УДК 811.111+81`3
DOI https://doi.org/10.24919/2308-4863/88-2-29

Наталія ПЛЮТА,
orcid.org/0009-0008-2254-3141
аспірант кафедри прикладної лінгвістики, порівняльного мовознавства та перекладу
Українського державного університету імені Михайла Драгоманова
(Київ, Україна) n.l.pliuta@npu.edu.ua

ЕКСПЕРИМЕНТАЛЬНЕ ДОСЛІДЖЕННЯ ВПЛИВУ СПОСОБУ ФОРМУВАННЯ КОРПУСНОЇ ВИБІРКИ НА РЕПРЕЗЕНТАТИВНІСТЬ РЕЗУЛЬТАТІВ

У статті розглянуто вплив способів формування корпусних вибірок, підвибірок (у тому числі їх розмірів) на результати їх застосування у корпусній лінгвістиці. Зокрема на похибку розрахунку частоти слововживання, а відповідно і репрезентативність отриманих результатів. У процесі дослідження було використано статистичні методи обробки результатів вимірювань. Як відомо, корпус повинен мати розмір, достатній для забезпечення статистично надійних результатів по частоті та дистрибуції. Дослідження проводилося на корпусі розміром 290 тисяч слів, сформованому з брифінгів прессекретаря Білого дому США Шона Спайсера, та двох корпусах розміром 277 тисяч слів і 271 тисячу слів, сформованих з брифінгів прессекретаря Білого дому США Арі Флейшера. Для кожної підвибірки вказаних корпусів були обраховані частоти дейктиків I, We, You, The, He, President, Americans (American people), а також абсолютні та відносні похибки кожного вимірювання і середнє значення частоти вказаних дейктиків для корпусу в цілому, розраховане на одну тисячу слів. Крім того, обчислювалися довірчі інтервали для довірчих ймовірностей $P = 0,68$ та $P = 0,95$. Ці дії було виконано для випадку розбивки корпусів на підвибірки по 10 тисяч слів, а також для випадку розбивки корпусів на підвибірки довільного розміру. Аналіз отриманих результатів дозволив зробити висновки про те, що результат визначення середнього значення частот слововживання для корпусу в цілому практично не залежить від способу розбивки корпусної вибірки на підвибірки, оскільки похибка знаходиться в межах довірчого інтервалу. З метою оцінки мінімального розміру вибірки, достатнього для отримання репрезентативних результатів, було проведено обрахунки для випадку формування підвибірок наростаючим підсумком. Отримані результати показали, що для довірчої ймовірності $P = 0,68$ мінімальний розмір корпусної вибірки для отримання репрезентативних результатів повинен бути не меншим, ніж 250 тисяч слів, тоді як для довірчої ймовірності $P = 0,95$ мінімальний розмір вибірки становить 200 тисяч слів.

Ключові слова: *корпус, вибірка, корпусна лінгвістика, статистичний метод, частота, розмір корпусу.*

Nataliia PLIUTA,
orcid.org/0009-0008-2254-3141
Postgraduate student at the Department of Applied Language Studies,
Comparative Linguistics and Translation
Mykhailo Dragomanov State University of Ukraine
(Kyiv, Ukraine) n.l.pliuta@npu.edu.ua

AN EXPERIMENTAL STUDY OF THE INFLUENCE OF THE METHOD OF FORMING A CORPUS SAMPLE ON THE REPRESENTATIVENESS OF THE RESULTS

The article deals with the influence of the methods of forming corpus samples and subsamples (including the sample size) on the results of their application in corpus linguistics. In particular, on the error in calculating the frequency of word usage, and accordingly, on the representativeness of the results obtained. In the course of the research, statistical methods of processing the measurement results were used. As is known, the corpus should be of a size sufficient to provide statistically reliable results in terms of frequency and distribution. The study was conducted on a corpus of 290 thousand words, formed from the briefings of the US White House Press Secretary Sean Spicer, and two corpora of 277 thousand words and 271 thousand words, formed from the briefings of the US White House Press Secretary Ari Fleischer. For each subsample of the corpora, the frequencies of the deictics I, We, You, They, He, President, Americans (American people) were calculated, as well as the absolute and relative errors of each measurement; the average frequency of the deictics for the corpus as a whole was calculated per a thousand words. In addition, confidence intervals were calculated for the confidence probabilities $P = 0,68$ and $P = 0,95$. These steps were performed for the case of splitting the corpora into subsamples of 10 thousand words each, as well as for the case of splitting the corpora into subsamples of arbitrary size. The analysis of the obtained results allowed us to conclude that the result of determining the average word frequency for the corpus as a whole depends little on the method of splitting the corpus into subsamples, since the error falls within the

confidence interval. In order to estimate the minimum sample size sufficient to obtain representative results, we performed calculations for the case of forming subsamples by cumulative total. The results showed that for a confidence probability of $P = 0,68$, the minimum corpus sample size to ensure representative results should be at least 250 thousand words, while for a confidence probability of $P = 0,95$ the minimum sample size should be at least 200 thousand words.

Key words: *corpus, sample, corpus linguistics, statistical method, quantity, sample size.*

Постановка проблеми. Використання корпусних інструментів у дослідженнях дискурсу, у тому числі політичного дискурсу, може давати і дає вагомі результати. У процесі таких досліджень доводиться не тільки обчислювати частоту вживання мовних одиниць, а і порівнювати знайдені частоти між собою. І у першому, і у другому випадку коректний аналіз обчислених частот неможливий без оцінки їх похибок. З цією метою в корпусних дослідженнях широко застосовують статистичні методи обробки результатів вимірювання. Без цього неможлива адекватна інтерпретація отриманих результатів.

Аналіз досліджень. За Bohm G. та Zech G. (Bohm, Zech, 2017) жодне обчислення частоти слововживання не може бути проведене з абсолютною точністю – воно обов'язково матиме похибку, спричинену певними факторами. Навіть якщо просто розраховувати частоту слововживання для ряду одиниць в обраному тексті, то вже за рахунок заокруглень буде виникати похибка. Для зручності використання загальновживаної у статистичному методі термінології в подальшому розрахунок частоти слововживань будемо називати вимірюванням.

Як відомо (Жуковська, 2013; Кульчицький, 2015; Перебийніс, 2013), при застосуванні корпусних методів у лінгвістиці важливе значення має спосіб формування вибірки, її розмір та інші параметри. Згідно з П. де Гааном (De Haan, 1992), корпус повинен мати розмір, достатній для того, щоб «забезпечити статистично надійні результати по частоті й дистрибуції». У той же час «прийнятність вибірок залежить від конкретного дослідження, яке проводиться» (De Haan, 1992). Як визначає В. В. Жуковська, «однією з вагомих проблем сучасної корпусної лінгвістики є визначення розміру корпусу, достатнього для отримання надійних висновків» (Жуковська, 2013). Це пов'язано з тим, що основою статистики є закон великих чисел. Його суть полягає в наступному: чим більший обсяг тексту (вибірки) ми досліджуємо, тим більш репрезентативними будуть результати (Заяць, Заяць, 2010; Bohm, Zech, 2017). У той же час цілком очевидно, що дослідити безкінечно великий обсяг тексту неможливо. Тому завдання полягає у визначенні мінімального розміру вибірки, достатнього для отримання репрезентативних результатів.

Проте, як відзначає В. В. Жуковська (Жуковська, 2013), деякі дослідники зауважують, що будь-який текст чи зібрання текстів теоретично може бути визнане корпусом, а тому корпусний аналіз можливо здійснювати навіть на дуже малих за обсягом текстах.

Вплив способу формування вибірки (підвибірки) на точність отримуваних результатів був предметом досліджень у науковому середовищі (Плюта, 2023a; Плюта, 2023b; De Haan, 1992). Досить детально це питання розглянуто В. І. Перебийніс (Перебийніс, 2013). Її робота показує, що величина відносної похибки отриманих результатів залежить від багатьох факторів, пов'язаних з формуванням вибірки. Зокрема це і величина вибірки, і кількість підвибірок, і абсолютні значення частот досліджуваних одиниць.

Для грубої оцінки обсягу вибірки В. І. Перебийніс (Перебийніс, 2013: 29) пропонує користуватися формулою:

$$\varepsilon = \frac{1,96}{\sqrt{m}}, \qquad (1)$$

де ε – відносна похибка, m – кількість вживань одиниці або її абсолютна частота.

Звідси В. І. Перебийніс (Перебийніс, 2013: 29) робить висновок «що величина відносної похибки буде зменшуватися зі збільшенням кількості випадків вживання досліджуваної одиниці і навпаки». Однак, як вказує сама В. І. Перебийніс, цей «спосіб оцінки точності дослідження досить грубий <…> використовувати його слід лише для першого, грубого визначення обсягу вибірки» (Перебийніс, 2013: 31).

Метою статті є визначення впливу розміру вибірки та підвибірок на репрезентативність результатів вимірювань.

Виклад основного матеріалу. У якості матеріалу дослідження слугували вибірки, сформовані з прес-брифінгів прес-секретаря Білого дому за першої каденції Президента США Дональда Трампа, Шона Спайсера розміром 290 тисяч слів, а також два корпуси розміром 277 тисяч слів та 271 тисяч слів відповідно, сформовані на основі прес-брифінгів прес-секретаря Білого дому за каденції Президента США Джорджа Буша, Арі Флейшера.

Досліджувалася частота вживання мовцями дейктиків *I, WE, YOU, THEY, HE, PRESIDENT,*

AMERICANS (AMERICAN PEOPLE), розрахована на тисячу слів. Для кожної підвибірки було обчислено абсолютні похибки δx_i та відносні похибки ε_i. Крім того, для корпусу визначалося середнє значення частоти $\underline{x}$ і модуль довірчого інтервалу σ для кожного дейктика.

Обрахунки проводились у такому порядку (Bohm, Zech, 2017):

Знаходимо вибіркове значення середнього $\underline{x}$, яке приблизно дорівнює середньому арифметичному (або середньому) $\langle x \rangle$:

$$\underline{x} \approx < x > = \frac{1}{n}\sum_{i=1}^{n} x_i , \qquad (2)$$

де n – кількість підвибірок, $i = 1 \ldots n$.

Для кожної i-тої підвибірки обчислюємо абсолютну похибку знайденого значення частоти слововживання:

$$\delta x_i = x_i - \langle x \rangle , \qquad (3)$$

$i = 1 \ldots n$.

Оцінюємо відносну похибку частоти вживання вибраного дейктика в i-тій підвибірці:

$$\varepsilon_i = \frac{|\delta x_i|}{\langle x \rangle} \qquad (4)$$

Обчислюємо стандартне відхилення середнього:

$$S = \sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}(x_i - \underline{x})^2} , \qquad (5)$$

де n – кількість підвибірок: $i = 1 \ldots n$.

Для заданої довірчої ймовірності P, числа ступенів свободи $k(k = n-1)$, по таблиці знаходимо значення коефіцієнта Стьюдента $t(P;k)$ і обчислюємо модуль довірчого інтервалу:

$$\sigma = S * t(P;k) \qquad (6)$$

Істинне значення частоти слововживання записуємо у вигляді:

$$x = \underline{x} \pm \sigma , \qquad (7)$$

або

$$\underline{x} - \sigma < x < \underline{x} + \sigma \qquad (8)$$

Для прикладу, в таблиці 1 наведено результати обрахунків для дейктика *I* на основі даних вибірки з прес-брифінгів прес-секретаря Білого дому Шона Спайсера розміром 290 тис. слововживань. Вибірка розбивалася на 29 підвибірок (n_i) по 10 тис. слів кожна (N_i). Підрахунки проводилися згідно формул (2) – (4).

Відповідно до формул (5) – (8) було отримано значення стандартного відхилення середнього $S = 0{,}8286$ (для $\underline{x} = 20{,}8144$), що дало змогу обрахувати наступні значення довірчих інтервалів σ.

Для довірчої ймовірності $P = 0{,}68$ і коефіцієнта Стьюдента $t = 1{,}0124$ модуль довірчого інтервалу $\sigma = 0{,}8389$.

Для довірчої ймовірності $P = 0{,}95$ і коефіцієнта Стьюдента $t = 2{,}0484$ модуль довірчого інтервалу $\sigma = 1{,}6973$.

Отже істинне значення частоти x дейктика *I* для довірчої ймовірності $P = 0{,}68$ знаходиться в межах інтервалу $20{,}8144 - 0{,}8389 \le x \le 20{,}8144 + 0{,}8389$

Таблиця 1

Значення частот x_i, їх абсолютних похибок δx_i та відносних похибок ε_i для кожної підвибірки корпусу Шона Спайсера

№ п/п	N_i	x_i	δx_i	ε_i	№ п/п	N_i	x_i	δx_i	ε_i
1.	10 000	17,295	3,5194	0,1691	16.	10 000	22,791	1,9766	0,095
2.	10 000	26,479	5,6646	0,2721	17.	10 000	22,791	1,9766	0,095
3.	10 000	27,295	6,4806	0,3113	18.	10 000	27,992	7,1776	0,3448
4.	10 000	18,1	2,7144	0,1304	19.	10 000	23,788	2,9736	0,1429
5.	10 000	22,098	1,2836	0,0617	20.	10 000	24,49	3,6756	0,1766
6.	10 000	20,9	0,0856	0,0041	21.	10 000	24,798	3,9836	0,1914
7.	10 000	19,594	1,2204	0,0586	22.	10 000	25,492	4,6776	0,2247
8.	10 000	16,292	4,5224	0,2173	23.	10 000	25,895	5,0806	0,2441
9.	10 000	14,596	6,2185	0,2988	24.	10 000	20,29	0,5244	0,0252
10.	10 000	14,099	6,7154	0,3226	25.	10 000	17,098	3,7164	0,1786
11.	10 000	14,593	6,2214	0,2989	26.	10 000	11,499	9,3154	0,4475
12.	10 000	17,898	2,9164	0,1401	27.	10 000	22,291	1,4766	0,0709
13.	10 000	19,898	0,9164	0,044	28.	10 000	22,231	1,4166	0,0681
14.	10 000	25,195	4,3806	0,2105	29.	10 000	14,943	5,8714	0,2821
15.	10 000	22,898	2,0836	0,1001					

для довірчої ймовірності $P = 0,95$ – в межах інтервалу $20,8144 - 1,6973 \le x \le 20,8144 + 1,6973$.

Розглянемо також вплив розміру підвибірок на оцінку середнього значення частоти слововживання. Для цього розіб'ємо корпус, сформований з текстів прес-брифінгів прессекретаря Білого дому Шона Спайсера, розміром 290 тисяч слів на 29 підвибірок довільного розміру. Результати відповідних обрахунків наведено в таблиці 2.

Як видно з таблиці, отримані значення частот слововживання для лексичної одиниці *I* практично не відрізняються від відповідних частот обчислених для підвибірок розміром 10 тисяч слів кожна.

Було отримано значення стандартного відхилення середнього $S = 0,9901$ (для $\underline{x} = 20,9777$), що дало змогу отримати наступні значення довірчих інтервалів σ.

Таблиця 2

Значення частот x_i, їх абсолютних похибок δx_i та відносних похибок ε_i для підвибірок довільного розміру корпусу Шона Спайсера

№ п/п	N_i	x_i	δx_i	ε_i	№ п/п	N_i	x_i	δx_i	ε_i
1.	7500	14,4	6,5777	0,3136	16.	1700	20,5882	0,3895	0,0186
2.	3400	23,5294	2,5517	0,1216	17.	5500	22,1818	1,2041	0,0574
3.	8600	26,5116	5,5339	0,2638	18.	9800	20	0,9777	0,0466
4.	9500	27,6842	6,7065	0,3197	19.	12450	24,5783	3,6006	0,1716
5.	1300	29,2308	8,2531	0,3934	20.	19500	21,6923	0,7146	0,0341
6.	2700	12,963	8,0147	0,3821	21.	14600	25,0685	4,0908	0,195
7.	9900	19,7979	1,1798	0,0562	22.	17800	25,7303	4,7526	0,2266
8.	11300	22,8319	1,8542	0,0884	23.	13450	24,684	3,7063	0,1767
9.	6800	19,8529	1,1248	0,0536	24.	19380	23,9938	3,0161	0,1438
10.	2200	29,5454	8,5677	0,4084	25.	8800	27,0454	6,0677	0,2892
11.	4580	10,4803	10,4974	0,5004	26.	4650	17,8495	3,1282	0,1491
12.	7900	20,5063	0,4714	0,0225	27.	30700	17,5896	3,3881	0,1615
13.	13700	14,2336	6,7441	0,3215	28.	11900	19,916	1,0617	0,0506
14.	16900	15,7988	5,1789	0,2468	29.	7190	20,8623	0,1154	0,0055
15.	6300	9,2063	11,7714	0,5611					

Для довірчої ймовірності $P = 0,68$ і коефіцієнта Стьюдента $t = 1,0124$ модуль довірчого інтервалу $\sigma = 1,0024$.

Для довірчої ймовірності $P = 0,95$ і коефіцієнта Стьюдента $t = 2,0484$ модуль довірчого інтервалу $\sigma = 2,0281$.

Отже істинне значення частоти x дейктика *I* для довірчої ймовірності $P = 0,68$ знаходиться в межах інтервалу $19,9753 \le x \le 21,9801$; для довірчої ймовірності $P = 0,95$ – в межах інтервалу $18,9496 \le x \le 23,0058$.

Порівнявши отримані істинні значення частот дейктика *I* для підвибірок розміром 10 тис. слів і підвибірок довільного розміру, доходимо висновку про те, що розмір підвибірок мало впливає на результат вимірювань.

Розрахунок частоти слововживання для випадку, коли весь корпус буде однією підвибіркою розміром 290 тис. слів для дейктика *I* дав результат $x_i = 20,9897$. Тобто отриманий результат близький до середніх значень, отриманих у двох попередніх випадках.

Отже, в окремих випадках, коли визначення довірчого інтервалу не є необхідним для дослідження, можна обмежитись одноразовим підрахунком частоти слововживань для всієї вибірки.

Наведене твердження вимагає більш детального розгляду. З цією метою корпус Шона Спайсера було розбито на 29 підвибірок, коли кожна наступна включає в себе попередню плюс 10 тис. слів. Для цих підвибірок було здійснено обрахунки частот всіх дейктиків. Результати підрахунків для частоти дейктика *I* наведені в таблиці 3.

З метою оцінки мінімального розміру вибірки, достатнього для отримання репрезентативних результатів, було розглянуто ще два корпуси розміром 276 700 слів та 270 767 слів відповідно, сформованих на основі прес-брифінгів прессекретаря Білого дому Арі Флейшера. Кожна підвибірка в цих корпусах відповідає 3–5 брифінгам, отже розміри підвибірок різні. Обрахунки проводилися за тим же алгоритмом, що і для корпусу Шона Спайсера.

Таблиця 3

Значення частот x_i дейктика *I* для підвибірок наростаючим підсумком корпусу Шона Спайсера

№ п/п	N_i	x_i	№ п/п	N_i	x_i
1.	10 000	17,3000	16.	160 000	20,0063
2.	20 000	21,9000	17.	170 000	20,1706
3.	30 000	23,7000	18.	180 000	20,6056
4.	40 000	22,3000	19.	190 000	20,7737
5.	50 000	22,2600	20.	200 000	20,9600
6.	60 000	22,0333	21.	210 000	21,1429
7.	70 000	21,6857	22.	220 000	21,3409
8.	80 000	21,0125	23.	230 000	21,5391
9.	90 000	20,3000	24.	240 000	21,4875
10.	100 000	19,6800	25.	250 000	21,3120
11.	110 000	19,2182	26.	260 000	20,9346
12.	120 000	19,1083	27.	270 000	20,9852
13.	130 000	19,1692	28.	280 000	21,0321
14.	140 000	19,6000	29.	290 000	20,9897
15.	150 000	19,8200			

За результатами обрахунків для всіх дейктиків у трьох корпусах було визначено мінімальні розміри корпусів, починаючи з яких частота вимірюваних одиниць не виходить за межі довірчого інтервалу. Отримані результати відображено в таблицях 4–6.

У кожному з трьох розглянутих корпусів обрахунки проводили для семи дейктиків: *I, WE, YOU, THEY, HE, PRESIDENT, AMERICANS (AMERICAN PEOPLE)*. Сумарно це 21 випадок. Для кожного з цих випадків було визначено мінімальний розмір вибірки, починаючи з якого частота вживання обраної лексичної одиниці не виходить за межі довірчого інтервалу для довірчої ймовірності $P = 0,95$ та довірчої ймовірності $P = 0,68$. Скориставшись цими даними, сформуємо таблиці 8–9 для $P = 0,95$ і $P = 0,68$ відповідно. При цьому у верхньому рядку таблиці будуть значення інтерва-

Таблиця 4

Значення мінімальних розмірів корпусу для вибраних дейктиків, обчислених за корпусом Шона Спайсера

	I	WE	YOU	THEY	HE	PRESIDENT	AMERICANS
P=0,95	130 000	100 000	250 000	280 000	280 000	100 000	220 000
P=0,68	160 000	140 000	250 000	280 000	290 000	220 000	260 000

Таблиця 5

Значення мінімальних розмірів корпусу для вибраних дейктиків, обчислених за корпусом Арі Флейшера №1

	I	WE	YOU	THEY	HE	PRESIDENT	AMERICANS
P=0,95	165 033	176 419	108 888	11 849	165 033	120 633	98 647
P=0,68	249 398	213 103	213 103	11 849	213 103	165 033	141 655

Таблиця 6

Значення мінімальних розмірів корпусу для вибраних дейктиків, обчислених за корпусом Арі Флейшера №5

	I	WE	YOU	THEY	HE	PRESIDENT	AMERICANS
P=0,95	115 485	157 723	94 711	41 114	61 890	82 994	51 341
P=0,68	137 661	216 847	137 661	106 282	226 081	106 282	82 994

лів розмірів вибірок у тисячу слів; в середньому рядку – кількість випадків, коли значення мінімального розміру вибірки потрапляє в цей інтервал; у нижньому рядку – відсоток вибірок, які потрапили у заданий інтервал.

Висновки. Аналіз результатів, наведених в таблицях 7–8, дозволяє стверджувати, що для довірчої ймовірності $P = 0,95$ мінімальний розмір корпусу для забезпечення репрезентативних результатів повинен бути не менше 200 тис. слів. Тобто 95% вимірювань значення частоти досліджуваних одиниць будуть знаходитись в межах інтервалу $\underline{x} - 2\sigma \le x_i \le \underline{x} + 2\sigma$. Для довірчої ймовірності $P = 0,68$ цей розмір становить приблизно 250 тис. слів. Отже 68% вимірювань значення частоти досліджуваних одиниць будуть знаходитись в межах інтервалу $\underline{x} - \sigma \le x_i \le \underline{x} + \sigma$.

Порівняння даних таблиці 1 і таблиці 2 показує, що розміри підвибірок не мають суттєвого впливу на результати вимірювань і можуть бути різними в межах однієї вибірки.

Таблиця 7

Розподіл кількості мінімальних обсягів вибірок, що потрапляють у заданий діапазон розмірів корпусів для P=0,95

Діапазон розміру корпусів (тис. сл.)	0 – 50	0 – 100	0 – 150	0 – 200	0 – 250	0 – 300
К-кість випадків, що потрапляють в інтервал	3	11	17	21	21	21
Відсоток випадків, що потрапляють в інтервал	14,27	52,38	80,95	100	100	100

Таблиця 8

Розподіл кількості мінімальних обсягів вибірок, що потрапляють у заданий діапазон розмірів корпусів для P=0,68

Діапазон інтервалів (тис. сл.)	0 – 50	0 – 100	0 – 150	0 – 200	0 – 250	0 – 300
К-кість випадків, що потрапляють в інтервал	1	3	10	13	21	21
Відсоток випадків, що потрапляють в інтервал	4,7	14,27	47,62	61,91	100	100

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Жуковська В. В. Вступ до корпусної лінгвістики: навчальний посібник. Житомир: Вид-во ЖДУ ім. І.Франка. 2013. 140 с.

2. Заяць В. М., Заяць М. М. Методи зіставлення статистичних характеристик при формуванні вибірок у лінгвістиці. Інформаційні системи та мережі: [збірник наукових праць] / відповідальний редактор В.В. Пасічник. Л.: Видавництво Львівської політехніки. Т. 673. № 1. 2010. С. 296–305. URL: https://science.lpnu.ua/sisn/all-volumes-and-issues/volume-673-2010/metodi-zistavlennya-statistichnih-harakteristik-pri

3. Кульчицький І. М. Вибір розміру вибірки для статистичних опрацювань текстів. Вісник Національного університету «Львівська політехніка». 2015. Вип. 814. № 1. С. 293–300. URL: https://science.lpnu.ua/uk/sisn/vsi-vypusky/vypusk-814-2015/vybir-rozmiru-vybirky-dlya-statystychnyh-opracyuvan-tekstiv

4. Перебийніс В. І. Статистичні методи для лінгвістів: посіб. Вид. 2-ге, випр. і доп. Вінниця: Нова Книга, 2013. 176 с.

5. Плюта Н. Л. До питання стратегії формування та роботи з корпусними вибірками при аналізі політичного дискурсу. Світ мови – світ у мові: матеріали VII Міжнародної наукової конференції (м. Київ, 27 жовтня 2023 р.; Український державний університет імені Михайла Драгоманова): тези доповідей / уклад. І. В. Воловенко, Т. В. Слива; за заг. ред. Ю. В. Кравцової. Київ: Вид-во УДУ імені Михайла Драгоманова. 2023. С. 221–224. URL: https://fif.udu.edu.ua/images/Conferences/2023/SM_2023_zb.pdf

6. Pliuta N. L. The estimation of the corpus size in the corpus-based analysis of the white house press secretary discourse. Актуальні проблеми романо-германської філології: матеріали IV Міжнародної наукової конференції. Київ: УДУ імені Михайла Драгоманова. 2023. С. 123–126 URL: https://fif.udu.edu.ua/images/Conferences/2023_05_11_12.pdf

7. Bohm G., Zech G. Introduction to Statistics and Data Analysis for Physicists, VerlagDeutsches Elektronen-Synchrotron, E-Book. 2017. 504 P. URL: https://bib-pubdb1.desy.de/record/389738/files/vstatmp_e17.pdf

8. De Haan, P. The optimum corpus sample size? New Directions in English Language Corpora: Methodology, Results, Software Developments, Berlin, Boston: De Gruyter Mouton. 1992. Pp. 3–20. URL: https://doi.org/10.1515/9783110878202.3

REFERENCES

1. Zhukovska V. V. (2013) Vstup do korpusnoi linhvistyky: navchalnyi posibnyk. [Introduction to corpus linguistics: a textbook] Zhytomyr: Vyd-vo ZhDU im. I.Franka. 140 p. [in Ukrainian].

2. Zaiats V. M., Zaiats M. M. (2010) Metody zistavlennia statystychnykh kharakterystyk pry formuvanni vybirok u linhvistytsi. [Methods of comparing statistical characteristics in the formation of samples in linguistics] Informatsiini systemy ta merezhi: [zbirnyk naukovykh prats] / vidpovidalnyi redaktor V.V. Pasichnyk. L.: Vydavnytstvo Lvivskoi politekhniky. T. 673. № 1. 296–305. URL: https://science.lpnu.ua/sisn/all-volumes-and-issues/volume-673-2010/metodi-zistavlennya-statistichnih-harakteristik-pri [in Ukrainian].

3. Kulchytskyi I. M. (2015) Vybir rozmiru vybirky dlia statystychnykh opratsiuvan tekstiv. [Selection of sample size for statistical processing of texts] Visnyk Natsionalnoho universytetu «Lvivska politekhnika». Vyp. 814. № 1. 293–300. URL: https://science.lpnu.ua/uk/sisn/vsi-vypusky/vypusk-814-2015/vybir-rozmiru-vybirky-dlya-statystychnyh-opracyuvan-tekstiv [in Ukrainian].

4. Perebyinis V. I. (2013) Statystychni metody dlia linhvistiv: posib. [Statistical methods for linguists: a guide] Vyd. 2-he, vypr. i dop. Vinnytsia: Nova Knyha. 176 P. [in Ukrainian].

5. Pliuta N. L. (2023) Do pytannia stratehii formuvannia ta roboty z korpusnymy vybirkamy pry analizi politychnoho dyskursu. [On the strategy of forming and working with corpus samples in the analysis of political discourse] Svit movy – svit u movi: materialy VII Mizhnarodnoi naukovoi konferentsii (m. Kyiv, 27 zhovtnia 2023 r.; Ukrainskyi derzhavnyi universytet imeni Mykhaila Drahomanova): tezy dopovidei / uklad. I. V. Volovenko, T. V. Slyva; za zah. red. Yu. V. Kravtsovoi. Kyiv: Vyd-vo UDU imeni Mykhaila Drahomanova. 221–224. URL: https://fif.udu.edu.ua/images/Conferences/2023/SM_2023_zb.pdf [in Ukrainian].

6. Pliuta N. L. (2023) The estimation of the corpus size in the corpus-based analysis of the white house press secretary discourse. Aktualni problemy romano-hermanskoi filolohii: materialy IV Mizhnarodnoi naukovoi konferentsii. Kyiv: UDU imeni Mykhaila Drahomanova. 123–126 URL: https://fif.udu.edu.ua/images/Conferences/2023_05_11_12.pdf

7. Bohm G., Zech G. (2017) Introduction to Statistics and Data Analysis for Physicists, VerlagDeutsches Elektronen-Synchrotron, E-Book. 504 P. URL: https://bib-pubdb1.desy.de/record/389738/files/vstatmp_e17.pdf

8. De Haan, P. (1992) The optimum corpus sample size? New Directions in English Language Corpora: Methodology, Results, Software Developments, Berlin, Boston: De Gruyter Mouton. Pp. 3–20. URL: https://doi.org/10.1515/9783110878202.3