Ганна Ситар

УДК 81'373.7:81'32

СТАТИСТИЧНИЙ АНАЛІЗ ПРИСЛІВ'ЇВ І ПРИКАЗОК: ПОКАЗНИК АСОЦІАЦІЇ *MUTUAL INFORMATION* (НА МАТЕРІАЛІ УКРАЇНСЬКОГО НАЦІОНАЛЬНОГО ЛІНГВІСТИЧНОГО КОРПУСУ)[1]

Стаття продовжує цикл публікацій, присвячених статистичному аналізу фразеологічних і фразеологізованих одиниць української мови. У ній з'ясовано ступінь невипадковості поєднання компонентів у складі українських прислів'їв і приказок за допомогою обчислення показника асоціації mutual information (MI).

Отримані результати обчислень для 53 прислів'їв і приказок, виконаних за даними Українського національного лінгвістичного корпусу, доводять, що всі проаналізовані одиниці мають високий ступінь невипадковості поєднання словоформ (MI перебуває в діапазоні від 24,5 до 95,27), що є кількісним підтвердженням стійкості їхнього зв'язку.

Зафіксовано статистично вірогідний зв'язок між кількістю компонентів прислів'я або приказки і величиною показника асоціації MI. Наведені результати загалом узгоджуються із статистичними даними, отриманими на попередніх етапах дослідження для інших типів фразеологічних одиниць – лексичних і синтаксичних фразеологізмів.

Ключові слова: показник асоціації, фразеологічна одиниця, mutual information, прислів'я, приказка, статистика, українська мова.

Постановка проблеми, актуальність дослідження. Сучасна лінгвістика позиціонує корпуснозорієнтованість як необхідну умову мовознавчого дослідження. Статистичний аналіз фразеологічних і фразеологізованих одиниць належить до актуальних завдань лінгвістичної статистики, оскільки за допомогою математичних методів і прийомів покликаний підтвердити або спростувати належність певної мовної одиниці до класу стійких. Процедуру такого аналізу на матеріалі синтаксичних фразеологізмів української мови запропоновано у працях (Sytar, "Statystychni Kryteriyi Analizu Syntaksychnykh Frazeolohizmiv"; Sytar, "Statystychnyi analiz frazeolohizovanykh rechen…"; Sytar, "Syntaksychni frazeolohizmy v rozrizi konstruktsiinoi hramatyky").

Обчислення показників асоціації як метод визначення невипадковості поєднання компонентів може бути застосований для різних типів конструкцій. Цю статтю присвячено аналізу прислів'їв і приказок, які в межах широкого підходу до розуміння обсягу фразеології кваліфікують як один із типів стійких одиниць (В. Л. Архангельський, А. М. Баранов, Д. О. Добровольський, Т. О. Туліна та ін.).

У розумінні прислів'їв і приказок спираємось на усталений в українському мовознавстві погляд, згідно з яким прислів'я визначають як «стійкий вислів переважно фольклорного походження, в якому зафіксований практичний досвід народу та його оцінка різних подій і явищ. Прислів'я на відміну від приказок, − це самостійні судження, граматично та інтонаційно оформлені як прості («Дружній череді і вовк не страшний») або складні («Біда тому волові, котрого корова коле») речення» (Ukrainska mova: Entsyklopediia: 530, автор статті М. Т. Демський). Оскільки в зібраннях прислів'їв і приказок їх подають без розмежування, до статистичного аналізу залучаємо їх разом як групу умовно однорідних одиниць.

Матеріал і методи дослідження. Об'єктом статистичного аналізу стали 53 прислів'я і приказки, дібрані з авторитетних джерел ("Ukrainski prykazky, pryslivia i take inshe"; "Pryslivia ta prykazky"), серед них 9 трикомпонентних, 17 − чотирикомпонентних, 14 − п'ятикомпонетних, 6 − шестикомпонентних і 7 − семикомпонентних одиниць. Вірогідність одержаних кількісних даних забезпечено виконанням обчислень на матеріалі значного за обсягом й індексованого корпусу текстів − Українського національного лінгвістичного корпусу (далі УНЛК) Українського мовно-інформаційного фонду НАН України.

В арсеналі сучасної статистики існує низка статистичних критеріїв (коефіцієнтів), об'єднаних терміном «показники асоціації» (англ. association measures, measures of association). За Кембриджським словником статистики Брайана Еверітта (Brian S. Everitt), «Показники асоціації – числові індекси, що обчислюють силу статистичної залежності двох або більше кваліативних змінних» (Everitt: 241).

У цьому авторитетному лексикографічному виданні термін «асоціація» витлумачено як «загальний термін, що використовується для опису відношення між двома змінними. Значною мірою є синонімічним до кореляції» (Everitt: 20).

У словнику статистики й методів дослідження Американської асоціації психологів (за ред. Sheldon Zedeck) асоціацію визначено як «ступінь статистичної залежності або відношення між двома або більше явищами», а кореляцію – як «ступінь відношення (зазвичай лінійного) між двома змінними, який може бути обрахований як коефіцієнт кореляції, сила асоціації» (APA: 65).

[1] Дослідження виконано в межах фундаментального наукового проекту «Об'єктивна і суб'єктивна мовносоціумна граматика: комунікативно-когнітивний та прагматико-лінгвокомп'ютерний виміри» (0118U003137).

Статистична залежність послідовності словоформ у корпусі визначувана за допомогою показника асоціації *mutual information* (далі MI) (буквально − взаємна, спільна інформація). Поняття MI ввів у теорію інформації Роберт Маріо Фано (Fano). У лінгвістичних дослідженнях його вперше застосували Кеннет Ворд Чарч (Kenneth Ward Church) і Патрік Генкс (Patrick Hanks) (Church, Hanks). Сутність спільної інформації вчені визначили так: «спільна інформація порівнює ймовірність спостереження *x* та *y* разом (поєднана ймовірність) з імовірностями спостереження *x* та *y* незалежно (випадкова)» (Church, Hanks: 23). Відповідно у дослідників мова йшла про невипадковість поєднання двох слів у тексті і про потребу залучення цього методу для лексикографії, укладання конкордансів, вивчення сполучуваності слів та ін.

Оскільки обраний об'єкт дослідження − прислів'я та приказки − є багатокомпонентними (три- і більше) одиницями, постає потреба врахувати у формулі MI більшу кількість компонентів. Тому обчислення здійснено за формулою (1), виведеною у працях (Petrovic, Snajder, Basic, Kolar: 323; Yagunova, Pivovarova: 586).

$$MI = log_2 \frac{f(c_1,c_2,\ldots,c_i) \times N^{(i-1)}}{f(c_1) \times f(c_2) \times \ldots \times f(c_i)}, \quad (1)$$

де MI – коефіцієнт mutual information;

i – це кількість компонентів конструкції;

c_1 – перша лексична одиниця;

c_2 – друга лексична одиниця;

c_i – *i*-а лексична одиниця;

$f(c_1,c_2,\ldots c_i)$ – абсолютна частота вживання конструкції $c_1, c_2, \ldots, c_i$ в корпусі (з урахуванням порядку одиниць усередині конструкції);

$f(c_1)$ – абсолютна частота c_1 в корпусі;

$f(c_2)$ – абсолютна частота c_2 в корпусі;

$f(c_i)$ – абсолютна частота c_i в корпусі;

N – загальна кількість словоформ у корпусі;

log_2 – логарифм числа за основою 2.

Мета цього дослідження – визначити ступінь невипадковості поєднання компонентів у складі українських прислів'їв і приказок за допомогою обчислення показника асоціації MI. Для досягнення поставленої мети розв'язано такі **завдання:**

1) укладено реєстр прислів'їв і приказок, що охоплює одиниці з різною кількістю компонентів і різну тематику;
2) з УНЛК отримано частотні дані для прислів'їв і приказок;
3) виконано обчислення за формулою MI для багатокомпонентних одиниць;
4) проаналізовано отримані результати.

Для коректного встановлення абсолютної частоти конструкції та абсолютної частоти окремих словоформ, що входять до її складу, в пошуковій формі УНЛК було задано визначений порядок словоформ та передбачено пошук словоформи, а не слова з урахуванням його парадигми. Оскільки цей корпус текстів є динамічним, зазначимо, що частотні дані подаємо станом на лютий 2018 року. Загальна кількість слововживань у корпусі в період здійснення підрахунків становила 189 200 000 одиниць.

Покажемо приклад здійснених підрахунків. Для обчислення ступеня невипадковості поєднання словоформ у межах прислів'я *Терпи, козаче, отаманом будеш* з УНЛК було отримано такі кількісні дані: абсолютна частота прислів'я становить 11, абсолютна частота словоформи *терпи* − 228; *козаче* − 307; *отаманом* − 239; *будеш* − 1968. Підставляючи ці дані до формули (1), отримуємо:

$$\text{MI (Терпи, козаче, отаманом будеш)} = log_2 \frac{11 \times (189200000)^3}{228 \times 307 \times 239 \times 1968} = 51{,}007076 \approx 51{,}01.$$

Коефіцієнт MI обраховували з точністю до двох знаків після коми. Отримані результати MI для трикомпонентних прислів'їв і приказок подано в таблиці 1.

Як видно з таблиці 1, коефіцієнт MI для трикомпонентних прислів'їв і приказок перебуває у межах від 24,5 (*Хліб усьому голова*) до 35,34 (*Горбатого могила виправить*).

Таблиця 1

Показник асоціації MI для трикомпонентних прислів'їв і приказок за даними УНЛК

№ з/п	Прислів'я або приказка	Абсолютна частота вживання прислів'я або приказки	Абсолютна частота вживання словоформ-компонентів прислів'я або приказки	Показник асоціації MI
1	*Береженого Бог береже*	38	*береженого 89;* *Бог 2 556;* *береже 569*	33,29
2	*Гол, як сокол*	4	*гол 457;* *як 6 274;* *сокол 53;*	29,81
3	*Голий, як кістка*	2	*голий 920;* *як 6 274;* *кістка 362*	25,03
4	*Горбатого могила виправить*	25	*горбатого 138;* *могила 963;* *виправить 155*	35,34
5	*Змерз, як собака*	2	*змерз 262;* *як 6 274;* *собака 1 465*	24,83
6	*На двох стільцях*	65	*на 6 331;* *двох 4 584;* *стільцях 283*	28,08
7	*Сила солому ломить*	7	*сила 3 032;* *солому 517;* *ломить 200*	29,57
8	*Собака на сіні*	25	*собака 1 465;* *на 6 331;* *сіні 342*	28,07
9	*Хліб усьому голова*	9	*хліб 2 067;* *усьому 2 111;* *голова 3 122*	24,5

Контрольна величина, починаючи від якої вважаємо зв'язок слів невипадковим, залежить від показників абсолютної частоти конструкції, від абсолютної частоти її окремих складників і від розміру корпусу. Для Українського національного лінгвістичного корпусу, розмір якого в лютому 2018 року становив 189 200 000 слововживань, ця контрольна величина становить 7,56 (детально процедуру виведення контрольної величини викладено у праці (Sytar 2017: 310-311)):

$$log_2 189 = 7{,}56377 \approx 7{,}56$$

Відповідно отримані результати можна кваліфікувати як такі, що відбивають високий ступінь невипадковості (зв'язаності) компонентів конструкції, оскільки вони більше ніж утричі перевищують контрольну величину.

Статистичні дані щодо чотири-, п'яти-, шести- й семикомпонентних прислів'їв і приказок наведено в таблицях 2, 3, 4 і 5 відповідно.

Таблиця 2

Показник асоціації MI для чотирикомпонентних прислів'їв і приказок за даними УНЛК

№ з/п	Прислів'я або приказка	Абсолютна частота вживання прислів'я або приказки	Абсолютна частота вживання словоформ-компонентів прислів'я або приказки	Показник асоціації MI
1	*Більшому й більше треба*[2]	0	–	–
2	*Велике дерево поволі росте*	1	*велике 3 063; дерево 2 016; поволі 1 629; росте 1 605*	38,61
3	*Вік живи — вік учись*	19	*вік 2 382; живи 662; учись 100*	48,28
4	*Гарна дівка, як маківка*	5	*гарна 1 697;дівка 613; як 6 274;маківка 125*	45,24
5	*Гусь свині не товариш*	4	*гусь 34;свині 750; не 6 308; товариш 1 421*	46,75
6	*Два українці − три гетьмани*	2	*два 4 780; українці 1 333; три 4 607; гетьмани 267*	40,65
7	*Двічі літа не буває*	2	*двічі 2 140; літа 2 050; не 6 308; буває 2 820*	37,34
8	*Знай, коза, своє стійло*	1	*знай 1 118; коза 523; своє 4 234; стійло 66*	45,24
9	*На двох стільцях сидить*	2	*на 6 331; двох 4 584; стільцях 283; сидить 2 241*	39,42
10	*Народ скаже, як зав'яже*	4	*народ 2 547; скаже 2 038; як 6 274; зав'яже 32*	44,56
11	*Рідна мова — не полова*	1	*рідна 1 371; мова 3 101; не 6 308; полова 210*	40,13
12	*Ситий голодному не вірить*	4	*ситий 499; голодному 220; не 6 308; вірить 1 337*	44,73
13	*Ситий голодному не товариш*	7	*ситий 499; голодному 220; не 6 308; товариш 1 421*	45,45
14	*Терпи, козаче, отаманом будеш*	11	*терпи 228; козаче 307; отаманом 239; будеш 1 968*	51,01
15	*Усі під Богом ходимо*	8	*усі 4 175;під 5 636; Богом 1 619; ходимо 423*	41,61
16	*Хвали мене, моя губонько*	3	*хвали 296;мене 3912; моя 3 214; губонько 6*	49,69
17	*Язик до Києва доведе*	20	*язик 1 642; до 6 281; Києва 1 979; доведе 592*	43,35

[2] У випадку можливої, але не зафіксованої в УНЛК конструкції (абсолютна частота 0), частоти окремих компонентів не наводимо через те, що обчислення MI не має смислу, оскільки логарифму 0 не існує. Обчислення показників асоціації для таких конструкцій не здійснювали, тому у відповідній графі таблиці стоїть знак «–».

Таблиця 3

Показник асоціації MI для п’ятикомпонентних прислів’їв і приказок за даними УНЛК

№ з/п	Прислів’я або приказка	Абсолютна частота вживання прислів’я або приказки	Абсолютна частота вживання словоформ-компонентів прислів’я або приказки	Показник асоціації MI
1	*Баба з воза – коням легше*	6	*баба 1 299; з 6 304;воза 841; коням 330; легше 2 310*	60,35
2	*Всяка пташка своє гніздо знає*	2	*всяка 762; пташка 844; своє 4 234; гніздо 953; знає 3 268*	58,07
3	*Гуртом добре й батька бити*	1	*гуртом 1 046; добре 4 220; й 6 092; батька 2 547; бити 1 552*	53,42
4	*Життя прожити — не поле перейти*	14	*життя 4 809; прожити 965; не 6 308; поле 2 858; перейти 1 968*	56,60
5	*Козак хороший, та немає грошей*	0	–	–
6	*Кому весілля, а курці смерть*	4	*кому 2 995; весілля 1 403; а 6 299; курці 217; смерть 2 986*	58,05
7	*Розуміється, як вовк на зорях*	3	*розуміється 854; як 6 274; вовк 1 221; на 6 331; зорях 218*	58,56
8	*Степ та воля – козацька доля*	4	*степ 953; та 6 216; воля 1 932; козацька 526; доля 2 727*	58,12
9	*Тепер життя панам та котам*	0	–	–
10	*Що село, то й сотник*	4	*що 6 293; село 2 162; то 5 532; й 6 092; сотник 329*	54,92
11	*Як дбаєш, так і маєш*	6	*як 6 274; дбаєш 116; так 6 074; і 6 189; маєш 1 879*	57,05
12	*Яке зіллячко, таке й сім'ячко*	1^2	*яке 4 644; зіллячко 55; таке 4 384; й 6 092; сім'ячко 3*	65,76
13	*Яке їхало, таке й здибало*	7	*яке 4 644; їхало 189; таке 4 384; й 6 092; здибало 26*	63,67
14	*Яке коріння, таке й насіння*	11	*яке 4644; коріння 1436; таке 4384; й 6 092; насіння 1173*	55,91

Таблиця 4

Показник асоціації MI для шестикомпонентних прислів'їв і приказок за даними УНЛК

№ з/п	Прислів'я або приказка	Абсолютна частота вживання прислів'я або приказки	Абсолютна частота вживання словоформ-компонентів прислів'я або приказки	Показник асоціації MI
1	*Боже поможи, а сам не лежи*	4	*Боже 2 257; поможи 430; а 6 299; сам 4 306; не 6 308; лежи 276*	74,16
2	*Дай серцю волю — заведе в неволю*	21	*дай 2 231; серцю 894; волю 2 349; заведе 309; в 6 333; неволю 465*	79,98
3	*Де два українці, там три гетьмани*	14	*де 5 679; два 4 780; українці 1 333; там 4 605; три 4 607; гетьмани 267*	73,81
4	*Під лежачий камінь вода не тече*	20	*під 5 636; лежачий 98; камінь 1 939; вода 2 814; не 6 308; тече 1 286*	77,39
5	*Що можна лялі, не можна мамі*	1	*що 6 293; можна 5 520; лялі 41; не 6 308; мамі 757*	72,45
6	*Як неділя, то й сорочка біла*	1	*як 6 274; неділя 739; то 5 532; й 6 092; сорочка 1 009; біла 1 808*	69,53

Таблиця 5

Показник асоціації MI для семикомпонентних прислів'їв і приказок за даними УНЛК

№ з/п	Прислів'я або приказка	Абсолютна частота вживання прислів'я або приказки	Абсолютна частота вживання словоформ-компонентів прислів'я або приказки	Показник асоціації MI
1	*Береженого Бог береже, а козака шабля стереже*	0	–	–
2	*Гол, як сокол, а гострий, як бритва*	2	*гол 457; як 6 274; сокол 53; а 6 299; гострий 1 425; бритва 225*	95,27
3	*Коли маєш сто кіп, то будеш піп*	1	*коли 5 545; маєш 1879; сто 2 631; кіп 152; то 5 532; будеш 1 968; піп 539*	90,6
4	*Коли убогому жениться, то й ніч мала*	1	*коли 5 545; убогому 97; жениться 173; то 5 532; й 6 092; ніч 2 985; мала 3 642*	90,12
5	*Пан з паном, а Іван з Іваном*	2	*пан 2 056; з 6 304; паном 916; а 6 299; Іван 2 018; Іваном 701*	86,83
6	*Під лежачий камінь і вода не тече*	5	*під 5 636; лежачий 98; камінь 1 939; і 6 189; вода 2 814; не 6 308; тече 1 286*	90,29
7	*Що вільно панові, те не можна Іванові*	0	–	–

Дані, наведені в таблицях 2–5 дають змогу констатувати, що коефіцієнт MI для чотирикомпонентних прислів'їв і приказок перебуває в межах від 37,34 (*Двічі літа не буває*) до 51,01 (*Терпи, козаче, отаманом будеш*), тобто в 5−6 разів більший за 7,56; для п'ятикомпонентних – від 53,42 (*Гуртом добре й батька бити*) до 65,76 (*Яке зіллячко, таке й сім'ячко*), тобто в 7–8 разів більший за контрольну величину; для шестикомпонентних − від 69,53 (*Як неділя, то й сорочка біла*) до 79,98 (*Дай серцю волю – заведе в неволю*), тобто в 9–10 разів вищий від контрольної величини; для семикомпонентних − від 86,83 (*Пан з паном, а Іван з Іваном*) до 95,27 (*Гол, як сокол, а гострий, як бритва*), тобто більший в 11−12 разів за контрольну величину.

Висновки. Отримані результати обчислень для 53 прислів'їв і приказок, виконаних за даними Українського національного лінгвістичного корпусу, доводять, що всі проаналізовані одиниці мають високий ступінь невипадковості поєднання словоформ: коефіцієнт MI перебуває в діапазоні від 24,5 до 95,27 (тобто є

втричі – удванадцятеро більшим, ніж контрольна величина), що є кількісним підтвердженням стійкості зв'язку словоформ у складі відповідних одиниць.

Зафіксовано статистично вірогідний зв'язок між кількістю компонентів прислів'я / приказки й величиною показника асоціації MI. Так, для трикомпонентних одиниць результат MI становить від 24,5 (*Хліб усьому голова*) до 35,34 (*Горбатого могила виправить*); для чотирикомпонентних – від 37,34 (*Двічі літа не буває*) до 51,01 (*Терпи, козаче, отаманом будеш*); для п'ятикомпонентних від 53,42 (*Гуртом добре й батька бити*) до 65,76 (*Яке зіллячко, таке й сім'ячко*); шестикомпонентних − від 69,53 (*Як неділя, то й сорочка біла*) до 79,98 (*Дай серцю волю – заведе в неволю*); семикомпонентних − від 86,83 (*Пан з паном, а Іван з Іваном*) до 95,27 (*Гол, як сокол, а гострий, як бритва*).

Серед нерозв'язаних на сьогодні проблем статистичного аналізу варто відзначити омонімію, зокрема, потребу залучення людини-експерта для розмежуванням випадків типу *гол* (пор. *перший гол* і *гол, як сокол*) або *на двох стільцях* (пор. стійка сполука *сидіти на двох стільцях* і вільний (нефразеологізований) вияв у реченні *Посеред кімнати* на двох стільцях *стоїть маленька з сірої бляхи ванночка* (В. Винниченко. Записки Кирпатого Мефістофеля).

Наведені результати загалом узгоджуються із статистичними даними, отриманими на попередніх етапах дослідження для інших типів фразеологічних одиниць – лексичних і синтаксичних фразеологізмів. Водночас вони засвідчують вищий ступінь невипадковості поєднання словоформ саме для прислів'їв і приказок, що є, очевидно, наслідком їхньої багатокомпонентності, стійкості й цілісності їхнього сприйняття носіями української мови.

Перспективним вважаємо статистичний аналіз інших типів стійких одиниць і зіставлення відповідних даних з результатами, отриманими для синтаксичних і лексичних фразеологізмів, прислів'їв і приказок.

References

APA Dictionary of Statistics and Research Methods. Sheldon Zedeck, PhD, editor in chief. Washington, DC: American Psychological Association, 2014. Print.

Church, Kenneth Ward, and Patrick Hanks. "Word Association Norms, Mutual Information, and Lexicography". *Computational Linguistics* 16(1) (1990): 22–29. Print.

Everitt, B. S. *The Cambridge Dictionary of Statistics*. 2nd edition. Cambridge: Cambridge University Press, 2002. Print.

Fano, Robert M. Transmission of *Information*: A *Statistical Theory* of *Communications.* The Technology Press, M.I.T., and John Wiley & Sons, Inc., New York, 1961. Print.

Petrovic, S., Snajder, J., Basic, B. D., Kolar, M. "Comparison of collocation extraction for document indexing". *Journal of Computing and information technology*, 14 (4) (2006): 321–327. Print.

Sytar, Hanna. "Statystychni Kryteriyi Analizu Syntaksychnykh Frazeolohizmiv (Statistical Criteria of Analysis of Syntactic Idioms)." *Visnyk Donets'koho Natsional'noho Universytetu. Seriya B. Humanitarni Nauky* (*The Bulletin of Donetsk National University. Series B. Humanities*) 1–2 (2015): 245–256. Print.

Sytar, Hanna. "Statystychnyi analiz frazeolohizovanykh rechen: pokaznyk asotsiatsii *mutual information* (Statistical Analysis of Sentences with Phraseological Structures: Association Measure of *Mutual Information*)". Ukrainske movoznavstvo (Ukrainian Linguistics). 1(46) (2016): 103–125. Print.

Sytar, Hanna. *Syntaksychni frazeolohizmy v rozrizi konstruktsiinoi hramatyky (Syntactic Idioms in the Context of Construction Grammar)*. Vinnytsya: TOV «Nilan-LTD», 2017. Print.

Ukrainska mova: Entsyklopediia (Ukrainian language: Encyclopedia). Redkol.: Rusanivskyi V. M. (spivholova), Taranenko O. O. (spivholova), Ziabliuk M. P. ta in. 2-he vyd., vypr. i dop. Kyiv: Vyd-vo "Ukrainska entsyklopediia" im. M. P. Bazhana, 2004. Print.

Yagunova, Ye. V., Pivovarova, L. M. "Ot kollokatsiy k konstruktsiyam (From Collocations to Constructions)". *ACTA LINGUISTICA PETROPOLITANA. Works of the Institute of Linguistic Researches of RAS, Russkiy yazyk: grammatika konstruktsiy i leksiko-semanticheskie podkhody* (*The Russian Language: Construction Grammar and Lexical and Semantic Approaches*): X, part 2. (2014) 568–617. Print.

List of Sources

Pryslivia ta prykazky (Proverbs and Sayings). Ukl. M. Paziak. Red. Myshanych S. V.; red. Berezovskyi I. P.; red. Hordiichuk M. M.; red. Zubkov S. D.; red. Sushko L. D.; red. Kuz V. P. Kyiv: Nauk. dumka, 1991. Print.

Ukrainski prykazky, pryslivia i take inshe (Ukrainian Sayings, Proverbs etc*).* Ukl. M. Nomys. Kyiv: Lybid, 1993. Print.

Надійшла до редакції 20 березня 2018 року.

STATISTICAL ANALYSIS OF PROVERBS AND SAYINGS: ASSOCIATION MEASURE OF *MUTUAL INFORMATION* (ON MATERIAL OF UKRAINIAN NATIONAL LINGUISTIC CORPUS)

Hanna Sytar

Department of General and Applied Linguistics and Slavonic Philology, Vasyl' Stus Donetsk National University, Vinnytsia, Ukraine

Abstract

Background: The author examines the statistical analysis of proverbs and sayings on the material of the Ukrainian National Linguistic Corpus of Ukrainian Lingua-Information Fund, NAS of Ukraine. Corpus-oriented statistical research of Ukrainian proverbs and sayings has not been carried out yet. The object of the analysis is 53 proverbs and sayings, selected from authoritative sources: Ukrainski prykazky, pryslivia i take inshe (Ukrainian proverbs, sayings etc) / Ukl. M.Nomys. Kyiv: Lybid, 1993; Pryslivia ta prykazky (Proverbs and sayings)/ ukl. M. Paziak Kyiv: Nauk. dumka, 1991. Among them, there are 9 three-component, 17 − four-component, 14 – five-component, 6 – six-component and 7 – seven-component units.

Purpose: The purpose of this study is to determine the degree of non-randomness of the components combination in Ukrainian proverbs and sayings by means of the calculating the association measure of *mutual information* (hereinafter MI).

Results: The obtained results of the calculations for 52 proverbs and sayings, done with the help of the Ukrainian National Linguistic Corpus data, prove that all analysed units have high degree of non-randomness combination of word forms (MI is in the range from 24.5 to 95.27), which is a quantitative confirmation of their connection stability. The reference value, from which the connection of words is considered to be non-random, depends not only on the indicators of the absolute frequency of the construction and its individual components but also on the size of the corpus. For the Ukrainian National Linguistic Corpus, which consisted of 189 200 000 words in February 2018, the reference value is 7.56.

There is a statistically probable connection between the number of components in the proverb or saying and the size of association measure *mutual information.* Thus, for the three-component units the result of the MI is from 24.5 (*Хліб усьому голова / Khlib usomu holova*)) to 35.34 (*Горбатого могила виправить / Horbatoho mohyla vypravyt*); for four-component units – from 37.34 (*Двічі літа не буває / Dvichi lita ne buvaie*)) to 51.01 (*Терпи, козаче, отаманом будеш / Terpy, kozache, otamanom budesh*); for five-component from 53.42 (*Гуртом добре й батька бити / Hurtom dobre y batka byty*)) to 65.76 (*Яке зіллячко, таке й сім'ячко / Yake zilliachko, take y simiachko*); for six-component – from 69,53 (*Як неділя, то й сорочка біла / Yak nedilia, to y sorochka bila*) to 79.98 (*Дай серцю волю – заведе в неволю / Dai sertsiu voliu – zavede v nevoliu*); for seven-component – from 86.83 (*Пан з паном, а Іван з Іваном / Pan z panom, a Ivan z Ivanom*) to 95.27 (*Гол, як сокол, а гострий, як бритва / Hol, yak sokol, a hostryi, yak brytva*).

Discussion: The presented results are broadly consistent with the statistical data obtained in the previous stages of the study of other types of phraseological units – lexical and syntactic idioms. At the same time, they prove the higher degree of non-randomness of the combination of word forms for proverbs and sayings, which, obviously, are the consequence of their multicomponent, as well as the stability and integrity of their perception by the speakers of the Ukrainian language. It is considered perspective to involve other types of phraseological units to the statistical analysis.

Keywords: association measure, phraseological units, mutual information, proverb, saying, statistics, the Ukrainian language.

Vitae

Hanna Sytar is PhD of Philology, Assosiate Professor, Assosiate Professor of Department of General and Applied Linguistics and Slavonic Philology at Donetsk National University named after Vasyl Stus. Her areas of research interests include syntax, semantics, pragmatics, construction grammar, applied linguistics.

Correspondence: h.v.sytar@donnu.edu.ua