КОНСТРУКЦІЙНА ГРАМАТИКА

УДК 81'367:81'373.7

Ганна Ситар, канд. філол. наук, доц., докторант
Донецький національний університет, Вінниця

СТАТИСТИЧНИЙ АНАЛІЗ ФРАЗЕОЛОГІЗОВАНИХ РЕЧЕНЬ: ПОКАЗНИК АСОЦІАЦІЇ *MUTUAL INFORMATION*

Статтю присвячено статистичному аналізу фразеологізованих речень української мови. Обґрунтовано доцільність застосування статистичного критерію mutual information для встановлення коефіцієнта невипадковості певної послідовності слів у тексті.

Наведено результати обчислення mutual information для моделей фразеологізованих речень за даними Українського національного лінгвістичного корпусу. Доведено, що всі проаналізовані моделі речень мають високий ступінь невипадковості компонентів, що входять до складу незмінної частини речення. Для обстежуваних одиниць запропоновано обчислення модифікованого показника MI – MI^3 (MI – mutual information).

Зіставлено отримані дані з відповідними показниками MI та MI^3 для лексичних фразеологізмів та нефразеологізованих речень. Виділено чинники, які впливають на коректність здійснених підрахунків.

Ключові слова: *конструкція, конструкційна граматика, корпус текстів, синтаксичний фразеологізм, статистичний аналіз, показник асоціації mutual information, показник асоціації MI^3, українська мова, фразеологізоване речення.*

Фразеологізовані речення (або синтаксичні фразеологізми) – особливий тип речення, у якому пов'язані ідіоматично постійний (незмінний) і змінний компоненти мають фіксовані позиції, граматичні зв'язки і прямі лексичні значення слів послаблені або втрачені на сучасному етапі розвитку мови. Такі речення є впливовим засобом вираження ставлення мовця до висловлюваного, характерним для розмовного мовлення, текстів художнього та публіцистичного стилів [Балобанова; Величко; Всеволодова, Лим Су; Русская; Шмелёв] (докладно про

ознаки та статус синтаксичних фразеологізмів у системі мовних одиниць див. у праці [Ситар 2011]).

Фразеологізовані речення переважно не містять дієслівних предикатів як організаційних центрів, вони формуються навколо поєднання службових і повнозначних компонентів, яким властиве семантичне спустошення або семантичний зсув (*чим не, що за, оце так, от тобі/вам і/й, яке там, куди там* і под.: *Чим не відповідь! Що за книга! Оце так подарунок! Яке там встигли!*), або навколо поєднання повторюваних повнозначних і службових компонентів (*як, так, і, а, не* і под.: *Дівчина як дівчина*) та компонентів-зв'язок (*Закон є закон*).

У світлі ідей конструкційної граматики (Construction Grammar) фразеологізоване речення кваліфікуємо як один із типів конструкцій – мовного знака, у якому певний аспект плану вираження або плану змісту не можна пояснити, спираючись на форму або зміст його компонентів [Fillmore; Fillmore, Kay, O'Connor; Goldberg 1995; Goldberg 2003]. Відповідно фразеологізовані речення вважаємо некомпозиційними синтаксичними одиницями з виразним прагматичним спрямуванням (докладніше див. [Ситар 2015]).

Статистичний етап будь-якого дослідження передбачає одержання кількісних даних, які підтверджують правильність зроблених припущень або спростовують їх. Під час виконання теоретичної частини дослідження фразеологізованих речень ми сформулювали дві робочі гіпотези, які потребують перевірки за допомогою статистичного аналізу:

1. Фразеологізовані моделі речень, як і будь-які інші стійкі одиниці, мають високий ступінь невипадковості поєднання компонентів, що входять до складу незмінної частини речення.

2. Оскільки існує взаємозв'язок між якісними ознаками та кількісними параметрами мовних одиниць (за Б. Головіним, В. Левицьким, В. Перебийніс та ін.), припускаємо, що показники невипадковості появи двох і більше компонентів відрізняються для таких типів мовних одиниць, як лексичні фразеологізми, нефразеологізовані речення і фразеологізовані речення (синтаксичні фразеологізми).

Актуальність дослідження вмотивована потребою здійснення статистичного аналізу фразеологізованих речень на матеріалі корпусу текстів як значного і реперезентативного мовного матеріалу, що разом із вибором адекватних статистичних показників, застосовуваних до відповідних типів мовних одиниць, забезпечує вірогідність отриманих даних.

Українська лінгвостатистика має значні здобутки в статистичному аналізі окремо взятих лексем та груп слів, а також текстів певних стилів (наукового, публіцистичного, художнього та ін.) (див. праці В. Перебийніс, Н. Дарчук, В. Левицького, С. Бук та ін. дослідників), проте майже не звертається до статистичного аналізу поєднань слів (словосполучень і речень).

Проведений аналіз наукової літератури [Залесская; Хохлова; Ягунова, Пивоварова; Church, Hanks; Evert; Petrovic; Seretan; Stubbs 1995] засвідчує, що найбільш часто застосовуваними і вже апробованими на матеріалі переважно англійської та російської мов є три показники асоціації (англ. association measure), за допомогою яких визначають статус мовної одиниці, що складається з кількох слів (лінгвісти на їх позначення вживають переважно терміни "конструкція" та "колокація"): mutual information, t–score та log–likelihood.

М. Хохлова визначає колокацію як "статистично стійке словосполучення" [Хохлова : 10], тобто поєднання слів, стійкість якого підтверджена відповідними кількісними показниками.

Мета статті – здійснити статистичний аналіз моделей фразеологізованих речень української мови шляхом визначення показника асоціації mutual information. Поставлена мета передбачає розв'язання таких завдань: 1) розглянути сутність статистичного критерію mutual information та обґрунтувати доцільність його застосування для аналізу фразеологізованих речень в українській мові; 2) обчислити показник mutual information для низки моделей фразеологізованих речень за даними Українського національного лінгвістичного корпусу; 3) зіставити отримані дані з відповідними показниками для лексичних фразеологізмів та нефразеологізованих речень; 4) визначити чинники, які впливають на коректність здійснених підрахунків.

Поняття mutual information (англ. mutual information – взаємна, спільна, повна інформація, далі – MI) запропонував відомий американський учений італійського походження Р. М. Фано у праці з теорії інформації "Transmission of Information: A Statistical Theory of Communications" ("Передача інформації: статистична теорія комунікацій") [Fano].

Mutual information (або MI-score) – це коефіцієнт, який відбиває невипадковість (залежність) певної послідовності слів у тексті. У лінгвістичний обіг формулу (1) для обчислення MI ввели американські дослідники К. В. Чарч та П. Хенкс [Church, Hanks : 23]:

$$I(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)}, \qquad (1)$$

де $I(x,y)$ – взаємна інформація;
x – перше слово;
y – друге слово;
$P(x,y)$ – імовірність поєднання слів x та y;
$P(x)$ – імовірність слова x;
$P(y)$ – імовірність слова y;
$\log_2$ – логарифм числа за основою 2 (двійковий логарифм).

Учені пояснюють сутність спільної інформації так: "<…> спільна інформація порівнює імовірність спостереження x та y разом (поєднана імовірність) з імовірностями спостереження x та у незалежно (випадкова). Якщо наявна справжня асоціація між x та y, тоді поєднана імовірність $P(x,y)$ буде значно більшою, ніж випадкова $P(x)P(y)$ та відповідно $I(x,y) >> 0$. Якщо нема жодних цікавих відношень між x та у, тоді $P(x,y) \approx P(x)P(y)$ і, таким чином, $I(x,y) \approx 0$. Якщо x та у перебувають у доповнювальній дистрибуції, тоді $P(x,y)$ буде значно меншим, ніж $P(x)P(y)$, витворюючи $I(x,y) << 0$" [Church, Hanks 1990 : 23].

Величину P (англ. probability – імовірність) автори визначають за допомогою встановлення f (англ. frequency – частота) та здійснення нормалізації, потреба у якій зумовлена отриманими підрахунками на матеріалі кількох корпусів різного розміру: "<…> імовірності слів $P(x)$ та $P(y)$ оцінюються через підрахунок

числа спостережень x та y в корпусі, *f(x)* та *f(y)*, та нормалізацію через *N* – розмір корпусу" [Church, Hanks : 23]. При цьому підрахунки ймовірностей здійснюють окремо для всіх можливих послідовностей слів (у цьому випадку – *x y* та *y x*).

З теорії ймовірності відомо, що ймовірність події (у нашому випадку вживання кількох слів разом) обраховують за формулою (2):

$$P(x) = \frac{f(x)}{N}, \tag{2}$$

де *P(x)* – імовірність *x*;
f(x) – частота *x*;
N – кількість усіх можливих уживань *x*.

Підставивши формулу (2) до формули (1), отримуємо:

$$MI(x, y) = \log_2 \frac{P(x, y) \times N}{f(x) \times f(y)}, \tag{3}$$

де *MI* – коефіцієнт mutual information;
x – перша лексична одиниця;
y – друга лексична одиниця;
f(x,y) – абсолютна частота вживання біграми *xy* в корпусі (з урахуванням порядку одиниць усередині біграми);
f(x) – абсолютна частота *x* у корпусі;
f(y) – абсолютна частота *y* в корпусі;
N – загальна кількість словоформ у корпусі;
$\log_2$ – логарифм числа за основою 2.

На матеріалі російської мови коефіцієнт MI застосувала М. Хохлова для з'ясування закономірностей синтаксичної поєднуваності слів у російській мові [Хохлова : 12]. Л. Пивоварова та О. Ягунова за допомогою цього показника здійснили спробу розмежувати та побудувати шкалу колокацій і конструкцій у російськомовних наукових та публіцистичних текстах [Ягунова, Пивоварова : 582–583].

Цікаво, що в праці К. В. Чарча та П. Хенкса число N не введене до формули обчислення MI, проте при поданні та інтерпретації результатів кількісних даних учені послідовно вказують обсяг обстеженого корпусу, що засвідчує важливість цього показника для дослідження [Church, Hanks : 24 і далі].

Щодо обрахунків коефіцієнта MI варто наголосити на кількох важливих моментах. По-перше, при обчисленні MI необхідно враховувати порядок слів у тексті [Ягунова, Пивоварова: 582], тобто брати показник частоти конструкції із заданим порядком компонентів.

По-друге, значення MI може варіюватися залежно від обсягу обстежуваного корпусу слововживань. Установлено таку закономірність: чим більшим є обсяг корпусу, тим вищими є отримувані значення MI [Evert; Stubbs; Ягунова, Пивоварова; Хохлова].

По-третє, формулу (3) використовують саме для двокомпонентних сполучень слів, а для триграм (трикомпонентних конструкцій) S. Petrovic, J. Snajder, B. D. Basic, M. Kolar до формули (1) вводять третій компонент *z*, унаслідок чого отримуємо формулу (4) [Petrovic, Snajder, Basic, Kolar : 323]:

$$I(x,y,z) = \log_2 \frac{P(x,y,z)}{P(x) \times P(y) \times P(z)}, \quad (4)$$

З формули (4) можна вивести формулу (5) для конструкцій із кількістю компонентів, більшою ніж 2 ($i > 2$) (формулу (5) подаємо за [Ягунова, Пивоварова : 586]):

$$MI = \log_2 \frac{f(c_1,c_2,...,c_i) \times N^{(i-1)}}{f(c_1) \times f(c_2) \times ... \times f(c_i)}, \quad (5)$$

де *MI* – коефіцієнт mutual information;
i – 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.

По-четверте, постає питання, як саме інтерпретувати отримані підрахунки. М. Стаббс цілком слушно зауважує: “<…> якими б не були спродуковані комп’ютером квантитативні знахідки або статистика, вони мають бути інтерпретовані людиною-аналітиком” [Stubbs 2001 : 75]. Кількісно підтверджену невипадковість кількох слів у тексті Л. Пивоварова та О. Ягунова кваліфікують як “непряму ознаку наявності стійкого семантичного і/або синтаксичного зв’язку між словами” [Ягунова, Пивоварова : 610], тобто як кількісний показник вірогідності виділення певної конструкції як окремої одиниці. Щодо інтерпретації значень MI дослідниці зазначають: “<…> якщо слова цілком незалежні, то ймовірність їхньої спільної появи дорівнює добутку ймовірностей появи кожного з них, тобто добутку частот, а значення показника MI дорівнює нулю” [Ягунова, Пивоварова : 583]. У цьому питанні вони спираються на думку К. В. Чарча та П. Хенкса, які визначають коефіцієнт $MI \approx 0$ як “нецікавий” для дослідження, $MI > 3$ відповідно як “цікавий” [Church, Hanks : 24]). Іншими словами, невипадковість поєднання лексем фіксуємо при $MI > 3$.

Статистичний аналіз синтаксичних фразеологізмів ми здійснювали за даними Українського національного лінгвістичного корпусу (далі УНЛК), створеного колективом Українського мовно-інформаційного фонду НАН України і розміщеного за адресою: http://unlc.icybcluster.org.ua/virt_unlc/[3]. Під час установлення абсолютних частот конструкції та її окремих складників у пошуковій формі УНЛК було задано визначений порядок словоформ і передбачено пошук саме словоформи (а не слова з усією можливою парадигмою) для отримання коректного результату. Оскільки цей корпус текстів є динамічним, зазначимо, що частотні дані подаємо станом на жовтень 2015 року (версія 5.4.24.1). Загальна кількість слововживань у період здійснення підрахунків становила 180 мільйонів слововживань.

[3] Дякуємо Директорові Українського мовно-інформаційного фонду НАН України академіку НАН України В. Широкову за наданий доступ до корпусу.

До аналізу залучено моделі фразеологізованих речень, з-поміж яких є дво-, три- і чотиричленні незмінні компоненти моделі, та враховано (обраховано окремо) варіанти моделі (наприклад, числові або родові модифікації). Обчислення показника асоціації MI здійснювали за формулами (3) для двочленних і (5) для багаточленних незмінних компонентів моделей синтаксичних фразеологізмів. Фразеологізовані речення з одночленним постійним компонентом (***Який*** *чоловік! Жінка є жінка. Україна –* ***не*** *Росія*) не залучалися до статистичного аналізу за показником MI.

Коефіцієнт MI обраховуємо з точністю до чотирьох знаків після коми, оскільки, як було з'ясовано під час виконання обчислень, саме така точність потрібна для коректності відбиття деяких інших показників асоціації (зокрема, t-score). Дужками позначено факультативність компонента моделі. Отримані результати для 10 різнотипних фразеологізованих моделей (за частиномовним статусом змінного й незмінного компонентів моделі, кількісним складом незмінного компонента, наявністю варіантів моделі та продуктивністю) подаємо в таблиці 1.

Таблиця 1

Показник асоціації MI для моделей синтаксичних фразеологізмів (СФ) в українській мові

№ з/п	Модель СФ	Абсолютна частота вживання незмінного компонента СФ в УНЛК	Абсолютна частота вживання словоформ, що входять до незмінного компонента СФ, в УНЛК	Показник асоціації MI
1	*Ати-бати, йшли* N_1 Cop_f /Inf/*в* N_4	0	*ати-бати* 6 *йшли 1840*	–
2	*Де (вже) там* N_1 Cop_f/Inf/Adj/Adv	*де там* 552	*де 4273* *там 3270*	12,7973
		де вже там 59	*вже 3876*	28,398
3	*Ну і/й* N_1 Cop_f	*ну і 823*	*ну 2423і 4731*	13,6588
		ну й 897	*й 4674*	13,8

Закінчення табл. 1

№ з/п	Модель СФ	Абсолютна частота вживання незмінного компонента СФ в УНЛК	Абсолютна частота вживання словоформ, що входять до незмінного компонента СФ, в УНЛК	Показник асоціації MI
4	*Оце так* N_1 Cop_f	400	*оце 1548* *так 4676*	13,2814
5	*Теж мені* N_1 Cop_f	201	*теж 2680* *мені 2582*	12,3535
6	*Чим не* N_1 Cop_f	478	*чим* 2938 *не* 4844	12,5621
7	*Що (ж це) за* N_1 Cop_f	*що за* 2646	*що 4843* *за 4831*	14,3143
		що це за 738	*це 4593*	27,7322
		що ж це за 206	*ж 4240*	41,2658
8	*Яке (вже/ж) там* N_1 Cop_f /Inf/Adv	*яке там 218*	*яке 3613* *там 3326*	11,6741
		яке вже там 17	*вже 3946*	23,4714
		яке ж там 7	*ж 4240*	22,088
9	*Який там* N_1 Cop_f	*який там 358*	*який 4345* *там 3326*	12,1239
		яка там 397	*яка 4326*	12,2791
		які там 474	*які 4282*	12,5498
10	N_1 (*він*) *і в Африці* N_1	*і в Африці* 30	*і* 4731 *в* 4864 *Африці* 510	26,3059
		він і в Африці 5	*він 4009*	39,1767
		вона і в Африці 5	*вона 4137*	39,1316
		воно і в Африці 0	–	–
		вони і в Африці 3	*вони 4366*	38,3166

Отримані результати дають змогу зробити такі висновки. Для всіх обстежених синтаксичних фразеологізмів показник МІ відбиває високий ступінь (МІ >> 3) невипадковості поєднання словоформ, що є кількісним підтвердженням стійкості зв'язку словоформ у складі незмінних компонентів фразеологізованих моделей речень.

Діапазон варіювання показника МІ для різних моделей з однаковою кількістю словоформ у складі незмінного компонента є невеликим. Так, для синтаксичних фразеологізмів з двочленним незмінним компонентом МІ перебуває в межах від 11,6741 (*яке там*) до 14,3143 (*що за*). МІ для трикомпонентних моделей є приблизно вдвічі вищим, коливається в межах від 22,088 (*яке ж там*) до 28,398 (*де вже там*). МІ для чотирикомпонентних моделей є ще вищим – від 38,31 (*вони і в Африці*) до 41,265 (*що ж це за*). Відповідно зафіксовано статистично вірогідний зв'язок між кількістю компонентів конструкції і величиною показника МІ. При цьому цікаво, що чим більшою є кількість компонентів, тим меншою є частота конструкції, водночас тим більшим є коефіцієнт МІ (саме через врахування абсолютної частоти більшої кількості словоформ).

Обстеження варіантів моделей фразеологізованих речень демонструє таку закономірність: родові або числові варіанти моделі мають близькі показники МІ (відмінності стосуються першого знака після коми, тому це не впливає на значення цілого числа), що є аргументом на користь їхньої кваліфікації як варіантів моделей, а не окремих моделей. Водночас варіантам моделей речення, пов'язаним з уведенням часток до складу незмінного компонента, тобто зі збільшенням кількості компонентів, властивий значно більший показник асоціації МІ.

Коефіцієнт МІ запропоновано в теорії інформації, і з погляду цієї теорії немає значення, що саме ми обраховуємо. Проте практика здійснення статистичних досліджень у лінгвістиці свідчить, що виконання таких обчислень для мовних одиниць різного типу має специфіку, зумовлену їхнім неоднаковим статусом і механізмом утворення.

У науковій літературі неодноразово наголошено, що показник MI виявляється невиправдано великим для поєднань слів із дуже низькою частотою (наприклад, частотою 1) [Evert; Seretan; Stubbs 1995; Залесская; Ягунова, Пивоварова; Хохлова]. Щоб уникнути підвищення значущості таких конструкцій, дослідники пропонують кілька модифікованих формул коефіцієнта MI – MI^2, MI^3, salience і под. Усі вони в той чи інший спосіб підвищують значущість частоти конструкції.

Щодо правомірності введення альтернативних формул В. Серетан (V. Seretan) зазначає: "Зрештою не існує єдиного показника, що міг би бути запропонований для всіх цілей. Фокус новітніх досліджень було зсунуто з порівняння окремих якостей стандартно використовуваних ПА (показників асоціації) до відкриття альтернативних не так широко розповсюджених ПА, що можуть надати оптимальні результати в певних умовах <…>" [Seretan : 43]. Ш. Еверт виділяє як окрему групу статистичних показників такі, що є похідними від базових, та називає їх "евристичними формулами" (грецьк. *ευρίσκω (heuristiko)* – знаходжу, відшукую, відкриваю) [Evert : 77, 89–91].

Б. Делль здійснила експериментальну перевірку модифікацій MI^k для k від 2 до 10 і дійшла до висновку, що саме MI^3 дає оптимальні результати, "хороший компроміс між урахуванням занадто рідкісної події та її нехтуванням" [Daille 1994 : 139]:

$$MI^3 = \log_2 \frac{(O_{11})^3}{E_{11}}, \tag{6}$$

де MI^3 – коефіцієнт MI^3,
O_{11} – спостережувана частота;
E_{11} – очікувана частота.

Відповідно MI^3 обраховуємо за формулою (7):

$$MI^3(x,y) = \log_2 \frac{f^3(x,y) \times N}{f(x) \times f(y)}, \tag{7}$$

де MI^3 – коефіцієнт MI^3, а решта компонентів ідентична до наведених у формулі (3).

Результати здійснених підрахунків відбито в таблиці 2, для наочності зіставлення подано й показник MI.

На думку Ш. Еверта, "MI^k Делль є простим прикладом **параметричного показника асоціації** [*виділення Ш. Еверта – Г.С.*]. Значення параметра k може бути обране вільно (у принципі можливий будь-який $k > 0$), для того щоб модифікувати властивості показника. У такий спосіб може бути можливим "налаштування" параметричних показників до потреб специфічних застосувань" [Evert : 90].

Таблиця 2

Показники асоціації MI та MI^3 для моделей синтаксичних фразеологізмів (СФ) в українській мові

№ з/п	Модель СФ	Абсолютна частота вживання незмінного компонента СФ в УНЛК	Показник асоціації MI	Показник асоціації MI^3
1	*Ати-бати, йшли* N_1 Cop_f /Inf/*в* N_4	0	–	–
2	*Де (вже) там* N_1 Cop_f/Inf/Adj/Adv	*де там* 552	12,7973	31,0156
		де вже там 59	28,398	40,1634
3	*Ну і/й* N_1 Cop_f	*ну і 823*	13,6588	33,0302
		ну й 897	13,8	33,4086
4	*Оце так* N_1 Cop_f	400	13,2814	30,5704
5	*Теж мені* N_1 Cop_f	201	12,3535	27,6568
6	*Чим не* N_1 Cop_f	478	12,5621	30,3661
7	*Що (ж це) за* N_1 Cop_f	*що за* 2646	14,3143	37,0558
		що це за 738	27,7322	43,4498
		що ж це за 206	41,2658	56,6392
8	*Яке (вже/ж) там* N_1 Cop_f /Inf/Adv	*яке там 218*	11,6741	27,2113
		яке вже там 17	23,4714	31,6472
		яке ж там 7	22,088	27,7033

Закінчення табл. 2

№ з/п	Модель СФ	Абсолютна частота вживання незмінного компонента СФ в УНЛК	Показник асоціації MI	Показник асоціації MI^3
9	*Який там* N_1 Cop_f	*який там 358*	12,1239	31,0525
		яка там 397	12,2791	29,5468
		які там 474	12,5498	30,3289
10	N_1 (*він*) *і в Африці* N_1	*і в Африці* 30	26,3059	36,1202
		він і в Африці 5	39,1767	43,8209
		вона і в Африці 5	39,1316	43,7754
		воно і в Африці 0	–	–
		вони і в Африці 3	38,3166	41,4867

З наведених у таблиці 2 даних видно, що відмінність коефіцієнта MI^3 для фразеологізованих речень із дво- і тричленним компонентом не є такою значною, як у випадку з показником MI. Для синтаксичних фразеологізмів із двочленним незмінним компонентом коефіцієнт MI^3 перебуває в межах від 27,2113 (*яке там*) до 37,0558 (*що за*), а з тричленним – від 27,7033 (*яке ж там*) до 40,1643 (*де вже там*). MI^3 для чотирикомпонентних моделей закономірно є вищим – від 41,4867 (*вони і в Африці*) до 56,6392 (*що ж це за*). При цьому важливою видається встановлена закономірність: показники MI^3 виявилися близькими для базової моделі та для її варіантів навіть за умови більшої кількості словоформ у складі незмінного компонента моделі.

Оскільки фразеологізовані речення є одиницями з особливим статусом, зумовленим їхньою подвійною природою як синтаксичних та фразеологічних одиниць водночас, цілком умотивованим вважаємо за необхідне зіставити отримані результати обчислення MI з відповідними показниками для традиційних (лексичних)

фразеологізмів та традиційних (нефразеологізованих) речень. До аналізу залучено по 10 одиниць обох типів (дво- і трикомпонентних). Результати наводимо в таблицях 3 і 4 відповідно.

Таблиця 3

Показники асоціації MI та MI^3 для лексичних фразеологізмів (ЛФ) в українській мові

№ з/п	ЛФ	Абсолютна частота вживання ЛФ в УНЛК	Абсолютна частота вживання в УНЛК словоформ-компоненті в ЛФ	Показник асоціації MI	Показник асоціації MI^3
1	*Блудний син*	61	*блудний 108 син 2002*	15,633	27,4960
2	*Горшки побили*	7	*горшки 200 побили 528*	13,543	19,1588
3	*Дала гарбуза*	3	*дала 1769 гарбуза 194*	10,6202	13,7926
4	*Наріжний камінь*	101	*наріжний 143 камінь 1367*	16,507	29,8242
5	*Пекти раків*	14	*пекти 321 раків 175*	15,455	23,0701
6	*Горобцям дулі давати*	3	*горобцям 31 дулі 138 давати 1584*	33,742	36,9122
7	*За царя Гороха*	21	*за 4830 царя 1040 Гороха 47*	31,4272	40,2123
8	*Кишки грають марш*	1	*кишки 338 грають 946 марш 569*	27,4059	27,4059
9	*Скакати в гречку*	6	*скакати 161 в 4864 гречку 290*	29,671	34,8411
10	*Ускочити в халепу*	9	*ускочити 65 в 4864 халепу 316*	31,441	37,7864

Таблиця 4

Показники асоціації MI і MI^3 для нефразеологізованих речень в українській мові

№ з/п	Речення	Абсолютна частота вживання речення в УНЛК	Абсолютна частота вживання в УНЛК словоформ-компоненті в речення	Показник асоціації MI	Показник асоціації MI^3
1	*Я співаю*	80	*я 3481 співаю 229*	14,142	26,7874
2	*Дівчина гарна*	21	*дівчина 1468 гарна 1189*	11,081	19,8671
3	*Соловей тьохкає*	3	*соловей 316 тьохкає 38*	15,458	18,6285
4	*Серце тьохкає*	8	*серце 2182 тьохкає 38*	14,085	20,0853
5	*Мені добре*	195	*мені 2581 добре 3129*	12,088	27,3026
6	*Прийшла весна*	75	*прийшла 1729 весна 1189*	12,682	25,1409
7	*Я його кохаю*	15	*я 3481 його 4702 кохаю 344*	26,366	34,1804
8	*Ти моя любов*	5	*ти 3165 моя 2287 любов 1879*	23,506	28,1521
9	*Я їду додому*	12	*я 3481 їду 630 додому 2086*	26,343	34,1106
10	*Був сонячний день*	6	*був 3900 сонячний 624 день 2960*	24,688	29,8584

З таблиць 1 і 3 видно, що частота вживання фразеологізованих речень у корпусі є суттєво вищою, ніж частота лексичних фразеологізмів, велика частина яких має частоту до 10. Зафіксовані високі показники MI для всіх лексичних фразеологізмів (MI >> 3) засвідчують невипадковість зв'язку словоформ у їхньому складі, високий ступінь їхнього злиття. Діапазон коефіцієнта MI є незначно більшим, ніж цей показник у фразеологізованих реченнях: для двокомпонентних лексичних фразеологізмів він перебуває в межах від 10,6202 (*дала гарбуза*) до 16,507 (*наріжний камінь*), для трикомпонентних – від 27,4059 (*кишки грають марш*) до 33,742 (*горобцям дулі давати*), при цьому для трикомпонентних моделей показники є значно вищими, що підтверджує більшу стійкість поєднання складників лексичних фразеологізмів, ніж синтаксичних.

Загалом можна зробити висновок, що коефіцієнт MI^3 зменшує показник спільної інформації для лексичних фразеологізмів як переважно низькочастотних мовних одиниць. Кількісні дані, отримані для стійких одиниць, що мають частоту 1, демонструють тотожність показників MI і MI^3, тобто застосування критерію MI^3 до таких одиниць не має смислу.

Для всіх проаналізованих нефразеологізованих речень зафіксовано високий коефіцієнт MI (MI >> 3). Проте аналіз отриманих даних виявив несподіваний для нас результат. Показники MI для двокомпонентних нефразеологізованих речень, лексичних фразеологізмів і синтаксичних фразеологізмів перебувають майже в одному діапазоні. Статистично вірогідні відмінності зафіксовано в показниках MI для трикомпонентних конструкцій: найнижчі показники мають нефразеологізовані речення, за ними йдуть синтаксичні фразеологізми, найвищі показники зафіксовано для лексичних фразеологізмів. Якщо у випадку з багатокомпонентними конструкціями таке розташування є цілком прогнозованим, то одержані дані для двокомпонентних одиниць змушують замислитись над причинами близькості показників спільної інформації для одиниць, що мають різний статус та різні механізми утворення. На нашу думку, близькі значення коефіцієнтів не можуть інтерпретуватися без

урахування різних параметрів, "вихідних даних", властивих проаналізованим одиницям. Так, для лексичних фразеологізмів визначальною особливістю є цілісність значення конструкції, нефразеологізовані речення мають граматичні передумови невипадковості поєднання словоформ (наприклад, узгодження в роді, числі, відмінку іменних частин мови), а синтаксичні фразеологізми, не маючи таких виражених передумов злиття компонентів, демонструють близькі показники спільної інформації.

Коефіцієнт MI^3 для двокомпонентних одиниць перебуває в межах: для синтаксичних фразеологізмів – від 27,6568 (*теж мені*) до 37,0558 (*що за*); для лексичних фразеологізмів – від 13,7926 (*дала гарбуза*) до 29,8242 (*наріжний камінь*); для нефразеологізованих речень – від 18,6285 (*Соловей тьохкає*) до 27,3026 (*Мені добре*). Для трикомпонентних конструкцій – відповідно від 27,7033 (*яке ж там*) до 43,4498 (*що це за*), від 27,4059 (*кишки грають марш*) до 40,2123 (*за царя Гороха*), від 28,1521 (*Ти моя любов*) до 34,1804 (*Я його кохаю*). Отже, найвищими виявилися значення коефіцієнта MI^3 для синтаксичних фразеологізмів.

Отже, показник асоціації MI видається придатним для визначення коректності виділення фразеологізованої моделі речення та вірогідності встановлення стійкості поєднання двох або більше словоформ у межах незмінного компонента моделі речення. Здійснений статистичний аналіз дав змогу підтвердити правильність висунутої гіпотези про наявність високого ступеня (>>3) невипадковості поєднання словоформ у межах незмінного компонента всіх обстежених моделей фразеологізованих речень.

Друга гіпотеза підтвердилася частково: відмінності в показниках невипадковості появи словоформ виявлено в різних групах мовних одиниць – лексичних фразеологізмів, нефразеологізованих речень і фразеологізованих речень – тільки для трикомпонентних конструкцій.

З-поміж чинників, які впливають на коректність здійснених підрахунків, виділяємо такі: неможливість у межах УНЛК обрахувати окремо речення з різним частиномовним наповненням

змінного компонента (наприклад, *Де там* N_1Cop_f, *Де там* Inf, *Де там* Adj, *Де там* Adv) та потребу залучення експерта для розмежування омонімів (наприклад, *там* з локативним та із заперечним значенням; *горох* як загальна і *Горох* як власна назва).

На значення показника MI впливає кілька чинників: частота конструкції, частота словоформ, кількість компонентів конструкції, обсяг корпусу та тип мовної одиниці (у нашому випадку – фразеологізоване речення / лексичний фразеологізм / нефразеологізоване речення). Модифікована формула MI^3 збільшує вагу частоти вживання конструкції в корпусі, унаслідок чого найвищі результати отримують саме фразеологізовані речення.

Перспективи подальших досліджень бачимо в обчисленні інших показників асоціації для моделей фразеологізованих речень української мови та в зіставленні отриманих результатів із коефіцієнтом MI.

ЛІТЕРАТУРА

1. *Балобанова Л. А*. Семантико-прагматический потенциал синтаксических фразеологизмов и их лексикографическое представление в словаре учебного типа : автореф. дисс. на соискание учёной степени канд. пед. наук / Л. А. Балобанова / Московский гос. ун-т имени М. В. Ломоносова. – М., 2004. – 28 с.

2. *Величко А. В*. Синтаксическая фразеология для русских и иностранцев : Учебное пособие / А. В. Величко. – М. : Изд-во МГУ, 1996. – 96 с.

3. *Всеволодова М. В., Лим Су Ён*. Принципы лингвистического описания синтаксических фразеологизмов: На материале синтаксических фразеологизмов со значением оценки / М. В. Всеволодова, Ён Лим Су. – М. : МАКС Пресс, 2002. – 164 с.

4. *Залесская В. В*. Программа выявления в тексте двучленных статистически значимых осмысленных коллокаций (на материале русского языка) / В. В. Залесская // Технологии информационного общества в науке, образовании и культуре : сборник научных статей. Труды XVII Всероссийской объединенной конференции “Интернет и современное общество” (IMS-2014), Санкт-Петербург, 19–20 ноября 2014 г. – СПб : Университет ИТМО, 2014. – С. 283–289.

5. *Личук М. І.* Ступені фразеологізації речень : автореф. дис. на здобуття наук. ступеня канд. філол. наук / М. І. Личук / НАН України; Інститут української мови. – К., 2001. – 16 с.

6. *Личук М. І., Шинкарук В. Д.* Ступені фразеологізації речень / М. І. Личук, В. Д. Шинкарук. – Чернівці : Рута, 2001. – 136 с.

7. Русская грамматика: В 2-х т. – Т. 2. Синтаксис / Под ред. Н. Ю. Шведовой. – М. : Наука, 1980. – 709 с.

8. *Ситар Г. В*. Статус синтаксичних фразеологізмів у системі фразеологічних одиниць / Г. В. Ситар // Вісник Донецького національного університету. Серія Б. Гуманітарні науки. – Донецьк : ДонНУ, 2011. – № 2. – С. 66–74.

9. *Ситар Ганна*. Конструкційна граматика як теоретичне підґрунтя дослідження фразеологізованих речень / Г. Ситар // Типологія та функції мовних одиниць : наук. журн. на пошану член-кореспондента НАН України І.Р. Вихованця / [редкол. : Н.М. Костусяк (гол. ред.) та ін.]. – Луцьк : Східноєвропейський нац. ун-т ім. Лесі Українки, 2015. – № 2 (4). – С. 192–205.

10. *Хохлова М. В*. Исследование лексико-синтаксической сочетаемости в русском языке с помощью статистических методов (на базе корпусов текстов) : автореф. дисс. на соискание ученой степени канд. филол. наук / М. В. Хохлова. – Санкт-Петербург, 2010. – 26 с.

11. *Шмелёв Д. Н*. Синтаксическая членимость высказывания в современном русском языке / Д. Н. Шмелёв. – М. : URSS, 2006. – 148 с.

12. *Ягунова Е. В., Пивоварова Л. М*. От коллокаций к конструкциям / Е. В. Ягунова, Л. М. Пивоварова // ACTA LINGUISTICA PETROPOLITANA. Труды Института лингвистических исследований РАН. – Т. X. – Ч. 2. Русский язык: грамматика конструкций и лексико-семантические подходы / Ред. тома С. С. Сай, М. А. Овсянникова, С. А. Оскольская. – СПб. : Наука, 2014. – С. 568–617.

13. *Church K., Hanks P*. Word association norms, mutual information, and lexicography / K. Church, P. Hanks // Computational Linguistics. – № 16(1). – 1990. – Pp. 22–29.

14. *Daille B*. Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques. Thèse de Doctorat en Informatique Fondamentale / B. Daille. – Université Paris 7, 1994. – 228 p.

15. *Evert S*. The Statistics of Word Cooccurrences: Word Pairs and Collocations / S. Evert : PhD dissertation, IMS, University of Stuttgart, 2004 (Published in 2005). – 353 P. – Free PDF available from http://purl.org/stefan.evert/PUB/Evert2004phd.pdf.

16. *Fano Robert M*. Transmission of Information: A Statistical Theory of Communications / Robert M. Fano // The Technology Press, M.I.T., and John Wiley & Sons, Inc. – New York, 1961. – 389 p.

17. *Fillmore Charles J.* The Mechanisms of "Construction Grammar" / Charles J. Fillmore // Proceedings of the Fourteenth Annual Meeting of the Berkeley Linguistics Society. – 1988. – P. 35–55.

18. *Fillmore C. J., Kay P., O'Connor M. C.* Regularity and Idiomaticity in Grammatical Constructions: the Case *of let alone* / C. J. Fillmore, P. Kay, M. C. O'Connor // Language. – 1988. – 64(3). – P. 501–538.

19. *Goldberg A. E*. Constructions: A Construction Grammar Approach to Argument Structure. 1 edition / A. E. Goldberg. – University Of Chicago Press, March 15, 1995. – 271 p.

20. *Goldberg Adele E*. Constructions : a New Theoretical Approach to Language / Adele E. Goldberg // Trends in Cognitive Sciences. – 2003. – Vol. 7 – No. 5 May. – P. 219–224.

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

22. *Seretan V*. Syntax-Based Collocation Extraction / V. Seretan // Text Speech and Language Technology. Series Editors Nancy Ide, Jean Véronis. – Volume 44. – Dordrecht – Heidelberg – London – New York : Springer, 2011. – 222 p.

23. *Stubbs M*. Collocations and semantic profiles: On the cause of the trouble with quantitative studies / M. Stubbs // Functions of Language. – 1995. – 2, 1. – P. 23–55.

24. *Stubbs M*. Words and Phrases: Corpus Studies of Lexical Semantics / M. Stubbs. – Blackwell, Oxford, 2001. – 288 p.

REFERENCES

1. Balobanova L.A. (2004) Semantiko-pragmaticheskiy potentsial sintaksicheskikh frazeologizmov i ikh leksikograficheskoe predstavlenie v slovare uchebnogo tipa [*Semantic and Pragmatic Potential of Syntactic Idioms and Their Lexicographic Representation in the Dictionary of Educational Type*] (PhD Thesis), Moscow: Moscow State University named after M.V. Lomonosov, 28 p. (in Russian).

2. Velichko A.V. (1996) Sintaksicheskaya frazeologiya dlya russkikh i inostrantsev [*Syntactic Phraseology for Foreigners*]. Moscow: Moscow State University named after M.V. Lomonosov, 96 p. (in Russian).

3. Vsevolodova M.V., Lim Su Yen. (2002) Printsipy lingvisticheskogo opisaniya sintaksicheskikh frazeologizmov: Na materiale sintaksicheskikh frazeologizmov so znacheniem otsenki [*Principles of Linguistic Description of Syntactic Idioms: On the Basis of Syntactic Idioms with Evaluative Meaning*]. Moscow : MAKS Press, 164 p. (in Russian).

4. Zalesskaya V.V. (2014) Programma vyyavleniya v tekste dvuchlennykh statisticheski znachimykh osmyslennykh kollokatsiy (na materiale russkogo yazyka) [*Programme of Identification of the Binominal Statistically Significant Meaningful Collocations in the Text (Based on the Russian Language*)]. Proceedings of the Internet i sovremennoe obshchestvo (Russia, Saint Petersburg, November 19-20, 2014), Saint Petersburg: ITMO University, pp. 283-289 (in Russian).

5. Lychuk M.I. (2001) Stupeni frazeologhizaciji rechenj [*Levels of Sentences Phraseologization*] (PhD Thesis), Kyiv: NAS of Ukraine; Institute of the Ukrainian Language, 16 p. (in Ukrainian).

6. Lychuk M.I., Shynkaruk V.D. (2001) Stupeni frazeologhizaciji rechenj [*Levels of Sentences Phraseologization*]. Chernivci: Ruta, 136 p. (in Ukrainian).

7. Shvedova N.Yu. (ed.) (1980) Russkaya grammatika: v 2-kh t [*Russian Grammar: In 2 Vol.*]. Moscow: Nauka, vol. 2, 709 p. (in Russian).

8. Sytar H.V. (2011) Status syntaksychnykh frazeologhizmiv u systemi frazeologhichnykh odynycj [*Status of Syntactic Idioms in the System of Phraseological Units]*. *Bulletin of Donetsk National University. Part B. Humanities*, no. 2, pp. 66-74. (in Ukrainian).

9. Sytar Hanna (2015) Konstrukcijna ghramatyka jak teoretychne pidgruntja doslidzhennja frazeologhizovanykh rechenj [*Construction Grammar as a Theoretical Background of Syntactic Idioms Studying*]. *Typology and Functions of Language Units : Scientific Journal to Respect of Corresponding Member of NAS of Ukraine I. R. Vykhovanets*, no. 2 (4), pp. 192-205 (in Ukrainian).

10. Khokhlova M.V. (2010) Issledovanie leksiko-sintaksicheskoy sochetaemosti v russkom yazyke s pomoshchyu statisticheskikh metodov (na baze korpusov tekstov) [*The Study of Lexical and Syntactic Compatibility in the Russian Language with the help of Statistical Methods (Based on the Corpus of Texts)*] (PhD Thesis), Saint Petersburg: Saint Petersburg State University, 26 p. (in Russian).

11. Shmelev D.N. (2006) Sintaksicheskaya chlenimost vyskazyvaniya v sovremennom russkom yazyke [*Syntactic Divisibility of Statement in the Modern Russian Language]*. Moscow: URSS, 148 p. (in Russian).

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

13. Church K., Hanks P. (1990) Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, no. 16(1), pp. 22-29.

14. Daille B. (1994) *Approche Mixte Pour L'extraction Automatique de Terminologie: Statistiques Lexicales et Filtres Linguistiques* (PhD Thesis), Paris : Université Paris, 228 p.

15. Evert S. (2004) (Published in 2005) *The Statistics of Word Cooccurrences: Word Pairs and Collocations* (PhD Thesis) (electronic source), Stuttgart: University of Stuttgart, 353 p. Available at: http://purl.org/stefan.evert/PUB/Evert2004phd.pdf.

16. Fano Robert M. (1961) *Transmission of Information*: *A Statistical Theory of Communications*. New York : The Technology Press, M.I.T., and John Wiley & Sons, Inc, 389 p.

17. Fillmore Charles J. (1988) *The Mechanisms of "Construction Grammar"*. Proceedings of the *Fourteenth Annual Meeting of the Berkeley Linguistics Society*, pp. 35-55.

18. Fillmore C.J., Kay P., O'Connor M.C. (1988) *Regularity and Idiomaticity in Grammatical Constructions: the Case of let alone. Language*, no. 64(3), pp. 501-538.

19. Goldberg A.E. (1995) *Constructions: A Construction Grammar Approach to Argument Structure*. University of Chicago, 271 p.

20. Goldberg Adele E. (2003) Constructions: a New Theoretical Approach to Language. *Trends in Cognitive Sciences*, vol. 7, no. 5, pp. 219-224.

21. Petrovic S., Snajder J., Basic B.D., Kolar M. (2006) Comparison of collocation extraction for document indexing. *Journal of Computing and information technology*, vol. 14 (4), pp. 321-327.

22. Seretan V. (2011) Syntax-Based Collocation Extraction. *Text Speech and Language Technology* (eds. Nancy Ide, Jean Véronis), vol. 44. Dordrecht - Heidelberg - London - New York : Springer, 222 p.

23. Stubbs M. (1995) Collocations and semantic profiles: On the cause of the trouble with quantitative studies. *Functions of Language*, vol. 2, 1, pp. 23-55.

24. Stubbs M. (2001) *Words and Phrases: Corpus Studies of Lexical Semantics*. Blackwell, Oxford, 288 p.

Стаття надійшла до редколегії 06.09.15

Анна Ситарь, канд. филол. наук, доц., докторант
Донецкий национальный университет, Винница

Статистический анализ фразеологизированных предложений: мера ассоциации *mutual information*

Статья посвящена статистическому анализу фразеологизированных предложений украинского языка. Обоснована целесообразность применения статистического критерия mutual information для установления коэффициента неслучайности определенной последовательности слов в тексте.

Приведены результаты вычисления mutual information для моделей фразеологизированных предложений по данным Украинского национального лингвистического корпуса. Доказано, что все проанализированные модели предложений обладают высокой степенью неслучайности компонентов, которые входят в состав неизменяемой части предложения. Для исследуемых единиц предложено вычисление модифицированного показателя MI – MI^3.

Полученные данные сопоставлены с соответствующими показателями MI и MI^3 для лексических фразеологизмов и нефразеологизированных предложений. Выделены факторы, влияющие на корректность проведенных расчетов.

Ключевые слова: *конструкция, конструкционная грамматика, корпус текстов, синтаксический фразеологизм, статистический анализ, мера ассоциации mutual information, мера ассоциации MI^3, украинский язык, фразеологизированное предложение.*

Hanna Sytar, PhD of Philology, Associate Professor, Doctoral Candidate
Donetsk National University, Vinnytsya

Statistical analysis of sentences with phraseological structures: association measure of *mutual information*

The article is devoted to the statistical analysis of the sentences with phraseological structures of the Ukrainian language. Expediency of application of statisti-

cal criterion of mutual information was substantiated for determining the nonrandom measure of a certain sequence of words in the text.

There were provided the results of mutual information computation for the models of sentences with phraseological structures according to the data of Ukrainian National Linguistic Corpus. There was proved that all the analyzed sentences models have a high degree of components non-randomness that make up invariable part of the sentence. There were suggested the computation of the modified measure of MI – MI^3 for the researched units.

Obtained data was compared with the corresponding measures of MI and MI^3 for lexical phraseologisms and non-idiomatic sentences. There were singled out the factors that affect the correctness of calculations performed.

Keywords: *construction, construction grammar, text corpus, syntactic idiom, statistical analysis, association measure of mutual information, association measure of MI^3, sentence with phraseological structure.*