
СТАТИСТИЧНІ МЕТОДИ В ЛІНГВІСТИЦІ (НА ПРИКЛАДІ ЧАСТКОВОГО СТАТИСТИЧНОГО АНАЛІЗУ ФРАЗЕОЛОГІЧНИХ ОДИНИЦЬ В РОМАНІ ЧАРЛЬЗА ДІККЕНСА "ОЛІВЕР ТВІСТ" ТА ЙОГО ПЕРЕКЛАДІ О. В. КРИВЦОВА)

УДК 81'324

Ратушняк Я. І.

Національний університет кораблебудування імені адмірала Макарова

Проаналізовано приклади вживання статистичних методів у лінгвістиці. Встановлено, що взаємозв'язок дисциплін призводить до ефективного функціонування. Визначені основні статистичні характеристики роману Чарльза Діккенса "Олівер Твіст" та перекладу О. В. Кривцова. Проаналізовано ступінь адекватності перекладу роману за його образним компонентом.

Ключові слова: статистика, лінгвістика, фразеологічна одиниця, генеральна сукупність, вибірка, варіанти, інтервальний варіаційний ряд, функція щільності статистичного розподілу.

Останнім часом все більшу популярність набувають прикладні науки та взаємодія різних наукових дисциплін для отримання дуже важливих спільних результатів, важливих для різних сфер науки та життєдіяльності людини та суспільства.

Все більша увага приділяється використанню, наприклад, статистичних методів для потреб лінгвістики. Першим звернувся до цього питання ще на початку ХХ століття відомий вчений Фердінанд де Соссюр, його ідеї розвивали науковці Л. Єльмслєв, Л. Блумфільд, детально вивчав статистичні методи в лінгвістиці вчений-генетик Георг Мендель, в Росії цим займалися Бодуен де Куртене та Ф. Фортунатов, значний внесок зробили вчені-лінгвісти Н. Хомські, Р. Фрумкіна та О. Реформатський. Це питання залишається актуальним і на сьогодні, що головним чином пов'язано зі стрімким розвитком інформаційних технологій та все більшою комп'ютеризацією суспільства.

Відштовхуючись від актуальності цієї проблеми, було проведене дане дослідження, метою якого було дослідити статистичні методи в лінгвістиці шляхом проведення часткового статистичного аналізу роману Чарльза Діккенса "Олівер Твіст" та його перекладу О.В. Кривцова.

Виходячи з мети дослідження, були сформульовані завдання:

- проаналізувати літературу з фразеології, статистики, лінгвістики, сферам застосування цих дисциплін;
- вивчити фразеологічні одиниці в романі та його перекладі;
- провести часткову статистичну обробку даних (увесь текст роману – генеральна сукупність, а фразеологічні одиниці – випадкові змінні);
- порівняти дані часткового статистичного аналізу по роману та його перекладу.

Як відомо, статистика – це математична наука, що займається збором, аналізом, тлумаченням чи поясненням, а також представленням даних. Цю науку використовують для різних академічних дисциплін, починаючи з природничих та суспільних наук, закінчуючи гуманітарними. Також статистику застосовують для прийняття важливих рішень в політиці та бізнесі.

Коли статистичні методи використовують для узагальнення та опису зібраних даних, то це має назву описова статистика. Зокрема, зразки даних можуть моделюватися особливим чином, який пояснює випадковість та непевність спостережень, а потім використовується для висновків з приводу досліджуваного процесу чи генеральної сукупності, це має назву статистика висновків. Обидві ці дисципліни формують прикладну статистику. [1]

Також, як було згадано вище, статистика може та повинна вносити свій внесок до лінгвістичних наук. Діапазон робіт, для яких можна застосувати статистичні методи, досить широкий. Наприклад, це може бути збір лінгвістичних даних, лінгвістична варіація та інші аспекти прикладної лінгвістики. Світова практика показує різні приклади застосування статистики в лінгвістиці шляхом збору статистичних даних та виведення закономірностей опрацьованої інформації. В лінгвістичних дослідженнях застосовують різні

статистичні методи: хі-квадратичний та варіативний аналіз, кластерний аналіз, аналіз головних компонентів та факторний аналіз [1].

Одним із прикладів може бути аналіз авторства та авторського стилю, що проводили Ліза Барбун та Арунава Чатрджі. Це дослідження є частково комп'ютерним трансформаційно-генеративним аналізом на основі вибраних драматичних робіт іспанського автора Алехандро Казона.

Шляхом підрахунку крос-кореляційних функцій статистичних змінних у вибраних роботах Казони дослідники змогли порівняти частоти цих змінних у літературних роботах Казони для дорослих та дітей.

Процедура даного дослідження – це визначення та вибір робіт Казони для порівняння, визначення вибірок та змінних, котрі можна буде проаналізувати за допомогою комп'ютерних програм, що були створені саме для цих цілей.

Наступний етап – це створення текстового файлу вибірки робіт Казони, які не входять до складу фінального аналізу, що вручну був проаналізований на наявність змінних, та запуск програми для визначення вірності обраних змінних [2].

Останній етап – введення вибірок в комп'ютерну програму, котру створили дослідники, запуск програми, групування результатів, підрахунок взаємозалежностей та аналіз подібності та різниці в роботах для порівняння.

П'ятдесят дев'ять змінних, що піддаються кількісному визначенню, були обрані за допомогою комп'ютерної програми, створеної дослідниками. Три з цих змінних – це середня довжина слова, середня кількість слів у реченні та кількість різних слів. Це єдині змінні, що не були проаналізовані за допомогою адаптованої трансформаційно-генеративної моделі. Також досліджувались перфектні та контінюусні конструкції. Ці конструкції пов'язані з можливим розповсюдженням предиката в іспанській мові та генеруються за правилами граматики, але до їх складу не входять трансформації. Було також декілька змінних, в тому числі заперечення та пасивний залог з серійністю, що включає до себе однорядні трансформації, додаткові змінні (разом з підляганням та порівняльні конструкції) – це дворядні трансформації [2].

В результаті проведення дослідження виявились три важливих моменти. Перший – те, що мова в роботах для дорослих структурно більш складна (тобто містить більше трансформацій), ніж в роботах для дітей. Другий – частота досліджуваних синтаксичних структур відрізняється в історичних та художніх творах. Третій – роботи для дітей в середньому містять менше різних лексичних форм, ніж роботи для дорослих. Всі ці результати вказують на простоту визначення стиля Казони та передбачуваність його синтаксичного стиля [2].

Пропонуємо інший приклад застосування статистичних методів в лінгвістиці шляхом проведення власного експерименту – статистичний аналіз фразеологічних одиниць в романі Чарльза Діккенса "Олівер Твіст" в оригіналі та в перекладі О.В. Кривцова.

Для цього були вивчені фразеологічні одиниці в романі та його перекладі, проведена часткова статистична обробка даних (весь текст роману – генеральна сукупність, а фразеологічні одиниці – випадкові змінні), а потім проведене порівняння даних часткового статистичного аналізу роману та його перекладу.

Були використані наступні методи: аналіз літератури з даного питання (фразеологія та статистика), статистичні методи для лінгвістів, методи статистичних розрахунків та побудови графіків.

Об'єктом були фразеологічні одиниці в романі та перекладі, а предметом дослідження – частковий статистичний аналіз фразеологічних одиниць в романі та перекладі з метою порівняння ступеню адекватності перекладу.

Практична цінність експерименту – це те, що дані цього дослідження можуть бути використані для більш поглибленого статистичного аналізу (наприклад, обчислення чисельних характеристик вибірки: дисперсії, середнього квадратичного, моди, медіани, коефіцієнту варіації, а також для детального дослідження всіх робіт Діккенса чи перекладів).

Далі наводиться статистична обробка та опис на основі роману "Олівер Твіст" Чарльза Діккенса в оригіналі та перекладі О.В. Кривцова.

Вся досліджена сукупність об'єктів (кількістю N) має назву генеральна сукупність.

Частина об'єктів кількістю n ($n \leq N$), випадково відібрана з генеральної сукупності має назву вибірка.

Нехай над випадковою величиною X проведено n незалежних випробувань; $x_1, x_2, \dots, x_n$ – її можливі значення. Дані заносяться у наведену нижче таблицю в порядку їх отримання [3].

1	2	...	i	...	n
x_1	x_2	...	x_i	...	x_n

Ця таблиця називається статистичним рядом, а x_i – варіантами. Серед варіант можуть бути однакові. Якщо кожній варіанті поставити у відповідність

$$P_i^* = \frac{n_i}{n}$$

її частоту (кількість повторень) n_i , а також відносну частоту P_i^* , а самі варіанти записувати у зростаючому (спадаючому) порядку, то наведену нижче таблицю називають варіаційним рядом [3].

X	x_1	x_2	...	x_i	...	x_k
N	n_1	n_2	...	n_i	...	n_k
P^*	P_1^*	P_2^*	...	P_i^*	...	P_k^*

Статистичною функцією розподілу $F^*(x)$ випадкової величини X називається закон зміни відносної частоти випадку $X < x$ в даному статистичному матеріалі:

$$F^*(x) = P^*(X < x).$$

Якщо об'єм вибірки n великий, то весь інтервал отриманих значень x_i розбивається на часткові (як правило рівні) інтервали $(x_1, x_2), (x_2, x_3), \dots, (x_k, x_{k+1})$. Обозначимо через n_i число значень випадкової величини X , котрі потрапили до інтервалу (x_i, x_{i+1}) . Припустимо, що початок інтервалу входить до інтервалу, а кінець – ні. Для кожного інтервалу визначимо відносну частоту P_i^* [3].

Наведена нижче таблиця, котра була отримана в результаті цих дій, має назву інтервальний варіаційний ряд.

I	(x_1, x_2)	(x_2, x_3)	...	(x_i, x_{i+1})	...	(x_k, x_{k+1})
---	--------------	--------------	-----	------------------	-----	------------------

N	n ₁	n ₂	...	n _i	...	n _k
P*	P ₁ *	P ₂ *	...	P _i *	...	P _k *

Якщо на кожному з відрізків $[x_i, x_{i+1}]$, взявши їх за основу, побудувати прямокутник, площа котрого дорівнюється P_i^* , то отримана фігура називається гістограма. Її площа дорівнює одиниці. Вважаємо, що крок $h = x_{i+1} - x_i = \text{const}$.

Якщо наближати $h \rightarrow 0$, то гістограма буде все більше й більше наближатися до деякої кривої, котра обмежує площу, рівну одиниці. Ця крива – графік деякої функції $f(x)$, котру називають щільністю розподілу випадкової величини X [3].

Charles Dickens "Oliver Twist" (оригінал)

Складаємо вибірку об'ємом $n = 55$ (таблиця 1.1).

Таблица 1.1

i	God bless you!	Do one's worst	Worse luck!	In the wind	Stuff and nonsense!	Upon my word!	Dear heart!	Lord save us!	Come it strong	Have the why and the wherefore
x	20	4	6	3	2	7	13	10	1	3
i	Make oneself at home	Bring to mind	Stock in trade	A long way off	Run to seed	A run of luck	What's the row?	Be left to the tender mercies of somebody	Go the wrong way	For that matter
x	5	5	1	3	2	1	1	1	3	1
i	Give leg-bail	As sober as a judge	At every pore	Make a fool of somebody	Put one's best leg first	On the lay	The kinchin lay	Jack Ketch	For God's sake!	Find one's way
x	2	0	5	4	1	2	1	1	6	3
i	Hue-and-cry	Pad it on the hoof	In the inmost recesses of the heart	Under one's hand and seal	Get the upper hand	Throw somebody off his guard	By Heaven!	Place in an awkward position	That's flat	As fit as a fiddle
x	1	1	1	1	1	1	7	4	1	0
i	Turn King's evidence	Have effect	Find oneself in the wrong box	I'll eat my head	The devil and all	As the crow flies	Give credence to	Lose cast	Take care of	Turn white-livered
x	2	1	4	1	1	2	1	1	1	1
i	Keep up appearances	Play booty	A bag of bones	One's blood is up	It is all up with him	-	-	-	-	-
x	4	2	1	2	1	-	-	-	-	-

На основі цих даних складаємо варіаційний ряд (таблиця 1.2)

Таблиця 1.2

x_i	0	1	2	3	4	5	6	7	10	13	20
n_i	2	25	8	5	5	3	2	2	1	1	1

1. Для впорядкування цих чисел ці дані подаємо графічно: на числовій вісі відмічаємо результати вимірювань, а над кожним виміром точками відмічаємо, скільки разів зустрілося це число.

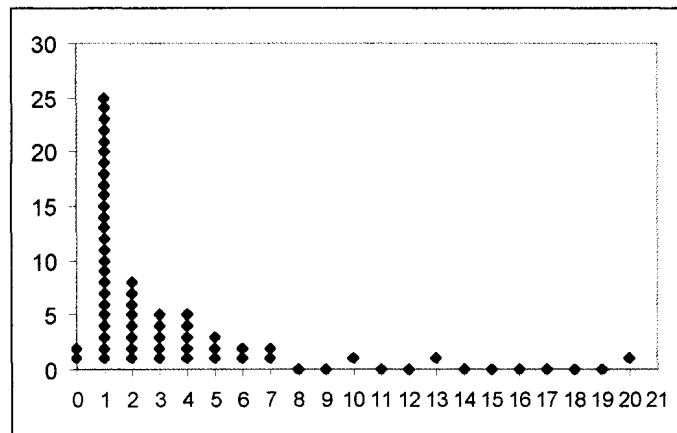


Рисунок 1.1

2. $x_{max} = 20, x_{min} = 0, n = 55$. Обираємо крок за формулою:

$$h = \frac{x_{max} - x_{min}}{1 + 3,32 \lg n} = \frac{20 - 0}{1 + 3,32 \lg 55} = 2,95$$

Беремо крок $h=3$.

3. Отримуємо інтервальний варіаційний ряд:

Таблиця 1.3

I_i	0-3	3-6	6-9	9-12	12-15	15-18	18-21
n_i	40	15	4	1	1	0	1
P_i^*	0,727	0,272	0,072	0,018	0,018	0	0,018
P_i^*/h	0,242	0,090	0,024	0,006	0,006	0	0,006

За отриманими даними будуємо гістограму (рис. 1.2)

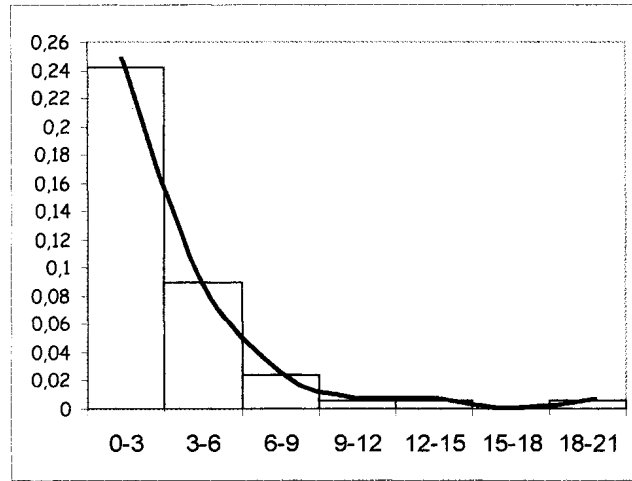


Рисунок 1.2

4. Будуємо приблизну криву щільності статистичного розподілу $f(x)$ (рис. 1.2)

5. Знаходимо статистичну функцію розподілу $F^*(x)$:

$$F^*(0) = 0;$$

$$F^*(3) = 0,727;$$

$$F^*(6) = 0,727 + 0,272 = 0,899;$$

$$F^*(9) = 0,899 + 0,072 = 0,971;$$

$$F^*(12) = 0,971 + 0,018 = 0,989;$$

$$F^*(15) = 0,989 + 0,018 = 1,007;$$

$$F^*(18) = 1,007;$$

$$F^*(21) = 1,007 + 0,018 = 1,025.$$

За цими даними будуємо графік (рис. 1.3):

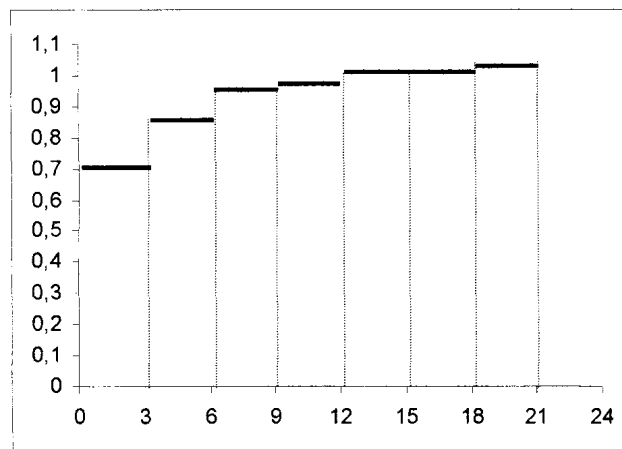


Рисунок 1.3

Charles Dickens "Oliver Twist" (переклад О.В. Кривцова)

Складаємо вибірку об'ємом $n = 42$ (таблиця 2.1).

Таблиця 2.1

i	Спаси и сохрани !	Что за шум ?	Замани вать в ловушку	Как огурчик	Даю голову на отсечение	Отпраздновать труса	Господи помилуй!	Рука руку моет	Заморыш	Навострять лыжи	В тайниках души	Приводить в содрогание	Из огня, да в полымя	Его песенка спета
x	6	2	1	0	4	1	18	0	1	3	2	1	0	5
i	Черт возьми!	С головы до ног	Разгорячиться	Подкрепиться словом делом	Боже мой!	По размять копыта	Кататься как сыр в масле	Чувствовать себя, как дома	Облапошивание младенца	Кровь кипит в жилах	Заморыш	Одурачить	Приходить на ум	Делать вид
x	8	2	1	1	12	1	0	2	3	1	1	1	2	1
i	Застигнуть врасплох	Ломать комедию	Застигнуть врасплох	Быть беде	Решительно и бесповоротно	Перегнуть палку	Поставить в неудобное положение	Уронить авторитет	Нестись, сломя голову	Из одного теста	Делать вид	Непромокает мое сердце	Проложить дорогу	Подложить вой
x	3	1	1	0	5	1	1	2	6	1	0	1	1	1

На основі цих даних складаємо варіаційний ряд (таблиця 2.2).

Таблиця 2.2

x_i	0	1	2	3	4	5	6	8	12	18
n_i	6	19	6	3	1	2	2	1	1	1

1. Для впорядкування цих чисел ці дані подаємо графічно: на числовій вісі відмічаємо результати вимірювань, а над кожним виміром точками відмічаємо, скільки разів зустрілося це число.

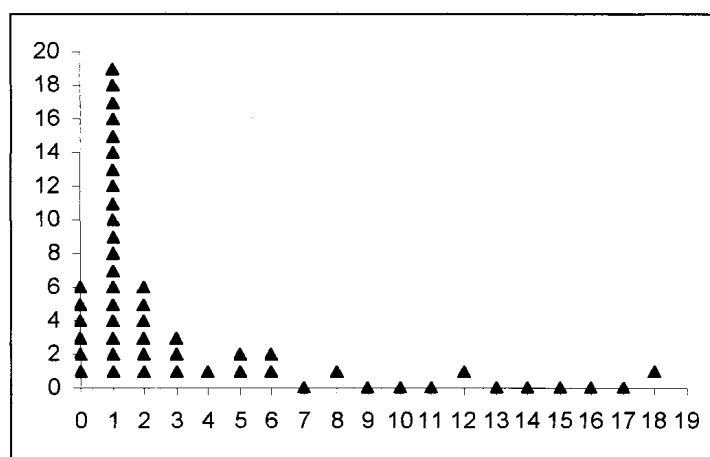


Рисунок 2.1

2. $x_{max} = 18, x_{min} = 0, n = 42$. Обираємо крок за формулою:

$$h = \frac{x_{max} - x_{min}}{1 + 3,32 \lg n} = \frac{18 - 0}{1 + 3,32 \lg 42} = 2,82$$

Беремо крок $h=3$.

3. Отримуємо інтервальний варіаційний ряд:

Таблиця 2.3

I_i	0-3	3-6	6-9	9-12	12-15	15-18
n_i	34	8	3	1	1	1
P_i^*	0,809	0,109	0,071	0,023	0,023	0,023
P_i^*/h	0,269	0,036	0,023	0,007	0,007	0,007

За отриманими даними будуюмо гістограму (рис. 2.2)

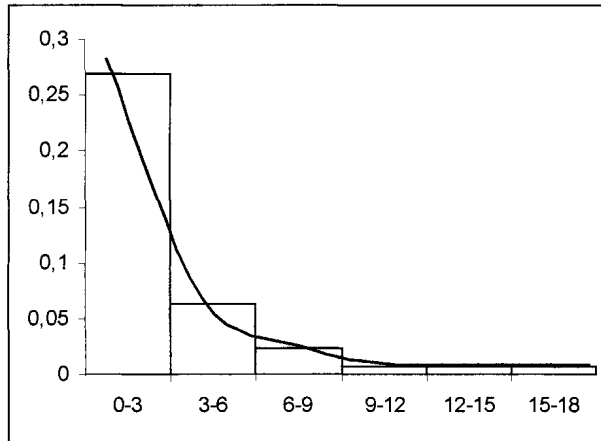


Рисунок 2.2

4. Будуюмо приблизну криву щільності статистичного розподілу $f(x)$ (рис. 2.2).
5. Знаходимо статистичну функцію розподілу $F^*(x)$:

$$F^*(0) = 0;$$

$$F^*(3) = 0,809;$$

$$F^*(6) = 0,809 + 0,109 = 0,918;$$

$$F^*(9) = 0,918 + 0,071 = 0,989;$$

$$F^*(12) = 0,989 + 0,023 = 1,012;$$

$$F^*(15) = 1,012 + 0,023 = 1,035;$$

$$F^*(18) = 1,058$$

За цими даними будуюмо графік (рис. 2.3):

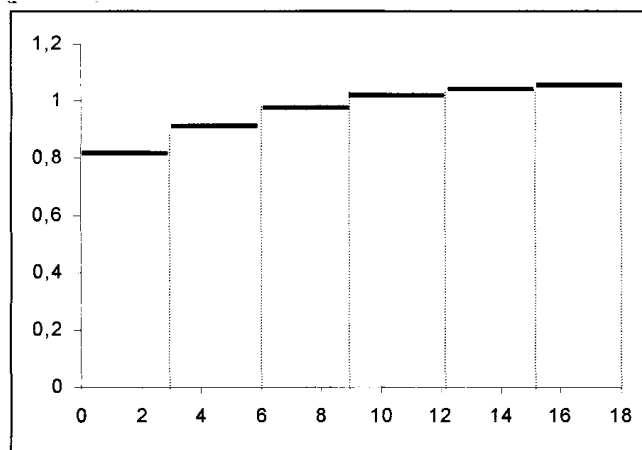


Рисунок 2.3

Отже, порівнюючи результати, отримані після статистичної обробки оригіналу роману та його перекладу, можна зробити певний висновок – переклад, зроблений Кривцовим, є адекватним та релевантним стосовно передачі образності оригіналу (а саме фразеологічних одиниць), частота їх вживання практично повністю збережена (наявне несуттєве зменшення, що може бути пов'язане з відсутністю у мові перекладу певних відповідних фразеологічних одиниць, що відносяться до реалій мови оригіналу), а всі інші показники використання фразеологічних одиниць майже ідентичні. Дані цього експерименту можуть бути використані для подальшого, більш детального статистичного аналізу цього роману, його перекладу, творчості Чарльза Діккенса тощо.

Таким чином, була проаналізована література з питань статистики, лінгвістики, фразеології, були вивчені фразеологічні одиниці в романі Чарльза Діккенса "Олівер Твіст" та його перекладі О.В. Кривцова, була проведена статистична обробка даних, виконаний частковий статистичний аналіз даних по роману та його перекладу, були сформульовані висновки зі ступеню адекватності перекладу цього роману за образним компонентом.

Література

1. *Головин Б.Н. Язык и статистика* / Б.Н. Головин– М.: Просвещение, 1971. – 190 с.
2. *Ермоленко Г.В. Лингвистическая статистика. Краткий очерк* / Г.В. Ермоленко. – М., 1975. – 155 с.

З. Перебийніс В.І. Статистичні методи для лінгвістів: навч. посіб. / В.І. Перебийніс. – Вінниця : Нова книга, 2001. – 168 с.

Ratushnyak Y. I. Statistics methods in linguistics (partial statistical analysis of the phraseological units in the novel "Oliver Twist" by Charles Dickens and its translation by A. V. Krivtsov)

Examples of the statistical methods in linguistics are analyzed. The fact that interconnection of the disciplines results in effective functioning is set. The basic statistical characteristics of the novel "Oliver Twist" by Charles Dickens and its translation by A.V. Krivtsov are defined. The degree of adequacy of the novel translation by its figurativeness component is analyzed.

Key words: *statistics, linguistics, phraseological unit, universal set, sample, variants, interval variation series, relative distribution function.*