Орися ДЕМСЬКА

ТЕКСТОВИЙ КОРПУС: ПОНЯТТЯ І ВИЗНАЧЕННЯ

Коли народжується дитина, найперше, про що думають батьки, це те, яке ім'я їй дати. Так само, коли з'являється якась річ у світі, то люди починають міркувати, як її назвати. А в науці ще постає додаткове завдання: дати визначення цій речі. Те ж було і з корпусом. Коли у 1962 році в Університеті Брауна (США) почали працювати над укладанням першого в історії комп'ютерного, чи електронного, корпусу текстів сучасного американського варіанта англійської мови *Brown Corpus,* чи *корпус В*[1], перед його творцями постало завдання, по-перше, дати назву і, по-друге, дати визначення тієї реалії, яку вони створюють. Науковці, що працювали над цим проектом, пішли звичним для мовознавців шляхом і звернулися до словника. Усвідомлюючи, що зібрання тексту є певним віддзеркаленням мови, сказати б, її корпусом, як корпус тіла людини, вони почали розглядати тлумачення лексеми *корпус* у словнику англійської мови і натрапили там на одне зі значень, марковане як лінгвістичний термін. Далі про те, яким чином це відбувалося пише У.Френсис, один із творців *корпусу В:* «Четвертому значенню слова corpus 'корпус' у словнику англійської мови *"Random House Dictionary of the English Language"* відповідає таке: *лінг.* Сукупність висловлювань або речень, яку можна вважати представницькою для певної мови чи діалекту й використовувати для граматичного аналізу цієї мови або цього діалекту. Це визначення загалом суголосне з визначенням 3б у словнику *"Webster's Third New International Dictionary of the English Languages"*. Та й воно надто вузьке. Я пропоную розширити його так: сукупність текстів, які можна вважати представницькими для певної мови, діалекту або іншої підмножини мов, призначена для лінгвістичного аналізу»[2]. До слова скажемо, що і в академічному одинадцятитомному *«Словнику української мови»* серед восьми різних значень реєстрового слова *корпус* є й таке: «Повна збірка яких-небудь текстів». І якщо зважити на те, що том на літеру К датований 1973 р., коли не лише в українському мовознавстві, а й практично в усіх інших і гадки не було про електронний текстовий корпус, то можемо говорити, що це слово було готове, аби назвати майбутнє комп'ютерне зібрання текстів не тільки в англійській мові, а й в українській. Так зібрання текстів, що отримали електронну форму, були певним чином організовані й віддзеркалювали англійську мову, з легкої руки У.Френсиса і його колеги Г.Кучери почали називати ***корпусом***.

І якщо сам термін *корпус* залишився без змін понині й увійшов у всі мови та національні лінгвістичні терміносистеми, де розвиваються корпусні студії, то його визначення від 1962–1964 рр. (часу укладання *корпусу В*) зазнало змін та уточнень. Це сталося насамперед тому, що змінився сам корпус, можна образно сказати: виріс із немовляти і став зрілим, а також набула розвитку і теоретична частина корпусної лінгвістики. Тепер, щоб довільне зібрання текстів могло називатися корпусом, воно повинно мати певні важливі характеристики. По-перше, мати електронну форму і бути стандартно організованим, тобто відповідати вимогам міжнародного стандарту кодування корпусу[3]. По-друге, бути репрезентативним і збалансованим. По-третє, бути вичерпним; у корпусі ідеально має бути не забагато і не замало певних текстів або їхніх уривків. Зазвичай саме ці ознаки, характеристики, критерії нині є основними у визначенні поняття *корпус,* саме їхня наявність робить *корпус* корпусом. Та це далеко не всі його ознаки, критерії. Часто й інші критерії беруть за основу визначення корпусу, вважаючи їх засадничими. Наприклад, п'ять доконечних ознак – відібраність, репрезентативність, скінченність щодо обсягу, машиночитаність і стандартність – лягли в основу визначення поняття *корпус* одночасно в англійських і російського корпусних лінгвістів Т.МакЕнері, А.Вилсона і В.Рикова. Вони корпусом уважають «зібрання відібраних репрезентативних текстових даних довільної природної мови, що має скінченний обсяг, машиночитану форму та стандартну організацію»[4].

Найчастіше основними, визначальними ознаками текстового корпусу вважають машиночитаність чи наявність, по-перше, електронної форми подання і, по-друге, спеціальної системи кодування текстових даних, та репрезентативність, чи здатність корпусу правильно відображати мову або її частину. Є кілька однотипних визначень корпусу, в основу яких покладено ознаки машиночитаності та репрезентативності, наприклад: «зібрання машиночитаних текстів, відібраних таким чином, щоб максимально репрезентувати мову та її різноманіття»[5]; «корпуси – це велика кількість текстів природної мови, що мають комп'ютерну форму і є об'єктом

---

[1] Так подано назву цього корпусу в перекладній статті У.Френсис: Проблемы формирования и машинного представления большого корпуса текстов. Далі у тексті ми використовуватимемо саме цю форму назви.

[2] Фрэнсис У. Проблемы формирования и машинного представления большого корпуса текстов / У.Фрэнсис // Новое в зарубежной лингвистике. – 1983. – Вып. XIV. – С. 334.

[3] Corpus Encoding Standard (CES) є набором стандартів для здійснення корпусних досліджень і робіт, результати і дані яких можуть застосовуватися у галузі інформаційних технологій, машинного перекладу, комп'ютерної лексикографії тощо. Див. також: Ide N. Corpus Encoding Standard [Електронний ресурс] / N.Ide. – 2000. – Режим доступу: http://lpl.univ.-aix.fr/projects/multext/CES.

[4] Див.: McEnery T. Corpus linguistics [Електронний ресурс] / T.McEnery, A.Wilson. – Режим доступу: http://www.comp.lancs.ac.uk; Рыков В. Корпусная лингвистика [Електронний ресурс] / В.Рыков. – Режим доступу: http://rykov-cl.narod.ru/lekcii.doc.

[5] Ball C. Concordances and corpora [Електронний ресурс] / C.Ball. – Режим доступу: http://www. georgetown.edu/faculty/ ballc/corpora/tutorial3.html

певного лінгвістичного дослідження, де під 'природний' розуміють усе, що фактично було висловлено в усній або писемній формі»[6]. А доповнивши машиночитаність і репрезентативність ознакою застосовності корпусу в лінгвістичних дослідженнях, Н.Деш і Б.Чаудхурі визначають корпус як «зібрання лінгвістичних даних, скомпонованих або з писемних текстів, або із транскрибованих усних текстів, основною метою якого є перевірка гіпотез про мову»[7].

Однак серед сучасних визначень текстового корпусу на особливу увагу заслуговує визначення, що його запропонував Дж.Синклер. Так, учений, уперше висловивши думку про те, що корпус є «певною моделлю подання реалізації мовної системи»[8], крім набору корпусних вимог, звертає ще увагу на низку лінгвістичних критеріїв створення текстового корпусу. Далі ці критерії він поділяє на позамовні (соціальний аспект мовної комунікації, характер комунікативних актів, учасники комунікації тощо) і внутрішньомовні (особливості функціонування самої мови). Враховуючи корпусні вимоги і критерії, Дж.Синклер стверджує, що довільне зібрання текстів або текстовий архів не є корпусом, оскільки такі комп'ютерні текстові ресурси «не потребують відбору чи впорядкування за лінгвістичними критеріями, що, власне, кардинально відрізняє їх від корпусів»[9].

Крім загальнотеоретичного визначення електронного текстового корпусу Дж.Синклер увів до практичного обігу коротке, так зване робоче визначення, згідно з яким електронний текстовий корпус – це «зібрання, колекція відібраних і впорядкованих текстових уривків, що їх використовують як взірець мови»[10].

У сучасному українському мовознавстві, де корпусна лінгвістика усе ще не сформувалася як окремий напрям, унаслідок чого її зазвичай розглядають у межах комп'ютерної лінгвістики або рідше – комп'ютерної лексикографії, крім понять *корпус текстів*, *корпус даних*, поширені також терміни *база даних* і *база знань*, між якими нерідко ставлять знак рівності. Важко з цим погодитися, оскільки базою даних зазвичай вважають сукупність даних, підтримуваних в активному режимі, що відображає властивості об'єктів реального світу. База даних містить не лише дані, а й їхній опис, і ці дані не завжди є даними мови. А коли вже йдеться про мову, то говоримо про різновид бази даних – про лінгвістичну базу даних. Цей термін поширений у комп'ютерній лінгвістиці та значно менше – у корпусній лінгвістиці. Доволі часто він функціонує без визначення, але якщо визначення все-таки є, то вони переважно трактують лінгвістичну базу даних як мовний інформаційний масив, спроектований таким чином, щоб спростити введення й пошук даних. А під поняттям *база знань* узагальнено розуміють «структурований і формалізований набір відомостей про об'єкти предметної галузі та відношення між ними, на основі яких можна будувати судження про них, здійснювати різноманітні операції логічних умовиводів»[11]. На перший погляд важко побачити істотну відмінність між різними типами баз даних і корпусом, але основною відмінною ознакою є те, що бази даних, бази знань можуть торкатися не лише мовних, текстових реалій, а й будь-яких інших. Натомість корпус – це завжди текстовий корпус, незалежно від того, яким цей текст є: словом чи твором. Крім того, формат подання текстів, призначення і функції баз даних і баз знань та корпусу є різними.

Говорячи про термін *корпус*, також маємо звернути увагу на те, що в сучасному українському мовознавстві паралельно з поняттям *текстовий корпус* існує поняття *корпус мови*. Однак *корпус мови*, так само як і корпусне планування, – це актуальні поняття соціолінгвістики, де під *корпусом мови* зазвичай розуміють «мову як таку»[12], а корпусне планування – одне з двох типів (іншим типом є статусне) мовного планування, яке «має справу з владними зусиллями, спрямованими на характеристики мови як такої»[13]. Очевидно, що за таких умов термін *корпус* стає багатозначним, що він виник і побутує у різних галузях науки про мову, зокрема у соціолінгвістиці й корпусній лінгвістиці. Але щоб така багатозначність не стала проблемою, варто скористатися думкою Н.Дзюбишиної-Мельник, яка слушно пропонує вживати на позначення корпусу мови термін *тіло мови / тіло національної мови*[14]. Цей погляд заслуговує на увагу ще й тому, що дослідниця пропонує питомий термін, який доволі глибоко віддзеркалює суть поняття на противагу запозиченню *корпус мови*. Таким чином маємо розрізнення: *текстовий корпус мови* – це комп'ютерний ресурс, а *тіло мови* – це «внутрішня структура мови як такої»[15].

Узагальнимо: **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.**

Залежно від мети, способу використання, загальної структури, обсягу, формату подання, принципів добору текстів, критеріїв класифікації тощо сучасні комп'ютерні текстові корпуси можуть бути такими:

---

[6] Bańko M. Korpus tekstów jako źródło wiedzy o języku [íŠÍŮÚ] / M.Bańko // Wykład na sesji MSH Uniwersytetu Warszawskiego (Rękop.). – Warszawa, 2003. – S. 2.

[7] Dash N. The process of designing a multidisciplinary monolingual sample corpus / N.Dash, B.Chaudhuri // International journal of corpus linguistics. – 2000. – Vol. 5. – № 2. – P. 78.

[8] Sinclair J. Corpus typology – A framework for classification / J.Sinclair // Studies in anglistics / ed. by G.Melchers, B.Warren. – Stockholm : Almqvist & Wiksell, 1995. – P. 17.

[9] Там само. – С. 18.

[10] Там само. – С. 18.

[11] Карпіловська Є. Вступ до комп'ютерної лінгвістики / Є.Карпіловська. – Донецьк : Юго-Восток, 2003. – С. 47.

[12] Фішман Дж. Не кидайте свою мову напризволяще / Дж. Фішман. – К. : К.І.С., 2009. – С. 29.

[13] Там само. – С. 29.

[14] Див.: Дзюбишина-Мельник Н. Тіло національної мови / Н.Дзюбишина-Мельник // Магістеріум. – Вип. 37 : Мовознавчі студії. – К. : ВД «Києво-Могилянська академія». – 2009. – С. 24–27.

[15] Там само. – С. 24.

За метою створення та призначенням	
ДОСЛІДНИЦЬКИЙ застосовується у лінгвістиці для формулювання нових теорій, концепцій, теоретичних положень чи гіпотез про мову	ІЛЮСТРАТИВНИЙ релевантний для підтвердження вже висловлених теоретичних положень чи гіпотез про мову
МОНІТОРИНГОВИЙ дає змогу перманентно відстежувати зміни в мові	РЕФЕРЕНЦІЙНИЙ забезпечує якомога різноманітнішу інформацію про мову на певному синхронному зрізі
СИНХРОННИЙ репрезентує мову певного (сучасного, історичного) часового проміжку	ДІАХРОННИЙ репрезентує мову впродовж певного (довшого) історичного проміжку часу
ДИНАМІЧНИЙ сфокусований на динаміці мови, розвиткові мови, змінах у мові з урахуванням діахронії	СТАТИЧНИЙ засвідчує стан мови на певному синхронному зрізі
ЗАГАЛЬНОМОВНИЙ репрезентує загальнонародну національну мову; скерований на розв'язання наукових дослідницьких завдань, зорієнтованих на мову народу в усіх її виявах	СПЕЦІАЛЬНИЙ репрезентує певний мовний зріз або рівень, фрагмент мови; скерований на розв'язання часткових, галузевих, специфічних наукових дослідницьких завдань
За типом текстового матеріалу	
ПОВНОТЕКСТОВИЙ тексти у корпусі подано повністю	ФРАГМЕНТНИЙ у корпусі подано уривки текстів
УСНИЙ конститутивним матеріалом корпусу є лише тексти усного мовлення	ПИСЕМНИЙ конститутивним матеріалом корпусу є лише писемні (друковані) тексти
ОДНОМОВНИЙ до корпусу ввійшли тексти лише однієї мови	КІЛЬКАМОВНИЙ до корпусу ввійшли тексти двох або більше мов
ПОРІВНЯЛЬНИЙ корпуси різних мов або різних варіантів однієї мови, структура, текстові дані, принципи опрацювання яких є однаковими	ПАРАЛЕЛЬНИЙ єдність підмножини оригінальних текстів та підмножини їх перекладів на іншу(і) мову(и)
За типом програмного оброблення	
АНОТОВАНИЙ наявна формалізована інформація щодо одиниць зберігання корпусу	НЕАНОТОВАНИЙ відсутня формалізована інформація щодо одиниць зберігання корпусу

Кожен з цих текстових корпусів національних мов може використовуватися у різних типах наукового і/або практичного вивчення мови. Скажімо, для історичних досліджень мови важливими, інформативними є передусім діахронні корпуси, які дають змогу з'ясувати появу в мові певного слова, його відмінкові форми, семантику, переміщення в межах лексичної системи, зникнення з мови і т.ін. Важливо, що дослідницьким об'єктом може бути і семантично повнозначна, і неповнозначна одиниця зі своїм контекстним оточенням. Для словникарства кожен із корпусів буде важливим. І важливість корпусу залежатиме від того, який саме словник укладатиметься за допомогою корпусу: загальномовний – відповідно основна увага словникарів зосереджуватиметься на загальномовному текстовому корпусі; діалектний – на діалектному корпусі; словник мови письменника – на т.зв. авторському корпусі і т.д. Крім того, за допомогою текстових корпусів можна здійснити низку інших корпусних наукових досліджень мови, зокрема типології відмінкових форм іменних частин мови, специфіки вживання слів із різним лексико-граматичним значенням у стилістично і жанрово різних текстах, структури синтаксичних конструкцій різного типу, мінімальної/максимальної/середньої довжини речення; обстежити лексичний склад мови, базовий словник мови, статистичні характеристики лексичних одиниць різного типу, неологізми/архаїзми, фразеологічні одиниці та специфіку їхнього функціонування тощо. І, врешті, усі ці наукові дослідження різних аспектів мови на основі корпусу так чи інакше втіляться у шкільних підручниках про мову.

Іншим поглядом на використання текстового комп'ютерного корпусу є, очевидно, дидактика. Тут особливо важливий корпус національної мови, що охоплює мову як таку в усіх її варіантах, може слугувати джерелом інформації про те, як правильно сполучувати, відмінювати, вживати слова, як формулювати речення, які фразеологізми в яких ситуаціях правильно вживати тощо. Так само корпусний матеріал можна використовувати для укладання практичних вправ з української мови, уривок корпусного тексту може бути й диктантом. Загалом найважливіше, що різні корпуси призначені забезпечувати найрізноманітніші потреби і вченого, і викладача, і вчителя, і звичайної людини, для якої мова є джерелом думки, натхнення, творчості. На превеликий жаль, наша мовознавча наука, хоча й повністю готова до того, щоб розпочати роботу над укладанням передусім великого загальномовного текстового корпусу сучасної української мови, а також і низки менших, спеціальних корпусів, проте не виявляє в цьому помітної активності. Та чим далі, тим жорсткіше стоятиме проблема текстового комп'ютерного корпусу в нашій науці та культурі. І це далеко не маргінальна проблема вузького кола адептів комп'ютера, як хотілося б багатьом вважати, а проблема, яка знову і знову ставить питання, порушене М.Хвильовим ще 1925 р.: «Європа чи просвіта?»[16]. І хоча ми воліємо уникати категоричності в цьому питанні, оскільки і «Європа», і «просвіта» є важливими складниками розвитку науки та культури, кожна з яких має своє місце, функцію, призначення, мету, сферу тощо, наука в сучасному світі не може бути «просвітянською». Тому, перефразовуючи питання «Європа чи просвіта?», ставимо питання «Корпус чи традиційний текст і/або картотека?». Що з цього буде – побачимо.

[16]Цит. за: Ш к а н д р і й М. Модерністи, марксисти і нація: Українська літературна дискусія 1920-х років / М.Шкандрій. – К. : Ніка-Центр, 2006. – С. 25.