# СТРУКТУРНА, ПРИКЛАДНА ТА МАТЕМАТИЧНА ЛІНГВІСТИКА

УДК 81’33

***Бобкова Т. В.***
Київський національний лінгвістичний університет

## КЛАСИФІКАЦІЯ УКРАЇНСЬКИХ КОЛОКАЦІЙ НА МАТЕРІАЛІ КОРПУСУ ТЕКСТІВ

*Стаття присвячена обґрунтуванню процедурних засад класифікації українських колокацій. Запропоновано бінарну типологію підходів до опису системних і функціональних ознак колокацій. Виявлено основні проблеми морфолого-синтаксичної класифікації українських колокацій. Установлено класифікаційні ряди та класи двослівних лексичних, предикативних і граматичних колокацій. Визначено морфолого-синтаксичні особливості функціонування колокацій у законодавчих текстах.*

***Ключові слова:*** *колокація, корпус, класифікація, лексичні колокації, предикативні колокації, граматичні колокації.*

Валентині Ісидорівні Перебийніс
присвячується

**Постановка проблеми.** Розроблення прикладних систем з автоматичного аналізу тексту й укладання словників потребує опрацювання великих обсягів інформації, представленої лінгвістичними корпусами. Відомості корпусів різноструктурних мов свідчать, що в текстах регулярно трапляються усталені словосполучення, які під час мовлення цілком видобуваються з пам’яті мовця і є неоднорідними за структурою та семантикою. Однак більшість прикладних лінгвістичних систем орієнтовані на традиційне визначення слова. На позначення співвідносного зі словом усталеного сполучення в сучасній лінгвістиці вживається термін колокація [4; 6; 7; 8; 10; 11; 13; 15; 20; 23; 24; 25; 26]. Окремі аспекти усталених сполучень були об’єктом досліджень уже в першій половині XX ст. (Ш. Баллі, Дж. Ферс, В. Порциг, В. Виноградов), а з 60-х років опис колокацій стає одним з основних завдань корпусної лінгвістики. Так, необхідність розв’язання завдань з автоматичного аналізу українського тексту призвела до появи досліджень, присвячених проблемам ідентифікації колокацій [5; 7; 9; 11]. Проте в українській лінгвістиці цей феномен потребує ґрунтовного аналізу, оскільки інвентар і класифікацію українських колокацій не встановлено.

**Аналіз останніх досліджень і публікацій.** Питання класифікації колокацій є дискусійним і безпосередньо пов’язане з традиціями певного напряму, критеріями ідентифікації й завданнями дослідження [4, с. 18, 13, 25]. Нині вживання терміна «колокація» фахівцями різних галузей призвело до серйозної плутанини [15, p. 9], що зумовлюється багатоаспектністю й невизначеним статусом усталених сполучень у мові [16, p. 133; 20, p. 89; 26, p. 206]. Залежно від критеріїв до колокацій часто зараховують інші види сполучень, на позначення одного поняття використовують різні терміни (граматична колокація – колігація), і це значно ускладнює класифікацію. Наразі критерії виділення колокацій поділяються на «необхідні умови та континууми» [17, p. 50]. Необхідні умови застосовуються для ідентифікації й класифікації колокацій на підставі усталених класів слів [18, p. 399; 24, p. 116]. Критерії-континууми визначають ступінь прояву класифікаційних ознак: 1) семантична транспарентність; 2) композиційність; 3) сполучуваність; 4) частота; 5) перед-

бачуваність [17, p. 50]. Розгалужена система з десяти критеріїв-континуумів передбачає градацію функціональних і структурно-семантичних ознак колокацій [21, p. 332], хоча викликає сумніви доцільність віднесення до колокацій граматично не пов'язаних слів, суміжно вживаних у межах однієї колокації (орган державної, вищий навчальний) або випадково вживаних разом (Україна щодо, Україна відповідно). Загалом критерії-континууми є підставою для градуйованої класифікації, хоча жоден із них не є абсолютним для виділення колокацій.

Використання критеріїв ідентифікації та класифікації колокацій повністю залежить від теоретико-методологічних підходів, які доцільно звести до бінарної типології. І. Системоорієнтований підхід – фразеологічний, семантичний, структурний, теорія «Смисл ↔ Текст», лексико-граматичний, синтаксичний. ІІ. Текстоорієнтований підхід – статистичний, лексико-композиційний, контекстно-орієнтований, корпусно-орієнтований [1, с. 8–9]. В аспекті структури й семантики колокації класифікуються як одиниці системи мови за системоорієнтованим підходом, як одиниці тексту – за текстоорієнтованим. Для системоорієнтованої класифікації колокацій вихідними є формальні та якісні ознаки, установлені автоматично чи вручну. Більше уваги при цьому приділяється опису лексичних колокацій – осмислених сполучень однорідних одиниць: згідно з П. Ньюмарком, установлено 7 типів лексичних колокацій з урахуванням синтаксичної функції та лексико-граматичних розрядів іменника [22, p. 114–115]. На підставі ієрархічних відношень між ядром і колокатом П. Нейшен запропонував більш чітку класифікацію в термінах граматичних класів і з виокремленням предикативних колокацій [18, p. 399]. Проте питання виділення граматичних колокацій залишається дискусійним. Систематизацію класифікаційних рядів колокацій можна здійснити на підставі виділених нами **системних ознак** [3]: 1) формальні (довжина, комбінаторні, морфологічні); 2) структурні (синтаксичний і підрядний зв'язок, граматична структура, стійкість); 3) семантичні (кількість значень, семантична структура, лексична варіативність, семантична композиційність); 4) системні функціональні (обмеженість сполучуваності, синтаксична функція, сфера вживання).

В українській прикладній лінгвістиці класифікація колокацій у термінах граматичних класів реалізується через виділення синтаксичних конструкцій [4, с. 16; 9, с. 147; 11, с. 31–33] або структурних моделей [5, с. 336–337]. Найбільш розгалуженою є класифікація колокацій на матеріалі Українського національного лінгвістичного корпусу [5, с. 336–337]: за морфологічними ознаками й без обмежень за довжиною встановлено 11 іменникових моделей – до дев'яти колокатів (ANАродNродАродNродNродАродNродNрод – *остання неділя останнього місяця п'ятого року повноважень Верховної Ради України*), 6 дієслівних – до чотирьох колокатів (VPrepзнахАзнахАзнахNзнах – *повідомити про можливі негативні наслідки*) і 5 ад'єктивних – до трьох колокатів (APrepзнахАзнахNзнах – *спрямований на швидку ліквідацію*). Однак окреслені вище моделі колокацій можуть бути представлені як двослівні, ускладнені послідовним і паралельним приєднанням колокатів до ядра або колокату, або до обох складників одночасно [12, с. 163]. Різноманіття виявлених в автентичних текстах колокацій не завжди вписується в межі усталених морфолого-синтаксичних класів [1, с. 13; 5, с. 337] і майже вдвічі збільшує кількість моделей сполучуваності [6, с. 177–180]. Тому традиційні теорії словосполучення, застосовані в методиці викладання й у перекладі, не знаходять широкого використання в розробленні прикладних систем розпізнавання колокацій [13; 19, p. 4]. У реальних текстових умовах ідентифікація колокацій є досить проблематичною, і верифікації підлягають усі виявлені словосполучення [19, p. 16]. Здійснений аналіз літератури виявляє відсутність загальноприйнятої класифікації українських колокацій; установлення принципів і класифікаційних рядів здійснюється фрагментарно й залежно від прикладних завдань [5; 8; 11]. Безумовно, системоорієнтовані класифікації характеризуються певною суб'єктивністю, тому питання доцільності виділення усталених класів колокацій залишається дискусійним [14, p. XXXI–XXXIV; 18; 24, p. 116]. Перспектива розв'язання окреслених проблем вбачається в застосуванні текстоорієнтованого підходу, що відповідає проміжному статусу колокацій у континуумі сполучуваності [3].

**Постановка завдання.** Метою пропонованого дослідження є теоретичне й експериментальне обґрунтування процедурних засад класифікації українських колокацій за даними корпусу текстів. Досягнення основної мети передбачає встановлення класифікаційних рядів і класів українських колокацій, релевантних для законодавчих текстів. Пропонована класифікація укра-

їнських колокацій базується на положеннях концепції Дж. Ферса, застосованої до морфологічно розвинутої мови [1; 3]. Опис колокації як текстової одиниці передбачає аналіз функціональних ознак, під якими розуміємо частоту, позицію в тексті, реалізацію системних характеристик, комунікативне, прагматичне, емотивне призначення та стилістичне забарвлення [3]. **Колокація** розуміється нами як невипадкове усталене сполучення слів, характерне для мови, використовуваної для усного чи письмового спілкування [1; 2; 3]. Для розпізнавання та подальшого опису колокацій вихідною стає уживаність у тексті [8; 13; 21; 25]. Першою спробою класифікації колокацій за частотою компонентів вважаємо систематику Дж. Синклера [24, p. 115–116]: більш уживане ядро сполучається з менш частотним колокатом за низхідним типом (*договірна 2409 сторона 3788*, *мати 3233 право 2472*), а менш уживане ядро з більш частотним колокатом – за висхідним (*відповідно 2575 до 10672*, *нова 1037 редакція 879)*. Будь-яке поточне слово тексту може бути ядром або колокатом, але не одночасно [24, p. 115]. Імовірно, більшість висхідного типу становлять граматичні колокації, а низхідного – лексичні. Проте відсутність загальноприйнятих стандартів статистичної значущості, низька частота типових колокацій і випадкова сполучуваність розмивають межі класифікації й спричинюють виділення між цими типами нейтральних колокацій [24, p. 116].

Загалом результати корпусного аналізу дозволяють уточнити структурно-семантичні ознаки виділених і описати відсутні в лексикографічних джерелах колокації на підставі таких **функціональних ознак**: 1) статистичні (частота, реалізація системних ознак, рівень аналізу); 2) сполучувальні (граматичний зв'язок, тип зв'язку, колокаційна спеціалізація, стилістична функція). Завдяки систематизації корпусних даних здійснене відокремлення від ідіом і уточнення статусу усталених сполучень у мові, а наявні класифікації доповнено типами граматичних і предикативних колокацій [14; 25; 26, p. 215]. Однак у корпусах текстів класифікація колокацій часто обмежується функціональними можливостями корпус-менеджера й необхідністю верифікації результатів вручну [2]. Розроблені для двослівних сполучень статистичні міри не можуть охопити все різноманіття колокацій [10] і диференціювати ядро та колокат [16, p. 33]. Необхідність узагальнення статистичних результатів і опису структурних відношень вимагає доповнення корпусних методів лінгвістичними. Розпізнавання колокацій розуміємо як частину загальної проблеми автоматичного аналізу тексту, що передбачає виокремлення одиниць безпосередньо з корпусу за допомогою спеціальної процедури запитів (автор програмного забезпечення – В. Сорокін) [2]. Процедура охоплює поетапний корпусний аналіз із використанням морфологічного аналізу Корпусу текстів української мови (http://www.mova.info/corpus.aspx?l1=209). Автоматична ідентифікація на підставі статистичного аналізу дає змогу уникнути свідомо встановлених обмежень щодо структури колокацій, але отримані результати потребують остаточної верифікації лінгвістом-експертом [3]. Редагування автоматично укладеного словника виявляє значну неоднорідність потенційних колокацій в аспекті системних, функціональних ознак і дистрибутивних преференцій. За результатами класифікації першої тисячі найчастотніших колокацій, різноманіття спостережуваних у корпусі усталених сполучень не обмежується іменними моделями [1, с. 13]. За морфологічними та синтаксичними ознаками (предикативність, спосіб зв'язку) доцільно виділити такі класифікаційні ряди колокацій: I) лексичні – іменні, дієслівні та прислівникові; II) предикативні – предметно- та безособово-предикативні; III) граматичні – ядрові й ад'юнктні прийменникові (див. рис. 1). Відповідно до максимально виражених сполучувальних потенцій найбільш розгалуженими є ряди іменних і дієслівних колокацій, які реалізуються в певних морфолого-синтаксичних дистрибутивних моделях (NNr – *надання згоди 19*, VNz – *регулювати здійснення 22*, PRVinf – *повинен перевищувати 22*). Під час установлення дистрибутивних моделей колокацій у протиріччя вступають традиційно визначені системні та текстові функціональні ознаки частин мови.

У проблемних випадках визначення частиномовної належності здійснюється за словником, а морфологічних характеристик – за контекстом: *керуючий санацією* – AVNo, де AV – дієприкметник теперішнього часу, No – іменник у формі орудного відмінка. За результатами класифікації встановлено дистрибутивні моделі граматичних колокацій, відсутні в наявних систематиках [6, с. 177–180; 14, p. XIX–XXX; 18, p. 399; 22]: APrp (*необхідний для 331*, *хворий на 134*) і ADPrp (*незалежно від 207, відповідно до 2361*). До основних проблем класифікації українських колокацій зараховуємо особливості функціонування прикметників, числівників і дієприкмет-

ників у тексті та структурну неоднозначність дієслівних та іменних моделей. Так, прикметник *хворий* у більшості колокацій функціонує як іменник (*лікування хворих 24, такий хворий 11*), що значно розширює сполучувальні можливості (*інфекційно хворий 20*). Подібно до цього дієприкметник функціонує як прикметник (*рекомендований лист 22*, *спеціалізована установа 29*), присудкове слово (*повинен бути 311, повинен відповідати 60*) або іменник (*керуючий санацією 14*, *ураховуючи викладене 10*). В ад'юнктних моделях числівники часто вживаються в значенні прикметників (*третя особа 81*, *перша черга 15*, *перша інстанція 12*), і це не фіксується словником. Критерієм класифікації при цьому може слугувати пре- і постпозиція колоката.

Українські колокації відрізняються можливістю як препозитивного (*спеціально обладнаний 19*, *раніше отриманий 10*), так і постпозитивного приєднання колокатів (*згаданий вище 24*, *засвідчений нотаріально 72*). При цьому позиційне розташування компонентів не відіграє значної ролі для встановлення ядра (на відміну від англійських конструкцій) [12, с. 164–165], хоча в українських колокаціях спостерігаються певні преференції щодо приєднання колоката: дієприкметникова модель із препозитивним прислівником – 14 колокацій, а з постпозитивним – 6. В окремих найменуваннях, термінологічних сполученнях (*порядок денний 117*, *засіб вимірювальний 10*) й іменниково-числівникових моделях частіше спостерігається постпозитивне приєднання колокатів (*частина третя 109, абзац третій 76*), яке можна вважати особливістю функціонального стилю законодавчих текстів. Значні проблеми класифікації українських колокацій виникають у зв'язку з омонімією граматичних форм іменника в сполученнях із дієсловами (*забезпечити можливість/можливістю, запитувана держава/державою*) і віддієслівними іменниками (*відповідність вимог/вимогам*, *звітність товариству/товариства*). У таких випадках розв'язання граматичної омонімії здійснюється через аналіз безпосереднього контексту вживання. На підставі здійсненої класифікації укладені корпусні словники українських предикативних (http://www.mova.info/Page.aspx?l1=208) і граматичних колокацій (http://www.mova.info/Page.aspx?l1=66).

**Висновки і пропозиції.** Не претендуючи на повний опис українських колокацій, ми усвідомлюємо, що для успішного вирішення цього завдання необхідно створити теоретичні передумови та корпуси з відповідним програмно-лінгвістичним забезпеченням. Представлена систематика не є ідеальною, оскільки визначення усталених класів колокацій утруднюється взаємодією граматичних категорій складників. Частково це пояснюється недосконалістю традиційного підходу, за яким не враховуються особливості функціонування частиномовних категорій в автентичному тексті. Визначені системні й функціональні ознаки українських колокацій, а також установлені на їх основі класифікаційні ряди й класи можуть слугувати підґрунтям для морфолого-синтаксичної класифікації.

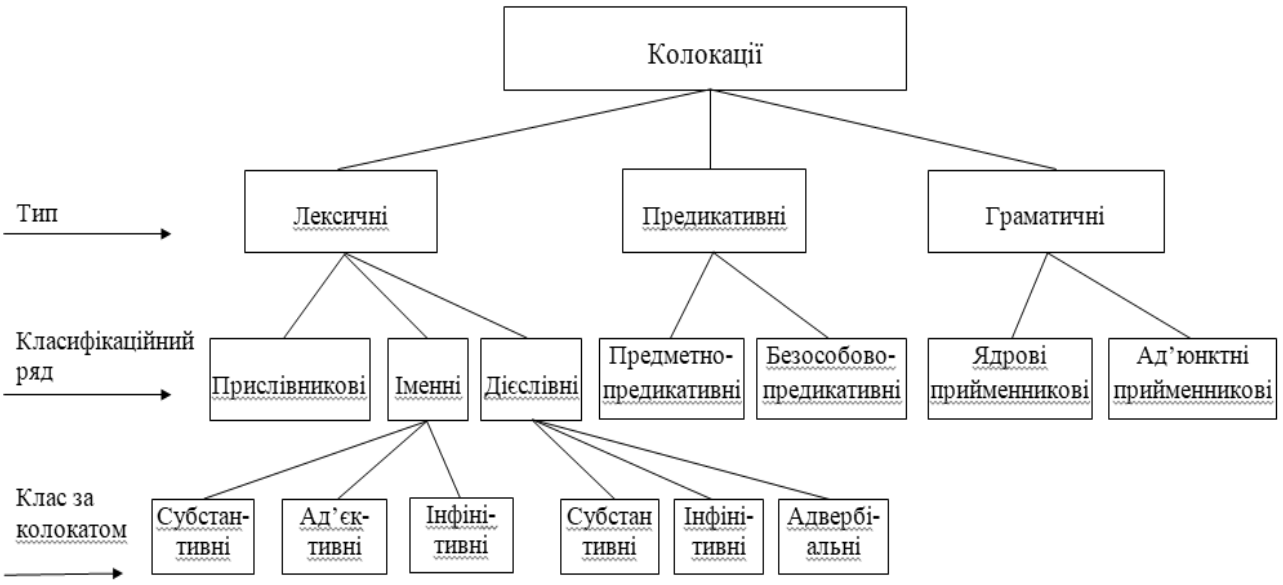


**Рис. 1. Морфолого-синтаксична класифікація колокацій**

**Список літератури:**

1. Бобкова Т. Класифікація колокацій українського юридичного дискурсу: попередні результати. Вісник Київського національного лінгвістичного університету. 2015. Т. 18. № 1. С. 7–15.
2. Бобкова Т. Методика извлечения коллокаций из корпуса украинских текстов. Kalbų Studijos. 2015. NR. 27. С. 93–105.
3. Бобкова Т. Классификация коллокаций: основные подходы и критерии. Respectus philologicus. 2016. № 29 (34). С. 87–98.
4. Гладка В. Структурно-синтаксичний підхід у вивченні колокацій (на матеріалі французької мови). Наукові записки Національного університету «Острозька академія». Серія «Філологічна». 2013. Вип. 39. С. 16–20.
5. Грязнухіна Т., Бугаков О., Любченко Т., Шкурко В. Текстові колокації в лексикографічній системі словника української мови. Українська лексикографія в загальнослов'янському контексті: теорія, практика, типологія. Київ, 2011. С. 336–345.
6. Дарчук Н. Комп'ютерне анотування українського тексту: результати і перспективи: монографія. Київ: Освіта України, 2013. 544 с.
7. Зацеркляний М., Узлов Д. Об'єктно-орієнтований тезаурус і словник колокацій для бази знань криміналістичних інформаційних систем. Системи обробки інформації. 2013. № 2. С. 183–186.
8. Романюк А., Кваснюк Г., Романишин М. Розпізнавання багатослівних конструкцій. Вісник Національного університету «Львівська політехніка». 2011. № 711. С. 158–165.
9. Хайрова Н., Узлов Д. Идентификация криминально значимых коллокаций в украиноязычных текстах. Збірник наукових праць Військового інституту Київського національного університету імені Тараса Шевченка. 2013. Вип. 44. С. 147–151.
10. Хохлова М. Исследование лексико-синтаксической сочетаемости в русском языке с помощью статистических методов (на базе корпусов текстов): автореф. дис. ... канд. филол. наук: 10.02.21. Санкт Петербург, 2010, 26 с.
11. Шкурко В. Лексикографічний агент екстракції колокацій у природно-мовному тексті. Вісник Київського національного університету імені Тараса Шевченка. 2012. № 28. С. 31–35.
12. Шутикова А. Синтаксическая неоднозначность в английских многокомпонентных именных группах. Прикладная лингвистика в науке и образовании: матер. III Междунар. науч. конф. (Санкт Петербург, 16–17 марта 2006 г.). Санкт Петербург, 2006. С. 161–166.
13. Ягунова Е., Пивоварова Л. От коллокаций к конструкциям. Труды Института лингвистических исследований РАН. 2011. URL: http://www.webground.su/datalit/pivovarova_yagunova/Ot_kollokatsiy_k_konstruktsiya.pdf (дата звернення: 1.08.2018).
14. Benson M., Benson E., Ilson R. Your guide to Collocations and Grammar. The BBI Combinatory Dictionary of English. Amsterdam, 2009. P. XIX–XXX.
15. Fontenelle Th. What on earth are collocations? English Today: the International Review of the English Language. 1994. Vol. 10 (40). No. 4. P. 42–48.
16. Durrant Ph., Doherty A. Are high-frequency collocations psychologically real? Corpus Linguistics and Linguistic Theory. 2010. Vol. 6. No 2. P. 125–155.
17. Handl S. Essential collocations for learners of English. Phraseology in Foreign Language Learning and Teaching / ed F. Meunier; S. Granger. Amsterdam, 2008. P. 43–66.
18. Hausmann F. Wortschatzlernen ist Kollokationslernen. Praxis des neusprachlichen Unterrichts. 1984. Vol. 31. P. 395–406.
19. Kimmes A., Koopman H. Collocation-Analyzer: an Electronic Tool for Collocation Retrieval and Verification. URL: http://www.t21n.com/homepage/ articles/T21N-2010-11-immes, Koopman. pdf (дата звернення: 1.08.2018).
20. Marcinkevičienė R. Lietuvių kalbos kolokacijos: monograija. Kaunas: Vytauto Didžiojo universiteto leidykla, 2010. 212 p.
21. Nation P. Learning vocabulary in another language. Cambridge, 2001. 477 p.
22. Newmark P. Approaches to Translation. London-New York-Toronto, 1988. P. 114–115.
23. Seretan V. Syntax-Based Collocation Extraction. Berlin, 2011. 232 p.
24. Sinclair J. Corpus, concordance, collocation (describing English language). Oxford, 1991. 200 p.
25. Smadja F. Retrieving collocations from Text: XTRACT. Computational Linguistics. 1993. No 19 (1). P. 143–177.
26. Tognini-Bonelli E. Corpus Classroom Currency. Darbai ir Dienos. 2000. No 24. P. 205–243.

**КЛАССИФИКАЦИЯ УКРАИНСКИХ КОЛЛОКАЦИЙ НА МАТЕРИАЛЕ КОРПУСА ТЕКСТОВ**

*В статье обосновываются принципы классификации коллокаций. Предлагается бинарная типология подходов к описанию системных и функциональных характеристик. Анализируются основные проблемы классификации украинских коллокаций. Установлены классификационные ряды и классы двухсловных лексических, грамматических и предикативных коллокаций. Определены морфолого-синтаксические особенности употребления коллокаций в законодательных текстах.*

***Ключевые слова:*** *коллокация, корпус, классификация, лексические коллокации, предикативные коллокации, грамматические коллокации.*

**CORPUS-BASED CLASSIFICATION OF UKRAINIAN COLLOCATIONS**

*The article deals with the principles of Ukrainian collocation classification. Binary topology of the approaches to the description of collocation system and functional features is suggested. The main problems of morphological-syntactical classification of Ukrainian fixed phrases are defined. The taxonomy of two-word lexical, grammatical and predicative collocations is determined. The morphological-syntactical peculiarities of collocation use at the Ukrainian Law Acts are defined.*

***Key words:*** *collocation, corpus, classification, lexical collocation, grammatical collocation, predicative collocation.*