ISSN 2224-087X. Електроніка та інформаційні технології. 2016. Випуск 6. С. 155–164
Electronics and information technologies. 2016. Issue 6. P. 155–164

УДК 004.6, 004.9, 538.9

ПРО СТАТИСТИКУ ВІДСТАНЕЙ МІЖ СЛОВАМИ В ТЕКСТІ ТА ПРОБЛЕМУ РОЗПІЗНАВАННЯ ЗМІСТОВИХ СЛІВ

О. Кушнір, А. Волоско, Л. Іваніцький, С. Рихлюк

Львівський національний університет імені Івана Франка
вул. Ген. Тарнавського, 107, 79017 Львів, Україна
o_kushnir@franko.lviv.ua

У роботі досліджено статистику відстаней між найближчими слововживаннями тих же словоформ в україномовному тексті. Виділено три граничні випадки – стохастичний режим, рівномірний розподіл відстаней і випадок кластеризації слів, які відповідають відсутності лексичних взаємодій, синтаксичному "відштовхуванню" слів і їхньому "притяганню". Розглянуто нульову статистичну гіпотезу, яка відповідає експоненційному розподілу ймовірності відстаней, а також спостережувані відхилення від неї. Доведено, що згадані три граничні випадки описуються "параметром асиметрії" R – відношенням стандартного відхилення відстані до його середнього значення, – яке приблизно дорівнює одиниці, є меншим або більшим за одиницю, відповідно. Показано, що великі значення R сигналізують про ключовий характер слова в тексті, а також проаналізовано переваги і недоліки цього методу розпізнавання змістових слів для україномовних текстів.

Ключові слова: статистичні розподіли, дискретні та неперервні розподіли, комп'ютерна лінгвістика, ключові слова.

Вступ. Статистичні характеристики текстів природними мовами викликають посилений інтерес у галузях пошуку інформації та інтелектуального аналізу даних. Добре відомо, що статистичні закономірності для структурних одиниць тексту непрямо відображають особливості морфології, синтаксису та семантики. Корисними в цьому плані є закони Ціпфа та Гіпса (див., наприклад, [1]), які визначають частотність вживання різних словоформ і лексем. Проте ці закони є недостатнім знаряддям досліджень, оскільки семантична складова слів впливає не лише на їхню частотність, але й на просторове положення цих слів і їхнє взаємне розміщення у тексті.

Завдяки жорстким синтактичним обмеженням на поєднання слів тексти виявляють виразні короткосяжні кореляції, а питання далекосяжних лексичних взаємодій на відстанях, істотно більших за типові довжини речень, потребують окремого вивчення. Корисним тут є аналіз статистики відстаней Δw (в одиницях слів) між найближчими в тексті слововживаннями тої ж словоформи. Відомо, що функціональні (або допоміжні чи службові) слова майже однорідно розподілені в тексті, тоді як просторовий розподіл змістових (або значущих чи ключових) слів неоднорідний і характеризується наявністю просторових флуктуацій у вигляді кластерів [2–6]. Отже, ефект кластеризації вказує на семантичну вагу слова та дає змогу визначити ключові слова даного тексту. Важливо, що такий метод індексування, реферування та встановлення тематики тексту не потребує додаткових референційних текстів.

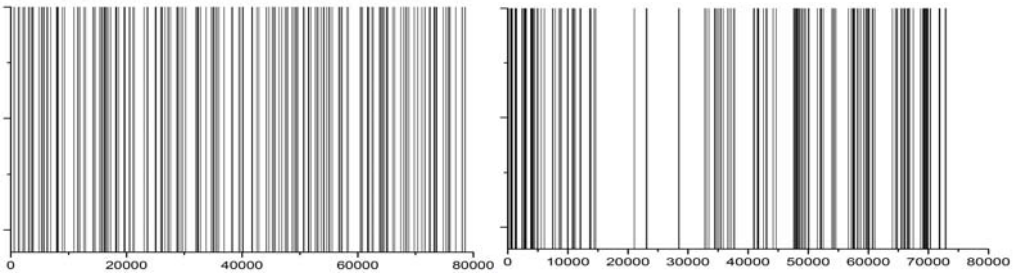
Хоча базові закономірності просторового розподілу слів у текстах відомі, основні успіхи тут стосуються англійської мови. Така ситуація типова для лінгвістики загалом і комп'ютерної лінгвістики зокрема (див., наприклад, зауваження авторів монографії [7]). Водночас, узагальнення встановлених закономірностей на інші мови часто нетривіальне, адже характеристики статистичних розподілів можуть кількісно і навіть якісно залежати від мови, а для української як синтетичної мови слід очікувати додаткових відмінностей, порівняно з аналітичною англійською мовою. Мета цієї роботи – вивчення статистичних розподілів інтервалів між функціональними та змістовими словами та з'ясування на цій підставі можливостей розмежування слів даних груп на прикладі україномовного тексту.

Методика досліджень. Об'єктом дослідження було обрано український переклад повісті Дж. Р. Р. Толкіна "Гобіт" за авторством О. Мокровольського [8]. Перед визначенням частотності слів у тексті з нього було видалено всі розділові знаки; великі та малі літери не розрізнялися. Текст вміщував $L \approx 79{,}9$ тис. слововживань і $V \approx 16{,}9$ тис. словоформ. Усі просторові позиції слів було пронумеровано за черговістю їхньої появи в тексті, а інтервали Δw_i між найближчими слововживаннями даної словоформи визначалися як $\Delta w_i = w_{i+1} - w_i$, де w_i – це позиції (порядкові номери) словоформи. На додаток, для всіх словоформ w ми визначали їхні абсолютні частоти (кількості слововживань) F, відносні частоти f ($f = F/L$) і ранги r (порядкові номери словоформ у списку всіх словоформ за спаданням їхньої частоти). Було визначено масові функції ймовірності $p(\Delta w)$ для статистичних вибірок $\{\Delta w\} = \{\Delta w_1, \Delta w_2, ...\}$. Оскільки статистичні характеристики для слів із низькою абсолютною частотою ненадійні, подальшому аналізові піддавалися лише слова з $F \geq 10$.

Результати та обговорення. Рис. 1 ілюструє закономірності просторових позицій у тексті службового слова *все* і змістового слова (імені одного з персонажів) *Торін*, які мають близькі абсолютні частоти ($F = 200$ і 185, відповідно). Очевидно, що функціональні слова більш чи менш однорідно розподілені в тексті, а змістові слова формують кластери, відстані між якими того ж порядку величини або й більші, ніж типові розміри самих кластерів. Корисним тут є порівняння статистики лексичної системи зі статистикою некорельованих або взаємодіючих енергетичних рівнів у невпорядкованих квантових системах. За цією аналогією, можемо говорити про відсутність взаємодій лексичних одиниць (своєрідна модель "ідеального газу" слів) або наявність цих взаємодій, які реалізуються як "притягання" або "відштовхування" слів [5]. За відсутності взаємодій очікуємо більш чи менш хаотичного розміщення слововживань у тексті, "відштовхування" в граничному випадку приведе до еквідистантного (рівномірного) просторового розподілу слів, а "притягання" сприятиме їхньому групуванню в кластери. Ще одна можлива аналогія – це формування планет шляхом поступової кластеризації протопланетного пилу. Отже, дані рис. 1а і рис. 1б грубо відповідають випадкам відсутності взаємодій лексичних одиниць і наявності "притягання" між ними.

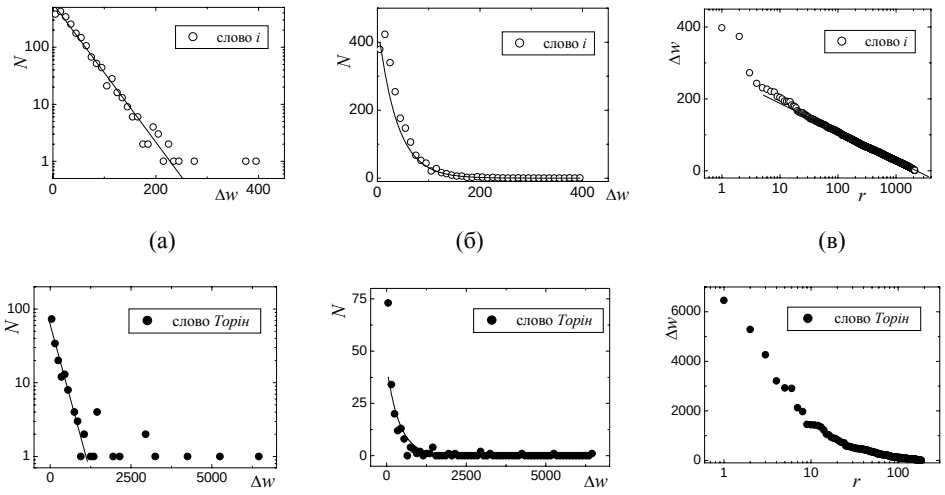
Ці явища відображають і функції розподілу $p(\Delta w)$, приклади яких для функціонального слова *і* та згаданого вище змістового слова *Торін* представлено на рис. 2. Дещо точніше, на рис. 2 фігурують гістограми для абсолютної кількості N значень Δw зі ста-

тистичної вибірки $\{\Delta w\}$, які потрапляють до різних бінів. Залежність $N(\Delta w)$ відповідає масовій функції $p(\Delta w)$ з точністю до сталого множника нормування ($\sum N$). Статистичні параметри деяких слів підсумовано в табл. 1 і 2. Зокрема, ці таблиці містять дані про ранг r і абсолютну та відносну частоти F і f слів, середнє значення ($\overline{\Delta w}$) і стандартне відхилення ($\sigma_{\Delta w} = \sqrt{\overline{\Delta w^2} - (\overline{\Delta w})^2}$) відстаней між цими словами, відповідне "середнє значення" відстані $\overline{\Delta w}'$, розраховане як $\overline{\Delta w}' = 1/f$, а також "параметр асиметрії" розподілу $R = \sigma_{\Delta w} / \overline{\Delta w}$ (див. нижче).



(а) (б)

Рис. 1. Абсолютні позиції слів (а) *все* (F = 200) і (б) *Торін* (F = 185) у тексті повісті "Гобіт" (довжина тексту $L \approx$ 79,9 тис).



(а) (б) (в)

(г) (д) (е)

Рис. 2. Гістограми $N(\Delta w)$ кількості випадань N різних відстаней Δw для слів *і* (а, б) і *Торін* (г, д) у звичайному (б, д) і напівлогарифмічному (а, г) масштабах, а також рангові залежності $\Delta w(r)$ для цих же слів (в, е). Прямі лінії на рисунках (а) і (г) відповідають лінійній апроксимації в обмеженому діапазоні змінної Δw. Неперервні криві на рисунках (б) і (д) розраховані за формулою (1) і значеннями $\overline{\Delta w}$ = 36,9 і 396,9, відповідно. Пряма лінія на рисунку (в) – це лінійна апроксимація для обмеженого діапазону r.

Як видно з рис. 2, функція $p(\Delta w)$ для обох слів спадає зі зростанням відстані Δw, за винятком області найменших Δw для слова *і*. Найімовірніша відстань для слова *і* – це значення $\Delta w_p \approx 15$, для якого маємо максимум функції $N(\Delta w)$. Це значення помітно менше за середнє $\overline{\Delta w} = 36{,}9$. Для слова *Торін* максимуму не спостерігаємо, а екстремум відповідає першому бінові $\Delta w \approx 50$.

Таблиця 1

Деякі статистичні параметри десяти слів з найвищими частотами у тексті та параметри статистичних розподілів відстаней між цими словами (див. текст)

Слово	Ранг *r*	Частота *F*	Відносна частота *f*	Середня відстань $\overline{\Delta w}$	Стандартне відхилення $\sigma_{\Delta w}$	Відношення $R = \sigma_{\Delta w} / \overline{\Delta w}$
і	1	2107	0,0264	36,9	35,8	0,97
не	2	1666	0,0209	46,9	51,0	1,09
й	3	1412	0,0177	55,6	58,3	1,05
на	4	1360	0,0170	57,6	59,1	1,03
що	5	1270	0,0159	61,8	61,4	0,99
з	6	1027	0,0129	76,8	84,5	1,10
в	7	984	0,0123	80,2	85,6	1,07
він	8	895	0,0112	87,8	114	1,30
а	9	805	0,0101	98,2	91,7	0,93
до	10	739	0,0093	107,0	115,1	1,08

Таблиця 2

Деякі статистичні параметри десяти слів з найвищим відношенням $R = \sigma_{\Delta w} / \overline{\Delta w}$ і параметри статистичних розподілів відстаней між цими словами (див. текст)

№	Слово	Ранг *r*	Частота *F*	Відносна частота *f*, 10^{-4}	Відстань $\overline{\Delta w}' = 1/f$	Середня відстань $\overline{\Delta w}$	Стандартне відхилення $\sigma_{\Delta w}$	Відношення $R = \sigma_{\Delta w} / \overline{\Delta w}$
1.	*Гам-Гам*	110	78	9,763	1024	324,7	2399	7,39
2.	*Беорн*	189	48	6,008	1664	1010	4287	4,24
3.	*Орли*	430	22	2,754	3631	2350	9291	3,95
4.	*Гандальф*	55	160	20,00	499,3	495	1894	3,83
5.	*гобліни*	77	102	12,80	783,2	624	2381	3,82
6.	*Берт*	414	22	2,754	3631	297,7	1052	3,53
7.	*Дракон*	126	68	8,512	1175	1095	3857	3,52
8.	*ельфи*	102	81	10,10	986,3	931	3172	3,41
9.	*тролі*	555	17	2,128	4699	1908	6440	3,38
10.	*тролів*	590	16	2,003	4993	4624	15484	3,35

У поясненні та кількісному описі залежностей $p(\Delta w)$ нульовою гіпотезою є відома модель “міху зі словами” (bag-of-words), якій відповідає стохастичний пуасонівський

процес як узагальнення бінарного процесу Бернуллі. Якщо ймовірність p одиничної події (витягування даного слова w із "міху") низька, кількість незалежних спроб необмежено велика, а середня "інтенсивність" подій фіксована, то кількість подій за певний час описуватиметься дискретним розподілом Пуасона (див. [9]). Відповідником часу тут слугує дискретна позиція, на якій з'являється дане слово в тексті (див. опис методики). Тоді "час} між двома послідовними подіями в процесі Пуасона (т. зв. "час очікування") у нас відповідатиме відстані Δw між двома найближчими появами деякого слова в тексті. Добре відомо, що час очікування для процесу Пуасона описується дискретним геометричним розподілом (див., наприклад, [4]). Більш зручним у практичних розрахунках є його неперервний аналог – експоненційний розподіл [*]

$$p(\Delta w) = (1/\overline{\Delta w})\exp(-\Delta w/\overline{\Delta w}), \qquad (1)$$

яким коректно користуватися замість геометричного для великих Δw ($\Delta w >> 1$; принаймні не за умови $\Delta w \sim 1$) і $p \to 0$. Оскільки ймовірність p у нас відповідає частоті f, умову $p \to 0$ надійно дотримано по суті для всіх слів, окрім, можливо, кількох найчастотніших (див. табл. 1 і 2). Тому ми не вважаємо, що перехід від експоненційного до точнішого геометричного розподілу дасть помітний виграш в описі залежності $p(\Delta w)$ [4, 5] і, відповідно, зростання точності методики розділення функціональних і змістових слів (див. нижче).

Непогана точність лінійної апроксимації на гістограмах $N(\Delta w)$, представлених на напівлогарифмічній шкалі $\lg N = f(\Delta w)$ (див. рис. 2а, г), загалом засвідчує якісну придатність наближення (1). Незадовільний опис даних для найбільших аргументів Δw тут не є принциповим недоліком. Він виражає стандартний наслідок ефекту скінченних розмірів тексту: якщо інтервали бінування однакові, то біни для великих Δw на гістограмах недостатньо заповнені. Виходом було би логарифмічне бінування із поступовим розширенням інтервалів бінування [10] або перехід від масової функції розподілу до інтегрального представлення, менш чутливого до флуктуацій, – використання кумулятивної функції розподілу $p_c(\Delta w) = \Pr(\Delta w \le \Delta w_0)$, як у роботі [3].

Неперервні криві на рис. 2б, д описуються формулою (1), єдиний параметр $\overline{\Delta w}$ у якій знайдено не підгонкою шляхом мінімізації суми квадратів відхилень теорія–експеримент, але з незалежних оцінок середніх значень $\overline{\Delta w}$ = 36,9 і 396,9 вибірок $\{\Delta w\}$ для слів *і* та *Торін*, відповідно (див. табл. 1). Ці криві наближено узгоджуються з емпіричними даними, позначеними на рис. 2б, д точками. Проте формула (1) не пояснює максимуму $N(\Delta w)$ в області найменших Δw для слова *і* (див. рис. 2б). Ця відмінність принципова, адже формула (1) базується на припущеннях про відсутність будь-яких взаємодій поміж однаковими словами в тексті і їхнє просторове розміщення, зумовлене стохастичним процесом. З іншого боку, розподіл $p(\Delta w)$ для слова *і* при малих Δw визначається правилами синтаксису, які забороняють надто "тісне" розміщення однакових слів. Відповідно, найімовірніше значення Δw_p є своєрідною "рівноважною" відстанню, так

[*] Автори праці [5] помилково стверджують, що розподіл Пуасона неперервний, а геометричний розподіл є його дискретним аналогом.

що при $\Delta w \le \Delta w_p$ починає діяти сила "відштовхування", внаслідок чого ймовірність $p(\Delta w)$ швидко спадає. Водночас, типові відстані Δw між слововживаннями *Торін* настільки великі ($\overline{\Delta w}$ = 396,9), що напевно не втрапляють до діапазону синтаксичного "відштовхування". Зазначимо, що згаданий діапазон грубо обмежений відстанями $\Delta w \sim \Delta w_p$, хоча за даними роботи [6] вплив синтаксису на розподіл імовірності $p(\Delta w)$ відчутний навіть при $\Delta w \sim 50$. Отже, "голова" розподілу для слова *і* потрапляє до області впливу синтаксису, а для слова *Торін* деяке неузгодження теорія–експеримент для найменших Δw повинне мати іншу природу.

Близькість статистичного розподілу $p(\Delta w)$ для слова *і* до експоненційного непрямо підтверджують дані рис. 2в. Справді, кумулятивний розподіл тоді теж експоненційний, а відповідна рангова залежність $\Delta w(r)$, яка визначається оберненою до нього функцією (див. [11]), повинна бути логарифмічною. Відхилення від апроксимаційної прямої $\Delta w \propto \lg r$ для найбільших значень Δw, швидше за все, є артефактом, зумовленим бідною статистикою (див. рис. 2а, б).

Зазначимо, що рангові залежності частот літер у текстах наближено логарифмічні [**] [13, 14], а тому графеми теж можна вважати наближено невзаємодіючими лінгвістичними елементами (див. виведення [13]). З іншого боку, представлені на рис. 2е дані $\Delta w(r)$ для слова *Торін* засвідчують помітні відхилення від логарифмічної залежності, а тому й відхилення від експоненційного розподілу $p(\Delta w)$. Відповідно, такі слова не можна вважати невзаємодіючими. Нарешті, при $\Delta w \sim \overline{\Delta w}$ експериментальні точки на рис. 2д розташовані нижче за теоретичну криву. Мовою кумулятивної функції розподілу $p_c(\Delta w)$ [3] це означає таке: ймовірність $\Pr(\Delta w \le \overline{\Delta w})$ того, що сусідні слова *Торін* у тексті розміщені на відстанях, менших за середню, є вищою за ймовірність, теоретично передбачену функцією $p_c(\Delta w)$ на підставі експоненційного розподілу. Іншими словами, спостерігаємо кластеризацію та "притягання".

Уведемо "параметр асиметрії" $R = \sigma_{\Delta w} / \overline{\Delta w}$ розподілу $p(\Delta w)$ – відношення його стандартного відхилення до середнього значення. Безпосередня перевірка за континуальними формулами $\overline{\Delta w} = \int_0^\infty \Delta w p(\Delta w) d(\Delta w)$ і $\sigma^2_{\Delta w} = \int_0^\infty (\Delta w - \overline{\Delta w})^2 p(\Delta w) d(\Delta w)$ засвідчує, що для розподілу (1) матимемо $\sigma_{\Delta w} = \overline{\Delta w}$, тобто $R = 1$. Вплив лексичних взаємодій і кластеризації відстаней між словами полягає в змінах статистичних характеристик розподілу $p(\Delta w)$, порівняно з експоненційним розподілом. Один із виявів цього – це модифікація середньої відстані $\overline{\Delta w}$, порівняно зі значенням, яке випливає з тривіального виразу $\overline{\Delta w}' = L/F = 1/f$ для випадку однорідного розподілу слів у тексті. Відмінності $\overline{\Delta w}$ і $\overline{\Delta w}'$ нехтовно малі для найбільш високочастотних функціональних слів (див. табл. 1), проте стають істотними для слів із найбільшими R (див. табл. 2). Для цих слів "довжину

[**] Твердження про те, що залежність $f(r)$ частоти фонем від рангу описується геометричним розподілом [12], насправді стосується кумулятивного розподілу ймовірності $p_c(f)$.

хвилі" $\overline{\Delta w}$, уведену ще Г. К. Ціпфом як середній "час" між найближчими появами слова в тексті (див. [6]), не можна визначати за формулою $\overline{\Delta w} = 1/f$.

За умови кластеризації відхилення $\sigma_{\Delta w}$ випадкової змінної Δw стає більшим за середнє $\overline{\Delta w}$, що добре видно з даних табл. 2. "Хвіст" розподілу тоді розширюється, тобто функція $p(\Delta w)$ затухає повільніше, ніж за законом експоненти (порівн. із формулою (1)). Ці явища можна моделювати [6] переходом до "розтягнутого" експоненційного ("stretched exponential") розподілу або розподілу Вейбуля. "Розтягнутий" експоненційний розподіл описується виразом (див. [15])

$$p(\Delta w) = \overline{\Delta w}^{-1}\Gamma^{-1}((\beta+1)/\beta)\exp[(-\Delta w/\overline{\Delta w})^{\beta}], \tag{2}$$

де $0 < \beta \le 1$, $\Gamma(x)$ – гамма-функція, а розподіл Вейбуля визначають як

$$p(\Delta w) \propto \beta\overline{\Delta w}^{-\beta}\Delta w^{\beta-1}\exp[(-\Delta w/\overline{\Delta w})^{\beta}]. \tag{3}$$

Граничний випадок експоненційного розподілу в формулах (2) і (3) маємо за умови $\beta = 1$, причому розподіл (2) є кумулятивною функцією розподілу (3). Зауважимо, що вирази (2) і (3) не єдино можливі в кількісному описі залежностей $p(\Delta w)$ [11]. Визначальним є факт підвищення "асиметрії" розподілу, де асиметрію розуміємо не в звичному сенсі (як відмінність розподілу справа та зліва від середнього значення випадкової величини), а як неоднаковість стандартного відхилення і середнього. Знайдемо, як приклад, відповідне відношення R для простішого випадку розподілу (2):

$$R^2 = \frac{\Gamma(3/\beta)\Gamma(1/\beta)}{\Gamma^2(2/\beta)} - 1. \tag{4}$$

Із відомих властивостей гамма-функції випливає, що для "розтягнутого" експоненційного розподілу справді $R \ge 1$, а рівність $R = 1$ відповідає границі $\beta \to 1$.

Отже, умова $R \approx 1$ відсутності лексичних взаємодій відповідає функціональним словам (див. табл. 1), умова $R > 1$ описує кластеризацію, притаманну змістовим словам (див. табл. 2), а $R < 1$ – це випадок синтаксичного "відштовхування" слів на коротких відстанях ($\Delta w \approx \overline{\Delta w}$ і $\sigma_{\Delta w} \to 0$; частковим прикладом може слугувати службове слово *і* в табл. 1). У роботі [3] було введено стандартне відхилення $\sigma_z = \sqrt{\overline{z^2} - \overline{z}^2}$ нормованої відстані між словами $z = \Delta w/\overline{\Delta w}$ як кількісний показник змістової значущості слова. З урахуванням зв'язку змінних z і Δw одержуємо $\sigma_z = \sigma_{\Delta w}/\overline{\Delta w}$, тобто $\sigma_z = R$. Проте сам факт кореляції явища кластеризації та зростання σ_z у роботі [3] по суті постулювався, а відповідні статистичні причини не з'ясовувалися.

Хоча окремі автори критикують часто суб'єктивні спроби "рукотворного" створення покажчиків ключових слів [16], ми таки створили відповідний покажчик для якісного дослідження можливостей автоматичного розпізнавання змістових слів у тексті за параметром R (див. також [3]). На рис. 3а представлено залежність параметра R від відносної частоти для функціональних і змістових слів. Загалом функціональні слова

(720 одиниць із $F \geq 10$) згруповані навколо рівня $R \approx 1$, а змістовим словам (265 одиниць) притаманні помітно більші значення *R*. Це додатково ілюструє рис. 3б, де показано ненормовані розподіли ймовірності параметра *R*. Для груп функціональних і змістових слів маємо відповідно $\overline{R} \approx 1{,}07$ (стандартне відхилення 0,26) і $\overline{R} \approx 1{,}64$ (стандартне відхилення 0,74). Відзначимо істотну асиметрію обох розподілів імовірності $p(R)$ із їхнім "затягуванням" у бік більших значень *R*, а також більшу ширину розподілу $p(R)$ для змістових слів.

Звісно, що в автоматичному розпізнаванні ключових слів до критерію величини *R* слід ще додати фільтрування порівняно низьких частот. Перевагою методики є відсутність потреби в розлогих "стоп-списках" службових слів, розташованих у "голові" ціпфівського розподілу (див. [17]). Недоліком є недостатня надійність методики, що зрозуміло хоча б із факту істотного перекривання розподілів $p(R)$ для обох типів слів (див. рис. 3б). Яскравою ілюстрацією є труднощі беззастережного віднесення імені головного героя *Більбо* ($r = 15$, $F = 532$, $R = 1{,}43$) до ключових слів.

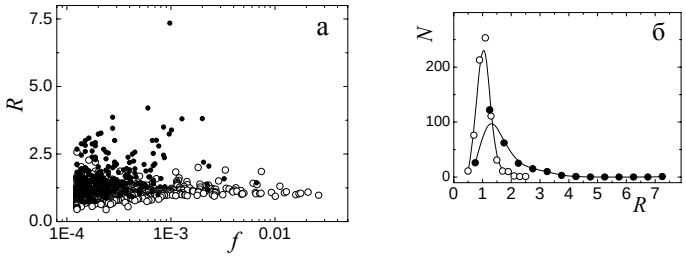


Рис. 3. (а) Залежності параметра асиметрії *R* для функціональних і змістових слів від їхньої частоти. (б) Гістограми ненормованого розподілу ймовірності параметра *R*. Символи o і • відповідають функціональним і змістовим словам.

Хоча подібні проблеми вже висвітлено в літературі [2–5], у разі україномовних текстів з'являються додаткові труднощі. Наприклад, словоформи *тролі* та *тролів* (див. табл. 2) є різними граматичними формами одного й того ж слова, проте практичні можливості об'єднаного аналізу їхньої статистики примарні. Справді, програмні пакети для автоматизованої лематизації української лексики, наскільки відомо авторам, досі відсутні, а "розпорошування" статистичних даних (наприклад, змістове слово *гоблін*, представлене в табл. 2, трапляється в тексті в дев'яти відмінках однини і множини) призводить до збіднення відповідної статистики та помилок інтерпретації. Можливо, саме синтетичний характер української мови є причиною того, що зареєстровані нами максимальні величини параметра *R* для перших двадцяти слів ($2{,}8 \div 7{,}4$) помітно менші за дані роботи [3], здобуті для тексту з не надто відмінними розмірами ($7{,}4 \div 24{,}2$). Одним із можливих виходів був би аналіз не для слів, а для N-грам (див. [5, 16]), хоча істотним недоліком цього підходу залишається різке зростання кількості необхідних обчислювальних операцій.

Висновки. Отже, у цій роботі з'ясовано можливості розрізнення змістових і функціональних слів в україномовних текстах на підставі методики дослідження функції розподілу ймовірності відстаней між найближчими слововживаннями. Як критерій семантичної значущості слова можна використовувати "параметр асиметрії" *R* розподілу відстаней між його найближчими слововживаннями – відмінність стандартного відхилення згаданої відстані від її середнього значення. На відміну від аналогічного

параметра – стандартного відхилення σ_z для нормованої відстані, відомого в літературі, – введений нами параметр R має прозору статистичну інтерпретацію. У роботі вказано на переваги методики розпізнавання ключових слів, її недоліки та шляхи їхнього усунення для випадку синтетичної української мови.

На завершення зазначимо, що припущення про однорідність розподілу окремих слів у тексті та неістотність флуктуацій їхньої частоти є принциповими і лежать в основі деяких підходів статистичної лінгвістики (див. [18, 19]). Як видно з результатів наших досліджень, це припущення надійне лише для високочастотних функціональних слів.

Цікавим об'єктом подальших досліджень є порівняння згаданих вище статистичних закономірностей із відповідною статистикою для рандомних і рандомізованих текстів різних типів, а також порівняльний аналіз статистики $p(\Delta w)$ для різних частин мови.

Автори висловлюють вдячність асист. каф. перекладознавства та контрастивної лінвістики ЛНУ Кушнір Л. О. за практичну допомогу та корисне обговорення даних.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. *Altmann E. G.* Statistical laws in linguistics / Altmann E. G., Gerlach M. // arXiv:1502.03296 [physics.soc-ph] 2015.
2. *Montemurro M. A.* Entropic analysis of the role of words in literary texts / Montemurro M. A., Zanette D. H. // Adv. Complex Systems. – 2002. – Vol. 05. – P. 7–17.
3. *Ortuno M.* Keyword detection in natural languages and DNA / Ortuno M., Carpena P., Bernaola-Galvan P., Munoz E., Somoza A. M. // Europhys. Lett. – 2002. – Vol. 57. – P. 759–764.
4. *Herrera J. P.* Statistical keyword detection in literary corpora / Herrera J. P., Pury P. A. // European Phys. J. – 2008. – Vol. 63. – P. 135–146.
5. *Carpena P.* Level statistics of words: finding keywords in literary texts and symbolic sequences / Carpena P., Bernaola-Galván P., Hackenberg M., Coronado A. V., Oliver J. L. // Phys. Rev. E. – 2009. – Vol. 79. – P. 035102(R).
6. *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. – P. e7678.
7. *Bolshakov I.* Computational linguistics. Models, resources, applications / Bolshakov I., Gelbukh A. – Mexico : Ciencia de la Computacion, 2004. – 198 p.
8. *Толкієн Дж. Р. Р.* Гобіт, або Мандрівка за Імлисті гори / З англ. переклав О. Мокровольський. – К. : Школа, 2002. – 352 с.
9. *Грабовський В. А.* Практикум з ядерної фізики: Навчальний посібник / Грабовський В. А., Дзендзелюк О. С., Кушнір О. С. – Львів: Видавн. центр ЛНУ імені Івана Франка, 2008. – 222 с.
10. *Newman M. E. J.* Power laws, Pareto distributions and Zipf's law / Newman M. E. J. // Contemporary Phys. – 2005. – Vol. 46. – P. 323–351.
11. *Li W.* Fitting ranked linguistic data with two-parameter functions / Li W., Miramontes P., Cocho G. // Entropy. – 2010. – Vol. 12. – P. 1743–1764.
12. *Mačutek J.* Discrete and continuous modelling in quantitative linguistics / Mačutek J., Altmann G. // J. Quant. Linguist. – 2007. – Vol. 14. – P. 81–94.
13. *Gusein-Zade S. M.* Frequency distribution of letters in the Russian language / Gusein-Zade S. M. // Probl. Peredachi Inform. – 1988. – Vol. 24. – P. 102–107.

14. *Kanter I.*, *Kessler D. A.* Markov processes: linguistics and Zipf's law / Kanter I., Kessler D. A. // Phys. Rev. Lett. – 1995. – Vol. 74. – P. 4559–4562.
15. *Grzywacz N. M.* Statistics of optical coherence tomography data from human retina / *Grzywacz N. M.*, *de Juan J.*, *Ferrone C.*, *Giannini D.*, *Huang D.*, *Koch G.*, *Russo V.*, *Ou Tan*, *Bruni C.* // IEEE Trans. Med. Imaging. – 2010. – Vol. 29. – P. 1224–1237.
16. *Damashek M.* Gauging similarity with n-grams: language-independent categorization of text / Damashek M. // Science. – 1995. – Vol. 267. – P. 843–848.
17. *Luhn H. P.* The automatic creation of literature abstracts / Luhn H. P. // IBM J. Res. Dev. – 1958. – Vol. 2. – P. 159–165.
18. *Bernhardsson S.* The meta book and size-dependent properties of written language / Bernhardsson S., Correa da Rocha L. E., Minnhagen P. // New J. Phys. – 2009. – Vol. 11. – 123015.
19. *Bernhardsson S.* Size-dependent word frequencies and translational invariance of books / Bernhardsson S., Correa da Rocha L. E., Minnhagen P. // Physica A. – 2010. – Vol. 389. – P. 330–341.

Стаття: надійшла до редакції 02.03.2016,
доопрацьована 10.03.2016,
прийнята до друку 16.03.2016.

ON THE STATISTICS OF INTER-WORD DISTANCES AND THE PROBLEM OF RECOGNITION OF CONTENT WORDS

O. Kushnir, A. Volosko, L. Ivanitskyi, S. Rykhlyuk

Ivan Franko National University of Lviv
107 Tarnavsky St., UA–79017 Lviv, Ukraine
o_kushnir@franko.lviv.ua

In this work we have studied the statistics of distances between the nearest word tokens of the same word types in a Ukrainian text. Three limiting cases have been distinguished – a stochastic regime, a uniform distribution of the distances, and a case of word clustering – which correspond to a lack of lexical interactions, a syntactic 'repulsion' of words, and their 'attraction'. A null statistical hypothesis has been considered, which corresponds to the exponential probability distribution of those distances, and deviations from that hypothesis observed empirically have been analyzed. It has been proved that the three limiting cases mentioned above can be described by the 'asymmetry parameter' R – the ratio of the standard deviation of the distances to their average value – which is roughly equal to one, less than one or greater than one, respectively. We have shown that large R values point to keywords in the text. The advantages and drawbacks of this method for recognizing content words have been analyzed for the Ukrainian texts.

Key words: statistical distributions, discrete and continuous distributions, computational linguistics, keywords.