Марія Шведова
Кандидат філологічних наук
Київський національний лінгвістичний університет
м. Київ

КОРПУСНІ МЕТОДИ ДОСЛІДЖЕННЯ РЕГІОНАЛЬНИХ ОСОБЛИВОСТЕЙ ЛІТЕРАТУРНОЇ МОВИ

Українська корпусна лінгвістика в останні роки розвивається досить активно. Корпусні дослідження проводять вчені інституту

філології КНУ ім. Т. Г. Шевченка, Києво-Могилянської академії, Донецького національного університету (м. Вінниця), Національного університету «Львівська політехніка». Триває робота над Браунським українським корпусом обсягом 1 мільйон слововживань (Корпусна група БрУК). В інтернеті українська мова представлена кількома лінгвістичними корпусами: 1) Корпус української мови лабораторії комп'ютерної лінгвістики інституту філології КНУ ім. Т. Г. Шевченка, під керівництвом Н. П. Дарчук КНУ (mova.info); 2) корпуси текстів української мови кафедри загального та прикладного мовознавства і слов'янської філології Донецького національного університету (м. Вінниця) на сайті corpora.donnu.edu.ua; 3) відкриті корпуси української мови групи lang-uk, доступні для завантаження за адресою: lang.org.ua/uk/corpora/: 229 текстів з українського браунівського корпусу, корпус UberText (корпус художніх та публіцистичних текстів обсягом близько 600 млн. словоформ, призначений для комп'ютерної обробки), корпус законів та нормативно-правових актів України 4) Паралельні українсько-російський та російсько-український корпуси (www.ruscorpora.ru/search-para-uk.html); 5) Польсько-український паралельний корпус (domeczek.pl/~polukr/index.php) та інші. Але досі бракувало такого корпусу, де можна було б формувати підкорпуси для окремих досліджень – за часом написання текстів, авторами, регіонами, до яких належать тексти – і який був би достатньо великим і репрезентативним для цього.

Генеральний регіонально анотований корпус української мови (ГРАК – uacorpus.org) було укладено нами для дослідження регіональної варіативності української мови. Він наразі має загальний обсяг понад 190 млн словоформ, містить понад 8 тисяч текстів 3440 авторів.

В корпусі представлені оригінальні українські тексти і переклади з 38 мов, найбільше – з англійської (488 текстів, 22 млн сл.) та російської (443 тексти, 10 млн сл.). Частка перекладів – 29% (55 млн словоформ).

Склад корпусу за жанром: художні тексти – 113 млн сл. (60%), наукові тексти – 45 млн сл. (24%), публіцистика – 17 млн сл. (9%), а також мемуари, щоденники (у тому числі тексти з Facebook), інтерв'ю, промови та ін.

Всі тексти корпусу датовані, отже є можливість укладати підкорпуси за часом написання. Обсяг текстів за періодами такий:

1819-1916 рр. – 5,9 млн сл.

1917-1932 рр. – 6,4 млн сл.

1933-1969 рр. – 25,3 млн сл.

1970-1990 рр. – 34 млн сл.

1991-2018 рр. – 102,4 млн сл.

Більшість текстів корпусу походить з великих міст, обласних центрів, тому в основу розмітки корпусу за регіонами покладено сучасний адміністративний поділ України за областями. Залежно від мети дослідження ці підкорпуси можна об'єднувати різним чином і робити більші підкорпуси текстів. Обсяг підкорпусів за областями наразі такий: Львівська 30,856924 млн сл., Харківська – 19,67 млн сл., Полтавська – 12,32 млн сл., Черкаська – 10,34 млн сл., Івано-Франківська – 8,73 млн сл., Чернігівська – 8,11 млн сл., Вінницька – 7,67 млн сл., Київська – 7,5 млн сл., Волинська – 7,5 млн сл., Тернопільська – 7,03 млн сл., Дніпропетровська – 6,48 млн сл., Хмельницька – 6,17 млн сл., Донецька – 5,24 млн сл., корпуси інших областей менші за обсягом (але не менше ніж 1,5 млн сл.).

Якщо автор довгий час жив за кордоном або емігрував, йому приписували також іншу країну, таким чином утворився підкорпус текстів авторів діаспори (США, Канада, Польща, Німеччина, Велика Британія, Франція та ін.). Це здебільшого тексти емігрантів 40-х років і, менша частка, 1917-20-х рр. Обсяг корпусу української мови діаспори – понад 10 млн. словоформ.

Регіональну варіативність, за даними корпусу, можна побачити зокрема на прикладі вживання форми родового відм. у значенні знахідного для неістот (схопити ножа, купити телевізора). Це явище, приблизний діалектний розподіл якого можна побачити за корпусом письмових текстів, адже воно є характерною рисою окремих говірок, але також не виходить за межі літературної норми, і носії не уникають його у мовленні. Ця граматична форма вживається по всій території України, але частотність її в різних регіонах неоднакова. Для статистичного дослідження було відібрано за корпусом 7062 контексти, які містять сполучення дієслів з назвами частин тіла у форми другого знахідного відм. Середня частотність знайдених одиниць у загальному корпусі – 36,78 ipm (одиниць на мільйон словоформ). Найвища частотність у центральних регіонах: Київська

– 54,49 ipm, Черкаська – 58,05 ipm (високий показник частотності у Рівненській та Закарпатській областях виник через недостатню збалансованість корпусу, більшу частину прикладів знайдено в текстах двох авторів); найнижча частотність на сході: Луганська обл. – 16,28 ipm, Донецька обл. – 17,92 ipm. Ці результати підтверджуються даними досліджень, які знаходимо в авторитетних мовознавчих та діалектологічних працях.

Регіональна, хронологічна, жанрова розмітка і лінгвістичний інструментарій корпусу ГРАК дозволяють не тільки знаходити слова, граматичні форми та їх сполучення, а також робити різноманітні статистичні дослідження, дослідження сполучуваності, частотності та динаміки розвитку лексичних, фразеологічних, граматичних мовних явищ і їх регіональних відмінностей.