, але гармонічно поєднаних елементів. В літературі – це слова, в музиці – звуки, в

живописі – фарби... Іноді ці елементи легко відділяються один від одного, іноді ні. Деякі зустрічаються досить часто, деякі – дуже рідко.

Автори належать до тих, хто вірить, що і мова, як складний процес поєднання звуків, букв, слів теж має тісний зв'язок з числами, тому прагнуть до пошуку істини у цьому питанні.

Теоретичні відомості.

На початку минулого століття француз Есту (J. Estoup) задумав створити систему стенографії на науковій основі. Його задум ґрунтувався на думці: чим частіше слово зустрічається, тим простішим має бути стенографічний знак, який позначає це слово. Для визначення частот слів, Есту склав словник, який в наш час отримав назву частотного. Слова в цьому словнику розміщувалися в порядку спадання частоти, з якою вони зустрічалися в досліджуваному тексті. Він помітив, що якщо порядковий номер слова помножити на його частоту, то отримані добутки майже не змінюються, тобто $p_i \cdot i \approx \text{const}$, де i - порядковий номер слова, p_i - частота i -го слова. Даний результат було опубліковано в 1916 році. Есту не помітив, що якщо цю залежність переписати інакше, тобто у вигляді функціональної залежності частоти p_i від порядкового номера i : $p_i \approx \frac{\text{const}}{i}$, то отримаємо «гармонійну послідовність чисел», яку отримав Піфагор для коливань струни [1].

Результат Есту залишився непоміченим, десять років по тому до схожого результату прийшов співробітник телефонних лабораторій фірми «Белл» Е. Кондон (E. Condon) – йому частоти слів знадобилися для оптимізації телефонних кодів. Ще через десять років вийшла книга американського лінгвіста Ципфа (Georg Zipf) «Психобіологія мови». Він намагався говорити про гуманітарну науку на мові чисел та графіків. Так на сьогодні закономірність знайдену ще Естом та Кондоном називають «законом Ципфа», так як саме Ципф розширив спостереження Еста і Кондона і дав їм змістовну інтерпретацію [1]. Свій закон Ципф вивів експериментально, досліджуючи різноманітні тексти на різних мовах і різної історичної давності. Отримана залежність графічно виражається гіперболою, а точніше має форму «гіперболічних

сходів». Ципф виявив, що при зміні одного текста іншим загальний характер розподілу не змінюється.

На початку 50-х років минулого століття кібернетик Б. Мандельброт (B. Mandelbrot), який досліджував процеси оптимізації кодування, запропонував теоретичне обґрунтування закону Ципфа. Він вважав, що можна порівняти письмову мову з кодуванням, причому всі знаки повинні мати певну «вартість». Б.

Мандельброт отримав таку залежність: $P_i = \frac{K}{(B + i)^\gamma}$, де γ , K , B -

const, які залежать від властивостей тексту та визначаються статистичними методами. Ця залежність включає в себе закон Ципфа, як частинний випадок [1,2].

Також було встановлено, що описаній залежності підпорядковані не тільки мови, але й явища соціального і біологічного характеру: розподіли населення за розмірами прибутків в капіталістичних країнах (В. Парето, 1897 р.), біологічних родів за чисельністю видів (Дж. Уїлліс, 1922 р.), вчених за числом опублікованих ними статей (А. Лотка, 1926 р.), документів в будь-якій галузі знань (С. Бредфорд), міст США по чисельності населення (Д. Ципф, 1949 р.) [2].

Дослідники, які працювали над перевіркою закону Ципфа в наш час, виявили, що додатки, які створюються для Linux, теж узгоджуються з цією теорією [3], крім того ця ж закономірність діє і для частки ринку програмного забезпечення, безалкогольних напоїв, автомобілів, цукерок, для частоти звернення до інтернет-сайтів [4].

Упорядкована сукупність, яка задається ранжуванням (порядком розміщення) найменувань елементів за частотою їх появи в порядку її спадання, називається ранговим розподілом. Прикладами рангових розподілів є розподіли, які вивчав Ципф. При побудові рангові розподіли в більшості випадків мають форму закону Ципфа з поправкою Мандельброта: $p_i \cdot i^\gamma \approx const$. При цьому коефіцієнт γ є змінною величиною, він не змінюється тільки на середній ділянці графіка розподілу. Якщо графік приведеної вище закономірності побудувати в логарифмічних координатах, то ця ділянка приймає форму прямої, і називається центральною зоною рангового розподілу, зліва від цієї зони знаходиться зона ядра рангового розподілу, а праворуч – зона утигання.

Якщо аналізувати терміни якої-небудь області знань, то ядерна зона містить найбільш загальноновживані, загальнонаукові терміни. Центральна зона містить терміни найбільш характерні і специфічні саме для даної області знань. В зоні утинання зосереджені терміни не характерні для даної області знань. Таким чином, основа лексики певної області знань зосереджена в центральній зоні рангового розподілу. За допомогою термінів ядерної зони ця область знань «стикується з більш загальними областями знань», а зона утинання грає роль авангарду, який налагоджує зв'язки з іншими галузями науки. За допомогою графіка рангового розподілу можна охарактеризувати досліджувану область знань. Графік, який має велику ядерну зону та малу зону утинання відноситься до консервативної області знань. Для динамічних галузей науки характерна збільшена зона утинання, для оригінальних областей знань – мала ядерна зона [2].

Закон Ципфа і українська мова.

Автори статті вирішили долучитися до числа тих дослідників, які працюють над перевіркою закону Ципфа. Для перевірки було взято однакового обсягу частини вузькопрофільних (харчова промисловість, економіка) та літературного текстів написаних українською мовою в один часовий період (5 років) (текст 1 – [5], текст 2 – [6], текст 3 – [7]). Для складання частотних словників використана програма TextAnalyzer, яка була створена спільно з Міжнародним науково-навчальним центром інформаційних технологій та систем НАН України та МОН України і пройшла апробацію в вищезгаданому центрі. Більш докладно остання версія програми описана у [8].

Ми не мали за мету порівнювати ці тексти між собою, лише перевірити на скільки для кожного виду текстів виконується закон Ципфа, на скільки справджується припущення щодо зон рангового розподілу при аналізі термінів, якщо розглядати тексти вузькопрофільного напрямку.

Приведемо основні статистичні характеристики текстів, які розглядаються (таблиця 1).

Таблиця 1.

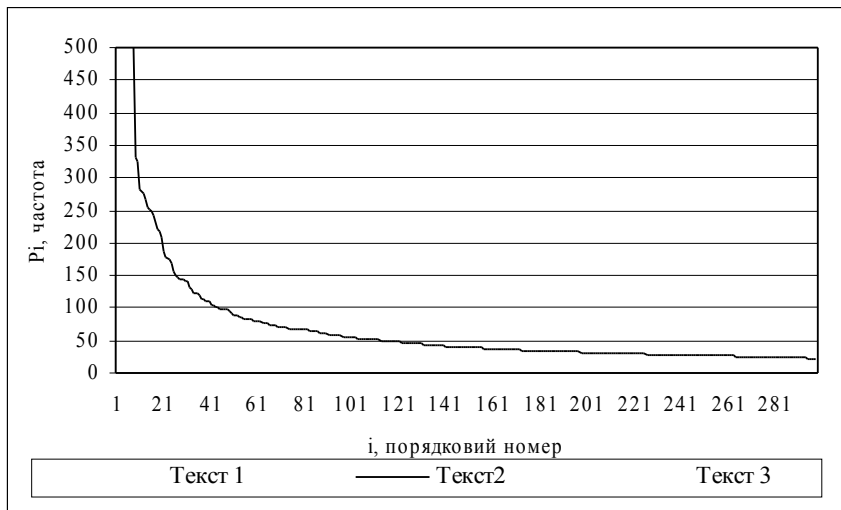
Характеристика	Текст 1	Текст 2	Текст 3
Всього літер у словнику	88 718	89 545	63 663
Всього літер у тексті	356 111	333 465	170 398
Всього слів у тексті	50 740	50 632	36 424

Продовження табл. 1.

Всього унікальних слів	9 886	10 289	8 709
Макс. кількість літер у слові	25	23	19
Мін. кількість літер у слові	1	1	1
Усього речень	2 608	3 105	3 132
Середня кількість літер у слові (в тексті)	7,02	6,59	4,68
Середня кількість літер у слові (в словнику)	8,97	8,70	7,31
Коефіцієнт багатства лексики	0,20	0,20	0,24
Середня кількість слів в реченні	19,46	16,31	11,63

Зобразимо графічно для кожного текста залежність, описану Ципфом (діаграма 1), утворивши з частотного словника ранговий розподіл (дане поняття описане в п. «Теоретичні відомості»). Для наочності вибрані обмеження по частоті 500, а по порядковому номеру 300, що не впливає на загальний вигляд діаграми. Дійсно, описана Ципфом залежність має форму гіперболи не залежно від виду текста. Більше того, для вузькопрофільних текстів (текст 1, текст 2), криві майже співпадають.

Діаграма 1.



Спробуємо описати характеристики областей знань (текст 1, текст 2). Якщо для даних текстів побудувати графіки рангових розподілів в логарифмічних координатах, то можна прослідкувати наступне для обох розподілів форму прямої графіки приймають наближено на ділянці від 90 до 320 (за порядковим номером), що відповідає центральній зоні. Тоді ядерній зоні відповідає ділянка від 1 до 89, а зоні утинання від 321 до 88 718 (текст 1) та 89 545 (текст 2), отже, оскільки зона утинання збільшена це свідчить про те, що ці два тексти представляють динамічні науки. І насправді, як харчова промисловість, так і інноваційний менеджмент є достатньо динамічними галузями науки.

На прикладі тексту 1, проаналізуємо терміни, які потрапили в різні зони. В ядерну зону потрапили терміни: управління, розвиток, проект, виробництво, інновація, в центральну: нововведення, менеджмент, прогнозування, у зону утинання: час, держава, відносини та інші. Оскільки провести межу між загальнонауковими та найбільш характерними для даної області знань термінами досить важко. А в зону утинання, дійсно, потрапили терміни, які не характерні для даної області знань.

Висновки.

На основі проведеного дослідження можна зробити такі висновки:

- для обраних текстів закон Ципфа не виконався для всього частотного словника, але на окремих ділянках рангового розподілу він має місце, що дає підстави для розгляду задачі про знаходження поправки Мандельброта.

- розглянувши графіки рангових розподілів в логарифмічних координатах, слова частотних словників були поділені на зони, при цьому не було виявлено чіткої межі між розподілом термінів на загальнонаукові та найбільш характерні для даної області знань, можливо це пов'язано з тим, що для підтвердження припущення щодо розподілу термінів по зонах потрібно взяти текст іншого більш вузького спрямування. Припущення щодо зони утинання підтвердилось, в ній знаходяться терміни не характерні для даної області знань.

Список літератури.

1. Ю.К. Орлов. Невидимая гармония. // Число и мысль. Вып. 3. – М.: Знание, 1980. С. 70 - 105.

2. Н.Н. Чурсин. Популярная информатика. – К.: «Техника», 1982. – 158 с.

3. T. Maillart, D. Sornette, S. Spaeth and G. von Krogh. Empirical Tests of Zipf's Law Mechanism in Open Source Linux Distribution. // Physical Review Letters (2008) (<http://www.physorg.com>).

4. Сет Годин. Идея-вирус? Эпидемия! Заставьте клиентов работать на ваш сбыт. – СПб.: "Питер", 2005 г. – 192 с.

5. П.П. Микитюк. Інноваційний менеджмент: Навчальний посібник. – Тернопіль: Економічна думка, 2006. – 295 с.

6. І.В. Сирохман, Т.М. Раситюк. Товарознавство м'яса і м'ясних товарів. Підручник. – К.: Центр навчальної літератури, 2004. – 384 с.

7. Юрко Покальчук. Озерний вітер. – К.: Клуб сімейного дозвілля, 2008. – 232 с.

8. Власенко Н.А., Кузьминская Н.Л., Максименко А.А. Текстометрические исследования многоязычных научных текстов. // УСиМ. – 2009. – № 2. – с. 43-47.