DOI https://doi.org/10.30525/978-9934-26-110-7-57

ОСНОВНІ ПІДХОДИ ДО ЕКСТРАЛІНГВІСТИЧНОЇ РОЗМІТКИ КОРПУСУ ТЕКСТІВ

Шкляревський В. Г.
інженер-програміст лабораторії комп'ютерної лінгвістики, здобувач кафедри англійської філології, перекладу і філософії мови імені професора О. М. Мороховського
Київський національний лінгвістичний університет
м. Київ, Україна

Предметом корпусної лінгвістики є створення та розробка загальних принципів побудови й використання корпусів текстів. Не викликає сумніву той факт, що для подальшого застосування в лінгвістичних дослідженнях тексти корпусу мають бути анотовані за певною системою.

Метарозмітка, або метаопис – це система структурованих даних екстралінгвістичного характеру (метаданих), що стосуються формальної сегментації і зовнішнього анотування тексту, а також фіксації технологічної обробки електронного файлу [3, с. 3]. Без такої інформації про тексти, що входять до складу корпусу, використання ресурсу в лінгвістичних дослідженнях і програмних застосуваннях є неможливим. Кодування, редагування й зберігання метаданих виділяють як окремий етап розробки корпусу поряд із плануванням, збиранням текстів і лінгвістичною розміткою.

Основним функційним призначенням метарозмітки є інформування користувача про тексти на предмет їхнього авторства, стилістично-жанрової специфіки, тематики, дати і місця написання [2, с. 89]. Отже, виділяють декілька видів метарозмітки: 1) зовнішня, або **екстралінгвістична розмітка** – бібліографічні, типологічні, тематичні і соціологічні характеристики; 2) формальна, або **структурна розмітка** текстів на частини; 3) **технологічна розмітка** – дані кодування, дати, виконавці, джерело електронної версії [3, с. 7]. У статті обговорюються основні підходи до екстралінгвістичної розмітки текстового масиву. На сьогодні у практиці корпусної лінгвістики існують два підходи до метарозмітки залежно від мети розробки та параметрів корпусу, яким відповідають два види набору метаданих: 1) стандартний, загальний метаопис корпусу [2, с. 89; 5; 9] і спеціалізований метаопис, спрямований на структуризацію метаданих корпусу, розробленого для розв'язання конкретних завдань [6, p. 35; 4, с. 30; 1, с. 224].

Загальний набір метаданих регулюється стандартами й рекомендаціями Text Encoding Initiative (TEI) [7], Open Language Archive Community (OLAC) [8], EAGLES [9], ISLE Metadata Initiative (IMDI) [10] і відповідає загальноприйнятій практиці кодування текстів. Стандартний метаопис застосовується в основному для універсальних національних корпусів, розроблених на засадах репрезентативності, збалансованості, автентичності, комп'ютерної підтримки, документованості та стандартності (BNC, ANC, НКРЯ). Орієнтація на якісно різного користувача національних корпусів потребує адаптації мовнонезалежного підходу кодування даних, забезпечуваного міжнародними стандартами та документами, головним чином принципів TEI [2, с. 7]. Зазвичай стандарти включають метадані про розробника корпусу, джерела, термін створення, мову текстів, формат файлів, авторські права.

Набір даних спеціалізованого метаопису мотивується передбачуваним призначенням корпусу, згідно з яким створюють індивідуальні набори метаданих або вносять до стандарту додаткові ознаки. Так, метарозмітка Корпусу українських текстів для вивчення граматичної службовості включає хронологічний параметр в аспекті п'яти періодів – з українсько-руського X-XV ст. до новітнього українського – з 1991 р., що дозволяє використовувати корпус для діахронічних досліджень [1, с. 227]. Метаопис спеціального навчального корпусу текстів – Ukrainian Corpus of Learner English (UCLE) забезпечує дослідження актуального словника студентів за лексичною різноманітністю з урахуванням параметрів статі, віку, рідної мови, курсу навчання й факультету [4, с. 30–31]. Метарозмітка корпусу текстів з комп'ютерної лінгвістики крім вихідних даних включає окреме кодування анотації тексту, що дає можливість у майбутньому створити на базі корпусу автоматичну систему реферування фахових текстів [6, p. 36]. Отже, встановлений набір метаданих поряд із функціями пошукового механізму й лінгвістичною розміткою визначає функційні можливості використання корпусу і його сумісність з іншими програмними продуктами.

Література:

1. Данилюк І. Корпус текстів для вивчення граматичної службовості. Лінгвістичні студії, 2013. № 26. С. 224–229.

2. Демська-Кульчицька О. Основи національного корпусу української мови. Національної академії наук України. Київ, 2005. 219 с.

3. Захаров В. П. Корпусная лингвистика: Учебно-метод. пособие. Санкт-Петербург: СПб, 2005. 48 с.

4. Коломієць В., Котик С. Спеціальний навчальний корпус текстів UCLE: сучасний стан і перспективи використання. Комп'ютерна лінгвістика: сучасне і майбутнє. Матеріали Міжнародної науково-практичної конференції. Київ: КНЛУ, 23-24 лютого, 2012. С. 29–32.

5. Савчук С. О. Метатекстовая разметка в Национальном корпусе русского языка: базовые принципы и основные функции. Национальный корпус русского языка: Результаты и перспективы. Москва, 2005. С. 62–88.

6. Bobkova T. Corpus of computational linguistic texts. Computer Treatment of Slavic and East European Languages. Bratislava: Tribun, 2009. P. 35–40.

7. Burnard L. Metadata for Corpus Work. URL: http://users.ox.ac.uk/~martinw/dlc/chapter3.htm (дата звернення: 21.06.2021)

8. Simons G. OLAC Metadata Usage Guidelines. URL: http://www.language-archives.org/NOTE/usage.html (дата звернення: 21.06.2021)

9. Sinclair J. Preliminary recommendations on text typology. EAGLES Document EAG-TCWG-TTYP/P, 1996. URL: http://www.ilc.cnr.it/EAGLES96/texttyp/texttyp.html (дата звернення: 21.06.2021)

10. Wittenburg P. Metadata Proposals for Corpora and Lexica. LREC: Max-Planck-Institute for Psycholinguistics, 2002. URL: http://www.mpi.nl/IMDI/documents/2002%20LREC/Metadata%20Proposals%20for (дата звернення: 21.06.2021)