Людмила Могельницька,
к. філол. н., доцент
Державний університет «Житомирська політехніка»

РОЛЬ ЛІНГВІСТИЧНИХ КОРПУСІВ В ОБРОБЦІ ПРИРОДНОЇ МОВИ

Серед векторів сучасної прикладної лінгвістики значне місце посідає корпусна лінгвістика, становлення якої датується 60-ми роками XX століття. Активне застосування комп'ютерних технологій та спеціального програмного забезпечення значною мірою полегшило збір лінгвістичних даних та викликало появу нових способів дослідження мови. Лише за декілька секунд стало можливим здійснювати пошук у багатомільйонних текстових масивах (лінгвістичних корпусах), будувати конкорданс для будь-якого слова, одержувати дані про частоту словоформ, лексем, граматичних категорій, синтаксичних конструкцій, відстежувати зміни в частоті й контексті мовної одиниці в різні хронологічні періоди, одержувати дані про лексичну (колокацію) та граматичну (колігацію) сполучуваність тощо. [2; 4]

Основним поняттям корпусної лінгвістики виступає «корпус текстів» або «лінгвістичний корпус», що трактується в мовознавстві як уніфікований, структурований, розмічений, філологічно компетентний, значний за обсягом масив мовних даних електронного формату, створений для вирішення конкретних лінгвістичних завдань [1; 6]. Мовний корпус становлять тексти, що походять з реальних першоджерел: книг, газет, журналів, інтерв'ю, промов тощо, та представляють мову в природному середовищі її функціонування [3; 5]. Наразі в мережі Інтеренет вільним і безкоштовним є доступ до великої кількості корпусів різних типів, розмірів та мов, зокрема, Британський національний корпус (British National Corpus) [8], Американський національний корпус (American National Corpus) [7], Корпус Берлінської Брандербузької академії наук (DWDS-Corpus) [10], Корпус французької мови (Corpus de Référence du Français parlé) [9].

Вивчення корпусів дозволяє одержувати точні дані про лексичний склад мови, певної підмови, жанру чи індивідуального стилю письменника, виявити останні зміни в лексичному складі мов, різні його варіації (наприклад, поява й зникнення неологізмів) та включити ці дані до словника. (лексикологічних досліджень. Корпуси широко використовуються для укладання довідкової літератури – граматик, навчальних словників та довідників. Серед корпусобазованих граматик англійської мови, зазначимо, наприклад, (Collins COBUILD English Grammar(1990); Longman Grammar of Spoken and Written English (1999). На основі лінгвістичних корпусів проводяться дослідження синтаксичних явищ і конструкцій, зокрема, вивчають валентність лексичних одиниць, типи керування в прийменникових словосполученнях, порядок слів, синтаксичні функції частин мови в реченні, тощо. У такому випадку лінгвістичний корпус виступає як результат, з одного боку, та джерело, з іншого, обробки природної мови.

Обробка природної мови спирається на машинне навчання, що відбувається шляхом автоматичного засвоєння правил через аналіз великих корпусів типових реальних прикладів. Навчати комп'ютери розуміти

природну мову є завданням комп’ютерної лінгвістики як вектора прикладної лінгвістики. Продуктами комп’ютерної лінгвістики, серед яких зазначимо голосові асистенти (Siri; Alexa; Ok, Google), сентимент-аналіз, що передбачає аналіз тексту з метою встановлення його емоційного забарвлення, чат-боти, системи перевірки орфографії, програми з коригування стилю письма, фільтрування спаму, машинний переклад користуються відомі компанії на кшталт Google, Amazon, Facebook, Apple, Twitter, банки тощо. Генерація тексту широко використовується в автоматизованій журналістиці. Процес обробки природної мови є складним з огляду на ряд причин, серед яких семантичні й орфографічні зміни, лексична та синтаксична полісемія, метафоричність, запозичення, неологізми. Комп’ютерні лінгвісти використовують готові датасети, власне, мовні корпуси, наприклад, Europarl, Coronavirus Corpus для тренування моделей. За допомогою анотованого лінгвістичного корпусу вивчають специфіку словозміни, аналізують морфологічні показники мови, наприклад, визначають найтиповіші синтаксичні функції різних частин мови, або встановлюють, з якими відмінками найчастіше використовується той чи інший прийменник [1; 6].

Аналіз даних корпусу відбувається із використанням відповідного комп’ютерного забезпечення, завдяки чому опрацювання великих обсягів матеріалів природної мови, пошук відповідних одиниць, сортування результатів пошуку й анотування текстів стає значно швидшим й ефективнішим. Аналіз «живого» мовлення, що функціонує у «природних» реальних ситуаціях сприяє розширенню словникового запасу ШІ та підвищує рівень розуміння ним різних мовленнєвих моделей [1; 6]. Корпуси уможливлюють розв’язання проблем із тлумаченням і використанням синонімів та «майже синонімів», тобто слів, які не є взаємозамінними. Значення лексем, що тісно пов’язані з контекстом, розрізняють за шаблонами або патернами (patterns) і фразами, у яких вони типово вживаються.

Підсумовуючи, наголосимо, що корпусні студії зосереджуються на аналізі природної мови в умовах реального функціонування з використанням комп'ютерних технологій на основі великих за обсягом, ретельно відібраних та впорядкованих текстових корпусів. Корпусна лінгвістика впровадила нові методи аналізу лінгвального матеріалу, розширила межі наукового досліджен ня, забезпечила зручні інструменти для ефективного опанування мовою. Комп'ютерні лінгвісти послуговуються статистичними й лінгвістичними закономірностями, що були виявлені на матеріалі корпусу, для створення комп'ютерних моделей мови. Корпуси використовуються для розробки й налаштування автоматизованих систем, наприклад, машинного перекладу, розпізнавання мовлення, інформаційного пошуку.

ЛІТЕРАТУРА

1. Жуковська, В. В. Вступ до корпусної лінгвістики: навчальнийпосібник / В. В. Жуковська. Житомир: Вид-во ЖДУ ім. І. Франка, 2013. 142 с.

2. Жуковська, В. В. Корпусний підхід у навчанні та вивченні англійської мови // Формування компетентності «Навчання впродовж життя» як ключової вимоги професійної підготовки вчителя XXI століття / навчально-методичний

посібник / Гирин О.В., Євченко В.В., Жуковська В.В, Калініна Л.В., Савчук І.І., Прокопчук Н.Р., Щерба Н.С. та ін. Житомир, 2018. 168 с.

3. Коцюк, Л. М., Коцюк, Ю. А. Класифікаційна парадигма корпусу текстів за особливостями його дизайну, структури та способами використання, а також способом фіксації та індексації текстових даних. Наукові записки Національного університету «Острозька академія»: серія «Філологія». Острог: Вид-во НаУОА, 2020. Вип. 9(77). С. 106–110.

4. Могельницька, Л. Ф. Реалізація концепту загроза в мові та мовленні. Наукові записки Національного університету «Острозька академія»: серія «Філологія». Острог : Видво НаУОА, 2018. Вип. 1(69), ч. 2, березень. С. 35–37.

5. Фокін, С. Корпуси текстів: здобутки України та перспективи врахування закордонного досвіду // Вісник національного університету імені Тараса Шевченка. Літературознавство. Мовознавство. Фольклористика. 1 (28)/2018. С. 51 – 54.

6. Широков, В.А. Корпусна лінгвістика / В.А. Широков, О.В. Бугаков, Т.О. Грязнухіна та ін. К.: Довіра, 2005. 471 с.

7. American National Corpus. – Режим доступу: http://americannationalcorpus.org/

8. British National Corpus. – Режим доступу: http://www.natcorp.ox.ac.uk/

9. Corpus de Référence du Français parlé – Режим доступу: http://sites.univprovence.fr/delic/corpus/index.htm

10. DWDS-Corpus – Режим доступу: http://www.dwds.de/pages/pages_textba/dwds_text ba.htm