ЛЕКСИКОГРАФІЯ, ЛЕКСИКОЛОГІЯ: ТЕОРІЯ ТА ПРАКТИКА

Орися Демська-Кульчицька, к. ф. н.[*]
Інститут української мови НАН України (Київ)
УДК 161.2.81'374.72'22

СЛОВ'ЯНСЬКІ КОРПУСНІ СТУДІЇ

У статті зроблено огляд розвитку корпусних досліджень у слов'янських мовознавчих традиціях. Проаналізовано корпуси слов'янських мов, їхні типологічні характеристики, сферу застосування та коло лінгвістичних завдань, які виконують на їх основі.

Стрімке входження корпусної лінгвістики в національні мовознавчі традиції не оминуло й славістику. Так, у слов'янських мовах (особливо за останні п'ять років) з'явилася низка загальномовних корпусів як національного, так і інших типів. Це стимулювало розвиток напряму корпусної лінгвістики в межах національних мовознавчих традицій Славії. І сьогодні вже можемо говорити про чеську, сербську, болгарську, польську, словенську, словацьку, російську корпусні лінгвістики. Кожна з цих корпусних лінгвославістик репрезентована передусім корпусами національного типу. На особливу увагу заслуговують такі слов'янські корпуси: *Český Národní Korpus, Slovenský Národý Korpus*, *Beseda, Nova Beseda, Korpus IPI PAN, Korpus PWN, FIDA: Korpus Slovenskega Jezika, Национальный корпус русского языка, Большой корпус русского языка* та ін.

Корпусні студії в славістиці не можна розглядати узагальнено, оскільки кожна з слов'янських корпусних лінгвістик перебуває на іншому етапі, має різний практичний досвід та історію розвитку. Так, створення *Корпусу сербської мови Дж. Костіча* розпочато в 1957 р., але впродовж 1962–1996 рр. робота над проектом фактично не проводилася. 1994 р. датовано початок досліджень і роботи над *Чеським національним корпусом.* На 1997 р. припадає початок укладання

Корпусу словенської мови FIDA. Пізнішими є, зокрема, такі слов'янські корпусні проекти: *Корпус текстів Інституту словенської мови ім. Франа Рамовша – БЕСЕДА*, датований 2000 р., і *Нова Беседа* – 2005 р., *Словацький національний корпус* – 2001 р., *Національний корпус російської мови* – 2003 р.

Незважаючи на різний рівень розбудованості корпусної лінгвістики в слов'янських мовознавчих традиціях, на сьогодні є підстави констатувати факт існування корпусних підходів до осмислення національних слов'янських мов, що передусім відображено в розробленій теорії побудови морфологічно анотованих корпусів слов'янських мов національного статусу. Теоретичні основи побудови корпусів слов'янських мов ґрунтуються на загальній теорії корпусної лінгвістики, міжнародних стандартах кодування корпусу, а також, коли йдеться про морфолого-синтаксичну анотацію, на результатах власних досліджень в галузі теоретичної, функціональної та формальної граматик. Це уможливлює зіставляти корпуси слов'янських мов та паралельно досліджувати ці мови із використанням корпусних методик. Зіставність текстових корпусів слов'янських мов – одна з важливих ознак, на підставі якої власне можна говорити про певну єдність корпусної лінгвославістики.

З погляду типологічних специфік слов'янські корпуси національного статусу детерміновані як *одномовні*, або *мономовні*, та *дослідницькі.* Двомовним – уведено тексти словенською й англійською мовами – є лише *Корпус текстів Інституту словенської мови ім. Франа Рамовша – БЕСЕДА. З*агалом, якщо говорити про типологію корпусів, то єдина відмінність між слов'янськими корпусами полягає в хронологічному параметрі, тобто йдеться про часові рамки текстів, залучених до корпусу – тексти сучасної мови чи історичні. Так, *Чеський національний корпус, Національний корпус російської мови, Корпус сербської мови Дж. Костіча* є діахронно-синхронними. У *Ческому національному корпусі* діахронна частина подана як система підкорпусів: ČNKDIA archive (*діахронний Чеський національний архів*) містить тексти старочеської мови обсягом 2 млн. сл. і фактично є архівом; ČNKDIA bank (*діахронний банк чеської мови*) – зібрання затранскрибованих текстів обсягом 2 млн. сл., транслітерованих текстів на 100 тис. сл., а також діалектних текстів на 200 тис. сл.; Database and dictionaries (DB) (бази даних і словники) – словники старочеської мови; DIAKORP

(*діахронний корпус*) – вибрані тексти старочеської мови від перших текстів і до сучасного періоду.

У *Національному корпусі російської мови* та *Корпусі сербської мови Дж. Костіча* діахронній і синхронній частинам не притаманна така категорична дискретність. Діахронна частина є в обох цих корпусах, хоча й оформлена як субкорпус, але це швидше зроблено для зручності побудови генерального корпусу і роботи з ним. У *Національному корпусі російської мови* виділено підкорпуси текстів сучасної російської літературної мови від початку XIX ст. і до сьогодні та давньоруської мови від XI до XIV ст. Крім того, передбачено введення до корпусу текстів так званих нелітературних форм сучасної російської мови: розмовної, просторічної, діалектної. А в *Корпусі сербської мови Дж. Костіча* – підкорпуси відповідно XII–XVII ст., XVIII–XIX ст., першої половини XIX ст. і сучасної сербської мови.

Словацький національний корпус, Корпус словенської мови, Нова Беседа, Корпус польської мови Національного наукового видавництва, Корпус польської мови Інституту основ інформатики Національної академії наук є синхронними корпусами, які охоплюють лише зріз сучасної мови.

На початках корпусної лінгвістики параметр обсягу корпусу вимагав особливою уваги. Тоді йшлося головним чином про один мільйон слововживань, і це була доволі недосяжна межа. Основною причиною такого обмеження були комп'ютерні ресурси, що, зокрема навіть мотивувало недовіру до корпусних студій Н. Хомського і спричинило певне уповільнення в розвиткові корпусної лінгвістики. На сьогодні вже йдеться про 100 млн. сл., що послідовно засвідчено слов'янськими корпусами. Більшість слов'янських корпусів, особливо ті, робота над якими почалася давніше, орієнтовані на мінімальний обсяг 100 млн. слів. А *Нова Беседа* уже минула цей кількісний поріг і налічує 162 млн. слів.

З погляду структури усі слов'янські корпуси, як і корпусні побудови загалом, можуть мати або моноструктуру, або поліструктуру, тобто творити єдиний корпус без виділення субкорпусів (такими є *Нова Беседа, Великий корпус російської мови, Корпус польської мови Інституту основ інформатики Національної академії наук*), чи структурованими й у межах генерального корпусу виділяти субкорпуси, які далі можуть бути ієрархічно, або лінійно корельовані. Субкорпусну

структуру мають, наприклад, уже згадувані *Чеський національний корпус, Національний корпус російської мови*.

Якщо на рівні типології, обсягу та структури більшість слов'янських корпусів є зіставними, то релятивно зіставними вони будуть на рівні джерельної бази. Джерельна база детермінована, по-перше, національною лінгвістичною традицією, яка, як правило, формувалася в процесі укладання лексичних картотек для загальномовних академічних словників національних мов, по-друге, періодизацією розвитку кожної конкретної слов'янської мови і, по-третє, колом тих завдань, які плановано виконувати на корпусі. Так, до синхронної частини *Словацького національного корпусу* загалом увійшли тексти за період від 1955 р. до 2005 р. з розрізненням текстів 1990 – 2003 рр. та 2003 – 2006 рр.; до *Корпусу словенської мови* – фрагменти текстів писемного варіанту та зафіксовані на письмі уривки усного мовлення сучасної словенської мови від другої половини XX ст. з акцентом на час від 1990 рр. до сьогодні; до *Корпусу сербської мови Дж. Костіча* – тексти авторських творів відповідно XII – XVII ст., XVIII – XIX ст., тексти завершених творів В. Караджича, тексти першої половини XIX ст. і синхронна частина, яка охоплює сучасну сербську мову, репрезентовану 1) новелами й есе; 2) поезією; 3) періодикою; 4) науковою прозою, 5) політичною прозою; 6) текстами белградських сюрреалістів.

Корпус польської мови Національного наукового видавництва охоплює белетристику, науково-популярну літературу, періодику, тексти усного мовлення й рекламні тексти; *Корпус текстів Інституту словенської мови ім. Франа Рамовша – БЕСЕДА* складається з чотирьох частин: 1) художня література; 2) переклади на словенську мову художніх творів; 3) щоденна газета Delo; 4) художня література англійською мовою, де частини A і B репрезентують 112 прозових творів, з яких 98 – оригінальні твори й 14 перекладних (з 1858 р до 1996 р.); частину D – електронні видання словенської щоденної газети Delo за період від січня 1998 р. до червня 2003 р.; частину E – зібрання творів В. Скотта, О. Вальда і М. Твена.

Важливою характеристикою сучасних текстових корпусів є їхня анотованість. Головним чином ідеться про морфологічну анотацію, але, наприклад, чеська й болгарська корпусні лінгвістики диспонують уже корпусним ресурсом, анотованим синтаксично. Слов'янські корпуси переважно є морфологічно анотованими, що логічно стимулювало наявність теоретичних студій із корпусних анотації. Так, окремими

дослідженнями в межах корпусного проекту є розроблення так званого морфологічного стандарту для *Національного корпусу російської мови.* Морфологічний стандарт обумовлює подання в корпусі інформації про морфологічні форми й значення, тобто частину мови, рід, відміну, відмінок, особу тощо. Рішення, прийняті в корпусі, головно спираються на морфологічну модель, подану в *Граматичному словнику російської мови* А. Залізняка [1].

Проте специфіка корпусу як універсального засобу вивчення мови диктує певні особливості рішень і, власне, саме цією специфікою мотивовані всі відхилення від моделі *Граматичного словника* в нашому стандарті. Загалом структура морфологічної інформації в *Національному корпусі російської мови* передбачає маркування:

- лексеми, до якої належить словоформа: „словниковий запис" цієї лексеми і її частиномовна характеристика;
- множини граматичних ознак цієї лексеми або словокласифікативних характеристик, наприклад: рід для іменника, перехідність для дієслова;
- множини граматичних ознак цієї словоформи або словозмінні характеристики, наприклад: відмінок для іменника, число для дієслова;
- інформацію про нестандартність граматичної форми, орфографічне спотворення тощо.

Важливим рішенням щодо оформлення засобів метамови граматичних міток є те, що в її основу, з огляду на широку міжнародну аудиторію користувачів *Національного корпусу російської мови*, покладено систему скорочених міток, або тегів, сформованих на основі латинського алфавіту.

Окремим завданням корпусної лінгвістики є лінгвістичне обґрунтування та створення програмного корпусного продукту, або програми роботи з корпусом. Існує багато цікавих розроблень у цій галузі, й однією з таких програм є мовнонезалежна універсальна програма *Poligrap*, створена в межах проекту *Корпусу польської мови Інституту основ інформатики Національної академії наук.* Цей програмний продукт можна застосовувати до корпусів довільних мов, у тім числі й української. Універсальність програми *Poligrap* досягнута низкою технологічних рішень, а саме: „використовуваний у корпусі тегсет не вбудовано в програму, його лише задано через зовнішній конфігураційний файл, натомість внутрішній, використовуваний

формат кодування символів – це універсальний формат UTF–8. Тому ніщо не перешкоджає використанню програми *Poligrap* для роботи з іншими корпусами, в тому числі й інших мов" [2, с. 41]. На сьогодні *Poligrap* існує в трьох варіантах: 1) інтернетному (дозволяє працювати з *Корпусом Інституту основ інформатики ПАН* в режимі on-line); 2) графічному, призначеному для операційних систем Windows 2000, Windows XP та GNU/Linux; 3) текстовому, призначеному для системи GNU/Linux. Усі три варіанти оперують доволі багатим синтаксисом запиту, який дозволяє сформулювати запит про текстові уривки, парадигматичні форми й морфолого-синтаксичні параметри слів. Крім того, пошукова програма *Poligrap* диспонує трьома варіантами інтерфейсу, а саме: найпростішим інтернетваріантом, графічним і текстовим під Linux, кожен з яких має власну специфіку на рівні нюансів користування. Базовим варіантом інтерфейсу аналізованої пошукової програми визначено графічний варіант, важливим атрибутом якого є можливість візуалізації парадигматичних форм слів та їхніх лем і морфолого-синтаксичних тегів. Така інформація вкрай важлива для граматичних, лексичних і особливо лексикографічних досліджень мови.

Хоча на сьогодні маємо різний рівень розбудови корпусної лінгвістики у слов'янських мовознавчих традиціях, теоретичні основи корпусної лінгвославістики загалом і побудови корпусів слов'янських мов зокрема базують на загальній теорії корпусної лінгвістики, міжнародних стандартах кодування первинних даних корпусу, а також, коли йдеться про морфолого-синтаксичну анотацію, на результатах власних досліджень у галузі теоретичної, функціональної та формальної граматик. Відштовхуючись від англо-саксонської корпусної лінгвістики, кожен зі слов'янських корпусних напрямів пішов власним шляхом, базованим на національних лінгвістичних традиціях, що певним чином вплинуло на виникнення ідеї впровадження корпусної лінгвістики в українську мовознавчу парадигму й сприяло розробленню теорії корпусної лінгвоукраїністики.

Література

1. Зализняк А. А. Грамматический словарь русского языка. – М.: Наука, 1987.
2. Przepiórkowski A. Korpus IPI PAN: wersja wstępna. – Warszawa: Instytut podstaw informatyki PAN, 2004.