УДК 811.161.2

А. Романюк, Г. Кваснюк, М. Романишин
Національний університет "Львівська політехніка",
кафедра систем автоматизованого проектування,
кафедра прикладної лінгвістики

РОЗПІЗНАВАННЯ БАГАТОСЛІВНИХ КОНСТРУКЦІЙ

Розглянуто проблему багатослівних конструкцій, яка відіграє дуже важливу роль у технології обробки природної мови. Багатослівні конструкції – це вирази, які складаються з щонайменше двох слів і можуть бути синтаксично і/або семантично ідіосинкратичними. Це зокрема складені іменники, ідіоми і фразові дієслова. У цій роботі досліджено сучасні підходи до класифікації багатослівних конструкцій, їхньої ідентифікації та видобування з текстів.

Ключові слова: багатослівні конструкції, опрацювання природної мови, складений іменник, ідіома, фразове дієслово, колокація, ідентифікація багатослівних конструкцій, видобування багатослівних конструкцій.

This paper surveys the problem of multiword expressions (MWE), which plays the important role in development of large-scale, linguistically sound natural language processing technology. Multiword expressions are expressions which are made up of at least 2 words and which can be syntactically and/or semantically idiosyncratic. This category includes such constructions as compound nouns, idioms and phrasal verbs.This paper deals with modern approaches to MWE stratification, extraction and identification.

Keywords: Multiword expressions, natural language processing, compound noun, idiom, phrasal verb, collocation, MWE identification, MWE extraction.

1. Постановка проблеми

Сьогодні існує протистояння між символічними і статистичними методами опрацювання природної мови. Є певні суперечності щодо доцільності використання статистичних методів лінгвістичного аналізу, однак сьогодні вченим необхідні нові і покращені мовні моделі для опрацювання природної мови. «Глибоке» (лінгвістично точне) опрацювання наразі перетнуло індустріальний поріг і почало слугувати основою для постійного розвитку в багатьох сферах використання.

При опрацюванні природної мови основними проблемами, серед інших, є мовна неоднозначність і багатослівні конструкції. Саме другій проблемі і присвячена ця робота. У широкому розумінні, словом ми називаємо лексичну одиницю, яка означає поняття. Наприклад, слова *train*,*water* і *ability* можна легко позначити відповідними концептами. Навіть у випадку слова *blackboard*, ми ментально сполучуємо його з певним концептом. А от колокація *Prime Minister* є багатослівною конструкцією (англійською Multiword expression або MWE), інтерпретація якої перевищує межі слова. Найпростіше багатослівні конструкції можна визначити як «ідіосинкратичні інтерпретації, що виходять за межі слова» [11]. Отже, цілу конструкцію потрібно обробляти як єдину лексичну одиницю. Ця тема є актуальною, оскільки сьогодні не існує високоефективних методів опрацювання багатослівних конструкцій. У цій статті ми аналізуємо основні методи ідентифікації, класифікації та видобування багатослівних конструкцій.

Поставлену проблему ускладнює і те, що кількість MWE в лексиконі мовця є майже такою самою, як кількість власне слів. А, як зазначає Фельбаум, 41% статей у WordNet 1.7 є

багатослівними конструкціями [6]. MWE є у текстах усіх жанрів, що знову ж таки доводить значимість цієї проблеми для усіх видів опрацювання природної мови.

2. Аналіз останніх досліджень

Сьогодні теорія багатослівних конструкцій є ще малорозвиненою і перебуває у процесі свого становлення, оскільки довгий час вона перебувала поза увагою науковців. Багато науковців працюють над цією проблемою. Сьогодні MWE опрацьовують у багатьох проектах, які розробляють великомасштабні, лінгвістично точні граматики, зокрема *ParGram Project (http://www.parc.xerox.com/ist1/groups/nltt/pargram/), XTAG Project* у Пенсильванському університеті *(http://www.cis.upenn.edu/~xtag/)*, роботу над *Combinatory Categorial Grammar* в Едінбургському університеті, *LinGO Project (http://lingo.stanford.edu), FrameNet Project (http://www.icsi.berkley.edu/~framenet/)* та ін. Усі ці проекти певною мірою залучені до лінгвістичного вивчення MWE [11].

3. Цілі статті

Мета цього дослідження – аналіз сучасних підходів до визначення багатослівних конструкцій та способів їх ідентифікації в тексті для подальшого видобування.

Отже, цілями статті є:

- ознайомлення з теорією MWE у сучасній комп'ютерній лінгвістиці;
- дослідження методів та підходів для ідентифікації MWE в текстах;
- визначення необхідних дій для практичної реалізації автоматизованого видобування MWE з текстів.

4. Основний матеріал

4.1. Які конструкції можна визначити як MWE?

Як було показано вище, не лише одне слово, а усі компоненти MWE стосуються одного концепту. Продовжуючи цю думку, можна подати таке визначення MWE:

«Послідовність слів або інших елементів, які є або здаються штучно створеними: тобто збережена і за потреби видобута з памяті єдність, а не предмет генерації або аналізу мовною граматикою» [12].

Таке визначення подає психолінгвістичний погляд на багатослівні конструкції, але не встановлює об'єктивного критерію для ідентифікації MWE. Таким критерієм повинна бути ознака, яку можна перевірити при спостереженні. Тому за основне визначення MWE візьмемо:

«Багатослівна конструкція, яка виходить за межі слова і є лексично, статистично, синтактично, семантично або прагматично ідіосинкратичною» [11].

Отже, існують такі рівні ідіосинкратичності (унікальності) у MWE:

1. Лексична. На цьому рівні колокації не є загальновживаними в мові, а є, ймовірно, запозиченими з інших мов та інституціалізованими внаслідок використання. Наприклад, *ad hoc, ad hominem, raison d'etre.*

2. Синтактична. У цьому випадку певні колокації можуть не слідувати правилам традиційної граматики, таким чином порушуючи спроби успішного синтаксичного аналізу і правильної інтерпретації. Наприклад, *by and large, wine and dine*. Ці приклади наведено на рис. 1.

3. Семантична. Одним з найважливіших класів MWE є ті сполуки, в яких семантика не є очевидною із зіставлення значень складових слів. Такі колокації є дуже поширеними в людській мові через їх метафоричне і фігуративне вживання. Обробка цього типу MWE є дуже важливою для обробки природної мови загалом. Для прикладу, *spill the beans, kick the bucket, run for office.*

4. Статистична. Деякі колокації явно не належать до вищезгаданих типів, тобто вони є ідеально скомпоновані і семантично, і синтаксично. Але такі колокації з'являються набагато частіше саме в такому компонуванні, ніж в будь-якому іншому. Наприклад, набагато більша ймовірність натрапити на вислів *traffic signal*, ніж *traffic lamp* або *traffic lights*, хоча усі вони означають одну і ту саму річ.

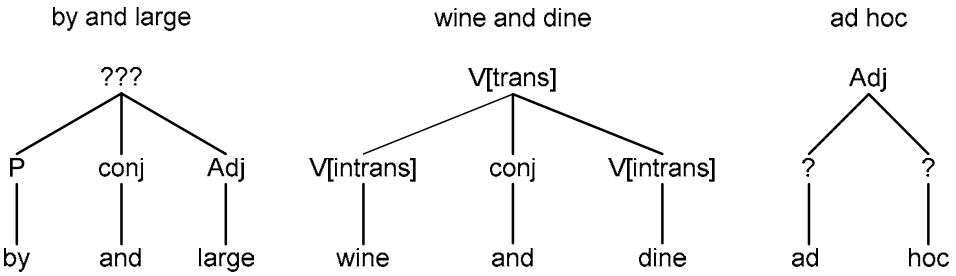

Рис. 1. Синтактична ідіосинкратичність [2].

Але чи можна точно стверджувати, що певна колокація є MWE? Наприклад, у випадку семантичної ідіоматичності колокація *let the cat out of the bag* є цілком не композиційною (значення не виявляється через складові), тоді як значення *lend a helping hand* можна частково визначити. У випадку *fall asleep* значення досить зрозуміле, але при буквальній інтерпретації воно означатиме *falling while sleeping*. Отже, слід пам'ятати, що насправді в багатьох експериментах навіть вчені не завжди згідні щодо ідіоматичності або неідіоматичності колокації.

Для успішного розпізнавання MWE необхідно знати їхні характеристики. А саме:

1. Інституціоналізація. MWE є виразом, який прийнятий до загального використання, а тому має статистичну значимість. Цей факт можна використовувати для ідентифікації потенційних MWE.

2. Здатність бути перефразованим. Оскільки MWE відповідає за один концепт, то є можливим перефразувати його за допомогою одного слова. Це одна з найважливіших характеристик. Наприклад, *leave out* означає *omit.*

3. Замінність. Внаслідок свого інституціоналізованого використання, MWE в більшості випадків протистоять замінюванню своїх складових подібними словами. Наприклад, *many thanks* не можна замінити на *several thanks* або *many gratitudes.*

4. Некомпозиційність, тобто прихованість значення. Наприклад, *blow hot and cold, spill the beans.*

5. Синтактична стійкість. Як і будь-яка мовна структура, MWE піддається змінам (для створення часової форми, плюралізації тощо). Наприклад, *traffic signal(s), promise(s/d/ing) (him/her/one) the moon.* Семантична композиційність впливає на кількість лексичних варіацій, яким можуть піддаватись колокації. Наприклад, *promising the moon* буде стійкою до будь-якого вставляння компонентів, але фраза *promise the pastry* дає змогу вставити нові слова, як *promise the chocolate pastry* [7].

4.2. Класифікація MWE

Загалом використовується така класифікація MWE відповідно до різних критеріїв: синтаксичні форми; семантична композиційність; синтаксична стійкість.

4.2.1. На основі синтактичних форм

MWE є загальним терміном для колокацій, які можуть набувати різних синтаксичних форм. Класифікація на основі синтаксичних форм необхідна для розвитку методів ідентифікації специфічних типів MWE. Ці типи є лінгвістично правильні і повинні бути основою для розроблення методів видобування MWE.

1. Складений іменник. Послідовність слів, що трактується як один іменник. Наприклад, *traffic signal,motor car, china tea cup.*

2. Фразове дієслово. Це колокації, які містять дієслово, за яким іде прийменник, що трактується як частка. Саме ця частка додає певного значення до колокації, модифікуючи значення самого дієслова. Наприклад, *broke up, gobble up.*

3. Нескладні дієслівні конструкції. Колокації, які складаються з дієслова та іменної частини, як *fall asleep* або *make a demo,* де семантика є не повністю композиційною. Іменник і надалі використовується в основному значенні.

4. Фразеологізми. Наприклад, *promise him the moon, blow hot and cold.*

5. Дієслівно-прийменникові фразові конструкції. У цьому випадку дієслово прикріплене до прийменникової фрази і має некомпозиційну семантику. Наприклад, *sweep under the carpet.*

4.2.2. На основі семантичної композиційності

MWE може бути цілком некомпозиційна або частково композиційна. В першому випадку семантика є повністю непрозора як *promise the moon.* У другому такі ідіоми, як *spill the beans*, є частково композиційні, оскільки spill використовується у значенні «виявити» і beans метафорично представляє «секрет».

4.2.3. На основі синтаксичної стійкості

MWE можуть бути синтаксично фіксованими, як у *by and large*, коли варіації взагалі не можливі. Деякі MWE дозволяють певну варіаційність флексій, але не дозволяють вставляння слів, як *promise one the moon.* Є й крайність, адже деякі MWE можуть піддаватись багатьом варіаціям, включаючи вставляння, як *keep an eye* може варіювати у *keep a sharp eye.*

За класифікацією, поданою у [11], MWE поділяються на лексикалізовані і інституціоналізовані фрази (адаптована термінологія згідно з Бауером (1983)). Лексикалізовані фрази мають щонайменше частково ідіосинкретичний синтаксис або семантику або містять слова, що не вживаються ізольовано. Далі вони можуть бути поділені на стійкі, напівстійкі і синтаксично вільні сполуки.

Інституціоналізовані фрази є синтаксично і семантично композиційними та трапляються в певному контексті з високою частотою.

1. Стійкі сполучення. В англійській мові існує великий клас незмінних сполучень, які не відповідають правилам традиційної граматики. До цього класу належать такі, як: *by and large, kingdom come, every which way.* Крім того, багато запозичених сполучень також належать до цього класу: *ad hoc (ad nauseum, ad libitum, ad hominem,…), Palo Alto (Los Altos, Alta Vista,…)*тощо. Такі колокації є повністю лексикалізованими і не піддаються ні морфосинтаксичним, ні внутрішнім варіаціям.

2. Напівстійкі сполучення. Ці вирази відповідають чітким правилам у порядку слів і композиції, але можуть піддаватись лексичним варіаціям певною мірою. Вони мають різноманітні форми, такі як нерозкладні ідіоми, певні складені іменники і власні назви. Розглянемо їх детальніше.

a. Нерозкладні ідіоми – це ідіоми, в яких значення не виявляється через сукупність значень складників. Наприклад, *kick the bucket.* У цьому випадку можна змінити закінчення відповідно до контексту, як у *he kickes the bucket.* Хоча зробити синтаксичну зміну не можна: *the bucket was kicked* (ідіоматичне значення втрачається).
b. Складені іменники – сполучення, які є синтаксично не змінними, але можуть змінюватись відповідно до граматичної категорії числа (*car park – 2 car parks*).
c. Власні назви – напівстійкі сполучення, оскільки також можуть набувати різних форм. Наприклад, назва американської спортивної команди – *the San Francisco 49ers* може використовуватись в мові, як *the 49ers* або як *a 49ers player.*

3. Синтаксично вільні сполучення, які можуть мати дуже велику кількість різноманітних конструкцій. В тексті вони виявляються як:

d. Розкладні ідіоми. Це ідіоми, які є синтаксично вільними певною мірою, тому при їх пошуку дуже складно розраховувати лише на синтаксичний аналіз. Наприклад, let the cat out of the bag, sweep under the rug.
e. Конструкції дієслово-частка. Це можуть бути семантично ідіосинкратичні вирази – brush up, або композиційні - break up (the meteorite broke up in the earth’s atmosphere).
f. Нескладні дієслівні конструкції – конструкції, в яких складно передбачити, з яким іменником буде комбінуватися дієслово. Хоч вони і є ідіосинкратичні, їх потрібно відрізяти від ідіом. У цій конструкції іменник вживається в нормальному сенсі, а значення дієслова не ідіоматичне (make a mistake, give a demo, *do a mistake, *make a demo).

4. Інституціалізовані фрази – це вирази, які стали загальновживаними (*salt and pepper, traffic light, to kindle excitement*). Вони є семантично і синтаксично композиційні, але статистично ідіосинкратичні.

4.3. Видобування MWE

У попередній частині статті було наведено характеристики і лінгвістичні риси MWE. У цій частині розглянемо різноманітні методи, запропоновані в науковій літературі для автоматизованого видобування багатослівних конструкцій з корпусів текстів. Метою насамперед є створення лексикону цих сполук, для чого буде створено програму, за допомогою якої лексикографи зможуть досягти цієї мети. Видобування є першою сходинкою для залучення MWE до мовного аналізу, а компільований словник стане корисним для граматики мови загалом.

Отже, будь-який метод вилучення MWE повинен виконувати такі завдання:

1. Знаходити потенційних кандидатів. Метод повинен надавати міру статистичної значимості колокації в корпусі як доказ її інституціоналізації.

2. Встановлювати лінгвістичну правильність кандидатів. Не усі колокації, які мають високу частоту використання, є багатослівними конструкціями. Наприклад, *the…of* є дуже частим сполученням, хоча воно й не має лінгвістичного сенсу.

3. Вимірювати семантичну некомпозиційність кандидатів. Семантична некомпозиційність є ключовою характеристикою більшості багатослівних конструкцій. Дуже важливою є здатність вимірювати некомпозиційність і розробляти об'єктивні критерії для ідентифікації виразу як MWE.

Перший етап – це знаходження потенційних кандидатів. Для цього існують методи, які ґрунтуються переважно на знаходженні статистичної значимості колокацій в корпусі, при чому фільтрування потенційних кандидатів повинне відбуватися на великих корпусах і бути повністю комп'ютеризованим.

1.1. Спільна точкова інформація

Спільна інформація – це та інформація, яка поділяється двома випадковими змінними. Слова в колокації вважаються цими двома змінними, і мірою їхньої близькості є точкова спільна інформація [3].

$$I(x,y) = \log_2 \frac{P(x,y)}{\overline{p}(x)\overline{p}(y)}, \tag{1}$$

де (x,y) – це пара, яку тестують; $I(x,y)$ – спільна точкова інформація між ними.

Спільна точкова інформація є достатньою для підрахунків, хоча її недоліком є висока переоцінка появи рідкісних подій. Але все ж таки цей метод може слугувати добрим початковим фільтром для ідентифікації потенційних багатослівних конструкцій.

1.2. Тест Хі-квадрата Пірсона

Тест Пірсона можна використовувати для того, щоб перевірити, чи слова в колокації є незалежними одне від одного. У цьому випадку нульова гіпотеза означає, що слова не залежать одне від одного. За допомогою частотного розподілу корпусу можна побудувати таблицю (табл. 1) ймовірностей для двох слів (w1 і w2).

Таблиця 1

w1w2	w1 ~ w2
~ w1w2	~ w1 ~ w2

~ – позначає відсутність слова, отже w1 ~ w2 – це частота колокації, яка починається з першого слова, але за ним не слідує друге слово.

Потім необхідно підрахувати статистику хі-квадрата:

$$x^2 = \sum_{i,j} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}, \tag{2}$$

де $O_{i,j}$ – частота з таблиці; $E_{i,j}$ – очікувана частота в кожній клітинці, коли w1 і w2 випадково з'являються разом. Очікувана частота для кожної клітинки дорівнює:

$$\frac{\text{сума за рядками*сума за стовпчиками}}{\text{загальна сума}}$$

Чим вища вартість статистики хі-квадрата, тим більший зв'язок між словами. Можна також встановити зріз згідно з бажаним рівнем значимості. Однак у випадку низьких частот, хі-квадрат не дає задовільних результатів, а отже, для його використання необхідні великі корпуси.

1.3. Коефіцієнт ймовірності

На відміну від попереднього алгоритму, цей тест працює краще на менших корпусах з різними дистрибутивними характеристиками. Цей тест надає коефіцієнт ймовірності, протиставляючи певний підпростір цілому простору параметрів. Вираховується за формулою:

$$\lambda = \frac{\max_{\omega \in \Omega_0} H(\omega; k)}{\max_{\omega \in \Omega} H(\omega; k)}, \quad (3)$$

де ω – точка в просторі параметрів Ω; k – точка в просторі спостережень K.

Кількість $2\log\lambda$ використовується як вимір для тесту, в той час як хі-квадрат асимптотично розподілений зі ступенем свободи, що дорівнює різниці в вимірах між Ω і Ω_0.

Цей тест можна застосувати для вирішення проблеми пошуку статистично значимих колокацій [5]. Для цього формується корпус з частотами слова в колокації як біномний розподіл. Для колокації w1w2, нульовою гіпотезою є : *P(w2|w1) = P(w2|~w1)*. Отже, ймовірність параметрів є:

$$H(p1, p2; k1, n1, k2, n2) = p1^{k1}(1-p1)^{n1-k1} p2^{k2}(1-p2)^{n2-k2}, \quad (4)$$

де *p1 = P(w2|w1), p2=P(w2|~w1), n=c1, k1=c12, n2=n-c1, k2=c2-c12;*

c1, c2 = частоти w1, w2 відповідно;

c12 = частота w1w2;

n= загальна кількість слів у корпусі.

В основу розглянутих тестів покладено пошук статистичної значимості. Для роботи з усіма статистичними методами необхідні великі корпуси, що і є основним недоліком таких тестів.

Другим важливим завданням для видобування багатослівних конструкцій є встановлення лінгвістичної правильності кандидатів. А саме: як нам вибрати колокації, які мають лінгвістичний сенс (як наприклад, складений іменник), уникаючи беззмістовних *the…of* ? Існують два загальновикористовувані методи: використання морфологічної розмітки (POS-тегів) і фільтрування неправильних колокацій, враховуючи рівень залежності. Ці методи не працюють для синтаксично ідіосинкратичних конструкцій, але в цьому випадку ми сконцентруємося на семантичній ідіосинкратичності.

2.1. Фільтри частин мови

Колокації, які містять певні комбінації тегів, можуть мати лінгвістичну цінність. У табл. 2 наведені деякі послідовності тегів, які можна визначити як потенційні багатослівні конструкції.

Таблиця 2

Послідовність тегів	Можлива багатослівна конструкція	Приклад
Noun-Noun	*Compound-Noun*	*motor car*
Adjective-Noun	*Compound-Noun*	*green card*
Verb-Noun	*Verbal Idioms*	*spill beans*
	Light Verbs	*fall asleep*
Verb-Preposition	*Phrasal Verbs*	*catch up*

2.2 Фільтри зв'язків залежності

Зв'язок залежності є більш структурованим методом для ідентифікації кандидатів. Аналізуючи, ми можемо визначити залежність між лексичними одиницями. Потрійна залежність вкаже на потенційну багатослівну конструкцію. Наприклад, імовірніше, що дієслово буде зв'язане з прямим додатком, аніж з будь-яким іншим іменником, а отже, можна вибирати дієслово і його прямий додаток. Можна говорити і про головний іменник та його модифікатори, що разом можуть складати складений іменник. Тобто, цей метод є доволі ефективним підходом, хоча під час аналізу також можуть траплятися помилки.

Третім етапом є семантична некомпозиційність, що є однією з найголовніших рис MWE. Сьогодні існує дуже велика кількість досліджень і методів, які вимірюють саме цю характеристику колокацій. Розглянемо найголовніші з них.

3.1. Заміна схожим складовим словом.

Якщо багатослівна конструкція не розкладається семантично, то заміна її складової подібним словом призведе до утворення нового виразу з іншими дистрибуційними характеристиками. Вимірюючи різницю між дистрибуційними характеристиками двох колокацій, можна отримати міру семантичної некомпозиційності виразу. Наприклад, спільну точкову інформацію між оригінальним виразом і заміненим можна використовувати як вимір семантичної некомпозиційності. Подібні слова для заміни одержують з автоматично згенерованого тезауруса, що анотований мірою подібністі між парою слів. Слово для заміни – це слово, що є найбільш подібне до слова кандидата в тезаурусі [4].

Цей метод вимагає аналізу для визначення залежності між словами, в якому теж можуть траплятися помилки. Використання корпусів, що автоматично навчаються, робить цей метод придатним і для нових доменів, але потребує великого корпусу для великого різноманіття побудованого тезауруса.

3.2. Латентне семантичне індексування

Певна колокація не обов'язково має ідіоматичний сенс. Іноді трапляється, що вона вживається в буквальному сенсі. Наприклад, *take someone for a ride* може мати обидва сенси. Тобто для вирішення чи колокація є ідіоматичною, важливий також і контекст. Контекст можна представити у вигляді сукупності слів з частотними характеристиками. Для колокацій і для їх складових будуються вектори. Ідея полягає в тому, що вектори, які представляють колокації, будуть відрізнятись від сукупності векторів, що показують складові. Подібність між векторами можна вирахувати за допомогою косинуса подібності. Найлегший шлях створення складових векторів – це додатковий вектор. Запропонований метод може вилучати можливі багатослівні конструкції в корпусі без звернення до певних прикладів колокацій [1, 8].

3.3. Використання мультилінгвального вирівнювання слів

Значення ідіоматичного виразу не може складатись із значень його складових слів. Ця проблема стає дуже суттєвою при перекладі з однієї мови на іншу. Послівний переклад, на відміну від композиційних фраз, де це більш імовірно, в цьому випадку не спрацює. Цю характеристику можна використати за методом вирівнювання слова для виміру семантичної некомпозиційності [10]. Ідея полягає в тому, що ідіоматична фраза імовірно матиме більше перекладів, ніж буквальна. А отже, некомпозиційність приховує в собі непевність під час перекладу, і цю непевність можна підрахувати як перекладну ентропію такого виразу:

$$H(T \mid s) = -\sum_{t \in T_S} P(t \mid s) \log P(t \mid s). \quad (5)$$

Можливі переклади слова можна отримати, вирівнюючи паралельні корпуси, використовуючи застосунки для вирівнювання слів (наприклад, GIZA++).

3.4. Перекриття частки для фразових дієслів

Цей метод можна застосовувати для фразових дієслів. Частка в конструкціях дієслово-частка дуже сильно впливає на семантику фрази (*climb up*). Однак у фразових дієсловах вона вживається більше для наголошення, ніж буквального значення (*speak up*). Цей факт може бути використаний під час заміни дієслова схожим дієсловом. Наприклад, заміна *climb* близькими дієсловами утворює *walk up, run up, limp up, crawl up,* які є правдоподібними. Але подібна заміна для *speak - talk up, chatter up* – дає інший результат, бо утворені конструкції позбавлені сенсу і малоймовірно, що будуть знайдені в корпусі. Отже, можна провести тест, який вимірює кількість близьких конструкцій дієслово-частка, які можуть бути вибрані з автоматично згенерованого тезауруса для даної конструкції дієслово-частка. Більша кількість дієслів з однією часткою вказує на вищу композиційність [9].

Висновки

Багатослівні конструкції, або MWE, є загальним поняттям для багатьох лінгвістичних феноменів. Саме ці колокації за останні два десятиліття стали важливим і зростаючим питанням у сфері комп'ютерної лінгвістики та опрацювання природної мови. Термін MWE використовується стосовно різних типів лінгвістичних одиниць і виразів, що містять ідіоми, складені іменники, фразові дієслова та інші звичні для використання колокації.

Сьогодні некомпозиційні MWE є викликом для автоматизованого аналізу, оскільки семантику цих виразів не можна виявити за їхніми складовими. Такі MWE і є осередком виникнення проблем. Узагальнюючи вищевикладений аналіз методів опрацювання MWE, можемо зазначити, що для забезпечення ефективного опрацювання некомпозитних MWE потрібно дотримуватись такого алгоритму: знайти потенційні MWE за допомогою статистичного аналізу, виокремити з потенційних MWE багатослівні конструкції, а тоді виміряти семантичну некомпозиційність MWE для визначення подальших дій.

1. Baldwin T. An empirical model of multiword expressions decomposability. In Proc. of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment / T. Baldwin, C. Bannard, T. Tanaka, D. Widdow. – 2003. 2. Baldwin T. Multiword Expressions (Presentation) / T. Baldwin. – Available from: www.csse.unimelb.edu.au/~tim/pubs/altss2004.pdf. 3. Church K. Word association norms, mutual information, and lexicography. Computational Linguistics / K. Church, P. Hanks. – 1990. 4. Dekang L. Automatic identification of non-compositional phrases. Proceedings of ACL-99 / L. Dekang. – 1999. 5. Dunning T. Accurate methods for the statistics of surprise and coincidence. Computational Linguistics / T. Dunning. – 1993. 6. Fellbaum C. WordNet: An Electronic Lexical Database / C. Fellbaum. – MIT Press, 1998. 7. Jurafsky D. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition / D. Jurafsky, J. H. Martin. – Upper Saddle River, NJ: Prentice Hall, 2008. – 988 p. – 2nd edition. 8.Katz G. Automatic identification of non-compositional multi-word expressions using latent semantic analysis. Proc. of the ACL-2006 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties / G. Katz, E. Giesbrechts. – 2006. 9. McCarthy D. Detecting a continuum of compositionality in phrasal verbs. Proc. of the ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment / D. McCarthy, B. Keller, J. Carroll. – 2003. 10. Moiron B. V. Identifying idiomatic expressions using automatic word alignment. Proceedings of the EACL 2006 Workshop on Multiword Expressions in a multilingual context / B. V. Moiron, J. Tiedemann. – 2006. 11. Sag I. Multiword expressions: A pain in the neck for nlp. Proceedings of CICLing / I. Sag, T. Baldwin, F. Bond, A. Copestake, D. Flickinger. – 2002. 12. Wray A. Formulaic Language and the Lexicon / A. Wray. – Cambridge University Press, 2002.