УДК 004.415.5
А. Луцків, В. Михалюк
(Тернопільський національний технічний університет імені Івана Пулюя)

АЛГОРИТМІЧНЕ ТА ПРОГРАМНЕ ЗАБЕЗПЕЧЕННЯ ПОБУДОВИ КОРПУСІВ МОВ CAT-ЗАСОБІВ

На сьогоднішній день, коли поглиблюється співпраця України із міжнародною спільнотою, вкрай важливим є подолання мовних бар'єрів, задля простішого доступу українців до інформації, що існує лише в іншомовному вигляді. Найефективнішим способом вирішення цієї проблеми є використання CAT (Computer Assisted Transltion). CAT, або автоматизований переклад – це такий вид перекладу, при якому людина-перекладач використовує спеціалізовані комп'ютерні програми і засоби для підтримки і спрощення процесу перекладу. Відомими CAT-системами на даний час є SDL Trados, memoQ, Crowdin, OmegaT та інші. CAT-системи зазвичай поєднують у собі такі програмні засоби: пам'ять перекладів, перевірки написання та граматики, глосарії, електронні словники, термінологічні бази, засоби пошуку по тексту, конкордансери та бітекстові інструменти. Для більшості цих засобів корисно використовувати методи корпусної лінгвістики. Корпус текстів – це збірка природніх текстів спільної тематики, що зберігаються у цифровому вигляді для зручних аналізу та обробки комп'ютерними програмами. Прикладами корпусів мов є Британський національний корпус, паралельний арабо-англомовний корпус новин, Міжнародний корпус англійської мови початківців тощо. З україномовних корпусів варто відзначити Корпус текстів української мови, а також корпус текстів електронної бібліотеки Чтиво.

При формуванні корпусів слід враховувати такі фактори:

- **Розмір** – чим більший корпус, тим він більш інформативний, проте у деяких випадках корисніше використовувати менші корпуси, щоб не мати справи з надмірністю даних (наприклад, для дослідження прийменників).
- **Збалансованість** – тексти у корпусі повинні бути збалансованими по довжині, типу, стилю, щоб уникнути виведення неправильних висновків.
- **Репрезентативність** – корпус є репрезентативним, якщо результат його дослідження буде справедливим і щодо усієї мови (або окремої її частини).

Основні етапи формування корпусу:

1. **Захоплення даних.** Проводиться збір усіх текстів, які треба внести до корпусу.
2. **Початкова перевірка та перетворення.** Тексти перевіряються на можливість внесення, та перетворюються у зручний формат.
3. **Лінгвоанотація**. Кожне слово у тексті означається його лінгвістичними властивостями (член речення, частина мови, склад слова).
4. **Каталогізація текстів та остаточна перевірка.** Тексти, що пройшли анотацію, конвертуються у остаточний формат, до кожного тексту формується заголовок, текст вноситься до корпусу.

Література

1. Corpus building and investigation for the Humanities [Електронний ресурс] // University of Nottingham – Режим доступу до ресурсу: https://www.birmingham.ac.uk/Documents/college-artslaw/corpus