

ПЕРСПЕКТИВИ РОЗВИТКУ ПАРАЛЕЛЬНИХ КОРПУСІВ

У цій статті ми розглянемо перспективи розвитку паралельних корпусів і роль Sketch Engine у вивченні та підвищенні якості перекладів. Корпусна лінгвістика є однією з найбільш інноваційних і перспективних галузей мовознавства, яка дає змогу аналізувати великі масиви текстової інформації для вивчення мови та культури. Завдяки сучасним інструментам, таким як Sketch Engine, ми можемо здійснювати аналіз паралельних корпусів, виявляти ключові слова, терміни та n-грами, що допомагає перекладачам знайти оптимальний варіант перекладу та покращити якість перекладу в цілому.

Ключові слова: корпусна лінгвістика, паралельний корпус, Sketch Engine, n-грами.

Corpus linguistics is one of the most innovative and promising branches of linguistics, which allows analyzing large arrays of textual information for the study of language and culture. Thanks to modern tools such as Sketch Engine, we can analyze parallel corpora, identify keywords, terms and n-grams, which helps translators find the optimal translation option and improve the quality of the translation as a whole. In this article, we will consider the prospects for the development of parallel corpora and the role of Sketch Engine in studying and improving the quality of translations.

Key words: corpus linguistics, parallel corpus, Sketch Engine, n-grams.

Корпусна лінгвістика як напрям прикладної лінгвістики означає загальні принципи побудови лінгвістичних корпусів даних із використанням сучасних комп'ютерних технологій,

окреслює методика відбору текстів та шляхи їх збереження та аналізу [1, с. 4].

Один із напрямів корпусної лінгвістики — створення і використання паралельних корпусів, що застосовуються для вирішення різноманітних завдань, таких як створення і налаштування систем машинного перекладу, порівняльне вивчення мов, розвиток теорії перекладознавства, навчання мов.

Українські дослідники визначають явище паралельних корпусів як «єдність підмножини оригінальних текстів та підмножини їх перекладів на іншу(і) мову(и)» [3, с. 90]; «корпус, який складається як мінімум з двох підкорпусів, один з яких є вихідним, а інший містить тексти-переклади вихідного корпусу» [2, с. 136]. Всі ці визначення не суперечать одне одному, а описують новий лінгвістичний феномен із різних боків.

Перекладачі можуть використовувати корпуси паралельних текстів для швидкого аналізування мовних одиниць і варіантів перекладу, використовуючи великий обсяг текстової та екстра-лінгвістичної інформації. Якість перекладу залежить від розуміння предметної області, відмінного знання мови, на яку перекладається, та досконалого володіння мовою оригіналу.

Згідно з Л. М. Томіленко, формування корпусів має такі етапи:

- 1) проектування;
- 2) забезпечення надходження текстів із відповідних джерел; підготовка технологічного опису;
- 3) перетворення текстів на зчитувану машиною форму;
- 4) конвертування та передоброблення текстів;
- 5) графематичний аналіз (токенізація);
- 6) метарозмітка;
- 7) лінгвістична розмітка (оскільки наявність різних типів розмітки дозволяє користувачам використовувати корпус як інформаційно-пошукову систему для розв'язання практичних завдань);
- 8) коригування результатів автоматичної розмітки;

9) завантаження розмічених текстів до структури корпус-менеджера; забезпечення доступу до корпусу (пошук); та створення документації для корпусу.

Sketch Engine є інструментом для проведення корпусних досліджень, які зазвичай здійснюються на базі корпусів і великих електронних колекцій текстів.

У користувача є можливість використовувати широкий спектр інструментів для аналізу текстів, починаючи від простого пошуку слова у тексті і закінчуючи використанням спеціальних фільтрів для пошуку пропозицій за певною схемою.

Інструменти Sketch Engine. Наприклад, інструмент «Word Sketch» шукає слова і словосполучення, котрі задає користувач, оцінює частоту їх появи і показує, в яких контекстах зустрічаються задані користувачем слова у корпусі «Concordance», дозволяє вивчати контекст вживання слова, фрази або конструкції в корпусі текстів. Корконкорданс дає можливість швидко знайти всі входження слова (або фрази) в текстах корпусу, а також вивчити їх контекст та конкорданс. Крім того, користувач може виконати розширений пошук і дослідити, як вживається дане слово в різних лінгвістичних контекстах, які сполучення частіше за все вживаються разом з цим словом, як воно вживається в певних стилістиках мовлення тощо. Інструмент «Concordance» допомагає вивчити структуру та вживання мови в реальних текстових матеріалах та зробити висновки про лінгвістичні та культурні особливості мови.

У програмі Sketch Engine доступні спеціальні інструменти, які використовують дистрибуцію лексем у корпусі для визначення рівня синтагматичних і парадигматичних зв'язків. Ці інструменти включають тезаурус, кластеризацію та диференціацію [8].

Інструмент KEYWORD у Sketch Engine використовують для порівняння корпусів і визначення унікальних та типових елементів. Для цього вибраний корпус порівнюють із довідковим,

щоб визначити такі ключові показники, як ключові слова, терміни та N-грамові ключові вирази. При цьому беруть до уваги лише ті елементи, які зустрічаються у вибраному корпусі частіше, ніж у довідковому. Інструмент може залучати будь-який токен, а також використовувати типовий формат термінології у мові корпусу для визначення ключових багатослівних виразів.

Паралельний корпус текстів дає змогу не лише порівнювати оригінальний текст з його перекладом, а й, навпаки, порівнювати переклад з оригінальним текстом. Довгий час цей аспект процесу перекладу залишався недостатньо вивченим, проте глибоке дослідження цих питань може допомогти краще зрозуміти процес перекладу з точки зору психології, когнітивістики, лінгвістики, кібернетики та інших наукових дисциплін.

Список використаних джерел

1. Leonidovich M. S. Genre divisions in official discourse and some linguistic features of each genre. *The 7 th International scientific and practical conference "Science, innovations and education: problems and prospects"* (February 9–11, 2022). CPN Publishing Group, Tokyo, Japan, 2022. P. 628 (67).
2. Anderman G., Rogers M. The Linguist and the Translator. *Incorporating Corpora. The Linguist and the Translator*. Clevedon : Multilingual Matters Ltd, 2008. Pp. 5–17.
3. MCenery T., Xiao R. Parallel and Comparable Corpora: What is Happening? *Incorporating Corpora. The Linguist and the Translator*. Clevedon : Multilingual Matters Ltd, 2008. Pp. 18–31.