РОЗДІЛ 8
ПОРІВНЯЛЬНО-ІСТОРИЧНЕ І ТИПОЛОГІЧНЕ МОВОЗНАВСТВО

УДК 81'373'37
DOI https://doi.org/10.32782/tps2663-4880/2023.30.45

ІНТЕРКОРПУСНИЙ АНАЛІЗ ЛЕКСИКО-СЕМАНТИЧНИХ ЗВ'ЯЗКІВ У СУЧАСНИХ МОВАХ

INTERCORPUS ANALYSIS OF LEXICAL AND SEMANTIC RELATIONS IN MODERN LANGUAGES

Мельник І.Є.,
orcid.org/0000-0001-5594-4269
кандидат педагогічних наук,
доцент кафедри англійської філології
Прикарпатського національного університету імені Василя Стефаника

Ткаченко Л.М.,
orcid.org/0000-0001-6286-6246
кандидат філологічних наук,
доцент кафедри іноземних мов та міжнародної комунікації
Черкаського державного технологічного університету

Калініченко Т.М.,
orcid.org/0000-0001-7672-5858
кандидат педагогічних наук,
доцент кафедри загального мовознавства і романо-германської філології
Харківського національного педагогічного університету імені Г.С. Сковороди

Корпусний аналіз має давню традицію в лексичній семантиці. Завдяки розвитку комп'ютерів і доступності великих електронних корпусів лексикологи та лексикографи мають у своєму арсеналі велику кількість надійних матеріалів, які формують значну емпіричну базу даних для описових досліджень. Для аналізу великих обсягів даних лексикографи використовують інструменти статистичного аналізу, які полегшують етап корпусного аналізу. Статистичні методи дають змогу виявити контекстуальні підказки, як-от паралельні слова (фрази) та синтаксичні моделі (граматичні вирази та словосполучення), щоб досліджувати лексичні значення з корпусних даних. З іншого боку, ці методи класифікують входження лексем за різними значеннями, які потім можна використовувати відповідно до контекстуальних підказок. В аналізі колокацій статистичний аналіз автоматизує ідентифікацію контекстуальних підказок, тоді як класифікація подій і типів залишається на розсуд інтеркорпусного аналізу. В цій статті представлено основні особливості та найважливіші референції різних підходів і підтипів, а також проілюстровано принципи інтеркорпусного розподільного аналізу на прикладі демомоделі розподілу. Розглянуто також лексичні програми для аналізу полісемії. До того ж, описано метод візуалізації, який дає змогу інтерпретувати семантичні структури, описані семантичною дистрибутивною моделлю. Ця стаття презентує метод дистрибутивного аналізу, який розглядається як логічне продовження статистичних методів, на яких ґрунтується аналіз лексико-семантичних відношень у сучасних мовах. В цій статті описано потенціал розподільного семантичного моделювання для міжкорпусного аналізу. Описано закономірності статистичного інструментарію, який використовується в лексичній семантиці. Інтеркорпусні моделі дають змогу вивчати слова або на рівні типів, або на рівні токенів. Моделі на рівні типів не тільки пояснюють, що одне слово відрізнятиметься від іншого, але й виявляють синонімію та пов'язані з ним лексичні зв'язки.

Ключові слова: корпусне навчання, корпусна лінгвістика, семантика, лексичні закономірності, синтаксичні закономірності.

The corpus analysis has a long tradition in lexical semantics. Thanks to computers and the availability of larger electronic corpora, lexicologists and lexicographers now have large amounts of reliable data at their disposal, which constitute a vast empirical database for their descriptive work. In order to analyse large amounts of data, lexical semantics researchers use statistical analysis tools that facilitate the stages of corpus analysis. Statistical methods allow identifying contextual clues in the corpus data to learn the meaning of a lexeme, such as parallel words (phrases) and syntactic patterns (grammatical phrases or collocations). On the other hand, these methods allow you to classify the occurrence of a token in different meanings and use it according to contextual indices. Collocation analysis has automated the identification of contextual clues through statistical analysis, but it leaves the classification of events and types to intercorpus analysis. The paper presents the main features and most important references for different approaches and subtypes, and explains the principles of intercorpus distributional analysis with an example of a demo distribution model. The lexicological application for polysemy analysis is discussed. Finally, a visualisation method is presented that allows for the interpretation of semantic

structures defined by semantic distribution models. This paper presents methods of distributional analysis, which we consider to be a logical continuation of the statistical state of the art on which the analysis of lexical-semantic relations in modern languages is based. The regularities of statistical tools used in lexical semantics are described. Intercorpus models allow learning words either at the level of types or at the level of tokens. Type-level models not only explain how one word is different from another, but also reveal synonymy and related lexical relations.

Key words: corpus-based learning, corpus linguistics, semantics, lexical regularities, syntactic regularities.

Постановка проблеми. Методи розподільного аналізу систематично поєднують асоціативні заходи та багатовимірні статистичні методи для дослідження лексичних семантичних структур у великих текстових корпусах. В результаті вдається доповнити сформовані закономірності в поданні статистичного інструментарію, який використовуються в лексичній семантиці (табл. 1).

Використання статистичних методів на обох етапах аналізу даних є логічним, навіть необхідним, доповненням з двох причин. По-перше, лексикологи та лексикографи отримують користь від додаткової підтримки статистичних методів виявлення закономірностей. Неможливо вручну кодувати, категоризувати та ідентифікувати тисячі словникових статей у словнику. Аналіз статистичних даних дає змогу мати репрезентативну вибірку різних додатків, яку потім можна детально проаналізувати. По-друге, великі дані та доступність більших обсягів інформації розширюють традиційний фокус лексикографічних та лексикографічних досліджень. Середовище великих даних дає змогу дослідникам вивчати тенденції та закономірності, які могли б залишитися непоміченими під час аналізу менших корпусів, наприклад, розповсюдження нових слів або нових уживань наявних слів у соціальних мережах. Для виявлення таких тенденцій важливо використовувати методи інтеркореляційного аналізу даних, а також відповідні кількісні методи.

Аналіз останніх досліджень і публікацій. Корпусний аналіз має давні традиції в лексичній семантиці, що ґрунтуються на появі великих словникових проєктів у XIX столітті. У XX столітті дослідники лексичної семантики спиралися на текстові підказки в реальних мовах для виявлення й організації різноманітних значень і вживань слів. У 1950-х роках синтаксичні дослідження відійшли від реальних даних, але ідеї Д.Р. Ферта (1957), З. Харріса (1954) і У. Вівера (1955) дали поштовх до застосування підходу, який розглядав реальні дані як природне емпіричне підґрунтя для семантичних ідентифікацій [1, с. 13]. Спочатку корпусні дані збирали й аналізували вручну. Завдяки розвитку комп'ютерів і наявності більших електронних корпусів лексикологи й лексикографи тепер мають у своєму розпорядженні велику кількість надійних даних, які становлять величезну емпіричну базу для описової роботи. Для аналізу цієї великої кількості даних лексикографи використовують інструменти статистичного аналізу, які полегшують етап корпусного аналізу. З одного боку, статистичні методи дають змогу ідентифікувати в корпусних даних контекстуальні підказки для вивчення значення лексеми, такі як співпаралельні слова (словосполучення) і синтаксичні патерни (граматичні словосполучення або коллогіації) [2]. З іншого боку, ці методи уможливлюють класифікацію входження лексеми в різні значення і використовувати відповідно до контекстуальних індексів. Перший підхід, тобто визначення контекстуальних показників завдяки статистичним методам, пов'язаний із традиційною корпусною лінгвістикою англійської мови, розробленою Д. Сінклером (1991), визначав лексичне значення слова відповідно до типового слова-словосполучення або синтаксичного зразка (колокації). К. Черч і П. Хенкс (1989) запровадили статистичні показники, такі як t-індекси, для визначення релевантних та інформативних словосполучень і колокацій за розподілом частотності в тексті [3]. Згодом ці показники були вдосконалені та оптимізовані [4]. Другий підхід – статистичне групування вживання – нині широко використовується в лінгвістиці, його можна побачити в найновіших напрацюваннях у царині когнітивної семантики. Зокрема, підхід поведінкового профілю нещодавно презентував багатовимірні статистичні методи, які автоматично класифікують вживання слів за кількома

Таблиця 1

Огляд статистичних засобів, що використовуються в лексичній семантиці

	Виявлення контекстуальних підказок	**Класифікація подій**
Класична філологія	вибірка	вибірка
Розбір словосполучень	статистика	вибірка
Аналіз поведінкових профілів	вибірка	статистика
Методи розподілу	статистика	статистика

Джерело: власне розроблення авторів

різними значеннями або вживаннями на основі корпусних даних [5, с. 301]. Наприклад, корпусна лінгвістика використовує чинниковий аналіз відповідності для візуалізації того, як вживання дієслів групуються в різні значення відповідно до їхньої синтаксичної поведінки та семантичних властивостей. Ці два підходи використовують незалежно в цих різних традиціях. У колокаційному аналізі ідентифікація контекстуальних підказок автоматизована за допомогою статистичного аналізу, але категоризація подій і типових контекстів залишена для ручного аналізу. Аналіз профілю поведінки, з іншого боку, є статистичною автоматизацією класифікації типових подій і контекстів маркерів. Зрозуміло, що навчання в класичній філології проводилося в ручному режимі. Дослідження колокації, з одного боку, і аналіз поведінкових профілів, з іншого, автоматизували лише один з двох кроків.

Методи інтеркорпусного розподільного аналізу, також відомі як моделі простору слів, спочатку використовувалися в когнітивній психології для моделювання лексичної пам'яті [6, с. 622]. Потім вони були розроблені в Computational Linguistics, де зараз широко використовуються для моделювання семантичного аналізу в галузі обробки природної мови (NLP).

Інтеркорпусний аналіз сьогодні є важливим способом представлення та використання лексики та лексичної семантики [7].

Постановка завдання. Методи інтеркорпусного аналізу ґрунтуються на гіпотезі розподілу про те, що слова, які трапляються в схожих контекстах, як правило, мають схожі значення. Мета цієї статті – розглянути підходи до аналізу зв'язків між значеннями слів та їхніх підтипів із трьох точок зору: типу контексту, рівня аналізу та обчислювальної репрезентації.

Виклад основного матеріалу. Методи і моделі розрізняються в залежності від типу контексту, який використовується для представлення значення слів. Розрізняють моделі на основі документів і моделі на основі слів. У моделях, що спираються на документи, семантичний аналіз слова *x* ґрунтується на спостереженні за вживанням *x* у контексті всього документа. Розрізняють моделі на основі документів і моделі на основі слів. У моделях, що спираються на документи, семантичний аналіз слова *x* ґрунтується на спостереженні за вживанням x у контексті всього документа. Англосаксонські прототипи, такі як Latent Semantic Analysis (LSA) [8], є моделями, основою яких є документи. Моделі, побудовані на словах, беруть за початок виникнення *x* в контексті інших слів. Деякі моделі, що ґрунтуються на словах, репрезентують кожне входження *x* у корпусі з «графічним входженням» у заданому контекстному вікні [9]. Інші моделі на основі слів пояснюють синтаксичні залежності між словом *x* і його «синтаксичним входженням» [10]. Моделі на основі документів більше підходять для моделювання синтаксичних і реляційних зв'язків, наприклад, між словом *лікар і лікарня* або *автомобіль і водій*. Моделі, засновані на словах, особливо ті, які покладаються на інформацію про синтаксичну залежність, є більш точними і краще відображають парадигматичні зв'язки, такі як близькі до синонімів *лікарня і клініка*. Деякі роботи виходять за рамки контекстів появи в текстах (словах або документах), щоб також інтегрувати в інтеркорпусний аналіз «візуальні слова» або зображення, витягнуті з обчислювальних візуальних методів, що призводить до «мультимодальної дистрибутивної семантики» [10].

Інтеркорпусні моделі дають змогу вивчати слова або на рівні типів, або на рівні токенів. Моделі на рівні типів не тільки пояснюють, що одне слово відрізнятиметься від іншого, аби виявити синонімію та пов'язані з нею лексичні зв'язки, а й дають уявлення про різні значення багатозначного слова. Однак, щоб визначити значення конкретного явища, моделі на рівні токенів вирізняють одне використання цього слова від іншого використання того самого слова, у такий спосіб вивчаючи його багатозначність.

Інтеркорпусні моделі розрізняються залежно від способу обчислювального представлення розподільчих даних, тобто представлення на основі графів або векторне представлення. Векторні моделі спираються на растрове представлення слів у векторному просторі. Наприклад, матриця слів×слова, яка має на меті охарактеризувати слова на основі їхніх графічних співструмів, або матриця відносин залежності від слів×залежностей, яка характеризує слова відповідно до їхніх синтаксичних співструмів. Важливо звести велику кількість контекстуальних ознак до обмеженої кількості «семантичних вимірів» або для ефективної обчислювальної обробки, або для візуалізації, що дає змогу дослідникам-лінгвістам легше здобути доступ до семантичних зв'язків, що виникають.

Для ілюстрації головних принципів дистрибутивної семантичної моделі використовується конкретна інтеркорпусна модель, тобто модель на основі слів, яка розглядає графічні входження на рівні типу як векторне представлення.

Цей приклад використовується для вивчення трьох, а саме *собака, кішка та кава*, у пробному демо корпусі з шести прикладів речень.

(1) Собака голосно гавкає на перехожих.

(2) Ветеринар бере собаку за шию.

(3) Кіт голосно муркоче.

(4) Ветеринар був подряпаний кішкою, коли брав її на руки.

(5) Ми п'ємо більше кави, ніж чаю.

(6) Ветеринар бере чашку і п'є каву.

Було використано найпростішу модель, яка ігнорує синтаксичні залежності та перераховує абсолютні найпоширеніші іменники, дієслова та прислівники тільки на основі семантики. Частоти показані в матриці частотності нижче (табл. 2).

Нехтування синтаксичними зв'язками на користь семантичних називається підходом до пошуку інформації за принципом «мішка слів». На реальних прикладах і великих корпусах контекстні слова (стовпчики) часто складаються з тисяч слів і, звісно, мають частоту входження набагато вищу за «2».

Частоту входження можна інтерпретувати як координати, що дають змогу позиціонувати лексеми *собака*, *кіт* і *кава* в багатовимірному семантичному просторі. Продовжуючи цю геометричну метафору, відстань між двома словами в цьому багатовимірному семантичному просторі також можна обчислити, щоб виміряти відстань між їхніми значеннями. На практиці для обчислення подібності між трьома цільовими словами використовується векторна алгебра для обчислення подібності. Фактично, кожне цільове слово може бути представлено вектором його входження або контекстом. Отже, дані розподілу подаються у вигляді векторів входжень. Цільові слова з однаковою кількістю входжень мають схожі векторні подання. Потім вектори порівнюються через показник схожості й обчислюється відстань у векторному просторі: схожість векторів трьох цільових слів обчислюється за косинусом кута між ними. Косинус є стандартною мірою подібності в дистрибутивній семантиці. Інтуїтивно зрозуміло, що кут між двома схожими поняттями (наприклад, *собака* і *кішка*) буде меншим, ніж кут між двома різними словами (наприклад, *собака* і *кава*). Іншими словами, якщо розподіл слів менш схожий, значення косинуса подібності буде меншим:

- cos(собака, кішка) = 0,55 cos(собака, кава) = 0,27
- cos(кішка, кава) = 0,30

Результат відповідає інтуїції: «*собака*» і «*кішка*» найбільш схожі, в той час як «*собака*» і «*кава*» мають найменшу кількість спільних випадків. Треба зазначити, що «*собака*» дещо менше схоже (0,27) на «*каву*», ніж на «*кішку*» (0,30), оскільки «*собака*» також з'являється з «*гавкотом*» і «*перехожим*», на відміну від «*кішки*» і «*кави*».

Застосування цієї методики до реальних прикладів і корпусів не ґрунтується на абсолютних частотах співвходження між цільовим словом і його співвходженнями. Найчастіші супутні випадки не обов'язково дають найбільше інформації про значення цільового слова. Як і в дослідженнях словосполучень, методи інтеркорпусного аналізу спираються на статистичні показники спорідненості, які можуть додавати значення збігам, що з'являються значно частіше з цільовим словом, ніж можна було б очікувати випадково. Ці вагомі збіги, незалежно від їхньої абсолютної частоти, надають більше інформації про значення цільового слова, ніж будь-які інші співпадіння. Наприклад, у випадку зі словом «*собака*» слово «*ветеринар*» семантично ближче до поняття «*тварина*», ніж слово «*брати*», хоча й менш частотне. Показники асоціації, які використовуються як вагові функції в дистрибутивній семантичній моделі, запозичені з аналізу словосполучень. Схема зважування, використана в цьому прикладі, базується на точковій мірі взаємної інформації. Вона спирається на частоти поширеності, зібрані на великому корпусі для обчислення значень схожості. Припустимо, що у ветеринара колокаційна вага становить 4,3, якщо вона з'являється у собаки, 3,5 – у кішки і 0,8 – у кави. Тоді обчислення косинусної подібності з колокаційних ваг вказує, з одного боку, на те, що *собака* і *кішка* більш схожі, ніж у незваженому розрахунку, а з іншого боку, що вони менш схожі на *каву*:

- cos(собака, кішка) = 0,87 cos(собака, кава) = 0,31
- cos(кішка, кава) = 0,32

Зважений розрахунок косинусної подібності між усіма цільовими парами слів генерує матрицю подібності з цільовими словами, як у рядках, так і в стовпцях, і зі значенням косинусної подібності на цільову пару слів у клітинках. Всі елементи діагоналі є «1», причому кожне цільове слово повністю схоже на себе. Матриця

Таблиця 2

Матриця спільного входження для цільових слів собака, кіт і кава

	лаяти	**прохожий**	**ветеринар**	**брати**	**шия**	**муркотіти**	**сильно**	**подряпина**	**пити**	**більше**	**чай**	**чашка**
Собака	1	1	1	1	1	0	1	0	0	0	0	0
Кіт	0	0	1	1	0	1	0	1	0	0	0	0
Кава	0	0	1	1	0	0	0	0	2	1	1	1

Джерело: власне розроблення авторів

симетрична, з однаковими значеннями по обидва боки від діагоналі, тому що косинусна подібність між словами A і B така ж, як і між словами B і A. В таблиці 3 нижче показана подібність косинусів для наших трьох прикладів цільових слів. Наприклад, у великих реальних корпусах для кожного цільового слова можна знайти найбільш схоже слово в решті лексики. Оскільки дуже схожі слова часто є (майже) синонімами, цей тип матриці схожості слів часто використовується для автоматичного вилучення синонімів в обчислювальній лінгвістиці (табл. 3).

Таблиця 3

Матриця значень подібності косинусів

	Кіт	**Кава**	**Собака**
Кіт	1	0,32	0,87
Кава	0,32	1	0,31
Собака	0,87	0,31	1

Джерело: власне розроблення авторів

Метод інтеркорпусного розподільного аналізу підходить для різноманітних лексикографічних та лексикографічних застосувань, як-от автоматичний відбір зразків речень у словниках, діахронічні семантичні дослідження та вивчення соціолінгвістичної варіативності.

Висновки. В цій статті описано потенціал розподільного семантичного моделювання для міжкорпусного аналізу. Завдяки використанню великих електронних корпусів лексикологи і лексикографи мають у своєму розпорядженні значну емпіричну базу даних. Для аналізу такої великої кількості даних можна використовувати інструменти статистичного аналізу та методи міжкорпусного дистрибутивного аналізу для виявлення закономірностей лексичних і семантичних зв'язків.

Методи інтеркорпусного розподільного аналізу ґрунтуються на припущенні, що слова, які зустрічаються в схожих контекстах, мають схожі значення. Залежно від типу контексту, рівня аналізу та комп'ютерного представлення даних вони поділяються на різні підходи та підтипи. Основні принципи цих моделей описано в цій роботі на основі конкретної дистрибутивної моделі, яка є порівняно інтуїтивно зрозумілою для лінгвістів, а саме моделі на основі слів, що бере до уваги графічну комбінованість на рівні жанрів у вигляді векторних уявлень. Ця розподільча модель дає змогу не лише виявляти семантичну (не)схожість між різними словами, як, наприклад, у прикладах із демонстраційного корпусу «*собака*», «*кіт*» та «*кава*», а й знаходити різні значення певного слова на основі семантичної (не)схожості між словами, що трапляються одночасно.

Мета цієї статті – продемонструвати корисність дистрибутивних семантичних моделей для лінгвістів, лексикологів і лексикографів, надавши нетехнічний вступ до основних принципів і конкретних лексикографічних застосувань. Як показують різні підходи та багато підтипів моделей, майбутні дослідження повинні бути зосереджені на розробленні набору параметрів для отримання моделі, яка найкраще відповідає конкретним потребам необхідного лексикографічного або лексикографічного дослідження.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ:

1. Cavasso L., Taboada M. A corpus analysis of online news comments using the Appraisal framework. *Journal of Corpora and Discourse Studies.* 2021. Vol. 4. P. 1–38. DOI: 10.18573/jcads.61

2. Kozlova T., Polyezhayev Y. A Cognitive-Pragmatic Study of Australian English Phraseology. *AD ALTA.* 2022. Vol. 12(1). P. 85–93. DOI: 10.33543/12018593

3. Lefter I., Baird A., Stappen L., Schuller B.W. A cross-corpus speech-based analysis of escalating negative interactions. *Frontiers in Computer Science.* 2022. Vol. 4. URL: https://www.frontiersin.org/articles/10.3389/fcomp.2022.749804/full

4. Lin P. ChatGPT: Friend or foe (to corpus linguists)?. *Applied Corpus Linguistics*. 2023. Vol. 3(3). DOI: 10.1016/j.acorp.2023.100065

5. Lin P., Adolphs S. Corpus linguistics. In The Routledge Handbook of Applied Linguistics. Routledge, 2023. P. 296–308. DOI: 10.4324/9780203856949.ch3

6. Messina C.M., Jones C.E., Poe M. Prompting Reflection: Using Corpus Linguistic Methods in the Local Assessment of Reflective Writing. *Written Communication.* 2023. Vol. 40(2). P. 620–650. DOI: 10.1177/0741088322114942

7. Newman-Griffis D., Sivaraman V., Perer A., Fosler-Lussier E., Hochheiser H. TextEssence: A Tool for Interactive Analysis of Semantic Shifts Between Corpora. In Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting. 2021. 106. NIH Public Access. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8212692/

8. Saddhono K., Rohmadi M., Setiawan B., Suhita R., Rakhmawati A., Hastuti S., Islahuddin I. Corpus Linguistics Use in Vocabulary Teaching Principle and Technique Application: A Study of Indonesian Language for Foreign Speakers. *International Journal of Society, Culture & Language*. 2023. Vol. 11(1). P. 231–245. DOI: 10.22034/ijscl.2022.1971972.2823

9. Wang G., Wang H., Wang L. Research trends in tourism and hospitality from 1991 to 2020: an integrated approach of corpus linguistics and bibliometrics. *Journal of Hospitality and Tourism Insights*. 2023. Vol. 6(2). P. 509–529. https://www.emerald.com/insight/content/doi/10.1108/JHTI-09-2021-0260/full/html

10. Xu J. Application Research of Cognitive Linguistics Based on Big Data Internet Corpus Construction. *Journal of Physics: Conference Series*. 2021. Vol. 1. IOP Publishing. DOI: 10.1088/1742-6596/1861/1/012028