

*До 100-річчя
Національної академії наук України*

НАЦІОНАЛЬНА АКАДЕМІЯ НАУК УКРАЇНИ
УКРАЇНСЬКИЙ МОВНО-ІНФОРМАЦІЙНИЙ ФОНД

ЛІНГВІСТИЧНО-ІНФОРМАЦІЙНІ СТУДІЇ

ПРАЦІ
УКРАЇНСЬКОГО МОВНО-ІНФОРМАЦІЙНОГО ФОНДУ
НАН УКРАЇНИ

У П'ЯТИ ТОМАХ

Дослідження 1991–2018 років

За загальною редакцією *В. А. Широкова*

КИЇВ 2018

НАЦІОНАЛЬНА АКАДЕМІЯ НАУК УКРАЇНИ
УКРАЇНСЬКИЙ МОВНО-ІНФОРМАЦІЙНИЙ ФОНД

ТОМ IV

**КОРПУСНА ТА КОГНІТИВНА
ЛІНГВІСТИКА**

УКРАЇНСЬКИЙ МОВНО-ІНФОРМАЦІЙНИЙ ФОНД

*Рекомендовано до друку вченою радою
Українського мовно-інформаційного фонду НАН України
(протокол № 9 від 23 жовтня 2018 р.)*

Рецензенти:

академік НАН України *О. В. Палагін*,
член-кореспондент НАН України *А. О. Загнітко*

Автори четвертого тому:

*В. А. Широков, О. В. Бугаков, Н. М. Заїка, Т. О. Грязнухіна, М. О. Кригін, М. В. Надутенко,
О. Г. Рабулець, Н. М. Сидорчук, Л. О. Терещенко, І. В. Шевченко, О. О. Шипнівська,
С. С. Ющенко, К. М. Якименко*

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.

Четвертий том «Корпусна та когнітивна лінгвістика» п'ятитомної монографії «Лінгвістично-інформаційні студії» присвячено викладу теоретичних і технологічних засад корпусної лінгвістики. Особливу увагу приділено розробленню Українського національного лінгвістичного корпусу (УНЛК) та його численним застосуванням у проведенні різноманітних лінгвістичних і когнітологічних досліджень та укладанні лексикографічних праць, насамперед академічного Словника української мови в 20 томах.

Погляд на лінгвістичний корпус (ЛК) як на велику колекцію інформаційно і лінгвістично анотованих масивів природномовних текстів спонукає до розроблення структури та архітектури цих лінгвоінформаційних систем, методів ідентифікації, стандартизації та експлікації (маркування) їх структурних елементів. У цьому зв'язку окреслено роль спеціальних мов маркування (SGML, HTML, XML etc.) і різноманітних форматів подання бібліографічної та лінгвістичної інформації в технологіях побудови ЛК.

Розглянуто лінгвістичні та системотехнічні засади побудови УНЛК та його основних структурних підсистем: бібліографічної та лінгвістичної, які забезпечують індексування об'єктів зберігання, створення, редагування та оперування повнотекстовим індексом, забезпечення бібліографічного і повнотекстового багатокритеріального пошуку, статистичні та сервісні функції маніпулювання корпусними даними.

Значну увагу приділено лінгвістичним дослідженням, проведеним у Фонді на матеріалі УНЛК: вивченню різних типів омонімії лексичної системи української мови, функціонуванню прийменникових конструкцій в українських текстах, одержанню статистичних параметрів та визначенню ступеня фонетичності писемного варіанту української мови. Отримані результати мають велике значення для розроблення цифрових систем опрацювання природної мови (орфокоректори, системи семантичної інтерпретації, лінгвістичної концептографії та машинного перекладу, аналізатори та озвучувачі текстів та ін.).

Окремий розділ присвячено застосуванню корпусних технологій у сфері теорії, та технології проведення різного роду лінгвістичних експертиз природномовних текстів.

ЗМІСТ

Розділ 19. УКРАЇНСЬКИЙ НАЦІОНАЛЬНИЙ ЛІНГВІСТИЧНИЙ КОРПУС	7
Вступ	7
19.1. Загальні аспекти корпусної лінгвістики	9
19.2. Лінгвістичні корпуси в історичному аспекті	12
19.3. Сучасний стан розвитку корпусної лінгвістики	16
19.4. Маркування ЛК. Ініціатива та стандарт кодування ЛК	24
19.4.1. Історія та логіка проблеми маркування природномовних текстів	24
19.4.2. Принципи SGML та приклади його застосування	26
19.4.3. Розвиток SGML: HTML, XML, XHTML, MathML, ChemML, UML	31
19.4.4. Text Encoding Initiative та Corpus Encoding Standart	32
19.4.5. Етапи кодування ЛК	37
19.4.6. Критерії застосування стандарту CES	38
19.5. Лінгвістичні засади побудови УНЛК	39
19.5.1. Парадигматизація та лематизація в ЕГС	39
19.5.2. Моделювання системи української транскрипції	41
19.5.3. Моделювання фонетики української мови	42
19.5.4. Моделювання наголошування в системі транскрипції	47
19.6. Система маркування УНЛК	48
19.6.1. Бібліографічний та структурний опис документів в УНЛК	48
19.6.2. Морфологічне маркування УНЛК	61
19.6.2.1. Структура бази даних ЕГС	62
19.6.2.2. Програмні засоби граматичної ЛБД	67
19.6.3. Синтаксичне маркування ЛК	73
19.6.4. Семантичне маркування УНЛК	80
19.6.4.1. Формальні аспекти відношення синонімії та аналогія з визначенням перекладних еквівалентів	85
19.7. Системотехнічні та технологічні засади побудови УНЛК	88
19.7.1. Лінгвістичний корпус як лексикографічна система	88
19.7.2. Архітектура та компоненти програмного забезпечення УНЛК	93
19.7.2.1. Організація збереження метаданих	95
19.7.2.2. Структура БД електронної бібліотеки	96
19.7.2.3. Рівень логіки застосунків	100
19.7.2.4. Рівень представлення даних	101
19.8. Системотехніка структурного маркування УНЛК	109
19.9. Системотехніка морфологічного аналізатора	111
19.10. Системотехніка семантичного аналізатора	118
Розділ 20. КОРПУСНІ ТА СТАТИСТИЧНІ МЕТОДИ ДОСЛІДЖЕННЯ ТЕКСТІВ	120
20.1. Корпусні методи дослідження текстів	120
20.1.1. Дослідження граматичної омонімії на морфологічно анотованому корпусі текстів	120
20.1.2. Визначення потенційних типів морфологічних омонімів в українській мові ...	121
20.1.3. Дослідження морфологічної омонімії на матеріалі УНЛК	126
20.1.4. Міжчастиномовні морфологічні омоніми в українських текстах	128

20.1.5. Частота вживання ТМО в текстах	131
20.1.6. Характеристика ТМО за наповнюваністю їхніх ОР	134
20.1.7. Лінгвістична база контекстів морфологічних омонімів.....	135
20.1.8. Дослідження функціонування українських прийменників у тексті на матеріалі УНЛК.....	139
20.1.9. Дослідження функціонування прийменників на морфологічному рівні.....	146
20.1.9.1. Усунення міжчастиномовної омонімії прийменників	146
20.1.9.2. Ідентифікування складених прийменників у тексті.....	151
20.1.10. Дослідження функціонування прийменників на синтаксичному рівні	159
20.1.11. Дослідження функціонування прийменників на семантичному рівні.....	167
Розділ 21. ДОСЛІДЖЕННЯ СТАТИСТИЧНИХ ПАРАМЕТРІВ УКРАЇНСЬКОЇ МОВИ	176
21.1. Постановка проблеми, мета та завдання статистичного дослідження мови. Історичний огляд досліджень зі статистики мови	176
21.1.1. Апарат статистичного дослідження мови. ЛК як експериментальна база дослідження	180
21.1.2. Статистика знакової системи.....	181
21.1.3. Транскрибована форма УНЛК та дослідження відповідності українського правопису фонетичному принципу.....	185
Розділ 22. КОРПУСНІ ТЕХНОЛОГІЇ В ЛІНГВОЕКСПЕРТНІЙ ДІЯЛЬНОСТІ.....	191
22.1. Лінгвістична експертиза: логіка проблеми	191
22.2. Лінгвістична експертиза: історія.....	192
22.2.1. Лінгвістична та судова лінгвістична експертизи: подібності та відмінності	196
22.3. Сучасний стан розвитку галузі лінгвістичних експертиз.....	199
22.3.1. Психолінгвістична судова експертиза	200
22.3.2. Лінгвістична судова експертиза в репутаційних справах	201
22.3.3. Семантична лінгвістична експертиза.....	202
22.3.4. Лінгвістична експертиза в справах щодо введення в оману.....	203
22.3.5. Логіко-лінгвістична експертиза.....	204
22.3.6. Лінгвістична експертиза законопроектів.....	204
22.3.7. Авторознавча експертиза	205
22.3.8. Лінгвістична експертиза на ознаки недоброчесного запозичення (плагіату)	207
22.3.9. Лінгвістична експертиза на предмет пропагування наркотиків.....	209
22.3.10. Лінгвістична експертиза на предмет ознак богохульства.....	210
22.3.11. Лінгвістична експертиза на предмет прихованих мотивів	210
22.4. Практичне застосування ЛК як інструменту лінгвоекспертної діяльності	212
22.4.1. Корпусні технології в логіко-лінгвістичних експертизах.....	213
22.4.2. Корпусні технології в авторознавчих експертизах.....	216
22.4.3. Корпусні технології в експертизах на предмет прихованих мотивів	221
ВИСНОВОК	230
ДОДАТОК А. Лінгвоекспертна діяльність Українського мовно-інформаційного фонду НАН України в публікаціях. Розвиток теорії, методики та програмних засобів	233
ЛІТЕРАТУРА.....	238

РОЗДІЛ 19

УКРАЇНСЬКИЙ НАЦІОНАЛЬНИЙ ЛІНГВІСТИЧНИЙ КОРПУС

Вступ

В останні десятиліття великого поширення у мовознавстві набули лінгвістичні корпуси – великі колекції лінгвістично анотованих масивів природномовних текстів. У цьому зв'язку навіть виник цілий мовознавчий напрям, який одержав назву «корпусна лінгвістика». Безпосереднім наслідком впровадження лінгвістичних корпусів як провідного технологічного інструментарію сучасного мовознавчого дослідження стало розширення меж лексикографічного опису мови. Невипадково, що навіть сам лінгвістичний корпус у його найпростішому варіанті та безпосередній формі за своєю структурою нагадує певний специфічний словник, де реєстровою одиницею і водночас «лівою частиною словникової статті» виступає певна лексема, а інтерпретаційною, «правою» частиною слугує сума «мікроконтекстів», до яких входить реєстрова лексема (у всіх її граматичних значеннях). Бажання дослідників експлікувати якомога повніше систему мовних одиниць, відношень та структур, які імпліцитно містяться в текстах корпусів, робить аналогію між лінгвістичними корпусами та лексикографічними системами – словниками, ще більш глибокою і повною. Забігаючи наперед, відзначимо, що всі лінгвістичні корпуси і насправді мають структуру певних лексикографічних систем¹.

Перші корпуси текстів являли собою колекції зразків відносно невеликих розмірів. Основним об'єктом такого колекціонування був нормативний писемний варіант мови. Корпуси текстів укладались за строгою схемою, виключне значення надавалось репрезентативності масиву.

Обсяги сучасних лінгвістичних корпусів, які сягають сотень мільйонів слововживань, залучення до корпусів не тільки писемних, але й усних джерел, розширення їхніх функціональних можливостей у бік багатомовності та вдосконалення мовно-інформаційного інструментарію свідчать про фундаментальність корпусної лінгвістики як окремої ділянки науки, що намагається охопити досліджувану предметну галузь (мову) в усіх її проявах та повноті явищ. Адже для побудови досконалих мовно-інформаційних артефактів необхідно мати такі експериментальні масиви лінгвістичних даних, на яких у достатньо репрезентативному вигляді були б представлені всі мовні феномени. Зазначене помітно розширило сферу інтересів укладачів лінгвістичних корпусів. Деякі відомі корпуси текстів (наприклад, BNC, COBUILD / Birmingham Corpus)

¹ Широков В. А. Феноменологія лексикографічних систем. Київ: Наукова думка, 2004. 327 с.

містять, крім писемної мови, також субкорпуси усного мовлення. Є й корпуси, укладені лише зі зразків усного мовлення (IBM-Lancaster Spoken English Corpus, Corpus of Spoken American English). Все більшу увагу розробники приділяють специфічним варіантам мови, до яких належить, наприклад, дитяча розмовна та письмова мова (корпус CHILDES), мовлення «неносіїв мови» (International Corpus of Learner English), інші різного роду підмови та ненормативні варіанти мови (сленги, девіації тощо). Ще однією тенденцією є поява так званих історичних (або діяхронічних) корпусів (*historical corpora*), які фіксують мову на різних етапах її розвитку (ARCHER Corpus, Helsinki Corpus, etc.)².

Відповідно до мети укладання розрізняють лінгвістичні корпуси: синхронічні, діяхронічні, розмовні, діалектні, письмової або усної форми мови, мови певних письменників, стилів або жанрів, проблемно-галузеві, фольклорні, соціолінгвістичні тощо. За лінгвістичними характеристиками корпуси поділяють на: фонетичні, орфографічні, морфологічні, синтаксичні, семантичні, одно- та багатомовні, паралельні, порівняльні та ін. Все розмаїття лінгвістичних корпусів зводиться до двох груп: загальні корпуси (що належать до мови загалом) та спеціальні (створені під конкретні дослідницькі завдання, вузькоспеціалізовані).

Із поширенням (або навіть – переважанням) цифрової форми інформації репрезентативні масиви текстів можуть бути зібраними лише у вигляді комп'ютерних *БД*, на матеріалі яких можна проводити різнопланові дослідження та створювати прикладні мовно-інформаційні системи. Отже, корпусна лінгвістика пов'язується із застосуванням комп'ютерних лінгвістичних корпусів як інструментів мовознавчого дослідження, що робить її важливою галуззю прикладної та комп'ютерної лінгвістики.

Надалі вживатимемо термін «лінгвістичний корпус» без додавання слова «комп'ютерний», а також аббревіатуру ЛК.

Одним із сучасних потужних інструментів вітчизняної лексикографії та лінгвістичних досліджень є Український національний лінгвістичний корпус (надалі паралельно вживатимемо аббревіатуру УНЛК), опису лінгвістичних та системотехнічних принципів формування якого присвячено цей розділ. УНЛК створюється та розвивається в Українському мовно-інформаційному фонді НАН України в межах робіт із розвитку Національної словникової бази України³.

² Докладніше про розвиток корпусної лінгвістики в різних країнах див., наприклад, Biber, Douglas and Conrad, Susan and Reppen, Randi: *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge University Press, 1998. Баранов А. Н., Добровольский Д. О. Немецкая корпусная лингвистика. Вестник МГУ. Сер. Иностранные языки. № 1, 1998. Баранов А. Н. Введение в прикладную лингвистику. Москва: Эдиториал УРСС, 2001. С. 121–128; McEnery T., Wilson A. *Corpus Linguistics: An Introduction*. 2nd Edition, Edinburgh: Edinburgh University Press, 2001. P. 20–24.

³ Указ Президента України «Про розвиток Національної словникової бази» від 7 серпня 1999 року № 967/99. URL: <http://zakon3.rada.gov.ua/laws/show/967/99>.

19.1. Загальні аспекти корпусної лінгвістики

Зміст поняття «корпусна лінгвістика» охоплює як теоретичне, так і прикладне мовознавство – напрацювання лінгвістичної теорії покладено в основу моделей та алгоритмів систем опрацювання природної мови, прикладні результати втілюються в лінгвістичні корпуси, що, в свою чергу, виступають інструментом досліджень у галузі теоретичного мовознавства. Наголошуємо, що йдеться саме про комп'ютерні корпуси, оскільки корпусами вважалися й певні традиційні паперові збірки інформації, призначені для цілей лінгвістичного дослідження (наприклад, лексичні або словникові картотеки), хоча такий напрям досліджень корпусною лінгвістикою зараз називати не прийнято. Сказане, а також потреба у створенні та практичному застосуванні в мовознавчих (і не тільки) дослідженнях комп'ютерних аналогів лексичних картотек, зрештою, спонукало до формулювання, так би мовити, рефлексивних визначень поняття лінгвістичного корпусу. Наведемо приклади найтипівіших уявлень самих представників корпусної лінгвістики, що ж таке є лінгвістичний корпус у сучасному розумінні.

Піонер корпусної лінгвістики, один з розробників Браунівського корпусу, про який йтиметься далі, В. Френсіс (William Nelson Francis) визначає лінгвістичний корпус як «зібрання текстів, яке вважається репрезентативним стосовно певної мови, діалекту або іншої ділянки мови і призначене для використання в лінгвістичних дослідженнях»⁴.

Сучасний російський теоретик корпусної лінгвістики, професор кафедри перекладознавства Університету Лідса (Велика Британія) С. А. Шаров наводить два підходи до визначення лінгвістичного корпусу⁵. Перший передбачає, що корпусом є будь-яка колекція текстів з певної тематики в електронній формі (*Корпус 1*). Інший – що корпус – це колекція текстів, зібрана у відповідності до явно сформульованих правил і, можливо, розмічена (annotated) на якомусь рівні лінгвістичного аналізу (*Корпус 2*). Цьому визначенню, зокрема, відповідає колекція текстів, зібраних у Машинному фонді російської мови⁶.

Поширеним є погляд на ЛК як на репрезентативну колекцію текстів у розумінні *Корпус 2*, яка може адекватно представляти потенційно необмежене коло текстів фіксованого типу в певному діахронічному зрізі⁷. Репрезентативність у цьому визначенні передбачає, що колекція текстів є збалансованою за жанрами та функціональними стилями і що вона має достатній обсяг вибірки за кількістю

⁴ Francis W. Language Corpora B.C. In: Jan Svartvik (ed.) *Direction in Corpus Linguistics. Proceeding of Nobel-82*. Stockholm, 4–8. August 1991. Berlin – New York. Mouton de Gruyter. P. 17–35.

⁵ Шаров С. А. Представительный корпус русского языка в контексте мирового опыта. НТИ. Сер. 2. № 6. 2003. С. 9.

⁶ URL: <http://cfri.ruslang.ru/>.

⁷ McEnery T., Wilson A. *Corpus Linguistics: An Introduction*. 2nd Edition, Edinburgh: Edinburgh University Press, 2001. P. 2–3.

текстів і авторів для того, щоб слугувати основою для статистично достовірних досліджень лінгвістичних феноменів у текстах певної тематики.

Російський лінгвіст В. П. Захаров, автор низки праць з корпусної лінгвістики, наводить таке визначення: «Під назвою “корпус текстів” будемо розуміти великий, уніфікований, структурований, розмічений масив мовних (мовленнєвих) даних в електронному вигляді, призначений для певних цілей»⁸. Пізніше ним уточнено: «Під лінгвістичним корпусом текстів розуміється великий, уніфікований, структурований, розмічений, філологічно компетентний масив мовних даних, представлений в електронному вигляді й призначений для розв’язання різних лінгвістичних завдань»⁹. До цього визначення близьке й сформульоване школою О. С. Герда (кафедра математичної лінгвістики філологічного факультету Санкт-Петербурзького державного університету), до якої належить і В. П. Захаров: «великий за обсягом, представлений в електронному вигляді, уніфікований, структурований, розмічений і філологічно компетентний масив мовних даних, доповнений системою керування даними – універсальними програмними засобами для пошуку різноманітної лінгвістичної інформації та зручного представлення її широкому користувачеві»¹⁰.

У наведених дефініціях усі детермінанти слів «масив» та «розмітка» є суттєвими і відмежовують лінгвістичний корпус від простого зібрання (колекції) електронних текстів.

Маркування лінгвістичного корпусу (*розмічання, анотування, тегування (tagging)*)¹¹ полягає в приписуванні текстам корпусу та їхнім компонентам спеціальних тегів (міток) різних типів: зовнішнього (елементи бібліографічного опису: видання, рік, автор тощо); структурного (описують структуру тексту: розділ, абзац, речення); лінгвістичного (лексикографічні, граматичні, семантичні, синтаксичні та інші параметри).

Постає питання: чи має корпусна лінгвістика якийсь окремий, специфічний, характерний саме для неї об’єкт та предмет дослідження? Це питання, власне, стосується і комп’ютерної лінгвістики загалом.

Відповідь не є однозначною. Звернімося до аналогії. Скажімо, у фізиці ділянку досліджень, які проводяться за допомогою спектроскопів, не називають «спектроскопічною фізикою», між тим як термін «спектроскопія» все-таки існує. Так само і ділянку астрономії, де дослідження проводяться за допомогою телескопів, не прийнято називати «телескопною астрономією», хоча термін

⁸ Захаров В. П. Чешский национальный корпус текстов: организация и способы использования. Корпусная лингвистика и лингвистические базы данных. Докл. научн. конфер. (РФ, Санкт-Петербург, 5–7 марта 2002 г.).

⁹ *Корпусная лингвистика – 2004*. Международная конференция по корпусной лингвистике. URL: <http://dialog-21.ru/>.

¹⁰ Герд А. С. Вступне слово на міжнародній конференції *Корпусная лингвистика – 2004*. URL: www.phil.pu.ru/news/kllbd/corpling.htm.

¹¹ Leech G. Corpus Annotation Schemes. *Literary and Linguistics Computing*. 1993. Vol. 8 # 4. P. 275–281.

«радіоастрономія» і, відповідно, ділянка астрономії, де дослідження проводяться в радіодіапазоні і за допомогою радіотелескопів, є загальноприйнятим.

Отже взагалі: чи коректно іменувати певну ділянку науки за тим експериментальним інструментарієм (об'єктом експериментальної техніки), за допомогою якого проводяться дослідження у цій ділянці? Мабуть, лише тією мірою, яка визначає, наскільки зазначений інструмент виділяє специфічний об'єкт дослідження з-поміж усього розмаїття об'єктів цієї науки. У більш загальному розумінні відповідь лежить у площині співвідношення предмета та методу¹².

З огляду на сказане, корпусна лінгвістика не відзначається яскраво вираженим об'єктом та предметом дослідження, принаймні в області їхніх системних ознак. Те, що насправді вирізняє корпусну лінгвістику з решти галузей лінгвістичної науки, – це обсяги мовного матеріалу, який залучається до мовознавчого дослідження, комплексність, оперативність опрацювання зазначеного матеріалу та можливість прямого доступу до великого числа лінгвістичних фактів. Як буде продемонстровано далі, обсяги окремих ЛК сягають сотень мільйонів (а то й багатьох мільярдів, як, наприклад, веб-простір – інформаційне наповнення мережі Інтернет) слововживань¹³. Такі обсяги у паперовому варіанті немислимі навіть фізично, не кажучи вже й про інші пошукові та сервісні можливості, що надають комп'ютерні інформаційні системи.

Природною ділянкою корпусної лінгвістики, безперечно, мають бути статистичні властивості мови, які проявляються лише на великих масивах текстової інформації. При цьому слід зважити на те, що саме поняття «великий обсяг інформації» не є інваріантним і дуже залежить від того, властивості яких об'єктів мовної системи вивчаються. Отже, на перший погляд, специфіка корпусної лінгвістики як окремої ділянки мовознавства може вважатися доконаним фактом лише тією мірою, якою справджується кваліфікація, що кількісні зміни у параметрах та обсягах її об'єктів дослідження зумовлюють зміни якісні.

Проте останнє положення є занадто загальним, але як визначення – бодай часткове – певних специфічних завдань цієї ділянки мовознавчої науки виявляється не дуже складним – до них належать насамперед ті, які «вручну» виконати неможливо вже фізично, хоча їхня природа і не є статистичною. Візьмімо, наприклад, таке просте і детерміністське за своєю принциповою постановкою завдання, як побудова фонетичної транскрипції для всіх українських словоформ на реєстрі, скажімо, 200 тисяч лексем, де повне число словоформ сягатиме майже чотирьох мільйонів. Виписати ці чотири мільйони слів, та ще й потім записати їх транскрипційними знаками (зрозуміло, з наголосами) – завдання неосяжне для традиційного, не «комп'ютерного» мовознавства. Зазначимо, що воно було

¹² Декарт Р. Рассуждение о методе. Метафизические размышления. Начала философии. Луцк: Вежа, 1998. 302 с.

¹³ Інформацію про різні ЛК, які зараз функціонують у світі, можна одержати на сайті: https://en.wikipedia.org/wiki/List_of_text_corpora.