Проект корпусного словника українських колокацій

Т.В. Бобкова*

Київський національний лінгвістичний університет, м. Київ, Україна
*Corresponding author. E-mail: tatva93@gmail.com

Paper received 14.07.15; Accepted for publication 31.07.15.

Анотація. Презентована методика виділення колокацій ґрунтується на застосуванні статистичних критеріїв і Корпусу українських текстів як лексикографічного джерела й програмного інструментарію. Аналізуються результати автоматичного вилучення реєстру колокацій з підкорпусу законодавчих текстів із застосуванням програм лематизації і постредагування в ручному режимі. Встановлені колокації репрезентують предметну галузь права і можуть слугувати основою словника українських колокацій.

Ключові слова: *колокація, текст, корпус текстів, словник колокацій, реєстр*

На сучасному етапі розвитку лінгвістики укладання словників ґрунтується на опрацюванні значних обсягів природно-мовного матеріалу корпусів текстів. Доступність текстового матеріалу і переваги його опрацювання корпусним інструментарієм докорінно змінили обличчя сучасної лексикографії [4, с. 281]. Зокрема, результати аналізу емпіричних даних корпусів свідчать, що при побудові тексту значну роль відіграють не традиційно визначені ізольовані лексичні одиниці, а регулярно відтворювані синтагматичні структури – готові фрази [14, с. 60]. Подібно до простих слів зазначені структури зберігаються в пам'яті мовців як цілісні утворення, і є важливим джерелом інформації як про мову, так і про навколишній світ [11, с. 15]. Значний інтерес насамперед становить опис високочастотних стійких синтагматичних послідовностей – колокацій [8; 9; 11; 15; 16]. Це пояснюється затребуваністю інформації про стійкі сполуки для вирішення різних прикладних завдань: укладання словників, тезаурусів, баз даних, автоматичного аналізу тексту, машинного перекладу, інформаційно-пошукових систем та ін. Зокрема, потреби розв'язання завдань автоматичного аналізу українського тексту сприяли появі низки досліджень, присвячених розробці лексикографічних систем і баз даних колокацій [2; 3; 6; 7].

Основну проблему формування реєстру спеціалізованого словника становить визначення критеріїв ідентифікації колокацій, що зумовлюється використовуваним у дослідженні підходом [1, с. 18; 4, с. 304; 5, с. 158; 13, p. 152; 15, p. 258; 16, p. 341-342]. Загалом традиційний – лінгвістичний підхід до відбору стійких сполук спирається на широке розуміння колокації і використанні наявних лексикографічних джерел: колокація розуміється як звичне використання певних лексем разом [14, p. 115–116]. У методичному відношенні цей підхід реалізується за допомогою методики закритого списку [8]: критерієм включення до реєстру є відповідність певної сполуки заздалегідь укладеному списку неоднослівних єдностей, встановлених за словником. Визнаються лише лексичні колокації – поєднання двох повнозначних слів, що насамперед сприяє дослідженню певних предметних галузей і терміносистем [13, p. 152]. В українській лексикографії на підставі закритого списку реалізовано пошук контекстів у тримовному тлумачному словнику термінів і терміносполучень з комп'ютерної лінгвістики [10, p. 39] і формування реєстру електронного словника фразеологізмів у системі автоматичного граматичного аналізу тексту АГАТ [2, с. 174]. Однак, укладання списку в такий спосіб обмежується набором колокацій, встановленим за наявними словниками, базами даних і певною предметною галуззю.

Більш складним у методичному відношенні є формування реєстру словника на підставі напівзакритого списку заздалегідь встановлених ознак: лексичних функцій, варіантів колокацій або моделей керування [8]. Так, на основі набору морфолого-синтаксичних моделей базується ідентифікація субстантивно-ад'єктивних і субстантивно-субстантивних колокацій в українських кримінально значимих текстах [6, с. 147] і розробка лексикографічної системи екстракції субстантивних, ад'єктивних і вербальних колокацій із законодавчих текстів [7, с. 32-33]. Вважається, що власне методика ідентифікації за структурними моделями відповідає завданням лексикографічного опису колокацій [9, p. 40]. Проте суб'єктивність критеріїв ідентифікації, обмежених традиціями певної школи і інтуїцією дослідника [8], значно звужує діапазон досліджуваних одиниць, зокрема можливості аналізу морфологічних і функціональних ознак. Перспективи встановлення об'єктивних диференційних ознак і автоматизації укладання реєстру словника з'являються з впровадженням корпусного, або статистичного підходу.

Корпусний підхід до виділення колокацій ґрунтується на вузькому розумінні колокації як сполучення слів, що зустрічаються у тексті частіше поруч, ніж за випадковою вірогідністю окремо [14, p. 115-116]. Укладання корпусного словника колокацій спирається на методику відкритого списку, яка передбачає безпосереднє спостереження мовних фактів [8]. На відміну від описаних вище методик вихідними для формування реєстру колокацій стають ознаки поведінки лексичної одиниці, встановлені за даними корпусу текстів [1, с. 16]. Певною мірою застосування корпусного підходу передбачає заміну покладених в основу ідентифікації лінгвістичних моделей математичними [6] або статистичними [3]. Так, на підставі статистичних даних про частоту, вагу та унікальність терміносполучень у колекціях корпусів укладено словник колокацій для криміналістичної інформаційної системи [3, с. 184-186]. Істотними перевагами методики відкритого списку вважаються відсутність заздалегідь заданих обмежень, покриття всього текстового матеріалу [8], можливість дослідження варіативності морфологічних ознак [13, p. 151-152] і виявлення колокацій, не зафіксованих наявними списками і словниками [4, с. 303].

Однак проблема полягає у неможливості вилучення всього діапазону колокацій [15, p. 258] і значному зростанні відсотку повторів і помилок ідентифікації, що потребує додаткового постредагування реєстру в ручному режимі. Крім того, застосовувані сьогодні статистичні міри не здатні до диференціації ядра та колоката, необхідної для формування словникової статті [9, p. 37]. Забезпечення об'єктивного й ґрунтовного лексикографічного опису колокацій вимагає доповнення статистичних результатів аналізом морфолого-синтаксичних ознак. Першою спробою всебічного опису колокацій в українській лінгвістиці можна вважати проект лексикону багатослівних сполук [5], у якому визначення "лінгвістичної правильності" статистично значущих багатослівних конструкцій планується здійснити за допомогою морфолого-синтаксичних фільтрів, а семантичної некомпозиційності – через латентне семантичне індексування [5, с. 163]. Отже, результати короткого огляду представлених словників і систем показують, що загалом увагу українських дослідників зосереджено не на розв'язанні проблем лексикографічного аналізу колокацій таких, як встановлення критеріїв ідентифікації, класифікація, а на прикладному застосуванні результатів, зокрема на ефективності автоматичного розпізнавання стійких сполук в природно-мовному тексті.

Презентована стаття продовжує цикл публікацій, присвячених лексикографічному опису колокацій, і має на меті обґрунтування критеріїв автоматичного виділення з корпусу текстів стійких сполук для формування реєстру словника. Як було зазначено вище, впровадження корпусного аналізу колокацій обмежується через недосконалість сучасних лінгвістичних процесорів, що вимагає обов'язкового постредагування реєстру. На сьогодні статистично-пошуковий апарат Корпусу текстів української мови не дозволяє автоматично укласти списки колокацій з текстів: теоретично можливим є здійснення пошуку окремих сполучень через задавання леми й відповідних граматичних обмежень [12]. Лексикографічним матеріалом для укладання корпусного словника українських колокацій слугують тексти підкорпусів публіцистики, наукового стилю й законодавчих текстів корпусу (http://www.mova.info/corpus.aspx?l1=209). Цей вибір пояснюється максимальною насиченістю в текстах підкорпусів стійких сполук різних видів: багатослівних найменувань, термінологічних сполучень, мовних кліше, вставних конструкцій, похідних прийменників і сполучників. Власне планування підкорпусу законодавчих текстів було підпорядковано меті укладання словника колокацій [1, с. 21], що дозволяє розглядати отримані на підставі підкорпусу емпіричні дані як основу реєстру майбутнього словника українських колокацій. Крім відповідних терміносполучень і найменувань юридичної підмови в очікуваний реєстр мають ввійти колокації, характерні також і для сучасної української літературної мови загалом.

Відсутність закритих списків колокацій і потреби автоматичного аналізу текстів різних стилів зумовлюють використання методики відкритого списку, яка ґрунтується на підході F. Smadja і передбачає поетапне застосування статистичного аналізу, лінгвістичних фільтрів (позиційного, синтаксичного й морфологічного) і завершального редагування з боку експерта-лінгвіста [15, с. 253]. Оскільки автоматичне укладання списку колокацій обмежено функціональними можливостями Корпусу української мови, то вилучення колокацій з законодавчих текстів на підставі статистичних даних сполук здійснюється через процедуру запитів мовою SQL (автор програмного забезпечення – В.М. Сорокін). В основу формування попереднього списку покладено статистичну інтерпретацію, базовану на формальному визначенні колокації як сполучення слів, зафіксованого в текстах принаймні двічі. Враховуючи найбільшу частоту в текстах різних функціональних стилів за основну одиницю аналізу обрано двослівне контактно розташоване сполучення. Відповідно до цього розглядаються як лексичні колокації – сполуки повнозначних слів (*договірна сторона*), так і граматичні – сполуки повнозначного й службового слова (*визначений в, у разі, що має*). Для перевірки достовірності статистичних даних і відбору до реєстру словника лексикографічно релевантних колокацій застосовано асимптотичну гіпотезу. Внаслідок автоматичного опрацювання вибірки законодавчих текстів обсягом в 1 млн. слововживань і лематизації компонентів укладено список, який нараховує 64.361 сполуку. Отримані результати укладання списку колокацій законодавчих текстів планується верифікувати на матеріалі відповідних за обсягом вибірок підкорпусів публіцистики і наукових текстів.

З метою включення до реєстру найчастотніших одиниць поріг частоти вживання колокацій обмежено до 10 разів на 1 млн., і для подальшого аналізу відібрано 9537 сполучень, ранжованих за частотою. Отже, для відбору стійкої сполуки до реєстру словника застосовано такі критерії: 1) колокацією є двослівне контактно розташоване сполучення; 2) кожна сполука має зустрітись принаймні 10 разів на 1 млн.; 3) ядром колокації може бути як повнозначне, так і службове слово (*час проведення, під час*); 4) кожна сполука має належати до безпосередніх складників (*виплата страховий – страховий поліс*). Відповідність встановленим критеріям потребує обов'язкового редагування автоматично укладеного реєстру у зв'язку з необхідністю: а) вилучення складових багатослівних сполук, б) відбору граматично правильних сполучень, в) зняття неразв'язаної омонімії, г) видалення повторень і д) коригування статистичних показників. Зокрема, підлягають вилученню з реєстру сполуки з сурядним зв'язком, відкритий характер якого передбачає наявність принаймні трьох компонентів (*акт та…*), а також аналітичні форми слів (*більш пізній, не перебувати*). Встановлено, що середній відсоток помилок ідентифікації колокацій програмним шляхом становить 22,4 %: у результаті постредагування список скоротився до 7393 сполук. Істотні проблеми автоматичного опрацювання українських текстів становлять: ідентифікація багатослівних сполучень, зняття граматичної і лексико-граматичної омонімії та аналіз сумісно вживаних слів, не пов'язаних між собою підрядним зв'язком. Застосування корпусного підходу дозволяє встановити розподіл частот безпосередньо спостережуваних у законодавчих текстах колокацій (див. Табл. 1).

Таблиця. 1. Розподіл частот колокацій

№	Кількість колокацій	Частота	Граматичні колокаї		Лексичні колокації
			предикативні	граматичні	
1	1	2361		1	
2	5	1999–1000		2	3
3	18	999–500		9	9
4	44	499–300	1	20	23
5	76	299–200	8	39	29
6	262	199–100	12	133	117
7	638	99–50	29	283	326
8	954	49–30	63	393	498
9	1252	29–20	84	489	679
9	4143	19–10	254	1470	2419
УСЬОГО:	7393		451	2839	4103

Аналіз розподілу частот показує, що статистична структура словника колокацій зумовлюється дією закону Ципфа: незначна кількість вилучених сполук (0,1 %) вживається в текстах з максимальною частотою (більше 1000 разів), а більше половини (56 %) – з частотою у 10 разів. Це цілком відповідає закономірностям розподілу частот ізольованих слів [16, p. 345]. Максимальний відсоток вилучених сполук становлять лексичні (55,5 %), більше третини – граматичні (38,4 %) і незначний відсоток – предикативні колокації (6,1 %). Істотне переважання в реєстрі лексичних колокацій пояснюється насамперед характеристиками досліджуваного матеріалу, і може слугувати статистичним параметром офіційного стилю. До закономірностей функціонування колокацій в українських законодавчих текстах також слід віднести зростання кількості лексичних колокацій із зменшенням показника частоти їх вживання (від 99 разів і менше). Гіпотетично можна припустити, що високочастотні колокації (крім терміносполучень і найменувань) є характерними власне для загального словника, а менш частотні – для певного функціонального стилю або предметної області. Однак, це припущення вимагає верифікації на матеріалі текстів інших функціональних стилів.

Різнотипність спостережуваних колокацій в автоматично укладеному реєстрі висуває вимоги до формування різних стандартів словникової статті. В основу відбору і формування словникової статті насамперед слід покласти морфологічні ознаки колокацій, зокрема щодо здатності утворювати форми [16, p. 341]. Апробована методика автоматичного укладання реєстру через лематизацію потребує певного доопрацювання, зокрема в аспекті аналізу уживаності форм компонентів колокацій (*в силі – в силу; цінні папери*). Очевидно, мають бути розроблені стандарти словникової статті для незмінних сполук (*відповідно до, слід мати*), для сполучень незмінного й змінного компонентів (*згідно пункту, щодо змін*) і для сполук змінних компонентів (*окремий вид, сплата внесок*). Результати статистичного аналізу змінних форм компонентів можуть бути представлені у вигляді дерева колокації. Встановлені критерії ідентифікації, а також отримані списки колокацій можуть бути використані для укладання словників різних предметних галузей і розв'язання прикладних завдань з автоматичного аналізу українського тексту: морфолого-синтаксичної розмітки, зняття омонімії і машинного навчання систем розпізнавання колокацій.

ЛІТЕРАТУРА

[1] Бобкова Т.В. Лексикографический корпус как источник для словарей нового типа // Žmogus kalbos erdvėje, 2015. Nr 8. – 15-26 с.

[2] Дарчук Н. Комп'ютерне анотування українського тексту: результати і перспективи // К. : Освіта України, 2013. – 544 с.

[3] Зацеркляний М.М., Узлов Д.Ю. Об'єктно-орієнтований тезаурус і словник колокацій для бази знань криміналістичних інформаційних систем // Системи обробки інформації, 2013. № 2. – 183-186 с.

[4] Лендау С.І. Словники : мистецтво та ремесло лексикографії // К. : К.І.С., 2012. – 480 с.

[5] Романюк А., Кваснюк Г., Романишин М. Розпізнавання багатослівних конструкцій // Вісник Нац. ун-ту "Львівська політехніка". Комп'ютерні системи проектування. Теорія і практика, 2011. № 711. – 158-165 с.

[6] Хайрова Н.Ф., Узлов Д.Ю. Идентификация криминально значимых коллокаций в украиноязычных текстах // Зб. наук. пр. Військового ін-ту Київського нац. ун-ту ім. Т. Шевченка 2013, № 44. – 147-151 с.

[7] Шкурко В.В. Лексикографічний агент екстракції колокацій у природномовному тексті // Вісник Київського нац. ун-ту ім. Т. Шевченка. Серія: Літературознавство. Мовознавство. Фольклористика, 2012, № 28. – 31-35 с.

[8] Ягунова Е.В., Пивоварова Л.М. От коллокаций к конструкциям // Acta linguistica petropolitana. Труды Института лингвистических исследований РАН, 2011. – URL: http://www.webground.su/data/lit/pivovarovayagunova/Ot_kollokatsiy_k_konstruktsiya.pdf

[9] Bartsch, S. Structural and Functional Properties of Collocations in English: A Corpus Study of Lexical and Pragmatic Constraints on Lexical Co-occurrence // Gunter Narr Verlag, 2004. – 244 p.

[10] Bobkova, T. etc. Corpus of computational linguistic texts // Computer treatment of Slavic and east European languages. – Bratislava : Tribun, 2009. – 35-40 p.

[11] Evert S. The Statistics of Word Cooccurrences Word Pairs and Collocations. Ph.D. thesis // Stuttgart, 2005. – URL: http://elib.uni-stuttgart.de/opus/volltexte/2005/%202371/pdf/Evert2005phd.pdf

[12] Kotsyba, N. Praktyczny przewodnik po korpusach języka ukraińskiego // Praktyczny przewodnik po korpusach języków słowiańskich. – Warsaw, 2013. – URL: http://www.domeczek.pl/~natko/papers/ przewodnik-korp-ukr2013.pdf

[13] Seljan, S. Gašpar, A. First Steps in Term and Collocation Extraction from English-Croatian Corpus // Computational Language Analysis, Computer-Assisted Translation and e-Language Learning. – Zagreb : Zavod za informacijske studije, 2012. – P. 149-156.

[14] Sinclair, J. Corpus, Concordance, Collocation. – Oxford : Oxford University Press, 1991. – 200 p.

[15] Smadja, F.A. McKeown K.R. Automatically Extracting and Representing Collocations for Language Generation // Proceedings on the 28-th Annual Meeting of the ACL. – Pittsburg : PA, 1990. – P. 252-259.

[16] Shin, D., Nation P. Beyond single words: the most frequent collocations in spoken English // ELT Journal. – 2008. – Vol. 62. – N 4. – P. 339-348.

REFERENCES

[1] Bobkova, T. Lexicographical corpus as the source for new dictionary compiling // Žmogus kalbos erdvėje, 2015. N. 8. – P. 15-26.

[2] Darchuk, N. Computational Annotation of Ukrainian Text: Results and Prospects // K. : Education of Ukraine, 2013. – 544p.

[3] Zacerklyaniy, M.M. Uzlov, D.Y. Object-oriented thesaurus and collocations dictionary for knowledge base forensic information systems // Sistemi obrobki informacii, 2013. № 2. – P. 183-186.

[4] Landau, S.I. Dictionaries: the Art and Craft of Lexicography // K.: K.I.S., 2012. – 480 p.

[5] Romaniuk, A., Kvasniuk, G., Romanyshyn M. Multiword expressions identification // Bulletine of National Universit "Lvivska politechnika". Computer systems of design. Theory and practice, 2011. № 711. – P. 158-165.

[6] Hayrova, N., Uzlov, D. Identification of criminal significant collocation in the Ukrainian Texts // Coll. Sc. Papers of Military Institute of the Kyiv National Univ. n.a. Taras Shevchenko, 2013. № 44. – P. 147-151.

[7] Shkurko, V.V. Lexicographical agent of extraction of collocations in natural language text // Bulletin of Kyiv National Univ. n.a. Taras Shevchenko. Series: Literature. Linguistic. Folklore, 2012. № 28. – P. 31-35.

[8] Jagunova, E.V., Pivovarova, L.M. From collocations to constructions // Acta linguistica petropolitana. Proceedings of the Institute of Linguistic Studies, RAS, 2011. – URL: http://www.webground.su/data/lit/pivovarovayagunova/Ot_kollokatsiy_k_konstruktsiya.pdf

[9] Bartsch, S. Structural and Functional Properties of Collocations in English: A Corpus Study of Lexical and Pragmatic Constraints on Lexical Co-occurrence // Gunter Narr Verlag, 2004. – 244 p.

[10] Bobkova, T. etc. Corpus of computational linguistic texts // Computer treatment of Slavic and east European languages. – Bratislava : Tribun, 2009. – 35-40 p.

[11] Evert, S. The Statistics of Word Cooccurrences Word Pairs and Collocations. Ph.D. thesis // Stuttgart, 2005. – URL: http://elib.uni-stuttgart.de/opus/volltexte/2005/%202371/pdf/Evert2005phd.pdf

[12] Kotsyba, N. A Practical Guide to the bodies of the Ukrainian language // A Practical Guide to the bodies of Slavic languages. – Warsaw, 2013. – URL: http://www.domeczek.pl/~natko/papers/przewodnik-korp-ukr2013.pdf

[13] Seljan, S. Gašpar, A. First Steps in Term and Collocation Extraction from English-Croatian Corpus // Computational Language Analysis, Computer-Assisted Translation and e-Language Learning. – Zagreb : Department of Information Studies, 2012. – P. 149-156.

[14] Sinclair, J. Corpus, Concordance, Collocation. – Oxford : Oxford University Press, 1991. – 200 p.

[15] Smadja, F.A. McKeown K.R. Automatically Extracting and Representing Collocations for Language Generation // Proceedings on the 28-th Annual Meeting of the ACL. – Pittsburg : PA, 1990. – P. 252-259.

[16] Shin, D., Nation P. Beyond single words: the most frequent collocations in spoken English // ELT Journal. – 2008. – Vol. 62. – N 4. – P. 339-348.

Project of Corpus Ukrainian Collocation Dictionary

T.V. Bobkova

Abstract. The method for collocation extraction is based on the applied statistical criteria and relies on Corpus of Ukrainian Language Texts as a lexicographical source and programming instrument for collocation identification. Statistically extracted collocation list of Law Acts Subcorpusi is filtered by linguistic engineering tool, though human post-editing is required. Extracted collocations tend to cover legislative domain and could serve as a base to Ukrainian collocation dictionary.

***Keywords*:** *collocation, text, corpus of texts, collocation dictionary, wordlist*