ПРИКЛАДНА (КОМП'ЮТЕРНА) ЛІНГВІСТИКА

КОРПУСНА ЛІНГВІСТИКА

О. Демська-Кульчицька, к. філол. н.[*]
Інститут української мови НАНУ, (Київ)
УДК 161.2.81'374.72'22.322

НАЦІОНАЛЬНИЙ КОРПУС УКРАЇНСЬКОЇ МОВИ: КОНЦЕПТУАЛЬНИЙ АСПЕКТ

У статті висвітлено засади формування та формалізації матеріалів Національного корпусу української: обґрунтовано принципи добору текстів, визначено кількість та обсяг текстових фрагментів, подано адаптовану систему кодування первинних даних.

Сучасні мовознавчі дослідження сьогодні активно послуговуються корпусним методом і для усіх європейських мов, серед яких не забуваємо про слов'янські, існує по кілька корпусів текстів різного типу, обсягу, структури, наповнення та призначення. И узагальнивши понад сорокалітню історію становлення корпусної лінгвістики та побудови корпусів, мовознавці визнали, що створення корпусу є „обов'язком щодо рідної мови" [4: 5].

Створення корпусів на сьогодні виформувало в межах загального напряму корпусної лінгвістки певний субнапрям, а саме *теорію і практику побудови текстових корпусів. В*иокремлення такого типу досліджень передовсім умотивовано тим, що процедура побудови текстового корпусу, в термінах корпусної лінгвістики, до моменту його експлуатації як лінгвістичного об'єкта, а також використання як програмістського ресурсу, вимагає розв'язання низки наукових проблем, зокрема типології корпусу, його призначення, обсягу, параметризації предметної галузі, репрезентативності, структурування та принципів відбору базових одиниць зберігання etc, тобто йдеться про процедуру концептуалізації корпусу, в нашому випадку – *Національного корпусу української мови (НКУМ)*.

Національний корпус української мови: теоретико-практичний аспект є науково-дослідною темою Інституту української мови НАН України, затвердженою Вченою радою ІУМ НАНУ від 26.01.06 протокол № 1, Бюро відділення ЛММ НАНУ від 02.03.06 протокол № 2. Це ставить ряд вимог до концепції *НКУМ. Н*асамперед ідеться про ідеологію проекту, яка чітко окреслює некомерційний, науковий статус корпусу, призначеного для здійснення наукових досліджень різних рівнів сучасної української мови в Інституті української мови НАН України. Здійснення цих досліджень сучасної української мови накладає особливі вимоги щодо параметризації предметної галузі корпусу, яка повинна базуватися на традиційній параметризації українських текстів у лінгвоукраїністиці, та щодо необхідності його морфологічної розмітки, яка також за основу має брати кваліфікації слів, вироблені в межах української граматичної традиції. Тобто науковий проект *НКУМ* повністю ґрунтується на досягненнях української мовознавчої науки.

Національний корпус – це зібрання текстів, що репрезентують національну мову на певному етапі(ах) її існування в усьому різноманітті жанрів, стилів, історичних, територіальних і соціальних варіантів. Електронний текстовий корпус додатково передбачає машиночитану форму існування текстового ресурсу, стандартне подання й можливість застосування у дослідженнях мови та в програмуванні. Виходячи зі сказаного, визначаємо **Національний корпус української мови (НКУМ)** як *перетворену на електронну форму, організовану згідно з корпусними стандартами, програмно опрацьовану, репрезентативну вибірку текстів сучасної української мови, призначена для лінгвістичного застосування. Л*інгвістична орієнтованість передбачає здійснення академічних досліджень а) української мови, зокрема лексики, фразеології,

термінології, морфології та синтаксису, словотвору, орфографії тощо; б) методики викладання / навчання / вивчення української мови як рідної та іноземної.

З огляду на призначення, типологію й коло завдань *НКУМ*, описуваний корпус повинен бути корпусом:

- **загальнонародної мови:** з урахуванням територіальної специфіки як у межах України, так і поза ними;
- **одномовним:** тексти, що ввійшли до корпусу, є результатом мовної діяльності носіїв української мови;
- **дослідницьким:** зорієнтованим на широкий спектр лінгвістичних завдань;
- **динамічним:** тобто передбачати постійне поповнення множини корпусних текстів;
- **синхронним:** має охоплювати рівень сучасної української мови;
- **фрагментним:** збудованим із текстових фрагментів, тобто уривків текстів, відібраних за попередньо визначеними засадами відбору текстових даних до корпусу;
- **мішаним:** передбачати введення до корпусу писемних і усних текстових фрагментів;
- **морфологічно розміченим:** усі текстові дані розмічені до рівня слова, і кожне слово передбачає маркування частиномовної належності та відповідних морфологічних форм.

Концепція довільного корпусного проекту вимагає параметризації предметної галузі, репрезентованої корпусом. Оскільки національний корпус є зібранням електронних текстів, які репрезентують національну мову на певному етапі(ах) її існування в усьому різноманітті жанрів, стилів, територіальних і соціальних варіантів, то предметною галуззю *НКУМ* буде українська загальнонародна мова в таких формах її існування, як літературна мова, територіальний діалект і соціальний діалект.

Найповніше в *НКУМ* повинна бути представлена літературна мова – основна наддіалектна форма існування природної мови, ознаками якої є опрацьованість, унормованість, поліфункціональність, стилістична здиференційованість, фіксованість. Найважливішим аспектом літературної мови є її співвіднесеність (унаслідок розвиненої здиференційованості) з усіма сферами людської діяльності, що, відповідно, забезпечує всі основні типи суспільної інформації.

Щодо літературної мови, то не підлягає сумніву потреба її відображення в корпусі національного типу. Натомість вимагає додаткової аргументації введення діалектних та соціолектних текстів.

Діалектна лексика, репрезентована діалектними текстами, як зазначає О. Гердт, є частиною лексики сучасного усного мовлення, однією з синхронних форм реалізації мовної системи, і, якщо йдеться про репрезентативність корпусу національного типу, то введення діалектного мовлення до його структури є беззаперечним. Важливою характеристикою територіального діалекту – різновиду національної мови, якому властива відносна структурна близькість і який є засобом спілкування людей, об'єднаних спільною територією, елементами матеріальної й духовної культури, традиціями й самосвідомістю, – це функціонування в синхронії та одночасно належність до діахронії. Належність до синхронії та діахронії власне є основним аргументом на користь уведення діалектного матеріалу до корпусу. Ще одним аргументом на користь цього є те, що будь-який діалект значно старший за будь-яку літературну мову. Крім того, територіальні діалекти охоплюють як найдавнішу питому мовну специфіку, так і результати інноваційних процесів, потенційних для літературної мови. Наявність діалектного матеріалу в *НКУМ* дає змогу досліджувати фактичний матеріал літературної мови в зіставленні з діалектними даними в межах однієї корпусної побудови.

Уведення до *НКУМ* текстів, які б репрезентували соціальні діалекти, пов'язано, з одного боку, з загальною лінгвістичною тенденцією до „термінологічної універсалізації соціально маркованої лексики" [1: 47], на що вказує Л. Ставицька, а з іншого – із завданням оптимальної репрезентації мови в корпусі загальномовного типу. Тобто йдеться про понятійне визначення соціально маркованої лексики і принципи відбору соціолектного текстового матеріалу. До соціальних діалектів, за Л. Ставицькою, зараховуємо арго (особлива мова певної відокремленої професійної чи соціальної групи, яка складається з видозмінених елементів однієї або двох природних мов), жаргон (напіввідкрита лексико-фразеологічна підсистема, яку застосовують та чи та соціальна група з метою відособлення від решти мовної спільноти) і сленг (різновид розмовної

мови, яку суспільство оцінює як підкреслено неофіційну („побутова", „фамільярна", „довірлива").

Очевидно, що в межах *НКУМ* соціолектні тексти не повинні кількісно переважати діалектні тексти а становити лише незначний відсоток від загального обсягу корпусу. Тут також варто зазначити, що соціолектні тексти матимуть переважно звукову форму, тобто входитимуть до усної частини корпусу.

Уведення до *НКУМ* діалектних і соціолектних текстів додатково зумовлює й літературна традиція вплетення в полотно художнього твору, відповідно, діалектних і соціолектних елементів для естетичної та культурно-змістової виразності твору. Крім того, наявність у корпусі цих форм загальнонародної мови є одним із шляхів досягнення репрезентативності *НКУМ.*

За аналогією до *Британського національного корпусу*, який розглядаємо як еталонний у сучасній корпусній лінгвістиці, вважаємо за доцільне, особливо на етапі впровадження напряму корпусної лінгвістики в українську мовознавчу традицію, визначити хронологічні межі *НКУМ* межами сучасної української літературної мови. Тобто від останніх років XVIII ст. і до сьогодні, залишаючи поза увагою всі попередні періоди існування української мови, а саме: давньоукраїнський період від середини XI ст. до кінця XIV ст., ранньосередньоукраїнський – від початку XV ст. до середини XVI ст., середньоукраїнський – від середини XVI ст. до перших років XVIII ст., пізньосередньоукраїнський – від середини і до кінця XVIII ст.

Змістова параметризація предметної галузі *НКУМ* передбачає визначення джерельної текстової бази корпусу з подальшою стилістично-жанровою стратифікацією відібраних текстів. Враховуючи традицію відбору та організації фактичного матеріалу для лінгвістичних досліджень у лінгвоукраїністиці, до *НКУМ* повинні передусім увійти літературно опрацьовані масові друковані тексти, які визначає багатство лексики й нормативність слововживання. Важливою ознакою таких текстів є відображення в них не лише писемного варіанта мови, а й усного. До основних вимог текстових джерел належать: а) адекватне представлення сучасної української мови в усіх найважливіших сферах функціонування; б) охоплення всіх жанрів, у яких реалізуються функціональні стилі мови; в) орієнтація на масового читача; г) достатність щодо кількісних параметрів. Остання вимога є надзвичайно важливою, оскільки вона, з одного боку, впливає на встановлення реального лексичного складу мови, а з іншого – на забезпечення кожного мовного факту достатньою кількістю контекстів. Таким чином, опрацюванню підлягатимуть: художня проза, драматургія, поезія, публіцистика, наукова, науково-популярна, суспільно-політична література, офіційні документи, підручники, посібники й листи.

Стилістично-жанровий аспект текстового ресурсу вимагає додаткового опису. Так, стильова диференціація мови можлива, по-перше, за характером мовної експресії, а по-друге, за характером суспільної функції мови. За характером мовної експресії традиційно розрізняють високий, середній і низький стилі. В основі такої тристильової диференціації лежить концепція залежності між предметом викладу, тематикою і добором мовних засобів та жанрів. Цей поділ, успадкований європейською традицією доби Відродження й бароко з александрійської філософської школи античного періоду, застосовано до параметризації предметної галузі в багатьох національних корпусах європейських мов.

В українській традиції тристильова диференціація мови, яку, зокрема, розробляли Ф. Прокопович, М. Довгалевський, Г. Кониський, втратила актуальність у процесі історичного розвитку стилістичної системи української мови, коли три традиційні стилі ('слоги') занепали, а основними одиницями стильової диференціації української мови стали функціональні стилі. Таким чином, в українську лінгвістичну традицію увійшов функціональний принцип стилістичної диференціації мови, і така диференціація покладена в основу змістової параметризації предметної галузі *НКУМ*.

Сучасна українська мова на рівні стильової диференціації вкладається в семиелементну систему: 1) художній, 2) науковий, 3) офіційно-діловий, 4) публіцистичний, 5) релігійний, 6) розмовний та 7) епістолярний стилі. Крім того, виділяємо усний і писемний варіанти мови. Художній, публіцистичний, науковий та офіційно-діловий функціональні стилі в межах *НКУМ* репрезентовані головно писемними текстами, а розмовний – усними.

Для багатьох галузей мовної практики писемна форма національної мови є пріоритетною, вона зазвичай реалізує себе в художньому, публіцистичному, науковому й офіційно-діловому стилях. Тому для *НКУМ* приймаємо, що фактичним матеріалом

писемного варіанта української мови будуть тексти таких функціональних стилів сучасної української мови:

(1) художнього;
(2) публіцистичного;
(3) наукового;
(4) офіційно-ділового;
(5) релігійного;
(6) епістолярного.

Художній стиль у *НКУМ* репрезентуватимуть прозові, поетичні й драматургічні тексти відповідних хронологічних меж (від останніх років XVIII ст. і до сьогодні) з такою жанровою диференціацією:

– проза: роман, повість, оповідання, новела;
– поезія: вірш, сонет, поема, балада, сатира;
– драматургія: драма, комедія, трагедія.

Публіцистичний стиль у *НКУМ* головно репрезентуватиме періодика і частково рекламні тексти. Тематично-функціональний спектр періодичних видань надзвичайно широкий і покриває всю стильову диференціацію мови, тому періодичні видання, незалежно від їхнього типу – газета, журнал, бюлетень, тижневик, місячник, квартальник тощо, – вкладемо в прийняту нами шестиелементну стильову систему і відповідно структуруватимемо фактичний матеріал. Крім текстів періодики, до *НКУМ* буде уведено фрагменти публіцистичних творів Івана Франка, Лесі Українки, Бориса Грінченка, Ліни Костенко, Юрія Мушкетика та ін.

Науковий стиль у *НКУМ* поділено на підстилі: власне науковий, науково-популярний і науково-методичний та здиференційовано за галузями наук – гуманітарні, природничі й технічні.

Офіційно-діловий стиль у *НКУМ* доцільно представляти як усним текстовим матеріалом, так і писемним, що засвідчуватиме його існування в усному й писемному варіантах української мови. Офіційно-діловий стиль репрезентуватимуть:

– законодавчо-правові державні документи, наприклад, Акт проголошення незалежності України, Конституція України, Постанови Верховної Ради, Укази Президента;
– організаційно-службові документи: протокол, наказ, розпорядження, ухвала, заява;
– суспільна документація, наприклад, колективний договір, статут державних або недержавних організацій тощо.

Уведення до *НКУМ* текстів релігійного стилю, сформованого на книжній і народнорозмовній основі, умотивоване, по-перше, розширенням сфери його побутування в сучасному суспільстві, по-друге, його статусом в історії розвитку української мови, і, по-третє, потребою доповнення лексичного складу мови цим пластом лексики. Джерелами фактичного матеріалу релігійного стилю в *НКУМ* будуть Біблія, богослужбові книги, богословська й духовна література.

Епістолярний стиль, який сьогодні існує в паперовій і електронній формі, відповідно, зумовлює виділення в *НКУМ* класичних епістолярних текстів, чи листів, і електронної кореспонденції. Крім того, класичні епістолярні тексти додатково поділено на: а) офіційно-ділове внутрішньодержавне листування, б) офіційне міждержавне листування і в) приватне листування.

Усну частину в *НКУМ* репрезентуватимуть лише усні тексти розмовного стилю, а саме: а) публічні виступи, тексти парламентських дебатів; б) інтерв'ю й коментарі; в) ділове мовлення; г) повідомлення в транспорті; ґ) аудіореклама; д) побутове діалогічне й монологічне спілкування; е) діалектне мовлення; д) соціолектне мовлення.

Отже, до *НКУМ* увійдуть тексти:

I. Писемного варіанта української мови.
   1. Художнього стилю:
   2. Наукового стилю (власне наукового, науково-популярного та науково-методичного підстилів):
   3. Офіційно-ділового стилю.
   4. Публіцистичного стилю.
   5. Релігійного стилю.
   6. Епістолярного стилю.

II. Усного варіанта української мови.

1. Розмовного стилю.

Щодо стратегій обсягу, то на початковому етапі *НКУМ* доцільно будувати як середній, визначивши нижню межу в **1 млн слів,** передбачивши можливості перманентного поповнення корпусу текстовими даними до **100 млн слів** і не ставлячи обмежень на верхню межу кількості слів у корпусі.

Концептуалізація процедури створення корпусу обов'язково ставить вимогу на окреслення кола дослідницьких завдань, здійснюваних на копусі. І на початковому етапі експлуатації *НКУМ* дослідницькі завдання доцільно згрупувати за: а) мовними рівнями й б) прикладними аспектами.

Щодо першої групи *НКУМ* передусім доцільно використати для морфологічних, синтаксичних та лексичних досліджень. Так, морфологічні дослідження можуть стосуватися, зокрема, вивчення специфіки реалізації морфологічних категорій і значень, типології флективних змін у парадигматичних формах іменника та прикметника, реальної іменної та дієслівної системності, прийменникової та безприйменникової сполучуваності слів у мовленні тощо. Синтаксичні – вивчення типології простих і складних речень сучасної української мови, довжини речень (мінімальної, максимальної, типової), функціональної специфіки активних і пасивних конструкцій, засад упорядкування слів у реченнях сучасної української мови тощо. Лексикологічні – вивчення динаміки лексичного складу української мови, аспектів неологізації vs архаїзації українського лексикону, механізмів поповнення лексичного ядра та периферії на сучасному етапі функціонування мови, шляхів збагачення лексичного складу мови. Крім наведених вище, не варто забувати також про можливість корпусного дослідження фонетики й словотвору.

Прикладний аспект є чи не найважливіщим застосуванням корпусів, оскільки опосередковано викликав і появу самих корпусів, і формування корпусної лінгвістики. Тут головно йдеться про лексикографію, а ще й, крім того, корпусну лексикографію та методику мови. У лексикографічних дослідженнях *НКУМ* можемо використовувати для укладання, наприклад, нового тлумачного словника української мови, шкільного тлумачного словника сучасної української мови, словника динаміки української лексики в XX ст., словника неологізмів, словника архаїзмів, словника дієслівних керувань, не кажучи вже про різні словники так званого реєстрового типу й частотні словники.

Щодо методики української мови як рідної та іноземної, то *НКУМ*, з одного боку, може слугувати базою для написання підручників і посібників, бути частиною інтерактивних методичних ресурсів, а з іншого боку – корпусні дані та можливість швидкого різнорівневого пошуку дадуть змогу підібрати дидактичний матеріал для практичного вивчення мови; а „здатність викликати комбінації слів, а не індивідуальні слова" [3] уможливлює працю з реальними контекстами слів, за допомогою яких простіше зрозуміти правила вживання різних лексичних одиниць і, відповідно, опановувати мову.

Створення будь-якого текстового корпусу, як ще у 60-х роках минулого століття зазначав У. Френсис, вимагає детального плану чи, висловлюючись лексиконом початку XXI ст. – концепції, а якщо „планування зроблено правильно, то методика відбору та формування корпусу стає абсолютно зрозумілою, а практичне його створення виявляється цілком формальною процедурою" [2: 348].

Література

1. Ставицька Л. Арго, жаргон, сленг. – К.: Критика, 2005. – 462 с.
2. Френсис У. Н. Проблемы формирования и машинного представления большого корпуса текстов // Новое в лингвистике. – 1983. – Вып. XIV. – С. 334–352.
3. McEnery T., Wilson A. Corpus Linguistics. – 1996. – http://www.comp.lancs.ac.uk.
4. Przepiórkowski A. Korpus IPI PAN: wersja wstępna. – Warszawa: Instytut podstaw informatyki PAN, 2004. – 89 st.