УДК 81’3.161.2
DOI https://doi.org/10.32782/2522-4077-2024-210-31

ЛІНГВІСТИЧНИЙ КОРПУС ЯК ІНСТРУМЕНТ СЕМАНТИЧНИХ ДОСЛІДЖЕНЬ

LINGUISTIC CORPUS AS A TOOL OF SEMANTIC RESEARCH

Олександрук І.В.,
orcid.org/0000-0003-2701-1714
доктор філософії,
молодший науковий співробітник відділу загального мовознавства
Інституту мовознавства імені О.О. Потебні НАН України

У статті розглянуто застосування лінгвістичного корпусу для аналізу лексичної сполучуваності відносних прикметників з іменниками у межах розроблення мовно-інформаційного інструментарію багатопараметричного опису лексичної семантики відносних прикметників української мови в електронному тлумачному словнику.

Мета дослідження: продемонструвати можливість та доцільність застосування корпусу українських текстів у сучасних семантичних дослідженнях, зокрема для аналізу лексичної сполучуваності відносних прикметників з іменниками.

Здійснено аналіз лексичної сполучуваності відносних прикметників з іменниками на матеріалі Українського національного лінгвістичного корпусу. За результатами аналізу виділено типи семантичних класів іменників, синтаксично пов’язаних із відносними прикметниками як інструмент введення ознаки семантичної валентності для опису лексичного значення відносних прикметників, тлумачення яких передано за відсильними неексплiцитними формулами тлумачення «Прикм. до…» та «Стос. до…» в електронній формі тлумачного Словника української мови у 20 томах.

На основі сполучуваності відносних прикметників з іменниками виділено типи найширше представлених та найчастотніших семантичних класів: «людина», «тварина», «рослина», «будівля/заклад/установа/споруда», «речовини та матеріали», «результат творчої/наукової діяльності», «час», «простір», «дія», «транспорт», «меблі», «обладнання/устаткування/прилад/пристрій», «одяг/взуття», «різні матеріальні предмети», «тканина», «посуд», «прикраса», «природні явища», «їжа/напої», «термін».

Результати аналізу входять до лінгвістичної бази даних «Семантика відносних прикметників», розробленої для розв’язання проблеми експліцитного представлення інформації в електронних тлумачних словниках, що сприятиме дальшому впровадженню комп’ютерних технологій у дослідження з лексичної семантики, орієнтованих на створення систем опрацювання природної мови.

Ключові слова: електронний тлумачний словник, корпус текстів української мови, лексикографія, лексична семантика, семантичний клас.

The article examines the use of a linguistic corpus for the analysis of the lexical conjugation of relative adjectives with nouns within the framework of the development of a linguistic and informational toolkit for a multi-parameter description of the lexical semantics of relative adjectives of the Ukrainian language in an electronic explanatory dictionary.

The aim of the study: to demonstrate the possibility and expediency of using the corpus of Ukrainian texts in modern semantic research, in particular for the analysis of the lexical compatibility of relative adjectives with nouns.

An analysis of the lexical compatibility of relative adjectives with nouns was carried out on the material of the Ukrainian National Linguistic Corpus. Based on the results of the analysis, the types of semantic classes of nouns syntactically related to relative adjectives were identified as a tool for introducing the sign of semantic valence to describe the lexical meaning of relative adjectives, the interpretation of which was given according to the strong non-explicit interpretation formulas «Adjective to ...» and «In relation to ...» in the electronic version of the explanatory dictionary of the Ukrainian language in 20 volumes.

On the basis of the compatibility of relative adjectives with nouns, the types of semantic classes are distinguished, the broadest of which are: «man», «animal», «plant», «building/facility/institution/structure», «substances and materials», «result of creative/scientific activity», «time», «space», «action», «transport», «furniture», «equipment/equipment/appliance/device», «clothing/shoes», «miscellaneous material items», «fabric», «dishes», «decoration», «natural phenomena», «food/drinks», «term».

The results of the analysis are included in the linguistic database «Semantics of relative adjectives», developed to solve the problem of explicit presentation of information in electronic explanatory dictionaries, which will contribute to the further implementation of computer technologies in research on lexical semantics, focused on the creation of natural language processing systems.

Key words: electronic explanatory dictionary, corpus texts of the Ukrainian language, lexicography, lexical semantics, semantic class.

Постановка проблеми. У сучасному мовознавстві електронні (цифрові) корпуси текстів слугують як інструментом, так і матеріалом для різних досліджень, зокрема семантичних. Корпуси текстів є потужними мовно-інформаційними системами, що активно використовують для розв'язання широкого кола дослідницьких завдань майже в усіх галузях мовознавства: лексикографії, граматиці, лексикології та семасіології, стилістиці, перекладацьких студіях, прагматиці, соціолінгвістиці, психолінгвістиці, дискурсології, когнітивній лінгвістиці, лінгвістичній варіантології, навчанні та вивченні мови (рідної та / або іноземної), літературознавчих дослідженнях та інших [1, с. 114].

Аналіз останніх досліджень і публікацій. Корпусна лінгвістика перебуває у колі уваги багатьох науковців. Учені аналізують як теоретичні питання корпусної лінгвістики, так і застосування корпусів текстів для розв'язання мовознавчих проблем. Н. П. Дарчук звертає увагу на можливості семантичної розмітки корпусу української мови [2], В. В. Жуковська розглядає лінгвістичний корпус як новітній інформаційно-дослідницький інструментарій сучасного мовознавства, описує проблеми застосування корпусних технологій у навчанні та вивченні іноземної мови [3; 1 та ін.], у працях М. О. Шведової представлено низку досліджень, здійснюваних за допомогою Генерального регіонально анотованого корпусу української мови [4; 5 та ін.]. Застосування Українського національного корпусу для розв'язання різних мовознавчих проблем розглянуто у працях співробітників Українського мовно-інформаційного фонду НАН України, зокрема в монографіях [6; 7] та ін.

Актуальність дослідження зумовлено інтересом до корпусної лінгвістики та застосування лінгвістичних корпусів для досліджень у сучасному мовознавстві. Як зазначає В. В. Жуковська, лінгвістичний корпус – це не просто джерело ілюстративних прикладів, а потужний інструмент, багатофункціональна лінгво-інформаційна система для проведення різнопланових мовознавчих досліджень [1, с. 117].

Мета дослідження – продемонструвати можливість та доцільність застосування лінгвістичного корпусу для аналізу лексичної сполучуваності відносних прикметників з іменниками у межах розроблення мовно-інформаційного інструментарію багатопараметричного опису лексичної семантики відносних прикметників української мови в електронному тлумачному словнику.

Виклад основного матеріалу дослідження. З-поміж основних переваг, які надає лінгвістичний корпус досліднику – обсяги мовного матеріалу, залучені до мовознавчого дослідження, комплексність, оперативність опрацювання, а також можливість прямого доступу до великої кількості лінгвістичних фактів. Лінгвістичний корпус виконує такі основні функції: 1) репрезентації (представлення) даних; 2) маркування (тегування, анотування; розмічання) текстів; 3) експлікації даних [7, с. 89].

Найважливішим поняттям, яке відмежовує корпуси текстів від зібрання електронних текстів, є поняття розмітки. Воно полягає в тому, що текстам корпусу, а також їхнім компонентам призначаються спеціальні мітки (індикатори) різних типів: зовнішні (описують елементи бібліографічного опису: видання, рік, автор тощо); структурні (описують структуру тексту: розділ, абзац, речення тощо); лінгвістичні (описують лексикографічні, граматичні, семантичні, синтаксичні та інші характеристики) [6, с. 13].

Лінгвістичний корпус дає змогу здійснювати дослідження не на окремих прикладах, а на репрезентативному матеріалі. У нашому дослідженні для аналізу лексичної сполучуваності

конструкцій «відносний прикметник + іменник» обрано Український національний лінгвістичний корпус (УНЛК) [8], обсягом понад 200 млн слововживань. УНЛК, розроблений в Українському мовно-інформаційному фонді НАН України, застосовують для мовознавчих досліджень та укладання сучасних словників, насамперед Словника української мови у 20 томах.

На основі аналізу лексичної сполучуваності виділено типи семантичних класів іменників, синтаксично пов'язаних із відносними прикметниками як інструмент введення ознаки семантичної валентності для опису лексичного значення відносних прикметників, тлумачення яких передано нееекспліцитними формулами в електронній формі тлумачного Словника української мови у 20 томах, наприклад: «Прикм. до…» та «Стос. до…» (**БАНАНОВИЙ,** а, е. 1. Прикм. до **бана́н**; **ЖУРНА́ЛЬНИЙ,** а, е. Стос. до журналу) [9]. Результати дослідження входять до лінгвістичної бази даних «Семантика відносних прикметників» [10, с. 123–129], розробленої для розв'язання проблеми експліцитного представлення інформації в електронних словниках, що є однією з основних умов для використання електронного тлумачного словника як ефективного інструментарію семантичних досліджень із застосуванням комп'ютерних технологій.

Експліцитне представлення для електронного тлумачного словника інтерпретується як реалізація формулою тлумачення такої форми представлення в словниковій статті інформації про лексичне значення, яка б давала змогу за допомогою комп'ютерних програм встановлювати семантичні характеристики текстової словоформи та слова зі словника, будувати в автоматизованому режимі різні семантичні класифікації слів, екстрагувати та передавати семантичну інформацію на вхід інших модулів лінгвістичних аналізаторів та ін. Дотримання вимоги експліцитного представлення інформації в електронному тлумачному словнику сприятиме застосуванню комп'ютерних технологій у дослідженнях з лексичної семантики, а також забезпечуватиме виконання словником функції бути автоматизованою інформаційною системою, розрахованою на широке коло користувачів.

Одним із етапів на шляху до розроблення мовно-інформаційного інструментарію багатопараметричного опису лексичної семантики відносних прикметників української мови в електронному тлумачному словнику було дослідження лексичної сполучуваності відносних прикметників з іменниками. Процес встановлення лексичної сполучуваності є надзвичайно складний. У нашому дослідженні, в межах Інтегрованої лексикографічної системи, компонентом якої є Український національний лінгвістичний корпус, вибір контекстів здійснювався автоматично із залученням програми автоматичного морфологічного аналізу тексту та програми лематизації. Застосування УНЛК дало змогу залучити тексти різних стилів для аналізу лексичної сполучуваності відносних прикметників з іменниками, а також отримати статистичні дані на основі великих масивів мовної інформації, що забезпечує достовірність результатів дослідження.

У мовознавчій літературі встановлення лексичної сполучуваності вчені розглядають як надійний метод представлення значення слова. На думку М. П. Кочергана, наведення типових контекстів у тлумачних словниках має важливіше значення, ніж саме тлумачення слова. Як зазначає вчений: «Опис сполучуваності слів для кожного із значень якраз і має кінцевою метою дати перелік слів, що сполучаються з аналізованими лексико-семантичними варіантами слова. Однак простий перелік слів є неекономним, незручним і в багатьох випадках (коли їх список незакритий чи надто об'ємний) нездійснений. Тому доречнішою була б їх класифікація на основі десигнатів (тематична чи семантична класифікація), тобто перелік тільки груп (класів), виділених на основі семантичної спільності» [11, с. 53].

У дослідженні лексичні одиниці з подібними значеннями об'єднано в семантичні класи, де кожен семантичний клас представляє певне лексичне значення слова. Такий принцип для синтезу найбільш загальної форми семантичних станів одиниць лексичного рівня представлено в монографії «Граматичні системи. Феноменологічний підхід»: «…теоретично для експлікації повного комплексу значень конкретного слова необхідно зібрати всі – у певному сенсі – його

контексти, де воно функціонує, розподілити їх за однорідними у певному (семантичному) відношенні групами, кожна з яких і є репрезентантом певного лексичного значення. Далі, вивчаючи ці групи контекстів, лексикограф виводить із кожної такої групи окреме лексичне значення аналізованої лексеми і кваліфікує відповідні граматичні значення» [12, с. 38–39].

Класифікація семантичних класів є відкритою. Ілюстрації подано з індексом та номером значення для багатозначних слів з електронного тлумачного Словника української мови у 20-ти томах. З-поміж виділених семантичних класів у дослідженні найчастіше траплялися такі:

Людина (*чоловік*, *жінка*, *дитина*, *дівчина*, *хлопець* та ін.); група людей (*гурток*, *колектив*, *товариство* та ін.); народи/племена/етнографічні групи (*англійці*, *гуцули*, *українці* та ін.); спільноти людей (*етнос*, *народ*, *плем'я* 1 та ін.); члени родини (*батько* 1, *мати* 1, *дочка* 1 та ін.); зовнішня характеристика людини (*врода* 1, 2, 3, *зріст* 2, *статура* та ін.); частина тіла людини (*вухо* 1, *плече* та ін.); риси характеру людини (*чемність*, *чуйність*, *щирість* та ін.); емоції/почуття (*веселість*, *радість* та ін.); рід діяльності/професія/посада (*вчитель* 1, *водій* 1, *директор* та ін.); інтелектуальна характеристика людини (*розум* 1, *тямущість*, *уміння* та ін.); властивість людини (*сміх*, *гідність*, *голос* та ін.);

Тварина. Хижак (*вовк* 1, *лис* 1, *лев* 1 та ін.); свійська тварина (*кіт*, *кінь* 1, *собака* 1 та ін.); комаха (*бджола*, *джміль*, *жук* 1 та ін.); птах (*ластівка*[1] 1, *лелека*, *соловей* 1 та ін.); риба (*карась*, *кілька*, *щука* та ін.); гризун (*білка* 1, *бобер* 1, шиншила 1 та ін.); група тварин (*зграя* 1, *отара* 1, *табун* 1 та ін.); частина тіла тварини (*дзьоб* 1, *лапа*, *хвіст* 1 та ін.).

Рослина. Дерево (*каштан* 1, *клен*, *яблуня* та ін.); кущ (*калина* 1, *малина* 1, *смородина* 1 та ін.); квітка (*півонія* 2, *ромашка*, *троянда* 2 та ін.); трава (*спориш*, *лаванда*, *чебрець* та ін.); плід рослини (*лимон* 2, *диня* 2, *кавун* 2 та ін.); частина рослини (*гілка* 1, *кора* 1, *листок*[1] та ін.); рослини, засаджені на одній площі (*бір*[1], *ліс* 1, *сад* та ін.); виріб, зроблений з рослин (*кошик* 1, *букет* 1, *стілець* та ін.); лікарська рослина (*звіробій*[2], *кропива*, *шавлія* та ін.); овочі (*буряк*, *картопля*, *морква* та ін.); фрукти (*вишня* 2, *малина* 2, *персик* 2, *черешня* 2 та ін.).

Будівля/заклад/установа/споруда. Заклад культури (*бібліотека* 1, *кінотеатр*, *театр*, *музей* 1 та ін.); навчальний заклад (*гімназія*, *коледж*, *університет* та ін.); підприємство (*завод*[1] 1, *комбінат* 1, *фабрика* 1); заклад громадського харчування (*їдальня* 2, *кафе*, *буфет* 2 та ін.); лікувальний заклад (*лікарня*, *профілакторій* 1, *санаторій* та ін.); торговельна установа (*бутик*, *крамниця*, *універмаг* та ін.); частина установи/підрозділ (*відділ* 2, *бухгалтерія* 2, *дирекція* та ін.); житлова будівля (*будинок* 1, *дача*[2], *хата* 1 та ін.); релігійний заклад (*храм* 1, *монастир* 2 та ін.); частина будівлі (*вікно* 1, *дах*, *кімната* 1 та ін.).

Простір. Ділянка землі (*клумба*, *сад*, *двір*[1] 1 та ін.); водойма (*басейн* 1, 2, *море* 1, 2, *ставок* та ін.); повітряний простір (*небо* 1, *небокрай* та ін.); населений пункт/його частина (*квартал* 2, *село* 1, *вулиця* та ін.); адміністративно-територіальна одиниця (*область*, *округ*, *район* та ін.); місце, лінія, що поділяє простір на частини (*кордон* 1, *лінія* 2, *рубіж* 1 та ін.); місце, з певними кліматичними умовами (*пояс* 4, *край*[1] 5, *місцевість* 2 та ін.).

Час. День тижня (*понеділок*, *вівторок*, *неділя* та ін.); місяць (*січень*, *лютий*, *березень* та ін.); пора року (*зима*, *весна*, *літо*, *осінь*); частина доби (*вечір* 1, *ніч*, *світанок* 1 та ін.); відрізок/проміжок часу (*доба*, *місяць*, *рік* та ін.).

Речовини та матеріали. Горюче (*бензин*, *гас*, *керосин* та ін.); рідина (*дьоготь*, *смола*, *спирт*, *фарба* 1 та ін.); будівельні матеріали (*бетон*, *скло*, *цемент*, *шифер* 2 та ін.); коштовне каміння (*алмаз* 2, *діамант* 1, *рубін* та ін.); корисні копалини (*вугілля*, *золото*, *руда*, *нафта* та ін.); мінеральні утворення (*кремінь*, *камінь* 1 та ін.); тверда речовина (*мінерал*, *сплав*[1] та ін.); хімічна речовина (*азот*, *кислота* 2, *натрій* та ін.).

Результат творчої/наукової діяльності. Літературний твір (*вірш* 1, *казка* 1, *роман* 1 та ін.); музичне мистецтво (*пісня* 1, *симфонія* 1, *соната* та ін.); театральне мистецтво (*балет* 1, *вистава*[1] 1, танець 1 та ін.); кіномистецтво (*кіно*, *фільм* 1, *кінофільм* та ін.); наукова праця (*дисертація*, *монографія*, *стаття* 1 та ін.).

Транспорт. Автомобільний (*автобус*, *автомобіль* та ін.); морський (*корабель* 1, *лайнер* 1, *човен* та ін.); повітряний (*лайнер* 2, *літак*, *вертоліт* та ін.); залізничний (*потяг*, *електричка* та ін.); частина транспортного засобу (*бампер*, *кермо*, *колесо* 1, 3 та ін.).

Меблі (*стіл*, *крісло* 1, комод, *диван*[1], *шафа* 1, *стелаж* та ін.).

Обладнання/устаткування/прилад/пристрій (*дзвінок*, *ліхтар*, *світильник*, *вентилятор*, *антена*, *апарат*, *детектор*, *далекомір*, *диктофон* та ін.).

Одяг/взуття (*сукня*, *сорочка*, *штани* та ін.); верхній одяг (*шуба*, *плащ* та ін.); головний убір (*бриль*, *капелюх* 1, *шапка* та ін.); взуття (*кеди*, *кросівки*, *туфлі* та ін.); частина одягу/взуття (*каблук*, *кишеня* 1, *рукав*, *комір* та ін.).

Дія (*біг*, *крок* 1, *стрибок* 1 та ін.); процес (*дослідження* 1, *виробництво* 2, *будівництво* 1); змагання (*конкурс*, *бій* 3, 4, *перегони* та ін.); спортивна гра (*баскетбол*, *волейбол*, *футбол* та ін.); поїздка з певною метою (*екскурсія* 1, *подорож* 1, *експедиція* 4 та ін.).

Різні матеріальні предмети (*блокнот*, *клавіша*, *відро*, *сідло* та ін).

Тканина (*батист*, *бавовна* 2, *шифон*, *кашемір*, *трикотаж* та ін.).

Термін (*косинус*, *індент* 1, *кліше* 2 та ін.).

Посуд (*бокал* 1, 2, *глечик*, *ложка*, *склянка* 1, *кухоль*, *тарілка*, *таця* та ін.).

Прикраса (*браслет* 1, *кольє*, *перстень*, *кулон*, *сережки*, *намисто* та ін.).

Природні явища (*дощ* 1, *злива* 1, *сніг*, *туман*, *веселка*, *блискавка* та ін.);

Їжа/напої (*борщ*, *бекон*, *лікер*, *печиво*, *хліб*, *варення*, *джем* та ін.).

За результатами аналізу простежуємо, що найширше представленими та найчастотнішими виявилися семантичні класи «Людина» (12%), «Рослина» (11%), «Тварина» (8%), «Будівля/заклад/установа/споруда» (9%), «Речовини та матеріали» (8%), «Простір» (7%), «Час» (6%), дещо меншими «Транспорт» (5%), «Дія» (5 %), «Результат творчої/наукової діяльності» (4%), «Їжа/напої» (3%), «Термін» (2%) і т. д.

Висновки. Результати дослідження дають змогу зробити висновок, що лінгвістичний корпус доцільно застосовувати для широкого кола мовознавчих проблем, зокрема для аналізу лексичної сполучуваності відносних прикметників з іменниками, оскільки корпус текстів – це не просто джерело ілюстрацій, а сучасний інструмент мовознавчих досліджень. Корпусне опрацювання дає змогу здійснити аналіз не лише на окремих прикладах, а на репрезентативному матеріалі, надати відомості про статистичні дані. Дослідження лексичної сполучуваності відносних прикметників з іменниками є одним із важливих завдань на етапі розроблення мовно-інформаційного інструментарію для багатопараметричного опису семантики відносних прикметників в електронному тлумачному словнику.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ:

1. Жуковська В. В. Лінгвістичний корпус як новітній інформаційно-дослідницький інструментарій сучасного мовознавства. *Вчені записки ТНУ імені В. І. Вернадського. Серія: Філологія. Соціальні комунікації*. Том 31 (70). № 3. Ч. 1. 2020. С. 113–119.

2. Дарчук Н. П. Можливості семантичної розмітки корпусу української мови (КУМ). *Науковий часопис Національного педагогічного університету імені М. П. Драгоманова. Серія 9: Сучасні тенденції розвитку мов* : зб. наук. пр. Київ : Вид-во НПУ імені М. П. Драгоманова, 2017. Вип. 15. С. 18–28.

3. Жуковська В. В. Застосування корпусних технологій у навчанні та вивченні іноземної мови. *Актуальні проблеми сучасної лінгвістики та методики викладання мови і літератури (за матеріалами онлайн конференції, проведеної кафедрою міжкультурної комунікації та прикладної лінгвістики Навчально-наукового інституту іноземної філології)*. 2018. Житомир : Вид-во ЖДУ ім. Івана Франка. С. 39–59.

4. Шведова М. О. Генеральний регіонально анотований корпус української мови (ГРАК) як інструмент дослідження лексико-граматичної варіативності. *Людина. Комп'ютер. Комунікація* : зб. наук. пр. Львів : Вид-во Львівської політехніки, 2019. С. 145–148.

5. Шведова М. О. Граматичне освоєння запозичених іменників із кінцевим -о в українській мові: корпусне дослідження. *Українська мова*. 2020. № 2 (74). С. 13–30.

6. Корпусна лінгвістика : монографія / В.А. Широков та ін. Український мовно-інформаційний фонд НАН України. Київ : Довіра, 2005. 472 с.

7. Лінгвістично-інформаційні студії : праці Українського мовно-інформаційного фонду НАН України : у 5 т. / В. А. Широков та ін. Т. 4 : Корпусна та когнітивна лінгвістика. Київ. Український мовно-інформаційний фонд НАН України. 2018. 246 с.

8. УНЛК: Український національний лінгвістичний корпус. URL: https://svc.ulif.org.ua/UNLC/virt_unlc_4.5/

9. Словник української мови onlinec. Томи 1-11 (А-ОБМІЛЬ). URL: https://services.ulif.org.ua/expl/Entry/index?wordid=25129&page=835 (дата звернення 17.06.2024).

10. Олександрук І.В. Моделювання семантики одиниць лексичного рівня в лексикографічній системі тлумачного типу (на матеріалі Словника української мови у 20 томах) : Доктор філософії : спец.. 035 – Філологія. Інститут мовознавства ім. О. О. Потебні Національної академії наук України. Київ. 220 с.

11. Кочерган М. П. Слово і контекст (Лексична сполучуваність і значення слова). Львів : Вища школа, 1980. 184 с.

12. Граматичні системи: феноменологічний підхід / В. А. Широков, Т. П. Любченко, І. В. Шевченко, К. В. Широков. Київ : Наукова думка, 2018. 310 с.