УДК 81.32

Бобкова Т. В.,
канд. філол. наук, доцент, Київський національний лінгвістичний університет, м. Київ, Україна

ОСНОВНІ ПІДХОДИ ДО ІДЕНТИФІКАЦІЇ Й ВИЛУЧЕННЯ КОЛОКАЦІЙ ІЗ ТЕКСТІВ

Ідентифікація й екстракція колокацій є частиною загальної проблеми автоматичного розпізнавання й опрацювання природно-мовного тексту й пошуку інформації. Існуючі методики та прийоми ідентифікації колокацій ґрунтуються на лінгвістичному, статистичному і комбінованому підходах, які диференціюються за критеріями ідентифікації і послідовністю застосовуваних процедур. Статистичний підхід передбачає використання корпусних методів та інтерпретацію колокації як статистично значущої одиниці. У роботі обґрунтовується необхідність використання комбінованого підходу до вилучення колокацій із текстів. Розроблено методику ідентифікації колокацій, яка на основі статистичного опрацювання корпусу та лематизації дозволяє автоматично вилучати двослівні сполуки з українських текстів. Результати екстракції потребують обов'язкового постредагування в аспекті зняття лексико-граматичної омонімії і відбору граматично коректних колокацій. Застосування більшого корпусу та лінгвістичних фільтрів забезпечить підвищення ефективності результатів.

***Ключові слова**: колокація; текст; вилучення; ідентифікація; лінгвістичний підхід; статистичний підхід; комбінований підхід.*

Зростання інтересу до вивчення колокацій пояснюється затребуваністю цієї інформації для створення пошукових систем, машинного перекладу, укладання словників, тезаурусів, побудови баз даних, семантичних мереж та розв'язання інших прикладних завдань. Ідентифікація та екстракція колокацій є частиною загальної проблеми автоматичного опрацювання й розпізнавання тексту природною мовою та пошуку інформації. Необхідність розв'язання нагальних завдань щодо ідентифікації колокацій у тексті призвела до появи на початку другого десятиріччя XXI ст. низки досліджень, присвячених проблемам опису й фразеологізації стійких сполук [5; 8], зокрема, й у галузі інформаційних технологій і автоматичного пошуку [6; 9; 11; 12].

Науковий інтерес з боку фахівців різних галузей сприяв формуванню різних способів і підходів до вилучення й опису колокацій. Наразі, за даними дослідницької літератури з прикладної лінгвістики [18, p. 153; 19, p. 252], пропонується три способи ідентифікації і формування реєстру колокацій: у ручному режимі, (напів)автоматичному – через вилучення відповідних одиниць з автентичних текстів і змішаному. Кожен з окреслених вище способів відбору спирається на відповідні підходи до ідентифікації колокацій, серед яких більшість фахівців [8; 10; 13; 14; 16; 18; 19; 20] виділяють лінгвістичний, нелінгвістичний і комбінований. Зазначені підходи диференціюються в термінах і послідовності застосовуваних методів і процедур дослідження [14, p. 1034; 20, p. 53–54].

Аналіз фахової літератури показує, що основні проблеми опису полягають у встановленні критеріїв ідентифікації колокацій [7, с. 304; 8, с. 158; 16, p. 509; 18, p. 152; 19, p. 258], класифікації їх кількісних і якісних ознак [16, p. 511; 19, p. 258], презентації отриманої інформації як частини словника [16, p. 509] й оцінці ефективності використаних прийомів [7, с. 303; 11, с. 46; 18, p. 149]. Отже, постає необхідність удосконалення методики розпізнавання колокацій у природномовному тексті на підставі об'єктивних критеріїв. **Мета** статті – дослідити основні підходи й критерії ідентифікації і вилучення колокацій із природномовних текстів. Досягнення поставленої мети передбачає виконання таких **завдань:** 1) здійснити аналіз окреслених у дослідницькій літературі підходів і критеріїв ідентифікації колокацій; 2) навести аргументи на користь комбінованого підходу й запропонувати відповідну методику вилучення колокацій з корпусу українських текстів.

Проаналізуємо окреслені вище способи й підходи з погляду пропонованих критеріїв ідентифікації колокацій. Аналіз троякої природи колокацій вимагає врахування комплексу лексичних, граматичних і статистичних ознак, що не сприяє встановленню єдиного критерію їх відбору. Суто ручному способу відповідають традиційні докорпусні й корпусно-інформативні дослідження колокацій на засадах **лінгвістичного підходу**, в межах якого залежно від інтерпретації зв'язності виокремлюють лексико-семантичний, лексико-функціональний, лексико-граматичний, синтаксичний і контекстно-орієнтований [3, с. 16–18]. За лінгвістичним підходом здійснюється формування реєстру на підставі лексикографічних даних наявних словників [18, p. 153] і/або корпусів текстів, використовуваних у ролі ілюстративного матеріалу для підтвердження концепції укладача. Основними **критеріями** ідентифікації колокацій при цьому виступають лінгвістичний та інтуїтивний (див. табл. 1).

Перевагою ручного способу є відсутність будь-яких обмежень за **формальними** (кількісними) і **якісними** ознаками колокацій, оскільки основним критерієм включення до реєстру є знання, досвід і дослідницька

інтуїція укладача. Зокрема, ручний спосіб дозволяє ідентифікувати багатослівні та дистантні колокації, можливості виявлення яких автоматичним способом значно обмежені [10, с. 18, 24; 16, p. 511; 19, p. 252]. Формування списку колокацій вручну й без урахування формальних критеріїв щодо кількості та позиційного розташування складників, сприяє, насамперед, вивченню **семантичних** ознак за лексико-семантичним і лексико-функціональним підходом. Так, на відміну від автоматичного, реєстр, укладений лінгвістом-експертом, переважає за кількістю семантично правильних, релевантних для певної предметної галузі колокацій [18, p. 152, 156]. Також ручний режим формування списку колокацій не передбачає обмежень щодо певних лінгвістичних – лексико-граматичних і синтаксичних моделей.

Хоча при контекстно-орієнтованому підході такі обмеження зумовлюються власне досліджуваним матеріалом. Так, з чисельної групи словосполучень, виявлених у науковому тексті «Лексическая семантика. Синонимические средства языка» (Ю. Д. Апресян), було відібрано для перекладу польською мовою близько 200 колокацій [4, с. 225]. Класифікацію колокацій обмежено вилученими з тексту оригіналу дієслівно-іменними, атрибутивними й адвербіальними словосполученнями й здійснено шляхом зіставлення лексичних ресурсів російської і польської мов. На думку Е. Бялек, основну проблему дослідження колокацій становить визначення відповідності відібраних одиниць встановленим автором критеріям через «нечіткість меж у фразеології» і специфіку сполучуваності слів [4, с. 225].

Таблиця 1

Підходи до ідентифікації і вилучення колокацій

Підхід		Методика		Обмеження			Критерії			Спосіб		
		(напів)закритий список	відкритий список	якісні	кількісні	формальні	лінгвістичні	статистичні	інтуїтивні	ручний	(напів) автоматичний	змішаний
Лінгвістичний	Лексико-семантичний	+	–	+	–	–	+	–	+	+	–	–
	Лексико-функціональний	+	–	+	+	+	+	–	+	+	–	+
	Лексико-граматичний	+	–	+	+	+	+	–	–	+	+	+
	Синтаксичний	+	–	+	+	+	+	–	–	+	–	+
	Контекстно-орієнтований	+	–	+	+	+	+	–	+	+	–	+
Нелінгвістичний (статистичний)		–	+	–	+	+	–	+	–	–	+	–
Корпусно-орієнтований		–	+	+	+	+	+	+	–	–	+	–
Комбінований		+	+	+	+	+	+	+	+	+	+	+

Проте через властивості людської пам'яті віддавати перевагу незвичним випадкам перед типовими [15, p. 3], базовані на інтуїції лінгвіста-дослідника висновки можуть бути ненадійними. Отже, існує велика ймовірність проігнорувати частотні «характерні схеми» вживання мови й «контекстові чинники», що впливають на варіативність [15, p. 3]. Зокрема, про це свідчить збільшення в автоматичному списку кількості найбільш частотних – двослівних сполук у середньому на 9,4–13,5 % [18, p. 152]. Крім того, якщо, в укладеному експертом реєстрі подано виключно вихідні, словникові форми колокацій, то в автоматично створеному – ті ж колокації представлені різними граматичними формами, зокрема множиною або в прийменникових конструкціях: *adopted by Member State – adopted by Member States* [18, p. 153]. Це суттєво для ідентифікації у таких випадках, коли граматичні форми диференціюють різні значення колокацій: *цінні папери – на папері, кредитні ризики – власний ризик*.

Отже, до загальних недоліків ручного способу слід віднести неповноту сформованого вручну реєстру и [19, p. 252] з погляду відображення **граматичних** (морфологічних) і **функціональних ознак** (частоти вживання) колокацій. Загалом, укладання реєстру колокацій вручну вважається доволі тривалим і витратним процесом, який вимагає об'єднання зусиль багатьох кваліфікованих фахівців [11, с. 41; 18, p. 156; 19, p. 252]. Саме тому навіть часткова автоматизація формування списку колокацій становить практичний інтерес для лексикографів [11, с. 41].

Під автоматичною екстракцією [18, p. 150] колокацій розуміється операція, на вході якої подаються опрацьовувані тексти, а на виході – укладений список потенційних колокацій. Таким чином, автоматична ідентифікація колокацій передбачає наявність репрезентативної колекції текстів з лінгвістично-пошуковим апаратом [14, p. 1034; 18, p. 156] і застосування методики (напів)закритого або відкритого списку (див. табл. 1). Терміни «закритий» / «напівзакритий» / «відкритий список» традиційно використовуються в дослідженнях з розпізнавання й сприйняття текстової інформації людиною або автоматичною системою. Методика (напів)закритого списку застосовується в основному в традиційних лінгвістичних дослідженнях колокацій і відкритого – в корпусно-керованих.

У загальному розумінні методика «закритого» або «напівзакритого списків» ґрунтується на лінгвістичному підході до ідентифікації колокацій. Методика закритого списку передбачає задавання й редагування переліку неоднослівних єдностей, заздалегідь встановлених за словником [1, с. 120; 13]. Тобто здійснення пошуку, ідентифікація та екстракція обмежуються наявними в закритому списку колокаціями. Зазначена методика ґрунтується на приналежності колокацій до певного

домену [16, p. 510] і, як правило, використовується для дослідження предметних галузей. Зокрема, на підставі закритого списку реалізовано пошук контекстів вживання термінів і терміносполучень у корпусі текстів з комп'ютерної лінгвістики [21] для тримовного тлумачного словника (див. рис. 1).

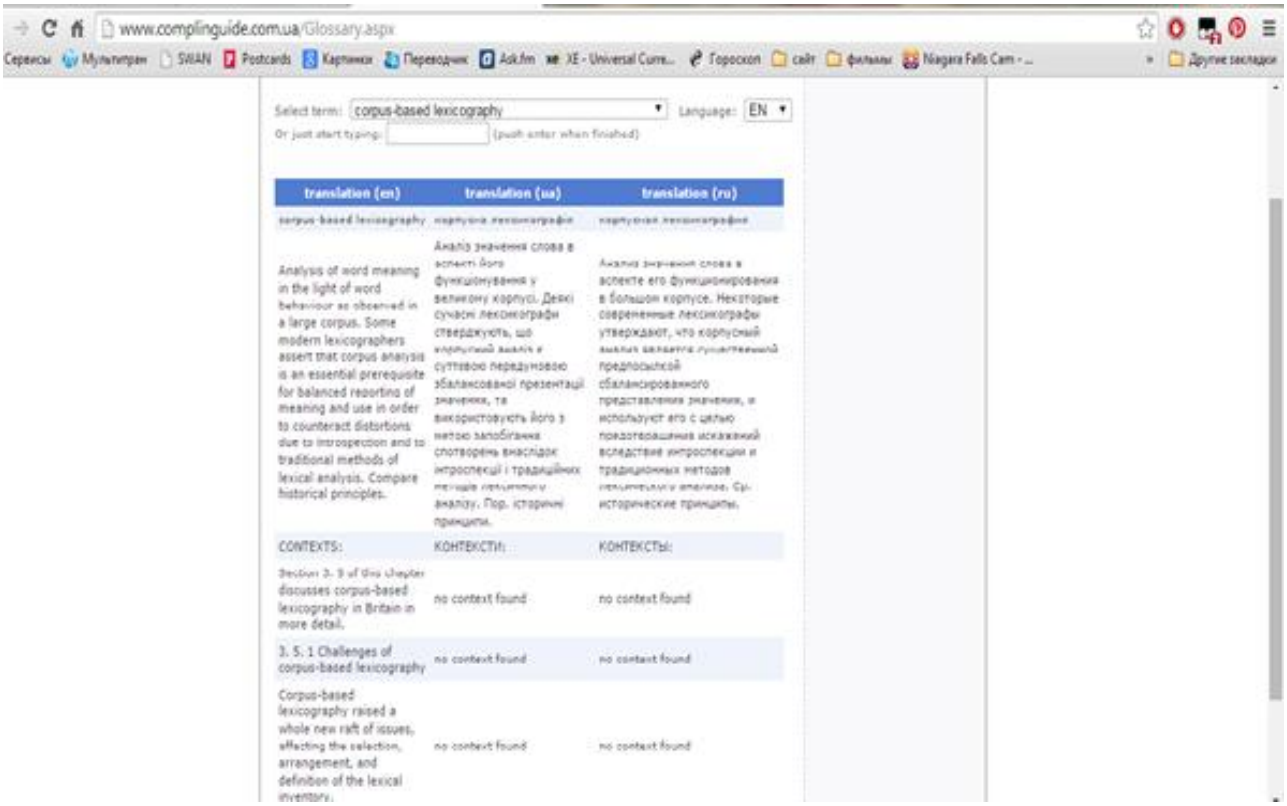

Рис. 1. Результати екстракції терміносполучення **corpus-based lexicography**

На подібних засадах укладено електронний словник фразеологізмів для системи автоматичного граматичного аналізу тексту (АГАТ) [5, с. 119–120]. Еквівалентність слову зумовлює включення колокацій до реєстру стійких сполук поруч з безумовними ідіомами: *брати участь, діаметрально протилежний* [5, с. 120]. Наразі спостерігається тенденція до фразеологізації вільних у загальновживаній мові сполук і утворення ними колокації у певній підмові: *create, delete, save + file* [16, p. 510]. Саме тому формування реєстру колокацій, що не належать до фахових баз термінів, вважається більш складним з погляду методики: *approve draft terms, enter into force* та ін. [18, p. 152].

Укладання більш складного в методичному відношенні напівзакритого списку зумовлюється попереднім встановленням обмеження за **формальними** (кількісними) і/або **якісними ознаками**: функціональними моделями в межах лексико-функціонального підходу теорії «Смисл↔Текст» [10, с. 9–10; 16, p. 511], лексичними варіантами [13], синтаксичними конструкціями [1, с. 22; 12, с. 33; 13] і морфолого-синтаксичними моделями [9, с. 147; 10, с. 3]. Зокрема, на засадах лексико-граматичного й синтаксичного підходів побудовано ідентифікацію субстантивно-ад'єктивних і субстантивно-субстантивних колокацій в українських «кримінально значимих текстах»: *вогнепальна зброя, умисний підпал, заволодіння майном* [9, с. 147] і скінчений автомат – лексикографічну систему для екстракції субстантивних, ад'єктивних і дієслівних колокацій із законодавчих текстів [12, с. 32–33]. У такий спосіб, ідентифікація на базі (напів)закритого списку передбачає певні умови щодо **якісних і кількісних ознак** колокацій. Свідомо встановлювані укладачами обмеження є підставою класифікувати ручне укладання й методику (напів)закритого списків ідентифікації колокацій як *інтуїтивні,* тобто такі, що відображають традиції певної школи, досвід і інтуїцію дослідника або обмежену тему дослідження [13].

На противагу інтерпретації, базованій на знаннях та інтуїції дослідника, автоматичне вилучення колокацій за **нелінгвістичним підходом** ґрунтується на статистичних (математичних) критеріях ідентифікації. Дослідження автентичних текстів різних функціональних стилів і предметних галузей здійснюється за відсутності заздалегідь встановленого реєстру колокацій і за умови повного опрацювання матеріалу. Зазначений підхід не передбачає жодних встановлених обмежень щодо якісних і формальних ознак і реалізується через статистичну інтерпретацію текстових даних, отриманих у вигляді відкритого списку [13]. Таким чином, критерії граматичної сполучуваності та лінгвістичної стійкості замінюються **статистичними критеріями**, вимірюваними відповідними параметрами, зокрема, частотою [11, с. 42] або встановленням відповідного статистичного порогу вживання колокації [16, p. 509]. Результатом статистичного аналізу даних текстів є автоматично укладений список кандидатів-колокацій [18, p. 150]. Сказане вище дозволяє класифікувати методику відкритого списку ідентифікації колокацій як статистичну.

На сучасному етапі нелінгвістичний (статистичний) підхід до відбору колокацій знаходить все більше застосування в розробці прикладних інформаційно-пошукових систем [6, с. 183], автоматизованого робочого місця лексикографа [11, с. 42] і машинного перекладу [8, с. 164]. Так, укладання реєстру для криміналістичної інформаційної системи здійснено на підставі «вектору статистичних даних про частоту» термінів-колокацій (*злочинна дія, злочинна бездіяльність суб'єкта*) в колекціях текстових корпусів [6, с. 184–186]. Підвищенню ефективності опрацювання текстів сприяє доповнення статистичного підходу більш жорстким критерієм «комбінаторної варіативності» [11, с. 42], тобто здійснення відбору колокацій на підставі даних аналізу вилучення, заміни або додавання слова до сполучення.

Статистичні критерії ідентифікації становлять переваги нелінгвістичного підходу в вивченні варіативності **граматичних форм** [18, p. 152] і **функціональних ознак** колокацій, зокрема, автоматичний список забезпечує більше покриття тексту порівняно з укладеними вручну [14, p. 2035]. Так, зіставлення отриманого в результаті статистичного опрацювання англо-хорватських текстів автоматичного, фільтрованого за лінгвістичними моделями, й укладеного вручну виявляє збіг списків колокацій всього на 20 % [18, p. 152]. Отже, статистична інтерпретація колокації і не обмежена інтуїцією дослідника методика відкритого списку уможливлюють виявлення відсутніх у наявних словниках усталених сполучень. Однак проблема полягає в тому, що жодна з розроблених на сьогодні систем не вилучає весь діапазон колокацій [19, p. 252]. Крім того, в автоматично створеному списку відсоток помилок ідентифікації значно вище: не завжди вилучені колокації є семантично правильними, більше половини двослівних сполучень становлять «фальшиві примітиви» й повторення, які потребують постредагування [18, p. 152]. Проте, незважаючи на вищесказане, на сьогодні процес автоматичного укладення списків колокацій залишається менш витратним порівняно з ручним способом.

Певним різновидом статистичного слід вважати **корпусно-орієнтований підхід**. У цьому аспекті еволюція статистичного підходу визначається переходом від частотного критерію ідентифікації колокації, зумовленого обсягом корпусу [16, p. 509, 512], до незалежного кореляційного критерію на базі статистичних мір (MI, t-score, log-likelihood, salience) [6, с. 184–186; 10, с. 4–5] і відповідних алгоритмів [11, с. 42]. Застосування кореляційного критерію відкриває можливості для вилучення з тексту розірваних колокацій, а побудова алгоритмів на базі статистичних параметрів – для виявлення багатослівних сполук. Укладені для певної підмови алгоритми можуть бути застосовувані для опису різних предметних галузей [11, с. 42; 16, p. 507] і текстів різних функціональних стилів. Так, за результатами тестування алгоритму на базі комбінаторної варіативності виявлено, що спільними для художнього й наукового текстів російською мовою є лише найчастотніші – двослівні усталені сполучення: *тот самый, при том, который был* та ін. [11, с. 45].

Зняття обмежень щодо якісних і кількісних ознак забезпечує переваги корпусного підходу в ідентифікації і вилученні найчастотніших колокацій, не встановлених заздалегідь синтаксичних конструкцій [18, p. 151], а також граматично некоректних стійких сполук. Так, зафіксовані в словниках колокації, як правило, зустрічаються в корпусі, натомість автоматично ідентифіковані навіть у невеликій колекції текстів сполуки можуть бути відсутні в наявних словниках через неповноту останніх [7, с. 303; 11, с. 46]. Застосування корпусного статистично-пошукового апарату забезпечує об'єктивність оцінювання відносного значення й рангу колокації в автоматично укладеному реєстрі [7, с. 303]. Проте використовувані статистичні міри не є всеосяжними для опису **формальних ознак**, оскільки розроблені тільки для двослівних сполучень [10, с. 12], для **семантичних характеристик** – не здатні диференціювати ядро й колокат (а місце в реєстрі словника визначається саме ядром), і являються недостатньо чуттєвими до **синтаксичних ознак** колокацій. Крім того, за статистичними мірами неможливо визначити приналежність колокації до певного домену, особливо якщо сполука є частотною, але не належить до професійної терміносистеми: *approve draft terms, enter into force, formed in accordance with* [18, p. 152]. Отже, трояка природа колокацій не дозволяє повністю відмовитись від «символічних методів» [8, с. 158] опрацювання природно-мовного тексту.

Загалом на сучасному етапі впровадження корпусно-орієнтованого підходу до ідентифікації колокацій обмежується через недосконалість лінгвістичних процесорів, яка вимагає обов'язкового постредагування. Зокрема, статистично-пошуковий апарат ані наявного у вільному доступі Корпусу української мови, ані обмеженого в доступі Українського Національного Лінгвістичного Корпусу на сьогодні не уможливлює вилучення колокацій [17]. Тому постає необхідність впровадження **комбінованого підходу**, який поєднає об'єктивний опис статистичних результатів і даних про лінгвістичні ознаки колокацій [8; 18; 19; 20]. Окремі спроби ідентифікації і вилучення колокацій на підставі лінгвістичних і статистичних характеристик базуються на застосуванні методики напівзакритого списку морфолого-синтаксичних моделей і побудови відповідних алгоритмів [8; 9]. Так, для опрацювання «криміналістично-значимих текстів» [9, с. 148] пропонується двоетапний метод, базований на використанні логіко-лінгвістичної моделі виділення іменних сполук і ймовірнісної моделі ідентифікації колокацій.

Уперше принципово новий підхід автоматичного вилучення всього діапазону колокацій реалізовано Ф. Смаджа в системі Xtract [19]. Новизна підходу Ф. Смаджа полягає у порушенні «канонічного порядку аналізу» усталених сполучень [20, p. 54]: уперше вихідними даними для ідентифікації колокацій стають не лінгвістичні ознаки, а статистичні характеристики. Програма екстракції колокацій (компілятор спільної появи) включає два компоненти: конкорданс опрацювання корпусу біржових звітів і їх статистичний аналіз [19, p. 253]. Для подальшого опрацювання лінгвістичними фільтрами зберігаються тільки статистично значущі пари слів. Крім зазначеного вище, впроваджений Ф. Смаджа підхід до ідентифікації колокацій суттєво відрізняється в деталях від попередніх власне поетапним

застосуванням статистичного аналізу, трьох лінгвістичних фільтрів (позиційного, синтаксичного, морфологічного) та оцінюванням точності експертом-лексикографом.

Подібний підхід до вилучення колокацій використано для автоматизованого укладання бази термінів на матеріалі паралельного корпусу англо-хорватських юридичних текстів [18]. Процес екстракції колокацій включає етапи автоматичного вилучення та ідентифікації засобами програм MultiTerm Extract, Lexterm і розпізнавання терміносполучень, включаючи верифікацію за заздалегідь укладеним експертом списком [18, p. 150–151]. Фінальний список колокацій опрацьовується за лінгвістичними фільтрами-моделями, аналізується з погляду синтаксису й семантики двома незалежними фахівцями в галузі перекладу та комп'ютерної лінгвістики.

В українській лінгвістиці комбінований підхід до вилучення й опису колокацій планується застосувати для укладання лексикону багатослівних сполук [8, с. 165]. Статистичні параметри потенційних кандидатів розглядаються як вихідні дані для встановлення синтаксичних, семантичних ознак та еквівалентів перекладу з англійської. Визначення «лінгвістичної правильності» виявлених багатослівних конструкцій пропонується здійснити за допомогою морфолого-синтаксичних фільтрів і латентного семантичного індексування для опису прихованого значення колокацій: *blow hot and cold, spill the beans* [8, с. 163].

Загалом, вилучені за комбінованим підходом колокації, як правило, не тільки покривають тексти досліджуваного домену, але й можуть слугувати додатковою підставою для укладання різних словників [18, p. 156]. Наявність обов'язкового етапу постредагування й доопрацювання результатів в описаних розробках дозволяє класифікувати аналізований підхід як **напівавтоматичний**. Підвищення ефективності результатів розпізнавання колокацій за зазначеним підходом вимагає використання більших за обсягом корпусів, програм лематизації і фільтрації вилучених сполук за синтаксичними моделями [18, p. 156].

Окреслені вище принципи комбінованого підходу враховано для укладання корпусного Словника колокацій українського юридичного дискурсу [2, с. 44]. З цією метою в межах Корпусу української мови створено підкорпус законодавчих текстів [22]. Відсутність реєстру потенційних колокацій і корпусний метод їх вилучення із текстів передбачає застосування методики відкритого списку. Вважається, що автоматично вилучені з текстів колокації, як правило, покривають певний досліджуваний домен [18, p. 156]. Однак гіпотетично очікуваний реєстр має, за винятком терміносполучень юридичної підмови, також містити й загальномовні колокації, встановлення яких уможливить застосування отриманих даних для вивчення текстів інших функціональних стилів.

Теоретично функціональні можливості Корпусу української мови дозволяють здійснити лише пошук окремих сполук через задавання леми й відповідних граматичних обмежень [17]. Однак безпосереднє здобуття інформації щодо статистичних ознак й укладання реєстру колокацій потребують розробки відповідного програмного забезпечення. За пропонованою методикою автоматичне вилучення й ідентифікація колокацій ґрунтується на джерельній базі підкорпусу законодавчих текстів, використанні пошукового інструментарію Корпусу української мови й додатково розробленого програмного забезпечення для ідентифікації і вилучення колокацій (автор – В. М. Сорокін). Розроблена програма являє собою серію запитів мовою SQL для екстракції пар слів за певним статистичним порогом їх сумісної появи в текстах організованої вибірки й подальшим обчисленням імовірнісних характеристик.

Використання **статистичного підходу** для ідентифікації й автоматичного вилучення двослівних сполучень дозволяє ігнорувати умови щодо якісного складу колокацій, які інтерпретуються як «невипадкові поєднання» двох слів, характерні як «для мови в цілому (текстів будь-якого типу), так і певного типу текстів або навіть (під)вибірки текстів» [13]. Для відбору потенційних кандидатів колокацій встановлено статистичний поріг, відповідно до якого сполука двох контактно розташованих слів має зустрітися в організованій вибірці текстів обсягом в 1 млн слів принаймні двічі [2, с. 44]. Таке рішення пояснюється максимальними показниками частоти біграмних сполук у текстах різних функціональних стилів [11, с. 44].

Унаслідок статистичного опрацювання текстів і лематизації укладено список 64.361 потенційної колокації у вигляді розділених комою пар лем з показниками частоти сполуки в аналізованому підкорпусі: *договірний, сторона, 1902; закон, Україна, 1032*. Отриманий у такий спосіб реєстр двослівних усталених сполучень потребує певного редагування на предмет зняття лексико-граматичної омонімії (*заробітний, платити; наглядовий, рад*) і відбору граматично коректних колокацій. Подальші дослідження мають наукову та практичну цінність з метою тестування описаної методики на матеріалі корпусу текстів різних стилів і встановлення спільних колокацій.

Здійснене дослідження дозволяє дійти таких **висновків**: 1) сучасні методики вилучення й ідентифікації колокацій базуються на лінгвістичному, статистичному й комбінованому підході; 2) основні проблеми опису колокацій полягають у встановленні критеріїв ідентифікації, класифікації, оцінюванні результатів і презентації інформації щодо кількісних і якісних ознак як частини словника; 3) трояка природа колокацій і недосконалість існуючих прийомів потребує впровадження **комбінованого підходу**, що поєднує об'єктивний опис статистичних і лінгвістичних ознак; 4) пропонована методика ідентифікації колокацій у корпусі українських текстів дозволяє автоматично вилучати й обчислювати ймовірнісні характеристики біграмних сполук. Результати екстракції потребують обов'язкового постредагування в аспекті зняття лексико-граматичної омонімії і встановлення граматично коректних колокацій.

ЛІТЕРАТУРА

1. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика : [учеб. пособ.] / [Большакова Е. И., Клышинский Э. С., Ландэ Д. В. и др.]. – М. : МИЭМ, 2011. – 272 с.
2. Бобкова Т. В. Концепція колокації: корпусний підхід / Т. В. Бобкова // Науковий вісник Міжнар. гуманітарного ун-ту. Серія «Філологія» : [зб. наук. праць]. – Одеса, 2014. – Вип. 10. – Т. 2 – С. 42–45.
3. Бобкова Т. В. Теоретико-методологічні підходи до вивчення колокацій / Т. В. Бобкова // Вісник Київського нац. лінгв. ун-ту. Серія : Філологія. – 2014. – Т. 17. – № 2. – С. 14–22.
4. Бялек Э. Коллокация как единица перевода / Э. Бялек // Cuadernos de Rusistica Espanola. – 2004. – № 1. – С. 223–231.
5. Дарчук Н. Комп'ютерне анотування українського тексту : результати і перспективи : [монографія] / Наталія Дарчук. – К. : Освіта України, 2013. – 544 с.
6. Зацеркляний М. М. Об'єктно-орієнтований тезаурус і словник колокацій для бази знань криміналістичних інформаційних систем / М. М. Зацеркляний, Д. Ю. Узлов // Системи обробки інформації. – 2013. – Вип. 2. – С. 183–186.
7. Лендау С. І. Словники : мистецтво та ремесло лексикографії / Сидні І. Лендау ; [пер. з англ.]. – [2 вид.]. – К. : К. І. С., 2012. – 480 с.
8. Романюк А. Розпізнавання багатослівних конструкцій / А. Романюк, Г. Кваснюк, М. Романишин // Вісник Нац. ун-ту «Львівська політехніка». Комп'ютерні системи проектування. Теорія і практика. – 2011. – № 711. – С. 158–165.
9. Хайрова Н. Ф. Идентификация криминально значимых коллокаций в украиноязычных текстах [Електронний ресурс] / Н. Ф. Хайрова, Д. Ю. Узлов // Зб. наук. праць Військового ін-ту Київського нац. ун-ту ім. Т. Шевченка. – 2013. – Вип. 44. – С. 147–151. – Режим доступу : http://nbuv.gov.ua/j-pdf/Znpviknu_2013_44 _27. pdf.
10. Хохлова М. В. Исследование лексико-синтаксической сочетаемости в русском языке с помощью статистических методов (на базе корпусов текстов) : автореф. дис. на соиск. уч. степ. канд. филол. наук : спец. 10.02.21 «Прикладная и математическая лингвистика»/ М. В. Хохлова. – СПб., 2010. – 26 с.
11. Червяк А. В. Об особенностях применения статистических алгоритмов выявления устойчивых словесных цепочек / А. В. Червяк, А. В. Вечур, Е. Л. Шевченко, В. Н. Ляпота // Восточно-Европейский журнал передовых технологий. – 2011. – 4/2(52). – С. 41–47.
12. Шкурко В. В. Лексикографічний агент екстракції колокацій у природномовному тексті / В. В. Шкурко // Вісник Київського нац. ун-ту ім. Т. Шевченка. Серія : Літературознавство. Мовознавство. Фольклористика. – 2012. – № 28. – С. 31–35.
13. Ягунова Е. В. От коллокаций к конструкциям [Електронний ресурс] / Е. В. Ягунова, Л. М. Пивоварова // Acta linguistica petropolitana. Тр. Ин-та лингв. исслед. РАН. – 2011. – Режим доступу : http://www.webground.su/data/lit/pivovarova yagunova/Ot kollokatsiy k konstruktsiyam.pdf.
14. Ananiadou S. A Methodology for Automatic Term Recognition / S. Ananiadou // Proceedings of the 15th conference on Computational linguistics. – 1994. – Vol. 2 – P. 1034–1038.
15. Biber D. Corpus linguistics : investigating language structure and use / D. Biber, S. Conrad, R. Reppen. – Cambridge : Cambridge University Press, 1998. – 310 p.
16. Kathleen R. Collocations / Kathleen R. McKeown and Dragomir R. Radev // A Handbook of Natural Language Processing / Ed R. Dale, H. Moisl, H. Somers. – 2000. – P. 507–523.
17. Kotsyba N. Praktyczny przewodnik po korpusach języka ukraińskiego [Electronic resource] N. Kotsyba // Praktyczny przewodnik po korpusach języków słowiańskich. – Warsaw, 2013. – Mode of access : http://www.domeczek.pl/~natko/ papers/przewodnik-korp-ukr2013.pdf.
18. Seljan S. First Steps in Term and Collocation Extraction from English-Croatian Corpus / S. Seljan, A. Gašpar // Computational Language Analysis, Computer-Assisted Translation and e-Language Learning / Ed S. Seljan. – Zagreb : Zavod za informacijske studije, 2012. – P. 149–156.
19. Smadja F. A. Automatically Extracting and Representing Collocations for Language Generation / F. A. Smadja, K. R. McKeown // Proceedings on the 28-th Annual Meeting of the ACL. – Pittsburg : PA, 1990. – P. 252–259.
20. Wermter J. Collocation and Term Extraction Using Linguistically Enhanced Statistical Methods / Joachim Wermter // Dissertation zur Erlangung des akademischen Grades Doctor philosophiae. – Jena : der Friedrich-Schiller-Universit, 2008. – 230 p.
21. Тримовний тлумачний словник термінів з комп'ютерної лінгвістики [Електронний ресурс] / Т. В. Бобкова, К. М. Лебедєв. – Київ, 2010. – Режим доступу : http://www.complinguide.com.ua/Glossary.aspx.
22. Законодавчі тексти [Електронний ресурс]. – Режим доступу : http://www.mova.info/ corpus2.aspx.

Бобкова Т. В., *Киевский национальный лингвистический университет, г. Киев, Украина*

ОСНОВНЫЕ ПОДХОДЫ К ИДЕНТИФИКАЦИИ И ИЗВЛЕЧЕНИЮ КОЛЛОКАЦИЙ ИЗ ТЕКСТА

Идентификация и экстракция коллокаций является частью общей проблемы автоматического распознавания и обработки текста на естественном языке и поиска информации. Существующие методики и приемы идентификации коллокаций основываются на лингвистическом, статистическом и комбинированном подходах, отличающихся критериями и последовательностью применяемых процедур. Статистический подход предусматривает использование корпусных методов и идентификацию коллокации как статистически значимой единицы. В работе обосновывается необходимость использования комбинированного подхода к извлечению коллокаций. Разработана методика идентификации коллокаций, позволяющая на основе статистической обработки и лемматизации автоматически извлекать устойчивые сочетания из корпуса текстов. Результаты экстракции требуют постредактирования с целью снятия лексико-грамматической омонимии и

установления грамматически правильных коллокаций. Применение большего по объему корпуса и лингвистических фильтров обеспечит повышение эффективности результатов идентификации коллокаций.

Ключевые слова: *коллокация; текст; извлечение; идентификация; лингвистический подход; статистический подход; комбинированный подход.*

Bobkova T. V., *Kyiv National Linguistics University, Kyiv, Ukraine*

THE MAIN APPROACHES TOWARD COLLOCATION IDENTIFICATION AND EXTRACTION

Collocation lists represent valuable resources covering specific domain and frequent multiword expressions, which can be used in informational retrieval, machine translation researches, compiling dictionaries, thesauri, semantic networks. Collocation recognition and extraction is based on linguistic, nonlinguistic and combining approaches. Based on a linguistic approach human-created collocation lists contain more meaningful candidates, but are more time-consuming. Statistically compiled lists have to be filtered by language patterns and require post-editing, but are still created at lower cost and time. There are multidirectional trends in the Ukrainian collocation study: in using both linguistic and statistical approaches in order to gain experience in building a collocation base. In the paper a necessity of combining approach use in collocation extraction is proved. The generic method for collocation extraction has been presented, relying on subset of Corpus of Ukrainian Law Acts used as a source and programming instrument for collocation identification. Statistical approach is based on the availability of text resources, corpus methods and earmarking a collocation as a statistical unit. Collocation extraction includes the following phases: automatic collocation acquisition and colocation recognition including a verification by an expert. Statistically created collocation list is filtered by linguistic engineering tool and lemmatization, though human post-editing is required. The results would be considerably improved if bigger corpus and language patterns are used. Extracted collocations tend to cover legislative domain and could serve as an additional base to dictionaries.

Keywords: *collocation; text; extraction; identification; linguistic approach; statistical approach; mixed approach.*

 Дата надходження статті до редколегії 19.03.2015