past, their stylistic markers and correlation with linguistic rules; define the role of color in the creation of vestizmov range of generalized symptoms of the individual, which also serves as a kind of indicator of the national and cultural identity of the ethnic group.

Keywords: mental-cognitive level, perception, cognitive structure, stereotype, stereotype touch component field «name clothing».

УДК 81'367:81'373.7=161.2

Г. В. Ситар

СТАТИСТИЧНІ КРИТЕРІЇ АНАЛІЗУ СИНТАКСИЧНИХ ФРАЗЕОЛОГІЗМІВ

Реферат. *Статтю присвячено статистичним критеріям аналізу синтаксичних фразеологізмів на матеріалі української мови. Синтаксичні фразеологізми розглянуто з позицій конструкційної граматики та інтерпретовано як один із типів некомпозиційних мовних знаків – конструкцій.*

Проаналізовано основні показники асоціації: MI, t-score, log-likelihood, Dice, gmean. Обрано показник асоціації MI як статистичний критерій, що дає змогу визначити коефіцієнт невипадковості поєднання двох і більше слів у тексті, враховує частоту конструкції, частоту її компонентів, розмір корпусу та має формулу в узагальненому вигляді для конструкцій з будь-якою кількістю компонентів.

Подано результати здійсненого статистичного аналізу моделей синтаксичних фразеологізмів української мови за даними Українського національного лінгвістичного корпусу. З'ясовано, що всі обстежені за показником асоціації MI моделі синтаксичних фразеологізмів мають високий (MI >>3) ступінь невипадковості поєднання компонентів, що входять до складу незмінної частини речення, тобто характеризуються статистично доведеною зв'язаністю.

Ключові слова: *конструкція, конструкційна граматика, корпус текстів, синтаксичний фразеологізм, статистичний аналіз, показник асоціації, фразеологізоване речення.*

Синтаксичні фразеологізми, або фразеологізовані речення, – специфічний тип речення, у якому фіксовано розташовані постійний (незмінний) і змінний компоненти пов'язані ідіоматично, граматичні зв'язки і прямі лексичні значення слів послаблені або втрачені на сучасному етапі розвитку мови. Синтаксичні фразеологізми є дієвим засобом вираження ставлення мовця до висловлюваного, властивим для розмовного мовлення, художніх і публіцистичних текстів [Балобанова 2004; Величко 1996; Всеволодова, Лим Су 2002; Русская грамматика 1980; Шмелёв 2006 та ін.] (докладно про ознаки та статус синтаксичних фразеологізмів у системі мовних одиниць див. у праці [Ситар 2011]).

Спираючись на положення конструкційної граматики (Construction Grammar) як напряму сучасних граматичних досліджень, започаткованого Чарльзом Філлмором (Charles J. Fillmore) та розвиненого Полом Кеєм (Paul Kay), Мері Кетрін О'Коннор (Mary Catherine O'Connor), Адель Голдберг (Adele E. Goldberg), Вільямом Крофтом (William Croft), Мір'ям Фрайд (Mirjam Fried) та іншими дослідниками, синтаксичний фразеологізм вважаємо одним із типів конструкції – мовного знаку, певний аспект плану вираження або плану змісту якого не можна пояснити через поєднання форми або змісту його компонентів [Fillmore 1988; Fillmore, Kay, O'Connor 1988; Goldberg 1995; Goldberg 2003; Fried 2010 та ін.]. Відповідно синтаксичний фразеологізм інтерпретуємо як некомпозиційну синтаксичну одиницю з виразним прагматичним спрямуванням (докладніше див. [Ситар 2015]).

Конструкційна граматика, як і багато інших напрямів сучасної лінгвістики, визначає необхідність статистичного етапу дослідження, що передбачає залучення статистичних методів для перевірки правильності висунутих гіпотез, отримання кількісних даних, що підтверджують або спростовують певні теоретичні положення (див., зокрема, праці представників колоструктційного аналізу (Collostructional Analysis) Анатоля Стефановича (Anatol Stefanowitsch), Стефана Ґріса (Stefan Th. Gries) та ін. [Gries, Stefanowitsch 2004; Stefanowitsch, Gries 2003; Stefanowitsch, Gries 2005]). При цьому єдиним надійним матеріалом для

статистичного аналізу будь-яких мовних або мовленнєвих одиниць вважають корпус текстів; іншими словами, саме корпусна зорієнтованість статистичного дослідження забезпечує вірогідність одержаних кількісних даних та переконливість зроблених висновків.

Корпус текстів (або повнотекстова база даних) – це «упорядкована сукупність текстів у цілісному вигляді» [Карпіловська 2006: 94], головною ознакою якої є наявність розмічення (індексування) – спеціально створених позначок для відображення лінгвістичної інформації (морфологічної, синтаксичної, семантичної, акцентної, метатекстової). Крім цього, на сьогодні в корпусах текстів передбачають можливість автоматичного здійснення статистичних підрахунків за визначеними формулами (див., напр., проект Sketch Engine [https://www.sketchengine.co.uk]; зазначимо також, що українські корпуси текстів подібних параметрів поки що не мають). Щодо ролі корпусу в мовознавчих дослідженнях і статистичних можливостей корпусів текстів Марія Хохлова зазначає, що «в сучасній лінгвістиці незамінним інструментом і водночас матеріалом для лінгвістичного дослідження і розв'язання прикладних завдань стали корпуси текстів. […] Статистичний апарат, застосовуваний у корпусах текстів, дає змогу користувачам ранжувати результати пошуку за різними параметрами та задавати порогові значення, що спричиняє отримання найбільш значущої інформації» [Хохлова 2010: 3].

Виконання теоретичної частини нашого дослідження дало змогу сформулювати таку робочу гіпотезу: синтаксичні фразеологізми (або фразеологізовані моделі речень), як і будь-які інші стійкі одиниці, мають високий ступінь невипадковості поєднання компонентів, що входять до складу незмінної частини речення. Спробуємо довести це твердження за допомогою статистичних методів.

В арсеналі сучасної статистики існує низка статистичних критеріїв (коефіцієнтів), об'єднаних терміном «показники асоціації» (англ. association measures, measures of association). Згідно з Кембріджським словником статистики Брайана Еверітта (Brian S. Everitt), «Показники асоціації – числові індекси, що обчислюють силу статистичної залежності двох або більше кваліативних змінних» [Everitt 2002 : 241].

За Штефаном Евертом (Stefan Evert), **«Показник асоціації** [шрифтові виділення Штефана Еверта – *Г.С.*] – це формула, що обчислює **величину асоціації** виходячи з інформації про частоту у факторній таблиці парного типу. Ця величина розглядається як індикатор того, наскільки сильною є асоціація між компонентами пари, з поправкою на випадкові ефекти […]. Я використовую конвенцію, що високі показники асоціації засвідчують сильну асоціацію» [Evert 2004: 75].

Показники асоціації призначені передусім для автоматичного виділення колокацій / конструкцій в тексті (корпусі текстів) на підставі встановлення випадковості / невипадковості певної послідовності слів у тексті (корпусі текстів). Саме з такою метою їх застосовують на матеріалі англійської, німецької, останнім часом російської мов [Залесская 2014; Хохлова 2008; Ягунова, Пивоварова 2014; Church, Hanks 1990; Evert 2004; Seretan 2011; Stubbs 1995 та ін.].

Термін «конструкція» вживаємо в розумінні, запропонованому представниками конструкційної граматики (див. вище). Щодо колокацій зазначимо, що в зарубіжному та вітчизняному мовознавстві сформувалось кілька підходів до кваліфікації колокації, умовно їх можна назвати структурним, семантичним, семантико-синтаксичним та статистичним (пор. Тетяна Бобкова виділяє чотири основні підходи до аналізу колокацій: когнітивний, семантико-синтаксичний (з подальшою диференціацією на лексико-семантичний і лексико-функціональний), контекстно-орієнтований (об'єднує лексико-граматичний, синтаксичний і корпусний) та психолінгвістичний [Бобкова 2014: 16 і далі]).

У межах лінгвістичної статистики (за Тетяною Бобковою, корпусного підходу) терміном «колокація» позначають «статистично стійкі поєднання» слів [Хохлова 2010: 4], тобто невипадковість їхнього поєднання в тексті підтверджена здійсненими обчисленнями спеціально обраного відповідно до мети і предмета дослідження показника асоціації (на сьогодні переважно кількох показників асоціації) [Залесская 2014; Хохлова 2008; Хохлова 2010; Ягунова, Пивоварова 2014; Evert 2004; Kilgarriff, Tugwell 2002; Seretan 2011 та ін.]. Із цього приводу Тетяна Бобкова зауважує, що «за корпусним підходом колокація розуміється як послідовність слів, що зустрічаються разом частіше, ніж можна було б очікувати випадково» [Бобкова 2014: 19]. Введення ймовірнісно-статистичних методів виявлення колокацій є

«розвитком ідей британського контекстуалізму» [Хохлова 2010: 10], найвідомішим представником якого є Джон Руперт Фьорз (John Rupert Firth).

На можливості застосування показників асоціації з різною метою наголошує Штефан Еверт: «Величини, обчислені показником асоціації, можуть бути інтерпретовані в різний спосіб: (i) Вони можуть бути використані прямо для оцінки величини асоціації між компонентами парного типу. (ii) Вони можуть бути використані для одержання ранжування парних типів у наборі даних. У цьому випадку абсолютна величина значень є нерелевантною. (iii) Вони також можуть бути використані для оцінки парних типів із визначеним першим або другим компонентом» [Evert 2004: 75].

На нашу думку, обчислення показників асоціації можна застосовувати у вивченні синтаксичних фразеологізмів – для кількісного підтвердження правомірності кваліфікації певної моделі речення як ідіоматичної (фразеологізованої, зв'язаної) з опертям на визначення коефіцієнта / коефіцієнтів, що відбиває / відбивають ступінь випадковості / невипадковості (залежності / незалежності, злитості / незлитості, зв'язаності / незв'язаності) певної послідовності слів у тексті. Іншими словами, обчислення показників асоціації дає змогу підтвердити або спростувати так звану «статистичну зв'язаність» конструкції. У цьому плані спираємось на диференціацію зв'язаності, запропоновану Тетяною Бобковою, яка розмежовує семантичну, формальну і статистичну зв'язаність колокації (у термінології дослідниці – «зв'язність») [Бобкова 2014: 16].

Предметом нашого статистичного аналізу стали синтаксичні фразеологізми, незмінний компонент яких складається з поєднання кількох лексем – службових і повнозначних слів, яким властиве семантичне спустошення або семантичний зсув (*що за, от тобі/вам і/й, не до, теж мені, яке там* і под.: ***Що за** тон! **От вам і** відповідь! **Теж мені** друг! **Яке там** гарна! **Не до** розмов тепер мені*). Відповідно застосування показників асоціації доцільне саме для дво- і більше компонентних незмінних комплексів, проте поза межами розгляду залишаються група синтаксичних фразеологізмів, побудованих шляхом поєднання повторюваних повнозначних і службового компонентів типу *як, так, і, а, не* і под. (*Хлопець **як** хлопець; Зробив **так** зробив*) або компонента-зв'язки (*Війна **є** війна*) та група фразеологізованих речень, що мають однослівний незмінний компонент (без повторів слів у змінній частині): ***Яка** пісня! **Оце** сказала!* і под.

Мета статті – проаналізувати основні показники асоціації, використовувані для статистичного аналізу конструкцій, та визначити, які з них придатні для синтаксичних фразеологізмів. Поставлена мета передбачає розв'язання таких завдань: 1) розглянути сутність основних показників асоціації; 2) обґрунтувати доцільність застосування відповідного статистичного критерію для аналізу синтаксичних фразеологізмів; 3) здійснити обчислення показників асоціації для низки моделей синтаксичних фразеологізмів за даними Українського національного лінгвістичного корпусу.

На сьогодні в лінгвістичній статистиці застосовують кілька десятків показників асоціації (див. [Evert 2004; Seretan 2011] та ін.). До найбільш часто використовуваних для вивчення колокацій / конструкцій належать такі показники асоціації, як MI, t-score, log-likelihood, Dice, gmean та ін.

Показник асоціації MI (або MI-score) – коефіцієнт, який відбиває невипадковість (залежність) певної послідовності слів у тексті. Поняття MI (англ. mutual information – взаємна, спільна, повна інформація) запропоноване в теорії інформації американським ученим італійського походження Робертом Маріо Фано (Robert Mario Fano) [Fano 1961]. У лінгвістичний обіг формулу (1) для обчислення MI ввели американські дослідники Кеннет Ворд Чарч (Kenneth Ward Church) та Патрік Хенкс (Patrick Hanks) [Church, Hanks 1990 : 23]:

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

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).

Величину імовірності *P* (англ. probability – імовірність) автори обраховують за допомогою визначення частоти *f* (англ. frequency – частота) та здійснення нормалізації, потреба у якій викликана отриманими результатами підрахунків на матеріалі кількох корпусів різного розміру.

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

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

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

Підставивши формулу (2) до формули (1) отримуємо формулу (3) для двокомпонентних конструкцій (біграм):

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

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* в корпусі;
f(y) – абсолютна частота *y* в корпусі;
N – загальна кількість словоформ у корпусі;
log_2 – логарифм числа за основою 2.

Кеннет Ворд Чарч та Патрік Хенкс подають таку інтерпретацію результатів обчислень: невипадковим вважається поєднання лексем, якщо MI > 3 [Church, Hanks 1990 : 24].

Оскільки значна частина синтаксичних фразеологізмів належить до багатокомпонентних одиниць, постає потреба у врахуванні більшої кількості компонентів, ніж 2. У науковій літературі запропоновано формулу для триграм – трикомпонентних поєднань слів. Так, Sasa Petrovic, Jan Snajder, Bojana Dalbelo Basic, Mladen Kolar до формули (1) вводять третій компонент z, унаслідок чого отримуємо формулу (4) [Petrovic, Snajder, Basic, Kolar 2006 : 323]:

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

З формули (4) можна вивести загальну формулу (5), за якою можна обрахувати коефіцієнт MI для конструкцій із будь-якою кількістю компонентів ($i \geq 2$). Формулу (5) подаємо за [Ягунова, Пивоварова 2014 : 586]:

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

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.

Статистичний показник t-score (або t-test, t-критерій, T-value) для вивчення біграм запропонували Кеннет Чарч (Kenneth Church), Вілліам Гейл (William Gale), Патрік Хенкс (Patrick Hanks), Доналд Хіндл (Donald Hindle) у праці [Church, Hanks, Hindle, Gale 1991: 6 і далі]. Він є застосуванням відомого у статистиці критерію Стьюдента для дослідження сполучень слів. Загалом критерій Стьюдента використовують на статистичному етапі дослідження в різних галузях знань передусім з метою перевірки рівності середніх значень у двох вибірках. У згаданій статті автори здійснюють аналіз сполучуваності англійській слів на матеріалі корпусу Брауна і корпусу Associated Press, що і мотивує доцільність зіставлення даних, і називають t-test «показником відмінності».

Сутність відмінностей у застосуванні MI та t-score американські дослідники сформулювали так: «Показник спільної інформації є кращим для віднайдення схожості; t-величини є кращими для встановлення відмінностей між близькими синонімами. Ми не намагаємося сказати, що один показник є кращим за інший, обидва є важливими. Іноді ми більше зацікавлені у знаходженні асоціацій, а іноді ми зацікавлені у зосередженні на найтонших відмінностях» [Church, Hanks, Hindle, Gale 1991: 28].

Тому формула (6), наведена у працях Марії Хохлової [Хохлова 2008: 348; Хохлова 2010: 12], та використана у вивченні колокацій російськими вченими [Залесская 2014; Ягунова, Пивоварова 2014 та ін.] як така, що «враховує частоту спільного вживання ключового слова та його колоката і відповідає на питання, наскільки не випадковою є сила асоціації (зв'язаності) між колокатами)» [Хохлова 2008: 347] (пор. аналогічну думку висловлено у праці [Пивоварова, Ягунова 2014: 585]), у такому тлумаченні втрачає основне призначення – зіставлення результатів для двох груп (вибірок, підвибірок) подібних одиниць або тих самих одиниць, але досліджуваних в різний період часу, і застосовується з іншою метою:

$$t - score = \frac{f(n,c) - \frac{f(n) \times f(c)}{N}}{\sqrt{f(n,c)}}, \quad (6)$$

де
n – ключове слово;
c – колокат;
f(n,c) – частота вживання слова n у парі з його колокатом c;
f (n) – частота вживання слова n у корпусі;
f(c) – частота вживання слова c у корпусі,
N – загальна кількість словоформ у корпусі.

Застосування показника t-score для нашого дослідження вважаємо недоцільним з урахуванням таких чинників:

1. Відповідно до специфіки об'єкта дослідження (синтаксичні фразеологізми української мови) та мети статистичного етапу дослідження, що полягає у з'ясуванні ступеня зв'язаності лексем у складі незмінного компонента синтаксичних фразеологізмів, зіставлення з іншими вибірками або підвибірками не передбачене.

2. Майкл Стаббс (Michael Stubbs) показав, що обчислення t-критерію дає результат, що приблизно дорівнює кореню частоти обстежуваної складеної одиниці: t-score $\approx \sqrt{f(n,c)}$ [Stubbs 1995]. Проведені нами обчислення конструкцій дають змогу зробити висновок, що це відповідає дійсності з точністю до другого або третього знака після коми, пор. значення кореня частоти конструкції та одержані за формулою (6) значення t-score: модель *Де там* N_1 Cop_f/Inf/Adj/Adv – відповідно 23,4946 і 23,4913, *Ну і* N_1 Cop_f – 28,6879 і 28,6857; *Чим не* N_1 Cop_f – 21,8632 і 21,8595, *Що за* N_1 Cop_f – 51,4392 і 51,4367 под. Наведені підрахунки засвідчують, що критерій t-score майже не враховує частоту вживання окремо взятих слів та загальну кількість словоформ у корпусі попри те, що ці компоненти входять до формули.

3. Показник асоціації t-score має обмеження щодо кількості компонентів обстежуваних мовних одиниць, він не призначений для багатокомпонентних конструкцій.

Статистичний критерій log-likelihood (буквально «логарифм імовірності») – це логарифметична функція правдоподібності поєднання кількох явищ. У випадку дослідження поєднань слів формула log-likelihood враховує спостережувані частоти мовної одиниці O_{ij} та її

очікувані частоти E_{ij}, що обчислюються за таблицею спряженості, та випадковість / невипадковість лівобічного і правобічного контексту для слова [Evert 2004: 83]:

$$log - likelihood = 2\sum_{ij} O_{ij}\, log \frac{O_{ij}}{E_{ij}} \quad (7)$$

Тед Даннінг (Ted E. Dunning) запропонував застосовувати log-likelihood як показник асоціації та ввів модифіковану формулу для її обчислення [Dunning 1993: 67].

Для нашого дослідження застосування коефіцієнта log-likelihood є не виправданим, оскільки переважна більшість синтаксичних фразеологізмів не має лівого контексту через те, що вживається на початку речення (досить часто також на початку тексту як заголовок). Якщо лівий контекст є, то мова йде про попереднє речення, але сусіднє речення, як і знаки пунктуації, не є об'єктом нашого дослідження.

Для виконання статистичних досліджень використовують також коефіцієнт Дайса (Dice, $k_{Dice,}$ індекс Дайса). Вживання в науковій літературі нетотожних назв цього показника – коефіцієнт Дайса, коефіцієнт Сьоренсена, коефіцієнт Дайса-Сьоренсена, коефіцієнт Дайса-Брея та ін. – зумовлене тим, що незалежно один від одного різні вчені з різних країн у різних галузях знань ввели індекс збігу (coincidence index) для визначення подібності двох явищ. У лінгвістиці під коефіцієнтом Дайса розуміють показник, який визначається як частота конструкції, поділена на середнє арифметичне частот її компонентів. Коефіцієнт Дайса розрахований на обстеження двох явищ, відповідно для двох компонентів формула має вигляд (8):

$$Dice(x, y) = \frac{2f(x, y)}{\mathrm{f(x)} + \mathrm{f(y)}} \quad (8)$$

Для завдань нашого дослідження виникла потреба встановити зв'язаність конструкцій з більшою кількістю компонентів. Тому на підставі формули (8) авторка статті вивела формулу для багатокомпонентних мовних одиниць (9):

$$Dice_n(x_1, x_2, \dots, x_n) = \frac{n\, f(x_1, x_2, \dots, x_n)}{\mathrm{f}(x_1) + \mathrm{f}(x_2) + \dots + \mathrm{f}(x_n)}, \quad (9)$$

де $Dice_n$ – модифікація показника Дайса для n-компонентних конструкцій;

n – кількість компонентів у складі конструкції;

x_1 – перша словоформа;

x_2 – друга словоформа;

…

x_n – n-а словоформа;

$f(x_1, x_2, \dots, x_n)$ – абсолютна частота вживання конструкції в корпусі (з урахуванням порядку одиниць усередині конструкції);

$f(x_1)$ – абсолютна частота x_1 в корпусі;

$f(x_2)$ – абсолютна частота x_2 в корпусі;

$f(x_n)$ – абсолютна частота x_n в корпусі.

Показник gmean (англ. geometric mean) означає частоту конструкції, поділену на середнє геометричне частот її компонентів [Evert 2004: 85].

$$gmean(x, y) = \frac{f(x, y)}{\sqrt{f(x) \times f(y)}} \quad (10)$$

Відповідно для багатокомпонентних конструкцій пропонуємо формулу (11):

$$gmean_n(x_1, x_2, \dots, x_n) = \frac{f(x_1, x_2, \dots, x_n)}{\sqrt[n]{f(x_1) \times f(x_2) \times \dots \times f(x_n)}}, \quad (11)$$

де $gmean_n$ – модифікація показника gmean для n-компонентних конструкцій, а решта позначок ідентична використаним у формулі (9).

Статистичний аналіз синтаксичних фразеологізмів ми здійснювали за даними Українського національного лінгвістичного корпусу (далі УНЛК), створеного колективом Українського мовно-інформаційного фонду НАН України та розміщеного за адресою

http://unlc.icybcluster.org.ua/virt_unlc/[1]. Для встановлення абсолютної частоти конструкції та абсолютної частоти окремих словоформ, що входять до складу конструкції, в пошуковій формі корпусу було задано визначений порядок словоформ та передбачено пошук словоформи, а не слова з урахуванням його парадигми. Загальна кількість слововживань на момент здійснення підрахунків становила 180 мільйонів слововживань.

Коефіцієнт MI обраховували з точністю до двох знаків після коми. Отримані результати за показником MI для 10 різнотипних (за частиномовним статусом змінного і незмінного компонентів моделі, кількісним складом незмінного компонента, наявністю варіантів моделі і продуктивністю) фразеологізованих моделей речень наведено в таблиці 1. Через скісну риску подано можливі варіанти в межах однієї моделі, дужками позначено факультативність компонента моделі. Показники асоціації обраховано окремо для кожного варіанта моделі.

Таблиця 1

Показник асоціації MI для моделей
синтаксичних фразеологізмів в українській мові за даними УНЛК

№ з/п	Модель синтаксичного фразеологізму	Абсолютна частота вживання незмінного компонента синтаксичного фразеологізму	Абсолютна частота вживання словоформ, що входять до незмінного компонента синтаксичного фразеологізму	Показник асоціації MI
1	*Де (вже) там* N_1 Cop_f/ Inf/Adj/Adv	*де там 552*	*де 4349* *там 3326*	12,75
		де вже там 59	*вже 3876*	28,4
2	*Ну і/й* N_1 Cop_f	*ну і 823*	*ну 2423* *і 4731*	13,66
		ну й 897	*й 4674*	13,8
3	*Оце так* N_1 Cop_f	*400*	*оце 1548* *так 4676*	13,28
4	*Теж мені* N_1 Cop_f	*201*	*теж 2680* *мені 2582*	12,35
5	*Чим не* N_1 Cop_f	*478*	*чим* 2938 *не* 4844	12,56
6	*Що (ж це) за* N_1 Cop_f	*що за 2646*	*що 4843* *за 4831*	14,31
		що це за 738	*це 4593*	27,73
		що ж це за 206	*ж 4240*	41,27
7	*Яке (вже/ж) там* N_1 Cop_f /Inf/Adv	*яке там 218*	*яке 3613* *там 3326*	11,67
		яке вже там 17	*вже 3946*	23,47
		яке ж там 7	*ж 4240*	22,09
8	*Який там* N_1 Cop_f	*який там 358*	*який 4345* *там 3326*	12,12
		яка там 397	*яка 4326*	12,28
		які там 474	*які 4282*	12,55
9	Cop_f *не до* N_2	*не до 1841*	*не* 4844 *до* 4845	13,79

[1] Дякуємо Директору Українського мовно-інформаційного фонду НАН України академіку НАН України Володимиру Широкову за наданий доступ до корпусу.

10	N_1 (*він*) *і в Африці* N_1 C op_f	*і в Африці 30*	*і* 4731 *в* 4864 *Африці* 510	26,30
		він і в Африці 5	*він 4009*	39,18
		вона і в Африці 5	*вона 4137*	39,13
		воно і в Африці 0	-	-
		вони і в Африці 3	*вони 4366*	38,32

Як видно з таблиці 1, діапазон варіювання показника MI для різних моделей з однаковою кількістю словоформ у складі незмінного компонента є невеликим. Для синтаксичних фразеологізмів із двочленним незмінним компонентом MI перебуває в межах від 11,67 (*яке там*) до 14,31 (*що за*), для трикомпонентних моделей – від 22,09 (*яке ж там*) до 28,4 (*де вже там*), для чотирикомпонентних моделей – від 38,32 (*вони і в Африці*) до 41,27 (*що ж це за*). Відповідно зафіксовано статистично вірогідний зв'язок між кількістю компонентів конструкції і величиною показника MI. При цьому цікаво, що чим більшою є кількість компонентів, тим меншою є частота конструкції, водночас тим більшим є коефіцієнт MI, що є цілком закономірним через урахування абсолютної частоти більшої кількості словоформ.

Отже, для всіх обстежених синтаксичних фразеологізмів показник MI відбиває високий ступінь (MI >> 3) невипадковості поєднання словоформ, що є кількісним підтвердженням стійкості зв'язку словоформ у складі незмінних компонентів фразеологізованих моделей речень.

Результати обчислень показників асоціації Dice та gmean подаємо в таблиці 2. Ці коефіцієнти обраховували з точністю до п'яти знаків після коми з метою забезпечення точності результатів, які є значно меншими за 1.

Таблиця 2

Показник асоціації Dice та gmean для моделей
синтаксичних фразеологізмів в українській мові за даними УНЛК

№ з/п	Модель синтаксичного фразеологізму	Абсолютна частота вживання незмінного компонента синтаксичного фразеологізму	Показник асоціації Dice	Показник асоціації gmean
1	*Де (вже) там* N_1Cop_f/ Inf/Adj/Adv	*де там 552*	0,14384	0,14518
		де вже там 59	0,01522	0,01532
2	*Ну і/й* N_1 Cop_f	*ну і 823*	0,23008	0,24308
		ну й 897	0,25278	0,26654
3	*Оце так* N_1 Cop_f	*400*	0,12853	0,14867
4	*Теж мені* N_1 Cop_f	*201*	0,07639	0,07641
5	*Чим не* N_1 Cop_f	*478*	0,12285	0,12671
6	*Що (ж це) за* N_1 Cop_f	*що за 2646*	0,54703	0,54703
		що це за 738	0,15518	0,15523
		що ж це за 206	0,04452	0,04459
7	*Яке (вже/ж) там* N_1 Cop_f /Inf/Adv	*яке там 218*	0,06283	0,06289
		яке вже там 17	0,00456	0,00469
		яке ж там 7	0,00188	0,00189
8	*Який там* N_1 Cop_f	*який там 358*	0,09334	0,10224
		яка там 397	0,10376	0,10466

		які там 474	0,12461	0,12561
9	Cop_f *не до* N_2	*не до 1841*	0,38002	0,38002
10	N_1 (*він*) *і в Африці* N_1 Cop_f	*і в Африці 30*	0,00891	0,01321
		він і в Африці 5	0,00142	0,00192
		вона і в Африці 5	0,001404	0,0019
		воно і в Африці 0	-	-
		вони і в Африці 3	0,00083	0,001127

На підставі викладених даних можна зробити висновок, що чим більшою є кількість компонентів конструкції, тим меншим є значення показників асоціації Dice і gmean. Привертають увагу випадки, коли отримані значення для двох коефіцієнтів є близькими і навіть тотожними (див. моделі *Що (ж це) за* N_1 Cop_f, Cop_f *не до* N_2 та ін.). Такі значення отримуємо за умови, коли абсолютні частоти словоформ у складі конструкції є близькими (напр., *що* 4843 *і за* 4831; *не* 4844 і *до* 4845).

Варто зазначити, що коефіцієнти Dice і gmean не враховують розмір обстежуваного корпусу, а отримувані значення перебувають у межах від 0 до 1, що ускладнює розмежування випадкових і невипадкових поєднань слів, переконливої інтерпретації яких у науковій літературі на сьогодні не запропоновано.

Отже, показник асоціації MI видається найбільш придатним для визначення коректності виділення моделі синтаксичного фразеологізму та вірогідності встановлення стійкості поєднання двох або більше словоформ у межах незмінного компонента моделі речення. Коефіцієнт MI враховує всі важливі параметри вживання конструкції (частоту конструкції, частоту кожної словоформи у її складі) та розмір корпусу, в межах якого здійснюється статистичне дослідження. Безумовною перевагою цього показника асоціації є можливість здійснення обчислень для будь-якої кількості словоформ.

Проведений 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. При цьому встановлена закономірність: чим більшою є кількість компонентів конструкції, тим більшим є ступінь їхньої зв'язаності.

Перспективою наступного етапу дослідження є обчислення модифікованих показників асоціації MI – MI^3, MI log Freq для моделей синтаксичних фразеологізмів української мови з метою збільшення значущості частоти конструкції, а не її окремих компонентів, та належного врахування значень низькочастотних конструкцій.

Список використаної літератури

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 : спец. 13.00.02 «Теория и методика обучения и воспитания (русский язык как иностранный)» / Л. А. Балобанова / Московский гос. ун-т имени М. В. Ломоносова. – М., 2004. – 28 с.

2. Бобкова Т. В. Теоретико-методологічні підходи до вивчення колокацій у сучасному мовознавстві / Т. В. Бобкова // Вісник КНЛУ. Серія Філологія. – 2014. – Том 17. № 2. – С. 14-22.

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

4. 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.

5. 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.

6. Карпіловська Є. А. Вступ до прикладної лінгвістики: комп'ютерна лінгвістика: Підручник / Є. А. Карпіловська. – Донецьк : ТОВ «Юго-Восток, Лтд», 2006. – 188 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 : спец. 10.02.21 «Прикладная и математическая лингвистика» / М. В. Хохлова / Санкт-Петербургский государственный университет. – Санкт-Петербург, 2010. – 26 с.

11. Хохлова М. В. Экспериментальная проверка методов выделения коллокаций / М. В. Хохлова // Slavica Helsingiensia 34. Инструментарий русистики: корпусные подходы. – Редколл.: А. Мустайоки, М. В. Копотев, Л. А. Бирюлин, Е. Ю. Протасова. – Helsinki : Department of Slavonic and Baltic Languages and Literatures, 2008. – С. 343-357.

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

13. 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.

14. Dunning Ted E. Accurate methods for the statistics of surprise and coincidence / Ted E. Dunning // Computational Linguistics. – 1993. – 19(1). – P. 61–74.

15. Church K., Hanks P. Word association norms, mutual information, and lexicography / K. Church, P. Hanks // Computational Linguistics. – #16(1). – 1990. – P. 22–29.

16. Church K., Hanks P., Hindle D., Gale W. Using Statistics in Lexical Analysis / K. Church, P. Hanks, D. Hindle, W. Gale, U. Zernik (ed) // Lexical Acquisition: Using On-line Resources to Build a Lexicon. – Lawrence Erlbaum, 1991: http://www.cs.jhu.edu/~kchurch/wwwfiles/publications.html.

17. Everitt B.S. The Cambridge Dictionary of Statistics. 2nd edition / B.S. Everitt. – Cambridge : Cambridge University Press, 2002. – 410 pp.

18. 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

19. 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 pp.

20. Fillmore Charles J. The Mechanisms of «Construction Grammar» / Charles J. Fillmore // Proceedings of the Fourteenth Annual Meeting of the Berkeley Linguistics Society. – 1988. – Pp. 35–55.

21. 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). – Pp. 501–538.

22. Fried Mirjam. Constructions and Frames as Interpretive Clues / Mirjam Fried // Belgian Journal of Linguistics. – 2010. – Vol. 24. Frames: from Grammar to Application, ed. by P. Sambre and C. Wermuth. – Pp. 83–102.

23. 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.

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

25. Gries S. Th., Stefanowitsch A. Extending Collostructional Analysis: a Corpus-Based Perspective on 'Alternations' / Anatol Stefanowitsch, Stefan Th. Gries // International Journal of Corpus Linguistics. – 2004. – 9(1). – Pp. 97–129.

26. 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.

27. 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 pp.

28. Stefanowitsch A., Gries S. Th. Collostructions: Investigating the Interaction between Words and Constructions / Anatol Stefanowitsch, Stefan Th. Gries // International Journal of Corpus Linguistics. – 2003. – 8–2. – Pp. 209–43.

29. Stefanowitsch A., Gries S. Th. Covarying Collexemes / Anatol Stefanowitsch, Stefan Th. Gries // Corpus Linguistics and Linguistic Theory. – 2005. – 1–1 – Pp. 1–43.

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

Аннотация

Ситарь А. В. Статистические критерии анализа синтаксических фразеологизмов.

Статья посвящена статистическим критериям анализа синтаксических фразеологизмов на материале украинского языка. Синтаксические фразеологизмы рассмотрены с позиций конструкционной грамматики и интерпретированы как один из типов некомпозиционных языковых знаков – конструкций.

Проанализированы основные меры ассоциации: MI, t-score, log-likelihood, Dice, gmean. Выбрана мера ассоциации MI в качестве статистического критерия, который дает возможность определить коэффициент неслучайности сочетания двух и более слов в тексте, учитывает частоту конструкции, частоту ее компонентов, размер корпуса и имеет формулу в общем виде для конструкций з любым количеством компонентов.

Представлены результаты проведенного статистического анализа моделей синтаксических фразеологизмов украинского языка по данным Украинского национального лингвистического корпуса. Определено, что все исследованные по мере ассоциации MI модели синтаксических фразеологизмов имеют высокую (MI >>3) степень неслучайности сочетания компонентов, входящих в состав неизменяемой части предложения, то есть характеризуются статистически доказанной связностью.

Ключевые слова: конструкция, конструкционная грамматика, корпус текстов, синтаксический фразеологизм, статистический анализ, мера ассоциации, фразеологизированное предложение.

Summary

Sytar H. V. Statistical Criteria of Analysis of Syntactic Idioms.

The article is devoted to the statistical criteria of analysis of syntactic idioms based on the Ukrainian language. Syntactic idioms are considered in terms of construction grammar and interpreted as one of the types of non-compositional language signs – constructions.

The main association measures were analyzed: MI, t-score, log-likelihood, Dice, gmean. Association measure of MI was chosen as a statistical criterion that enables determination of the non-randomness coefficient of two or more words combination in the text, takes into account the frequency of construction, frequency of its components, and size of the corpus and has a formula in general form for constructions with any amount of components.

Paper reports the results of the conducted statistical analysis of syntactic idioms models of the Ukrainian language according to the Ukrainian National Linguistic Corpus. It was found that all the analyzed in terms of MI association measure syntactic idioms models are of high (MI >>3), degree of the non-randomness of components combination that make up the constant part of the sentence, that is characterized by statistically proven coherence.

Keywords: construction, construction grammar, text corpus, syntactic idiom, statistical analysis, association measure, sentence with phraseological structure.

УДК 821.161.2-5

О. Є. Соловей

ДО ПРОБЛЕМИ ВІДЧУЖЕННЯ У НОВЕЛІ ІГОРЯ КОСТЕЦЬКОГО «МИ З НЕДЖ»

Реферат. Стаття присвячена проблемі відчуження в малій прозі Ігоря Костецького раннього (абсурдистського) періоду творчости й розглядається конкретно на матеріялі новели «Ми з Недж» (написана 1944-го року в Німеччині). Центральною проблемою в малій прозі письменника цього періоду є проблема відчуження (алієнації), реалізована як на змістовому рівні, так і на рівні форми (за влучним висловом письменника, «це гротескова оптика й учуднена пластика»). Чи не найзаповітнішою метою письменника було запропонувати шляхи подолання цієї понад гострої проблеми, пов'язаної з руйнівними наслідками Другої світової війни: розривом міжлюдських стосунків, кризою віри та гуманізму, цинізмом, надмірним (егоїстичним) раціоналізмом, недовірою, зневірою у вічних цінностях тощо. Пробиваючись крізь тугі шари колючої (учудненої) мови твору, персонажі насправді долають перешкоди, що виникають перед ними, заважаючи відчути в людині людину (так звану суб'єктність людської особистости) і поділитися з нею власним теплом і любов'ю. З дрібних деталей і туманних реплік раз-у-раз виступають і надалі лише увиразнюються чіткі контури гуманістичного материка художнього світу І.Костецького, письменника, місією якого було довести власним ідеалізмом (життєвим і творчим прикладом), що неможливе – можливе.

Ключові слова: *відчуження, експресіонізм, абсурдизм, гуманізм, ідеалізм, етика, свобода.*

Якби в сорокових роках з нами
хтось почав розмову про реалізм,
ми ревнули б з люті. Реалізм був
синонімом Чапленка, а таке
для нас виключалося заздалегідь.
Ігор Костецький, 1962 р.

Українська сучасна література має
повне право фігурувати на рівних засадах
серед інших літератур світу, а ми
свідомо знижуємо її культурний рівень.
В.Петров. «Злидні днів. До питання
про "багату літературу"», 1946 р.

Існує думка, витворена ще в 40-х – 50-х роках минулого століття серед української еміґрації, що творчість Ігоря Костецького призначена не для кожного реципієнта, тобто не кожному читачеві доступна [див. хоча б: 1; 8; 13]. Про відсутність *свого* читача говорив і сам письменник у відомому інтерв'ю [2, с. 109]. Відтак, на перших підступах до відкриття масштабного доробку письменника в авторитетній і резонансній монографії С.Павличко з'явилося визначення письменника як *нігілістичного модерніста* [див.: 6]. Насправді ж, це зовсім не так, що спокійно та аргументовано довели україністи з діяспори М.Стех [10, с. 13] і Г.Грабович [див.: 3]. Гостро учуднена форма творів І.Костецького з його «золотих сорокових років» [5, с. 129], що є об'єктивним (сказати б, іманентним) явищем експресіоністичної поетики, у жодному разі не дегуманізує зміст його творів, радше навпаки [див.: 14]. Гальмування процесу читання, сприйняття та рецепції внаслідок виникнення учудненої поетики призводить до переорієнтації з механістичного читання до цілком усвідомленого взаємнення з «другою реальністю» (мистецьким простором літератури). Упродовж останніх п'ятнадцяти років з'явилося декілька студій, у яких здійснено плідну спробу відчутного