Наші інтерв'ю

УКРАЇНСЬКА КОРПУСНА ЛІНГВІСТИКА: СЬОГОДЕННЯ І ПЕРСПЕКТИВИ

Корпусна лінгвістика – порівняно новий, мало розроблений напрям сучасного мовознавства як у світі, так і в Україні. Відповідно здобутків у нас у цій галузі негусто. Тим цінніша поява праць, що активізують розвиток української корпусної лінгвістики. Одна з них – монографія Орисі Демської «Текстовий корпус: ідея іншої форми», яка нещодавно побачила світ і яка, власне, визначила тему нашої розмови.

Орися Мар'янівна Демська – кандидат філологічних наук, доцент кафедри української мови Національного університету «Києво-Могилянська академія». Коло наукових інтересів: лексикологія і фразеологія, лексикографія, корпусна лінгвістика.

Закінчила Дрогобицький державний педагогічний інститут (тепер – університет), працювала у Львіському національному університеті ім. Івана Франка, Вроцлавському університеті, Інституті української мови НАН України, викладала у Варшавському університеті та Московському державному університеті ім. Михайла Ломоносова.

Автор і співавтор близько 70-ти наукових праць, що вийшли друком в українських, російських, польських, французьких збірниках, а також окремими виданнями «Словник омонімів української мови» (1996), «Українська мова у XX сторіччі: історія лінгвоциду» (2005), «Українсько-польський тематичний словник» (2007), «Фразеологія» (2008), «Вступ до лексикографії» (2010), «Текстовий корпус: ідея іншої форми» (2011).

– Спробуймо означити корпусну лінгвістику (КЛ), аби дати нашим читачам, а це переважно вчителі-словесники, найзагальніше уявлення про те, на чому вона зосереджує свою увагу, що є її завданням?

– Ще з часів давнього світу прийшла до нас ідея вивчення людської мови, яку сприймаємо опосередковано – за допомогою тексту, усного або друкованого чи писаного. Спочатку вивчали писані, згодом друковані тексти різних творів: теологічних, філософських, фольклорних і так далі. Потім таких текстів стало дуже багато, і мовознавці почали укладати картотеки – зібрання цитат з переважно класичних творів національної літератури, записані на спеціальні картки, передусім для укладання словників, але й також для вивчення граматичних особливостей мови, її лексичного і фразеологічного складу. Однак ці картотеки мали певні обмеження. І одним з таких обмежень був час, який треба було згаяти, аби дістатися до картотечних матеріалів й опрацювати їх. Саме ці обмеження подолав електронний *текстовий корпус* – спеціально організоване зібрання текстів природної мови, призначене для наукового і практичного вивчення її. Згодом дослідження, вивчення, описування людської мови, що здійснюється на основі такого електронного текстового корпусу, об'єдналися під назвою *корпусна лінгвістика*. Таким чином, *корпусна лінгвістика* – це розділ мовознавчої науки, що зосереджений, з одного боку, на питаннях теорії і практики побудови різних комп'ютерних текстових корпусів національної мови, а з іншого – на питаннях теорії і практики наукового і практичного вивчення цієї мови через текстовий корпус, який можна вважати, за умови правильної його побудови, певною моделлю природної мови.

– Учені не мають єдиної думки щодо початків КЛ, тож хотілося б знати Вашу думку. І ще: що зумовило появу цього напряму мовознавства?

– Справді, дещо по-різному трактують витоки корпусної лінгвістики різні науковці. Одні розширюють її межі до початку XX століття, говорячи про протокорпусний (вступний корпусний) і власне корпусний періоди в лінгвістиці, інші – кажуть, що початки цього напряму можна співвіднести з появою перших електронних текстових власне корпусів (мова не йде про комп'ютерні текстові ресурси на зразок лінгвістичних баз даних, бібліотек тощо), тобто датувати його 60-ми роками XX століття, коли й з'явилися перші корпуси англійської мови, і, врешті, є й такі, що схильні говорити про корпусну лінгвістику від моменту появи терміна *корпусна лінгвістика* у 1984 році, коли вперше побачив світ науковий збірник під назвою «Корпусна лінгвістика».

Якщо ж робити якісь висновки чи узагальнення, то все-таки варто говорити про корпусну лінгвістику в широкому сенсі цього слова і відносити її появу до початку XX століття, як це роблять, зокрема, британські вчені Т.МакЕнері та

А.Вилсон, виділяючи в її історії початковий або підготовчий чи протокорпусний період і власне корпусну лінгвістику, базовану саме на електронних текстових корпусах, яка починає свій відлік із 60-х років минулого століття, і остаточно оформлюється як самостійний напрям у мовознавстві на середину 80-х років того ж XX століття.

– З однієї Вашої статті я довідалася, що інтенсивним розвитком КЛ позначені 1990-ті роки. У цей час з'являються національні корпуси майже всіх європейських мов. Розкажіть про це докладніше.

– Справді, 90-ті роки минулого століття позначені активністю корпусних студій на Європейському континенті й у Північній Америці. Не кажу, що таких студій немає за межами цього простору, але вони залишилися трохи поза моєю увагою. І це піднесення також можна трактувати як підтвердження того, що саме у 80-х роках корпусна лінгвістика остаточно оформилася як самостійний напрям.

Отже, що ж відбулося у 90-х роках з корпусною лінгвістикою? По-перше, вона перестала бути попелюшкою в науці про мову. Навіть ті мовознавці, які гостро її критикували, припинили це робити, а подекуди й навернулися до ідеї текстового корпусу як певної моделі мови, яку можна досліджувати й отримувати якісні результати. По-друге, у цей час з'являються комп'ютерні текстові корпуси національних мов не тільки, скажімо, англійської, німецької, французької, але й слов'янських. Зокрема один за одним побачили світ загальномовні корпуси чеської, польської, російської та й інших національних мов Славії. А такій появі конче мало передувати теоретичне обґрунтування засад корпусної організації національної мови, що й дало поштовх до нового етапу в розвиткові корпусної лінгвістики. І, по-третє, у цей період започатковано низку глобальних корпусних проектів, зокрема таких, як MULTEXT, MULTEXT-EAST, PAROLE, а також створення робочої групи EAGLES (Expert Advisory Group on Language Engineering Standards) і консорціуму TEI (Text Encoding Initiative). Отже, найзагальніше можемо казати, що саме цей період звершив етап теорії і практики побудови електронних текстових корпусів і став початком нового етапу – теорії і практики наукового й практичного вивчення мови на базі корпусу.

– Крім питань виникнення та розвитку, нашим читачам, гадаю, цікаво буде дізнатися й про теоретичне підґрунтя сучасної КЛ.

– Це надзвичайно широке запитання. Очевидно, що в нашому інтерв'ю я не зможу докладно розповісти про це, але загалом можу висловити свій власний погляд, своє наукове переконання, що теоретичним підґрунтям корпусної лінгвістики найбільше є структуралізм, чи, за словами В.Виноградова: «сукупність поглядів на мову та методів її дослідження, в основі яких лежить розуміння мови як знакової системи, у межах якої виділяються структурні елементи (одиниці мови, їхні класи тощо)». Крім цього, структуралізм виявляє прагнення до чіткого, наближеного до точних наук, формального опису мови. І саме це чи не найбільше вирішило долю структуралізму як теоретичного підґрунтя корпусної лінгвістики.

– Розкрийте ширше поняття текстового корпусу: суть, принципи його створення і, звичайно ж, практичне застосування.

– Про це я вже сказала, відповідаючи на перше запитання. Отже, трошки повторюся, текстовий корпус – це перетворена на електронну форму, стандартно організована, програмно опрацьована, репрезентативна вибірка текстів української мови, призначена для наукового і практичного вивчення нашої мови. Під це визначення найбільше підходить загальномовний корпус, чи корпус, до складу якого входять тексти, що відображають мову загалом: її літературний варіант, територіальні та соціальні діалекти, термінологічні системи тощо. Однак можливі також і так звані спеціальні, не загальномовні, а часткові корпуси, зокрема дитячої мови, молодіжного сленгу, окремим може бути і корпус мови письменника, наприклад Тараса Шевченка чи Івана Франка.

Проте далеко не кожен текстовий комп'ютерний ресурс можна вважати корпусом. Аби зібрання текстів, що мають електронну форму, можна було вважати корпусом, воно повинно відповідати певним критеріям. Цих критеріїв багато, але далеко не всі є обов'язковими. Загалом обов'язковими вважаються репрезентативність, стандартність, вичерпність. І якщо говорити про ці критерії, то вони означають таке. *Репрезентативність,* чи, надзвичайно спрощено, здатність корпусу правильно відображати всі властивості або мови загалом, або певного її рівня, через тексти, що входять до корпусу. *Вичерпність,* чи якомога повніше відображення або мови загалом, або певного її рівня у корпусі. І, врешті, *стандартність,* чи така організація текстових ресурсів у корпусі, що відповідає засадничим вимогам міжнародного стандарту кодування корпусів. Якщо ці вимоги не дотримані, то маємо справу з будь-яким іншим добрим комп'ютерним текстовим ресурсом, але, на превеликий жаль, не корпусом.

З вашого дозволу омину принципи створення корпусу, бо для цього треба переказати свою останню книжку, а також думки багатьох інших корпус-

них лінгвістів. Але із задоволенням розкажу про практичне застосування такого комп'ютерного об'єкта, як корпус. Отже, найперше, загальний корпус національної мови можемо використовувати в навчанні української мови і як рідної, і як чужої. Наприклад, учитель помітив, що учні роблять певні помилки в письмових роботах і в усному мовленні. Щоб подолати цю помилку, вчитель може дати завдання знайти такий приклад вживання слова у словосполученні чи реченні, вдавшись до корпусу. Якщо такого прикладу не буде, доречно доповнити завдання: знайти його відповідник, який в академічному загальномовному корпусі матиме правильну, тобто без помилки, форму, і скласти з ним речення. За допомогою корпусу можна перевірити правильність вживання відмінкових форм іменних частин мови, дієвідмінювання, сполучування слів, як-от *згідно з* чи *згідно до, навчати мови* чи *мові* і багато іншого. Надзвичайно корисним є корпус тоді, коли ви вивчаєте чужу мову або коли викладаєте рідну мову як чужу, оскільки за його допомогою найпростіше побачити, як правильно поєднувати й узгоджувати слова в реченнях, які словосполучення є стійкими, а які взагалі не притаманні виучуваній мові й утворені за моделлю рідної, і ще багато іншого.

Не можу оминути завдання корпусу, яке практично безпосередньо й зумовило його появу. Ідеться про те, що корпус є центром всесвіту словникарства. Без текстів нема словників. І одним з основних застосувань корпусу є використання його під час укладання словників. В ідеалі сучасні словникарі з корпусу добирають слова й визначають їхні значення, звідти ж приходять і словникові цитати чи приклади вживання слова з конкретним значенням. На основі авторських корпусів укладають словники мови письменників, на основі корпусів дитячої мови – словники для дітей. І так далі... Але треба пам'ятати, що з розвитком мови має розвиватися і корпус.

– Ми з Вами говоримо про практичний бік справи, а можливо, є ще якісь інші аспекти застосування чи існування корпусу?

– Так. Є один надзвичайно важливий аспект існування електронного текстового корпусу – це його загальнокультурне значення. Колись вважалося, що розвинена національна культура повинна мати у своєму арсеналі великий академічний словник національної мови й теоретичну граматику. Нині до цих двох китів додався класичний третій – національний корпус. Погляньте на континентальні народи – німців, французів, англійців, чехів, поляків... – усі вони мають академічний словник, теоретичну граматику і національний корпус, не кажучи вже про низку інших корпусів. Зазвичай у розбудованих національних мовознавчих традиціях існує по кілька, а подекуди по багато найрізноманітніших текстових корпусів.

– У Вашій монографії є розділ «Лінгвістична ідеологія національного корпусу української мови», де запропоновано його проект. Чи набула ця ідея реального втілення, якщо так, то якою мірою реалізована?

– Первинно цей проект мав виконуватися в Інституті української мови НАН України. Але маємо розуміти, що, по-перше, це доволі дороге задоволення, і, по-друге, воно потребує багато часу й людської праці. На момент появи та обґрунтування цієї ідеї ще не було готових кадрів для її реалізації. Напевно, навіть ще не було й самого розуміння, для чого потрібен цей проект. Однак, зауважимо, що через такий стан речей у різний час проходили практично всі національні корпусні лінгвістики. Яка ситуація нині з проектом побудови національного корпусу української мови в Інституті української мови НАНУ, я не дуже знаю, оскільки працюю в Національному університеті «Києво-Могилянська академія» й не можу осягнути неосяжне. Проте думаю, що поява корпусу української мови – це лише питання часу.

– Наші читачі – це переважно вчителі-словесники, їх, найімовірніше, цікавить питання про те, який шкільний вимір цієї справи? Чи є він?

– Працюючи над своїми книжками, присвяченими теоретичному аспектові корпусних студій, я не надто зважала на цей аспект. Мене більше цікавила наука. Так-так, саме наука задля науки. Але погоджуюся з вами, що має бути і шкільний вимір цієї справи. Очевидно, що зараз я не дам якоїсь чіткої відповіді, бо тільки почну про це системно думати. Проте, поза сумнівом, що шкільний, методичний аспект текстового корпусу, особливо загальномовного, є чи не одним із найважливіших. І що найцікавіше, найперший комп'ютерний текстовий корпус англійської мови – так званий Браун Корпус (укладений в Університеті Брауна, США) – у 1962 році був створений на замовлення Міністерства освіти США. Цікаво, але цей факт тільки під час нашого інтерв'ю я актуалізувала для себе. Отже – шлях на освіту...

– Останнім часом змінюються шкільні програми, приходять нові підручники – можливо, є сенс увести в них КЛ?

– На все повинна бути повнота часу, як казав Іван Павло II. Навіть в академічних підручниках аспекти корпусної лінгвістики лише починають з'являтися. Українська корпусна лінгвістика ще дитя. Треба вирости. Головне, щоб цьому ростові дали право.

Спілкувалася ***Мирослава-Марія РИБАЛКО***