СТАТИСТИЧНІ ЗАКОНОМІРНОСТІ ЛІНГВІСТИКИ КОМП'ЮТЕРНИХ ПРОГРАМ

О. С. Кушнір, М. А. Альфавіцький, В. І. Богданець, Л. Б. Іваніцький, І. М. Катеринчук

кафедра оптоелектроніки та інформаційних технологій,
Львівський національний університет імені Івана Франка
вул. Тарнавського, 107, 79017 м. Львів, Україна
e-mail: o_kushnir@franko.lviv.ua

Відомо, що статистика текстів, написаних природними мовами, виявляє низку емпіричних закономірностей [1, 2]. Зокрема, залежність абсолютної частоти F випадання того чи іншого слова в тексті від його рангу r (порядкового номера в списку всіх слів за спаданням їхньої частоти) описується степеневим законом Ціпфа:

$$F(r) \propto r^{-\alpha}, \tag{1}$$

де $\alpha \sim 1$ – стала. Частотну характеристику (залежність імовірності p слів од їхньої частоти F) також представляють степеневою формулою:

$$p(F) \propto F^{-\beta}, \tag{2}$$

де $\beta \sim 2$. Іноді замість масової функції розподілу $p(F)$ використовують кумулятивну (інтегральну) функцію $P(F)$, яку описують законом Парето $P(F) \propto F^{-k}$ (k – стала). Нарешті, незважаючи на наявність більш строгих, але громіздких теоретичних виразів (див. [3, 4]), розміри словника V (кількість різних слів без урахування їхніх повторень у тексті) залежно від розмірів L тексту (кількості всіх слів, з урахуванням їхніх повторень) переважно описують у рамках емпіричного закону Гіпса [5]:

$$V(L) \propto L^{\theta}, \tag{3}$$

Питання про те, чи виконуються закономірності (1)–(3) для штучних текстів, викликають традиційний інтерес дослідників. Йдеться, зокрема, про рандомні або рандомізовані тексти різних типів, тексти на штучних мовах, а також коди програм, написані різними мовами (див. [6, 7]). Зокрема, відомо, що закон Ціпфа у формі (1) або схожій формі виконується для текстів окремих програм, які входять до ядра операційної системи Linux [7], а також для низки програм, написаних на різних мовах програмування [8–10]. Це ж стосується й закону Гіпса [10]. Було порівняно параметри ентропії, складності [11], флуктуацій і довгосяжних кореляцій [12–14] для природних і програмних мов, а також проаналізовано масштабно інваріантні мережі, сформовані програмами на різних об'єктно-орієнтованих мовах, які теж описуються степеневими законами [15].

Мета цієї роботи – подальші дослідження емпіричних законів статистичної лінгвістики для текстів комп'ютерних програм і порівняння з природними текстами. Зокрема, ми вивчали програми, включені до ядра операційної системи Linux (мова C) та до інструментарію Swing для створення графічних інтерфейсів (мова Java), який входить до пакету JDK. Основні відмінності дослідження від попередніх такі: (1) більші довжини текстів програм, (2) комплексне вивчення всіх емпіричних залежностей $F(r)$, $p(F)$, $P(F)$ і $V(L)$ і (3) чітке з'ясування впливу коментарів до програм на статистичні дані. Ці особливості підвищують надійність порівняння теорія–експеримент і достовірність висновків.

Для реалізації цих завдань було створено програму для розрахунку даних $F(r)$, $p(F)$, $P(F)$ і $V(L)$ для текстів програм на мовах C/C++ і Java. На відміну від попередніх робіт за предметом, для послаблення шумів ми визначали залежність $V(L)$ із усередненням за

рухомим вікном зі змінними розмірами. Програма передбачала об'єднання всіх файлів *.c, *.cpp або *.java, зчитаних із заданої папки (включно із вкладеними папками), в єдиний вихідний файл для подальшої роботи, розрахунки частотних таблиць і словника, створення списку абсолютних позицій обраного слова і його т. зв. часів очікування, а також збереження вихідних даних у форматі *.csv. Програма виводила графіки залежностей $F(r)$, $p(F)$, $P(F)$ і $V(L)$ в подвійному логарифмічному масштабі, лінійність яких відповідає виконанню першого та другого законів Ціпфа, закону Парето і закону Гіпса. Було передбачено режими роботи з текстами програм із видаленими коментарями, текстами коментарів і текстами, що містять програми разом з коментарями.

Дефініція слів у мовах програмування відмінна від природної мови, де слова ідентифікують як послідовності літер і окремих символів між сусідніми пробілами. Як і в [10], ми розрізняли ідентифікатори, ключові слова, літерали, оператори і роздільники як різні типи "слів" у текстах програм. Було розроблено і реалізовано алгоритми ідентифікації всіх цих типів. Разом досліджено 251 файл із ядра Linux і 508 файлів із бібліотеки Swing, загальний обсяг яких складав 6,53 МБ і 0,75 МБ, відповідно. У табл. 1, як приклад, наведено рангові списки слів ($r = 1 \div 12$) для двох конкретних програм із Linux і Swing.

Таблиця 1. Рангові списки слів для текстів окремих комп'ютерних програм, що входять до ядра Linux і бібліотеки Swing ($r = 1 \div 12$), з урахуванням і без урахування коментарів.

	1	2	3	4	5	6	7	8	9	10	11	12
Linux Kernel with comments	()	,	;	*	->	=	.	{}	if	the	&	/**/
Linux Kernel without comments	()	;	,	->	=	*	{}	struct	if	&	.	return
Java Swing with comments	.	()	*	;	the	{}	=	>	code	<	0	if
Java Swing without comments	()	.	;	,	{}	=	0	int	if	public	return	null

У текстах без коментарів найбільш уживані слова – це роздільники ('()', '.', ';' тощо), оператори ('*', '->', '=') і ключові слова ('struct', 'int', 'if'). Їх можна трактувати як "функціональні слова" (див. [10]). Серед літералів найвищий ранг має '0', а серед ідентифікаторів – назви змінних 'i' та 'g'. Можна очікувати, що представники двох останніх класів можуть виявитися "ключовими словами" (в термінах лінгвістики, а не програмування) в текстах програм. Проте проблема з'ясування "функціональних" і "ключових" слів у програмах виходить за рамки даного дослідження.

Загалом ранги слів із бібліотеки Swing, наведені в табл. 1, непогано відповідають даним роботи [10]. Проте слід також враховувати, що в [10] використовували версію JDK 1.5, а не новішу версію 1.8, як у нас, а з тієї ж роботи [10] відомо, що статистика текстів програми помітно залежить від її версії (наприклад, значення степенів α і θ змінюються в межах 7% і 12%, відповідно). Зазначимо також відчутні зміни в рангових списках, які стаються внаслідок урахування коментарів. Наприклад, додавання природної мови до тексту програм із ядра Linux приводить до появи в списку табл. 1 слова 'the' і позначення коментарів '/**/'. Для тексту бібліотеки Swing слово '/**/' має ранг, нижчий за $r = 12$, а тому не потрапляє до табл. 1.

На рис. 1, 2 і 3 наведено емпіричні залежності $F(r)$, $p(F)$ і $V(L)$ для текстів програм, що входять до ядра Linux і бібліотеки Swing, без урахування коментарів. Для "змішаних" текстів програм із коментарями (дані не представлено на рис. 1–3) відносні частоти f ($f = F/L$) перших за рангами слів складають $f = 7{,}5\%$ для '()' (мова C) і $f = 7{,}0\%$ для '.' (мова Java). Ці дані принципово не відрізняються від відповідних параметрів для при-

родних текстів. Проте ці частоти помітно зростають (f = 10,3% і 12,3% для '()' в обох мовах) у випадку "чистих" текстів без коментарів. Отже, текстам програм притаманний підвищений уміст найчастотніших слів, тобто слова з найвищими рангами домінують.

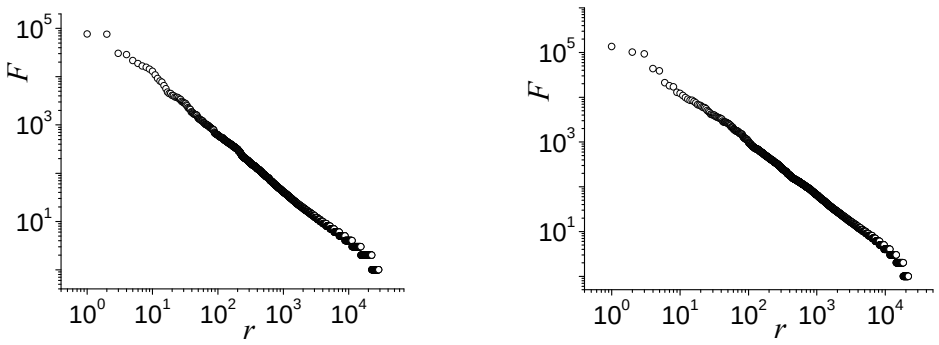

Рис. 1. Рангові залежності для лексичних одиниць із текстів програм ядра Linux (а) і Swing (б) без коментарів у подвійному логарифмічному масштабі.

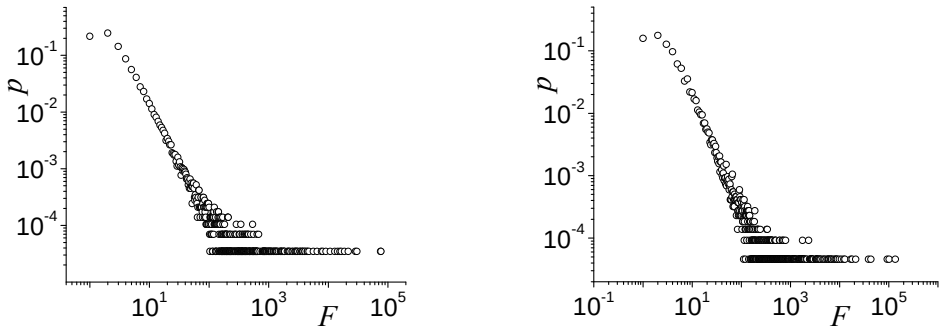

Рис. 2. Частотні залежності для лексичних одиниць із текстів програм ядра Linux (а) і Swing (б) без коментарів у подвійному логарифмічному масштабі.

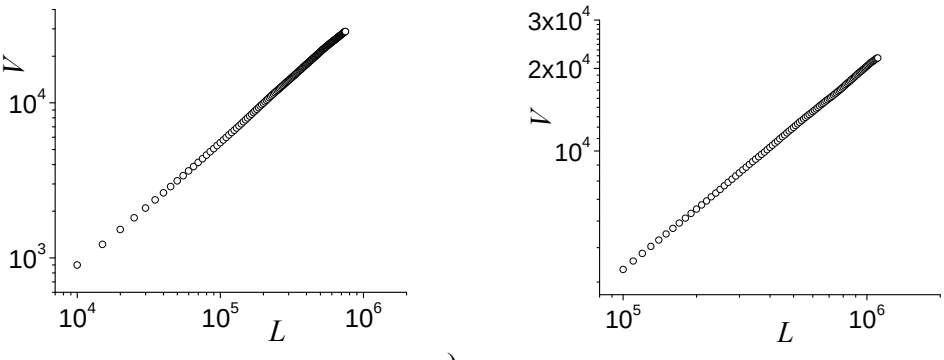

Рис. 3. Залежності словника лексичних одиниць від довжини тексту для програм ядра Linux (а) і Swing (б) без коментарів у подвійному логарифмічному масштабі.

На рис. 1 бачимо нетипову для природних текстів ситуацію: ширини "полиць", сформованих словами з частотами $F = 1$ (hapax legomena), є дещо вужчими, ніж для слів із $F = 2$ (dis legomena). Це відповідає аномалії, присутній на частотних залежностях рис. 2 – немонотонності кривої $p(F)$ на низькочастотній ділянці: кількість слів із частотою $F = 2$ перевищує кількість слів із частотою $F = 1$. Наскільки відомо авторам, таке явище досі не спостерігали ні в природних текстах, ні в текстах програм (див. [7, 10]), а його пояснення потребує додаткових досліджень. Проте аналог уже відомий в описі мережі взаємодіючих Java-модулів, що входять до великих програмних систем [16]. Цікаво, що для

обох груп програм із коментарями ця аномалія відсутня – на залежностях $p(F)$ можна спостерігати лише деяке локальне зменшення нахилу на низькочастотній ділянці.

Наші результати підтверджують висновки [9, 10] про те, що початкова ділянка рангових залежностей дещо пологіша, аніж передбачає закон Ціпфа в формі (1). Ці відхилення описують поправкою Мандельброта $F(r) \propto (r + r_0)^{-\alpha}$, де r_0 – стала. За плавністю і регулярністю наші залежності $F(r)$ ближчі до даних [10], але не даних [9].

Для перевірки виконання законів (1)–(3) ми використали стандартні методи лінійної апроксимації емпіричних даних $F(r)$, $P(F)$, $p(F)$ і $V(L)$ у подвійному логарифмічному масштабі. У табл. 2 подано статистичні параметри, які описують апроксимацію основних залежностей $F(r)$, $p(F)$ і $V(L)$ формулами (1)–(3) для пакетів програм, що входять до ядра Linux і бібліотеки Swing. Тут же наведено сумарну довжину L текстів програм з кожного пакету в одиницях кількості слів. Для словника діапазон апроксимації відповідав усьому діапазонові незалежної змінної, а для рангової та паретівської залежностей, а особливо для частотної залежності було відкинуто обмежену кількість початкових даних на ділянці "голови" і більшість даних із зашумленого "хвоста". Відповідно, найнижча якість апроксимації притаманна частотній залежності.

Таблиця 2. Деякі параметри рангової і частотної залежності, а також залежності розмірів словника від розмірів тексту для текстів комп'ютерних програм, що входять до ядра Linux і бібліотеки Swing: α, β і θ – показники степенів відповідно залежностей $F(r)$, $p(F)$ і $V(L)$, R і SD з індексами α, β і θ – коефіцієнти детермінації і нормовані середньоквадратичні відхилення теорія–експеримент для відповідних лінійних апроксимацій, L – сумарна довжина текстів програм.

Параметр	C	C без коментарів	Java	Java без коментарів
α	1,15	1,10	1,33	1,17
R_α	0,9991	0,9980	0,9992	0,9992
SD_α	0,019	0,028	0,021	0,019
β	1,90	2,01	1,73	1,91
R_β	0,9828	0,9852	0,9726	0,9851
SD_β	0,113	0,117	0,143	0,110
θ	0,73	0,82	0,66	0,74
R_θ	0,9999	0,9999	0,9999	0,9999
SD_θ	0,003	0,005	0,003	0,001
L	1 091 217	745 626	2 066 483	1 102 135

Знайдені нами для бібліотеки Swing (мова Java) і ядра Linux (мова C) степені Ціпфа α (відповідно 1,17 і 1,10) грубо корелюють із даними літератури. Автор [10] одержав $\alpha = 1{,}28$ для Swing; знайдене нами усереднене для 12 вивчених у праці [10] програм мовою Java значення α складає 1,25, а середнє для шести вивчених в [10] програм мовою C дорівнює $\alpha \approx 1{,}20$. Крім того, менш точна глобальна апроксимація залежностей $F(r)$ у всьому діапазоні рангів дає значення $\alpha = 1{,}29$ для Swing і $\alpha = 1{,}11$ для ядра Linux, які ближчі до даних [10]. У будь-якому разі, за нашими результатами і за даними [10] можна констатувати, що мові Java притаманні вищі степені α, порівняно з мовою C.

На відміну від результатів [10], залежності $V(L)$ на рис. 3 не виявляють нерегулярностей, що, мабуть, зумовлено вжитою нами процедурою усереднення. Багато дослідників констатують, що емпіричний закон Гіпса є лише порівняно грубим наближенням в описі словника природних текстів, особливо в разі великих за обсягом текстів [3–5, 17]. Зокрема, у стилометрії загальновідомою є масштабна неінваріантність C-параметра Герда-

на [18], яка прямо пов'язана з порушенням припущення про степеневу залежність (3). Проте ситуація з текстами програм інша: якість апроксимації словника $V(L)$ степеневою залежністю (3) досить висока (див. рис. 3 і табл. 2 і порівн. із даними R [10]).

Відомо, що існують доволі загальні теоретичні співвідношення поміж степенями α, β, k і θ, які можна вважати аналогами співвідношень скейлінгу для різних критичних індексів у теорії критичних явищ: $\beta = 1 + 1/\alpha$, $k = 1/\alpha$, $\beta = 1 + \theta$ і $\theta = 1/\alpha$ [3, 5]. Наприклад, для програм на мові С із коментарями з даних $F(r)$ маємо $\alpha = 1{,}15$, а з даних $P(F)$ – значення $k = 0{,}90$, звідки одержуємо близьке значення $\alpha = 1/k = 1{,}11$. Загалом же з урахуванням усіх зв'язків величини незалежно визначених емпіричних параметрів α, β і θ із табл. 2 узгоджуються одна з одною з орієнтовною точністю від 5 до 20%. Це цілком задовільний результат, ураховуючи значні джерела похибок розрахунку цих степенів за грубим графічним методом апроксимації, а також шуми залежності $p(F)$ [1].

Порівняння даних табл. 2 для степенів Гіпса програм на мовах С і Java (відповідно $\theta = 0{,}82$ і 0,74) засвідчує, що перша з цих мов володіє багатшим словником. Щоправда, цей висновок не підтверджується літературними даними [10]: одержаний автором [10] степінь θ для бібліотеки Swing дорівнює 0,744, знайдене нами усереднене значення θ для 12 вивчених у [10] текстів мовою Java складає 0,745, а середнє для шести вивчених у [10] текстів мовою С дорівнює $\theta \approx 0{,}750$. Проте дані [10] для θ ми кваліфікуємо як менш надійні: виходячи зі зв'язків степенів α і θ, більша усереднена величина α, одержана в [10] для мови Java, порівняно з мовою С, повинна означати менше середнє θ для Java. Іншими словами, фактичний збіг величин степенів θ для мов С і Java [10] не узгоджується з тим, що степені α для С і Java в цій праці різні.

Цікаво, що наше дослідження стимулював пошук мов із обмеженим словником на зразок ієрогліфічних, для яких закони статистичної лінгвістики (1)–(3) значно модифікуються (див. [19]): власне, початкова гіпотеза полягала в тому, що внаслідок вузькоспеціалізованого характеру текстів програм їхній словник повинен бути бідним, а закон зростання $V(L)$ – повільнішим, аніж степеневий. Проте кінцеві дані повністю спростували це припущення: степінь θ для програм без коментарів виявився доволі високим (див. табл. 2). Хоча обмежена кількість "функціональних слів" і справді охоплює значний обсяг тексту, у програмах присутній значний пласт низькочастотної лексики, пов'язаної з іменами змінних, процедур тощо. Відомо також, що степінь Гіпса для аналітичних мов на зразок англійської ($\theta \sim 0{,}5 \div 0{,}6$) нижчий, аніж для синтетичних мов на зразок української ($\theta \sim 0{,}8$ або й більше), що пов'язано з численними флексіями в цих мовах. Хоча флексії не притаманні морфології мов програмування і ці мови повністю аналітичні, їхній степінь Гіпса усе ж ближчий до значень, типових для синтетичних мов.

Із табл. 2 видно, що коментарі займають вагому частку професійних програм. Вони складають від 30 до 50% обсягу програм, які формують Linux Kernel або бібліотеку Swing. Що стосується статистичних наслідків наявності коментарів, їхнє врахування в текстах програм збільшує параметр α і зменшує параметри β і θ (див. табл. 2). Згадаймо, що класичні значення α і β, передбачувані більшістю теоретичних моделей для природних текстів, складають відповідно $\alpha = 1$ і $\beta = 2$ [20–22]. Отже, ми приходимо до парадоксального на перший погляд висновку: врахування коментарів як вкраплень природної мови в текстах програм не наближає, а віддаляє характеристики текстів програм від характеристик "класичних" природних текстів. Пояснення, очевидно, ґрунтується на тому, що самі тексти коментарів специфічні – вони вузькоспеціалізовані, тематично дуже однотипні та мовно вбогі. Відомо, що степінь Ціпфа α для таких текстів може істотно переважати одиницю, а степінь β частотної залежності, відповідно, ставати меншим за двійку [20, 21]. Непрямо це підтверджують і дані θ, які фактично описують різнома-

ніття словника: як це не дивно, але динаміка зростання словника текстів програм без коментарів швидша за відповідну динаміку для програм із коментарями (див. табл. 2). Отже, текст коментарів справді повинен характеризуватися вбогим словником.

Сформулюємо основні висновки: (1) на значному статистичному матеріалі підтверджено, що основні закономірності статистичної лінгвістики природних текстів – закони Ціпфа, Парето і Гіпса – виконуються і для кодів, написаних мовами програмування C і Java; (2) виявлено явище "недозаселеності" низькочастотної "голови" розподілу ймовірності лексичних одиниць у текстах програм, аналог якого відсутній для природних текстів; (3) хоча мови програмування за морфологією суто аналітичні, порівняно високі показники степеня Гіпса роблять їх спорідненими із синтетичними мовами; (4) показано, що закон Гіпса для комп'ютерних програм виконується з високою точністю, не притаманною більшості природних текстів. Результати даного дослідження можуть бути корисними у підготовці програмістів, наприклад, для перевірки відповідності вживання тих чи інших програмних методів чи засобів зі стандартами професійних програм.

Нарешті, серед перспективних об'єктів цього дослідження вбачаємо масштабну поведінку флуктуацій словника та частот слів і знаків, довгосяжні кореляції в текстах програм і характеристики складності програмних кодів як складних систем. Цікавим є й питання "ключових слів" у текстах програм, які визначають "семантику" програмного коду. Його з'ясування потребує вивчення статистики часів очікування лексичних одиниць і параметрів неоднорідності розподілу знаків, слів і лексичних n-грам по тексту.

[1] M. E. J. Newman. Contemp. Phys., 46, 323 (2005).
[2] D. H. Zanette. Centro Atomico Bariloche. 2012. http://fisica.cab.cnea.gov.ar/ estadistica/2te/
[3] F. Font-Clos, G. Boleda, A. Corral. New J. Phys., 15, 093033 (2013).
[4] F. Font-Clos, A. Corral. Phys. Rev. Lett., 114, 238701 (2015).
[5] D. C. van Leijenhorst, Th. P. van der Weide. Inform. Sci., 170, 263 (2005).
[6] W. Li. IEEE Trans. Inform. Theory, 38, 1842 (1992).
[7] A. Krause, A. Zollmann. Algorithms for Information Networks – Project Report. 2005. http://www.cs.cmu.edu/~zollmann/publications.html
[8] P. Kokol, T. Kokol. J. Amer. Soc. Inform. Sci., 47, 781 (1996).
[9] D. Pierret, D. Poshyvanyk. 17th IEEE International Conf. ICPC'09, Program Comprehension, 228 (2009).
[10] Hongyu Zhang. Inform. Processing & Management, 45, 477 (2009).
[11] G. Febres, K. Jaffé, C. Gershenson. arXiv, 1311.5427 (2015).
[12] P. Kokol, J. Brest, V. Žumer. Cybernetics and Systems, 28, 43 (1997).
[13] P. Kokol, V. Podgorelec, M. Zorman, T. Kokol, T. Njivar. J. Amer. Soc. Inform. Sci., 50, 1295 (1999).
[14] K. Kosmidis, A. Kalampokis, P. Argyrakis. Physica A, 370, 808 (2006).
[15] A. Potanin, J. Noble, M. Frean, R. Biddle. Commun. ACM, 48, 99 (2005).
[16] P. Louridas, D. Spinellis, V. Vlachos. ACM Trans. Softw. Engin. Method., 18, Art. 2 (2008).
[17] S. Bernhardsson, L. E. Correa da Rocha, P. Minnhagen. New J. Phys., 11, 123015 (2009).
[18] F. J. Tweedie, R. H. Baayen. Computers and the Humanities, 32, 323 (1998).
[19] Linyuan Lu, Zi-Ke Zhang, Tao Zhou. Sci. Rep., 3, 1082 (2013).
[20] R. Ferrer i Cancho. Eur. Phys. J. B, 44, 249 (2005).
[21] R. Ferrer i Cancho. Eur. Phys. J. B, 47, 449 (2005).
[22] R. Ferrer i Cancho. BioSystems, 84, 242 (2006).