УДК 81’322

Василь Старко
Східноєвропейський національний університет імені Лесі Українки,
Наталія Чейлитко
Київський національний університет імені Тараса Шевченка

ПАРАМЕТРИЗАЦІЯ КОРПУСУ ЯК СПОСІБ ПІДВИЩЕННЯ ЙОГО РЕПРЕЗЕНТАТИВНОСТІ ТА ЗБАЛАНСОВАНОСТІ

Розглянуто різні методи параметризації корпусу з метою підвищення його репрезентативності та збалансованості. Обґрунтовано, що оптимального ефекту можна досягти завдяки виваженій комбінації різних підходів до параметризації корпусу з огляду на його специфіку.

Ключові слова: корпус, корпусна лінгвістика, параметризація, репрезентативність, збалансованість, вибірка.

Із-поміж вимог, що їх висувають до корпусів, постійно повторюються дві ключові – репрезентативність і збалансованість. Попри те, що вимоги ці були сформульовані ще на етапі становлення корпусної лінгвістики як окремої галузі знань, вони анітрохи не втратили актуальності. За цей час постали нові типи корпусів, вироблено нові підходи до їх укладання, а тому виникла потреба узагальнити, систематизувати напрацювання корпусної лінгвістики.

Основна теза цієї праці полягає в тому, що розв’язання центральних завдань, пов’язаних із репрезентативністю й збалансованістю корпусу, передбачає його докладну та виважену параметризацію. Серед українських мовознавців на потребі параметризувати різні види корпусів наголошували, зокрема, Є. Карпіловська [Карпіловська 2003 : 76] та О. Демська-Кульчицька [Демська-Кульчицька 2005 : 59]. Під параметризацією корпусу маємо на увазі його побудову за визначеним набором параметрів. Параметризація корпусу відбиває (попередньо проведену) параметризацію досліджуваної ділянки мови й має щонайменше такі виміри: 1) широту (охоплення максимальної кількості типів

текстів першого рівня); 2) глибину (виокремлення підкатегорій на нижчих рівнях); 3) пропорційність (заповнення окремих "клітинок" текстами, дібраними в певній пропорції); 4) часовий вимір.

Без репрезентативності будь-які результати дослідження мовного матеріалу, вміщеного в корпусі, чинні лише щодо цього корпусу – їх не можна екстраполювати на відповідну ділянку мови, бо незрозуміло, як корпус із нею співвідноситься. Відтак корпус втрачає свою роль інструмента дослідження мови. Ідеальної репрезентативності можна досягти за умови, що відома чітко визначена й параметризована генеральна сукупність, з якої належить зробити вибірку. Позаяк подати достовірний опис такої сукупності, зокрема визначити реальні пропорції всіх можливих типів текстів, неможливо, творці корпусів мусять задовольнятися наближенням до ідеальної репрезентативності [Leech 2007].

Однією з необхідних складових репрезентативності корпусу є його збалансованість. Збалансованість розуміють переважно як пропорційність різних частин корпусу: частка кожного типу текстів у корпусі має бути пропорційною – однак щодо чого? Справді пропорційний корпус можна побудувати з огляду на три вихідні позиції [Biber 1993]: 1) творці тексту (у цьому разі пропорції визначають відносно кількості створених текстів за певний період часу); 2) реципієнти (пропорції встановлюють відповідно до кількісного співвідношення текстів, сприйнятих носіями мови під час комунікативної діяльності – усної чи письмової); 3) самі тексти (пропорції визначають відносно розподілу текстів за різними жанрами). Однак у жодному серйозному корпусі не зроблено спроби досягти такої пропорційності. Річ у тім, що в кожен момент часу універсум живої мови складається переважно з усного приватного мовлення й повністю пропорційний корпус містив би близько 90 % записів бесід, 3 % листів і нотаток і 7 % решти жанрів [Biber 1993]. На практиці в корпусах усна частина якщо й наявна, то значно поступається писемній. Проте навіть у межах писемного корпусу досягти ідеальної збалансованості фактично неможливо.

Репрезентативність і збалансованість можуть мати множинні виміри й потребують певної точки відліку. Іншими словами, ціл-

ком резонно запитати: Корпус репрезентативний / збалансований щодо чого? Корпус збалансований відносно чого? Один із аспектів, який майже не враховують під час планування корпусів, – розмежування функцій мовця й слухача. Зазвичай два тексти вважають рівноцінними – як результати мовної діяльності їх було створено один раз. Однак із погляду реципієнтів текстів можна говорити про "індекс сприймання", тобто скількох слухачів / читачів текст досягнув. До того ж, різні аспекти цих понять можуть входити в суперечність. Наприклад, корпус, репрезентативний і збалансований щодо типів текстів за обсягом їх продукування, не буде репрезентативним за типами мовних явищ [Biber 1993]. Отже, поняття репрезентативності й збалансованості не лише градуйовані й багатовимірні, а й певною мірою неоднозначні.

Одним із поширених підходів до здійснення параметризації корпусу є **орієнтування на авторитетний корпус**, як-от Браунський корпус (Brown Corpus), цебто побудова нового корпусу за параметрами взірцевого.

Параметризація на основі фахової оцінки. Експертів залучають для оцінювання різних величин, які є базисом вироблених згодом параметрів корпусу. Індивідуальні оцінки бажано усереднювати.

Параметризація із застосуванням формальних методів. В обмежених випадках, коли відома генеральна сукупність текстів, можна скористатися статистичними формулами отримання достовірної вибірки. Коли ж цю сукупність неможливо проаналізувати, послуговуються непрямими методами. Приміром, Д. Байбер [Biber 1989; 1993] запропонував будувати типологію текстів, а отже й параметризувати корпус, на основі статистичних показників, які характеризують дистрибуцію кластерів мовних явищ у різних типах текстів. Соціологічні методи побудови вибірки застосовують, коли вдаються до демографічної моделі (див. нижче).

Параметризація з опертям на списки джерел. За цього підходу використовують національні бібліографічні індекси, списки бестселерів, списки найчастіше запитуваних книжок у бібліотеках, опис фондів конкретної бібліотеки. Можна використати

стохастичні методи й статистичні формули, щоб визначити, уривки зі скількох книжок потрібно ввести до корпусу, щоб отримати репрезентативну вибірку.

Параметризація з позицій отримувачів тексту. Пропорційність можна забезпечити шляхом відтворення кількісного співвідношення різних типів текстів, сприйнятих їх реципієнтами: застосовуючи звичні методи соціологічного опитування, побудувати репрезентативну демографічну вибірку й дослідити, скільки часу реципієнти сприймають усні повідомлення та письмові тексти різних жанрів.

Параметризація на основі внутрішньомовних критеріїв. Єдина відома нам концепція такого типу розроблена в працях Д. Байбера [Biber 1989; 1993]. Автор розробив метод оцінки розподілу мовних явищ за типами текстів на основі статистичних критеріїв із використанням факторного й кластерного аналізу. Суть методу полягає в тому, що типи текстів можна визначати за кластерами типових для них мовних ознак. Статистичні показники також придатні для визначення мінімального обсягу корпусу, репрезентативного щодо однієї або низки мовних ознак.

Параметризація на основі зовнішніх (ситуаційних) критеріїв. На сьогодні це найпоширеніший метод параметризації, застосований у багатьох значних корпусах. За такого підходу важить як широта, так і глибина типології [Leech 2007]. Кожна клітинка утвореної матриці підлягає заповненню відповідними текстами, зазвичай у певній пропорції. Визначення цих пропорцій становить окрему проблему.

Параметризація відповідно до потреб користувачів. Чимало корпусів, зокрема національних, покликані забезпечити найрізноманітніші пошукові запити користувачів. У такому разі важливо забезпечити гнучкість корпусу, зокрема дати користувачам змогу формувати підкорпуси за необхідними параметрами.

Динамічний підхід передбачає методику побудови моніторингових корпусів. Дж. Синклер [Sinclair 1991 : 24–26] обстоював ту ідею, що з лавиноподібним зростанням електронних текстів стане непотрібним конструювання малих корпусів на основі ви-

падкових виборок для інтенсивного дослідження. Натомість доцільним виявиться укладання відкритого корпусу, який постійно поповнюватиметься. Автор назвав корпуси такого типу *моніторинговими* (monitor corpora), позаяк вони схоплюють поточний "стан мови" за кожен інтервал часу. На основі моніторингового корпусу можна формувати менші корпуси за заданими параметрами.

Циклічний підхід. На основі попереднього досвіду й теоретичних положень формується початковий корпус (наприклад, визначається матриця типів текстів і пропорція кожного різновиду), після чого здійснюється емпіричне дослідження вміщеного в ньому матеріалу й визначаються шляхи розвитку корпусу. Д. Байбер запропонував циклічну модель створення корпусу: первісний проект корпусу, збирання текстів, емпіричне дослідження мовної варіативності (ключове поняття для автора), перегляд структури корпусу і наступний цикл [Biber 1993].

Використання ресурсів Інтернету. Нове покоління надвеликих корпусів створюють із використанням ресурсів Інтернету, зокрема за допомогою пошукових сервісів. Про збалансованість чи репрезентативність у цьому випадку не йдеться. Достовірність таких корпусів забезпечена їхнім обсягом. Однак параметризація текстів має бути запроваджена, оскільки вона забезпечує виконання корпусом низки важливих функцій.

Метод стратифікації полягає у формуванні класифікаційної матриці, кожне вічко якої має бути заповненим. У такий спосіб підвищують збалансованість. Цей загальний метод набуває конкретного втілення залежно від об'єкта стратифікації: текстів, суб'єктів усного мовлення чи релевантних ознак досліджуваної ділянки мови.

Метод пропорційної вибірки. Відомий також як метод пропорційного звуження, він має на меті досягти пропорційної репрезентативності, за якої відносна частота досліджуваного явища в корпусі близька до його відносної частоти в досліджуваній ділянці. Однак вищезгаданий метод наражається на кілька перешкод: обсяг ділянки мови найчастіше неможливо встановити (вона, як правило, постійно зростає), існує поріг відображення

(тобто за межами вибірки можуть виявитися менш частотні мовні явища), послідовне застосування пропорційності проблематичне (через значну перевагу усного мовлення).

Метод суцільної вибірки. Цей метод технологічно найпростіший, але призводить до розбалансованості. Повноту охоплення (наприклад, певних мовних явищ) можна оцінити лише постфактум. Проте такий підхід за умови здійснення додаткового редагування корпусу може виявитися дуже корисним, зокрема, під час укладання ілюстративного корпусу.

Метод випадкової вибірки. Проста випадкова вибірка може призвести до того, що периферійні шари лексики буде охоплено недостатньо. Одним із шляхів розв'язання цієї проблеми є застосування *стратифікованої випадкової вибірки*, коли спершу визначають категорії текстів (та їхню частку в загальному корпусі), а потім ці категорії наповнюють випадково вибраними текстами.

Демографічна вибірка полягає в соціальній стратифікації респондентів за віком, статтю, соціальним класом та географічним регіоном. За такими параметрами формують репрезентативну вибірку мовців.

"Опортуністичний" метод полягає в тому, що укладачі корпусу просто користуються з наявних на даний момент технологій і доступних мовних матеріалів і не звертають особливої уваги на принципи побудови корпусів. Це особливо показово щодо корпусів, укладених до настання ери електронних публікацій.

Метод цілеспрямованого доповнення. Мається на увазі використання спеціальних методів здобуття мовного матеріалу від мовців, наприклад, психолінгвістичних методик, які допомагають встановили наявні в репертуарі носіїв мови одиниці, що навряд чи трапляться в корпусі [Фрэнсис 1983].

Доповнення спеціалізованими корпусами. Корпуси загального призначення не завжди достатньо повно охоплюють певні типи текстів, які необхідно скрупульозно досліджувати. Тому укладають спеціалізовані корпуси – вужчі, але водночас репрезентативніші щодо обраної ділянки мови.

Критерій культурної значущості тексту. Творці корпусів застосовують оцінні критерії, щоб розмежувати тексти за важ-

ливістю. У сучасних умовах функціонування української мови (зокрема засилля суржику) такий відбір вкрай потрібен, інакше корпус ризикує перетворитися на смітник. Цей критерій може бути застосовано й до цілих жанрів, наприклад у випадку встановлення бажаного відсотку художньої літератури чи перекладів.

Критерій чистоти вибірки. Проблеми чистоти вибірки австралійського й новозеландського варіантів англійської мови описує Ґ. Кеннеді [Kennedy 1998 : 64–66]. Не менш актуальними вони є і для укладачів українських корпусів, особливо усних, – ідеться про розрізнення явища двомовності й суржику, або ж засміченої, неохайної мови. У випадку писемних творів ідеться, зокрема, про відсіювання неграмотних текстів та текстів із явними ознаками машинного перекладу, переважно з російської мови.

Насамкінець завважмо, що логістичні, фінансові, організаційні, часові та інші обставини накладають обмеження на формування корпусу й змушують його творців до компромісу між бажаним і здійсненним.

Розгляньмо коротко параметризацію **Корпусу сучасної американської англійської мови** (Corpus of Contemporary American English, COCA), що його уклав М. Дейвіс [Davies 2010], – чи не єдиний справді моніторинговий корпус англійської мови. Пропорційність дотримано щодо жанрів, підкатегорій текстів (наприклад, "медицина" в наукових публікаціях) і в часі! Корпус містить у рівній пропорції (по 20 %) п’ять жанрів – спонтанне усне мовлення, художня література, популярні журнали, газети й наукові журнали. Щороку автор корпусу додає приблизно 4 мільйони слововживань до кожного жанру, й наразі на кожен із них припадає по 80 з лишком мільйонів слововживань, що разом становить 450 млн. Подвійна збалансованість (синхронійна між жанрами й піджанрами та діахронійна в межах жанру та піджанру) дає змогу здійснювати статистично надійні порівняльні дослідження. Корпус вдало поєднує параметризацію за зовнішніми критеріями, пропорційність, динамічний підхід, уможливлює порівняння між кількома корпусами, втілює загальну тенденцію до репрезентативності й збалансованості, що дає йому незаперечну перевагу над менш систематично організованими великими корпусами.

Ми окреслили, звісно, лише деякі аспекти складної проблеми. Окремі підходи до параметризації потребують глибшого вивчення, і жоден не є панацеєю. Проте зрозуміло, що створення якомога репрезентативніших і збалансованіших корпусів безпосередньо залежить від продуманої, виваженої їх параметризації, а саме оптимального поєднання різних підходів, методів і критеріїв із врахуванням специфіки конкретного корпусу. Параметризація корпусу залишається частково суб'єктивною, а тому бажано, щоб рішення, покладені в її основу, були прорефлексовані, чітко проартикульовані й спиралися на консенсусну оцінку фахівців.

1. *Демська-Кульчицька О.* Основи національного корпусу української мови: монографія / О. Демська-Кульчицька. – К., 2005. 2. *Карпіловська Є. А.* Вступ до комп'ютерної лінгвістики / Є. А. Карпіловська. – Донецьк, 2003. 3. *Фрэнсис У. Н.* Проблемы формирования и машинного представления большого корпуса текстов / У. Н. Фрэнсис // Новое в зарубежной лингвистике. – М., 1983. – Вып. XIV. Проблемы и методы лексикографии. – С. 334–352. 4. *Biber D.* A typology of English texts / D. Biber // Linguistics. – 1989. – Vol. 27. – P. 3–43. 5. *Biber D.* Representativeness in corpus design / D. Biber // Literary and Linguistic Computing. – 1993. – 8 (4). – P. 243–257. 6. *Davies M.* The Corpus of Contemporary American English as the first reliable monitor corpus of English / M. Davies // Literary and Linguistic Computing. – 2010. – Vol. 25. – № 4. – P. 447–464. 7. *Kennedy G. D.* An introduction to corpus linguistics / G. D. Kennedy. – London, New York, 1998. 8. *Leech G.* New resources, or just better old ones? / G. Leech // Corpus Linguistics and the Web. – Amsterdam, 2007. – P. 134–149. 9. *Sinclair J.* Corpus, Concordance, Collocation / J. Sinclair. – Oxford, 1991.

Рассмотрены разные методы параметризации корпуса как средство повышения его репрезентативности и сбалансированности. Обосновано, что оптимального эффекта можно достичь благодаря взвешенному соотношению различных подходов к параметризации корпуса с учетом его специфики.

Ключевые слова: корпус, корпусная лингвистика, параметризация, репрезентативность, сбалансированность, выборка.

The article considers various methods of corpus parametrization as ways to enhance its representativity and balance. The authors argue that an optimal effect may be achieved through a carefully calculated combination of various approaches to corpus parametrization in view of the special aspects of its nature.

Keywords: corpus, corpus linguistics, parametrization, representativity, balance, sample.

Стаття надійшла до редакції 20.09.2012