ФЛУКТУАЦІЇ ЧАСТОТ ЛІНГВІСТИЧНИХ ЕЛЕМЕНТІВ У ТЕКСТОВІЙ БАЗІ

О. С. Кушнір, І. Я. Довгань, М. А. Альфавіцький, Л. Б. Іваніцький, В. В. Яремків

кафедра оптоелектроніки та інформаційних технологій,
Львівський національний університет імені Івана Франка
вул. Тарнавського, 107, 79017 м. Львів, Україна
e-mail: vv.yaremkiv@gmail.com

Складні динамічні властивості текстів, написаних природними мовами, і відповідні флуктуаційні явища привертають значну увагу дослідників [1–4]. Відповідно до сучасних поглядів, просторові (або, у більш традиційних термінах, часові) позиції тих чи інших лінгвістичних об'єктів (знаків, букв, слів або їхніх послідовностей – n-грам) у тексті не є хаотичними; вони скорельовані на значних масштабах, сумірних із повними розмірами самого тексту. Попри деяку очевидність цього твердження і відповідне розуміння обмеженості нульової стохастичної гіпотези про текст як випадкову послідовність лінгвістичних знаків, не всі вияви та віддалені наслідки наявності довгосяжних кореляцій у текстах на сьогодні добре вивчені та усвідомлені. Наприклад, хоча в статистичній лінгвістиці часто трактують відносну частоту слів $f = F/L$ (де F – абсолютна частота цих слів, L – довжина або «тривалість» тексту) як своєрідну імовірність появи слів ($p = \lim_{L\to\infty}(F/L)$), насправді така границя за наявності довгосяжних кореляцій може й не існувати (див., наприклад, доведення [5]). Іншими словами, тоді звичне макроскопічне наближення, в якому нехтують флуктуаційними ефектами, може взагалі стати недосяжним, а для відповідних параметрів системи буде відсутнє «самоусереднення» [4].

У припущенні про ергодичність системи, розгляд її часової динаміки можна замінити на аналіз «статичного зрізу» ансамблю систем. Проте приклади такого аналізу для текстів рідкісні (див. [4]), а більшість досліджень флуктуацій виконано за стандартним методом рухомого вікна шириною w в тексті. У цій праці ми коротко опишемо емпіричні дані альтернативних статистичних досліджень для низки лінгвістичних елементів у корпусах текстів. Предметом вивчення є поведінка абсолютних частот цих елементів зі зміною довжини текстів, коли параметр F змінюється не через зміни w в єдиному тексті ($0 < w < L$), а внаслідок розгляду ансамблю текстів із різними довжинами L ($0 < L < L_{\max}$).

Як і за умови динамічної еволюції одного тексту, за згаданих умов можна зробити аналогічні припущення про те, що усереднена абсолютна частота $\overline{F}$ та її середньоквадратична флуктуація ΔF степенево залежать від L і одна від одної:

$$\overline{F}(L) \propto L^{\alpha_1},\ \Delta F(L) \propto L^{\alpha_2},\ \Delta F(L) \propto \overline{F}^{\gamma}, \qquad (1)$$

де α_1, α_2 і γ – постійні. Остання з формул (1) відповідає відомому законові Тейлора [6]. Природно припустити, що кількість деяких лінгвістичних елементів пропорційна до довжини тексту ($\alpha_1 = 1$ або принаймні $\alpha_1 \approx 1$), а межі для параметра α_2 визначаються відсутністю кореляцій у складній лінгвістичній системі ($\alpha_2 = ½$) або наявністю додатних довгосяжних кореляцій ($½ < \alpha_2 < 1$). Тоді для параметра γ матимемо $\gamma = \alpha_2/\alpha_1 \approx \alpha_2$.

Перевірку наведених теоретичних положень ми проводили для великих англомовного та україномовного корпусів, які налічували відповідно 4838 і 1280 текстів. У кодуванні UTF-8 повний обсяг англійських текстів становив 2,3 ГБ, а українських текстів – 0,5 ГБ.

Серед інших статистичних даних згадаємо, що довжини англійських текстів змінювалися в межах від 83 до 562 тис. слів, а середня довжина текстів складала 78,6 тис. слів.

Мовою C# було написано програмне забезпечення, що давало змогу знаходити абсолютні частоти таких лінгвістичних об'єктів: заданих користувачем букв, символів і відповідних n-грам, слів і відповідних n-грам, а також повного словника V тексту і парціальних внесків V_1, V_2, … слів із різними частотами $F = 1, 2, \ldots$ до словника. Для знаходження параметрів $\overline{F}$ і ΔF на підставі первинних даних $F(L)$ було створено програму, яка розраховувала гістограми $\overline{F}(L)$ і $\Delta F(L)$ шляхом бінування, тобто поділу всього діапазону довжин текстів $L_i = L_{\min} \div L_{\max}$ на n інтервалів (бінів).

Ми використовували три методи бінування: 1) з однаковими довжинами бінів ($\Delta L_n = \text{const}$); 2) з різними довжинами бінів, кожен з яких містить однакову кількість емпіричних точок L_i; 3) експоненційне (або логарифмічне) бінування, коли довжина n-го біну зростає за законом $\Delta L_n = a^n$, де $a > 1$ – стала. Останній тип бінування часто застосовують для обробки статистичних даних, якщо передбачається масштабно інваріантні степеневі залежності параметрів [7]. Зазначимо, що тип бінування (1) не давав статистично надійних результатів, а дані методів бінування (2) і (3), хоча дещо й відрізнялися, проте давали більш узгоджені та надійні результати. Було вивчено вплив кількості бінів на дані, одержані для степенів α_1, α_2 і γ, і показано, що ці дані найменш критичні в діапазоні $n_{\max} = 15 \div 40$. Нарешті, характеристичні степені α_1, α_2 і γ визначалися за наближеним методом графічної лінійної апроксимації емпіричних залежностей $\overline{F}(L)$, $\Delta F(L)$ і $\Delta F(\overline{F})$ у подвійному логарифмічному масштабі.

На рис. 1 наведено кілька прикладів вихідних залежностей $F(L)$ для корпусу англійських текстів, а на рис. 2 – залежності $\overline{F}(L)$, $\Delta F(L)$ і $\Delta F(\overline{F})$ для 2-грами «of the».

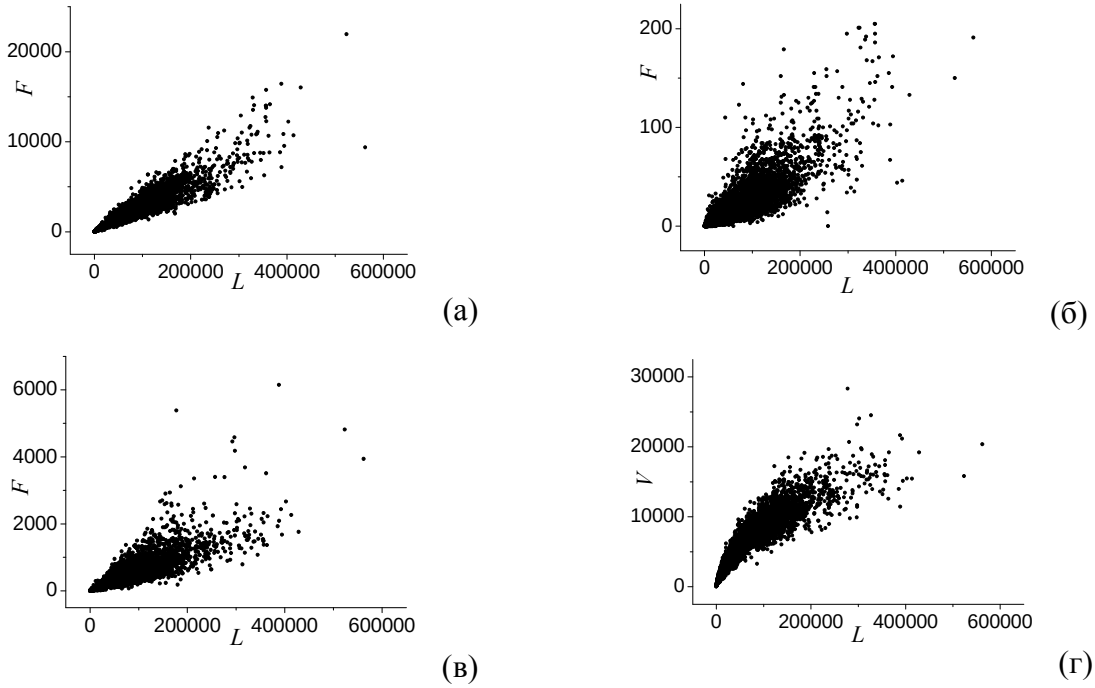

Рис. 1. Емпіричні залежності $F(L)$ для слова «and» (а), слова «word» (б) і словесної біграми «of the» (в), а також залежність словника $V(L)$ (г), здобуті для англомовного корпусу (4838 точок).

У Таблицях 1 і 2 представлено основні результати цього дослідження – коефіцієнти α_1, α_2 і γ степеневих законів (1), одержані для низки лінгвістичних об'єктів в англомовно-

му та україномовному корпусах. Унаслідок відмінності даних, здобутих за методами бінування (2) і (3), тут подано інтервали для емпіричних значень коефіцієнтів.

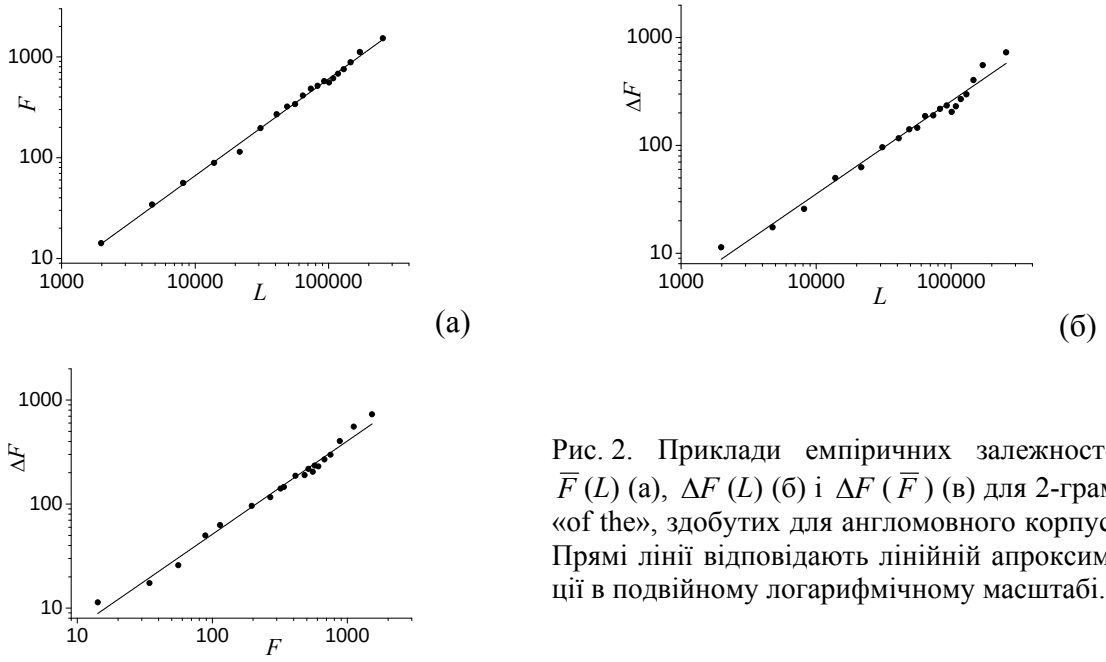

Рис. 2. Приклади емпіричних залежностей $\overline{F}(L)$ (а), $\Delta F(L)$ (б) і $\Delta F(\overline{F})$ (в) для 2-грами «of the», здобутих для англомовного корпусу. Прямі лінії відповідають лінійній апроксимації в подвійному логарифмічному масштабі.

Таблиця 1. Значення параметрів α_1, α_2 і γ степеневих законів (1), одержані для досліджених лінгвістичних елементів англомовного корпусу за допомогою методів бінування (2) і (3)

Лінгвістичний елемент	Степінь α_1	Степінь α_2	Степінь γ
лексична монограма «and»	0,93–0,98	0,81–0,91	0,86–0,93
лексична монограма «word»	0,99–1,13	0,71–0,80	0,70–0,71
лексична монограма «you»	1,01–1,20	0,82–0,97	0,76–0,81
лексична біграма «of the»	0,95–0,99	0,85–0,86	0,90–0,91
лексична біграма «he had»	0,99–1,06	0,85–0,88	0,82–0,83
Повний словник V	0,62–0,66	0,55–0,71	0,89–1,07
Парціальний словник V_1	0,54–0,62	0,42–0,63	0,85–1,06
Парціальний словник V_3	0,67–0,76	0,57–0,67	0,80–0,98

Таблиця 2. Значення параметрів α_1, α_2 і γ степеневих законів (1), одержані для досліджених лінгвістичних елементів україномовного корпусу за допомогою методів бінування (2) і (3)

Лінгвістичний елемент	Степінь α_1	Степінь α_2	Степінь γ
буквена монограма «и»	0,99–1,01	0,91–0,94	0,91–0,94
буквена монограма «і»	0,98–1,00	0,89–0,91	0,91–0,93
буквена біграма «ел»	0,97–0,99	0,76–0,77	0,78–0,80
буквена триграма «стр»	0,99–1,01	0,74–0,75	0,74–0,75
Повний словник V	0,82–0,83	0,82–0,87	0,93–1,05

Зазначимо, що розкид даних для окремих коефіцієнтів і різних досліджуваних корпусів (типова похибка $\pm(0{,}01\div0{,}05)$) схиляє до думки проте, що ці корпуси все ще недостатні для одержання точних результатів для α_1, α_2 і γ. На додаток, графічний метод апроксимації, мабуть, програє в точності безпосередній нелінійній апроксимації функцій (1).

Загалом наші дані підтверджують інтуїтивне припущення про лінійне зростання середньої частоти $\overline{F}$ від довжини тексту L ($\alpha_1 = 1$), хоча точність дотримання цієї рівності

для англійського корпусу недостатня навіть для такого високочастотного слова як «and». Очевидним винятком є повний словник і його парціальні складові, для яких добре відомий сублінійний закон зростання ($\alpha_1 < 1$). Зрозумілими є відмінності коефіцієнтів α_1 для українських і англійських текстів: український словник зростає швидше через синтетичний характер мови. Водночас, прості феноменологічні викладки прогнозують однакові швидкості зростання, тобто однакові α_1 для словника V і його складових V_1, V_2, ... [8]. Проте для з'ясування цього моменту нашим даним трохи не вистарчає точності. У будь-якому разі, точність для повного словника V, мабуть, вища через багатшу статистику відповідних лінгвістичних елементів.

Крім даних для словників англійської мови, для яких $\alpha_2 \sim ½$, залежності $\Delta F\,(L)$ для всіх решти лінгвістичних елементів виявляють аномальні флуктуації із $\alpha_2 > ½$. Це може засвідчувати присутність довгосяжних кореляцій у послідовностях відповідних елементів, унаслідок згаданої вище властивості ергодичності лінгвістичних систем. З іншого боку, автори [4] вважають, що це пояснюється тематичними відмінностями текстів.

Ще одним цікавим моментом є значна відмінність коефіцієнтів γ від класичного значення ½, яка має місце фактично для всіх лінгвістичних елементів і корпусів. Зокрема, для випадку словника це непогано корелює з результатами авторів [4], які одержали $\gamma = 1$. Факт $\gamma > ½$ означає, що відносні флуктуації повільно (за законом $\Delta F / \overline{F}(L) \propto L^{\gamma-1}$) загасають зі зростанням масштабів системи L. За умови $\gamma \to 1$, яка наближено виконується для словника, вони взагалі не прямують до нуля, тобто макроскопічного наближення не можна досягнути. Іншими словами, врахування мікроскопічних флуктуацій словника принципове навіть для безмежно великої лінгвістичної системи (див. також [9]).

У праці [3] зроблено висновок, що довгосяжні кореляції позицій лінгвістичних елементів у тексті та кластеризація цих елементів є відносно незалежними явищами, а для елементів з істотнішою семантикою ці явища посилюються. Хоча ми вивчаємо не єдиний текст, а їхній ансамбль, усе ж аналогії тут корисні. У зв'язку з цим зауважимо, що наші дані для степеня α_2 у залежностях $\Delta F\,(L)$ для різних лінгвістичних об'єктів засвідчують практичну відсутність кореляції між семантикою об'єкта та величиною α_2. Справді, найбільшу семантику в англійському корпусі слід приписувати слову «word» або, принаймні, слову «you». Водночас, степені α_2 для таких функціональних слів або n-грам як «and» або «of the» не нижчі, або й вищі, ніж для слова «word». В українському корпусі схожі висновки можна зробити при порівнянні степенів α_2 для буквених n-грам, семантичне навантаження яких очевидно вище, і окремих букв. Не спостерігаємо й тенденції до зростання α_2 при переході від монограм до біграм в англійському корпусі.

[1] W. Ebeling, A. Neiman. Physica A, 215, 233 (1995).
[2] D. Y. Manin. arXiv:0809.0103 [cs.CL] (2013).
[3] E. G. Altmann, G. Cristadoro, M. D. Esposti. Proc. Nat. Acad. Sci., 109, 11582 (2012).
[4] M. Gerlach, E. G. Altmann. New J. Phys., 16, 113010 (2014).
[5] Ф. Реиф. Статистическая физика. Москва, 1977. 352 с.
[6] Z. Eisler, I. Bartos, J. Kertész. Adv. Phys., 57, 89 (2008).
[7] M. E. J. Newman. Contemp. Phys., 46, 323 (2005).
[8] A. Kornai. Glottometrics, 4, 60 (2002).
[9] О. С. Кушнір, О. С. Брик, В. Є. Дзіковський, Л. Б. Іваніцький, І. М. Катеринчук, Я. П. Кісь. Вісн. нац. ун-ту «Львівська політехніка", серія «Інф. системи та мережі», №854, 228 (2016).