Презентації

ЩО ТАКЕ КОРПУСНО-БАЗОВАНІ ДОСЛІДЖЕННЯ МОВИ

Орися ДЕМСЬКА-КУЛЬЧИЦЬКА, кандидат філологічних наук, м. Київ

Наприкінці минулого століття в мовознавчих дослідних методиках сформувався новий підхід до відбору та організації мовного матеріалу для подальшого його вивчення, опису, аналізу. Його назва — корпусний, що походить від назви об'єкту, в який організовано фактичний матеріал — ***корпус текстів.***

Говорячи про корпус, маємо на увазі електронний вигляд системно організованої та програмно обробленої вибірки текстів, що представляють всі історичні й географічні варіанти та форми існування довільної мови. Він призначений для мовознавчих досліджень і програмних застосувань.

Появу цього дослідного об'єкта спричинили два чинники: складність та тривалість ручного відбору фактичного матеріалу та можливість спростити цей процес, застосувавши комп'ютер.

У так званий дотехнологічний період збір мовних даних у мовознавстві завжди був складним технічним завданням, пов'язаним з ручним довготривалим опрацюванням письмових текстів як історичних, так і сучасних, опитуванням інформаторів, анкетуванням, створенням традиційних лексичних картотек тощо. Так, лексичну картотеку ***Словника української мови в 11-ти томах*** було започатковано у 20-х рр. минулого століття, а перший том словника побачив світ 1971 року, тобто через півстоліття.

У традиційному доборі мовного матеріалу є й інші проблеми, зокрема поновлення фактичного матеріалу, пошук необхідних одиниць у кількамільйонних картотеках, неможливий віддалений доступ до неелектронних ресурсів. Як слушно зазначає А.Баранов, до появи комп'ютера і відповідного програмного забезпечення для оброблення даних природної мови подолати проблеми, пов'язані зі збором, організацією та доступом до матеріалу для лінгвістичних досліджень, було майже неможливо'.

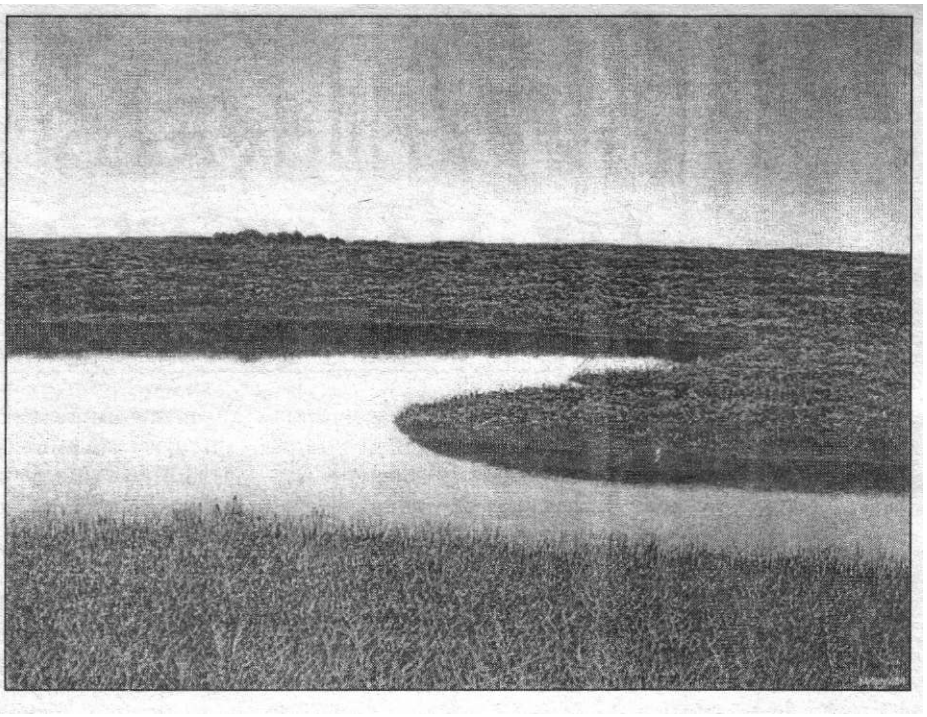
Завдяки застосуванню комп'ютера та програмних засобів оброблення природної мови до відбору та організації дослідного матеріалу з'явилися мовні комп'ютерні об'єкти різного типу та архітектури. Конститутивними елементами таких об'єктів залежно від типології стали одиниці природної мови різних реалізаційних рівнів — морфеми, лексеми, фраземи, словосполучення, речення тощо. Одним із таких лінгвістичних об'єктів став корпус текстів природної мови.

Суттєвою відмінністю корпусної організації мовного матеріалу від старих, головно картотечних, принципів є не лише технологічна відмінність, а насамперед високий рівень точності та надійності зберігання всієї мовної інформації й великі обсяги цієї інформації. Йдеться про зібрання текстів чи текстових уривків на 100 і більше мільйонів слововживань. Крім того, важливою перевагою є можливість отримати різноаспекгні відповіді щодо конкретної мовної одиниці впродовж короткого проміжку часу. Наприклад, працюючи з корпусом, за один сеанс роботи можна паралельно отримати інформацію про частоту вживання слова, час його появи у мові і/або зникнення з мови, усі можливі контексти, в яких функціонує слово чи фразеологізм, словоформи, варіанти слова тощо.

Говорячи про корпус, маємо на увазі електронний вигляд системно організованої та програмно обробленої вибірки текстів, що представляють всі історичні й географічні варіанти та форми існування довільної мови. Він призначений для мовознавчих досліджень і програмних застосувань

Загалом корпусно-базовані дослідження зорієнтовані на автоматичне добування лінгвальної, а також лінгвістичної інформації з корпусних текстів, оброблення даних переважно математичними процедурами і перевірку правильності інтерпретації цих даних.

З досвіду тих мов, для дослідження яких уже збудовано корпуси текстів[2], найчастіше корпусно-базовані дослідження стосуються вивчення лексики, граматики та історії мови. Також корпусні ресурси активно використовують у вивченні мови як рідної, так й іноземної.



Досліджуючи лексику, передовсім з'ясовують статистичні параметри лексичних одиниць мови. Наприклад, у CD-версії одного із корпусів польської мови (*Korpus języka polskiego PWN),* яка охоплює 4 млн. слововживань. частота лексеми *Україна* становить 800 вживань, а *Росія* - 607. Далі, дослідника зацікавлюють контексти лексеми *Україна*, серед яких, наприклад:

(1)... iż celem KPU ma być ustanowienie na Ukrainie «władzy ludu» w formie rad deputowanych ludowych. zapewnienie «socjalistycznej, ogólnonarodowej...»

(2)... «Ruch» Kierowany przez Wiaczesława Czornowiła Ludowy Ruch Ukrainy jest — historycznie rzecz biorąc — pienvszą, a zarazem największą demokratycznq organizacją ...

(3')... wszyscy powołani piłkarze byli przygotowani do meczu z Ukrainą w Kijowie...

(4)... W istocie to Rosja jest młodszą siostrą Ukrainy, która dala jej pierwiastki kultury i cywilizacji typu bizantyjsko-rzymskiego...

З контекстів можна з'ясувати вживання відмінкових закінчень досліджуваного слова, його синтасичну роль, можливості сполучуваності тощо. Аналогічні дослідження можливі також і на базі традиційних мовних ресурсів, але, щоб проаналізувати тексти на 4 млн. слововживань, вичленувати з них 800 необхідних прикладів і відібрати наведені, потрібно значно більше за 15 хвилин, витрачених на зазначені процедури при користуванні електронним корпусом.

Лексикологічні дослідження безпосередньо пов'язані з лексикографічним опрацюванням мовних одиниць. Укладаючи довільний словник на базі одномільйонного корпусу, і тим паче класичного стомільйонного корпусу, автор значно швидше сформує реєстр словника, визначить варіанти реєстрових одиниць, перевірить їх значення за контекстами, добере ілюстративний матеріал тощо.

Граматичні дослідження можуть стосуватися вивчення типології відмінювання та дієвідмінювання слів у мові, наприклад, змін у відмінкових закінченнях іменника. Для української мови цікавим і важливим буде аналіз функціонування закінчень родового відмінка в абстрактних та конкретних іменниках чоловічого роду, вивчення специфіки функціонування слів з однаковим граматичним значенням у текстах різного типу або у словосполученнях з різним граматичним зв'язком, визначення структури синтаксичних конструкцій різного типу, мінімальної — максимальної — класичної довжини речення залежно від типу тексту. Така інформація буде корисною для методики навчання мови і рідної, й іноземної. Навчаючи учнів творити текст, можна чітко визначити оптимальні параметри синтаксичних одиниць залежно від стилістично-жанрових характеристик тексту.

Для вивчення історії мови створюють так звані хронологічні корпуси. Аналіз хронологічних корпусів дозволяє з'ясувати, наприклад, першу фіксацію лесеми у мові, набір її парадигматичних форм та їхні реалізаційні моделі в різний історичний період, розвиток її семантики, переміщення в межах лексичної системи, зникнення з мови тощо. Важливо, що дослідним об'єктом може бути і семантично повнозначна, і неповнозначна одиниця зі своїм оточенням.

Будь-які мовознавчі дослідження ґрунтуються на мовному матеріалі. їх результати безпосередньо залежать від обсягу цього матеріалу, його достовірності, кількості параметрів, залучених до його добору та систематизації, об'єму сфери дії мовних закономірностей і явищ, які є об'єктом вивчення у кожному конкретному випадку. Таких вимог легко дотриматися до корпусно-базованому вивчення та дослідження мови, їх якість залежить передовсім від коректності побудови самого корпусу, яка має бути пов'язаною з національною філологічною традицією і, як справедливо зазначає В.Риков, повинна «мати власну концепцію, бути чітко детермінованою щодо призначення, а фактичний матеріал повинен репрезентувати реальні тексти національної мови, враховуючи весь історичний період її існування, варіантність мовної реалізації, стилістичну диференціацію і територію поширення»[3].

Примітки

[1] Баранов А.Н. Введение в прикладную лингвистику. — Москва, 2001. — С. 116.

[2] Див., наприклад British National Corpus, FRANTEXT, Český národní korpus, Polski Korpus Narodowy, Korpus IPI PAN, Korpus PWN, Korpus Slovenskega Jezika, Hrvatski elektronski tekstovni arhiv, Национальный корпус русского языка, Большой корпус русского языка та ін.

[3] Там само. — С. 118.