УДК 81'33

В. Г. Шкляревський,
Київський національний лінгвістичний університет, м. Київ

МЕТАРОЗМІТКА КОРПУСІВ ТЕКСТІВ: СТАНДАРТИ І РЕАЛІЗАЦІЯ

У статті досліджуються основні стандарти представлення метаданих від перших і до сучасних корпусів текстів. Аналізується сучасний стан розробки проблеми метарозмітки корпусів і вимоги до метаданих. Окреслюються принципи метарозмітки, реалізовані у вітчизняних корпусах англомовних текстів.
***Ключові слова**: корпус текстів, розмітка, метарозмітка, метамова.*

В статье исследуются основные стандарты презентации метаданных от первых и до современных корпусов текстов. Анализируется современное состояние разработки проблемы метаразметки корпусов и требования к метаданным. Формулируются основные принципы, реализованные в отечественных корпусах англоязычных текстов.
***Ключевые слова**: корпус текстов, разметка, метаразметка, метазык.*

This paper considers the main trends of corpora data presentation for encoding from the very first corpuses till modern corpora projects. The state of art in elaborating corpus encoding issues and requirements to the encoding data are analyzed. The main principles of encoding used in Ukrainian English-language corpuses are outlined.
***Keywords**: corpus, markup, text encoding, metalanguage.*

Актуальність аналізу стану розробки та застосування розмітки і метарозмітки різноманітних корпусів текстів пояснюється широким використанням корпусів у різних галузях лінгвістики, включаючи її найновітніші напрями. Із розвитком корпусної лінгвістики виникає потреба у формулюванні чіткіших вимог до метаданих, задіяних у корпусах, необхідність побудови класифікацій параметрів розмітки для опрацювання широкого спектру текстової інформації. Метарозмітка, або інтелектуальна розмітка, як опис позамовних властивостей тексту дає змогу не лише виявити фактори впливу на словесну тканину текстів, виявити залежність мовних (лінгвальних) характеристик текстів від середовища їх побутування, а й окреслити тенденції розбудови тієї чи іншої дисципліни за допомогою аналізу відповідних фахових текстів. Так, одним з передбачуваних результатів нашого дослідження метаданих англомовних текстів з лінгвокогнітології є систематизація основних чинних та прогнозованих напрямів розвитку цієї дисципліни. Встановлення типології метаданих сприяє запобіганню дослідницьких непорозумінь у процесі зіставлення і опрацювання корпусів різних текстів.

Мета статті – визначити стандарти метарозмітки як невід'ємної частини укладання корпусів текстів та окреслити шляхи їх дотримання. Досягнення поставленої мети передбачає виконання таких **завдань**: 1) визначити передумови формування стандартів метарозмітки; 2) окреслити етапи еволюції стандартів метаданих від перших корпусів текстів до новітніх; 3) простежити реалізацію стандартів метарозмітки в англомовних корпусах, розроблених в Україні.

На сучасному етапі розвитку інформаційних технологій застосування комп'ютерів стає все інтенсивнішим в усіх сферах діяльності людини. Зростання потреб суспільства в пошуку й опрацюванні інформації зумовило активне впровадження комп'ютерних технологій у лінгвістичних дослідженнях. Укладання і використання корпусів текстів – одна з провідних тенденцій світового мовознавства – ґрунтується на посиленому використанні інформаційних технологій. Указана тенденція зумовлена максимальною точністю і об'єктивністю наукових розвідок, які здійснюються із використанням комп'ютерів. Наступним кроком після укладання є розмітка корпусу – надання додаткової інформації про тексти корпусів й анотування текстових елементів. Ця інформація може бути екстралінгвістичною та лінгвістичною [1, с. 87], тобто такою, що стосується позамовних (екстралінгвальних) і власне мовних (лінгвальних) тексту. Якщо лінгвістична розмітка передбачає характеристику тексту з позицій морфології, синтаксису, семантики, просодії тощо, екстралінгвістична розмітка (метарозмітка) у традиційному розумінні – це його зовнішня характеризація, тобто така, що описує обставини створення тексту. Всі «нелінгвістичні» елементи метарозмітки можна поділити на елементи «формальної» структурної розмітки (текст, розділ тощо) і на елементи «зовнішньої», екстралінгвістичної метарозмітки, які характеризують текст в цілому – інформація про автора, умови створення текстів, інформація про тематику, стиль, жанр текстів [1, с. 87].

Потреба в екстралінгвістичній розмітці 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, метарозмітка 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.

Формування стандартів метарозмітки починається з середини 1960-х років після появи першого корпусу текстів – Браунівського корпусу в 1963 році. На цьому етапі розвитку корпусної лінгвістики для метарозмітки використовується термін «описова розмітка» (descriptive markup) [11, p. 219]. На той час проблеми метарозмітки корпусів текстів досліджувалися в аспекті загального опису письмових текстів. Метарозмітка здійснювалася за допомогою макросів – певної короткої команди, яка передбачала реалізацію великої кількості інструкцій нижчого рівня. Основним параметром опису тексту була його формальна композиційна структура: виділення частин тексту, назви, рядка тощо. Ця розмітка на початку використовувалась у сфері книговидавництва, в процесі використання систем створення текстів. Праці Уильяма Таникліфа, Стенлі Райса й Чарльза

Голдфарба [11, p. 220] про створення системи позначень розмітки (тегів) та мови розмітки наприкінці 1960-х років стали передумовами формування стандарту машиночитаних визначень мови описової розмітки SGML. Окрім вищевказаних проектів до перших систем розмітки належать the Brown University PRESS hypertext system, проект Брайана Райда SCRIBE й розробки опрацювання текстів Д. Енгельбарта.

Цей досвід надихнув дослідників на створення єдиного середовища для описової розмітки як підготовчого етапу здійснення метарозмітки. Частково результати досліджень були опубліковані в теоретичних працях К. Голдфарба, Б. Райда на початку 1980-х років [11]. Однак більшість документації проекту розробки систем розмітки залишилася неопублікованою, з'являючись лише в технічних документах до експериментальних систем. На цьому етапі так звана описова розмітка, результатом якої було визначення архітектури тексту, мала велику кількість переваг для широкого спектру діяльності, об'єктом якої був текст, зокрема, редагування тексту, накопичення інформації про стандарти оформлення тексту залежно від стилю, жанру і подальше здійснення метарозмітки тексту.

Автоматичне редагування тексту під час написання здійснювалась шляхом застосування процедури розпізнавання типу тексту і завдяки відповідному типу форматування. Структурнозорієнтоване редагування запобігало появі помилок у побудові текстів. Крім того, файли із розміченими текстами були оптимальними для їх перенесення до систем опрацювання текстів, створених на базі іншого програмного забезпечення. Здійснення описової розмітки як найбільш оптимального підходу в дослідженнях текстів забезпечується завдяки визначенню тексту як впорядкованої ієрархії об'єктів змісту [7]. Отже, описова розмітка мала на меті виключно композиційну характеристику текстів і стала проміжним етапом між лінгвістичною розміткою і метарозміткою.

Наступним етапом становлення апарату метарозмітки стало створення метамови опису – інструменту метарозмітки. Розробка метамови як мови для характеристики машиночитаних визначень мов описової розмітки було метою розпочатого наприкінці 1960-х років в Асоціації візуальних комунікацій проекту GenCode, очолюваного Н. Шарпфом. Пізніше, проект було передано до Інституту американських національних стандартів, де його очолив Ч. Голдфарб. Спроба створити метамову на базі GML завершилася презентацією першого варіанту SGML у 1980 році. Співпраця з Міжнародною організацією стандартизації (ISO) ознаменувалася публікацією ISO 8879 в 1986 році: «Information Processing – Text and Office Systems – Standard Generalized Markup Language (SGML)» [10]. У такий спосіб було завершено формування і засад здійснення метарозмітки корпусів текстів в світовій практиці.

Середина 1990-х років характеризувалася наявністю кількох стандартів метамови і метарозмітки корпусів текстів. Ці стандарти у вдосконаленому варіанті використовуються і сьогодні. Варто зазначити, що метамова, така як SGML (Standard Generalized Markup Language) чи XML, не є набором тегів розмітки документу для опису його елементів. Метамова використовується лише для опису кодів документа. Отже, такі метамови, як SGML чи XML, протиставляються власне мовам розмітки, таким як HTML, XHTML, TEI, DocBook.

Укладання словника мови розмітки, який містить стандартний набір кодів для опису елементів текстів, видається надзвичайно складним завданням через альтернативні класифікації текстів (використання відмінної термінології, розбіжності у сприйнятті характеристик текстів), множинність мов розмітки текстів. Тому стандартизація метамови розмітки відіграє ключову роль у забезпеченні інтероперабельності комп'ютерних програм і вирішує питання створення єдиної мови розмітки для індустрії книговидавництва [11, p. 230]. За цим підходом мови розмітки, розроблені відповідно до стандартів SGML, можуть бути розпізнані програмним забезпеченням, що підтримує SGML. Це дає змогу майбутнім незалежним розробникам працювати над мовою розмітки без узгодження її параметрів з розробниками програмного забезпечення за умови, що ця мова відповідає стандартам SGML, а програмне забезпечення підтримує SGML.

Невдовзі після проекту SGML розпочалася робота над визначенням основних правил форматування і створення мов метарозмітки текстів. Найбільш визначними для формування сучасних стандартів метамови стали такі проекти створення метамов: Document Style and Semantics Specification Language (DSSSL), The World Wide Web Consortium (W3C)'s Extensible Stylesheet Language (XSL), що має в своїй основі DSSSL, і W3C Cascading Style Sheet зі схожими, проте більш обмеженими функціями.

Розробники SGML представляють свій стандарт як набір понять і положень. До цього набору належать, перш за все, поняття елементу та його типу [10]. Елементом вважається певний компонент документу. Під типом елементу розуміється його специфічна характеристика: назва, абзац, рядок тощо. Іншим важливим поняттям є дефініція типу документу, або DTD (document type definition). Частина DTD подається у визначенні типу документу (document type declaration), яке містить сукупність визначень розмітки SGML для попереднього розпізнавання типу словника й синтаксису мови розмітки. Ці визначення типу документу зберігають відомості про тип елементів документу, їхні можливі комбінації, характеристики, абревіатури для даних тощо.

Однак семантика елементів не може бути виражена формально в SGML; для надання додаткової необхідної інформації використовується коментування або уточнення цієї інформації у документації мови розмітки. Важливо розрізняти поняття дефініції типу документу (document type definition) і визначення типу документу (document type declaration): дефініція характеризує мову розмітки; частина дефініції представлена в термінах, які використовуються під час розмітки; додаткова інформація міститься в коментарях або документації до мови розмітки. Паспорт документу становлять його дані й розмітка відповідно до дефініції типу документа та структури елементів.

Потреба у створенні мови метарозмітки для мережі Інтернет зумовила появу проекту XML (eXtensible Markup Language) у 1998 році. Для спрощення концепції метамови розробники відмовились від обов'язкової дефініції документу за наявності всіх необхідних тегів. Однак документ має відповідати ряду вимог, дотримання яких забезпечує чітке розпізнавання структури елементів документу.

На сьогоднішній день серед розробників найбільш поширені стандарти TEI (Text Encoding Initiative) [13]. Починаючи з 1987 року об'єднання трьох організацій – The Association for Computers in the Humanities (Асоціація з використання комп'ютерів у гуманітарних науках), the Association for Literary and Linguistic Computing (Асоціація з використання комп'ютерів у літературознавстві й лінгвістиці) і the Association for Computational Linguistics (Асоціація з прикладної лінгвістики) – працювало над формулюванням стандартів метарозмітки текстів. Перший варіант рекомендацій опубліковано в 1990 році, остання версія датується 2007 роком. Після певних змін проект TEI Guidelines був визнаний як найбільш використовуваний серед розробників систем опрацювання текстів. Зокрема, в останньому випуску TEI Guidelines [13] встановлено правила метарозмітки корпусів.

Оскільки анотування великого за обсягом корпусу є працемістким процесом, розробники пропонують класифікацію елементів корпусу за параметром доцільності анотування: обов'язкові до анотування, рекомендовані, анотовані за рішенням розробника й такі, що не підлягають анотуванню [13]. Розмітка корпусу здійснюється вручну, в автоматичному або напівавтоматичному режимах. Способи розмітки мають бути зазначені у відповідній документації до корпусу. Загалом, метарозмітка корпусу текстів проводиться аналогічно до метарозмітки документів за винятком відмінних елементів структури корпуса.

Більш детальні стандарти метарозмітки корпусів текстів представлено ініціативою EAGLES (European Advisory Group on Language Engineering Standards). Оскільки сучасні проекти з укладання корпусів орієнтовані на метатекстову розмітку, розробникам слід дотримуватися рекомендацій щодо класифікації текстів корпусу.

Досвід укладачів Британського національного корпусу (BNC) свідчить про доцільність використання класифікації Дж. Синклера–Шарова [12] як міжнародного варіанту метарозмітки. Класифікація передбачає виділення зовнішніх (позамовних) чинників, серед яких розрізняють три групи – походження тексту, зовнішні ознаки й мету написання тексту, та внутрішніх (змістовних і мовних) чинників – предметної області й стилістичних особливостей тексту. В результаті укладається паспорт з чотирма групами елементів метаопису тексту (бібліографічні дані, відомості про жанр, автора, інформація про структуру, особливості розмітки й модифікації тексту).

Стандартів укладання й розмітки корпусів за EAGLES дотримуються і розробники Американського національного корпуса (ANC). Метарозмітка корпусу здійснюється згідно із специфікою XML, а саме, з версією метарозмітки Corpus Encoding Standard (XCES)3 за рекомендаціями EAGLES. Належність XCES до програмного забезпечення XML гарантує доступ і можливість використання даних корпусу в мережі Інтернет [9, p. 4–5].

Стандарт, представлений EAGLES, є найбільш поширеним серед розробників корпусів. Проте набір елементів опису текстів не є достатньо повним для наукових текстів, оскільки він не дає змогу визначити їхню поняттєву структуру, створити когнітивну мапу текстів. Саме тому для укладачів корпусів фахових текстів доцільніше використовувати індивідуалізований стандарт метарозмітки, а саме, тематичну метарозмітку, яка сфокусована на тематичній належності тексту і пов'язаними з цим особливостями мови тексту.

Проект CES має на меті формування стандартів метарозмітки корпусів текстів. За словами Н. Іде, проект, який грунтується на засадах TEI, відрізняється більш конкретним об'єктом і філософією здійснення метарозмітки [8]. Серед його ключових переваг захист текстів корпусу від копіювання й одночасна доступність розмітки, можливість об'єднання її альтернативних варіантів. Здійснення метарозмітки корпусу передбачає три рівні розмітки: 1) базове анотування, відсутність іншомовної розмітки й попереднє зазначення всіх деталей анотування; 2) детальна синтаксична й морфологічна розмітка; 3) максимально повна синтаксична й морфологічна розмітка виключно за стандартом [8].

Прикладом реалізації стандартів метарозмітки в українській корпусній лінгвістиці є два корпуси текстів української мови [2; 3], тримовний ілюстративний Корпус текстів з комп'ютерної лінгвістики [6], англомовний Корпус анотацій наукових статей із комп'ютерної лінгвістики [4] і Багатомовний паралельний корпус усного мовлення, укладений на базі субтитрів до сучасних телевізійних серіалів [5]. Корпус текстів української мови побудований таким чином, щоб користувач міг одержати різноаспектну інформацію, що передбачає зовнішнє анотування за бібліографічними параметрами. Така метатекстова розмітка є міжнародною – стандарт EAGLE Дж. Синклера з додатками для слов'янських текстів С. Шарова [3, с. 47]. До метатекстової розмітки Корпусу текстів для вивчення граматичної службовості закладено параметри трьох форм мовлення – писемної, усної і мережевої; шести функціональних стилів з підстилями, п'яти часових періодів [2, с. 227–228].

Досвід метарозмітки англомовних текстів презентовано в українській корпусній лінгвістиці трьома проектами. Метатекстова розмітка англомовного Корпусу анотацій наукових статей із комп'ютерної лінгвістики включає інформацію про автора (ім'я, місце роботи, чи є англійська рідною мовою) і текст (тема, назва, обсяг, джерело, дата публікації) [4, с. 32-33]. Для створення ілюстративного англо-українсько-російського Корпусу текстів з комп'ютерної лінгвістики було розроблено допоміжну програму адресації, яка дозволяє опрацьовувати тексти за вихідними даними: назва тексту, прізвище та ініціали автора, назва видавництва й рік видання, анотація тексту [6, p. 36]. Така метарозмітка забезпечує базовий рівень опису фахових текстів, однак не є достатньо інформативною в аспекті дослідження тенденцій розбудови відповідної дисципліни. Метадані Багатомовного корпусу паралельних текстів включають загальну адресну інформацію: назву серіалу, номер сезону та епізоду, жанр серіалу, дату виходу серії на екрани [5, с. 37].

Отже, огляд наявних у вільному доступі українських корпусів показує, що більшість розробників корпусів текстів обирає власний варіант розмітки з урахуванням вимог міжнародного стандарту. Здійснена розвідка дає змогу зробити такі висновки: 1) потреба в швидкому опрацюванні великих за обсягом корпусів текстів

зумовила необхідність їх метарозмітки; 2) на початковому етапі проблеми метарозмітки корпусів текстів розглядалися в аспекті розмітки письмових текстів у цілому; 3) множинність стандартів метарозмітки вимагає створення єдиного або принаймні стандартизованого середовища для опрацювання корпусів текстів; 4) міжнародним стандартом метарозмітки корпусів текстів доцільно вважати той, що був розроблений і затверджений у результаті спільного проекту SGML і ISO.

Література:

1. Волков С. Св. Метаразметка в историческом корпусе XIX века / С. Св. Волков // Труды международной конференции «Корпусная лингвистика–2004». – Санкт-Петербург : Изд. Санкт-Петерб. ун-та, 2004. – С. 86–98.

2. Данилюк І. Корпус текстів для вивчення граматичної службовості / І. Данилюк // Лінгвістичні студії. – Вип. 26. – Донецьк : ДонНУ, 2013. – С. 224–229.

3. Дарчук Н. П. Дослідницький корпус української мови : основні засади і перспективи / Н. П. Дарчук // Вісник Київського національного університету ім. Тараса Шевченка. Серія : Літературознавство. Мовознавство. Фольклористика. – К. : ВПЦ «Київський університет», 2010. – № 21. – С. 45–49.

4. Коломієць 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.

5. 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.

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

7. DeRose S. J. What is text, really? / S. J. DeRose, D. Durand, E. Mylonas, H. A. Renear // Journal of Computer Documentation. – #21. 3. – 1990. – P. 1–24.

8. Ide N. Encoding linguistic corpora / N. Ide // Proceedings of the Sixth Workshop on Very Large Corpora. – Montreal, 1998. – P. 9–17.

9. Ide N. The American National Corpus : a standardized resource for American English / N. Ide // Proceedings of the Second International Conference on Language Resources and Evaluation, LREC, May 31–June 2, 2000. – Athens, 2000. – [Access Mode] : http://www.lrec-conf.org/proceedings/ lrec2000/ pdf/196.pdf

10. ISO Information Processing – Text and Office Systems – Standard Generalized Markup Language (SGML). ISO 8879. 1986(E). – Geneva, 1986. – [Access Mode] : http://www.w3.org/XML

11. Renear H. A. Text Encoding. / H. A. Renear // A Companion to Digital Humanities by Ed S. Schrelbman, R. Siemens, J. Unsworth. – Oxford : Blackwell, 2007. – P. 218–239.

12. Sinclair J. Preliminary recommendations on text typology / J. Sinclair // EAGLES Document EAG-TCWG-TTYP/P, 1996. – [Access Mode] : http:// www. ilc. cnr. it/EAGLES96/texttyp/texttyp. html

13. TEI Consortium, eds. TEI P5: Guidelines for Electronic Text Encoding and Interchange. TEI Consortium. – [Access Mode] : http://www.tei-c.org/Guidelines/P5/