УДК 811.161

Орися Демська
(Київ)

УКРАЇНСЬКА МОВА І ТЕКСТОВИЙ КОРПУС

У статті описано кореляцію між лексичною картотекою і корпусом, загальні підходи до корпусного відображення сучасної української мови й засади подання, зокрема, таких її ідіомів як літературна мова і територіальний діалект. Також сказано про значення праць Г. Шила для корпусного аспекту діалектології

Ключові слова: лексична картотека, корпус, територіальний діалект.

Електронний текстовий *корпус* є еволюційним кроком в організації текстової емпірики для мовознавчих студій, порівняно з своєю попередницею лінгвістичною картотекою, не кажучи вже про той чи той текст як безпосередньо дане для наукового вивчення мови. У чому ж полягає ця еволющйиість?

Лінгвістична картотека зазвичай є лексичною і далі, залежно від того чи того аспекту лексичного матеріалу, зібраного у картотеці, це може бути власне лексична картотека, матеріали якої віддзеркалюють літературну мову, або ж діалектна, фраземна, ономастична того, відповідно, матеріали яких віддзеркалюють діалектну лексику, власні назви, фраземи тощо. Однак, так чи так йдеться передусім і головно про лексичний рівень мови, що й відображено у самій назві - лексична. Літературний, діалектний, ономаетичний, фраземний тощо текстовий матеріал може бути поданим синтетично в одній загальномовній картотеці, або ж окремо і становити, відповідно, діалектну, ономастичну, фраземну тощо картотеки. Основною функцією лексичної картотеки є «практично забезпечувати як повноту добору слів (!) до реєстру різних типів словників, так і наукового опрацювання словникових статей, виконуючи таким чином основне інформативно ілюстративне завдання - давати у розпорядження лексикографа певний довідковий матеріал, повідомляти про наявність слова у мові, кількість його значень та особливості функціонування» [1,307]. Додатковою ж функцією є її здатність відбивати «особливості вживання граматичних форм, синтаксичних конструкцій, правопису тощо» [1, 307]. Хоча таке звуження і основної, і додаткової функцій є аж надто категоричним, оскільки картотека може бути джерелом для наукового вивчення мови не лише у лексикографічному аспекті. Однак, навіть розширивши функції лексичної картотеки годі буде кардинально вийти, по-перше, за межі лексикологічного різня мови, навіть коли йдеться про інші аспекти слова, ніж лексичні; по-друге, за межі лексикографічності; і, по-третє, що й, власне, задає задекларовану еволюційну відмінність *корпусу* від картотеки, за межі «мови як такої», мови як системи, тобто поєднати, за Г.Степановим, зовнішні (функціональні) системи як вияв форм існування мови

та внутрішню структуру «мови як такої» [2, 20]. На відміну від картотеки, текстовий *корпус,* можливо первинно й не ставлячи перед собою такого завдання, на сьогодні здатен не просто певним чином організовувати мовний матеріал для здійснення лінгвістичних студій, а відображати мову як цілісну комунікативну систему в усіх її ідіомах. І, з одного боку, класичний загальномовний текстовий *корпус,* структура якого на сьогодні вже усталена в корпусній лінгвістиці, відображає саме «тіло мови», ідеологічну модель якого для сучасної української мови (і така модель прийнятна для більшості природних мов світу) вдало сформулювала Н.Дзюбишина-Мельник: «тіло сучасної української національної мови твориться двома повними системами, якими є: природна форма її існування - це територіальні діалекти та штучна форма, якою є літературна, стандартна форма національної мови» [3, 27]. А з іншого - і загальномовний, і спеціальні, часткові *корпуси* задають функціональну стратифікацію мови: *корпус* літературної мови, загальномовний *корпус,* діалектний *корпус,* усний, писемний, мішаний *корпус, корпус* дитячої мови, *ідіолект ний* корпус тощо.

На сьогодні типологічна палітра *корпусів,* що відповідає функціональній стратифікації природної мови, надзвичайно різноманітна, особливо для англійської мови. Так, корпусну стратифікацію англійської мови задають *корпуси:* британського, американського, австралійського, новозеландського, південноафриканського тощо варіантів англійської мови; далі історичні страти англійської мови, задані *корпусами* давньоанглійської, середньоанглійеької та сучасної англійської мови (наприклад, *Helsinki Corpus, Corpus of Early English Correspondence Sampler);* а у межах регіональної стратифікації додатково можуть існувати *корпуси* усні та писемні (наприклад, *Wellington Corpus, Wellington Spoken Corpus (New Zealand)*) тощо.

Корпусне відображення будь-якої природної мови найперше торкається форм її побутування, чи екзистенційної специфіки мови - писемної та усної і, відповідно, існування класифікаційного типу *корпусу: усного - писемного мішаного.* Далі корпусне відображення мови, з одного боку, пов'язане з її рівнями: фонологічним, морфологічним, лексичним та синтаксичним, а з іншого, - з функціональною специфікою мови. Рівнева специфіка мови зазвичай есплікується через корпусну анотацію, чи розмітку: введення у полотно корпусного тексту відповідної (фонетичної, морфологічної, лексичної і/або синтаксичної) інформації. Може йтися про фонетично анотований корпус, морфологічно і / або синтаксично анотований. Однак не можемо говорити про лексично анотований корпус, оскільки на практиці немає синтетичної лексичної розмітки. Як правило на лексичному рівні радше можемо говорити про певну анотаційну аналітичність, суть якої полягає в існуванні різних типів розмітки орієнтованої на слово, його значення, етимологію, специфіку функціонування тощо, а також маркування системних відношень у межах лексики: синонімії, антонімії, полісемії, омонімії.

В інший спосіб відбувається корпусна репрезентація функціональної специфіки мови. Фактично йдеться про функціональну стратифікацію мови. Це поняття прийнято трактувати двояко: як, по-перше, членування континууму мови на літературну мову, побутово-розмовну, просторіччя, територіальні та соціальні діалекти, а, по-друге, функціонально-стильове варіювання мови «зумовлене вживанням мови у різни комунікативних сферах (публіцистика, наука, школа, державне керівництво, право, художня література тощо), а також у різних літературних жанрах (мова поезії, прози; стильові особливості лірики, епічної поезії)» [4,3-4], тобто членування континууму мови за функціональними сталями.

На спосіб відображення мови у текстовому *корпусі* також впливає ще й принцип організації корпусних даних. Йдеться про синтетичне відображення рівневої, стратифікаційної та екзистенційної специфіки мови у межах одного загальномовного текстового *корпусу,* або ж про аналітичне відображення, особливо щодо функціональної стратифікації мови, різних сегментів мови у різних часткових *корпусах,* у наслідок чого постають окремі, часто самостійні, *корпуси* літературної мови, діалектні (говіркові), койие, молодіжного сленгу, так звані стилістичні *корпуси:* наукового стилю, публіцистики, конфесійного тощо. Можливий також синтетично-аналітичний підхід, коли у межах загальномовного *корпусу* текстові дані організовані ієрархічно як генеральний *корпус,* що об'єднує систему субкорпусів різних мовних функціональних страт у тім і субкорпус говорів чи навіть говірок, і кожен з цих субкорпусів містить рівневу (фонологічну, морфологічну, синтаксичну тощо) анотацію текстових даних, або ж функціональна стратифікація лежить в основі параметр изації предметної галузі загальномовного текстового *корпусу.* Саме у такий спосіб, через функціональну стратифікацію предметної галузі *корпусу,* зазвичай відображають мову в *корпусах* національного типу.

Заеадничі характеристики *Корпусу сучасної української мови, дослідницький* (орієнтований на широкий спектр лінгвістичних завдань), *загальномовний* (предметну галузь якого становить сучасна українська мова), *фрагментний* (збудований із текстових фрагментів, чи уривків текстів), *мішаний* (передбачає введення до корпусу писемних і усних текстових фрагментів), *динамічний* (передбачає поповнення множини корпусних текстів), *синхронний* (охоплює рівень сучасної української мови з урахуванням територіальної специфіки як у межах України, так і поза Україною на історичних етнічних землях та в діаспорі) й *одномовний* (корпусні тексти є результатом мовної діяльності носіїв сучасної української мови), - накладають вимоги на його *предметну галузь* та *структуру,* оскільки, згідно з теоретичними засадами корпусної лінгвістики щодо побудови текстових корпусів, і *предметна галузь,* і *структура* корпусу безпосередньо Детерміновані типом та основними характеристиками *корпусу.* Таким чином, якщо *Корпус сучасної української мови* - це зібрання електронних текстів, що

репрезентують національну мову на певному етапі її існування (сучасна українська мова) в усьому різноманітті жанрів, стилів, територіальних і соціальних варіантів, то його *предметну галузь* становитиме *сучасна українська загальнонародна мова* у таких формах її існування, як *літературна мова, територіальний діалект* і, частково, *соціальний діалект*. Додатково може постати питання про *койне*.

На думку Н.Дзюбишиної-Мельник, у сучасній українській мові немає *коїте* у *sensu strict,* натомість є «близьке до нього явище сленгу» [3,27], яке вчена дефініює за Л. Ставицькою: «різновид розмовної мови, яка оцінюється суспільством як підкреслено неофіційна («побутова», «фамільярна», «довірлива»)» [5, *40;* див. також: 3, *25].* Повністю погодитися на відмову від *койне* сучасної української мови не можна. Однак можна говорити про те, що *койне* наближене до сленгу є явищем притаманним колишній підросійській Україні, де не було українського міста. Натомість підавстрійській Україні, де маємо український урбанний сегмент, відоме *койне,* оскільки саме чинник міста є детермінативним для появи *койне:* «виникнення койне, - зазначає Л. Масенко, - є типовим явищем мовного життя міст, де відбувається особливо інтенсивна взаємодія різних діалектних стихій, наслідком чого стає нівелювання різких діалектних відмінностей» [6, 55]. Про існування галицького та гилицько-буковинського *койне* говорить і Ю. Шевельов: «Спілкування між освіченими людьми, що розмовляли західноукраїнськими говірками, найжвавіше відбувалося у Львові. Можна припустити, що на 1900 рік в місті витворилося галицьке чи галицько-буковинське койне» [7,27], і слідом за ним Л. Ткач, визнає існування і також акцентує на інтелектуалізмі описуваного *койне:* «у галицько-буковинському' койне поєдналися традиції української освіченої верстви, що формувалася від кінця XVI ст., народно розмовне мовлення у його різноманітних виявах, а також елементи новоутворюваних соціолектів, що відображали професійне розгалуження соціалем української мови» [8, 522]. Однак через відсутність комплексних досліджень *койне* сучасної української мови уведення текстів *койне* до *Корпусу сучасної української мови* є радше перспективним завданням. Тут доцільно почати з часткового *корпусу койне* а вже потім уводити цю страту до загальномовного *корпусу.*

Абсолютно інший стан речей щодо територіальних *діалектів.* Жодна з національних корпусних лінгвістик не ставить під сумнів обов'язковість уведення до загальномовного національного корпусу діалектних текстів, оскільки, як слушно зауважує Н.Дзюбишина-Мельник, «жива природна мова, якими є її територіальні різновиди, залишається невичерпним джерелом для подальшого розвитку мови у будь-якій її формі: чи то штучного страту (ідіома), а власне, стандартної (тобто літературної) мови, чи то природних, але соціально зумовлених різновидів мови, або корпоративних субмов» [3, 25]. Крім того, діалектна лексика, репрезентована діалектними текстами є частиною лексики сучасного усного мовлення, однією з синхронних форм реалізації мовної

системи, і, якщо йдеться про репрезентативність дослідницького *корпусу* загальнонародної мови, то введення діалектного мовлення, діалектних текстів до його структури є не просто бажаним, а однією з важливих передумов досягнення репрезентативності та вичерпності *Корпусу сучасної української мови*. Що також мотивовано такою важливою характеристикою територіальних діалектів, що під ними розуміємо різновиди національної мови, яким властива відносна структурна близькість і які є засобом спілкування людей, об'єднаних спільною територією, елементами матеріальної й духовної культури, традиціями й самосвідомістю, - як, за О.Гердтом, їх функціонування у синхронії та одночасна належність до діахронії [9, 71 - 72]. І, власне, одночасна належність до синхронії та діахронії, крім інших, є найпершим аргументом на користь уведення діалектного матеріалу до *Корпусу сучасної української мови*. Крім того, будь-який сучасний діалект значно старший за будь-яку сучасну літературну мову, і територіальні діалекти охоплюють як найдавнішу питому мовну специфіку, так і результати інноваційних процесів, потенційних для літературної мови, що дає змогу розширити коло дослідницьких завдань *корпусу,* не кажучи вже про те, що для сучасної української мови діалект є основою її творення та джерелом поповнення. Відповідно, корпусні студії літературної мови без аналізу діалектного матеріалу, без зіставлення з діалектним матеріалом завжди залишатимуться неповними, а наявність діалектного матеріалу в *Корпусі сучасної української мови* власне й уможливлює дослідження літературної мови у зіставленні з діалектними даними і, що важливо, такого типу дослідження реальні у межах однієї корпусної побудови, а не на розрізнених картотечних ресурсах чи друкованих текстах літературної мови і діалектних записах різної якості. Але, пам'ятаючи, що таке дослідження можливе за умови обов'язкового уведення до *корпусу* текстів *літературної мови* - основної наддіалектної форми існування природної мови, ознаками якої є опрацьованість, укормованість, поліфункціональність, стильова диференційованість, фіксованість. Найважливішою характеристикою *літературної мови,* яка перетворює її на центральний сегмент *Корпусу сучасної української мови,* є її співвіднесеність (унаслідок розвиненої диференційованості) з усіма сферами людської діяльності, що, відповідно, забезпечує всі основні типи суспільної інформації. І лише за наявності текстів літературної мови у *корпусі* можна досягнути репрезентативності та потенційно вичерпності загальномовного дослідницького *корпусу,* призначеного для широкого кола лінгвістичних студій над сучасною мовою.

Отже, українська мова у корпусному відображенні, це подання, поряд з літературною мовою, її основного джерела - говірки, говору, і годі сказати краще про цей текстовий пласт у *корпусі,* ніж сказав про російські говіркові тексти у *корпусі* О.Гердт: «це наше культурно-історичне національне надбання, Це фонди, споріднені зі скарбами Ермітажу, Музею антропології, рукописним

відділом бібліотеки ім. В.Леніна або публічної бібліотеки ім. М.Салтикова-Щедріна»[9, 71 -72].

Очевидно, що дбати про збереження мовних скарбів має не одне покоління. Але, якщо йдеться про покоління Г.Шила, то власне завдяки Гаврилові Федоровичу ми можемо сьогодні обговорювати на рівні корпусної лінгвістики аспекти подання говіркового тексту в корпусі української мови, бо саме Гаврило Федорович залишив нам і великий фактичний діалектних матеріал, і теоретичне його осмислення, без чого неможливо забезпечити еволюційних перехід від класичного до корпусного мовознавства. І особливо важливими з цього огляду є, зокрема, такі праці вченого: «Проблеми класифікації наддністрянських і західнополіських говорів» (1971), «Південно-західні говори УРСР на північ від Дністра» (1957), «Із лексики українських говорів» (1960), що й сьогодні не втратили актуальності.

Література

1. Бурячок А. А. Лексична картотека // Енциклопедія Українська мова. - К.: Видавництво «Українська енциклопедія» ім. М.П.Бажана, 2004. - С. 307.

2. Степанов Г. В. Типология языковых состояний и ситуаций в странах романской речи. - М.: Наука, 1976. - 224 с.

3. Дзюбишина-Мельник Н. Тіло національної мови // Магістеріум: Мовознавчі студії. - Вип. 37. - К., 2009. - С. 24 - 27.

4. Глухман М. М. Введение // В: Функциональная стратификация языка. М.: Наука, 1985.-С. 3-8.

5. Ставицька Л. Арго, жаргон, сленґ. - К.: Критика, 2005. - 462 с.

6. Масенко Л. Нариси з соціолінгвістики. - К.: Видавничий дім «КМ Академія», 2010. - 242 с.

7. Шевельов Ю. Українська мова в першій половині двадцятого століття (1900 - 1941): стан і статус // Сучасність. - 1987. - 294 с.

8. Ткач Л. Українська літературна мова на Буковині в кінці XIX - на початку XX ст. - Частина 2: Джерела і соціокультурні чинники розвитку. - Чернівці: Книги - XXI, 2007. - 704 с.

9. Герд А. С. Типы русских текстов и организация Машинного фонда русского языка // Машинный фонд русского языка: идеи и суждения. - М.: Наука, 1986.-С. 67-75.

The article considers: the correlation between the lexical file and the text corpus; general approach to the corpus representation the modern Ukrainian in their variants - literature language ant dialects. Some words in the article are dedicated to professor H.Shylo and his scientific works.

Key words: lexical file, corpus, dialect.