КОГНІТИВНА ЛІНГВІСТИКА

УДК 81’378

Дем’янчук Ю. І.,
кандидат економічних наук,
Львівський державний університет безпеки життєдіяльності
y.demianchuk@gmail.com

МЕТОДИ ВИДІЛЕННЯ КОЛОКАЦІЙ ІЗ ВІЙСЬКОВИХ ТЕРМІНОЛОГІЧНИХ СПОЛУЧЕНЬ НАТО

Постановка проблеми та стан дослідження. Актуальною лінгвістичною проблемою, можна вважати корпусно-орієнтовані дослідження колокацій сталих термінологічних виразів. Саме виділення стійких словосполучень застосовується в багатьох сферах, серед яких семантичні і лексикографічні дослідження (зокрема, створення електронних словників для Національних корпусів), у сфері автоматизованого перекладу та аналізу спеціалізованих термінів.

В даному випадку, використання статистичного підходу для виділення стійких словосполучень можна вважати найбільш простим способом виявлення колокацій в тексті. За його допомогою складаються частотні списки слів, які опинилися ліворуч або праворуч від ключової леми в межах заданого діапазону. Також застосовуються статистичні заходи асоціації (log- likelihood, MI, t-score), які засновані на формулах, що використовують частоту спільного явища слів в колокації, частоти кожного компонента словосполучення, обсяг корпусу тощо. При цьому, ці частоти можуть обраховуватися також в межах потрібного діапазону. Дослідженням колокації займалися Т. В. Бобкова [2], М. В. Хохлова [7], Е. В. Ягунова [8], Л. М. Пивоварова [8] тощо. Проте проблема виділення колокацій зі спеціалізованих військових термінів НАТО, залишається недослідженою проблемою.

Мета статті – розглянути застосування статистичного методу для виділення колокацій з військових термінів НАТО. **Концептуальні завдання:** розглянути популярні корпусні методи виділення колокацій зі сталих спеціалізованих термінологічних сполучень; проаналізувати статистичний метод, для виділення колокацій із військових термінів НАТО; статистичні результати дослідження запропонувати для паралельного корпусу НКРМ.

Виклад основного матеріалу, результати дослідження. У корпусній лінгвістиці існує велика кількість досліджень і розробок, присвячених стійким сполученням які торкаються як теоретичних аспектів статистичного підходу до даного поняття, так і практичних методів виявлення колокацій. Ці методи ґрунтуються на дослідженні n-грам (зазвичай це біграми або триграми) в межах заданого контекстного вікна (діапазону). Проте найпростішим способом виявлення сполучуваності лексичних одиниць є складання частотних списків словосполучень.

Дослідник П. Браславський [1] виділяє три базові підходи до виділення термінів і термінологічних словосполучень: 1) лінгвістичний підхід, 2) статистичний підхід і 3) змішаний підхід, заснований на застосуванні статистичної, і лінгвістичної інформації.

Натомість, математичним апаратом для встановлення синтагматичного зв’язку між словами в тексті служать методи асоціації (association measures). Це,

передусім, статистичні асоціації, які поєднуються між собою синтаксичним та лексичним зв'язками. Лінійна близькість є важливою передумовою для утворення стійких поєднань. Заходи асоціації враховують як частоту спільних властивостей, так і інші параметри, насамперед частоту в даному корпусі кожного окремого елемента. Значення заходів асоціації можна вважати показниками сили синтагматичною зв'язку між елементами словосполучень.

Для вилучення складових найменувань зі сталих військових термінологічних виразів НАТО, пропонується застосувати кілька статистичних методів. Найбільш важливим, на нашу думку, є так званий критерій сили зв'язку, який використовується для визначення сили залежності між компонентами вираження. Загальна кількість цих заходів між зв'язками обраховується біграмами. Значення заходів асоціації можна вважати показниками сили синтагматичною зв'язку між елементами словосполучень.

Для опису найбільш поширених заходів найчастіше застосовуються критерії MI, t-score і log-likelihood. Окремі корпусні менеджери надають можливість обрахунку потрібних заходів. Зокрема, міра MI (mutual information), порівнює залежні контекстно-пов'язані частоти з незалежними, та словом, який з'являлося в тексті випадково. Якщо значення MI (n, c) більше визначеного значення, тоді дане поєднання слів можна вважати статистично значущим.

Міра t-score також враховує частоту спільного утворення ключового слова і його колокації, відповідаючи на запитання, наскільки невипадковою є сила асоціації (пов'язаності) між колокаціями. Також досить часто застосовується міра, log-likelihood, або логарифмічна функція правдоподібності.

Загалом, застосування статистичних заходів (MI і t-score) дають можливість охарактеризувати предметну сферу і стилістичні спеціалізованих текстів. Списки колокацій із окремих військовий термінів НАТО, отриманих за допомогою MI і t-score, принципово різні. Наприклад колокації, що виділяються за допомогою MI, дають можливість визначати назви об'єктів, терміни, складні номінації, що відображають предметну сферу, а критерій t-score спрямований на виділення "загальномовних стійких сполучень" (похідних службових слів, дискурсивних слів) і "стійких конструкцій", де і ті та інші характеризують стилістичні особливості спеціалізованих текстів (Рисунок 1, 2, 3).

Основним матеріалом дослідження є військовий термін tactical (тактичний, оперативно-тактичний) із офіційних документів НАТО. Експеримент полягав в находженні біграм, одним з компонентів яких є іменник, інший – прикметник (дієслово) (Таблиця 1, 2) [5].

Табл. 1

Статистичне виділення колокацій
із терміна tactical (тактичний, оперативно-тактичний)

1. tactical air force (тактичне авіаційне сполучення);
2. tactical air operation (тактичні дії авіації);
3. tactical battle management functions (TBMF) (функціональні обов'язки командира на тактичному рівні);
4. tactical command (TACOM) (тактичне командування);

5. tactical control (TACON) (1. тактичне управління військами (силами); 2. оперативно-тактичне управління військами (силами);
6. tactical employment (тактичний епізод, зі сценарію навчань);
7. tactical equipment (польове обладнання);
8. tactical formation (з'єднання, військове формування, що складається з декількох однорідних частин або з'єднань меншого складу);
9. tactical fuel handling equipment (TFHE) (польові технічні засоби забезпечення пальним; польові технічні засоби Служби пального);
10. tactical mission preparation (комплекс заходів, що забезпечують своєчасну готовність військ (сил) до бойових дій;
11. tactical mobility (тактична мобільність);
12. tactical support (тактичне забезпечення);
13. tactical training (бойова підготовка; тактична підготовка; навчальна підготовка на тактичному рівні);
14. tactical unit (військовий підрозділ)

Табл. 2

Частотні дані і заходи асоціації для терміна tactical "тактичний" (Перше значення для леми / друге значення для форми-додатка)

Collocation	Joint	Freq1	MI score	LL score	T score 1,96
tactical air force	2/5	1362	1,34/2,82	3,19/5,09	0,74/1,16
tactical air operation	16/11	1162	0,98/1,11	4,19/1,11	1,62/2,83
tactical battle management functions	24/14	890	2,34/1,77	2,49/0,15	1,19/2,11
tactical command	10/7	1562	2,14/0,88	2,04/4,18	1,32/2,46
tactical control	12/31	564	3,11/1,80	5,40/4,55	2,53/3,16
tactical employment	32/14	1139	2,77/2,42	3,06/2,32	2,54/0,32
tactical equipment	33/21	372	1,63/2,33	4,09/2,44	3,42/1,68
tactical formation	18/10	264	2,87/1,74	1,77/2,43	2,58/1,19
tactical fuel handling equipment	37/5	742	1,44/0,15	4,09/0, 02	1,78/0,83
tactical mission preparation	34/5	836	2,16/2,42	2,90/5,11	1,42/0,63
tactical mobility	26/7	1118	2,09/1,93	1,40/3,85	2,66/3,82
tactical support	21/18	932	0,34/1,82	2,99/4,22	3,44/3,36
tactical training	32/15	429	3,11/1,82	1,66/3,47	1,84/1,10
tactical unit	35/3	552	1,34/2,06	2,19/5,31	3,72/2,28

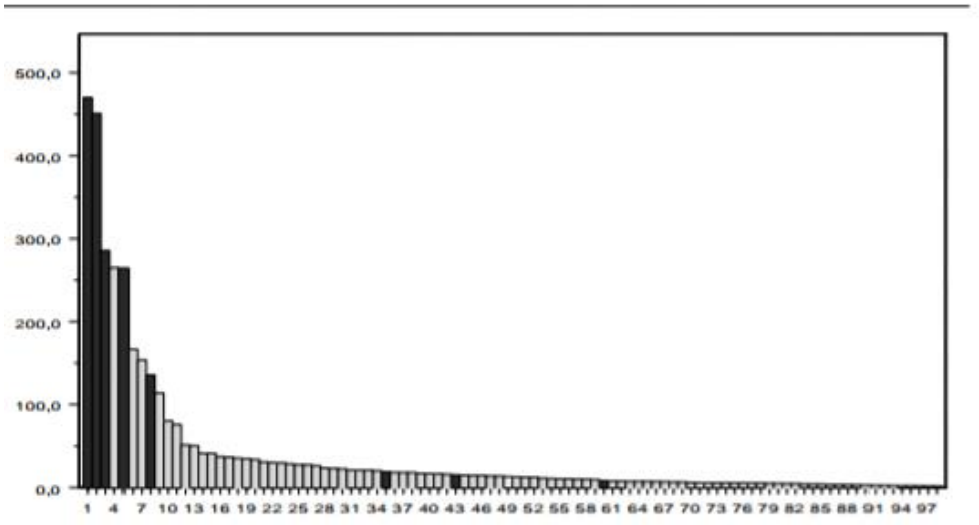

Рис.1. log-likelihood схема колокації до терміна *tactical*

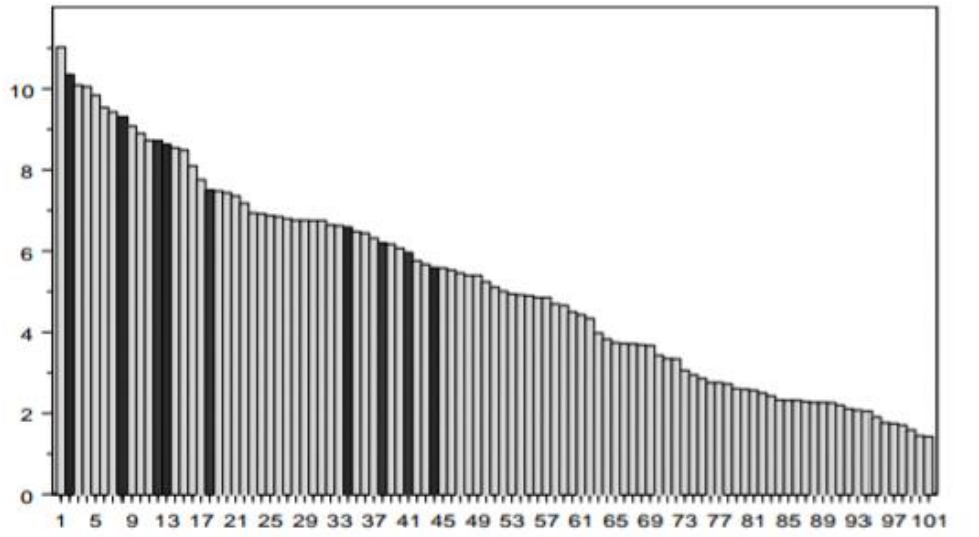

Рис. 2. MI схема колокації до терміна *tactical*

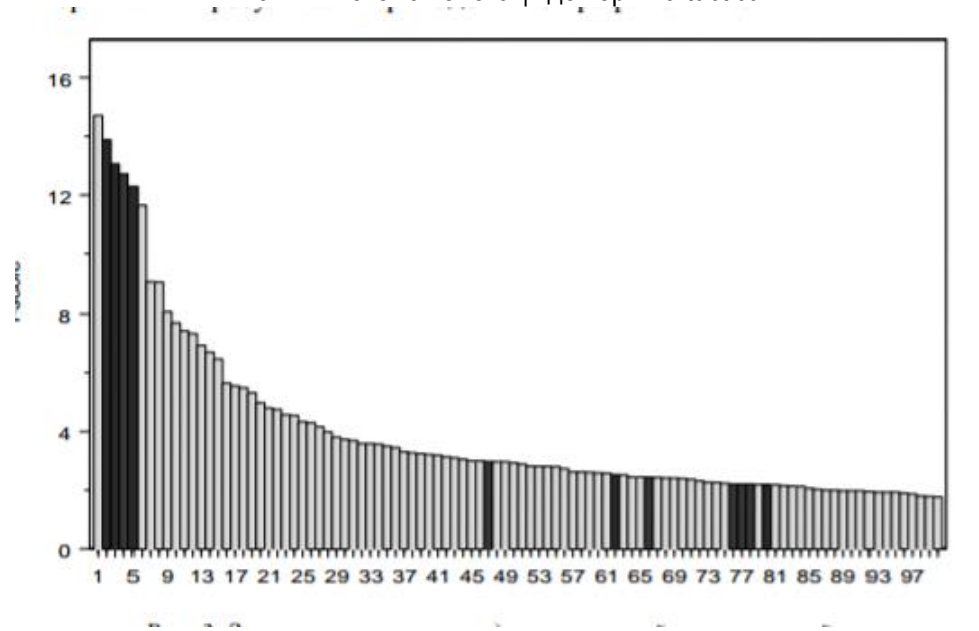

Рис. 3. t-score схема колокації до терміна *tactical*

Для всіх розглянутих поєднань характерна однакова тенденція: чим менше значення заходів, тим більша ймовірність, що ці словосполучення зафіксовано як стійкий вираз в електронному словнику. Загалом, дані про сполучуваність, які наведені в словниках, збігаються з даними, отриманими на основі заходів асоціації.

Важливим результатом дослідження є той факт, що в процесі експерименту були виділені сталі військові вирази які незафіксовані ні в одному зі словників. Аналіз таких поєднань показав, що відомі біграми, знаходяться вгорі списку (відсортованого за зменшенням). Невідомі вирази, з деякою часткою ймовірності виявляються стійкими і, відповідно, можуть бути внесені в електронний словник НКРМ [6].

Висновки. Здійснивши виділення колокації з військового терміна *tactical,* можна стверджувати, що статистичні заходи (MI і t-score) дозволяють охарактеризувати предметну сферу спеціалізованого тексту. Міра MI, дає найкращі нейтральні результати. Вона дозволяє виділити сталі спеціалізовані словосполучення та лексичні поєднання, де в якості колокацій виділяються військові назви, а також низькочастотні спеціальні терміни. До недоліків використання заходів t-score можна віднести те, що вона виділяє колокації з великою кількістю частотних слів-колокацій (стоп-слів). Тому для t-score необхідно задавати список стоп-слів, щоб відкинути непотрібні частотні слова. Багатозначні ж колокації, характеризуються високими значеннями заходів t-score. Тому, здійснене дослідження стане важливим додатком до НКРМ, як головного джерела для паралельного перекладу офіційних документів НАТО.

Література

1. Браславский П., Соколов Е. Сравнение четырех методов автоматического извлечения двухсловных терминов из текса / П. Браславский, Е. Соколов // Компьютерная лингвистика и интеллектуальные технологи : Труды международной конференции “Диалог 2006” (Бекасово, 31 мая – 4 июня 2006 г.) / [под ред. Лауфер Н. И., Нариньяни А. С., Селегея В. П.]. – М. : Изд-во РГГУ, 2006. – С. 23–36.
2. Бобкова Т. В., Теоретико-методологічні підходи до вивчення колокацій. Вісник Київського нац. лінгвістичного ун-ту. Серія: Філологія. – 2014. – Т. 17, № 2. – С. 14–22.
3. Добров Б. В. Формирование базы терминологических словосочетаний по текстам предметной области / Добров Б. В., Лукашевич Н. В., Сыромятников С. В. // Труды пятой Всероссийской научной конференции “Электронные библиотеки: перспективные методы и технологии, электронные коллекции” – RCDL2003 (Санкт-Петербург, 2003). – М. : Языки славянских культур, 2007. – С. 201–210.
4. Кобрицов Б. П. Поверхностные фильтры для разрешения семантической омонимии в текстовом корпусе / Кобрицов Б. П., Ляшевская О. Н., Шеманаева О. Ю. // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции “Диалог’2005” (Звенигород, 1–6 июня, 2005 г.) / [под ред. Кобозевой И. М., Нариньяни А. С., Селегея В. П.]. – М. : Наука, 2005. – С. 72–89.
5. Офіційний сайт архівних документів “НАТО” [Електронний ресурс]. – Режим доступу : http://www.nato.int/cps/ru/natohq/official_texts.htm.
6. Сайт Национального корпуса русского язика [Електронний ресурс]. – Режим доступу : http://www.ruscorpora.ru/corpora-biblio.html.
7. Хохлова М. В. Экспериментальная проверка методов выделения коллокаций / М. В. Хохлова // Slavica Helsingiensia 34. Инструментарий русистики: Корпусные подходы / [под ред. Мустайоки А., Копотева М. В., Бирюлина Л. А., Протасовой Е. Ю.]. – Хельсинки, 2008. – С. 343–357.

8. Ягунова Е. В., Пивоварова Л. М. Извлечение и классификация коллокаций на материале научных текстов. Предварительные наблюдения / Е. В. Ягунова, Л. М. Пивоварова. – СПб., 2010. – 250 с.

Анотація

У статті розглядаються загальновідомі теорії, принципи та методи виділення колокацій із спеціалізованої термінології. Особлива увага приділяється статистичному методу, його структурним частинам log-likelihood, MI, t-score. На думку автора, використання статистичного підходу для виділення стійких словосполучень можна вважати найбільш простим способом виявлення колокацій в тексті. За його допомогою складаються частотні списки слів, які опинилися ліворуч або праворуч від ключової леми в межах заданого діапазону. Також застосовуються статистичні заходи асоціації, які засновані на формулах, що використовують частоту спільного явища слів в колокації, частоти кожного компонента словосполучення, обсяг корпусу тощо. В процесі паралельного перекладу військових термінів із офіційних документів НАТО, вказується на можливість застосування статистичного апарату для виділення сталих виразів в Національному корпусі російської мови.

Ключові слова: колокація, спеціалізована термінологія, статистичний метод, log-likelihood, MI, t-score, паралельний переклад, НКРМ.

Summary

The article deals with the well-known theory, principles and methods of allocation of the collocation of specialized terminology. Particular attention is paid to the statistical method, its structural parts of the log-likelihood, MI, t-score. The author notes that the allocation of set phrases used in many fields, including semantic and lexicographical research (in particular, the creation of electronic dictionaries for National buildings), in automated translation and analysis of specialized terms. In this case, the use of a statistical approach to isolate stable combinations can be considered as the easiest way to identify the collocation in the text. With it consists frequency lists words that turned left or right of lemmas key within a predetermined range.It is alleged that the existing software tools automatically detect collocations based on statistical methods should be developed. During a parallel translation of the terms of the NATO military official documents indicated the possibility of using statistical tools for the isolation of stable expression in the Russian National Corpus.

Keywords: Colocation, specialized terminology, the statistical method, log-likelihood, MI, t-score, a parallel translation, NKRM.

УДК 811.111’367.7’373.7:398

Гаврилова В. В.,
аспірантка,
Бердянський державний педагогічний університет
havrilova.vika@gmail.com

СТРУКТУРНИЙ АСПЕКТ АНГЛІЙСЬКИХ ПАРЕМІЙ ІЗ КОМПОНЕНТОМ “НАЗВА СВЯТА”

Паремії як втілення багатовікового народного досвіду є важливою частиною фразеологічного складу мови. Паремії вже не одне десятиліття різнобічно досліджуються в традиційних мовознавчих, пареміологічних, фольклористичних студіях, результатом яких є сотні праць, що висвітлюють особливості структури, семантики, стилістики та функціонування паремійних одиниць на матеріалі різних мов. Питання світової та слов’янської пареміології знайшли відображення у працях багатьох лінгвістів: О. Потебні, Л. Сміта, Л. Булаховського, В. Мокієнка, М. Алефіренка, Й. Млацек, Л. Скрипник, П. Савін та ін.