Vitae
Natalia Darchuk, Doctor of Philology, professor, professor of Department of Ukrainian Language and Applied Linguistic in Taras Shevchenko National University of Kyiv. Her areas of research interests include applied linguistics and computational linguistics.
Correspondence: nataliadarchuk@gmail.com

Ганна Ситар

DOI 10.31558/1815-3070.2018.36.24
УДК 81'373.7:81'32

СТАТИСТИЧНИЙ АНАЛІЗ ЦІЛІСНИХ СЛОВОСПОЛУЧЕНЬ: НА МАТЕРІАЛІ УКРАЇНСЬКОГО НАЦІОНАЛЬНОГО ЛІНГВІСТИЧНОГО КОРПУСУ[1]

Стаття продовжує цикл публікацій, присвячених статистичному аналізу стійких одиниць української мови. За даними Українського національного лінгвістичного корпусу визначено ступінь невипадковості поєднання словоформ дво-, три- і чотирикомпонентних цілісних словосполучень української мови шляхом обчислення показника асоціації mutual information (MI).

Для всіх обстежених цілісних одиниць властива невипадковість поєднання словоформ (результати MI перебувають у діапазоні від 8,64 до 44,63). У межах одного корпусу текстів величина MI залежить від таких чинників, як абсолютна частота конструкції, абсолютна частота її компонентів, кількість компонентів і тип цілісного словосполучення.

Ключові слова: показник асоціації, фразеологічна одиниця, mutual information, статистика, цілісне словосполучення, українська мова.

Постановка проблеми, актуальність дослідження. Сучасна лінгвістика визначає статистичні дослідження як виключно корпуснобазовані, тобто вчені вважають, що вірогідні статистичні результати можна одержати тільки на підставі аналізу репрезентативного корпусу текстів. Статистичний аналіз стійких одиниць різних типів, виконаний на корпусному матеріалі, є важливим завданням лінгвістичної статистики, здатним дати чіткі кількісні критерії зарахування мовних одиниць до класу стійких, що можуть бути використані для їх автоматичної ідентифікації в корпусі текстів. Процедуру такого аналізу на матеріалі синтаксичних фразеологізмів української мови викладено у працях (Sytar, "Statystychni Kryteriyi Analizu Syntaksychnykh Frazeolohizmiv"; Sytar, "Statystychnyi analiz frazeolohizovanykh rechen…"; Sytar, "Syntaksychni frazeolohizmy v rozrizi konstruktsiinoi hramatyky").

Обчислення показників (індексів) асоціації (англ. association measures, measures of association) як метод визначення випадковості / невипадковості поєднання певних одиниць може бути застосований для різних типів конструкцій. Пропонована стаття присвячена статистичному аналізу цілісних словосполучень.

У трактуванні словосполучення маємо опертям традиційний підхід, згідно з яким його визначають як непредикативну синтаксичну одиницю, «компонентами якої є слово та форма слова або кілька форм слів, з'єднаних між собою підрядним синтаксичним зв'язком» (Zahnitko, "Slovnyk suchasnoyi linhvistyky: ponyattya i terminy", 1040).

У розгалуженій класифікації словосполучень цілісні[2] одиниці посідають особливе місце й охоплюють кілька структурних і семантичних різновидів. Вони становлять один із трьох типів словосполучень, які виділяють за ступенем злиття компонентів. За цією ознакою з-поміж словосполучень Анатолій Загнітко розмежовує:

1) вільні словосполучення;

2) синтаксично зв'язані словосполучення;

3) фразеологічно зв'язані словосполучення (Zahnitko, "*Teoretychna hramatyka ukrayins'koyi movy: Syntaksys*", 63).

На думку мовознавця, визначальними ознаками синтаксично зв'язаних (або нечленованих, неподільних, цілісних) словосполучень є такі: виконання ролі одного члена речення, наявність структури і граматичного

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

[2] В україністиці, крім терміна «цілісні словосполучення», у межах різних підходів до кваліфікації цих одиниць та створення різних класифікацій дослідники використовують також терміни «неподільні словосполучення», «синтаксично неподільні словосполучення», «семантично неподільні словосполучення», «нерозкладні словосполучення», «нечленовані словосполучення», «синтаксично нечленовані словосполучення» і под. (Zahnitko, Balko, Maksymiuk, Lychuk та ін.).

значення, пов'язаність компонентів одним з різновидів підрядного зв'язку, наприклад: *три роки, один із нас, дівчина з гарними очима* (Zahnitko, "*Teoretychna hramatyka ukrayins'koyi movy: Syntaksys*", 63-64).

Оксана Максим'юк звертає увагу на інформативну недостатність, синсемантичність стрижневих слів, що входять до складу цілісних словосполучень, а також на комплетивні (доповнювальні) відношення як типові для словосполучень цього різновиду, наприклад: *зграя лебедів, тарілка супу* (Maksymiuk, 8 і далі). За нашими спостереженнями, крім комплетивних відношень, цілісним словосполученням властиві також атрибутивні відношення, наприклад: *чоловік високого зросту, дівчина з карими очима, суддя міжнародної категорії.*

Класифікація цілісних словосполучень на сьогодні залишається дискусійною проблемою синтаксису. У цьому дослідженні враховуємо 10 типів цілісних словосполучень, виділених Мариною Балко:

1) словосполучення з кількісним значенням: *два хлопці, кілька дівчат*;
2) словосполучення зі значенням вибірковості: *дехто із студентів, четверо з них*;
3) словосполучення зі значенням сумісності: *Микола з другом, діти з учителем*;
4) словосполучення характеризувальної семантики: *людина інтелектуальної праці*, *чоловік високого зросту*;
5) словосполучення фонової семантики: *тим часом, зимового вечора;*
6) словосполучення якісної і станової семантики: *надзвичайна подія, добрий господар*;
7) словосполучення з обмежувальною семантикою: *у період з 2014 року по 2018 рік; дорога з міста до села*;
8) сполуки фазових, модальних дієслів, слів категорії стану з інфінітивом (складені присудки): *мав працювати, перестати ходити*;
9) словосполучення із семантикою невизначеності: *щось біле, щось радикальне*;
10) метафоричні й перифрастичні конструкції: *сурма великого горя, геній українського народу* (Balko, 164-168).

Матеріал і методи дослідження. Об'єктом статистичного аналізу стали 56 цілісних словосполучень різних структурних (за кількістю компонентів, їх частиномовним вираженням) і семантичних типів. Серед них 20 двокомпонентних одиниць, 20 – трикомпонентних і 16 – чотирикомпонентних. Вірогідність одержаних результатів забезпечено виконанням обчислень на матеріалі значного за обсягом й індексованого корпусу текстів − Українського національного лінгвістичного корпусу (далі УНЛК) Українського мовно-інформаційного фонду НАН України. Загальна кількість слововживань у корпусі в період здійснення підрахунків становила 189200000 одиниць.

У статистиці відомим є індекс асоціації *mutual information* (англ. взаємна, спільна інформація, далі MI). За допомогою MI визначають невипадковість (залежність) послідовності певних явищ або подій, у нашому випадку − словоформ у корпусі текстів (Fano). Вперше в лінгвістичних дослідженнях його застосували Кеннет Ворд Чарч (Kenneth Ward Church) та Патрік Генкс (Patrick Hanks) для виявлення залежності поєднання двох слів в англійських корпусах текстів (Church, Hanks).

Через багатокомпонентність значної частини цілісних словосполучень використовуємо формулу (1), наведену у працях (Petrovic, Snajder, Basic, Kolar: 323; Yagunova, Pivovarova: 586) і призначену для конструкцій з будь-якою кількістю компонентів:

(1)

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

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 слововживань у корпусі;
log_2 – логарифм числа за основою 2.

Мета пропонованого дослідження – установити ступінь невипадковості поєднання компонентів цілісних словосполучень шляхом обчислення показника асоціації MI. Для досягнення поставленої мети розв'язано такі **завдання:**

1) укладено робочий варіант реєстру цілісних словосполучень, що охоплює одиниці з різною кількістю компонентів, різним морфологічним наповненням та нетотожними семантико-синтаксичними відношеннями;
2) з УНЛК отримано частотні дані для цілісних словосполучень;
3) виконано обчислення за формулою MI для багатокомпонентних одиниць;

4) проаналізовано отримані результати та виявлено взаємозв'язки величини MI й типу цілісного словосполучення.

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.

Наведемо приклади виконаних обчислень. Для визначення ступеня невипадковості поєднання складників цілісного словосполучення *другого дня* з УНЛК було отримано такі кількісні дані: абсолютна частота конструкції становить 723, абсолютна частота словоформи *другого* − 3656; *дня* − 3623. Підставляючи ці дані до формули (1), отримуємо:

$$\text{MI (другого дня)} = log_2 \frac{723 \times 189200000}{3656 \times 3623} = 13{,}334169239 \approx 13{,}33$$

Коефіцієнт MI обраховували з точністю до двох знаків після коми. Отримані результати MI для двокомпонентних цілісних словосполучень подано в таблиці 1.

Таблиця 1

Показник асоціації MI для двокомпонентних цілісних словосполучень за даними УНЛК

№ з/п	Цілісне словосполучення	Абсолютна частота вживання цілісного словосполучення	Абсолютна частота вживання словоформ-компонентів цілісного словосполучення	Показник асоціації MI
1	*Багато грошей*	355	*багато 4499; грошей 2537*	12,52
2	*два хлопці*	80	*два 4778; хлопці 1887*	10,71
3	*дещо дивне*	14	*дещо 3103; дивне 1327*	9,33
4	*добрий господар*	67	*добрий 2391; господар 1773*	11,55
5	*другого дня*	723	*другого 3656*; *дня 3623*	13,33
6	*зграя птахів*	17	*зграя 709; птахів 1262*	11,81
7	*зимового вечора*	54	*зимового 571; вечора 2135*	13,03
8	*кілька книжок*	110	*кілька 3673; книжок 1496*	11,89
9	*літр молока*	25	*літр 249; молока 1320*	13,81
10	*мав працювати*	21	*мав 3731; працювати 2663*	8,64
11	*надзвичайна подія*	71	*надзвичайна 580; подія 1471*	13,94
12	*повинен відповісти*	27	*повинен 2980; відповісти 1812*	9,89
13	*почав робити*	163	*почав 3114; робити 3498*	11,47
14	*почав ходити*	231	*почав 3114; ходити 2106*	12,70
15	*тим часом*	2508	*тим 5127; часом 4209*	14,42
16	*троє дівчат*	37	*троє 1959; дівчат 1680*	11,05
17	*хочу піти*	80	*хочу 2868; піти 2341*	11,14
18	*шматок хліба*	371	*шматок 1488; хліба 1962*	14,55
19	*щось біле*	98	*щось 3759; біле 1409*	11,77
20	*щось радикальне*	11	*щось 3759; радикальне 200*	11,43

Як видно з таблиці 1, коефіцієнт MI для двокомпонентних цілісних словосполучень становить від 8,64 (*мав працювати*) до 14,55 (*шматок хліба*). При цьому показово, що різні типи цілісних словосполучень мають нетотожні результати, зокрема, низький індекс MI мають складені присудки, високий − словосполучення фонової (часової) семантики та конструкції з партитивним значенням (останній тип цілісних словосполучень виділяє Оксана Максим'юк (Maksymiuk, 8). Важливим моментом також вважаємо вищі абсолютні частоти як цілісних словосполучень загалом, так і їх складників, порівняно з відповідними частотами лексичних фразеологізмів, прислів'їв та приказок (пор. результати, наведені у працях (Sytar, "Syntaksychni frazeolohizmy v rozrizi konstruktsiinoi hramatyky"; Sytar, "Statystychnyy analiz prysliv"yiv i prykazok...").

Контрольна величина, починаючи від якої кваліфікуємо поєдання слів як невипадкове, залежить від низки чинників: абсолютної частоти конструкції, абсолютної частоти її окремих складників і від розміру корпусу. Для Українського національного лінгвістичного корпусу, розмір якого під час виконання підрахунків складав 189 200 000 слововживань, ця контрольна величина дорівнює 7,56 (процедуру виведення контрольної величини викладено у праці (Sytar, "Syntaksychni frazeolohizmy v rozrizi konstruktsiinoi hramatyky", 310-311)):

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

Відповідно результати MI, одержані для всіх проаналізованих двокомпонентних цілісних словосполучень, засвідчують невипадковість поєднання компонентів у їх складі.

Покажемо приклад обчислень для трикомпонентних словосполучень. УНЛК дає такі кількісні дані для конструкції *ніхто з нас*: абсолютна частота конструкції становить 354, абсолютна частота словоформи *ніхто* – 3438, *з* – 6301, *нас* − 4291. Підставляючи ці дані до формули (1), отримуємо:

$$\text{MI (ніхто з нас)} = log_2 \frac{354 \times (189200000)^2}{3438 \times 6301 \times 4291} = 27{,}0224621 \approx 27{,}02$$

Статистичні дані, одержані для три- і чотирикомпонентних цілісних словосполучень, наведено в таблицях 2 і 3 відповідно.

Таблиця 2

Показник асоціації MI для трикомпонентних цілісних словосполучень за даними УНЛК

№ з/п	Цілісне словосполучення	Абсолютна частота вживання цілісного словосполучення	Абсолютна частота вживання словоформ-компонентів цілісного словосполучення	Показник асоціації MI
1	*Батько з матір'ю*	30	*батько 2703; з 6301; матір'ю 577*	26,70
2	*борошно вищого ґатунку*	4	*борошно 603; вищого 1528; ґатунку 332*	28,80
3	*геній українського народу*	4	*геній 684; українського 2653; народу 2776*	24,76
4	*дехто зі студентів*	5	*дехто 1866; зі 4872; студентів 1656*	23,50
5	*діти шкільного віку*	9	*діти 2979; шкільного 763; віку 2659*	25,67
6	*до пізнього вечора*	212	*до 6278; пізнього 759; вечора 2135*	29,47
7	*кращий з учнів*[3]	0	-	-
8	*людина інтелектуальної праці*	3	*людина 3660; інтелектуальної 763; праці 3467*	23,40
9	*ми з другом*	12	*ми 5060; з 6301; другом 1175*	21,57
10	*найкращий з усіх*	30	*найкращий 1612; з 6301; усіх 4321*	24,54
11	*на другий день*	838	*на 6328; другий 4156; день 4066*	28,06

[3] 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 «-».

12	*ніхто з нас*	354	*ніхто 3438; з 6301; нас 4291*	27,02
13	*один із нас*	78	*один 4977; із 5792; нас 4291*	24,43
14	*речі великого розміру*	1	*речі 3121; великого 3198; розміру 1265*	21,44
15	*система шкільних закладів*	0	-	-
16	*суддя міжнародної категорії*	3	*суддя 1071; міжнародної 1385; категорії 1755*	25,30
17	*сурма великого горя*	1	*сурма 233; великого 3198; горя 1290*	25,15
18	*четверо з них*	69	*четверо 1282; з 6301; них 5527*	25,72
19	*чоловік високого зросту*	7	*чоловік 3019; високого 2169; зросту 888*	25,36
20	*я з подругою*	2	*я 4815; з 6301; подругою 389*	22,53

Таблиця 3

Показник асоціації MI для чотирикомпонентних цілісних словосполучень за даними УНЛК

№ з/п	Цілісне словосполучення	Абсолютна частота вживання цілісного словосполучення	Абсолютна частота вживання словоформ-компонентів цілісного словосполучення	Показник асоціації MI
1	*Багато з моїх знайомих*	1	*багато 4499 з 6301; моїх 2395; знайомих 1452*	36,00
2	*дехто з моїх друзів*	5	*дехто 1866; з 6301; моїх 2395; друзів 1885*	39,21
3	*дехто з моїх колег*	4	*дехто 1866; з 6301; моїх 2395; колег 1082*	39,69
4	*дівчина з синіми очима*	3	*дівчина 2035; з 6301; синіми 585; очима 2904*	39,76
5	*дівчина з карими очима*	3	*дівчина 2035; з 6301; карими 232; очима 2904*	41,10
6	*жінки з довгими ногами*	1	*жінки 2672; з 6301; довгими 1286; ногами 2294*	36,99
7	*з вересня по грудень*	3	*з 6301; вересня 1888; по 6024; грудень 660*	38,64
8	*з Києва до Вінниці*	1	*з 6301; Києва 1979; до 6278; Вінниці 434*	37,54
9	*з міста до села*	6	*з 6301; міста 3063; до 6278; села 2396*	37,03
10	*з січня по жовтень*	6	*з 6301; січня 1703; по 6024; жовтень 806*	39,50
11	*із села до міста*	7	*із 5792; села 2396; до 6278; міста 3063*	37,37
12	*людина з великим серцем*	4	*людина 3660; з 6301; великим 2738; серцем 1702*	37,88
13	*людина з добрим серцем*	2	*людина 3660; з 6301; добрим 1513; серцем 1702*	37,73
14	*майстер спорту міжнародного класу*	13	*майстер 1303; спорту 826; міжнародного 1446; класу 2075*	44,63
15	*хтось з наших людей*	2	*хтось 3186; з 6301; наших 3266; людей 4272*	35,49
16	*чоловік з високим чолом*	1	*чоловік 3019; з 6301; високим 1942; чолом 709*	37,91

Дані, наведені в таблицях 2 і 3, засвідчують, що коефіцієнт MI для всіх цілісних словосполучень є високим: для трикомпонентних одиниць він перебуває в межах від 21,44 (*речі великого розміру*) до 29,47 (*до пізнього вечора*), тобто у 2,8 – 3,9 разу більший за 7,56; для чотирикомпонентних – від 35,49 (*хтось з наших людей*) до 44,63 (*майстер спорту міжнародного класу*), тобто в 4,7 – 5,9 разу більший за контрольну величину.

З-поміж трикомпонентних одиниць низькі результати MI мають словосполучення зі значенням сумісності, у складі яких є займенник, середні – словосполучення зі значенням вибірковості, високі – словосполучення темпоральної семантики, зокрема внаслідок високої абсолютної частоти конструкцій у корпусі текстів. Цікаво, що метафоричні словосполучення як одиничні, часто оказіональні, утворення мають низьку абсолютну частоту, проте досить високий показник MI, очевидно, внаслідок порівняно невисоких частот їх складників.

Усі чотирикомпонентні цілісні словосполучення мають низьку абсолютну частоту вживання в корпусі текстів. Діапазон індексу MI для цих одиниць є невеликим, близькі результати MI зафіксовано для словосполучень зі значенням вибірковості та словосполучень з обмежувальною семантикою, вищі – для словосполучень характеризувальної семантики за умови невисокої частоти хоча б одного з її компонентів.

Статистично вірогідного зв’язку між результатами MI та типом переданого семантико-синтаксичного відношення не зафіксовано.

Висновки. Отримані результати обчислень для цілісних словосполучень, виконаних за даними Українського національного лінгвістичного корпусу, доводять, що для всіх проаналізованих одиниць властива невипадковість поєднання словоформ: коефіцієнт MI перебуває в діапазоні від 8,64 до 44,63. У межах одного корпусу текстів величина MI залежить від таких чинників, як абсолютна частота конструкції, абсолютна частота її компонентів, кількість компонентів і тип цілісного словосполучення.

Наведені статистичні дані корелюють із результатами, отриманими для інших типів фразеологічних одиниць – лексичних фразеологізмів, синтаксичних фразеологізмів, прислів’їв і приказок. Результати MI для цілісних словосполучень є прогнозовано нижчими порівняно з іншими стійкими одиницями через меншу кількість компонентів у їх складі.

Перспективи. Подальший етап дослідження передбачає виконання статистичного аналізу інших типів стійких одиниць, зокрема складених прийменникових еквівалентів та ін.

References

Balko, Maryna. “*Semantyko-syntaksychni i strukturni aspekty tsilisnykh slovospoluchen' suchasnoyi ukrayins'koyi movy (Semantic-syntactical and Structural Aspects of Indivisible Word-combinations of Modern Ukrainian Language)*”: Diss. Zaporizhzhya National U, 2004. Abstract. Print.

Balko, Maryna. *Aktual'ni problemy teoriyi slovospoluchennya suchasnoyi ukrayins'koyi movy (Actual Problems of the Word-combination Theory of Modern Ukrainian Language)*: [monohrafiya]. Dnipropetrovs'k: Svidler, 2014. Print.

Church, Kenneth Ward, and Hanks, Patrick. “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.

Lychuk, Mariya. “Syntaksychno nechlenovani slovospoluchennya: ustalenist' termina, istoriya doslidzhennya (Syntactically Nondivided Word-combinations: Term Sustainability, History of Research)”. *Linguistic Bulletin*, 21 (2016): 142-148. Print.

Maksymiuk, Oksana. “*Koreferentnist' nerozkladnykh komponentiv u strukturi rechennya (Co-reference of Stable Components in the Structure of the Sentence).*” Diss. Chernivtsi National U, 2005. Abstract. 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 prysliv"yiv i prykazok: pokaznyk asotsiatsiyi mutual information (na materiali Ukrayins'koho natsional'noho linhvistychnoho korpusu) (Statistical Analysis of Proverbs and Sayings: Association Measure of Mutual Information (on material of Ukrainian National Linguistic Corpus).” *Лінгвістичні студії / Linguistic Studies* 35 (2018): 170-177. 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.

Zahnitko, Anatoliy. *Slovnyk suchasnoyi linhvistyky: ponyattya i terminy (Dictionary of Modern Linguistics: Concepts and Terms)*. Donets'k: DonNU, 2013. Print.

Zahnitko, Anatoliy. *Teoretychna hramatyka ukrayins'koyi movy: Syntaksys (Theoretical Grammar of the Ukrainian Language: Syntax).* Donets'k: DonNU, 2001. Print.

Надійшла до редакції 23 листопада 2018 року.

STATISTICAL ANALYSIS OF INDIVISIBLE WORD-COMBINATIONS: 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 article is devoted to the statistical analysis of indivisible word-combinations in the Ukrainian. The research was performed on the material of the Ukrainian National Linguistic Corprs of the Ukrainian Language and Information Fund of the National Academy of Sciences of Ukraine. The object of analysis is 56 homogeneous word-combinations with different components, different morphological content and non-identical semantic-syntactic relations. Among them, there are 20 two-component units, 20 three-component, and 16 – four-component ones.

Purpose: The purpose of the research is to establish the degree of indivisible word-combinations components sequence non-randomness in Ukrainian by calculating the association measure *mutual information* (MI).

Results: the frequency data for word-combination from the Ukrainian National Linguistic Corpus is received, the MI for multicomponent units is computed, the obtained results are analyzed, the correlation between the MI value and the type of indivisible word-combinations is revealed.

For the analyzed indivisible word-combinations, the MI association measure is in the range of 8.64 (Ukr. *мав працювати*) to 44.63 (Ukr. *майстер спорту міжнародного класу*), that means that the components combination in all these units is non-randomness. Within a single text corpus, the MI value depends on such factors as the absolute construction frequency, the absolute frequency of its components, the number of components and the type of indivisible word-combinations.

Discussion: MI results for indivisible word-combinations are predicted to be lower than other phraseological units due to the smaller number of components in their composition. It is important to find out that the predefined MI value depends on the type of indivisible word-combinations. The next stage of the study involves performing a statistical analysis for other types of phraseological units, in particular, composed prepositional equivalents.

Keywords: association measure, phraseological units, mutual information, indivisible word-combination, statistics, the Ukrainian language.

Vitae

Hanna Sytar is PhD of Philology, Associate Professor, Associate 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