УДК 811.161.2’33:519.25

І. М. Кульчицький
Національний університет “Львівська політехніка”,
кафедра прикладної лінгвістики

ВИБІР РОЗМІРУ ВИБІРКИ ДЛЯ СТАТИСТИЧНИХ ОПРАЦЮВАНЬ ТЕКСТІВ

Працю присвячено одному із важливих напрямків квантитативних досліджень мови та мовлення – вивченню інформаційно-статистичних властивостей тексту. Здійснено спробу встановити для творів Марка Черемшини відсоток авторських текстів, який достатній для аналізу вірогідних відносних частот символів у його творах та дослідити стійкість цих частот. Зроблено низку висновків про розмір уривків тексту, з яких формується текст-вибірка для статистичних обстежень.

Ключові слова: квантитативні дослідження, вибірка, обсяг вибірки, частота, Марко Черемшина.

The article is dedicated to one of the most important areas of quantitative studies of language and speech that is the study of information and statistical properties of text. An attempt was made to establish Cheremshyna’s literary works percentage sufficient to analyze relative frequencies of characters in his works and to investigate the stability of these frequencies. A number of conclusions were made about the size of text passages which may form text-sample for statistical surveys.

Key words: quantitative study, sample, sample size, frequency, Marko Cheremshyna.

Вступ

Застосування квантитативних методів у лінгвістиці значно розширює та модифікує наші знання як про саму мовну систему, так і про можливості її функціонування [5]. Використовують їх під час дослідження явищ мови та мовлення у трьох умовних напрямках [13]:

- отримання різноманітних кількісних відомостей;
- використання методів теорії ймовірності для побудови лінгвістичних моделей, здебільшого з використанням даних попереднього напрямку;
- статистична перевірка гіпотез про ті чи інші явища.

З огляду на це, вивчення способів та методів кількісних досліджень у лінгвістиці і на сьогоднішній день залишаються актуальними.

Загальна постановка проблеми

Один із важливих напрямків квантитативних досліджень мови та мовлення – вивчення інформаційно-статистичних властивостей тексту [8; 16]. Їх широко використовують під час атрибуції текстів [10; 14], дешифрування [1] тощо. З іншого боку, для будь-якого статистичного обстеження тексту важливо правильно вибрати спосіб та величину вибірки [12, 27, 28]. Мета цього дослідження — встановити для творів Марка Черемшини відсоток авторських текстів, який достатній для визначення вірогідних відносних частот символів у його творах та дослідити стійкість цих частот. Стосовно останнього зазначимо, що літературні джерела, наприклад [1; 2], стверджують про стабільність та характерність відносних частот букв у текстах для кожної мови.

Аналіз досліджень та публікацій

Хоча використання математичних методів у лінгвістиці передбачили ще Ф. де Соссюр та І. А. Бодуен де Куртене, фактично застосовувати їх почали з середини минулого століття [6, 95].

Дослідження в цьому напрямі проводили [3; 4] такі зарубіжні вчені, як Ґабріель Альтман (Gabriel Altmann), Рейнгард Кьолер (Reinhard Köhler) (Німеччина); Петер Ґжибек (Peter Grzybek), Еммеріх Келіх (Emmerich Kelih) (Австрія); Гейза Віммер (Geiza Wimmer) (Словаччина); Адам Павловскі (Adam Pawłowski), Ядвіґа Самбор (Jadwiga Sambor) (Польща); Юхан Тулдава (Естонія); Раймунд Піотровский, Анатолій Шайкевич (Росія) та інші. В Україні квантитативними дослідженнями мовних явищ займаються Володимир Широков [16], Максим Кригін [8], Валентина Перебийніс [12], Соломія Бук [4] та ін.

Аналіз наукових результатів

Вибір та організація матеріалу дослідження

Дослідницьку збірку текстів створено на основі повного зібрання творів Марка Черемшини 1937 р. [15]. Матеріал було вибрано й організовано, враховуючи такі припущення [8, 12, 16, с. 109–116; 27, 28]. До статистичних обстежень зазвичай із різних причин залучають не всі тексти творів письменника, а вибрані з них окремі уривки. Величина уривків та спосіб їх вибору впливають на результати статистичного дослідження і залежать від його мети. Здебільшого їх вибирають емпіричним шляхом [12, 28]. У пропонованому дослідженні основний об’єкт – відносна частота букв української абетки. Позаяк в українських текстах окрім букв у їхньому статусі використовують знаки дефісу, апострофа та пробілу (останній ділить текст на слова), то під час обчислень тексти творів Марка Черемшини інтерпретували як множину символів розширеної української абетки, до якої окрім її букв додано знаки апострофа, дефіса та пробілу. Враховуючи, що створення текстів-вибірок у дослідженні передбачено програмно, розмір абзаца співвіднесено з розміром уривку тексту. Матеріал підготовлено послідовним виконанням таких кроків.

Крок 1. Тексти творів повного видання [15] перетворено на електронну форму та нормалізовано [9]. Список творів та довжину кожного з них у символах подано у табл. 1–3. Для цих творів визначено відносні частоти символів розширеного алфавіту (табл. 4). Ці завдання в межах кваліфікаційної праці виконала студентка кафедри прикладної лінгвістики Львівської політехніки 2014 року випуску Христина Созанська [17].

Таблиця 1

Кількість символів у творах М. Черемшини першого тому повного видання

Назва	Кількість символів	Назва	Кількість символів
Муха	420	Дід	3618
Сумно одинокому	428	На-боже	4692
Ледові квіти	542	Святий Николай у гарті	5785
Плач квітів	660	Сльоза	6118
Заморожені фіялки	696	Лік	6535
Обнова	722	Нечаяна смерть	7148
Море	820	Грушка	7354
Ґердан	840	Чічка	8348
Верба	873	Злодія зловили	8990
Симфонія	1077	Хіба даруймо воду	10110
Весна	1099	Керманич	11836
Щоб не тії гори	1152	Карби	12781
Осінь	1677	Основини	15485
Горнець	2293	Раз мати родила	16308
Зведениця	2389	Більмо	18807
Бабин хід	3247		

Таблиця 2

Кількість символів у творах М. Черемшини другого тому повного видання

Назва	Кількість символів	Назва	Кількість символів
Желання	501	Зрадник	6502
Під осінь	699	На Купала-на Івана	7533
Стефаникові мужики	1157	Ласка	7535
Колядникам науки	1337	Бо як дим підоймається	10276
Село потерпає	3480	Добрий вечір пане брате	11310
Після бою	3620	Перші стріли	12746
Йордан	3869	Село вигибає	22639
Його кров	4336	Поменник	27718
Туга	4754	Бодай їм путь пропала	30947
Коляда	5853	Верховина	37048
Писанки	6122		

Таблиця 3

Кількість символів у творах М. Черемшини третього тому повного видання

Назва	Кількість символів	Назва	Кількість символів
Параска	681	Інвалідка	3723
Та не думай Марусенько	753	Зарікайся мід-горівку пити	5103
Недописані книги	857	Парубоцька справа	9687
Сокіл	1098	Козак	10305
Хитрість	2472	Парасочка	10512
Анюта	2868	Марічку головка болить	15085
Любість	3376	За мачуху молоденьку	22644

Крок 2. З отриманих на попередньому кроці текстів утворено п’ять однакових дослідницьких текстів-репрезентантів завдовжки близько 470 тисяч символів, які відрізнялись лише розмірами абзаців (близько 100, 200, 300, 400 та 500 символів відповідно). Розмір абзаців обрано довільно. При цьому дотримано таких правил [16]:

- усі малі букви українського алфавіту замінено великими;
- символи тексту, що не входять до розширеної абетки, замінено пробілом;
- між словами залишено лише один пробіл;
- текст поділено на абзаци фіксованої довжини з точністю до слова, тобто якщо після додавання чергового слова до абзацу його довжина ставала більшою від наперед заданої, то слово не обрізали й залишали абзац більшим на декілька символів;
- під час обчислення символ кінця абзацу рахують як пробіл.

Організація та результати дослідження

Дослідження проводили на персональному комп’ютері класу IBM PC під управлінням операційної системи Windows у програмних середовищах Python та MS Office.

Теоретичним підґрунтям основної частини дослідження був критерій згоди К. Пірсона (χ^2) [11]. За гіпотетичну теоретичну функцію розподілу прийнято частотний розподіл символів у дослідницькій збірці (таб. 4). Експериментальні тексти-вибірки отримано з тексту-репрезентанта випадковим вибором абзаців. За нульову гіпотезу H_0 прийнято твердження: “у тексті-вибірці

розподіл частот символів розширеної української абетки не відрізняється від відповідного розподілу в тексті-репрезентанті". Для її перевірки виконано такі кроки:

- для тексту-вибірки обчислено абсолютну частоту символів розширеної абетки;
- отримані частоти використано для обчислення статистики критерію $\chi^2_{експ.}$;
- $t_{кр} = \chi^2_{1-\alpha,k-1} = 49.802$ визначали за відповідною таблицею [11] за рівня значущості α=0,05 та ступенях свободи для k=36 (кількість символів розширеної української абетки);
- якщо отримували, що $\chi^2_{âêñïï.} \geq t_{êð}$, то гіпотезу відхиляли, інакше її приймали.

Паралельно для тексту-вибірки визначено ранг кожного символу в частотному розподілі.

Таблиця 4

Кількість та частота символів у творах Марка Черемшини

Символ	Кількість	Частота	Символ	Кількість	Частота	Символ	Кількість	Частота
А	37232	0,079780	Ї	2352	0,005040	Ф	0,000936	0,000936
Б	8498	0,018209	Й	6591	0,014123	Х	0,008421	0,008421
В	19357	0,041478	К	15514	0,033243	Ц	0,005850	0,005850
Г	6526	0,013984	Л	15598	0,033423	Ч	0,010774	0,010774
Ґ	698	0,001496	М	11288	0,024188	Ш	0,008554	0,008554
Д	14774	0,031657	Н	20362	0,043631	Щ	0,003306	0,003306
Е	18537	0,039721	О	33540	0,071869	Ь	0,012038	0,012038
Є	3333	0,007142	П	10473	0,022441	Ю	0,008183	0,008183
Ж	3340	0,007157	Р	16073	0,034441	Я	0,014828	0,014828
З	8369	0,017933	С	15130	0,032420	Апостроф	209	0,000448
И	27234	0,058356	Т	20166	0,043211	Дефіс	790	0,001693
І	20966	0,044925	У	13702	0,029360	Пробіл	82015	0,175740

Для кожного тексту-репрезентанта з відповідним розміром абзацу проводили ті самі дослідження у два етапи. На першому етапі почергово задавали величину тексту-вибірки від 2 % до 98 % розміру тексту-репрезентанта з кроком у 2 % та визначали необхідну приблизну кількість абзаців, щоби отримати вибірку заданого розміру. За отриманими параметрами 2000 разів генерували текст-вибірку заданого розміру. Кожний отриманий текст перевіряли вищеописаним способом на гіпотезу H_0. Фіксували прийняття чи відхилення гіпотези, величину $\chi^2_{експ.}$ та частотний ранг кожного символу. Результати перевірки гіпотези для різних розмірів вибірки та довжин уривків (абзаців) подано в табл. 5.

На другому етапі уточнено розмір тексту-вибірки, для якого гіпотезу H_0 буде прийнято завжди. З цією метою для кожного тексту-репрезентанта на основі результатів першого етапу визначено діапазон розмірів вибірок, для яких прийняття гіпотези переходить з категорії "майже завжди" в категорію "завжди". Для кожного діапазону повторено перший етап, де крок зміни розміру вибірки визначався як 0,05 від різниці між максимальним і мінімальним значенням, а кількість генерацій тексту-вибірки кожного розміру дорівнює 5000. Результати подано в табл. 6. Графа "Розмір вибірки" подає обсяг тексту-вибірки у відсотках до розміру тексту-репрезентанта, графа "Збіг" – відсоток випадків з 5000 тисяч експериментів, коли гіпотезу було прийнято.

На завершення дослідження підсумовано результати фіксації частотного рангу символів розширеного алфавіту для всіх генерованих текстів-вибірок. Загалом їх було згенеровано 965000. Результатний ранг кожного символу встановлено за тим місцем, яке він займав найбільшу кількість разів (табл. 7).

Таблиця 5

Перевірка гіпотези H_0 для вибірок різни обсягів та різних довжин уривків тексту

№	Розмір вибірки у відсотках до розміру тексту-репрезентанта	Для гіпотези кількість									
		Прийняття					Відхилення				
		Розмір уривку тексту					Розмір уривку тексту				
		100	200	300	400	500	100	200	300	400	500
1	2	3	4	5	6	7	8	9	10	11	12
1	2 %	1748	1487	1323	1080	951	252	513	677	920	1049
2	4 %	1796	1572	1351	1175	1047	204	428	649	825	953
3	6 %	1874	1624	1426	1238	1120	126	376	574	762	880
4	8 %	1856	1633	1496	1331	1202	144	367	504	669	798
5	10 %	1870	1711	1548	1367	1249	130	289	452	633	751
6	12 %	1903	1754	1607	1450	1280	97	246	393	550	720
7	14 %	1916	1778	1644	1512	1338	84	222	356	488	662
8	16 %	1928	1825	1667	1534	1436	72	175	333	466	564
9	18 %	1955	1853	1719	1630	1471	45	147	281	370	529
10	20 %	1965	1866	1779	1627	1543	35	134	221	373	457
11	22 %	1974	1909	1798	1692	1562	26	91	202	308	438
12	24 %	1992	1935	1850	1742	1662	8	65	150	258	338
13	26 %	1988	1940	1873	1784	1701	12	60	127	216	299
14	28 %	1987	1959	1893	1826	1743	13	41	107	174	257
15	30 %	1991	1970	1920	1865	1804	9	30	80	135	196
16	32 %	1997	1980	1930	1878	1822	3	20	70	122	178
17	34 %	1999	1986	1953	1914	1856	1	14	47	86	144
18	36 %	1998	1988	1966	1922	1886	2	12	34	78	114
19	38 %	1998	1992	1963	1933	1913	2	8	37	67	87
20	40 %	1998	1992	1982	1970	1933	2	8	18	30	67
21	42 %	2000	1998	1986	1980	1936	0	2	14	20	64
22	44 %	2000	1998	1992	1990	1964	0	2	8	10	36
23	46 %	1999	2000	1995	1983	1978	1	0	5	17	22
24	48 %	2000	2000	1997	1983	1986	0	0	3	17	14
25	50 %	2000	2000	1999	1998	1992	0	0	1	2	8
26	52 %	2000	2000	1999	1996	1987	0	0	1	4	13
27	54 %	2000	2000	1999	1998	1996	0	0	1	2	4
28	56 %	2000	2000	2000	1997	2000	0	0	0	3	0
29	58 %	2000	2000	2000	1999	1999	0	0	0	1	1
30	60 %	2000	2000	1999	2000	1998	0	0	1	0	2
31	62 %	2000	2000	2000	1999	2000	0	0	0	1	0
32	64 %	2000	2000	2000	2000	2000	0	0	0	0	0
33	66 %	2000	2000	2000	2000	2000	0	0	0	0	0
34	68 %	2000	2000	2000	2000	2000	0	0	0	0	0
35	70 %	2000	2000	2000	2000	2000	0	0	0	0	0
36	72 %	2000	2000	2000	2000	2000	0	0	0	0	0
37	74 %	2000	2000	2000	2000	2000	0	0	0	0	0

Продовження табл. 5

1	2	3	4	5	6	7	8	9	10	11	12
38	76 %	2000	2000	2000	2000	2000	0	0	0	0	0
39	78 %	2000	2000	2000	2000	2000	0	0	0	0	0
40	80 %	2000	2000	2000	2000	2000	0	0	0	0	0
41	82 %	2000	2000	2000	2000	2000	0	0	0	0	0
42	84 %	2000	2000	2000	2000	2000	0	0	0	0	0
43	86 %	2000	2000	2000	2000	2000	0	0	0	0	0
44	88 %	2000	2000	2000	2000	2000	0	0	0	0	0
45	90 %	2000	2000	2000	2000	2000	0	0	0	0	0
46	92 %	2000	2000	2000	2000	2000	0	0	0	0	0
47	94 %	2000	2000	2000	2000	2000	0	0	0	0	0
48	96 %	2000	2000	2000	2000	2000	0	0	0	0	0
49	98 %	2000	2000	2000	2000	2000	0	0	0	0	0

Таблиця 6

Перевірка гіпотези H_0 для вибірок уточнених розмірів та різних довжин уривків тексту

Розмір уривку тексту									
100		200		300		400		500	
Розмір вибірки	Збіг	Розмір вибірки	Збіг	Розмір вибірки	Збіг	Розмір вибірки	Збіг	Розмір вибірки	Збіг
29,4 %	99,50 %	35,1 %	99,26 %	41,4 %	99,28 %	45,3 %	99,14 %	49,1 %	99,3 %
30,8 %	99,76 %	36,2 %	99,42 %	42,8 %	99,52 %	46,6 %	99,28 %	50,2 %	99,4 %
32,2 %	99,80 %	37,3 %	99,42 %	44,2 %	99,74 %	47,9 %	99,62 %	51,2 %	99,7 %
33,6 %	99,82 %	38,4 %	99,56 %	45,5 %	99,70 %	49,2 %	99,68 %	52,4 %	99,6 %
35,0 %	99,86 %	39,5 %	99,78 %	47,0 %	99,88 %	50,5 %	99,72 %	53,5 %	99,8 %
36,4 %	99,94 %	40,6 %	99,74 %	48,4 %	99,84 %	51,8 %	99,80 %	54,6 %	99,8 %
37,8 %	99,90 %	41,7 %	99,88 %	49,8 %	99,92 %	53,1 %	99,74 %	55,6 %	99,9 %
39,2 %	99,98 %	42,8 %	99,94 %	51,2 %	99,96 %	54,3 %	99,94 %	56,8 %	99,8 %
40,6 %	99,98 %	43,9 %	99,88 %	52,6 %	99,98 %	55,7 %	99,94 %	57,9 %	99,9 %
42,0 %	99,98 %	45,0 %	99,88 %	54,0 %	99,96 %	57,0 %	99,98 %	59,0 %	100 %
43,4 %	100 %	46,1 %	99,98 %	55,4 %	99,94 %	58,3 %	99,94 %	60,0 %	100 %
44,8 %	100 %	47,2 %	99,94 %	56,8 %	99,98 %	59,6 %	99,98 %	61,2 %	100 %
46,2 %	100 %	48,3 %	100 %	58,2 %	99,96 %	60,9 %	100 %	62,3 %	100 %
47,6 %	100 %	49,4 %	100 %	59,5 %	100 %	62,2 %	100 %	63,4 %	100 %
49,0 %	99,98 %	50,5 %	100 %	61,0 %	100 %	63,5 %	100 %	64,4 %	100 %
50,4 %	100 %	51,6 %	99,92 %	62,4 %	100 %	64,7 %	100 %	65,5 %	100 %
51,8 %	100 %	52,7 %	100 %	63,8 %	100 %	66,1 %	100 %	66,7 %	100 %
53,2 %	100 %	53,8 %	99,98 %	65,2 %	100 %	67,4 %	100 %	67,8 %	100 %
54,6 %	100 %	54,9 %	100 %	66,6 %	100 %	68,7 %	100 %	68,9 %	100 %

Таблиця 7

Результатний частотний ранг символів розширеної української абетки

Ранг	Символ	Кількість	Ранг	Символ	Кількість	Ранг	Символ	Кількість
1	Пробіл	965000	13	С	850330	25	Ш	677283
2	А	964330	14	Д	893949	26	Х	597547
3	О	964315	15	У	957430	27	Ю	777114
4	И	964981	16	М	957659	28	Ж	505639
5	І	933228	17	П	957379	29	Є	507453
6	Н	767183	18	Б	776434	30	Ц	949971
7	Т	769190	19	З	776124	31	Ї	953732
8	В	928458	20	Я	917336	32	Щ	963415
9	Е	946288	21	Й	662289	33	Дефіс	907832
10	Р	917748	22	Г	683924	34	Ґ	905061
11	Л	626545	23	Ь	950992	35	Ф	957131
12	К	615816	24	Ч	953590	36	Апостроф	962393

Висновки та перспективи подальших наукових розвідок

У результаті проведених досліджень стосовно творів Марка Черемшини можна дійти таких висновків:

- За розміром уривків тексту, з яких формується текст-вибірка для статистичних обстежень, найкращі результати дають уривки довжиною в сто символів з точністю до слова.
- За обсягом вибірки бажаних результатів можна досягти, починаючи з такого, що становить 29,4 % від обсягу всіх текстів письменника, однак надійний результат отримуємо, починаючи від 50,4 %.
- Зі зростанням розміру уривків тексту обсяг "надійної" вибірки зростає.
- Частотні ранги символів, особливо для найчастотніших, показують певну стабільність, однак тільки для груп найчастотніших, середньочастотних та низькочастотних символів.

Отримані результати стосуються лише творчості Марка Черемшини. Для узагальнень щодо інших українськомовних художніх творів необхідні подальші дослідження.

1. Алферов А. П. Основы криптографии: учеб. пособие, 2-е изд., испр. и доп. / А. П. Алферов, А. Ю. Зубов, А. С. Кузьмин, А. В. Черемушкин. – М: Гелиос АРВ, 2002.— 480 с. 2. Архипова О. О. Частотний аналіз використання букв української мови / О. О. Архипова, В. М. Журавльов [Електронний ресурс]. – Режим доступу: http://kudin.net/r/index.php/ukr-frequency-analysis 3. Бук С. Лінгвостатистичний опис "Не спитавши броду" Івана Франка / С. Бук [Електронний ресурс]. – Режим доступу: http://www.lnu.edu.ua/faculty/ Philol/www/visnyk/55_2011/55_2011_Buk.pdf 4. Бук С. Сучасні методи дослідження мови письменника у слов'янознавстві / С. Бук [Електронний ресурс]. – Режим доступу: http://www.lnu.edu.ua/page/n61/010.pdf 5. Верхозин С. С. О статусе количественных методов в лингвистике / С. С. Верхозин [Електронний ресурс]. – Режим доступу: http://cyberleninka.ru/article/n/o-statuse-kolichestvennyh-metodov-v-lingvistike 6. Гладкий А. В. Математические методы изучения естественных языков, Математическая логика, теория алгоритмов и теория множеств: Сборник работ. Посвящается академику Петру Сергеевичу Новикову к его семидесятилетию, Тр. МИАН СССР, 133, 1973, 95–108 [Електронний ресурс]. – Режим доступу: http://www.mathnet.ru/links/ 3ff1f6b395ed41df319615fb89072f40/ tm2737.pdf 7. Головин Б. Н. О вероятностно-статистическом изучении стилевой дифференциации языка / Б. Н. Головин. – М.: Наука, 1988. – 258 с. 8. Кригін М. Ю., Широков В. А. Дослідження інформаційно-статистичних властивостей українського тексту / М. Ю. Кригін, В. А. Широков // Математические машины и