

УДК 81'322

DOI: 10.31471/2304-7402-2025-21(77)-33-48

**ТЕОРЕТИЧНІ ОСНОВИ КОРПУСНОЇ ЛІНГВІСТИКИ:
ПРИНЦИПИ ПОБУДОВИ ТЕКСТОВИХ КОРПУСІВ****Світлана Луцак, Галина Юрчак***Івано-Франківський національний медичний університет;**76018, м. Івано-Франківськ, вул. Галицька, 2;**e-mail: sv.lutsak@gmail.com, galuna.yurchak@ukr.net*

У науковій статті розглянуто теоретичні засади корпусної лінгвістики, окреслено основні принципи побудови текстових корпусів, визначено провідні напрями сучасних корпусних досліджень. На основі аналізу праць українських та зарубіжних дослідників уточнено поняття «текстовий корпус», визначено його ключові параметри (машиночитаність, автентичність, репрезентативність, збалансованість, відібраність, стандартизованість, оперативність опрацювання матеріалу тощо). Розглянуто типові етапи побудови корпусу, зокрема токенизацію, лематизацію, синтаксичний аналіз та розмітку, а також труднощі, пов'язані з формуванням масштабних корпусних ресурсів. Особливу увагу зосереджено на практичному застосуванні корпусів у різних галузях: лексикографії, перекладознавстві, соціолінгвістиці, лінгводидактиці, стилістиці, судовій лінгвістиці та мовній варіантології. У статті також окреслено значення корпусного підходу для сучасної лінгвістики, що передбачає відхід від інтроспекції на користь емпіричних методів аналізу великих обсягів реального мовного матеріалу. Систематизовано основні вимоги до корпусів, подано огляд типологічних характеристик та перспектив подальшого розвитку корпусних досліджень в Україні та світі.

Ключові слова: *корпусна лінгвістика, текстовий корпус, репрезентативність, машиночитаність, корпусна розмітка, цифрові технології.*

Постановка проблеми. Упродовж останніх десятиліть корпуси текстів набули широкого застосування в мовознавчих дослідженнях. Ці великі масиви анотованих природномовних текстів стали основою для нового напрямку – корпусної лінгвістики, яка ґрунтується на використанні електронних корпусів як інструменту дослідження мовної системи.

У науковому дискурсі існують різні підходи до визначення статусу корпусної лінгвістики. Деякі дослідники розглядають її як самостійну галузь мовознавства з власною методологією та теоретичними засадами. Так, Тоні Макенері та Ендрю Гарді зазначають, що корпусна лінгвістика

не є лише набором методів, а «має потенціал переорієнтувати наш підхід до вивчення мови. Вона може уточнити та переосмислити низку мовних теорій, а також сприяти розвитку нових теорій мови, що ґрунтуються на засвідченому мовному вживанні та отриманих з нього висновках» [17, с. 2].

Аналіз досліджень. У сучасному мовознавстві спостерігається динамічне зростання зацікавлення корпусними дослідженнями, що виявляється в постійному оновленні та розширенні лінгвістичних корпусів, збільшенні кількості наукових праць. Основні напрями корпусної лінгвістики охоплюють як загальнотеоретичні аспекти (Д. Байбер, М. Бейкер, Р. Буса, Дж. Синклер, В. Тойберт, С. Хюстон, В. Жуковська), так і взаємодію корпусного підходу з іншими галузями лінгвістики, зокрема соціолінгвістикою, когнітивною лінгвістикою, перекладознавством тощо (Д. Бібер, М. Бейкер, С. Конрад, Р. Реппен, Т. Макенері, А. Харді). Значну увагу дослідники приділяють типології корпусів, розробленню методів інтерпретації корпусних даних (С. Гріс), а також принципам формування корпусів природних мов для лінгвістичного аналізу (Б. Алтенберг, А. Вілсон, У. Френсіс, Є. Карпіловська, Г. Кеннеді, А. Таран, В. Широков). Зростання кількості та різноформатності корпусних проєктів свідчить не лише про міждисциплінарну інтеграцію корпусної лінгвістики, а й про її ключову роль у формуванні сучасної парадигми прикладних і теоретичних мовознавчих досліджень на міжнародному рівні.

Актуальність теми зумовлена потребою систематизації знань про теоретичні засади корпусної лінгвістики, зокрема принципів побудови текстових корпусів, які є базовими інструментами у сфері прикладного мовознавства, лексикографії, перекладознавства, соціолінгвістики тощо. З огляду на цифровізацію науки, питання ефективного формування та розмітки корпусів набуває особливої ваги.

Метою статті є узагальнення історико-теоретичних аспектів корпусної лінгвістики, визначення основних принципів побудови текстових корпусів та окреслення перспектив майбутніх досліджень.

Корпусну лінгвістику часто розглядають як напрям прикладної лінгвістики, оскільки вона активно використовується для вирішення практичних завдань у лексикографії, перекладознавстві, лінгводидактиці та інших сферах. У цьому контексті вона слугує інструментом для аналізу великих обсягів мовних даних, що дозволяє отримувати емпірично обґрунтовані результати.

«Одним із ключових аспектів розвитку лінгвістичних досліджень є міждисциплінарна взаємодія, тісний зв'язок з іншими науками» [8, с. 258]. Так, розвиток мовознавства супроводжується інтеграцією з іншими науками та активним застосуванням сучасних технологій, що зумовило появу таких міждисциплінарних галузей, як математична, комп'ютерна, кібернетична, корпусна лінгвістика тощо [8, с. 59]. Корпу-

сна лінгвістика має міждисциплінарний характер: вона є самостійною галуззю мовознавства з власними теоретичними засадами, водночас тісно інтегрованою в прикладне мовознавство як потужний інструмент для вирішення практичних лінгвістичних завдань. Текстові корпуси – структуровані електронні зібрання текстів – слугують не лише джерелом емпіричних даних, а й підґрунтям для нових методологічних підходів.

Сучасна корпусна лінгвістика – це багатовимірна галузь знань, що поєднує мовознавчий аналіз із цифровими технологіями для вивчення природної мови на основі масивів текстів. Її головне завдання полягає у створенні, структуризації та інтерпретації електронних текстових корпусів, які слугують об'єктивною базою для лінгвістичних досліджень. За словами О. Демської, «текстовий корпус – це певним чином організоване комп'ютерне зібрання писемних і/або усних текстів будь-якої національної мови, якому притаманні обов'язкові ознаки і який призначений для наукового та практичного вивчення цієї мови» [2, с. 36]. Дослідниця стверджує, що корпус – це «машиночитане, стандартно подане зібрання репрезентативних для певної мови, діалекту або іншої підмножини мов писемних або усних текстів, призначених для лінгвістичного аналізу й опису, відібраних і впорядкованих згідно з екстра- та інтралінгвістичними критеріями» [3, с. 89]. Тобто, корпус – це не просто «цифрова бібліотека», а спеціально підготовлений інструмент, який відображає живу мову в усіх її аспектах і призначений для наукового аналізу.

Корпусна лінгвістика, як наукова галузь, не виникла на порожньому місці. Її витоки пов'язані з традицією лексикографічних картотек – паперових зібрань лексичних одиниць, які слугували джерелом для укладання словників і здійснення лінгвістичних досліджень. Такі картотеки, які могли налічувати мільйони карток, відображали систематизовану вибірку мовних одиниць із текстів, відібраних за заданими критеріями. Відомим прикладом є картотека для Оксфордського словника англійської мови (понад 10 млн карток) [7, с. 13].

Лексична картотека Інституту української мови НАН України, як зазначено у 4 томі «Корпусна та когнітивна лінгвістика» п'ятитомного видання «Лінгвістично-інформаційні студії» Українського мовно-інформаційного фонду НАН України за редакцією В.А. Широкова, була започаткована у 1941 році, зараз налічує понад 6 млн карточок. Вона стала основою укладання багатьох словників, зокрема і «Словника української мови» в 11 томах [7, с. 13].

Перехід до електронної обробки текстів розпочався орієнтовно в середині ХХ століття. Корпусна лінгвістика як дисципліна сформувалась у 1960-х роках, коли в Університеті Брауна (США) було створено Brown Corpus – перший масштабний електронний корпус сучасної американської англійської мови. Його автори В. Френсіс і Г. Кучера тлумачили корпус як сукупність текстів, які можна вважати представницькими для

певної мови, діалекту або іншої підмножини мов, призначену для лінгвістичного аналізу [7, с. 12-13]. Це визначення стало фундаментальним для наступних десятиліть.

Браунівський корпус містив тексти стандартного американського англomовного письма, зібрані для проведення кількісного аналізу мови. Так, Браунівський корпус текстів (Brown Corpus) вважають першим у світі комп'ютеризованим корпусом природної мови, створеним у США в середині ХХ століття. Його розробка була ініційована В. Н. Френсісом і Генрі Кучерою в Університеті Брауна (Brown University, Rhode Island, США) у 1961 році. Остаточне формування корпусу було завершено в 1964 році.

Цей корпус був ретельно збалансований: зразки добиралися з урахуванням жанрового розмаїття, функціональних стилів, соціального контексту. Таким чином, корпус відображав типову норму письмової американської англійської мови першої половини 1960-х років.

Браунівський корпус започаткував традицію корпусної репрезентативності – включення до корпусу текстів, які максимально відбивають мовні явища певного типу, періоду й стилю. Цей ресурс став підґрунтям для перших частотних словників, граматичних описів, колокаційних аналізів тощо. Обробка текстів здійснювалася на великих обчислювальних машинах того часу, що потребувало ручного введення текстів, їх маркування й сортування. Саме Браунівський корпус став першим ресурсом, де було систематизовано частотність слів, синтаксичні конструкції, морфологічні ознаки, що дало змогу будувати об'єктивні моделі мови [19, с. 384]. Згодом за його зразком були створені аналогічні корпуси: Lancaster-Oslo/Bergen Corpus (LOB) для британської англійської, Australian Corpus of English (ACE) тощо.

Одночасно подібні ініціативи виникали в Європі: зокрема, у 1956 р. в Італії було започатковано Центр автоматизації вивчення літературних текстів у Галареті. Центр став однією з перших європейських інституцій, яка системно застосовувала комп'ютерні технології для лінгвістичних і гуманітарних досліджень. Центр був створений з ініціативи італійського єзуїтського священника, філософа та філолога Роберто Буза (Roberto Busa), який вважається відкривачем цифрової гуманітаристики та одним з основоположників корпусної лінгвістики [13, с. 83-90]. Роботу над проектом «Index Thomisticus» – цифровим індексом творів Фоми Аквінського – Р. Буза розпочав ще в 1949 році, однак повноцінна реалізація стала можливою лише після 1956 року, коли проєкт отримав технічну підтримку з боку IBM (США), зокрема у вигляді машин для обробки текстових даних [18, с. 20-21].

Дослідник корпусної лінгвістики В. Широков зауважує, що Браунівський корпус та Центр автоматизації вивчення літературних текстів у Галареті започаткували напрям корпусної лінгвістики, однак їхні прак-

тичні можливості були обмежені через відсутність обчислювальних комплексів та обмежений спектр функціональних можливостей інформаційних систем [7, с. 15-16].

Незважаючи на технологічні обмеження того часу, процес формування корпусів продовжувався. Уже в 1980-х роках, у зв'язку з удосконаленням обчислювальних засобів, стало можливим включення до корпусів значно більших обсягів текстового матеріалу. Саме в цей період починає усталюватися практика створення електронних корпусів, орієнтованих на потреби гуманітарних міждисциплінарних досліджень – передусім у таких галузях, як богослов'я, юриспруденція та історія культури.

Один із найвідоміших теоретиків мовознавства ХХ століття Н. Хомський у своїх ранніх працях виступав із різкою критикою корпусів текстів. Він уважав, що лінгвістичне дослідження має ґрунтуватися не на вивченні фактичних мовленнєвих реалізацій, а на моделюванні ідеальної мовної компетенції носія мови – внутрішнього знання мовної системи. У фундаментальній праці «Syntactic Structures» (1957 р.) Н. Хомський стверджував, що збирання мовних даних у вигляді текстів чи корпусів не дозволяє адекватно описати глибинні граматичні структури, оскільки емпіричні дані завжди обмежені, варіативні та можуть містити відхилення від норми [14].

У своїй студії автор навів приклад граматично правильного, але семантично абсурдного речення: «Colorless green ideas sleep furiously». Цей вислів, за задумом Н. Хомського, мав ілюструвати те, що граматичність не залежить від наявності або відсутності мовної конструкції в реальних текстах, а отже – і в корпусах текстів [14, с. 18].

Хоча таке речення, найімовірніше, ніколи не фіксувалося в природному мовленні, носії мови однозначно сприймають його як граматично коректне. Такий приклад засвідчує обмеження статистичного підходу: частотність або неприсутність мовної одиниці в корпусі не може бути єдиним критерієм для її граматичного статусу. З цієї позиції Н. Хомський піддав сумніву методологічну цінність корпусних даних для побудови лінгвістичних моделей, які мають відображати не лише мовну поведінку (performance), а передусім мовну компетенцію (competence).

У цьому контексті корпуси розглядалися ним як недостатнє джерело для формулювання універсальних граматичних правил, оскільки вони завжди репрезентують лише обмежену, емпірично зафіксовану частину мовного досвіду, тоді як внутрішня мовна здатність носія мови виходить за межі корпусно зафіксованих реалізацій.

Наприкінці ХХ століття багато мовознавців почали критично переосмислювати ідеї Н. Хомського, вказуючи на необхідність емпіричної верифікації лінгвістичних гіпотез (зокрема, Т. Макенері, А. Вілсон,

Г. Кеннеді). У сучасному мовознавстві обидва підходи – раціоналістичний (хомськіанський) та емпіричний (корпусний) – часто застосовуються паралельно або комбінуються для розв’язання конкретних завдань.

Таким чином, критика Н. Хомського сприяла теоретичній дискусії про природу лінгвістичного знання, але не зупинила розвиток корпусної лінгвістики. Навпаки, вона сприяла глибшому усвідомленню її меж і потенціалу.

У кінці XX – на початку XXI століття корпуси стали важливим ресурсом для автоматизованої обробки мовних даних. Було розроблено десятки нових корпусів, які охоплювали різні мови, стилі, жанри, регіони, часові зрізи. Паралельно виникали стандарти корпусної розмітки та технології лінгвістичної анотації.

Можна вважати, що у сучасному розумінні історія корпусної лінгвістики охоплює близько 60 років і поєднує дві лінії розвитку:

- традицію картотечних зібрань (аналогова лексикографія);
- цифрову обробку текстових даних як основу для лінгвістичних, прикладних і міждисциплінарних досліджень.

Термін «корпусна лінгвістика», як стверджує В. Жуковська, увійшов до наукового вжитку лише наприкінці XX століття «з публікацією у 1983 році збірника наукових праць «Corpus Linguistics: Recent Developments in the Use of Computer Corpora in English Language Research» за матеріалами конференції ICAME “Conference on the Use of Computer Corpora in English Language Research”» [10, с. 8]. Широке впровадження комп’ютерних технологій і спеціалізованого програмного забезпечення докорінно трансформувало підходи до лінгвістичних досліджень, зробивши процес збору й аналізу мовних даних значно ефективнішим. Завдяки автоматизованим засобам стало можливим оперативне опрацювання багатомільйонних текстових масивів, включно з побудовою конкордансів, аналізом частотності словоформ, лексем, граматичних і синтаксичних конструкцій. Корпусні інструменти дозволяють відстежувати динаміку мовних одиниць у різних часових періодах, досліджувати їх лексичну (колокаційну) та граматичну (колігаційну) сполучуваність.

В. Жуковська стверджує, що дослідницька програма «корпусної лінгвістики, яка, будучи суто емпіричною дисципліною, при аналізі лінгвального матеріалу покладається на реальне функціонування мови з метою встановлення правил та вивчення особливостей продукування мови людиною, на відміну від тих досліджень, які опираються на вигадані приклади чи інтроспекцію» [10, с. 13].

Використання комп’ютерних технологій забезпечує швидку й ефективну обробку значних обсягів мовного матеріалу, що дає змогу отримати повний набір прикладів уживання певної мовної одиниці в межах конкретного корпусу. Це створює для дослідника об’єктивну кількісну

основу, яка сприяє формулюванню більш обґрунтованих і верифікованих лінгвістичних висновків. Завдяки корпусному підходу мовознавці мають змогу перевіряти достовірність теоретичних гіпотез, що стосуються функціонування мовних елементів, а також відкривати нові дослідницькі перспективи, які раніше залишалися поза увагою через обмеженість доступу до великих масивів текстових даних.

Пройшовши етапи критичного сприйняття та маргіналізації в 1960–1970-х роках, корпусна лінгвістика з часом утвердилася як один із ключових напрямів сучасного мовознавства. Нині вона по праву, як стверджує О. Демська-Кульчицька, розглядається як самостійна галузь прикладної лінгвістики, що має власну теоретичну основу, методологію та сферу практичного застосування [4, с. 122]. Корпусна лінгвістика, як переконливо доводить Г. Лук'янець, ґрунтується на розробленні теоретичних принципів і практичних методик створення, комп'ютерного опрацювання, використання та аналізу мовних даних, представлених у вигляді текстових корпусів. А лінгвістичний корпус – це електронна база текстів, що має стандартизовану структуру й містить репрезентативний набір усних і писемних мовних матеріалів, характерних для певної мови, діалекту або мовної підсистеми. Такі корпуси виконують функцію інформаційно-довідкових систем, які забезпечують дослідника емпіричною основою для аналізу мовних одиниць на різних рівнях мовної організації – від фонетичного й морфологічного до лексико-семантичного та синтаксичного. Вони також дають змогу вивчати особливості функціонування природної мови як у писемній, так і в усній формах, а також досліджувати стилістичну варіативність мовлення у художньому, офіційно-діловому, публіцистичному, науковому та розмовному стилях [6, с. 127-128].

Так, орієнтуючись переважно на емпіричний аналіз мовного матеріалу, корпусна лінгвістика забезпечує зменшення впливу дослідницької суб'єктивності та сприяє наближенню до об'єктивного, науково обґрунтованого вивчення мовних явищ.

Корпус у корпусній лінгвістиці – це не просто зібрання текстів, а структурований і функціонально обґрунтований мовний ресурс. Найбільш поширене визначення корпусу, запропоноване Т. Макенері та А. Вілсоном: за їхнім тлумаченням, корпус – це «зібрання відібраних репрезентативних текстових даних довільної природної мови, що має скінченний обсяг, машиночитану форму та стандартну організацію» [18, с. 5]. О. Демська слушно зауважує: для того, щоб певне зібрання творів могло називатися корпусом, воно повинно мати певні характеристики. Серед них – «мати електронну форму і бути стандартно організованим, тобто відповідати вимогам міжнародного стандарту кодування корпусу; ... бути репрезентативним і збалансованим; ... бути вичерпним; у корпусі

сі ідеально має бути не забагато і не замало певних текстів або їхніх уривків» [2, с. 35].

З-поміж численних характеристик корпусу виокремлюємо такі ключові параметри: машиночитаність, автентичність, репрезентативність, збалансованість, відібраність, стандартизованість, оперативність опрацювання матеріалу тощо.

Машиночитаність – обов’язкова наявність електронної форми, яка забезпечує обробку корпусу спеціалізованими програмами.

Автентичність передбачає, що до корпусу включаються реальні тексти або їх фрагменти, створені носіями мови в умовах природної комунікації – як усної, так і писемної. Дотримання цієї вимоги забезпечує емпіричну достовірність корпусного матеріалу, що є необхідною умовою для об’єктивного аналізу мовних явищ. Автентичність текстів, на думку В. Жуковської, становить одну з ключових передумов наукової валідності корпусного підходу [10, с. 56].

Репрезентативність полягає в тому, що корпус має відображати типові мовні явища, характерні для певної мовної системи або підсистеми. Репрезентативність досягається через жанрову, стилістичну, часову, територіальну та соціальну збалансованість текстів.

Збалансованість – це урахування жанрової, стильової, часової, регіональної, соціальної диференціації мовлення. Наприклад, у національних корпусах враховують пропорції художніх, публіцистичних, наукових, офіційно-ділових і розмовних текстів.

Відібраність передбачає формування корпусу на основі чітких критеріїв (не всі тексти можуть бути включені до корпусу без фільтрації). Корпус формується не шляхом суцільного охоплення всього мовного матеріалу, а через цілеспрямований і мотивований добір текстів із безмежного мовного простору. Жоден корпус, незалежно від його обсягу, не може відтворити всю множину усних і писемних реалізацій мови, що виникають навіть у межах одного дня. Навіть за умов сучасного рівня цифрових технологій повне охоплення мовного континууму залишається недосяжним. Саме тому, як стверджує В. Жуковська, формування корпусу потребує обґрунтованої стратегії відбору, яка ґрунтується на заздалегідь визначених принципах, пов’язаних із типом корпусу та його функціональним призначенням [10, с. 56]. Такий підхід забезпечує репрезентативність і релевантність корпусних даних для конкретних лінгвістичних завдань.

Стандартизованість передбачає, що всі тексти в корпусі оформлені за єдиною системою структурних та лінгвістичних параметрів. Це стосується не лише формату представлення текстових даних, але й принципів розмітки, поділу на одиниці аналізу (речення, абзаци, слововформи), позначення метаданих, таких як авторство, жанр, дата створення, стиль тощо. Дотримання стандартів забезпечує уніфікованість кор-

пусу, що дозволяє здійснювати системний, порівняльний і автоматизований аналіз у межах усього мовного масиву. Стандартизованість також є необхідною умовою для інтеграції корпусу в ширші лінгвістичні платформи й для подальшого повторного використання даних у різних дослідницьких контекстах.

Вагомою ознакою лінгвістичного корпусу, на думку В. Широкова, є «оперативність опрацювання зазначеного матеріалу та можливість прямого доступу до великого числа лінгвістичних фактів» [7, с. 11]. За допомогою корпусних інструментів (конкордансів, частотних словників, тегерів, пошукових запитів) дослідник миттєво отримує доступ до великого обсягу мовних даних. Тобто не потрібно самостійно перечитувати десятки текстів у пошуках мовного явища; замість тривалого ручного аналізу, програмні засоби швидко фільтрують потрібні приклади; результати подаються структуровано й у зручному для аналізу вигляді. Корпус дає змогу дослідникові безпосередньо працювати з емпіричними мовними одиницями: словоформами, фразами, конструкціями – у реальному контексті їх уживання. Це відкриває широкі можливості для спостереження закономірностей, забезпечує репрезентативну базу для лінгвістичних узагальнень, дозволяє перевіряти гіпотези на масиві даних, а не на одиничних прикладах.

Ознаки статичності та динамічності корпусу були вперше сформульовані Г. Кеннеді як важливі типологічні параметри корпусної організації. Згідно з його підходом, статичний корпус репрезентує фіксований обсяг текстів, що охоплюють чітко визначений часовий проміжок і відображають мовний стан на конкретному етапі розвитку мови. Такий корпус, як правило, не поповнюється новими текстами після завершення формування і використовується для синхронного аналізу [16].

Натомість динамічний (моніторинговий) корпус постійно оновлюється – у нього регулярно додаються нові тексти, що дозволяє відстежувати мовні зміни в діахронії. Основна функція такого корпусу – фіксація нових мовних явищ: неологізмів, інновацій у синтаксисі, змін у частотності лексем, колокацій та порядку слів. Завдяки спеціалізованому програмному забезпеченню динамічний корпус дозволяє відслідковувати еволюцію мови в режимі реального часу, що є особливо цінним для соціолінгвістичних і лексикографічних досліджень.

Отже, лінгвістичний корпус – це не просто сховище текстів, а динамічний аналітичний інструмент, що поєднує швидкість обробки і масовість доступу до мовних фактів, що якісно відрізняє його від традиційних методів дослідження.

Крім цих основних параметрів, важливу роль відіграють обсяг, документованість, структурованість і економність, тобто здатність корпусу з максимальною ефективністю відображати мовне середовище без надмірних витрат ресурсів.

Створення повноцінного текстового корпусу – це багаторівневий процес, що потребує як глибокої мовознавчої експертизи, так і залучення цифрових інструментів. Згідно з В. Жуковською, типовий цикл формування корпусу охоплює такі етапи: «визначення джерел лінгвального матеріалу, ...введення даних, ...попереднє опрацювання тексту, конвертування й графематичний аналіз, ...розмітка тексту, ...коректування результатів автоматичної розмітки, ... конвертування розмічених текстів, ... забезпечення доступу до корпусу» [10, с. 85].

Залежно від конкретної мети та умов проєкту, склад і послідовність процедур зі створення корпусу можуть варіюватися, а технологічний процес – бути значно складнішим, ніж наведено у загальних описах. Проте варто враховувати, що формування масштабних довідкових корпусів потребує значних фінансових та технічних ресурсів, тому в найближчій перспективі малоймовірно, що буде започатковано створення принципово нових великих корпусів. Замість цього, як зауважує В. Жуковська, існуючі корпуси, як правило, доповнюються за рахунок приєднання малих спеціалізованих корпусів, які розробляються окремими дослідниками відповідно до вузько окреслених цілей. У процесі побудови корпусу застосовується комплекс лінгвістичних і технічних процедур, зокрема токенізація (виділення мовних одиниць), лематизація (наведення слів до початкової форми), стеммінг (скорочення слів до основи), синтаксичний аналіз (парсинг) та інші [10, с. 88-90]. Ці етапи є важливими для подальшої розмітки й автоматизованого аналізу корпусного матеріалу.

Використання корпусів текстів сьогодні виходить далеко за межі лінгвістики. У своїй сукупності корпуси становлять не лише об'єкт, а й інструмент мовознавчого дослідження, забезпечуючи доказовість, масштабність і надійність лінгвістичного аналізу. Як наголошує Г. Войтко, ці корпуси відіграють ключову роль у мовній освіті, розробці перекладацьких систем, автоматизованому реферуванні, створенні словників, лінгвістичних експертизах [1, с. 210-211].

У сучасному мовознавстві корпусна лінгвістика посідає провідне місце серед інноваційних підходів до аналізу мови, адже вона забезпечує поєднання емпіричної достовірності, масштабності даних і об'єктивності висновків. Однією з найрозвиненіших сфер її застосування є лексикографія. Завдяки корпусам створюються частотні, тлумачні, двомовні та колокаційні словники, у яких відображено реальні вживання слів у різних комунікативних ситуаціях. Це дозволяє уникнути суб'єктивізму й забезпечити репрезентативне та функціональне тлумачення лексичних одиниць. Корпуси також використовуються для вивчення варіантів слововживання, аналізу лексичних новотворів, виявлення змін у значенні слів, що особливо важливо в умовах динамічних мовних процесів.

У перекладознавстві корпусний підхід дав змогу перейти від інтуїтивної до доказової моделі аналізу перекладу. Паралельні корпуси, які містять оригінали й переклади, використовуються для вивчення перекладацьких стратегій, закономірностей відповідності мовних конструкцій, виявлення трансформацій, а також для автоматизації процесів машинного перекладу. Окрім того, корпуси слугують навчальною базою для підготовки перекладачів, даючи змогу простежити стилістичну особливність слововживання.

У межах соціолінгвістики корпуси забезпечують реальний емпіричний матеріал для вивчення мовної поведінки представників різних вікових, соціальних, професійних і територіальних груп. Це дозволяє досліджувати діалекти, професійні жаргони, гендерні особливості мовлення та взаємодію мови й соціального середовища. Такі корпуси стають основою для аналізу мовної динаміки в суспільстві, зокрема змін у мовленнєвій нормі, популяризації нових форм комунікації, зумовлених впливом медіа, інтернету тощо.

Корпусний підхід є надзвичайно цінним і в лінгводидактиці, особливо у викладанні іноземних мов. Використання частотних словників і корпусів дозволяє створювати навчальні матеріали на основі автентичного мовлення, добирати реальні приклади вживання слів і граматичних конструкцій, формувати навчальні тексти відповідно до мовних норм. Це сприяє оптимізації змісту навчання та наближенню викладання до живого мовного середовища.

У стилістиці корпуси відкривають нові можливості для аналізу мовних засобів, характерних для різних жанрів і авторських стилів. Дослідники мають змогу кількісно вимірювати частоту вживання риторичних засобів, синтаксичних конструкцій, стилістичних маркерів, порівнювати авторські стратегії та жанрову специфіку. Корпусна стилістика дозволяє виявляти складні функціональні закономірності в організації мовного тексту, що раніше залишалися поза межами традиційних методів аналізу.

Однією з важливих прикладних сфер застосування корпусної лінгвістики є судова лінгвістика – міждисциплінарна галузь, що перебуває на перетині мовознавства та юриспруденції. У цьому контексті корпуси виступають інструментом для аналізу мовних даних у правовій практиці, зокрема в таких завданнях, як експертиза авторства тексту, виявлення мовної маніпуляції, оцінка лексичної або стилістичної відповідності нормативним зразкам, а також для інтерпретації правових термінів і формулювань. Корпусні методи дозволяють зіставити спірні мовні одиниці з широким контекстом їх уживання у типових ситуаціях, а також кількісно обґрунтувати експертні висновки.

Зокрема, як зазначають Т. Макенері та А. Гарді, корпуси можуть бути використані для виявлення аномальних або нехарактерних мовних патернів, що є ключовим під час визначення ймовірного автора докуме-

нта або у справі про плагіат [17, с. 103–105]. Окрім того, судова лінгвістика в умовах цифрової доби активно залучає корпуси для аналізу мови в мережі Інтернет, зокрема у випадках мовленнєвої агресії, кібербулінгу, поширення мови ворожнечі тощо. У подібних справах корпус дозволяє обґрунтовано порівняти спірне висловлювання з типовими мовними шаблонами або нормами публічного дискурсу.

Варто зазначити, що корпусний аналіз має перевагу над традиційною інтуїтивною лінгвістичною експертизою, оскільки дає змогу об'єктивізувати висновки через статистичну репрезентативність. Це особливо важливо в правовому середовищі, де від результатів мовної експертизи можуть залежати судові рішення.

Не менш важливою є роль корпусної лінгвістики в мовній варіантології та нормотворенні. Корпусні дані використовуються для вивчення внутрішньомовної варіативності на всіх мовних рівнях – лексичному, морфологічному, синтаксичному, а також для формулювання обґрунтованих мовних стандартів. Це особливо актуально в контексті кодифікації мовної норми та укладання академічних граматик, словників, правописних стандартів [4].

У сучасному мовознавстві лінгвістичні корпуси дедалі частіше розглядаються не лише як інструмент академічного аналізу, а й як засіб реалізації державної мовної політики. Створення корпусних ресурсів забезпечує емпіричну базу для розробки, кодифікації й уніфікації мовної норми, а також для моніторингу стану мови в публічному просторі, освіті, медіа, діловій та науковій сферах.

Таким чином, корпусна лінгвістика значно розширює інструментарій мовознавства, забезпечуючи кількісну базу, можливість об'єктивного спостереження, повторюваність результатів і доступ до масивів реального мовлення. Її застосування охоплює не лише класичні філологічні напрями, а й активно впроваджується в цифрову гуманітаристику, лінгвістичну інженерію та міждисциплінарні дослідження.

Як зазначає дослідниця А. Таран, створення сучасної моделі системи української мови вимагає не лише аналізу формальних та семантичних характеристик лексичних одиниць і їх функціонування в текстах, а й розроблення ефективних інструментів для моніторингу мовної динаміки. На її думку, «важливою особливістю даних Інтернету як лінгвістичного джерела є можливість безпосереднього звернення до контекстів, у яких зафіксовані вживання потрібного слова» [9, с. 102]. Це відкриває шлях до переходу від суто кількісних характеристик до якісного аналізу мовних одиниць, включно з їх семантичними та стилістичними варіантами. Водночас дослідниця підкреслює, що навіть потужні корпуси та пошукові системи не здатні повною мірою відобразити всю мовну динаміку. Нинішні пошукові інструменти, за її спостереженням, потребують суттєвого вдосконалення, зокрема у впорядкуванні результатів пошуку й

реалізації фільтраційних механізмів [9, с. 102]. У зв'язку з цим А. Таран вважає за доцільне створення спеціалізованих пошукових систем, орієнтованих на виконання наукових завдань мовознавства.

Корпусна лінгвістика в Україні перебуває на етапі активного становлення. Найвідомішим проєктом є Національний корпус української мови, який містить понад 100 мільйонів словоформ, має розмітку, систему пошуку, а також розділи для мовних рівнів. Український національний лінгвістичний корпус (УНЛК) – це масштабний проєкт, що реалізується під егідою Інституту української мови НАН України з метою створення інформаційної лінгвістичної платформи для наукового та прикладного використання. Корпус покликаний забезпечити репрезентативне зібрання текстів української мови, які структуровано й опрацьовано відповідно до сучасних вимог корпусної лінгвістики – зокрема з урахуванням репрезентативності, стандартизованості, уніфікованості, багаторівневої розмітки та відкритого доступу. УНЛК включає тексти різних жанрів і функціональних стилів (художні, наукові, публіцистичні, офіційно-ділові, розмовні), охоплює різні часові періоди, що дає змогу здійснювати діахронічний аналіз мовних змін. Корпус також містить морфологічну й частиномовну розмітку та будується на основі принципів відібраності, автентичності й системної репрезентативності. Проєкт спрямований на формування нормативної бази для створення словників, граматики, правописних стандартів і навчальних матеріалів, а також на забезпечення повноцінної участі українського мовознавства в загальноєвропейських корпусних ініціативах. О. Демська зазначає, що «створення національного корпусу української мови є необхідною умовою для повноцінної участі українського мовознавства в загальноєвропейських корпусних ініціативах» [4, с. 9].

Отже, корпусна лінгвістика на сучасному етапі розвитку постає як одна з найдинамічніших і найперспективніших галузей мовознавства, що поєднує теоретичну глибину з прикладною ефективністю. Її поява стала можливою завдяки стрімкому розвитку комп'ютерних технологій, які відкрили нові горизонти для лінгвістичного аналізу, забезпечивши роботу з великими масивами автентичного мовного матеріалу. В основі корпусного підходу лежить принцип емпіричності, що дозволяє мінімізувати вплив суб'єктивізму дослідника та базувати лінгвістичні висновки на об'єктивно зафіксованих фактах мовного вжитку.

Отже, корпусна лінгвістика є не лише інструментом для аналізу мовних явищ, а й потужним чинником трансформації самого мовознавства, його методів, підходів та практичних реалізацій. Її подальший розвиток тісно пов'язаний із міждисциплінарною співпрацею, технологічним прогресом та розширенням доступу до цифрових мовних ресурсів. В умовах цифровізації науки корпусна лінгвістика стає ключовим джерелом для формування нової епістемології в лінгвістиці XXI століття.

В українській та у світовій лінгвістичній практиці простежується виразна тенденція до розширення типів корпусів, удосконалення методологій розмітки та глибшої міждисциплінарної інтеграції. Очікується активне формування нових корпусів усного мовлення, регіональних діалектів, історичних текстів і професійних жаргонів, що дозволить значно поглибити розуміння мовної варіативності. Водночас дедалі більшого значення набуває поєднання корпусних підходів із технологіями штучного інтелекту, машинного навчання та глибинного аналізу даних, що відкриває нові можливості для автоматичного синтаксичного, семантичного та прагматичного аналізу. У глобальному контексті зростає роль національних корпусних платформ як інструментів державної мовної політики, лексикографії та стандартизації. Для України важливим є не лише продовження розробки й розширення корпусів, зокрема Українського національного лінгвістичного корпусу, а й забезпечення відкритого доступу до корпусних ресурсів, їх популяризація в академічному середовищі, інтеграція в освітній процес, участь у міжнародних ініціативах. Корпусний підхід також дедалі ширше впроваджується в мовну освіту: створюються частотні словники, корпуси учнівських текстів, автоматизовані системи оцінювання. Усе це свідчить про перехід корпусної лінгвістики до нового етапу – як фундаментальної, так і прикладної науки, яка визначатиме цифрову парадигму розвитку лінгвістики XXI століття.

Література

1. Войтко Г. Корпусна лінгвістика: історія формування і перспективи розвитку. *Теоретичні засади лінгвістичних досліджень*. 2014. № 6. С. 208-214.
2. Демська О. М. Текстовий корпус: поняття і визначення. *Дивослово*. 2011. № 10. С. 35-37.
3. Демська-Кульчицька О.М. Британський національний корпус і національний корпус української мови. *Філологічні студії*. Луцьк, 2004. № 4. С. 89-95.
4. Демська-Кульчицька О. Основи національного корпусу української мови: монографія. К.: Інститут української мови НАН України, 2005. 218 с.
5. Карпіловська Є. Лексична картотека і корпус як інструмент лексикографічного моделювання слова. *Лексикографічний бюлетень*. 2015. Вип. 24. С. 15-22. Режим доступу: http://nbuv.gov.ua/UJRN/lecbu_2015_24_5
6. Лук'янець Г. Основні напрямки сучасних корпусних досліджень мови. *Наукові праці НУХТ*. 2012. № 44. С. 127–132.
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. Луцак С. М., Юрчак Г. М. Тенденції розвитку прикладної лінгвістики в контексті розвитку наукових галузей. *Digital transformations in culture: Scientific monograph*. Riga, Latvia: «Baltija Publishing», 2023. Р. 255-268.
9. Таран А. Пошукова система й корпус як інструмент дослідження семантики слів. *Вісник Черкаського національного університету імені Богдана Хмельницького*. Серія: Філологічні науки. 2019. Вип. 1. С. 99-103. Режим доступу: http://nbuv.gov.ua/UJRN/VchuF_2019_1_16
10. Жуковська В. В. Вступ до корпусної лінгвістики. Житомир: Вид-во ЖДУ ім. І. Франка, 2013. 142 с. Режим доступу: <https://doi.org/10.30525/978-9934-26-319-4-15>
11. Baker M. Corpora in translation studies. An overview and suggestions for future research. *Target*. 1995. № 7(2). Р. 223-243.
12. Biber D., Conrad S., Reppen R. *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press, 2001. 312 p.
13. Busa R. The Annals of the Index Thomisticus. In: *Computers and the Humanities*. 1980. 14(2), pp. 83–90.
14. Chomsky N. *Syntactic Structures*. Second Edition With an Introduction by David W. Lightfoot. Mouton de Gruyter. Berlin, New York, 2002. 119 p. https://tallinzen.net/media/readings/chomsky_syntactic_structures.pdf
15. Hunston S. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press. 2012. 477 p.
16. Kennedy G. *An Introduction to Corpus Linguistics*. II. Series: Studies in language and linguistics (London, England). Routledge, 1998. 328 p.
17. McEnery T., Hardie, A. *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press, 2012. 294 p.
18. McEnery T., Wilson A. *Corpus Linguistics. An Introduction*. Edinburgh: Edinburgh University Press, 2001. 234 p. https://uogbooks.wordpress.com/wp-content/uploads/2014/10/tony_mcenery_andrew_wilson_corpus_linguisticsbook4you-org.pdf
19. *The Brown Standart Corpus of American English*. Brown, 1964. 386 p.

THEORETICAL FOUNDATIONS OF CORPUS LINGUISTICS: PRINCIPLES OF TEXT CORPUS CONSTRUCTION

Svitlana Lutsak, Halyna Yurchak

Ivano-Frankivsk National Medical University;

76018, Ivano-Frankivsk, Halytska St., 2;

e-mail: sv.lutsak@gmail.com, galuna.yurchak@ukr.net

This research article explores the theoretical foundations of corpus linguistics, outlines key principles of text corpus construction, and identifies the

main directions of contemporary corpus-based research. Based on the analysis of works by Ukrainian and international scholars, the definition of a "text corpus" is clarified, and its key parameters are specified (machine-readability, authenticity, representativeness, balance, selectivity, standardization, and processing efficiency, among others). The article reviews the typical stages of corpus development such as tokenization, lemmatization, syntactic parsing, and annotation, as well as challenges in building large-scale corpus resources. Special attention is given to practical applications of corpora across fields such as lexicography, translation studies, sociolinguistics, language education, stylistics, forensic linguistics, and language variation analysis. Emphasis is placed on the significance of the corpus-based approach in modern linguistics, particularly the shift from introspection toward empirical analysis of authentic language data. The article also outlines the main requirements for corpora, provides an overview of typological features, and discusses the prospects for further development of corpus research in Ukraine and worldwide.

Keywords: corpus linguistics, text corpus, representativeness, machine-readability, corpus annotation, digital technologies.