ПРИКЛАДНА ЛІНГВІСТИКА

8. Корпусна лінгвістика: історія становлення та сучасний стан

Жуковська В.В.

Сутність, предмет і завдання корпусної лінгвістики

Корпусна лінгвістика є одним із найбільш перспективних напрямків сучасного теоретичного і прикладного мовознавства. Ця відносно нова лінгвістична галузь розпочала своє активне становлення у 60-х роках XX століття у зв'язку із інтенсивним розвитком комп'ютерних технологій. Сам термін "корпусна лінгвістика" міцно ввійшов до наукового вжитку лише в останні десятиліття XX століття з публікацією у 1983 році збірника наукових праць "Corpus Linguisitcs: 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". Звичайно, застосування комп'ютерів та спеціального програмного забезпечення суттєво змінило спосіб дослідження мови та значно полегшило роботу по збору лінгвістичних даних. Без зусиль, лише за декілька секунд, стало можливим здійснювати пошук у багатомільйонних текстових масивах (лінгвістичних корпусах), будувати конкорданс для будь-якого слова, одержувати дані про частоту словоформ, лексем, граматичних категорій, синтаксичних конструкцій, відстежувати зміни у частоті і контексті мовної одиниці у різні хронологічні періоди, одержувати дані про лексичну (колокацію) та граматичну (колігацію) сполучуваність і т. ін. Створення емпіричної бази дослідження таким способом забезпечує мінімум витрат зусиль і часу на зібрання даних, обсяг яких водночас суттєво збільшується.

Сьогодні дані корпусів масштабно використовуються в лексикографії, стилістиці, судовій лінгвістиці, лінгвістичній варіантології, перекладознавстві, соціолінгвістиці, методиці навчання і вивчення іноземної мови та в багатьох інших лінгвістичних дослідженнях [Baker 2006, с. 2-3].

Корпусна лінгвістика як галузь прикладного мовознавства займається визначенням загальних принципів побудови, обробки та експлуатації даних лінгвістичних корпусів (корпусів текстів) із використанням сучасних комп'ютерних технологій, розробленням методики збору реальних мовних явищ – писемних та усних текстів, а також способів їх збереження та аналізу. Під ***корпусом текстів*** розуміється значний за обсягом, представлений в електронному вигляді, уніфікований, структурований, розмічений, філологічно компетентний масив мовних даних, створений для вирішення конкретних лінгвістичних завдань [Захаров 2005, с. 3].

Корпусний аналіз вирізняється низкою характерних ознак: 1) емпіричний підхід до аналізу мовних даних (досліджуються реальні моделі мовної реалізації у природних текстах); 2) використання великих за обсягом, структурованих колекцій природних текстів (корпусів) як основи для аналізу; 3) широке залучення комп'ютерних технологій для дослідження лінгвального матеріалу; 4) застосування кваліативних і квантитативних аналітичних методик, з суттєвою перевагою останніх (вивчення частоти вживання лінгвістичних одиниць, статистичні дослідження сполучуваності і т.ін.). Отримані в результаті корпусного аналізу дані не тільки сприяють формулюванню якісно нових висновків про мову, але й окреслюють такі напрями досліджень, які до появи корпусів не привертали уваги дослідників. Спираючись головним чином на реальний "живий" мовний матеріал, а

не на мовну інтуїцію та інтроспекцію, корпусні дослідження дозволяють абстрагуватися від суб'єктивності дослідника і наблизитися до об'єктивного вивчення мови. І, як стверджує Л. Ричкова, матеріал корпусу дозволяє не лише оптимізувати і об'єктивізувати лінгвістичні дослідження, але і по-новому окреслити багато традиційних лінгвістичних понять [Ричкова 2003, с. 185]. Корпусні розвідки переорієнтовують традиційний підхід до вивчення мови, а результати аналізу даних корпусу сприяють переоцінці низки лінгвістичних теорій [MacEnery, Hardie 2012, с. 1].

Численні розвідки в галузі сучасного корпусного мовознавства відбуваються у двох магістральних напрямках. Перший напрям зосереджений на розробці проблем, що стосуються теорії та практики створення корпусів, іншими словами, концептуалізації корпусу: типологія корпусу, його призначення, обсяг, параметризація предметної галузі, репрезентативність, структурування та принципи відбору базових одиниць, зберігання і т. ін. А другий напрям спрямований на дослідження саме лінгвістичних корпусів, тобто вивчення мови за допомогою корпусних методів [Копотев, Мустайоки 2008, с. 12]. Проте слід зауважити, що чіткої межі між зазначеними напрямами не існує, адже практично всі укладачі корпусів в той же час здійснюють і лінгвістичні дослідження на їх основі. Така двовекторність корпусної лінгвістики зумовлюється подвійною природою ***об'єкта*** її дослідження – текстового корпусу, який, з одного боку, виступає в якості вихідного мовленнєвого матеріалу для корпусної лінгвістики, а з іншого, є результатом діяльності цього мовознавчого напряму. ***Предметом*** корпусної лінгвістики виступають теоретичні основи і практичні механізми створення та експлуатації мовних корпусів [Захаров, Богданова 2011, с. 10].

Головне ***завдання*** корпусної лінгвістики вбачається у повному й системному відображенні змістовного спілкування мовою. Важливою особливістю цього інформаційно-семіотичного напряму лінгвістичних досліджень є підхід до розгляду прикладних проблем лінгвістики конкретно в комунікативних процесах. При цьому в центрі уваги виявляється не мова як система, і не проблема її формалізації, а процес змістовного спілкування мовою, і по можливості точний її опис, який може бути використаний для розв'язання науково-технічних завдань інформатики.

Першочерговою ***метою*** цієї мовознавчої галузі є об'єктивний лінгвістичний опис мовної системи, причому до цього опису корпусна лінгвістика підходить від вивчення конкретної людської комунікації. У якості другорядної цілі розглядається вироблення особливого способу відображення мовного матеріалу в корпусі текстів. Цей спосіб, у свою чергу, може використовуватися й іншими лінгвістичними дисциплінами.

Теоретичним підґрунтям корпусної лінгвістики є структуралізм, який декларує примат реального тексту в лінгвістичному дослідженні. Для корпусних розвідок головним є постулат, що мова як об'єкт дослідження може бути вивчена лише у формі писемних та усних текстів [Демьска 2010, с. 6]. У тексті, як формі існування мовлення, існує лише те, що передбачено системою мови, тому корпусна лінгвістика розглядає текст як засіб експлікації специфіки міжелементних мовних зв'язків. Дослідження тексту поглиблює розуміння суті природної мови.

Все вище сказане чітко окреслює дослідницьку програму 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. Застосування комп'ютерів дозволяє миттєво обробити величезний обсяг мовного матеріалу і відібрати всі можливі у конкретному корпусі приклади вживання необхідних для аналізу одиниць. У розпорядження лінгвіста надаються об'єктивні кількісні дані, забезпечуючи досягнення більш ґрунтовних та переконливих висновків. Корпусна лінгвістика дозволяє вченим підтвердити або спростувати гіпотези про функціонування мови, а також окреслити нові напрями дослідження, які до застосування корпусних методів не попадали до фокусу уваги дослідників.

Корпусна лінгвістика в системі мовознавчих наук

Незважаючи на значні досягнення та популярність, статус корпусної лінгвістики у сучасній мовознавчій парадигмі ще не є однозначно визначеним (див. наприклад: [Biber, Conrad, Reppen 1998, с. 3; Kennedy 1998, с. 2; McEnery, Wilson 2001, с. 2; Meyer 2002, с. xi]. Так, Т. МакЕненрі, Е. Вільсон, Р. Ксіао та інші лінгвісти характеризують її як особливу "методологію" (*methodology*) дослідження мовного матеріалу, на противагу таким традиційним галузям лінгвістики, як фонетика, синтаксис, семантика чи граматика. Зазначені мовознавчі галузі досліджують окремий аспект мовної системи, натомість корпусна лінгвістика не обмежується вивченням лише одного аспекту мовної реалізації. Більш того, корпусні методи можуть використовуватися для вивчення мовних одиниць будь-якого мовного рівня. Наприклад, синтаксичні явища можливо дослідити як із застосуванням корпусних методик, так і без них, так само існують корпусні і некорпусні семантичні дослідження [McEnery, Wilson 2001; McEnery, Xiao, Tono 2006, с. 6].

Відзначається, що корпусна лінгвістика як система методів і принципів використання корпусів для вивчення мови та для вивчення/навчання мові має теоретичне підґрунтя, але саме по собі теоретичне підґрунтя ще не є науковою теорією. Звідси, корпусна лінгвістика визнається методологією вивчення мови з широкими можливостями застосування в багатьох лінгвістичних напрямах та теоріях [McEnery, Xiao, Tono 2006, с. 7].

Проте більшість вітчизняних та зарубіжних лінгвістів не погоджується з цією точкою зору і розглядає корпусну лінгвістику як самостійну дисципліну прикладного мовознавства з власним предметом, об'єктом і метою дослідження, терміноапаратом, теоретичним і практичним підґрунтям, власними дослідними прийомами [Демська 2005, с. 8, 12; Баранов 2001, с. 112–117]. На переконання російських лінгвістів В.П. Захарова та С.Ю. Богданової, корпусна лінгвістика має принаймні дві ознаки, що дають їй підставу претендувати на статус самостійної дисципліни: 1) характер аналізованого словесного матеріалу; 2) специфіка інструментарію [Захаров, Богданова 2011, с. 9]. Сучасна корпусна лінгвістика вже "вийшла зі статусу методології" та стала окремою "дисципліною" [Tognini-Bonelli 2001, с. 1], сформувалася в самостійний науковий напрямок, досягнення якого знаменують новий етап розвитку наукової думки [Колпакова 2011, с. 1].

Корпусна лінгвістика як емпіричний мовознавчий напрям суттєво відрізняється від традиційної лінгвістики підходами та методами вивчення мовного матеріалу та демонструє виразну переорієнтацію об'єкта дослідження з "системи" на "узус", з "мови" на "мовлення".

Оскільки корпусні дослідження проводяться із широким залученням комп'ютерів та спеціально розроблених комп'ютерних програм, корпусна лінгвістика часто ототожнюється з іншою галуззю прикладної

лінгвістики – комп'ютерною лінгвістикою. Хоча ці два напрями мають багато спільного, проте вони не ідентичні. Комп'ютерна лінгвістика, як окрема прикладна лінгвістична дисципліна, характеризується обов'язковістю використання комп'ютерних засобів до оброблення лінгвальних даних, це широка сфера використання комп'ютерних інструментів – програм, комп'ютерних технологій організації та обробки даних – для моделювання функціонування мови в різних умовах, ситуаціях, проблемних галузях, а також сфера застосування комп'ютерних моделей не лише в лінгвістиці, а й у суміжних із нею дисциплінах [Баранов 2001, с. 13]. Комп'ютерна лінгвістика займається вирішенням таких проблем, як автоматичний переклад, автоматизоване добування інформації з природних текстів, конструювання зручних інтерфейсів між людиною та машиною, кількісний опис спілкування на природних мовах.

Для корпусної лінгвістики застосування комп'ютерних інструментів не є визначальним критерієм, але, на відміну від комп'ютерної лінгвістики, корпусна лінгвістика покликана не моделювати функціонування мови в різних умовах, ситуаціях, проблемних галузях та послуговуватися цими моделями, а лише фіксувати всі аспекти функціонування мови, зберігаючи як інтра-, так і екстралінгвістичну специфіку, забезпечуючи оптимальну адекватність лінгвальних даних [Демська 2005, с. 14].

Однак в певних аспектах комп'ютерна і корпусна лінгвістики тісно взаємодіють і доповнюють одна одну. Так, комп'ютерна лінгвістика створює необхідні інструменти для корпусної лінгвістики, тобто спеціалізовані комп'ютерні програми для аналізу та обробки великих масивів мовних даних. Це так звані парсери, тегери, конкордансери та інш.

Типологія досліджень у царині корпусного мовознавства

Сьогоднішня корпусна лінгвістика – це гетерогенна область дослідження мови, всередині якої виокремлюються окремі піднапрями, що різняться підходами до конструкції, експлуатації корпусів та аналізу корпусних даних. В основі виділення цих під напрямів знаходяться такі параметри [McEnery, Hardie 2012, с. 3-21]:

- ***формат представлення текстів у корпусі (mode of communication)***. Корпуси можуть містити тексти, представлені в усній чи писемній формі. В залежності від форми представлення текстів виділяють корпуси усного мовлення, корпуси писемного мовлення та корпуси змішаного типу. Нові типи корпусів, такі як мультимедійні корпуси та корпуси кінетичного мовлення, реєструють також і паралінгвістичні засоби, які супроводжують спілкування або є безпосереднім способом спілкування. Відмінності у формі представлення текстів у корпусі обумовлюють особливі підходи до відбору та обробки лінгвального матеріалу, використання спеціального програмного забезпечення для аналізу даних та виявляють суттєві лінгвістичні відмінності між отриманими даними в результаті аналізу певного типу корпусу;

- ***корпуснобазовані (corpus-based)*** vs. ***корпуснокеровані (corpus-driven) дослідження.*** У корпуснобазованих дослідженнях дані корпусу використовуються для доведення, спростування чи уточнення визнаної на певному етапі розвитку наукової думки теорії чи гіпотези. Цей тип корпусних розвідок трактує корпусну лінгвістику як метод дослідження мови. Натомість корпуснокерована лінгвістика відмовляється від визнання корпусної лінгвістики як методу і стверджує, що корпус сам по собі є єдиним джерелом гіпотез про мову та втілює свою теорію мови;

- ***режим накопичення даних у корпусі (data collection regimes).*** Відбір даних до корпусу та їх кількість безпосередньо залежать від тих лінгвістичних завдань, заради вирішення яких і створюється корпус. Звідси питання обмеженості / необмеженості обсягу корпусу є одним із важливих питань корпусної лінгвістики. Виділяють два широкі підходи до режиму накопичення даних у корпусі: 1) підхід моніторингового корпусу (Дж. Синклер [Sinclair 1991]), згідно якого розмір корпусу пропорційно зростає з плином часу; 2) підхід статичного корпусу (Д. Байбер [Biber 1993] та Дж. Ліч [Leech 2007]), згідно якого ретельно організований корпус має кінцевий розмір і засвідчує стан мови на певному синхронному зрізі;

- ***використання анотованих (annotated) / неанотованих (unannotated) корпусів***. Головною відмінністю сучасного корпусу є наявність анотації, тобто спеціальних міток, що приписуються словам у текстах корпусу та позначають різноманітні лінгвістичні категорії, наприклад, граматичні, синтаксичні і т. інш. Анотація може бути внесена безпосередньо до корпусу, а може супроводжувати корпус окремим документом;

- ***повне врахування (total accountability)*** vs ***відбір даних (data selection)***. Принцип повного врахування в корпусному дослідженні полягає у неприпустимості вмотивованого відбору даних із корпусу з метою уникнення фальсифікації відомостей для підтвердження досліджуваної гіпотези/теорії.

Згідно принципу відбору пошук у корпусі здійснюється з метою підбору специфічного прикладу чи низки ретельно відібраних прикладів для спростування висунутої гіпотези. Такий спосіб експлуатації корпусу є сумісним із принципом повного врахування. Тож використання лише певних частин корпусу в якості банку прикладів для ілюстрації теорії, що розвивається

дослідником (так звані корпусноінформовані дослідження (corpus-informed)), без врахування решти даних корпусу може спотворити реальний стан речей і дозволить досліднику маніпулювати даними корпусу заради підтвердження власної (можливо і хибної) теорії;

- ***багатомовні (multilingual)*** vs ***одномовні (monolingual) корпуси.*** Іншим критерієм, що розрізняє типи корпусів є кількість мов, представлених у корпусі. Більшість корпусів є одномовними, в тому сенсі, що вони репрезентують лінгвістичну варіативність певної однієї мови. Натомість багатомовні корпуси – це корпуси, що побудовані на матеріалі двох або більше мов. У вузькому розумінні багатомовним корпусом взагалі вважається корпус, укладений принаймні на трьох мовах, а корпуси, що містять дані двох мов, називаються двомовними (*bilingual*) корпусами. Багатомовні корпуси у свою чергу підрозділяються на порівняльні корпуси (*comparable*) та паралельні (*parallel*) корпуси.

Історія становлення корпусної лінгвістики: від паперових конкордансів і картотек до перших електронних корпусів

Традиційно вважається, що корпусна лінгвістика почала активно формуватися як самостійна галузь у 60-х роках XX століття, коли з'явилися перші корпуси у комп'ютерному форматі. У. МакЕнері та А. Вільсон значно віддаляють у часі дату зародження корпусної лінгвістики. Вчені виокремлюють два періоди в корпусних розвідках. Перший період – це стадія ранньої корпусної лінгвістики (1910–1960-ті рр.), коли відбувається формування теоретичного підґрунтя та прагматичних передумов виникнення напряму й створення текстових зібрань для лінгвістичного дослідження переважно на паперових носіях. Другий період (починається з 1960 рр.) характеризується

інтенсивним піднесенням корпусних студій і безпосередньо пов'язаний із значним розвитком комп'ютерних технологій.

Проте, якщо корпусна лінгвістика – це вивчення мови, що базується на прикладах реального вживання мови [McEnery, Wilson 2001, с. 1], то необхідно констатувати, що масштабні емпіричні дослідження мови з використанням великого обсягу реальних мовних даних здійснювалися і в більш ранні періоди розвитку мовознавчої науки. Наука про мову поступово наближалася до ідеї текстового корпусу і до самого корпусу в такому вигляді, в якому ми зараз його знаємо. Як дотепно висловився Я. Свартвік [Svartvik 2007, с. 12], в історії корпусів спочатку був так званий "кам'яний вік", або докомп'ютерний період, коли корпуси укладалися вручну на папері. Ці перші паперові доелектронні корпуси по суті були конкордансами, тобто алфавітними списками всіх ужитих у певному тексті/текстах слів у їх контекстному оточенні. Укладання таких паперових корпусів-конкордансів займало багато часу та вимагало напруженого аналізу, який здійснювався вручну. Найвагомішу роль доелектронні корпуси відіграли в таких лінгвістичних проектах, як укладання конкордансів Біблії й літературних творів, а також написання граматик і словників [Meyer 2008, с. 1].

Найважливішим та найвпливовішим доелектронним корпусом вважається The Survey of English Usage, укладений Рендольфом Квірком у 1959 р. в University College London. Корпус представляв собою велику базу даних на картонних картках, які містили зразки щоденного спілкування (як писемного, так і усного) звичайних громадян. Цей проект став перехідним етапом у становленні корпусної лінгвістики. Спочатку збереження даних в електронному форматі не передбачалося і лише через багато років по тому відбулося перенесення даних

корпусу в електронну форму. Сьогодні базу даних корпусу можна переглянути на офіційному сайті проекту[1].

Корпуси першого покоління. Ідея створення корпусу (уже в сучасному його розумінні) зародилася в 60-х роках XX століття під значним впливом здійснених масштабних емпіричних досліджень, про які ми вже зазначали. До кінця 1960-х існувало декілька невеликих корпусів, укладених на різних принципах. За сучасними стандартами вони були дуже малими, але вирізнялися ретельно продуманою організацією.

Першим комп'ютерним корпусом є одномільйонний Браунівський корпус (the Brown Corpus)[2], укладений у Браунівському університеті (США) лінгвістами Нельсоном Френсісом та Генрі Кучероюю Створення корпусу мало на меті дослідження лінгвістичних особливостей американського варіанту англійської мови. Він містив 500 текстових уривків обсягом по 2 000 слововживань загальним обсягом біля 1 млн. слів. Тексти булі відібрані із п'ятнадцяти найбільш масових жанрів англомовної друкованої прози США, що були опубліковані в 1961 році. Корпус супроводжувався значною кількістю матеріалів його первинної статистичної обробки – частотний і алфавітно-частотний словник, різноманітні статистичні розподіли. Перша версія корпусу була представлена простим текстовим форматом (з невеликою кількістю структурної розмітки для виділення абзаців, заголовків, цитованих фрагментів і т. ін.). Пізніше у 1980 році корпус був доповнений розміткою частин мови та морфологічних ознак слів. Поява Браунівського корпуса викликала загальний інтерес у науковій спільноті та стимулювала наукову дискусію, в якій обговорювалися критерії відбору текстів та потенційні можливості корпусу для наукового

[1] http://www.ucl.ac.uk/english–usage
[2] http://www.helsinki.fi/varieng/CoRD/corpora/BROWN/index.html

аналізу. Корпус швидко перетворився у популярний об'єкт лінгвістичних досліджень. Поступово в процесі його використання вчені дійшли до розуміння того, що провести певні порівняння і виявити конкретні закономірності можливо лише шляхом аналізу значних за розміром масивів текстів, які організовані за визначеними правилами. Так почали проводитися нові дослідження мови вже на більш високому і надійному рівні в межах нового напряму в лінгвістиці, яким стала корпусна лінгвістика.

Браунівський корпус перетворився у певний стандарт для створення інших аналогічних корпусів. Услід за ним з'явився його британський аналог – Ланкастерсько-Осло-Бергенський корпус (the Lancaster-Oslo-Bergen Corpus[3]), що був названий за місцем його створення. Його укладачі притримувались основних принципів укладання Браунівського корпусу – така ж сама середня довжина текстів, такі ж самі способи його представлення на машинному носії (15 жанрів вже британської масової друкованої продукції – 1961 року видання). Анотована версія корпусу з'явилася у 1985 році. Створення Браунівського та Ланкастерського корпусів дало можливість проводити різноаспектні філологічні порівняння двох варіантів англійської мови (американського і британського) на текстах різних жанрів, доступних комп'ютерній обробці.

У 1970-і за моделлю близькою до Браунівського корпусу був створений частотний словник російської мови, укладений на основі корпусу текстів обсягом також в 1 млн. слів, що включав приблизно в однаковому співвідношенні суспільно-політичні, художні, наукові та науково-популярні тексти і драматургію. Корпус містив внесену вручну лематизацію та частиномовну розмітку.

[3] http://www.helsinki.fi/varieng/CoRD/corpora/LOB/index.html

За форматом Браунівського та Ланкастерсько-Осло-Бергенського корпусів з деякими модифікаціями було укладено низку інших корпусів, серед яких the Kolhapur Corpus of Indian English, the Wellington Corpus of Written New Zealand English, the Australian Corpus of English, the Corpus of English-Canadian Writing, the Standard Corpus of Present-day English Language Usage, the London-Lund Corpus (LLC) (детальніше див.: http://www.uow.edu.au/~dlee/CBLLinks.htm).

Корпуси другого покоління – це продукти Інтернету і характеризуються значним обсягом. Так, у кінці 80-х років у Великобританії був створений перший мега-корпус, що задав новий стандарт для представницьких корпусів – Британський національний корпус (British National Corpus[4]). Цей корпус характеризується обсягом 100 млн. слів, використанням повних текстів, а не вибірок з текстів, підкорпусом усного мовлення (10 млн. слів), наявністю частиномовної розмітки та доступом через Інтернет. Для корпусу використовувалася детальна класифікація документів за декількома параметрами: вид мовлення (писемне, усне приватне і усне публічне), для писемного за тематикою, типом видання (книги, періодика, машинописні тексти і т.п.), параметром утворення очікуваної аудиторії (високий, середній чи довільний) та складністю мови (складний, середній, простий).

За заданим Британським національним корпусом стандартом були укладені представницькі корпуси багатьох європейських мов. За цією моделлю були створені національні корпуси іспанської, італійської, хорватської, чеської мов.

Подібний проект Банк англійської мови (the Bank of English) розпочався у 1980-і рр. У 1989 році його обсяг був 20 млн. слів, а у 2012 – 650 млн. слів. Банк англійської

[4] http://www.natcorp.ox.ac.uk/

мови – це так званий моніторинговий корпус, що покликаний відслідковувати мовні зміни шляхом регулярного поповнення новими текстами та порівняння частотних параметрів, наприклад, таких, як зміна частоти слів та граматичних конструкцій, поява нових слів і т.ін. Він охоплює англійське писемне та усне мовлення, а також різні територіальні варіанти англійської мови.

Ще один вид корпусів представлений Інтернаціональним корпусом англійської мови (the International Corpus of English[5]). Цей проект розпочався у 1990 році з метою здійснення порівняльного дослідження існуючих варіантів англійської мови. Двадцять центрів у всіх куточках світу працюють над створенням підкорпусів власного національного чи територіального варіанту англійської мови, дотримуючись єдиного принципу побудови та граматичної анотації.

Успіх Британського національного корпусу у дослідженні британського варіанту англійської мови дав поштовх для створення американського відповідника. У 1999 році було засновано консорціум Американського національного корпусу (the American National Corpus)[6]. У 2003 році була випущена перша частина (близько 10 млн. слів) корпусу, а за декілька років по тому була випущена і друга частина (близько 20 млн. слів). Особливою прикметою Американського національного корпусу є те, що він містить жанри, які ще не були поширеними за часу створення Британського національного корпусу. Це особисті електронні тести, Інтернет чати та веб сторінки.

У колишньому Радянському Союзі таким мега-проектом став Машинний Фонд російської мови, створений під керівництвом А.П. Єршова та В.М. Андрющенко. Головними цілями цього фонду були

[5] http://ice-corpora.net/ice/
[6] http://americannationalcorpus.org/

укладання представницького корпусу і підкорпусів різних жанрів та розробка комп'ютерних програм опрацювання природної мови, з метою сприяння комплексній інформатизації лінгвістичних досліджень, включаючи створення граматик та словників. На жаль, цей проект не завершився створенням власне представницького корпусу, хоча були зібрані колекції текстів різних типів. Сьогодні з Машинного Фонду доступні деякі тексти XIX століття та газетний корпус за 1997 рік.

Сучасний розвиток корпусної лінгвістики (пост 2000-і роки) дуже бурхливий, що підтверджується величезною кількістю нових досліджень у галузі. Про популярність та інтенсивність розвитку корпусної лінгвістики свідчить як чимале число теоретичних і технологічних праць, підручників, так і самих корпусів. Йдеться про дослідження у галузі лексичної граматики [Stubbs 1996; Hunston, Francis 2000; Renouf 2001; Nesselhauf 2005; Exploring the Lexis-Grammar Interface 2009], лексикографії та навчання мові [McEnery, Kifle 2002, Altenberg, Granger 2002; McEnery, Xiao 2004, Максимів 2008], когнітивної лінгвістики [Corpora in Cognitive Linguistics 2006; Gilquin 2003; Gries 2003; Gries, Stefanowitch 2004; Schmidt 2000; Schonefeld 1999], прагматики та дискурс-аналізу [Aijmer and Stentström 2004; Archer 2005; Baker 2005; Baker, McEnery 2005; Hardt-Mautner 1995; Koller, Mautner 2004; McEnery 2005; Orpin 2005; Partington et al. 2004; Vivanco 2005; Wang 2005], стилістики [Burrows 2002; Charteris-Black 2004; Corpus-Based Approaches to Metaphor and Metonymy 2006; Deignan 2005; Semino and Short 2004; Stubbs 2005], перекладознавства [Malmkjær 1998; Zanettin 1998; Incorporating Corpora. The Linguist and the Translator 2008]. Корпусно-базовані дослідження відбуваються для вивчення значення слова [Partington 2004], фразеології [Hunston 2001; Лозинська 2009], синтаксичних властивостей граматичних структур [Duffley 2003],

дистрибуції граматичних категорій [Biber 2001] (детальніше див.: Corpus Linguistics and Written Language Resources Bibliography[7]).

У цей час корпуси створені для багатьох мов світу (див. веб сайт Дейвіда Лі[8], на якому розміщено дані про існуючі корпуси та ресурси корпусної лінгвістики).

Найновіші досягнення в царині корпусного мовознавства друкуються у визнаних міжнародних наукових журналах: *Corpus* (2001–) (Nice: Laboratoire "Bases, Corpus, Langage", Université de Nice – Sophia Antipolis)[9]; *Corpus Linguistics and Linguistic Theory* (2005–) (Berlin – New York: Mouton De Gruyter)[10]; *ICAME Journal*, Journal of the International Computer Archive of Modern English (1987–) (Bergen: Norwegian Computer Centre for the Humanities)[11]; *International Journal of Corpus Linguistics* (1996–) (Amsterdam: John Benjamins)[12]; *Language Resources and Evaluation* (2005–) (Dordrecht: Springer)[13]; *Literary and Linguistic Computing* (1986–) (Oxford: Oxford University Press)[14].

На сьогоднішній день практично для усіх європейських мов існує по кілька корпусів текстів різного типу, обсягу та структури. Створення національного корпусу вважається виявом поваги до рідної мови. Хоча для українського мовознавства укладання національного корпусу і впровадження корпусного мовознавства ще перебуває на етапі становлення, у царині вітчизняної корпусної лінгвістики вже здійснено вагомі доробки. На

[7] liceu.uab.es/~joaquim/language_resources/lang_res/biblio_corpus.html
[8] http://www.uow.edu.au/~dlee/CBLLinks.htm
[9] http://corpus.revues.org/
[10] http://www.degruyter.com/view/j/cllt
[11] http://nora.hd.uib.no/journal.html
[12] http://www.benjamins.com/#catalog/journals/ijcl
[13] http://www.springerlink.com/content/1574-020x/
[14] http://llc.oxfordjournals.org

разі у складі Національної словникової бази Українського мовно-інформаційного фонду НАН України функціонує і постійно розвивається Український національний лінгвістичний корпус (УНЛК), що розробляється під керівництвом академіка НАН України В.А. Широкова [Корпусна лінгвістика 2005, с. 103]. Розпорядженням Кабінету Міністрів України від 11.02.2004 р. № 73-р Національну словникову базу Українського мовно-інформаційного фонду НАН України внесено до державного реєстру наукових об'єктів, що становлять національне надбання. На даний момент обсяг корпусу понад 100 млн. слововживань. Оскільки об'єктом дослідження УНЛК є сучасна українська літературна мова, то стратегія формування корпусу полягає у максимально повному охопленні літературних українських джерел із початку XIX до початку XXI століття, тобто за двісті років існування та розвитку сучасної української літературної мови. Корпус представлений текстами різних стилів і жанрів без збереження пропорцій. За необхідності дослідник може самостійно генерувати підкорпуси окремих стилів із врахуванням статистичних параметрів.

В Україні вже декілька років поспіль відбувається щорічна конференція MegaLing[15], присвячена різноаспектним корпусним дослідженням. У центрі уваги конференції знаходяться питання, пов'язані з теоретичними засадами корпусної лінгвістики, лексикографії, термінографії, нотографії, концептографії і т.ін. З'являються оглядово-ознайомлюючі публікації з питань корпусного мовознавства, а також здійснюються теоретичні та прикладні дослідження у цьому річищі.

[15] http://megaling.ulif.org.ua/

Конститутивні параметри корпусу текстів

Центральним для корпусної лінгвістики є поняття "корпус текстів". Адже, одного боку, він виступає головною ціллю, яку ставить перед собою корпусна лінгвістика, а з іншого, постає об'єктом дослідження цієї мовознавчої галузі. Перед тим, як увійти до лінгвістичної термінології, слово "корпус" широко використовувалося на позначення збірки або колекції певних текстів. Лише у 1984 році Дж. Аартс та В. Мейс у статті "Corpus Linguistics: Recent Developments in the Use of Computer Corpora in English Language Research" [Aarts, Meijs 1984]. вперше використали його як термін корпусної лінгвістики. До кінця 90-х років термін "корпус" міцно увійшов до лінгвістичного вжитку вже у його традиційному значенні як сукупність стандартно організованих електронних текстів, призначених для проведення конкретних лінгвістичних досліджень.

Лінгвістичний корпус характеризується низкою дистинктивних ознак, що відрізняють його від звичайних колекцій текстів в електронній формі (електронних бібліотек, архівів): репрезентативність, автентичність, відібраність, збалансованість, машиночитаність.

Репрезентативність полягає в здатності корпусу відображати всі властивості предметної галузі. Під предметною галуззю розуміється рівень реалізації мовної системи, яка містить феномени, що підлягають лінгвістичному описові. Предметна галузь для корпусу може бути як завгодно великою або як завгодно малою. Так, якщо йдеться про авторський корпус і лінгвістичний опис стосується авторської мови, то предметна галузь – авторська мова – не буде надто великою, наприклад, у порівнянні з предметною галуззю загальномовного корпусу. Але якщо завданням є побудова національного корпусу, то, відповідно, предметна галузь буде значно більшою [Демська-Кульчицька 2005, с. 101-102].

Автентичність передбачає відбір реально створеного носієм(ями) мови писемного або усного тексту(ів), уривка(ів) тексту(ів) у процесі реальної комунікації. Дотримання вимоги автентичності є однією зі складових емпіризації фактичного корпусного матеріалу.

Відібраність ставить вимогу обмеження фактичного матеріалу шляхом відбору певних фрагментів мови з усього мовного континууму. Навіть найбільший за обсягом корпус природної мови завжди залишається лише крихітним взірцем усіх усних і писемних текстів, створених усіма носіями мови навіть упродовж одного дня, і навіть сучасні інформаційні технології не дають змоги подати весь цей мовний матеріал, тому необхідна певна вибірка, яка передбачає застосування чітких правил екстрагування даних, що відповідають обраній стратегії побудови корпусу, мотивовані типом корпусу і метою його створення.

Збалансованість полягає у введенні до корпусу пропорційної кількості текстових ресурсів. На практиці, де традиційно використовують різні методики відбору текстового матеріалу до корпусу, одним із доволі складних завдань є досягнення збалансованості. Для досягнення збалансованості корпусу необхідні мінімальні критерії відбору текстів, які мають включати розрізнення між художньою літературою і нехудожньою літературою; книжкою, журналом або газетою; нормативним і ненормативним варіантом мови; з контролем віку, статі та походження авторів.

Машиночитаність є визначальною ознакою до сучасного електронного текстового корпусу природної мови. Крім електронної форми подання, ця вимога передбачає наявність кодування первинних корпусних даних та лінгвістичну анотацію, хоча на сьогодні це вже параметр "за про мовчанням", тобто іншим сучасний корпус не повинен бути. До найбільших переваг комп'ютеризації корпусу відноситься швидкість обробки та легкість оперування великими масивами мовних даних (пошук, відбір, сортування, форматування і т. ін.). По-друге, комп'ютери надають точні та об'єктивні результати обробки машиночитаних даних. По-третє, комп'ютери виключають можливість впливу людського фактору на аналіз даних, що сприяє отриманню об'єктивних результатів. Нарешті, машиночитаність уможливлює подальшу автоматичну обробку даних корпусу, що в свою чергу збагачує корпус різноманітними метаданими (різними типами розміток).

Таким чином, корпус текстів – це машиночитане, збалансоване, репрезентативне зібрання особливо розмічених (анотованих) текстів, відібраних згідно фіксованих параметрів для досягнення визначеної лінгвістичної мети та досліджуваних нелінійно за принципом гіпертексту.

Сучасне корпусне мовознавство має у своєму дослідницькому арсеналі значну кількість корпусів різних типів і розмірів. Таке розмаїття існуючих корпусів визначається широким спектром дослідницьких і прикладних завдань, для розв'язання яких вони створюються, та особливостями мовного матеріалу, покладеного в основу [Демська-Кульчицька 2004, с. 156-157] (див рис. 1):


Повнотекстові
(тексти в корпусі подані повністю)
Фрагментарні
(подані фрагменти текстів)
Дослідницькі
(застосовують у лінгвістичних дослідженнях із метою формулювання нових теорій, концепцій тощо)
Ілюстративні
(застосовують для підтвердження уже висловлених теоретичних положень чи гіпотез про мову)
Дослідницькі
(подають тексти як цілісні об'єкти, як факт реалізації мовної системи)
Інтерпретаційні
(становлять інформаційно-довідкові та дослідницькі системи)
Діахронні
(репрезентують мову в понад часовому зрізі)
Синхронні
(репрезентують мову або тип тексту певного визначеного часового проміжку)
Моніторингові
(динамічні)
(*забезпечують можливість відстежувати зміни у мові, враховуючи аспект діахронії)*
Статичні
(засвідчують стан мови на певному синхронному зрізі)
Загальномовні
(репрезентують загальнонародну, національну мову)
Спеціалізовані
(скеровані на розв'язання часткових, особливих, специфічних науково-дослідних завдань)

Незважаючи на різноманітність корпусів, можна виділити два основні способи розподілу корпусів на класи: 1) протиставлення корпусів, що укладаються до всієї мови (так звані національні корпуси), корпусам, що укладаються до певної підмови (жанру, стилю, мови певної вікової або соціальної групи, мови письменника або вченого й т.п.); 2) поділ корпусів за типом лінгвістичної розмітки. Хоча існує безліч типів розмітки, більшість реально існуючих корпусів відноситься до корпусів морфологічного або синтаксичного типу (англійською treebanks, що можна перекласти як "банки синтаксичних структур"). При цьому слід підкреслити, що корпус із синтаксичною розміткою експліцитно або імпліцитно містить у собі й морфологічні характеристики лексичних одиниць [Захаров, Богданова 2011, с. 20-21].

Національний корпус – це великий за обсягом корпус, що прагне до об'єднання у своїй структурі текстів найрізноманітніших жанрів і типів (сучасні технології дозволяють вбудовувати в національні корпуси аудіо- і відоматеріали, наприклад, Національний корпус російської мови[16]). Чим більший розмір такого корпусу, чим різноманітніша та точніша розмітка, чим ефективніше програмне забезпечення такого корпусу, тим вища його цінність як лінгвістичного ресурсу. Важливість та необхідність створення таких корпусів важко переоцінити, адже, перш за все, вони збирають і зберігають мову для сучасників та майбутніх поколінь, дані цих корпусів дозволяють аналізувати стан багатьох мов світу у синхронічному та діахронічному аспектах (якщо корпус включає історичні підкорпуси, а сама можливість поповнення корпусу вже передбачає діахронію).

[16] http://www.ruscorpora.ru/

Серед відомих національних корпусів назвемо the British National Corpus (обсяг 100 млн. слововживань)[17], the American National Corpus (22 млн.)[18], the PELCRA Referenc Corpus of Polish Corpus (100 млн.), the Czech National Corpus (більше 100 млн.)[19], the Hungarian National Corpus (187,6 млн.)[20], the Hellenic National Corpus (корпус сучасної грецької мови, загальним обсягом 47 млн. слововживань)[21], the DWDS corpus (обсяг 100 млн. слововживань)[22], the Slovak National Corpus (339 млн.)[23], the Modern Chinese Language Corpus (100 млн. знаків)[24]. Інші національні корпуси the FRANTEXT Database for French[25], the Croatian National Corpus (101,3 млн.)[26], Korpus 2000 for Danish (28 млн.)[27], the National Corpus of Irish (30 млн.)[28] та інші [детальніше про інші корпуси див.: http://www.lancaster.ac.uk/fass/projects/corpus/cbls/corpora.asp]. Як ми вже зазначали, на базі Національної словникової бази Українського мовно-інформаційного фонду НАН України також активно розробляється Український національний лінгвістичний корпус (УНЛК).

Національним корпусам протиставляються спеціальні, які створюються для вирішення конкретних лінгвістичних задач. ***Спеціалізований корпус*** – це жанрово чи галузево специфічний корпус, що має на меті відобразити певну підмову. Наприклад, the Guangzhou Petroleum English

---

[17] http://www.natcorp.ox.ac.uk/
[18] http://americannationalcorpus.org/
[19] http://korpus.cz/english/co_je_korpus.php
[20] http://corpus.nytud.hu/mnsz/index_eng.html
[21] http://hnc.ilsp.gr/en/default.asp
[22] http://www.dwds.de/
[23] http://korpus.juls.savba.sk/index_en.html
[24] http://rocling.iis.sinica.edu.tw/CKIP/engversion/20corpus.htm
[25] http://www.lib.uchicago.edu/efts/ARTFL/databases/TLF/
[26] http://www.hnk.ffzg.hr/default_en.htm
[27] http://korpus.dsl.dk/korpus2000/engelsk_summary.php?languk
[28] http://www.focloir.ie/corpus/

Corpus містить 411 612 слововживань писемного варіанту англійської мови, відібраних із текстів нафтохімічної тематики. Сьогодні спостерігається підвищений інтерес до створення та використання спеціалізованих корпусів в освітній та професійних сферах. Наприклад, the Corpus of Professional Spoken American English (CPSA) складається з транскриптів комунікативних ситуацій з академічної та політичної професійних галузей. The Michigan Corpus of Academic Spoken English (MICASE) містить біля 1,7 млн. слововживань (близько 200 годин записів) сучасного усного університетського мовлення, що було записано в Мічиганському університеті.

Надзвичайно ефективними для визначення рівня адекватності перекладу, удосконалення систем машинного перекладу і гармонізації термінології є ***паралельні корпуси***, які складаються з оригінальних та перекладних текстів, що вирівняні по реченнях.

Особливий тип спеціальних корпусів – це так звані корпуси учнівського мовлення або ***учнівські корпуси*** (*learner corpora*), які укладаються з усних і/або писемних текстів, спродукованих особами, що вивчають мову як іноземну. Такі корпуси почали створюватися ще в кінці 80-х на початку 90-х років XX століття. Серед найбільших учнівських корпусів слід назвати the International Corpus of Learner English та 25-мільйонний the Hong Kong University of Science and Technology Learner Corpus, що містить тексти англійською мовою, створені винятково носіями китайської мови. Меншими за обсягом є учнівські корпуси носіїв норвезької мови (the EVA Corpus of Speech), іспанської мови (the APU Spanish Learner Corpus). Слід зауважити, що кількість учнівських корпусів для інших мов, окрім англійської, невпинно зростає.

Іншим типом спеціалізованих корпусів, що виділяються за хронологічним параметром і зорієнтовані на вивчення та аналіз еволюційних процесів у конкретній

мові, є ***історичні корпуси***. Такий тип корпусів переважним чином укладається на матеріалі текстів однієї мови, відібраних у різні часові проміжки з метою вивчення історичної динаміки мовних змін [Xiao 2008: 401].

За типом мовних даних корпуси поділяються на корпуси писемного, усного мовлення та змішаного типу. У ***корпусах писемного мовлення*** усний варіант реалізації мовної системи не представлений. Найпершим сучасним корпусом писемного англійського мовлення був корпус американського варіанту англійської мови the Brown University Standard Corpus of Present-Day American English, услід за яким було укладено низку подібних корпусів. Іншими корпусами такого ж типу є the Australian Corpus of English, що репрезентує австралійський писемний варіант англійської мови (1986-). The Wellington Corpus of Written NZ English (WWC) представляє новозеландський писемний варіант англійської мови (1986-1990 pp.), a the Kolhapur Corpus відображає індійський писемний варіант англійської мови (1978-).

Корпуси усного мовлення включають транскрибовані версії реальних усних комунікативних ситуацій. ***Корпуси усного мовлення*** – це спеціальні колекції ретельно відібраних текстових уривків (слів, фраз, речень), вимовлених численними мовцями за різних акустичних умов [Teubert 2007, c. 126]. Зважаючи на трудомісткість і напруженість збору усних даних порівняно із писемними, ці корпуси є значно меншими за обсягом. Серед корпусів усного мовлення назвемо the London-Lund Corpus (LLC), the Lancaster/IBM Spoken English Corpus (SEC), the Cambridge and Nottingham Corpus of Discourse in English (CANCO DE), the Santa Barbara Corpus of Spoken American English (SBCSAE) та the Wellington Corpus of Spoken New Zealand English (WSC).

Підводячи підсумок огляду існуючих корпусів, зазначимо, що різноманіття наявних на сьогоднішній день корпусів аж ніяким чином не обмежується наведеними вище даними. З огляду на розширення спектру лінгвістичних завдань, що ставлять перед собою дослідники мови, арсенал корпусної лінгвістики постійно поповнюється новими, більш досконалими розробками.

Типи корпусної розмітки

Лінгвістичний корпус за визначенням є такою колекцією природно мовних текстів, де здійснено розмітку (маркування) хоча б за одним лінгвістичним параметром. Ця ознака є такою, що вирізняє лінгвістичний корпус з-поміж великого числа інших лінгвістичних інформаційно-інструментальних систем, баз даних та знань [Корпусна лінгвістика 2005, с. 33]. Процес розмітки (*tagging, annotation*) полягає в приписуванні текстам і їх компонентам спеціальних міток (*tag, tags*):

- *зовнішніх, екстралінгвістичних* (відомості про автора й відомості про текст: автор, назва, рік і місце видання, жанр, тематика; відомості про автора можуть включати не тільки його ім'я, але також вік, стать, роки життя й багато чого іншого (це кодування інформації має назву *метарозмітка*);
- *структурних* (розділ, абзац, речення, словоформа);
- власне *лінгвістичних*, що описують лексичні, граматичні та інші характеристики елементів тексту.

Набір цих метаданих багато в чому визначає можливості, що надаються корпусами дослідникам. Анотація додає цінності корпусу, адже суттєво розширює діапазон дослідницьких питань, що можуть бути досліджені на матеріалі даного корпусу. При виборі цих даних необхідно керуватися цілями дослідження й потребами лінгвістів, а також можливостями по внесенню в текст тих або інших додаткових ознак.

Зауважимо, що у корпусних студіях розрізняють анотацію (*annotation*) та структурне маркування (*markup*) корпусу/корпусних даних. Розмежування цих двох понять виходить з визначення анотації Дж. Лічем, згідно якого «процес анотування корпусних даних – це додавання інтерпретованої, лінгвістичної інформації до електронного корпусу усного чи/або писемного мовлення» [Leech 1997, с. 2]. Під поняттям "інтерпретована лінгвістична інформація" розуміється така анотація, яка в деякій мірі є продуктом розуміння тексту людським розумом (наприклад, анотація за частинами мови). Маркування надає відносно об'єктивну верифіковану інформацію про частини корпусу та структуру кожного тексту [McEnery, Xiao, Tono 2006, с. 29].

Під **лінгвістичною анотацією** у корпусній лінгвістиці традиційно розуміють: а) довільну лінгвістичну інформацію про лінгвально релевантні одиниці текстових даних, поданих через формальний код; б) практику введення формалізованої лінгвістичної інформації в електронний текст; в) наявність такої інформації у тексті [Демська-Кульчицька 2004, с. 26].

Лінгвістична анотація може здійснюватися на різних рівнях та набувати різних форм. Так, серед лінгвістичних типів розмітки виділяються:

- **Морфологічна розмітка**. В іноземній термінології вживається термін *part-of-speech tagging* (POS-tagging), дослівно – частиномовна розмітка. У дійсності морфологічні мітки включають не тільки ознаки частини мови, але й ознаки граматичних категорій, властивих даній частині мови. Це основний тип розмітки, оскільки, по-перше, більшість великих корпусів є саме морфологічно розміченими корпусами, по-друге, морфологічний аналіз розглядається як основа для подальших форм аналізу – синтаксичного й семантичного,

і, по-третє, успіхи в комп'ютерній морфології дозволяють автоматично розмічати корпуси великих розмірів.

- **Синтаксична розмітка**, що є результатом синтаксичного аналізу, або *парсинга* (*parsing*), виконуваного на основі даних морфологічного аналізу. Цей вид розмітки описує синтаксичні зв'язки між лексичними одиницями й різні синтаксичні конструкції (наприклад, підрядне речення, дієслівне словосполучення й т.п.).
- **Семантична розмітка**. Хоча для семантики немає єдиної семантичної теорії, найчастіше семантичні теги позначають семантичні категорії, до яких відноситься дане слово або словосполучення, і більш вузькі підкатегорії, що специфікують його значення.
- **Анафорична розмітка**. Фіксує референтні зв'язки, наприклад, займенникові.
- **Просодична розмітка**. У просодичних корпусах застосовуються мітки, що описують наголос та інтонацію. У корпусах усного розмовного мовлення просодична розмітка часто супроводжується так званою *дискурсною* розміткою, яка служить для позначення пауз, повторів, застережень, і т.д.

Інший тип інформації також може бути закодований у корпусі. Наприклад, у корпусах розмовного мовлення може бути представлена соціолінгвістична інформація про такі риси як стать, вік, приналежність до соціального прошарку та місце проживання.

Під час створення корпусу використовується низка процедур і програм, як-от: токенізація, лематизація, стеммінг, парсинг [Захаров 2011, с. 38-41]. **Токенізація** – це розбиття потоку символів природної мови на окремі значимі одиниці (токени, словоформи). **Лематизація** – процес утворення початкової форми слова, виходячи з інших його словоформ. С**теммінг** полягає в знаходженні стеми (основи) слова. **Парсинг** – це процес аналізу

синтаксичної структури тексту чи частини тексту, що ґрунтується на зіставленні лінійної послідовності лексем (слів, токенів) мови з її формальною граматикою.

Побудова автоматичних синтаксичних аналізаторів (парсерів) для великих корпусів є однією із найважливіших областей комп'ютерної лінгвістики.

Комп'ютерні програми, що використовуються для аналізу даних корпусу

Оскільки корпусна лінгвістика послуговується у своїх дослідженнях великими за обсягом та репрезентативними вибірками текстів природної мови, існує декілька типів програмного забезпечення, що можуть використовуватися під час дослідження: комерційні комп'ютерні програми – конкордансери (н.д., LEXA, MonoConc, MicroConcord, TACT, WordSmith, WordCruncher, Manatee (Bonito), IMS Corpus Workbench (CQP), XAIRA, LEXA, Virtual Corpus Manager (VMC), EXMARaLDA Corpus-Manager (Co-Ma)), а також комп'ютерні програми, розроблені для специфічних процедур аналізу. Ці типи програм застосовуються для складних граматичних моделей.

Робота з корпусом здійснюється за допомогою спеціальних програмних засобів – конкордансерів (більш простий тип програм) і корпусних менеджерів, що надають різноманітні можливості по одержанню необхідної інформації з корпуса. **Корпусний менеджер** – це спеціальна пошукова система, що включає програмні засоби для пошуку даних у корпусі, отримання статистичної інформації й надання результатів користувачеві в зручній формі. Результати цієї процедури подаються у вигляді горизонтальних рядків із пошуковим словом посередині. Ця процедура має назву KWIC (Key Word In Context).

Пошукові можливості корпусних менеджерів включають у себе пошук конкретних словоформ, пошук

словоформ по лемах, пошук групи словоформ у вигляді розривної або нерозривної синтагми, пошук словоформ по набору морфологічних ознак і т.д., висновок результатів пошуку із вказівкою контексту заданої довжини, збереження відібраних матеріалів в окремому файлі. Використання корпусів дозволяє не тільки вивчати лексичні одиниці в контекстах, але й одержувати дані про частоту словоформ, частоту лексем, граматичних категорій, про спільну зустрічальність лексичних одиниць, особливості їх сполучуваності, управління і т.ін.

Результати пошуку видаються у вигляді конкордансу (тому корпусні менеджери і називають конкордансерами), де аналізована одиниця представлена в її контекстному оточенні та супроводжується статистичною інформацією. Останні можуть фіксувати частотні характеристики окремих мовних одиниць, або грамем, або можуть характеризувати сполучуваність декількох лексичних одиниць. Багато систем дозволяють коректувати формат представлення даних (міняти довжину лівого й правого контексту, задавати обсяг показу й порядок сортування даних, відображати або не відображати лінгвістичні й екстралінгвістичні характеристики і т.д.).

Сфери застосування лінгвістичних корпусів

Поява електронних корпусів суттєво змінила спосіб дослідження мовних явищ. Репрезентативні, збалансовані, значні за обсягом корпуси надають інформацію про динаміку процесів зміни лексичного складу мови, функціонування граматичних та синтаксичних одиниць, дозволяють здійснювати аналіз вживання мовних одиниць у різних стилях, жанрах, у різних авторів і т.ін. Корпуси надають недоступні для безпосереднього спостереження дані про реалізацію мовної системи, що в свою чергу відображається у словниках, граматиках та навчальних матеріалах. Якщо ранні корпусні дослідження

зосереджувалися головним чином лише на вживаності мовних одиниць (н. д., частота вживання певних іменників, дієслів та прикметників), то з появою потужних комп'ютерних технологій дослідники отримали змогу визначати та аналізувати більш складні лексичні, граматичні та синтаксичні явища.

У галузі лексикографії корпусна лінгвістика не лише впровадила нові методи аналізу мовного матеріалу, але і значно розширила межі наукового дослідження. Використання корпусу в лексикографічній практиці замістило паперові картки автоматичним пошуком слів у корпусі та суттєво переорієнтувало підхід до відбору слів до сло́вника, замінивши відбір слів, обумовлений особистими перевагами лексикографа, більш об'єктивними критеріями: даними про частоту та дистрибуцію слів у певних текстах чи жанрах. Переважна більшість сучасних словників є корпусоінформованими, а не "укладачеінформованими" [McEnery, Gabrielatos 2006, с. 48]. Одним із останніх нововведень у лексикографічній практиці стало розміщення у словниковій статті крім традиційної інформації про значення та семантичні відношення одиниці також й інформації про її сполучуваність та дистрибуцію.

Вже звичною практикою стало використання корпусів для укладання довідкової літератури – граматик, навчальних словників та довідників як для носіїв мови, так і для тих, хто її вивчає. Визнано, що традиційні шкільні граматики та підручники досить часто проілюстровані штучно створеними чи відредактованими прикладами вживання певної лінгвістичної одиниці чи явища. Такі приклади є малоефективними для успішного навчання учнів використанню мови як засобу комунікації, оскільки вивчене ними може кардинально різнитися від реального функціонування мови. У цьому відношенні дослідження корпусів природньої мови надає більш

точний та деталізований опис мови, що у свою чергу відображається у змісті педагогічних матеріалів, граматик та словників, а також дизайні навчальних програм та підручників. На даний момент вже існує декілька корпусобазованих граматик англійської мови, наприклад, загальних (Collins COBUILD English Grammar, 1990; Longman Grammar of Spoken and Written English, 1999) та таких, що акцентують увагу на певних аспектах (Collins COBUILD Grammar Patterns 1: Verbs, 1996). Таким чином, хоча це і не швидкий процес, проте мало-помалу все більше навчального матеріалу на сьогоднішній день стає "корпусоінформованим". Зростає тенденція до вкладання коштів у розробку значних за обсягом корпусів такими великими видавцями і розробниками навчальних матеріалів, як Cambridge University Press, Oxford University Press, Pearson-Longman, Collins-COBUILD and Macmillan.

Використання корпусів стає дедалі популярнішим у навчанні та вивченні мови. Вчителі та учні все частіше звертаються до корпусу для здійснення швидкого різнорівневого пошуку та аналізу найрізноманітнішої лінгвальної інформації, а конкорданс визнається зручним інструментом для ефективного опанування мовою. Корпуси текстів слугують неоціненним ресурсом реальних прикладів вживання тих чи інших лексичних та граматичних одиниць, за умови, що учні і вчителі мають доступ до корпусів та обізнані з технікою роботи з корпусом.

Орієнтоване на дані навчання (*data-driven learning*) іноземній мові, коли корпус стає головним джерелом та способом отримання лінгвістичних знань й умінь, отримує все більше прихильників серед сучасних студентів та вчителів. При цьому підході базоване на корпусі програмне забезпечення використовується для підтримки інтерактивної навчальної діяльності, а залучення "живого"

мовлення з корпусу сприяє пробудженню у студентів інтересу до процесу вивчення мови та заохоченню їх до самостійного вивчення автентичного мовного матеріалу. Студент виконує активну роль дослідника мови, а вчитель лише направляє та сприяє навчанню [Zhukovska 2011; Zhukovska 2012].

Іншою сферою інновацій, що безпосередньо пов'язана із корпусною лінгвістикою – це укладання та аналіз учнівських корпусів (*learner corpora*), які надають важливу для вчителя інформацію про те, як учні вивчають та засвоюють мову, про типові помилки у вживанні лексичних,граматичних та синтаксичних одиниць. Вагомим аспектом для методики навчання мови є також використання паралельних корпусів, що уможливлюють доступ до безпосередніх перекладних еквівалентів слова, фрази, синтаксичної конструкції. Такий інструмент не тільки полегшує створення перекладних словників, базованих на реальних прикладах перекладацької практики, а і сприяє легшому вивченню та навчанню іноземної мови відповідно до комунікативної стратегії навчання. Застосовуючи двомовні конкорданційні програми, комп'ютерна система висвітлює обидва тексти поряд, і студент легко та швидко перекладає текст, оскільки не змушений шукати кожне невідоме слово у словнику. А от корпуси підручників з іноземної мови надають можливість порівняти використання рідної мови в різних контекстах та мови, з якою мають справу учні у посібниках (див напр.: Harwood 2005; Römer 2004). Аналіз таких невідповідностей допоможе при відборі текстів, які повинні бути включені до навчальних матеріалів. Огляд корпусів начальних текстів також може виявити, якою мірою тексти підручників впливають на усне та писемне мовлення учнів.

Крім того, останнім часом корпуси стали активно використовуватися для створення матеріалів для

тестування та навчання вчителів з метою підвищенню рівня їх професіоналізму.

Перелічені вище можливості використання корпусів не вичерпують всі сфери їх експлуатації. Із збільшенням обсягів та удосконаленням технічного оснащення дослідницький потенціал корпусів невпинно зростає, що робить їх цінним лінгвістичним ресурсом.

Отож, підсумовуючи огляд головних досягнень у царині корпусного мовознавства, зауважимо, що з часу свого становлення до сьогодні корпусна лінгвістика пройшла складний шлях від повного неприйняття у 60-ті роки XX століття до магістрального напряму сучасного мовознавства. На сьогодні в науковій спільноті також немає однозначного ставлення до корпусних досліджень: у них є як віддані прибічники, так і суворі опоненти. Проте необхідно пам'ятати, що незважаючи на революційні можливості корпусної лінгвістики, вона є лише частиною численного методологічного інструментарію сучасної лінгвістичної науки. І як зазначив Ч. Філмор [Fillmore 1992, с. 35], навіть значні за обсягом корпуси не в змозі відобразити все можливе у мові, натомість і невеликі за обсягом корпуси можуть надати інформацію, яку б нереально було отримати, не звертаючись до корпусних даних.

Рекомендована література з теми

1. Демська–Кульчицька О. Основи національного корпусу української мови [монографія] / О. М. Демська-Кульчицька – Київ, 2005.

2. Захаров В.П., Богданова С.Ю. Корпусная лингвистика: учебник для студентов гуманитарных вузов / В.П. Захаров, С.Ю. Богданова. – Иркутск: ИГЛУ, 2011.

3. Жуковська В.В. Вступ до корпусної лінгвістики [навчальний посібник] / В.В. Жуковська – Житомир: Вид-во ЖДУ ім.І.Франка, 2013.

4. Широков В.А., Бугаков О.В., Грязнухіна Т.О. та ін. Корпусна лінгвістика / В.А. Широков, О.В. Бугаков, Т.О. Грязнухіна та ін. – К.: Довіра, 2005.

5. MacEnery T., Hardie A. Corpus Linguistics: Method, Theory and Practice / T. MacEnery, A. Hardie. – Cambridge University Press, 2012.

6. McEnery T., Xiao R., Tono Y. Corpus-based Language Studies: an Advanced Resource Book / T. McEnery, R. Xiao, Y. Tono. – London: Routledge, 2006.

7. Meyer Ch. P. English Corpus Linguistics. An introduction / Ch.P. Meyer. – Cambridge University Press, 2004.

Питання і завдання для самоперевірки

1. Поясніть сутність корпусної лінгвістики.
2. Яким чином застосування лінгвістичних корпусів змінило спосіб дослідження мови?
3. Сформулюйте предмет і завдання корпусної лінгвістики як лінгвістичної науки.
4. Охарактеризуйте зв’язки корпусної лінгвістики з іншими мовознавчими науками.
5. Що є теоретичним підґрунтям корпусної лінгвістики?
6. Наведіть аргументи на користь визнання корпусної лінгвістики особливою методологією дослідження лінгвального матеріалу/ самостійною мовознавчою наукою.
7. У чому полягають основні відмінності між корпусною лінгвістикою та традиційним мовознавством; корпусною лінгвістикою і комп’ютерною лінгвістикою?
8. Перелічіть параметри, що лежать в основі виділення піднапрямів у корпусних дослідженнях.
9. Охарактеризуйте історичний розвиток корпусних досліджень.

10. У чому виявляється специфіка доелектронних корпусів? Назвіть основні лінгвістичні проекти, в яких доелектронні корпуси відіграли значну роль.
11. Окресліть головні характеристики корпусів першого покоління, назвіть їх.
12. Визначте основні характеристики корпусів другого покоління.
13. Назвіть найвідоміші мегакорпуси сучасності.
14. Охарактеризуйте сучасний етап розвитку корпусних студій у світі та в Україні.
15. Які, на Вашу думку, перспективи розбудови корпусної лінгвістики в українському мовознавстві?
16. Дайте визначення терміну «лінгвістичний корпус текстів»?
17. Перерахуйте дистинктивні ознаки, за якими корпус текстів відрізняється від електронної бібліотеки та архіву. Охарактеризуйте кожну з них.
18. Назвіть і схарактеризуйте існуючі типи корпусів. Наведіть приклади до кожного з типів та коротко опишіть основні характеристики цих корпусів.
19. Поясніть у чому полягає суть корпусної розмітки.
20. У чому полягає відмінність між анотацією та структурним маркуванням корпусних даних?
21. Дайте визначення лінгвістичній анотації. Назвіть типи лінгвістичної анотації.
22. Які на Вашу думку переваги та недоліки анотованих корпусів?
23. Наведіть приклади корпусних менеджерів. У чому полягають їх особливості?
24. Які переваги використання даних корпусів у навчанні та вивченні іноземної мови? Які труднощі/недоліки Ви вбачаєте у впровадженні корпуснобазованого підходу до вивчення іноземної мови?
25. Чи користуєтесь Ви корпусом? Якого типу інформацію Ви шукаєте в корпусі?