ОРИСЯ ДЕМСЬКА-КУЛЬЧИЦЬКА

ЩО НОВОГО В НАУЦІ ПРО МОВУ?

Відомо, що наука про мову виникла за кілька століть до нашої ери в Давній Греції та Індії, Китаї і Арабському Сході. З тих часів дійшли до нас граматика індійського вченого Паніні (5—4 ст. до н.е.), праці європейців Платона й Арістотеля про природу і сутність назв, початки фонетики й синтаксису, вчення про частини мови.

Сьогодні маємо у мовознавстві розгалужену систему розділів, виокремлених відповідно до того, який саме тип мовних одиниць є предметом її вивчення: фонетика і фонологія, акцентологія, словотвір, морфологія, синтаксис, лексикологія, фразеологія та ономастика, лексикографія, семасіологія, стилістика, діалектологія, лінгвогеографія. Усі вони уже давно вважаються класичними мовознавчими дисциплінами. Але у 20 ст. окреслюється цікава тенденція у розвитку науки про мову. Так, крім власних теорій і методик, мовознавство все частіше звертається до теоретичної бази і методологічних принципів інших наук: психології, соціології, етнології, біології, математики, кібернетики тощо, внаслідок чого виникають нові, так звані стикові напрямки і дисципліни.

У 50-х роках минулого століття на межі психології та лінгвістики формується психолінгвістика, предметом вивчення якої є психічні процеси творення комунікативного акту засобами мови. Тобто глобально йдеться про природу і механізми формування висловлювань, їх сприйняття та розуміння мовцем і слухачем. На перетині соціології та лінгвістики виникає соціолінгвістика, що в центр уваги ставить соціальні аспекти функціонування людської мови. Із поєднання етнології та лінгвістики постала етнолінгвістика, яка простежує взаємодію мовних, етнокультурних й етнопсихологічних чинників у функціонуванні та еволюції мови. Контакт мовознавства з біологією стимулював розвиток нейролінгвістики. Ґрунтуючись на лінгвістичних даних, ця галузь досліджує функції та зони центральної нервової системи, пов’язані з мовною діяльністю людини.

Новаторським напрямком розвитку мовознавства XX ст. стала математична лінгвістика, яка виникла на перетині лінгвістики і математики у 20-х роках. Використовуючи методи математичної логіки для формального опису категорій природної мови, мовознавство виявилося тією гуманітарною наукою, яка вперше почала залучати не лише інструментальні методи спостереження над мовними одиницями, а й систематично застосовувати математичні прийоми разом з ЕОМ у вивченні мови. Отже, у межах математичної лінгвістики почали розвиватися часткові напрямки, зокрема, обчислювальна лінгвістика, метою якої є створення складних систем обслуговування комп’ютера, які уможливили б спілкування в системі «людина — машина» за допомогою людської мови. Використання комп’ютера спричинило також виникнення корпусної лінгвістики, яка вивчає теорію і практику формування електронних мовних баз — *корпусів*.

У нашій мові давно стали звичними словосполучення на зразок: *корпус миру, військовий корпус, корпус годинника, дипломатичний корпус, корпус державних службовців.* Натомість словосполучення *корпус мови* звучить по-новому і зрозуміле не кожному. Отже, спробуймо з’ясувати, що таке *корпус мови*.

Приблизно з середини 60-х років минулого століття у комп’ютерному мовознавстві існує практика перетворення текстів у електронний вигляд і зберігання таких текстових даних у пам’яті машини. На початках, коли електронно-обчислювальні машини були громіздкими і займали величезні приміщення, така практика не була популярною. І лише з появою персонального комп’ютера вона значно активізується, а з початку 90-х років 20 ст. учені починають створювати великі систематизовані зібрання текстів національних мов на машинних носіях. Останні дістали назву *корпус* (*corpus*). Однак довільне зібрання машинних текстів природної мови ще не можна називати *корпусом*. Тексти повинні бути відібрані згідно з визначеними критеріями, відповідати певним вимогам, бути систематизованими, закодованими й організованими відповідно до вимог Стандарту кодування корпусу (1996).

Критерії відбору текстів до корпусу залежать від того, який саме корпус ми будуємо: усієї сучасної мови, корпус періодики, діалектний чи історичний тощо. Наприклад, якщо йдеться про створення Українського Національного корпусу, то критерії відбору текстів повинні передбачати діахронний аспект (які тексти і якого часового відтинка відібрати), стилістичний (доцільно репрезентувати усі стилістичні системи національної мови), територіальний (слід врахувати специфіку літературної мови залежно від регіону України та у діаспорі), квантитативний аспект (чітко зумовлює кількість слів у тексті чи уривку, внесених до корпусу). Загалом критерії відбору текстових уривків є окремою проблемою однойменної теорії відбору матеріалу до корпусів різних типів.

Відібрані та внесені до комп'ютера текстові дані об'єднуються згідно з такими вимогами:

а) автентичність (забороняє будь-які модифікації текстового матеріалу);

б) еталонність (уведені до корпусу твори художньої літератури, політичні есе, наукові тексти, епістолярій видатних людей тощо повинні відповідати нормам літературної мови);

в) квантитативність (обсяг сучасних корпусів традиційно починається від одного мільйона слів і сягає понад п'ятсот мільйонів);

г) закодованість (передбачає додавання до тексту певної формальної інформації шляхом вписування у класичний текст відповідних символів, зрозумілих для комп'ютерних програм оброблення текстових даних).

Тексти чи уривки текстів у будь-якому корпусі повинні бути систематизованими. Тобто усі тексти в корпусі об'єднуються в більші структурні одиниці на підставі певних характеристик, наприклад: системномовних (загальнонародна мова — діалект — професійна мова); стилістичних (художній твір — наукова монографія — дитяче оповідання); тематичних (лінгвістика — право — медицина — міжнародні відносини — ...); часових (... — 19 ст. — 20 ст. — 21 ст.) тощо.

Кожен текст — група текстів — підкорпус в межах корпусу повинна мати детальну паспортизацію. Вона передбачає інформацію про назву твору, його частини та розділи, підзаголовки, автора, дату видання (перевидання), місце видання, видавця, кількість сторінок, обсяг твору (як правило подається кількість слів і обсяг у байтах), методи і ресурси кодування. Часто якісні і кількісні бібліографічні параметри автори корпусу встановлюють індивідуально, але передумовою паспортизації додаткових бібліографічних параметрів є основні, спільні для всіх текстів і корпусів дані: автор, назва видання, назва твору, місце і рік видання, кількість сторінок.

Напрямок науки, пов'язаний з комп'ютерними технологіями, вводить у поле мовознавчих досліджень категорію стандартності. Так, якщо ми будуватимемо корпус, не дотримуючись певних вимог, то стандартні програми, призначені для роботи з текстами природної мови, не працюватимуть з нашим корпусом, жоден пошук не буде реалізований. Тобто ми не зможемо дати машині команду знайти потрібні нам слова, синтаксичні конструкції різного типу, цитати тощо.

Застосування електронних корпусів національних мов у сучасному технологічному світі найрізноманітніше. Залежно від сфери застосування, корпус мови може бути використаний для:

- підготовки правопису і організації довідково-консультативної правописної служби в Інтернеті;
- укладання класичних словників та їх комп'ютерних варіантів;
- написання підручників, посібників з української мови для середньої та вищої школи, а також підручників з української мови як іноземної;
- забезпечення сучасним методичним матеріалом курсу ділової української мови;
- впровадження найсучасніших методик навчання української мови та літератури в школі;
- організації методичної допомоги вчителям української мови та літератури в режимі on-line і на компакт-дисках;
- забезпечення учнів і вчителів середніх шкіл програмними текстами з літератури в автоматизованому вигляді;

- створення довідкових засобів у мережі Інтернет для іноземних посольств, консульств, представництв, українських державних і громадських осередків за кордоном;
- українізації комп'ютера і програмних продуктів в Україні;
- побудови машинної мовної моделі як технологічної бази для розробок у галузі інформаційних технологій;
- створення програм автоматичного розпізнавання і синтезу мовлення;
- створення програм машинного перекладу з мови на мову;
- мовної підтримки процесу створення аудіовізуальної реклами і товарних етикеток українською мовою.

Як бачимо, нові напрямки розвитку мовознавчої науки характеризуються високим рівнем технологічності, що забезпечує лінгвістиці особливе місце в сучасному інформаційному світі.