УДК 81'33

Т. В. Бобкова,
Київський національний лінгвістичний університет, м. Київ

КОРПУСНИЙ СЛОВНИК КОЛОКАЦІЙ: МЕТОДИКА УКЛАДАННЯ

У статті аналізується методичні етапи укладання словника колокацій на базі корпусу текстів. У якості лексикографічного джерела словника колокацій розглядається підкорпус українських законодавчих текстів. Обґрунтовується необхідність застосування багатометодного підходу опису колокацій.
Ключові слова: *колокація, корпус текстів, метод, корпусний підхід, словник колокацій.*

В статье описывается методические этапы составления словаря коллокаций на основе корпуса текстов. В качестве лексикографического источника словаря коллокаций рассматривается подкорпус украинских законодательных текстов. Обосновывается необходимость использования многометодного подхода к описанию коллокаций.
Ключевые слова: *коллокация, корпус текстов, метод, корпусный подход, словарь коллокаций.*

The article deals with the methodology order of the corpus collocations dictionary compiling. As an example of lexicographical source for dictionary compiling the Subcorpus of Ukrainian Law Acts is analyzed. A necessity to use multimethod corpus-based approach in collocation study is proved.
Keywords: *collocation, corpus, method, corpus-based approach, collocations dictionary.*

Актуальність дослідження пов'язана з загальною проблематикою автоматичного аналізу тексту й укладання словників. Завдання розробки й удосконалення прикладних систем аналізу тексту потребують дослідження великих обсягів природно-мовного матеріалу. Сучасні дослідження автентичних текстів ґрунтується на об'єктивних корпусних даних, вилучених і опрацьованих достовірними методами. Емпіричні дані спостереження корпусів свідчать, що близько 80 % текстів складають не ізольовані лексичні одиниці, а регулярно відтворювані структури [20, p. 62], опис яких становить значний інтерес для прикладних лінгвістів. На позначення стійких синтагматичних послідовностей у корпусній лінгвістиці традиційно вживається термін «колокація» [23; 24; 25]. **Мета статті** – дослідити методологічні засади корпусного підходу укладання словника колокацій. Досягнення поставленої мети передбачає виконання таких **завдань**: 1) дати визначення терміна «колокація»; 2) окреслити основні засади корпусного підходу вивчення колокацій; 3) обґрунтувати необхідність застосування комплексної методики опису; 4) встановити методологічні етапи укладання корпусного словника колокацій.

З самого початку формування корпусної лінгвістки ключовою концепцією нової парадигми стає колокація (English Collocation Studies, 1966), а її опис як «лексичної структури» – основною метою розробки корпусів [24, p. 137–138]. Подальші дослідження в галузі корпусної лінгвістики призвели до поширення терміна в лексикології, деривatології, синтаксисі, психолінгвістиці, лексикографії, теорії перекладу, лінгвістиці тексту [20, p. 77] й актуалізували необхідність: а) точного **визначення** поняття, б) окреслення **статусу** колокацій серед стійких сполук і в) розв'язання проблем, пов'язаних з **вилученням** та **описом** колокацій. Певною мірою поширеність терміна пояснюється різними значеннями, що вкладаються в поняття «колокація» з метою відображення лексико-граматичних і статистичних ознак. Багатовимірність феномену не сприяла точному визначенню й спонукала до аспектуалізації поняття колокації: виділення широкого значення – в лінгвістичному й вузького – в статистичному аспекті [20, p. 40]. На сучасному етапі в широкому сенсі під колокацією розуміється звичне використання двох слів разом [20, p. 40; 23, p. 115; 8, с. 303]. У вузькому значенні колокація – це слова, що зустрічаються у тексті частіше разом, ніж за випадковою вірогідністю окремо [23, p. 116]. Залежно від інтерпретації аналізованого поняття визначається статус, якісний і кількісний склад колокацій, зокрема питання щодо «граматичних колокацій» – сполучень лексичної одиниці й службового слова [19, p. 49]. Троякa природа феномену зумовлює широкий спектр прикладних досліджень і необхідність розробки комплексної методології укладання спеціалізованих словників колокацій (Г. Кустова; А. Романюк; М. Зацеркляний; J. Sinclair; M. Benson; F. Čermàk; F. J. Hausmann; R. Marcinkevičienė; C. Müller-Spitzer).

Залежно від інтерпретації концепції колокації у традиціях певної школи й застосовуваної методології опису виділяють семантико-синтаксичний, контекстно-орієнтований і корпусно-базований підходи [3, с. 16]. В основу традиційних досліджень колокацій покладено семантико-синтаксичний підхід: зв'язність стійких сполук визначається семантичною сумісністю компонентів [2, с. 77; 19, с. 19-20] і / або певною синтаксичною моделлю [14, с. 3; 16; 1, с. 22; 22, p. 27; 13, с. 31; 3, с. 16]. Методологічним підґрунтям традиційних досліджень колокацій слугують контекстологічний [7, с. 205-206; 20, p. 75; 15, с. 10], дистрибутивний аналіз [15, с. 10; 3, с. 16], моделювання [5, с. 8; 14, с. 10-11] й квантитативні методи [7, с. 202-206; 14, с. 10-11]. Поява й прогрес у розробці електронних корпусів, їх принципова відмінність від тексту як емпіричної бази [25, p. 207-208] призвела до необхідності впровадження нового підходу дослідження колокацій. За корпусним підходом колокація розглядається як вияв ідіоматичності, яка разом з принципом необмеженого вибору зумовлює побудову тексту [23, p. 110] на рівні структури й значення. На відміну від семантико-синтаксичного в межах корпусного підходу під ідіоматичністю розуміється наявність у мові великої кількості готових фраз, у широкому сенсі – сумісно вживаних слів, які не обов'язково визначаються семантичною неподільністю, й використовуються мовцем згідно до індивідуальних преференцій [20, с. 61].

У межах корпусного підходу поняття колокації набуває нового виміру: визначальними для зв'язності сполуки визнаються статистичні чинники. Саме тому корпусний підхід класифікують як **статистичний**, а колокації визначають як «статистично стійкі» словосполучення [14, с. 4-5], або навіть «чисто статистичний феномен» [18, с. 47]. У сучасній корпусній лексикографії виявлення й опис колокацій здійснюється на базі **статистичних методів** [14, с. 12; 22, p. 28]. Так, для визначення сили смислового [16, с. 137] й синтагматичного зв'язку [14, с. 4] між елементами словосполучення використовуються статистичні показники – міри асоціацій. Зняття обмежень щодо якісного й кількісного складу забезпечує переваги в вилученні колокацій з тексту, виявленні граматично некоректних структур [18, p. 47] і, зрештою – визначенні лексикографічно релевантних сполук, що становлять основу реєстру словника. Як статистично сталі сполуки колокації розглядаються як певний континуум, шкала без чітких меж між вільними й фразеологізованими сполученнями [18, p. 48; 14, с. 10]. На основі узагальнення спостережень великої кількості корпусів Дж. Синклером [23, p. 112–113] встановлено **характерні ознаки** колокацій: 1) невизначеність обсягу (*set eyes on*); 2–4) можливість внутрішньої лексичної (*in some cases / in some instances*) і синтактичної варіативності (*recriminate isn't in his nature / isn't in his nature of an academic*); 5–6) висока частотність компонентів (*hard work, hard evidence*), тенденція до сумісної вживаності слів у певних граматичних структурах (*set about testing*) і в певному семантичному оточенні (*happen – accident*). Отже, перспектива об'єктивного й ґрунтовного опису колокацій стає можливою лише з удосконаленням корпусної методології, базованої на відповідному програмно-статистичному забезпеченні, здатному виявити всі колокації, спосіб і частоту сполучення компонентів. У такий спосіб,

впровадження корпусного підходу уможливлює встановлення об'єктивних критеріїв опису колокацій, що збігається з метою їх лексикографічного аналізу [24, p. 138].

У сучасній українській лінгвістиці укладання словників колокацій ґрунтуються на засадах або семантико-синтаксичного – колокаційна лексикографічна система [13, с. 32], або статистичного підходу – лексикон багатослівних сполук [12, с. 163], словник колокацій для бази знань криміналістичних інформаційних систем [6, с. 184]. Однак слід розуміти, що здійснення опису колокацій у межах лише одного підходу вкрай утруднено [18, p. 47]. Так, застосовувані в корпусному підході статистичні міри не є всеосяжними, оскільки не диференціюють ядро й колокат стійкої сполуки [17, p. 131–132]. Водночас поєднання корпусного аналізу з лінгвістичними методами створює перспективи виходу не тільки за межі простої статистики, а й деяких традиційних уявлень про мову [20, p. 53–57]. У сучасних дослідженнях колокацій можливості корпусного підходу суміщаються з синтаксичним [22; 1; 13], семантико-синтаксичним [5; 14], когнітивним [20; 15] і психолінгвістичним підходом [17]. Отже, вивчення колокацій має ґрунтуватись на багатоаспектному підході з використанням корпусного методу в поєднанні з дистрибутивним, контекстологічним аналізом і моделюванням, а ні в якому разі не як єдино вірної методологічної основи.

Вибір корпусного підходу як базового для укладання словника колокацій українського юридичного дискурсу пояснюється можливостями виявлення в корпусі одиниць, «зазвичай відсутніх у цитатних картотеках», й обчислення «відносної ваги їх зв'язку» [8, с. 303]. Відповідно до методологічної схеми [20, p. 106], створення корпусного словника ґрунтується на застосуванні програмних засобів опрацювання текстів: корпус → конкорданс → типова модель використання одиниці → словник. При цьому під типовою моделлю використання розуміється колокація – характерне оточення, або звичний контекст одиниці. В укладанні корпусного словника колокацій слід виділити такі основні **методологічні етапи**: 1) розробка лексикографічного корпусу, 2) формування реєстру словника на основі статистичного аналізу колокацій, 3) встановлення стандарту словникової статті й 4) визначення лексико-синтаксичних характеристик колокацій. Джерелом корпусного словника колокацій слугує розроблений нами підкорпус українських законодавчих текстів [26]. Відповідно до дослідного призначення в основу планування підкорпусу покладено детермінативні параметри скінченності обсягу, репрезентативності, збалансованості, анотованості й автентичності текстів, які становлять ключові вимоги до корпусних об'єктів. Фіксований обсяг підкорпусу (1.157 375 млн. слів) дозволяє оптимально детермінувати поріг відображення предметної галузі [8, с. 332] й відповідає меті укладання корпусного словника: під **колокацією** розуміється сполука слів, що зустрілась у масиві текстів обсягом в 1 млн. принаймні двічі [20, p. 104].

Вибір у якості документального джерела підкорпусу «Зібрання законодавства України» [9] зумовлено максимальною насиченістю колокацій в законодавчих текстах: багатослівних найменувань (*асамблея союзу, Гаазька угода*), терміносполучень (*міжнародна реєстрація, юридична особа*), мовленнєвих формул і кліше (*перешкоджати застосуванню, додержуватись положень*), похідних прийменників і сполучників (*у відповідності з, у силу того, що*). Репрезентативність [8, с. 325] підкорпусу забезпечується документальною базою, діапазоном 43 текстових типів, загальним обсягом текстів і охопленим часовим інтервалом (з 1991 р. по сьогодні). Відібрані згідно до експліцитних та імпліцитних критеріїв тексти забезпечують адекватну фіксацію юридичної підмови й дозволяють класифікувати розроблений підкорпус як представницький для дослідження колокацій сучасного українського юридичного дискурсу. Як складова Корпусу української мови підкорпус законодавчих текстів забезпечено міжнародною метатекстовою розміткою (категорія – *Нехудожні тексти*, жанр – *Законодавчі тексти*, тип тексту – *Договір, Інструкція* тощо) й набором програм морфологічної, синтаксичної і лексико-семантичної розмітки в автоматичному режимі з подальшим зняттям омонімії, необхідним для укладання корпусного словника.

Автоматичне вилучення колокацій і формування реєстру словника здійснюється на основі джерельної бази та програмно-статистичного інструментарія розробленого підкорпусу законодавчих текстів. Підґрунтям укладання корпусного словника колокацій слугує **комплексна методика**, базована на поєднанні корпусного, дистрибутивного й контекстологічного методів. При цьому основним методом отримання лексикографічних даних слугує корпусний, а в якості методів аналізу застосовуються статистичний, дистрибутивний і контекстологічний методи. Корпусний метод реалізується за допомогою **конкордансингу** – основного прийому програмного опрацювання текстів й накопичення «сировини для лексикографічного опису» [25, p. 207]. Застосування конкордансингу вирішує проблему здобуття лексикографічних даних у необхідному й достатньому обсязі [21, p. 142], а для усунення їх надлишку й узагальнення даних корпусного аналізу використовується система статистичних фільтрів корпусу, яка дозволяє задавати порогові значення, ранжувати результати пошуку за різними параметрами й отримувати статистично значиму інформацію [21, p. 140–141].

Відбір конкордансів для формування реєстру словника обмежується, насамперед, через **статистичну інтерпретацію** колокації, покладену в основу «формального, структурного визначення» одиниці [10, с. 9]. У якості порогового значення встановлено абсолютну частоту сполуки: колокацією визнається поєднання двох слів, зафіксоване в текстах підкорпусу принаймні двічі [23, p. 57]. Вибір в якості колокації біграмних сполук пояснюється дією семіотичного закону простоти Дж. К. Ципфа: «двочленні синтаксичні моделі вживаються частіше, ніж моделі з більшою кількістю складників» [11, с. 446]. Використання корпусного підходу для автоматичного вилучення біграм дозволяє ігнорувати обмеження щодо якісного складу колокацій, які розглядаються як «звичне поєднання двох слів» [23, p. 116; 20, p. 40; 8, с. 303]. У такий спосіб, функціональні можливості корпус-менеджера уможливлюють генерування словника через програмне опрацювання корпусу з мінімальним втручанням лексикографа [21, p. 139]. Статистична інтерпретація колокації визначає також і організацію **зональної вибірки** текстів обсягом в 1 млн. слів [20, p. 104]. Генеральною сукупністю для зональної вибірки слугує підкорпус законодавчих текстів, а зоною – «сукупність одиниць, встановлена за певною ознакою» [10, с. 14], тобто частотою сполуки ($f \geq 2$) в текстах підкорпусу. У результаті зональної вибірки сформовано дослідний масив, до якого включено 128 зразків текстів 31 текстового типу. До вибірки увійшли тексти від 3 тис. слів (діапазон обсягу – від 3.014 до 44.008 тис. слів), повністю вилучено такі текстові типи, як *Вказівки, Декларація, Закон, Заява, Звернення, Норми, Резолюція, Розпорядження, Указ, Ухвала*, що не задовольняють окресленим вимогам.

На етапі конкордансингу автоматичний корпусний аналіз завершується, й починається робота лінгвіста з визначення «лексикографічно релевантних» [21, p. 141], або типових моделей, які становлять основу реєстру словника. Подальший розгляд колокацій на предмет включення до реєстру здійснюється на підставі **статистичного аналізу** – ранжування й обчислення ймовірнісних характеристик вилучених з підкорпусу колокацій. Ранжування вилучених пар слів здійснюється відповідно до їхньої відносної ймовірності бути колокацією [17, p. 129]. Для статистичного узагальнення попередніх результатів використано методику асимптотичної гіпотези (Ch. Manning, H. Schütze, 1999). Нульовою гіпотезою є твердження, що слова в текстах з'являються разом не частіше, ніж можна очікувати їх випадкову появу поодинці. Для перевірки нульової гіпотези встановлюється частота колокації та окремих компонентів. Ймовірність того, що відібрані пари є колокаціями, обчислюється за формулою: $P(w_1w_2) = P(w_1) \times P(w_2)$, де ймовірність появи (P) певного слова (w) вираховується як його частота поділена на загальну кількість слів у зональній вибірці (1 млн.). Так, обчислення для сполуки *державна*

адміністрація – $P(w_1w_2)$ = 0,003541 × 0,000408 = 0,000001444728, показує, що випадково вибрані з масиву слова будуть цією парою з доволі низькою ймовірністю в 1,444728e-06, тобто можна передбачити випадкову появу цього сполучення в масиві 1,44 рази. Однак, оскільки за даними спостережень, абсолютна частота сполуки *державна адміністрація* дорівнює 200, то можна стверджувати, що ці слова зустрічаються разом частіше, ніж випадково, а значить являються колокацією.

Статистична інтерпретація колокації, покладена в основу автоматичного вилучення сполук з тексту, значно розширює межі аналізованого об'єкту. У цьому розумінні для формування реєстру словника після постредагування й ранжування потрібно вирішити низку проблем, пов'язаних з ідентифікацією колокацій серед вилучених сполук, а саме щодо: 1) сурядного vs. підрядного зв'язку між компонентами сполук, 2) предикативних структур і 3) лексичних vs. граматичних колокацій. Троя́ка природа колокації вимагає доповнення корпусного аналізу **контекстологічним** і **дистрибутивним методом**. Так, при встановленні стандарту словникової статті корпусного словника колокацій мають бути враховані як статистичні, так і суто лінгвістичні ознаки сполук. Саме тому ранжовані колокації мають бути проаналізовані в аспекті класифікації лексичних і синтаксичних відношень [17, p. 133] традиційними методами. Якщо колокації з підрядним зв'язком розглядати з точки зору формальної структури як вільні сполучення, то словникова стаття має включати інформацію про граматичні характеристики кожного з членів словосполучення й тип синтаксичного зв'язку [4, с. 177]. Так, для леми *адміністрація*: серед лексичних колокацій – найчастотнішою є ядрова модель – ад'єктив + іменник, зафіксована 286 разів (*державна/в'язнична/ податкова/ тимчасова/центральна адміністрація*), ад'юнктні моделі – **дієслово** + іменник – зустрілась 4 рази (*здійснюватись адміністрацією*), **іменник** + іменник – 20 разів (*адреса/глава/іменем/листування/показання адміністрації, виконання/ невидача адміністрацією*), а також граматичні колокації **прийменник** + іменник (*на/через адміністрацію, з/від адміністрації*) – 17 разів.

Розроблений підкорпус законодавчих текстів й описана методика укладання словника колокацій мають значний дослідний потенціал у лексикографічній практиці й прикладній лінгвістиці для створення онтологій, розв'язання завдань машинного перекладу, налагодження лінгвістичних процесорів та інформаційного пошуку. Здійснена розвідка дозволяє дійти таких **висновків**: 1. За корпусним підходом ідіоматичність інтерпретується як наявність у мові великої кількості сумісно вживаних сполук, які не обов'язково характеризується семантичною зв'язаністю й використовуються мовцем відповідно до індивідуальних преференцій. 2. Під колокаціями слід розуміти статистично стійкі сполучення, що становлять певний континуум між вільними й сталими сполуками. 3. Основними ознаками колокацій доцільно вважати: ідіоматичність, невизначеність обсягу, можливість лексичної і синтаксичної варіативності, функціонування в певних граматичних структурах і семантичному оточенні. 4. Методика опису колокацій має ґрунтуватись на врахуванні лінгвістичних і статистичних чинників. 5. Укладання словника колокацій передбачає поєднання корпусного методу з дистрибутивним і констекстологічним аналізом.

Література:

1. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика : [учеб. пособ.] / Е. И. Большакова и др. – М. : МИЭМ, 2011. – 272 с.
2. Борисова Е.Г. Фразеологическое значение в устойчивых словосочетаниях / Е.Г. Борисова, О.В. Захарова // Филологические науки. – 1994. – № 4. – С. 77–84.
3. Гладка В.А. Структурно-синтаксичний підхід у вивченні колокацій (на матеріалі французької мови) / В. А. Гладка // Наукові записки нац. ун-ту «Острозька академія». Серія «Філологічна». – 2013. – Вип. 39.– С. 16–20.
4. Дарчук Н. Комп'ютерне анотування українського тексту : результати і перспективи : [монографія] / Наталія Дарчук. – К. : Освіта України, 2013. – 544 с.
5. Дяченко П. В. Разработка компьютерных методов обучения владению языком с помощью аппарата лексических функцій : автореф. дис. на соиск. уч. степ. канд. тех. наук : спец. 05.13.17 «Теоретические основы информатики» / П. В. Дяченко. – М., 2008. – 20 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. Левицкий В.В. Семасиология / Виктор Васильевич Левицкий. – Винница : Нова Книга, 2006. – 508 с.
8. Лендау С. І. Словники : мистецтво та ремесло лексикографії / Сидні І. Лендау; [пер. з англ.]. – К. : К. І. С., 2012. – 480 с.
9. Омега : Зібрання законодавства України. – [Електронний ресурс]. – К. : УІПЦ, 2009.
10. Перебийніс В. І. Статистичні методи для лінгвістів : [посіб.] / Валентина Ісидорівна Перебийніс. – Вінниця : Нова Книга, 2002. – 171 с.
11. Перебийніс В.І. Частота мовних одиниць як відображення їхніх системних характеристик / В. І. Перебийніс, Т.В. Бобкова // Проблеми загального, германського та слов'янського мовознавства : [зб. наук. праць]. – Чернівці : Книги – ХХІ, 2008. – С. 446–453.
12. 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.
13. 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.
14. Хохлова М. В. Исследование лексико-синтаксической сочетаемости в русском языке с помощью статистических методов (на базе корпусов текстов) : автореф. дис.... канд. филол. наук / М. В. Хохлова, С.-Петербургский ун-т. – СПб., 2010. – 26 с.
15. Шилихина К. М. Дискурсивная практика иронии : когнитивный, семантический и прагматический аспекты : дис.... докт. филол. наук : 10.02.19 «Теория языка» / Ксения Михайловна Шилихина. – Воронеж, 2014. – 399 с.
16. Ягунова Е. В. От коллокаций к конструкциям / Е. В. Ягунова, Л. М. Пивоварова // Труды Института лингвистических исследований РАН. – 2011. – [Электронный ресурс] – Режим доступа : http://www.webground.su/data lit/pivovarova_yagunova/Ot_kollokatsiy_k_konstruktsiya.pdf
17. Durrant Ph. Are high-frequency collocations psychologically real? / Ph. Durrant, A. Doherty // Corpus Linguistics and Linguistic Theory. – 2010. – Vol. 6, No 2. – P. 125–155.
18. Fontenelle Th. What on earth are collocations? / Th. Fontenelle // English today. – 1994. – Vol. 10 (40), No. 4. – P. 42–48.
19. Kapstad M. Faste uttrykk i russisk og norsk med henblikk på russiskundervisning for nordmenn / Masteroppgave i russisk språk ved Institutt for litteratur / Maria Kapstad. – Våren, 2006. – 124 c.
20. Marcinkevičienė R. Lietuvių kalbos kolokacijos : [monografija] / Ruta Marcinkevičienė. – Kaunas : Vytauto Didžiojo universitet, 2010. – 212 p.
21. Rundell M. Good Old-fashioned Lexicography : Human Judgmen and the Limits of Automation / M. Rundell // Lexicography and Natural Language Processing : a Festschrift in Honour of B.T.S. Atkins. – Grenoble : EURALEX, 2002. – P. 138–155.
22. Seretan V. Syntax-Based Collocation Extraction / Violeta Seretan. – Berlin : Springer Science & Business Media, 2011. – 232 p.
23. Sinclair J. Corpus, Concordance, Collocation / John Sinclair. – Oxford : Oxford University Press, 1991. – 200 p.
24. Teubert W. Linguistique de corpus : un alternative / W. Teubert // Semen. Critical Discourse Analysis I. Les notions de contexte et d'acteurs sociaux / – 2009. – Vol. 27. – P. 130–152.
25. Tognini-Bonelli E. Corpus Classroom Currency / E. Tognini-Bonelli // Darbai ir Dienos. – 2000. – No 24. – P. 205–243.
26. Законодавчі тексти // Корпус текстів української мови. – [Електронний ресурс] – Режим доступу : http://www.mova.info/corpus2.aspx