УДК 004.6, 004.9, 538.9

О. С. Кушнір [1], О. С. Брик [1], В. Є. Дзіковський [2], Л. Б. Іваніцький [1], І. М. Катеринчук [1], Я. П. Кісь [3]

[1] Львівський національний університет імені Івана Франка,
кафедра оптоелектроніки та інформаційних технологій,
[2] Природничий коледж Львівського національного університету імені Івана Франка
[3] Національний університет "Львівська політехніка",
кафедра інформаційних систем і мереж

СТАТИСТИЧНИЙ РОЗПОДІЛ І ФЛУКТУАЦІЇ ДОВЖИН РЕЧЕНЬ В УКРАЇНСЬКИХ, РОСІЙСЬКИХ І АНГЛІЙСЬКИХ КОРПУСАХ

© Кушнір О. С., Брик О. С., Дзіковський В. Є., Іваніцький Л. Б., Катеринчук І. М., Кісь Я. П., 2016

Вивчено розподіли частот речень за їхньою довжиною для українського, російського та англійського корпусів і з'ясовано середні довжини речень в одиницях знаків, літер і слів. Показано, що хвости статистичних розподілів задовільно описуються експоненційною або близькими до неї функціями, що узгоджуються зі стохастичним характером довжини речень. Доведено, що залежність флуктуацій частоти речень різних довжин від середніх значень цієї частоти визначається степеневим законом Тейлора. Значні відносні флуктуації частот і відносні зміни середньої довжини речення підтверджують важливість урахування флуктуаційних явищ у статистичній лінгвістиці.

Ключові слова: комп'ютерна лінгвістика, корпуси, статистичні розподіли, довжина речення, флуктуації.

We have studied statistical distributions of the frequency of sentences over their length for Ukrainian, Russian and English corpora and found the average sentence lengths in terms of linguistic signs, letters and words. It has been shown that the tails of the statistical distributions are satisfactorily described by the exponential function or the related ones, which is consistent with random nature of the sentence length. We have proven that the fluctuations of the frequency of sentences of different lengths depend on the average values of that frequency according to the Taylor's power law. Significant relative fluctuations of the frequency and the relative changes in the average sentence length confirm the importance of fluctuation phenomena in the statistical linguistics.

Key words: computational linguistics, corpora, statistical distributions, sentence length, fluctuations.

1. Вступ. Загальна постановка проблеми та аналіз літератури

Вивчення статистичних закономірностей для лінгвістичних елементів різних рівнів – на зразок довжин слів або речень – цікаві в фундаментальному плані, зокрема з погляду встановлення структурних особливостей природних мов і, загальніше, особливостей співвідношень і взаємодій між об'єктами різних класів, які формують складні лінгвістичні системи (див., наприклад, [1, 2]). Відомо, що лінгвістичні системи часто демонструють цікаві статистичні властивості, зокрема степеневі або близькі до них розподіли з "важкими хвостами", а також масштабну інваріантність і фрактальність [1]. Дослідження статистики довжин слів і речень актуальні й у прикладному плані з огляду на автоматизовану обробку та класифікацію текстових документів, встановлення авторства, тематики, стилю, жанру та мови текстів [3, 4], зокрема з використанням систем із машинним навчанням.

Додатковий інтерес до досліджень розподілів довжин слів зумовлений перевіркою відомих "закону скорочення" Ціпфа (за яким довжина слова та його частота негативно скорельовані, виходячи з міркувань оптимізації та ефективності спілкування [1, 5–7]), закону Менцерата–

Lviv Polytechnic National University Institutional Repository http://ena.lp.edu.ua

Альтмана (який пов'язує довжини лінгвістичних елементів різних рівнів [8, 9]), визначенням та практичним використанням "індексу читабельності" [10, 11] та іншими міркуваннями [12].

Хоча перше дослідження статистики довжин речень з'явилося ще в 1938 році [13], а розвиток напряму тісно пов'язаний зі стилеметрією та оцінюванням тієї самої читабельності текстів, статистику речень сьогодні вивчено помітно слабше, аніж для слів (див. праці [14–20]). Зазвичай дослідники наводять лише середню довжину речення, рідше – стандартне відхилення, не звертаючи уваги на відповідний розподіл ймовірності та не з'ясовуючи, чи він має нормальний характер (див. обговорення в [14]). Часто вивчають окремі тексти [21], хоча найповнішу статистику можна здобути лише на корпусі. Нарешті, статистичні розміри досліджуваних текстів або корпусів текстів зазвичай були недостатньо великими. Так, у праці [14] аналізували корпус розміром ~ 10^6 слів, що явно недостатньо з огляду на можливості сучасних комп'ютерів. Більше того, часто в літературі навіть не вказують розмірів текстової вибірки, для якої одержано числові дані. Нарешті, досі досліджували статистику речень, не враховуючи флуктуації відповідних частот од тексту до тексту. Водночас з літератури відомо, що флуктуаційні явища можуть відігравати важливу роль у лінгвістичних системах [9, 21].

Висловлені міркування обґрунтовують дослідження статистичних розподілів довжин речень для помірно великих за розмірами українських, російських і англійських текстових баз з урахуванням флуктуаційних явищ. Ці дослідження і є основною метою роботи.

2. Об'єкти та методика досліджень

Об'єктами досліджень були електронні корпуси класичної художньої літератури українською, російською та англійською мовами. Як видно з формальних характеристик, наведених у табл. 1, ці корпуси можна кваліфікувати як помірно значні за розмірами; у разі англійської текстової бази можна говорити про т. зв. "гігакорпус" (сума довжин текстів $\sum_i L_i \sim 10^9$). Це сприяє репрезентативності корпусів і підвищує надійність статистичних даних.

На відміну від праць [18, 19], але схоже до підходу [14, 20], ми не вважали знаки ":" ";" "()" і лапки ознаками закінчення речення, визначаючи довжину речення як кількість знаків (l_s, включно з розділовими знаками, але без пробілів), літер (l_c) або слів (l_w), які трапляються між двома сусідніми пунктуаційними знаками ".", "?", "!" або їхніми комбінаціями. Для обробки текстів у середовищі Visual Studio було створено програму для визначення довжин речень у всіх текстах корпусів із експортуванням даних.

Для кожного текстового файла визначалися емпіричні абсолютні (F_{ij}) і відносні ($f_{ij} = F_{ij} / L_s$, де L_s – кількість усіх речень) частоти речень з різними довжинами l_j, а також такі *інтратекстуальні* параметри: середня довжина речень в *i*-му тексті $\bar{l}_i$ і середньоквадратичне відхилення (СКВ) Δl_i довжин речень у цьому тексті від середнього значення

$$\bar{l}_i = \sum_j l_j f_{ij}\,,\ \Delta l_i = [(\overline{l^2})_i - (\bar{l}_i)^2]^{1/2}\,, \tag{1}$$

де $(\overline{l^2})_i = \sum_j l_j^2 f_{ij}$. Крім того, на підставі даних для частот f_{ij} окремих текстів ми визначали такі *інтертекстуальні* частотні параметри: зважені середні за корпусом відносні частоти речень $\overline{f}_j$ із різними довжинами l_j і зважені за корпусом СКВ Δf_j:

$$\overline{f}_j = \sum_i f_{ij} w_i\,,\ \Delta f_j = [\overline{f_j^2} - \overline{f}_j^2]^{1/2}\,, \tag{2}$$

де $\overline{f_j^2} = \sum_i f_{ij}^2 w_i$ – це середні квадрати кожної з довжин речення в корпусі, а $w_i = L_i / \sum_i L_i$ – статистичні вагові коефіцієнти текстів.

Нарешті такі інтертекстуальні параметри, як середня довжина речення в корпусі $\bar{l}$ і відхилення Δl довжини речення від середнього по всьому корпусі було знайдено за формулами

$$\bar{l} = \sum_i \bar{l}_i w_i\,,\ \Delta l = [\overline{l^2} - \bar{l}^2]^{1/2}\,, \tag{3}$$

Lviv Polytechnic National University Institutional Repository http://ena.lp.edu.ua

де $\overline{l^2} = \sum_i \bar{l}_i^2 w_i$ – це середній квадрат довжини речення в корпусі. Зазначимо, що альтернативні визначення $\bar{l} = \sum_j l_j \overline{f}_j$, $\Delta l = [\overline{l^2} - \bar{l}^2]^{1/2}$ (із $\overline{l^2} = \sum_j l_j^2 \overline{f}_j$) давали дещо інші результати для $\bar{l}$ і Δl.

Таблиця 1

Деякі загальні характеристики досліджуваних українського, російського і англійського корпусів літературних текстів

Характеристика корпусу	Український корпус	Російський корпус	Англійський корпус
Кількість текстів N	750	850	2990
Загальна довжина ΣL всіх текстів відповідно в знаках, літерах і словах	$6{,}87 \cdot 10^7$ $6{,}40 \cdot 10^7$ $1{,}28 \cdot 10^7$	$1{,}56 \cdot 10^8$ $1{,}46 \cdot 10^8$ $2{,}83 \cdot 10^7$	$8{,}67 \cdot 10^8$ $8{,}17 \cdot 10^8$ $1{,}90 \cdot 10^8$
Середня довжина $\bar{L}$ тексту відповідно в знаках, літерах і словах	$92{,}0 \cdot 10^3$ $85{,}7 \cdot 10^3$ $17{,}1 \cdot 10^3$	$183 \cdot 10^3$ $171 \cdot 10^3$ $33{,}3 \cdot 10^3$	$290 \cdot 10^3$ $273 \cdot 10^3$ $63{,}5 \cdot 10^3$

Варто зауважити, що залежності частот f(l) для окремого тексту і параметри $\bar{l}_i$, Δl_i, з одного боку, а також залежності f(l) для всього корпусу і відповідні параметри $\bar{l}$, Δl, з іншого боку, мають різний зміст. Різною може виявитися й їхня аналітична поведінка, а тому було би помилкою автоматично переносити висновки для інтратекстуальних залежностей f(l) (наприклад, взятих із дослідження [14]) на відповідні інтертекстуальні залежності.

3. Емпіричні дані та їхнє обговорення

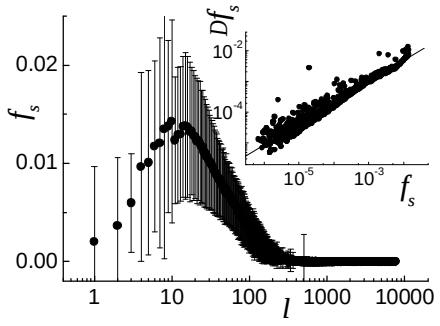
Залежності частот речень з різними довжинами у корпусах від довжини речення, вираженої в різних одиницях, наведено на рис. 1–3. Тут вжито скорочені позначення $\overline{f}_j \equiv f$, $\Delta f_j \equiv \Delta f$ і $l_j \equiv l$. Натомість у позначеннях підкреслено те, яких саме одиниць довжин речення стосується частота – f_s, f_c чи f_w.

Логарифмічний масштаб по осях абсцис на рис. 1–3 вжито для кращої візуалізації області малих l в умовах дуже істотної асиметрії розподілів $f(l)$, а вертикальні риски на рис. 1–3 позначають флуктуації Δf частот речень за корпусом, виражені як СКВ. Оскільки відносні флуктуації $\Delta f/f$ не надто відмінні від одиниці, вони відіграють ключову роль, а нехтування ними неприпустиме. Цей факт підтверджує загальні міркування [9]: флуктуаційні відхилення від усереднених параметрів лінгвістичних систем такі ж важливі, як і самі лінгвістичні закони, які формулюють для цих параметрів. Нарешті, за нашими спостереженнями, обмеженість обсягу вибірки часто приводить не так до недооцінки СКВ, як до його переоцінки (порівн. параметри Δf для різних корпусів).

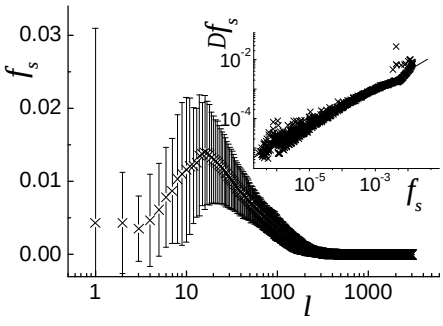
На залежностях $f(l)$, найперше для частот f_s і f_c, помітні деякі нерегулярності в області найменших l. Серед можливих причин – скорочення в текстах (наприклад, “Mr.” в англійській, “ін.” в українській тощо) і залишкові неточності оцифрування текстів. Вплив цих факторів важко усунути на практиці. Можна помітити й менш істотні “викиди” функцій $f(l)$ за великих l, мабуть, внаслідок помилкової відсутності пунктуації. Тим не менше, автори праці [14] трактували схожі нерегулярності, знайдені у залежностях $f(l)$ для довжини слів в окремих текстів як наслідок суперпозиції кількох внесків різних класів слів, функціональні залежності $f(l)$ яких різні. Це нагадує стандартний підхід у спектроскопії, де нерегулярності спектральних профілів приписують внескам елементарних ґауссових кривих. Такий механізм нерегулярностей не виключений, оскільки трактування численних спостережуваних на рис. 1–3 “викидів” у функції $f(l)$ для гігакорпусів як виключного результату помилок оцифрування є непевним. За браком місця нижче ми не наводимо докладніших даних для інтратекстуальних параметрів $\bar{l}_i$ і Δl_i. Дані для інтертекстуальних параметрів $\bar{l}$, Δl наведено в табл. 2. Тут ми уникаємо стандартної форми представлення

Lviv Polytechnic National University Institutional Repository http://ena.lp.edu.ua

результатів $\bar{l} \pm \Delta l$, яка коректна лише для нормального розподілу $f(l)$, але не для різко асиметричних функцій $f(l)$ на зразок показаних на рис. 1–3. Так, для логнормального розподілу випадкової величини, що формально нагадує одержані нами дані $f(l)$, коректною формою представлення є $\bar{l} / \Delta l \div \bar{l} \Delta l$ [22]. Внаслідок асиметрії середня та найбільш імовірна довжини слів істотно різні. Наприклад, для англійської мови перша становить $\bar{l}_c \approx 79$, а друга – $l_{c\max} \approx 19$.

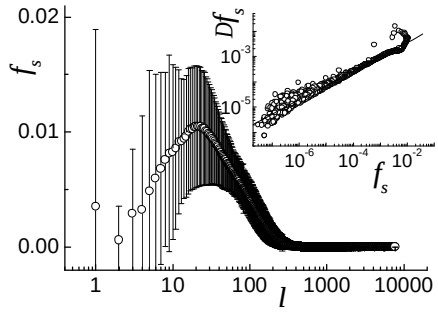


а

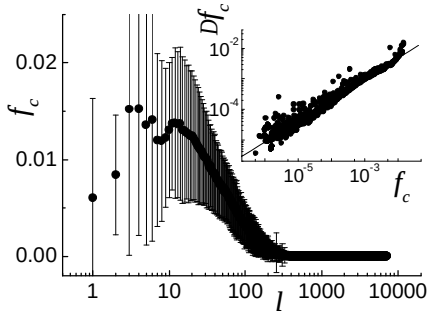


б

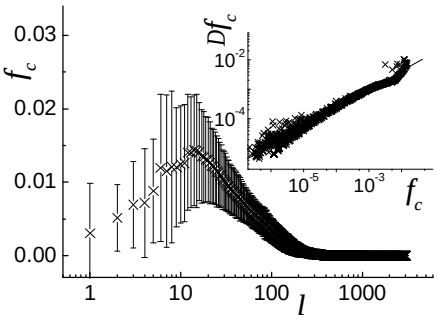
Рис. 1. Залежності середньої за корпусом частоти речень f_s від їхньої довжини l_s, вираженої кількістю знаків, для українського (а), російського (б) і англійського (в) корпусів. Вертикальні риски позначають флуктуації частоти за корпусом, виражені величиною СКВ. На вставках – залежності флуктуацій за корпусом Δf_s від середнього значення f_s



в

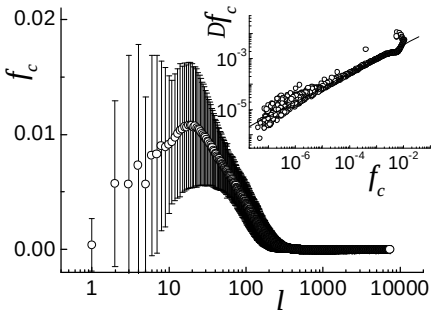


а



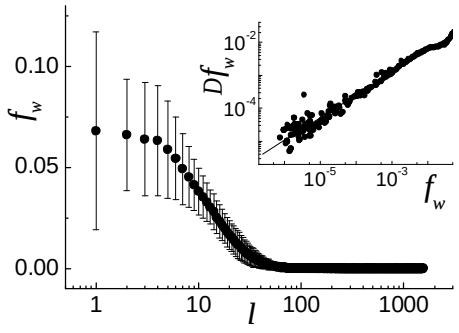
б

Рис. 2. Залежності середньої за корпусом частоти речень f_c від їхньої довжини l_c, вираженої кількістю літер, для українського (а), російського (б) і англійського (в) корпусів. Вертикальні риски позначають флуктуації частоти за корпусом, виражені величиною СКВ. На вставках – залежності флуктуацій за корпусом Δf_c від середнього значення f_c

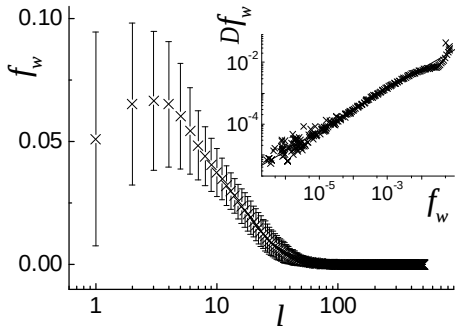


в

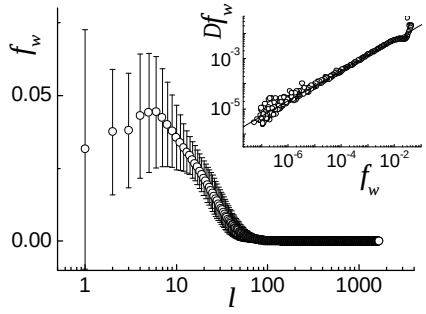
Lviv Polytechnic National University Institutional Repository http://ena.lp.edu.ua



а



б



в

Рис. 3. Залежності середньої за корпусом частоти речень f_w від їхньої довжини l_w, вираженої кількістю слів, для українського (а), російського (б) і англійського (в) корпусів. Вертикальні риски позначають флуктуації частоти за корпусом, виражені величиною СКВ. На вставках – залежності флуктуацій за корпусом Δf_w від середнього значення f_w

Таблиця 2

Середні за корпусом довжини речень в одиницях кількості знаків ($\bar{l}_s$), літер ($\bar{l}_c$) і слів ($\bar{l}_w$), а також відповідні СКВ Δl_s, Δl_c і Δl_w для українського, російського і англійського корпусів літературних текстів

Одиниця довжини речення	Український корпус			Російський корпус			Англійський корпус		
	$\bar{l}$	Δl	$\Delta l / \bar{l}$	$\bar{l}$	Δl	$\Delta l / \bar{l}$	$\bar{l}$	Δl	$\Delta l / \bar{l}$
Знаки (*s*)	72,4	26,3	0,36	76,5	33,5	0,44	83,5	30,2	0,36
Літери (*c*)	67,7	24,7	0,37	72,3	32,8	0,45	79,2	29,3	0,37
Слова (*w*)	13,1	4,38	0,33	13,7	5,64	0,41	18,2	6,42	0,35

Дані табл. 2 дають змогу оцінити довжину "українського речення", принаймні для мови художньої літератури, а також знайти середню насиченість тексту розділовими знаками (в перерахунку на одну літеру або одне слово – відповідно параметри $(\bar{l}_s - \bar{l}_c)/\bar{l}_c$ або $(\bar{l}_s - \bar{l}_c)/\bar{l}_w$). Точність оцінок визначають відносні зміни $\Delta l / \bar{l}$, які наближаються до 40 % майже незалежно від мови. За табл. 2 за формулою $\bar{l}_{cw} = \bar{l}_c / \bar{l}_w$ легко одержати середню довжину слова у літерах. Для українського, російського та англійського корпусів маємо відповідно $\bar{l}_{cw}$ = 5,2, 5,3 і 4,4, що непогано узгоджується і з численними даними літератури, і з нашими даними $\bar{l}_{cw}$, здобутими прямим методом (4,9, 5,2 і 4,3, відповідно). Очевидно, що точність збігу цих даних зростає зі зростанням розмірів корпусів. Незалежно від одиниць вимірювання, середня довжина речення зростає для такої послідовності мов: українська–російська–англійська. Хоча слова в англійській мові коротші, це компенсується більшою кількістю слів у реченні, так що середнє англійське речення містить приблизно на 15 % більше знаків і на 17 % більше літер. Для окремо взятих слів у різних мовах часто немає взаємно однозначної відповідності "кількості інформації" (порівн. "I have taken my keys out of my bag" і "Я витяг ключі з сумки" – 9 проти 5 слів, 26 проти 17 літер і 34 проти 21 знака, відповідно), навіть попри невраховані вище її експліцитну та імпліцитну, а також лінгвістичну й екстралінгвістичну складові. У будь-якому разі, речення видається дещо кращим, аніж слово, мірилом інформації, переданої мовленням. Тоді факт довшого середнього українського

Lviv Polytechnic National University Institutional Repository http://ena.lp.edu.ua

речення суперечить звичним поглядам, за якими англійська “коротша” за українську а чи російську мови. Проте висновок про більшу “стислість” української мови, який напрошується на підставі наших даних, мабуть передчасний і ненадійний – хоча би з огляду на значну, навіть принципову (див. нижче) роль відносних флуктуацій $\Delta l / \bar{l}$.

4. Аналіз та інтерпретація результатів

4.1. Флуктуації частот

Перейдемо до кількісного аналізу одержаних результатів. Найперше зупинимося на залежностях СКВ частоти Δf від середньої частоти f (див. вставки на рис. 1–3). Попри помітні високочастотні викиди, для всіх частот (f_s, f_c і f_w) і всіх корпусів ці залежності добре описуються степеневою функцією:

$$\Delta f = A f^{\gamma}, \tag{4}$$

де A, γ – константи. Таку степеневу залежність флуктуації деякого параметра від його середнього значення називають законом Тейлора (Taylor) [23, 24] – за іменем еколога, що вперше відкрив цю закономірність для флуктуації та середньої кількості особин або біологічних видів.

Параметри лінійної апроксимації залежностей $\Delta f(f)$ у подвійному логарифмічному масштабі наведено в табл. 3. За даними R і SD, стандартні показники якості та точності регресії зростають зі зростанням обсягу статистичної вибірки (розміру корпусу $\sum_i L_i$ – див. дані табл. 1). Це також вияв ефекту скінченних розмірів вибірки. Показники степеня γ мало залежать від одиниць, у яких вимірюють довжину речення, а визначаються переважно мовою: найбільша γ притаманна українській, а найменша – англійській мові. Оскільки ці степені близькі для споріднених слов’янських мов, які мають синтетичний характер, помітно відрізняючись від γ для аналітичної за своїм характером англійської мови, то спокусливим виглядало би припущення про кореляцію показника степеня γ із мірою синтетичності/аналітичності: черговість і порядки величин γ (0,68; 0,65 і 0,58) і справді можна розглядати як наслідок зростання аналітичності мови в послідовності українська–російська–англійська. Проте на підставі тільки наших даних цю гіпотезу довести неможливо. Зокрема не виключено, що параметр γ насправді не залежить (або слабко залежить) від мови, але непрямо визначається розмірами статистичних вибірок (корпусів), послідовність зростання яких така сама. Нарешті, знайдені нами степені γ для довжин речень помітно менші, ніж дані робіт [25, 26], одержані відповідно для флуктуацій лексичного словника та флуктуацій відносних частот літер в українському, російському та англійському корпусах ($\gamma \approx 1$).

Таблиця 3

Параметри лінійної апроксимації залежностей $\Delta f(f)$ на вставках рис. 1–3 для українського, російського і англійського корпусів літературних текстів

Корпус: одиниця довжини речення	Коефіцієнт A	Показник степеня γ	Коефіцієнт детермінації R	Квадратичне відхилення SD
Український:				
знаки	–0,953	0,685	0,979	0,183
літери	–0,951	0,687	0,980	0,177
слова	–0,883	0,682	0,985	0,180
Російський:				
знаки	–1,122	0,636	0,983	0,152
літери	–1,121	0,639	0,983	0,151
слова	–0,893	0,679	0,990	0,143
Англійський:				
знаки	–0,953	0,564	0,987	0,141
літери	–1,122	0,569	0,988	0,134
слова	–1,320	0,601	0,992	0,131

Отже, для українського, російського та англійського корпусів масштабування (або скейлінг) флуктуацій частоти речень зі зміною середньої частоти відповідає законові Тейлора зі степенями $\gamma = 0{,}56 \div 0{,}69$. Зазначимо, що в описі відповідних явищ базовим механізмом зростання незалежної змінної явно чи неявно вважають зміни обсягу вибірки, що слушно для абсолютних частот F, але не застосовно до нашого випадку відносних частот f. Але степеневу залежність $\Delta f(f)$ усе ж можна

Lviv Polytechnic National University Institutional Repository http://ena.lp.edu.ua

трактувати як своєрідний прояв закону Тейлора. Добре відомо (див., наприклад, [27]), що статистика лічби випадкових незалежних подій (у статистичній лінгвістиці випадковість і незалежність подій – випадання деякого слова, речення певної довжини тощо – переважно слугує нульовою гіпотезою), описується розподілом Пуассона. СКВ і середнє значення для нього пов'язані з формулою $\Delta f = \sqrt{f}$, тобто $\gamma = ½$.

У теорії складних систем скейлінг флуктуацій (4) із $\gamma > ½$ вважають аномальним і таким, що вказує на присутність взаємодій або кореляцій елементів системи [24]. На додачу, факт $½ < \gamma < 1$ уповільнює загасання відносних флуктуацій в границі високих частот, яке відбувається за законом $\Delta f / f \propto f^{\gamma-1} \xrightarrow{f\to\infty} 0$ ($\gamma - 1 < 0$). Інакше кажучи, зростання степеня γ робить частоту f принципово флуктувальним параметром, а в границі при $\gamma \to 1$ вона перестає бути "самоусереднюваним" параметром: "макроскопічне" наближення, у якому флуктуації неістотні, стає недосяжним (див. емпіричні дані [25, 26]). Для порівняння, відносні флуктуації в статистичній фізиці (наприклад, флуктуації кількості мікрочастинок у деякому об'ємі) загасають з показником степеня $\gamma - 1 = -½$. Отже, наші висновки загалом підтверджують гіпотезу [9] про важливість опису флуктуацій у формулюванні закономірностей комп'ютерної лінгвістики (див. розділ 1). Це актуалізує вивчення можливих взаємодій і кореляцій у поведінці досліджуваних систем.

4.2. Характер статистичних розподілів f(l)

Спробуємо дослідити, яку аналітичну форму має функція $f(l)$. Хоча ми й працюватимемо з нею як зі статистичним розподілом, тобто із диференційною функцією розподілу ймовірності, надалі все ж уникатимемо терміна "розподіл ймовірності", оскільки вживання останнього передбачає постулювання певної асимптотичної поведінки частоти f із необмеженим зростанням обсягу статистичної вибірки (або самої вимірюваної величини f), а також існування відповідної границі. Проте наведені вище результати для флуктуацій засвідчують, що відповідна поведінка частоти в статистичній лінгвістиці залишається під питанням.

Згідно з загальними міркуваннями авторів праці [14], частота коротких слів або речень зростає зі зростанням l за законами комбінаторики (збільшення можливих комбінацій літер у слові чи слів у реченні) і тому має степеневий характер ($f \propto l^a$, де a – константа), а потім (при $l > l_{\max}$ або $l > \bar{l}$) кількість довгих слів або речень різко зменшується з міркувань зниження економності та ефективності комунікації [14]. Останній фактор найпростіше виразити як $f \propto \exp(-bl)$ (де b – константа). Враховуючи обидва фактори водночас, отримаємо загальний вираз [14]

$$f(l) \propto l^a \exp(-bl) . \qquad (5)$$

Формула (5) відповідає розподілові Ґуда (Good) – дискретному аналогові відомого гамма-розподілу для неперервних випадкових змінних (див. [12]). Його асимптотична поведінка на ділянці "хвоста" (при $l >> \bar{l}$ або $l \to \infty$) експоненційна. Зазначимо, що в формулі (5) і в усіх подальших виразах для статистичних розподілів фігурує лише знак пропорційності "$\propto$", а не знак "=". Це означає, що ми не виписуємо доволі громіздких виразів для постійних множників, які визначаються зі стандартної умови нормування ймовірності $\int f(l)dl = 1$. Ця стала нормування неістотна у процедурі подальшої графічної апроксимації емпіричних даних за формулами на зразок (5).

Хоча з літератури відомі й складніші дискретні розподіли для довжин речень (наприклад, від'ємний біноміальний розподіл [20]), надалі ми будемо використовувати практично зручніші та аналітично простіші неперервні функції розподілу. Так, для випадку довжин слів часто вживають представлення функції $f(l)$ неперервним логнормальним (lognormal) розподілом [1]:

$$f(\log l) \propto \exp[-(\log l - \mu)^2 / 2\sigma^2] , \qquad (6)$$

де m і s – відповідно середнє значення і СКВ для $\log l$. Фактично це розподіл випадкової змінної, логарифм якої описується нормальним розподілом. Відомо також, що нормальний і логнормальний розподіли відповідають адитивному та мультиплікативному стохастичним процесам.

Корисним підходом до нашої проблеми є інтерпретація розподілу довжин речень у тексті мовою розподілу "часів очікування" – часових (чи, цілком еквівалентно, просторових) інтервалів між двома наступними подіями (див. [28–32]). У нашому випадку – це події, що полягають у появі в тексті розділового знаку ("стоп-символу" – наприклад, крапки), який сигналізує про закінчення

Lviv Polytechnic National University Institutional Repository http://ena.lp.edu.ua

речення. Нульовій статистичній гіпотезі (моделі "міху" з лінгвістичними елементами, з якого навгад і незалежними послідовними спробами витягують ці елементи) тоді відповідатиме стохастичний пуассонівський розподіл розділових знаків. Часи очікування цих знаків, тобто розподіл речень за довжинами, для процесу Пуассона описуються геометричним розподілом. Його неперервний аналог – це експоненційний (exponential) розподіл

$$f(l) \propto \exp(-l/\bar{l}), \tag{7}$$

яким коректно користуватися замість геометричного для великих l (при $l >> 1$, $l >> l_{\max}$ або принаймні $l > \bar{l}$), тобто на хвості емпіричного розподілу. Експоненційному розподілові не притаманний "важкий" хвіст, оскільки функція exp(–*l*) швидко загасає для великих *l*. Характерною ознакою цього розподілу і його необхідною (хоча й недостатньою) умовою є рівність $\Delta l = \bar{l}$, яку легко одержати з формули (7). Такому стохастичному режимові формування речень відповідає відсутність жодних "взаємодій" між розділовими знаками або модель "ідеального газу" речень.

Наявність "взаємодій" у формі "відштовхування" стоп-символів означатиме нерівність $\Delta l / \bar{l} < 1$, а в границі дасть детермінований, строго періодичний процес (однакові довжини всіх речень і СКВ $\Delta l = 0$). З іншого боку, "притягання" стоп-символів сприятиме їхньому групуванню в кластери (тобто, значно більшому розкидові довжин речень). Явище кластеризації лінгвістичних елементів ($\Delta l > \bar{l}$ і $\Delta l / \bar{l} > 1$), яке ще називають "пульсаціями" [29, 31, 32], означатиме наявність важкого хвоста у розподілі *f*(*l*). Його переважно описують "розтягнутим" експоненційним (stretched exponential) розподілом [32]

$$f(l) \propto \exp[(-l/\bar{l})^{\beta}] \ (0 < \beta \le 1), \tag{8}$$

або найзагальнішим варіантом розподілу Вейбуля (Weibull) [31]. Опускаючи, як і всюди вище, множник нормування, останній розподіл представляють у вигляді

$$f(l) \propto l^{\beta-1} \exp[(-l/\bar{l})^{\beta}]. \tag{9}$$

Зокрема, при *b* = 1 формули (8) і (9) зводяться до експоненційного розподілу (7), а в границі $\beta \to 0$ із розподілу Вейбуля (9) приходимо до конкретного випадку степеневого розподілу [1] із показником степеня $\alpha = 1 - \beta = 1$. Загалом же стандартний степеневий розподіл описують виразом

$$f(l) \propto l^{-\alpha}, \tag{10}$$

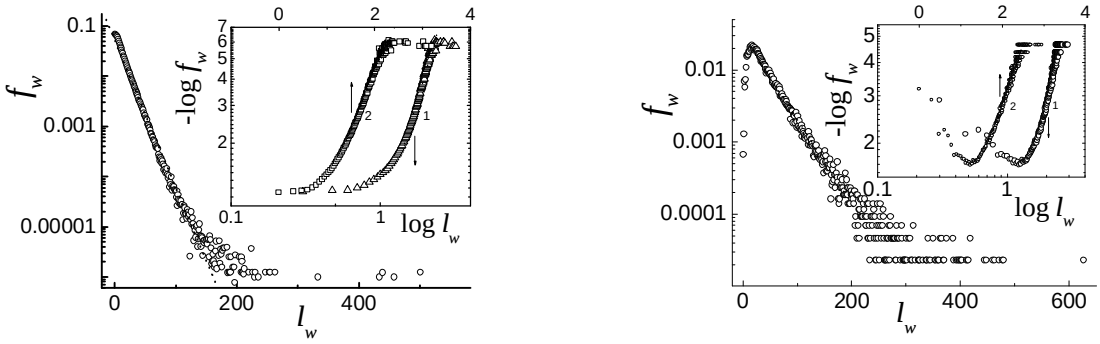
де стала α необов'язково дорівнює одиниці. За умови $\alpha \le 3$, що виконується чи не для всіх емпіричних даних, з якими мають справу дослідники (див. огляд [1]), із формули (10) випливає розбіжність для СКВ випадкової змінної, а за достатньо загальних умов $2 < \alpha \le 3$ одержуємо $\Delta l / \bar{l} \to \infty$. Отже, степеневий розподіл відповідає граничному випадку максимальних пульсацій.

Підсумовуючи всі розглянуті вище теоретичні можливості для розподілів *f*(*l*), наголосимо, що розподіли (5) і (7) характеризуються швидко загасаючими хвостами, а розподіли (8), (9) і (10) мають все важчі хвости. Хвіст же логнормального розподілу принципово "легкий" асимптотично, проте на практиці (за умови аналізу на скінченних інтервалах абсцис) його характер визначається співвідношенням параметрів *m* і *s* ($f(l) \propto l^{-1} \exp[-(\log l - \mu)^2 / 2\sigma^2]$): формула (6) далека від степеневої функції з *a* = 1 або близька до неї, якщо параметр *s* відповідно малий або великий (див. [1]). Якісно це відповідає закономірностям, описаним вище для відношення $\Delta l / \bar{l}$.

Тут і надалі ми будемо цікавитися лише монотонно спадним хвостом емпіричних розподілів (див. дані рис. 1–3 за порівняно великих *l*). Статистичні ж механізми, які формують основну частину розподілів (т. зв. "голову" – ділянку порівняно малих *l*), є дещо іншими (див. обговорення формули (5)). Так, зі статистики часів очікування букв або слів випливає, що зростаюча залежність *f*(*l*) на ділянці голови розподілу зумовлена "відштовхуванням" цих елементів на малих відстанях ($l < \bar{l}$) унаслідок законів граматики або синтаксису [32]. У разі розподілу довжин речень теж можна говорити про своєрідне "відштовхування", яке стосується розділових знаків в областях $l << \bar{l}$ і $l < \bar{l}$. Фактично це відповідає малій імовірності надто коротких речень. Нарешті, найбільш низькочастотну частину хвоста розподілу (наприклад, точки на рис. 1–3 при $l_w > 500$ слів) теж можна відкинути, оскільки на ній домінують помилки оцифрування текстових даних.

Lviv Polytechnic National University Institutional Repository http://ena.lp.edu.ua

Як видно з табл. 2, відносні інтертекстуальні флуктуації $\Delta l / \bar{l}$ для корпусів усіх трьох мов займають діапазон 0,33÷0,45. Якісно це відповідає проміжному режимові між детермінованим і стохастичним характерами формування довжин речень, який далекий від режиму кластеризації. Цікаво порівняти ці дані з інтратекстуальними даними $\Delta l / \bar{l}$. Для функцій $f_w(l)$, притаманних різним текстам з українського корпусу, це відношення змінюється від 0,05 до 3,52, формуючи деякий статистичний розподіл $p(\Delta l / \bar{l})$ із середнім значенням $\Delta l / \bar{l} \approx 0{,}71$ і СКВ 0,25. Середній інтратекстуальний параметр $(\Delta l / \bar{l})_{\text{intra}} \approx 0{,}7$ помітно більший, аніж його інтертекстуальний аналог $(\Delta l / \bar{l})_{\text{inter}} \approx 0{,}4$. Тому функції $f(l)$ для окремих текстів якісно не надто відрізняються од відповідних функцій для корпусів (в обох випадках маємо $\Delta l / \bar{l} < 1$), але їх можна вважати ближчими до експоненційного розподілу (7).



а *б*

Рис. 4. Залежності частоти речень f_w від їхньої довжини l_w, вираженої кількістю слів, для всього українського корпусу (а) і тексту роману У. Самчука "Волинь" (б) у напівлогарифмічному масштабі. На вставках – відповідні залежності $\log f_w$ від $\log l_w$, представлені в подвійному логарифмічному (крива 1) і напівлогарифмічному (крива 2) масштабах. Штриховані прямі лінії відповідають лінійній апроксимації у відповідних діапазонах по осях абсцис

Рис. 4, *а* ілюструє застосування простих графічних методів до апроксимації залежності $f_w(l_w)$ для українського корпусу, а для порівняння рис. 4, *б* відображає залежності $f_w(l_w)$ у різних масштабах для одного з текстів українського корпусу – роману У. Самчука "Волинь", для якого маємо $\Delta l / \bar{l} \approx 0{,}9$. Тут масштаби по осях абсцис і ординат підібрано так, аби лінеаризувати залежності (6)–(8). Наприклад, якщо формула (7) добре описує емпіричні дані, то одержимо пряму лінію в координатах l і $\log f$; якщо формула (8) добре описує емпіричні дані, то вони мали би лягати на пряму лінію в координатах $\log l$ і $\log(\log f)$ і т. ін.

Лінії 1 і 2 на вставках відповідають логнормальному та розтягнутому експоненційному розподілам. Найкраща апроксимація за формулою (6) описується параметрами R = 0,992 для корпусу і 0,985 для тексту, хоча візуально якість її виглядає чи не найгірше (див. лінії 1 на вставках рис. 4). На додаток, ця апроксимація дає коефіцієнти нахилу, які відповідають степеням c = 2,4 і 2,1 біля $\log l$ для корпусу та тексту, замість очікуваного c = 2 (див. формулу (6)). Можливо, дещо краща ситуація з апроксимацією даних за формулою (8): тут маємо R = 0,995, $b \approx 0{,}54$ для корпусу та R = 0,986, b = 0,45 для тексту (див. лінії 2 на вставках рис. 4). Нарешті, експоненційний розподіл (7) на рис. 4 дещо виграє за діапазоном застосовності і, частково, за якістю апроксимації (R = 0,993 для корпусу і 0,991 для тексту). Зазначимо, що "полиці" на рис. 4б для тексту відповідають абсолютним частотам $F_w = 1, 2, \ldots$ і є виявом ефекту скінченних розмірів. Залежностей $f_w(l)$ у подвійному логарифмічному масштабі, які відповідають степеневому розподілові (10), не наведено на рис. 4. Емпіричні залежності $\log f_w$ від $\log l$ виявляють помітну кривизну, коефіцієнти детермінації тут чи не найнижчі (R ≈ 0,97 або 0,98), а одержані степені неправдоподібно великі ($a \approx 5{,}4$ для корпусу та $a \approx 3{,}0$ для тексту). Нарешті, для розподілів Ґуда і Вейбуля (див. формули (5) і (9)) прості графічні методи апроксимації запропонувати важко.

Lviv Polytechnic National University Institutional Repository http://ena.lp.edu.ua

Дані, якісно схожі на дані рис. 4, *а*, одержано для інших корпусів (див. рис. 1–3), а також для інших текстів. Зокрема, для англійського корпусу якість апроксимації за формулами (6)–(8) дещо нижча. Отже, на підставі простих графічних підходів важко однозначно обрати аналітичну формулу, яка би якнайкраще описувала всі експерименти. Можливо, за спаданням якості апроксимації це експоненційний, розтягнутий експоненційний або логнормальний розподіли. Додаткові міркування про типові величини $\Delta l / \bar{l} < 1$ більше схиляють до висновку про експоненційний, а не розтягнутий експоненційний розподіл, хвіст якого важчий. Зазначимо, що тісно пов'язаний з експоненційним гамма-розподіл також добре (R = 0,992) описував залежності $f_w(l)$ для невеликого корпусу англійських текстів у роботі [14].

Значна ширина статистичного розподілу $f(l)$ деякого лінгвістичного елемента в тексті, домінування параметра Δl над $\bar{l}$ і наявність важкого хвоста розподілу, які породжують пульсації, зумовлені значною інтратекстуальною неоднорідністю входження цього елемента в текст та його кластеризацією на окремих ділянках тексту. Всі ці явища сигналізують про деяку семантику, яку переносить лінгвістичний елемент. Наприклад, розподіли часів очікування елементів нижчих рівнів (літер) або функціональних слів, яким бракує семантики, не виявляють важких хвостів і добре описуються експоненційною функцією; кластеризуються лише ключові слова, які визначають семантичне навантаження тексту [32–34]. Отже, за явищем пульсацій повинна стояти наявність і важливість семантичної складової. Оскільки довжину слова важко безпосередньо пов'язати з його змістом [1], мабуть, марно очікувати пульсацій розподілу слів за довжиною. Це підтверджують і емпіричні дані [12, 14]. Міркування про пульсації розподілу речень за довжиною загалом схожі, а тому й гіпотеза про відсутність важкого хвоста розподілу $f(l)$ теж слушна. З іншого боку, перехід від групи коротких до групи довгих речень на різних ділянках тексту можна непрямо пов'язувати з переходом від розмовних до описових або більш офіційних фрагментів тексту, а тому семантичну складову і відповідну кластеризацію стоп-символів у тексті не виключено (див. також примітку). Отже, загальні лінгвістичні міркування, на жаль, не є настільки однозначними та безперечними, аби підтвердити чи заперечити статистичне явище існування важкого хвоста функції $f(l)$. Наші дані підтверджують швидше форму хвоста, близьку до експоненційної.

Нарешті, поняття розподілу часів очікування стоп-символів і можлива кластеризація останніх мають безпосередній зміст лише для інтратекстуальних розподілів $f(l)$. У разі інтертекстуальних характеристик корпусів дані $f(l)$ для окремих текстів усереднюються за корпусом, причому статистичні наслідки такого усереднення і відповідного спотворення початкових розподілів $f(l)$ загалом не очевидні. Інакше, під час переходу від інтратекстуальної функції $f(l)$ до інтратекстуальної додатково відбувається деякий стохастичний процес, характер і природа якого достеменно не відомі. Із наших даних випливає, що, незважаючи на порівняно незначні кількісні відмінності цих двох типів розподілів, функції $f(l)$ для корпусів усе ж більш віддалені від експоненційної, так що їхні хвости дещо легші. Можливо, це наслідок впливу деякого детермінованого періодичного процесу.

5. Заключні зауваження та висновки

Отже, в цій роботі вперше досліджено розподіли f(l) частот речень за їхньою довжиною для великих українського, російського та англійського корпусів. Виявлено дуже істотну асиметрію функцій f(l) відносно значення l із максимальною частотою. Обговорено можливі причини знайдених нерегулярностей на кривих f(l). З'ясовано середні довжини речень в одиницях знаків, літер і слів. Встановлено, що в термінах речення як мірила інформативності комунікації інформаційна ємність українського речення дещо вища за відповідні величини для англійського або російського речень, незалежно від одиниць вимірювань довжини речення.

Обговорено можливі аналітичні форми розподілів f(l) для корпусів і окремих текстів, а також зроблено спробу встановити відповідні параметри за методами графічної лінійної апроксимації.

[1] Узагалі кажучи, це питання потребує докладнішого вивчення (див. працю [21]). Так, лексика розмовної англійської мови та діалогів переважно коротка, а в текстах, наприклад, наукового стилю частіше трапляються довші слова латинського походження. Відповідно, переходові від першого до другого стилю, мабуть, може відповідати деяка семантика, що залишає питання можливості появи важкого хвоста розподілу слів за довжиною відкритим.

Lviv Polytechnic National University Institutional Repository http://ena.lp.edu.ua

Показано, що хвости функцій f(l) задовільно описуються експоненційним, розтягнутим експоненційним або логнормальним розподілами. За підходом до довжин речень як часів очікування між послідовними розділовими знаками, які сигналізують завершення речення, показано, що експоненційна або близькі до неї функції узгоджуються зі стохастичним характером довжини речень, фактом відсутності важкого хвоста розподілів f(l) і обмеженою семантикою, яка стоїть за довжиною речення.

Доведено, що залежність флуктуацій частоти Δf речень різних довжин від середніх значень цієї частоти f визначається степеневим законом Тейлора. Відмінності відповідних степенів γ для української, російської та англійської мов може бути зумовлена різними позиціями цих мов на шкалі синтетичність/аналітичність, хоча не виключено простіше пояснення цих відмінностей як наслідку ефекту скінченних розмірів статистичної вибірки. Відмінність знайдених нами величин $\gamma = 0{,}56 \div 0{,}69$ від класичного значення ½ вказує на аномальний скейлінг флуктуацій, відхилення від стохастичного пуассонівського процесу формування довжин речень у текстах і присутність взаємодій або кореляцій елементів у цих складних лінгвістичних системах.

Виявлені нами значні відносні флуктуації частот і значні відносні зміни $\Delta l / \bar{l}$ середньої довжини речення, які наближаються до 40 % майже незалежно від мови, підтверджують важливість і принциповість урахування флуктуаційних явищ у статистичній лінгвістиці. Навіть якби скейлінг флуктуацій був нормальним ($\gamma \approx ½$), за формулою (4) і нашими даними для параметра A ($A \sim 1$) одержуємо, що відносна флуктуація надто повільно становить зі зростанням f, аби можна було прийти до надійної "макроскопічної" границі: $\Delta f/f$ становить не менше 7 % навіть для таких високих відносних частот, які недосяжні на практиці ($f \sim 0{,}5$). А при $\gamma \approx 0{,}6$ одержуємо $\Delta f/f \sim 12$ %, що ніяк не можна вважати нехтівно малою величиною. Отже, і відсутність взаємодій чи кореляцій у лінгвістичній системі не гарантує стабільного макроусереднення її характеристик. На нашу думку, це пов'язано з принципово "мезоскопічним" характером лінгвістичних систем: навіть у разі порівняно великих текстів ($L \sim 10^6$) або корпусів ($L \sim 10^9$) ці системи безнадійно малі порівняно з типовими макроскопічними системами в статистичній фізиці, які можуть містити ~ 10^{20} або істотно більше частинок. Отже, для мезоскопічних лінгвістичних систем принципово не можна позбутися впливу флуктуаційних явищ.

Описані вище оцінки аналітичної форми статистичних розподілів довжин речень на підставі лінійної апроксимації із залученням графічних даних мали швидше якісний характер. Подальша робота в цьому напрямі передбачає використання складних і трудомістких, але строгих підходів до нелінійної апроксимації даних f(l) (див. також [1, 35]). Більше того, наші результати переконують, що апроксимація обов'язково повинна бути зваженою. З урахуванням величини флуктуацій вагові коефіцієнти w експериментальних точок для частоти f можна обрати у вигляді $w \sim \Delta f^{-1} \sim f^{-g}$.

Крім того, з літератури відомо [31, 33], що виявами кореляцій елементів у лінгвістичній системі є не лише кластеризація та важкі хвости розподілу цих елементів, але й ефект "пам'яті" у послідовності відповідних часів очікування. Останній аналізують за методами флуктуаційного аналізу, що базуються на підході "рандомних прогулянок" (див. [19, 36]). Тому іншим предметом подальшої роботи вбачаємо дослідження довгосяжних кореляцій у розподілах довжин речень для текстів і корпусів.

1. Newman M. E. J. Power laws, Pareto distributions and Zipf's law / M. E. J. Newman // Contemp. Phys. – 2005. – Vol. 46. – P. 323–351. 2. Towards a theory of world length distribution / Wimmer G., Köhler R., Grotjahn R., Altmann G. // J. Quant. Linguist. – 1994. – Vol. 1. – P. 98–106. 3. Forsyth R. S. Feature-finding for text classification */ Forsyth R. S., Holmes D. I. // Literary and Linguistic Computing. – 1996. – Vol. 11. – P. 163–174. 4. / Stamatatos E. Computer-based authorship attribution without lexical measures / Stamatatos E., Fakotakis N., Kokkinakis G. // Computers and the Humanities. – 2001. – Vol. 35. – P. 193–214. 5. Piantadosi S. T. Word lengths are optimized for efficient communication / Piantadosi S. T., Tily H., Gibson E. // Proc. Natl. Acad. Sci. USA. – 2011. – Vol. 108. – P. 3526–3529. 6. Griffiths T. L. Rethinking language: How probabilities shape the words we use / Griffiths T. L. // Proc. Natl. Acad. Sci. USA. – 2011. – Vol. 108, N 10. – P. 3825–3826. 7. Bentz C. Zipf's law of abbreviation as a language universal / Bentz C., Ferrer-i-Cancho R. // Proc. Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics (ed. by C. Bentz et al.). – Режим доступу:*

Lviv Polytechnic National University Institutional Repository http://ena.lp.edu.ua

https://publikationen.uni-tuebingen.de/xmlui/handle/10900/68558. 8. When is Menzerath-Altmann law mathematically trivial? A new approach / Ferrer-i-Cancho R., Hernández-Fernández A., Baixeries J., Dębowski L., Mačutek J. // Statist. Appl. in Genetics and Molecular Biology. – 2014. – Vol. 13, issue 6. – P. 12. 9. Altmann E. G., Gerlach M. Proc. Flow Machines Workshop: Creativity and Universality in Language (Paris, 2014). – arXiv:1502.03296 (2015). 10. Gunning R. The fog index after twenty years/ Gunning R. // J. Business Commun. – 1969. – Vol. 6. – P. 3–13. 11. Yasseri T. A Practical Approach to Language Complexity: A Wikipedia Case Study / Yasseri T., Kornai A., Kertész J. // Plos one. – 2012. – Vol. 7. –P.8. 12. Eeg-Olofsson M. Why is the Good distribution so good? Towards an explanation of word length regularity / Mats Eeg-Olofsson // Working Papers in Linguistics / Lund University, Dept. of Linguistics and Phonetics. – 2008. – Vol. 53. – P. 15–21. 13. Yule U. G. Notes of Karl Pearson's Lectures on the Theory of Statistics, 1884-96 / Yule U. G. // Biometrika. – 1938. – Vol. 30 (1-2). – P. 198-203. 14. Sigurd B., Eeg-Olofsson M., van de Weijer J. // Studia Linguistica. – 2004. – Vol. 58. – P. 37–52. 15. Grzybek P. Exact Methods in the Study of Language and Text / Grzybek P., Stadlober E. ;ed. P. Grzybek, R. Köhler. – Berlin; New York: Mouton de Gruyter, 2007. – P. 205–217. 16. Buk S. Menzerath–Altmann law for syntactic structures in Ukrainian / Buk S., Rovenchak A. // Glottotheory. – 2008. – Vol. 1. – P. 10–17. 17. Grzybek P. The relation between word length and sentence length: An intra-systemtic perspective in the core data structure / Grzybek P., Kelih E., Stadlober E. // Glottometrics. – 2008. – Vol. 16. – P. 111–121. 18. Ausloos M. Punctuation effects in english and esperanto texts / Ausloos M. // Physica A. – 2010. – Vol. 389. – P. 2835–2840. 19. Multifractal Analysis of Sentence Lengths in English Literary Texts / Grabska-Gradzioska I., Kulig A., Kwapieo J., Oświęcimka P., Drożdż S. Proc. 3rd World Conf. on Information Technol. (WCIT-2012), 3-14 December 2012, Dubai, United Arab Emirates. – Dubai,2013. – Vol. 03. – P. 1700–1706. 20. Grzybek P. Methods and Applications of Quantitative Linguistics / Grzybek P.; ed. by I. Obradović, E. Kelih, R. Köhler. – Belgrade: Academic Mind, 2013. – P. 44–58. 21. Ebeling W. Long-range correlations between letters and sen-ences in texts / Ebeling W., Neiman A. // Physica A. – 1995. – Vol. 215. – P. 233–241. 22. Limpert E. Log-normal Distributions across the Sciences: Keys and Clues / Limpert E., Stahel W. A., Abbt M. // BioScience. – 2001. – Vol. 51. – P. 341–352. 23. Taylor L. R. Aggregation, Variance and the Mean // Nature. – 1961. – Vol. 189. – P. 732–735. 24. Eisler Z. Fluctuation scaling in complex systems: Taylor's law and beyond / Eisler Z., Bartos I., Kertész J. // Adv. Phys. – 2008. – Vol. 57. – P. 89–142. 25. Gerlach M. Stochastic Model for the Vocabulary Growth in Natural Languages / Gerlach M., Altmann E. G. // New J. Phys. – 2014. – Vol. 16. – P.19 . 26. Рангові залежності та лексичні частотні спектри для підгруп слів тексту з різними довжинами / О. С. Кушнір, Л. Б. Іваніцький, М. Я. Максисько, С. В. Рихлюк // Мат-ли VII Укр.-польськ. наук.-практ. конф. "Електроніка та інформаційні технології" (ЕлІТ-2015). – Львів : Вид-во Львів. ун-ту, 2015. – Випуск 5. – С. 167-174. 27. Грабовський В. А. Практикум з ядерної фізики: навч. посібн. / Грабовський В. А., Дзендзелюк О. С., Кушнір О. С. – Львів: Видавн. центр ЛНУ ім. І. Франка, 2008. – 222 с. 28. Corral A. Local distributions and rate fluctuations in a unified scaling law for earthquakes / Corral A. // Phys. Rev. E. – 2003. – Vol. 68. –P.4 . 29. Barabási A.-L. The origin of bursts and heavy tails in human dynamics / Barabási A.-L. // Nature. – 2005. – Vol. 435. – P. 207–211. 30. Vázquez A. Exact Results for the Barabási Model of Human Dynamics / Vázquez A. // Phys. Rev. Lett. – 2005. – Vol. 95. – 248701 (4 pp.). 31. Goh K.-I. Burstiness and memory in complex systems / Goh K.-I., Barabá-si A.-L. // Europhys. Lett. – 2008. – Vol. 81, №4. – 48002 (5 pp.). 32. Altmann E. G. Beyond word frequency: Bursts, lulls, and scaling in the temporal distributions of words / Altmann E. G., Pierrehumbert J. B., Motter A. E. // PLOS ONE. – 2009. – Vol. 4, № 11. – e7678 (7 pp.). 33. Altmann E. G. On the origin of long-range correlations in texts / Altmann E. G., Cristadorob G., Esposti M. D. // Proc. Natl. Acad. Sci. USA. – 2012. – Vol. 109. – P. 11582–11587. 34. Про статистику відстаней між словами в тексті та проблему розпізнавання змістових слів / Кушнір О. С., Волоско А. В., Іваніцький Л. Б., Рихлюк С. В. // Мат-ли VII Укр.-польськ. наук.-практ. конф. "Електроніка та інформаційні технології" (ЕлІТ-2016). – Львів : Вид-во Львів. ун-ту, 2016. – Т. 6. – С. 155-164. *35. Perline R. Strong, Weak and False Inverse Power Laws / Perline R. // Statist. Sci. – 2005. – Vol. 20, №1. – P. 68–88. 36. Hřebíček L. Persiscence Other Aspects of Sentence-Length Series / Hřebíček L. // J. Quant. Linguist. – 1997. – Vol. 4, №1–3. – P. 103–109.*

Lviv Polytechnic National University Institutional Repository http://ena.lp.edu.ua