ОРИСЯ ДЕМСЬКА-КУЛЬЧИЦЬКА

ЩО ТАКЕ КОРПУС ТЕКСТІВ?

Український мовець давно звик до словосполучень на кшталт: *корпус миру, військовий корпус, дипломатичний корпус, корпус державних службовців, корпус годинника.* Тому доволі дивними йому можуть видатися словосполучення *корпус текстів, національний корпус, текстовий корпус національної мови.* Отже, спробуймо з'ясувати, що таке *корпус текстів.*

Уявімо, що на уроці української мови вчитель поставив завдання вибрати з літературних текстів ЗО типових словосполучень зі словом *день.* Скільки часу потрібно для виконання? Це залежатиме від того, яким чином ми його виконуватимемо. Якщо просто озброїтися олівцем і виписувати потрібні словосполучення з книжок, журналів і т. ін., це може забрати кілька тижнів, може, й місяців... Але сучасний технологічний досвід дозволяє досягти того самого результату іншим шляхом. А саме: перш ніж відібрати тексти, прочитати їх і виписати словосполучення, варто скористатися комп'ютером і організувати новий матеріал, репрезентований як *корпус текстів.* Приблизно з середини 60-х років XX ст. у мовознавстві поширилася практика перетворення текстів природної мови в електронний вигляд і зберігання їх у пам'яті машини. Особливої популярності набуває цей досвід із появою персонального комп'ютера, а з початку 90-х років учені починають активно створювати великі систематизовані зібрання текстів національних мов на машинних носіях. Такі зібрання отримали назву *корпус* (corpus). Однак довільне зібрання машинних текстів природної мови ще не може називатися корпусом. Для цього тексти мають бути відібрані згідно з визначеними критеріями, відповідати певним вимогам, бути систематизованими, закодованими і організованими відповідно до вимог Стандарту кодування корпусу.

Критерії відбору текстів визначають залежно від специфіки створюваного корпусу: усієї мови (загальномовний, національний корпус), корпус періодики (охоплює тільки тексти періодичних видань, написані літературною мовою), діалектний (репрезентує діалектні тексти), історичний (репрезентує тексти певного історичного періоду існування мови) тощо. Скажімо, 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 також підстилі національної мови); територіальний (не слід нехтувати специфікою літературної мови залежно від регіону України, як і тим фактом, що українська мова є засобом творення літературного усного або писемного тексту за межами України); квантитативний (чітко обумовлює кількість слів у тексті чи уривку, внесеному до корпусу, кількість самих текстів і / або уривків). Загалом у корпусних дослідженнях критерії відбору текстових уривків є окремою проблемою, в межах якої сьогодні розробляються лінгвістичні та технічні критерії, а в науковій літературі роботи з цієї проблематики узагальнюються як теорія критеріїв відбору текстового матеріалу до корпусів різних типів.

Відібрані та внесені до комп'ютера текстові дані об'єднують у єдине ціле — *корпус.* Як цілісний об'єкт, корпус має відповідати певним вимогам, як-от: а) *автентичність,* що забороняє будь-які модифікації текстового матеріалу; б)*репрезентативність,* тобто введені до корпусу твори художньої літератури, політичні есе, наукові тексти, епістолярій тощо мають відтворювати реальний стан мови в кожен конкретний період її функціонування; в) *квантитативність* — відлік обсягу сучасних корпусів починається від одного мільйона слів і сягає понад п'ятсот мільйонів; г) *закодованість* — передбачає, що до звичного для нас тексту додається певна формальна інформація шляхом вписування у класичний текст відповідних символів, за допомогою яких машина розуміє цей текст і зможе відповісти на питання щодо тексту, наприклад, видати ЗО типових прикладів слововживання лексеми *день.*

Тексти чи уривки текстів у будь-якому корпусі систематизуються Тобто всі тексти 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 (*... — XIX ст. — XX ст. – XXI ст.)* тощо. Кожен текст /уривок або групу текстів /уривків детально паспортизують відповідно до спеціально розроблених з цією метою стандартів. Паспортизація передбачає фіксацію інформації про назву твору, його структуризацію на частини, розділи, відомості про автора, дату видання, можливо, перевидання, місце видання і перевидання, видавця, кількість сторінок, обсяг тощо. Часто якісні та кількісні паспортні параметри встановлюють індивідуально автори корпусів, але обов'язковими є основні, спільні для всіх текстів і корпусів дані: автор, назва видання, назва твору, місце і рік видання, кількість сторінок.

Напрямок науки, пов'язаний із комп'ютерними технологіями, упроваджує в мовознавчий обіг категорію стандартності. Так, якщо ми будуватимемо корпус, не дотримуючись стандартних вимог, то програми, призначені для роботи з корпусами, не працюватимуть із нашим корпусом, ми не зможемо дати машині команду знайти потрібні словосполучення. Жоден пошук не буде реалізований, отже, ми не отримаємо необхідної інформації (наприклад, про слово *день).* Будь-який стандартно закодований корпус повинен, крім інших, обов'язково складатися з двох частин: 1) електронного заголовка, 2) власне закодованого тексту.

Електронний заголовок — це початкова частина електронного документа, де через систему спеціальних символів-кодів здійснено паспортизацію текстового документа. Загалом його можна співвіднести з титульною сторінкою книжки.

Закодований текст передбачає, що у межах звичного для нас тексту мають з'явитися спеціальні позначки, які містять інформацію про структуру самого тексту, про речення і слова, іноді — про звуки чи інтонацію. Така позначка дозволяє машині ідентифікувати у процесі пошуку відмінкові форми *дня, днем, дню* із лексемою *день.* Важливо, що такі операції пошуку відбуваються упродовж 20-30 секунд. Наприклад, пошук слова *Україна* в польському корпусі періодики на один мільйон слововживань тривав 18 секунд, забезпечивши 711 прикладів.

Отже, *корпус текстів, національний корпус, текстовий корпус національної мови* — це систематизоване, структуроване, прорамно оброблене зібрання репрезентативних текстів природної мови різних варіантів та форм її існування. Корпусний (технологічний) спосіб отримання інформації про мову передбачає застосування комп'ютера і корпусних методик для наступного аналізу чи опису. Він дає змогу, раз увівши текстовий матеріал у машину, оформивши його як корпус, надалі безперешкодно, швидко і максимально повно отримувати найрізноманітнішу інформацію.