Гавриленко В. М., аспірантка
Демиденко О. П., к. пед. н., доц.
Національний технічний університет України
«Київський політехнічний інститут ім. Ігоря Сікорського»

СПЕЦИФІКА ВИКОРИСТАННЯ КОРПУСНОГО МЕТОДУ У РОБОТІ З ДВОМОВНИМИ ТЕРМІНОЛОГІЧНИМИ ГЛОСАРІЯМИ

Словниковий запас мови сьогодні росте з експоненціальною швидкістю, що обумовлено рядом факторів, серед яких збільшення кількості міждисциплінарних досліджень, загальна глобалізація наукового процесу, використання сучасних інформаційних технологій. Враховуючи величезні об'єми інформації, які потребують обробки, з'являються нові методи їх аналізу.

В дослідженнях термінології перспективними є методи, розроблені в межах корпусної лінгвістики [3, с.146]. В їх основі лежить робота зі значними за обсягом текстовими масивами – корпусами текстів, в яких здійснюється автоматизована вибірка слів за заданими параметрами, їх кількісна і якісна обробка. Під корпусом текстів мається на увазі значний масив мовних даних, створений для вирішення конкретних лінгвістичних завдань [2, с. 3].

Методи корпусної лінгвістики є оптимальними для побудови глосаріїв і словників, оскільки дозволяють виділити з текстів лексичні одиниці, які відповідають певним параметрам. А за допомогою інструментів типу Term Extraction і Alignment можна створювати дво-і навіть трьохмовні словникові добірки. Двомовний корпус, що складається з текстів, буде містити терміни, типові для конкретної предметної галузі, які в процесі роботи з текстом можна виділити, проаналізувати і зберегти [4]. Найзручнішими для створення двомовного глосарію є так звані паралельні корпуси (Parallel Corpora). В них співставляються тексти-оригінали і тексти перекладу, вирівняні (aligned) по реченням (або сегментам) [1, с. 8].

Під час роботи з майбутнім глосарієм слід враховувати ряд особливостей. В роботі з англомовними текстами необхідно виконати розмітку (морфологічну/синтаксичну або частиномовну (POS-tagging) розмітку), щоб виключити з добірки клішовані виразити, артиклі, слова-зв'язки, але зберегти у добірці слова і вирази, які, наприклад, вживаються з прийменниками. Після розмітки програма виконує токенізацію корпусу – виділення одиниць, які відповідають параметрам заданим у програмі (токен – одиниця, відокремлена у тексті пробілами). Після цього вручну виконується вичитка, під час з якої з частотної вибірки необхідно прибрати «негативний словниковий матеріал», коли з готової частотної вибірки прибираються залишки – допоміжні дієслова, сполучники, вигуки, власні назви (якщо вони не є термінологічними абревіатурами) та їх похідні. Крім цього в процесі проводиться очищення вибірки від словоформ. В кінці з багатомільйонного текстового масиву можна отримати термінологічний глосарій розміром у декілька десятків тисяч слів.

Ще однією особливістю роботи з двомовними термінологічними глосаріями може міждисциплінарність текстового масиву. В такому випадку при вирівнюванні текстів і подальшому застосуванні корпусного методу може виникнути проблема різночитання терміну (якщо тексти не були уніфіковані або якщо значення терміну змінюється в залежності від контексту).

Отже, можна зробити висновок про доцільність використання методів корпусної лінгвістики у процесі створення двомовних глосаріїв, які в свою чергу можуть послужити основою для укладання лексикографічних джерел у обраній предметній області.

Література

1. Беляева Л. Н. Лингвистические технологии в современном сетевом пространстве: language worker в индустрии локализации. СПб. : Книжный дом, 2016. 134 с.

2. Захаров В.П. Корпусная лингвистика: учебно-метод. пособие. СПб., 2005. 48 с.

3. Саєнко Н. С. Корпусний підхід у навчанні іноземних мов у технічному університеті // Педагогічні науки: теорія, історія, інноваційні технології. 2016. № 1. С. 142-151.

4. Khurshid A., Rogers M. Terminology management: a corpus-based approach // Translating and the Computer. 1992. URL : http://www.mt-archive.info/90/Aslib-1992-Ahmad.pdf.