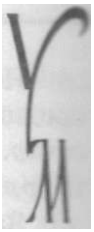
Корпусна лінгвістика



Орися Демська-Кульчицька

РЕПРЕЗЕНТАТИВНІСТЬ ЯК ОЗНАКА ТЕКСТОВОГО КОРПУСУ

Сучасні дослідження писемних або розмовних текстів не обмежені лише корпусною лінгвістикою. Окремий текст часто, а інколи виключно, є об'єктом літературознавчого чи мовознавчого аналізу, який не обов'язково повинен здійснюватися в межах корпусної лінгвістики. Також не доцільно розглядати як корпусний, підхід до вивчення мовних феноменів з використанням електронних бібліотек чи повно текстових баз даних, яким не притаманні ті ознаки, на підставі яких текстове зібрання вважають корпусом.

На сьогодні корпусна лінгвістика ще не завершила процес вироблення єдиного погляду на корпусні ознаки, наявність яких у електронного зібрання текстів перетворює його на корпус. Уперше про необхідність визначення набору релевантних корпусних ознак сказав Дж. Синклер: „Корпус повинен мати характеристики, значення яких є „значенням за промовчанням". Йдеться про ті параметри, які апріорі повинно мати зібрання текстів природної мови, щоби вважатися корпусом" [9:27]. До таких детермінативних параметрів текстового корпусу, за Дж. Синклером, належать обсяг, автентичність, машиннеподанняідокументованість. Цейперелікмодифікує Ґ. Кеннеді [7], який обов'язковими корпусними ознаками визначив статичність або динамічність, репрезентативність, збалансованість, обсяг. Вимогу репрезентативності крім Ґ. Кеннеді, визнають також А. Баранов і В. Риков. Останній вважає, що саме ця властивість, за задумом основоположників корпусної лінгвістики, перетворює набір текстів на машинному носії на унікальну словесну єдність — корпус текстів. Обсяг вважають важливою ознакою корпусу А. Баранов, В. Риков, Дж. Синклер. Крім

обсягу, А. Баранов додатково пропонує параметри повноти, економності, структурованості матеріалу та комп'ютерної підтримки [1:116]. А В. Риков — ще розміщення корпусу на машинному носії, стандартне подання словесного матеріалу на цьому машинному носії, що дозволяє застосовувати стандартні програми до його оброблення.

Отже, перелік релевантних корпусних ознак складатимуть:

1) обсяг;
2) автентичність;
3) репрезентативність;
4) збалансованість;
5) електронна форма або комп'ютерна підтримка;
6) документованість;
7) простота подання;
8) повнота;
9) економність;
10) структурованість;
11) статичність або динамічність.

Проте цей перелік вимагає певного критичного аналізу, оскільки не всі з ознак є однаково важливими для сучасного електронного текстового корпусу.

Ідея створення корпусу текстів природної мови полягала в тому, щоби оперувати великими обсягами слів у реальних контекстах, і фактичний матеріал обсягом 1 млн слововживань вважався достатньо великим на початках так званої електронної корпусної лінгвістики. Але сьогодні комп'ютерні технології дозволяють працювати з набагато більшими текстовими ресурсами і, за умови існування надвеликих текстових корпусів, обсяг яких сягає понад за 100 млн. слововживань і рідко зупиняється на 1 млн, дотримання вимоги обсягу стає непринциповим. Аналогічно не буде дотримано вимоги обсягу, коли йтиметься про корпуси, предметна галузь яких не досягатиме 1 млн слововживань.

Частково релевантною у великих за обсягом корпусах є повнота, яка, за А. Барановим, корелює з репрезентативністю. І ця кореляція має такий характер: „Репрезентативність корпусу вказує на те, що одиниці проблемної галузі відображено пропорційно в корпусі даних, але при певному порозі деякі релевантні явища зникнуть з корпусу. Повнота вимагає врахування релевантних явищ, навіть якщо це не відповідає ідеї пропорційного звуження проблемної галузі" [1:117]. Власне ця кореляція нівелює параметр повноти, оскільки із підтримкою репрезентативності одночасно забезпечується повнота текстового матеріалу. Вимога повноти релевантна для спеціальних і малих корпусів, де проблемою перших є якомога повніше відобразити аналізоване мовне явище, а других — не пропустити якихось мовних явищ через обмеженість обсягу корпусу.

Економність як „економія зусиль при вивченні предметної галузі, відображеної в корпусі", детермінує, за визначенням А. Баранова, „поріг відображення предметної галузі" [1:116-117], тобто чим еко-

номнішим є корпус, тим вищий поріг відображення й тим зручніший він у користуванні. Погоджуємося, що надвеликий текстовий корпус справді ускладнює роботу користувача, але завжди можна працювати з підкорпусом або ж удосконалити пошукову програму.

Корпусні ознаки статичності або динамічності формулює лише Ґ. Кеннеді. Але тут очевидно швидше йдеться про методики відображення предметного домену в корпусі. Так, якщо в корпусі репрезентовано якийсь чітко детермінований проміжок часу і стан предметного домену, то корпус буде статичним. Альтернативою є динамічний, чи моніторинговий корпус, у якому предметний домен не обмежений часово. Проте, на наш погляд, статичність та динамічність не слід розглядати як релевантну ознаку корпусу, оскільки це, все-таки, типологічна характеристика текстового корпусу.

Отже, обсяг, структурованість, повнота, економність і статичність або динамічність, на наш погляд, не вимагають повної реалізації в корпусному об'єкті. Якщо ці ознаки є необов'язковими або частково обов'язковими, то репрезентативність, автентичність, відібраність, збалансованість, машиночитаність — вважаємо критичними корпусними ознаками, без яких текстове зібрання не можна кваліфікувати як корпус, де:

— **репрезентативність** полягає в здатності корпусу відображати всі властивості предметної галузі. Під предметною галуззю тут розуміємо 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 при 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;

— **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 використо-

вують різні методики відбору текстового матеріалу до корпусу, одним із доволі складних завдань є досягнення збалансованості. Щоби досягнути збалансованості корпусу, Дж. Синклер пропонує мінімальні критерії для відбору текстів, які „повинні включати розрізнення між художньою літературою і нехудожньою літературою; книжкою, журналом або газетою; нормативним і ненормативним варіантом мови; з контролем віку, статі та походження авторів" [10]. У цьому полягає стратифікаційний підхід Дж. Синклера;

— **машиночитаність** є визначальною ознакою 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 уже параметр „за промовчанням", тобто іншим сучасний текстовий корпус не може бути.

Не випадково репрезентативність стоїть на першому місці. Це та невід'ємна ознака сучасного електронного текстового корпусу, без якої текстове зібрання назване корпусом не виконуватиме покладених на нього завдань, а категоричніше — не вважатиметься корпусом у термінах корпусної лінгвістики [3].

Щодо репрезентативності А. Баранов, зокрема, зазначає: „створення репрезентативного корпусу даних є необхідною умовою практично будь-якого лінгвістичного дослідження. Незважаючи на очевидність цього положення, далеко не в усіх лінгвістичних дисциплінах на це взагалі звертають увагу. І зовсім рідко порушують питання про методологію досягнення репрезентативності. Тут лінгвістика, на жаль, істотно відстає не тільки від таких природничих наук, як фізика, хімія, біологія, але й навіть від таких гуманітарних спеціальностей, як соціологія й психологія" [2]. На відміну від традиційного мовознавства репрезентативність, яка є тією важливою ознакою, що дистанціонує корпус від інших електронних текстових ресурсів, в корпусній лінгвістиці займає центральне місце.

Під **репрезентативністю,** на наш погляд, дещо спрощено розуміють „те, що ми вважаємо типовими й центральними аспектами мови, що дає достатню кількість уживань слів і висловлювань для лексикографів і тих, хто вивчає мову як іноземну, і що дозволяє одержувати з корпусу достатню кількість даних, які дають реалістичну картину функціонування лексики" [10]. Таке обережне й доволі нечітке тлумачення переформулюємо, визначивши **репрезентативність** як *релевантне відображення предметної галузі — певного реалізаційного рівня(ів) мовної системи — в корпусних даних, де задано в пропорції, детермінованій реальною частотою досліджуваного явища, всі властивості відтворюваного в корпусі домену.* Тобто відносна частота довільного явища в корпусі має наближатися до відносної частоти цього явища в досліджуваній мові чи субмові.

Категорія репрезентативності корпусу є релятивною. Не можна говорити про репрезентативність загалом, оскільки остання залежить від типу корпусу, його призначення та мети створення, що мотивує

відповідно різні принципи добору фактичного матеріалу. Так, якщо доменом є мова окремого автора, інакше кажучи якщо йдеться про спеціалізований корпус, то репрезентативність останнього буде досягнута лише внаслідок уведення до нього всіх автентичних текстів, створених автором. Таким чином, реалізовуватиметься принцип тотальної колекції. Але очевидно, що загальномовний корпус через надвелике число реальних текстів не може охопити всі створені носіями мови тексти, тому принцип тотальної колекції не працюватиме в досягненні репрезентативності загальномовного корпусу. Тут стоїть завдання зробити коректну вибірку, що базуватиметься на загальних і / або індивідуальних критеріях відбору емпіричного матеріалу.

Обговорюючи проблему репрезентативності корпусу, традиційно йдеться про розмір уривка як важливий аспект досягнення репрезентативності, про кількість текстів з такою, а не іншою стилістично-жанровою, тематичною, структурною тощо характеристикою в корпусі та кількість слів у кожному текстовому фрагменті. Хоча Д. Байбер не вважає обсяг вибірки головним аспектом репрезентативності й слушно зазначає, що швидше визначення домену та рішення щодо методу вибірки є важливішими [6].

Загалом, передумовою досягнення репрезентативності є коректна параметризація відображуваного в корпусі рівня мови. Тут виникають питання, без відповіді на які неможливо як здійснити саму параметризацію, так і, відповідно, досягти репрезентативності. Маємо на увазі цільовий аспект, тобто для чого, для яких саме досліджень будований корпус має бути репрезентативним. Цільовий аспект безпосередньо пов'язаний з кореляцію між реалізаційним рівнем / підрівнем мови, забезпеченим у корпусі. Наприклад, можливість адекватного застосування корпусу термінолекту журналістів для дослідження динаміки лексичних змін у мові. Очевидно, що таке застосування термінолектного корпусу експлікує нерепрезентативність його щодо застосування, яке вимагатиме значно ширшого охоплення мови. Про це також говорить В. Риков: „Чи можна двомовний навчальний корпус текстів застосувати в системах машинного перекладу? Наскільки він буде адекватний (репрезентативний) для цього завдання? Або чи відбиває складений за всіма канонами корпус текстів газетних політичних метафор усе мовне різноманіття газетної прози? Чи буде цей корпус репрезентативний для лінгвістичних досліджень газетної прози?" [4].

Найчастіше безпосередньо визначити обсяг предметної галузі, відображуваної в загальномовному корпусі, неможливо, оскільки остання не обмежена, тобто не обмежена кількість реальних писемних і / або усних текстів, згенерованих носіями певної мови. За таких умов доцільно звернутися до так званих непрямих класифікаційних методів досягнення репрезентативності корпусної побудови, які хоча й „не дають стовідсоткового результату, однак дозволяють максимально врахувати всі релевантні явища предметної галузі, навіть якщо немає гарантії відбиття відносної частоти досліджуваного феномена" [1:118].

Застосування класифікаційних методів досягнення репрезентативності корпусу ґрунтується на встановленні максимальної кількості релевантних параметрів та можливих їхніх комбінацій, на підставі яких ідентифікують функціонування мовних одиниць, що вводяться до корпусу. Наприклад, А. Баранов пропонує такий набір релевантних параметрів та їхніх комбінацій для корпусу заголовків газет: „1) часовий параметр (дата d1… dn); 2) рубрика газети ("Передові статті", "Листи з місць", "Культура", "Спорт" і т. ін.); 3) синтаксична структура заголовка ("Розповідне речення", "Питальне речення"; "Номінативне речення" і т. ін.); 4) наявність / відсутність мовної гри тощо. Усі можливі комбінації значень параметрів задають клітинку матриці, яку заповнює заголовок, знайдений у проблемній галузі. Наприклад, заголовок у корпусі може мати такий набір значень параметрів: <d1; „Листи з місць"; „Розповідне речення"; „Без мовної гри" [2].

Технічно досягнення репрезентативності текстового корпусу можливе через застосування однієї з чотирьох методик відбору:

— пропорційного;

— стратифікованого;

— суцільного з наступною оцінкою репрезентативності;

— випадкової вибірки.

Суть пропорційного відбору полягає у введені фактичного матеріалу до корпусу з урахуванням значимості та ваги кожного типу тексту в мові. Пропорційна стратегія хоча і є продуктивною для здійснення деяких важливих досліджень мови, зокрема пов'язаних із квантитативністю, все ж має низку слабких місць. Так, за умови застосування цієї стратегії в корпусі домінуватиме мова засобів масової інформації та офіційних документів, натомість такі важливі для розвитку й функціонування мови тексти, як художня проза, поезія й драматургія, наукова й навчальна література або виявляться поданими мінімально, або взагалі не потраплять до корпусу.

Ідея стратифікованої вибірки, яка полягає в поданні однакових за обсягом фрагментів різних типів текстів, незалежно від їх значення і ваги в мові, належить Д. Байберу. Стратифікаційний підхід вимагає каталогізації різних категорій реальних мовних текстів, які на підставі ієрархічних критеріїв (таких як жанр, комунікативний канал, стиль і т. ін.), об'єднані в "страти" (від лат. stratum — настил, шар, послідовність). І наступний відбір фактичного матеріалу відбуватиметься не з довільних текстів з певними параметрами, а зі "страт", тобто сукупностей текстів з однаковими типологічними ознаками. Стратифікована вибірка виявляє більшу перспективність, ніж пропорційний відбір, але слід пам'ятати, що „ефективна стратифікація і, відповідно, досягнення репрезентативності через застосування методики вибірки може знову ж таки бути виконана лише на основі чіткого і добре обґрунтованого бачення генеральної вибірки корпусу" [6].

Можна погодитися, що стратифікаційна методика продуктивніша щодо пропорційної в досягненні репрезентативності корпусної побу-

дови, але вона також недосконала, оскільки визначення "страт" є доволі суб'єктивним процесом. Тут погоджуємося з думкою А. Мерменсолін, яка, виходячи з досвіду побудови національного корпусу шотландської мови, ставить під сумнів застосування стратифікаційного методу, акцентуючи увагу власне на суб'єктивності самої стратифікації, її основне питання: „Чи вдалося нам застосувати неупереджені принципи відбору усних і писемних текстів у межах різних типів комунікацій в різних культурах?" [8:617]. Виділення або невиділення тих або інших "страт" може значно вплинути на репрезентативність корпусу загалом.

Найчастіше застосовують технологічно простіший спосіб досягнення репрезентативності — суцільного відбору даних з наступною оцінкою репрезентативності. За умови суцільного відбору до корпусної побудови послідовно вводять усі релевантні явища цільового домену, які становлять інтерес у зв'язку з призначенням корпусу або завданнями дослідника. Основна проблема застосування технології суцільного відбору даних полягає в тому, що, по-перше, оцінку репрезентативності здійснюють post factum, і, по-друге, способи оцінки репрезентативності корпусу неможливо спрогнозувати, вони безпосередньо залежать від тих феноменів, які введені до корпусу.

Прийом суцільного відбору з наступною оцінкою репрезентативності має свої переваги перед пропорційним та стратифікаційним, особливо коли йдеться про спеціальні корпуси, але його доволі складно застосувати до великих і надвеликих загальномовних корпусів. Можливий, правда, варіант застосування суцільного відбору з наступною оцінкою до субкорпусів генерального корпусу, якщо такі передбачені в його структурі.

Крім методик пропорційного, стратифікаційного і суцільноп відбору, можливе також застосування методики випадкової вибірки Власне ця методика була застосована під час побудови *Браунівського корпусу.* Суть випадкової вибірки полягає в попередньому визначенні загального обсягу і структури корпусу та обсягу конститутивних фрагментів, інакше кажучи в параметризації цільового домену і наступній випадковій процедурі добору фактичного матеріалу [5]. Не зважаючи на певну псевдометодичність, випадкова вибірка дала свої результати: *Браунівський корпус* і до сьогодні є певним еталоном корпусної побудови, а прийом випадкової вибірки використовують у процесі створення сучасних текстових електронних корпусів. Але слід пам'ятати, що випадкова вибірка не завжди даватиме бажаний результат, особливо коли йдеться про обсяги менші, ніж один мільйон слов - вживань.

Отже, репрезентативність як детермінативний параметр сучасного електронного корпусу текстів становить окрему проблему в корпусній лінгвістиці, яка пов'язана з: а) розумінням репрезентативності б) характером репрезентативності й в) методами досягнення репрезентативності.

1. *Баранов А. Н.* Введение в прикладную лингвистику. — М., 2001. — 358 с.
2. *Баранов А. Н.* Проблема репрезентативности корпуса текстов // Труды Международного семинара Диалог-2001 по компьютерной лингвистике и ее приложениям. — 2001. — http://www.dialog-21.ru.
3. *Рыков В. В.* Корпус текстов как отражение состояния русского языка. — 2002. — http://rykov-cl.narod.ru.
4. *Рыков В. В.* Корпусная лингвистика. — 2001. — http://rykov-cl.narod.ru/lekcii.doc.
5. *Френсис У. Н.* Проблемы формирования и машинного представления большого корпуса текстов // Новое в лингвистике. — 1983. — Вып. XIV. — С. 334—352.
6. *Biber D.* Representativeness in corpus design // Literary and Linguistic Computing. — 1993. — Vol. 8. — №4. — P. 243—257.
7. *Kennedy G.* Introduction to Corpus Linguistics. — London — New-York, 1998. — 309 p.
8. *Meurman-Solin A.* On the morphology of verbs in Middle Scots: present and present perfect indicative // History of Englishes: New methods and interpretations in historical linguistics. — Berlin, 1992. — P. 611—623.
9. *Sinclair J.* Corpus, Concordance, Collocation. — Oxford, 1991. — 137 p.
10. *Sinclair J.* Corpus Typology Draft. — 1996. — http://www.icl.pi.cnr.it.