МЕТОДИ КОРПУСНОЇ ЛІНГВІСТИКИ В ПІДГОТОВЦІ ФАХІВЦІВ-ФІЛОЛОГІВ

Таценко Н. В.
доктор філологічних наук, доцент,
завідувач кафедри іноземних мов
Сумського державного університету
м. Суми, Україна

Сьогодення позначене розвитком інформаційного суспільства знань, що спричинило бурхливий прогрес у галузі комп'ютерних технологій опрацювання природної мови. Перед мовознавством постали нові завдання щодо аналізу різних властивостей мовної системи, які потребують дослідження різнопланових мовних одиниць, структур, мовленнєвих явищ не на окремих показових прикладах, а в їхньому повному репрезентативному обсязі. Зрозуміло, що це вимагає застосування спеціального комп'ютерного інструментарію, зокрема, залучення методів корпусної лінгвістики. Попит на корпусні дані збігся з появою відповідних технічних можливостей.

Корпусна система даних, з одного боку, істотно спрощує пошук матеріалу, однак, з іншого боку, вимагає глибокого знання і творчого використання різних підходів і методик лінгвістичних розвідок. Корпусний дослідницький напрям продовжує виформовуватись, а його проблематика пов'язана з розробкою теоретичних засад і практичних прийомів побудови, машинного опрацювання й експлуатації лінгвальних даних, оформлених як корпус текстів [2]. Наразі є безліч визначень поняття «корпус», тому ми наведемо лише декілька з них. Представники Ланкастерської школи корпусної лінгвістики визначають його як збірник мовних фрагментів, відібраних відповідно до чітких мовних критеріїв для використання як моделі мови [5]; крім того, він являє собою репрезентативну збірку текстів, зазвичай у машиночитаному форматі, що вміщує інформацію про

ситуацію текстотворення, тобто інформацію про мовця, автора, адресата й аудиторію [3].

Незважаючи на поширене переконання, що корпус мовних даних повинен оброблятися кількісно (статистично), доводячи чи спростовуючи гіпотетичні припущення за умови високої частоти випадків, він цілком здатний бути й цінним джерелом якісної інформації – прикладів уживання мови в природній комунікації [5, с. 75-77].

Зауважимо, що в сучасних визначеннях простежуються основні риси корпусу текстів – мета або логічна ідея, машиночитаний формат, репрезентативність як результат певної відбіркової процедури, а також наявність металінгвістичної інформації. Це уможливлює розуміння **корпусу** як стандартно поданого зібрання писемних або усних репрезентативних машиночитаних текстів фіксованого розміру, призначених для лінгвістичного опису та аналізу певної мови або діалекту й відібраних та впорядкованих відповідно до інтра- й екстралінгвальних критеріїв. Стандартизоване подання словесного матеріалу на машинному носії дозволяє використовувати стандартні програми його обробки.

Корпуси бувають по-різному організовані залежно від прагматичної мети їхніх творців, хоча використовуються і для цілей, не передбачених авторами. Їх поділяють на одномовні та багатомовні, загальні і спеціалізовані, паралельні й порівняльні (перекладацькі корпуси), синхронні та діахронні тощо. Тексти як складові елементи корпусу можуть бути цілісними оригінальними словесними творами або їхніми частинами. Матеріали, зібрані в текстах, – це зразки мови, яка використовується в комунікативній діяльності спільноти, вони охоплюють різні жанри й періоди.

Доцільність створення та смисл використання корпусів визначаються такими передумовами:

– достатньо великий (репрезентативний) і збалансований обсяг корпусу гарантує типовість даних і забезпечує повноту подання всього спектра мовних явищ;

– різнотипні дані перебувають у корпусі в своїй природній контекстуальній формі, що уможливлює їхнє всестороннє та об'єктивне дослідження;

– один раз створений і підготовлений масив даних здатний використовуватися багаторазово, різними дослідниками із різною метою [1, с. 6].

В останні роки кількість і різноманітність створюваних комп'ютерних корпусів значно зросла й на сьогодні їх зафіксовано понад 600. З-поміж сучасних корпусів англійської мови (як американського, так і британського варіантів) найбільш відомими є Британський національний корпус (*British National Corpus – BNC*), Міжнародний корпус англійської мови (*International Corpus of English – ICE*), Лінгвістичний банк англійської мови (*Bank of English*), Корпус сучасної американської англійської мови (*Corpus of Contemporary American English – COCA*) тощо.

Корпуси уможливлюють розв'язання проблем із тлумаченням і використанням синонімів та «майже синонімів», тобто слів, які не є взаємозамінними [4, с. 45]. Оскільки лексеми тісно пов'язані з контекстом, їхнє значення розрізняють за шаблонами або патернами (*patterns*) і фразами, у яких вони типово з'являються. Співвідношення значень і патернів необхідно розглядати з допомогою полісемантичних слів. Водночас слова з подібними значеннями вживаються в однакових патернах.

Окрім загального аналізу даних щодо використання лексем та їхніх значень, які асоціюються з певними патернами, з допомогою конкордансів можливо спостерігати за їхнім статусом і функціями, сполученням з іншими словами й за тим, що ці сполучення означають [4, с. 51]. Маємо зауважити, що патерни є імпульсом дослідження конституентів мови, оскільки виявлення типу і значень патернів прирівнюються до процесу розуміння людиною граматичної структури мови.

Для роботи з корпусами великих розмірів, де кількість даних для опрацювання є досить об'ємною, науковці пропонують досліджувати щоразу по 30 випадково вибраних ліній конкордансу до тих пір, поки подальша вибірка не

перестане видавати чогось нового. Такий аналіз є «гіпотетичним тестуванням», в якому мала вибірка ліній стає основою для створення низки гіпотез про патерни. Проте такий спосіб дослідження застосовують лише для слів із дуже високою частотністю вживання [6, с. 157]. Важливим є те, що з обраних патернів отримуються контури семантичного поля певної досліджуваної лексеми.

Виявлення й аналіз таких структур є надзвичайно корисним для створення комп'ютерних програм, за допомогою яких запрограмовані патерни видаватимуться автоматично, безвідносно до попередніх знань і уявлень про те, якими вони повинні бути. Фонетичні, морфологічні, синтаксичні та стильові патерни репрезентують типове для корпусу та є більш уживаними, ніж сталі вирази. Вони можуть бути підґрунтям для узагальнення багатьох лінгвістичних теорій, зроблених мовознавцями раніше, і надихати їх на нові ідеї про сполучуваність кожного окремого слова у мовленні.

Необхідно додати, що корпуси є джерелом багатовекторних лексикографічних робіт зі створення сучасних та історичних словників. Якщо традиційна лінгвістика, когнітивна лінгвістика, лінгвопрагматика, соціолінгвістика мають на меті опис, оцінювання й використання мовної структури, то корпусна лінгвістика – це методологія, яку можливо використовувати в усіх наведених сферах мовознавства. Двоїстий характер корпусної лінгвістики (налаштованість як на створення, так і на використання корпусів) обумовлюється подвійним характером її об'єкта – текстів, що являють собою вихідний мовленнєвий матеріал для мовознавців і водночас є продуктом корпусних досліджень. Окремий аспект складає теорія і практика програмного опрацювання корпусних ресурсів.

Корпуси допомагають проаналізувати спонтанне мовлення, мовлення різних гендерних, соціальних, етнічних та вікових груп, інформацію про особливості певного діалекту; істотно змінюють уявлення про мовленнєву культуру й мовні норми, являючи собою надійні критерії для оцінювання й визначення прийнятності певних узусних явищ. Наявний масив мовленнєвих даних за певний період уможливлює здійснення аналізу лексико-граматичних

характеристик мовлення різних авторів та різних жанрів, перевірки автентичності текстів, компаративних розвідок, інтертекстуальності, перекладу, прихованих (потенційних) моделей уживання лексики, її квантитативного навантаження, розвитку й динаміки концептів у часі і навіть використання у розслідуванні злочинів (forensic linguistic analysis).

Отже, у процесі навчання фахівців-філологів корпусні ресурси забезпечують викладачів емпіричним матеріалом для підтвердження їхніх гіпотез, а також екстралінгвальною інформацією (вік, рід автора чи мовця, часові та просторові параметри походження тексту тощо).

Література

1. Захаров В. П., Богданова С. Ю. Корпусная лингвистика: учебник для студентов направления «Лингвистика». 2-е изд., перераб. и доп. Санкт-Петербург: СПбГУ. РИО. Филологический факультет, 2013. 148 с.

2. Таценко Н. В. Корпусна лінгвістика як методологія сучасних філологічних наукових розвідок // Наукові записки Вінницького державного педагогічного університету імені Михайла Коцюбинського. Серія: Філологія (мовознавство): збірник наукових праць. 2016. Вип. 23. С. 281-286.

3. Finegan E. Language: its structure and use. New York: Harcourt Brace College Publishers, 2004. 575 p.

4. Hunston S. Corpora in applied linguistics. Cambridge: Cambridge University Press, 2002. 254 p.

5. McEnery T., Wilson A. Corpus linguistics: an introduction. 2nd ed. Edinburgh: Edinburgh University Press, 2001. 235 p.

6. Sinclair J. M. A way with common words // Out of corpora: studies in honour of Stig Johansson / eds. H. Hasselgard and Oskefjell. Amsterdam: Rodopi, 1999. P. 157-179.

Таценко Н.В. Методи корпусної лінгвістики в підготовці фахівців-філологів. *Scientific and pedagogic internship "Organization of educational process in the field of philological sciences in Ukraine and EU countries": Internship proceedings, August 24 – October 2*: тези доповідей. Венеція: Venice : Izdevnieciba "Baltija Publishing", 2020. С. 177-181.