Корпусна лінгвістика

Орися Демська-Кульчицька

ДЕЯКІ АСПЕКТИ КОРПУСНОЇ ЛІНГВІСТИКИ

Виникнення і розвиток машинної форми існування природної мови є незаперечною реальністю сьогодення. Її існування можна трактувати по- різному, але не можна не погодитися з твердженням А. Єршова, який відносив існування машинної форми мовленнєвих процесів „до тих безальтернативних інновацій в суспільстві, які не можуть не відбутися і визначатимуть розвиток суспільства у майбутньому" [2:27]. Тому сьогодні можемо говорити про те, що доля національних мов на зламі XX—XXI ст. вирішується у сфері комп'ютеризації, як колись у сфері книгодрукування. Звідси невипадковість появи так званих суміжних дисциплін, ідеєю яких є застосування теоретичних положень і практичних розв'язків комп'ютерних наук і самого комп'ютера у власних дослідженнях.

Лінгвістика виявилася тією гуманітарною наукою, яка, не поривaючи зв'язків з іншими науками про людину та її культуру, першою рішуче почала використовувати не лише інструментальні методи спостереження та експериментальні прийоми, але й систематично застосовувати математичні способи, в тім числі й комп'ютери для формулювання та фіксації своїх висновків.

Внаслідок застосування комп'ютерних засобів до збирання, організації та програмного оброблення мовного матеріалу виник напрямок корпусної лінгвістики, який інтенсивно розвивається впродовж останнього двадцятиліття фактично в усіх національних мовознавчих науках. І, як стверджує Л. Ричкова, чинники, що мотивують стрімку динаміку розвитку корпусної лінгвістики, очевидні — це впровадження до лінгвістичного вжитку спеціально приготовленого матеріалу, що „дозволяє не лише оптимізувати і об'єктивізувати лінгвістичні дослідження, але і по-новому окреслити багато традиційних лінгвістичних понять" [7:185].

© ДЕМСЬКА-КУЛЬЧИЦЬКА О.М.

Корпусна лінгвістика — напрямок, завданнями якого є розроблення теоретичних засад і практичних прийомів побудови, машинного опрацювання та експлуатації лінгвальних даних, оформлених як корпус текстів. **Об'єктом** корпусної лінгвістики є корпус, а предметом - текст

Єдиного погляду на витоки цього напрямку все ще немає. Його початки або пов'язують з *Браунівським* корпусом, датованим 60-ми роками XX ст. Це по суті перший електронний корпус текстів, який відповідав критеріям корпусної побудови, тобто тим критеріям, які перетворюють довільне електронне зібрання текстів природної мови на корпус. Або ж зараховують до корпусної лінгвістики так звані докорпусні дослідження, пов'язані зі збиранням, організацією та описом лінгвістичного матеріалу з або без використання машинних ресурсів. Йдеться головно про укладання лексичних картотек, записи дитячого мовлення чи фіксацію неписемних мов. Найвідоміші з таких досліджень: вивчення частотної дистрибуції і узгодження букв німецької мови Й. Кадінгом на 11-мільйонному систематизованому текстовому матеріалі (без машинного ресурсу); створення лабораторій механізації лексикологічної та лексикографічної роботи в Європі, а саме — Лабораторії лексикографічного аналізу при Центрі вивчення словника французької мови у Безансоні в 1957-му році та Лабораторії Центру з автоматизації філологічного аналізу в італійському Галараті 1953-го року. Підготовка до створення останньої розпочалася ще у 1949-му році, де першою дослідною роботою стало реєстрування змісту творів Фоми Аквінського (з використанням машинного ресурсу).

Датування корпусної лінгвістики початком минулого століття, а не серединою чи початком другої половини XX ст. усе частіше зустрічається в корпусних дослідженнях. Зокрема, Т. МакЕнері та А. Вилсон зазначають, що „думка про те, що корпуси текстів з'явилися в 1960-х роках і особливо інтенсивно почали розвиватися в 1980-х є помилковою. До появи генеративної граматики у лінгвістиці якраз домінувало вивчення масивів емпіричних даних, тобто корпусів. Інша справа, що аналіз виконувано вручну, внаслідок чого обсяги даних були надзвичайно обмеженими" [10]. Розвиваючи свою думку, вчені розширюють хронологічні межі аналізованого напрямку і пропонують виділяти два періоди у розвитку корпусних студій (1) так звану ранню і (2) власне корпусну лінгвістику.

Рання або **протокорпусна лінгвістика** припадає на етап формування теоретичного підґрунтя та прагматичних передумов виникнення напрямку і створення текстових зібрань для лінгвістичного дослідження на переважно паперових носіях (1910-60 рр.). З 1960-х р. починається період **корпусної лінгвістики,** безпосередньо пов'язаної з машинними носіями, в межах якої до 1990-х років сформувалися три напрямки теорії та практики: (1) побудова електронних текстових корпусів, (2) програмне опрацювання текстових корпусів, (3) екстрагування, аналіз і опис корпусних даних.

Але слід зазначити, що використання самого терміна 'корпусна лінгвістика' поширилося лише в останнє двадцятиріччя, після публікації у 1984-му році збірника під назвою *„Corpus Linguistics: Recent Developments in the Use of Computer Corpora in English Language Research"*.

Виходячи з такої періодизації розвитку напрямку, визнаємо 1960-ті роки часом переходу від протокорпусної до корпусної лінгвістики. І сьогодні уже можна стверджувати, що це був не найкращий момент, оскільки на той час в лінгвістичній науці домінували ідеї раціоналізму та породжувальної граматики Н. Хомського, які опосередковано спричинили гальмування розвитку корпусних досліджень у світі.

Н. Хомський критикував і самі корпуси як емпіричну базу лінгвістичного дослідження, і продуктивність корпусного підходу до вивчення мови. Щодо текстового корпусу Н. Хомський висловив таку думку: „у будь-якому корпусі природної мови існують спотворення. Деяких речень у них не буде, бо вони очевидні, інших — тому, що вони хибні, ще інших — тому, що вони неввічливі. Таким чином, природномовний корпус дасть настільки сильно спотворену картину, що базований на ньому опис виявиться звичайним списком мовних одиниць" [9:67]. Заперечення доцільності використання емпірики у лінгвістичному дослідженні мотивовано раціоналістичною позицією Н. Хомського. Як послідовник раціоналізму вчений вважав, що дослідження емпіричних даних — абсолютно беззмістовне заняття, оскільки суть лінгвістики полягає у вивченні мовної компетенції, а не її відображення — мовної діяльності.

Головним чином недосконалість комп'ютерів та критичний погляд Н. Хомського на доелектронні текстові корпуси і стали причиною відходу корпусної лінгвістики у 1950-х — 80-х роках на периферію лінгвістичної дослідної парадигми. Проте такий відхід не означав припинення корпусних робіт і корпусно-базованих досліджень мови, головним чином зосереджених у цей період на англійській мові та її варіантах. Внаслідок цього з 1962-го до 1989-го року з'явилися еталонні корпуси різних варіантів англійської мови (див. таблицю).

У 1990-х рр. наступає новий етап розвитку корпусної лінгвістики, пов'язаний з її поширенням на національні лінгвістичні традиції та стимульований низкою корпусних проектів, до яких належать CRATER, MULTEXT, MULTEXT-EAST, PAROLE і створенням робочої групи EAGLES (Expert Advisory Group on Language Engineering Standards) та консорціуму TEI (Text Encoding Initiative). Внаслідок цього процесу з'являються національні корпуси майже усіх європейських мов, починаючи від еталонного *Британського національного корпусу* і закінчуючи слов'янськими корпусами: *Cesky nârodni korpus, Polski Korpus Narodowy, FIDA: Korpus Slovenskega Jezika, HETA: Hrvatski elektronski tekstovni arhiv, Национальный корпус русского языка* etc. Крім національних мов, у поле зору корпусної лінгвістики усе частіше потрапляють мови з іншим, ніж національний, статусом. Серед цих мов! є:

Динаміка появи корпусів англійської мови у XX ст.

№	Назва корпусу	Роки	Характеристики
1.	Brown Corpus (Brown University Standard Corpus of Present-Day American English)	1962-1964	корпус друк, текстів амер. варіанта англ. мови; 1 ООО ООО сл.
2.	American Heritage Intermediate Corpus	1971	корпус друк, текстів амер. варіанта англ. мови; 5 000 000 сл.
3.	Lancaster-Oslo-Bergen Corpus	1978	корпус друк, текстів брит, варіанта англ. мови; 1 000 000 сл.
4.	Birmingham University International Language Database / Birmingham Corpus	1987	корпус друк, текстів брит, варіанта англ. мови; 200 000 000 сл.
5.	London-Lund Corpus / LLC	1980	корпус усних текстів брит, варіанта англ. мови; 500 000 сл.
6.	Kolhapur Corpus	1988	корпус друк, текстів індійського варіанта англ. мови; 1 000 000 сл.
7.	Tools for Syntactic Corpus Analysis / TOSCA	1988	корпус друк, текстів брит, варіанта англ. мови; 1 500 000 сл.
8.	Survey of English Usage / SEU	1989	корпус друк, текстів брит, варі-

* слововживань

— офіційні регіональні мови без національного статуту, наприклад, *уельська мова;*

— діалекти, чи мовленнєві варіанти великих національних мов, наприклад, *варіанти китайської;*

— мови з великою кількістю мовців, але без статусу національної мови, наприклад, *суахілі;*

— національні мови з малим числом носіїв, наприклад, *ірландська;*

— мова меншості з відносно невеликою кількістю носіїв, наприклад, *мова мешканців острова Мен;*

— мови емігрантських меншин, наприклад, *Пенджабі* у Великобританії;

— мови, які перебувають під загрозою зникнення;

— кінетичні мови.

Таким чином, сучасна корпусна лінгвістика виходить за межі національних мов, орієнтуючись на діалекти, соціолекти, дитяче мовлення і т. ін., що є одним з чинників стимулювання її розвитку.

Поряд з питанням виникнення та розвитку корпусної лінгвістики не менш важливим є і питання теоретичного підґрунтя останньої. Осмислюючи теоретичні аспекти аналізованого напрямку приходимо до висновку, що теоретичним підґрунтям сучасної корпусної лінгвістики можна вважати структуралізм чи структурну лінгвістику: систему поглядів на мову та методи її дослідження, в основі яких лежить розуміння мови як знакової системи з дискретними структурними еле-

ментами (одиницями мови, їх підкласами і класами), та використання формальних прийомів опису мови.

Концептуально, структурний підхід до мови передбачає аналіз реального тексту, який дозволяє виділити узагальнені інваріантні одиниці (схеми речень, морфем, фонем) і співвіднести їх з конкретними мовленнєвими сегментами, виходячи з детермінованих правил реалізації, які визначають межі варіювання мовних одиниць у мовленні, у такий спосіб декларуючи примат реального тексту в лінгвістичному дослідженні, що власне і є ідеєю корпусної лінгвістики.

Структуралізм ніколи не був однорідним, але для усіх його шкіл, за твердженням Е. Бенвеніст, можна виділити такі спільні положення [6:26; 8:38-39]:

— основним принципом є твердження, що мова творить систему, усі частини якої об'єднані відношеннями спільності та залежності;

— система організовує одиниці, які є відмінними і розмежованими партикулярними знаками;

— система домінує над елементами;

— відношення елементів як у мовленнєвому ланцюгу, так і в формальних парадигмах розкривають структуру системи.

Тобто, елементи системи, залежно від їхнього типу, характеру, рівня, значення etc. таким, а не іншим чином функціонують у мовленні, яким є усний або писемний текст — предмет корпусу, виявляючи загальну і/або часткову специфіку самої системи. Іншими словами, текст як результат мовленнєвого акту, включає в себе певний інвентар мовних елементів, відібраних і скомбінованих згідно з граматикою мови та регульованих нормою. У тексті, як формі існування мовлення, існує лише те, що передбачено системою мови, тому текст, який є засобом експлікації специфіки міжелементних зв'язків, корпусна лінгвістика розглядає як предмет, дослідження якого поглиблює розуміння суті природної мови.

Крім засадничого положення про системність мови та її реалізацію у мовленні-тексті, корпусна лінгвістика свідомо або інколи підсвідомо, приймає окремі положення різних напрямків структуралізму. Наприклад, з теоретичних поглядів Празької лінгвістичної школи для корпусної лінгвістики релевантними є твердження про те, що „мову (langue) слід розуміти як абстрактну систему норм, яка є необхідною умовою взаєморозуміння, але не має самостійної форми існування і може бути пізнаною лише на основі конкретних висловлювань", і мову як функціональну систему, тобто „систему засобів вираження, яка служить певній визначеній меті" [6:65]. Особливо останнє положення чітко реалізоване у теорії побудови текстових корпусів, коли йдеться про прагматику тексту або ж про прагматику умов генерування тексту. Так само важливим для корпусної лінгвістики є положення Лондонської школи з приводу опозиційності мовлення та мови як системи. Так, „мовленнєвому відрізку протиставлено поняття системи і структури як категорії, яка є результатом аналізу мовленнєвого відрізка і наступно-

її узагальнення і абстрагування, зробленого дослідником на парадигматичному і синтагматичному рівні" і власне корпусна лінгвістика пропонує розглядати текст, усний або писемний, як форму екзистенції мовлення для аналізу, узагальнення, абстрагування і наступного висновку. Тобто дослідна процедура корпусної лінгвістики лежить у межах моделі „мовлення мова".

Якщо положення Празької та Лондонської шкіл адаптовані спорадично, то глосематику, зокрема положення, сформульовані Л. Єльмслєвим у *„Пролегоменах до теорії мови"*, можна вважати програмними для корпусної лінгвістики, яка de facto сповідує більшість з цих принципів [1]. Йдеться про таке:

— єдине, що дано досліднику мови як вихідний пункт, то це текст у своїй нерозчленованій і абсолютній цілісності;

— оскільки тексти природної мови надзвичайно великі числом і тривалістю, слід задовольнятися деякою вибіркою з них;

— лінгвістична теорія починається з тексту як єдиного даного і намагається прийти до несунеречливого і вичерпного опису цього тексту шляхом аналізу чи послідовного розділу: „користуючись інструментом лінгвістичної теорії, ми можемо видобути з вибірки текстів запас знань, який знову можна використовувати на інших текстах. ... Ці знання стосуються не так процесів чи текстів, з яких вони отримані, як системи чи мови, на основі якої будуються всі тексти";

— на наступному кроці процедури великі частини тексту повинні бути поділені на твори окремих авторів, окремі праці, частини, параграфи і т.д., а потім уже на складні та прості речення;

— робота над лінгвістичною теорією завжди емпірична, через свою довільність вона пов'язана з обчисленнями.

Однак, не зважаючи не те, що європейський структуралізм підготував ґрунт для виникнення корпусного мовознавства, останнє усе ж розвивається у межах північноамериканської лінгвістичної традиції, зокрема дескриптивізму. І це має своє тлумачення. По-перше, реальність доктрини дескриптивної лінгвістики, у якій більше, ніж в інших напрямках структуралізму виявляється тенденція до введення імовірнісних та статистичних методів. По-друге, очевидна „утилітарна скерованість цілого ряду галузей лінгвістики в США" [6:183]. По-третє, динаміка зникнення мов американських аборигенів, які вимагали опису, змусили американських мовознавців застосовувати особливі методи збирання, організації та зберігання цього фактичного матеріалу, будувати власні підходи та прийоми до його систематизації й опису, багато з яких згодом було перенесено на процедури аналізу природних мов взагалі. Так, сформувалися передумови для виникнення двох відокремлених аналітичних технік, одна з яких скерована на розроблення дескриптивної методики виявлення і класифікації лінгвістичних одиниць, яка зазвичай розрахована на аналіз уже зафіксованого корпусу мови.

Від моменту виникнення до сьогодні проблематика корпусного напрямку зазнала значної диференціації і передбачає опрацювання

загальної теорії корпусної лінгвістики, над якою працюють зокрема Д. Байбер, Дж. Синклер, В. Тойберт, кореляцію корпусної лінгвістики та інших лінгвістичних дисциплін, типологію корпусів та методики інтерпретації корпусних даних, засад створення текстових корпусів природних мов, — ідеться про доробок Б. Алтенберга, М. Баньки, У. Френсиса, М. Ґелерстема, Г. Кеннеді, Г. Ліча, А. Баранова, М. Михайлова, В. Рикова, Л. Ричкової, С. Шарова та ін. Теорія і практика програмного оброблення корпусних ресурсів — окремий напрямок.

На відміну від англо-саксоністики, романо-германістики та славістики корпусне мовознавство в сучасній лінгвоукраїністиці усе ще не оформилося ні як метод, ні як напрямок, ні як дисципліна. Існує кілька публікацій з корпусної лінгвістики, — розділ „Корпусна лінгвістика: предмет дослідження і завдання" у підручнику Є. Карпіловської „Вступ до комп'ютерної лінгвістики" [4], український сегмент у проекті TRACTOR. Це зумовлено тим, що у вітчизняному мовознавстві на момент виникнення корпусної лінгвістики чи точніше ідеї власне корпусної організації текстового матеріалу, домінували інші методики комп'ютерного подання та оброблення мовного матеріалу. Основну увагу сконцентровано на створенні формалізованих моделей опису та інтерпретації мовних явищ, побудові систем автоматичного морфологічного та синтаксичного аналізу тексту, машинного перекладу. А з кінця 70-х років XX ст. у Радянському Союзі дослідження і роботи з комп'ютерної лінгвістики головно зорієнтовані на створення машинних фондів національних мов [5], в архітектурі яких окреме місце займали бази текстових ресурсів, хоча й відмінні від електронних текстових корпусів, але по суті співвідносні з ними. Як слушно зазначає Л. Ричкова „для текстів на машинних носіях там (у машинних мовних фондах) відводилася роль виключно ілюстративному матеріалу для використання з метою об'єктивізації лінгвістичних даних. Незалежно від долі мегапроекту Машинного фонду російської мови, треба зазначити, що його концепція ілюструвала загальну для радянської лінгвістики того часу „зневагу" до корпусної підтримки лінгвістичних досліджень" [3:64; 4:189], що опосередковано зумовило відсутність корпусного напрямку і національного корпусу в лінгвоукраїністиці.

Отже, перед сучасною лінгвоукраїністикою, аналогічно до інших слов'янських лінгвістичних традицій, стоїть завдання, по-перше, увести у власну наукову парадигму напрямок корпусної лінгвістики, забезпечивши йому передумови розвитку. По-друге, створити концепцію *Національного корпусу української мови* як моделі організації емпіричних даних для лінгвістичних студій над українською мовою в умовах технологічного суспільства. По-третє, розробити процедуру його створення від етапу проектування до реалізації корпусно-базованих досліджень української мови різних періодів і форм її існування.

1. *Ельмслев Л.* Пролегомены к теории языка // Новое в лингвистике. — М., 1960 — Вып. 1. - С. 264-390.
2. *Ершов А. П.* Машинный фонд русского языка — внешняя постановка // ВЯ. — 1985. - № 2. - С. 27-34.
3. *Захаров В. П., Коваль С. А.* Корпусная лингвистика и лингвистические базы данных // НТИ: Информационные процессы и системы. — 2002. — № 7. — С. 62-69.
4. *Карпіловська Є. А.* Вступ до комп'ютерної лінгвістики. — Донецьк, 2003. — 183 с.
5. *Машинный* фонд русского языка: идеи и суждения. — М., 1986. — 239 с.
6. *Основные* направления структурализма / Отв. ред. М.М. Глухман, В.Н. Ярцева. - М., 1964. - 358 с.
7. *Рычкова Л.В.* Праблема састауных аб'ектау у корпусах славянскімоу і лінгвістычных базах дадзеных // Мовознауства. Літаратура. Культуралогія. Фалькларыстыка. XIII Міжнароднъї з'езд славыстау. Даклады беларускай дзлегацыі. — Мінськ, 2003. - С. 184-195.
8. *Benveniste E.* "Structure" en linguistique // Sens et usages du terme *structure* dans les sciences humaines et sociales. — 's-Gravenhage, 1962. — P. 38—39.
9. *Chomsky N.* The Acquisition of Language. — New York, 1964. — 187 p.
10. *McEnery T., WilsonA.* Corpus Linguistics. — http://www.comp.lancs.ac.uk, 1996.

Мовна мозаїка

НАШ **БЮЛЕТЕ́НЬ** ЗДОБУВАЄ ЩАСНИЙ ДЕНЬ

Вибори Президента України 2004 року – вікопомна подія для нашої Держави і всього світу. Вперше український народ повстав проти злочинної влади, проти її сваволі і корумпованості. Наш народ виборов право провести демократичні вибори й обрав Народного Президента Віктора Ющенка. І здається, що може бути простіше і могутніше за виборчий бюлетень?! За допомогою цього виборчого засобу можна змінити Долю нашої Країни, проголосувати за омріяну нами Свободу, за жадану Волю!

Отож про іменник іншомовного походження – тепер любе нам слово **бюлете́нь**. Віднині це слово заслуговує, щоб його вимовляли й писали правильно. В іменникові **бюлетень** наголос припадає на останній склад, а не на перший, що нерідко спостерігаємо в мовленні. Тобто вимовляймо не **бю́летень**, а **бюлете́нь**. У відмінкових формах маємо зберігати **е.** Тут **е** не випадне. Запам'ятаймо правильне наголошування слова **бюлете́нь** і його відмінювання. Наприклад: *Приходить він, навіки славний день, Коли зведем в мільйонах рук свобідних Наш виборчий, наш вольний* ***бюлете́нь*** (М.Бажан); *На засіданні суду відзначалось, що в Законі про вибори чітко не сформульоване положення про контроль видачі, одержання і погашення виборчих* ***бюлете́нів*** (газета "Сільські вісті").

Іван Вихованець