Наталя Лотоцька
(наук. керівник: к.техн.наук, доц. І. М. Кульчицький)
Національний університет «Львівська політехніка»

МЕТАТЕКСТОВЕ РОЗЗНАЧЕННЯ ТЕКСТІВ (НА МАТЕРІАЛІ ТВОРІВ РОМАНА ІВАНИЧУКА)

У статті розглянуто поняття та роль метатекстового роззначення у корпусі текстів (КТ), окреслено стандарти метаопису текстів та запропоновано інформаційні параметри й позначки для опису позамовної інформації у КТ.

Ключові слова: корпус текстів (КТ), роззначення, метатекстове роззначення, метаопис, інформаційні параметри .

Постановка проблеми. Укладання і використання корпусів текстів – одна з провідних тенденцій світового мовознавства, що ґрунтується на посиленому використанні інформаційних технологій. Для надання доступу користувачам, опрацювання широкого спектру текстової інформації, багаторазового використання і забезпечення сумісності корпусу з іншими корпусами виникає потреба у формулюванні відповідного стандартного оформлення текстів. Роззначення – головна характеристика корпусу, що відрізняє його від простих колекцій текстів, оскільки надає особливу додаткову інформацію про властивості текстів, які входять до корпусу. А власне метатекстове роззначення слугує описом позамовних властивостей тексту, дає змогу виявити фактори впливу на словесну тканину текстів та залежність мовних характеристик текстів від середовища їх побутування.

Аналіз останніх досліджень і публікацій. У сучасному мовознавстві спостерігаємо стрімкий розвиток корпусних студій.

Огляд зарубіжних публікацій про корпусну лінгвістику, історію її становлення, подано у працях В. Широкова, О. Демської, В. Жуковської, О. Ванівської [2: 46; 3: 7-29; 11: 70]. Висвітлення того, чим є корпуси текстів, їх визначальні риси, класифікація, методи та галузі використання корпусів, способи та галузі їх застосування подано у працях В. Широкова, А. Загнітка, Н. Дарчук, О. Демської, В. Жуковської [9: 11-17; 2: 45-46; 3: 82-142].

У своїй праці О. Демська [3] досліджує текстовий корпус національної мови, передумови, засади планування та укладання корпусів, зокрема корпусу сучасної української мови. Дослідження Н. Дарчук [2] висвітлює основні засади та перспективи дослідницького корпусу української мови. Монографія «Корпусна лінгвістика» В. Широкова [9] присвячена найбільшому в Україні Українському національному лінгвістичному корпусу (УНЛК). У монографічному дослідженні І. Кульчицького [10] опрацьовано технічні аспекти підготовки текстів до залучення їх до корпусу.

Мета статті – дослідити поняття та значення метатекстової розмітки, визначити стандарти (параметри) метатекстового роззначення як невід'ємної частини укладання корпусів текстів. Для досягнення поставленої мети необхідно розробити систему позначок для опису метатекстової (позамовної) інформації.

Виклад основного матеріалу дослідження. Вивчення мовних явищ із використанням КТ уможливлює не тільки підтвердження чи спростування окремих аспектів теорії, а й дозволяє здійснити відкриття невідомих закономірностей у функціонуванні мовних одиниць різних рівнів.

За І. Кульчицьким, КТ – це призначена для різноманітних лінгвістичних досліджень лінгвістично описана сукупність мовних чи мовленнєвих даних, які подані в електронній формі та

лінгвістику, історію її
кова, О. Демської,
3: 7-29; 11: 70].
їх визначальні риси,
корпусів, способи та
ціях В. Широкова,
В. Жуковської [9:

є текстовий корпус
вання та укладання
української мови.
сновні засади та
країнської мови.
кова [9] присвячена
національному
чному дослідженні
спекти підготовки

ення метатекстової
) метатекстового
корпусів текстів.
розробити систему
нформації.
кення. Вивчення
ивлює не тільки
ектів теорії, а й
кономірностей у

ля різноманітних
укупність мовних
онній формі та

споряджені відповідним спеціалізованим програмним забезпеченням [10: 29].

Одним із важливих кроків укладання КТ є роззначення корпусу – надання додаткової інформації про тексти в корпусі. Роззначення надає особливу додаткову інформацію про властивості текстів, які входять до корпусу, вона дозволяє швидко та ефективно знайти ті слова, форми та конструкції, які необхідні дослідникові. Роззначення полягає у приписуванні текстам і їх компонентам спеціальних позначок: зовнішніх, або екстралінгвістичних (відомості про автора й відомості про текст); структурних (розділ, абзац, речення, словоформа) власне лінгвістичних, що описують лексичні, граматичні та інші характеристики елементів тексту.

Лінгвістичне роззначення передбачає характеристику тексту з позицій морфології, синтаксису, семантики, просодії тощо; структурне – відображає інформацію про структуру тексту, дає змогу відокремити одне слово від іншого, виділити розділові знаки, окремі слова, межі словосполучення, речення, абзацу, глави, розділу, заголовка та ін.; екстралінгвістичне (метатекстове) роззначення у традиційному розумінні – це його зовнішня характеризація, тобто така, що описує обставини створення тексту.

На сьогодні в корпусній лінгвістиці розроблено кілька систем стандартів метарозмітки корпусів, серед яких найчастіше використовуваним є Text Encoding Initiative (TEI). Рекомендації TEI передбачають багаторівневий склад метаданих [15]. За О. Демською, основним функціональним призначенням корпусних метаданих є інформування якісно різних користувачів про тексти на предмет їхнього авторства, стилістично-жанрової специфіки, тематики, дати і місця написання.

Потреба в екстралінгвістичному роззначенні зумовлена кількома чинниками. Перш за все, це великий обсяг корпусу

текстів. Можливість здійснення пошуку в корпусі за допомогою
заданих параметрів забезпечує швидкий вибір необхідних для
лінгвістичного дослідження підкорпусів. Крім того, метаописи
слугує інструментом формування архітектури корпусу, а також
дозволяє, з одного боку, характеризувати тексти статичних
корпусів, а з іншого, контролювати процес оновлення даних
динамічних корпусів із дотриманням вимог збалансованості й
репрезентативності.

Прикладом реалізації стандартів метатекстового розмічення
в українській корпусній лінгвістиці є два корпуси текстів
української мови [1; 2], тримовний ілюстративний Корпус текстів з
комп'ютерної лінгвістики [14], англомовний Корпус анотацій
наукових статей із комп'ютерної лінгвістики [4] і Багатомовний
паралельний корпус усного мовлення, укладений на базі субтитрів
до сучасних телевізійних серіалів [12]. Корпус текстів української
мови побудований таким чином, щоб користувач міг одержати
різноаспектну інформацію, що передбачає зовнішнє анотування
за бібліографічними параметрами.

Завданням нашого дослідження полягає у розробленні
технології формування корпусу текстів письменника Романа
Іваничука — створенні системи позначок для корпусу текстів з
метою здійснення його метатекстового розмічення. Розробляючи
систему позначок, ми оператимемося на міжнародні стандарти TEI
[15].

Щоб перевірити ефективність системи позначок, шість
романів, а саме «Червоне вино» [7], «Манускрипт з вулиці
Руській» [7], «Вогненні стовпи» [4], «Хресна проща» [6], «Вода з
каменю» [5] та «Сакура у піжмах» [5] було перетворено в
електронну форму та здійснено їхнє метатекстове розмічення.

Для кодування метаопису документа стандарт TEI
рекомендує засоби для опису таких інформаційних одиниць: автор,

назва твору, час та місце створення, час та місце видання, жанрова та стилістична належність твору, його обсяг, відповідальний за якість електронної копії та за виправлення різних неточностей, спосіб отримання електронної копії та ін. Обов'язковими є дві компоненти: опис безпосередньо тексту (автор, назва твору, його обсяг) та подання профілю тексту (час створення, цільова аудиторія, час створення та ін.), які разом компонують заголовок документа.

Загальними параметрами заголовка опису документа є: унікальний ідентифікатор тексту в корпусі; ім'я файлу, що містить документ; мова тексту; бібліографічний опис; дані про автора, загальний клас документа, його стиль, жанр та ін.; відомості про розмітку — або її структура, або поклик на стандартну; історія змін та модифікацій документа. Бібліографічний опис включає бібліографічні дані про текст, бібліографічні відомості про видання, дані про джерело електронної форми тексту.

Бібліографічні дані про текст містять такі складові, як назва тексту (за необхідності окремо повна, окремо коротка), автор тексту, обсяг у певних умовних одиницях (наприклад слова), відомості про інших, крім автора, осіб, які долучилися до створення документа (у вигляді «ім'я» — «спосіб доручення, наприклад «сканування»). Указують ім'я чи псевдонім автора (тип зазначають відповідним параметром) залежно від того, під яким іменем твір є відомий. Персональні дані про автора зазначають іншим параметром.

Відомості про видання документа щонайменше повинні містити такі складові: видавець редакції документа, яку включено в корпус, місце та рік видання. Якщо текст, доданий у корпус, узято із збірки, наприклад окрема новела, повість чи вірш, то вносять бібліографічний поклик на назву та редактора збірки. Якщо потрібно, можна вказати й іншу інформацію, зокрема

особливості конкретного видання твору, графіка чи гарнітура тексту, редакторські чи авторські коментарі, виправлення, примітки. Також можна додати довідкову інформацію про ISBN, ББК, УДК, власника авторських прав та ін.

Для кодування профілю документа передбачено такі параметри: дані про час і місце створення тексту, класифікаційні ознаки твору за різними класифікаційними схемами: система ключових слів, стандартні таксономії, власноруч розроблені схеми для конкретного корпусу; умови створення тексту — канал комунікації (усне чи писемне мовлення), тип документа (книжка, газета, факс, лист та ін.); фрагментаційні ознаки; відношення до інших текстів та ін.

Описуючи появу тексту, передусім необхідно зазначити час, місце, а також інші обставини створення тексту. Зазвичай достатньо зафіксувати рік створення або інтервал у роках, якщо точна дата невідома. У персональних даних автора доцільно подавати інформацію: стать, вік на момент написання твору, дата та місце народження (дата смерті здебільшого не має впливу на твір, хоча може інформувати видання, опубліковані за життя автора, чи неможливість наступної редакції), рідна мова автора та знання інших мов, місце перебування на момент створення тексту, рівень освіти, вид занять та соціально-економічний статус, який рекомендують описувати з дотриманням певних стандартів класифікації.

З огляду на те, що корпус творів Р. Іваничука має початкову стадію формування, пропонуємо для метаопису текстів спрощений перелік позначок, який повністю задовольняє сьогочасні потреби опрацювання творів письменника. Надалі перелік позначок необхідно буде розширити. Отже, спрощений метаопис творів Р. Іваничука компонують такі інформаційні параметри:

- TextTitle — назва твору;

- TextAuthor — автор твору (у нашому випадку — Роман Іваничук);
- TextStyle — стиль твору (художній, науковий, офіційно-діловий, публіцистичний, конфесійний, розмовний, епістолярний);
- TextKin — літературний рід твору (лірика, драма, епос);
- TextMode — жанр твору в межах роду (наприклад для епосу — епопея, казка, байка, легенда, оповідання, повість, роман, новела, художні мемуари);
- TextVariety — жанровий різновид у межах виду (наприклад для роману — історичний);
- TextSpecies — тип творення твору (оригінал/переклад);
- TextLanguage — мова оригіналу;
- TextTranslator — автор перекладу;
- PublishWay — спосіб видання (ціле/частина);
- PublicationType — тип видання (книжка, журнал, газета та ін.);
- CompleteTitle — назва цілого (для частин);
- PublicationCity — місто видання;
- Publisher — видавець;
- PublicationYear — рік видання;
- PublicationNumber — номер для періодичних видань;
- PageFirst — перша сторінка частини загалом, для цілого — кількість сторінок;
- PageLast — остання сторінка частини загалом.

Висновки і перспективи подальших розвідок у цьому напрямку. Запропоновані нами інформаційні параметри та позначки для опису метатекстової інформації дадуть змогу прослідкувати позамовні властивості тексту, виявити фактори впливу на словесну тканину текстів та залежність мовних

характеристик текстів від середовища їхнього існування, забезпечити швидкий вибір необхідних для лінгвістичного дослідження підкорпусів, а також представити архітектуру корпусу текстів.

Література

1. Данилюк І. Корпус текстів для вивчення граматичної службовості / І. Данилюк // Лінгвістичні студії. – Вип. 26. – Донецьк : ДонНУ, 2013. – С. 224–229. **2. Дарчук Н.** Дослідницький корпус української мови: основні засади і перспективи / Н. Дарчук // Вісник Київського національного університету імені Тараса Шевченка. Літературознавство, мовознавство, фольклористика . — 2010.— Вип. 21 .— С. 45-49. **3. Демська О. М.** Текстовий корпус: ідея іншої форми / О. М. Демська .— Нац. ун-т "Києв.-Могил. акад.". - К. : Вид. дім "Києво-Могилянська академія", 2011. - 284с. **4. Іваничук Роман.** Вогненні стовпи : тетралогія / Роман Іваничук .— Харків: Фоліо, 2011 .— 505с. **5. Іваничук Роман.** Вода з каменю ; Саксаул у пісках : романи / Роман Іваничук .— Харків : Фоліо, 2015 .— 412 с. **6. Іваничук Роман.** Хресна проща : романний триптих / Роман Іваничук .— Львів : Піраміда, 2011 .— 281с. **7. Іваничук Роман.** Черлене вино : романи / Роман Іваничук .— Львів : Каменяр, 1979 .— 350 с. **8. Коломієць В.**, Орел В. Корпус анотацій наукових статей із комп'ютерної лінгвістики / В. Коломієць, В. Орел // Комп'ютерна лінгвістика : сучасне і майбутнє. Матеріали Міжнародної науково-практичної конференції. – К. : КНЛУ, 2012. – С. 32–34. **9. Корпусна лінгвістика:** Моногр. / В. А. Широков, О. В. Бугаков, Т. О. Грязнухіна, О. М. Костишин, М. Ю. Кригін; НАН України, Укр. мов.-інформ. фонд. — К.: Довіра, 2005. — 472 с. **10. Кульчицький І. М.** Технологічні аспекти укладання корпусів текстів / І. М. Кульчицький // Дані текстових корпусів у лінгвістичних дослідженнях : монографія / В. А. Широков,

І. В. Шевченко, А. П. Загнітко та ін. ; за ред. О. П. Левченко .— Львів : Видавництво Львівської політехніки, 2015 .— С. 29-45. **11. Кульчицький Ігор.** Корпуси текстів як лінгвотехнологічне підґрунтя виявлення змін в українській мові // XX–XXI століття: жанрово-стильові й лінгвістичні метаморфози в українській мові та літературі: Монографія / А. Архангельська (гол. ред.), О. Левченко, О. Тищенко, та ін .— Оломоуць: Університет ім. Ф. Палацького, 2016 .— С. 269-298. **12. Лебедєв К.** Створення Багатомовного корпусу паралельних текстів / К. Лебедєв // Комп'ютерна лінгвістика: сучасне і майбутнє. Матеріали Міжнародної науково-практичної конференції. – К. : КНЛУ, 2012. – С. 36–37. **13. Лотоцька Н. Я.** Структурне роззначення речень у корпусі текстів Р. Іваничука / Н. Я. Лотоцька // Збірнику наукових праць «Актуальні проблеми філології та перекладознавства» Хмельницького національного університету, – 2017. – Вип. № 12. – С. 147-151. **14. Bobkova T.** Corpus of computational linguistic texts / T. Bobkova // Computer Treatment of Slavic and East European Languages. – Bratislava: Tribun, 2009. – P. 35–40. **15. TEI:** Text Encoding Initiative [Електронний ресурс]. — Режим доступу : http://www.tei-c.org/index.xml.

Данило Островський

(наук. керівник: к. філол. н., доц. Г. В. Ситар)

Донецький національний університет імені Василя Стуса

БАЗА ДАНИХ «СПОСОБИ ВІДТВОРЕННЯ АНГЛІЙСЬКИХ ВЛАСНИХ НАЗВ УКРАЇНСЬКОЮ ТА РОСІЙСЬКОЮ МОВАМИ» (НА МАТЕРІАЛІ РОМАНУ ДЖ. К. РОЛІНҐ «ГАРРІ ПОТТЕР І ФІЛОСОФСЬКИЙ КАМІНЬ»)